Computational Needs for the Next Generation Electric Grid

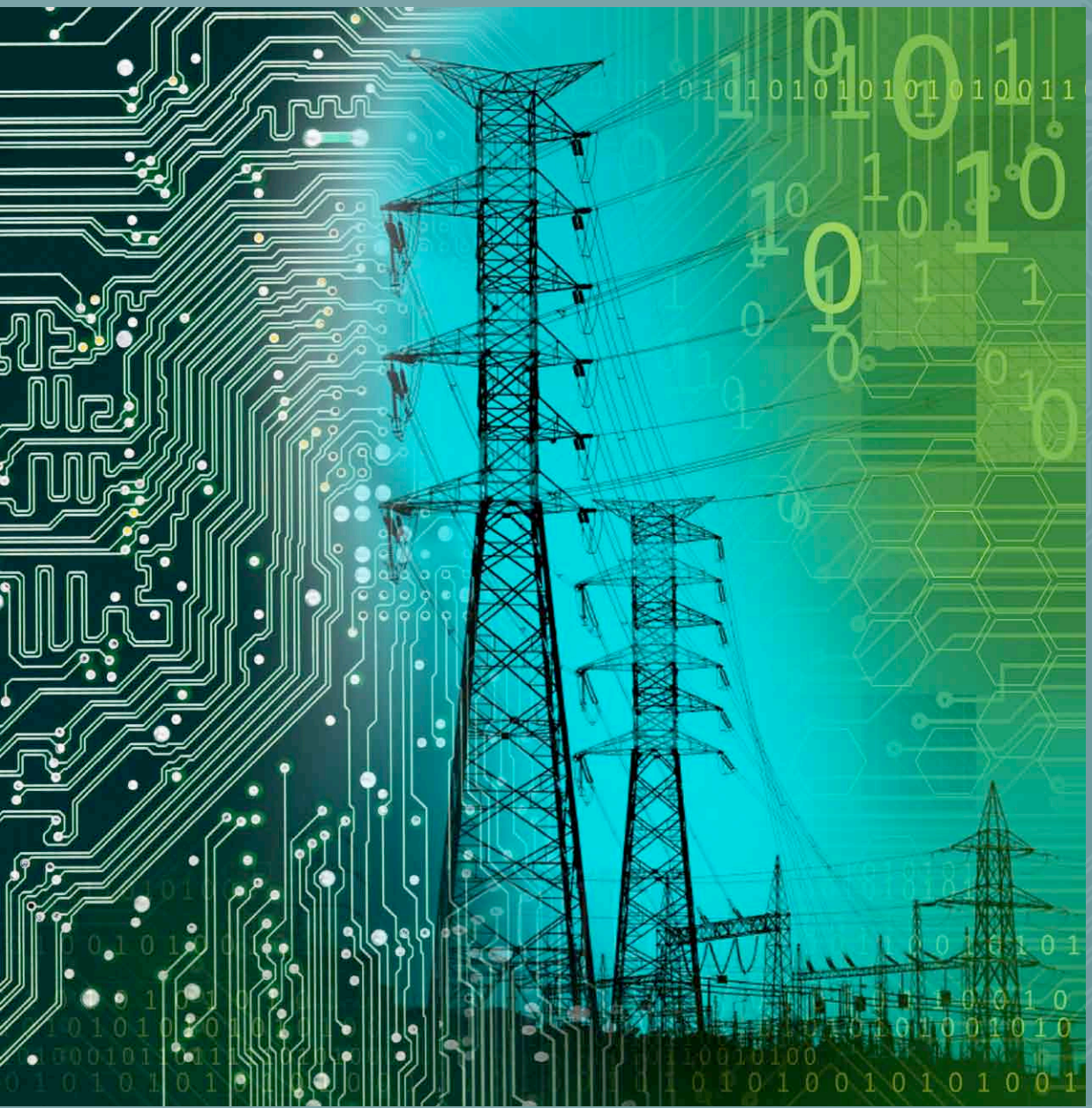
Proceedings

April 19-20, 2011

Editors:

Joseph H. Eto
Lawrence Berkeley National Laboratory

Robert J. Thomas
Cornell University



LBNL-5105E

Computational Needs for the Next Generation Electric Grid

Proceedings

April 19-20, 2011

Editors:
Joseph H. Eto, Lawrence Berkeley National Laboratory
Robert J. Thomas, Cornell University

The work described in this report was funded by the Office of Electricity Delivery and Energy Reliability of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Disclaimer

Table of Contents

Acknowledgements

The work described in this report was funded by the Office of Electricity Delivery and Energy Reliability of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The editors wish to thank the following individuals for their help in organizing the workshop and preparing the proceedings: Katherine Behrend; Sally Bird; Allan Chen; Nancy J. Lewis; Anthony Ma; Andrew Mills; Annika Todd.

Foreword

The April 2011 DOE workshop, "Computational Needs for the Next Generation Electric Grid", was the culmination of a year-long process to bring together some of the Nation's leading researchers and experts to identify computational challenges associated with the operation and planning of the electric power system. The attached papers provide a journey into these experts' insights, highlighting a class of mathematical and computational problems relevant for potential power systems research.

While each paper defines a specific problem area, there were several recurrent themes. First, the breadth and depth of power system data has expanded tremendously over the past decade. This provides the potential for new control approaches and operator tools that can enhance system efficiencies and improve reliability. However, the large volume of data poses its own challenges, and could benefit from application of advances in computer networking and architecture, as well as data base structures.

Second, the computational complexity of the underlying system problems is growing. Transmitting electricity from clean, domestic energy resources in remote regions to urban consumers, for example, requires broader, regional planning over multi-decade time horizons. Yet, it may also mean operational focus on local solutions and shorter timescales, as reactive power and system dynamics (including fast switching and controls) play an increasingly critical role in achieving stability and ultimately reliability. The expected growth in reliance on variable renewable sources of electricity generation places an exclamation point on both of these observations, and highlights the need for new focus in areas such as stochastic optimization to accommodate the increased uncertainty that is occurring in both planning and operations. Application of research advances in algorithms (especially related to optimization techniques and uncertainty quantification) could accelerate power system software tool performance, i.e. speed to solution, and enhance applicability for new and existing real-time operation and control approaches, as well as large-scale planning analysis.

Finally, models are becoming increasingly essential for improved decision-making across the electric system, from resource forecasting to adaptive real-time controls to on-line dynamics analysis. The importance of data is thus reinforced by their inescapable role in validating, high-fidelity models that lead to deeper system understanding. Traditional boundaries (reflecting geographic, institutional, and market differences) are becoming blurred, and thus, it is increasingly important to address these seams in model formulation and utilization to ensure accuracy in the results and achieve predictability necessary for reliable operations.

Each paper also embodies the philosophy that our energy challenges require interdisciplinary solutions – drawing on the latest developments in fields such as mathematics, computation, economics, as well as power systems. In this vein, the

workshop should be viewed not as the end product, but the beginning of what DOE seeks to establish as a vibrant, on-going dialogue among these various communities. Bridging communication gaps among these communities will yield opportunities for innovation and advancement.

The papers and workshop discussion provide the opportunity to learn from experts on the current state-of the-art on computational approaches for electric power systems, and where one may focus to accelerate progress. It has been extremely valuable to me as I better understand this space, and consider future programmatic activities. I am confident that you too will enjoy the discussion, and certainly learn from the many experts. I would like to thank the authors of the papers for sharing their perspectives, as well as the paper discussants, session recorders, and participants. The meeting would not have been as successful without your commitment and engagement. I also would like to thank Joe Eto and Bob Thomas for their vision and leadership in bringing together such a well-structured and productive forum.

Sincerely,

Gil Bindewald
Program Manager, Advanced Computation & Grid Modeling
Office of Electricity Delivery and Energy Reliability
United States Department of Energy

Introduction

Background

The US electric power system has undergone substantial change since the late 1980's and promises to continue to change into the foreseeable future. The changes began in 1989 with the restructuring of the way industry procured electric supply. The restructuring sought to replace centralized decision-making by the traditional vertically integrated utility with decentralized decision-making by market participants that establish prices through market forces, not regulation. Today, that transformation is still in progress. Vertically integrated firms continue to serve customers in regions of the country. A sustainable means for ensuring adequate transmission has not been demonstrated. And the demand side of the equation is not able to compete fully in an open market environment.

In the midst of this restructuring, advances in electric transportation, a greater awareness of the environmental effects of electricity production, and a desire on the part of the US to eliminate its dependence on foreign oil has prompted a movement to again re-invent the electric system. The new objectives include better accommodation of planning and operational uncertainty, especially that associated with variable renewable generation sources such as wind and solar, accommodation of major reductions in CO2 and other pollutants harmful to air quality, and economic and reliable operation of existing assets with less margin than in the past. It is now agreed that the "smart grid" that will be needed to achieve these objectives will involve the confluence of new sensing, communication, control, and computing as a unique blend of technologies that must be designed specifically to manage the requirements for a future electric power system based on competitive markets. Fundamental to this agreement is the idea that significant advances are needed in the areas of large-scale computation, modeling, and data handling.

Approach

To begin to address the large-scale computation, modeling, and data handling challenges of the future grid, seven survey papers were commissioned in 2010 from eminently qualified authorities conducting research activities in problem areas of interest. Each paper was to define a problem area, concisely review industry practice in this area up to the present time, and provide an objective, critical, and comparative assessment of research needed during the next 5 to 10 years. Authors were asked to identify seminal papers or reports that had motivated later, generic, related work.

Electric power system computational needs appropriate for discussion in the survey papers were to include:

1. New algorithms that are scalable and robust for solving large nonlinear mixed-integer optimization problems and methods for efficiently solving (in real-time) large sets of ordinary differential equations with algebraic constraints, including delays, parameter uncertainties, and monitored data as inputs. These new algorithms should accommodate randomness for capturing appropriate notions of security and incorporate recent results on improving deterministic and randomized algorithms for computationally hard problems.
2. A new mathematics for characterizing uncertainty in information created from large volumes of data as well as for characterizing uncertainty in models used for prediction.
3. New methods to enable efficient use of high-bandwidth networks by dynamically identifying only the data relevant to the current information need and discarding the rest. This would be especially useful for wide-area dynamic control where data volume and latency are barriers.
4. New software architectures and new rapid development tools for merging legacy and new code without disrupting operation. Software should be open source, modular, and transparent. Security is a high priority.

We assume that designing and building larger and faster computers and faster communications will not be sufficient to solve the electric grid computational problems, although these improvements might ultimately be helpful. Instead, our expectation is that fundamental advances are needed in the areas of algorithms, computer networking and architecture, databases and data overwhelm, simulation and modeling, and computational security; perhaps most importantly, these advances must be achievable in a time frame that will be useful to the industry.

On April 19-20, 2011, a two-day workshop was held on the campus of Cornell University to explore critical computational needs for future electric power systems. Workshop participants provided input based on the presentation of the seven papers. The seven papers were not expected to be exhaustive, but acted as a framework within which to explore a rich range of topics associated with the overall issue.

The collection of materials in this volume is intended to provide as complete a record as possible of the workshop proceedings. The volume contains the final versions of the seven papers that were presented, along with discussions of the papers' focus that were prepared ahead of time to stimulate discussion at the workshop, and the reports of the discussions that took place among workshop participants. The authors of the seven papers reviewed and approved the reports of the workshop discussions, which include the reporters' interpretations.

Summary of the Papers

In developing the workshop our focus has been on a class of problems that have been neglected but will have to be solved if we are to move in a timely way to a new smart architecture capable of accommodating our vision for the grid of the future. We summarize below the contributions each of the seven papers makes to the discussion of this class of problems.

The first paper by Kenneth P. Birman, Lakshmi Ganesh, and Robbert van Renesse explores the relatively new paradigm of cloud computing in relation to future electric power system needs. The authors note that future needs will demand scalability of a kind that only cloud computing can offer. Their thesis is that there will be power system requirements (real-time, consistency, privacy, security, etc.) that cloud computing cannot currently support and that many of these needs, which are specific to the expected future electric power paradigm, will not soon be filled by the cloud industry.

The second paper by Michael Ferris is about modeling. This paper argues that decision processes are predominantly hierarchical and that, as a result, models to support such decision processes should also be layered or hierarchical. Ferris contends that, although advice can be provided from the perspective of mathematical optimization on managing complex systems, that advice must be integrated into an interactive debate with informed decision makers. He also agrees that treating uncertainties in large scale planning projects will become even more critical as the smart grid evolves because of the increase in volatility of both supply and demand. Optimization models with flexible systems design can help address these uncertainties not only during the planning and construction phases, but also during the operational phase of an installed system.

The third paper by Andreas G. Hofmann and Brian C. Williams focuses on the twin problems of: 1) increasing the level of automation in the analysis and planning for contingencies in response to unexpected events, and 2) the problem of incorporating considerations of optimality into contingency planning and the overall energy management process. With regard to the first problem, the authors note that, although the level of anticipated automation is still advisory and humans remain in the loop, use of automation would reduce the drudgery and error prone nature of the current labor-intensive approach. Automation would also guarantee the completeness of an analysis and validity of the contingency plans. With regard to the second problem, the optimization would include establishing risk bounds on actions taken to achieve optimal performance.

The fourth paper by Janos Sztipanovits, Graham Hemingway, Anjan Bose, and Anurag Srivastava is also about modeling. The thesis is that the future electric system will require "the efficient integration of digital information, communication systems, real time-data delivery, embedded software and real-time control decision-making." The authors posit that no high-fidelity models are capable of simulating electric grid interactions with communication and control infrastructure elements for large systems.

They also conclude that it is a challenge to model infrastructure interdependencies related to the power grid, including the networks and software for sensors, controls, and communication.

The fifth paper by Santiago Grijalva argues that future electric system problems are multi-scale and that there is a need to develop multi-scale simulation models and methods to the level that exists in other engineering disciplines. The paper discusses 18 areas of multi-scale, multi-dimensional power system research that are needed to provide a framework for addressing emerging power system problems.

The sixth paper by Sarah M. Ryan, James D. McCalley, and David L. Woodruff describes computational tools that are needed in the area of optimization for large-scale planning models that account for uncertainty. The authors present, as an example, a proposed model for electric system planning that includes linkages with transportation systems. The paper addresses multi-objective planning in the presence of uncertainty where decision makers must balance, for example, sustainability, costs (investment and operational), long-term system resiliency, and solution robustness.

The seventh and final paper, by Jinjun Xiong and his associates, explores computational challenges in the context of security-constrained unit commitment and economic dispatch with stochastic analysis, management of massive data sets, and concepts related to large-scale grid simulation. Although other papers address simulation and optimization, this paper is unique in its exploration of emerging substantive data management issues.

Conclusions

The April 2011 workshop touched on important research and development needs for the future electric power system, but was not exhaustive. We hope that it has created the basis for the formation of a community of researchers who will focus on these very substantial and interesting needs.

We wish to thank the authors of the papers for their outstanding contributions and for providing important food for thought. In addition to the authors of the papers, we wish to acknowledge the contributions of the Paper Discussants: James Nutaro, Ali Pinar, Bernard Lesieutre, Henry Huang, Roman Samulyak, Jason Stamp, and Loren Toole; and the Session Recorders: Ghaleb Abdulla, Alejandro Dominguez-Garcia, Hyung-Seon Oh, Victor Zavala, Sven Leyffer Chao Yang, and Jeff Dagle. Finally, we are grateful for the active contributions of the industry participants and invited guests. Everyone's participation led to a very successful enterprise.

Joseph H. Eto
Robert J. Thomas

White Paper

Running Smart Grid Control Software on Cloud Computing Architectures

Kenneth P. Birman, Lakshmi Ganesh, and Robbert van Renesse
Cornell University

Abstract

There are pressing economic as well as environmental arguments for the overhaul of the current outdated power grid, and its replacement with a Smart Grid that integrates new kinds of green power generating systems, monitors power use, and adapts consumption to match power costs and system load. This paper identifies some of the computing needs for building this smart grid, and examines the current computing infrastructure to see whether it can address these needs. Under the assumption that the power community is not in a position to develop its own Internet or create its own computing platforms from scratch, and hence must work with generally accepted standards and commercially successful hardware and software platforms, we then ask to what extent these existing options can be used to address the requirements of the smart grid. Our conclusions should come as a wakeup call: many promising power management ideas demand scalability of a kind that only cloud computing can offer, but also have additional requirements (real-time, consistency, privacy, security, etc.) that cloud computing would not currently support. Some of these gaps will not soon be filled by the cloud industry, for reasons stemming from underlying economic drivers that have shaped the industry and will continue to do so. On the other hand, we don't see this as a looming catastrophe: a focused federal research program could create the needed scalability solutions and then work with the cloud computing industry to transition the needed technologies into standard cloud settings. We'll argue that once these steps are taken, the solutions should be sufficiently monetized to endure as long-term options because they are also of high likely value in other settings such as cloud-based health-care, financial systems, and for other critical computing infrastructure purposes.

1 Introduction: The Evolving Power Grid

The evolution of the power grid has been compared, unfavorably, with the evolution of modern telephony; while Edison, one of the architects of the former, would recognize most components of the current grid, Bell, the inventor of the latter, would find telephony unrecognizably advanced since his time [40]. It is not surprising, then, that the power grid is under immense pressure today from inability to scale to current demands, and is growing increasingly fragile, even as the repercussions of power outages grow ever more serious. Upgrading to a smarter grid has escalated from being a desirable vision, to an urgent imperative. Clearly, the computing industry will have a key role to play in enabling the smart grid, and our goal in this paper is to evaluate its readiness, in its current state, for supporting this vision.

EXECUTIVE SUMMARY

HIGH ASSURANCE CLOUD COMPUTING REQUIREMENTS OF THE FUTURE SMART GRID

- **Support for scalable real-time services.** A *real-time service* will meet its timing requirements even if some limited number of node (server) failures occurs. Today's cloud systems do support services that require rapid responses, but their response time can be disrupted by transient Internet congestion events, or even a single server failure.
- **Support for scalable, consistency guaranteed, fault-tolerant services.** The term *consistency* covers a range of cloud-hosted services that support database ACID guarantees, state machine replication behavior, virtual synchrony, or other strong, formally specified consistency models, up to some limited number of server failures. At the extreme of this spectrum one finds Byzantine Fault Tolerance services, which can even tolerate compromise (e.g. by a virus) of some service members. Today's cloud computing systems often "embrace inconsistency"[31][37], making it hard to implement a scalable consistency-preserving service.
- **Protection of Private Data.** Current cloud platforms do such a poor job of protecting private data that most cloud companies must remind their employees to "not be evil". Needed are protective mechanisms strong enough so that cloud systems could be entrusted with sensitive data, even when competing power producers or consumers share a single cloud data center.
- **Highly Assured Internet Routing.** In today's Internet, consumers often experience brief periods of loss of connectivity. However, research is underway on mechanisms for providing secured multipath Internet routes from points of access to cloud services. Duplicated, highly available routes will enable critical components of the future smart grid to maintain connectivity with the cloud-hosted services on which they depend.

Figure 1: Summary of findings. A more technical list of specific research topics appears in Figure 6.

We shall start with a brief review to establish common terminology and background. For our purposes here, the power grid can be understood in terms of three periods [34],[10]. The "early" grid arose as the industry neared the end of an extended period of monopoly control. Power systems were owned and operated by autonomous, vertically-integrated, regional entities that generated power, bought and sold power to neighboring regions, and implemented proprietary Supervisory Control And Data Acquisition (SCADA) systems. These systems mix hardware and software. The hardware components collect status data (line frequency, phase angle, voltage, state of fault-isolation relays, etc.), transmit this information to programs that clean the input of any bad data, and then perform state estimation. Having computed the optimal system configuration, the SCADA platform determines a control policy for the managed region, and then sends instructions to actuators such as generator control systems, transmission lines with adjustable capacity and other devices to increase or decrease power generation, increase or decrease power sharing with neighboring regions, shed loads, etc. The SCADA system also plays key roles in preventing grid collapse by shedding busses if regional security[1] requires such an action.

The "restructuring" period began in the 1990's and was triggered by a wave of regulatory reforms aimed at increasing competitiveness [19]. Regional monopolies fragmented into power generating companies, Independent System Operators (ISOs) responsible for long-distance power transmission and grid safety, and exchanges in which power could be bought and sold somewhat in the manner of other commodities (although the details of power auctions are specific to the industry, and the difficulty of storing power also distances power markets from other kinds of commodity markets). Small power producers entered the market, increasing competitive pressures in some regions. Greater inter-regional connectivity emerged as transmission lines were built to facilitate transfer of power from areas with less expensive power, or excess generating capacity into regions with more costly power, or less capacity.

One side effect of deregulation was to create new economic pressures to optimize the grid, matching line capacity to the pattern of use. Margins of excess power generating capacity, and excess transmission capacity, narrowed significantly, hence the restructured grid operates much nearer its security limits. SCADA systems play key roles, performing adjustments in real-time that are vital for grid security. The cost of these systems can be substantial; even modest SCADA product deployments often represent investments of tens or hundreds of millions of dollars, and because federal regulatory policies require full redundancy, most such systems are fully replicated at two locations, so that no single fault can result in a loss of control.

[1] *Security* here is to mean the safety and stability of the power grid, rather than protection against malice.

This review was prepared during the very first years of a new era in power production and delivery: the dawn of the "smart" power grid. Inefficient power generation, unbalanced consumption patterns that lead to underutilization of expensive infrastructure on the one hand, and severe overload on the other, as well as urgent issues of national and global concern such as power system security and climate change are all driving this evolution [40]. As the smart grid concept matures, we'll see dramatic growth in green power production: small production devices such as wind turbines and solar panels or solar farms, which have fluctuating capacity outside of the control of grid operators. Small companies that specialize in producing power under just certain conditions (price regimes, certain times of the day, etc.) will become more and more common. Power consumers are becoming more sophisticated about pricing, shifting consumption from peak periods to off-peak periods; viewed at a global scale, this represents a potentially non-linear feedback behavior. Electric vehicles are likely to become important over the coming decade, at least in dense urban settings, and could shift a substantial new load into the grid, even as they decrease the national demand for petroleum products. The operation of the grid itself will continue to grow in complexity, because the effect of these changing modalities of generation and consumption will be to further fragment the grid into smaller regions, but also to expand the higher level grid of long-distance transmission lines. Clearly, a lot of work is required to transition from the 50-year-old legacy grid of today to the smart grid of the future. Our purpose in this paper is to see how far the computing industry is ready to meet the needs of this transition.

2 The Computational Needs of the Smart Grid

We present a few representative examples that show how large-scale computing must play a key role in the smart power grid. In the next sections, we shall see whether current computing platforms are well suited to play this role.

i. **The smart home.** In this vision, the home of the future might be equipped with a variety of power use meters and monitoring devices, adapting behavior to match cost of power, load on the grid, and activities of the residents. For example, a hot-water heater might heat when power is cheap but allow water to cool when hot water is unlikely to be needed. A washing machine might turn itself on when the cost of power drops sufficiently. Air conditioning might time itself to match use patterns, power costs, and overall grid state. Over time, one might imagine ways that a SCADA system could reach directly into the home, for example to coordinate air conditioning or water heating cycles so that instead of being random and uniform, they occur at times and in patterns convenient to the utility.

ii. **Ultra-responsive SCADA for improved grid efficiency and security**. In this area, the focus is on improving the security margins for existing regional control systems (which, as noted earlier, are running with slim margins today), and on developing new SCADA paradigms for incorporating micro-power generation

into the overall grid. One difficult issue is that the power produced by a wind farm might not be consumed right next to that farm, yet we lack grid control paradigms capable of dealing with the fluctuating production and relatively unpredictable behavior of large numbers of small power generating systems. One recent study [2] suggested that to support such uses, it would be necessary to create a new kind of grid-stat system, tracking status at a fine-grained level. Such approaches are likely to have big benefits, hence future SCADA systems may need to deal with orders of magnitude more information than current SCADA approaches handle.

iii. **Wide area grid state estimation.** Blackouts such as the NorthEast and Swiss/Italian blackouts (both in 2003), originated with minor environmental events (line trips caused by downed trees), but that snowballed through SCADA system confusions that in turn caused operator errors (see "Northeast_Blackout_of_2003" and "2003_Italy_blackout" in Wikipedia). Appealing though it may be to blame the humans, those operator errors may have been difficult to avoid. They reflected the inability of regional operators to directly observe the state of the broader power grids to which their regions are linked; lacking that ability, a hodgepodge of guesswork and telephone calls are often the only way to figure out what a neighboring power region is experiencing. Moreover, the ability to put a telephone call through during a spreading crisis that involves loss of power over huge areas is clearly not something one can necessarily count upon in any future system design. As the power grid continues to fracture into smaller and smaller entities, this wide area control problem will grow in importance, with ISOs and other operators needing to continuously track the evolution of the state of the grid and, especially important, to sense abnormal events such as bus trips or equipment failures. Data about power contracts might inform decisions, hence the grid state really includes not just the data captured from sensors but also the intent represented in the collection of power production and consumption contracts.

What are the computational needs implied by these kinds of examples?

i. **Decentralization**. Information currently captured and consumed in a single regional power system will increasingly need to be visible to neighboring power systems and perhaps even visible on a national scale. An interesting discussion of this topic appears in [2].

ii. **Scalability**. Every smart grid concept we've reviewed brings huge numbers of new controllable entities to the table. In some ideas, every consumer's home or office becomes an independent point for potential SCADA control. In others, the homes and offices behave autonomously but still must tap into dynamic data generated by the power provider, such as pricing or load predictions. Other ideas integrate enormous numbers of small power producing entities into the grid and require non-trivial control adjustments to keep the grid stable. Thus scalability will be a key requirement – scalability of a kind that dwarfs what the

industry has done up to now, and demands a shift to new computational approaches [25][26][2][40].

iii. **Time criticality**. Some kinds of information need to be fresh. For example, studies have shown that correct SCADA systems can malfunction when presented with stale data, and some studies have even shown that SCADA systems operated over Internet standards like the ubiquitous TCP/IP protocols can malfunction [25][26][2][12], because out-of-the-box TCP delays data for purposes of flow control and to correct data loss. Future smart-grid solutions will demand real-time response even in the presence of failures.

iv. **Consistency**. Some kinds of information will need to be consistent [5][6][7][8][25][19], in the sense that if multiple devices are communicating with a SCADA system at the same time, they should be receiving the same instructions, even if they happen to connect to the SCADA system over different network paths that lead to different servers that provide the control information. Notice that we're not saying that control data must be computed in some sort of radically new, decentralized manner: the SCADA computation itself could be localized, just as today's cloud systems often start with one copy of a video of an important news event. But the key to scalability is to replicate data and computation, and consistency issues arise when a client platform requests data from a service replica: is this really the most current version of the control policy? Further, notice that consistency and real-time guarantees are in some ways at odds. If we want to provide identical data to some set of clients, failures may cause delays: we lose real-time guarantees of minimal delay. If we want minimal delay, we run the risk that a lost packet or a sudden crash could leave some clients without the most recent data.

v. **Data Security**. Several kinds of data mentioned above might be of interest to criminals, terrorists, or entities seeking an edge in the power commodities market. Adequate protection will be a critical requirement of future SCADA systems.

vi. **Reliability**. Power systems that lose their control layer, even briefly, are at grave risk of damage or complete meltdown. Thus any SCADA solution for the future smart grid needs to have high reliability.

vii. **Ability to tolerate compromise**. The most critical subsystems and services may need to operate even while under attack by intruders, viruses, or when some servers are malfunctioning. The technical term for this form of extreme reliability is Byzantine Fault Tolerance; the area is a rich one and many solutions are known, but deployments are rare and little is known about their scalability.

3 The Evolving Computing Industry: An Economic Story

We shall now describe the current state of the computing industry, and examine its ability to provide the properties described above for the future smart grid. We begin by giving a brief history of the computing industry and the economic drivers of its

evolution. These same drivers are likely to determine whether the power community can use current computing platforms for its needs, or not.

Prior to the late 1990's, the computing industry was a world of client computers that received data and instructions from servers. Client-server computing represented a relatively wrenching transition from an even earlier model (mainframe computing), and the necessary architecture and tools were slow to mature; in some sense, the excitement associated with the area anticipated the actual quality of the technology by five to ten years. Yet the client-server architecture slowly gained acceptance and became the basis of widely adopted standards, until finally, within the last decade or so, software tools for creating these kinds of applications have made it possible for a typical programmer to create and deploy such applications with relative ease.

Today, client-server computing is the norm, yet the power industry retains legacies from the mainframe computing era. For example, SCADA systems use high performance computing (HPC) techniques but play roles similar to SCADA solutions in older mainframe architectures, which featured a big computer in the middle of a slaved network of sensors and actuators. This is in contrast to cloud architectures, which take the client-server model and push it even further: the client is now supported by multiple data centers, each of which might be composed of a vast number of relatively simple servers, with second and even third tiers of support layered behind them. But the issues are also social: power is a critical infrastructure sector – one that affects nearly every other sector – and understandably, the power community is traditionally risk-averse and slow in adopting new technology trends.

The computing industry has seen three recent technical epochs, each succeeding the prior one in as little as five years. Looking first at the period up to around the centennial, we saw a game-changing transition as the early Internet emerged, blossomed, briefly crashed (the .com boom and bust), and then dramatically expanded again. That first boom and bust cycle could be called the early Internet and was dominated by the emergence of web browsers and by human-oriented Internet enterprises. The Internet architecture became universal during this period. Prior to the period in question, we had a number of networking technologies, with some specialized ones used in settings such as wireless networks, or in support of communications overlaid on power transmission lines. Many power companies still use those old, specialized, communication technologies. But today, the Internet architecture has become standard. This standardization is useful. For example modern power companies visualize the status of sensors and actuators through small web pages that provide quick access to parameter settings and controls. Software on those devices can be quickly and easily patched by upgrading to new versions over the network. But these same capabilities have also created the potential for unintended connectivity to the Internet as a whole. Attackers can exploit these opportunities: we saw this in the widely publicized "Eligible Receiver" exercises, in which the government demonstrated that a technically

savvy but non-expert team could use publicly available information to take control of power systems and inflict serious damage on transformers, generators, and other critical equipment [39].

We now arrive at a period covering roughly the past five years, which witnessed a breathtaking advance in the penetration and adoption of web technologies. Standardization around web protocols and the ease of adding web interfaces even to older mainframe or client-server applications meant that pretty much any computing entity could access any other computing entity, be it hardware or software. Outsourcing boomed as companies in India, China, and elsewhere competed to offer inexpensive software development services. Penetration of the Internet into the public and private sector triggered explosive revenue growth in all forms of Internet advertising. New computing platforms (mobile phones, tablet computers) began to displace traditional ones, triggering a further boom associated with mobility and "app" computing models. Rarely have so many changes been compressed into so short a period of time.

Perhaps most unsettling of all, completely new companies like Facebook and Google displaced well established ones like IBM, HP, and Microsoft, seemingly overnight. One might reasonably argue that the power industry should be immune to this sort of turmoil, yet the impact of restructuring has caused an equal shakeup on the business side of the power community, even if the technical side remains less impacted. And there is good reason to believe that this will soon change. For example, the team that created Google is prominent among industry leaders promoting a smarter power grid. It is hard to imagine them being content to do things in the usual ways.

Cloud computing, our primary focus in this paper, is an overarching term covering the technologies that support the most recent five-years or so of the Internet, with different specific meanings for different cloud operators. The term means different things to different cloud owner/operators, but some form of cloud computing can be expected in any future Internet. A recent document laying out a Federal Cloud Computing Strategy, drafted by the CIO of the United States government (Dr. Vivek Kundra) recently called for spending about $20 billion of the $80 billion federal IT budget on cloud computing initiatives [28] and urged all government agencies to develop Cloud-based computing strategies. About a third of the cost would come from reductions in infrastructure cost through data center consolidation.

The perspective that sheds the most light on the form that cloud computing takes today starts by recognizing that cloud computing is an intelligent response to a highly monetized demand, shaped by the economics of the sectors from which that demand emanated [8]. These systems guarantee the properties needed to make money in these sectors; properties not required (or useful only in less economically important applications) tend not to be.

What are these requirements? Perhaps the most important emerges from the pressure to aggregate data in physically concentrated places. The rise of lightweight, mobile devices, and of clients who routinely interact with multiple devices, shifts the emphasis from personal computing (email on the user's own machine, pictures in my private folder, etc.) towards data center hosting models, for example Hotmail, Gmail, Flickr, and YouTube. Social networking sites gained in popularity, for example Facebook, YouTube, Flickr, and Twitter; they revolve around sharing information: my data and your data need to be in the same "place" if we're to share and to network in a sharing-driven manner. Moreover, because cloud platforms make money by performing search and placing advertising, cloud providers routinely need to index these vast collections of data, creating pre-computed tables that are used to rapidly personalize responses to queries.

Thus, cloud computing systems have exceptional capabilities for moving data from the Internet into the cloud (web crawlers), indexing and searching that data (MapReduce [16], Chord [3], Dynamo [17], etc.), managing files that might contain petabytes of information (BigTable [13], the Google File System [20], Astrolabe[35]), coordinating actions (Chubby [12], Zookeeper [26], DryadLINQ [38]), and implementing cloud-scale databases (PNUTS [15]). These are just a few of many examples.

Massive data sets are just one respect in which cloud systems are specialized in response to the economics of the field. Massive data centers are expensive, and this creates a powerful incentive to drive the costs down and to keep the data center as busy and efficient as possible. Accordingly, cost factors such as management, power use, and other considerations have received enormous attention [21]. Incentives can cut both ways: social networking sites are popular, hence cloud computing tools for sharing are highly evolved; privacy is less popular, hence little is known about protecting data once we move it into the cloud [29].

It should not be surprising that cloud computing has been shaped by the "hidden hand of the market," but it is important to reflect on the implications of this observation. The specific attributes of the modern data center and its cloud computing tools are matched closely to the ways that companies like Amazon, Microsoft, Google and Facebook use them: those kinds of companies invested literally hundreds of billions of dollars to enable the capabilities with which they earn revenue. Cloud computing emerged overnight, but not painlessly, and the capabilities we have today reflect the urgent needs of the companies operating the cloud platforms.

How then will we deal with situations in which the power grid community needs a cloud capability lacking in today's platforms? Our market-based perspective argues for three possible answers. If there is a clear reason that the capability is or will soon be central to an economically important cloud computing application, a watch and wait approach would suffice. Sooner or later, the train would come down the track. If a

capability somehow would be costly to own and operate, even if it were to exist, it might rapidly be abandoned and actively rejected by the community. We'll see that there is an instance of this nature associated with consistency. Here, only by finding a more effective way to support the property could one hope to see it adopted in cloud settings (hence, using the same economic metrics the community uses to make its own go/no-go decisions). Finally, there are capabilities that the commercial cloud community would find valuable, but hasn't needed so urgently as to incentivize the community to actually create the needed technology. In such cases, solving the problem in a useful prototype form might suffice to see it become part of the standards.

4 The Case for Hosting the Smart Grid on Cloud Computing Infrastructures

Cloud computing is of interest to the power community for several business reasons. Some parallel the green energy considerations that have stimulated such dramatic change in the power industry: cloud computing is a remarkably efficient and green way to achieve its capabilities. Others reflect pricing: cloud computing turns out to be quite inexpensive in dollar terms, relative to older models of computing. And still others are stories of robustness: by geographically replicating services, companies like Google and Microsoft are achieving fraction of a second responsiveness for clients worldwide, even when failures or regional power outages occur. Cloud systems can be managed cheaply and in highly automated ways, and protected against attack more easily than traditional systems [31]. Finally, cloud computing offers astonishing capacity and elasticity: a modern cloud computing system is often hosted on a few data centers any one of which might have more computing and storage and networking capacity than all of the world's supercomputing centers added together, and can often turn on a dime, redeploying services to accommodate instantaneous load shifts. We shall enumerate some of the issues in the debate about using the cloud for building the smart grid.

4.1 The Cloud Computing Scalability Advantage

The cloud and its transformation of the computing industry have resulted in the displacement of previous key industry players like Intel, IBM, and Microsoft by new players like Google, Facebook, and Amazon. Technology these new-age companies created is becoming irreversibly dominant for any form of computing involving scalability: a term that can mean direct contact with large numbers of sensors, actuators or customers, but can also refer to the ability of a technical solution to run on large numbers of lightweight, inexpensive servers within a data center. Earlier generations of approaches were often abandoned precisely because they scaled poorly. And this has critical implications for the smart grid community, because it implies that to the extent that we launch a smart grid development effort in the near term, and to the extent that the grid includes components that will be operated at large scale, those elements will be

built on the same platforms that are supporting the Facebooks and Amazons of today's computing world. In Figure 2 and Figure 3, we look at the scalability needs of two scenarios representative of the future smart grid.

4.2 The Cloud Cost Advantage

The Smart Grid needs a national-scale, pervasive network that connects every electricity producer in the market, from coal and nuclear plants to hydroelectric, solar, and wind farms, and small independent producers, with every electricity consumer, from industrial manufacturing plants to residences, and to every device plugged into the wall. This network should enable the interconnected devices to exchange status information and control power generation and consumption. The scale of such an undertaking is mind boggling. Yet, the key enabler, in the form of the network itself, already exists. Indeed, the Internet already allows household refrigerators to communicate with supermarkets and transact purchases [30]. It won't be difficult to build applications ("apps") that inform the washing machine of the right time to run its load, based on power pricing information from the appropriate generators. Whatever their weaknesses, the public Internet and cloud offer such a strong cost advantage that the power community cannot realistically ignore them in favor of building a private, dedicated network for the smart grid.

4.3 Migrating High Performance Computing (HPC) to the Cloud

We noted that SCADA systems are instances of "high performance computing" applications. It therefore makes sense to ask how the cloud will impact HPC. Prior to the 1990s, HPC revolved around special computing hardware with unique processing capabilities. These devices were simply too expensive, and around 1990 gave way to massive parallelism. The shift represented a big step backward for some kinds of users, because these new systems were inferior to the ones they replaced for some kinds of computation. Yet like it or not, the economics of the marketplace tore down the old model and installed the new one, and HPC users were forced to migrate. Today, even parallel HPC systems face a similar situation. A single cloud computing data center might have storage and computing capabilities tens or hundreds of times greater than all of the world's supercomputing facilities combined. Naturally, this incentivizes the HPC community to look to the cloud. Moreover, to the extent that HPC applications do migrate into the cloud, the community willing to pay to use dedicated HPC (non-cloud HPC) shrinks. This leaves a smaller market and, over time, represents a counter-incentive for industry investment in faster HPC systems. The trend is far from clear today, but one can reasonably ask whether someday, HPC as we currently know it (on fast parallel computers) will vanish in favor of some new HPC model more closely matched to the properties of cloud computing data centers.

Scenario one: National Scale Phasor Data Collection

A *phasor* is a complex number representing the magnitude and phase angle of a wave. Phasors are measured at different locations at a synchronized time (within one microsecond of one another). The required accuracy can be obtained from GPS. For 60 Hz systems, each Phasor Measurement Unit (PMU) takes about 10 to 30 such measurements per second. The data from various (up to about 60) PMUs is collected by a Phasor Data Concentrator (PDC) (transmitted over phone lines), and then forwarded along a Wide Area Measurement System (WAMS) to a SCADA system. The SCADA system must receive the data within 2 to 10 seconds.

It has been suggested that as the future power grid becomes increasingly interconnected to promote sharing so as to reduce wasted power and smooth the regional impact of erratic wind and solar power generation, we will also expose the grid to rolling outages. A possible remedy is for the regional operators to track the national grid by collecting phasor data locally and sharing it globally. We now suggest that the scale of the resulting problem is similar to the scale of computational challenges that motivated web search engines to move to the modern cloud computing model.

Simple back-of-the-envelope-calculations lead to a cloud computing model: Today's largest PMU deployment has about 120 PMUs, but for the purposes outlined here, one could imagine a deployment consisting of at least 10,000 PMUs. If we have 25 PMUs per PDC, then such a system would require 400 PDCs. Each PDC would deliver 30 measurements per second. If a measurement is 256 bytes in size (including magnitude, phase angle, timestamp, origin information, and perhaps a digital signature to protect against tampering or other forms of data corruption), then each PDC would deliver 25 x 256 x 30 = 192 KBytes/sec. The 400 PDCs combined would contribute about 77 Mbytes/sec, or about 615 Mbits/sec. The data would probably have to be shared on a national scale with perhaps 25 regional SCADA systems, located throughout the country, hence the aggregate data transmission volume would be approximately 15 Gbit/sec, more than the full capacity of a state of the art optical network link today[2].

While it would be feasible to build a semi-dedicated national-scale phasor-data Internet for this purpose, operated solely for and by the power community, we posit that sharing the existing infrastructure would be so much cheaper that it is nearly inevitable that the power community will follow that path. Doing so leverages the huge investment underway in cloud computing systems to distribute movies and Internet video; indeed, the data rates are actually "comparable" (a single streamed HD DVD is about 40 Mbits/second). But it also forces us to ask what the implications of monitoring and controlling the power grid "over" the Internet might be; these questions are at the core of our study (we pose, but don't actually answer them).

Figure 2: Tracking Phasor Data on a National Scale

[2] The 10Gbit rate quoted is near the physical limits for a single optical network link operated over long distances (as determined by the Shannon coding theory). But it is important to keep in mind that Internet providers, having invested in optical networking capacity, can often run multiple side-by-side optical links on the same physical path. Thus, the core Internet backbone runs at 40Gbits, and this is achieved using 4 side-by-side 10Gbit optical links. Moreover, network providers often set aside dedicated bandwidth under business arrangements with particular enterprises: Google or MSN, for example, or Netflix. Thus even if the future power grid runs "over" the Internet, this does not imply that grid control traffic could be disrupted or squeezed out by other kinds of public traffic.

Scenario Two: Power Aware Appliances in a Smart Home

According to the most recent US government census report, the United States had approximately 115 million households in 2010. Appliance ownership is widely but variably estimated. Reports on the web suggest that more than 95% of all households have major kitchen equipment such as a refrigerator and range, that 40 to 60% own a dishwasher, between 60 and 95% have a dedicated washer and dryer, and that as many as 80% or more have their own hot water heaters (the quality of these statistics may be erratic). These homes are heated, air conditioned, artificially lighted, and contain many powered devices (TVs, radios, etc.). Some will soon own electric vehicles.

Such numbers make clear the tremendous opportunity for smart energy management in the home. Current industry trends suggest the following mode: the consumer will probably gravitate towards mobile phone "apps" that provide access to home energy management software, simply because this model has recently gained so much commercial traction through wide adoption of devices such as the iPhone, BlackBerry, and Android phones, all of which adopt this particular model; apps are easy to build, easy to market, have remarkable market penetration, and are familiar to the end user. As they evolve, power-aware apps will coordinate action to operate appliances in intelligent ways that reduce end-user costs but also smooth out power demands, reduce load when the grid comes under stress, etc.

Thus, one might imagine a homeowner who loads the dishwasher but doesn't mind it running later, needs hot water early in the morning (or perhaps in the evening; the pattern will vary but could be learned on a per-household basis), etc. Ideally, the local power grid would wish to "schedule" these tasks in a price-aware, capacity-aware, energy efficient manner.

In one popular vision the grid simply publishes varying prices, which devices track. But this approach is poorly controlled: it is hard to know how many households will be responsive to price variability, and while one could imagine a poorly subscribed service failing for lack of popularity, one can also imagine the other extreme, in which a small price change drives a massive load shift and actually destabilizes the grid. Some degree of "fine grained" control would be better.

Thus, we suspect that over time, a different model will emerge: utilities will be motivated to create their own power management "apps" that offer beneficial pricing in exchange for direct grid control over some of these tasks: the grid operator might, for example, schedule dishwashing and clothes washing at times convenient to the grid, vary household heating to match patterns of use, heat water for showers close to when that hot water will be needed, etc.

But ***these are cloud computing concepts***: the iPhone, Blackberry, and Android are all so tightly linked to the cloud that it is just not meaningful to imagine them operating in any other way. Smarter homes can save power, but the applications enabling these steps must be designed to run on cloud computing systems, which will necessarily handle sensitive data, be placed into life-critical roles, and must be capable of digital "dialog" with the utility itself. All of these are the kinds of issues that motivate our recommendation that the power community start now to think about how such problems can be solved in a safe, trustworthy, and private manner.

Figure 3: Power-Aware Home Using Cloud-Hosted Power Management Applications ("Apps")

The big challenge for HPC in the cloud revolves around what some call the checkpoint barrier. The issue is this: modern HPC tools aren't designed to continue executions during failures. Instead, a computation running on n nodes will typically stop and

restart if one of the n fails. To ensure that progress is made, periodic checkpoints are needed. As we scale an application up, it must checkpoint more often to make progress. But checkpointing takes time. It should be clear that there is a number of nodes beyond which all time will be spent checkpointing and hence no progress can be made at all. On traditional HPC hardware platforms, the checkpoint barrier has not been relevant: failure rates are low. But cloud computing systems often have relatively high rates of node and storage server failures: having designed the systems to tolerate failures, it becomes a cost-benefit optimization decision to decide whether to buy a more reliable, but more costly server, or to buy a larger number of cheaper but less reliable ones. This then suggests that HPC in the current form may not migrate easily to the cloud, and also that it may not be possible to just run today's standard SCADA algorithms on large numbers of nodes as the scale of the problems we confront grows in response to the trends discussed earlier. New SCADA solutions may be needed in any case; versions matched closely to the cloud model may be most cost-effective.

4.4 High Assurance Applications and the Cloud Computing Dilemma

The cloud was not designed for high-assurance applications, and therefore poses several challenges for hosting a critical infrastructure service like the smart grid. One complicating factor is that many of the cost-savings aspects of the cloud reflect forms of sharing: multiple companies (even competitors) often share the same data center, so as to keep the servers more evenly loaded and to amortize costs. Multiple applications invariably run in a single data center. Thus, whereas the power community has always owned and operated its own proprietary technologies, successful exploitation of the cloud will force the industry to learn to share. This is worrying, because there have been episodes in which unscrupulous competition within the power industry has manifested itself through corporate espionage, attempts to manipulate power pricing, etc. (ENRON being only the most widely known example). Thus, for a shared computing infrastructure to succeed, it will need to have ironclad barriers preventing concurrent users from seeing one-another's data and network traffic.

The network, indeed, would be a shared resource even if grid operators were to run private, dedicated data centers. The problem here is that while one might imagine creating some form of separate Internet specifically for power industry use, the costs of doing so appear to be prohibitive. Meanwhile, the existing Internet has universal reach and is highly cost-effective. Clearly, just as the cloud has inadequacies today, the existing Internet raises concerns because of its own deficiencies. But rather than assuming that these rule out the use of the Internet for smart grid applications, we should first ask if those deficiencies could somehow be fixed. If the Internet can be enhanced to improve robustness (for example, with multiple routing paths), and if data is encrypted to safeguard it against eavesdroppers (using different keys for different grid operators), it is entirely plausible that the shared public Internet could emerge as the cheapest and most effective communication option for the power grid. Indeed, so cost-effective is the public Internet that the grid seems certain to end up using it even in

its current inadequate form. Thus, it becomes necessary to undertake the research that would eliminate the technical gaps.

We've discussed two aspects of the cloud in enough detail to illustrate the mindset with which one approaches these kinds of problems, using a market-based perspective to understand why cloud computing takes the form it does, and then using that same point of view to conceive of ways that technical improvements might also become self-sustaining cloud computing options once created, evaluated, and demonstrated in a convincing manner. But it is important to understand that these were just two of many such issues. Let's touch briefly on a few other important ones. Cloud computing is also peculiar in its access control and privacy capabilities [18][27][33]. Google's motto is "Don't be Evil", because in the cloud, the providers all must be trusted; if Google (or any of its thousands of employees) actually are evil, we may already be in a difficult situation. The cloud just doesn't have a serious notion of private data and, indeed, many in the industry have gone to lengths to point out that in a detailed, technical, legally binding sense, terms like privacy are very much up in the air today [33]. What precisely does it mean to ensure the privacy of an email, or a video, in a world where people casually send unencrypted messages over the public network, or share details of their personal histories with "friends" they know only as user-names on Facebook?

So extreme is this situation, and so pervasive the reach of the cloud, that it is already possible that any technical remedy could be out of reach. At minimum, the law lags the technology [29]. An editorial in the New York Times goes further, suggesting that the era of individual privacy may already be over [27], a sobering thought for those who hope to live unobserved, private lives.

Today's cloud technology is also weak in the area of reliability: the cloud is always up, but data centers often suffer from brief episodes of amnesia, literally forgetting something as soon as they learn it, and then (perhaps) rediscovering the lost information later. Sometimes, data is uploaded into a cloud, promptly lost, and never rediscovered at all. This can lead to a number of forms of inconsistency, a term used in the distributed computing community to refer to a system that violates intuitive notions of server correctness in ways that reveal the presence of multiple server replicas that are acting in uncoordinated ways, or using stale and incomplete data [4]. A consistency-preserving guarantee would eliminate such issues, but today's cloud systems manage well enough with weak consistency (after all, how much consistency is really required for a search query, or to play a video?) By imposing weak consistency as an industry standard, the cloud platforms become simpler and hence cheaper to build and to manage. Thus, yet again, we see economic considerations emerging as a primary determinant of what the cloud does and does not offer.

The issue goes well beyond service consistency. Cloud computing also places far greater emphasis on the robustness of the data center as a whole than on the robustness of any

of the hundreds of thousands of servers it may have within it: data centers casually shut servers down if they seem to be causing trouble. No reliability assumptions at all are made about client systems, in part because viruses, worms, and other malware have hopelessly compromised the technologies we run on client platforms. By some estimates [14][18], fully 80% of home computers are slaves in one or more Botnets, basically seeming normal (maybe slow) to the owner yet actually under remote control by shadowy forces, who can use the hijacked machines as armies in the Internet's version of warfare (for example, Estonia and Ukraine have both been taken off the network in recent years [14]), use them as host sites for illicit materials, or simply harness them as sources for waves of spam. In his fascinating analysis of the cyber-attack risks associated with network-based terrorism, Richard Clarke discusses the risks to today's power grid at some length [14]. In a nutshell, he shows that power control systems are poorly secured and can be attacked via the Internet or, using public information, attacked by cutting wires. Either outcome could be disastrous. Worst among his scenarios are attacks that use "logic bombs" planted long ahead of the event; he conjectures that such threats may already be widely disseminated in today's power grid control systems.

Clearly, this situation will need to change. The smart grid will play a wide range of safety and life-critical roles, and it is completely reasonable to invest more money to create a more robust technology base. For example, it is possible to use automated code verification techniques to prove that modest sized computing systems are correct. We can use hardware roots of trust to create small systems that cannot be compromised by viruses. By composing such components, we can create fully trustworthy applications. Such steps might not work for the full range of today's cloud computing uses (and might not be warranted for the cloud applications that run Twitter or Facebook), but with targeted investment, the smart grid community can reach a point of being able to create them and to deploy them into cloud environments.

To summarize, let's again ask what cloud computing is "really about". The past few pages should make it clear that the term is really about many things: a great variety of assumptions that can seem surprising, or even shocking, when stated explicitly. We have a model in which all data finds its way into one or more massive storage systems, which are comprised of large numbers of individually expendable servers and storage units. Cloud platforms always guarantee that the data center will be operational, and try to keep the main applications running, but are far weaker in their guarantees for individual data items, or individual computations. The cloud security and privacy guarantees are particularly erratic, leaving room for cloud operators to be evil if they were to decide to do so, and even leaving open the worry that in a cloud shared with one's competitors, there might be a way for the competition to spy on one's proprietary data or control activities. Yet there seem to be few hard technical reasons for these limitations: they stem more from economic considerations than from science. Given the life-critical role of the power grid, some way of operating with strong guarantees in all

of these respects would be needed, at least for the grid and for other "safety critical" purposes.

SUMMARY OF CLOUD PROPERTIES
CHARACTERISTICS OF TODAYS CLOUD COMPUTING AND INTERNET INFRASTRUCTURE

- **Inexpensive to own and operate.** Economies of scale, sharing, and automation are pervasive within cloud systems and central to the model.
- **Emphasis on rapid response and scalability.** Modern cloud computing systems are designed to ensure that every request from the client to the cloud receives a timely response, even if the response might be "incorrect".
- **Self-Managed, Power-Efficient, Self-Repairing.** Cloud computing systems are astonishingly green: they use power efficiently, keep machines busy, and dynamically adapt under all sorts of stresses, including load surges, failures, upgrades/downgrades, etc.
- **Weak Consistency Guarantees.** The embrace of the CAP theorem (see Section 6.4) has been used to justify a number of weak guarantees [31][37]. In a nutshell, most cloud services are capable of using stale data to respond to requests and the client is expected to deal with this. Cloud services are also unable to hide failures: the client must anticipate sudden faults and should reissue requests or otherwise compensate to mask such events.
- **Internet as a weak point.** The modern Internet experiences a surprising number of brief outages. Cloud computing systems are expected to ride them out. Multi-homing is offered for the cloud but not the average client (a cloud can be addressed by two or more distinct IP addresses), but we lack true multi-path routing options, so even with multi-homing, some clients may experience long periods of disrupted connectivity.

Figure 4: Summary of Assurance Properties

5 Three Styles of Power Computing

We now concretize the foregoing discussion by grouping smart grid computing into three loosely defined categories. These are as follows:

i. Applications with weak requirements. Some applications have relatively relaxed needs. For example, because it takes a long time to install new transmission lines, applications that maintain maps of the physical infrastructure in a power delivery region will change relatively rarely, much as road maps rarely change. They can be understood as systems that provide guarantees but against easy constraints. Today's cloud is well matched to these uses.

ii. Real-time applications. This group of applications needs extremely rapid communication, for example to move sensor readings or SCADA control information fast enough to avoid actions based on stale data. Some studies suggest that for many SCADA control policies, even 50ms of excess delay relative to the minimum can be enough to result in incorrect control decisions [23][25][1]. Today's cloud is tuned to provide fast responses, but little attention has been given to maintaining speed during failures of individual server nodes or brief Internet connectivity disruptions.

iii. Applications with strong requirements. A final class of applications requires high assurance, strong access control and security policy enforcement, privacy, fault-tolerance, consistent behavior over collections of endpoints at which actions occur, or other kinds of properties. We will argue that the applications in this class share common platform requirements, and that those differ (are incomparable with) the platform properties needed for real-time purposes [4][5][36][23]. Today's cloud lacks the technology for hosting such applications.

We've argued that the cloud takes the form seen today for economic reasons. The industry has boomed, and yet has been so focused on rolling out new competitively exciting technologies and products that it has been limited by the relative dearth of superb engineers capable of creating and deploying new possibilities. The smart grid would have a tough time competing head to head for the same engineers who are focused on inventing the next Google, or the next iPad. However, by tapping into the academic research community, it may be possible to bring some of the brightest minds in the next generation of researchers to focus on these critical needs.

Figure 5 summarizes our observations. One primary conclusion is that quite a bit of research is needed simply to clarify the choices we confront. Yet the broader picture is one in which a number of significant technology gaps clearly exist. Our strong belief is that these gaps can be bridged, but we also see strong evidence that today's cloud developers and vendors have little incentive to do so and, for that reason, that a watch-and-wait approach would not succeed.

	Fits current cloud computing model	*Poses demanding HPC computing challenges*	*Collects data at many locations*	*Takes actions at many locations*	*Requires rapid real-time response*	*Requires strong consistency*	*Needs to protect personal or proprietary data*	*Must protect against attack*
Smart home	**Varies (1)**						✓	**? (3)**
Next generation SCADA with alternative power generation		✓	✓	✓	✓	**? (3)**		**? (3)**
Support for wide-area power contracts		✓	✓	✓		✓	✓	✓(2)
Grid protection			✓	✓	✓	**? (3)**		✓
Grid status monitoring	**? (3)**		✓			**? (3)**		✓(2)

Figure 5: Cloud-Hosted Smart Grid Applications: Summary of Assurance Requirements

Notes:

(1) Some prototypical "smart home" systems operate by using small computing devices to poll cloud-hosted web sites that track power pricing, then adapt actions accordingly. However not all proposed home-adaptation mechanisms are this simple; many would require closer coordination and might not fit the current cloud model so closely.

(2) Concerns here include the risk that disclosure of too much information could give some producers opportunities to manipulate pricing during transient generation shortages, and concerns that without publishing information about power system status it may be hard to implement wide-area contracts, yet that same information could be used by terrorists to disrupt the grid.

(3) Further research required to answer the question.

6 Technical Analysis of Cloud Computing Options

Some technical questions need more justification than was offered in the preceding pages. This section undertakes a slightly deeper analysis on a few particularly important issues. We reiterate claims made earlier, but now offer a more specific explanation of precisely why these claims are valid and what, if anything, might be done about the issues identified.

6.1 Rebooting a cloud-controlled smart grid

One place to start is with a question that many readers are no doubt puzzled by: the seeming conundrum of implementing a smart grid control solution on top of an Internet that would be incapable of functioning without power. How could one restart such a system in the event of a loss of regional power? There are two basic elements to our response. First: geographic diversity. Cloud computing makes it relatively easy to replicate control functionality at two or more locations that operate far from one another and hence, if one is lost, the other can step in. As for the Internet, it automatically reroutes around failures within a few minutes. Thus, for many kinds of plausible outages that impact a SCADA system at one location, having a software backup at a modest distance is sufficient: shipping photons is cheap and fast. In the Internet, nobody knows if their SCADA system is running next door, or two states over. Geographic diversity is also interesting because, at least for cloud operators, it offers an inexpensive way to obtain redundancy. Rather than building dual systems, as occurs in many of today's SCADA platforms for the existing power grid, one could imagine cloud-hosted SCADA solutions that amortize costs in a similar manner to today's major cloud applications, and in this way halve the cost of deploying a fault-tolerant solution. But one can imagine faults in which a remote SCADA platform would be inaccessible because the wide-area network would be down, due to a lack of power to run its routers and switches. Thus, the second part of the answer involves fail-safe designs. The smart grid will need to implement a safe, "dumb" mode of operation that would be used when restarting after a regional outage and require little or no fine-grained SCADA control. As the system comes back up, more sophisticated control technologies could be phased back in. Thus, the seeming cycle of dependencies is broken: first, one restores the power; next, the Internet; last, the more elaborate forms of smart behavior.

6.2 Adapting standard cloud solutions to support more demanding applications

We've repeatedly asserted that the cloud is cheap. But why is this the case, and to what extent do the features of today's cloud platforms relate to the lower cost of those platforms?

Cloud computing can be understood as an approach that starts with client-server computing as its basis, and then scales it up dramatically – whereas server systems of the past might have run on 32 nodes, cloud systems often have hundreds of thousands of machines, each of which may have as many as 8 to 16 computational cores. Thus a cloud computing system is a truly massive structure. Some are as large as 4-5 football fields, packed so densely with computing and storage nodes that machines are purchased by the container-truck load and the entire container is literally "plugged in" as a unit. Yet as vast as these numbers may be, they are dwarfed by the even larger number of client systems. Today, it is no exaggeration to say that every laptop, desktop, pad, and even mobile telephone is a cloud-computing client system. Many have literally dozens of cloud applications running at a time. Thus the cloud is a world of billions of end user systems linked, over the Internet, to tens of millions of servers, residing in data centers that individually house perhaps hundreds of thousands or millions of machines. The cost advantage associated with this model relates to economies of scale. First, simply because of their scale, cloud computing systems turn out to be remarkably inexpensive to own and operate when compared with a small rack of servers such as one finds in most power industry control centers. James Hamilton, in his widely cited blog at http://mvdirona.com, has talked about the "cost of a cloud." He concludes that relative to other types of scalable infrastructure, the overall cost of ownership is generally a factor of 10 to 15 lower when all costs are considered (human, infrastructure, servers, power, software development, etc.). This is a dramatic advantage. Cloud systems also run "hot": with buildings packed with machines, rather than humans, the need for cool temperatures is greatly reduced. The machines themselves are designed to tolerate these elevated temperatures without an increased failure rate. The approach is to simply draw ambient air and blow it through the data center, without any form of air conditioning. Interior temperatures of 100°+F are common, and there has been talk of running clouds at 120°F. Since cooling costs money, such options can significantly reduce costs.

Furthermore, cloud systems often operate in places where labor costs and electric power costs are cheap: if a large power consumer is close to the generator, the excess power needs associated with transmission line loss are eliminated and the power itself becomes cheaper. Thus, one doesn't find these systems in the basement of the local bank; they would more often be situated near a dam on a river in the Pacific Northwest. The developers reason that moving information (such as data from the client computing system) to the cloud, computing in a remote place, and moving the results back is a relatively cheap and fast option today, and the speed and growth trends of the Internet certainly support the view that as time passes, this approach might even do better and better.

6.3 The Internet as a weak link

We've asserted that the Internet is "unreliable," yet this may not make sense at first glance; all of us have become dependent on a diversity of Internet-based mechanisms.

Yet upon reflection, the concern makes more sense: anyone who uses an Internet radio, or who owns a television adapter that supports watching movies on demand, quickly realizes that while these technologies "usually" are quite robust, "sometimes" outages do occur. The authors of this white paper own a number of such technologies and have sometimes experienced multiple brief outages daily, some lasting just seconds, and others perhaps minutes. Voice over IP telephony is a similar experience: users of Skype think nothing of needing to try a call a few times before it goes through. Moreover, all of these are consequences of mundane issues: studies reveal that the Internet glitches we've been talking about are mostly triggered by operator error, brief load surges that cause congestion, or by failures of the routers that support the network; a typical network route today passes through 30 or more routers and when one goes offline, the Internet may need as much as 90 seconds to recover full connectivity. Genuinely long Internet outages have occurred more rarely, but they do happen from time to time, and the root causes can be surprising: in one event, an undersea cable got severed off Egypt, and India experienced disrupted network connectivity for some several days [1].

When the Internet has actually come under attack, the situation is much worse. Experience with outright attacks on the network is less limited than one might realize: recent events include so-called distributed denial of service attacks that have taken entire small countries (such as Estonia) off the network for weeks, disrupted government and military web sites, and harassed companies like Google (when that company complained about China's political policies recently). A wave of intrusions into Department of Defense (DOD) classified systems resulted in the theft of what may have been terabytes of data [14]. Researchers who have studied the problem have concluded that the Internet is really a very fragile and trusting infrastructure, even when the most secure protocols are in use. The network could be literally shut down, and there are many ways to do it; some entirely based on software that can be launched from anywhere in the world (fortunately, complex software not yet in the hands of terrorists); other attacks might deliberately target key components such as high-traffic optical cables, using low-tech methods such as bolt cutters. Thus any system that becomes dependent upon the Internet represents a kind of bet that the Internet itself will be up to the task.

Thus the Internet is one "weak link" in the cloud computing story. We tolerate this weak link when we use our web phones to get directions to a good restaurant because glitches are so unimportant in such situations. But if the future smart grid is to be controlled over a network, the question poses itself: would this be the Internet, in a literal sense? Or some other network to be constructed in the future? On this the answer is probably obvious: building a private Internet for the power grid would be a hugely expensive proposition. The nation might well contemplate that option, but when the day comes to make the decision, we are not likely to summon the political will to invest on the needed scale. Moreover, that private Internet would become an extension of the public Internet the moment that some enterprising hacker manages to

compromise even a single machine that has an Internet connection and also has a way to talk to the power network.

This is why we've concluded that the best hope is for a technical advance that would let us operate applications that need a secure, reliable Internet over today's less secure, less reliable one. Achieving such a capability would entail improving handling of failures within today's core Internet routers (which often are built as clusters but can be slow to handle failures of even just a single router component), and also offering enhanced options for building secure routes and for creating redundant routes that share as few links as possible, so that if one route becomes disrupted or overloaded, a second route might still be available. In addition, the power grid can make use of leased connections to further improve reliability and performance.

6.4 Brewer's CAP Conjecture and the Gilbert/Lynch CAP Theorem

We've discussed the relatively weak consistency properties offered by today's cloud computing platforms and even commented that cloud providers "embrace inconsistency" as a virtue [31][37]. Why is this the case, and can we hope to do anything about it? Cloud computing systems are so massive (and yet built with such relatively "weak" computers) that the core challenge in building cloud applications is to find ways to scale those applications up, so that the application (a term that connotes a single thing) might actually be implemented by thousands or even tens of thousands of computers, with the user's requests vectored to an appropriate machine.

How can this form of scaling be accomplished? It turns out that the answer depends much on the extent to which different user systems need to share data:

- At the easiest end of the spectrum we find what might be called "shared nothing" applications. A good example would be the Amazon shopping web pages. As long as the server my computer is communicating with has a reasonable approximation of the state of the Amazon warehouse systems, it can give me reasonable answers to my queries. I won't notice if a product shows slightly different popularity answers to two identical queries reaching different servers at the same time, and if the number of copies of a book is shown as 3 in stock, but when I place my order suddenly changes to 1, or to 4, no great harm occurs. Indeed, many of us have had the experience of Amazon filling a single order twice, and a few have seen orders vanish entirely. All are manifestations of what is called "weak consistency" by cloud developers: a model in which pretty good answers are considered to be good enough. Interestingly, the computations underlying web search fall solidly into this category – so much so that entire programming systems aimed at these kinds of computing problems have become one of the hottest topics for contemporary research; examples include MapReduce [16] and other similar systems, file systems such as Google's GFS [20] and the associated BigTable database layered on top of it [13], etc.

These are systems designed with loose coupling, asynchronous operation and weak consistency as fundamental parts of their model.

- A slightly harder (but not much harder) problem arises in social networking sites like Twitter or Facebook where groups of users share data, sometimes in real-time. Here, the trick turns out to be to control the network routing protocols and the so-called Domain Name Service (DNS) so that people who share data end up talking to the same server. While a server far away might pull up the wrong version of a page, or be slow to report a Tweet, the users talking to that single server would be unaware that the cloud has split its workload into perhaps millions of distinct user groupings.
- Gaming and Virtual Reality systems such as Second Life are similar to this second category of systems: as much as possible, groups of users are mapped to shared servers. Here, a greater degree of sophistication is sometimes needed and computer gaming developers publish extensively on their solutions: one doesn't want to overload the server, and yet one does want to support games with thousands of players. eBay faces a related challenge when an auction draws a large number of bidders. Such systems often play small tricks: perhaps not every bidder sees the identical bid sequence on a hotly contended-for item. As long as we agree on the winner of the auction, the system is probably consistent enough.
- Hardest of all are applications that really can't be broken up in these ways. Air Traffic Control would be one example: while individual controllers do "own" portions of the air space, because airplanes traverse many such portions in short periods of time, only an approach that treats the whole airspace as a single place and shows data in a consistent manner can possibly be safe. The "my account" portion of many web sites has a similar flavor: Amazon may use tricks to improve performance while one shops, but when an actual purchase occurs, their system locks down to a much more careful mode of operation.

The trade-offs between consistency and scalable performance are sometimes summarized using what Eric Brewer has called the Consistency Availability and Partitioning (CAP) theorem [11]. Brewer, a researcher at UC Berkeley and co-founder of Inktomi, argued in a widely cited keynote talk at PODC 2000 that to achieve high performance and for servers to be able to respond in an uncoordinated, independent manner to requests they receive from independent clients, those servers must weaken the consistency properties they offer. In effect, Brewer argues that weak consistency scales well and strong consistency scales poorly. A formalization of CAP was later proved under certain weak assumptions by MIT's Gilbert and Lynch, but data centers can often make stronger assumptions in practice, and consequently provide stronger properties. Moreover, there are many definitions of consistency, and CAP is only a theorem for the specific definition that was used in the proof. Thus CAP is something of a folk-theorem: a convenient paradigm that some data centers cite as a reason for offering weak consistency guarantees (guarantees adequate for their own needs,

although inadequate for high assurance purposes), yet not a "law of nature" that cannot be circumvented under any circumstances.

We believe that more investigation is needed into the scalability and robustness options that weaker consistency models might offer. CAP holds under specific conditions; perhaps data centers can be designed to invalidate those conditions most closely tied to the impossibility result. Hardware assistance might be helpful, for example in supporting better forms of cloud security. Thus CAP stands as an issue, but not one that should discourage further work.

6.5 Hidden Costs: Security Implications of Weak Consistency

Cloud security illustrates one of the dangers of casual acceptance of the CAP principles. We build secure systems starting with specifying a security policy that the system is expected to obey. Typically, these policies consist of rules and those rules are represented as a kind of database; the data in the database gives the logical basis for making security decisions and also identifies the users of the system and the categories of data. As the system runs, it can be thought of as proving theorems: Joe is permitted to access Sally's financial data because they are a couple; Sandra can do so because she is Sally's banker. John, Sally's ex-husband, is not permitted to access those records. The data evolves over time, and correct behavior of the system depends upon correct inference over the current versions of the underlying rules and the underlying data.

Cloud systems have real difficulty with these forms of security, because the same embrace of weak consistency that makes them so scalable also implies that data may often be stale or even outright wrong when the system tries to operate on it. Perhaps some node will be slow to learn about Sally's divorce – maybe it will never learn of it. Cloud systems don't provide absolute guarantees about such things, on the whole, and this makes them easier to scale up. But it also makes them deeply – perhaps fundamentally – untrustworthy.

The term "trustworthy" deliberately goes beyond security. Suppose that a smart grid control device needs to handle some event: perhaps line cycles drop or increase slightly, or a current surge is sensed. To coordinate the reaction appropriately, that device might consult with its cloud server. But even if connectivity is not disrupted and the cloud server is running, we run into the risk that the server instance that responds – perhaps one of a bank of instances that could number in the thousands – might have stale data and hence respond in an incorrect manner. Thus it is entirely possible for 99 servers to "know" about some new load on the grid, and yet for 1 server to be unaware of this, or to have data that is incorrect ("inconsistent") in a plethora of other ways.

Cloud systems are also quite casual about restarting servers even while they are actively handling client requests – this, too, is part of the scalability model (it reduces the human cost of management, because one doesn't need to gracefully shut things down before

restarting them or migrating them). Thus our smart grid control device might find itself working off instructions that reflect faulty data, or deprived of control in an abrupt, silent manner, or suddenly talking to a new controlling server with no memory of the recent past.

7 Pretty Good is Sometimes Good Enough

Cloud computing is a world of very large scale systems in which most components are working correctly even if a few are lagging behind, working with stale data, restarting after an unplanned and sudden outage, or otherwise disrupted. Yet it is vital to realize that for many purposes these properties are good enough. Facebook, Youtube, Yahoo, Amazon, Google, MSN Live – all are examples of systems that host vast numbers of services that work perfectly well against this sort of erratic model. Google's difficulties repelling hacker attacks (apparently from China) do give pause; this event illustrates the downside of the cloud model; it is actually quite hard for Google to secure its systems for the same reasons we discussed earlier: security seems to be at odds with the mechanisms that make those systems scalable. Moreover, the cloud model would seem to create loopholes that hackers can exploit (including the massive and remote nature of the cloud centers themselves: ready targets for agents of foreign powers who might wish to intrude and introduce virus or other undesired technical components).

The frustration for many in the field today is that we simply don't know enough about what can be solved in the standard cloud model. We also don't know enough about mapping stronger models onto cloud-like substrates or onto the Internet. Could the same hardware that runs the Internet not host software that might have better network security and reliability characteristics? One would be foolish to assert that this cannot be done. Could the same platforms we use in cloud settings not support applications with stronger properties? Very possibly. We simply don't know how to do so, yet, in part for the reason just cited: Google, Microsoft, Yahoo, and others haven't had much need to do this, and so the huge investment that gave us the cloud hasn't seen a corresponding investment to create a highly assured cloud for mission-critical roles.

Moreover, one can turn the problem on its head and ask whether control of the future smart grid actually requires consistency and coherency. Very possibly, one can control a smart grid in a manner that relies on a "mostly consistent" behavior by huge numbers of relatively loosely coupled, autonomous control agents. Perhaps centralized servers aren't even needed or, if they are needed, they don't need to behave in a manner one would normally think of as reflecting central control – terminology that already evokes the image of a single entity that makes the control decisions.

Finally, it is worthwhile to recognize that while the smart grid community may be confronting these problems for its own reasons, the community is certainly not alone. A future work of increasingly automated health care systems will surely have similar

needs (imagine, for example, a substantial community of elderly home-care diabetic patients who depend upon remote control of their insulin pumps: the picture is comparable and the same concerns apply). Electronic medical health records will demand a strong model, at least as far as security, privacy, and rapid accurate data reporting are concerned. The same is true of banking systems, systems controlling infrastructure such as water or traffic lights, and indeed a plethora of socially sensitive, critical applications and services. Cloud computing beckons through its attractive price-point, but to benefit from that price point, we need to learn to move applications with sensitive requirements onto the cloud.

8 A Research Agenda

This paper was written to expose a problem, but not to solve it. The problem, as we've now seen, is that many of the most exciting ideas for the future smart grid presuppose models of computing that have become outmoded and are being replaced by cloud computing. Others require a kind of scalability that only cloud computing can offer. And even mundane ideas sometimes have failed to grapple with the implications of an industry shift in which cloud computing has become a universal answer to every need: a commodity standard that is sweeping all other standards to the side. Familiar, successful computing models of the recent past may be the unsupported legacy challenges of the near-term future.

Yet cloud computing, as we've shown, lacks key properties that power control and similar smart grid functionality will need. These include security, consistency, real-time assurances, ways to protect the privacy of sensitive data, and other needs.

A doom-and-gloom story would, at this point, predict catastrophe. But the authors of this survey believe that every problem we've noted can probably be solved. The key is to incentivize researchers to work on these problems. Somewhat astonishingly, that research is not occurring today. With the exception of work on computer security, the government has largely pulled back from funding what could be called "basic systems" research, and there are no major research programs focused on highly assured cloud computing at NSF, DARPA, or other major government research agencies today. In effect, we're making a wager that industry will solve these problems on its own. Yet as noted above, cloud computing systems are under at most modest economic incentives to tackle these needs. They don't impact the bottom line revenue stream in major ways, and cloud computing has been shaped, up to now, by the revenue stream. To us this suggests that such a wager might fail.

Accordingly, we recommend that the nation embark on a broad-reaching and multi-faceted research effort. This effort would have elements specific to the smart electric power grid, but other elements that are cross-cutting and that would seem equally

beneficial in future medical systems, banking systems, and a wide range of other application areas:

i. Quantify the kinds of guarantees that cloud computing solutions can offer. The goal of this effort would be to create a scientific foundation for cloud computing, with the mathematical and practical tools one associates with any scientifically rigorous foundation.

ii. Quantify the kinds of guarantees that are required for a new generation of smart grid control paradigms. This effort would seek to develop new strategies for control of a smart power grid, perhaps including such elements as decentralized control points and some degree of autonomous local control for smaller devices such as home units that might adapt their power consumption to better exploit off-peak power and reduce peak needs. It would then look at various ways to implement those strategies on cloud platforms.

iii. Learn to reintroduce strong trust properties in cloud settings. Perhaps the conclusion from these first efforts would be that today's CAP-conjecture-based cloud is ideally suited to some new style of weakly consistent control paradigm. But we may also find that some applications simply require cloud applications that can scale well and be administered cheaply, and yet that offer strong guarantees of security, consistency, availability, fault-tolerance, etc. If so, it will be incumbent upon us to learn to host such applications in cloud settings.

iv. Better quantify the possible attacks against a computer-controlled smart grid. We've seen that energy producers might be motivated to manipulate power markets (cf. the Enron situation of some years ago), and Clarke's book points to the possibility of hackers or even foreign powers that might single out the power grid as their target. Needed are a careful analysis of the threats – all forms of threats – and a considered strategy for building systems that might defend against such attacks.

v. Learn to build an Internet with better availability properties, even under attack. Today's Internet has one primary role: it needs to get the data from the sender to the receivers and while reliability isn't a need on a packet-by-packet basis, we are learning that reliability does matter for "flows" that occur over longer periods of time. But the Internet isn't reliable in this second sense, and is easily attacked. We need to find ways to evolve the network to have much higher reliability for packet flows that need stronger assurance properties, and to do so even when the network comes under attack.

vi. Improve attack tolerance. If we are to build nationally critical infrastructures on the Internet, the day may come when adversaries attack those infrastructures. Today, this would result in serious disruption; tomorrow, as the dependencies enlarge, the results may be devastating. Thus it is obligatory to learn to build attack-tolerant versions of the key components of the future infrastructure. This is a tall order, but short of rejecting the Internet and the cloud as inappropriate for critical use, there really is no alternative but to find ways to secure what we build against major, deliberate, coordinated, sophisticated attacks.

It would not be honest to offer such a list without also observing that this is a tremendously ambitious, difficult agenda. Saying that such-and-such a problem "must be solved" is easy; estimating the time and resource needs to solve it is another matter. Worse still, the topics we've listed aren't typical of the areas receiving the most research energy and enthusiasm today.

A further observation, of a similar nature, is that computer security has been a source of frustration for decades; we've made huge progress and yet the landscape has shifted beneath our feet in such a way that the problems we've solved seem like issues that haven't mattered in decades, while the problems of the day seem far beyond reach. So to say that we need to "find a way" to create trustworthy cloud computing applications is facile and perhaps unrealistic. It may be that we will never reach a point at which computing can really be trusted in the senses required!

Yet it would also seem premature to give up. While there is a CAP theorem, we've commented that it holds only under very weak assumptions and there is no hard-and-fast reason that data centers can't make stronger assumptions. For this reason, CAP is more of a folk theorem: those who wish to build weakly consistent systems use CAP to justify their approach, elevating it to the status of a theorem perhaps as much to justify their own endorsement of weak properties as for any mathematically rigorous reason. Meanwhile, the theory community points to the theorem as an impossibility result, seemingly unaware that many cloud systems wouldn't match the assumptions used in proving the result, and hence aren't "bound" by it. And this same comment could be made in a much broader way: There is little concrete evidence that the obstacles to highly assured cloud computing are even all that hard. Perhaps all that is needed is new, talented minds and new kinds of applications, such as the smart grid, to help motivate the work and to ground it in reality. Lack of funding has impeded this entire area for almost a decade (triggered by a DARPA pull-back under the Bush administration). Thus, with more resources, an exciting and important problem, and perhaps some really bright young researchers, it may actually be possible to move mountains.

SUMMARY OF HIGHEST PRIORITY RESEARCH TOPICS

1. Quantify the kinds of guarantees that cloud computing solutions can offer.
2. Quantify the kinds of guarantees that are required for a new generation of smart grid control paradigms.
3. Learn to reintroduce strong trust properties in cloud settings.
4. Better quantify the possible attacks against a computer-controlled smart grid.
5. Learn to build an Internet with better availability properties.
6. Improve attack tolerance.

Figure 6: Summary of the most urgent research topics

9 Conclusions

The smart grid challenges us today: creating it could be the first and perhaps most important step towards a future of dramatically improved energy efficiency and flexibility. The Internet and the Cloud Computing model around which it has coalesced appear to be natural partners in this undertaking, representing the culmination of decades of work on high-productivity, low-cost computing in a distributed model. But only if the gap between the needs of the smart grid and the properties of the cloud can be bridged can these apparent opportunities be safely realized.

References

[1] Murad Ahmed. India Suffers Massive Internet Disruption After Undersea Cables Break. The Times. December 19, 2008.

[2] David E. Bakken, , Anjan Bose, Carl H. Hauser, Edmond O. Schweitzer III, David E. Whitehead, and Gregary C. Zweigle. Smart Generation and Transmission with Coherent, Real-Time Data. Technical Report TR-GS-015. School of Electrical Engineering and Computer Science. Washington State University; Pullman, Washington, USA. August, 2010

[3] Hari Balakrishnan, M. Frans Kaashoek, David Karger, Robert Morris, Ion Stoica. Looking up data in P2P systems. February 2003 Communications of the ACM, Volume 46 Issue 2

[4] Kenneth Birman. Reliable Distributed Systems Technologies, Web Services, and Applications. P. 2005, XXXVI, 668 p. 145 illus., Hardcover ISBN: 0-387-21509-3.

[5] Ken Birman. History of the Virtual Synchrony Replication Model. Appears in Replication: Theory and Practice. B. Charron-Bost, F. Pedone, A. Schiper (Eds) Springer Verlag, 2010. LNCS 5959, pp. 91–120, 2010.

[6] Ken Birman and Robbert van Renesse, eds. Reliable Distributed Computing with the Isis Toolkit. IEEE Computer Society Press, 1994, Los Alamitos, Ca.

[7] Ken Birman, Gregory Chockler, Robbert van Renesse. Towards A Cloud Computing Research Agenda. SIGACT News Distributed Computing Column. June 2009.

[8] Ken Birman, Coimbatore Chandersekaran, Danny Dolev, and Robbert van Renesse. How the Hidden Hand Shapes the Market for Software Reliability. In Proceedings of the First IEEE Workshop on Applied Software Reliability, Philadelphia, PA. June 2006.

[9] Kenneth P. Birman, Jie Chen, Kenneth M. Hopkinson, Robert J. Thomas, James S. Thorp, Robbert van Renesse, and Werner Vogels. Overcoming Communications Challenges in Software for Monitoring and Controlling Power Systems. Proceedings of the IEEE. Vol. 9, No. 5. May 2005.

[10] Anjan Bose. Electric Power Systems. Chapter in The Electrical Engineering Handbook, Wai Kai Chen Ed., Elsevier Academic Press, 2005.

[11] Eric A. Brewer. Towards Robust Distributed Systems Distributed Systems. Keynote presentation, ACM PODC, July 2000. www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf

[12] Mike Burrows. The Chubby lock service for loosely-coupled distributed systems. OSDI '06, November 2006.

[13] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data. Transactions on Computer Systems (TOCS) , Volume 26 Issue 2, June 2008.

[14] Richard Clarke and Robert Knake. Cyber War: The Next Threat to National Security and What to Do About. Ecco (April 20, 2010).

[15] Brian F. Cooper, Raghu Ramakrishnan, Utkarsh Srivastava, Adam Silberstein, Philip Bohannon, Hans-Arno Jacobsen, Nick Puz, Daniel Weaver, Ramana Yerneni. PNUTS: Yahoo!'s hosted data serving platform. Proceedings of the VLDB, Volume 1 Issue 2 August 2008.
[16] Jeffrey Dean, Sanjay Ghemawat. MapReduce: a flexible data processing tool. January 2010 Communications of the ACM, Volume 53 Issue 1.
[17] Joseph DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Werner Vogels. Dynamo: Amazon's highly available key-value store. SOSP 2007, Oct 2007.
[18] Whitfield Diffie, Susan Eva Landau. Privacy on the Line: The politics of wiretapping and encryption. MIT Press, 1999.
[19] Kira Fabrizio, Nancy Rose, Catherine Wolfram. Do Markets Reduce Costs? Assessing the Impact of Regulatory Restructuring on the U.S. Electric Generation Efficiency. American Economic Review. 2007.
[20] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google File System. SOSP '03 (December 2003), Brighton Beach, England.
[21] James Hamilton. Internet Scale Efficiency. Keynote presentation at LADIS 2008, the 3rd Workshop on Large Scale Distributed Systems (IBM Hawthorne Research Facilities, Sept 2008). Presentation materials available at http://www.cs.cornell.edu/projects/ladis2008/materials/JamesRH_Ladis2008.pdf
[22] Maya Haridasan, Robbert van Renesse. Defense Against Intrusion in a Live Streaming Multicast System. In Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing (P2P2006), Cambridge, UK, September 2006.
[23] Chi Ho, Robbert van Renesse, Mark Bickford, Danny Dolev. Nysiad: Practical Protocol Transformation to Tolerate Byzantine Failures. USENIX Symposium on Networked System Design and Implementation (NSDI 08). San Francisco, CA. April 2008.
[24] Kenneth Hopkinson, Kate Jenkins, Kenneth Birman, James Thorp, Gregory Toussaint, and Manu Parashar. Adaptive Gravitational Gossip: A Gossip-Based Communication Protocol with User-Selectable Rates. IEEE Transactions on Parallel and Distributed Systems. December 2009 (vol. 20 no. 12) pp. 1830-1843
[25] Hopkinson, Ken M.; Giovanini, Renan; Wang, Xaioru; Birman, Ken P.; Coury, Denise V.; Thorp, James S. IEEE Transactions on Power Systems. EPOCHS: Integrated Cots Software For Agent-Based Electric Power And Communication Simulation (Earlier version published as Winter Simulation Conference.7-10 of December 2003, New Orleans, USA.)
[26] Flavio P. Junqueira, Benjamin C. Reed. The life and times of a Zookeeper (Talk abstract). PODC '09, August 2009.
[27] Paul Kirn. Little Brother is Watching. New York Times, Oct. 15 2010.

[28] Vivek Kundra. Federal Cloud Computing Strategy. http://www.cio.gov/documents/Federal-Cloud-Computing-Strategy.pdf. February 2011.
[29] Laurence Lessig. Code and Other Laws of Cyberspace (2000)ISBN 978-0-465-03913-5.
[30] LG Appliances. Internet Refrigerator. http://us.lge.com/www/product/refrigerator_demo.html
[31] Dan Pritchett. BASE, an ACID Alternative. ACM Queue, July 2008.
[32] Fred B. Schneider and Ken Birman. The Monoculture Risk Put into Context. IEEE Security & Privacy. Volume 7, Number 1. Pages 14-17. January/February 2009.
[33] Polly Spenger. Sun on Privacy: "Get Over It.". Wired Magazine, Jan 26, 1999.
[34] Robert J Thomas. Issue Papers on Reliability and Competition: Managing Relationships Between Electric Power Industry Restructuring and Grid Reliability. PSERC technical report Aug 2005. http://www.pserc.wisc.edu/documents/publications/papers/2005_general_publications/thomas_grid_reliability_aug2005.pdf. School of Electrical and Computer Engineering, Cornell University.
[35] Robbert van Renesse, Kenneth Birman, Werner Vogels. Astrolabe: A Robust and Scalable Technology for Distributed System Monitoring, Management, and Data Mining. ACM Transactions on Computer Systems, May 2003, Vol.21, No. 2, pp 164-206.
[36] Robbert van Renesse, Rodrigo Rodrigues, Mike Spreitzer, Christopher Stewart, Doug Terry, Franco Travostino. Challenges Facing Tomorrow's Data center: Summary of the LADiS Workshop. Large-Scale Distributed Systems and Middleware (LADIS 2008). September 2008.
[37] Werner Vogels. Eventually Consistent. ACM Queue. December 2008.
[38] Yuan Yu, Michael Isard, Dennis Fetterly, Mihai Budiu, Ulfar Erlingsson, Pradeep Kumar Gunda, Jon Currey. DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language. Symposium on Operating System Design and Implementation (OSDI), San Diego, CA, December 8-10, 2008.
[39] US Department of Defense. Eligible Receiver. June 9-13, 1997. http://www.globalsecurity.org/military/ops/eligible-receiver.htm
[40] US Department of Energy. The Smart Grid: An Introduction. 2008. http:// www.oe.energy.gov/DocumentsandMedia/DOE_SG_Book_Single_Pages(1).pdf

Discussant Narrative

Running Smart Grid Control Software on Cloud Computing Architectures

James Nutaro
Oak Ridge National Laboratory

The smart grid is a vision of how recent innovations in computing and communications will be used to meet energy challenges in the coming decades. Central to this vision are sophisticated software systems that make a "smart grid" smart. Kenneth P. Birman, Lakshmi Ganesh, and Robbert van Renesse in their paper "Running Smart Grid Control Software on Cloud Computing Architectures" argue that cloud computing will provide the scalable, reliable, and affordable computing resources needed to operate a smarter, software-intensive grid. First, however, engineers must design this grid, and cloud computing may affordably provide the abundant computational resources that will be needed to do it.

The power industry, with its growing reliance on software for essential operations, is pursuing a technical revolution with successful precedents in the industries of aviation, defense, and industrial automation. An unmistakable feature of these industries is the sophisticated simulators used to design and test their products. There are three factors that drive the extensive use of simulators in these software-intensive industries. First is the prohibitive cost of prototyping hardware that software will control. It is infeasible to build major elements of an electrical power system for the purpose of programming its next generation of controllers. Second is the cost of testing software that controls critical systems. The failure of software in a critical control system could have catastrophic effects: the only safe and affordable vehicle for testing is a simulator. Third is the cost of training a system's operators. Each mundane task that automation takes from an operator is replaced by a new, more sophisticated task enabled by the same automation technology. Again, simulators are the only cost-effective way to train the operators of these new, more sophisticated systems.

Because of these issues, sophisticated simulators are inseparable from the engineering of highly automated, "smart" systems. The benefits of automation – in reduced operating costs, improved reliability, and better performance – justify in turn a substantial investment in computing and simulation technology. In every industry that has made the transition to software-intensive control, powerful and sophisticated simulation technology has led the way, and as the power industry transforms itself into a software-

intensive business, it too will discover that powerful and sophisticated simulators are inseparable from affordable and reliable energy systems. Unlike other industries, however, power systems pose uniquely challenging computational problems of unprecedented scale and consequence: a computational problem that cloud computing may help to solve.

Making cloud computing into an engineering tool is a necessary step towards the vision of Birman, Ganesh, and van Renesse. To make this step requires advances in two of the four areas identified by Birman and his collaborators. First is the protection of data: the simulators, the data that they operate on, and the information that they produce contain trade secrets and other information that utility companies and engineering firms are unwilling to share. Theft of this data could have serious repercussions for the business that owns them, and the use of cloud computing as an engineering tool is therefore predicated on the protection of data. Second is in support for scalable, consistency guaranteed, fault-tolerant services or, inversely, the development of new simulation methods that are appropriate to the computing environments offered by clouds today. Present approaches to simulation for engineering assume a computing platform that is fault-free and consistent. Scalability, traditionally less significant in engineering simulations, is a crucial issue for the smart grid. Advances in these two areas will move us closer to the vision of a smart grid managed in the Cloud by making the design of such a smart grid possible.

Recorder Summary

Running Smart Grid Control Software on Cloud Computing Architectures

Ghaleb Abdulla
Lawrence Livermore National Laboratory

Summary

Because of its low cost, flexible and redundant architecture, and fast response time, cloud computing is an attractive choice for implementing the new smart grid computing architecture. Cloud computing is the provision of computing resources on demand via a network. For an average user, services such as Facebook or Flickr are examples of cloud computing resources. Google's PowerMeter or Microsoft's smart home are cloud computing services that allow users to upload, analyze, and use data to make decisions and monitor energy consumption.

This paper examines the computational requirements for building successful smart electric grid control software that runs on cloud computing architectures and proposes six research elements that need to be addressed to achieve the stated goal. The authors identify several deficiencies in current cloud computing architectures and a general lack of support for addressing specific smart grid computational requirements, such as real-time data acquisition and analysis, consistency, privacy, and security. The authors' proposed research elements, however, would bridge the gap and provide a usable cloud computing architecture that will satisfy smart electric grid requirements.

Discussion

James Nutaro, the paper discussant, brings up an important issue related to the design of large engineered systems such as the smart grid. In order to avoid the prohibitive cost of hardware and software system testing, large-scale simulations are built to test deployment scenarios and outlier cases that could strain the system. Nutaro argues that the energy industry will be dealing with a challenging engineering project that requires simulations to design, build, and deploy efficient smart grid hardware and software. In order for cloud computing to support the needed simulations, advances in the areas of data protection and support for scalable, consistency guaranteed, fault-tolerant services

must be achieved. An alternative option would be to investigate new approaches for simulation methods that work on the current cloud computing infrastructure.

The comments during the discussion session can be classified into three categories:

1. General comments about the smart grid implementation in general, but useful for this paper:
    a. Smart grid is expected to push the envelope in several areas of data-centric computation, including storage, data integration, simulations for market pricing, and planning of generation and transmission lines deployment.
    b. Security will be a challenge for the smart grid due to the new peer-to-peer wireless communication for data collection from the advanced meters. Wireless communication provides opportunities for hackers and criminals to tap into the system and try to steal energy or information or bring the system down.
    c. Data security, privacy, and integrity present conflicting requirements. For example, data security might require data sharing, while privacy requires minimum data sharing.
    d. In some cases, smart grid applications will be embarrassingly parallel; in other cases, tightly coupled simulations require high-speed data communication.
    e. The smart grid will have to integrate user response to help manage the system. User response can be captured by monitoring the use of smart appliances or by direct user response reacting to peak load and pricing schedules. This is a potential use case for cloud computing, where computing agents for the smart appliances communicate with the cloud and use the services without knowing where and how these services are running.
    f. One participant from an ISO organization argued that cost of the computing infrastructure is not an issue. Reduction of production cost and improving the operator's efficiency will outweigh the cost of the computing infrastructure.
2. A discussion about the suitability of cloud computing for the smart grid versus dedicated hardware that can satisfy the strict computing and data handling requirements:
    a. Energy use case is important in deciding if the cloud or dedicated computing infrastructure should be used.
    b. For tightly coupled applications, the cloud might not be cost effective. A dedicated HPC solution would be better in these cases. The authors response was that new large-scale problems defy intuition, and this new class of problems needs special attention. Virtualization on the cloud might be a way to help dedicate special hardware configuration to solve large-scale problems (with highly coupled and decoupled systems).
    c. A concern was raised that we are moving to the solution space without defining some real use cases from the smart grid domain.

d. Can we trust the control system to run over the cloud? This is an example where a dedicated hardware might be needed, however, the paper is trying to motivate the research to address such gaps in current cloud computing and to support this type of use case.
e. One participant thought the cloud might be a good medium to test new technologies and pilot projects and quantify their impact fairly quickly.

3. The possibility of approaching the problem in stages, thus allowing the use of the cloud to evolve over time. For example, UCLA has a smart grid project, and certain parts of the project are running on the cloud.

Conclusions

The paper discusses an important and timely topic. The traditional approach to high performance computing is being reexamined. The community has realized that there are several large-scale applications that have conflicting requirements with respect to memory, storage, and communication usage. A memory-based and embarrassingly parallel application will not require huge secondary storage or high-speed networking. The cost of customized hardware is becoming less attractive compared to cloud computing, which can provide cost-effective solutions.

Participants agreed that this is an interesting area of research and we should explore the possibility of using the cloud because start up cost is minimal compared to acquiring dedicated data centers for the ISOs or Utilities. Naturally, there were concerns raised about how to conduct this research in an effective way. These concerns validate the authors' claim that this is an interesting and worthwhile research problem.

To make the cloud a medium for computing, attention must be paid to areas that currently prohibit the cloud from meeting the requirements of the smart grid. Advances in data protection and support for scalable, consistency guaranteed, fault-tolerant services must be made, and new simulation approaches that can run on the cloud must be investigated.

White Paper

Coupled Optimization Models for Planning and Operation of Power Systems on Multiple Scales

Michael Ferris
University of Wisconsin

Abstract

- Decision processes are predominantly hierarchical. Models to support such decision processes should also be layered or hierachical.
- Coupling collections of (sub)-models with well defined (information sharing) interfaces facilitates:
  - appropriate detail and consistency of sub-model formulation (each of which may be very large scale, of different types (mixed integer, semidefinite, nonlinear, variational, etc) with different properties (linear, convex, discrete, smooth, etc))
  - ability for individual subproblem solution verification and engagement of decision makers
  - ability to treat uncertainty by stochastic and robust optimization at submodel level and with evolving resolution
  - ability to solve submodels to global optimality (by exploiting size, structure and model format specificity)

  (A monster model that mixes several modeling formats loses its ability to exploit the underlying structure and provide guarantees on solution quality)
- Developing interfaces and exploiting hierarchical structure using computationally tractable algorithms will provide overall solution speed, understanding of localized effects, and value for the coupling of the system.

1 Problem Hierarchies and Timescales

Technological and economic trends imply significant growth in our nation's reliance on the power grid in the coming decades; well-accepted estimates cite 35% growth in electricity demand over the next 20 years [60]. Planning and operating the Next

Generation Electric Grid involves decisions ranging from time scales of perhaps 15 years, for major grid expansion, to time scales of 5-minute markets, and must also account for phenomena at time scales down to fractions of a second. A representation of the decision process over timescales of interest is shown in Figure 1.

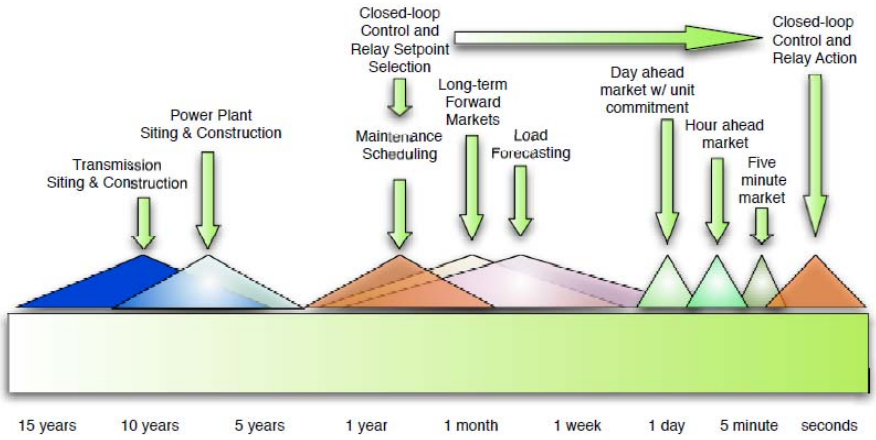


Figure 1: Representative decision-making timescales in electric power systems

What makes this setting particularly interesting is that behaviors at very fast time scales (e.g., requirements for grid resilience against cascading failures) potentially impose constraints on longer time scale decisions, such as maintenance scheduling and grid expansion. We argue here against building a single "monster model" that tries to capture all these scales, but propose using a collection of coupled or layered models for both planning and operation, interfacing via information/solution sharing over multiple time scales and layers of decision making. Such approaches have been successful in other application domains [18].

In addition to the multiple time scales in the decision process, the problem is confounded by uncertainties in estimates and structural makeup of the system. For example, plug hybrid electric vehicles are a visible technology that could dramatically alter the patterns, nature, and quantity of U.S. electricity use, and yet the ultimate market penetration of such technology is highly uncertain. Similarly, future grid penetration for non-traditional energy sources such as wind and solar, and for carbon-sequestration-equipped coal plants, also remains highly uncertain. These structural uncertainties present profound challenges to decision methodologies, and to the optimization tools that inform them. While traditional optimization approaches might seek to build a large-scale model that combines all instances together, such approaches are impractical as the size and ranges of the spatial and temporal scales expand, let alone treating the uncertainties that are inherently present in these decision problem settings.

As one moves between time scales, some of that uncertainty gets resolved, and some new uncertainties become relevant. The decision problems need to capture that uncertainty and allow a decision maker the flexibility to structurally change the system to the new environment. All encompassing models are typically not nimble enough to facilitate adaptation of the decisions as the real process evolves (both structurally and data-wise). Thus, our thesis is not simply about solution speed, but hinges on the added value that arises from modeling and solution in a structured (and better scaled and theoretically richer) setting.

The decision timeline of Figure 1 is intended to highlight the severe challenges the electric power environment presents. As an example of coupling of decisions across time scales, consider decisions related to the siting of major interstate transmission lines. These require economic forecasts, supply and demand forecasts, and an interplay between political and engineering concerns. Typically, relatively few possible choices are available – not only due to engineering or even economic constraints – but arising from public and political concerns that are often hard to justify rationally, but severely limit the possible layouts. Models that demonstrate the benefits and drawbacks of a particular siting decision at an aggregate level are of critical importance for informing discussions and decision makers. The key issue is to facilitate appropriate aggregations (or summarizations of details irrelevant to the decision at hand) that enable a quick, even interactive, and thorough exploration of the actual decision space. It is, and will remain, a significant challenge to identify and manage the interface between a given model and the other models that are connected – from a conceptual, modeling and computational viewpoint.

Note that transmission expansion decisions influence the capital investment decisions made by generating companies, again at a 1-10 year time scale. Much of the same data used for transmission expansion is pertinent to these models, but the decisions are made by independent agents without overall system control, so different types of models (game theoretic for example) more readily capture the decision process here. Specific decision models are typically governed by a overriding principle and can be formulated using the most appropriate modeling tools. The monster model is more likely to be a conglomeration of multiple principles, and becomes unmanageable, intractable and hard to understand the driving issues.

New generation capabilities subsequently affect bids into the power market which are then balanced using economic and reliability objectives on a day-ahead or 5-minute time scale. At this level, models are needed for electric pricing and market control to determine which units are to be deployed and at what price and quantity, accounting for the uncertainties in new forms of energy provision such as wind and solar. Such planning, deployment and commitment of specific resources must be carried out to ensure both operation reserves are sufficient, and to provide robust solutions for these choices that ensure security of the overall system (when confronted with the vast number of uncertainties that can confound the efficient operation of the electric grid).

Note that long term planners will not be able to accurately forecast all the structural (and potentially disruptive) changes to the system, and thus wind speed and weather patterns at fine scales are largely irrelevant – efficient sampling and (automated) information aggregation are key to allow informed decision making at widely different time scales.

Finally, power grid dynamics are operating at the millisecond to minutes time scales and involve decisions for settings of protective relays that remove lines and generators from service when operating thresholds are exceeded to guard against cascading failures. At this level, efficient nonlinear optimization must be carried out to match the varying demand for electricity with the ever increasing and uncertain supply of energy, without interruptions or catastrophic cascading failures of the system. While the underlying question may well be "Is there a better choice for the transmission line expansion to reduce the probability of a major blackout?", it is contended here that the additional knowledge gained from understanding the effects of one decision upon another in a structured fashion will facilitate better management and operation of the system when it is built and provide understanding and information to the operators as to the consequences of their decisions.

In addition to this coupling across time scales, one has the challenge of structural differences amongst classes of decision makers and their goals. At the longest time frame, it is often the Independent System Operator (ISO), in collaboration with Regional Transmission Organizations (RTO) and regulatory agencies, that are charged with the transmission design and siting decisions. These decisions are in the hands of regulated monopolies and their regulator. From the next longest time frame through the middle time frame, the decisions are dominated by capital investment and market decisions made by for-profit, competitive generation owners. At the shortest time frames, key decisions fall back into the hands of the Independent System Operator, the entity typically charged with balancing markets at the shortest time scale (e.g., day-ahead to 5-minute ahead), and with making any out-of-market corrections to maintain reliable operation in real time.

Each of these problems involves coordinating a large number of decision making agents in an uncertain environment. The computational needs for such solutions are immense and will require both modeling sophistication and decompostion methodology to exploit problem structure, and a large array of computing devices – whose power is seamlessly provided and available to critical decision makers (not just optimization or computational science experts) – to process resulting subproblems. Subsets of these subproblems may be solved repeatedly when resulting information in their interfaces changes. Model updates can then be coupled to evolving information flow. Quite apart from the efficiency gains achieved, the smaller coupled modelds are more easily verified by their owners, and otherwise hidden deficiencies of a monster model formulation are quickly detected and fixed.

Traditional optimization approaches are no longer effective in solving the practical, large scale, complex problems that require robust answers in such domains. While the study of linear programming, convex optimization, mixed integer and stochastic programming have in themselves led to significant advances in our abilities to solve large scale instances of these problems, typical application problems such as those outlined above require a sophisticated coupling of a number of these approaches with specific domain knowledge and expertize to generate solutions in a timely manner that are robust to uncertainties in an operating environment and in the data that feeds the model. Rather than attempting to model all these features together, we propose a methodology that utilizes layering and information sharing interfaces between collections of models, that allows decisions to be made using appropriately scaled problems, each of which approximates external features by aggregate variables and constraints. It could be argued that by using a collection of coupled models, we are leaving some optimization possibilities "on the table". Clearly, poorly defined interfaces will have this issue. The challenge for modelers and algorithms is to define these interfaces correctly, manage them automatically, understand the hierarchy of decision makers and match this to the model, thereby facilitating solution of the overall system by processing of (modified) problems at each level.

In short, there is clearly a need for optimization tools that effectively inform and integrate decisions across widely separated time scales, by different agents who have differing individual objectives, in the presence of uncertainty.

2 Motivating Problems

The purpose of the electric power industry is to generate and transport electric energy to consumers [56]. At time frames beyond those of electromechanical transients (i.e. beyond perhaps, 10's of seconds), the core of almost all power system representations is a set of equilibrium equations known as the power flow model. This set of nonlinear equations relates bus (nodal) voltages to the flow of active and reactive power through the network and to power injections into the network. With specified load (consumer) active and reactive powers, generator (supplier) active power injections and voltage magnitude, the power flow equations may be solved to determine network power flows, load bus voltages, and generator reactive powers. Current research is still ongoing to reliably solve these equations; approaches involve Newton based methods and techniques from semidefinite programming.

At the next level of sophistication, an Optimal Power Flow (OPF) can be used to determine least cost generation dispatch, subject to physical grid constraints such as power flow equations, power line flow limits, generator active and reactive power limits, and bus voltage limits. Typically the problem is solved by the ISO, and is characterized by a number of different methods that generator firm's can make bids to supply

electricity. For simplicity here, we assume that bids are characterized by a decision variable α, resulting in the following optimization problem:

$$\begin{array}{lll} \text{OPF}(\alpha): & \min_q & \text{energy dispatch cost } (q,\alpha) \\ & \text{s.t.} & \text{conservation of power flow at nodes} \\ & & \text{Kirchoff's voltage law, and simple bound constraints} \end{array}$$

Note that since α are (given) price bids, this problem is a parametric optimization for dispatch quantities q. We assume this problem has a unique solution for each α for ease of exposition.

Each generator firm i has to determine its bid α_i. Assuming no generator has market power (perhaps an unreasonable assumption), the problem faced by firm i is

$$\begin{array}{lll} \text{Bid}(\bar{\alpha}_{-i}): & \max_{\alpha_i,q,p} & \text{firm } i\text{'s profit } (\alpha_i,q,p) \\ & \text{s.t.} & 0 \le \alpha_i \le \hat{\alpha}_i \\ & & q \text{ solves OPF}(\alpha_i,\bar{\alpha}_{-i}) \end{array}$$

where the objective function involves the multiplier p determined from the OPF problem. This multiplier is not exposed to the decision maker. To overcome this issue, we can replace the lower level optimization problem by its first order (KKT) conditions and thus expose the multipliers directly to the upper level optimization problem:

$$\begin{array}{lll} \text{Bid}(\bar{\alpha}_{-i}): & \max_{\alpha_i,q,p} & \text{firm } i\text{'s profit } (\alpha_i,q,p) \\ & \text{s.t.} & 0 \le \alpha_i \le \hat{\alpha}_i \\ & & q,p \text{ solves } \text{KKT}(\text{OPF}(\alpha_i,\bar{\alpha}_{-i})) \end{array}$$

This process takes a bilevel program and converts it to a mathematical program with complementarity constraints (MPCC) since the KKT conditions form what are called complementarity constraints. We outline in the sequel methods to write down and solve problems of this form, but note that they are computationally difficult, and theoretically the MPCC is hard due to the lack of a constraint qualification. It may even be the case that this transformation is incorrect: the KKT may not be necessary and sufficient for global optimality of the lower level problem.

This problem is a single firm's problem. Adventurous modelers require further conditions, in that the firms collectively should have no incentive to change their bids in equilibrium: $(\bar{\alpha}_1,\bar{\alpha}_2,\ldots,\bar{\alpha}_m)$ is an equilibrium if

$$\bar{\alpha}_i \text{ solves Bid}(\bar{\alpha}_{-i}),\ \forall i.$$

This is an example of a (Nonlinear) Nash Equilibrium where each player solves an MPCC. It is known that such a Nash Equilibrium is PPAD-complete [17, 19]. While

complexity results of this nature lead to an appreciation of the extreme difficulty of the underlying problem, it is clear that such problems must be solved (repeatedly) for effective operation of the power system.

Unfortunately, in practice, a Security Constrained Optimal Power Flow (SCOPF) adds the additional constraint that the solution for powers and voltages must remain within limits for a user-specified set of contingencies (scenarios) [57, 70]. To some extent, this is a simplification made to the problem to gain tractability. Even so, such problems are currently beyond the state of the art for solution methodologies. We outline an extended mathematical programming (EMP) framework that allows such problems to be written down. We firmly believe that the underlying structure in these models will be necessary to exploit for any realistic solution method to be successful.

The constraints in the OPF and SCOPF problems make them more difficult to solve [80], and some programs use simplified models to quickly "solve" these equations. A common simple model is the so-called DC power flow, which is a simple linearized form of the true power flow equations. The industry uses this form extensively. However, "proxy constraints" are often added to the formulations to recapture effects that linearization (or approximations) lose. This is fraught with danger and possibilities for exploitation (of the difference in the approximate model and the nonlinear physics) when such constraints are used for pricing in markets for reactive power, for example. Nonlinear models are necessary and can and should be reliably solved at appropriate levels of detail.

Another example that motivates our work is the notion of transmission line switching [31, 41]. In this setting, an optimization problem of the form:

$\min_{g,f,\theta}$	$c^T g$	generation cost
s.t.	$g - d = Af, f = BA^T\theta$	A is node-arc incidence
	$\overline{\theta}_L \le \theta \le \overline{\theta}_U$	bus angle constraints
	$\overline{g}_L \le g \le \overline{g}_U$	generator capacities
	$\overline{f}_L \le f \le \overline{f}_U$	transmission capacities

can be solved to determine the flows and generation with a simplified DC power flow model. Transmission switching is a design philosophy that allows a subset of the lines to be opened to improve the dispatch cost. The additional discrete choice is whether the line *i* is open or closed and can be modeled using the following disjunction:

$$
\begin{array}{lll}
\min_{g,f,\theta} & c^T g & \\
\text{s.t.} & g - d = Af & \\
 & \overline{\theta}_L \le \theta \le \overline{\theta}_U & \\
 & \overline{g}_L \le g \le \overline{g}_U & \\
\text{either} & f_i = (BA^T\theta)_i, \overline{f}_{L,i} \le f_i \le \overline{f}_{U,i} & \text{if } i \text{ closed} \\
\text{or} & f_i = 0 & \text{if } i \text{ open}
\end{array}
$$

This disjunction is not a typical constraint for an optimization problem, but can be directly modeled in EMP. The framework allows automatic problem reformulations - in the above case, this can generate mixed integer programming problems, or indeed nonlinear mixed integer programs when the linearized DC model is replaced by the full AC model.

The final example concerns the transmission line expansion outlined in the introduction [48]. We suggest considering a hierarchical approach to this problem formulated as follows. If we let x represent the transmission line expansion decision and assume the RTO can postulate a (discrete) distribution of future demand scenarios (at the decade level scale), then the RTO problem is:

$$\min_{x \in X} \sum_{\omega} \pi_\omega \sum_{i \in N} d_i^\omega p_i^\omega(x)$$

where ω runs over the scenarios, π are the probabilities, and d_i^ω is the resulting demand in such a scenario at a given node in the network. The constraints $x \in X$ reflect budgetary and other constraints on the RTO's decision, and the function $p_i^\omega(x)$ is a response price in the given scenario to the expansion by x. Clearly, the key to solving this problem is to generate a good approximation to this response price since this is the interface to the lower levels of the hierarchy. We believe a lower level equilibrium model involving both generator firms and OPF problem solution in every scenario is one way to generate such a response. Optimization techniques based on derivative free methodology, or noisy function optimization may be a practical way to solve the RTO problem, requesting evaluations of this response price function. Alternatively, automatic differentiation could play a role in generating derivatives for the response price function, but that may require techniques to deal with its inherent nonsmoothness.

As outlined above, the transmission line expansion is likely to foster generator expansion. For each firm f, we denote the generator expansion by y_f and propose that this will be determined by an optimization principle:

$$\min_{y_f \in Y_f} \sum_{\omega} \pi_\omega \sum_{j \in F_f} c_j(y_j, q_j^\omega)$$

Here F_f denotes the generators in firm f's portfolio, and Y_f represents budgetary and other constraints faced by the generator firm. Each firm thus expends its budget to minimize the expected cost of supply. Note that q_j^ω is a parameter to this problem - the actual dispatch is determined by a scenario dependent OPF problem:

$$\forall \omega \quad \min_{z,\theta,q} \sum_{j \in F_f} c_j(y_j, q_j)$$

which is subject to flow balance constraints, line data constraints, line capacity constraints, generator capacity constraints and regulatory constraints. The multiplier on the flow balance constraints is $p_i^\omega(x)$. This problem may involve integer variables as well to model switching or commitment features of the problem, but that could raise issues of the integrity of the multipliers.

Note that the collection of all these optimization models (generator expansion and scenario dependent OPF) forms an equilibrium problem, once each optimization model is replaced by its KKT conditions. The specific interface between them is defined by the variables y and q. The equilibrium problem determines all the variables in all the models at one time, whereas the models assume price taking behavior or knowledge of generator expansions in parameteric form. (In fact, this is an example of an embedded complementarity system, details of which follow in the sequel.) In this case, the equilibrium problem may be replaced by a large scale optimization problem. This fact is useful in formally proving convergence of a decomposition algorithm that iteratively solves the small optimization problems and updates the linking variables in a Jacobi sense. All of this then generates the response price $p_i^\omega(x)$ for a given transmission expansion x in a computationally tractable way and allows extension to power system models at the regional or national scale. Other models that can be captured by these concepts include the following: [47, 46, 4, 61].

3 Extended Mathematical Programs

We believe that the design, operation and ennhancement of the Next Generation Electric Grid will rely critically on tools and algorithms from optimization. Accessing these optimization solvers, and many of the other algorithms that have been developed over the past three decades has been made easier by the advent of modeling languages. A modeling language [11, 33] provides a natural, convenient way to represent mathematical programs and provides an interface between a given model and multiple different solvers for its solution. The many advantages of using a modeling language are well known. They typically have efficient automatic procedures to handle vast amounts of data, take advantage of the numerous options for solvers and model types, and can quickly generate a large number of models. For this reason, and the fact that they eliminate many errors that occur without automation, modeling languages are heavily

used in practical applications. Although we will use GAMS [13] in our descriptions here, much of what will be said could as well be applied to other algebra based modeling systems like AIMMS [10], AMPL [34], MOSEL, MPL [55] and OPL [75] .

While much progress has also been made in developing new modeling paradigms (such as stochastic and robust programming, mixed integer nonlinear optimization, second order cone programming, and optimization of noisy functions), the ability for application experts to utilize these advances from within modeling systems has remained limited. The purpose of this work is to extend the classical problem from the traditional optimization model:

$$\min_x f(x) \text{ s.t. } g(x) \leq 0, h(x) = 0, \tag{1}$$

where f, g and h are assumed sufficiently smooth, to a more general format that allows new constraint types and problem features to be specified precisely. The extended mathematical programming (EMP) framework exists to provide these same benefits for problems that fall outside the classical framework [26]. A high-level description of these models in an algebraic modeling language, along with tools to automatically create the different realizations or extensions possible, pass them on to the appropriate solvers, and interpret the results in the context of the original model, makes it possible to model more easily, to conduct experiments with formulations otherwise too time-consuming to consider, and to avoid errors that can make results meaningless or worse.

We believe that further advancements in the application of optimization to electricity grid problems can be best achieved via identification of specific problem structures within planning and operational models, coupled with automatic reformulation techniques that lead to problems that are theoretically better defined and more ameable to rigorous computation. The ability to describe such structures in an application domain context will have benefits on several levels. Firstly, we think this will make the modelers task easier, in that the model can be described more naturally and (for example) soft or probabilistic constraints can be expressed explicitly. Secondly, if an algorithm is given additional structure, it may be able to exploit that in an effective computational manner; indeed, the availability of such structures to a solver may well foster the generation of new features to existing solvers or drive the development of new classes of algorithms. Specific structures that we believe are relevant to this application domain include mathematical programs with equilibrium constraints, second order cone programs (that facilitate the use of "robust optimization" principles), semidefinite programming, bilevel and hierarchical programs, extended nonlinear programs (with richer classes of penalty functions) and embedded optimization models. The EMP framework provides an extensible way to utilize such features.

Some extensions of the traditional format have already been incorporated into modeling systems. There is support for integer, semiinteger, and semicontinuous variables, and some limited support for logical constructs including special ordered sets (SOS). GAMS ,

AMPL and AIMMS have support for complementarity constraints [29, 28], and there are some extensions that allow the formulation of second-order cone programs within GAMS. AMPL has specific syntax to model piecewise linear functions. Much of this development is tailored to particular constructs within a model. We describe the development of the more general EMP annotation schemes that allow extended mathematical programs to be written clearly and succinctly.

3.1 Complementarity Problems

The EMP framework allows annotation to existing functions and variables within a model. We begin with the example of complementarity, which in its simplest form, is the relationship between nonnegative variables with the additional constraint that at least one must be zero. A variety of models in electricity markets use complementarity at their core, including [43, 44, 45, 54, 58, 69, 81, 84, 83] .

A first simple example are the necessary and sufficient optimality conditions for the linear program:

$$\begin{array}{ll} \min_x \; c^T x & \\ s.t. \quad Ax \geq b, x \geq 0 & \end{array} \tag{2}$$

which state that x and some λ satisfy the complementarity relationships:

$$\begin{array}{lll} 0 \leq c - A^T\lambda & \perp & x \geq 0 \\ 0 \leq Ax - b & \perp & \lambda \geq 0 \end{array} \tag{3}$$

Here, the "⊥" sign signifies (for example) that in addition to the constraints $0 \leq Ax - b$ and $\lambda \geq 0$, each of the products $(Ax-b)_i\lambda_i$ is constrained to be zero. An equivalent viewpoint is that either $(Ax-b)_i = 0$ or $\lambda_i = 0$, a disjunction. Within GAMS, these constraints can be modeled simply as:

```
positive variables lambda, x;
model complp / defd.x, defp.lambda /;
```

where defp and defd are the equations that define general primal and dual feasibility constraints $(Ax \geq b, c \geq A^T\lambda)$ respectively.

Complementarity problems do not have to arise as the optimality conditions of a linear program; the optimality conditions of the nonlinear program (1) constitute the following MCP:

$$
\begin{aligned}
0 &= \nabla f(x) + \lambda^T \nabla g(x) + \mu^T \nabla h(x) \quad &\perp \quad & x \text{ free} \\
0 &\leq -g(x) \quad &\perp \quad & \lambda \geq 0 \\
0 &= -h(x) \quad &\perp \quad & \mu \text{ free.}
\end{aligned} \tag{4}
$$

Many examples are no longer simply the optimality conditions of an optimization problem. The paper [30] catalogues a number of other applications both in engineering and economics that can be written in a similar format.

It should be noted that robust large scale solvers exist for such problems; see [29] for example, where a description is given of the PATH solver.

3.2 Disjunctive Constraints

A simple example to highlight disjunctions is the notion of an ordering of tasks, namely that either job *i* comes before job *j* or the converse. Such a disjunction can be specified using an annotation:

```
disjuncton * seq(i,j) else seq(j,i)
```

In such an example, one can implement a Big-M method, employ indicator variables or constraints, or utilize a convex hull reformulation.

In fact, there is a growing literature on reformulations of mixed integer nonlinear programs that describe new convex hull descriptions of structured constraint sets. This work includes disjunctive cutting planes [71], Gomory cuts [16] and perspective cuts and reformulations [35, 39].

More complicated (nonlinear) examples make the utility of this approach clearer. The design of a multiproduct batch plan with intermediate storage described in [77] and a synthesis problem involving 8 processes from [74] are included in the EMP model library. As a final example, the gasoline emission model outlined in [36] is precisely in the form that could exploit the features of EMP related to (nonlinear) disjunctive programming. Finally, the work by Grossmann and colleagues on generalized disjunctive programming [74, 77, 78] involves both nonlinear equations and optimization primitives coupled with pure logic relations; this has been used extensively in the synthesis and design of process networks.

3.3 Mathematical Programs with Complementarity Constraints

A mathematical program with complementarity constraints embeds a parametric MCP into the constraint set of a nonlinear program as indicated in the following problem:

$$\min_{x\in\Re^n, y\in\Re^m} f(x,y) \tag{5}$$

$$\text{s.t.} \quad g(x,y) \le 0 \tag{6}$$

$$0 \le y \perp h(x,y) \ge 0. \tag{7}$$

The objective function (5) needs no further description, except to state that the solution techniques we are intending to apply require that f (g and h) are at least once differentiable, and for many modern solvers twice differentiable.

The constraints that are of interest here are the complementarity constraints (7). Essentially, these are parametric constraints (parameterized by x) on the variable y, and encode the structure that y is a solution to the nonlinear complementarity problem defined by $h(x,\cdot)$. Within the GAMS modeling system, this can be written simply and directly as:

```
model mpecmod / deff, defg, defh.y /;
option mpec=nlpec;
solve mpecmod using mpec minimizing obj;
```

Here it is assumed that the objective (5) is defined in the equation deff, the general constraints (6) are defined in defg and the function h is described by defh. The complementarity relationship is defined by the bounds on y and the orthogonality relationship shown in the model declaration using ".". AMPL provides a slightly different but equivalent syntax for this, see [28]. The problem is frequently called a mathematical program with complementarity constraints (MPCC). Section 2 provided a specific example.

Some solvers can process complementarity constraints explicitly. In many cases, this is achieved by a reformulation of the constraints (7) into the classical nonlinear programming form given as (1). The paper [37] outlines a variety of ways to carry this out, all of which have been encoded in a solver package called NLPEC. Similar strategies are outlined in section 3 of [6]. While there are large numbers of different reformulations possible, the following parametric approach, coupled with the use of the nonlinear programming solver CONOPT or SNOPT, has proven effective in a large number of applications:

$$\begin{aligned} \min_{x\in\Re^n, y\in\Re^m, s\in\Re^m} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) \le 0 \\ & s = h(x,y) \\ & y \ge 0, s \ge 0 \\ & y_i s_i \le \mu, \ i = 1,....,m. \end{aligned}$$

Note that a series of approximate problems are produced, parameterized by μ>0; each of these approximate problems have stronger theoretical properties than the problem with μ=0 [62]. A solution procedure whereby μ is successively reduced can be implemented as a simple option file to NLPEC, and this has proven very effective. Further details can be found in the NLPEC documentation [37]. The approach has been used to effectively optimize the rig in a sailboat design [79] and to solve a variety of distillation optimization problems [6]. A key point is that other solution methodology may work better with different reformulations – this is the domain of the algorithmic developer and should remain decoupled from the model description. NLPEC is one way to facilitate this.

It is also possible to generalize the above complementarity condition to a mixed complementarity condition; details can be found in [27]. Underlying the NLPEC "solver package" is an automatic conversion of the original problem into a standard nonlinear program which is carried out at a scalar model level. The technology to perform this conversion forms the core of the codes that we use to implement the model extensions herein.

3.4 Variational Inequalities

A variational inequality VI(F,X) is to find $x \in X$:

$$F(x)^T (z - x) \geq 0, \text{ for all } z \in X.$$

Here X is a closed (frequently assumed convex) set, defined for example as

$$X = \{x \mid x \geq 0, h(x) \leq 0\}. \tag{8}$$

Note that the first-order (minimum principle) conditions of a nonlinear program

$$\min_{z \in X} f(z)$$

are precisely of this form with $F(x)=\nabla f(x)$. For a concrete example, note that these conditions are necessary and sufficient for the optimality of a linear programming problem: solving the linear program (2) is equivalent to solving the variational inequality given by

$$F(x) = c, \quad X = \{x \mid Ax \geq b, x \geq 0\}. \tag{9}$$

In this case, F is simply a constant function. While there are a large number of instances of the problem that arise from optimization applications, there are many cases where F is not the gradient of any function f. For example, asymmetric traffic equilibrium problems have this format, where the asymmetry arises for example due to different costs associated with left or right hand turns. A complete treatment of the theory and algorithms in this domain can be found in [25] .

Variational inequalities are intimately connected with the concept of a normal cone to a set S, for which a number of authors have provided a rich calculus. Instead of overloading a reader with more notation, however, we simply refer to the seminal work in this area, [67]. While the theoretical development of this area is very rich, the practical application has been somewhat limited. The notable exception to this is in traffic analysis, see for example [40] .

It is well known that such problems can be reformulated as complementarity problems when the set X has the representation (8) by introducing multipliers λ on the constraints h:

$$\begin{array}{lll} 0 \le F(x) + \lambda^T \nabla h(x) & \perp & x \ge 0 \\ 0 \le -h(x) & \perp & \lambda \ge 0. \end{array}$$

If X has a different representation, this construction would be modified appropriately. In the linear programming example (9), these conditions are precisely those already given as (3) .

When X is the nonnegative orthant, the VI is just an alternative way to state a complementarity problem. However, when X is a more general set, it may be possible to treat it differently than simply introducing multipliers, see [15] for example. In particular, when X is a polyhedral set, algorithms may wish to generate iterates via projection onto X.

Bimatrix games can also be formulated as a variational inequality. In this setting, two players have I and J pure strategies, and p and q (the strategy probabilities) belong to unit simplex $\triangle_I$ and $\triangle_J$ respectively. Payoff matrices $A \in R^{J \times I}$ and $B \in R^{I \times J}$ are defined, where $A_{j,i}$ is the profit received by the first player if strategy i is selected by the first player and j by the second. The expected profit for the first and the second players are then $q^T A p$ and $p^T B q$ respectively. A Nash equilibrium is reached by the pair of strategies (p^*, q^*) if and only if

$$p^* \in \arg\ \min_{p \in \triangle_I} \langle A q^*, p \rangle \ \text{ and } \ q^* \in \arg\ \min_{q \in \triangle_J} \langle B^T p^*, q \rangle$$

Letting x be the combined probability vector (p,q), the (coupled) optimality conditions for the above problems constitute the variational inequality:

$$F\left(\begin{bmatrix} p \\ q \end{bmatrix}\right) = \begin{bmatrix} 0 & A \\ B^T & 0 \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix}, \ X = \triangle_I \times \triangle_J .$$

Algorithms to solve this problem can exploit the fact that X is a compact set. The thesis [50] contains Newton based algorithms to solve these problems effectively.

3.5 Bilevel Programs

Mathematical programs with optimization problems in their constraints have a long history in operations research including [12, 32, 5]. Hierarchical optimization has recently become important in a number of different applications and new codes are being developed that exploit this structure, at least for simple hierarchies, and attempt to define and implement algorithms for their solution.

The simplest case is that of bilevel programming, where an upper level problem depends on the solution of a lower level optimization. For example:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) \le 0, \\ & y \text{ solves } \min_y v(x,y) \text{ s.t. } h(x,y) \le 0. \end{aligned}$$

This problem can be reformulated as an MPCC by replacing the lower level optimization problem by its optimality conditions:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) \le 0, \\ & 0 = \nabla_y v(x,y) + \lambda^T \nabla_y h(x,y) \perp x \text{ free} \\ & 0 \le -h(x,y) \perp \lambda \ge 0. \end{aligned}$$

This approach then allows such problems to be solved using the NLPEC code, for example. However, there are several possible deficiencies that should be noted. Firstly, the optimality conditions encompassed in the complementarity constraints may not have a solution, or the solution may only be necessary (and not sufficient) for optimality. Secondly, the MPCC solver may only find local solutions to the problem. The quest for practical optimality conditions and robust global solvers remains an active area of research. Importantly, the EMP tool will provide the underlying structure of the model to a solver if these advances determine appropriate ways to exploit this.

We can model this bilevel program in GAMS by:

```
model bilev /deff,defg,defv,defh/;
solve bilev using emp min f;
```

along with some extra annotations to a subset of the model defining equations. Specifically, within an "empinfo" file we state that the lower level problem involves the objective v which is to be minimized subject to the constraints specified in defv and defh.

```
bilevel x
min v defv defh
```

Note that the variables x are declared to be variables of the upper level problem and that defg will be an upper level constraint. The specific syntax is described in [38]. Having written the problem in this way, the MPCC is generated automatically, and passed on to a solver. In the case where that solver is NLPEC, a further reformulation of the model is carried out to convert the MPCC into an equivalent NLP or a parametric sequence of NLP's. A key extension to the bilevel format allows multiple lower level problems to be specified within the bilevel format.

3.6 Embedded Complementarity Systems

A different type of embedded optimization model that arises frequently in applications is:

$$\begin{array}{lll} \min_x & f(x,y) & \\ \text{s.t.} & g(x,y) \leq 0 & (\perp \lambda \leq 0) \\ H(x,y,\lambda) & = 0 & (\perp y \text{ free}) \end{array}$$

Note the difference here: the optimization problem is over the variable x, and is parameterized by the variable y. The choice of y is fixed by the (auxiliary) complementarity relationships depicted here by H. Note that the "H" equations are not part of the optimization problem, but are essentially auxiliary constraints to tie down remaining variables in the model.

Within GAMS, this is modeled as:

```
model ecp /deff,defg,defH/;
solve ecp using emp;
```

Again, so this model can be processed correctly as an EMP, the modeler provides additional annotations to the model defining equations in an "empinfo" file, namely that the function H that is defined in defH is complementary to the variable y (and hence the variable y is a parameter to the optimization problem), and furthermore that the dual variable associated with the equation defg in the optimization problem is one and the same as the variable λ used to define H:

```
min f x deff defg
vifunc defH y
dualvar lambda defg
```

Armed with this additional information, the EMP tool automatically creates the following MCP:

$$
\begin{aligned}
0 &= \nabla_x \mathcal{L}(x, y, \lambda) && \perp \quad x \text{ free} \\
0 &\geq -\nabla_\lambda \mathcal{L}(x, y, \lambda) && \perp \quad \lambda \leq 0 \\
0 &= H(x, y, \lambda) && \perp \quad y \text{ free},
\end{aligned}
$$

where the Lagrangian is defined as:

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda^T g(x, y).$$

Perhaps the most popular use of this formulation is where competition is allowed between agents. A standard method to deal with such cases is via the concept of Nash Games. In this setting x^* is a Nash Equilibrium if

$$x_i^* \in \arg\min_{x_i \in X_i} \ell_i(x_i, x_{-i}^*, q), \forall i \in \mathcal{I}$$

where x_{-i} are other players decisions and the quantities q are given exogenously, or via complementarity:

$$0 \leq H(x, q) \quad \perp \quad q \geq 0.$$

This mechanism is extremely popular in economics, and Nash famously won the Nobel Prize for his contributions to this literature.

This format is again an EMP, more general than the example given above in two respects. Firstly, there is more than one optimization problem specified in the embedded complementarity system. Secondly, the parameters in each optimization problem consist of two types. Firstly, there are the variables q that are tied down by the auxiliary complementarity condition and hence are treated as parameters by the ith Nash player. Also there are the variables x_{-i} that are treated as parameters by the ith Nash player, but are treated as variables by a different player j. While we do not specify the syntax here for these issues, [38] provides examples that outline how to carry out this matching within GAMS. Finally, two points of note: first it is clear that the resulting model is a complementarity problem and can be solved using PATH, for example. Secondly, performing the conversion from an embedded complementarity system or a Nash Game automatically is a critical step in making such models practically useful.

Many of the energy market pricing models are based in the economic theory of general equilibria. In this context, there are a number of consumers each maximizing some utility function U_k, income is determined by the market price of the endowment and production shares, production is a technology constrained optimization of profit, and the market clears by choosing appropriate prices:

$$(C): \quad \max_{x_k \in X_k} U_k(x_k) \text{ s.t. } p^T x_k \le i_k(y,p)$$

$$(I): \quad i_k(y,p) = p^T \omega_k + \sum_j \alpha_{kj} p^T g_j(y_j)$$

$$(P): \quad \max_{y_j \in Y_j} p^T g_j(y_j)$$

$$(M): \quad \max_{p \ge 0} p^T \left(\sum_k x_k - \sum_k \omega_k - \sum_j g_j(y_j) \right) \text{ s.t. } \sum_l p_l = 1$$

Note that this model is an example of a Nash Game, with four different types of agents. Note that in each problem, some of the variables are under the control of the agent, and some are given as parameters. One way to solve this problem is to form the KKT conditions of each agent problem, and combine them to make a large scale complementarity problem.

Alternatively, the problem can be reformulated as embedded complementarity system (see [24]) of the form:

$$\max_{x \in X, y \in Y} \quad \sum_k \frac{t_k}{\beta_k} \log U_k(x_k)$$

$$s.t. \quad \sum_k x_k \le \sum_k \omega_k + \sum_j g_j(y_j)$$

$$t_k = i_k(y,p) \quad \text{where } p \text{ is multiplier on NLP constraint}$$

Note that the consumers and producers have been aggregated together into a large nonlinear program parameterized by t_k, a variable that is updated using the external condition. In practice, such problems can then be solved using the sequential joint maximization algorithm [68] .

We note that there is a large literature on discrete-time finite-state stochastic games: this has become a central tool in analysis of strategic interactions among forward-looking players in dynamic environments. The Ericson-Pakes model of dynamic competition [23] in an oligopolistic industry is exactly in the format described above, and has been used extensively in applications such as advertising, collusion, mergers, technology adoption, international trade and finance.

For stylized models of this type, where a game is played over a grid of dimension *S*, the results of applying the PATH solver to the resulting complementarity problems are as follows:

S	Size	non-zero	dense (%)	Steps	Time (m:s)
20	2568	31536	0.48	5	0 : 03
50	15408	195816	0.08	5	0 : 19
100	60808	781616	0.02	5	1 : 16
200	241608	3123216	0.01	5	5 : 12

Note the number of Newton steps is constant, but the model size is increasing rapidly. For the largest grid size, the residual at each iteration is $1.56 * 10^{4}$, $1.06 * 10^{1}$, 1.34, $2.04 * 10^{-2}$, $1.74 * 10^{-5}$, and $2.97 * 10^{-11}$ respectively, demonstrating quadratic convergence. It is clear that it is much easier to generate correctly reformulated models quickly using the automation of the EMP tool.

3.7 Semidefinite Programs

Semidefinite programming is a relatively new optimization format that has found application in many areas of control and signal processing, and is now being more widely utilized due to its inherent modeling power. Excellent survey articles of the background to this area and its applications can be found in [76, 82].

In the context of OPF problems, Lavaei and Low [49] convexify the problem and apply an SDP approach to the Dual OPF. Instead of solving the OPF problem directly, this approach solves the Lagrangian dual problem, and recovers a primal solution from a dual optimal solution. It is proved in the paper that the dual problem is a convex semidefinite program and therefore can be solved efficiently using interior point solvers such as those described in [8, 73, 72]. In the general case, the optimal objective value of the dual problem is only a lower bound on the optimal value of the original OPF problem and the lower bound may not be tight (nonzero duality gap). If the primal solution computed from an optimal dual solution indeed satisfies all the constraints of the OPF problem and the resulting objective value equals the optimal dual objective value (zero duality gap), then strong duality holds and the primal solution is indeed (globally) optimal for the original OPF problem. The paper provides a sufficient condition that guarantees zero duality gap and global optimality of the resulting OPF solution.

However, applying the SDP approach outlined above shows that much more development in solution methodology is needed for this to be competitive for practical modeling. For larger OPF models described via [85] with solutions implemented via YALMIP [51] as a modeling tool and SeDuMi [72] as the solver, reported solutions were not feasible for the original nonlinear program and took significantly longer than alternative nonlinear programming approaches (CONOPT, IPOPT or SNOPT). Research is active in this area, however, and it is likely that methods exploiting underlying

structure in the SDP will become practical in the near future. Indeed, the first order methods for specially structured SDPs described in [42, 59] have already proven effective in eigenvalue optimization problems.

3.8 Extended Nonlinear Programs

Optimization models have traditionally been of the form (1). Specialized codes have allowed certain problem structures to be exploited algorithmically, for example simple bounds on variables. However, for the most part, assumptions of smoothness of f, g and h are required for many solvers to process these problems effectively.

In a series of papers, Rockafellar and colleagues [64, 65, 63] have introduced the notion of extended nonlinear programming, where the (primal) problem has the form:

$$\min_{x \in X} f_0(x) + \theta(f_1(x), \ldots, f_m(x)). \tag{10}$$

In this setting, X is assumed to be a nonempty polyhedral set, and the functions $f_0, f_1, \ldots, f_m$ are smooth. The function θ can be thought of as a generalized penalty function that may well be nonsmooth. However, when θ has the following form

$$\theta(u) = \sup_{y \in Y} \{y^T u - k(y)\}, \tag{11}$$

a computationally exploitable and theoretically powerful framework can be developed based on conjugate duality. A key point for computation and modeling is that the function θ can be fully described by defining the set Y and the function k. Furthermore, as is detailed in [26], different choices lead to a rich variety of functions θ, many of which are extremely useful for modeling.

The EMP model type works in this setting by providing a library of functions θ that specify a variety of choices for k and Y. Once a modeler determines which constraints are treated via which choice of k and Y, the EMP model interface automatically forms an equivalent variational inequality or complementarity problem. There may be alternative formulations that are computationally more appealing; such reformulations can be generated using different options to EMP.

Note that the Lagrangian L is smooth - all the nonsmoothness is captured in the θ function. The theory is an elegant combination of calculus arguments related to f_i and its derivatives, and variational analysis for features related to θ. Exploitable structure is thus communicated directly to the computational engine that can solve the model.

It is shown in [64] that under a standard constraint qualification, the first-order conditions of (10) are precisely in the form of the following variational inequality:

$$\mathrm{VI}\left(\begin{bmatrix} \nabla_x \mathcal{L}(x,y) \\ -\nabla_y \mathcal{L}(x,y) \end{bmatrix}, X \times Y\right), \tag{12}$$

where the Lagrangian L is defined by

$$\mathcal{L}(x,y) = f_0(x) + \sum_{i=1}^{m} y_i f_i(x) - k(y)$$
$$x \in X, y \in Y$$

When X and Y are simple bound sets, this is simply a complementarity problem.

Note that EMP exploits this result. In particular, if an extended nonlinear program of the form (10) is given to EMP, then the optimality conditions (12) are formed as a variational inequality problem and can be processed as outlined above. Under appropriate convexity assumptions on this Lagrangian, it can be shown that a solution of the VI (12) is a saddle point for the Lagrangian on $X \times Y$. Furthermore, in this setting, the saddle point generates solutions to the primal problem (10) and its dual problem:

$$\max_{y \in Y} g(y), \text{ where } g(y) = \inf_{x \in X} \mathcal{L}(x,y),$$

with no duality gap.

In [26], an alternative solution method is proposed, based on a reformulation as an NLP to solve (12). By communicating the appropriate underlying structure to the solver interface, it is possible to reformulate the nonsmooth problem as smooth optimization problems in a variety of ways. We believe that specifying Y and k is a theoretically sound way to do this. Another example showing formulation of an extended nonlinear program as a complementarity problem within GAMS can be found in [22].

4 Stochastic and Robust Optimization

In order to effectively model many of the problems resulting from electricty grid design, EMP will require new syntax to allow specification of problems such as stochastic recourse programs.

Consider, for example, the two stage stochastic recourse problem:

$$\min c'x + \sum_i p_i Q_i(x) \text{ s.t. } x \in X,$$

where $Q_i(x) = \min_y d_i' y$ s.t. $T_i x + W_i y \geq h_i, y \in Y$. Standard modeling notation allows the specification of both X and Y, and the equations that define the feasible set of the recourse (Q_i) problems. The empinfo file would describe:

- the probability distribution p_i
- what stage is each variable in
- what stage is each constraint in
- what parameters in the original model are random (ie T_i , h_i , etc)
- how to sample these random parameters (using a library of sampling functions)
- what problem to generate and solve (ie the equivalent deterministic linear program, or a format necessary for decomposition approaches).

Within the modeling system, there is no need for the underlying problems to be linear. The automatic system would need to check that no random parameters appear in first stage equations, and that no second stage variables appear in first stage equations (and recursively for multi-stage problems) .

An extension to chance constraints is also possible, where the problem is now:

$$\min c'x \text{ s.t. } x \in X, \sum_i p_i \mathcal{I}(T_i x + W_i y_i \geq h_i) \geq 1 - \epsilon,$$

Where $\mathcal{I}(\cdot)$ is the indicator function (1 or 0) for its argument. Clearly, not only should the information needed above be generated, but also the annotation must specify the fraction of the constraints that can be violated (ϵ) .

By employing variable annotations, it would also be possible to extend EMP to model risk measures such as CVaR. An additional variable (which represents a convex function of the decision variables x) would be used in the appropriate constraint or in the objective to be minimized. Extensions of solvers to perform subgradient optimization would be needed, or alternative decomposition approaches could be implemented "behind the scenes".

4.1 Optimization of noisy functions

Over the past few decades, computer simulation has become a powerful tool for developing predictive outcome of real systems. For example, simulations consisting of dynamic econometric models of travel behavior are used for nationwide demographic and travel demand forecasting. The choice of optimal simulation parameters can lead to improved operation, but configuring them remains a challenging problem. Traditionally, the parameters are chosen by heuristics with expert advice, or by selecting the best from a set of candidate parameter settings. Simulation-based optimization is an emerging field which integrates optimization techniques into the simulation analysis. The corresponding objective function is an associated measurement of an experimental simulation. Due to the complexity of simulation, the objective function may act as a black-box function and be time-consuming to evaluate. Moreover, since the derivative information of the objective function is typically unavailable, many derivative-

dependent methods are not applicable. The third example of Section 2 fits nicely into this framework.

We can think of such approaches as a mechanism for coordinating existing optimization technologies. Each simulation evaluation corresponds to the solution of a parameterized problem (maybe in prices, or in variables that link together competing agents) that may be extremely time consuming to compute, and that may incorporate complex domain information, and may be subject to errors due to uncertainties. The noisy function optimization will determine (at a coordination level) what parameters are appropriate and where to concentrate attention in the search space.

Although real world problems have many forms, many optimization strategies consider the following bounded stochastic formulation:

$$\min_{x\in\Omega} f(x) = \mathrm{E}\big[F(x,\xi(\omega))\big], \tag{13}$$

where

$$\Omega = \{x \in \Re^n : l \le x \le u\}.$$

Here, l and u are the lower and upper bounds for the input parameter x, respectively. The specific application of this framework to the electricity grid arises from considering the variables x to be the "design" or line capacity expansion variables. The function F would then model the response of the underlying system (as a large complex computer simulation) to those design decisions $p_i^{\omega}(x)$ and allow for uncertainties in the operating environment via the set Ω. An approach for solving these problems using Bayesian statistics is outlined in [20]. Such an approach balances the computational needs of the optimization against the need for much more accurate evaluations of the simulation. The facilitation of these extremely time intensive solutions using grid computational resources to couple the underlying optimization problems is critical for efficient solution.

4.2 Conic Programming

A problem of significant recent interest (due to its applications in robust optimization and optimal control) involves conic constraints [52, 1, 7]:

$$\min_{x\in X} p^T x \text{ s.t. } Ax - b \le 0, x \in C,$$

where C is a convex cone. For specific cones, such as the Lorentz (ice-cream) cone defined by

$$C = \left\{ x \in \Re^n \mid x_1 \ge \sqrt{\sum_{i=2}^{n} x_i^2} \right\},$$

or the rotated quadratic cone, there are efficient implementations of interior point algorithms for their solution [3]. It is also possible to reformulate the problem in the form (1) for example by adding the constraint

$$x_1 \geq \sqrt{\sum_{i=2}^{n} x_i^2}. \tag{14}$$

Annotating the variables that must lie in a particular cone using a "empinfo" file allows solvers like MOSEK [2] to receive the problem as a cone program, while standard NLP solvers would see a reformulation of the problem as a nonlinear program. It is also easy to see that (14) can be replaced by the following equivalent constraints

$$x_1^2 \geq \sum_{i=2}^{n} x_i^2, x_1 \geq 0.$$

Such constraints can be added to a nonlinear programming formulation or a quadratically constrained (QCP) formulation. This automatic reformulation allows the interior point algorithms to solve these problems since they can process constraints of the form

$$y^2 \geq x^T Q x, y \geq 0, Q \text{ PSD}.$$

Details on the options that implement these approaches can be found in [38] .

These approaches have been been adapted to robust optimization models and applied to unit commitment problems [9]. It is straightforward to facilitate the use of stochastic constraints that have become very popular in financial applications. Specifically, we mention the work of [66] on conditional value at risk, and the recent papers by [21], and [53] on stochastic dominance constraints. All of these formulations are easily cast as constraints on decision variables annotated by additional (in this case distributional) information.

5 Computational Needs

It is imperative that we provide a framework for modeling optimization problems for solution on the computing resources that are available to the decision maker. The framework must be easy to adapt to multiple grid engines or cloud computing devices, and should seamlessly integrate evolving mechanisms from particular computing platforms into specific application models. The design must be flexible and powerful enough for a large variety of optimization applications. The attraction of a grid computing environment is that it can provide an enormous amount of computing resources, many of which are simply commodity computing devices, with the ability to run commercial quality codes, to a larger community of users.

We strongly believe that grid computational resources are not enough to make parallel optimization mainstream. Setting aside the issue of data collection, it is imperative that we provide simple and easy to use tools that allow distributed algorithms to be developed without knowledge of the underlying compute engine. In large scale optimization, there are many techniques that can be used to effectively decompose a problem into smaller computational tasks that can then be controlled by a "coordinator" - essentially a master-worker approach. The above sections have outlined a variety of ways in which this could be accomplished within our optimization framework. We believe that in particular power flow applications, the decomposition approach can be significantly enhanced using specific domain knowledge, but these strategies may be complex to describe for a particular model. This approach is more general that current modeling systems allow, requiring extensions to current languages to facilate interfaces and interactions between model components, their solutions and the resources used to compute in a potentially diverse and distributed environment.

While it is clear that efficiency may well depend on what resources are available and the degree of synchronization required by the algorithm, it must be easy to generate structured large scale optimization problems, and high level implementations of methods to solve them. Stochastic programming (an underlying problem of particular interest within the design of the Next Generation Electric Grid) is perhaps a key example, where in most of the known solution techniques, large numbers of scenario subproblems need to be generated and solved.

A prototype grid facility [14] allows multiple optimization problems to be instantiated or generated from a given set of models. Each of these problems is solved concurrently in a grid computing environment. This grid computing environment can just be a laptop or desktop computer with one or more CPUs. Today's operating systems offer excellent multi-processing scheduling facilities and provide a low cost grid computing environment. Other alternatives include the Condor system, a resource management scheme developed at the University of Wisconsin, or commercial systems such as the cloud or supercomputing facilities avaiable at the National Laboratories. We must facilitate the use of new and evolving modeling paradigms for optimization on the rapidly changing and diverse computing environments that are available to different classes of decision makers.

6 Conclusions

Optimization can provide advice on managing complex systems. Such advice needs to be part of an interactive debate with informed decision makers. Many practical decision problems are carried out in a competitive environment without overall system control. Mechanisms to allow decision making in such circumstances can be informed by game theory and techniques from distributed computing. To answer the major design questions, small dynamic models need to be developed that are "level of detail" specific,

and provide interfaces to other “subservient" models that provide appropriate aggregation of data and understanding of underlying complex features.

A number of new modeling formats involving complementarity and variational inequalities have been described in this paper and a framework, EMP, that allows such problems to be specified has been outlined. Such extensions facilitate modeling with competitive agents and automatic problem reformulations. We believe this will make a modeler’s task easier by allowing model structure to be described succinctly in such a setting, and will make model generation more reliable and automatic. In this way, algorithms can exploit model structure to improve solution speed and robustness. Furthermore, models dedicated to well defined decisions can be formulated using new formats such as semidefinite programming and stochastic optimization, and such descriptions carry the potential of global optimality. EMP is only a first step in this vein.

Solution methods for the resulting optimization problems are available within modeling systems, and the electric power industry could exploit these methods in a flexible manner using a combination of different model formats and solution techniques. Recent advances in stochastic optimization and conic programming are readily available within such systems. Treating uncertainties in large scale planning projects will become even more critical over the next decade due to the increase in volatility of the supply side as well as the demand. Optimization models with flexible systems design can help combat these uncertainties in the construction phase, the operational phase of the installed system, and in the long term demand for the provided electricity.

Acknowledgements

The material in this paper draws heavily on discussions with Jim Luedtke, Jesse Holzer, Lisa Tang, Yanchao Liu, Sven Leyffer, Ben Recht, Chris De Marco and Bernie Lesieutre and I acknowledge their insights and contributions to this work. I am grateful to Steven Dirkse, Jan Jagla and Alex Meeraus for implementing the ideas in the EMP framework in the GAMS modeling system. The opinions outlined are my own, however.

References

[1] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, 95:3–51, 2003.

[2] E. D. Andersen and K. D. Andersen. The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In H. Frenk, K. Roos, T. Terlaky, and S. Zhang, editors, *High Performance Optimization*, pages 197–232. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[3] E. D. Andersen, C. Roos, and T. Terlaky. On implementing a primal-dual interior-point method for conic quadratic optimization. *Mathematical Programming*, 95(2):249–277, 2003.

[4] Göran Andersson. *Modelling and Analysis of Electric Power Systems*, 2008.

[5] J. F. Bard. *Practical Bilevel Optimization: Algorithms and Applications*, volume 30 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

[6] B. T. Baumrucker, J. G. Renfro, and L. T. Biegler. MPEC problem formulations and solution strategies with chemical engineering applications. *Computers and Chemical Engineering*, 32:2903–2913, 2008.

[7] A. Ben-Tal and A. Nemirovskii. *Lectures on Modern Convex Optimization*. MPS-SIAM Series on Optimization. SIAM, Philadelphia, PA, 2001.

[8] Steven J. Benson and Yinyu Ye. Dsdp3: Dual-scaling algorithm for semidefinite programming. Technical report, Preprint ANL/MCS-P851-1000, 2001.

[9] D. Bertsimas, E. Litvinov, X. Sun, J. Zhao, and T. Zheng. Two-stage adaptive robust optimization for the unit commitment problem. Technical report, MIT, 2011.

[10] J. Bisschop and R. Entriken. *AIMMS – The Modeling System*. Paragon Decision Technology, Haarlem, The Netherlands, 1993.

[11] J. Bisschop and A. Meeraus. On the development of a general algebraic modeling system in a strategic planning environment. *Mathematical Programming Study*, 20:1–29, 1982.

[12] J. Bracken and J. T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21:37–44, 1973.

[13] A. Brooke, D. Kendrick, and A. Meeraus. *GAMS: A User's Guide*. The Scientific Press, South San Francisco, California, 1988.

[14] M. R. Bussieck, M. C. Ferris, and A. Meeraus. Grid enabled optimization with GAMS. *INFORMS Journal on Computing*, 21(3):349–362, 2009.

[15] M. Cao and M. C. Ferris. A pivotal method for affine variational inequalities. *Mathematics of Operations Research*, 21:44–64, 1996.

[16] M.T. Cezik and G. Iyengar. Cuts for mixed 0-1 conic programming. *Mathematical Programming*, 104:179–202, 2005.

[17] X. Chen and X. Deng. 3-NASH is PPAD-complete. *Electronic Colloquium on Computational Complexity*, 134, 2005.

[18] Mung Chiang, Stephen H. Low, A. Robert Calderbank, and John C. Doyle. Layering as optimization decomposition: A mathematical theory of network architectures. *Proceedings of the IEEE*, 95(1):255–312, 2007.
[19] C. Daskalakis, P.W. Goldberg, and C.H. Papadimitriou. The complexity of computing a Nash equilibrium. *Electronic Colloquium on Computational Complexity*, 115, 2005.
[20] G. Deng and M. C. Ferris. Variable-number sample-path optimization. *Mathematical Programming*, 117:81–109, 2009.
[21] D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14:548–566, 2003.
[22] S. P. Dirkse and M. C. Ferris. MCPLIB: A collection of nonlinear mixed complementarity problems. *Optimization Methods and Software*, 5:319–345, 1995.
[23] R. Ericson and A. Pakes. Markov perfect industry dynamics: A framework for empirical analysis. *Review of Economic Studies*, 62:53–82, 1995.
[24] Yuri Ermoliev, Yuri Ermoliev, GГjnther Fischer, Gunther Fischer, Vladimir Norkin, and Vladimir Norkin. On the convergence of the sequential joint maximization method for applied equilibrium problems. Technical report, IIASI, Laxenburg, Austria, 1996.
[25] F. Facchinei and J. S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, New York, New York, 2003.
[26] M. C. Ferris, S. P. Dirkse, J.-H. Jagla, and A. Meeraus. An extended mathematical programming framework. *Computers and Chemical Engineering*, 33:1973–1982, 2009.
[27] M. C. Ferris, S. P. Dirkse, and A. Meeraus. Mathematical programs with equilibrium constraints: Automatic reformulation and solution via constrained optimization. In T. J. Kehoe, T. N. Srinivasan, and J. Whalley, editors, *Frontiers in Applied General Equilibrium Modeling*, pages 67–93. Cambridge University Press, 2005.
[28] M. C. Ferris, R. Fourer, and D. M. Gay. Expressing complementarity problems and communicating them to solvers. *SIAM Journal on Optimization*, 9:991–1009, 1999.
[29] M. C. Ferris and T. S. Munson. Complementarity problems in GAMS and the PATH solver. *Journal of Economic Dynamics and Control*, 24:165–188, 2000.
[30] M. C. Ferris and J. S. Pang. Engineering and economic applications of complementarity problems. *SIAM Review*, 39:669–713, 1997.
[31] E. Bartholomew Fisher, R. O'Neill, and M. C. Ferris. Optimal transmission switching. *IEEE Transactions on Power Systems*, 23:1346–1355, 2008.
[32] J. Fortuny-Amat and B. McCarl. A representation and economic interpretation of a two-level programming problem. *Journal of the Operational Research Society*, 32(9):783–792, 1981.
[33] R. Fourer, D. M. Gay, and B. W. Kernighan. A modeling language for mathematical programming. *Management Science*, 36:519–554, 1990.
[34] R. Fourer, D. M. Gay, and B. W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Duxbury Press, Pacific Grove, California, 1993.
[35] A. Frangioni and C. Gentile. Perspective cuts for a class of convex 0-1 mixed integer programs. *Mathematical Programming*, 106:225–236, 2006.

[36] K. Furman and I. P. Androulakis. A novel MINLP-based representation of the original complex model for predicting gasoline emissions. *Computers and Chemical Engineering*, 32:2857–2876, 2008.
[37] GAMS Development Corporation. *NLPEC,* User's Manual, 2008. http://www.gams.com/solvers/nlpec.pdf.
[38] GAMS Development Corporation. *EMP,* User's Manual, 2009. http://www.gams.com/solvers/emp.pdf.
[39] O. Gunluk and J. Linderoth. Perspective reformulation of mixed integer nonlinear programs with indicator variables. Technical report, University of Wisconsin, 2009.
[40] P. T. Harker. *Lectures on Computation of Equilibria with Equation–Based Methods*. CORE Lecture Series. CORE Foundation, Louvain–la–Neuve, Université Catholique de Louvain, 1993.
[41] K. W. Hedman, M. C. Ferris, R. P. O'Neill, E. B. Fisher, and S. S. Oren. Co-optimization of generation unit commitment and transmission switching with n-1 reliability. *IEEE Transactions on Power Systems*, page forthcoming, 2010.
[42] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.
[43] B.F. Hobbs. Linear complementarity models of Nash-Cournot competition in bilateral and POOLCO power markets. *IEEE Transactions on Power Systems*, 16:194–202, 2001.
[44] B.F. Hobbs, C. B. Metzler, and J.S. Pang. Strategic gaming analysis for electric power systems: An MPEC approach. *IEEE Transactions on Power Systems*, 15:638–645, 2000.
[45] W. W. Hogan. Energy policy models for Project Independence. *Computers & Operations Research*, 2:251–271, 1975.
[46] William W. Hogan. *Financial Transmission Right Formulations*, 2002.
[47] William W. Hogan, Scott M. Harvey, and Susan L. Pope. *Transmission Capacity Reservations and Transmission Congestion Contracts*, 1997.
[48] Tarjei Kristiansen and Juan Rosellón. A merchant mechanism for electricity transmission expansion. *Journal of Regulatory Economics*, 29:167–193, 2006.
[49] Javad Lavaei and Steven H. Low. Convexification of optimal power flow problem. In *Forty-Eighth Annual Allerton Conference*, 2010. Available from https://www.cds.caltech.edu/~lavaei/Allerton_2010.pdf.
[50] Q. Li. *Large-Scale Computing for Complementarity and Variational Inequalities*. PhD thesis, University of Wisconsin, Madison, Wisconsin, January 2010.
[51] J. Löberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.
[52] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.
[53] J. Luedtke. New formulations for optimization under stochastic dominance comstraints. *SIAM Journal on Optimization*, 19:1433–1450, 2008.
[54] A. S. Manne and T. F. Rutherford. International trade in oil, gas and carbon emission rights: An intertemporal general equilibrium model. *The Energy Journal*, 14:1–20, 1993.

[55] Maximal Software. MPL. http://www.maximal-usa.com/mpl/what.html, 2009.
[56] Robert H. Miller and James H. Malinowski. *Power System Operation*. McGraw-Hill Professional, third edition, 1994.
[57] A. Monticelli, M.V.F. Pereira, and S. Granville. Security-constrained optimal power flow with post-contingency corrective rescheduling. *IEEE Transactions on Power Systems*, PWRS-2(1), 1987.
[58] Frederic H. Murphy, Hanif D. Sherali, and Allen L. Soyster. A mathematical programming approach for determining oligopolistic market equilibrium. *Mathematical Programming*, 24:92–106, 1982.
[59] Yurii Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming, Series A*, 110(2):245–259, 2006.
[60] Office of Integrated Analysis and U.S. Department of Energy Forecasting. Energy information administration, international energy outlook 2007, 2007.
[61] Thomas J. Overbye, Xu Cheng, and Yan Sun. A comparison of the ac and dc power flow models for lmp calculations. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, January 2004.
[62] D. Ralph and S. J. Wright. Some properties of regularization and penalization schemes for MPECs. *Optimization Methods and Software*, 19(5):527–556, 2004.
[63] R. T. Rockafellar. Linear-quadratic programming and optimal control. *SIAM Journal on Control and Optimization*, 25:781–814, 1987.
[64] R. T. Rockafellar. Lagrange multipliers and optimality. *SIAM Review*, 35:183–238, 1993.
[65] R. T. Rockafellar. Extended nonlinear programming. In G. Di Pillo and F. Giannnessi, editors, *Nonlinear Optimization and Related Topics*, pages 381–399. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
[66] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2:21–41, 2000.
[67] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer-Verlag, 1998.
[68] T.F. Rutherford. Sequential joint maximization. In John Weyant, editor, *Energy and Environmental Policy Modeling*, volume 18 of *International Series in Operations Research and Management Science*. Kluwer, 1999.
[69] Y. Smeers. Computable equilibrium models and the restructuring of the European electricity and gas markets. *The Energy Journal*, 1997.
[70] Brian Stott, Ongun Alsac, and Alcir J. Monticelli. Security analysis and optimization. *Proceedings of the IEEE*, 75(12), 1987.
[71] R. Stubbs and S. Mehrotra. A branch-and-cut method for 0-1 mixed convex programming. *Mathematical Programming*, 86:515–532, 1999.
[72] Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12:625–653, 1999.
[73] K. C. Toh, M.J. Todd, and R. H. Tütüncü. *SDPT3: A MATLAB software package for semidefinite-quadratic-linear programming*. Available from http://www.math.nus.edu.sg/~mattohkc/sdpt3.html.

[74] M. Turkay and I. E. Grossmann. Logic-based MINLP algorithms for the optimal synthesis of process networks. *Computers and Chemical Engineering*, 20:959–978, 1996.
[75] P. Van Hentenryck. *The OPL Optimization Programming Language*. MIT Press, 1999.
[76] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.
[77] A. Vecchietti and I. E. Grossmann. LOGMIP: a disjunctive 0-1 non-linear optimizaer for process system models. *Computers and Chemical Engineering*, 23:555–565, 1999.
[78] A. Vecchietti, S. Lee, and I. E. Grossmann. Modeling of discrete/continuous optimization problems: Characterizaton and formulations of disjunctions and their relaxations. *Computers and Chemical Engineering*, 27:433–448, 2003.
[79] J. Wallace, A. Philpott, M. O'Sullivan, and M. Ferris. Optimal rig design using mathematical programming. In *2nd High Performance Yacht Design Conference, Auckland, 14–16 February, 2006*, pages 185–192, 2006.
[80] Hongye Wang, Carlos E. Murillo-Sánchez, Ray D. Zimmerman, and Robert J. Thomas. On computational issues of market-based optimal power flow. *IEEE Transactions On Power Systems*, 22(3):1185–1193, 2007.
[81] J.-Y. Wei and Y. Smeers. Spatial oligopolistic electricity models with Cournot firms and regulated transmission prices. *Operations Research*, 47:102–112, 1999.
[82] H. Wolkowitz, R. Saigal, and L. Vandenberghe, editors. *Handbook of semidefinite programming: theory, algorithms and applications*. Kluwer, 2000.
[83] J. Yao, B. Willems, S. Oren, and I. Adler. Cournot equilibrium in price-capped two-settlement electricity markets. In *Proceedings of the 38th Hawaii International Conference on Systems Sciences HICSS38*, Big Island, Hawaii, January 3-6 2005.
[84] Jian Yao, Shmuel S. Oren, and Ilan Adler. Computing cournot equilibria in two settlement electricity markets. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, January 2004.
[85] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas. MATPOWER's extensible optimal power flow architecture. In *Power and Energy Society General Meeting*. IEEE, July 26-30 2009.

Discussant Narrative

Coupled Optimization Models for Planning and Operation of Power Systems on Multiple Scales

Ali Pinar
Sandia National Laboratory

What More Do We Want From Michael?

Optimization Offers Critical Capabilities

- Almost all problems above the 5-seconds scale involve optimization.
- Significant impact in the short term is possible for a lot of problems.
- Basic research is required for many other problems.
- One of the biggest challenges is in integrating
  - Various time scales
  - Separate processes
  - Decision making entities.

Proposed ideas

- Are scientifically challenging.
- Have the potential to make a huge impact in practice.
- Bring a domain-specific flavor to be a signature for an OE-based research program.

Success Will Be Determined By Our Ability to Model

- Interdisciplinary teams that solve *the right problem* is only the first step.
- Formulation should define the problem in mathematical terms, not try to solve it.
- Extended mathematical is a good step towards this.
- How we structure the models is the main question.
- We learned a lot working with legacy codes.
  - legacy codes: old codes where the problem, model, assumptions, different time scales, different components, algorithms, their implementations, analysis of the output is interleaved.
  - Products of years of incremental process after starting on a less than ideal track.
- We are starting now. Let's do it right.

Living Interfaces Are Essential For Long-Term Success

- Standard interfaces boost productivity in the short term.
- Any good interface turns into a bottleneck.
- Case Study: National Energy Modeling System (NEMS)
  - Started with different components for modeling flexibility.
  - Now has each component pulling to a different direction.
- Open Questions
  - How do we handle evolving interfaces?
  - How do we handle rapid growing complexity?
  - In a hierarchical model, each layer will multiply complexity (even if we do not generate a single giant problem).
  - Abstractions are necessary for tractability (e.g., knowing that a solution exists may be sufficient without the solution itself).

Better Uncertainty Models Are Required

- A convenient assumption: Given a set of scenarios that represent the uncertainties…
- Uncertainties should be included in decision making, and better models means smarter decisions.
- The space of uncertainties can be huge, so intelligent models are required.
- We cannot walk around this problem by increasing the number of samples.
  - It may generate a lot of redundant samples.
  - Without any guarantee for good coverage.

High-Performance Computing

- HPC is a double-edged sword.
  - Enables solution of extremely hard problems.
  - Can be an excuse to use poor algorithms.
- New trends in high-performance computing work in favor of optimization algorithms.
  - Your laptop is already a parallel computer.
  - And it will have many, many more cores.
- Long-term planning problems will require bigger computational resources.
  - Some other government agencies have a lot of experience in this area.
  - Cost analysis between cloud computing and other platforms will be beneficial.
- For shorter term problems, extensive computational resources will be widely available.
  - Build the models, and the compute resources will come.

Other Optimization Challenges

- Multi-objective optimization.
- Optimization with soft constraints.
  - Conservation of energy is a hard constraint.
  - "I don't want to spend more than $100" is a soft constraint.
- Setting on the constraints.
  - How much reserve do we need?
  - What capacity constraints do we impose on the lines?
- Herding cats:
  - How do we make sure the full system moves in the right direction, while its entities are maximizing their gains?
  - What operational constraints do we set for this purpose?
  - What would be the return of a government incentive?

Recorder Summary

Coupled Optimization Models for Planning and Operation of the Power Systems on Multiple Scales

Alejandro D. Dominguez-Garcia
University of Illinois at Urbana-Champaign

The discussion on the paper got kicked off on the subject of trade-offs between encapsulating the problem into small pieces with appropriate interfaces between them and trying to formulate the problem as a full-blown optimization program. Encapsulation helps keeping a tractable model but there might some subtleties of the problem addressed that are not captured by the resulting (encapsulated) model. The question that is important to address is whether or not those subtleties might be important in the solution.

Tractability and validation of so-called monster models came up next. In other words, while it is possible to formulate a full-blown model of a particular problem, the complexity and size of the problems addressed makes the resulting model no longer tractable and questions arose on how to validate it. This is one of the reasons for breaking the "monster problem" into smaller sub-problems with interfaces between them. The result is a hierarchical optimization problem, the individual pieces of which are understood by the modeler. In this regard, modeler experience is an invaluable asset on evaluating whether or not the solutions of each piece are meaningful or not. It is clear that no experienced modeler can qualitatively validate the solution of a "monster model."

A discussant brought up the fact that a "monster model" should get a more accurate answer than the hierarchical optimization approach. While this is obviously true, the question that arose is whether the accuracy improvements provided by the "monster model" are significant. Additionally, in the hierarchical approach where smaller models are solved with appropriate interfaces can be used to guide the development of a "monster model." In this regard, once the results of small pieces of the model are (qualitatively) validated, it is possible to remove their interfaces. Ultimately, by removing all interfaces, the "monster model" is recovered. This bottom-up approach has a built-in validation mechanism (through qualitative verification each sub-problem solution) that might give the modeler more confidence on the results provided by the resulting "monster model"

A discussant pointed out that the electric industry is not interested in developing "monster models" as it is impossible to validate their results. Furthermore, a special appeal of the hierarchical modeling approach is that each sub-problem is of different nature and their characteristics might call for different solution methods. Understanding the features of each individual sub-problem that might improve the efficiency of solution techniques is much more important that building a "monster model" the solution of which (given the model is correct) is perhaps more accurate but far more difficult to obtain. In the context of this discussion, the importance of developing a library of benchmark models (and associated solution methods) that capture the nature of specific sub-problems came up. By having this library, it is possible to understand the similarities between certain problems and the weakness and strengths of solution methods to address them.

Questions of stochasticity and uncertainty came up next. A discussant brought up the fact that there was no material for handling uncertainty. The discussion then centered around the fact that there is a large body of work in optimization to deal with stochastictiy and uncertainty, from multi-stage dynamic programming to robust optimization and stochastic programming. The key issue here is not the solution technique but how stochasticity and uncertainty enters the model at different time scales.

A portion of the discussion centered around the issue of power system stability and the fact that there are not analytical expressions to capture power system stability limits when setting up an optimization problem. For example, when solving the OPF problem, the constraints cannot capture transient stability considerations. In this regard, although it might not be possible to include them as constraints, it might be feasible to verify whether or not the solution to the OPF problem meets transient stability criteria and other dynamic performance requirements. Again, the idea of using one the solution to one sub-problem to inform the formulation of another sub-problem can be used here. For example, if the solution to the OPF problem does not meet transient stability requirements, the OPF formulation can be modified so as some additional constraints act as a surrogate for dynamic performance criteria and then check whether or not the resulting solution meets these performance criteria.

A discussant remarked about the fact that there are multi-disciplinary optimization techniques developed in the context of aerospace problems that might be relevant to power systems. There was no specific answer to this remark, so I want to add my **personal token to the discussion**. As Fred Schweppe put it in the discussions section of his 1970 seminal paper on state estimation: **"I worked on aerospace problems for many years before converting to power systems, and, in my opinion at least, power problems are tougher in many respects ... The number of variables [in a power system] is huge, and many types of uncertainties are present... Few if any aerospace problems yield such a challenging set of conditions."** Enough said.

There was some discussion on the validity of the interfaces between sub-problems, basically captured through the KKT conditions of a particular sub-problem. A discussant (erroneously) brought up that the KKT conditions are only sufficient in the case of linear programs (they are also sufficient in convex programs). However, if KKT conditions are not sufficient in certain sub-problems, the solution (to the larger optimization problem that contains the sub-problem) that results from replacing the sub-problem with the corresponding KKT conditions will be a local maximum and minimum and a good starting point to search for other solutions.

White Paper

Mode Reconfiguration, Hybrid Mode Estimation, and Risk-bounded Optimization for the Next Generation Electric Grid

Andreas G. Hofmann and Brian Williams
MIT Computer Science and Artificial Intelligence Laboratory

1 Introduction

1.1 Background and Motivation

Achieving reliable and efficient power control in a large network is crucial for the nation's power grid. As demands on the national power grid continue to increase, and as the infrastructure and technology is upgraded and renovated, the problem of maintaining and guaranteeing reliable power, and producing and transmitting this power efficiently, will become increasingly challenging. While a number of advanced technologies have been employed to solve some aspects of this problem, others are currently not solved, or are addressed using ad-hoc methods. This RFI response identifies a number of unsolved problems, explains why current methods for addressing them are inadequate, and discusses promising research directions that will yield solutions.

An electric grid consists of a network of transmission lines. Power through this network is controlled in two ways. First, the power produced by generators can be adjusted. This will have direct and indirect consequences throughout the network. Second, circuit breakers can be used to open particular transmission lines. Using circuit breakers in this way is the only method to directly control power flow in a transmission line, but the effects on other parts of the network are indirect, and often hard to determine. Thus, key questions are: 1) how do changes in generated power change flow in the network, and 2) how does the sudden opening of a line change power flow in the network (Fig. 1). The latter question is important not only for understanding the consequences of a control action, but also, for understanding the consequences of unexpected line open events, and planning contingencies for them.

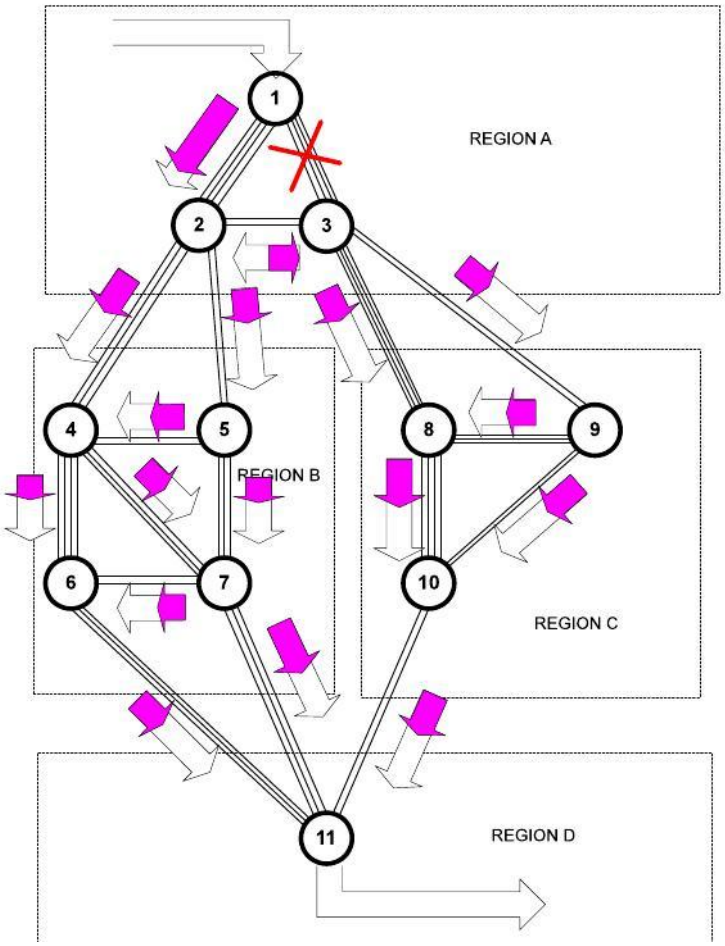


Figure 1: When a line opens, its power flow in that line stops almost instantaneously. The flow must go elsewhere. Most lines in the network will see change. (from NDIST 2009, Bob Thomas presentation)

In current practice, electric utilities strive to maintain an (n-1) readiness level. This means that they are always prepared to instantly take required compensating actions if any single line opens unexpectedly, in order to avoid service interruptions. The utilities achieve this readiness level by generating and testing what if scenarios corresponding to each possible line open event. For each scenario, compensating actions and action sequences are proposed and evaluated to ensure that safety requirements are met throughout the scenario. Safety requirements include balance of generation and demand, power flows being below limits (thermal safety), and power and voltage levels being within stability limits.

A shortcoming of this approach is its ad hoc and human labor intensive nature. Each scenario and the corresponding set of compensating actions, must be manually generated. Thus, completeness of the analysis is a potential problem; some scenarios, or some actions for a scenario may be missed. Evaluation of the compensating actions is accomplished using ad hoc rules, and by running simulations of the scenarios and

checking that all safety constraints are met. This is also a labor-intensive process. The entire process must be run continually to regenerate contingency plans as the overall situation changes. Thus, in reality, it is difficult to maintain the (n-1) guarantee for all times. If a situation changes rapidly, it takes time to re-generate the contingency plans. During this time, the (n-1) readiness level may not be met. Finally, in a real situation, there may be multiple faults; multiple lines may open at the same time. The (n-1) readiness criterion does not address this.

Besides the basic problem of maintaining reliable power, there are a number of related problems of interest. First, any kind of what-if analysis to develop contingency plans requires knowledge of the current situation or state. Computation of current state is based on sensors, and is called state estimation. A key challenge is determining hidden state, such as values that are not directly measured by sensors. This is accomplished using a model that predicts the hidden values from the observed ones. Such models have parameters that are time invariant, or change slowly with time. These parameters typically must also be estimated. In current practice, simple weighted least squares techniques are used to estimate parameters and states that minimize discrepancies in observed values. These techniques do not account for discrete mode changes, such as open and closing circuit breakers, which cause significant discontinuities in behavior.

A second related problem of interest is that of optimization. Beyond basic guarantees for maintaining reliable power to customers, it is desirable to generate and distribute the power in the most cost-effective way. A solution to this problem requires knowledge of demand forecasts, as well as knowledge of power generation and distribution costs.

1.2 Problem Summary

This RFI response focuses on two problems. The first is increasing the level of automation in the analysis and planning of contingencies in response to unexpected open line and other events. This level of automation would still be advisory, with humans in the loop, but it would reduce the drudgery and error-prone nature of the current labor-intensive approach. It would provide guarantees of completeness of the analysis, and the validity of the contingency plans. The second is incorporating optimality considerations into the contingency planning and overall energy management process. Such optimization would include risk bounds on actions taken to achieve optimal performance.

2 Key Insights

For the contingency analysis and planning problem, a key insight is that this is a type of Mode Reconfiguration problem. Mode Reconfiguration problems involve computing sequences of discrete actions to move a hybrid discrete/continuous system from an initial

state to a goal state. Recent advances in the solution of Mode Reconfiguration problems for control of spacecraft, naval, and chemical process systems should be further investigated to determine applicability to power grid problems. The appropriate leveraging of these new technologies would have a major impact on the way in which power grid contingency planning is achieved.

For the optimization problem, a key insight is that this is a problem in which optimal performance is desired, but only within specified failure risk bounds. This type of problem has received significant attention in recent years, driven by the increasing demand for autonomous systems that behave optimally in complex, uncertain environments. In particular, recent work in this area has resulted in risk-bounded optimal controllers for autonomous ground, air, and underwater vehicles, as well as for local, decentralized microgrids. Application of these new technologies to optimal power flow problems should be further investigated.

The following sections state the two problems in more detail, provide an overview of how these problems are currently solved, and then describe recently developed techniques for optimal estimation and control of hybrid systems, and requirements and challenges for applying these techniques to electrical grids.

3 Automated Contingency Planning

3.1 Automated Contingency Planning Problem and Requirements

Safe operation of an electrical grid implies that specific operating requirements are met. These include balance of generation and demand, balance of reactive power supply and demand, power flows being within limits (thermal safety), and power and voltage levels being within stability limits.

During operation, events may occur, such as line open or other failure events, that threaten safe operation. Such events, if not attended to appropriately, can put the system into a mode that results in a state evolution that violates safe operation requirements. Contingency planning involves anticipating likely events of this nature, and planning appropriate responses so that the state evolution remains in safe operating regions. The primary means for responding (for controlling the network) are switching of line circuit breakers, and adjusting generation. The former controls flow in an individual line directly. The latter is indirect, in that it can take some time to take effect.

Comprehensive contingency planning requires addressing five related sub-problems. First, it is necessary to identify acceptable, safe target modes for the electrical grid. Second, it is necessary to properly estimate the current mode of the system, including possible fault status of equipment. Third, is the planning itself, which involves

generation of control actions that result in a sequence of acceptable mode transitions leading to a safe target mode. Fourth, given a current situation, it is important to identify the most likely faults, and the probability of contingency plan success, so that contingency planning is focused on solving the most likely problems, and so that when alternative contingency plans are available, the one with the highest likelihood of success can be chosen. Finally, it is useful to analyze how to take pre-emptive actions that make the current mode, and possible target modes safer and more robust to disturbances. We now discuss current approaches to contingency planning. Subsequently we present promising research directions for addressing the five sub-problems.

3.2 Current Approaches to Contingency Planning

As stated previously, in current practice, electric utilities strive to maintain an (n-1) readiness level. This is currently achieved by generating a list of single fault contingencies, proposing associated control sequences that will deal with the contingency, and then testing the control actions using forward simulations.

A shortcoming of this approach is its ad hoc and human labor intensive nature. Each scenario and the corresponding set of compensating actions, must be manually generated. Even for single faults, this can result in a very large number of contingencies. When multiple faults are considered, the problem grows exponentially. While generation of single fault contingencies is relatively straightforward, generation of corresponding control sequences is not. Rules and guidelines are currently used, and then tested using simulations.

Thus, completeness of the analysis is a potential problem; some scenarios, or some actions for a scenario may be missed. Evaluation of the compensating actions is accomplished by running simulations of the scenarios, and by checking that all safety constraints are met. This is also a labor-intensive process. The entire process must be run continually to regenerate contingency plans as the overall situation changes. Thus, in reality, it is difficult to maintain the (n-1) guarantee for all times. If a situation changes rapidly, it takes time to re-generate the contingency plans. During this time, the (n-1) readiness level may not be met. In addition to these problems, the general lack of systematic probabilistic analysis makes it difficult to focus on the problems that are most likely to occur, to know the probability of success of contingency plans, and to know how to improve the robustness of modes through pre-emptive actions.

Rule-based expert system approaches have also been applied to the contingency planning problem [14, 21, 20]. These approaches attempt to replicate the guidelines and rules that operators use to handle contingencies. Similar to the manual approaches, there is no systematic guarantee that the actions taken will be correct, or that coverage is over all possible contingencies.

3.3 Promising Research Directions

The following subsections discuss promising research directions for the five contingency planning sub-problems introduced previously.

3.3.1 Identifying Safe Target Modes

Safe operating modes for an electrical grid are ultimately determined by safety constraints such as line power flow limits (see subsequent discussion). These constraints imply other constraints, including ones that directly constrain the network topology through constraints on switch settings. Thus, making the implicit constraints explicit is a means of identifying switch setting combinations that result in safe operating modes.

An important question is how much computational effort should go into making implicit constraints explicit, and how much of this should be done ahead of time vs. in real time. Pre-computing safe operating modes and caching them can be useful for fast emergency operation. In addition to the safe modes themselves, it is also useful to compute transitions between them, as this can provide fast, automated guidance to operators in emergency situations. More generally, it is useful to compute a transition graph representing possible transitions from emergency modes to safe modes.

Given the combinatorics in large networks, it may not be feasible to precompute and cache all modes and transition graphs. Therefore, an important area of research is identifying the best safe modes and transition paths to them for the most likely emergency modes.

3.3.2 Estimating Current Mode and Diagnosing Faults

Although the instrumentation of the electrical grid continues to improve, there is still a strong need for improving situational awareness. This means computing good estimates of states that can't be measured directly, and accounting for sensor noise and malfunction. This will improve any analysis that requires knowledge of the current situation. The improvement is achieved through improvement of the estimates themselves, and also through improved knowledge of the uncertainty of such estimates.

In electrical power system applications, state estimation is used to compute best estimates of the system's state variables, including bus voltages and angles, and line flows [1]. In current practice, automated state estimation algorithms focus on steady state analysis, and on estimation of continuous quantities (such as bus voltages and angles) using weighted least square error methods. Dynamic analysis is avoided due to its complexity and performance requirements. Separate, ad hoc modules are used to process measurements of discrete (logical) values, such as switch status, or general equipment status, in order to determine the current network configuration and availability of equipment.

Errors in the measurement of discrete values (due to sensor failure, for example) can have serious consequences for the state estimation process. In particular, weighted least square error methods are susceptible to large errors due to inclusion of bad data. Thus, detection and removal of such data is of primary importance. Current practice uses a combination of statistical and ad hoc methods to detect the presence of bad data. A more comprehensive approach, fully integrated with the state estimation process is needed.

Even when bad sensor data has been removed, there is still the possibility for incorrect results if the model parameters used in the weighted least square error methods have errors. Often, manufacturers data and one line drawings are used to determine parameter values. This can be problematic if the parameter values of the fielded equipment are different, or change over time. In current practice, extensive, continual measurement of parameters in the field is impractical. Error reduction in model parameters used for state estimation is an area requiring further work.

For these reasons, three important research directions for the state estimation problem are: 1) extending the algorithms to include dynamics; 2) systematically incorporating estimation of discrete values such as network topology and equipment status; and 3) automating and integrating the model parameter estimation process with the state estimation process. The goal should be to develop a comprehensive computational framework that incorporates all these considerations and requirements, rather than handling them separately.

Recent advances in the estimation of hidden discrete modes [11, 12] are a useful starting point when considering these extensions. These algorithms provide a novel combination of Hidden Markov Model (HMM) techniques [19], with Optimal SAT solver technology [27], resulting in a system that models probabilistic mode transitions with combinatoric guard constraints. These algorithms have been applied to autonomous control of spacecraft, shipboard naval systems, and other autonomous robots.

Techniques for estimating hidden discrete modes have recently been extended to incorporate continuous variables and constraints [23]. These *Hybrid Mode Estimation* (HME) techniques provide a comprehensive framework that could be applied to electrical grids to estimate both continuous state, such as bus voltage magnitudes and angles, and discrete mode, such as equipment status. Hybrid Mode Estimation algorithms also consider the dynamic evolution of the system over time.

In the HME framework, the behavior of discrete modes is governed by state transitions between modes, which may be probabilistic, and may be conditioned (guarded) on continuous state variables and the discrete states of other components. Each component is modeled by a probabilistic hybrid automaton (PHA), comprised of a set of component

modes, continuous state variables, observables, guarded probabilistic transitions, and stochastic dynamic equations associated with each mode.

Estimation of PHA state builds upon the theory of Bayesian Filtering, a versatile tool for framing hidden state interpretation problems. A Bayes Filter operates on a model described by an initial state function, which represents the probability that the PHA is in a particular initial state, a state transition function, which represents the conditional probability of a state transition, and an observation function, which represents the conditional probability of an observation given a state. The current belief state is updated using prediction/correction equations that first predict the future state using the transition function, and then correcting this using the observations and observation function.

HME computes trajectories of state estimates, maintaining multiple trajectories corresponding to multiple hypotheses for what is happening. Tracking the complete set of hybrid states is computationally intractable; hence HME instead maintains a set of most likely discrete and continuous state trajectories. Roughly speaking, continuous state estimates are updated through a generalization of Kalman Filter update, while discrete mode estimates are updated through a generalization of HMM update, while these continuous and discrete sub-systems pass their respective estimates to each other.

To date, hybrid estimation methods have been applied to individual and small sets of concurrently operating probabilistic hybrid automata. An open challenge is the task of scaling HME to larger systems. One approach to this worth exploring further is application of model decomposition methods based on Causal Ordering [5] to factor HME into a set of lower dimensional hybrid estimation problems. Further details on Hybrid Mode Estimation are provided in Appendix I.

Hybrid Mode Estimation and related techniques depend on system models of relationships between state variables to augment and filter the sensor information. These models contain parameters that must be accurate in order for the state estimates to be accurate. Such parameters are often obtained from manufacturer's equipment specifications, but this can be inaccurate in that there may be significant variance in parameter values in actual fielded equipment. Also, parameter values may change over time.

Thus, it is worth investigating computational frameworks that use online sensor data to adjust estimated values of parameters, as well as the state estimates themselves. A key assumption here is that parameter values change slowly over time, if at all, whereas state values change quickly. One promising approach to learning model parameter values, while also estimating state, is the use of *Expectation Maximization* (EM) methods [2]. EM methods alternate between using the best estimate thus far for parameter values to estimate state values, and then using the estimated state to gradually adjust parameter value estimates.

Another interesting research direction for learning parameters is the use of *Active Learning* algorithms [ref. Lars]. These transition the state of a system with uncertain parameters through a sequence of "maneuvers". This provides a means of exploring the parameter space, particularly in areas that are not usually encountered in normal operation. These techniques have demonstrated interesting results in the control of air vehicles [3] and underwater vehicles [8]. When using these techniques, care must be taken to avoid unsafe states when performing maneuver sequences.

An additional challenge in parameter estimation is the fact that models for large systems like electrical grids typically have a large number of parameters. One approach to this problem is to decompose a large system into smaller systems that can be solved separately in much less time than the full system [25].

The problem of hidden mode estimation is closely related to that of fault diagnosis; if the hidden modes include component fault status, then mode estimation techniques inherently perform fault diagnosis. This approach has been developed over the past two decades, and has yielded significant capabilities and results [6, 24]. Future research directions include improved estimation of probabilities of component failures when this information is not readily available from component specifications, and solution of very large problems through decomposition techniques.

3.3.3 Mode Reconfiguration Planning and Execution

Mode reconfiguration involves generating a sequence of discrete control actions that take a system from an initial discrete state (mode) to a goal mode [26]. For example, in a spacecraft, this might involve opening and closing valves to change the path of fuel from tanks to rocket motors, in order to avoid a path through faulty components, as shown in Fig. 2. In a naval application, this might involve opening and closing valves and circuit breakers to switch from primary to auxiliary systems after an attack. In an electrical grid, it might involve opening and closing circuit breakers to control power flows, as described previously.

The *Livingstone* system [26] was developed to perform reactive planning on spacecraft, through a reactive planner called *Burton*. This planner is capable of quickly generating control action sequences that take a complex system from an initial to goal mode. Key features of the planner are its speed, which allow it to be used for contingency planning situations, its ease of use due to its model-based approach, and the completeness of its analysis. Although originally developed for spacecraft applications, the planning algorithm is generally applicable.

Burton is a generative planner in that it takes a general specification of allowed control actions and transitions, and assembles a sequence of actions. Burton uses a model-based

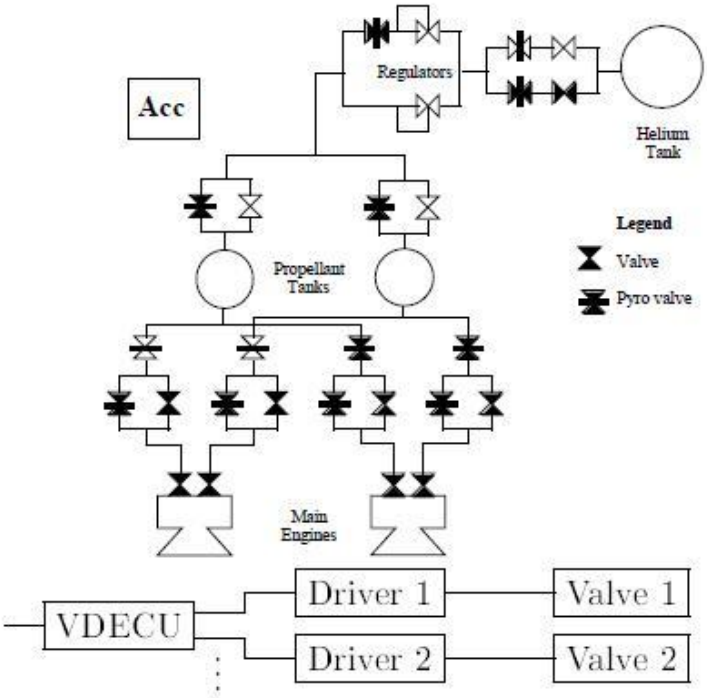


Figure 2: In a spacecraft, valves are opened and closed to change the path of fuel from tanks to rocket motors.

approach, where component models are defined individually, and then multiple instances of these models are assembled into a network. This planner achieves speed that is much better than other generative planners through the use of a causal graph, which is based on knowledge of the component topology in the network. Burton is also easy to use in that once component models have been defined, they can be easily used as building blocks to form large, complex networks. Through the use of the causal graph, and the associated speed improvements, Burton is able to provide guarantees of the completeness of its analysis.

The application of reactive planners like Burton to electrical power grid management problems will require enhancements in order to meet this application domain's requirements. Three areas require particular attention. First, it will be necessary to extend the planning representation from a purely discrete one to include hybrid discrete/continuous systems, and support planning over both discrete modes and continuous variables. This will allow for integration of AC power flow tools that are currently used in the industry to perform continuous power calculations. Second, current reactive planners accept a target state as an input. A significant improvement would be the ability to automatically generate and evaluate multiple target states. Third, there remain significant challenges in applying reactive planners to very large systems. Further performance improvements, achieved through novel network decompositions, are a promising way of solving this problem.

The following paragraphs describe reactive planning in more detail, and discuss requirements and research challenges for applying this algorithm to the problem of contingency planning for electrical grid faults.

Discrete Mode Reconfiguration

Transitioning a system's current discrete mode to a goal mode is called *mode reconfiguration*. Mode reconfiguration is accomplished by a sequence of control actions that step the system through a sequence of valid modes. The rules for which modes are valid, and how control actions accomplish transitions can be expressed using logical constraints. The set of such rules can become large and complex for large systems.

Model-based reactive planners, such as Burton, are able to plan transition sequences that accomplish mode reconfiguration. These planners use component models that are combined to model large, complex systems. By making a number of key assumptions, and by using a novel compilation approach, the planners are able to generate plans quickly, allowing for their use in real-time applications.

Burton is based on the concept of a transition system. A transition system, S, is a tuple $\langle \Pi \Sigma \tau \rangle$. Π is a set of finite domain variables, partitioned into the set Π_s of state variables, the set Π_c of control variables, and the set Π_d of dependent variables. Σ is the set of feasible assignments to the variables in Π. τ is the set of transition rules that define how the system evolves from one assignment to another. The set Σ is specified using propositional logic formulas. The set τ is specified using transition rules of the form $\Phi_i \Rightarrow \bigcirc y_i = e_i$, where Φ_i is a propositional formula, y_i is a state variable, e_i is a value in the domain of y_i, and $\bigcirc$ is the *next* operator, denoting truth in the next state of a trajectory of assignments.

Burton takes as input a transition system, S, an initial mode s_i, and a target mode t_i. It generates a control action, μ_i, that results in a new mode, s_{i+1}, that is consistent with the constraints and transitions rules, and this is part of a trajectory that leads to t_i.

A key difference between Burton, and traditional *Strips* planners is that the latter modify state directly. In contrast, our planner modifies state indirectly, through control actions. These affect state through the control and dependent variables. This greatly increases expressive power of the model, but it also introduces complications. Intractibility is eliminated through an automated model compilation method, and a set of simple assumptions.

In order to simplify the transition system, all dependent variables and associated interactions are compiled away by generating all *prime implicants* of $\bigcirc y_i = e_i$. An implicant of $\bigcirc y_i = e_i$ is a conjunction I of propositions involving state and control

variables, but no dependent variables, such that the transition specification entails the formula $I \Rightarrow \bigcirc y_i = e_i$. I is a prime implicant if no sub-conjunction of I is an implicant. The set of prime implicants constitutes the compiled transition specification, where each implicant specifies a transition for a single state variable.

Given a compiled transition system, Burton quickly generates the first control action of a valid plan, given an initial state assignment, and a set of goal assignments. Burton avoids runtime search, and expends no effort determining future actions that are not supported by the first action. Burton accomplishes this speedup by exploiting certain topological properties of component connectivity that frequently occur in designed systems. In particular, the input/output connections of a compiled plant frequently do not contain feedback loops. When they do occur, they are typically local and can easily be eliminated through careful modeling. Thus, a *causal graph* G for (compiled) transition system S is a directed graph whose vertices are the state variables of S. G contains an edge from v_1 to v_2 if v_1 occurs in the antecedent of one of the transitions of v_2. The causal graph must be acyclic, as, for example, the valve network in Fig. 2.

The basic idea underlying Burton is to solve a conjunction of goals by working "'upstream"' along the acyclic causal graph. Thus, the planning algorithm completes a conjunct $\bigcirc y_i = e_i$ before conjunct $\bigcirc y_i = e_i$ when precedes y_i in the causal graph. Burton avoids generating destructive control actions by exploiting the acyclic nature of the causal graph. The only variables needed to achieve an assignment to y are y's ancestors in the graph. Thus, Burton can achieve a conjunction of goal assignments in an order that moves along the causal graph from descendents to ancestors.

The following sub-section discusses challenges and required enhancements for applying this approach to electrical grid contingency planning problems.

Application to Electric Grid Contingency Planning: Hybrid Mode Reconfiguration

In applying the Burton approach to the problem of electrical grid contingency planning, it will be desirable to preserve properties that allow for Burton's fast performance. However, electrical grid contingency planning problems present additional requirements and characteristics beyond the ones that occur in the component network problems that Burton has previously been used for.

A first question to consider is what the discrete state (mode) variables for an electrical grid should be. For contingency planning problems, the on/off state of circuit breakers, generators, and loads must be included. Additionally, we might want to include the functional status of these components (ok or faulty), and of related components such as sensors, relays and transformers, in order to support contingency planning in the presence of components that have been diagnosed to be faulty.

A second question to consider is what the discrete control and dependent variables are, and what logical relations exist between these variables, and the discrete state variables. For example, it is important to consider whether circuit breakers should be modeled as being directly controllable, or as being indirectly controlled through a SCADA system, relay, or other controller. The latter implies a set of logical relations between control and state variables for a circuit breaker system. Additionally, there may be a priori specifications for valid combinations of circuit breaker settings that must be enforced. Finally, transition relations between discrete control, dependent, and state variables must be modeled.

When modeling the transition relations, the question of the acyclic nature of the transition model becomes important, as described in the previous sub-section. In particular, the Burton algorithm requires an acyclic transition graph to function properly. Therefore, a key question is whether electrical grid contingency planning problems can be formulated such that there are no cycles in the transition relations. An interesting possibility to investigate is whether it is beneficial to employ discrete, qualitative models of line states. Such qualitative models would consider not only whether a line is open or closed, but also its directionality of power flow. Such directionality would support the use of acyclic transition models. For example, if a line whose circuit breakers are open is to provide power to a consumer, then there must exist upstream "'input"' lines that provide power to this line. If such lines do not exist, or are not in the proper qualitative state, then there is no point in closing the circuit breakers for the line.

A complete model that supports contingency planning will be hybrid; it will include continuous variables and constraints as well as discrete ones. For example, in order to properly accomplish contingency planning, the continuous power flow and voltage levels, must be considered. Thus, a key open question is how to model the continuous aspect, and how the continuous part of the model should be processed by the planning algorithm. One comprehensive approach to this, that also includes optimal power flow calculations, is the "'SuperOPF"' system [9]. The SuperOPF system considers contingencies as well as optimization, but could benefit from incorporation of the Mode Estimation and Reconfiguration concepts discussed previously.

One option to consider is the use of steady state network power flow models. These models are constrained by the discrete mode of the circuit breakers, resulting in a set of linear equalities that relate power flows and voltages, and a corresponding set of linear inequalities that set upper limits on power flows, voltages, and other quantities of interest. With this approach, there are no continuous state variables, and therefore, no continuous state transitions. The continuous variables are essentially dependent variables that participate in constraints that define valid behavior of the system. Thus, a key question is how these additional constraints should be handled. For the discrete aspect, the prime implicant compilation approach was used to remove the dependent variables. Is something similar possible for the continuous variables? If not, is it possible

to incorporate traditional linear equation solvers into the transition constraint checking in an efficient way, so that speed is not significantly compromised? Should the combined logical and continuous constraint system be solved using a *Mixed Logic Linear Programming* (MLLP) approach [10]?

For some contingency planning problems, it may not be sufficient to use steady state network power flow models. In such problems, it may be necessary to consider dynamics associated with transient conditions, such as the short-term effects of circuit breakers opening or closing, or the start-up or shut-down of rotating machinery. This further complicates the planning problem, as there are now continuous state variables as well as discrete ones. The continuous state variable transitions are governed by a set of dynamics equations, usually expressed as ordinary difference or differential equations. A key question at this point is whether these continuous dynamics include continuous control actions that must be planned and optimized, or whether they are completely determined by the discrete control actions. In the former case, the continuous state variables can be easily updated using the difference equations, and their values included in the overall continuous constraint system. In the latter case, it is necessary to introduce an additional planning aspect that generates optimal control trajectories for the continuous control inputs. These must be coordinated with the planning for the discrete control actions.

It is clear that there are significant open questions and challenges in developing a comprehensive contingency planning capability. However, given the significant benefits of the Burton approach in terms of fast performance and completeness of analysis, it is well worth investigating these questions to understand how the approach might be extended.

3.3.4 Contingency Planning under Uncertainty

Given a current situation, it is important to identify the most likely faults, and the probability of contingency plan success, so that contingency planning is focused on solving the most likely problems, and so that when alternative contingency plans are available, the one with the highest likelihood of success can be chosen. This would leverage cached safe mode and mode transition graphs, as discussed previously, but it would inform decisions by using knowledge of probabilities to focus attention on the most likely faults, and the control actions most likely to successfully handle the faults.

Considerable research has been performed in the area of automatically computing likely faults. Phadke and Thorp [18, 22] studied hidden failures and failure modes, and defined the concept of *regions of vulnerability* and *vulnerability indices*. A hidden failure is a permanent defect that will cause a relay or a relay system to incorrectly and inappropriately remove circuit elements as a direct consequence of another switching event.

Each hidden failure has a region of vulnerability associated with it. The region of vulnerability is the region in which, if a fault occurs, the hidden failure will be exposed. A large region of vulnerability implies that a component, which may have a hidden failure, is critical to correct operation of the power system. The relative importance of each region of vulnerability is assigned a vulnerability index, which is a measure of the priority or sensitivity of the region. One of the measures that can be used to establish this index is the stability or instability of the system following some power system contingencies. Through this analysis, critical protection systems in which hidden failures would have a major impact on cascading failures or blackouts of the power system can be identified. These protection systems are candidates for increased monitoring.

In related work, McCalley [13] defines a security index that is a combination of probability of instability due to a contingency event, and the impact of this instability. The former is a combination of the probability of instability given the event, and the probability of the event itself. The impact factor is a combination of cost associated with replacing unsupplied energy with more expensive sources, and political/regulatory cost, particularly if consumer energy demand is not met. The goal is to choose limits in terms of risk associated with defined events known to potentially cause specific instabilities. The limits are characterized in terms of operating parameters. This approach can be leveraged in optimization algorithms that take risk into account, such as the Iterative Risk Allocation algorithm, which will be discussed in Section 4.

Most of the work on security indices has focused on steady state analysis. An example of use of security indices for dynamic analysis is given in [4]. In this approach, rather than using a single security index, multiple indices were developed to capture the change in various aspects of the power system state. These indices describe the change in the conditions of the power system between the pre-fault steady-state and the post-fault clearing state. Indices include change in generator rotor angles, change in bus voltages, and change in generator rotor speeds. These indices are used to rank contingencies, and ultimately, to classify contingencies into two distinct groups: definitely harmless, and potentially harmful. The latter are the ones that require further investigation. An additional aspect of this particular research is the use of a Neural Network to attempt to automatically perform this classification, based on the indices. The advantage of this is that the classification can be performed quickly. However, a key disadvantage is that the answer may be wrong; Neural Networks require exhaustive training that provides adequate coverage of all possible situations. If the training data set is missing some situations, then the Neural Network may not give the right classification.

Besides determining security indices based on probabilistic calculations, it is also important to estimate the probability of a contingency plan's success. Work in this area has been very limited. Nguyen and Pai [15] have developed an approach based on trajectory sensitivity analysis and forward simulation of dynamics. In this approach, it is assumed that a set of anticipated contingencies is known. For each contingency in this

set, a forward simulation is performed, and sensitivity analysis is performed. If there is a danger of violating a constraint, a load transfer calculation is performed. While this is a useful tool, it does not provide a comprehensive approach in that the discrete aspects of the system are not included. For example, there is no notion of probabilistic mode transitions, which would allow for a systematic analysis of past, current, and possible future contingency events.

Recent work in the autonomous robot planning community has developed the concept of dynamic controllability of flexible plans [7]. This concept utilizes plan flexibility, as expressed through operating and goal constraints, to provide guarantees of successful plan execution given specified bounds on disturbances. Associated with these guarantees are automatically derived control policies that must be used for the guarantees to be true. Although this new work is promising for control of general autonomous systems, significant work remains to investigate how it might be applied to electrical networks. The basic approach would be to represent contingency handling using plans to be executed. Plan execution results in control actions that return the system to a safe state.

Further work in the areas of likely fault identification, and likely successful plan actions would focus on systematic computation of failure probabilities of equipment, systematic consolidation of this information into vulnerability indices to help focus attention, and systematic evaluation of success probabilities for contingency plans.

3.3.5 Pre-emptive Actions

Related to, but distinct from the previously discussed contingency security concepts is the use of pre-emptive or pro-active actions that make the current mode, and possible target modes safer and more robust to disturbances. Little work has been done in this area, although the work of Nguyen and Pai [15] and Effinger [7] provide useful starting points. In particular, the notion of maximizing dynamic controllability at all times should be explored. This could be used to take the system from a current safe state, to an even safer state, according to some appropriate security metric. Risks and costs associated with such maneuvers would have to be taken into account and weighed against the probabilistic improvement in security.

4 Risk-bounded Optimal Power Flow and Unit Commitment

4.1 Optimal Power Flow Problem and Requirements

Optimal power flow and unit commitment problems involve deciding which generation sources should be active, and at what level, over a particular time interval, in order to optimally meet demand, as well as deciding how to optimally route the generated power to the loads. Demand is provided as a probabilistic forecast. Besides demand constraints,

an algorithm solving the optimal power flow and unit commitment problem must consider network transmission constraints, and cost considerations. Good decisions can result in large cost savings for electric power providers, and for customers.

4.2 Current Approaches to Optimal Power Flow

Optimal power flow and unit commitment problems are currently solved using MILP and MINLP problem formulations. These formulations divide a time horizon of interest into fixed time intervals, and calculate optimal unit activation and activation levels for each time interval. The formulations include constraints for demand, generation capacity of each generating unit, transmission constraints, and costs for generation and transmission.

A key shortcoming of the current approaches is that they do not adequately deal with uncertainty in demand, and also in generation capacity. In particular, the current emphasis on renewable energy sources requires incorporation of uncertainty models for these volatile energy sources. For example, the energy output of wind turbines and solar panels is highly dependent on the weather.

A second shortcoming is that these approaches are not fully integrated with tools that are used for contingency planning. A common framework that incorporates optimization of power flow into risk-bounded operating procedures and contingency handling is needed.

4.3 Promising Research Directions

4.3.1 Risk-bounded Optimization

There has been significant, recent work in the development of risk-sensitive optimization algorithms for control of autonomous vehicles and also, of micro-grids [16, 17]. Results indicate that these approaches merit further investigation into how they may be applied generally to macro-grid problems. These algorithms optimize performance while managing risk due to uncertainty of demand, weather conditions, equipment failures, and other uncontrollable events. This will address the fundamental problem of instability of supply and demand, which is exacerbated by the introduction of renewable resources. A risk-sensitive optimization algorithm anticipates and meets user goals and employs resources as efficiently as possible, while managing risk of failure to user specified levels.

The Iterative Risk Allocation algorithm [16, 17] ensures that the risk of failing to meet demand is always kept within user specified bounds. This algorithm can also be distributed, resulting in a market-based approach to risk allocation that will utilize the symmetric connectivity of the grid as an insurance mechanism, by allocating and

distributing risk appropriately throughout the network. Thus, the risk-sensitive optimization algorithm dynamically and autonomously optimizes energy performance based on changing environmental conditions, usage profiles, and user preferences, which include specifications of acceptable failure risk.

The approach is based on the Robust Model-Predictive Control (RMPC) paradigm. RMPC uses a stochastic dynamics model to generate the predicted outcome of current control inputs, and optimizes the sequence of control inputs based on prediction to ensure that state constraints will be satisfied with a certain probability over a time window (called a planning horizon).

The more recent approaches to optimal planning and control with risk management feature two key innovations. First, is the concept of iterative risk allocation, which ensures that the risk of goal failure is at or below user specified levels, while maximizing expected efficiency by allocating risk where the greatest savings is realized. Second, is enabling connected entities on the grid to reduce risk and improve sustainability through collaboration, by implementing risk allocation in a distributed manner using a market-based formulation of risk allocation.

4.3.2 Risk-bounded Optimization through Iterative Risk Allocation

A key challenge in optimizing energy management systems is balancing operating cost optimization with risk of failure events. This challenge can be addressed through a risk-sensitive energy management system that is robust to uncertainty associated with generation and consumption. Such an energy management system could be implemented by leveraging a novel technology called Iterative Risk Allocation, which intelligently allocates risk to operational tasks in such a way that operation is optimized, and overall risk is within user-specified bounds.

Sustainable operation of macro and micro energy grids is of critical importance, given the increasing cost and scarcity of non-renewable energy resources. A crucial goal is efficient, optimal operation of such grids. This includes optimal matching of energy sources and consumers. However, given the costs associated with unacceptable performance, component failure, and system failure, limiting the risk of such events is also key to optimizing energy costs.

A key challenge is that optimal operation tends to push a system to its limits, and the addition of uncertainty can cause the system to cross these limits into failure modes. For sustainable communities, uncertainty in user patterns, weather and available resources can increase the risk of failing to meet demand. For example, an automated home heating system that keeps the temperature low during the winter until occupants come home from work runs the risk of failing to meet demand if an occupant gets home earlier than expected. Similarly, a microgrid that turns off diesel generators and relies on wind

energy runs the risk of failing to meet demand if the wind dies suddenly, and the generators cannot be started quickly enough.

This challenge could be addressed through development of a risk-sensitive energy management system that is robust to uncertainty associated with generation and consumption. Such a system would address the fundamental problem of instability of supply and demand, which is exacerbated by the introduction of renewable resources. The system would use a distributed, market-based approach to risk allocation that utilizes the symmetric connectivity of the community as an insurance mechanism, by allocating and distributing risk appropriately throughout the network. Thus, the control architecture dynamically optimizes energy performance based on changing environmental conditions, usage profiles, and user preferences, which include specifications of acceptable failure risk.

A key feature of the risk-sensitive energy management system is the ability to maximize efficiency and utility with respect to user specified goals. Achieving this capability would leverage previous work on goal-directed optimal planning and control of autonomous systems. The control architecture would employ learned models to achieve user goals over time, to anticipate user needs, to shape this demand, and to plan how resources may be most effectively utilized towards achieving these goals.

In the Iterative Risk Allocation approach, operating constraints are transformed into chance constraints, to represent limits on risk. For example, they can be used to require that the probability that the room temperature is within the comfortable range $T_{in}(k) \in [T_{in,min}, T_{in,max}]$ over the planning horizon must be more than $1 - \Delta$ where Δ is a *risk bound*, which is typically a small number such as 0.01. To represent uncertain plant dynamics, the formulation uses a stochastic dynamics model, expressed as an equality constraint. This is used, for example, to express that the room temperature at a certain time step is a function of the room and outside temperature at the previous time step, solar heat input that can be controlled by adjusting the state $o(k)$ of shades, for each time increment k, and the heat input from the HVAC system. Note that the outside temperature $T_{out}(k)$ and the solar heat input $Q_{Solar}(k)$ are random variables for all k. Therefore the future room temperatures $T_{in}(k)$ are also random variables. It is assumed that the mean value (i.e. prediction) and the probability distributions are known.

Specifically, stochastic optimal control problem using chance constraints is formulated as

$$Pr\left[\bigwedge_{k=k_0}^{k_0+K}\left(T_{in,min} \le T_{in}(k) \le T_{in,max}\right)\right] \ge 1-\Delta \qquad (1)$$

and stochastic dynamics equations

$$T_{in}(k+1) = f\left(T_{in}(k), T_{out}(k), o(k), Q_{Solar}(k), Q_{HVAC}(k)\right) \qquad (2)$$

where the first term is the chance constraint, and the second is the stochastic dynamics model.

The difficulty of solving the optimization problem lies in the chance constraints. Since the probability is defined on the conjunction of the constraints on temperature at all time steps in the time window, its probability distribution is multi-dimensional. In general, integrating a probabilistic distribution over high-dimensional space is computationally difficult.

Iterative Risk Allocation addresses this issue by decomposing the chance constraint using the concept of risk allocation. The chance constraint can be decomposed as follows:

$$\begin{aligned} &\bigwedge_{k=k_0}^{k_0+K} Pr\left[T_{in,min} \le T_{in}(k) \le T_{in,max}\right] \ge 1-\delta_k \\ &\sum_{k=k_0}^{k_0+K} \delta_k \le \Delta \end{aligned} \qquad (3)$$

It can be proven that 3 is a sufficient condition of the original chance constraint by using Boole's inequality: $Pr[A \wedge B] \le Pr[A] + Pr[B]$. Therefore, any solution for the optimization problem with a decomposed chance constraint (3) is guaranteed to be a solution for the original optimization problem. Note that in (3), probabilities are defined on a single state variable. Therefore, it has a univariate probabilistic distribution, which is computationally easy to handle.

In (3), δ_k is an individual risk bound for the kth time step. In other words, it is an allocation of risk to the time step. The total amount of risk is bounded by the original risk bound Δ, just as in a resource allocation problem. Since $\delta_{k0} \ldots \delta_{k0+K}$ are under-constrained, they also have to be optimized.

The IRA algorithm can optimize the risk allocation as well as the control inputs. It starts from an arbitrary risk allocation, and improves it by iteration. In each iteration, redundant risk at inactive constraints is reallocated to active constraints. By allowing more risk at critical constraints (i.e. active constraints), performance improvement is achieved. Mathematically, IRA exploits the convexity of the optimization problem, and improves the risk allocation using a descent algorithm.

To scale the IRA algorithm to a large distributed system, the Market-based Iterative Risk Allocation (MIRA) has been developed recently. This is a market-based decentralized optimization algorithm for multiagent systems under stochastic uncertainty. It builds upon the paradigm of risk allocation, in which the planner optimizes not only the sequence of actions, but also its allocation of risk among state constraints. The concept of risk allocation is extended to multi-agent systems by highlighting risk as a resource that is traded in a computational market.

Significant open issues remain to be addressed with Iterative Risk Allocation approaches. These include adapting the IRA algorithm so that it can be used in a goal-directed context, adapting the algorithm so that it can be used in an on-line receding-horizon context (rather than off-line), and validating and improving the distributed version of the algorithm. Additionally, open challenges include integration of these algorithms with the afore-mentioned Mode Reconfiguration algorithms, resulting in an integrated system for optimal power flow and contingency handling.

5 Conclusions and Future Work

The goal of energy independence through the use of renewable energy sources is now clearly established as a national priority. Exploiting these new resources requires a reformulation of power grids so they can incorporate a range of alternative energy sources, such as wind, solar and biomass, and alternative storage capabilities, to be incorporated wherever they are most effectively generated and utilized. In addition, to dramatically reduce our nation's carbon footprint, our consumption of energy resources, independent of their form, must become dramatically more efficient. This requires that our communities be highly efficient in their utilization of these resources.

The problems described in the previous sections are among the most important for achieving this goal. We have proposed a set of novel, recently developed technologies that may be well suited for addressing these problems. In particular, they address shortcomings of current approaches by incorporating advanced automation techniques that have been successfully applied in the control of spacecraft, underwater vehicles, and other autonomous agents. We welcome the opportunity to investigate further how these approaches might be used in electrical grid applications.

6 Appendix

There are two variations of Hybrid Mode Estimation algorithms: static and dynamic. These are discussed in more detail in the following two subsections.

6.1 Hybrid Mode Estimation with Static Numerical Constraints

In order to control a large, complex system such as an electrical grid, it is necessary to estimate its state. This is accomplished by combining information from system sensors with information from a model of how the system behaves. The information provided by the model allows for estimating hidden states that are not directly measured by sensors, and also, for correcting sensor errors. The state of an electrical grid is hybrid: it has discrete and continuous components. For example, the settings of circuit breakers in an electrical network define discrete operating modes for the system. Continuous state includes quantities like power and voltage in the lines of the network.

In current practice, simple weighted least squares techniques are used to estimate parameters and states that minimize discrepancies in observed values. These techniques compute estimates for the continuous state, but they do not account for discrete mode changes, such as the opening and closing of circuit breakers, which cause significant discontinuities in behavior. In order to be accurate, models for electrical grids must include discrete and continuous state, and this hybrid state must be estimated.

Hybrid Mode Estimation (HME) uses hybrid models to estimate hybrid state. In the HME framework, the behavior of discrete modes is governed by state transitions between modes, which may be probabilistic, and may be conditioned (guarded) on continuous state variables and the discrete states of other components. Each component will then be modeled by a probabilistic hybrid automaton (PHA), comprised of a set of component modes, continuous state variables, observables, guarded probabilistic transitions, and continuous equations associated with each mode.

To estimate PHA state, we build upon the theory of *Bayesian Filtering*, a versatile tool for framing hidden state interpretation problems. A Bayes Filter operates on a model described by a tuple $\langle \Sigma, O, P_\Theta, P_\tau, P_O \rangle$, where Σ and O denote sets of feasible modes s_i and observations o_j. The initial state function, $P_\Theta(s_i)$, denotes the probability that s_i is the initial mode. The state transition function, $P_\tau(s_i(t), s_j(t+1))$ denotes the conditional probability of a mode transition, and observation function, $P_O(s_i(t), o_j(t))$, denotes the conditional probability of an observation given a mode.

The current belief state is updated according to the following prediction/correction equations:

$$b_{t|t-1}(s_i) = \sum_j P_\tau\left(s_i(t) \mid s_j(t-1)\right) b_{t-1}(s_j)$$

$$b_t(s_i) = \frac{P_O(o_k \mid s_i) b_{t|t-1}(s_i)}{\sum_j P_O(o_k \mid s_j) b_{t|t-1}(s_j)} \qquad (4)$$

Here, $b_t(s_i)$ indicates the belief that the system is in mode *i*.

HME [6, 22, 23, 26 - 29] performs Bayesian Filtering for Probabilistic Hybrid Automata. This is implemented using the architecture shown in Fig. 3. The Hybrid Markov Observer accepts a PHA and a set of observations as input, and computes the belief state update according to (4). It uses a *Mixed Logic Linear Program* (MLLP) solver to find the most likely combinations of discrete mode and continuous state that satisfy all logical and continuous constraints. The latter are provided by a continuous model, which specifies continuous constraints for each discrete mode.

HME computes a sequence of state estimates, each of which is a tuple $\hat{x}_k = \langle \hat{x}_{d,k}, p_{c,k} \rangle$, where $\hat{x}_{d,k}$ is the estimate of the discrete mode, and $p_{c,k}$ is the continuous state estimate, expressed as a multi-variate PDF with mean $\hat{x}_{c,k}$ and covariance matrix P_k. Tracking the complete set of hybrid states is computationally intractable; hence HME instead maintains a set of most likely discrete and continuous state trajectories. In the architecture shown in Fig. 3, discrete mode is updated by the Hybrid Markov Observer through a generalization of HMM update. It formulates a set of MLLP problems where the value function corresponds to likelihood, and the constraints are the logical and continuous constraints that must be satisfied. The MLLP solver thus computes the most likely feasible combinations of discrete mode and continuous state.

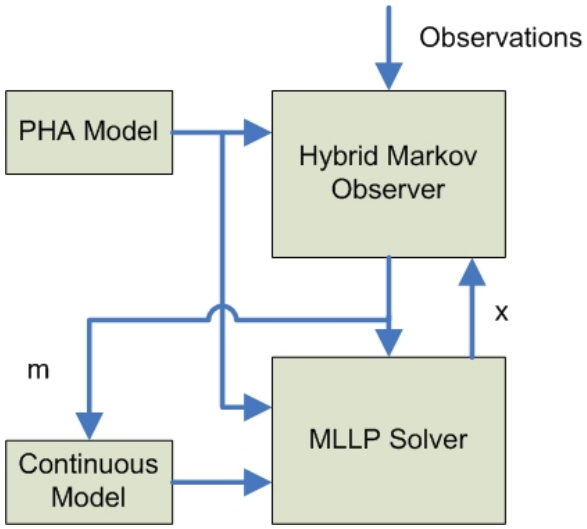


Figure 3: Hybrid Mode Estimator Architecture.

This approach is applicable to problems where the continuous constraints are algebraic equations. For example, steady state network power flow computations can be expressed this way. Thus, the approach could be used to estimate the on/off state of circuit breakers, generators, and loads in a network, as well as the continuous power flows. It could also be used to detect the presence of line faults (breaks and short circuits), though not the short-term transient aspects of such conditions.

Problems that require consideration of dynamics require inclusion of differential equations in the continuous model. Such problems are handled by an extension of HME, described in the next sub-section.

6.2 Hybrid Mode Estimation with Dynamic Numerical Constraints

In the extended HME framework, dynamics are modeled using stochastic difference equations of the form

$$\begin{aligned} x_{c,k} &= f\left(x_{c,k-1}, u_{c,k-1}, x_{d,k-1}\right) + v_{s,k-1} \\ y_{c,k} &= g\left(x_{c,k}, u_{c,k}, x_{d,k}\right) + v_{o,k} \end{aligned} \tag{5}$$

where $v_{s,k-1}$ is a random variable representing the model error, and $v_{o,k}$ is a random variable representing the sensor error. Estimation of dynamic continuous state is accomplished by incorporating a bank of extended Kalman filters into the HME architecture, as shown in Fig. 4.

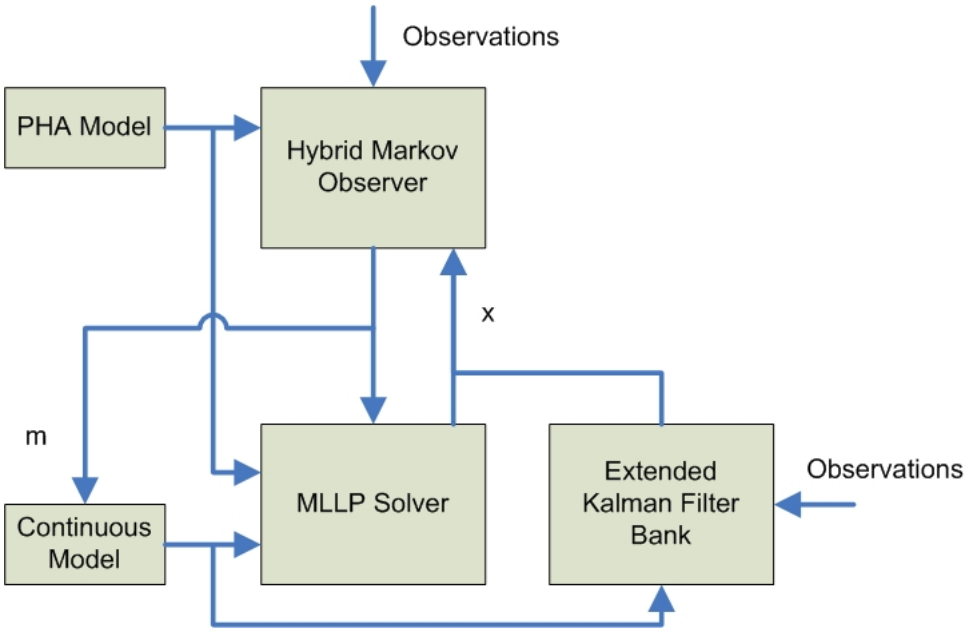


Figure 4: Hybrid Mode Estimator Architecture with Extended Kalman Filter Bank.

In this architecture, continuous state estimates are updated through a generalization of Kalman Filter update, with the Hybrid Markov Observer switching in the appropriate dynamics model based on the current mode estimate. Thus, the continuous and discrete sub-systems pass their respective estimates to each other.

This approach allows for modeling the dynamics of transient conditions. For example, the transient effects of line breaks and short circuits, start-up or shut-down of generators, and start-up or shut-down of loads can be modeled this way.

References Cited

[1] N. Balu, T. Bertram, A. Bose, V. Brandwajn, G. Cauley, D. Curtice, A. Fouad, L. Fink, MG Lauby, BF Wollenberg, et al. On-line power system security analysis. *Proceedings of the IEEE*, 80(2):262–282, 2002.

[2] L. Blackmore, S. Gil, S. Chung, and B. Williams. Model learning for switching linear systems with autonomous mode transitions. In *Decision and Control, 2007 46th IEEE Conference on*, pages 4648–4655. IEEE, 2008.

[3] L. Blackmore, S. Rajamanoharan, and B.C. Williams. Active estimation for switching linear dynamic systems. In *Decision and Control, 2006 45th IEEE Conference on*, pages 137–144. IEEE, 2007.

[4] V. Brandwajn, ABR Kumar, A. Ipakchi, A. Bose, and SD Kuo. Severity indices for contingency screening in dynamic security assessment. *Power Systems, IEEE Transactions on*, 12(3):1136–1142, 2002.

[5] S.H. Chung and B.C. Williams. *A decomposed symbolic approach to reactive planning*. Citeseer, 2003.

[6] J. De Kleer and B.C. Williams. Diagnosing multiple faults. *ARTIFICIAL INTELLIG.*, 32(1):97–130, 1987.

[7] R. Effinger, B. Williams, G. Kelly, and M. Sheehy. Dynamic Controllability of Temporally-flexible Reactive Programs. ICAPS, 2009.

[8] M. Greytak and F. Hover. Robust motion planning for marine vehicle navigation. In *Proceedings of the 18th International Offshore and Polar Engineering Conference*, pages 399–406, 2008.

[9] A. Lamadrid, S. Maneevitjit, T.D. Mount, C. Murillo-Sánchez, R.J. Thomas, and R.D. Zimmerman. A "SuperOPF" Framework. 2008.

[10] H. Li and B. Williams. Generalized conflict learning for hybrid discrete/linear optimization. *Principles and Practice of Constraint Programming-CP 2005*, pages 415–429, 2005.

[11] O.B. Martin, S.H. Chung, and B.C. Williams. A tractable approach to probabilistically accurate mode estimation. In *Proceedings of the Eigth International Symposium on Artifical Intelligence, Robotics, and Automation in Space*. Citeseer, 2005.

[12] O.B. Martin, B.C. Williams, and M.D. Ingham. Diagnosis as approximate belief state enumeration for probabilistic concurrent constraint automata. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 20, page 321. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

[13] JD McCalley, AA Fouad, V. Vittal, AA Irizarry-Rivera, BL Agrawal, and RG Farmer. A risk-based security index for determining operating limits in stability-limited electric power systems. *Power Systems, IEEE Transactions on*, 12(3):1210–1219, 2002.

[14] T. Nagata, H. Watanabe, M. Ohno, and H. Sasaki. A multi-agent approach to power system restoration. In *Power System Technology, 2000. Proceedings. PowerCon 2000. International Conference on*, volume 3, pages 1551–1556. IEEE, 2002.

[15] T.B. Nguyen and MA Pai. Dynamic security-constrained rescheduling of power systems using trajectory sensitivities. *Power Systems, IEEE Transactions on*, 18(2):848–854, 2003.
[16] M. Ono and B.C. Williams. An efficient motion planning algorithm for stochastic dynamic systems with constraints on probability of failure. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 2008.
[17] M. Ono and B.C. Williams. Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint. In *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*, pages 3427–3432. IEEE, 2009.
[18] AG Phadke and JS Thorp. Expose hidden failures to prevent cascading outages [in power systems]. *Computer Applications in Power, IEEE*, 9(3):20–23, 2002.
[19] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
[20] T. Sakaguchi and K. Matsumoto. Development of a knowledge based system for power system restoration. *Power Apparatus and Systems, IEEE Transactions on*, (2):320–329, 2007.
[21] DJ Sobajic and Y.H. Pao. An artificial intelligence system for power system contingency screening. *Power Systems, IEEE Transactions on*, 3(2):647–653, 2002.
[22] S. Tamronglak, SH Horowitz, AG Phadke, and JS Thorp. Anatomy of power system blackouts: preventive relaying strategies. *Power Delivery, IEEE Transactions on*, 11(2):708–715, 2002.
[23] B.C. Williams, M. Hofbaur, and T. Jones. Mode estimation of probabilistic hybrid systems. In *Proceedings of the International Joint Conference on Artificial Intelligence*. Citeseer, 2001.
[24] B.C. Williams, M.D. Ingham, S. Chung, P. Elliott, M. Hofbaur, and G.T. Sullivan. Model-based programming of fault-aware systems. *AI Magazine*, 24(4):61, 2003.
[25] B.C. Williams and W. Millar. Decompositional, model-based learning and its analogy to diagnosis. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 197–204. JOHN WILEY & SONS LTD, 1998.
[26] B.C. Williams and P.P. Nayak. A model-based approach to reactive self-configuring systems. In *Proceedings of the National Conference on Artificial Intelligence*, pages 971–978, 1996.
[27] B.C. Williams and R.J. Ragno. Conflict-directed A* and its role in model-based embedded systems. *Discrete Applied Mathematics*, 155(12):1562–1595, 2007.

Discussant Narrative

Mode Reconfiguration, Hybrid Mode Estimation, and Risk-bounded Optimization for the Next Generation Electric Grid

Bernie Lesieutre
University of Wisconsin-Madison

This paper sets out to introduce a variety of advanced network optimization techniques for potential application to electric power systems. The tools are intended to aid the decision making process for a range of power system problems. Roughly, the techniques listed in the title of the paper are associated with the following list of analyses and decision making activities; they may also apply more broadly:

1. Mode Reconfiguration. Determining the optimal steps to move from one system state to a more favorable state, employing novel reactive planning tools.
2. Hybrid Mode Estimation. Enhancing traditional state estimation to include additional temporal information.
3. Risk-bounded Optimization. Optimally allocating resources while maintaining the risk of failure within bounds.

The techniques are interesting and potentially useful, and the problems plausible. The paper suffers significantly from an inadequate review of traditional and present practices in the industry. The reader is led to believe that most power system analyses are ad hoc. It is difficult to compare the novel techniques presented in the paper to those in practice or that appear in the literature.

Mode Reconfiguration

Mode reconfiguration problems arise naturally in systems in which the transition from one operating state to another may be decomposed into a series of discrete steps. The optimization problem is then to determine the best set of steps to move from a starting state to a target state, accounting for risk of failure. The paper quotes application to spacecraft, naval, and chemical process systems. At a high level one can easily conceive of situations in which a tool that identifies a series of control actions could be of use to the grid. This may be directly applicable to changing configurations of certain meshed distribution system, designating sequences of start up from a black start, and

highlighting possible actions to move from an insecure state to a secure state after some contingency. The latter is the topic discussed in the paper.

The method for mode reconfiguration presented in the paper considers purely discrete steps. The authors note that the extension to power system applications will need to allow direct control over purely discrete variables, and continuous variables. That does need to be emphasized again; many actions taken in operation involve redispatching generating resources: changing set points on continuous variables. The authors note two other needed extensions. It is typical in prior application that a specific target state is identified. In the power system area, an acceptable target space need be allowed to move from an insecure state to any secure state. And lastly, the authors note the well-known issue in power systems: scale. It is often challenging to modify tools to work for very large systems.

From the presentation of the Burton tool in the paper it is difficult to initially assess the ultimate applicability of that tool to power system problems. This should be a topic for further discussion and clarification.

Hybrid Mode Estimation

Hybrid mode estimation is offered as tools that may enhance situational awareness by improving techniques for state estimation. In this area, the proposed technique would compute multiple sequences of state estimates to track the most probable system configurations. The most probable are determined using Bayesian tests.

This is an interesting approach, and should provide for interesting discussion. In particular, certain probabilities are needed to pursue this approach. If these were known for all the components in the system, then a means to sift through the many likely similarly probable states is needed. Perhaps the temporal data will be rich enough to do so. Discussion also needs to address whether there exist significant deficiencies in current practice.

Risk-bounded Optimization

The goal of risk-bounded optimization is to find an optimal solution to a particular problem while maintaining the risk of failure within a user-specified bound. This is a perfectly reasonable goal, and this type of approach is continually discussed in relation to power system security. In particular, should we blindly apply an (N-1) or similar contingency-based, security-constrained optimization when a probabilistic approach seemingly makes more sense? An (N-1) approach may be computationally tractable, but most contingencies do not merit consideration, while many (N-2) and higher contingencies warrant consideration due to high likelihood and/or terrific consequence.

In practice, in the planning stages, reliability organizations do study well beyond (N-1), and do consider multiple events that are deemed likely to occur. Nevertheless, a firm probabilistic approach has not been pursued. This is likely due to both, the lack of confidence that accurate probabilities can be established, and the difficulty in assessing the severity of events. Identifying key vulnerabilities, and anticipating events that could lead to large cascading outages, are vibrant areas of research. Discussion could address the practical implementation of a probabilistic approach to risk assessment at this time.

Recorder Summary

Mode Reconfiguration, Hybrid Mode Estimation, and Risk-bounded Optimization for the Next Generation Electric Grid

HyungSeon Oh
National Energy Renewable Laboratory (NREL)

This paper proposes a tool to integrate recent advancements in mode reconfiguration, estimation, and optimization for power system operation and planning. Shortcomings of the current techniques are presented, and a new tool that addresses these shortcomings is proposed.

This summary includes comments and discussions between the author and audience in the following order:

1. Logistics behind the current tools
2. Requirements to build the new tool suggested
3. Open questions for future research

1 Why do we use the current tools?

Due to the preventive paradigm, (N-1) contingency is widely used. The consideration of the contingencies yields an emergency control for reliable operation of power systems. It is worthwhile to note that the emergency control actions do not require probability information that is difficult to obtain.

2 Suggested requirements for building the new tool

2.1 Mode reconfiguration

Major issues: probability assessment, importance of contingency

The new tool provides a way with no exhaustive search for contingency planning and therefore, it would decrease the computation cost for large-scaled power system operation. For robust modeling, it is necessary to define a safe mode in terms of risk, cost, and safety.

There are many questions related to modeling rare events and constructing joint probability distribution that may have a low possibility with significant impact such as the contingency of the Japanese nuclear power plants. Computation of probabilities for such rare events is very difficult, mainly for two reasons: 1) limited understanding in modeling and 2) unavailability of the data. Conventional normalization technique with a long-term horizon may yield a very low probability. Even if possible, the evaluation of the joint probability distribution would increase the computation cost significantly.

A well-designed visualization would help to construct the contingency plan along with providing reasons for contingency suggested by the authors.

2.2 Hybrid Estimation

Major issue: learning algorithm, dynamic and static modeling

It would be important for a model to include properties related to power systems such as dynamic behavior and high voltages. This can be achieved from machine learning. The major concerns for the learning algorithm are effectiveness in computation and correctness of the model. Advances in technology and knowledge from human experts can improve model-based learning significantly. A rule-based statistical learning algorithm in social networking and hybrid method outlined in this paper may be also extended to model dynamic behavior.

2.3 Risk-bounded OPF

Major issue: formulation of the OPF, computation cost

There are various planning schemes in different time scales. For example, power system expansion planning is for a 5 to 10 year time horizon while LMP- and ancillary service markets are less than a day. It is not clear which planning horizon to that this risk-bounded OPF is applicable to. Once the time scale is defined, relevant constraints to the problem need to be formulated, which may include an emergency plan.

The new tool does not fully address the importance of contingencies that are taken into significant consideration in the current (N-1) contingency planning. A stochastic optimization provides a way to combine the probability and the importance of contingencies, which also increase the computation cost significantly. The author suggests that a contingency action for combinatory operation should be constructed by developing a metric based on vulnerability and probability.

3 Open questions for future research

Evaluation of joint probability distribution and modeling a rare event, which play a key role in the construction and the applicability of this tool, are not available. Optimal power flow should be evaluated in terms of economic value. It can be improved in terms of robustness to find a solution and inclusiveness of stochastic behavior. The tool needs to address the decision making process when the optimality limits contingency actions. The approach suggested in this paper may be combined with the proposal titled "Coupled Models for Planning and Operation of Power Systems on Multiple Scales" by Michael Ferris at the University of Wisconsin.

White Paper

Model-based Integration Technology for Next Generation Electric Grid Simulations

Janos Sztipanovits and Graham Hemingway, ISIS, Vanderbilt University
Anjan Bose and Anurag Srivastava, Washington State University

1 Introduction

The national electric power grid is going through transformational reform to be efficient, reliable and secure smart electric grid in line with the national energy security mission [1, 2, 3, 4]. The electric power grid constitutes the fundamental infrastructure of modern society and can be defined as the entire apparatus of wires and machines that connects the sources of electricity with customers. The smart electric grid utilizes enhanced communication, digital information and control technology to improve efficiency and reliability of system. There have been several smart grid technologies, and algorithms developed in view of new smart grid framework, which needs to be validated and tested. *Testing and validation of these smart grid algorithms can be done through modeling and simulation of integrated smart electric grid* models [5, 6].

The electric grid is composed of a wide range of networked physical, computational and human/organization components. Heterogeneity is pervasive:

Physical components include transmission lines, power generation equipment, substations, distribution lines, control devices, sensors, actuators and communication lines resulting in dynamic multi-layered physical networks.
Computational (cyber) components include Energy Management Systems (EMS), Supervisory Control and Data Acquisition (SCADA) systems, control algorithms, planning and operational tools, and communication data processing algorithms.
Human and human organization components include operators, distributed command structure for decision making, policies and organizations.

In addition, heterogeneity of the required models is increased by the multitude of simulations spanning a wide range from network planning and formulating operating strategies, to operation management of the transmission system, dispatching of generation units, design of control strategies, vulnerability and dependability analysis and many more. As the electric grid's aggregate behavior is influenced by the physical, computational and human/organizational components, modeling of system operation

and dynamics is very complex. With ongoing smart grid activities, complexity and heterogeneity is supposed to increase in multiple ways. According to a recent report [1] current grid modeling and simulation efforts typically focus on a narrow set of issues (such as physical system dynamics). It appears that raw computational horsepower is available, but the modeling and simulations software tools that allow the effective harnessing of that power are lagging behind.

Achieving the Smart Grid vision requires the efficient integration of digital information, communication systems, real time-data delivery, embedded software and real-time control decision-making. These domains expand beyond traditional power engineering algorithmic advancements. There remains a lack of high-fidelity models capable of simulating interactions of electric grid with communication and control infrastructure elements for large systems. Modeling interdependencies of infrastructure accompanying the power grid, including sensor, control, communication network and software, is a challenge. Currently, there are several gaps to efficiently model the integrated smart grid system in efficient way:

1. There is no single tool to model the power engineering, communication, and control system.
2. There is no single tool to model the system at transmission and distribution level with attention to all details.
3. There is need of tools to combine data that is multi-rate, multi-scale, multi-data, multi-user, multi-model from different domains.

In summary, simulation-based evaluation of the behavior of the electric grid is complex, as it involves multiple, heterogeneous, interacting domains. Each simulation domain has sophisticated tools, but their integration into a coherent framework is a very difficult, time-consuming, labor-intensive, and error-prone task. This means that computational studies cannot be done rapidly and the process does not provide timely answers to the planners, operators and policy makers. Furthermore, grid behavior has to be tested against a number of scenarios and situations, meaning that a huge number of simulations must be executed covering the potential space of possibilities. Designing and efficiently deploying such computational 'experiments' by utilizing multi-domain tools for integrated smart grid is a major challenge.

This paper addresses this important challenge by integrating smart grid modeling tools from diverse domains in a single coherent framework for integrated simulation of smart grid.

2 Modeling and Simulation of Electric Power Grid

2.1 Overview

Power system researchers have devoted significant efforts in the past to the development of sophisticated computer models for the analysis of various electric power grid applications [7, 8, 9, 10]. Mathematical formulation is an important consideration in model selection, because it has significant implications on solution requirements, computer memory and speed requirements, and overall complexity. Attributes that affect model complexity include [7]:

- Linear vs. nonlinear
- Discrete vs. continuous vs. hybrid
- Static vs. dynamic
- Plain simulation vs. directed (constrained) optimization
- Probabilistic vs. deterministic
- Aggregated (equivalent) vs. disaggregated (detailed) network and component representations
- Regional (e.g., pools, control areas, independent system operators, or utilities) vs. national scope

Modeling and simulations tools that use these existing models are generally limited to specific analysis. Power systems are multi-scaled in the time domain, from nanoseconds to decades, as shown in Table 1 [11].

Action/operation	Time frame
Wave effects (fast dynamics, lightning caused over voltages)	Microseconds to milliseconds
Switching over voltages	Milliseconds
Fault protection	100 milliseconds or a few cycles
Electromagnetic effects in machine windings	Milliseconds to seconds
Stability	60 cycles or 1 second
Stability Augmentation	Seconds
Electromechanical effects of oscillations in motors & generators	Milliseconds to minutes
Tie line load frequency control	1 to 10 seconds; ongoing
Economic load dispatch	10 seconds to 1 hour; ongoing
Thermodynamic changes from boiler control action (slow dynamics)	Seconds to hours
System security monitoring	Steady state; on-going
Load Management, load forecasting, generation scheduling	1 hour to 1 day or longer, ongoing
Maintenance scheduling	Months to 1 year; ongoing
Expansion planning	Years; ongoing
Power plant site selection, design, construction, environmental impact, etc.	10 years or longer

Table 1: Multi-scaled dynamics in power grids

The overall behavior of electric grid systems emerges through the simpler, more independent behavior of many individual components. However, modeling of power system components for these various time scales requires a variety of models for the different applications for functions [7, 8, 9, 10, 11, 12, 13]:

- *Planning tools* are used for determining the optimal long-term capacity expansion in generation, transmission and distribution, and are also used for evaluating policy measures proposed for the energy sector. From the power system modeling point of view, the components themselves change during the planning time frame of years
- *Operational support tools* are designed to support real-time operations or short-term operational planning and reliability assessments, as well as for operator training. From the modeling viewpoint the physical power system does not change during this short timeframe but the operational variables like loads and generation change widely.
- *Operation tools* are used in real time for doing automatic control or providing visual monitoring and manual control by the operator. From the modeling viewpoint the real time conditions have to be modeled.

2.2 Planning Tools

Planning tools are needed to simulate future scenarios of large additions of generation, transmission lines and distributed power to the grids. Models must deal with the intermittency of these sources, optimize the transmission and resource mix, quantify important metrics (e.g., economics, security, and environmental impact), and assess alternative solutions. Planning tools also look at addition of conventional generations for generation expansion analysis; transmission line additions and distribution systems in expanding neighborhoods. Some of the available tools for planning are NEPLAN, ADICA, PROMOD, TPLAN, PSS/E [12]. The main objective is to optimize the capital investment required to meet the long term growth in demand and policy changes.

2.3 Operational Support Tools

Operation support tools are generally used to decide operations strategy for the next few hours to few days. The physical components of the power grid do not change but the operation has to be optimized for the short term prediction of the expected scenarios. Some of these tools are DIgSILENT, ATP, EMTP, GE PSLF, PSCAD, PSS/E, RTDS, SimPowerSystems, ETAP, PSAT, TSAT, VSAT, POM, ASPEN etc. [12]. The main objective is the cost optimization of the operation for the near future. This is different from the planning tools in that the optimization is of operational cost rather than capital cost.

2.4 Operation Tools

These tools are available in the control centers to handle real time measurements to operate the power system in real time reliably and cost optimally. The tools must

support the automateic controls and operation as well as the manual monitoring and operation by the human operator. Some main features of these tools include: SCADA, State Estimation, Voltage Control, AGC, Optimal Power Flow, Contingency analysis. Although optimization of cost is still important, the larger focus is on reliable operation, especially the withstanding of disturbances with minimum disruption of supply to customers.

2.5 Simulation Tools

Although the tools mentioned above are software packages that perform certain functions, there is tremendous overlap of models and simulation among these different tools. From the viewpoint of modeling and simulation, it is more convenient to look at the mathematical analysis for the different models and simulations. One has to keep in mind that the many different mechanical and electrical components in the power grid can be non-linear, non-continous and have to be simulated in both steady-state and transient conditions, the latter in various time frames. Thus closed form analysis is not possible and the analytical methods are always by simulation in different time frames. Some analysis methods are as follows [12]:

- Steady-State Analysis: assumes sinusoidal voltages and currents used to determine power flow; voltage profiles, losses (active and reactive), reactive power compensation, and transformer tap positions. The program in this category are power flow, state estimation, contingency analysis, optimal power flow as well as for abnormal system conditions, such as short circuit and harmonics. PowerWorld software is one good example of steady state analysis tool.
- Electromechanical Dynamic Analysis: used to verify that the power system will not become unstable or even collapse during major disturbances, and to determine the operating limits of the system. The main program in this category is the transient stability program which is essentially the simulation of nonlinear ODEs with time steps in milliseconds. PSAT software is good example of this type of tool.
- Electromagnetic Dynamic Analysis: used to simulate electromagnetic behavior like switching transients, lightning strikes, etc. Program in this category are EMTP, ATP, RTDS etc. Essentially this requires solving nonlinear partial differential equations in microseconds time steps.

In addition some of the customized tools has been developed by different utilities. As indicated in this section, current power grid modeling and simulation efforts are often piecemeal and application specific, because the individual tools help to address a narrow set of issues.

3 Modeling and Simulation of Control and Communication in Electric Grid

Current electric grid systems have two way communication and control systems limited to mainly transmission system. Smart grid innovations will expand the use of digital information, computers, automated control and communications to transmission, generation and distribution system to much extent [8, 14, 15]. This will also lead to new sensor technologies, database management systems, data processing capabilities, computer networking facilities, means of cyber security, and visualization tools for asset operators and managers. The information technology systems that comprise a major part of this change need to address a number of security and management challenges for the data generated by power systems and for resources consumed by distributed applications. With smart grid activities, the number of devices installed throughout the network, the amount of data to be handled, the rate of data delivery, latency requirements and other interdependencies all need to be revisited.

In addition, achieving an adequate level of cyber-security and protection of ownership and access rights will be increasingly difficult, because underlying protocols will need to constantly change as they respond and adapt to an increasingly complex energy infrastructure and a highly dynamic risk environment. As shown in the Fig. 1, smart grid will have the integrated communication, computation, sensor and actuator networks embedded within physical power system including transmission, distributed energy resources (DER), substation and distribution.

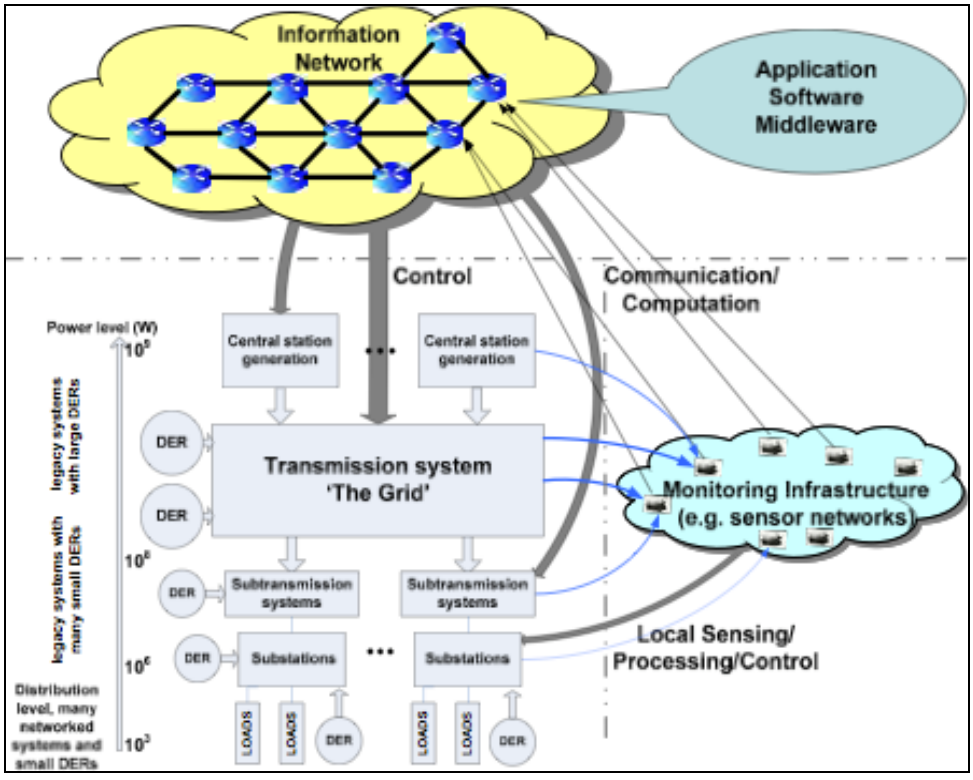


Figure 1: Infrastructure for future grid

The information network will merge the capabilities of traditional EMS and SCADA with the next generation of substation automation solutions. It will:

1. Enable multi scale networked sensing and processing.
2. Allow timely information exchange across the grid.
3. Facilitate the closing of a large number of control loops in real time.

This will ensure the responsiveness of the command and control infrastructure in achieving overall system reliability and performance objectives.

3.1 Computing Infrastructure for Data Management and Communication

The future power grid will have significantly greater data and resource management requirements driven by a large number of data sources and resource-consuming applications that process, store, and share data. The data will be generated by an increasing number of field devices such as IEDs (Intelligent Electronic Devices), automated stations that pre-process the data, integrated EMS systems at control centers, and inter control center communications. Even today IEDs generate a lot of information that is stored in substations but not sent to control centers due to band- width limitations and proprietary non-compatible data formats. Control centers archive a lot of historical data that is available but rarely used. The transition to a smarter grid means that an enormous amount of data will be collected that will require analysis and verification so that it can be transformed into usable information.

Currently, communication network in power system utilizes radio, microwave, broadband over power line, fiber optics, wireless and Ethernet [14, 15, 16]. It is by no means a new observation that the existing communication infrastructure is inadequate to handle smart grid requirements for data transfer [14]. The inadequacy is manifest in several ways. Some of the examples as pointed out in [14] are i) the slow response of grid operators to contingencies— as occurred during the 14 August 2003 blackout—is partly due to inadequate situational awareness across company and regional boundaries, ii) special protection scheme (SPS) deployment is extremely expensive, due largely to the cost of the point-to-point links between substations and iii) New and potentially beneficial approaches for control of the grid that allow faster adaptation and protection (such as hierarchical control) are not feasible without better communications

.

In short, the existing communication infrastructure limits the types of controls and protection that can be deployed. There is increasing consensus that inflexibility is the existing communication system's principal deficiency. GridStat [14, 15] is a middleware framework that provides a simple API based on abstractions for publishing and subscribing to status variables and status alerts. Achieving the required communication flexibility will require a new communication architecture. A high-bandwidth network

capable of intra-substation, inter-substation, and control center communication will be required to facilitate large-scale data collection, local processing, and distilled information transfer. A key feature of this architecture will be communications management via advanced middleware to provide coherent real time data delivery [15]. Several tools are available to model communication network other than GridStat, for example NS-2, OPNET and OMNET++ [17]. Each tool has its own set of advantages and limitations which must be accounted for during tool evaluation.

3.2 Computing Infrastructure for Control

Data management for the future grid will be one of the biggest challenges in real time networked operation and control. For example, massive amounts of data can introduce unacceptable latency (slow down communications) into energy systems if processed, stored, or utilized on today's architectures for power system control. Furthermore, as redundant, missing, and poor quality data substantially increases, so will the likelihood of faulty readings, which could cause energy systems to crash with wrong or suboptimal control action. With presence of higher rate phasor data from phasor measurement unit provides enhanced dimension of control for power system [18]. Efficient and timely control actions based on power system analysis tool can help to avoid blackouts [18].

Most of the power system tool can also be used to model the power system control including RTDS, PSS/E, PSCAD, ETAP, Simulink etc. Most of these tools can model the control components of the electric grid with limitations, but can capture characteristics of the control components depending on quality of the developed model. Developing a good model for some of the control components such as FACTS can be challenging, especially in dynamic condition.

Modeling and simulation are key tools to model properties of information network, communication and control to evaluate their interaction with the power system. In the following section we review technologies that have the potential for scaling up to the requirements of smart grid research and operation management.

4 Modeling and Simulation of Heterogeneous Systems

The smart grid is a heterogeneous Cyber Physical System (CPS) composed of different physical and computational components interacting through communications networks (see Fig. 1). They represent different simulation domains that are modeled by means of domain-specific modeling languages (DSML) with unique semantics and simulation tools. We cannot assume that a single "universal" modeling language will fit the needs of all smart grid domains. That approach would require building new modeling and simulation tool suites which is cost prohibitive, particularly in light of existing validated model libraries in sub-domains (such as networking or grid dynamics) that frequently represent major previous investment. The only practical solution is a multi-model

simulation approach that facilitates the precise integration of heterogeneous, multi-model simulations. However, multi-modeling creates three fundamental challenges:

1. *Compositional modeling.* There are numerous examples for the tremendous cost and time required for developing large, monolithic universal modeling language standards. In addition, the process never stops because modeling languages are intimately intertwined with progress in design and analysis technologies - rendering hard fought standards quickly obsolete. The rapidly changing, extremely dynamic field of smart grid design and operation requires an agile, lightweight solution for the rapid construction and adjustment of heterogeneous models using well understood modeling constructs – without compromising semantic precision.
2. *Precise, explicit specification of multi-aspect DSMLs*. Without rigorous, formal specification of constituent modeling languages, integrated simulations are hard to build. System simulations disintegrate into regimes/stovepipes (such as grid dynamics, network dynamics, controller dynamics, and system-level architecture models) that are formed around tool suites developed and used in relative isolation. Since end-to-end simulations require both automation and iteration across these regimes, the separation into independent islands of simulation tools is costly, leads to inconsistencies, and lacks the capability to represent cross-domain interactions. Furthermore, it locks companies into vendors-specific solutions whose interest is to maintain this isolation.
3. *Tool infrastructure for multi-model simulation.* Development of dedicated simulation tool suites for rapidly changing heterogeneous application domains is cost prohibitive. Without resolving the dichotomy between domain specificity and reusability, end-to-end simulation of heterogeneous systems will be restricted to slowly changing and relatively large, homogeneous application domains where a viable tool market can form.

The only practical solution is a multi-model simulation approach that facilitates the precise integration of heterogeneous, multi-model simulations. In this section we describe integration approaches starting with a short summary of component simulators for smart grids.

4.1 Simulators for Physical and IT Components

Modeling and simulation applications require a wide range of domain specific simulators, as described in section 2. Listed below are examples for widely used simulators. The scope of this description is restricted to illustrate the challenges of creating multi-modeling simulation environments.

4.1.1 Power Grid Simulators

Other than tools discussed in section 2, the *Real time Digital Simulator* (RTDS) is a example of power system simulator [51] designed for continuous real time operation. Its time resolution enables the simulation of electromagnetic transients. The simulator uses discrete time semantics with time steps that can be adjusted. The RTDS architecture enables users to interact with a running real time simulation via digital and analog I/O channels. Models of the simulated power system are defined using a graphical modeling language, RSCAD. The language includes predefined control blocks and scripts for controlling the simulation execution. Due to its real time operation, RTDS can be connected to external devices that allow closed loop testing of real-world physical equipment.

Connecting RTDS to a heterogeneous simulation environment requires handling it as a "physical component", since the RTDS simulation clock is strictly a real time clock. Consequently, multi-modeling simulation environments that include RTDS require time management that synchronizes *logical time* of discrete event simulators with real time.

MATLAB/ Simulink is example of non-real time power system simulation for transient analysis.

4.1.2 Network Simulators

Network simulators compute the behavior of a network by calculating the interaction between the different network components (computer nodes, routers, switches, data links, etc) using mathematical formulas. Most network simulators use *discrete event* semantics, in which "events" are ordered according to their "time stamps", and processed in chronological order. Network simulators include a modeling language and a simulation engine that maintain the event queues and schedules event processing by forwarding the simulation clock. Modeling languages enable a user to define a network topology, specifying the nodes on the network, the links between those nodes, the traffic between the nodes, and to facilitate the specification of various network protocols. Other than tools discussed in section 3, some well known examples for network simulators are NS-2/NS-3 [52], OMNeT++ [53] and OPNET [54].

Network simulators use logical time while computing progress. Since simulating network protocols requires detailed packet-level traffic simulation, simulation performance for large-scale, broadband and heterogeneous networks can be a major issue.

4.1.3 Monitoring and Control Simulators

Power system monitoring and control tools are discussed in section 2. In addition, Simulink is a widely used modeling and simulation tool for dynamic systems. The simulator uses continuous time semantics with component interactions modeled via

shared variables. Simulink models can be integrated with StateFlow models providing a hybrid system simulation tool. A key requirement for integrating a simulator with HLA is the ability for controlling time progression. The key mechanism for synchronizing the clock progression of the Simulink model is the basic time-progression model for S-function blocks. During its execution, the Simulink engine consults each block in a model about when it can generate an output. With all S-function blocks, code must be supplied, via an implementation of the *mdlGetTimeOfNextVarHit()* method, to respond to this request from the simulation engine. The integration code in an S-function block uses this method to synchronize the model with the simulation integration framework and allow simulation time within Simulink to progress only when the time management allows it to proceed.

An additional advantage of Simulink is the wide selection of libraries for the simulation of hardware and software components of digital relays. Diverse libraries of components are available in domains such as power systems, communications networks, and control systems.

4.1.4 DEVS

DEVS is a widely used formalism for modeling and analysis of generic discrete event systems (DES). DEVS has several extended formalisms for combined continuous and discrete event systems, for parallel DES, for real-time DESs and many others. Extensive run-time support and numerous simulation engines and libraries make DEVS a common tool for simulating software systems based on DEVS models.

Additionally, human in the loop can also be modeled using appropriate tool for human decision for power system control. In previous work we have modeled human organizations and decision processes using Colored Petri Nets [44].

4.2 Heterogeneous Simulation Example

The goal of heterogeneous simulation is to tie together disparate models of individual aspects into a cohesive and coordinated simulation of a larger system. This process starts with the individual models, each possibly residing in distinct native domains. An example may help to illustrate this process.

Fig. 2 provides an overview of our example. One model contains a simulation of a small-scale power grid with sensing and control equipment. A distributed data network relays the sensor data to a centralized location. The collected information is fed into a control center model that manages grid operation and sends command information, via the network, back to the grid.

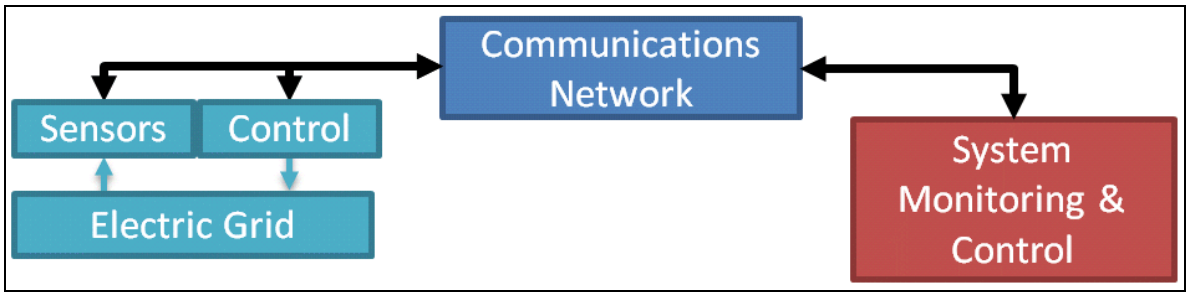


Figure 2: Heterogeneous simulation example

This figure shows three different models interacting to create a single simulation of the system. The black lines denote where two models interact. It is these interactions that the heterogeneous simulation infrastructure must coordinate.

In this example, a small power system grid is simulated using a Simulink model, though it easily could use other power system simulation software tools in real time (e.g. RTDS) or in non-real time. Fig. 3 shows the details of this small 5-bus model including two generators, three transmission lines, two transformers and three loads. In addition, three loads are modeled to be switched ON at given time during the run time simulation as a step change. Focus is on integration of simulation tools instead of modeling an actual large power system. A power system model with a large number of components can be integrated in same manner. The concept is to use existing libraries in different domains (that represent significant prior investment) in order to minimize cost and to reduce the effort needed for integrating simulations. The reuse of models can achieve reductions the cost, complexity, and time required to develop accurate integrated simulations.

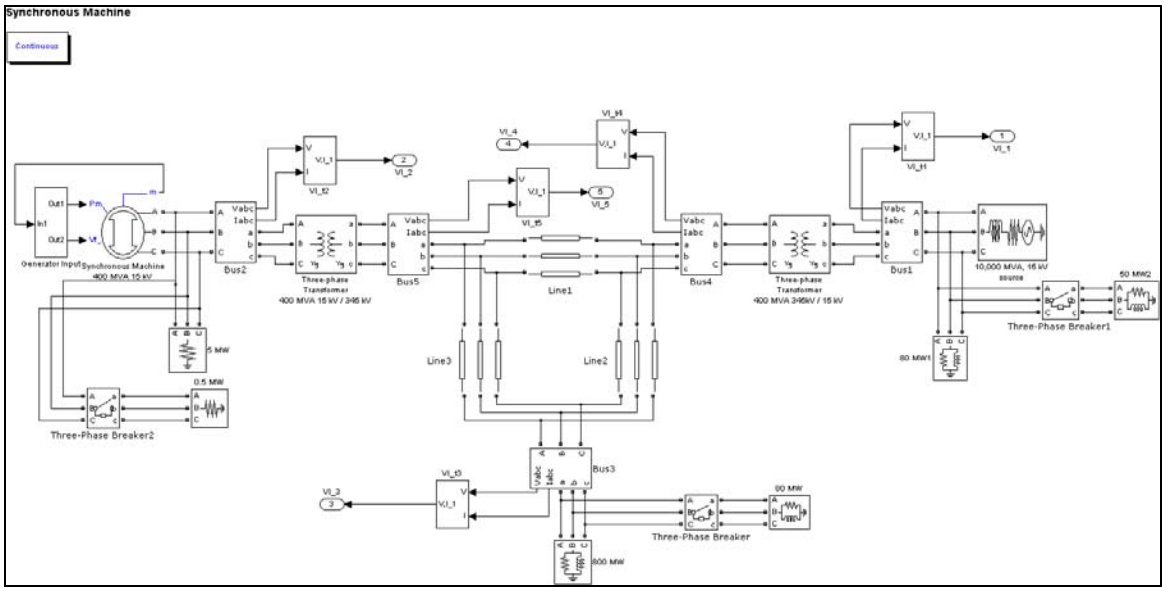


Figure 3: Example power grid model

The communications network for the example system is modeled using NS-2 and is shown in Fig. 4. It is comprised of several end-point nodes connected via LAN links and routers. The experiment runs for 15 seconds and during the time interval {7, 12}, TCP background traffic is injected into the system from node 'backtraf1' to 'backtraf2'.

Using NS-2, we are able to provide the following simulation capabilities over wireless and wired networks with different topologies. For wired networks, different link bandwidth, delay, and loss configurations can be modeled. For wireless network, different node deployments, channel capacity, and radio assignment configurations, different MAC protocols can be modeled. For both wireless and wired networks, we can experiment over a variety of transport layer protocols, such as different versions of TCP protocols and UDP, queue management schemes, and routing algorithms.

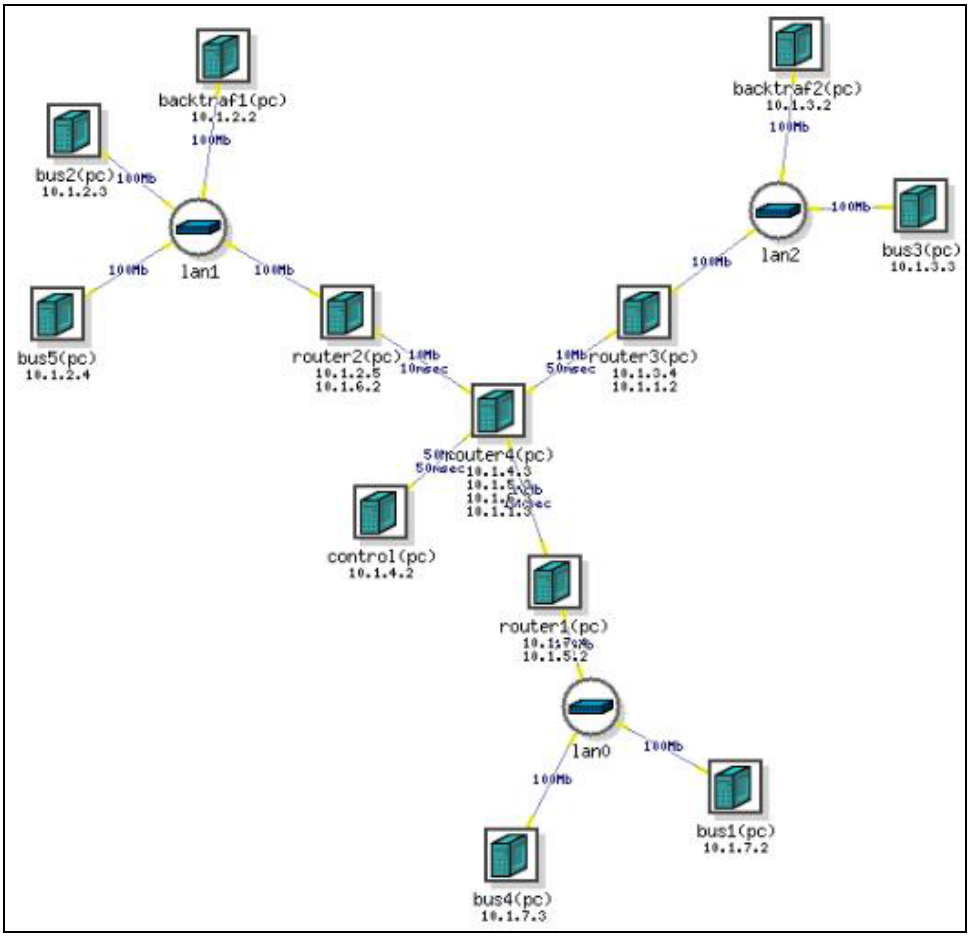


Figure 4: Example communications network model

Finally, the grid monitoring and control model is again simulated using Simulink. Fig. 5 illustrates the root level design for the power grid monitor using a state estimation tool. Note that, in this experiment, we did not model the feedback control from the control center back to the power grid. Sensor data, relayed via the network, is used to determine operational parameters for the grid. Control information is can be fed back to the grid via the network if needed in the similar manner.

These three models must be integrated together to form a consistent model that faithfully simulates the behavior of the whole system. The integration framework employed must bridge the gap between the varying semantics underlying the simulation platforms and should provide the essential interaction dynamics between them such as the shared messages and objects and coordinated evolution of time.

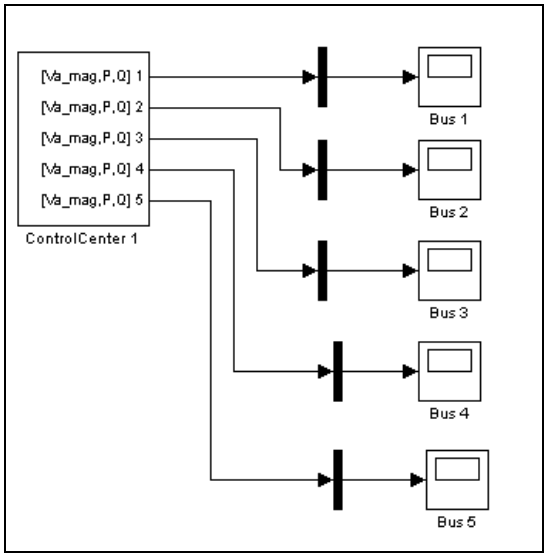


Figure 5: Infrastructure for future grid

4.3 Simulation Integration Frameworks

Heterogeneity and complexity of simulations in smart grid applications defines the following challenges for the M&S infrastructure:

1. Rapid construction and integration of component simulators
2. Effective deployment and real-time/faster-than-real-time execution of those integrated simulators to rapidly evaluate multiple experiments in parallel

The problem is compounded by the heterogeneity of the simulation tools to be used, and the inherent complexities of individual tools. Users need tools to rapidly set up and configure simulation studies, with minimal effort. Below we discuss three simulation integration frameworks that have been used successfully primarily in large-scale military applications.

4.3.1 SIMNET

Formalized frameworks for integrating simulations can greatly reduce development workload while simultaneously allowing for more flexible and scalable simulations. The SIMNET project [20] is credited with initiating research into scalable frameworks for distributed simulation. Prior research efforts focused primarily on rigid integration of a small number of stand-alone simulators. While not a truly heterogeneous platform,

SIMNET was able to integrate simulations for disparate purposes into a single virtual environment.

The SIMNET domain was discrete simulation of battlefields and combat vehicles. Soldiers used SIMNET as a live training environment. Full-scale vehicle mockups were installed in various locations around the world and networked together. The controls of these vehicles and visual displays for their occupants were directly connected to and controlled via SIMNET. The SIMNET language was a simple transaction modeling language with primitives for discrete event simulation and basic conditional logic. The SIMNET language described both the individual simulations and their network interactions.

One contribution of the SIMNET language was its use of dead reckoning, or incremental updates, as a means to provide position updates throughout the simulation. Use of incremental updates simplified network communications and allowed greater scaling of the overall simulation. SIMNET also developed the use of objects and events as a means to communicate environment changes [21]. Objects were owned by one node in the simulation and that node was responsible for communicating incremental state changes of the object. Similarly, events occurred as objects interacted, were created, or were destroyed.

The SIMNET architecture was strictly real-time, in the sense that the overall simulation always assumed human-in-the-loop and was not able to run at either slower or faster than real-time. In part this was due to the nature of the simulation. Individual nodes, and their human occupants, could enter or exit the virtual environment at any time, and each node was expected to meet specific performance guarantees about issuing object and event updates.

4.3.2 DIS

Due to its specific use in military simulation, the SIMNET language and framework did not gain significant acceptance outside of the U.S. military. The follow-on research project to SIMNET was the Distributed Interactive Simulation (DIS) project [22, 23]. Again, DIS was developed by DARPA to support large-scale battlefield simulations. DIS built on the core SIMNET network protocols, but completely divorced the network protocols from the individual simulations. Now, individual simulations could be built using any simulation engine and integrated with the rest of the virtual environment via standardized network protocols.

As was for SIMNET, in DIS individual simulations are responsible for maintaining their own internal state as well as their understanding of the greater environment. Via the integration framework a standardized format is used for communicating object status and all state changes. Individual simulations are able to query the state of an object via the network and are able to filter object update information as desired. Again, time

coordination in DIS is negligible. Time is always assumed to be strictly real-time and individual simulations are held to performance guarantees.

DIS protocol data units (PDU) define the data format and semantics of all interactions. The definition of PDUs is standardized by the DIS controlling body. There have been a number of updates to the collection of standard PDUs, but the context for the standardized PDUs is almost entirely military in nature (warfare, logistics, simulation management, minefield, etc.).

The use of standardized PDUs closely mirrors but predates the current research into ontology based integration methods, see Section 4.3.3 below. The PDUs reflect the view of an individual simulation encompassing all aspects of one whole entity, such as a tank, and not a single subsystem, such as an engine. DIS is not designed for, and not very capable of, integrating simulations of subsystems into a larger system, only of representing entities in a larger virtual environment.

4.3.3 HLA

Though still in use and under development today, DIS has mostly been eclipsed by the High- Level Architecture (HLA) [24]. While the development of the HLA standard was again guided by the military, no aspect of the HLA framework is distinctly military. The HLA is a completely general use integration framework for simulations. An HLA simulation is composed of a federation of individual simulation federates. Shared objects and interactions are defined to which any federate may publish or subscribe. Objects are analogous to OS-style shared memory and are owned by one federate. Interactions correspond to message passing. All federation configuration information is stored in a standardized format text file called the FED file. In theory, this configuration file is portable across different HLA run-time infrastructure (RTI) implementations. The HLA standard does not proscribe a set of standardized interactions or objects.

The U.S. military's OneSAF project [25], the inheritor of SIMNET and DIS's purpose, has defined a standard set of interactions and object for their use of HLA in battlefield simulation. Other organizations have similarly published standardized federation configurations. These standardized sets of interactions and objects again correspond to the ontological approach discussed below.

The HLA standard advanced the flexibility of time control in an integrated simulation. Fundamentally, every federate must maintain a clock corresponding to the logical time internal to its simulation. This clock is distinct from any real-world “wall-time". The standard provides numerous schemes for coordinating logical clock evolution among federates. These can range from completely lacking time synchronization, where one federate can execute arbitrarily far into the future, to completely synchronized, where all federates evolve time within a tightly bound window. Each federate can be configured to be time-constrained or time-regulating, both, or neither. A time-regulating federate's

progression constrains all time-constrained federates. Likewise, a time-constrained federate's advance is controlled by all time-regulating federates. A federate that is neither constrained nor regulating is free to evolve time on its own. Federates that are both, evolve time in tight lockstep. If, for example, all federates can run at least as fast as real-time, and one federate tightly correlates its time advance requests to wall-clock time, then the entire federation can be made to run in real-time. Otherwise, it is possible to execute simulations both faster and slower than real-time.

Each HLA federate defines a step size, lookahead interval and minimum time stamp. When a federate requests to evolve its internal simulation time is does so in increments of step size, which may vary in size from step to step. Lookahead corresponds to the amount of time into the future which the federate guarantees it will not issue an interaction or object update and is generally small compared to step size. When the federate is in a Time Advance Request (TAR) state, minimum time stamp is defined as the federate's requested time plus lookahead. When the federate is in a Time Advance Granted (TAG) state, minimum time stamp is the federate's logical clock time plus lookahead. It is also important to understand that each federate maintains an understanding of all of all other federate's minimum time stamps.

Fig. 6 adapted from [24] illustrates how time advances happen in a federation of two time-regulating and time-constrained federates. In this example, federate A always seeks to advance its clock in steps of size s, while federate B steps are of size 3s. Wall-clock time runs to the right, but has no units to reinforce that there is no mandatory correlation between logical and wall time. Event 1 is federate A issuing a time advance request (TAR) to the HLA run-time infrastructure (RTI) to advance its logical clock by its step size. It cannot advance its logical clock until federate B's minimum time is greater than its requested time. Event 2 is federate B issuing a TAR which immediately causes its minimum to go to T+3s since its step size is 3s. This allows federate A to change to Time Advance Granted (TAG) state and progress its logical clock to T+s. At events 4 & 5 and 6 & 7 federate A issues TAR followed immediately by TAG since federate B's minimum time is still greater. Finally, once event 6 has occurred, federate B can move into a TAG state and advance its logical clock. The whole sequence then begins to repeat itself.

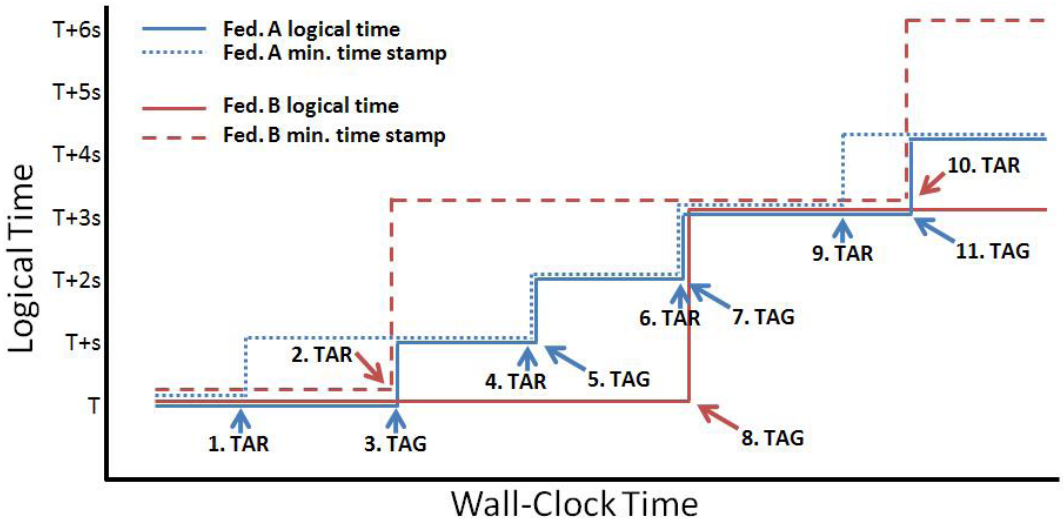


Figure 6: Synchronized time advancement in HLA

The HLA standard still lacks numerous facilities that would be useful for developing integrated distributed heterogeneous simulations. Foremost is any formalized methodology for developing collections of interactions and objects. While a significant advantage of HLA over DIS and SIMNET is its complete divorce of the framework from the subject being simulated, the lack of a standardized lexicon of interactions and objects places a greater burden on integration designers. Another problem stems from HLA's flexibility. SIMNET and DIS simulations tended to be relatively static in the sense that the deployment configuration for a given simulation would change relatively little over time. Because HLA-based simulations are so exible it is desirable to move a simulation from one network of computers to another for development or execution. The HLA standard does not directly address this ability and leaves such functionality up to HLA implementations or integration designers.

Each simulation model and simulation engine must be individually integrated at both the HLA API level and at overall simulation level. Simulation engine-to-HLA is simple technical integration, but simulation model-to-overall simulation integration is completely contextual, depending entirely upon its role in the greater simulated environment. Many research efforts exist relating to integrating various simulation engines via HLA, including: OPNET [26], SLX [27], and DEVSJAVA [28]. As mentioned earlier, the HLA APIs provide run-time support but the problem of model integration is not addressed in these efforts.

Relevant commercial integration software does exist, such as the HLA Toolbox [29] for MATLAB federates by ForwardSim Inc., MATLAB and Simulink interfaces to HLA and DIS based distributed simulation environments by MAK Technologies [30]. These products focus on integration of models running on the same simulation platform and do not provide support for heterogeneous simulations. Additionally, there have also been some efforts on enhancing HLA support by complementary simulation

methodologies such as in [31] and [32]. However, these efforts, similar to those above, pursue HLA integration of isolated simulation tools. Moreover, these efforts do not have any support for model-based rapid integration of simulation tools, and limited, or no, support for automated deployment and execution of the resulting simulation environment.

4.4 Model Integration Frameworks

A model-based approach holds the promise of easing the burden of developing an integrated simulation. An obvious artifact from any such effort would be a federation configuration specifying a set of interactions and objects valid for all possible constituent models. The problem then lies with understanding the domain and capabilities of each constituent simulation model and domain and determining a composition, or integration, of these that meets the goals of the overall simulation.

Multi-paradigm modeling (MPM) [33], or multi-formalism modeling [34, 35], research concerns the methods, tools, and applications related to engineering problems involving the composition of models from multiple different domains, specifically when the set of models derives from multiple modeling formalisms. Similar to MPM, the most challenging problem for distributed heterogeneous simulation is composing domain-specific models in a meaningful way. In order to analyze complex multi-formalism systems, MPM advocates finding methods to convert each constituent formalism into some common formalism. This can be accomplished in several possible ways. First, a "super-formalism" could be found that encompasses the expressiveness of the individual formalisms. Finding such a super formalism is difficult, especially if the sub-formalisms are diverse in their nature. Second, each sub-model could be transformed into some common formalism. Little distinction can be drawn between this approach and the super-formalism approach, in that both require transforming models from one formalism to another while attempting to not alter the semantics of the model. As long as little or no loss of semantic understanding occurs, these approaches are feasible. But, complex systems may involve such diverse formalisms that these approaches do not work. The use of the Formalism Transformation Graph (FTG) [36] somewhat simplifies the process of understanding which transformations are feasible or not. The final approach discussed in MPM does not involve model or formalism transformation. Co-simulation is the coordinated execution of simulations of all sub-models, using their respective formalism. This approach is the basis for systems such as HLA, discussed above.

One of the other previous efforts that relates to heterogeneous simulation is the MILAN framework [37]. MILAN provides a model-based multi-domain simulation of System-On-Chip (Soc) designs and integrates several simulation tools in order to optimize performance versus power consumption. This approach does not leverage the HLA as

an integration framework, is very SoC-specific and does not attempt to be a general engine for heterogeneous simulation.

4.4.1 Ontology-Based Model Integration Approach

Over time communities build a common understanding by using an agreed upon vocabulary. Specific terms and phrases take on special contextual meaning within that community. Ontology-based approaches [38, 39] for simulation integration build upon this concept. A domain ontology is built using expertise and knowledge focused on a tightly-scoped common interest or topic, such as power generation, communication networks or biology. A community of interest ontology is built around a community with a shared set of objectives, such as command and control. Finally, a simulation tool ontology is built to describe the functionality encapsulated within a particular simulation.

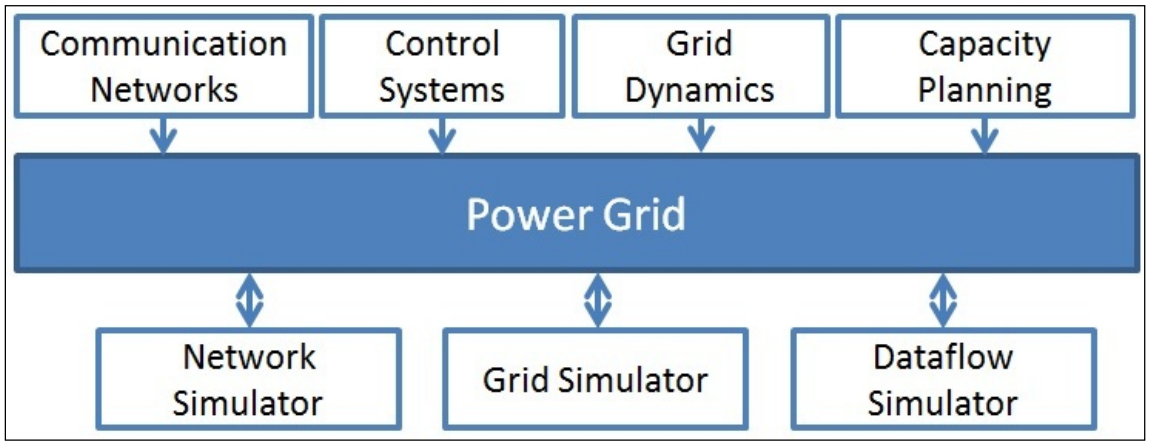


Figure 7: MPM simulation for power grids

Fig. 7 illustrates how these three types of ontologies work together using power grids as the context. At the top, ontologies for specific domains are captured. These individual domains correspond to integral parts of a power grid but are somewhat atomic in their semantics: communication networks, control systems, capacity planning and grid dynamics. Individuals may be experts in these areas but are unlikely to be experts in all. At the center, a community of interest ontology incorporates the domain ontologies and is able to capture the semantics of entire embedded systems. This ontology is not necessarily a super-set of all incorporated domain ontologies as it may discard unneeded aspects from a domain. At the bottom of Fig. 7 are the simulation tool domains. Each tool may have its own formalism and semantics, in our example a network simulator, dataflow simulator and power grid simulator. The community of interest ontology must be able to map onto a set of simulation tool ontologies in order to be simulated. A community must organize itself and come to agreement on the important terms, events and objects in order to develop a common ontology.

The OneSAF HLA federation configuration mentioned above is an excellent example of this. The military defined all of the interactions and object of interest to them and all

simulator suppliers must conform to this standard. But, the simple definition of objects and interactions does not necessarily alleviate the problem of heterogeneous integration.

A deeper understanding of the semantics of these items is necessary and is not addressed in the ontology-based research. Ontologies allow everyone to speak the same words, but not always have the same understanding. An emerging new approach, called *metamodeling*, provides a range of effective tools for placing model-integration on a solid semantic foundation.

4.4.2 Metamodeling-Based Model Integration Approaches

In the most widely accepted terminology, modeling languages are specified using metamodels [40] (the name expresses the fact that metamodels are models of modeling languages). Their purpose is to specify abstractions and their interpretations (semantics) in a formal, mathematically solid way using a metamodeling language.

Metamodels characterize the *abstract syntax* of the domain-specific modeling language (DSML), defining which objects (i.e. boxes, connections, and attributes) are permissible in the language and how they compose. Another way to view this is that the metamodel is a schema or data model for all of the possible models that can be expressed by a language. Using finite state machines as an example, the DSML would support states and transitions. From these elements any finite state machine can be defined.

There are several essential aspects of metamodeling frameworks that make them effective for model integration in heterogeneous simulation environments:

1. Advanced metamodeling frameworks include extensive support for specifying modeling languages. For example, the Model-Integrated Computing (MIC) framework [40] provides methods and tools for the explicit specification of *structural* and *behavioral* semantics of DSMLs using model transformations. Structural semantics defines the set of well-formed models that can be created in a DSML [42]. Behavioral semantics is represented as a mapping of the model into a mathematical domain that is sufficiently rich for capturing essential aspects of the behavior (such as dynamics) [45].
2. Metamodeling enables the development of metaprogrammable tool suites. Metaprogrammability means that complex tools (such as modeling, model management, and model transformation tools) can be customized (effectively "programmed") to suit the needs of specific domains, and thus after this customization they become (domain-specific) tools [46].

There are several alternate metamodeling frameworks that have been developed during the past decade. Examples besides the MIC tool suite are AToM [47], MetaCase [48], Microsoft DSL [49] and the Eclipse Modeling Framework [50].

4.5 Combining Simulation and Model Integration

The simulation and model integration frameworks described above address two key aspects of creating multi-modeling simulation environments. *Simulation integration frameworks*, such as DIS and HLA, address the composition of simulators using discrete event semantics. *Model integration frameworks*, such as specification of ontologies and metamodeling, provide methods and tools for integrating constituent models of heterogeneous systems that are represented in different domain specific modeling languages. These two aspects need to be combined to obtain an integrated multi-modeling simulation environment. In this section we discuss technical solutions using a notional architecture for smart grid simulation. In the discussion we use HLA as a simulation integration platform and metamodeling as a model integration approach.

Building on the example system described in Section 4.2, the left side of Fig. 8 captures the configuration for a distributed smart grid simulator. The configuration includes four simulators: Simulink for power system models, Simulink power system monitoring and control models, NS-2 for communication networks, and a Java application for overall simulation control. The individual simulators are set up using their native modeling languages and connected to the HLA-RTI via an HLA Federate. The HLA federates are responsible for managing information traffic among the simulators and provide services for time management across the distributed simulators. From the point of view of time management, power grid simulation the heaviest computational burden on the simulation architecture, as such its simulation clock evolution tends to be the constraining factor for the entire federate. Controllers are simulated using Simulink, which offers a wide selection of libraries for digital filtering, signal conditioning and other functions relevant to smart grid experiments. Properties of digital communication networks that connect monitors and controllers are essential for understanding overall system dynamics. Therefore, communication networks are simulated using the NS-2 network simulator that provides a wide range of network protocols. Other components of the system, including management of the overall simulation federation, are handled by a custom Java application.

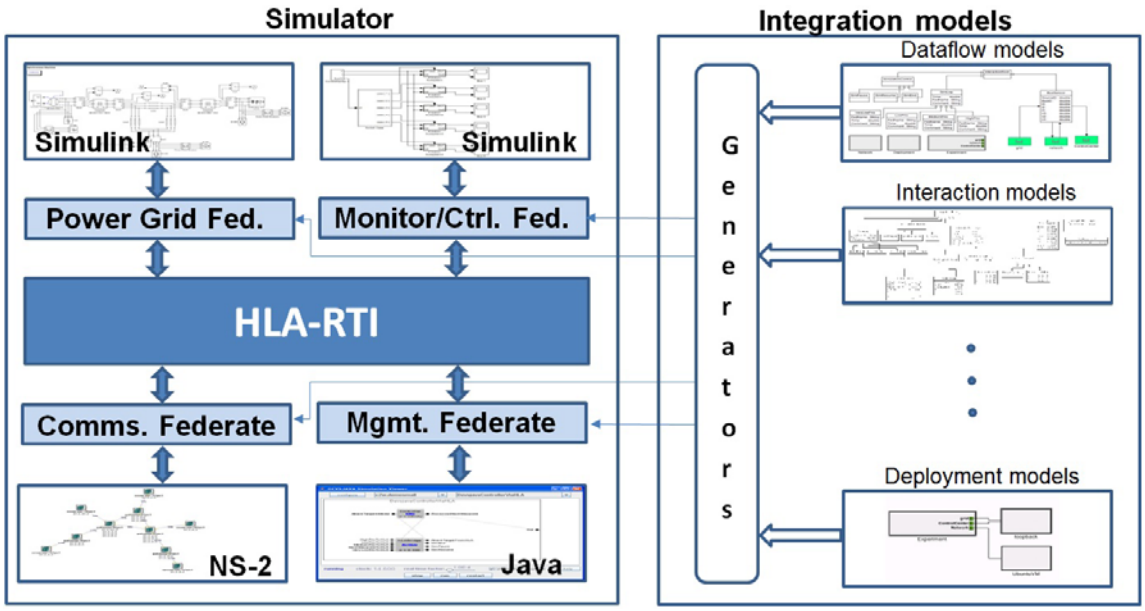


Figure 8: Smart grid simulation with multi-modeling

The right side of Fig. 8 represents components for model integration. The domain specific models describing the power system, controllers, networks and various software components need to be extended with *integration models* that specify the overall system architecture (dataflow models), and interactions among the simulators (interaction models). In the case of large-scale simulations there may be a need for deployment models that provide mapping between the simulation engines and the underlying computation platform. We discuss the challenges in more detail below.

4.5.1 Model Integration Challenge

An integrated simulation that includes models for heterogeneous simulation, including Simulink, NS-2, RTDS and DEVS simulators, requires precise modeling of system level interactions among these components. Since these interactions are implemented by the simulation federates (Power Grid Federate, Monitoring & Control Federate, Communications Federate and Control Federate in our example) interfacing the simulators to the HLA-RTI, the integration model needs to be precise enough for generating the federate code from the models. Below we provide a brief summary of considerations for creating a model integration layer for large-scale heterogeneous simulations. Detailed description of an implementation of the model-based integration approach is available at [44], and an open-source implementation of a prototype system can be found at [55].

The key challenge in model integration is to construct a Model Integration Language (MIL) that provides all of the modeling primitives required to specify the interactions in

a federated simulation. Once the integration model has been defined for a given environment using MIL, a set of reusable model interpreters are executed to automatically generate engine-specific glue-code (the federates) and all deployment and execution artifacts. Since in the example we utilize HLA as the distributed simulation integration framework, the MIL uses discrete event semantics. There are well known techniques for combining discrete event semantics with continuous time semantics (see e.g. [56]), therefore the HLA framework is sufficient for integrating continuous and discrete time simulators (such as Simulink and RTDS) with discrete event simulators (such as NS-2 and DEVS).

Modeling an integrated simulation (or federation in HLA terminology) requires the following categories of modeling constructs: *HLA-Interactions*, *HLA-Objects* and *Federates*. As defined by the HLA standard, *Federates* in a federation communicate among each other using *HLA-Interactions* and *HLA-Objects* – both of which are managed by the RTI. *Interactions* and *Objects*, in an analogy with inter-process communication, correspond to message passing and shared memory respectively. Attributes of these communication elements define delivery method, message order (timestamp or receive order), and other parameters.

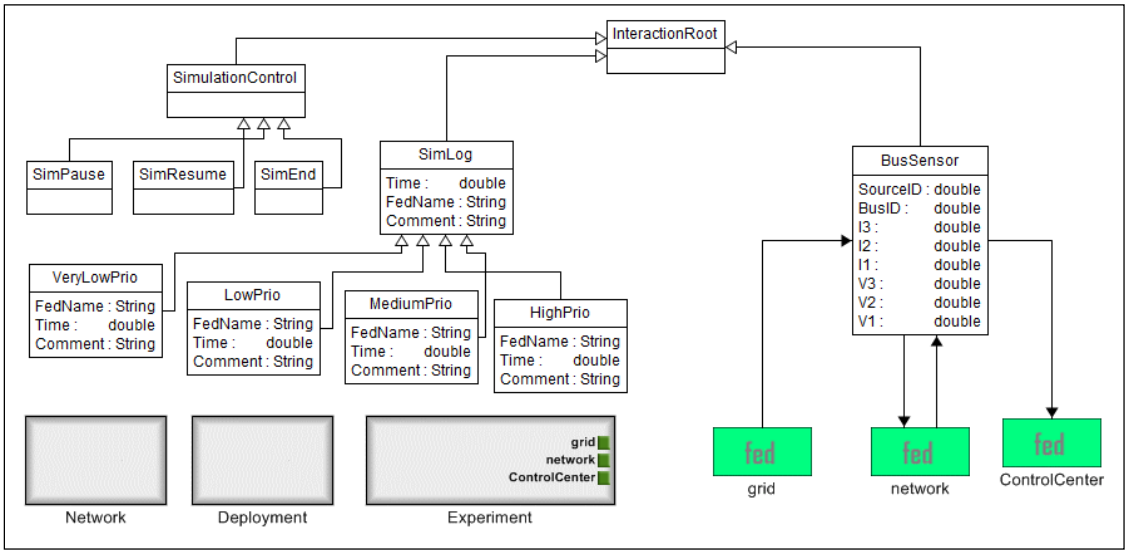


Figure 9: Example publish/subscribe relationship in an integration model

The *Federate* element directly corresponds to any single instance of a simulation tool involved in the federation. Once HLA-interaction and HLA-object models are created, the modeler must define publish-subscribe relations with federates. This is accomplished by connecting federates to interactions or object attributes with directional links. The connections to federates are defined in terms of modeling concepts imported from the native modeling language of the corresponding simulator. Thus, integrating domain-specific models using the MIL models is simply a matter of connecting federates to those interactions and objects with which they have a publish or subscribe

relationship. This greatly simplifies the designer's job, since they no longer need to directly incorporate simulation engine-specific considerations and can focus solely on the high-level interactions. Fig. 9 shows a simple integration model for our example. The model specifies the publish and subscribe relationships between federates (elements shaded) and interactions (elements not shaded).

Once the integration model has been defined for a given environment, a set of reusable model interpreters are executed to automatically generate platform-specific glue-code and all deployment and execution artifacts.

4.5.2 Simulation Integration Challenges

The central issue in integrating simulators in a multi-modeling environment is time management. This is a particularly challenging problem in smart grid simulation, where real-time components (RTDS) are combined with simulated components using logical time, the time resolution spans several order of magnitudes and the simulations need to be executed on a massive distributed computing platform in a tightly coordinated manner.

The HLA standard provides numerous schemes for coordinating time among federates. These can range from completely lacking time synchronization, where one federate can execute arbitrarily far into the future, to completely synchronized, where all federates evolve time within a tightly bound window. Using HLA terminology, all federates must be both "time-regulating" and "time-constrained". As mentioned in Section 4.3.3, a time-regulating federate's progression constrains all time-constrained federates. Likewise, a time-constrained federate's advance is controlled by all time-regulating federates. For time management, the primary attribute of a federate is its *Lookahead* – the interval into the future during which that federate guarantees that it will not send an interaction or update an object. Because in time coordination sensitive simulations all federates are both time-regulating and time-constrained, time progresses only when all federates are ready to proceed and then only as far as the smallest permissible time step across all of the federates. Without both characteristics for all federates, the overall behavior of the simulation can become non-deterministic due to reordering of events in time. Determinism is necessary if scenarios must be executed multiple times without variance in the events or outcomes, such as for scenario analysis. Other uses of simulations, such as training, may not, and therefore the requirement for federates to be both time-constrained and time-regulating could be relaxed.

Time sensitive simulations also assume that all interactions and object updates are strictly time-ordered and must have timestamps. The HLA standard specifies that messages can be sent at any time but may only be received while the federate is waiting for a time-advance-request to be granted. This ensures that all incoming messages will have a timestamp greater than or equal to a federate's current time, i.e. no timestamps are allowed on a message that make a message appear that it was received in the past.

Once a time-advance-request is granted a federate can simulate forward in time and processes incoming messages according to their timestamp order.

Given these assumptions, the operational semantics of a federation become straightforward. Each federate operates in a loop consisting of two steps: request a time advance from the RTI and wait, receive a time advance grant from the RTI, and simulate up to that time. The glue code generated from the integration model must be able to control the simulation engine execution to abide by this scheme.

A complete integration model does not contain information about the detailed execution of each federate. Nor does it replace in any way the internal operational semantics of any simulation engine. Every federate references an engine-specific model (e.g. a Simulink-based grid monitoring and control model) and it is within this model that the details of the internal semantics of the federate are contained. The multi-model simulation environment only builds upon the standard time management and message passing mechanisms provided by the underlying HLA infrastructure.

4.5.3 Deployment Modeling Challenges

Deployment, i.e. mapping a multi-model simulation on an underlying distributed computation platform, seems to be a mundane problem – while the scenarios are simple. As the complexity grows, manual deployment processes quickly began to consume more time than the actual execution of scenarios. A potential solution to this problem is to extend the model integration environment with deployment models. Fig. 7 shows the metamodel of a simple deployment modeling language that extends MIL (we have not shown example metamodel for MIL, see [44]). The language includes the new concepts *Experiment*, *Host*, *Computer*, *Network*, *Deployment* and the concept of FederateProxy imported from MIL. With this extension, a model interpreter automatically generates all of the necessary scripts and files, copies the files to the appropriate computers, and prepares the environment for execution.

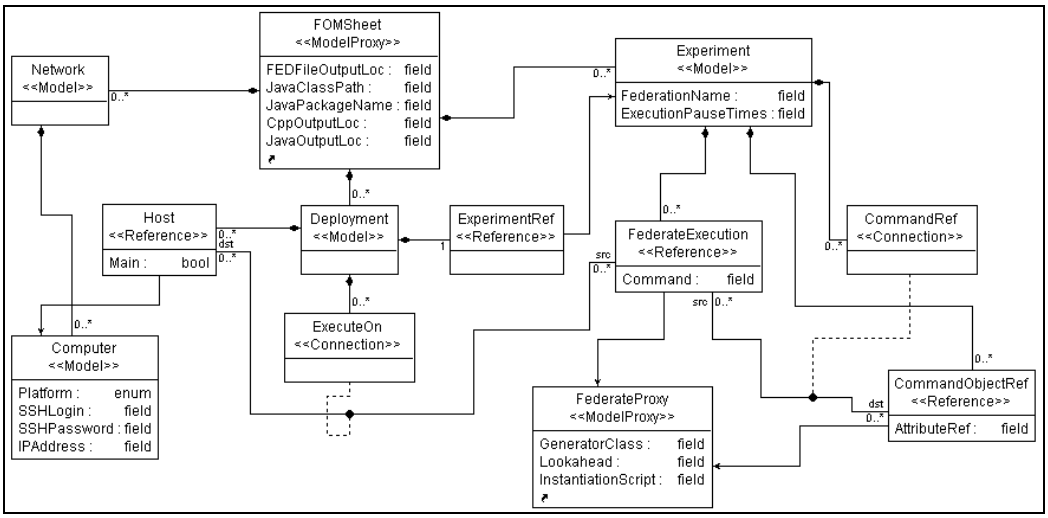


Figure 10: Deployment metamodel

As discussed in previous sections, the overall simulation model is a composition of federates and their relationships via interactions and objects. For any given experiment, a simulation scenario may only utilize a subset of the federates defined in the model. Similarly, each domain-specific model may be parameterized to allow for run-time flexibility. Example for such parameters may be the duration of network disturbance (for the network simulator engine). An experiment is the set of federates included in a specific deployment and their run-time parameterization.

Frequently, an experiment is run on more than one hardware setup. A designer may run the entire simulation on one machine during development, while deploying the simulation onto a cluster for full-scale demonstrations. The network element of the language extension is the physical set of computers involved in a specific deployment.

The deployment element is where an experiment configuration is mapped to a network configuration. Specific federates are assigned to hosts in the network. Thus, allowing complete flexibility in defining which simulation tools execute on which hardware.

A model interpreter reads the deployment configuration from the model and generates all of the script files necessary to support the deployment. In cases where modeling deployments may only be partially specified, for example in very large-scale or rapidly changing environments, the interpreter generates the deployment for whatever portion is defined. Once generated, the environment is fully prepared for experiment execution.

Finally, the generated scripts manage the actual movement of files and code to the various hosts being used. Upon invocation, the scripts remotely connect to each machine and create local copies of all necessary files before the simulation begins execution. The scripts then coordinate the execution of all federates. After the experiment is concluded, the scripts remotely stop all processes, collect output files, and clean all local copies to restore the hardware to its original state.

4.5.4 Customization Levels

The model-based integration framework shown in Fig. 8 serves as an infrastructure for conducting computational experiments. The critical issue from the point of view of usability is the complexity of technologies required for expanding the *infrastructure*, defining *scenarios* and running *experiments*. Table 2 below summarizes these levels, needed expertise, goal of activities and their frequency.

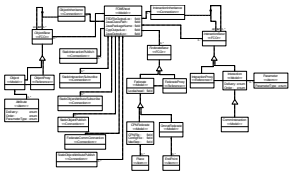
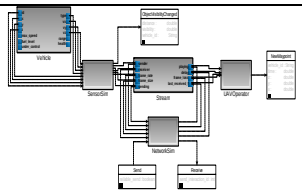
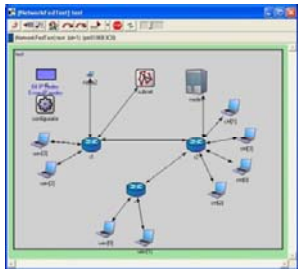
Level	Example	Expertise	Function	Frequency
Infrastructure		Metamodeling. Model transformation.	Adding new model types, integrating new tools.	Incorporating a new discipline or model type in the infrastructure.
Scenario		Power systems, control and networking, interaction modeling.	End-to-end modeling of architectures and experimental scenarios.	Startup activities for new computational studies.
Experiment		Component modeling. Experiment design. Data analysis.	Configure and parameterize component models for individual experiments and measure selected performance data	Performing studies and evaluations by running experiments

Table 2: Customization Levels

1. Extension of the *infrastructure* means the addition of simulation (and analysis) tools with their domain specific modeling languages (DSMLs). For example, the integration of MATLAB/Simulink to the infrastructure is an infrastructure-level customization. According to the discussion above, this requires the introduction of the model element in MIL that represents the MATLAB/Simulink engine, selection of concepts that need to be exported to the MIL metamodel (i.e. signal ports), and the development of the model interpreters that generate the interface code for HLA and the configuration files for the MATLAB/Simulink models.

2. The first step in the development of *experimental* scenarios is the specification of events and combination of events that capture the essential aspects of the scenario. For example, an experimental scenario designed for studying system-level resilience may include disruptions in distribution, increased delays on the communication network caused by cyber attack and a load pattern. The scenario is precisely specified using power system, network, control and other component models using the appropriate modeling languages and the integration models.
3. Experimental scenarios define a repeatable environment for configuring and executing *experiments*. Experiment execution can be highly automated after specifying configurations and parameter ranges for component models.

Computational experiments involve many controllable variables. Conditions for success include a sophisticated infrastructure for information management that is responsible for keeping the models and experiment results consistent and accessible.

4.6 Experimental Results

Using the HLA based integration technology provided by the C2WindTunnel, we constructed an example integrated simulation using the power grid, communication network, and the system monitoring models as described in Section 4.2. The integration model for this experiment is shown in Fig. 9. This example involves a non-real-time power grid model in Simulink sending sensor updates of power measurements via a simulated communication network (using NS-2 models as described previously) to the power grid system monitoring federate (modeled in Simulink). The total run-time for the experiment was 15 seconds and the sensor updates were provided every 100 milliseconds. Three-phase breakers were added to inject an extra load on the system during the time interval {5...10} seconds. The communication network uses the topology described in Section 4.2. Apart from the simulation of network packets that belong to the power gird sensor updates the NS-2 communication model also injects background TCP traffic during the interval {7...12} seconds. The power grid and system monitoring federates (Simulink) were running on a Windows XP machine, and the communication network (NS-2) was running in Ubuntu Virtual Machine. The resultant sensor updates were received at the grid monitoring station after being simulated over the network as shown in Fig. 11, below:

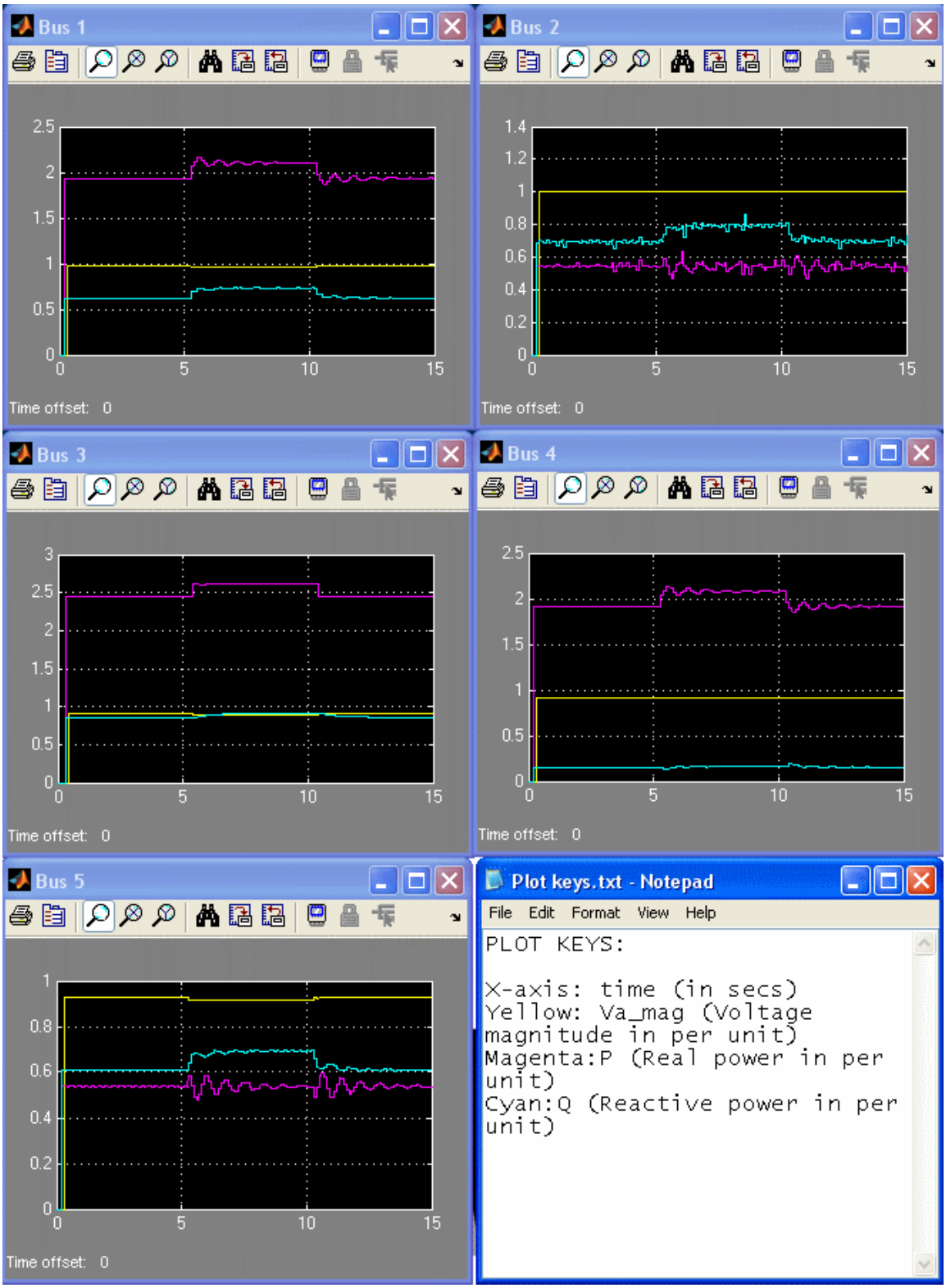


Figure 11: Integrated simulation experimental results

As can be seen from the figure above, the system reaches steady state soon in the simulation (around 1 second). Also, due to the extra load added during the interval {5…10} seconds, the voltage , real power and reactive power varies as expected.

4.7 Problems, Issues and Needs

Based on our understanding of the specific needs for smart grid modeling and simulation and our previous experience with multi-model simulation integration, the next generation smart grid simulation must have the following properties:

1. *It is a robust, easy-to-use software toolsuite that can be used by researchers as well as trained planners and operators alike to evaluate various technical solutions and plans in operationally relevant scenarios.* The key here is that the users of the toolsuite will not be required to address low-level software engineering details, and the configuration of integrated evaluation scenarios should be as simple and intuitive as possible.

   The focus on the software toolsuite is to support the rapid configuration and execution of evaluations. We envision that the end-users will be able to:

   (a) Select the specific simulation tools and the models used in those tools needed in a specific evaluation, e.g., NS-2 models or OMNeT++ models for networks
   (b) Select or import the specific model interfaces needed for a scenario, e.g., relevant simulated hosts in an NS-2 model, and interfaces to control the effects of faults or cyber exploits, e.g., effects of jamming on communication links
   (c) Configure the simulation and evaluation architecture by visually modeling the interactions among the simulated models using a simple publish-subscribe approach (i.e. connecting 'publisher' models to 'subscriber' models – but not necessarily specifying the low-level details of the messages being used)
   (d) Define the full scenario for the evaluation using a time-line: what events happen at what points in time, how those events interact with the component models, e.g. the effect of cyber exploits on power availability in specific regions or customers
   (e) Define the evaluation techniques used: what data gets displayed during the study and in what form, what data gets collected for further analysis or is archived
   (f) Specify what computational resources are requested for the simulation and evaluation (without specifying the low-level details of those resources, like host name, IP address, etc.)
   (g) Analyze the models for correctness and completeness and receive feedback if discrepancies are detected

(h) Execute the study by running multiple simulators in a coordinated manner, possibly under automatic control to collect data for immediate display and off-line analysis.

2. *The toolsuite is based on a carefully chosen set of simulation and analysis tools that are of 'industrial' quality, 'best of class', can incorporate new components, and – preferably – are familiar to the analysts using the system.* The planners are either familiar with the models and the 'interface points' where these models interact with other models and tools, or relevant documentation and training from experts is available to the planners.

3. *The toolsuite is open for integration: other, future simulation engines can be added to it via a well-defined (software) engineering process, based on meta-modeling.* Meta-modeling in this context means the analysis and formal description of the interfaces of a simulation and/or analysis tool, such that it becomes part of the integrated tools in a Smart Grid Simulator instance. We anticipate that such meta-modeling is done by skilled developers; it is important that the meta-models be hidden from the end-user: s/he should not deal with low-level details. The meta-modeling process embeds the interfaces of a new modeling/simulation/analysis tool in the Model Integration Language, it defines the formal integrity constraints used in correctness and completeness analysis and establishes the high-level abstractions through which the analysts can integrate the models in new scenarios. The meta-modeling process is fully supported by tools.

4. *The execution of the analysis can take advantage of available computing platforms, including distributed platforms.* To evaluate multiple scenarios, each one under various assumptions, one needs computational power that is available in the most cost-effective form today, i.e., as clusters of server nodes in a secure environment. Configuring and running multiple simulations is a challenging task that should be supported by appropriate tools that shield the end-users from low-level details. This concept has been successfully implemented and used in the network community in the form of the Emulab/DETER platform that can serve as a starting point for a next generation smart grid simulator execution platform (see Fig. 12).

Figure 12: Emulab cluster at Vanderbilt

5. *The toolsuite is accessible via a secure web interface.* For practical purposes, the toolsuite must be easy to access and users shall not be required to install and configure complex software packages on their (desktop) systems. The planners and analysts have access to all tool-supported activities (simulation configuration, execution, and analysis) via a web interface, while the tools run on a high-performance platform. In a sense, this makes the Smart Grid Simulator similar to cloud computing services: computations are executed on some hosts on a network, completely transparently for an end-user.

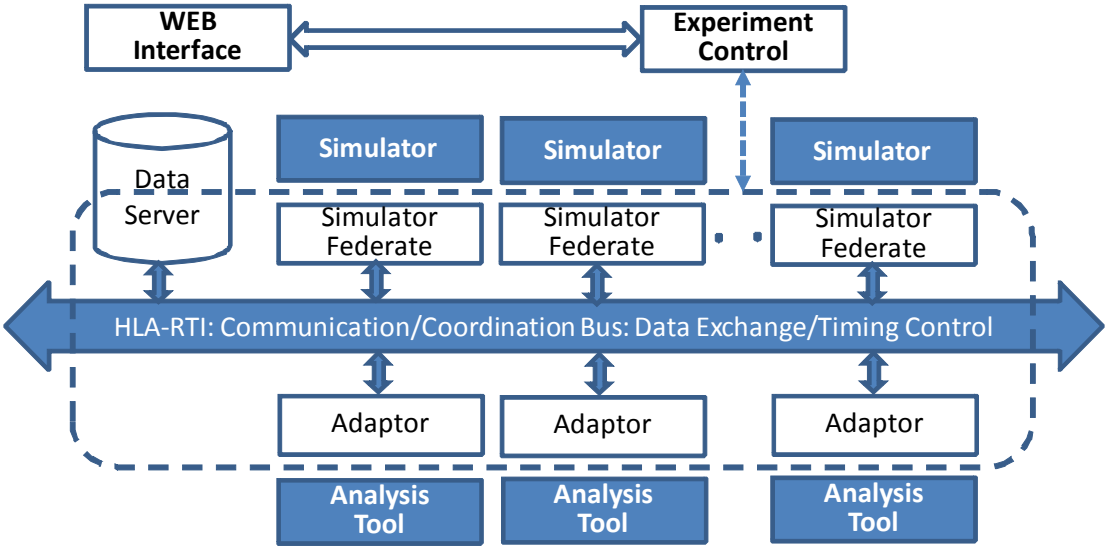


Figure 13: Notional Architecture for Next Generation Smart Grid Simulator

The notional architecture of the tool suite is shown on 13. In order to implement the above vision, the following software tools need to be developed:

1. Model Integration Language
2. Meta-models for simulator tools
3. Federates and integration adaptors for simulator and analysis tools
4. Software generators for experiment configuration and execution on distributed platforms
5. Experiment controller
6. Web interface to the tool suite

5 Roadmap and Research Needed

Modeling and simulation capabilities for the next generation electric grid require technology advancements for satisfying the drastically increased heterogeneity of models and the wide spectrum of use cases from research to planning, to training, and on to live operations. The pervasive use of computing and networking in Department of Defense (DoD) systems has created similar challenges for the modeling and simulation

infrastructure for defense systems. We believe that a wide range of approaches and tools can be leveraged in electric grid modeling and simulation that will significantly decrease the cost and shorten the development time. Below, we discuss research challenges in the areas of (1) Models and Model Integration, (2) Simulation and (3) Infrastructure. After summarizing the research challenges, we divide them into three categories, short-term (1-2 years), mid-term (3-5 years) and long-term (5-10 years). Research in the short term category means that proven results exist in other application domains and the task is adopting and validating the methods and tools to the next generation electric grid problem area. Research in the mid-term category implies that only preliminary results exist or existing methods and tools need to be significantly changed for applicability. Long-term research topics refer to challenges that are open and significant effort needs to be invested in finding solutions.

5.1 Models and Model Integration

Creating and integrating models that are multi-scale, multi-data and multi-resolution is one of the major challenge areas. These models must capture multiple domains using domain-specific modeling tools, address the needs of multiple uses, and include tightly integrated computing and networking models characterizing smart grid applications.

1. *Multi-scale modeling*: The objective of multi-scale modeling is to enable "zooming in" on selected geographic areas with high fidelity models, while abstracting the environment using equivalence models. Mathematical approaches for solving the problem exist and there are a variety of equivalencing packages providing the necessary software tools. The primary challenge is in developing practical solutions for managing and accessing data.
   Research needs:
   - Scalable information management system that can provide access to distributed, heterogeneous data sets. (Mid-term)
   - Standardized access mechanisms for searching and indexing data using techniques such as metadata tagging. (Mid-term)
   - Solutions for providing robust security for data access using encryption, authentication and auditing is an essential requirement. (Mid-term)
2. *Multi-data/Multi-resolution modeling*: Models are developed for different purposes (transient analysis, stability analysis, steady state analysis, etc.) and are constructed from component models. It is common that the same system components (generators, lines, transformers, circuit breakers/switches, substations, sensors, relays, controllers, etc.) have a range of different models using abstractions that correspond to phenomena intended to be captured. A common challenge is to make these abstractions consistent with each other.
   Research needs:

- Formally well defined abstraction layers for component modeling. (Short-term)
- Methods and tools for verifying model consistency and extracting more abstract models from high fidelity first principle-based models. (Short-term/Mid-term)

3. *Multi-modeling and model integration*: Models are typically prepared by different vendors using a variety of modeling formalisms and tools. It is not practical to expect that all vendors will use the same modeling technique and modeling languages, still, the models need to be integratable into multi-models. This requires a formalism to model the "modeling languages" used for their modeling.
   Research needs:
   - Application of metamodeling. (Short-term)
   - Introduction of formally defined semantics, semantically precise model integration frameworks and model translators. (Short-term/Mid-term)
4. *Multi-use modeling*: Models are developed for different purposes, such as Planning (focus on capital cost optimization), Operations Planning (operational cost management, training), and Real-time Operation Management (reliability and optimization). These usage scenarios require integrating outputs of different simulators with other data sources such as cost, reliability and real-time data streams.
   Research needs:
   - Fusing algorithms for different simulators and data sources to adjust the computations to the purpose we want to satisfy. (Long-term)
5. *Smart-grid implications:* "Smart grid" expresses a move towards the deeper integration of power systems operation with distributed, networked control systems deployed in a sensor and data rich grid environment. Fundamental properties of the grid, such as stability, transient behavior, optimality depend on the behavior of the underlying computation and communication services. Co-modeling of the dynamics of these IT components with the physical system components and factoring in the operational reliability with their failure modes (natural and human induced) causes significant new challenges.
   Research needs:
   - Integration of heterogeneous modeling languages and models. (Short-term)
   - Significantly extending the time-scale of the simulations. (Short-term)
   - Managing increased complexity of models & simulators. (Short-term/Mid-term)
   - Introduction of complex, highly dynamic simulation scenarios targeting active defense against intelligent adversaries and improving resilience against cyber-attacks (Long-term)

5.2 Simulators

Simulation engines are the workhorse of simulators. The integration and operation of heterogeneous simulation engines includes major challenges.

1. *Composable Multi-engine Simulators*

   Currently, a number of proprietary simulation packages are used in the industry, while academic researchers often use simpler, open-source packages. Typically each software package uses its own simulation engine, in isolation. In order to run these simulations together they have to be composed and operated in a coordinated manner. The current integration approach happens via a labor intensive, manual process scales poorly.

   Research needs:

   - Run-time simulation integration framework, possibly based on the High-Level Architecture of the DoD, that allows running multiple simulators in coordination. This would be a software framework to build integrated experiments that span across multiple simulators. (Short-term)
   - Integration of specific simulation packages into the simulation framework. For each simulation package specific 'adaptor' software is needed that brings it into compliance with the integration framework. Such adaptors need to be designed and developed based on the analysis of the specific simulation engines. (Short-term)
   - Computer network simulations often provide high fidelity but they are slow for practical use, or they are incapable of simulating the entire software stack running on a host (i.e. the application, the middleware, the network stack, and the operating system). In this case emulators are typically used. They allow the real application/middleware/operating system to be used with an emulated network. Research is needed on how these can be integrated with completely simulation-based systems, and on how physical devices can be used in conjunction with simulators. (Short-term/Mid-term)
   - Multi-core platforms are the becoming available and high-performance simulators have to take advantage of these novel platforms. There is very little research today on implementing power system simulators on multi-core processors or General Purpose Graphics Processing Units (GPGPU) that are also high-performance platforms. (Mid-term)

2. *Scalable simulation architecture*

   There is a need for a highly scalable simulation architecture that is easily expandable to large systems. There are several architectural patterns and

platforms that can be potentially applied, but the exact means of their application needs to be developed.

Research needs:

- HLA federates can be hierarchically structured (lower-level federates running at high fidelity, higher-level federates running at lower fidelity) but the exact details of how to do this and how to coordinate across layers are not clear, and they have to be analyzed and techniques developed. (Mid-term)
- The EMULAB platform that serves as the network emulator in typical networking experiments can also be used as a high-performance, highly networked cluster (typically an EMULAB host has 4x1GB/s network ports). It is an interesting research topic how to create a scalable simulation architecture that uses the EMULAB platform. (Mid-term)
- A data routing network is essential for network experimentation, and the information networks used in the power system domain may not be based on the 'public' Internet. Research is needed on how to model, simulate, and emulate this data distribution network in a scalable manner. (Mid-term)

3. *Decentralized simulation*

   In power systems data is typically available only locally, so in order to drive federated simulations with that data the simulations may have to be geographically distributed.

   Research needs:

   - The engineering of the networking and the configuration of the decentralized, distributed simulations. (Long-term)
   - Ensuring time synchronization across geographically distributed simulators. DoD has such systems for their application domains (e.g. battle simulation), but it is not clear how the same techniques can be applied in the power system domain. (Long-term)

4. *Model-based integration and configuration*

   When a suite of simulators have to cooperate and interact in a complex experiment, their configuration is a non-trivial task. There is a need to simplify this work such that engineers and researchers can rapidly construct experimental studies for large-scale systems. Model-based integration of the simulators and model-based configuration of the experiments is a viable answer to such problems, and has been demonstrated in various DoD projects (which were related to Command and Control systems).

   Research needs:

- High-level models can be used to describe the desired configurations, but often are too detailed and require too much work to construct. Research is needed on how to represent the desired deployment in a compact manner, and then automatically generate platform-specific deployments. (Short-term/Mid-term)
- Validation of an ensemble of simulators. Assuming the individual simulators are validated, how can one validate their federation where multiple simulators work in conjunction? (Mid-term)
- Fault management during the simulation experiments. Individual simulators may fail during an experiment, the network infrastructure may fail, or the simulation/emulation computing engines may fail. We need techniques and tools to detect, isolate, and manage such faults, such that the experimentation remains productive. (Mid-term)

5. *Open source and proprietary alternatives*
   Both high-fidelity proprietary simulators and low-fidelity open source simulators can be used for power systems simulations, often as alternatives. Research is needed to evaluate these packages and determine what capabilities they can provide, how well are they prepared for integration, and how well can they be used together in large-scale simulations.

5.3 Infrastructure

It would be a major mistake to believe that modeling and simulation challenges in the next generation electric grid are all about heterogeneous models and simulation engines. The sheer problem size poses major challenges regarding the management of models, software tools, systems configurations, and data sets not in terms of volume but of complexity and their interrelationships and consistency. We believe that there is a strong need to introduce a sufficiently robust infrastructure that can serve these needs.

1. *Collaborative, distributed modeling infrastructure*
   Our assumption is that progress toward the next generation electric grid will establish a vibrant R&D community with strong links to vendors and operators. Much of the activity will focus on modeling: creating, validating, sharing and experimenting with a wide range of models responding to changing requirements and technology advancements. A stimulating example for building infrastructure for similar purpose is software forges. In just over a decade, SourceForge alone has become one of the most active hubs of human collaboration in history. Software forges have pioneered many tools that have proven their value in contributing to widely distributed collaborative work. Structurally and functionally, they provide an excellent example for creating a collaborative distributed modeling infrastructure for the development of the next

generation electric grid.

Research needs:

- Establishment a Model Repository with multi-level security. The Model Repository includes public branches for which access is not limited, semi-public branches with access based on export control requirements and proprietary/restricted access branches for which access is limited to authenticated persons or organizations. (Short-term/Mid-term)
- Model version control system that supports check-out, check-in, roll back and issue tracking (Short-term)
- Credentialing and verification system for ensuring trustworthiness of the models. The methods adopted should provide assurance again malicious manipulation of the models and protect against unauthorized access to protected models. (Mid-term/Long-term)
- Metadata tagging of models and components that enables authorized entities for indexing distributed (regional) model repositories and providing authenticated access to models according to the required privileges. (Mid-term)

2. *Collaborative, distributed simulation infrastructure*

   The current state of practice makes the use of multi-model simulators difficult or impossible. High fidelity dynamic simulators (such as RTDS) are extremely expensive, integration of heterogeneous simulations requires substantial systems and software expertise, and challenge problems are not accessible. Changing this situation is essential.

   Research needs:

   - Scalable simulators need to be developed that provide a continuum from low-end simulators that are still capable for multi-modeling with lower performance to high performance simulators running on large computing clusters. Low-end simulators should be using open-source, low-cost but reliable software packages whose operation does not require specialized expertise. These simulators should be distributed via open-source repositories and made widely available. (Short-term)
   - Complex studies and computational experiments require Open Virtual Experimental Platforms (OVIP) that run on large, highly configurable computing clusters with open web interfaces. Similarly to the widely known EMULAB platform, OVIP-s need to be prepared for accepting experiment specifications, running the experiments and providing the results through easy to use interfaces. (Short-term/Mid-term)
   - Challenge problems need to be made available for the research community via a Scenario Repository. Experiment scenarios include

validated model suites that accept new model components (controllers, networking protocols, etc.) using standard modeling languages for evaluation. (Short-term/Mid-term)

6 Conclusion

The next generation electric grid will bring about the tight integration of physical systems and processes with computations and networking. Functionality and salient system characteristics will emerge through the interaction of physical and computational objects, computers and networks, devices, and their physical/social environment. Technical challenges to satisfy the rapidly increasing need for multi-resolution modeling and simulation will be significant due to the increased heterogeneity and complexity of systems comprising the smart grid.

The last decade has brought major advances in modeling and simulation technologies both in electric power systems and in heterogeneous system-of-systems driven by DoD needs. In this review paper we argue that the time is right for leveraging these advancements in the development of a highly scalable modeling and simulation infrastructure that is model-based, cyber-physical, scalable and highly automated. Feasibility of the proposed approach largely depends on the pervasive application of advanced software technology in modeling and model integration, heterogeneous multi-model simulation and information management infrastructure.

References

[1] S. Massoud Amin and B.F. Wollenberg, "Toward a smart grid: power delivery for the 21st century," IEEE Power and Energy Magazine, Vol. 3, No. 5 Sept.-Oct. 2005, pp. 34-41.

[2] E. Santacana, G. Rackliffe, L. Tang, and X Feng, "Getting Smart", IEEE Power and Energy Magazine, March-April 2010, vol. 8, issue 2, pp. 41 – 48

[3] S. E. Collier, "Ten Steps to a Smarter Grid", IEEE Industry Applications Magazine, March-April 2010, vol. 16, issue 2, pp. 62-68

[4] Thomas H. Morris, Sherif Abdelwahed, Anurag K Srivastava, Ray Vaughn and Yoginder Dandass, "Secure and Reliable Cyber Physical Systems", National Workshop on new research directions for future Cyber-Physical Energy Systems, June 3-4, 2009, Baltimore, Maryland

[5] Ram Mohan Reddi, and Anurag K Srivastava, "Real Time Test Bed Development for Power System Operation, Control and Cyber Security", North American Power Symposium, October, 2010, Arlington, TX

[6] T. Morris, A. Srivastava, B. Reaves, K. Pavurapu, R. Vaughn, W. McGrew, Y. Dandass, "Engineering Future Cyber-Physical Energy Systems: Challenges, Research Needs, and Roadmap", North American Power Symposium, Mississippi State, MS, October 4-6, 2009

[7] National Power Grid Simulation Capability: Needs and Issues. A report from the National Power Grid Simulator Workshop, December 9-10, 2008, Argonne, IL

[8] A. Bose, "Smart Transmission Grid Applications and Their Supporting Infrastructure", IEEE Transactions on Smart Grid, June 2010, vol. 1, no. 1, pp. 11 – 19.

[9] A. Bose, "Models and techniques for the reliability analysis of the smart grid", IEEE Power and Energy Society General Meeting, 25-29 July 2010, pp. 1 – 5, Minneapolis, MN

[10] A. Bose, "New smart grid applications for power system operations", IEEE Power and Energy Society General Meeting, 25-29 July 2010, pp. 1 – 5, Minneapolis, MN

[11] M. Amin, "National Infrastructures as Complex Interactive Networks", Automation, Control, and Complexity: An Integrated Approach, Samad & Weyrauch (Eds.), John Wiley and Sons, pp. 263-286, 2000

[12] L. Bam and W. Jewell, "Review: Power Systems Analysis Software Tools," in Proc. 2005 IEEE Power Engineering Society General Meeting, pp. 139-144.

[13] S. E. Widergren, J. M. Roop, R. T. Guttromson, and Z. Huang, "Simulating the Dynamic Coupling of Market and Physical System Operations", IEEE PES General Meeting, 2004

[14] C. Hauser, D. Bakken, and A. Bose, "A failure to communicate: next generation communication requirements, technologies, and architecture for the electric power grid", IEEE Power and Energy Magazine, March-April 2005, vol. 3, no. 2, pp. 47 – 55.

[15] D. Bakken, A. Bose, C. Hauser, D. Whitehead, and G. Zweigle. “Smart Generation and Transmission with Coherent, Real-Time Data”. Proceedings of the IEEE (Special Issue on the Smart Grid), 2011
[16] G. N. Ericsson, “Modeling of power system communications-recognition of technology maturity levels and case study of a migration scenario”, IEEE Transactions on Power Delivery, vol. 19 , no. 1, 2004 , pp. 105 – 110
[17] R. Mora, A. López, D. Román, A Sendín and I Berganza, “Communications architecture of smart grids to manage the electrical demand”, Third Workshop on Power Line Communications, October 1-2, 2009, Udine, Italy.
[18] M. D. Ilić , E. H. Allen , J. W. Chapman, C. A. King, J. H. Lang, E. Litvinov, “Preventing Future Blackouts by Means of Enhanced Electric Power Systems Control: From Complexity to Order”, Proceedings of the IEEE, Vol. 93, No. 11, November 2005.
[19] P. J. Mosterman, G. Biswas and J. Sztipanovits, “Hybrid Modeling and Verification for Embedded Control Systems”, Control Engineering Practice, vol. 6, pp. 511-521, 1998.
[20] H. A. Taha, Simulation modeling and SIMNET, ser. Prentice-Hall International Series in Industrial and Systems Engineering. Prentice-Hall, 1988
[21] D. C. Miller and J. A. Thorpe, “Simnet: The advent of simulator networking," Proceedings of the IEEE, vol. 83, no. 8, pp. 1114-1123, 1995.
[22] R. Hofer and M. L. Loper, “Dis today [distributed interactive simulation]," Proceedings of the IEEE, vol. 83, no. 8, pp. 1124-1137, 1995
[23] P. K. Davis, “Distributed interactive simulation in the evolution of DoD warfare modeling and simulation," Proceedings of the IEEE, vol. 83, no. 8, 1995.
[24] F. Kuhl, R. Weatherly, and J. Dahmann, Creating computer simulation systems: an introduction to the high level architecture. Prentice Hall PTR Upper Saddle River, NJ, USA, 1999.
[25] R. L. Wittman and A. J. Courtemanche, “The ONESAF product line architecture: An overview of the products and process." Citeseer, 2002
[26] F. Zhang and B. Huang, “HLA-based network simulation for interactive communication system," in Modelling & Simulation, 2007. AMS'07. First Asia International Conference on, 2007, pp. 177-180 J.
[27] Henriksen, “SLX: The x is for extensibility," in Proceedings of the 32nd conference on Winter simulation. Society for Computer Simulation International, 2000, p. 190
[28] B. Zeigler, G. Ball, H. Cho, J. Lee, and H. Sarjoughian, \Implementation of the devs formalism over the HLA/RTI: Problems and solutions," in Spring Simulation Interoperability Workshop, 1999.
[29] [Online]. http://www.mathworks.com/products/connections/product_main.html?prod id=696&prod name=HLA%20Toolbox
[30] [Online]. Available: http://www.mak.com

[31] H. Sarjoughian and B. Zeigler, “DEVS and HLA: Complementary paradigms for modeling & simulation?" Transactions-Society for Modeling and Simulation International, vol. 17, no. 4, pp. 187-197, 2000
[32] T. Lu, C. Lee, W. Hsia, and M. Lin, “Supporting large-scale distributed simulation using HLA," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 10, no. 3, p. 294, 2000
[33] H. Vangheluwe, J. D. Lara, and P. Mosterman, “An introduction to multi-paradigm modelling and simulation," in Proc. AIS2002. pp. 9-20.
[34] J. de Lara, H. Vangheluwe, and M. Alfonseca, “Meta-modelling and graph grammars for multi-paradigm modelling in atom 3," Software and Systems Modeling, vol. 3, no. 3, pp. 194-209, 2004.
[35] V. Vittorini, M. Iacono, N. Mazzocca, and G. Franceschinis, “The OSMOSYS approach to multiformalism modeling of systems," Software and Systems Modeling, vol. 3, no. 1, pp. 68-81, 2004.
[36] H. Vangheluwe, “DEVS as a common denominator for multi-formalism hybrid systems modelling," in IEEE International Symposium on Computer-Aided Control System Design, vol. 134. Citeseer, 2000.
[37] A. Baksai, V. Prasanna, and A. Ledeczi, “Milan: A model based integrated simulation framework for design of embedded systems," Proceedings of the 2001 ACM SIGPLAN workshop on Optimization of middleware and distributed systems, p. 93, 2001
[38] P. Benjamin, K. Akella, and A. Verma, “Using ontologies for simulation integration," in Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come. IEEE Press, 2007, pp. 1081-1089.
[39] J. A. Miller, G. T. Baramidze, A. P. Sheth, and P. A. Fishwick, “Investigating ontologies for simulation modeling," in Simulation Symposium, 2004. Proceedings. 37th Annual, 2004, pp. 55-63.
[40] Karsai, G., Maroti, M., Ledeczi, A., Gray, J., and Sztipanovits, J., 2004. “Composition and Cloning in Modeling and Meta-Modeling”, Control Systems Technology, IEEE Transactions, 12 pp. 263-278
[41] Jackson, E., Porter, J., Sztipanovits, J.: “Semantics of Domain Specific Modeling Languages” in P. Mosterman, G. Nicolescu: Model-Based Design of Heterogeneous Embedded Systems. pp. 437-486, CRC Press, November 24, 2009
[42] Jackson, E., Sztipanovits, J.: ‘Formalizing the Structural Semantics of Domain-Specific Modeling Languages,” Journal of Software and Systems Modeling pp. 451-478, September 2009
[43] J. Sztipanovits, G. Karsai, S. Neema, H. Nine, J. Porter, R. Thibodeaux, and P. Volgyesi, “Towards a model-based toolchain for the high-confidence design of embedded systems," 2007.
[44] G. Hemingway, H. Neema, H. Nine, J. Sztipanovits, and G. Karsai, “Rapid synthesis of HLA-based heterogeneous simulation: A model-based integration approach," Simulation: Transaction of the Society for Modeling and Simulation International (In press, to be published in 2011.)

[45] Kai Chen, Janos Sztipanovits, Sandeep Neema: "Compositional Specification of Behavioral Semantics," in Design, Automation, and Test in Europe: The Most Influential Papers of 10 Years DATE, Rudy Lauwereins and Jan Madsen (Eds), Springer 2008
[46] Karsai, G., Ledeczi, A., Neema, S., Sztipanovits, J.: The Model-Integrated Computing Toolsuite: Metaprogrammable Tools for Embedded Control System Design, Proc. of the IEEE Joint Conference CCA, ISIC and CACSD, Munich, Germany, 2006.
[47] de Laura, J., and Vangheluwe, H., 2002. "AToM3: A Tool for Multi-formalism and Meta-Modeling", Lecture Notes in Computer Science, 2306 (174-188).
[48] Tolvanen, J.P., and Lyytinen, K. 1993. "Flexible Method Adaptation in CASE. The Metamodeling Approach", Scandinavian Journal of Information Science, v5 n1 (71-77).
[49] Cook, S., Jones, G., Kent, S., and Wills, A. 2007. "Domain-specific Development with Visual Studio DSL Tools", Addison-Wesley Professional.
[50] Eclipse Modeling Framework http://www.eclipse.org/modeling/emf/
[51] McLaren PG, Kuffel R, Wierckx R, et al. A real time digital simulator for testing relays. IEEE Transactions on Power Systems 1992; 7(1):207–213.
[52] [online] http://nsnam.isi.edu/nsnam/index.php/Main_Page
[53] http://www.omnetpp.org/
[54] http://www.opnet.com/
[55] https://wiki.isis.vanderbilt.edu/OpenC2WT/index.php/Main_Page
[56] Andreas Franck and Volker Zerbe: A Combined Continuous-Time/Discrete-Event Computation Model for Heterogeneous Simulation Systems, Advanced Parallel Processing Technologies Lecture Notes in Computer Science, 2003, Volume 2834/2003, 565-576

Discussant Narrative

Model-based Integration Technology for Next Generation Electric Grid Simulations

Henry Huang
Pacific Northwest National Laboratory

This paper presents a good overview of current modeling and simulation techniques. Challenges in modeling and simulation are presented. The key point of this paper is to integrate multiple elements (grid, communication) into one single simulation framework to address the complexity of future power grids. This is an agreed direction.

The paper is sort of assuming this new modeling and simulation framework is needed. It could be made stronger if the driving forces are stated more clearly. This can be achieved via use case examples. Use case examples can include wide-area control and its design, vulnerability analysis of extreme events (solar storm, hurricane, etc.). For example, wide-area control of future power grid would need to study the impact of communication of signals on the control performance, and thus requires the combined simulation of grid and communication networks.

I liked the way the authors divided the problems to be modeling and simulation integration problems. To further clarify the points, it is worth discussing how the military systems would be applicable to power grids and how metamodeling would be developed for power grids.

Two kinds of power grid simulation need to be discussed and differentiated: real-time simulation and off-line simulation. Real-time simulation examples include simulation tools for real-time grid operation and hardware-in-the-loop testing. Off-line simulation examples include expansion planning and hypothetic scenario studies. The proposed approach (esp. cloud computing approach) is suitable for off-line simulation. For real-time simulation, performance requirements outweigh the convenience in development and deployment. Generalized high-level integration may not be adequate. Cloud computing as predicted by Microsoft is not suitable for real-time purposes.

Data sharing and system security issues are not addressed. Military environments control every piece of data and every step of the simulation. Data sharing and system security may not be an issue. In collaborative power grid environments, there are mixed

pieces: some are open information but other may be sensitive. The modeling and simulation framework should provide a means to ensure the data are protected and the whole modeling and simulation environment is secure.

Multi-scale, multi-resolution modeling is a necessary step. The emphasis is not only the data management between scales/resolutions but more importantly the representation of the coupling between different scales/resolutions.

Recorder Summary

Model-based Integration Technology for Next Generation Electric Grid Simulations

Victor M. Zavala
Argonne National Laboratory

Paper Summary

Understanding and operating the next-generation power grid in an efficient and secure manner will require of multi-scale and multi-physics simulations capable of integrating heterogenous components and spatio-temporal domains. According to the authors, there exist several technical bottlenecks in achieving this goal:

- There is no single simulation framework that integrates heterogeneous power engineering, communication, and control components.
- Simulation tools have different semantics, physical phenomena, time and spatial domains, and execution logic. For instance, there is no single tool that integrates transmission and distribution models at a sufficient detail.
- There is no tool that enables the combination of models to data that is multi-rate, multi-scale, and multi-user.

The authors propose the development of an integrative simulation and present preliminary results.

Discussant Summary

The key need raised by this paper is the integration of multiple simulation elements (grid, communication, data) into one single simulation framework to address the complexity of the future power grid. This is an agreed direction. It would be worth discussing how existing military (i.e., DARPA) simulation systems and know-how would be applicable in the power grid context.

The authors do not clearly motivate what is the need for integration. Operational scenarios and economic impact are needed to illustrate this. In addition, the authors should distinguish between real-time simulation and off-line simulation. Real-time simulation examples include simulation tools for grid operation, visualization, and hardware-in-the-loop testing. Off-line simulation examples include expansion planning and hypothetic scenario studies. The proposed approach is suitable for off-line

simulation. For real-time simulation, performance requirements outweigh the convenience in development and deployment.

Data sharing and system security issues are not addressed. Military environments control every piece of data and every step of the simulation. In collaborative power grid environments, there are mixed pieces: some are open information but other may be sensitive. The integrative simulation framework should provide a means to ensure that the data are protected and the whole modeling and simulation environment is secure.

Discussion Report

The discussion focused on two areas: 1) need and motivation and 2) emerging physical and numerical phenomena. These are discussed below.

Need and Motivation

It is not clear what is the need for an integrated multi-physics modeling framework. For instance, what type of phenomena can be captured and how we can use that knowledge for decision-making. It is necessary to establish a more organized modeling and simulation research agenda to define what problems and phenomena are critical to be addressed and thus how simulation tools should be designed and coupled. In addition, it is necessary to define what phenomena need to be captured and at what precisions to address different scientific and operational questions.

Multi-physics modeling and simulation is insufficient for understanding new phenomena in the smart grid. There is a need for modeling and simulation of the behavior of communication networks and computational components. This requirement is emerging from the fact that overall grid dynamics will emerge from the interaction of physical and computational components and networks. Multi-model simulation that includes both physical and cyber components is a significant and new challenge.

It is necessary to decide if we should aim for a single simulation framework coordinated and standardized by design or if we should design multiple simulation tools independently and develop a framework to coordinate and standardized them. For instance, the National Oceanic and Atmospheric Agency (NOAA) has been successful in coordinating a team of meteorologists, mathematicians, computer scientists, and software engineers for the development of WRF, the most advanced numerical weather forecasting system in the world. In the power grid industry, simulation tools have been largely developed and supported by industrial practitioners that might not have an incentive to integrate, standardize, and expand them.

It is not clear what is manageable with existing simulation and computational capabilities (time-scales and domain). This misleads research priorities. For instance, the

dynamic analysis and conclusions presented in the 2003 Eastern Interconnection blackout report were limited by computational times. Therefore, before moving to more complex simulations, it is necessary to analyze bottlenecks arising in each physical simulation domain and analysis task that help us explain previous and current grid catastrophic events.

Emerging Physical and Numerical Phenomena

Integrating modeling components where different phenomena (both physical and computational) occur at different time scales can lead to the discovery of emerging phenomena. These phenomena might be physical or spurious (i.e., due to numerical issues). Existing simulation tools do not provide systematic capabilities to detect these phenomena and thus advise the modeler. For instance, it is not clear what phenomena can arise from integrating electromagnetic and electromechanic effects. Similarly, it is not clear how time varying delays and pocket drops in communication networks used by networked controllers can influence the dynamics and transient phenomena in power grids. To validate these phenomena, it is necessary to incorporate multi-modeling, validation and verification and uncertainty quantification capabilities.

Numerical complexity and robustness are critical issues arising in multi-physics simulations. Robustness issues arise in various forms such as: stiffness due to drastically different time scales, non-differentiability due to discrete events, and convergence issues due to nonlinearities and cycling. It is thus necessary to develop robust and scalable numerical algorithms.

Power grid simulation tools can benefit from high-performance computing systems available at DOE national laboratories and numerical algorithms and tools developed from Office of Science initiatives such as PDEs, computational differentiation, hybrid systems and differential variational inequalities, time-stepping, and uncertainty quantification which are currently used in complex physics simulations. In addition, smoothing techniques for physical transitions are needed from power systems experts.

In conclusion, the integration of simulation tools into a single multi-physics/multi-modeling framework might lead to the discovery and understanding of physical phenomena, increased predictive capabilities and to a more systematic model and data management. It is not clear, however, what sophistication level and physical elements are needed. In addition, numerical algorithms are needed to enable a robust and computationally efficient integration of complex physics.

White Paper

Research Needs in Multi-Dimensional, Multi-Scale Modeling and Algorithms for Next Generation Electricity Grids

Santiago Grijalva
Georgia Institute of Technology

Abstract

The planning, simulation, operation, and control of electrical energy systems involve dealing with information and phenomena that is naturally associated with temporal, spatial, and scenario dimensions. Each one of these dimensions has specific scales and granularity requirements associated with relevant system behavior. The current forces affecting the electric industry such as deployment of smart grid technologies, growing penetration of distributed renewable energy sources, energy efficiency programs, and availability of massive information, result in relevant system behavior appearing on scales different from those traditionally studied. In order to model and simulate power system behavior, several assumptions are commonly made about the relevant dimensions and scales of specific problems. In many cases, these assumptions were needed in order to make the computational problem tractable. Because existing computation uses fixed assumptions regarding those dimensions and scales, the current analytical and control methods are inherently limited in their ability to capture emerging relevant phenomena.

Research targeted to increase the understanding and more formal handling of multi-scale and multi-dimensional power system behavior can provide significant benefits to the evolution of the electricity industry. This paper provides an overview of emerging multi-dimensional, multi-scale phenomena and suggests several areas where further research can be of service. The underlying topic of the paper is that a holistic, multi-dimensional, multi-scale framework may be required in order to address several emerging engineering challenges.

Multiscale modeling and computation has been the subject of great interest of computational science and applied mathematics. In the last decade there have been great advances in multiscale methods applied to material science, chemistry, fluid dynamics, and biology. Power systems engineering, though, has not developed multi-scale methods to the level that it exists in other engineering disciplines, except for specific

applications such as singular perturbation and hierarchical control. This paper identifies eighteen areas of multi-scale, multi-dimensional power system research that are needed to provide a solid framework to address emerging problems.

1 Emerging Behavior

The electricity industry is in the midst of an unprecedented transformation driven by growing penetration of renewable energy and storage, availability of massive information, new power and control technologies, and consumer expectations in the quest to realize a smart electric grid that can support a highly efficient, secure, and sustainable society. As a result, the complexity of the electricity infrastructure, its operation, and the associated planning processes is growing dramatically. At the same time, investment in substation automation, deployment of intelligent electronic devices, and installation of smart metering will produce amounts of multi-dimensional data several orders of magnitude larger than today's data acquisition. The existing models and analytical tools were not designed for and are not prepared to support and fully exploit massive multi-dimensional and multi-scale data.

Several engineering and computational challenges have emerged in recent years that cannot be addressed unless the emerging phenomena are accurately represented by the models. Further, the computational tools must be designed to formally ensure that the analytics is capable of comprehensively simulating the physical phenomena. To illustrate the need of multi-dimensional, multi-scale analysis, let us list a few examples of emerging issues that involve complex behavior and the limitations of current computational methods:

a) *Dispatch with Large Penetration of Wind Generation*: Wind energy is highly variable and difficult to predict. Wind variability becomes particularly critical in the seconds to minutes timeframe because of the limitations of conventional reserve to respond to drastic changes of wind production. Various systems in the U.S. have experienced emergency conditions and events due to unexpected reduction of wind production, which resulted in automatic load shedding [1, 2]. The limitations of the existing wind forecasting tools and the computational intensity of scheduling software has resulted in most utilities with growing penetration of wind deciding to operate the systems very conservatively, using high levels of regulating reserve [3, 4]. In particular, unit commitment has traditionally been done using a one-hour or half an hour temporal granularity, while real-time economic dispatch runs every five minutes based on snapshot information only. A better understanding of the temporal multi-scale behavior of wind, including how to incorporate significant uncertainty in reserve scheduling to ensure reachability, is highly needed. The economic dispatch software must be replaced with short-term stochastic scheduling software with granularity in the range of several seconds and a horizon from minutes to a few hours [5, 6].

b) *Demand Response in Smart Grids*: Demand response today encompasses a set of consumer actions to disconnect, reduce, or shift demand in time upon critical events or based on economic incentives such as temporally differentiated electricity pricing [7, 8, 9, 10]. A significant portion of the technologies deployed under the spectrum of smart grid activities are targeted to enable demand response. Demand response should provide significant benefits, the most important being increased reliability through load reduction at critical times and increased efficiency through higher utilization factors of distribution and generation capacity. Demand response can broadly involve actions from the protections range (millisecond and second) to day ahead scheduling. Several project reports indicate a level of response by customers much lower than predicted, which would question the feasibility of smart grid-enabled demand response programs. It has been suggested that a deeper understanding of the customer objectives and response constraints is essential to realize demand response [11]. A significant element of this process is the recognition that homes, buildings, microgrids, etc. may have different objectives from the utility or provider [12, 13, 14]. The fact that smart grid enables the consumer to make decisions implies that the consumer is now a strategic decision maker, and hence the interactions with the utility can exhibit Nash Equilibrium. However, the utility and the consumer operate at entirely different spatial (and power) scales, and these strategic interactions are far from being understood.

c) *Transmission and Distribution Multi-Level Information*: Independent System Operators (ISOs) typically model delivery points to distribution utilities as single loads that aggregate distribution networks. When distribution systems experience internal events or dangerous conditions, these problems are in the vast majority of cases not observable to the ISOs. Various ISOs have indicated the need to model and gather real-time information from the sub-transmission and distribution system in order to enhance reliability and awareness [15, 16], i.e., provide finer granularity modeling. While ISOs have traditionally been able to forecast load within a 2% error, deployment of distributed energy resources and utility-scale storage may increase the error substantially [17, 18]. There are several challenges that limit seamless integration of transmission and distribution models, including single- versus three-phase modeling, non-standardization of operational and planning models, node/breaker versus bus/branch modeling, data privacy, and operational liability. Utility engineers and academics have begun to question the historical separation of transmission and distribution scales, including the fundamental assumption of a balanced system at the bulk transmission level [19].

d) *Scenario Analysis in Transmission Planning*: Maintaining a high level of power system reliability requires the evaluation of a large number of plausible

contingencies, planning alternatives, or operational scenarios [20, 21]. NERC requires *N*–2 contingency evaluation of bulk transmission system security, which is computationally very expensive. Methods for reducing the problem dimensionality by filtering non-relevant scenarios are highly needed. In particular, systematic ways to prioritize searches, reduce scenario spaces [22], and dynamically update multi-dimensional information are sought after.

The previous examples illustrate some of the emerging power system behavior, which involves multiple dimensions and multiple scales. Addressing those problems requires a deeper characterization of power system phenomena and, in some cases, might require a reformulation of some of the fundamental simplifications commonly used. The reminder of this paper is organized as follows: in Section 2, we describe the dimensions and scales that are relevant in power system computation. In Section 3 we describe the needs associated with multi-dimensional and multi-scale modeling and analysis. Section 4 presents the research needs in optimization, dynamics, and control, involving multiple scales. Section 5 addresses the multi-dimensional and multi-scale research needs in data handling and visualization. Section 6 presents some concluding remarks of this paper.

2 Power System Dimensions and Scales

2.1 Power System Dimensions

Power system dimensions can be broadly classified into three groups: temporal, spatial, and scenario dimensions. Spatial dimensions and past temporal dimensions are continuous and well-established, but future temporal and scenario dimensions (contingency, pattern, profile, etc.) exhibit nuances, including sparsity, variable granularity, and combinatorial properties, depending on the engineering problem.

The Spatial Dimensions

Spatial dimensions in power systems are associated with the topology, geography, and power level scales of electric networks: transmission, distribution, microgrid, commercial, industrial, or residential. For some computational problems such as a power flow, topology and device parameters are essential while geo-referencing is often not relevant. However, there are many contexts in which spatial dimensions are required by the analytics, such as power flows as part of transmission planning. Emerging problems require close integration with spatial dimensions. For instance, consider the case of electric vehicle integration, which requires knowledge of where the sources, storage, and sinks of electric energy may move in the spatial dimensions. Another system that requires detailed geo-referenced information is distribution outage management system (OMS) [23, 24].

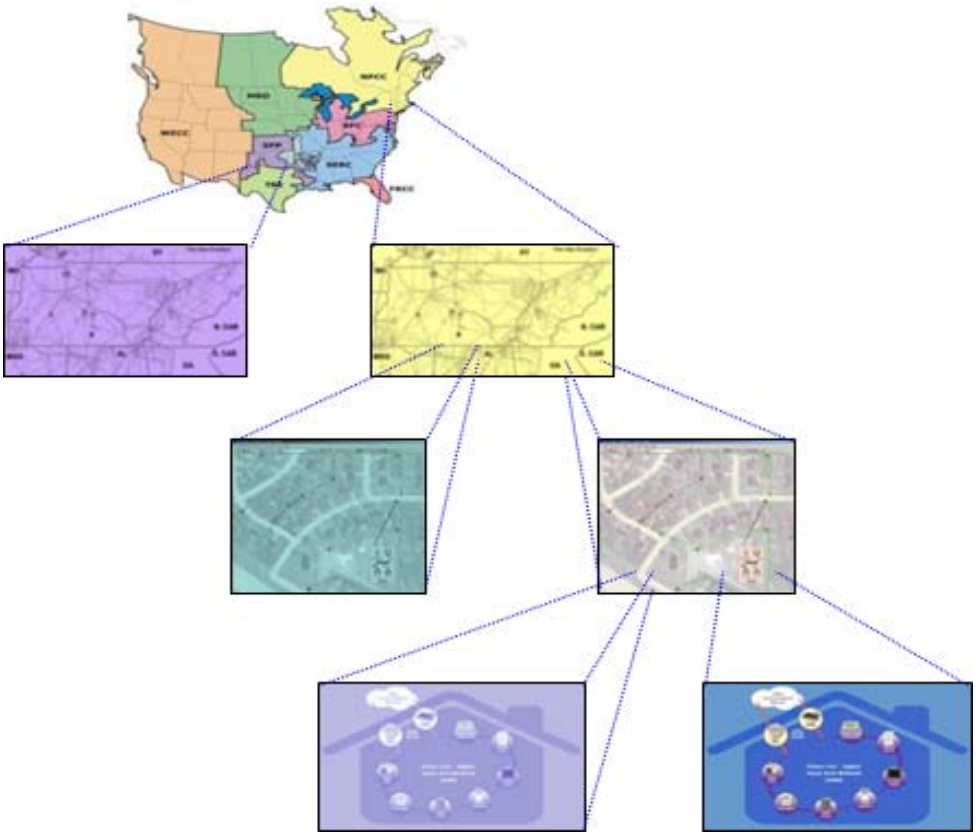


Figure 1: The Spatial Dimension.
Illustration of the spatial dimensions in power systems from interconnections and Independent System Operators (ISOs) levels down to distribution utilities, small cooperatives and municipals, aggregators, microgrids and homes. Although not shown, the spatial scale can go even further to represent individual devices such as electric vehicles and appliances.
For coordination and organization purposes, most of the electricity industry around the world has been organized in a hierarchical manner, with varying levels of discretion regarding ownership, voltage levels, aggregation and control.

Aggregation of information is at the core of spatial dimension handling in electricity systems. The load of customers is aggregated at the feeder level; a town's total load is aggregated as a single 69 kV bus in a transmission model; generation reserve in ancillary service markets is provided at the zonal level, etc. Power systems at various spatial scales are modeled differently. Information exchange and seamless modeling integration requires significant processing. Often, the data from more detailed models has to be aggregated and transferred "by hand" to other applications. Current technologies do not allow information to be dynamically aggregated at different levels without resorting to model conversions [25]. An example is intelligent fault detection based on IED recording and integration with bus/branch models [26].

The Temporal Dimension

Temporal dimensions are associated with natural frequencies of relevant behavior. An example is electricity markets, which operate with long-term contracts, the day-ahead

(hourly) energy market, and the real-time or spot market [27]. Simultaneously, power system transient effects cover the range from lightning propagation (microsecond) to transient stability (ms/sec) to long-term dynamics (min/hour) [28].

The Scenario Dimensions

Due to increased uncertainty about future system conditions, utilities must analyze a large number of scenarios to ensure secure and reliable system operation. Consider for instance a bulk transmission operations planning process that requires determining interface limits for the following week. Those limits will depend on hour of the day, temperature, *N*–2 contingency considerations, RAS modeling, generation profile including wind and solar scenarios, hydro levels, and load patterns [29, 30]. For a typical study the number of contingencies may be on the order of thousands, and the total number of scenarios that need to be evaluated can be on the order of 10^6 to 10^8. The resulting security dimension datasets contain the system conditions and operting regions that should be avoided. To illustrate the scenario dimension, consider Figure 2, which shows the visual pattern of thermal overloads of transmission devices in a very small system obtained by evaluating *N*–1 contingencies. Large-scale systems are hundreds of times larger, and *N*–2 security analysis is usually required.

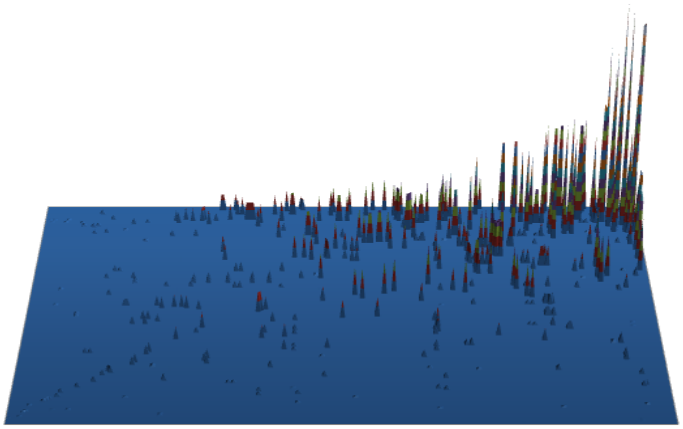


Figure 2: The Scenario Dimension.
Clustered *N*–1 thermal overloads in a 112-branch electric system. The size of the horizontal plane is 112×112. The vertical axis represents the megawatt overload of a transmission component upon the single outage of another transmission device. Branch/Outage dimensional sorting based on total megawatt of overload results in the shown clustering.
Realistic system size is on the order of 10^4 branches, which results in 10^6 to 10^8 data points for a *N*-2 security analysis of the system at a single point in the temporal dimension.

Linear techniques are often used as part of contingency analysis, available transfer capability (ATC), and SCOPF. Usually contingency dimensions are decoupled from scenario dimensions. This means that the contingency analysis must be repeated for every scenario. However, as it was described in Section 1, currently used instantaneous deterministic optimization (SCOPF) may need to evolve into dynamic stochastic

scheduling, which would create temporal inter-dependencies and further increase the number of scenarios [31].

2.2 Power System Scales

Table 1 summarizes the spatial scales of emerging power systems, ranging from continental interconnections to homes and appliances. The table also illustrates the emerging systems for energy control that are progressively being developed and deployed at each level, from Wide-Area Energy Management Systems (WA-EMS) [32, 33] to appliance and device power management [34, 35].

Level	Area Covered	Management	
Interconnection	Country	WA-EMS	Wide-Area EMS
ISO/Control Area	State/Region	EMS	Energy Management System (EMS)
Distribution Utility	County/City	DMS	Distribution Management System
Microgrid	Campus, Neighborhood	μGEMS	Microgrid EMS
Building/Facility	Block, Lot	BEMS	Building EMS
Home	Lot	HEMS	Home EMS
Appliance	Room/outlet	DPM	Device Power Management

Table 1: Spatial Scales

Table 2 lists the common temporal scales associated with relevant power system phenomena ranging from electromagnetic effects to long-term planning [36, 37, 38].

Stage	Time
Planning and Investment	years
Scheduling and Maintenance	months/week
Day Ahead Operations	days/hour
Real-Time Operation and Control	min/sec
Transient Effects and Dynamics	sec/ms

Table 2: Temporal Scales

3 Modeling and Analysis

3.1 Multiple Scales, Parameterization, and Scale-Invariance

Multi-scale processes are physical, engineering, biological, or social processes characterized by coupled phenomena that occur in disparate spatial and temporal scales. Usually, interactions at all scales affect the ultimate behavior of the complete system. Multiscale modeling and analysis considers the following [39, 40]:

a) A library of models that provide a consistent description of phenomena at the various relevant scales,
b) Simulation methods that account for steady state and dynamic scale interactions,
c) Efficient computational analysis, optimization, and control methods,
d) The representation of uncertainty and its propagation.

The first challenge and need for power system multi-scale analysis is to determine what the new relevant scales are, how likely these new scales are to remain contextually relevant in the future, and their important characteristics. As it was described in Section 1, renewable energy variability acts at new temporal scales, requiring changes to the unit commitment and dispatch algorithms. PMU and smart meters require data storage at novel temporal granularity. Power electronics-driven sources operate with ramp rates much faster than conventional generation. These are a few examples of new temporal scales. By the same token there are new spatial scales driven by distributed renewable energy and storage and the industry interest in microgrids, buildings, and homes. Finally, some of the emerging challenges in computation are associated with behavior that is spatio-temporal and multi-scale.

The theory of wavelet transforms has developed enormously in the last decade and is one of the tools that can provide time and scale decomposition of signals. However, there has been slow progress in the corresponding statistical framework to support the development of optimal, multi-scale analysis, in particular spatio-temporal statistical signal processing algorithms.

Research Need 1.
Develop a methodology to identify, correlate, and model relevant emerging behavior in spatial and temporal scales. This method should determine the following:
- The relevant scales for the system,
- How system behavior at a certain scale is communicated or propagated to other scales,
- A mapping of behavior to the various scales.

One of the central themes in multi-scale modeling is the determination of what information on the finer scale is needed to formulate equations or relations for the effective behavior on the coarser scale. Multi-scale modeling provides a balance so that the models become computationally feasible without losing much information [41, 42, 43]. This requires making rational decisions about the appropriate scales used, a process called *parameterization*. An example of this is weather forecasting: phenomena on the micron scale have impacts on planetary scale weather through the physical processes of cloud formation, precipitation, and radiation in conjunction with nonlinear advective processes linking disparate scales. In global computational weather models, all processes beyond the resolution of the numerical model are accounted for through the incorporation of parameterizations. Traditionally these process models are deterministic and empirical.

Recently, there has been interest in moving toward a more proper stochastic treatment of the problem of parameterization of weather processes [44]. A practical example is multi-scale aspects in forecasting of solar power. Accurate forecasting of solar power is of paramount importance to achieve the sustainability objectives of the industry. Solar power production forecasting for utility applications is usually based on satellite images and computation that uses granularity of 3-9 km^2 per cell, and is usually available in the day ahead and few hour-ahead scales. On the other hand, home energy management systems of customers equipped with a solar panel require much higher spatial and temporal accuracy. At the utility scale, the aggregated solar production is likely to be more accurate and less variable compared to the solar fluctuations that a home or a facility can experience [45]. Local, higher resolution spatio-temporal solar forecasting is needed for end consumer applications. Inaccurate forecasting can result in highly suboptimal energy scheduling of both conventional and flexible loads. Further accuracy in forecasting can be achieved by retrieving higher-resolution images and by utilizing adaptive multi-scale computational methods. An adaptive mechanism for adjustment of the resolution in various regions at various spatial and temporal scales can provide an optimal utilization of data to maximize accuracy and forecasting utilization.

Research Need 2.
Develop multi-scale forecasting models and tools that can ensure that information is useful for consumers at various scales. Adaptation to different scales promises to significantly improve the usefulness of forecasting.

3.2 Multi-scale Frameworks

Several analytical frameworks are commonly used in order to represent multi-scale systems [46, 47, 48, 49, 50, 51]. For instance, the power system spatial scales can be organized in a hierarchical structure compatible with the spatial scales listed in Table 1, where each lower level subsystem is seen as a component belonging to the upper level in the hierarchy. A network bus modeled by an ISO can represent the distribution system of a town or even an entire city. It should be noted that the electric system in the lower level is governed by physics similar to the higher level system, but at a lower scale. When the various levels of the hierarchy have similar characteristics, the system is said to exhibit *self-similarity* [52, 53]. One of the advantages of multi-scale systems with self-similarity is that the modeling of the system at various levels can be done by using the same reference model, which enables expansion and collapse of lower level information. Power systems exhibit self-similarity, although this property has not been fully studied. In particular, any power system from large interconnection to home networks can in theory be modeled for steady-state analysis using a generic three-phase, four-wire model.

Research Need 3.
Study the formal spatial self-similarity properties of electricity networks at levels from interconnection to homes and its implications for power system analysis.

This author did not find information describing why the current industry structure is largely hierarchical compare to other networks, such as the internet. In particular, the hierarchy of bulk transmission and distribution is based on voltage level and the distinction of interconnection versus delivery to consumer. Several utilities own both transmission and distribution networks. Often, the differentiation is not clear, and the term "sub-transmission" is used. It would seem that the current hierarchical structures have been implemented historically to address organization issues that arise in multiple spatial scales as opposed to different electrical phenomena. It is to be noted that, contrary to electricity systems, other types of multi-scale systems such as those encountered in physics, materials, and chemistry exhibit fundamentally different physics at various scales [54]. This leads us to the question of whether the electricity network can be seen instead as a non-hierarchical, unified network with a continuous range of scales of generation, transportation, load and storage. Because similar physics takes place at various hierarchical levels of power systems, data at different aggregation levels can theoretically be collapsed or expanded depending on the granularity needed. For instance, a combined SCOPF application may start with transmission level congestion management, but certain buses (distribution systems) may be dynamically expanded depending on the solution state to address distribution system internal constraints or conditions.

Research Need 4.
Study the potential benefits of transmission operators having access to information about the internal operating state of distributions systems: algorithmic issues, data availability and latency aspects, data privacy, and legal aspects.

The need for information and coordination of systems at different scales is apparent during coordinated maintenance and restoration processes involving transmission and distribution. It can be argued that multi-scale self-restoration and self-healing objectives [55, 56] would require automatic interchange of such information and decision logic. It can also be argued that ambitious objectives such as EV-based frequency regulation [57, 58] will require capabilities for seamless expansion of the lower level models into the upper coarser model.

Research Need 5.
Study unified models and algorithms that would allow dynamic expansion and collapse to multiple levels for network applications. Study the implementation of multi-level hierarchical network applications as well as the issues arising from dynamic model expansion and collapse.

Complex systems such as electric power networks are usually comprised of multiple subsystems that exhibit both highly nonlinear stochastic characteristics and are regulated hierarchically. These systems generate signals that exhibit behavior such as dependence on small disturbances and nonstationarity. For example, electricity prices are the result of multi-period decisions made by marketers including long-term bilateral transactions, day-ahead bidding, and real-time market dispatch, and thus they exhibit spikes. Multi-scale signals behave differently depending upon the scale at which the data are examined, e.g., spot prices compared to future prices of energy and reserve. How can the behaviors of such signals on a wide range of scales be simultaneously characterized? One strategy is to develop models that explicitly incorporate the concept of scale so that different behaviors on varying scales can be simultaneously characterized by the scale-dependent measure [59]. The analytics of wind energy production would benefit strongly if explicit methods that embed multi-scale information are developed. Wavelet transforms have been applied to the study of wind production in the temporal scale in the last few years. Reference [60] provides a recent analysis in the range from one second to one hour on wind data at the ISO system level. The study concludes that significant additional data collection would be required from a minimum of 50 sites, distributed over an area of several hundred square miles, to quantify power law aspects. Validation of the power law shape and estimation of the shape parameters would be of service in improving the models used to simulate wind power in large-scale integration studies.

Wavelet transformation methods have been used in various aspects of power system analysis and applications such as fault and disturbance detection, protections, parameter estimation, model identification, and power quality [61, 62, 63]. [64] analyzes power system transients using scale selection wavelet transform methods in the transient frequency range to estimate the transient frequency, scale, and energy. This method shows good performance for detecting, localizing, segmenting, and estimating the energy, frequency, magnitude, duration and over voltage of the transient disturbance, thus enabling the differentiation between oscillatory and impulsive transients. [65] addresses the use of the continuous wavelet transform (CWT) for low-frequency electromechanical oscillations. While there have been powerful developments in wavelet theory and applications to engineering systems in the last decade, the application to power system multi-scale phenomena has been limited to specific ranges. In particular, there is little literature in applications to spatial analysis of distributed renewable energy variability. For instance, in the case of distributed solar, the relevant spatial scales are in the range from ISO to the home level.

Research Need 6.
Investigate applications of multi-scale spatio-temporal wavelet analysis for distributed renewable energy production including wind and solar energy. This analysis can provide a modeling framework for coordinated multi-scale control and optimization of highly distributed systems.

3.3 Scenario Reduction

A problem that arises in the context of multi-dimensional modeling and analysis is the combinatorial nature on the scenario dimension. With increasing penetration of renewables, the number of scenarios that must be tested in planning and operations settings will grow dramatically. The computational effort for solving the problem over each horizon is highly dependent on the number of scenarios considered. Existing methods to reduce the number of scenarios include sampling, clustering, and formal scenario reduction heuristics [66]. Computationally difficult problems that may use scenario reduction arise in areas such as *N–k* contingency analysis, available transfer capability, power system optimization, and power system planning. Contingency coupling screening methods have been proposed to determine the combined effect of multiple event contingencies, which provides a promising method for determining the effect of double contingencies by exploring a reduced number of scenarios [67, 68].

Another method which focuses on reducing solution spaces is variational multi-scale analysis. This method is based on the decomposition of the solution spaces into resolved and unresolved scales and a subsequent derivation of exact governing equations for each scale. In conjunction with appropriate modeling assumptions, these equations serve to provide the basis for variational formulations capable of representing multiscale phenomena. The methods can be applied to a broad class of multiscale phenomena, ranging from problems where scale separation is induced by choice of a discrete space of resolved scales to problems where resolved and unresolved scales may represent concurrent models operating at different space and/or time scales.

Research Need 7.
In the contexts of planning, operations planning, and short term scheduling there is growing need of mechanisms to reduce the solution space of various computation and optimization applications. Methods that inherently capture the multi-dimensional nature of the solution space can be of great service in reducing computational burden.

3.4 Inference and System Identification

In many power system applications it is desirable to identify the properties or parameters of a system without knowing the detailed model of the system. Examples of such applications are, among others: external model determination in ISO systems, load modeling in utility and industrial applications, load transient behavior, and PMU-based disturbance detection [69, 70, 71]. While single-scale system identification is reasonably mature, and there are powerful platforms available, inference and system identification in multi-scale power systems has not been pursued. Such development will include the specification of broader assumptions about the scales or the identification of the actual relevant scales of an unknown system. This result points to a critical need for operational inference related to events and disturbances in external systems. For instance, a utility can detect an internal failure in a microgrid network [72]. A microgrid

can potentially detect cascading problems in the utility system and take preventive actions, including disconnection from the grid upon impending blackout. Similar methods should be investigated to provide highly desired information about the parameters and behavior of the load for a variety of applications, including control and scheduling. Smart meter data may be leveraged with this purpose.

Research Need 8.
Investigation of multi-scale system identification methods in emerging power systems. This includes external multi-scale modeling, load identification, and parameter identification.

4 Optimization, Dynamics, and Control

4.1 Control Paradigm

Multi-scale and multi-dimensional control issues will be pervasive in emerging electricity grids. To describe why this is the case, let us first describe the ongoing transition of electricity control infrastructure to a more distributed paradigm. Figure 3 illustrates the traditional real-time control loop implemented through EMS systems for bulk electricity. Several levels of refinement of the control loop can be realized from SCADA-only control to State Estimation-based control, economic dispatch, and security constrained automatic dispatch. Complex decisions are made both automatically (through the AGC system) and through operator decision making. Figure 3 represents the operational paradigm which has successfully addressed electricity system operations for several decades. EMS systems are often integrated with Market Management Systems (MMS) and Wide-Area Monitoring systems. The EMS system ultimately implements a sophisticated control loop of sensing, estimation, optimization and actuation.

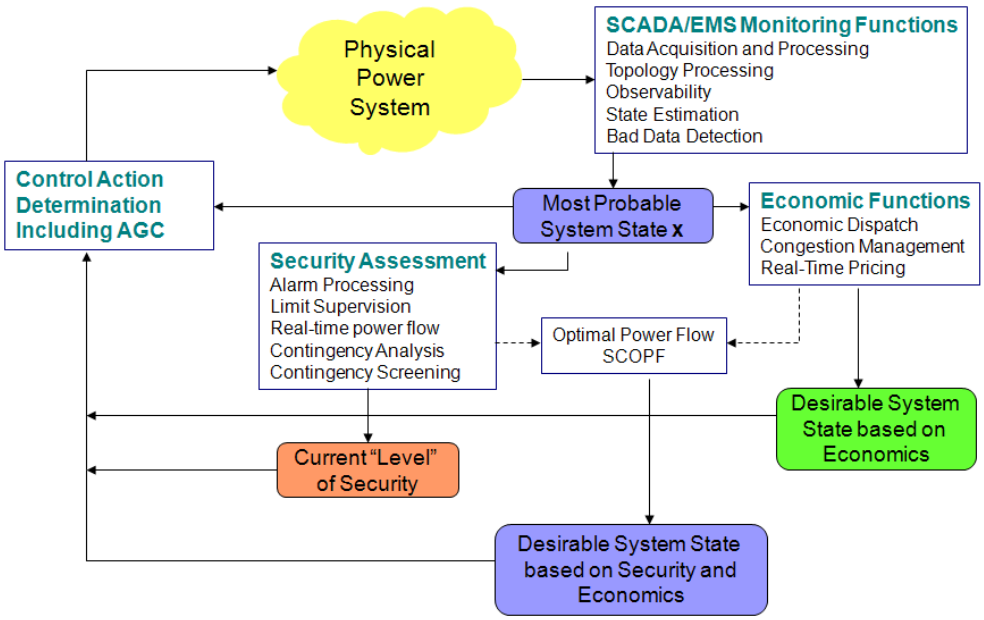


Figure 3: 20th Century Bulk Power System Operating Paradigm.

Operation is based on a control loop implemented as an Energy Management System (EMS). The EMS can provide various levels of feedback sophistication going from SCADA-only operation to instantaneous, deterministic optimization though the security-constrained optimal power flow (SCOPF). Except for instances of AGC control, all the control centers require human operation at the core of operational decisions.

An equivalent control loop takes place in real-time in distribution grids through Distribution Management System (DMS). The corresponding functions of DMS systems have yet to achieve the same level of penetration as EMS systems. For instance, many distribution utilities operate with SCADA-only systems. Ongoing distribution systems automation initiatives [73, 74] will realize higher penetration and automation at the distribution system level.

The next logical step down in the spatial dimension, which has received significant attention by the power system community in the last few years, is the microgrid [75, 76, 77, 78]. The natural technological trend is the development of the equivalent of microgrid energy management systems (μGEMS). Microgrids pose a number of new control and operational requirements, such as capability of isolated operation, self-restoration, integrated control of industrial loads, and self-reconnection. In addition, the microgrid must be able to operate unmanned, e.g., the μGEMS must reach a level of sophistication sufficient to bypass the human operator altogether. The interest in microgrids has been catalyzed by the goal of implementing renewable energy-based systems. There is significant interest in deploying microgrids that can operate in an isolated manner supplied by renewable sources and storage. Renewable energy sources are also highly variable at the microgrid level, which requires more accurate, local DER forecasting and advanced real-time dispatch of inverter-based sources. Beyond

microgrids, similar control systems are expected to be developed and deployed at the building and residential levels through Building Energy Management Systems (BEMS) and Home Energy Management Systems (HEMS) [79,80], respectively. These systems also have the requirement of being unmanned. Deployment of the various categories of EMSs at various scales is the basis for a shift towards distributed control in the electricity industry. The emerging distributed intelligence paradigm is illustrated in Figure 4.

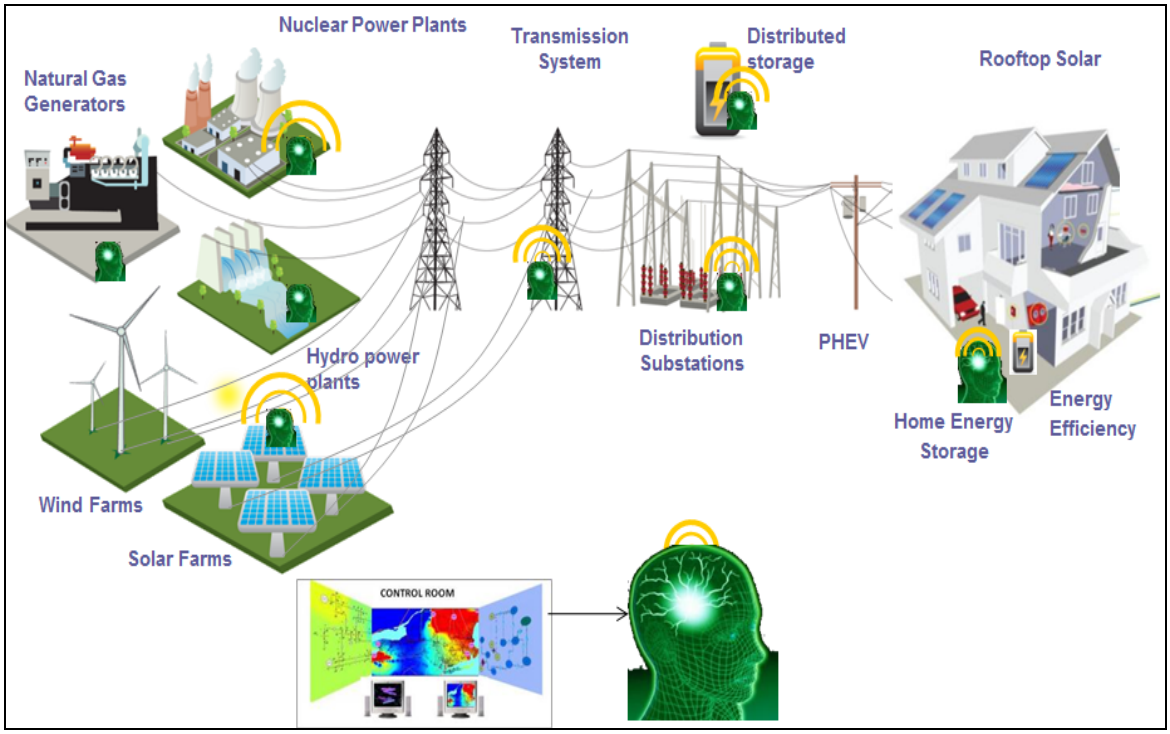


Figure 4: Emerging Distributed Intelligence Paradigm

The level of operational intelligence currently available in bulk EMS systems is brought down the spatial scale to DMS systems, microgrids, buildings and homes. Each one of these elements has its own economic objectives and is equipped with control and logic to make decisions strategically. Instantaneous deterministic optimization is replaced by dynamic stochastic energy scheduling.

EMS systems were designed based on the notion of real-time control of a just-in-time product and its mostly deterministic instantaneous optimization embodied by the Security-Constrained Optimal Power Flow (SCOPF) shown in Figure 3. There are three emerging aspects associated with the temporal scale that will alter the operational philosophy of EMS systems at all scales from bulk operation to homes:

a) Renewable energy variability,
b) Ability to store growing amounts of energy, and
c) Information-driven demand response and demand shifting.

In addition to these strong drivers, both renewable energy source output and demand response capabilities exhibit high uncertainty. Thus the traditional instantaneous

optimization paradigm and algorithms must be replaced by stochastic dynamic scheduling, fully capable of handling the various temporal issues and uncertainty [81]. There are several research needs associated with the development of the next generation dynamic scheduling algorithms, but this paper focuses only on the multi-scale and multi-dimensional aspects.

There is a number of control questions that emerge from the development and deployment of EMS systems at various spatial scales from interconnection to home. To better illustrate the control implications, consider the following example: a very small utility owns conventional and renewable generation and storage and serves only two microgrids. The microgrids also have their own generation and storage. These three systems can be equipped with advanced energy management systems including dynamic stochastic scheduling. The three systems attempt to minimize a cost function subject to multiple security and sustainability constraints. Among several others, the following questions can be formulated:

a) What control framework ensures maximizing global welfare? Is a hierarchical control scheme optimal?
b) If the utility and the microgrids operate individual control systems with homogenous objective functions such as cost minimization, would the control be stable?
c) What information should be exchanged and at what rates to ensure stable distributed control? What happens if the microgrids are allowed to exchange information with each other?
d) How would the results to the previous questions change if instead of two microgrids there were one hundred or one thousand? What is the sensitivity of the control performance as the heterogeneity of the microgrids change?
e) Assuming a market environment, does equilibrium exist? How stable is it?

4.2 Optimization

The area of multi-scale optimization has been the subject of considerable research among the finite element, multigrid, and combinatorial optimization communities. Hierarchical optimization, which is a type of multi-scale optimization, has been the subject of research by the power system community in the last decade [82, 83, 84, 85]. Hierarchical optimization is useful for handling large-scale systems that exhibit scale or organizational properties that can be modeled using a hierarchical arrangement. Computational solutions are obtained by decomposing a system into subsystems and optimizing them while considering the coordination and interactions between the subsystems. Hierarchical optimization takes the interaction into account by placing a *coordinator* above the subsystems to manage it. As a result, the entire system is optimized while considering the interaction between the scales by means of coordinated information. Most of the work related to power system hierarchical control and

optimization addresses two-level problems. There has been little research in multi-level hierarchical optimization and spatio-temporal power system optimization in general.

Research Need 9.
Study of the relevant industry subsystems that would participate in multi-level hierarchical control including staged coordination from ISO to utility to microgrid to end consumer and even to individual devices. Applications such as EV-supported frequency regulation would require such a level of coordination. Phenomena of interest include modeling the heterogeneity of the actors at the various layers, the coordination signals that need to be passed, and the protocols for information passing.

Scale-decomposition is an important concept in multi-scale optimization, which consists in obtaining the solution to loosely coupled optimization problems at various scales through relaxation of interdependencies among scales [86, 87, 88, 89, 90]. A set of similar methods falls under the category *multi-scale relaxation*, which has been applied extensively in multigrid algorithms. Scale decomposition in power systems is relevant when dealing with spatial and temporal dimensions. While little literature on multi-scale decomposition for power networks has been identified, multi-scale relaxation methods could be applied to problems such as wind generation including reserve considerations. The basic idea is to use wavelet decomposition of wind variability to identify and model wind behavior coupled to system control at various frequency ranges. Then the outer coordination control would utilize scale decomposition to solve the global coordinated problem.

Research Need 10.
Modeling of the system in the relevant frequency ranges of DER needs to be fully studied and integrated in a coordinated optimization process. Investigation of scale decomposition and multi-scale relaxation methods can prove to be of service in complex problems such as DER scheduling and dispatch.

Multi-scale decision making is a growing area in multi-scale optimization, which fuses decision theory and multiscale mathematics. The multiscale decision making approach draws upon the analogies between physical systems and complex man-made systems. On the theoretical side, the aim is to develop a generalized statistical formalism that describes a large variety of complex systems in an effective way. Rather than taking into account every detail of the complex system, one seeks for an effective description with few relevant variables. An example of the need of such formalism is decisions associated with transmission expansion investment for a portfolio of large-scale distributed wind resources. Formulating and selecting transmission expansion alternatives is a function of not only the expected amount of power produced and its variance, but also the overall system conditions, including simultaneous operation of other renewable and conventional generation, the existing and planned transmission, $N–k$ security, and load profiles. The level of required transmission expansion also depends on the combined

requirements of transmission for reserve provision from other locations and system production changes in the minutes range, as well as the effect of load growth and policy, acting up to the multi-year range. The application of rather simple transmission reliability metrics has demonstrated that the existing transmission planning methods do not fully capture the optimization problem complexity and reliability constraints. As a result, several systems in the U.S. are less reliable today than in previous years [91]. Large-scale installation of renewable energy can compound this problem considerably if better methods to determine optimal transmission capacity and expansion needs are not developed.

Research Need 11.
Development of multi-scale decision making methodologies for problems such as integrated transmission investment for large-scale renewable energy.

Finally, a key aspect of multi-scale optimization is *multi-scale metrics*, which are measures of relevant quantities involved in the objective function and constraints such as cost, energy, vulnerability, security, time, etc. Appropriate metrics for decision making in many power system planning and operations problems have not been developed. There is growing need of standardized multi-dimensional reliability metrics such as total system $N-k$ security. Such metrics would allow optimization algorithms and decision makers to formulate claims such as: "subsystem A's instantaneous security level is *x units*", or "system A today is 10% less secure than it was 5 years ago". Values of relevant metrics are then passed on and coordinated among the various components at various scales.

Research Need 12.
Formulation of standardized multi-scale reliability metrics that provide differentiated security indices at multiple scales.

4.3 Dynamics

Nonlinear differential equations are used throughout science and engineering. *Singular perturbation* theory examines limiting behavior of multiple time scale systems in which there is an infinite separation of fast and slow time scales. In the last few decades, power system dynamics has studied the issue of two-time scales as part of singular perturbation and fast and slow manifolds associated with transient stability equations [92, 93, 94]. Significant Lyapunov theory has been developed around the concept of singular perturbation for the two-time scale problem. Power system dynamic behavior, though, includes more than two time scales, and emerging behavior may include several other relevant scales. [95] proposes a method that enables modeling synchronous machines systems over diverse time scales in the range covering electromagnetic and electromechanical transients. The models use frequency-adaptive simulation of transients where analytic signals are found to lend themselves to the shifting of the

Fourier spectra. The shift frequency appears as a simulation parameter in addition to the time step size. When setting the shift to the common carrier frequency, the method emulates phasor-based simulation that is very suitable for extracting envelope information at relatively large time step sizes. [96] proposes a three-time scale robust redesign technique, which recovers the trajectories of a nominal control design in the presence of input uncertainties by using two sets of high gain filters. The trajectories of the resulting three-time scale redesigned system approach those of the nominal system when the filter gains are increased. In [97] a method is proposed to detect multi-scale transients and events using non-intrusive load monitoring. Although interest is growing in adaptive frequency methods, most of the work reported deals with two-scale dynamics and is limited to transient stability.

Research Need 13.
Study multi-scale and adaptive multi-scale dynamics methods that capture the emerging behavior in the entire temporal dimension, from the electromechanical to long-term dynamics ranges. These methods shall consider new power electronics and DER technologies. A highly desirable tool is dynamic security-constrained optimal power scheduling.

Wavelet transforms can be applied in the setting of spatio-temporal dynamical systems, where the evolution of a set of dynamical variables is of interest. In this setting, space is usually discrete and time is continuous, which allows for a multi-resolution decomposition [98, 99]. The interdependencies between scales are introduced through the action of wavelet decompositions. The multiscale system dynamics can be modeled by explicit representation of the subsystem dynamics within local components. The system dynamics does not require that the subsystem has reached its asymptotic state. To construct a model for each scale, the individual components can be coupled together to form an ensemble of nonlinear oscillators, a method used in modeling neural systems [100, 101].

The last decade has witnessed significant developments in the application of wavelets to multi-grid dynamics problems that arise in the various topics of material science, flow simulation, chemistry, and biology. Multigrid methods are a group of algorithms for solving differential equations using a hierarchy of discretizations. They are an example of a class of techniques called multiresolution methods, which are very useful in problems exhibiting complex behavior at multiple scales. For example, many basic relaxation methods exhibit different rates of convergence for short- and long-wavelength components, suggesting these different scales be treated differently. The main idea of multigrid methods is to accelerate the convergence of a basic iterative method by global correction from time to time, accomplished by solving a coarse problem. While multigrid methods have found a large number of applications in partial differential equations, they have not been, to the author's knowledge, used in conjunction with multi-scale power system dynamics modeling. The electricity network

shares some graph-fundamental properties with other types of networks such as communications and transportation, but Kirchhoff Laws usually make some of the graph-theoretical results available to multi-grid applications not directly applicable to electric networks.

Research Need 14.
Investigation of the feasibility to integrate multigrid methods with power system spatial multi-scale dynamics, where the multi-grid methods are used to enable multi-scale coordination and multi-scale relaxation.

4.4 Control

Multi-scale system control refers to a formal framework that uses multi-scale modeling and theory to realize desirable control and system response at all scales. Hierarchical control is gaining impetus among the smart grid and automation community [46, 85, 83, 85], but a comprehensive framework at all the relevant spatial levels is yet to be established. Furthermore, the system control must address control at all the relevant temporal scales. This is necessary for a variety of emergent problems such as distributed frequency regulation, demand response, ancillary services provision, disturbance detection, wholesale/retail markets, etc.

Research Need 15.
Design of a spatio-temporal multi-scale control framework including:
- Definition of relevant space and time dimensions, their modeling and parameters.
- Information exchange requirements between the various spatial scales and communication requirements.
- Control property assurance: observability, controllability, stability, reachability, etc.

We have mentioned that the properties of self-similarity and scale-free systems may have implications for power system multi-scale control methods. Normalization, traditionally used in power system analysis, suggests scale-free characteristics. This would also pose the question of whether multi-scale hierarchical control, networked control, or a fusion of the two is possible in multi-scale electricity networks.

Research Need 16.
Investigation of multi-scale networked control properties in power systems, including their advantages compared to hierarchical control.

5 Data Handling and Visualization

5.1 Multi-scale Data

The electricity industry transformation is already resulting in massive data being acquired and stored in areas such as smart metering, substation IEDs, and PMUs. This data should not be understood as being isolated, but as part of a multi-dimensional, multi-scale data set that will ultimately enable the realization of the various grid objectives and lead to the future electricity industry [102, 103, 104]. Information architecture and data set modeling are initiatives under the umbrella of "smart grid" that have not universally been adopted with a system-wide perspective. For instance, while some utilities read smart meters every 15 minutes and are able to pass the readings to the control center every hour, others are able to read once per hour and report once a day. The type of control and the realized demand response that can be achieved will depend on the speed of feedback information. Nevertheless, there is tremendous value in the emerging data for understanding the behavior of new subsystems such as consumer behavior and, more fundamentally, to better understand the type of data acquisition that is ultimately needed to design controls and realize future grid objectives.

Research Need 17.
There is growing need to organize the emerging data in structured, flexible ways, building scalable information architectures that can evolve as more data at the various temporal and spatial scales is acquired and as the future grid requirements evolve. There is also a need to architect the security, availability, and integrity of the data.

5.2 Visualization

Electricity grid visualization has become increasingly important and a topic of significant research in the last decade [105, 106, 107]. Visualization has been the mechanism to drastically increase situational awareness in bulk energy control centers. Systems with 2D-spatial visualization of the controlled region are the norm. Typically, the visualization requirements for these systems are at the level of 10^3 to 10^4 measurements every 5-10 seconds. Zooming and panning is usually too slow for real-time response in this environment. In general, interactive visualization in real-time has not been possible due to slow performance. Spatio-temporal and spatial-security 3D visualization has been proposed, but it has not been deployed for either operations or planning due to performance limitations even with Graphics Processing Units (GPUs) [108, 109, 110].

Visualization at the distribution level is driven by geo-referencing requirements and hence is coupled with Geographic Information Systems (GIS). The norm is a 2D-spatial GIS representation. Neither spatio-temporal nor spatial-security techniques have been proposed for distribution systems. Cluster GPU applications have been proposed but

have not been implemented at control centers. 4D (spatial + temporal + security) has been proposed by the author for small data sets. Multi-scale 2D spatial visualization is rudimentary, available only for snapshot data sets and in off-line modes.

Because of the performance and design of the GPU, highly parallelized and efficient computation is possible. Several General Purpose GPU (GPGPU) libraries have been developed to provide a better match to the computing domain. In recent years, numerous algorithms and data structures have been ported to run on GPUs. Scientific visualization generally has a spatial 2D or 3D mapping and can thus be expressed in terms of the graphics primitives of shader languages. In general, the current GPU programming model is well-suited to managing large amounts of spatial data. An example of a GPU generated image is shown in Figure 5. GPU visualization for large grids has been reported to be about 30 times faster than CPU-based visualization [111, 112, 113]. GPU clusters can theoretically yield visualization speeds in the order of 10^3 times the current levels, opening a large number of research and application possibilities. Multi-scale, multi-dimensional visualization represents an enabling technology for the understanding and study of emerging power systems. While in the past, visualization has been seen as a "presentation" tool, there is growing acknowledgement that visualization tools are becoming an enabler for advanced analytics, especially in multi-scale, multi-dimensional systems.

Figure 5: Conceptual Multi-scale Visualization for Coordination of Renewables and Distributed Storage A scattered cloud system results in highly variable solar production near afternoon commuting time. The Atlanta area exhibits potential problems due to loss of stationary storage (red contouring driven by aggregated security metrics). Variability absorption capabilities at 2x2 square mile resolution are illustrated in the upper 3D layer using cylinders (virtual generation). Fuchsia stacks are EV V2G energy, blue stacks are frequency regulation, and green and yellow stacks are 5 and 12 minute EV cleared reserve respectively.

Research Need 18.
Development of integrated multi-scale, multi-dimensional dynamic visualization tools with applications to multi-scale power system dynamics, PMU and smart meter massive data visualization, *N–k* contingency information visualization, and 4-D dynamic navigation.

6 Concluding Remarks

There is a significant number of emerging challenges in power system engineering that are associated with phenomena that is essentially multi-dimensional and multi-scale. This paper has conducted a literature review and exploration of research needs in multi-dimensional and multi-scale electric power modeling and computation. While multi-scale methods have found a large number of applications in other domains such as physics, materials engineering, chemistry and biology, our investigation reveals that there is substantially less literature in multi-scale methods applied to power systems,

control. It is fair to say that the level of multi-scale understanding in other engineering domains is far ahead compared that of electrical energy systems.

This paper has identified and discussed eighteen research needs classified in three broad areas: a) modeling and analysis, b) optimization, dynamics and control, and c) data handling and visualization. A significant set of the research needs are fundamental and would address basic questions such as, "What are the novel relevant scales in space and time?" or, "What are the properties of multi-scale electricity systems with regard to self-similarity and aggregation. A significant challenge for the community will be multi-scale modeling and unification, because for historical reasons vastly different modeling assumptions, objectives, and formats have been established at the various scales.

The topic of multi-scale statistics and inference has great relevance to what will be possible in the future electricity industry. In particular, multi-scale system identification and multi-scale forecasting can provide solid models and input data to the optimization and control stages. Wavelet theory stands out as one of the most powerful tools to address multi-scale analysis, but there is limited literature on spatio-temporal wavelet applications to power.

Acquiring a greater understanding of multi-scale dynamic behavior, including novel power electronics devices and DER variability, is essential to the power system community. Issues such as multi-dimensional dynamic stochastic scheduling are of paramount relevance for efforts associated with integration of renewable energy.

Finally, there are numerous unresolved aspects of multi-scale, multi-dimensional data acquisition, processing, and storage. Part of the challenge is to determine what data is needed in the first place, which a) depends on the assumed control architecture, and b) determines the sensing systems deployed. Because what is possible in the control arena depends on what the community learns about emerging multi-scale behavior, this will be an iterative and evolving topic of research. Multi-dimensional, multi-scale visualization can play a significant role in assisting researchers in obtaining a better understanding of the emerging complex phenomena.

References

[1] E. Ela, B. Kirby, "ERCOT Event on February 26, 2008: Lessons Learned", Technical Report, NREL/TP-500-43373, July 2008

[2] NERC, "Electric Industry Concerns on the Reliability Impacts of Climate Change Initiatives" Special Report, November, 2008.

[3] E.M. Constantinescu, V.M. Zavala, M. Rocklin, S. Lee, M. Anitescu, "Unit Commitment with Wind Power Generation: Integrating Wind Forecast Uncertainty and Stochastic Programming", Argonne National Laboratory, September, 2009.

[4] M.H. Albadi, E. F. El-Saadany. Overview of wind power intermittency impacts on power systems. Electric Power Systems Research, 80(6):627-632, 2010.

[5] P. A. Ruiz, C. R. Philbrick and P. W. Sauer, "Day-ahead uncertainty management through stochastic unit commitment policies", in Proc. 2009 IEEE Power Systems", Seattle, WA, March 2009.

[6] Powell, W., George, A., Lamont, A., & Stewart, J. "SMART: A Stochastic Multiscale Model for the Analysis of Energy Resources, Technology and Policy", 2009.

[7] S.D. Braithwait et al. "Load Impact Evaluation of California Statewide Critical Peak Pricing (CPP) for Non-Residential Customers Ex- Post and Ex Ante Report", Christensen Associates, May 1, 2009.

[8] A. Faruqui, R. Hledik, J. Tsoukalis, "The Power of Dynamic Pricing," The Electricity Journal, April 2009.

[9] P. Cappers, C. Goldman, D. Kathan, "Demand response in U.S. electricity markets: Empirical evidence", Energy 35(4), 1526-1535., 2009.

[10] P. Faria, Z.A. Vale, J. Ferreira, "Demsi: A demand response simulator in the context of intensive use of distributed generation", 2010 IEEE International Conference on Systems Man and Cybernetics (SMC), Page(s): 2025 – 2032, 2010.

[11] DOE Request for Information on Smart Grid Policy, September, 2010.

[12] H. Vogt, H. Weiss, P. Spiess, A.P. Karduck, A.P., "Market-based prosumer participation in the smart grid", 2010 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST), Page(s): 592 – 597, 2010.

[13] I. Lampropoulos, G.M.A. Vanalme, W.L. Kling, "A methodology for modeling the behavior of electricity prosumers within the smart grid", 2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe), Page(s): 1 – 8, 2010.

[14] N. Kakogiannis, P. Kontogiorgos, E. Sarri, G.P. Papavassilopoulos, "Games among long and short term electricity producers and users", 7th Mediterranean Conference and Exhibition on Power Generation, Transmission, Distribution and Energy Conversion (MedPower 2010), 2010.

[15] CEATI International Inc, Power System Planning and Operations Interest Group, Invitation for Proposals, "Improved T&D Interfaces and Information", Dec. 2010.

[16] L. Castanheira, et al., "Coordination of transmission and distribution planning and operations to maximise efficiency in future power systems", 2005 International

Conference on Future Power Systems, Digital Object Identifier: 10.1109/FPS.2005.204305, 2005, pp. 1-5.
[17] S. Fan, L. Chen, and W. J. Lee, "Short-term load forecasting using comprehensive combination based on multi-meteorological information," in Proc. IEEE Ind. Commercial Power Syst. Tech. Conf., May 4–8, 2008, pp. 1–7.
[18] W. Chu et al. "Multi-region Short-Term Load Forecasting in Consideration of HI and Load/Weather Diversity", IEEE Transactions on Industry Applications, Vol. 47, No. 1, Digital Object Identifier: 10.1109/TIA.2010.2090440, 2011, pp. 232-237.
[19] A.P.S. Meliopoulos et al., "Advances in the Super Calibrator Concept - Practical Implementations", HICSS 40th Hawaii International Conference on System Sciences, 2007, Digital Object Identifier: 10.1109/HICSS.2007.49, 2007.
[20] P. Maghouli, S.H. Hosseini, M.O. Buygi, M. Shahidehpour, "A Scenario-Based Multi-Objective Model for Multi-Stage Transmission Expansion Planning", IEEE Transactions on Power Systems, Vol. 26, No. 1, pp. 470-478, 2011.
[21] A.H. Escobar, R.A. Romero, R.A. Gallego, "Transmission network expansion planning considering multiple generation scenarios", 2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America, Digital Object Identifier: 10.1109/TDC-LA.2008.4641804, 2008.
[22] N.M. Razali, A.H. Hashim, "Backward reduction application for minimizing wind power scenarios in stochastic programming", 2010 4th International Power Engineering and Optimization Conference (PEOCO), Digital Object Identifier: 10.1109/PEOCO.2010.5559252, pp. 430-434, 2010.
[23] K. Sridharan, N. Schulz, "Outage Management through AMR Systems Using an Intelligent Data Filter", IEEE Power Engineering Review, Vol 21, No. 8, Digital Object Identifier: 10.1109/MPER.2001.4311582, 2001.
[24] Y. Dong, V. Aravinthan, M. Kezunovic, W. Jewell, "Integration of asset and outage management tasks for distribution systems", IEEE Power & Energy Society General Meeting, Digital Object Identifier: 10.1109/PES.2009.5275527, 2009.
[25] C. M. Marzinzik, S. Grijalva, J. D. Weber, "Experience Using Planning Software to Solve Real-Time Systems", Proceedings of the Hawaii International Conference on System Sciences,2009.
[26] Dong, Y.; Aravinthan, V.; Kezunovic, M.; Jewell, W., "Integration of asset and outage management tasks for distribution systems", IEEE Power & Energy Society General Meeting, Digital Object Identifier: 10.1109/PES.2009.5275527, 2009.
[27] S. Grijalva, S. Dahman, K. Patten, A. Visnesky, "Large-Scale Integration of Wind Generation Including Network Temporal Security Analysis", IEEE Transactions on Energy Conversion, Vol. 22, No.1, March, 2007.
[28] K. Sakellaris, "Modeling electricity markets as two-stage capacity constrained price competition games under uncertainty", 2010 7th International Conference on the European Energy Market (EEM), Digital Object Identifier: 10.1109/EEM.2010.5558755, 2010.
[29] P. W. Sauer, M. A. Pai, "Power System Dynamics and Stability", Prentice Hall, 1998.

[30] R.D. Dosano, S. Hwachang, L. Byongjun, "Network centrality based N-k contingency scenario generation", IEEE Transmission & Distribution Conference & Exposition: Asia and Pacific, Digital Object Identifier: 10.1109/TD-ASIA.2009.5356963, 2009.

[31] C. Qiming, J.D. McCalley, "Identifying high risk N-k contingencies for online security assessment", IEEE Transactions on Power Systems, Vol. 20, No. 2, pp. 823-834, 2005.

[32] N. Issarachai, "A Robust SMES Controller Design for Stabilization of Inter-area Oscillations Based on Wide Area Synchronized Phasor Measurements", Electric Power Systems Research, Vol. 79 Issue 12, p1738-1749. Dec. 2009

[33] D. Novosel, "Requirements of large-scale wide area monitoring, protection and control systems," in Proc. Fault and Disturbance Analysis Conference, Atlanta, GA, Apr. 2007.

[34] T. Hubert, S. Grijalva "Realizing Smart Grid Benefits Requires Energy Optimization Algorithms at Residential Level", *IEEE Conference on Innovative Smart Grid Technologies* (*ISGT*), Anaheim, California, January 17-19, 2011

[35] T. Ikegami, Y. Iwafune, K. Ogimoto, "Optimum operation scheduling model of domestic electric appliances for balancing power supply and demand", 2010 International Conference on Power System Technology (POWERCON), Digital Object Identifier: 10.1109/POWERCON.2010.5666592, 2010, pp. 1-8.

[36] J.B. Cardell, C.L. Anderson, "Analysis of the System Costs of Wind Variability through Monte Carlo Simulation," 2010 43rd Hawaii International Conference on System Sciences, pp. 1-8, 2010.

[37] H. Holttinen, "Estimating the impacts of wind power on power systems—summary of IEA Wind collaboration", Environmental Research Letters, 28 April, 2008

[38] T. Niknam, A. Khodaei. F. Fallahi, "A new decomposition approach for the thermal unit commitment problem." Applied Energy 86 (9) pp.1667-1674, 2007

[39] A. Benveniste, R. Nikoukhah, A.S. Willsky, "Multiscale System Theory", IEEE Transactions on Circuits and Systems I, No. 1, pp. 2-15, 1994.

[40] G. Guenard, P. Legendre, D. Boisclair, M. Bilodeau, "Multiscale Codependence Analysis", Ecology Oct2010, No. 10, pp. 2952-2964, 2010.

[41] H. Zhao, "Multiscale Analysis and Prediction of Network Traffic", 2009 IEEE 28th International Performance Computing and Communications Conference, No. 28, 388-393, 2009.

[42] S. J. Benson, L. C. McInnes, J. J. Mor´e, and J. Sarich, Scalable Algorithms in Optimization: Computational Experiments, Preprint ANL/MCS-P1175-0604, Mathematics and Computer Science, Argonne National Laboratory, Argonne, IL, 2004; in Proceedings of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization (MA&O) Conference, Albany, NY, 2004.

[43] R. M. Lewis and S. G. Nash, "Model problems for the multigrid optimization of systems governed by differential equations", SIAM J. Sci. Comput., 26 (2005), pp. 1811–1837.

[44] K. Uchida, T. Senjyu, N. Urasaki, A. Yona, "Installation effect by solar heater system using solar radiation forecasting", 2009 IEEE Transmission & Distribution Conference & Exposition: Asia and Pacific, Digital Object Identifier: 10.1109/TD-ASIA.2009.5356904, 2009, pp. 1 – 4.

[45] M.A.A. Pedrasa, T.D. Spooner, I.F. MacGill, "The value of accurate forecasts and a probabilistic method for robust scheduling of residential distributed energy resources", 2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Digital Object Identifier: 10.1109/PMAPS.2010.5528952, 2010, pp. 587 – 592.

[46] H. Mori and M. Saito, "Hierarchical OO-PTS for Voltage and Reactive Power Control in Distribution Systems," Proc. of 2003 IEEE PES Power Tech, Vol. 2, pp. 353 - 358, Bologna, Italy, June 2003.

[47] S. Grijalva, A. Visnesky, "The Effect of Generation on Network Security: Spatial Representation, Metrics, and Policy", Vol 21, No. 23, August 2006

[48] M.A. Serrano, M. Boguna, A. Vespignani, "Extracting the Multiscale Backbone of Complex Weighted Networks", Proceedings of the National Academy of Sciences of the United States of America 4/21/2009, No. 16, 6483-6488, 2009.

[49] J.H. Brown, V.K. Gupta , Bai-Lian Li , Bruce T Milne , Carla Restrepo , Geoffrey B West, " The Fractal Nature of Nature", Philosophical Transactions, No. 1421,619-626, 2002.

[50] Y. Kajikawa, Y. Takeda, I. Sakata, K. Matsushima, "Multiscale Analysis of Interfirm Networks in Regional Clusters", Technovation 30 (3), No. 3, 168-180, 2010.

[51] R Criado , J Pello , M Romance , M Vela-Perez, "A Node-based Multiscale Vulnerability of Complex Networks", International Journal of Bifurcation & Chaos in Applied Sciences & Engineering Feb2009, No. 2, 703-710, 2009.

[52] J. Maver, "Self-Similarity and Points of Interest", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 32, No. 7, pp. 1211 – 1226, 2010.

[53] C. Vanderkerckhove, "Macroscopic Simulation of Multiscale Systems Within the Equation-Free Framework", Ph.D. Thesis, Katholieke Universiteit Leuven, June 2008.

[54] S. Boccaletti et al., "Multiscale Vulnerability of Complex Networks", Chaos Dec2007, No. 4, 43110-43113, 2007.

[55] K. Rasmussen, "A real case of a self-healing network", CIRED 20th International Conference and Exhibition on Electricity Distribution - Part 2, 2009.

[56] M. Amin, "Challenges in reliability, security, efficiency, and resilience of energy infrastructure: Toward smart self-healing electric power grid", 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, Digital Object Identifier: 10.1109/PES.2008.4596791, 2008,pp. 1-5.,

[57] Y. Gu; L. Xie, "Look-ahead coordination of wind energy and electric vehicles: A market-based approach", North American Power Symposium (NAPS), 2010.

[58] Pillai, J.R.; Bak-Jensen, B., "Integration of Vehicle-to-Grid in the Western Danish Power System", IEEE Transactions on Sustainable Energy, Vol. 2, No. 1., pp. 12-19, 2011.

[59] T.T. Ho, A.B. Frakt, A.S. Willsky, "Multiscale Realization and Estimation for Space and Space-Time Problems", ISIT. Cambridge. MA. USA, August 16 - 21, 1998.

[60] Katie Coughlin and Joseph H. Eto, "Analysis of Wind Power and Load Data at Multiple Time Scales", LBNL, 2010.

[61] L. Lin; H. Huang; W. Huang, "Review of power quality signal compression based on wavelet theory", ICTM '09. International Conference on Test and Measurement, 2009, Digital Object Identifier: 10.1109/ICTM.2009.5412951, pp. 235 – 238, 2009.

[62] W. Kanitpanyacharoean, S. Premrudeepreechacharn, "Power quality problem classification using wavelet transformation and artificial neural networks", IEEE PES Power Systems Conference and Exposition, 2004, Digital Object Identifier: 10.1109/PSCE.2004.1397630, pp. 1496 – 1501.

[63] J. Wang, H. Shu, X. Chen, " Multiscale Morphology Analysis of Dynamic Power Quality Disturbances", Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering VOL 24 ISSU 4 PAGE 63-67 DATE April 2004, No. 4,63-67, 2004.

[64] J. Zhou; Z. Wang, "Detection of Dynamic Power Quality Disturbance Based on Lifting Wavelet", 2010 International Conference on Artificial Intelligence and Computational Intelligence (AICI), Digital Object Identifier: 10.1109/AICI.2010.142, pp. 93-95, 2010.

[65] J.L. Rueda, C.A. Juárez, I. Erlich, "Wavelet-Based Analysis of Power System Low-Frequency Electromechanical Oscillations", IEEE Transactions on Power Systems, Digital Object Identifier: 10.1109/TPWRS.2010.2104164, 2011.

[66] A.K. Noha, "Effects of integrating wind power for representative load scenarios in a us electric power system: Operational costs and environmental impacts", 2010 9th International Conference on Environment and Electrical Engineering (EEEIC), Digital Object Identifier: 10.1109/EEEIC.2010.5489969, pp. 186-189, 2010.

[67] V. Donde, "Severe Multiple Contingency Screening in Electric Power Systems", IEEE Transactions on Power Systems, Vol. 23, No. 2, pp. 406-417, May 2008.

[68] C.Davis, T.J. Overbye, "Linear Analysis of Multiple Outage Interaction," 42nd Hawaii International Conference on System Sciences, 2009.

[69] A. Borghetti et al., "Continuous-Wavelet transform for fault location in distribution power networks: definition of mother wavelets inferred from fault originated transients," IEEE Trans. Power Systems, vol. 23, pp. 380-388, 2008.

[70] B. Tianshu et al., "On-line parameter identification for excitation system based on PMU data", CRIS 2009 - Fourth International Conference on Critical Infrastructures, 2009, Digital Object Identifier: 10.1109/CRIS.2009.5071494.

[71] G. Kalogridis et al., "Privacy for Smart Meters: Towards Undetectable Appliance Load Signatures", 2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm), Digital Object Identifier: 10.1109/SMARTGRID.2010.5622047, pp. 232-237.

[72] M. Arenas-Martínez et al., "A Comparative Study of Data Storage and Processing Architectures for the Smart Grid", 2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm), Digital Object Identifier: 10.1109/SMARTGRID.2010.5622058, pp. 285-290.

[73] D. Haughton, G.T. Heydt, G.T., "Smart distribution system design: Automatic reconfiguration for improved reliability", 2010 IEEE Power and Energy Society General Meeting, Digital Object Identifier: 10.1109/PES.2010.5589678, pp, 1-8, 2010.

[74] H.M. Gill, "Smart Grid distribution automation for public power", 2010 IEEE PES Transmission and Distribution Conference and Exposition, Digital Object Identifier: 10.1109/TDC.2010.5484298, 2010.

[75] C.M. Colson, M.H. Nehrir, "A Review of Challenges to Real-Time Power Management of Microgrids", 2009 IEEE Power & Energy Society General Meeting, Calgary, Canada, July 2009.

[76] Eto, J. et al., "Overview of the CERTS Microgrid laboratory Test Bed", 2009 CIGRE/IEEE PES Joint Symposium on Integration of Wide-Scale Renewable Resources Into the Power Delivery System, 2009.

[77] P. Basak, A.K. Saha, S. Chowdhury, "Microgrid: Control techniques and modeling", 2009 Proceedings of the 44th International Universities Power Engineering Conference (UPEC), 2009.

[78] A.A. Zaidi, F. Kupzog, "Microgrid automation - a self-configuring approach", IEEE International Multitopic Conference, (INMIC), 2008, Digital Object Identifier: 10.1109/INMIC.2008.4777802, pp. 565- 570, 2008.

[79] M. Xudong et al., "Supervisory and Energy Management System of large public buildings", 2010 International Conference on Mechatronics and Automation (ICMA), Digital Object Identifier: 10.1109/ICMA.2010.5589969, pp. 928-933, 2010.

[80] P. Zhao, S. Suryanarayanan, M.G. Simões, "An Energy Management System for Building Structures Using a Multi-Agent Decision-Making Control Methodology", 2010 IEEE Industry Applications Society Annual Meeting (IAS), Digital Object Identifier: 10.1109/IAS.2010.5615412.

[81] D. Wunsch, W. Powell, A. Barto, J. Si, "Toward Dynamic Stochastic Optimal Power Flow", Handbook of Learning and Approximate Dynamic Programming, Digital Object Identifier: 10.1109/9780470544785 .ch22, pp. 561-598, 2004.

[82] S. Han, S. H. Han, K. Sezaki, "Design of an optimal aggregator for vehicle-to-grid regulation service", 2010 Innovative Smart Grid Technologies (ISGT), Digital Object Identifier: 10.1109/ISGT.2010.5434773.

[83] Ming Chen et al., "Hierarchical Utilization Control for Real-Time and Resilient Power Grid", 21st Euromicro Conference on Real-Time Systems, 2009. ECRTS '09, Digital Object Identifier: 10.1109/ECRTS.2009.19.

[84] M.E. Torres-Hernandez, M. Velez-Reyes, "Hierarchical control of Hybrid Power Systems",. 11th IEEE International Power Electronics Congress, CIEP 2008, Digital Object Identifier: 10.1109/CIEP.2008.4653837, pp. 169-176, 2008.

[85] Careri, F.; Genesi, C.; Marannino, P.; Montagna, M.; Rossi, S.; Siviero, I., "Definition of a zonal reactive power market based on the adoption of a

hierarchical voltage control", 2010 7th International Conference on the European Energy Market (EEM), Digital Object Identifier: 10.1109/EEM.2010.5558672, 2010.

[86] A. Chakrabortty, M.A. Arcak, "Three-time-scale redesign for robust stabilization and performance recovery of nonlinear systems with input uncertainties", 46th IEEE Conference on Decision and Control, 12-14 Dec. 2007, pp. 3484 - 3489 .

[87] S. Attinger and P.D. Koumoutsakos, Multiscale Modeling and Simulation, Springer 2004.

[88] K. Muralidharam, S.K. Mishra, G. Frantziskonis, P.A.Deymier, P. Nukala, S. Simunovic, S. Pannala, "Dynamic compound wavelet matrix for multiphysics and multiscale problems", Physical Review E, 77, pp. 026714-1 to 026714-14, (2008).

[89] C. Liu, M. Shahidehpour, and Le Wu, "Extended Benders Decomposition for Two Stage SCUC", Vol. 25, No. 2, May. 2010.

[90] J.E. Geiser, "Decomposition Methods in Multiphysics & Multiscale Problems: Models & Simulation", Nova Science Publishers Inc., ISBN 9781617286117, 2011.

[91] S. Grijalva, S. Dahman, K. Patten, A. Visnesky, "Large-Scale Integration of Wind Generation Including Network Temporal Security Analysis", IEEE Transactions on Energy Conversion, Vol. 22, No.1, March, 2007.

[92] Alberto, L.F.C. Chiang, H.D, "Theoretical foundation of CUEP method for two-time scale power system models", IEEE Power & Energy Society General Meeting, 26-30 July 2009, Calgary, AB, Canada, ISBN: 978-1-4244-4241-6.

[93] F Gao , K Strunz, " Frequency-Adaptive Power System Modeling for Multiscale Simulation of Transients", IEEE TRANSACTIONS ON POWER SYSTEMS 24 (2), No. 2, 561-571, 2009.

[94] L. Zhang, Q. Pan, P. Bao, H. Zhang, "The Discrete Kalman Filtering of a Class of Dynamic Multiscale Systems", IEEE Transaction on Circuits and Systems II: Analog and Diginal Signal processing, Vol. 49, No. 10, October, 2002.

[95] F. Gao , K. Strunz, " Multi-scale Simulation of Multi-machine Power Systems", International Journal of Electrical Power & Energy Systems Oct2009, No. 9, 538-545, 2009.

[96] A. Chakrabortty, M. Arcak, "A Three-time-scale redesign for robust stabilization and performance recovery of nonlinear systems with input uncertainties", 46th IEEE Conference on Decision and Control, New Orleans, LA, 2008, pp. 3484 – 3489.

[97] S. B. Leeb , J. L Kirtley, " Multiscale Transient Event Detector for Nonintrusive Load Monitoring", IECON Proceedings (Industrial Electronics Conference) VOL 1 ISSU PAGE 354-359 DATE 1993, No. 1,354-359, 1993.

[98] K. Kikuchi , B. Wang, " Spatiotemporal Wavelet Transform and the Multiscale Behavior of the Madden-Julian Oscillation*", Journal of Climate 23(14) 3814 Jul 15, No. 14, 3814-2010.

[99] R.H.G Tan, V.K. Ramachandaramurthy, "Power system transient analysis using scale selection wavelet transform", 2009 IEEE Region 10 TENCON Conference, 23-26 Jan. 2009, Singapour.

[100] C. Selle, M. West, "Multiscale Networks for Distributed Consensus Algorithms", Proceedings of the IEEE Conference on Decision and Control, No. 1, 4753-4758, 2009.
[101] M. Stork, J. Hrusak, D. Mayer, "Nonlinear dynamics of coupled oscillators: State space energy approach 2010 International Conference on Applied Electronics (AE), 2010 , Page(s): 1 – 4.
[102] Bin Lu; Wei Song, "Research on heterogeneous data integration for Smart Grid", 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Digital Object Identifier: 10.1109/ICCSIT.2010.5564620, 2010, pp.: 52 – 56.
[103] G.N. Srinivasa Prasanna, A. Lakshmi, S. Sumanth, V. Simha, J. Bapat, G. Koomullil, "Data communication over the smart grid", 2009 IEEE International Symposium on Power Line Communications and Its Applications (ISPLC), Digital Object Identifier: 10.1109/ISPLC.2009.4913442, 2009, pp.: 273 – 279.
[104] J.H. Shin, D.J. Yi, Y.I. Kim, H.G. Lee, K.H. Ryu, "Spatiotemporal Load-Analysis Model for Electric Power Distribution Facilities Using Consumer Meter-Reading Data", IEEE Transactions on Power Delivery, Digital Object Identifier: 10.1109/TPWRD.2010.2091973, 2011.
[105] T.J. Overbye, R. Klump, J.D. Weber, "Interactive 3D Visualization of Power System Information", Electric Power Components and Systems, Volume 31, Issue 12 December 2003 , pages 1205 – 1215.
[106] A.Venkatesh, G. Cokkinides, A. P. Meliopoulos, "3D-Visualization of Power System Data Using Triangulation and Subdivision Techniques," Hawaii International Conference on System Sciences, pp. 1-8, 42nd Hawaii International Conference on System Sciences, 2009.
[107] F. Milano, "Three-dimensional Visualization and Animation for Power Systems Analysis", Electric Power Systems Research, Vol. 79 Issue 12, p1638-1647, Dec. 2009
[108] S. Ingram, T. Munzner, M. Olano, "Glimmer: Multilevel MDS on the GPU", IEEE Transactions on Visualization and Computer Graphics, 15(2):249-261, Mar/Apr 2009.
[109] S.A. Shalom, M. Dash, M. Tue, "Efficient k-means clustering using accelerated graphics processors", In DaWaK '08: 10th international conference on Data Warehousing and Knowledge Discovery, 2008.
[110] D.S. Guo, "Mapping and Multivariate Visualization of Large Spatial Interaction Data", IEEE Transactions on Visualization and Computer Graphics, Vol. 15, Issue 6, pp 1041-1048, Nov.-Dec. 2009.
[111] J. Han, et al. "Stream cube: Architecture for multi-dimensional analysis of data streams", Distributed and Parallel Databases, 18(2):173–197, 2005
[112] C.B. Hurley, "Clustering Visualizations of Multidimensional Data", Journal of Computational and Graphical Statistics, Vol. 13, No. 4 pp. 788-806, 2006.
[113] V.G. Grishin, "Pictorial Analysis: A Multi-resolution Data Visualization Approach for Monitoring and Diagnosis of Complex Systems", Information Sciences, Vol. 152, pp. 1-24, Jun. 2003.

Discussant Narrative

Research Needs in Multi-Dimensional, Multi-Scale Modeling and Algorithms for Next Generation Electricity Grids

Roman Samulyak
Brookhaven National Laboratory

The planning and safe and optimal operation of electric power system is an intrinsically multi-dimensional and multi-scale problem. The paper by S. Grijalva proposes research directions enabling a better understanding and handling of power systems. The main message of the paper is that a multi-dimensional, multi-scale framework is required in order to address several emerging engineering challenges. The paper outlines 18 research needs in the area of:

- Modeling and analysis
- Optimization of dynamics and control
- Data Handling and visualization

Main discussion topics include:

- Development of methodology to identify, correlate and model relevant emerging behavior in spatial and temporal scales
- Multiscale forecasting models
- Self-similarity properties of electric networks
- Unified models and algorithms capable of dynamic expansion and collapse to multiple levels for network applications
- Wavelet analysis
- Reduction of the solution space
- Parameter identification of multiscale systems
- Optimization and decision making methodologies
- Reliability metrics
- Adaptive and multigrid methods
- Specialized parallel numerical algorithms for real time control
- Strong scalability on modern supercomputing architectures
- Multiscale control framework
- Scalable data structures and information architectures
- Integrated multiscale, multidimensional dynamic visualization tools

Recorder Summary

Research Needs in Multi-Dimensional, Multi-Scale Modeling and Algorithms for Next Generation Electricity Grids

Sven Leyffer
Argonne National Laboratory

Paper Summary

Modeling of the next-generation power grid will involve models that span multiple temporal, spatial, and scenario scales, motivating research needs in multi-scale modeling and algorithms for power grids. Spatial scales span a building/facility, microgrid, distribution, ISO control area, up to interconnection level. Temporal scales range from modeling transient effects (seconds), over real-time operation (minutes), and day-ahead operation, all the way to scheduling and maintenance (months) and planning and investment (years). In addition, models include uncertainty through high penetration of renewable energy such as wind or solar, and large scenario spaces as part of security constraints, or N-k contingency analysis.

Power grid models differ from traditional multi-scale phenomena in material science and chemistry, which are described by different physical models at different scales (such as DFT, MD, FEM). Instead, power grid models exhibit similar physical models at the different scales, which can be exploited in the development of more efficient algorithms. The unique multi-scale nature of next-generation power grid models motivates a range of research questions that include a library of models with a consistent description at the scales; efficient multi-scale methods in simulation, analysis, optimization, and control; and a proper treatment of uncertainties.

Discussant Summary

Existing power grid modeling and solution tools are insufficient to deal with the emerging challenges of smart grids. The coupling of scales and the introduction of multi-scale methods can introduce instabilities or non-convergence that requires new mathematical tools. The special network topology of grids requires specialized algorithms that exploit the network structure. Supercomputing and multi-core computing architectures together with hybrid programming models (e.g. MPI + multi-

threading) provide opportunities that can be exploited in the solution of next-generation grid models.

Discussion Report

The discussion focused on two areas: data-driven versus physics-based models, and extensions to the multi-scale models that were presented. These discussion points are summarized below.

Data-Driven versus Physics-Based Models

The computer science (CS) community has developed systems that capture data and build data-driven models that do not require physics-based models. Instead, control paradigms are learned from the data. Can this concept be used in power industry, e.g. by augmenting visualization with data analysis? Examples, where data-driven models have been successful include the detection of shockwaves in Euler flows using an equation-free, multi-scale approach that is based solely on particle flows.

Data acquisition is used in the power industry to obtain state estimators or learn load patterns. However, unlike some of the CS applications, the power industry has physical models that can be leveraged effectively. One concern is that there is typically only very little data, if things go wrong, and hence, data-driven models may be less useful in extreme situations. Another concern is that data only provides an estimate of the current state, and may have little predictive value, even though prediction based on time-series analysis might mitigate this effect.

One concern raised by the CS community was that one only trusts the model to the extent to which it has been verified by data. On the other hand, electrical engineers trust their physics-based models, since they have been shown to be accurate in a range of situations.

Extension to Multi-Scale Models

Many participants raised additional areas that could benefit from multi-scale analysis, including joint transmission management between RTOs, the role of weather forecast models, and the impact of electric vehicles (EVs). An important issue is to ensure that models are consistent across scales.

The coordination between RTOs (joint transmission management) can be improved by combing real-time dispatch with day-ahead market analysis, resulting in a multi-scale model. Increasing the coordination between RTOs, requires better real-time power-flow models, and the solution of day-ahead market models for multiple RTOs, which is a challenging problem. The long-term planning at PJM and MISO in terms of wind energy affects both ISOs. This requires a coarser than real-time operational model to understand

how the different RTOs are affected. This question has been addressed in collaboration with California Energy Commission, where each bus in the system is assigned a sensitivity metric for renewable energy needs, resulting in a large combinatorial problem. Using cutting-plane techniques can help makes this problem tractable.

The deployment of weather forecast models from NOAA suffers from a lack of resolution that is too high for the needs of wind or solar generators. This lack of resolution could be resolved by adding high-resolution data such as video feeds, or other measurements.

The widespread adoption of EVs provides novel research challenges to model charge and discharge. These models form a barrier to understanding how EVs impact the grid, and may require detailed simulations before EVs are widely deployed. A related issue is the question of which objective drives the development. There are not only techno-economical questions, but also questions that impact the automotive industry, resulting in more complex planning models. Some of these objectives are not well defined, like smart-grid and energy independence, which are not synonymous. Some models may also require the inclusion of atmospheric climate models.

Other Discussion Points

- DOE's SciDAC program has provided many useful concepts and tools that can be adopted by the power industry to address the emerging high-performance computing requirements.
- One challenge in implementing distributed controls is the fact that the quantity that is controlled (electricity) moves at the same speed as the wireless control instructions. This challenge may require more autonomous control with less communication.
- The multi-scale system may not be a single physics system. For example, there is no "nice" transition between a full-blown 60Hz model and the phasor domain. Switching inverters raise similar issue, and require modeling of powerful devices and controls.
- The assumption that the system and architecture are hierarchical in spatial and temporal terms may not hold, if we consider a more distributed system. To re-solve this issue, we may need to look at the industry in a network way, or apply a game theoretic approach, where the "prosumer" becomes a Nash player at the same level as the utility.
- Contingency planning remains a key challenge with emphasis on scenario reduction to make this problem tractable.

Overall, we need to be smarter about what we model, and how we overlap models, and transfer data between models. This challenge may involve replacing legacy (Fortran) code, and building experimental facilities to test data and models.

White Paper

Long Term Resource Planning for Electric Power Systems Under Uncertainty

Sarah M. Ryan and James D. McCalley
Iowa State University
David Woodruff
University of California Davis

Abstract

Electric power systems are subject to uncertainties of many types and at many levels. In addition to the uncertain fuel prices, demand growth, and equipment outages included in traditional models, current trends portend increasing uncertainty due to such factors as the growth in generation from renewable sources which may be intermittent, the possibility of regulations to control carbon emissions, and the increasing price responsiveness of demand encouraged by smart grid technologies. This paper begins by discussing issues that confront electric system planners and explains the requirements for useful tools. We continue with a description of academic and commercial planning models for the energy grid and conclude with a description of computational tools to support optimization for large scale planning models in the presence of uncertainty. For illustration, we include a proposed model for electric system planning that includes linkages with transportation systems.

1 Issues and Attributes for Electric System Investment Planning

Existing planning tools, models, and procedures are inadequate for identifying how to economically provide a sustainable and resilient low-carbon supply of energy. New planning capabilities are required to account for emerging issues and attributes of energy systems. The following is a summary of planning concerns adapted from PSERC (2009) and Budhraja et al. (2009), using some categorization of van Beeck (1999).

1.1 Model Scope

The scope of planning models must be enlarged in several dimensions:

1. *Multi-sector planning*: Growth in the coupling between different industry sectors, e.g., electric and associated fuel systems (natural gas, coal, uranium) and electrified transportation systems, requires planning processes, procedures, and tools that accommodate this broader view of the energy system. This is different from single-industry planning that has been the state of the art for many years now.

2. *Geographical scope*: Because of the variability in availability of different energy forms around the nation and need to move certain forms from one region to another, and because emissions must be treated nationally, planning models must span the nation. This is in contrast to the local and regional models that are used today. We observe that transmission infrastructure that is national in scope has been recently proposed (American Electric Power, 2010; EnerNex Corporation, 2010; Hansen, 2010; American Superconductor, 2009). This is motivated by the likely heavy growth in low-CO2 producing generation technologies, particularly wind, solar (both photovoltaic and concentrated solar power), geothermal, biomass, nuclear, clean-coal, and natural gas combined cycle. With the exception of nuclear, the cost of energy production from each of these technologies depends on the local richness or availability of the resource, and thus on geographical location. A key objective, then, is to identify at a national level whether development of each region's most attractive low-CO2 producing technology is more economic, sustainable, and resilient than use of long-distance transmission to bring in other regions' low-CO2 produced energy. Achieving this objective requires analytic tools capable of accounting for geographical variability of each resource's richness in assessing the cost of energy from each technology. It also requires the ability to simultaneously search the space of both generation and transmission options to identify optimal generation and transmission designs. We envision that such work would produce several national transmission overlay designs, each attractive in terms of the following criteria:

    - Facilitate consumption of energy from low-CO2 producing energy resources
    - Move energy to load centers
    - Lower cost of energy seen by consumers and reduce aggregate national energy costs
    - Avoid "pockets" of high energy costs
    - Have minimal environmental impact
    - Be resilient to large-scale disruptions
    - Have flexibility to adapt to future infrastructure changes

3. *Planning horizon*: Infrastructure elements; e.g., gas wells, pipelines, power plants, transmission, and transportation infrastructure/fleets, may have economic

lifetimes exceeding 50 years. In addition, the effect of emissions is cumulative over time. As a result, a 40 year planning horizon is more appropriate than the 10-20 year planning horizon typical of industry today.

Transmission projects may be justified based on their immediate benefits in connecting planned generation resources, but they also open up possibilities for additional generation investments to become attractive. The combination of a sufficiently long planning horizon and a wide geographic scope allow recognition of such dynamic impacts.

1.2 Breadth of Consideration

New developments have expanded the range of considerations that should be included in planning models, such as:

1. *Fuel diversity*: Transmission investments may accrue strategic benefits by allowing renewable generation to displace both baseload and marginal fossil generation. In particular, Budhraja et al. (2009) point out that, by accounting for the price-elasticity of natural gas demand, a wider impact of transmission for renewable energy can be quantified. That is, in addition to directly reducing consumption of natural gas, the fossil fuel displacement results in lower natural gas prices and corresponding savings across the electricity system.

2. *Technology selection*: Future technology change will come in the form of technologies expected to mature in the near or distant future, and technologies not yet known. Of these, the planning function should have embedded intelligence to determine which technology to implement and which to forego, accounting for uncertainties in technological maturation. This applies not only to generation technologies (e.g., wind, solar, natural gas combined cycle, integrated gasification, combined heat/power, geothermal, small hydro, nuclear, fuel cells, and storage), but also to transmission (HVDC or HVAC, voltage level, tower design and conductor type; e.g., high-temperature low-sag conductors, high surge-impedance loading conductors, gas-insulated lines, or high-temperature superconductors). It will be essential to develop and encode expertise and understanding necessary to integrate HVDC and EHVAC technologies within transmission design (Fleeman et al., 2009).

3. *Multiple objectives*: It is essential that decision-makers understand, value, and plan for tradeoffs among (1) sustainability, (2) costs (investment and operational), (3) long-term system resiliency, and (4) solution robustness (both solution variance to high-frequency uncertainties and solution performance under various futures associated with low-frequency uncertainties). Doing so requires multi-objective optimization, which implies that the work will not result in a single identified "best" solution but rather in a number of "good" solutions, typically called a

Pareto-optimal frontier; i.e., a set of solutions, each of which is characterized by the property that no change may be made to it that will improve one objective without degrading at least one other objective. Selection of the one solution to actually implement is a socio-political process that the multiobjective model seeks to inform. Methods for measuring some objectives are issues in their own right:

(a) *Measuring long-term resiliency*: To optimize on long-term resiliency and solution robustness, new tools must utilize measures to properly gauge the "strength" of 40 year plans that are national in scope. We view that traditional measures of reliability, which gauge energy unavailability following contingencies, are insufficient. New measures are needed. For example, one approach is to compute the monthly post-disturbance time-integral of nodal electricity prices for certain kinds of extreme events. Such events might include, for example, loss of one or more of the following: US Gulf natural gas supply, Powder River Basin coal, Middle Eastern oil, US uranium supply, Canadian gas, Mississippi barge traffic, US hydro, half of US wind resources. Figure 1 illustrates such a calculation for one Northeastern US electric node following the Katrina hurricanes (Gil, 2007; Gil and McCalley, forthcoming). Systems having low price variability for extreme events are said to be resilient.

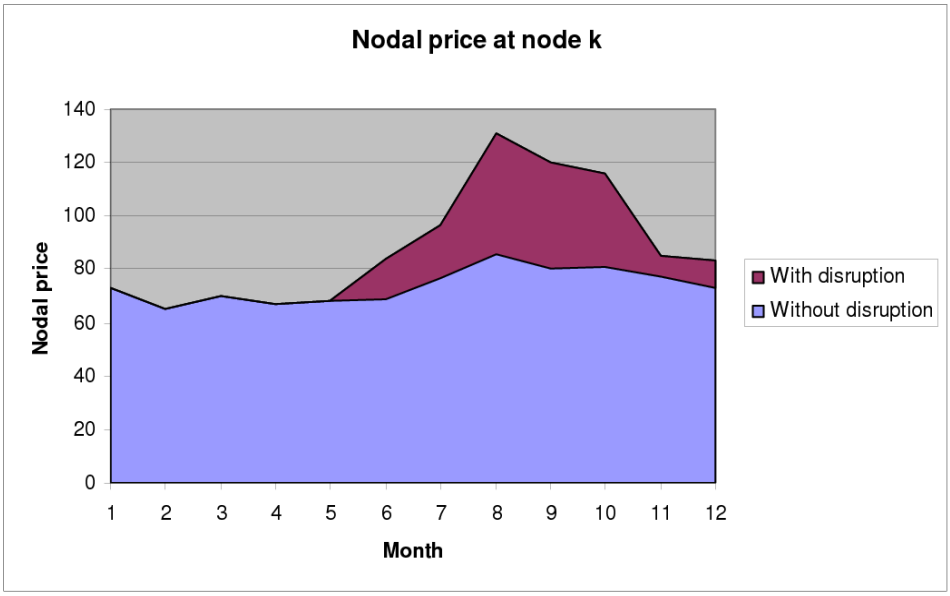


Figure 1: Illustration of price variability for Northeastern US electric node following Hurricane Katrina.

(b) *Measuring long-term sustainability*: As with resiliency metrics, optimizing on sustainability requires identifying appropriate metrics to characterize it. To this end, we define sustainability as the ability to sustain a resource or technology. Sustainability metrics include reserves to production ratios of fossil resources such as petroleum, coal, and natural gas, and also of raw

materials needed to produce certain technologies such as concrete, steel, and lithium. These metrics also include reductions in carbon emissions and other environmental impacts such as land and water usage, and nuclear waste.

4. *Value-based planning*: Traditional infrastructure planning was driven by the need to acquire sufficient generation and transmission capacity at least cost to satisfy peak load conditions; such a planning effort was initiated when violation of reliability criteria was impending. Although there is still need for this type of planning today, much of the industry is also embracing value-based planning where the objective is to identify opportunities that will lower the overall cost of energy. The Midwest Independent System Operator has led efforts to implement value-based planning in regional and national planning exercises (Hecker et al., 2009). Elements of value-based planning, including analysis of all loading conditions rather than just peak, and investigation of alternative infrastructure designs motivated by long-term net-present value of operational and investment costs, should be built into new investment planning tools.

1.3 Depth of Detail

The size and complexity of a model is greatly influenced by the level of detail with which it accounts for decision impacts.

1. *Operational issues*: Long-term planning tools should account for operational impacts that new technologies will have, such as MW variability in certain forms of renewable energy and the addition of new transportation load (e.g., electric vehicles, urban light rail, and railways), as such impacts may impose needs for additional investment in energy storage, demand response, or additional reserve. To this point, a recent study on wind integration (American Electric Power, 2010) concluded that "The within-the-hour impacts of varying load and wind generation are accounted for, at least approximately, in the production simulations by setting constraints on the unit-commitment and economic-dispatch algorithms. In each hour, specified amounts of reserves must be set aside and not used to serve load." Another operational issue concerns the electricity markets, their two-settlement structure, and the congestion management that ensues from the resulting locational marginal prices.

2. *Modeling granularity*: We require modeling granularity sufficient to identify electric system infrastructures that comprise solutions to the next 40 years' energy and emissions problems at the national level. Such a requirement may be specified in terms of element capacity for each type of infrastructure. For example, one may stipulate inclusion of all generation facilities having capacity in excess of 50 MW and all transmission facilities having voltage rating equal to or above 161 kV. Other granularity levels may be mode-appropriate, an observation that motivates study and analysis of this issue by, for example, choosing a granularity level,

performing the optimization identifying solutions, modifying the granularity level, and repeating to see if the solution changes significantly. Similar comments apply to temporal granularity; i.e., the choice of time step for each sector.

3. *Model fidelity*: Whereas granularity refers to the level of the system that is modeled, fidelity refers to the extent to which the actual physics and operational characteristics of the system are captured. For example, the electric grid may be modeled as a "flow" network, a "DC" network, or an "AC" network. A flow model respects capacity and nodal balance, but not Kirchhoff's voltage law (KVL), by assuming any desired electric flow may be achieved on a particular line independent of the impedance of that line. A DC model respects capacity, nodal balance, and KVL (under simplifying assumptions of zero resistance, unity voltages, and small angular differences across lines), but gives no information about reactive power and voltage magnitudes. An AC model respects capacity, nodal balance, and KVL (without the simplifying assumptions), and provides reactive/voltage magnitude information. These three models have computational requirements increasing with fidelity according to the order described: flow, DC, and AC. In addition, standard transmission investment planning optimization formulations become nonlinear if impedance is modeled, a situation that occurs even in linearized DC models.

1.4 Model Type

The type of a model describes its fundamental structure and goal.

1. *Optimization vs. equilibrium*: All previous discussion assumes an optimization model is desired, an assumption which needs clarification. Whereas optimization models result in a decision on how to accomplish a specific objective subject to specific constraints, equilibrium models result in an identification (and possibly evaluation) of the equilibrium state corresponding to a set of inputs sufficient to uniquely specify that state. A classic example of each of these, familiar to all power engineers, is the optimal power flow (an optimization model), vs. the power flow (an equilibrium model). An optimization model is preferred for the national long-term planning problem because optimization serves to drive policy, as opposed to equilibrium models which are policy-driven.

2. *Centralized vs. decentralized decisions*: Assuming adequate granularity and fidelity, results from an optimization model relate well to what actually happens if there exists a single decision-maker who has access to all of the model's required input data and complete authority to implement the outcome of the model. This is not the case for long-term, national planning, where the result is played out over decades among a multitude of individual and organizational elements. Thus, whereas the model represents a centralized planner, the reality is that decision-making is highly distributed, and one can rightfully argue that the optimization

model provides results that will never be realized due to the fact that distributed decisions must necessarily provide a sub-optimal outcome. Yet, this does not mean the result of the centralized model cannot be used. In fact, by identifying the best possible outcomes, the optimization model provides targets. Once these targets are identified, one then works with the realities so as to find solutions that approach one of them. On the other hand, game theoretic or multi-level models can help planners understand how distributed decision-makers may react to their plans – for example, how competing generating companies may choose investments in response to an expansion in the transmission network.

3. *Policy design*: There has been much activity in the US recently, at the state and federal levels, in regard to crafting production tax credits, investment tax credits, and renewable portfolio standards; future policies might also include emissions taxes or cap and trade mechanisms. Such devices are examples of policies intended to move society from a less desirable equilibrium to a more desirable one, according to some particular set of criteria. Once a desirable solution is identified (a task achieved by the investment planning application), then a policy design tool is needed to identify the most effective way to move from the existing equilibrium to a desired one. New analytic tools must be developed to accomplish this.

1.5 Treatment of Uncertainty and Risk

1. *Modeling uncertainties*: Developing investment plans for 40 year decision horizons requires extensive uncertainty modeling. Such modeling may be divided into two classes that have been labeled "random" and "non-random" (Buygi et al., 2004) or, as we term them in Section 3, high- and low-frequency. The former occur repeatedly and can be captured by fitting probability functions to historical data. Examples of these include uncertainties in fuel prices and component outages. The latter uncertainties do not occur repeatedly; therefore, their statistical behavior cannot be derived from historical data, but they may have great impact. Such uncertainties include, for example, dramatic shifts in weather and/or load forecasts, major policy changes, technological leaps, or extreme events (catastrophic contingencies). Furthermore, to handle uncertainty, solution robustness to changing conditions should be considered as an objective. See Section 3 for a detailed discussion.

2. *Risk mitigation*: Power system risks arise from insufficient diversification of supply, vulnerability to extreme events, and exposure to market dysfunction. Optimization models of sufficient scale and scope that properly account for uncertainty can identify portfolios of resources that both perform well on average and control the level of risk that results from volatility or extremes.

1.6 Computational Needs

New algorithmic and computational methods are needed to address (1) the high dimensionality of an optimization problem having a 40 year decision horizon, national geographical scope, multi-sector infrastructure representation, and high uncertainty; (2) a need to provide solutions in terms of tradeoffs among multiple objectives; and (3) the discrete nature of investment decisions.

In the following sections, we review existing planning models and discuss recent advances in techniques for optimization under uncertainty that could be exploited to incorporate more of the desired attributes in the planning tools.

2 Review of Long Term Resource Planning in Power Systems

This section summarizes research to date in models and tools to support long term resource planning in electric systems. Consistent with the traditional order in which resource types have been considered, we consider generation expansion first, followed by transmission expansion, and lastly some efforts to study the combination of both.

2.1 Generation Expansion Planning

The traditional approach to generation expansion planning by regulated utilities was to minimize the combination of investment and operational costs to meet demand. Optimization models accounted for uncertainty to varying degrees.

Bloom's method of using generalized Benders decomposition to incorporate probabilistic production cost simulation into long-term generation expansion planning was a significant breakthrough (Bloom, 1982, 1983). It constituted a major portion of the EGEAS software, which is still in use today. The decomposition divides the decision variables into two sets. The generating unit capacities are optimized in a master problem that minimizes the sum of investment and expected operating costs over the planning horizon subject to a constraint on expected unserved energy. For a candidate set of capacities, the expected operating costs and unserved energy are found by solving a production costing subproblem for each time period in the horizon. The probabilistic production costing simulation in each subproblem accounts for uncertain availability of generating units by constructing equivalent load duration curves for each unit in a fixed loading order derived from their variable operating costs. A unit's forced outage rate is used to represent the probability that it would not be available in a given time period, in which case a more expensive unit would be called upon to meet a portion of the demand equal to the cheaper unit's capacity.

Uncertainty in the load was considered by Murphy et al. (1982) in a two-stage stochastic program, assuming perfectly reliable equipment. The first-stage variables represented

capacities of various types of generating units while the second-stage variables represented allocations of those capacities to segments of the load duration curves. Load uncertainty was represented by multiple scenarios for the load duration curve, with a probability assigned to each. The overall objective was to minimize expected total annualized cost, including investment, operating and unserved energy. Their unusual result was that the optimal solution for the stochastic program could by found by solving a deterministic problem with a "average" load duration curve, which was constructed by carefully combining the scenario curves. This model was extended by Malcolm and Zenios (1994) to incorporate a notion of robustness. They augmented the expected cost in the objective function with a parameterized function of the cost variance along with a penalty for violating constraints in some scenarios. An expansion plan was termed "solution robust" if it was nearly optimal for every load scenario and "model robust" if it had nearly no excess capacity in every scenario. They explored tradeoffs between these two types of robustness by varying the weights of cost variance and infeasibility in the objective function. Because the cost variance was driven primarily by capacity shortages, the tradeoffs were essentially between risks associated with having too much or too little installed capacity to meet the realized demand.

Dynamic programming is another alternative for optimizing sequences of decisions over an uncertain future. Borison et al. (1984) used it in the context of a decomposition based on both time period and outcome scenario. As in the stochastic programming formulations, technology types and installation dates were determined in a master problem and the subproblems evaluated the operating costs. This approach allowed them to include uncertain fuel prices and demands, with the capability to also account for uncertain available capacity. Mo et al. (1991) used stochastic dynamic programming to incorporate discrete expansion sizes, correlations and autocorrelations among uncertain variables including fuel prices and demand, and economies of scale in investment costs. Due to the discrepancy in time scales, the overall planning problem was decomposed into investment and operational pieces.

As the restructuring of the electric utility industry began overseas and appeared on the horizon in the US, Hobbs (1995) reviewed the use of optimization models for use by utilities in their resource planning, with a particular emphasis on including demand side management as well as new generation resources. Anticipating increasing influences of competition and environmental regulation, he discussed additional uncertainties that should be included. He distinguished between "short-term" uncertainties such as generator availability and loads, which have been incorporated into optimization models as described above, and "long-run" uncertainties, such as government regulations, environmental concerns, and economic conditions. The latter are more commonly considered outside of an optimization model as "futures" and decision makers are hesitant to assign them probabilities. The anticipated growth of competitive markets suggested the need for models that account for demand price sensitivity, competition among generators, and the impact of transmission constraints.

With the advent of restructured markets, the focus has shifted from cost minimization to profit maximization and from centralized to decentralized decision-making. Game theoretic and bilevel optimization models have been developed in the past decade to examine the expansion plans that different profit-maximizing generating companies might concoct. The limited number of powerful producers in any wholesale electricity market has motivated the application of Cournot oligopoly theory. Chuang et al. (2001) discussed the applicability of the Cournot model and formulated a game among different technology options such as nuclear, coal, or gas turbine. Demand was assumed known in the form of an LDC and unit outages were taken into acount when determining each player's expected revenue from energy and real-time markets based on their capacity expansion and reserve decisions. Murphy and Smeers (2005) compared an open-loop Cournot model, to represent simultaneous expansion and long-term contractual sales, with a closed-loop Cournot model to represent expansion decisions followed by energy sales in a spot market. Their game was between two players, a baseload generator and a peaker generator, facing a known LDC with perfect reliability and no transmission constraints. Even in this simple model, an equilibrium does not necessarily exist in the closed-loop model; however, if it does exist, its electricity prices and quantities are closer to those derived from perfect competition than are the prices and quantities found in the open-loop model. Other two-tier game models have been used to model generator investments in capacity followed by participation in energy and/or reserve markets. Wang et al. (2009) used a co-evolutionary algorithm to search for a Nash equilibrium when each generator had a probability distribution for its competitors' expansion decisions. Nanduri et al. (2009) used reinforcement learning to solve a supply function game in the second tier and used conditional value at risk (CVaR, see Section 3.1) to assess the risk of capacity investments.

Stochastic dynamic programming has also been applied to compare centralized with decentralized decision-making for investments in generating capacity (Botterud et al., 2005). Short-term uncertainties in demand, such as due to weather patterns, were used to evaluate expected operating costs in each year, while long-term uncertainties, such as from economic development, governed the transitions among states in different years. Centralized decision-making was represented by an objective to maximize social welfare derived from linear demand and supply curves. Decentralized decision-making was modeled by considering only generator profits in the objective. A simulation of the decisions found by each model suggested that profit-maximizing generators would under-invest in baseload capacity compared to the social welfare solution, but that increasing price sensitivity of demand would reduce the differences between results of the centralized and decentralized models.

2.2 Transmission Expansion Planning

Models for transmission expansion planning have followed a similar history as those for generation expansion planning, though the problem has received less attention outside the power engineering community. Constraints imposed by the physical characteristics

of power transmisssion complicate the problem significantly. First, transmission capacity is inherently discontinuous: either there is a line connecting two buses or there is not; and expansions along a given right-of-way must also be in discrete increments of additional parallel lines or upgrades to higher voltage levels. While generation expansions also are discrete, they are more amenable to approximation with continuous variables than transmission expansions are. Second, for a given grid topology, the power flows are governed by Kirchhoff's current and voltage laws, which dictate how power travels from points where it is injected to points where it is withdrawn. As explained in Section 1.3, the full AC power flow equations are nonlinear. These can be approximated by linearized DC equations but, when decision variables concerning grid expansion are included, the resulting set of constraints imposed by the line flow capacities are rendered nonlinear.

Much effort has been devoted to solving a static version of transmission expansion planning for the regulated environment. Here, the objective is to minimize the investment cost required to satisfy demand using the available generating capacity. As Choi et al. (2005) noted, "It is difficult to obtain the optimal [expansion] solution of a composite power system considering the generators and transmission lines simultaneously in an actual system, and therefore, transmission system expansion is usually performed after generation expansion planning" (p. 1606). In the static version, the planner determines where and how to expand capacity to minimize the sum of investment cost and the cost of unsatisfied demand in a typical load scenario. The problem naturally decomposes into an investment master problem and operational subproblems, and many authors have applied Benders decomposition to solve variations of it. Romero and Monticelli (1994) devised a hierarchical approach to mitigate the risk of finding only a local optimal solution by initially solving a continuous linear approximation and including more nonlinear constraints and finally discrete variables in later stages, as the procedure presumably approached the global optimum. Latorre-Bayona and Pérez-Arriaga (1994) included reliability submodels as well as the operational ones to evaluate unserved power in different scenarios of equipment availability. Binato et al. (2001) replaced the nonlinear constraints with disjunctive ones that employ large constants ("big M"), found minimal values of the disjunctive constants to reduce numerical difficulties, and also employed traditional integer programming cutting planes in the master (investment) problem. Latorre et al. (2003) surveyed the work done to that point and highlighted the computational difficulties caused by nonlinearity and nonconvexity due to DC load flow constraints, the discrete character of investments, and stochastic modeling.

In the same cost-minimization vein, more recent research has included security constraints, probabilistic reliability criteria, and dynamic aspects of the problem. The traditional "N–1" security standard is that the network is able to meet the load even if any single one of the N major elements (e.g., generating unit or transmission line) becomes unavailable. de Silva et al. (2006) added a set of DC power flow constraints for each of the N contingencies representing the loss of an element and devised a genetic

algorithm to solve it. This is a deterministic approach to planning for uncertain events. In contrast, Choi et al. (2005) assigned probabilities to outages of system elements and compared solutions found according to various probabilistic reliability constraints. They measured reliability in terms of loss of load expectation (LOLE, expected number of hours per year in which demand is not satisfied) and expected energy not served (EENS, expected number of unmet MWh per year) at either individual buses or in the transmission system as a whole. The expectations were taken by constructing effective load duration curves that accounted for outages of generators, transformers and transmission lines, with a simplified transportation model for energy flows, and the problem was solved by a branch and bound method. Escobar et al. (2004) used a genetic algorithm to solve a dynamic expansion planning mixed integer nonlinear program to determine the timing of investments over a multi-year horizon under DC power flow constraints.

The uncertainty of load and generation patterns increases in the competitive market environment. de la Torre et al. (1999) conducted an early examination of the effects of uncertainty in generation plans due to decisions by independent power producers in a case study of the Central America region. Adopting a decision analysis approach, they sought a "robust" transmission expansion plan, which would incur no regret over different possible futures representing rates of load growth and the addition of generation resources. Fang and Hill (2003) also measured risk in terms of regret associated with the cost of expansion plans to meet possible future power flow patterns motivated by market considerations. They found an optimal expansion plan for each future using a model similar to that of Romero and Monticelli (1994) and then selected the plan with the minimum maximum regret over all futures. Regrets could also be weighted by future probabilities estimated from market simulation. Game theory has been applied to discover coalitions of agents such as generators, loads, or independent third parties, who would have the greatest cost allocation incentives to cooperate on constructing transmission resources (Contreras and Wu, 1999) and their actions have been studied in multi-agent simulations (Contreras and Wu, 2000).

2.3 Combined Generation and Transmission Expansion Planning

To plan resources for larger geographical areas, expansions in generation and transmission should be considered together. In the late 1980s, the Electric Power Research Institute funded a study of a computationally tractable method to do so, using the West Coast Power System as a model. Dantzig et al. (1989) decomposed the large-scale optimal resource planning problem into a deterministic facility expansion plan covering multiple years and its operation under uncertainties concerning loss of transmission or generation. By considering expansion sizes as continuous quantities, they formulated the problem as a two-stage stochastic linear program. To handle reliability constraints that prevented immediate decomposition by scenarios, they used two levels of decomposition – first, a primal Dantzig-Wolfe decomposition that consisted of a reliability master problem with resource planning as subproblems and second, a

Benders decomposition of each resource planning subproblem into separate operational subproblems under each scenario for the uncertain quantities. They used importance sampling to reduce the number of contingencies to consider and suggested parallel processing of the scenario subproblems.

Baughman et al. (1995a); Baughman et al. (1995b) extended the approach of Dantzig et al. (1989) to handle discrete expansion decisions in the form of either candidate generation and transmission facilities or demand side management initiatives. They omitted the reliability constraint but included customer outage costs in their objective to minimize expected discounted costs of investment and operation. The inclusion of reactive as well as real power constraints rendered the problem nonlinear and nonconvex as well as combinatorial, but they reported success on a 24-bus case study in applying Benders decomposition with a heuristic version of importance sampling to solve it. Siddiqi and Baughman (1995) investigated the effect of including different levels of approximation of the AC power flow constraints in the same model.

Roh et al. (2007) proposed a model to simulate the interactions among generating companies (gencos), transmission owners (transcos) and the independent system operator (ISO) in a market environment. Under the assumption that generation and transmission investors would not anticipate each others' decisions, they devised an iterative scheme to simulate capacity planning decisions by the generation and transmission facility owners, the ISO's check of transmission security, and the ISO's operational optimization problem to minimize the cost (as-bid) of meeting fixed loads. Solving the ISO's latter problem produced locational marginal prices (LMPs), which served as signals to the gencos for where to expand generation capacity and "flowgate marginal prices" (FMPs), which would analogously signal transcos where to expand transmission capacity. Roh et al. (2009) included random outages of generating units and transmission lines along with errors in long-term load forecasting in the same model. They used Monte Carlo simulation to create scenarios of outages and random load components and applied the scenario reduction technique of Dupačová et al. (2003) to reduce the computational burden.

Sauma and Oren (2007) formulated and solved a three period model in which the transmission planner, acting as Stackelberg leader, chooses one from among a set of candidate transmission lines to build or expand. In the second period, generating companies simultaneously invest in new capacity to maximize expected profit while anticipating market equilibrium outcomes in the third period. The market consists of an operator that solves a redispatch problem to maximize social welfare while generators simultaneously choose their production quantities to maximize profits. Expectations are taken over contingencies that represent uncertain demand or resource unavailability. They used a small example and a case study to illustrate how different planning objectives of maximizing expected social welfare or its components or minimizing local market power would result in different transmission expansion decisions.

2.4 Existing Tools

In this section, we describe types of commercially available tools that address the needs of long-term infrastructure planning for planning of both electric systems and multi-sector systems.

2.4.1 Electric System

There are three main types of planning tools for electric infrastructure: reliability, production costing, and resource optimization. Reliability tools do not identify solutions but just evaluate them. Both deterministic and probabilistic tools exist and are heavily used in the planning process. Deterministic tools include power flow, stability, and short-circuit programs, providing yes/no answers for specified conditions. Probabilistic tools compute indices such as loss-of-load probability, loss of load expectation, or expected unserved energy, associated with a particular investment plan. A representative list of commercial-grade reliability evaluation models include CRUSE (Lu et al., 2005), MARS (General Electric, 2010), TPLAN (Siemens, 2006), and TRELSS (Tran et al., 2006).

Production cost programs have become the workhorse of long-term planning. These programs perform chronological optimizations, often hour-by-hour, of the electric system operation, where the optimization simulates the electricity markets, providing an annual cost of producing energy. Although production cost models make use of optimization, it is for performing dispatch, and not for selection of infrastructure investments. Therefore, production cost models are equilibrium/evaluation models with respect to an investment plan. A representative list of commercial grade production cost models include GenTrader (Powercosts, Inc., 2010), MAPS (Bastian et al., 1999), GTMax (Koritanov and Veselka, 2003), ProMod (Ventyx, 2010a), and ProSym (Ventyx, 2010b), and Plexos (Energy Exemplar, 2010). Production cost programs usually incorporate one or more reliability evaluation methods.

Resource optimization models select a minimum cost set of generation investments from a range of technologies and sizes to satisfy constraints on load, reserve, environmental concerns, and reliability levels; they usually incorporate a simplified production cost evaluation, which includes a reliability evaluation. These optimization models identify the best generation investment subject to the constraints. However, at this point in time, these models generally do not represent transmission, or they represent it but do not consider transmission investments. A representative list of resource optimization models includes EGEAS (Head and Nguyen, 1990), Plexos (Energy Exemplar, 2010), Strategist (Yamin et al., 2001), and WASP-IV (Adica, 2010). The US Environmental Protection Agency Integrated Planning Model (IPM) is another resource optimization model (US Environmental Protection Agency, 2010); a linear programming model of the electric sector used to evaluate the projected impact of environmental policies at a national level. It provides forecasts of least-cost capacity expansion, electricity dispatch, and emission

control strategies for meeting energy demand and environmental, transmission, dispatch, and reliability constraints. IPM can be used to evaluate the cost and emissions impacts of proposed policies to limit emissions of sulfur dioxide (SO2), nitrogen oxides (NOx), carbon dioxide (CO2), and mercury (Hg) from the electric power sector. The Regional Energy Deployment System (ReEDS) model (Short et al., 2009) is a multiregional, multiperiod linear programming model of capacity expansion in the electric sector of the United States. Developed by the National Renewable Energy Laboratory, the model performs capacity expansion but with detailed treatment of the full potential of conventional and renewable electricity generating technologies as well as electricity storage. The principal issues addressed include access to and cost of transmission, access to and quality of renewable resources, the variability of wind and solar power, and the influence of variability on the reliability of the grid.

2.4.2 Multisector

The term "multisector" has been used to refer to the ability of a model to address more than just the energy sector of the economy; very few models are indeed multisector in this sense (van Beeck, 1999). There are three exceptions, all of which are intended to perform analysis at the national level. The first two have existed for several years and include the National Energy Modeling System (NEMS) (Energy Information Administration, 2003) and the MARKAL/TIMES suite (Energy Technology Network, 2010). Of these, NEMS is an equilibrium model and although effective in evaluating policy, it is ineffective in identifying policy. MARKAL/TIMES, an optimization model, comes closest to fulfilling the desired modeling attributes set forth in Section 1. However, it does not admit representation of energy transportation (e.g., transmission, natural gas pipeline, liquid fuel pipeline) or commodity/passenger transportation (rail, highway, air). The third multi-sector tool, NETPLAN, has been developed more recently by researchers at Iowa State University and is described in the next section.

2.4.3 NETPLAN – A new multi-sector investment planning tool

NETPLAN is a national multiobjective investment modeling framework to study long-term planning for the energy and transportation systems in an integrated fashion. We provide a brief overview of this model here but include an analytic description in Appendix A; additional information may be found in Ibanez et al. (2008); Ibanez et al. (2010); Ibanez and McCalley (2010). We consider the US energy system in terms of fuel production and transportation, electric generation and transmission, and vehicular, rail, and airline transportation systems for commodities and passengers. In 2008, the total US energy use was 100 Quads (1 Quad=1.015 BTUs), and of this, approximately 69% was consumed electrically or for transportation purposes (Riehmann et al., 2005). An annual publication from the US Department of Energy Information Administration on emissions (Energy Information Administration, 2008) indicates that 74% of all greenhouse gas emissions in the US is from the electric and transportation sectors. These facts suggest

that developing an approach to meeting the nation's energy needs with low-cost energy while reducing emissions must focus on both the electric and transportation sectors. Today, in the US, these two sectors utilize very different fuel sources, with 96% of electric resources obtained via coal, natural gas, nuclear, and hydro, while 95% of transportation resources are petroleum based. The only major interdependency between them is rail, used to transport many different commodities, among which is coal. If electric and transportation sectors remain lightly coupled, then they represent separate problems. However, it is likely that reducing greenhouse gas emissions must involve electrification of the transportation sector, while "greening" electric resources. Indeed, the growth of plug-in electric vehicles on the transportation side and wind energy on the electric side are unambiguous indications that the US is moving in this direction already, and there are expectations that this movement will require transformational changes to the electric infrastructure over the next 40 years.

Energy Modeling: The energy system is captured by representing the energy cycle from its multiple origins (e.g., coal mines, natural gas wells, wind, sun) to its final destination, accounting for possible conversion (power plants, wind farms, refineries) and transportation (natural gas pipelines, coal transport, high-voltage lines). The result is a generalized capacitated transportation model in which a single commodity, energy, is being delivered. The arcs that form the network are assigned fixed parameters. Arcs resulting from source nodes are assigned maximum extraction rate (MBtu/month) and extraction cost ($/MBtu). Conversion and transportation are endowed with: capacity (MBtu-capacity/month), efficiency (%), operational cost ($/MBtu-flow/month), investment cost ($/MBtu-capacity/month), emissions of "criteria" pollutants (Scorecard, 2010) (sulfur dioxide, nitrogen oxides, carbon monoxide, ozone, particulate matter, and lead) and carbon dioxide (tons/MBtu-flow/month), water use (gallons/MBtu-flow/month), and component reliability levels (percent of year component is expected to be forced out of service).

Transportation Modeling: There are three fundamental differences between the transportation formulation and the energy formulation. First, whereas the energy formulation must restrict energy flows of specific forms to particular networks (for example, natural gas or hydrogen cannot move through electric lines or liquid fuel lines), commodities may be transported over any of the transport modes (rail, barge, truck). Second, whereas energy movement requires only infrastructure (electric lines, liquid fuel pipelines, gas pipelines), commodity movement requires infrastructure (rail, locks/dams, roads, ports) and fleet (trains, barges, trucks), and there may be different kinds of fleets for each mode (e.g., diesel trains or electric trains). Third, the energy formulation (including coal transportation) is a nodal-demand model, implying that loading inputs are specified at the nodes, and the optimization identifies injections at the supply nodes together with flows on the links. In contrast, the transportation formulation is a link-demand model, implying that loading inputs are specified at the links, and therefore commodity movement routes do not change over time. The optimization identifies the particular fleet and energy form to use in transporting the commodities. We model

interstate passenger transportation in the same way as described for interstate commodity transportation. To accommodate these differences, the transportation is modeled using a multicommodity flow formulation (U.S. Department of Transportation, 1997; Bazaraa et al., 1992), with fleet and infrastructure capacity constraints treated as decision variables so that they may increase over time through investment to accommodate higher transportation demand as well as retirement of fleet and infrastructure. Although passenger and commodity flows are specified exogenously, coal flows are not, necessitating capacitation on transportation infrastructure. Flows are in units of tons of each major commodity. A commodity is considered major if its transportation requirements comprise at least 2% of the total freight ton-miles. Data available to make this determination indicates this criterion includes 23 commodities that comprise 90% of total ton-miles (e.g., the top eight, comprising 55%, are in descending order: coal, cereal grains, foodstuffs, gasoline and aviation fuel, chemicals, gravel, wood products, and base metals).

The described modeling of the energy and transportation systems enables simultaneous cost minimization of the integrated investment and operational problem for both infrastructure sectors, using a network flow linear program (LP). This leads to national investment portfolios for the electric system including transmission and generation (conventional and renewables), natural gas pipelines, liquid fuel treatment facilities nation's and pipelines, as well as investment on railroad, highways, ports, and fleet technologies such as diesel, electric, hybrid, and hydrogen powered vehicles.

Solution approach: Recognizing the need to identify investments that minimize cost while maximizing sustainability and resiliency, the formulation is multiobjective, using an evolutionary program based on the NSGA-II algorithm (Deb et al., 2002). Here, each individual of a population of problem instances is represented by a string of minimum investments to be made during the simulation span. That string is used to create lower bound constraints for the investment decision variables in the cost minimization engine. The fitness of each individual is computed as the net present value of investment and operational costs that result from the cost minimization LP, together with sustainability and resiliency evaluations which are performed following each LP. A parallelized computational approach is deployed so each cost minimization solution (which corresponds to a single individual within an NSGA-II iteration's population of solutions) is performed by a dedicated processor to yield cost, sustainability, and resiliency values for the individual. The process continues with the communication of the objective values to the search and selection algorithm in order to create the next generation of solutions. This process is illustrated in Figure 2.

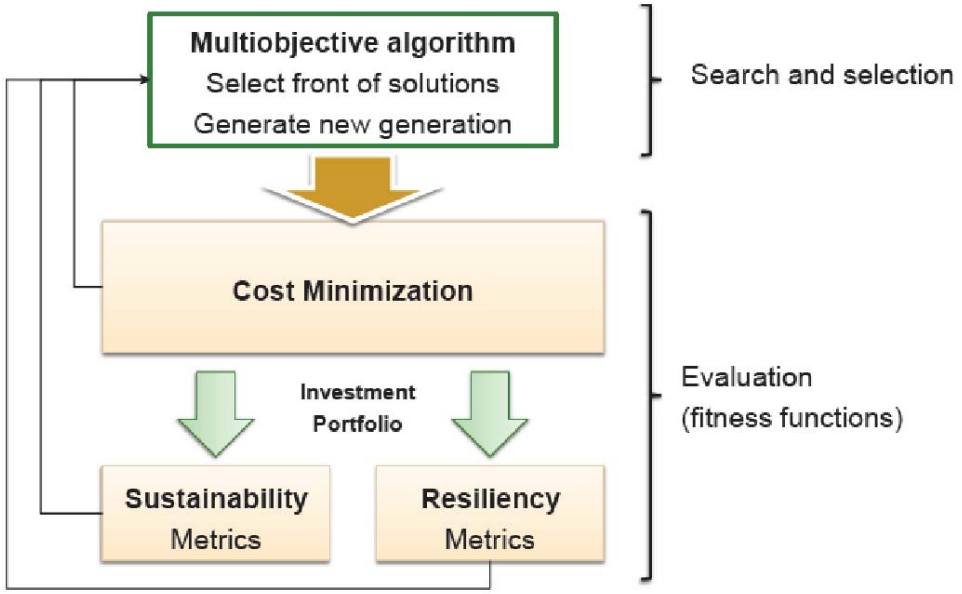


Figure 2: Multiobjective algorithm design.

3 State of the Art in Optimization Under Uncertainty for Planning

Uncertainty is widely recognized as an integral component in most real-world problems including the design or operation of energy systems. Uncertainty is associated with the problem input parameters; e.g., consumer demands, construction times, wind speeds, etc. Approaches to formulating optimization problems under uncertainty vary according to the time scale of decisions, the types of uncertainties, and the decision maker's attitude towards risk. If we distinguish between scheduling with a short time horizon and planning with a longer time horizon, we see that the modeling of uncertainty differs. In the short term, there are uncertain events that occur with high frequency and that can reasonably be modeled with probability distributions. For example, hourly demand for electricity in the next few days, given a weather forecast, fits this description. Another example is wind power availability in the short term. Profit maximizers who schedule in the short term can reasonably optimize expected value since there is no chance of ruin, and the expected values will be achieved due to large numbers. For those charged with public safety considerations (such as scheduling hospital generators), worst case analysis is often appropriate.

Planning is somewhat more complicated. One source of complication is that scheduling is often a sub-problem for the planning problem and so high frequency stochastics enter at this level. At the planning level, there are higher impact, lower frequency sources of uncertainty, such as major drifts in demand or supply, changes in technology, etc. Furthermore, for profit maximizing firms there is a risk of ruin and for planners charged with the public welfare, the risk of catastrophe must be considered. For some engineering design problems (e.g., sizing transmission towers), it makes sense to design

against the so-called worst case, but for most capacity design decisions that is not an option because the worst case is not well-defined. Since there is a risk of ruin or catastrophe it is not appropriate to consider expected values only. Furthermore, since these decisions and the stochastic realizations have low frequency, there is no reason to assume that expected values will be achieved.

When low-frequency uncertainties are important, decision approaches that do not involve probability may be applicable. Very conservative decision-makers may focus on the worst case and seek to minimize the damages that would result if this case occurred – this is the classical minimax approach. Minimax regret is a related approach based on hindsight, in which the decision maker optimizes over each different future, then computes the regret for a particular decision and future as the difference in objective function values under that future between the given decision and the optimal decision for that future. For each decision, the maximum regret over futures is computed, and the decision with the smallest value of maximum regret is chosen.

The words *robust optimization* are used in many ways, but the term as used by Ben-Tal and co-authors (see, e.g., Ben-Tal and Nemirovski (2002)) are of interest for grid planning. In this form of robust optimization – that we will call RO – the intersection of all constraints generated by an *uncertainty set* are considered. These sets can be constructed so as to facilitate efficient solution of the resulting optimization problem and they can be based on little or no knowledge of the probability distributions of uncertain parameters. The size of the uncertainty set controls the extent to which the resulting solution is likely to be feasible for realized values. For large uncertainty sets, RO requires almost sure feasibility to the extent that the modeler can indicate the boundaries of the parameter space. When the uncertainty set is less than the full worst-case set of possible events, this method results in a form of chance-constraints (Chen et al., 2007).

Some stochastic programming problems include so-called chance constraints (Prekopa, 2003; Watson et al., 2010b). Some, or all, constraints are required to hold only with a specified probability rather than for all realizable scenarios. This is particularly relevant for electric grid planning because it is one way to model service level requirements.

When comparing stochastic programming and RO for chance constrained stochastic optimization, each offers advantages and disadvantage. The main advantage of RO is that uncertainty sets can often be given without knowing much about the probability distributions of the parameters. This results in the main disadvantage: any sort of probabilistic statements about feasibility rely on strong (albeit succinct) assumptions about the probability distributions and these statements come only in the form of bounds. Stochastic programming enables full use of distributions or scenarios, but also requires them. In addition, probabilistic statements and confidence intervals can be given. When used for chance constraints, stochastic programming has the property that it "cherry picks" the portion of the distribution to ignore. That is, it simultaneously selects scenarios to ignore while setting variable values. Whether this is desirable

depends on the application. In contrast, with RO, the events that are ignored are determined without regard to their effect on the objective function. See Philpott (2010) for an overview of stochastic modeling approaches.

Consider linear constraints of the form

$$ax \leq b$$

where b is a scalar, a is a row vector of data and decision column vector is x, which has the same dimension as a. When given as a chance constraint, this becomes a constraint that holds with probability $1-\alpha$ where α is given as data:

$$Pr(ax \leq b) \geq 1-\alpha$$

Computationally effective stochastic programming approaches (Ruszczyński, 2002; Watson et al., 2010b) for chance constraints make use of the set of scenarios S with probability $Pr(s)$ for each $s \in S$ and introduce a binary scenario selection variable d_s so the constraint can be written as

$$ax \leq b \quad \forall s \in \{\mathcal{S} : d_s = 1\}$$

and

$$\sum_{s \in \mathcal{S}} Pr(s) d_s \geq (1-\alpha)$$

In contrast, RO requires introduction a compact set U_ω where ω is a parameter to control the size of U; then the constraint is written as

$$ax \leq b \quad \forall (a,b) \in \mathcal{U}_\omega$$

In the common case where probabilities can be attached to possible parameter values and the timing of the resolution of parameter uncertainty can be modeled as independent of the decisions, stochastic programming is an appropriate and widely studied mathematical framework to express and solve uncertain decision problems (Birge and Louveaux, 1997; Wallace and Ziemba, 2005; Kall and Mayer, 2005a; Shapiro et al., 2009). Many scholarly publications describe applications to power system planning, as highlighted in Section 2. We will focus here on methods collectively referred to as *stochastic programming* rather than those referred to as dynamic programming (Ross, 1983). Although in principle, either method could be used to solve stochastic optimization problems, they are usually associated with differing classes of problem formulations. Traditionally, dynamic programming has been applied to problems with perhaps a large number of time stages, but not very many decision variables. Conversely, stochastic programming is usually applied to problems with relatively few time stages but perhaps very many (millions) of decision variables. Recently,

Approximate Dynamic Programming (Powell, 2010) has been proposed as a way to use heuristic ideas from artificial intelligence applied to dynamic programming to address problems with a large number of decision variables.

Scenario-based planning provides an opportunity to simulate some of the sources of uncertainty and to purposely insert extreme, important cases. The use of multiple time stages enables the creation of plans that take into account the fact that there will be future recourse. The use of a risk measure, such as Conditional Value at Risk (CVaR), combined with the expected value gives the decision maker a chance to map the efficient frontier of planning decisions that balance risk and expected reward. Since CVaR is an expectation, it can be included in many software packages for stochastic programming.

3.1 Stochastic Programming: Definition and Notation

We concern ourselves with stochastic optimization problems where uncertain parameters (data) can be represented by a set of scenarios S, each of which specifies both (1) a full set of random variable realizations and (2) a corresponding probability of occurrence. The random variables in question specify the evolution of uncertain parameters over time. We index the scenario set by s and refer to the probability of occurrence of s (or, more accurately, a realization "near" scenario s) as $\Pr(s)$. Let the number of scenarios be given by $|S|$. Scenarios are frequently obtained via simulation, generated from statistics that are based on historical data, and/or formed from expert opinions. We assume that the decision process of interest consists of a sequence of discrete time stages, the set of which is denoted T. We index T by t, and denote the number of time stages by $|T|$.

Although we can model many types of non-linear constraints and objectives, we develop the notation primarily for the linear case in the interest of simplicity. For each scenario s and time stage t, $t \in \{1,\ldots,|T|\}$, we are given a row vector $c(s,t)$ of length $n(t)$, a $m(t) \times n(t)$ matrix $A(s,t)$, and a column vector $b(s,t)$ of length $m(t)$. Let $N(t)$ be the index set $\{1,\ldots,n(t)\}$ and $M(t)$ be the index set $\{1,\ldots,m(t)\}$. For notational convenience, let $A(s)$ denote $(A(s,1),\ldots,A(s,|T|))$ and let $b(s)$ denote $(b(s,1),\ldots,b(s,|T|))$.

The decision variables in a stochastic program consist of a set of $n(t)$ vectors $x(t)$; one vector for each scenario $s \in S$. Let $X(s)$ be $(x(s,1),\ldots,x(s,T))$. We will use X as shorthand for the entire solution system of x vectors, i.e., $X = x(1,1),\ldots,x(|S|,|T|)$.

If we were prescient enough to know which scenario $s \in S$ would be ultimately realized, our optimization objective would be to minimize

$$f_s(X(s)) \equiv \sum_{t \in T} \sum_{i \in N(t)} \left[c_i(s,t) x_i(s,t) \right] \quad (\mathrm{P}_s)$$

subject to the constraint

$$X \in \Omega s.$$

We use Ω_s as an abstract notation to express all constraints for scenario s, including requirements that some decision vector elements are discrete or more general requirements such as

$$A(s)X(s) \geq b(s).$$

The notation $A(s)X(s)$ is used to capture the usual sorts of single period and inter-period linking constraints that one typically finds in multi-stage mathematical programming formulations.

We must obtain solutions that do not require foreknowledge and that will be feasible, independently of which scenario is ultimately realized. We refer to solution systems that satisfy constraints for all scenarios as *admissible*. We refer to a system of solution vectors as *implementable* if for all pairs of scenario s and s' that are indistinguishable up to time t, $x_i(s,t') = x_i(s',t')$ for all $1 \leq t' \leq t$ and each i in each $N(t)$; these are also called non-anticipative. We refer to the set of all implementable solution systems as N_S for a given set of scenarios, S

To achieve admissible and implementable solutions, the expected value minimization problem then becomes:

$$\min \sum_{s \in S} [Pr(s) f(s; X(s))] \quad \text{(P)}$$

subject to

$$X \in \Omega s$$
$$X \in \mathcal{N}_S.$$

Only solutions that are implementable are useful since the future cannot be known in advance. Solutions that are not admissible, on the other hand, may have some value because while some constraints may represent laws of physics, others may be violated slightly without serious consequence.

Formulation (P) is known as a stochastic mathematical program. If all decision variables are continuous, we refer to the problem simply as a stochastic program. If some of the decision variables are discrete, we refer to the problem as a stochastic mixed-integer program.

In practice, the parameter uncertainty in stochastic programs is often encoded via a *scenario tree*, in which a node specifies the parameter values $b(s,t)$, $c(s,t)$, and $A(s,t)$ for all $t \in T$ and $s,s' \in S$ such that s and s' are indistinguishable up to time t. Scenario trees are discussed in more detail in Section 3.3.

For operational problems where there is no chance of extreme disasters, minimizing (or maximizing) the expected value makes sense because by the law of large numbers, the long run performance is dictated by the expected value. For strategic planning problems, some measure of risk should also be considered, perhaps in addition to the expected value. For some engineering design problems, it make sense to consider the worst-case scenario/

To capture these ideas in an abstract way, we introduce the notation $E[f(X)] \equiv \sum_{s \in S} [\Pr(s) f(s; X(s))]$ to represent the expected value and the notation $R[f(X)]$ to denote a risk measure. We can then write our problem as

$$\min E[f(X)] + \beta R[f(X)]$$

subject to

$$X \in \Omega s$$
$$X \in \mathcal{N}_S.$$

where β is a parameter that controls the relative importance of the two terms. Figure 3 illustrates various risk measures. Note that the Tail-Conditional Expectation is often referred to as the Conditional Value at Risk (CVaR). CVaR (Acerbi and Tasche, 2002; Rockafellar and Uryasev, 2002; Schultz and Tiedemann, 2006) is a *coherent* risk measure, which satisfies a set of axiomatic properties of risk measures given by Artzner et al. (1999).

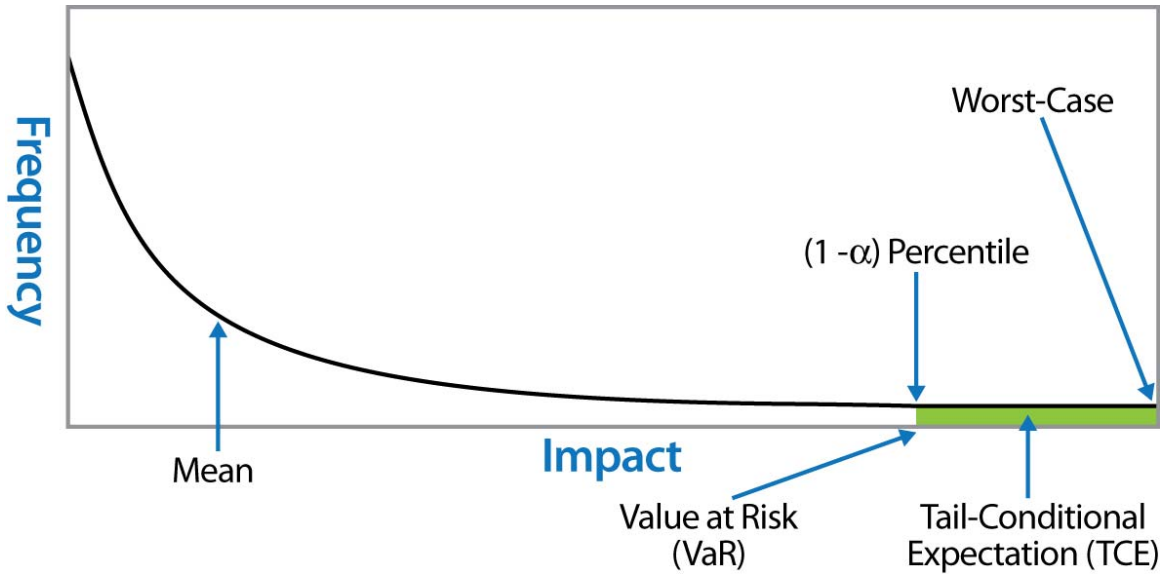


Figure 3: Various Risk Functions

3.2 *Proposals for Stochastic Programming Software Packages*

As noted in (Watson et al., 2010a), stochastic programming has not yet seen widespread, routine use in industrial applications – despite the significant benefit such techniques can confer over deterministic mathematical programming models. The growing practical importance of stochastic programming is underscored by the recent proposals for and additions of capabilities in many commercial algebraic modeling languages (AIMMS, ; Valente et al., 2009; Xpress-Mosel, 2010; Maximal Software, 2010).

Some commercial vendors have recently introduced modeling capabilities for stochastic programming, e.g., LINDO Systems (LINDO, 2010), FICO (XpressMP, 2010), and AIMMS (AIMMS,). On the open-source front, the options are even more limited. Part of COIN-OR, FLOPC++ (FLOPCPP, 2010) provides an algebraic modeling environment in C++ that allows for specification of stochastic linear programs. APLEpy provides similar functionality in a Python programming language environment.

The landscape for solvers (open-source or otherwise) devoted to generic stochastic programs is extremely sparse. Those modeling packages that do provide stochastic programming facilities with few exceptions rely on the translation of problem into the *extensive form* – a deterministic mathematical programming representation of the stochastic program in which all scenarios are explicitly represented and solved simultaneously. The extensive form can then be supplied as input to a standard (mixed-integer) linear solver. Unfortunately, direct solution of the extensive form is unrealistic in all but the simplest cases. Some combination of either the number of scenarios, the number of decision stages, or the presence of discrete decision variables typically leads to extensive forms that are either too difficult to solve or exhaust available system memory. Iterative decomposition strategies such as the L-shaped method, a version of Benders decomposition (Slyke and Wets, 1969), or Progressive Hedging (Rockafellar and Wets, 1991) directly address both of these scalability issues, but introduce fundamental parameters and algorithmic challenges. Other approaches include coordinated branch-and-cut procedures (Alonso-Ayuso et al., 2003). In general, the solution of difficult stochastic programs requires both experimentation with and customization of alternative algorithmic paradigms – necessitating the need for generic and configurable solvers.

3.2.1 *PySP*

An open-source software package – PySP – which was developed by Hart and Watson at Sandia and Woodruff at UC Davis, provides a generic and customizable stochastic programming solvers. At the same time, it provides modeling capabilities for expressing stochastic programs. Beyond the obvious need to somehow express a problem instance to a solver, there are fundamental characteristics of the modeling language that are necessary to achieve the objective of generic and customizable stochastic programming solvers. In particular, the modeling layer must provide mechanisms for accessing

components via direct introspection (Python, 2010), such that solvers can generically and programmatically access and manipulate model components. Further, from a user standpoint, the modeling layer should differentiate between abstract models and model instances – following best practices from the deterministic mathematical modeling community (Fourer et al., 1990).

To express a stochastic program in PySP, the user specifies both the deterministic base model and the scenario tree with associated uncertain parameters in the Pyomo open-source algebraic modeling language (Hart et al., 2010). This separation of deterministic and stochastic problem components is similar to the mechanism proposed in SMPS (Birge et al., 1987; Gassmann and Schweitzer, 2001). Pyomo is a Python-based modeling language, and provides the ability to model both abstract problems and concrete problem instances. The embedding of Pyomo in Python enables model introspection and the construction of generic solvers.

Given the deterministic and scenario tree models, PySP provides two paths for the solution of the corresponding stochastic program. The first alternative involves writing the extensive form and invoking a deterministic (mixed-integer) linear solver. For more complex stochastic programs, we provide a generic implementation of Rockafellar and Wets' Progressive Hedging algorithm (Rockafellar and Wets, 1991). Our particular focus is on the use of Progressive Hedging as an effective heuristic for approximating general multi-stage, mixed-integer programs. By leveraging the combination of a high-level programming language (Python) and the embedding of the base deterministic model in that language (Pyomo), we are able to provide completely generic and highly configurable solver implementations. Karabuk and Grant (2007) describe the benefits of Python for such model-building and solving in more detail.

Both PySP and the Pyomo algebraic modeling language upon which PySP is based are actively developed and maintained by Sandia National Laboratories along with an academic developer community. PySP is distributed with the Coopr open-source Python project for optimization, which is now part of the COIN-OR open-source initiative (COIN-OR, 2010).

3.2.2 Other Efforts

We now briefly survey other prior and on-going efforts to develop software packages supporting the specification and solution of stochastic programs, with the objective of placing the capabilities of PySP in this broader context. Numerous extensions to existing AMLs to support the specification of stochastic programs have been proposed in the literature; Gassmann and Irelan (1996) is an early example. Similarly, various solver interfaces have been proposed, with the dominant mechanism being the direct solution of the extensive form. Here, we primarily focus on specification and solver efforts associated with open-source and academic initiatives, which generally share the same distribution goals and user community targets as PySP.

StAMPL

Fourer and Lopes (2009) describe an extension to AMPL, called StAMPL, whose goal is to simplify the modeling process associated with stochastic program specification. One key objective of StAMPL is to explicitly avoid the use of scenario and stage indices when specifying the core algebraic model, thus separating the specification of the stochastic process from the underlying deterministic optimization model. The authors describe a preprocessor that translates a StAMPL problem description into the fully indexed AMPL model, which in turn is written in SMPS format for solution.

STRUMS

STRUMS is a system for performing and managing decomposition and relaxation strategies in stochastic programming (Fourer and Lopes, 2006). Input problems are specified in the SMPS format, and the package provides mechanisms for writing the extensive form, performing basic and nested Benders decomposition (i.e., the L-shaped method), and implementing Lagrangian relaxation; only stochastic linear programs are considered. The design objective of STRUM – to provide mechanisms facilitating automatic problem decomposition – is consistent with the design of PySP. However, PySP currently provides mechanisms for scenario-based decomposition, in contrast to stage-oriented decomposition.

DET2STO

Thénié et al. (2007) describe an extension of AMPL to support the specification of stochastic programs, noting that (at the time the effort was initiated) no AMLs were available with stochastic programming support. In particular, they provide a script – called DET2STO, available from http://apps.ordecsys.com/det2sto – taking an augmented AMPL model as input and generating the extensive form via an SMPS output file. The research focus is on the automated generation of the extensive form, with the authors noting: "We recall here that, while it is relatively easy to describe the two base components - the underlying deterministic model and the stochastic process - it is tedious to define the contingent variables and constraints and build the deterministic equivalent" (Thénié et al., 2007, p.35). While subtle modeling differences do exist between DET2STO and PySP (e.g., in the way scenario-based and transition-based representations are processed), they provide identical functionality in terms of ability to model stochastic programs and generate the extensive form.

SMI

Part of COIN-OR, the Stochastic Modeling Interface (SMI) (SMI, 2010) provides a set of C++ classes to (1) either to programmatically create a stochastic program or to load a stochastic program specified in SMPS, and (2) to write the extensive form of the resulting program. SMI provides no solvers, instead focusing on generation of the extensive form for solution by external solvers. Connections to FLOPC++ (FLOPCPP, 2010) do exist, providing a mechanism for problem description via an AML. While providing a subset

of PySP functionality, the need to express models in a compiled, technically sophisticated programming language (C++) is a significant drawback for many users.

APLEpy

Karabuk (2008) describes the design of classes and methods to implement stochastic programming extensions to his Python-based APLEpy (Karabuk, 2005) environment for mathematical programming, with a specific emphasis on stochastic linear programs. Karabuk's primary focus is on supporting relaxation-based decompositions in general, and the L-shaped method in particular, although his design would create elements that could be used to construct other algorithms as well. The vision expressed in (Karabuk, 2008) is one where the boundary between model and algorithm must be crossed so that the algorithm can be expressed in terms of model elements.

SPInE

SPInE (Mitra et al., 2005) provides an integrated modeling and solver environment for stochastic programming. Models are specified in an extension to AMPL called SAMPL (other base modeling languages are provided), which can in turn be solved via a number of built-in solvers. In contrast to PySP, the solvers are not specifically designed to be customizable, and are generally limited to specific problem classes. For example, multi-stage stochastic linear programs are solved via nested Benders decomposition, while Lagrangian relaxation is the only option for two-stage mixed-integer stochastic programs. SPInE is primarily focused on providing an out-of-the-box solution for stochastic linear programs, which is consistent with the lack of emphasis on customizable solution strategies.

SLP-IOR

Similar to SPInE, SLP-IOR (Kall and Mayer, 2005b) is an integrated modeling and solver environment for stochastic programming, with a strong emphasis on the linear case. In contrast to SPInE, SLP-IOR is based on the GAMS AML, and provides a broader range of solvers. However, as with SPInE, the focus is not on easily customizable solvers (most of the solver codes are written in FORTRAN). Further, the solvers for the integer case are largely ignored.

3.3 Progressive Hedging: A Generic Decomposition Strategy

We now transition from modeling stochastic programs and solving them via the extensive form to decomposition-based strategies, which are in practice typically required to efficiently solve large-scale instances with large numbers of scenarios, discrete variables, or decision stages. There are two broad classes of decomposition-based strategies: horizontal and vertical. *Vertical* strategies decompose a stochastic program by time stages; Van Slyke and Wets' L-shaped method is the primary method in this class (Slyke and Wets, 1969). In contrast, *horizontal* strategies decompose a stochastic program by scenario; Rockafellar and Wets' Progressive Hedging algorithm (Rockafellar

and Wets, 1991) and Caroe and Schultz's Dual Decomposition (DD) algorithm (Caroe and Schultz, 1999) are the two notable methods in this class.

Currently, there is not a large body of literature to provide an understanding of practical, computational aspects of stochastic programming solvers, particularly in the mixed integer case. For any given problem class, there are few heuristics to guide selection of the algorithm likely to be most effective. Similarly, while stochastic programming solvers are typically parameterized and/or configurable, there is little guidance available regarding how to select particular parameter values or configurations for a specific problem. Lacking such knowledge, the interface to solver libraries must provide facilities to allow for rapid parameterization and configuration.

Beyond the need for highly configurable solvers, solvers should also be generic, i.e., independent of any particular AML description. Decomposition strategies are non-trivial to implement, requiring significant development time – especially when more advanced features are considered. The lack of generic decomposition solvers is a known impediment to the broader adoption of stochastic programming. Thénié et al. (2007) concisely summarize the challenge as follows: "Devising efficient solution methods is still an open field. It is thus important to give the user the opportunity to experiment with solution methods of his choice." By introducing both customizable and generic solvers, our goal is to promote the broader use of and experimentation with stochastic programming by significantly reducing the barrier to entry.

3.3.1 The Progressive Hedging Algorithm

Progressive Hedging (PH) was initially introduced as a decomposition strategy for solving large-scale stochastic linear programs (Rockafellar and Wets, 1991). PH is a horizontal or scenario-based decomposition technique, and possesses theoretical convergence properties when all decision variables are continuous. In particular, the algorithm converges in linear time given a convex reference scenario optimization model.

Despite its introduction in the context of stochastic linear programs, PH has proven to be a very effective heuristic for solving stochastic mixed-integer programs. PH is particularly effective in this context when there exist computationally efficient techniques for solving the deterministic single-scenario optimization problems. A key advantage of PH in the mixed-integer case is the absence of requirements concerning the number of stages or the type of variables allowed in each stage – as is common for many proposed stochastic mixed-integer algorithms. A disadvantage is the current lack of provable convergence and optimality results. However, PH has been used as an effective heuristic for a broad range of stochastic mixed-integer programs (Fan and Liu, 2010; Listes and Dekker, 2005; Løkketangen and Woodruff, 1996; Cranic et al., 2009; Hvattum and Løkketangen, 2009). For large, real-world stochastic mixed-integer programs, the determination of optimal solutions is generally not computationally tractable.

The basic idea of PH for the linear case is as follows:

1. For each scenario s, solutions are obtained for Formulation P_s, the problem of minimizing, subject to the problem constraints, the deterministic f_s.
2. The variable values for an implementable – but likely not admissible – solution are obtained by averaging over all scenarios at a scenario tree node.
3. For each scenario s, solutions are obtained for Formulation P_s with additional terms that penalize the lack of implementability using a sub-gradient estimator for the non-anticipativity constraints and a squared proximal term.
4. If the solutions have not converged sufficiently and the allocated compute time is not exceeded, goto Step .
5. Post-process, if needed, to produce a fully admissible and implementable solution.

To begin the PH implementation for solving formulation (P), we first organize the scenarios and decision times into a tree. The leaves correspond to scenario realizations, such that each leaf is connected to exactly one node at time $t \in \mathcal{T}$ and each of these nodes represents a unique realization up to time t. The leaf nodes are connected to nodes at time $t-1$, such that each scenario associated with a node at time $t-1$ has the same realization up to time $t-1$. This process is iterated back to time 1 (i.e., "now"). Two scenarios whose leaves are both connected to the same node at time t have the same realization up to time t. Consequently, in order for a solution to be implementable it must be true that if two scenarios are connected to the same node at some time t, then the values of $x_i(t')$ must be the same under both scenarios for all i and for $t' \le t$.

Progressive Hedging is a technique to iteratively and gradually enforce implementability, while maintaining admissibility at each step in the process. For each scenario s, approximate solutions are obtained for the problem of minimizing, subject to the constraints, the deterministic f_s plus terms that penalize the lack of implementability.

These terms strongly resemble those found when the method of augmented Lagrangians is used (Bertsekas, 1996). The method makes use of a system of row vectors, w, that have the same dimension as the column vector system X, so we use the same shorthand notation. For example, $w(s)$ denotes $(w(s,1),\ldots,w(s,|\mathcal{T}|))$ in the multiplier system.

To provide a formal algorithm statement of PH, we first formalize some of the scenario tree concepts. We use $\Pr(\mathcal{A})$ to denote the sum of $\Pr(s)$ over all s for scenarios emanating from node $\mathcal{A}$ (i.e., those s that are the leaves of the sub-tree having $\mathcal{A}$ as a root, also referred to as $s \in \mathcal{A}$). We use $t(\mathcal{A})$ to indicate the time index for node $\mathcal{A}$ (i.e., node $\mathcal{A}$

corresponds to time t). We use $X(t;\mathcal{A})$ on the left hand side of a statement to indicate assignment to the vector $(x_1(s,t),\ldots,x_{N(t)}(s,|\mathcal{T}|))$ for each $s\in\mathcal{A}$. We refer to vectors at each iteration of PH using a superscript; e.g., $w^{(0)}(s)$ is the multiplier vector for scenario s at PH iteration zero. The PH iteration counter is k.

If we briefly defer the discussion of termination criteria, a formal version of the algorithm (with step numbering that matches in the informal statement just given) can be stated as follows, taking $\varrho>0$ as a parameter.

1. $k\leftarrow 0$
2. For all scenario indices, $s\in S$:

$$X^{(0)}(s)\leftarrow \text{argmin} f_s(X(s)):X(s)\in\Omega_s \qquad (1)$$

and

$$w^{(0)}(s)\leftarrow 0$$

3. $k\leftarrow k+1$
4. For each node, $\mathcal{A}$, in the scenario tree, and for all $t=t(\mathcal{A})$:

$$\overline{X}^{(k-1)}(t;A)\leftarrow \sum_{s\in A}\Pr(s)X(t;s)^{(k-1)}/\Pr(A)$$

5. For all scenario indices, $s\in S$:

$$w^{(k)}(s)\leftarrow w^{(k-1)}(s)+(\varrho)\left(X^{(k-1)}(s)-\overline{X}^{(k-1)}\right)$$

and

$$X^{k}(s)\leftarrow \text{argmin} f_s(X(s))+w^{(k)}(s)X(s)+\varrho/2\left|X(s)-\overline{X}^{k-1}\right|^2 :X(s)\in\Omega_s. \qquad (2)$$

6. If the termination criteria are not met (e.g., solution discrepancies quantified via a metric $g^{(k)}$), then goto Step 3.

The termination criteria are based mainly on convergence, but we must also allow for the use of time-based termination because non-convergence is a possibility. Iterations are continued until k reaches some pre-determined limit or the algorithm has *converged* – which we take to indicate that the set of scenario solutions s is sufficiently homogeneous.

One possible definition is to require the inter-solution distance (e.g., Euclidean) to be less than some parameter.

The value of the perturbation vector ϱ strongly influences the actual convergence rate of PH: if ϱ is small, the penalty coefficients will vary little between consecutive iterations. To achieve tractable PH run-times, significant tuning and problem-dependent strategies for computing ϱ are often required; mechanisms to support such tuning are described in Watson and Woodruff (2010).

4 Conclusions and Research Needs

Many of the issues identified in Section 1 can be addressed with techiniques described in Section 3 but substantial research needs exist. Here we identify priority areas for research and problem aspects that need more attention.

4.1 Resource Planning Goals

To decide whether an optimization or an equilibrium model is more appropriate, the goal of long-term resource planning in the restructured markets must be clarified. The proper model type depends on who is doing the planning and for whom the model is intended. Traditional optimization models may be of interest to policy makers but are based on an unrealistic assumption of a single centralized planner. More recent game theoretic models of generation expansion planning do not identify optimal expansion plans, but rather attempt to predict how competitive generating companies will decide on investments. Whether an optimization or an equilibrium model is more appropriate depends on the planning context and audience.

4.2 Incorporating Uncertainty

Substantial modeling and computational development are needed to exploit the benefits of stochastic optimization for long term resource planning.

The robust optimization approach has been developed to "immunize" problem solutions to uncertainties for which probability distributions cannot be estimated. Recent enhancements incorporate multiple decision stages where later decisions can adjust to realizations of some of the uncertain quantities (Ben-Tal et al., 2004). This approach is appropriate for planning investments in long-lived transmission and/or generation facilities whose long term value depends on unprecedented future outcomes of climate change, structural changes in demand such as from electric vehicles, carbon regulation policies, technological advancements, and extreme disruptive events. Here, adjustable decisions would include generation expansions that may become attractive after transmission is built and may be able to respond to the resolution of some uncertainties. A rigorous method for incorporating such uncertainties in stages should outperform the current industry approach of composing possible futures, optimizing an investment plan

for each future, and implementing their common elements. At the same time, stochastic programming is a proven method for solving high-dimensional, multistage decision problems when distributional knowledge of uncertain elements does exist. Discrete as well as continuous decisions can be considered in portfolios and shorter term risk is controlled by CVaR or chance constraints. Stochastic programming formulations naturally decompose into planning decisions that must be taken before realizations of random variables are known and operational decisions that incorporate recourse to those realizations. Computationally efficient solution methods will enable consideration of increased operational variability due to growing penetration of renewable generation, volatility of fuel prices, and expansion of wholesale electricity markets. Careful modeling that judiciously combines these approaches – perhaps hierarchically, as robust optimization with stochastic subproblems – could result in tractable formulations whose solutions perform well under both the low- and high-frequency uncertainties.

Modeling large numbers of uncertain variables that evolve over a long time horizon may result in an enormous number of scenarios to consider in a stochastic programming formulation. Research is needed to determine how many scenarios must be generated to capture the stochastic aspects of various planning problems. A related arises when, for the sake of computational tractability, only a sample of the set of scenarios is used. The sample size must be chosen carefully to ensure that replicating the entire sampling and solving procedure results in reasonably stable solutions and objective function values. General methods for reducing a large set of scenarios to a smaller representative set based on probability measures are increasingly used, but may require significant computational time in themselves. More effective methods may be devised by considering more information about the problem structure and the impact of different scenarios on first-stage decisions. A new scenario reduction heuristic applied in the context of medium-term simulation of energy flows has shown promise for improving the tradeoff between accuracy and computation time (Wang, 2010; Wang and Ryan, 2010). Finally, effective use of massively parallel computing resources should be investigated.

4.3 Operational Constraints

In the current long term resource planning models, significant computational effort is required to adequately capture the effects of operational constraints on the planning function. This need will only become more intensive as the amount of variable resources (mainly wind and solar) grows. Two important examples follow:

- Effects on fossil plant cycling: Increased penetrations of variable generation will necessarily cause cycling of fossil-fired power plants, where their loading cycles between various levels, or for the highest penetration levels, they are de-committed and then committed from hot, warm, and cold starts. These plants were not designed to experience the resulting temperature variability. This variability

will initially cause maintenance costs and forced outage rates to increase, and will eventually cause significant reduction in plant life.

- Valuing fast-response plants: Variability in net load will place increased value on fast response gas turbines, storage, and load control, but this increased value will be difficult to capture within production cost models implemented with one hour time intervals. Greater time granularity of production cost models will facilitate capture of these effects. A key research area will be to enable such capture, within both production cost models and resource optimization models, without increasing model complexity and computational burden.

4.4 Linear Programming Formulations for Optimization of Transmission with Continuous Capacity

A key computational barrier for transmission optimization results when branch reactance is modeled as a function of branch capacity, because this requires multiplication of two variables, susceptance (the inverse of reactance when neglecting resistance) and angle, to obtain power flow, thus forcing the use of the more computationally intensive nonlinear solvers. One method to handle this is to use a transportation flow model, but this approach does not account for the important effect of reactances on network power flows. Another method (Wang and McDonald, 1994; Jin and Ryan, 2011) provides for discrete selection of branches having pre-designed capacity and reactance, but this approach optimizes only with respect to the candidate solutions provided; in addition, it requires a linear integer solution which can be computationally intensive. Methods for identifying continuous capacity values which maintain corresponding reactance and are solvable as a linear program are not available and represent an important problem to be addressed.

References

[1] C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking and Finance*, 26: 1487—1503, 2002.

[2] Adica. http://www.adica.com/wasp\s\do5(i)v.html, 2010.

[3] AIMMS. Optimization software for operations research applications. http://www.aimms.com/operations-research/mathematical-programming/stochastic-programming, July 2010.

[4] Antonio Alonso-Ayuso, Laureano F. Escudero, and M. Teresa Ortuno. BFC, a branch-and-fix coordination algorithmic framework for solving some types of stochastic pure and mixed 0-1 programs. *European Journal of Operational Research*, 151 (3): 503–519, 2003.

[5] American Electric Power. Interstate transmission vision for wind integration. www.aep.com/about/i765project/docs/WindTransmissionVisionWhitePaper.pdf, 2010.

[6] American Superconductor. Superconductor electricity pipelines: Carrying renewable electricity across the u.s.a. http://www.amsc.com/pdf/PL\s\do5(W)P\s\do5(0)810\s\do5(l)ores.pdf, 2009.

[7] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9: 203—228, 1999.

[8] J. Bastian, J. Zhu, V. Banunarayanan, and R. Mukerji. Forecasting energy prices in a competitive market. *IEEE Computer Applications in Power Magazine*, 12: 40–45, July 1999.

[9] M. L. Baughman, S. N. Siddiqi, and J. W. Zarnikau. Integrating transmission into IRP part I: analytical approach. *IEEE Transactions on Power Systems*, 10: 1652–1659, 1995a.

[10] M. L. Baughman, S. N. Siddiqi, and J. W. Zarnikau. Integrating transmission into IRP part II: case study results. *IEEE Transactions on Power Systems*, 10: 1660–1666, 1995b.

[11] M. Bazaraa, J. Jarvis, and H. Sherali. *Linear Programming and Network Flows*. Wiley, New York, 1992.

[12] Ben-Tal and A. Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92 (3): 453–480, 2002.

[13] Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99 (2): 351–376, 2004.

[14] Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.

[15] Silvio Binato, Mário Veiga F. Pereira, and Sérgio Granville. A new Benders decomposition approach to solve power transmission network design problems. *IEEE Transactions on Power Systems*, 16: 235–240, 2001.

[16] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 1997.

[17] J.R. Birge, M.A. Dempster, H.I. Gassmann, E.A. Gunn, A.J. King, and S.W. Wallace. A standard input format for multiperiod stochastic linear program. *COAL (Math. Prog. Soc., Comm. on Algorithms) Newsletter*, 17: 1–19, 1987.

[18] Jeremy A. Bloom. Long-range generation planning using decomposition and probabilistic simulation. *IEEE Transactions on Power Apparatus and Systems*, PAS-101: 797–802, 1982.

[19] Jeremy A. Bloom. Solving an electricity generating capacity expansion planning problem by generalized Benders' decomposition. *Operations Research*, 31: 84–100, 1983.

[20] Adam B. Borison, Peter A. Morris, and Shmuel S. Oren. A state-of-the-world decomposition approach to dynamics and uncertainty in electric utility generation expansion planning. *Operations Research*, 32: 1052–1068, 1984.

[21] Audun Botterud, Marija D. Ilic, and Ivar Wangensteen. Optimal investments in power generation under centralized and decentralized decision making. *IEEE Transactions on Power Systems*, 20: 254–263, 2005.

[22] Vikram S. Budhraja, Fred Mobasheri, John Ballance, Jim Dyer, Alison Silverstein, and Joseph H. Eto. Improving electricity resource-planning processes by considering the strategic benefits of transmission. *The Electricity Journal*, 22 (2): 54–63, 2009.

[23] M. Buygi, G. Balzer, H. Shanechi, and M. Shahidehpour. Market-based transmission expansion planning. *IEEE Trans Pwr Sys*, 19: 2060–2067, November 2004.

[24] C.C. Caroe and R. Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24 (1–2): 37–45, 1999.

[25] X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55 (6): 1058–1071, 2007.

[26] J. Choi, T. Tran, A. El-Keib, R. Thomas, H. Oh, and R. Billinton. A method for transmission system expansion planning considering probabilistic reliability criteria. *IEEE Transactions on Power Systems*, 20: 1606–1615, 2005.

[27] Angela S. Chuang, Felix Wu, and Pravin Varaiya. A game-theoretic model for generation expansion planning: problem formulation and numerical comparisons. *IEEE Transactions on Power Systems*, 16: 885–891, 2001.

[28] COIN-OR. COmputational INfrastructure for Operations Research. http://www.coin-or.org, July 2010.

[29] J. Contreras and F. F. Wu. Coalition formation in transmission expansion planning. *IEEE Transactions on Power Systems*, 14: 1144–1152, 1999.

[30] J. Contreras and F. F. Wu. A kernel-oriented algorithm for transmission expansion planning. *IEEE Transactions on Power Systems*, 15: 1434–1440, 2000.

[31] T.G. Cranic, X. Fu, M. Gendreau, W. Rei, and S.W. Wallace. Progressive hedging-based meta-heuristics for stochastic network design. Technical Report CIRRELT-2009-03, University of Montreal CIRRELT, January 2009.

[32] G. B. Dantzig, P. W. Glynn, M. Avriel, J. C. Stone, R. Entriken, and M. Nakayama. Decomposition techniques for multi-area generation and transmission planning under uncertainty. Technical Report EL-6484, Electric Power Research Institute, 1989.

[33] T. de la Torre, J. W. Feltes, T. Gómez, and H. M. Merril. Deregulation, privatization, and competition: transmission planning under uncertainty. *IEEE Transactions on Power Systems*, 14: 460–465, 1999.

[34] de Silva, M. J. Rider, R. Romero, A. V. Garcia, and C. A. Murari. Transmission network expansion planning with security constraints. *IEEE Transactions on Power Systems*, 21: 1565–1573, 2006.

[35] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6: 182–197, 2002.

[36] J. Dupačová, N. Gröwe-Kuska, and W. Römisch. Scenario reduction in stochastic programming. *Mathematical Programming*, 95 (3): 493–511, 2003.

[37] Energy Exemplar. http://www.energyexemplar.com, 2010.

[38] Energy Information Administration. The national energy modeling system: An overview. http://www.eia.doe.gov/oiaf/aeo/overview/index.html, 2003.

[39] Energy Information Administration. Emissions of greenhouse gases in the united states. www.eia.doe.gov/oiaf/1605/ggrpt/carbon.html, 2008.

[40] Energy Technology Network. MARKAL and TIMES model documentation. www.etsap.org/documentation.asp, 2010.

[41] EnerNex Corporation. Eastern wind integration and transmission study, January 2010.

[42] Antonio H. Escobar, Ramón Alfonso Gallego, and Rubén Romero. Multistage and coordinated planning of the expansion of transmission systems. *IEEE Transactions on Power Systems*, 19: 735–744, 2004.

[43] Y. Fan and C. Liu. Solving stochastic transportation network protection problems using the progressive hedging-based method. *Networks and Spatial Economics*, 10 (2): 193–208, 2010.

[44] Risheng Fang and David J. Hill. A new strategy for transmission expansion in competitive electricity markets. *IEEE Transactions on Power Systems*, 18: 374–380, 2003.

[45] J. Fleeman, R. Gutman, M. Heyeck, M. Bahrman, and B. Normark. EHV AC and HVDC transmission working together to integrate renewable power. Technical report, CIGRE, 2009. Paper 978-2-85873-080-3.

[46] FLOPCPP. Flopc++: Formulation of linear optimization problems in c++. https://projects.coin-or.org/FlopC++, August 2010.

[47] R. Fourer and L. Lopes. A management system for decompositions in stochastic programming. *Annals of Operations Research*, 142: 99–118, 2006.

[48] R. Fourer and L. Lopes. Stampl: A filtration-oriented modeling tool for multistage recourse problems. *INFORMS Journal on Computing*, 21 (2): 242–256, 2009.

[49] Robert Fourer, David M. Gay, and Brian W. Kernighan. AMPL: A mathematical programming language. *Management Science*, 36: 519–554, 1990.

[50] H. I. Gassmann and E. Schweitzer. A comprehensive input format for stochastic linear programs. *Annals of Operations Research*, 104: 89–125, 2001.

[51] H.I. Gassmann and A.M. Irelan. On the formulation of stochastic linear programs using algebraic modeling languages. *Annals of Operations Research*, 64: 83–112, 1996.

[52] General Electric. http://www.gepower.com/prod\s\do5(s)erv/products/utility\s\do5(s)oftware/software/en/ge\s\do5(m)ars.htm., 2010.

[53] M. Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10 (4): 237–260, 1972.

[54] E. Gil. *Integrated network flow model for a reliability assessment of the national electric energy system*. PhD thesis, Iowa State University, 2007.

[55] E. Gil and J. McCalley. A US energy system model for disruption analysis: Validation using effects of 2005 hurricanes. *IEEE Transactions on Power Systems*, forthcoming.

[56] T. Hansen. Tres Amigas technology holds promise for expanding renewable energy supply. *Electric Light & Power*, January 2010.

[57] William E. Hart, Jean-Paul Watson, and David L. Woodruff. Python optimization modeling objects (Pyomo). *submitted to Mathematical Programming Computation*, 2010.

[58] W. Head and H. Nguyen. The procedure used to assess the long range generation and transmission resources in the mid-continent area power pool. *IEEE Transactions on Power Systems*, 5: 1137 – 1145, 1990.

[59] L. Hecker, Z. Zhou, D. Osborn, and J. Lawhorn. Value-based transmission planning process for joint coordinated system plan. In *Proc. of the 2009 IEEE PES General Meeting*, 2009.

[60] Benjamin F. Hobbs. Optimization methods for electric utility resource planning. *European Journal of Operational Research*, 83: 1–20, 1995.

[61] L.M. Hvattum and A. Løkketangen. Using scenario trees and progressive hedging for stochastic inventory routing problems. *Journal of Heuristics*, 15 (6): 527–557, 2009.

[62] E. Ibanez and J. McCalley. Multiobjective evolutionary algorithm for long-term planning of the national energy and transportation systems. Technical report, Iowa State University, Feb. 2010. under revision for *Energy Systems*.

[63] E. Ibanez, J. McCalley, D. Aliprantis, R. Brown, K. Gkritza, A. Somani, and L. Wang. National energy and transportation systems: Interdependencies within a long term planning model. In *Proc. of IEEE Energy 2030 Conference*, Nov 2008.

[64] E. Ibanez, K. Gkritza, J. McCalley, D. Aliprantis, A. Somani R. Brown, and L. Wang. Interdependencies between energy and transportation systems for national long term planning. In K. Gopalakrishnan and S. Peeta, editors,

Sustainable Infrastructure Systems: Simulation, Imaging, and Intelligent Engineering. Springer-Verlag, Berlin, 2010.

[65] Shan Jin and Sarah M. Ryan. Capacity expansion in the integrated supply network for an electricity market. *IEEE Transactions on Power Systems*, in press, 2011.

[66] Peter Kall and Janos Mayer. *Stochastic Linear Programming: Models, Theory, and Computation*. Springer, 2005a.

[67] Peter Kall and Janos Mayer. *Applications of Stochastic Programming*, chapter Building and Solving Stochastic Linear Programming Models with SLP-IOR. MPS-SIAM, 2005b.

[68] S. Karabuk. An open source algebraic modeling and programming software. Technical report, University of Oklahoma, School of Industrial Engineering, Norman, OK, 2005.

[69] S. Karabuk. Extending algebraic modeling languages to support algorithm development for solving stochastic programming models. *IMA Journal of Management Mathematics*, 19: 325–345, 2008.

[70] S. Karabuk and F.H. Grant. A common medium for programming operations-research models. *IEEE Software*, 24 (5): 39–47, 2007.

[71] V. Koritanov and T. Veselka. Modeling the regional electricity network in Southeast Europe. In *Proc. of the Pwr Engr Society General Meeting*, 2003.

[72] Gerardo Latorre, Rubén Darío Cruz, Jorge Mauricio Areiza, and Andrés Villegas. Classification of publications and models on transmission expansion planning. *IEEE Transactions on Power Systems*, 18: 938–946, 2003.

[73] G. Latorre-Bayona and I. J. Pérez-Arriaga. Chopin, a heuristic model for long term transmission expansion planning. *IEEE Transactions on Power Systems*, 9: 1886–1894, 1994.

[74] LINDO. LINDO systems, August 2010.

[75] O. Listes and R. Dekker. A scenario aggregation based approach for determining a robust airline fleet composition. *Transportation Science*, 39: 367–382, 2005.

[76] Løkketangen and D. L. Woodruff. Progressive hedging and tabu search applied to mixed integer (0,1) multistage stochastic programming. *Journal of Heuristics*, 2: 111–128, 1996.

[77] M. Lu, Z. Dong, and T. Saha. A framework for transmission planning in a competitive electricity market. In *Proc. of the Transmission and Distribution Conference and Exhibition: Asia and Pacific*, pages 1–6, 2005.

[78] Scott A. Malcolm and Stavros A. Zenios. Robust optimization for power systems capacity expansion under uncertainty. *Journal of the Operational Research Society*, 45: 1040–1049, 1994.

[79] Maximal Software. http://www.maximal-usa.com/maximal/news/stochastic.html, July 2010.

[80] G. Mitra, P. Valente, and C.A. Poojari. *Applications of Stochastic Programming*, chapter The SPInE Stochastic Programming System. MPS-SIAM, 2005.

[81] Birger Mo, Jan Hegge, and Ivar Wangensteen. Stochastic generation expansion planning by means of stochastic dynamic programming. *IEEE Transactions on Power Systems*, 6: 662–668, 1991.

[82] F. H. Murphy, S. Sen, and A. L. Soyster. Electric utility capacity expansion planning with uncertain load forecasts. *AIIE Transactions*, 14: 52–59, 1982.

[83] Frederic H. Murphy and Yves Smeers. Generation capacity expansion in imperfectly competitive restructured electricity markets. *Operations Research*, 53: 646–661, 2005.

[84] Vishnu Nanduri, Tapas K. Das, and Patricio Rocha. Generation capacity expansion in energy markets using a two-level game-theoretic model. *IEEE Transactions on Power Systems*, 24: 1165–1172, 2009.

[85] Andy Philpott. Modeling uncertainty in optimization problems. http://www.epoc.org.nz/papers/PhilpottEORMSRevision.pdf, June 2010.

[86] Warren B. Powell. Feature article - merging AI and OR to solve high-dimensional stochastic optimization problems using approximate dynamic programming. *INFORMS Journal on Computing*, 22 (1): 2–17, 2010.

[87] Powercosts, Inc. http://www.powercosts.com/products/gentrader.asp, 2010.

[88] Prekopa. *Handbooks in Operations Research and Management Science, Volume 10: Stochastic Programming*, chapter Probabilistic Programming. Elsevier, 2003.

[89] PSERC. U.S. energy infrastructure investment: Long-term strategic planning to inform policy development. www.pserc.wisc.edu/documents/publications/papers/2009\s\do5(g)eneral\s\do5(p)ublications/pserc\s\do5(e)nergy\s\do5(m)odeling\s\do5(w)hite\s\do5(p)aper\s\do5(m)arch\s\do5(2)009\s\do5(a)dobe7.pdf, March 2009. White paper developed by researchers of the Power Systems Engineering Research Center (PSERC).

[90] Dive Into Python. http://diveintopython.org/power\s\do5(o)f\s\do5(i)ntrospection/index.html, July 2010.

[91] P. Riehmann, M. Hanfler, and B. Froehlich. Interactive sankey diagrams. In *Proc. of IEEE Symp on Info. Visualization*, 2005.

[92] R. T. Rockafellar and R. J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16 (1): 119–147, 1991.

[93] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26, 2002.

[94] Jae Hyung Roh, Mohammad Shahidehpour, and Yong Fu. Market-based coordination of transmission and generation capacity planning. *IEEE Transactions on Power Systems*, 22: 1406–1419, 2007.

[95] Jae Hyung Roh, Mohammad Shahidehpour, and Lei Wu. Market-based generation and transmission planning with uncertainties. *IEEE Transactions on Power Systems*, 24: 1587–1598, 2009.

[96] R. Romero and A. Monticelli. A hierarchical decomposition approach for transmission network expansion planning. *IEEE Transactions on Power Systems*, 9: 373–380, 1994.

[97] Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Academic Press, Inc., Orlando, FL, USA, 1983. ISBN 0125984200.

[98] Ruszczyński. Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra. *Mathematical Programming*, 93: 195–215, 2002.

[99] Enzo E. Sauma and Shmuel S. Oren. Economic criteria for planning transmission investment in restructured electricity markets. *IEEE Transactions on Power Systems*, 22: 1394–1405, 2007.

[100] R. Schultz and S. Tiedemann. Conditional value-at-risk in stochastic programs with mixed-integer recourse. *Math. Program. (B)*, 105: 365–386, 2006.

[101] Scorecard. www.scorecard.org/env-releases/cap/pollutant-desc.tcl, 2010.

[102] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2009.

[103] W. Short, N. Blair, P. Sullivan, and T. Mai. ReEDS model documentation: Base case data and model description. Technical report, National Renewable Energy Laboratory, August 2009.

[104] S. N. Siddiqi and M. L. Baughman. Value-based transmission planning and the effects of network models. *IEEE Transactions on Power Systems*, 10: 1835 – 1842, 1995.

[105] Siemens. http://www.pti-us.com/pti/software/tplan/index.cfm, 2006.

[106] R. M. Van Slyke and R. J. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17: 638–663, 1969.

[107] SMI. SMI. https://projects.coin-org.org/Smi, August 2010.

[108] J. Thénié, Ch. van Delft, and J.-Ph. Vial. Automatic formulation of stochastic programs via an algebraic modeling language. *Computational Management Science*, 4 (1): 17–40, January 2007.

[109] T. Tran, J. Kwon, D. Jeon, and K. Han. Sensitivity analysis of probabilitstic reliablility evaluation of KEPCO system using TRELSS. In *International Conf on Probabilistic Methods Applied to Power Systems*, 2006.

[110] U.S. Department of Transportation. Commodity Flow Survey United States. http://www.bts.gov/publications/commodity\s\do5(f)low\s\do5(s)urvey/, 1997.

[111] US Environmental Protection Agency. www.epa.gov/airmarkets/progsregs/epa-ipm/index.html, 2010.

[112] Valente, G. Mitra, M. Sadki, and R. Fourer. Extending algebraic modelling languages for stochastic programming. *INFORMS Journal On Computing*, 21 (1): 107–122, 2009.

[113] N. van Beeck. *Classification of energy models*. Tilburg University and Eindhoven University of Technology, May 1999.
[114] Ventyx. http://www1.ventyx.com/analytics/promod.asp, 2010a.
[115] Ventyx. http://www.ventyx.com/analytics/market-analytics.asp, 2010b.
[116] Stein W. Wallace and William T. Ziemba, editors. *Applications of Stochastic Programming*. Society for Industrial and Applied Mathematics, 2005.
[117] Jianhui Wang, Mohammad Shahidehpour, Zuyi Li, and Audun Botterud. Strategic generation capacity expansion planning with incomplete information. *IEEE Transactions on Power Systems*, 24: 1002–1010, 2009.
[118] X. Wang and J. McDonald. *Modern Power System Planning*. McGraw Hill, 1994.
[119] Y. Wang. *Scenario reduction heuristics for a rolling stochastic programming simulation of bulk energy flows with uncertain fuel costs*. PhD thesis, Iowa State University, 2010.
[120] Y. Wang and S. M. Ryan. Effects of uncertain fuel costs on optimal energy flows in the U.S. *Energy Systems*, 1: 209–243, 2010.
[121] J.-P. Watson, , and D.L. Woodruff. Pysp: Modeling and solving stochastic programs in Python. Technical report, Sandia National Laboratories, Albuqueque, NM, USA, 2010a.
[122] J.-P. Watson, R. J.-B. Wets, and D.L. Woodruff. Scalable heuristics for a class of chance constrained stochastic programs. *INFORMS Journal on Computing – to appear*, 2010b.
[123] J.P. Watson and D.L. Woodruff. Progressive hedging innovations for a class of stochastic mixed-integer resource allocation problems. *Computational Management Science (to appear)*, 2010.
[124] Xpress-Mosel. http://www.dashopt.com/home/products/products\s\do5(s)p.html, July 2010.
[125] XpressMP. FICO express optimization suite. http://www.fico.com/en/products/DMTools/pages/FICO-Xpress-Optimization-Suite.aspx, July 2010.
[126] H. Yamin, J. Finney, W. Rutz, and H. Holly. GENCOs portfolio management using "STRATEGIST" in deregulated power markets. In *Large Engineering Systems Conference on Power Engineering*, July 2001.

Appendix A: NETPLAN cost minimization formulation

The optimization problem associated with this model can be conceptually described by (3),

$$\begin{aligned} &\textbf{min} \quad CostOp + CostInv \\ &\textbf{subject to}: \\ &Meet\ energy\ demand, \\ &Meet\ transportation\ demand, \\ &Capacity\ constraints \\ &Power\ \ flow\ constraints\ on\ electric\ transmission \end{aligned} \tag{3}$$

The objective is to minimize the combined energy and a transportation cost with constraints of meeting demands on energy, and freight transport while following the capacity constraints. The operational characteristics of the model provide additional constraints on the system, such as the inclusion of power flow relations for electric transmission lines (we used the so-called "DC" power flow approximation for this purpose).

This section contains a rigorous description of the formulation needed to achieve the characteristics described above. The explanation of the formulation is preceded by an introduction to the nomenclature used. Computational experience is reported in Ibanez and McCalley (2010); in addition, the model has been implemented in PySP (see section 3.2.1).

Nomenclature

Decision Variables

$e_{(i,j)}(t)$: Operational flow of energy arc from node i to node ,j, for time step t (MWh)

$eInv_{(i,j)}(t)$: Capacity investment on energy arc from node i to node j, for time step t. (MW)

$f_{(i,j,k,m)}(t)$: Operational flow of transportation arc from node i to node j for commodity k using transportation mode m during time step t (ton)

$fleetInv_{(i,j,m)}(t)$: Fleet m capacity investment for transportation arc from node i to node j during time step t (ton/hour)

$infInv_{(i,j,l)}(t)$: Infrastructure l capacity investment for transportation arc from node i to node j for time step t (ton/hour)

$\Theta_i(t)$: Phase angle at node *i*, used to model DC power flow (radians)

Sets and networks

N^E: Set of energy nodes

A^E: Set of energy arcs

$A^E_{DC} \subset A^E$: Set of AC electric transmission arcs, which satisfy DC power flow equations

N^T: Set of transportation nodes

A^T: Set of transportation arcs

$A^T_j \subset A^T$: Subset of transportation arcs that create an energy demand at energy node *j*

K: Set of commodities

$K_c \subset K$: Subset of commodities used by the energy system, e.g. coal

M: Set of fleet or transportation modes

$M_l \subset M$: Subset of fleet that can use transportation infrastructure *l*

$M_j \subset M$: Subset of fleet that require energy from energy node *j*

$n^E_{(i,k)} \in N^E$: Energy node that corresponds to the geographic location *i* and commodity *k* in the transportation system

Energy Parameters

$\eta_{(i,j)}(t)$: Efficiency of arc (*i,j*) during time *t* (unitless)

$lbe_{(i,j)}(t)$: Lower bound for flow in arc (*i,j*) for time *t* (MWh)

$ube_{(i,j)}(t)$: Upper bound for flow in arc (*i,j*) during time *t* due to the initial existing infrastructure (MWh)

$lbeInv_{(i,j)}(t)$: Minimum allowed capacity increase in arc (*i,j*) at time *t* (MW)

$ubeInv_{(i,j)}(t)$: Maximum allowed capacity increase in arc arc (*i,j*) at time *t* (MW)

$costOp_{(i,j)}(t)$: Operational cost for flow in arc (i,j) during time t ($/MWh)

$costInv_{(i,j)}(t)$: Investment cost for capacity increase in arc (i,j) in network l, at time t ($/MW)

$heatContent_k(t)$: Heat content of commodity k, at time t (MWh/ton)

$d_j^E(t)$: Fixed energy demand at node j during time t (MWh)

$b_{(i,j)}$: Susceptance of transmission line between nodes i and j (Ω^{-1})

Transportation Parameters

$ubFleet_{(i,j,m)}(t)$: Upper bound for total transportation flow for fleet m in arc (i,j) during time t due to the initial existing fleet (ton)

$lbFleetInv_{(i,j,m)}(t)$: Minimum allowed capacity increase in arc (i,j) for fleet m at time t (ton/h)

$ubFleetInv_{(i,j,m)}(t)$: Maximum allowed capacity increase in arc arc (i,j) for fleet m at time t (ton/h)

$ubInf_{(i,j,l)}(t)$: Upper bound for total transportation flow across infrastructure l in arc (i,j) during time t due to the initial existing infrastructure (ton)

$lbInfInv_{(i,j,l)}(t)$: Minimum allowed capacity increase in arc (i,j) for transportation infrastructure l at time t (ton/h)

$ubInfInv_{(i,j,l)}(t)$: Maximum allowed capacity increase in arc arc (i,j) for transportation infrastructure l at time t (ton/h)

$costOp_{(i,j,k,m)}(t)$: Operational cost for transportation flow in arc (i,j) in fleet m, commodity k and time t ($/ton)

$costFleetInv_{(i,j,m)}(t)$: Investment cost for capacity increase in fleet m for arc (i,j) and time t ($ h/ton)

$costInfInv_{(i,j,l)}(t)$: Investment cost for capacity increase in transportation infrastructure l for arc (i,j) and time t ($ h/ton)

$fuelCons_{(i,j,m)}(t)$: Fuel consumption for transportation mode m for transportation arc (i,j) during time step t (MWh/ton)

$d^{T}_{(i,j,k)}(t)$: Fixed transportation demand of commodity k for arc (i,j) during time t (ton)

Auxiliary parameters

These parameters are calculated as a combination of decision variables and predetermined parameters.

$CostOp^{E}$: Operational cost from the energy system ($)

$CostInv^{E}$: Cost due to the investment upgrades on the energy system ($)

$CostOp^{T}$: Operational cost for transportation system ($)

$CostFleetInv^{T}$: Investement cost on transportation fleet ($)

$CostInfInv^{T}$: Cost on transportation infrastructure ($)

$d^{ET}_{j}(t)$: Energy demand at node j during time t due to the transportation of commodities (MWh)

Other parameters

r: Discount rate

Δt: Length of time step t (h)

General formulation

The following formulation (4) incorporates the modeling capabilities that have been previously described in this paper.

The energy sector is represented by a set of nodes N^{E} and arcs A^{E}. Each node represents an energy subnetwork in a geographic location. For example, for a particular location, there might be three energy nodes; one for electricity, one for gas, and one for petroleum. Energy arcs link these various nodes, both within same subnetwork (representing transmission lines, natural gas pipelines, or petroleum pipelines) or different, where conversion takes place (e.g., power plants).

The flow across these arcs, $e_{(i,j)}$, are part of the decision variables of the problem. These flows must be such that they satisfy the demand for energy (4g), part of which is due to

the energy required to perform the movement of commodities in the transportation system (4n). Energy flows are also required to meet lower and upper bound constraints (4h). The upper bound is determined by the capacity of the existing energy infrastructure and is a combination of the initial capacity, $ube_{(i,j)}(t)$, and investment on upgrades, $eInv_{(i,j)}(t)$. The first is a parameter while the second is another decision variable, with its corresponding lower and upper bound (4p). Energy flow must also satisfy other constraints, such as DC power flow equations for electric transmission (4i).

The transportation system is also formed by a set of nodes N^T and a set of arcs A^T, although the nomenclature is slightly different. Here, each node represents a unique geographic location, while the transportation flows are referred to as $f_{(i,j,k,m)}$, where (i,j) represent the origin and destination nodes, k the commodity that is being transported and m the mode of transportation used.

These flows must satisfy the transportation demand, which is predetermined (4j) for all commodities except for those that are energy related (e.g., coal, uranium). In that case, the transportation demand depends on the use of that commodity in the energy system (4k). Transportation flows are limited by the capacity of the available fleet (4l) and transportation infrastructure(4m). Both can be increased by the their respective investments, $fleetInv_{(i,j,m)}(t)$ and $infInv_{(i,j,m)}(t)$, which have upper and lower bound constraints (4r,4s).

$$\textbf{min} \quad \{CostOp^E + CostInv^E + CostOp^T + CostFleetInv^T + CostInfInv^T\} \tag{4a}$$

where:

$$CostOp^E = \sum_t \sum_{(i,j)} (1+r)^{-t} costOp^E_{(i,j)}(t)\, e_{(i,j)}(t) \tag{4b}$$

$$CostInv^E = \sum_t \sum_{(i,j)} (1+r)^{-t} costInv^E_{(i,j)}(t)\, eInv_{(i,j)}(t) \tag{4c}$$

$$CostOp^T = \sum_t \sum_{(i,j,k,m)} (1+r)^{-t} costOp^T_{(i,j,k,m)}(t)\, f_{(i,j,k,m)}(t) \tag{4d}$$

$$CostFleetInv^T = \sum_t \sum_{(i,j,m)} (1+r)^{-t} costFleetInv_{(i,j,m)}(t)\, fleetInv_{(i,j,m)}(t) \tag{4e}$$

$$CostInfInv^T = \sum_t \sum_{(i,j,l)} (1+r)^{-t} costInfInv_{(i,j,l)}(t)\, infInv_{(i,j,l)}(t) \tag{4f}$$

subject to:

$$\sum_i \eta_{(i,j)}(t) e_{(i,j)}(t) - \sum_k e_{(j,k)}(t) = d^E_j(t) + d^{ET}_j(t) \tag{4g}$$

$$lbe_{(i,j)}(t) \le e_{(i,j)}(t) \le ube_{(i,j)}(t)\,\Delta t + \sum_{z=0}^{t} eInv_{(i,j)}(z)\,\Delta z \tag{4h}$$

$$e_{(i,j)}(t) = b_{(i,j)}\Big(\theta_i(t) - \theta_j(t)\Big), \qquad \forall (i,j) \in \mathcal{A}^E_{DC} \tag{4i}$$

$$\sum_m f_{(i,j,k,m)}(t) = d^T_{(i,j,k)}(t), \quad k \in \mathcal{K}\backslash\mathcal{K}_e \tag{4j}$$

$$\sum_m f_{(i,j,k,m)}(t) = heatContent_k(t)\, e_{(n^E_{(i,k)}, n^E_{(j,k)})}(t), \quad k \in \mathcal{K}_e \tag{4k}$$

$$\sum_k f_{(i,j,k,m)}(t) \le ubFleet_{(i,j,m)}(t)\,\Delta t + \sum_{z=0}^{t} fleetInv_{(i,j,m)}(z)\,\Delta z \tag{4l}$$

$$\sum_k \sum_{m \in \mathcal{M}_l} f_{(i,j,k,m)}(t) \le ubInf_{(i,j,l)}(t) + \sum_{z=0}^{t} infInv_{(i,j,l)}(z)\,\Delta z \tag{4m}$$

$$d^{ET}_j(t) = \sum_{(a,b) \in \mathcal{A}^T_j} \sum_{m \in \mathcal{M}_j} fuelCons_{(a,b,m)}(t) \sum_k f_{(a,b,k,m)}(t) \tag{4n}$$

$$e_{(i,j)}(t) \ge 0 \tag{4o}$$

$$lbeInv_{(i,j)}(t) \le eInv_{(i,j)}(t) \le ubeInv_{(i,j)}(t) \tag{4p}$$

$$f_{(i,j,k,m)} \ge 0 \tag{4q}$$

$$lbFleetInv_{(i,j,m)}(t) \le fleetInv_{(i,j,m)}(t) \le ubFleetInv_{(i,j,m)}(t) \tag{4r}$$

$$lbInfInv_{(i,j,l)}(t) \le infInv_{(i,j,l)}(t) \le ubInfInv_{(i,j,l)}(t) \tag{4s}$$

$$-\pi \le \theta_i(t) \le \pi \tag{4t}$$

The constraints are as follows: Equation (4g) requires that energy demand be met at every node. Inequalities (4h) are lower and upper bounds on the energy flows. Equations (4i) are the DC power flow equations. Equations (4j) represent transportation demands for non-energy commodities, while equations (4k) are transportation demands for energy commodities. Inequality (4l) is a fleet upper bound for transportation flows, inequality (4m) is an infrastructure upper bound for transportation flows, and equation (4n) represents energy demand from the transportation system. Bounds on the decision variables are given by (4o) for the energy flows, (4p) for the energy capacity investments, (4q) for the transportation flows, (4r) for the fleet investments, (4s) for the infrastructure investments, and (4t) for the voltage phase angles.

Compact notation

A more compact version of (4) can be produced using vector and matrix notation. First of all, the vector $\vec{capInv}$ includes all capacity investment variables, while the operational variables are grouped in vectors called $\vec{flows}_t$. The subscript t in this section represents the operational year to which a variable belongs. When using this notation we can reduce the model to the expressions (5).

$$\textbf{min} \quad \begin{bmatrix} \vec{Cost}\vec{Inv}^{\top} & \vec{Cost}\vec{Op}_1^{\top} & \vec{Cost}\vec{Op}_2^{\top} & \dots \end{bmatrix} \begin{bmatrix} \vec{capInv} \\ \vec{flows}_1 \\ \vec{flows}_2 \\ \vdots \end{bmatrix}$$

$$\textbf{subject to:}$$

$$\begin{bmatrix} -\vec{I} & \vec{0} & \vec{0} & \dots \\ \vec{I} & \vec{0} & \vec{0} & \dots \\ \vec{0} & \vec{A}_1 & \vec{0} & \dots \\ \vec{0} & -\vec{I} & \vec{0} & \dots \\ \vec{L}_1 & \vec{I} & \vec{0} & \dots \\ \vec{0} & \vec{0} & \vec{A}_2 & \dots \\ \vec{0} & \vec{0} & -\vec{I} & \dots \\ \vec{L}_2 & \vec{0} & \vec{I} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vec{capInv} \\ \vec{flows}_1 \\ \vec{flows}_2 \\ \vdots \end{bmatrix} \leq \begin{bmatrix} -\vec{lbInv} \\ \vec{ubInv} \\ \vec{d}_1 \\ -\vec{lb}_1 \\ \vec{ub}_1 \\ \vec{d}_2 \\ -\vec{lb}_2 \\ \vec{ub}_2 \\ \vdots \end{bmatrix} \tag{5}$$

The objective value is (4a), which is developed in expressions (4b-4f). The first two rows in the constraints are directly related to the minimum and maximum investment allowed (4p,4r,4s).

Following that there are three rows that are repeated for each year. The first two correspond to operational constraints, such as demand balances (4g,4j,4k) or DC power flow constraints (4i). Such expressions are summarized in the matrix $\vec{A}_t$ and the vector $\vec{d}_t$. The next two lines correspond to the minimum and maximum operational flows from (4h,4l,4m). The matrix $\vec{L}_t$ accounts for the fact that only investments done prior to the operational year can account for available capacity.

Inspection of (5) reveals an underlying structure (6) that would allow the use of Benders decomposition methods (Geoffrion, 1972). This approach is not new to the field, and it is implemented in EGEAS (Bloom, 1982, 1983), a very extended investment software used in the electric industry. This special structure could be used to reduce the complexity of the linear problem and possibly use parallelization to increase solution speeds, an issue we are currently studying.

$$\textbf{min} \quad \begin{bmatrix} \vec{CostInv}^\top & \vec{CostOp_1}^\top & \vec{CostOp_2}^\top & \dots \end{bmatrix} \begin{bmatrix} flows_1 \\ \vec{flows_2} \\ \vdots \end{bmatrix}$$

subject to: (6)

$$\begin{bmatrix} \vec{C}_0 & \vec{0} & \vec{0} & \dots \\ \vec{C}_1 & \vec{D}_1 & \vec{0} & \dots \\ \vec{C}_2 & \vec{0} & \vec{D}_2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \vec{capInv} \\ \vec{flows_1} \\ \vec{flows_2} \\ \vdots \end{bmatrix} \leq \begin{bmatrix} \vec{b}_0 \\ \vec{b}_1 \\ \vec{b}_2 \\ \vdots \end{bmatrix}$$

Appendix B: Getting Started with PySP

Both PySP and the underlying Pyomo modeling language are distributed with the Coopr software package. All documentation, information, and source code related to Coopr is maintained on the following web page: https://software.sandia.gov/trac/coopr. Installation instructions are found by clicking on the *Download* tab found on the Coopr main page, or by navigating directly to https://software.sandia.gov/trac/coopr/wiki/

GettingStarted. A variety of installation options are available, including directly from the project SVN repositories or via release snapshots found on PyPi. Two Google group e-mail lists are associated with Coopr and all contained sub-projects, including PySP. The *coopr-forum* group is the main source for user assistance and release announcements. The *coopr-developers* group is the main source for discussions regarding code issues, including bug fixes and enhancements. Much of the information found at https://software.sandia.

gov/trac/coopr is mirrored on the COIN-OR web site (https://projects.coin-or.org/Coopr); similarly, a mirror of the Sandia SVN repository is maintained by COIN-OR.

Discussant Narrative

Long-Term Resource Planning for Electric Power Systems Under Uncertainty

Jason Stamp
Sandia National Laboratories

Congratulations are due to the authors. This article has one of the finest background sections I have read in a long while. There is both a good summary of problem space as well as an effective introduction to stochastic optimization.

Discussion topics include:

1. Pursuant to the risk term in the formulation introduced on page 16, can the authors expand on the sources and measures for risk as considered by likely users of the proposed planning process?

2. From for various government agencies (both federal and state), how might effective optimization under uncertainty inform policy decisions and their planned evolution over time? Particularly with regard to technology incentives?

3. What are other key stakeholder groups and their interests? How are these interests modeled in the proposed formulations? How are they at odds? Is this conflict included in the stochastic formulation, or is that even the best approach? If not, what is the proper way to represent these?

4. What are the measures and sources of regret in grid planning? How are they perceived differently by different stakeholder groups?

5. Are there examples of similar planning problems that have been described and analyzed using stochastic programming similar to those described in this paper?

6. Can the authors expand upon their treatment of regulatory uncertainty as it relates to key issues (for example, nuclear power vs. renewable energy etc.)? How does the proposed planning analysis account for ongoing technology development?

Recorder Summary

Long-Term Resource Planning for Electric Power Systems Under Uncertainty

Chao Yang
Lawrence Berkeley National Laboratory

The discussion of the paper centered around two main themes:

1. The quality of the planning model and its formulation as a mathematical program
2. Techniques for solving such a mathematical program

The discussant Jason Stamp praised the quality of the paper in terms of its comprehensive introduction of the background and its presentation of stochastic optimization. He started the discussion by raising a number of questions on various aspects of the planning model such as the sources and measures of the risks and regret in the model, the targeted decision maker, how the model may evolve over time, other stakeholders whose interests should be included in the model, as well as the treatment of regulatory uncertainty and ongoing technology development in the model.

The discussion of the first theme (regarding the modeling aspect of the problem) covered the following aspects:

- **The scope of the model**. Although most participants agreed that the paper already considered a wide range of important issues such as multi-sector planning, geographical diversity, planning horizon, there appears to be some additional factors that a model should include. One such factor is what one participant called "social hedging" capability. The participant gave an example in term of a new transmission expansion put together recently for New York City (nycedc.com). Although the original plan considered several emission and economic factors, the final plan was ultimately decided by the strength of the policy drivers rather than economic drivers. Another participant asked whether the objective function for planning should be value based, if so, what the value should it be, e.g., fuel cost or reliability? If the latter, how would one build that into the model? The presenter suggested that the objective for planning should not only include reliability but also price volatility (in economics terms) and that one may formulate the problem as a multi-objective optimization problem.

- **The complexity of the model**. The paper discussed the possibility of combined generation and transmission planning although these two planning problems are normally treated separately. One of the participants asked whether by separating the generation and transmission planning, one creates a "chicken and egg" problem that is too difficult to tackle. The presenters' view was that this type of problem may be addressed by iterating back and forth between these two decompositions of the planning problem or by taking a multi-level approach. Another participant asked whether one should add a nonlinear feedback to the model. The presenters addressed this question by pointing out that nonlinear feedback is more important for short-term planning than it is for long term planning. For long-term planning, one may want to avoid "over-designing" the planning model and making it difficult to implement.
- **The treatment of uncertainty in the model**. One of the participants pointed out that the distinctions between the low-frequency and high-frequency uncertainties described in the paper and proposed methodologies for dealing with them were quite instructive. The notion of regret is also important. However, he questioned how one may assign weights to low frequency uncertainties in the problem formulation in the context of tail-conditioned expectation (TCE), and how one defines the TCE precisely. The presenters acknowledged that, although objective functions based on TCE had been used in the finance community, its precise definition and choice of weights were usually somewhat subjective.
- **The stability of the model.** One of the participants asked whether the solution to the planning problem is stable. He commented that, drawing his experience from working with the finance community, solutions to similar problems in finance are often unstable. Whether this is an issue from the ISO safety point of view, whether the restructuring of the grid resulting from the solution to the planning problem will guarantee the stability of the emerging grid are questions that require more thoughts. The presenters acknowledged that this problem is not addressed in the current model, and the issue of stability can perhaps be better identified and addressed upfront. However, as far as planning is concerned, it appears to be difficult to determine the stability of the solution in advance. There may be models that emphasize more on the reliability instead of economic impact. But transmission planning cannot in general wait for stability studies. Another participant pointed out that the economic stability problem has been addressed, to a large degree, by the electric vehicle community. One should think about what changes need to be made to affect stability. It may require reformulation of the problem. The power system community should take advantage of prior work in other communities.
- **The validation of the model.** One of the questions raised by a participant during the discussion concerns how to update the model in a horizontally decomposed scheme that can be solved by stochastic programming techniques, and how one may validate the results and obtain trustful information. This question is also

somewhat related to the tractability of the model, and whether one can validate the result retrospectively. The presenters' view was that validating long term planning is always difficult. One can perhaps validate some of the "local pieces".

- **The targeted decision maker.** One of the participants questioned whether the objective function proposed in the paper and the presentation was intended for a centralized decision maker or decentralized decision makers. He also asked how one can determine future prices from the solution of the planning problem. In his opinion, this problem is worse than the "chicken and egg" problem, and may require nested optimization. He also commented that there appears to be a trend of increasing amount of decentralized decision making for which most of the existing models are ill-suited.
- **Comparable models in other communities.** The discussant raised a question on whether there exists planning problems similar to those in electric power system that have been described and analyzed by stochastic programming (SP) techniques. The presenter pointed out that SP has been successfully used in supply chain management (SCM) problems, which have a similar flavor to electric power system planning. However, there are some major differences between the two types of problems in terms of physical constraints. For example, hierarchical planning in a SCM model is easier because inventories can buffer uncertainty and a top-level aggregate plan is not subject to operational constraints derived from Kirchhoff's laws.

The discussion of the second theme (techniques for obtaining a solution of model problem) can be summarized by the following points:

- **The computational tractability of stochastic programming.** A number of people asked whether it is realistic to expect that problems formulated as an SP can be solved computationally, especially when the problem involves many scenarios or stages and whether one should consider other alternatives such as high dimensional dynamic programming techniques or other adaptive algorithms for solving problems that are inherently stochastic, nonlinear and non-convex. One of the possibilities suggested was some type of approximate dynamic programming.
- **Progressive hedging.** One participant had some concern over the use of progressive hedging techniques. He pointed out that progressive hedging does not seem to provide predictability, and a more robust approach that is not based on robust optimization is needed. For example, an artificial intelligence based learning algorithm (adaptive dynamic programming) that is similar to optimal power flow (for planning with many attributes) is perhaps more appropriate. He also thought that the value of reliability cannot be captured by the use of a SP which only provides a cost benefit analysis. The presenter disagreed with this view, and offered to discuss more offline.
- **The role of nonlinear programming.** One participant suggested that perhaps the reason why some of the planning problems are formulated as LPs or MILPs is

that there is a lack of good nonlinear programming (NLP) solvers. He suggested that perhaps more investments should be made in NLP solvers. His view was concurred by another participant who added that the power system community desperately needs good quality solvers that can address equilibrium problems, MCD and optimal control. To ensure there are adequate solvers in the hands of experts require continued funding for collaboration.

White Paper

Framework for Large-Scale Modeling and Simulation of Electricity Systems for Planning, Monitoring, and Secure Operations of Next-generation Electricity Grids

Jinjun Xiong, Emrah Acar, Bhavna Agrawal, Andrew R. Conn, Gary Ditlow, Peter Feldmann, Ulrich Finkler, Brian Gaucher, Anshul Gupta, Fook-Luen Heng, Jayant R Kalagnanam Ali Koc, David Kung, Dung Phan, Amith Singhee, Basil Smith
IBM Smarter Energy Research

1 Introduction

According to US DOE publication, GridVision 2030, the US Power grid, arguably the most complex machine ever built is aging, is inefficient, congested and incapable of meeting the future needs of the information economy. The current power grid was designed to handle the maximum peak capacity with graceful responses to the contingencies. However, the investment in the infrastructure has lagged far behind the increase in the energy demand, causing grids to operate much closer to their transfer limits. This new operating mode with the same standards of reliability requires significantly improved situational awareness of the grid.

One way to improve situational awareness is to create better mathematical representations of the grid operation (models and analysis) that are as close as possible to the actual grid, and use them to understand the operation of the grid. Another aspect of improved situational awareness is the integration of various instrumentation/ measurement data into the analysis. The data are usually generated from heterogeneous systems with different formats, and the amount of data is increasing exponentially with new smart devices coming on grid. Addressing these challenges will require us to confront many significant computational challenges in this area.

This RFI report will explore the computational challenges in the context of the following three areas of smart grid analysis – security-constrained unit commitment and economic dispatch with stochastic analysis, massive data set management, and large scale grid simulation.

Chapter 2 is about security-constrained unit commitment and economic dispatch. Some of the key challenges identified there are: (1) Solving real life instances of unit commitment and economic dispatch problems with thousands of generators and substations in real-time; (2) Simultaneously solving both unit commitment and economic dispatch with nonlinear transmission constraints; (3) Security-constrained unit commitment and economic dispatch with stochastic analysis; and (4) Large-scale optimization with guaranteed convergence to high quality solutions. To address these challenges, the fundamental computational problems that need to be resolved are: (1) Parallel evaluation of multiple scenarios (resulting from different contingences or loading/generation profiles); (2) Parallel execution of decomposable formulations for large-scale optimization problems; and (3) Acceleration of solving large-scale linear system of equations.

Chapter 3 discusses the massive data set management problem. Grid operators have traditionally dealt with data coming in from heterogeneous sources in multiple formats, with multiple time scales, and varying on-line availability (off- or on-line). It has also become obvious that data collection and processing requirements will only increase significantly with time, and with the installation of more PMU's and other smart devices on the grid. Some of the key computational challenges identified in this area are: (1) distributed and parallel processing of large amount of stored data (data-at-rest); (2) real time processing of high-frequency streaming data (data-in-motion); and (3) highly scalable database systems to store and query petabyte scale data efficiently. The techniques to address these challenges, like MapReduce for processing distributed massive data, stream computing for fast data-in-motion processing, and hardware accelerated data warehousing, are fairly well developed in computer science areas, but need to be applied to power grid systems.

Chapter 4 deals with the large scale grid simulation problems for power systems analysis. The key challenges identified are: (1) real-time large network power flow analysis based on AC formulations; (2) faster than real-time dynamic simulation and stability assessment to improve situational awareness; and (3) faster than real time static and dynamic state estimation. The associated key computational challenges identified in this area are: (1) large scale parallelization of nonlinear system of equation solvers; (2) large scale parallelization of nonlinear differential algebraic equation solvers; (3) fast and parallel eigen-analysis of large systems; and (4) large scale implementation of state tracking approaches.

Addressing these fundamental issues will enable real-time solutions of large-scale networks and systems and allow integration of massive data from heterogeneous sources into the analysis. These solutions will greatly enhance visibility into the grid, thus reducing the need for additional sensors and allowing faster than real-time offline analysis, experiments, and insights without impacting the operational systems.

2 Security-constrained Unit Commitment and Economic Dispatch with Stochastic Analysis

2.1 Introduction

At the heart of the future smart grid lie two related challenging optimization problems: unit commitment (UC) and economic dispatch (ED). When operational and physical constraints are considered not only under normal operating conditions, but also under contingency conditions, the UC and ED problem becomes the security-constrained UC and ED problem. We focus on UC and ED in this chapter for two obvious reasons: (1) both problems are most relevant to ISOs/RTOs daily operation as they need to be solved on a daily basis, and (2) both problems are computationally intensive tasks that require significant improvement on the performance to meet real-time operational requirements.

Although these two problems are intermingled with each other, most of the current theoretical and practical effort treats them separately, due to the computational difficulty of solving a single unified problem. As Figure 1 illustrates, present solutions to the unit commitment problem considers only a direct current (DC) approximation of the alternate current (AC) transmission constraints. This problem observes any generator-related constraints, demand constraints, and linear transmission constraints. The output of this is an optimal schedule for generators in a twenty four hour time horizon, and is given as input to the economic dispatch problem. Economic dispatch problem then handles the original AC power flow constraints and outputs a dispatch plan: how much power to produce from each generator, and how to transmit the power over the network. To account for unexpected failure of generators and transmission lines, current unit commitment practices enforce spinning reserve requirements, allocating a fraction of a generator's capacity to reserves. Similar contingency analysis is also performed in economic dispatch, making sure that the load at each node of the network can be satisfied in the case of a failure of one of the generators, transmission lines, or other devices, which is also called N-1 contingency analysis.

There are several points that current practice is missing and that need to be handled in the very near future (in about 5 years): integration of renewable energy into the grid, considering failure of more than one generator and/or transmission line (also called N-k contingency analysis), considering stochastic in the problem, such as generation costs and load profile, and being able to solve larger instances of both unit commitment and economic dispatch problems. Theoretical and practical efforts along these lines have gained momentum and will likely keep increasing.

The ultimate goal is to solve practical instance of these two problems together, considering other relevant issues such as "demand response" and "energy storage". As will be shown, this definitely requires more algorithmic advancement that uses high-performance and parallel computing environments. Today, a typical commercial integer

programming solver can handle a unit commitment problem with 100 units, 24 time periods, and 50 uncertainty scenarios. Real life instances consist of several thousand buses, more than one thousand generators, 48 to 72 time periods, more than one hundred contingencies, and a few hundreds of scenarios. Solving such a large-scale real life instance of the unit commitment and economic dispatch problem together, along with all other relevant issues, is a grand challenge that will reinforce need for fast and parallelizable decomposition algorithms.

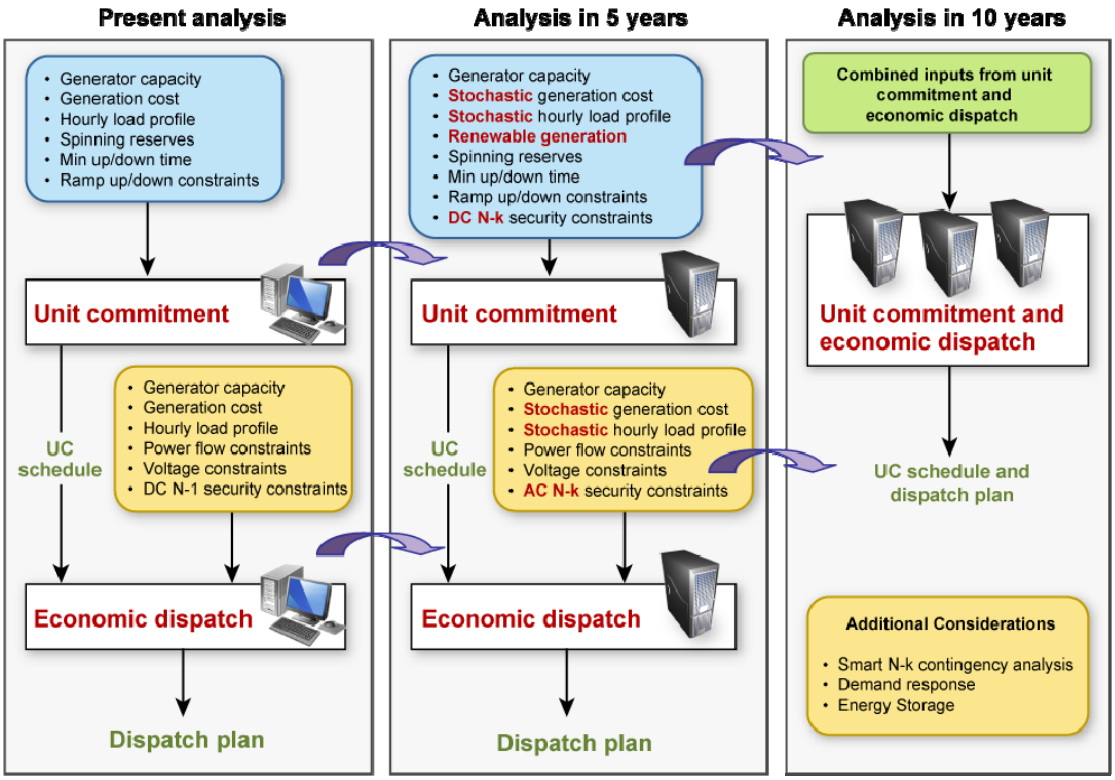


Figure 1: Unit commitment and economic dispatch: present and future

Both UC and ED can be formulated as nonlinear optimization problems (NLP) and mixed-integer NLPs that are, in general, non-convex and nonlinear. Existing industrial solutions to these two problems have been traditionally dominated by the Lagrangian relaxation methods, and only recently they have switched to general purpose integer programming solvers. Academic solutions are more diverse, but they are typically demonstrated on much smaller IEEE bus cases than the real life scenarios.

There are a number of limitations to existing solutions: (1) the inability to solve large-scale real-life power system operating problems in a given time, (2) the sub-optimality of the solutions – globally optimal solutions are seldom attained, (3) the lack of guarantee of convergence to a feasible solution, (4) limited consideration of contingency scenarios (typically focusing on N-1 but not on N-k contingencies), (5) the use of deterministic formalism for handling unpredictable loading and generation, and (6) the lack of support for real-time operation.

These limitations force current power systems operate very conservatively and maintain excessive margins in order to handle all types of un-modeled uncertainties. The future needs of the power system, such as increasing integration of renewable energy resources and growing demand for electricity, will only exacerbate this problem. These excessive margins will substantially limit the efficiency and capacity of the power grid if the limitations are not addressed properly.

To achieve significant integration of intermittent renewable energy sources and enable demand response will require that the formulations of both unit commitment and economic dispatch for system operation be enhanced. Advanced optimization techniques targeting globally optimal solutions and with guaranteed convergence need to be developed. To support real-time secure operations, N-k contingencies with renewable integration must be considered. We believe future solutions to both unit commitment and economic dispatch problems need to be implemented in a hybrid-computing environment that supports:

1. Parallel evaluation of multiple scenarios (resulting from different contingences or loading/generation profiles)
2. Parallel execution of decomposable formulations for large-scale optimization problems
3. Large-scale optimization with guaranteed convergence to high quality solutions.

In this context, the following references are deemed to be important. Reference [1] is the first paper that models the unit commitment (UC) problem as a mathematical program. References [2] and [3] are two excellent reviews on the UC problem. Two important techniques, dynamic programming [4] and Lagrangian relaxation [5], are known in the field to solve the UC problem. The earliest reference on economic dispatch (ED) is [6] in the 1960s. Some interesting references on solution techniques for ED are [7] where it presents an implementation of the interior-point method to solve the ED problem, and [8] where it discusses an approach on how to convexify the non-convex ED problem. One of the earliest formulations on security-constrained economic dispatch (SCED) is [9] in 1970s. An algorithm based on contingency filtering techniques was discussed in [10] to solve SCED by identifying and adding only those potentially binding contingencies into the formulation. A parallel solution of SCED using a nonlinear interior-point method was shown in [11]. Reference [12] should help readers to understand the stochastic economic dispatch problem formulation.

In the rest of this chapter, we first survey results on state-of-the-art solutions to unit commitment in Section 2.2, then address economic dispatch in Section 2.3, and security-constrained economic dispatch in Section 2.4. Because of their close relevance to optimization, we also conduct a survey on the general optimization methods and their latest advancement, which is presented in Section 2.5. Since solving systems of linear equations is at the core of most optimization engines, we review the techniques for this

in Section 2.6. In Sections 2.5 and 2.6, we emphasize solutions that are scalable and amenable to parallelization. We end the chapter with an outline for a number of future research directions in Section 2.7.

2.2 Unit Commitment Problem

UC is the problem of finding an optimal ramp up and down schedule and corresponding generation amounts for a set of generators over a planning horizon so that the total cost of generation and transmission is minimized, and a set of constraints, such as demand requirement, upper and lower limits of generation, minimum up/down time limits, ramp up/down constraints, transmission constraints, and so forth, are observed. We refer to [13] for the definition of "unit commitment."

Unit commitment (UC) is at the core of planning and operational decisions faced by independent system operators (ISOs), regional transmission organizations (RTOs), and utility companies. Hence it has received a good deal of attention in the industry.

The academic literature on the unit commitment problem dates back to the 1960s. Enumerating all possible combinations of generators and choosing the one that meets all the system constraints and that has the least cost that may solve the problem. Kerr et al. [14] is one of the first studies that reduced the unit commitment problem to mathematical terms. An integer programming approach proposed by Dillon et al. [1] was one of the earliest optimization-based approaches to the unit commitment problem. The authors address the unit commitment problem of hydro-thermal systems with reserve requirements. It was one of the earliest papers that could solve real life problems with 20 units. The authors developed two sets of valid inequalities that were globally valid to the problem and used these inequalities in the branch-and-bound algorithm. The dynamic programming approach developed by Snyder et al. [4] was one of the earliest successful dynamic programming algorithms. The algorithm featured a classification of units so as to reduce the number of states. The authors addressed the problem at San Diego Gas & Electric System with 30 generators. Birge and Takriti [16] realized that the demand constraint in the problem coupled all the units in the problem, which would otherwise decompose into separate small problems for each unit. The authors developed a Lagrangian relaxation algorithm based on relaxing the demand constraint. They also proposed a refining procedure that took Lagrangian solutions and produced feasible solutions to the problem. Expert systems [17], fuzzy logic [18], meta-heuristic algorithms [19], and ant colony systems [20] are among the other approaches that have been applied to the unit commitment problem.

Surveys by Sheble and Fahd [3] and by Padhy [2] review the academic literature on the unit commitment problem, and the book by Wood and Wollenberg [21] addresses several operational and planning problems in the energy industry, including the unit commitment problem.

In practice, unit commitment is solved either in a centralized or decentralized manner [13]. In western and southern regions, most small utility companies generate and distribute their power in a decentralized manner, with California ISO (CAISO) in the west and ERCOT in the south as two exceptions. In the Midwest, PJM and MISO, and in Northeast, NYISO, and ISO-NE handle most of the power system planning and transmission in a centralized manner. All these organizations work in a centralized manner to solve their unit commitment and economic dispatch problems [13] [22] [23] [24].

Specifically, unit commitment practices in PJM involve more than 600 utility companies in 13 states [23]. It manages 1210 generators with a total capacity of 164,905 MW, and it faces a peak load of 144,644 MW. PJM takes around 50,000 hourly demand bids from consumers one day ahead of the planning horizon. It then solves the day-ahead unit commitment problem considering around 10,000 contingencies, and publishes generation schedules for the companies and locational marginal prices (LMPs) for the consumers. Consumers revise their bids based on the prices, and submit their final bids. Around 10,000 demand bids are submitted, and 8,700 of them are considered eligible. PJM updates the generation schedules and redoes its security analysis based on the final bids. In real-time, PJM solves a one-period security-constrained unit commitment problem using full a transmission model and turns on (off) some peaker units if the total planned amount that can be transmitted securely falls behind (exceeds) the demand.

ISOs have used Lagrangian relaxation to solve their unit commitment problems until recently, and switched to general-purpose integer programming solvers [22][23][24]. For example, IBM ILOG CPLEX Mixed Integer Optimizer is one of the few solvers that are able to handle long planning horizons for large power systems.

2.3 Economic Dispatch Problem

Economic dispatch is the problem of determining the most efficient, low-cost and reliable operation of a power system by dispatching the available electricity generation resources to the load on the system. The primary objective of economic dispatch is to minimize the total cost of generation while satisfying the physical constraints and operational limits. We refer to [25] for definition of “economic dispatch”.

The economic dispatch problem plays an important role in power system analysis [21][26], especially for planning, operation and control of power systems. Carpentier originally introduced it in the early 1960s [6].

The economic dispatch problem is formulated as a nonlinear programming problem with the primary objective to minimize the total generation cost. Other forms of objective functions can be considered as well, such as the reactive power loss on transmission lines and the total system active power losses. Conventional constraints include either the DC or AC power flow equations, physical limits of the control

variables, physical limits of the state variables and other limits such as power factor limits [27]. Such constraints act as steady-state constraints without considering system contingencies.

The economic dispatch problem becomes a mixed-integer nonlinear programming problem when discrete control variables such as transformer taps, shunt capacitor banks, and flexible AC transmission system (FACTS) devices are taken into account [28]. Furthermore, if transient stability constraints are considered as the dynamic stability conditions, the problem will consist of a set of large-scale differential-algebraic equations [29][30].

Economic dispatch is a non-convex, nonlinear optimization problem, which is in general difficult to solve. Some efforts to convexify the problem have been considered by various authors. In [31], the author shows that the load flow problem of a radial distribution system can be modeled as a convex optimization problem in the form of a conic program. In a meshed network, nevertheless, the convexity cannot be derived, and the problem is then formulated as an extended conic quadratic format [32]. Recently, Lavaei and Low [8] proposed to use semi-definite programming relaxations for the economic dispatch problem. They introduce a sufficient condition that is satisfied by some specific network systems after the data sets are slightly perturbed, hence it guarantees that semi-definite programming can solve the problem. Motivated by distributed multi-processor environments, Kim and Baldick [122] presented some decomposition algorithms based on the auxiliary problem principle, the proximal point method and the alternating direction method of multipliers.

Because of the non-convexity of the problem, most solution approaches focus on the development of fast and robust local algorithms for the economic dispatch problem. In other words, they aim at finding a local optimum to the problem that satisfies the first-order optimality conditions. Popular numerical techniques include successive linear/ quadratic programming, trust-region based methods, Lagrangian-Newton methods and interior-point methods.

Successive linear/quadratic programming

This approach approximates the objective function by a linear or quadratic function whereas the constraints are successively linearized. Wells [33] utilized this idea to derive a sequence of linear programming to solve the economic dispatch problem with security constraints. Contaxis et al. [34] decomposed the economic dispatch problem into real and reactive subproblems. Then these two subproblems are solved using quadratic programming in each iteration. The algorithm by Amerongen [35] is obtained by rigorously linearizing the necessary Karush-Kuhn-Tucker conditions, and then transforming them into a sequence of related quadratic programming problems. The sparsity structure of the problem was exploited to speed up the computations by using explicit reduction of some of the variables. One basic disadvantage of these methods is

the poor computational results and convergence rate. They often fail to handle very large-scale problems.

Trust-region based methods

Trust-region methods belong to a class of optimization algorithms that minimize the approximation (typically a quadratic) of the objective function within a closed region, called the trust region. Various methods differ in the way to choose the trust region. For example, Min et al. [36] proposed a trust-region interior-point algorithm. There are two types of iterations in the algorithm: the main iteration and linear programming (LP) inner-iteration. The economic dispatch problem is linearized to form a trust-region subproblem in the main iteration. In the LP inner-iteration, the trust-region LP subproblem is solved by the multiple centrality corrections primal dual interior-point method. The trust region controls the linear step size. It was shown to be superior to successive linear programming methods. The method of Sausa et al. [15][37] is based on an infinity-norm trust region approach, using interior-point methods to solve the trust-region subproblems. The convergence robustness was tested and verified by using different starting points.

Lagrangian-Newton methods

Sun et al. [38] proposed one of earliest Lagrangian-Newton method approaches, which solves the Lagrangian function by minimizing a quadratic approximation. Santos and Costa [39] studied an augmented Lagrangian method that incorporates all the equality and inequality constraints into the objective function. The first-order optimality conditions are achieved by Newton's method together with a multiplier update technique. Baptista et al. [40] considered the application of logarithmic barrier method to voltage magnitude and tap-changing transformer variables and the other constraints are treated by an augmented Lagrangian method. In [41], Wang et al. treat inequality constraints in the context of the quadratic penalty method by using squared additional variables to form a sequence of unconstrained optimization problems. A trust-region method based on a 2-norm subproblem was used to solve the unconstrained problems.

Interior-point methods

A number of efficient methods based on the Karush-Kuhn-Tucker necessary conditions are interior-point methods [7][42][43][44][41]. Interior-point methods (IPMs) have been widely applied to solve the economic dispatch problem in the last decade because of its well-known excellent properties for large-scale optimization problems. In particular, its local quadratic convergence is established, and a class of polynomial-time algorithms has been designed. The methods transform the inequality constraints into equality constraints by introducing nonnegative slack variables, then, typically, treat the slack variables via a logarithmic barrier term. The main idea of these algorithms is to combine three relatively standard powerful techniques: logarithmic barrier function to handle inequality constraints, Lagrange theory of optimization subject to equality constraints,

and Newton's method for a system of equations. In addition, an implementation of the automatic differentiation technique for the economic dispatch problem is proposed in the work [45]. The interior-point based algorithm, in our opinion, is likely to be one of the most practical approaches designed for the very large-scale, real-world electric networks.

2.4 Security-constrained Economic Dispatch

In this section, we investigate some mathematical formulations for the security-constrained economic dispatch (SCED) problem and its deterministic solution methods in the literature. Unlike classical economic dispatch problems, the SCED problems take into account both the pre-contingency (base-case) constraints and post-contingency constraints. Our review focuses on the widely used N-1 contingencies, for example, even if one component (such as a line transmission, generator, or transformer) is out of service, the power system should still satisfy the load requirements without any operating violations.

The first type of SCED formulation is the preventive SCED [9], i.e., it minimizes some generation cost function by acting only on the base-case (such as, contingency-free) control variables subject to both the normal and abnormal (with one of the N-1 contingencies) operating constraints. For k contingency scenarios, the problem size of the preventive SCED is roughly k + 1 times larger than the classical (base-case) economic dispatch problem. Solving this problem directly for large-scale power systems with numerous contingencies would lead to prohibitive memory requirements and execution times.

In real-world applications, however, many post-contingency constraints are redundant, that is, their absence does not affect the optimal value [10]. Consequently, a class of algorithms based on contingency filtering techniques has been developed [9][47][48][49][50][10] that identify and only add those potentially binding contingencies into the formulation. For example, the contingency ranking schemes from [48] are achieved by investigating a relaxed preventive SCED problem, where a single contingency along with the base-case is considered one at a time. The ranking methods rely on the information of Lagrangian multipliers or the decrease factor of penalized objective function values, and then select contingencies with a severity index above some threshold for further consideration. Other contingency filtering methodologies [10] aim to efficiently identify a minimal subset of contingencies to be added based upon the comparison of post-contingency violations. In addition, an approach using the generalized Benders decomposition to construct the feasibility cut from the Lagrangian multiplier vector of constraints is introduced in [51]. It showed a significant speedup in terms of computation time. However, the drawback of applying the decomposition technique from convex optimization to the highly non-convex problem is the lack of mathematical convergence analysis. There is no established theoretical guarantee that it can provide a local minimizer.

The second type of SCED problem is called the corrective SCED, with the assumption that post-contingency constraint violations can be endured up to several minutes without damaging the equipment [50]. The corrective SCED allows post-contingency control variables to be rescheduled, so that it is easier to eliminate violations of contingency constraints than the preventive SCED. The optimal value of corrective SCED is often smaller than that of preventive SCED, but its solution is often harder to obtain, since it introduces additional decision variables and nonlinear constraints. Monticelli et al. [50] tackled the optimization problem by rewriting it in terms of only the contingency-free state variables and control variables, while constraint reductions are represented as implicit functions of these contingency-free state and control variables, which in turn are related to the infeasibility post-contingency operating subproblems. The solution algorithm then becomes an application of the generalized Benders decomposition [52] that iteratively solves a base-case economic dispatch and separate contingency analysis. Moreover, an extension of a contingency filtering technique from [10] was studied in [53], which features an additional optimal power flow module to verify the controllability of post-contingency states.

The third type of SCED problem is an improvement of the aforementioned corrective SCED. Capitanescu et al. [120] recognized that the system could face voltage collapse and/or cascading overload right after a contingency and before corrective action is taken. Therefore, their improved formulation imposes existence and viability constraints on the short-term equilibrium reached just after contingency occurrence and before corrective controls are applied. There are very few solutions to this formulation, but one exception is Yi [54] that utilized the Benders decomposition method to solve this problem.

The inclusion of contingencies beyond N-1 for future power grid operation may increase the complexity and scale of the problem by several orders of magnitude. Therefore, great effort has been devoted to the development of parallel algorithms for large-scale problem formulations. In this case the SCED problem is decomposed and distributed on a number of processors with each one handling independently a subset of the post-contingency analysis.

There are currently two promising approaches for parallelism: one related to interior-point methods [11] and one using Benders decomposition [56][54][57]. For interior-point methods, at each primal-dual iteration, we need to solve a large-scale system of linear equations. Because the matrix associated with these linear equations has a blocked diagonal bordered structure, by exploiting this fact, researchers have shown that the system of linear equations can be solved efficiently in parallel. For example, more than 10 times speedup can be obtained on a system with 16 processors [51]. On the other hand, Benders decomposition is a two-stage solution method consisting of a base-case problem and a list of contingency subproblems. Since the evaluation of different contingencies can be done independently, this formulation is amenable for parallelism.

One obvious benefit of exploiting parallelism in solving these problems is that it makes the complexity linearly dependent on the size of the problem as opposed to the quadratic growth for sequential computation [54]. All the above aforementioned algorithms, however, have appeared in the form of academic papers. Their practical use has not been validated.

Another challenge to the existing SCED formulation comes from the emergence of deregulated electricity markets, where the market-clearing process, pricing mechanism, for electricity trading should be included as part of the solution process. New formulations and algorithms need to be developed to address this challenge, and the solution should guarantee a satisfactory worst-case performance to meet the real-time dispatching requirements.

2.5 Optimization Methods in General

Real world optimization problems are often large and nonlinear. There has been significant progress in nonlinear optimization algorithms and software in recent years. Computing and algorithmic advances have made it possible for researchers and practitioners to develop tools to solve very large-scale optimization problems. We review some of those developments that may be of use toward the development of next generation power system planning and operational aspects.

One of the interesting developments in optimization is the use of algebraic modeling languages, such as AIMMS, AMPL, GAMS, LPL, Mathematica, MATLAB, MPL and OPL, to describe mathematic optimization problems of high complexity. Because the algebraic modeling languages have syntax similar to the mathematical notation of optimization problems, a large-scale optimization problem can be easily and concisely described. For example, some of these languages, such as AMPL and GAMS, even provide automatic differentiation capabilities for first and second derivatives.

Another interesting development is the wide acceptance of interior-point methods to solve various optimization problems. Since its inception in 1984 by Karmarkar [58] for solving linear programming problems, the interior-point method has gained significant attention. It was claimed [59] that the interior-point method is about 50 times faster than the most powerful simplex method, and has achieved excellent performance for large-scale problems. The interior-point methods have begun to challenge the existing status quo optimization techniques that are dominant in the power system analysis packages and traditionally known for their performance advantages, such as sequential quadratic programming and augmented Lagrangian methods. One of the best implementations of interior-point methods using line searches based on filter methods [60][61] for nonlinear optimization is IPOPT [62], which can handle a problem with millions of variables.

A rather new development in the optimization field is semi-definite programming (SDP) with guaranteed global optimality. SDP is a special case of convex optimization, and has

gained attention for its polynomial-time solvability [63] to achieve the globally optimal solution. Many practical problems in operations research and combinatorial optimization can be modeled or approximated as SDP problems. We can view SDP as a generalization of the well-known linear programming with its variables forming an n×n symmetric matrix as compared to an n×1 vector. In SDP, we try to minimize an affine function of this symmetric matrix subject to both linear constraints and semi-definiteness constraints. Most of robust and efficient methods are based on interior-point methods, including CSDP [64], SeDuMi [65], SDPT3 [66], DSDP [67], SDPA [68]. One exception is the ConicBundle method that implements the bundle method of Helmberg and Rendl [69], Helmberg and Kiwiel [70] for minimizing the sum of convex functions that are given by first-order oracles or arise from Lagrangian relaxation of particular conic linear programs.

We now survey some literature that might help the reader to approach the state-of-the-art optimization techniques. A relatively recent text that is well-written and accessible to the non-expert is [71]. For recent techniques on successive linear/quadratic programming and successive quadratic programming, the reader is referred to [46][72][60][73][74]. For trust-region methods, there is the encyclopedic text [75] for which some chapters are accessible to the non-expert. As an alternative, Chapter 4 in [71] is also very readable. For augmented Lagrangian and barrier methods, see for example [76][77]. The augmented Lagrangian methods solve the optimization problem by a sequence of subproblems that minimize the augmented Lagrangian. The two most popular packages, ALGENCAN [78][79] and LANCELOT [80], are based on the different ways to solve the subproblems: either a quasi-Newton method or a trust-region method. In addition, a comprehensive review of software for nonlinearly constrained optimization can be found in [81], and benchmarks for a wide range of optimization software are accessible in [82]. Recent references for mixed-integer nonlinear programming are [121][83][5].

For a given problem, many important issues may affect which types of optimization algorithms are the most appropriate. For example, if derivatives are not directly available, one has no choice but to use a derivative-free method to approximate derivatives via automatic differentiation or function differencing. In this case, one typically cannot expect to solve problems on a serial machine with more than about 200 variables. A recent text on derivative-free methods that is intended to be both comprehensive and readily accessible to the non-expert is [84]. An implementation of an interior-point method that uses automatic differentiation [85] in the context of the optimal power flow problem is given in [45].

The choice of optimization method is important if one wants to solve larger problems in a parallel environment. What one can say is that simulated annealing [86], genetic algorithms [87], or the popular Nelder-Meade method [88] (see, for example [89]) are rarely the best ones to use, although they, as well as similar methods such as particle

swarm, are often implemented [90][91] to solve the power system optimization problems.

If significant noise is present in the formulation, it also affects the choice of the optimization algorithm. For example, in this case, using finite difference gradients is not a good idea.

2.6 Solving Linear System of Equations in General

Since solving systems of linear equations is at the core of most optimization engine and analysis tools (such as power flow computation), we review it briefly with the emphasis on scalability and parallelization.

Direct methods for solving linear systems of the form Ax = b are based on computing A = LU, where L and U are lower and upper triangular matrices, respectively. Computing the triangular factors of the coefficient matrix A is also known as LU decomposition.

When A is sparse, the triangular factors L and U typically have nonzero entries in many more locations than A does. This phenomenon is known as fill-in, and results in a superlinear growth in the memory and time requirements of a direct method to solve a sparse system with respect to the size of the system. Despite a high memory requirement, direct methods are often used in many real applications due to their generality and robustness. In applications requiring solutions with respect to several right-hand side vectors and the same coefficient matrix, direct methods are often the solvers of choice because the one-time cost of factorization can be amortized over several inexpensive triangular solves.

Books by George and Liu [92] and Duff, Erisman and Reid [93] are excellent sources for a background on sparse direct methods. A comprehensive survey by Demmel, Heath and van der Vorst [94] sums up the developments in parallel sparse direct solvers until the early 1990s. Some remarkable progress was made in the development of parallel algorithms and software for sparse direct methods during a decade starting in the early 1990s. Gupta et al. [95][96] developed the framework for highly scalable parallel formulations of symmetric sparse factorization based on the multifrontal method (see tutorial by Liu [97] for details), and recently demonstrated scalable performance of an industrial strength implementation of their algorithms on thousands of cores [98]. Demmel, Gilbert and Li [99] developed one of the first scalable algorithms and software for solving unsymmetric sparse systems without partial pivoting. Amestoy et al. [100][101] developed parallel algorithms and software that incorporated partial pivoting for solving unsymmetric systems with (either natural or forced) symmetric pattern. Hadfield [55] and Gupta [102] laid the theoretical foundation for a general unsymmetric pattern parallel multifrontal algorithm with partial pivoting, with the latter following up with a practical implementation [103].

Iterative methods for solving sparse systems of linear equations are potentially less memory and computation intensive than direct methods, but often experience slow convergence or fail to converge. The robustness and the speed of Krylov subspace iterative methods are often dramatically improved by preconditioning.

Preconditioning is a technique for transforming the original system of equations into one with an improved distribution (clustering) of eigenvalues so that the transformed system can be solved in fewer iterations. A key step in preconditioning a linear system Ax = b is to find a nonsingular, preconditioner matrix M, such that the inverse of M is as close to the inverse of A as possible and solving My = b, which is significantly less expensive than solving Ax = b. Other practical requirements for successful preconditioning are that the cost of computing M itself must be low and the memory required to compute and apply M must be significantly less than that for solving Ax = b via direct factorization.

Most modern preconditoners are based on (i) stationary methods, (ii) incomplete factorization, (iii) sparse approximate inverse, (iv) geometric/algebraic multigrid, or (v) the physics of the underlying problem.

The readers are referred to Saad's book [104] and the survey by Benzi [105] for a fairly comprehensive introduction to various preconditioning techniques. These do not cover parallel preconditioners and may not include some of the most recent work in preconditioning. However, these are excellent resources for gaining an insight into the state-of-the-art of preconditioning circa 2002.

Hysom and Pothen [106] and Karypis and Kumar [107] cover the fundamentals of scalable parallelization of incomplete factorization based preconditioners. The work of Grote and Huckle [108] and Edmond Chow [109][110] is the basis of modern parallel sparse approximate inverse preconditioners. Chow et al.'s survey [111] should give the readers a good overview of parallelization techniques for geometric and algebraic multigrid methods.

Last, but not the least, almost all parallel preconditioning techniques relies on effective parallel heuristics for two critical combinatorial problems - graph partitioning and graph coloring. The readers are referred to papers by Karypis and Kumar [112][113] and Bozdag et al. [114] for an overview of these.

2.7 Future Research

Many of the grid system changes are inherently stochastic in nature, and they cannot be understood through the deterministic approaches currently in use. Future U.S. power grid operation should incorporate stochastic modeling of both generation and loading and comprehensive contingency analysis beyond existing N-1 practices in the formulation of both unit commitment and economic dispatch problems. Extremely fast,

security-constrained unit commitment and economic dispatch with stochastic analysis with integrated real-time N-k contingency analysis will be a critical capability for the evolving smart grid.

Existing power system optimization problems (such as UC and SCED) are mostly based on DC power flow formulations because of the difficulties related to AC power flow computation, including poor convergence rate and non-robust solution quality. Since the AC-based formulation captures the physical power flow more realistically than its DC counterpart, it is desirable to incorporate AC-based formulation (that is, nonlinear models) into existing power system optimization problems.

We point out a number of important avenues of research that will receive noteworthy attention in the coming decade.

Solving real life instances of unit commitment and economic dispatch problems with thousands of generators and substations in real-time

Current practice is that many ISOs use general purpose integer programming solvers for their unit commitment problems, and these solvers can handle up to around 100 generators with around 72 time periods. But in reality, ISOs need to deal with thousands of generators, and probably even more in the future as more distributed generation sources come on line. Hence it is very critical to be able to solve large-scale unit commitment and economic dispatch problems in real-time. In this respect, approximation algorithms that guarantee finding near optimal solutions and decomposition algorithms that are scalable will be among the most promising techniques. Algorithms that lend themselves to parallelization on distributed or shared memory computing environments will gain importance. In addition, it is also crucial to develop algorithms that are robust with fast convergence rate to achieve high quality solutions for the nonlinear AC-based UC/SCED problems.

Simultaneously solving both unit commitment and economic dispatch with nonlinear transmission constraints

In many real life practices, linear approximations of transmission constraints (that is, DC-based power flow constraints) are used in the unit commitment problem [33]. Also, there is some research on handling nonlinear transmission constraints of power flow (such as, AC-based power flow constraints) in economic dispatch, independent of the unit commitment decisions [26]. Ideally, we would like to solve the unit commitment problem together with the AC-based economic dispatch along with the original nonlinear transmission constraints to obtain more realistic results.

The problem can be formulated as a mixed-integer nonlinear programming (MINLP) problem, a subclass of both mixed-integer programming (MIP) and nonlinear programming. The state-of-the-art MIP solvers are now able to handle problem sizes up

to hundreds of integer variables efficiently. But MINLP has yet to reach that level of maturity. For the future power system applications, we should exploit the problem structure and leverage the recent advancements in solving large-scale NLP problems so that we can achieve good quality solutions to MINLP at least as good as MIP. In particular, it will be an important area of research to develop decomposition algorithms that handle transmission constraints and unit commitment decisions separately, and combine them in a way that commitment and transmission decisions optimize a centralized objective function. This will exploit the algorithms developed separately for the unit commitment and AC-based economic dispatch problems. Note that the presence of discrete control variables, such as transformer taps, shunt capacitor banks, and FACTS devices, also makes even the ED problem a mixed-integer nonlinear programming problem, which will also benefit from above research.

Security-constrained unit commitment and economic dispatch with stochastic analysis

In reality, many problems should be treated stochastically [115][12][91], which severely complicates the optimization. At the very least, it makes the size of any deterministic reformulation much larger. But most of these problems are highly structured. To develop a good solver, one should try to exploit such structures from both optimization and linear algebra perspectives. This has been done to a limited extent in [116] using [117][52].

Accounting for the contingencies of operations, such as generator and transmission outages, is critical in maintaining a reliable and secure power system. These contingencies can be addressed in two ways: First is a robust system design where its probability of failure is minimized. The second one is to overdesign the system where some excess capacities in the generators and transmission lines are designed in so that in the event of an outage, the system can meet the demand using its excess reserves [124]. This results in an optimization problem determining how much of the capacity to allocate to reserves so that the probability of failure is kept under control and the total cost of the system operation is minimized.

Countries around the globe have been encouraging and continue to encourage the integration of renewable energy into their power systems [25]. However, renewable energy sources, such as solar and wind power, are highly intermittent and unpredictable. For instance, there are several issues with the wind power: wind speed ramps up and down very quickly leading to fluctuations in the wind power, wind power can be generated only when the wind speed is between a lower and upper limit, and it is very difficult to provide an accurate day-ahead forecast for the wind speed. Similar problems, although less severe, exist with solar and other renewable energy sources. Thus, it is very critical to be able to plan the unit commitment decisions and economic dispatch considering these uncertainties.

There are two ways to incorporate intermittent renewables into the market. One approach is to assume that all of the renewable energy has to be used to meet the demand. This leads to a unit commitment problem with stochastic demand values, which requires stochastic programming [5] and robust optimization [123] techniques. In addition, stochastic programming requires building a representative probability distribution for the renewable energy source. Another approach is to buffer against these fluctuations by storing energy. This leads to a problem similar to the production planning problems with inventory decisions.

We envision stochastic analysis will be a key enabler to account for the stochastic nature of operating environments, including meteorology, fuel prices, load requirements, evolving demand response, distributed and intermittent energy generation, unit forced outages, line and transmission component outages, and virtual bids. An effective, next-generation UC/SCED tool should be developed to enable real-time multiple-scenario analysis and what-if evaluation, ensure the consistency of longer forecast horizons with day-ahead markets to anticipate strong dynamic variations of supply and demand (wind ramps, for example). It should also enable convergence of day-ahead and real-time prices, include market based mechanisms of trading, virtual demand, and supply bids in the dispatch formulation, account for detailed transmission congestion constraints and costs, and enable efficient and adaptive operational reserve management.

These capabilities will become even more critical when large amounts of renewable sources and demand response programs are integrated into the market, and thus the loads are more elastic, dynamic, and uncertain. These capabilities will also increase the computational complexity significantly. Recent advances in high performance computing and mathematics should be leveraged to provide these capabilities.

Large-scale optimization with guaranteed convergence to high quality solutions

Most theoretical results with respect to optimization algorithms are oriented toward the asymptotics, whereas in practice the engineers rarely have the luxury of operating that domain. In addition, the practitioners are always interested in determining the global optimum, whereas, except in special cases, such solutions cannot be algorithmically obtained in a reasonable length of time, if at all.

Fortunately, there is anecdotal evidence that the best asymptotic algorithms are also the best in practice. In some important cases, there is also evidence that global optima can be found (see for example [8]). But even in more general cases, if one is able to determine better solutions (approximations to local optima) than those that are already known, the results can be considered successful.

Both unit commitment and economic dispatch problems are formulated as non-convex nonlinear optimization problems. As mentioned in the previous section, popular solution approaches are devoted to finding a local optimum to the problem, which include successive linear/quadratic programming and successive quadratic programming [34][73][35], trust-region based methods [36][37] and augmented Lagrangian and barrier methods [39][40]. Other interior-point methods for optimal power flow include [7][118][42][119][43].

Because of the special structure of the SCED formulations, decomposition techniques could be very powerful when combined with parallel algorithms development. But there are a number of challenges that need to be addressed. First, theoretically, Benders decomposition cannot guarantee either convergence or optimality for many cases because of the non-convexity arising from the AC power flow equations of the problem. The duality theory has a gap when the algorithm is applied to subproblems. As a result, the algorithm may generate a sequence that does not converge either to a global or local optimum. Second, the iterative process requires us to solve globally, a number of non-convex and nonlinear contingency subproblems, which is a very difficult task. Third, to the best of our knowledge, no existing mathematical theory proves the existence of a feasible solution from the estimated sequence, even in the limit.

Unfortunately, as is typical, many a time engineers use algorithms that do not represent state-of-the-art optimization techniques since it is not their field of specialization. Therefore, there is a need to bring together the best of the mathematicians and engineers to solve the realistic problems. That is not to say that the state-of-the-art optimization methods are guaranteed to converge for such problems.

In conclusion, unit commitment and economic dispatch are critical for secure power grid operations and one of their main objectives is to maximize market efficiency. Today's unit commitment and economic dispatch analysis, however, cannot account for stochastic phenomena that affect real power grid system behaviors. This limitation comes largely from the lack of sophisticated software tools that can take advantage of the latest developments in mathematics and high-performance computing. It will be necessary to develop a hybrid-computing framework and software tools that can utilize new algorithms and mathematics to address these challenges efficiently.

3 Massive Data Processing

3.1 Introduction

Electricity grid operators have traditionally had to contend with large amounts of data, from sources such as SCADA systems, disturbance monitoring equipment, outage logs, weather logs, and meter readings, to deduce useful information about the condition of the grid that would help them to better understand the grid and make correct decisions.

Future electricity grids will, however, take the volume of data up by orders of magnitude both in terms of size and frequency of measurement as high resolution data sources are integrated into the system. Some of the key sources of this future massive data are:

1. Synchrophasor data from increased phasor measurement unit (PMU) deployment [1]
2. Energy user data from the millions of smart meters [2]
3. High-rate digital data from increasingly networked digital fault recorders (DFR)
4. Fine-grained weather data [3]
5. Energy market pricing signals and bid information [4]

This high resolution data on grid, environmental and market conditions has the potential to tremendously improve our understanding of the interaction of the grid with its operating environment and our situational awareness during grid operation, enabling much closer-to-optimal planning and operation of the grid.

One key link required to realize this vision is the ability to process this large amount of heterogeneous data, both off-line and in real-time, to extract useful information from it; information such as trends and patterns in massive historical data, anomalies in grid behavior, the likelihood of imminent disturbances and models for load forecasting.

The critical need for such capability was highlighted by the investigation into the causes of the August 14, 2003 North America blackout. Much of the investigation relied on historical data recorded before and during the events of the blackout and the resulting reports highlight the need to collect and use standardized and detailed data from the grid both for real-time system reliability awareness and for post-blackout forensic analysis [5]. Recent years have seen numerous other investigations and research that have further underscored this need and accelerated the development and deployment of technologies that can collect such data [6].

A scalable high performance computational infrastructure that can handle these high volumes of data and efficiently perform the data analytic tasks required for extracting relevant information from the data, must satisfy special processing requirements of these massive data analytic tasks. In general, we foresee three kinds of data processing needs:

1. Distributed and parallel processing of large amounts of stored historical data ("data at rest"). An example of this type of processing would be post-event diagnosis.
2. Real-time low-latency processing of streaming data, or stream computing ("data-in-motion"). An example of this type of processing would be in a wide-area monitoring system (WAMS) where high-frequency real-time PMU and other grid data are continuously processed to compute system reliability indicators.

3. Along with these, there will also be a need for highly scalable database systems that can quickly perform a variety of queries on petabyte-scale data.

Traditional data systems from computer science that use conventional database systems along with analysis tools such as spreadsheets and statistical analysis packages are designed with a strong focus on massive data storage and accessibility, but not for large scale parallel and distributed data processing or high-frequency data stream processing. Fortunately, researchers in the field of computer science have been grappling with these issues for the last couple of decades and much progress has been made in developing scalable computational frameworks that address these needs.

An example for distributed processing of massive data-at-rest is MapReduce [7]. It is particularly attractive because of its ease of development and deployment, and the availability of the open source implementations [8][9]. For data-in-motion, there are several implementations of stream computing, including the open-source S4 [10]. A third relevant technology is hardware-accelerated data warehousing [11][12], which typically exploits field-programmable gate array (FPGA) based special-purpose processors or graphics processing units (GPUs) to implement extremely fast database queries.

All these technologies are at a stage that they can be effectively adapted and applied to the massive data processing tasks required by smart grid applications. There are excellent reviews such as [13] that describe the communications and computational power requirements for a smart grid. However, they stop short of exploring the needs and technologies for the massive scale data analytics piece of a smart grid. The focus of this chapter will be to discuss these applications and technologies, highlighting relevance of the latter to the former, and outline what an effective massive data management system for the smart grid may look like.

The rest of this chapter is organized as follows. Section 3.2 will discuss in detail the massive data applications and computational tasks that will be relevant to a smart grid. Section 3.3 will describe the recently developed computational technologies that can be applied to address the needs of these massive data applications. Section 3.4 will discuss the complete massive data management infrastructure for electricity grids using an illustrative example and Section 3.5 will offer concluding remarks.

3.2 Applications and Algorithms

Two key enablers of a responsive, resilient and self-healing smart grid are wide-area monitoring system (WAMS) and situational awareness. Wide-area monitoring would require management and processing of detailed monitor data from a large number of data source spread out across the grid geography, substantially increasing the volume of data from today. Situational awareness would require management and near-instantaneous processing of data streaming in from the WAMS, substantially increasing

the rate of data processing from today. An illustrating example for synchrophasor data is available from [1]: 100 PMUs sampling at 60 samples per second will generate 4,170 MB/s of data and the tolerable latency of transporting and processing the data for real-time situational awareness is in the order of only 10s-100s of milliseconds. Furthermore, the PMU data collected by the Tennessee Valley Authority from 90 PMUs over 2009 amounted to roughly 11 terabytes (TBs). With projections of increasing PMU sampling rate and the number of PMUs deployed [14], just the PMU data across the North American ISOs can amount to 100s of terabytes to a petabyte (PB) over a year [1]. Similarly large volumes of data are expected from smart meters in the Advanced Metering Infrastructure. This is orders of magnitude more in volume and processing rate compared to today's SCADA systems: TVA's SCADA system accumulated only 90 GB of data in 2009 from 105,000 points of measurement [1]. These high resolution data sources are being deployed with a vision to enable new applications that will make the grid significantly smarter. A key requirement for enabling this vision is a data processing infrastructure that can scale gracefully to handle the increasingly large volumes and variety of data, and increasingly complex and numerous applications that will consume this data in parallel.

Some of the "killer" applications enabled over the next decade by high resolution synchrophasor data, as foreseen by the North-American SynchroPhasor Intiative (NASPI) [14], and the corresponding type of data processing needed, are

1) dynamic state estimation – high rate streaming data
2) oscillation monitoring – high rate streaming data
3) real-time controls – high rate streaming data
4) post-disturbance analysis – high volume stored data

Although these are examples for synchrophasor data, other sources of data, such as the AMI, digital fault recorders and fine-grained weather data, can also be exploited to support the needs of a smart grid. The other sources of data will also pose challenges with handling high volume stored data or low latency processing of streaming data.

This section will discuss some applications, along with some underlying algorithms, that will enable a highly instrumented, reliable and responsive grid. It will attempt to highlight the needs these applications will place on the computational framework. Taking a massive data processing view, these applications can be classified into the following three broad categories, which will be discussed in more detail next: 1) Pattern discovery, 2) Data analytics, and 3) Learning.

3.2.1 Pattern discovery

A large number of smart grid applications will look for patterns or features in the data that provide specific and useful information regarding the grid. One example is the class of methods use to detect anomalies in the measurement time series data that may be streaming in from PMUs, like the event detector from Oak Ridge National Lab (ORNL)

that uses k-median clustering [15]. Within a sliding window, PMU data points are grouped into two clusters – points before and after a hypothetical "event". If the distance between the two clusters is large enough then the event is flagged as having the potential to propagate and cause cascading failures. The challenge here is to develop an anomaly detector on streaming data by having an efficient implementation of the k-median clustering algorithm that has extremely low processing latency. The pattern of interest here is a large distance between pre-event and post-event data points.

A more general time-series anomaly detector, like the one in [16], may be one that uses classifiers from machine learning, for example neural networks or support vector machines. The classifier is trained off-line to predict with high confidence whether streaming time-series data has anomalies. This may be done by creating thousands of waveforms with disturbances injected into them, which are then used to train the classifier to detect if an incoming signal has a voltage, frequency or harmonic disturbance. Although there is no explicit pattern of interest here, the classifier is looking for data from specific regions in the data space that would be classified as anomalies. [16], for instance, uses neural network classifiers to detect anomalies in the voltage, frequency and harmonics of the incoming signal, based on features computed from wavelet coefficients [25] of the signal.

Simulation-free grid instability monitoring methods can also be viewed as performing pattern discovery at an abstract level. These methods attempt to perform the important task of detecting and/or predicting undesirable instabilities, but by only analyzing the incoming measurement data, without simulating any model of the grid. For example, in [18], the authors propose using singular value decomposition (SVD) on the observed PMU data directly to predict voltage instability. This differs from the more traditional network model based approach which detects instabilities by finding the minimum singular value of the Jacobian matrix of the load flow equations [19]. Similarly, SVD was used in [20] on streaming PMU data to find modal frequencies and damping ratios and to detect poorly damped oscillations. The real-time evaluation of large SVD matrices with very low sub-second latency can be challenging and requires efficient data stream processing frameworks which can exploit parallel processing. An alternative is to use some machine learning algorithm to train a state prediction model. For instance, [21] matches generator angle data to a set of standard waveform and uses the best matching waveforms to predict the near-future trajectory and identify any potential transient instability. An interesting approach is to use classifiers to detect post-fault instability, as proposed in [17]. The method in [17] uses decision trees to classify post-fault evolving transients as either stable or unstable. These decision trees are trained off-line on features such as load variability from the peak, and velocities and accelerations of generator angles, using numerous simulated fault transients under different operating conditions.

All examples discussed above are for real-time applications that consume streaming data and must perform complex processing on it with very low latency to enable timely response and control. A software framework that is custom-designed for such applications is stream computing, and is discussed further in Section 3.3.2.

Apart from the real-time applications, pattern discovery is also performed on high volume stored historical data. A good example is post-disturbance analysis where the goal is to identify events and trends in historical data from a variety of sources and even regions that help explain the cause of the disturbance. Such analyses provide deep, useful insight into the behavior of the grid and are critical for developing mechanisms to avoid future occurrences of the same type of disturbances. Indeed much investigative effort was invested after the 2003 blackout in the Northeast U.S. and Ontario and the 2007 Florida blackout to trace the sequence of events and reasons for those disturbances [22][1][23]. The availability of high resolution, time-synchronized PMU data brings the potential for fast and detailed forensic analysis, but also the challenge of efficiently processing high volumes of data on distributed storage systems. A high performance processing system that can scale well with the increasing data volumes and needs of the increasingly "smarter" grid over the next few decades must be able to perform highly parallel and distributed processing, maintain data reliability with heterogeneous storage media and avoid data bandwidth bottlenecks. A promising candidate is MapReduce, a framework developed over the last decade to handle internet-scale data processing needs [7]. Section 3.3.1 discusses it in further detail.

3.2.2 Data analytics

Another class of applications that may be applied to grid data is data analytics. It includes tasks where the data is manipulated into a different format usually for further consumption by other applications and may be required both in real-time streaming data applications and offline high volume data applications. Some common examples of generic analytic tasks are sorting data based on one or more fields, searching for some combination of criteria and computing aggregation functions like max() and sum(). Some analytic tasks specific to time series are moving average and autoregressive integrated moving average (ARIMA). The latter is used, for example, in [24] to predict day-ahead electricity prices based on recent historical price series data.

A more specialized analytic task is signal decomposition, such as wavelet transform [25], Fourier transform or Prony analysis. In the previous section we saw an example of wavelet transform being used to obtain the features from which a neural network classifier can detect anomalies in the incoming signal. A more generic use of these transforms can be to compress signal data by keeping only the components from the decomposition of the signal that has the dominant coefficients. For example, if a transient signal window can be reconstructed with acceptable accuracy using only the first 10 wavelet components, then we only need 10 double precision numbers to represent that window of signal, instead of every sampled signal value in the window.

Compression factors of 4-5× in time-series data are possible using wavelet compression. Massive data compressed with using these methods can have significant storage and I/O bandwidth savings, especially when coupled with the data processing frameworks discussed in Section 3.3.

A potentially useful discretized representation of time series, developed recently in the computer science community, is symbolic aggregate approximation (SAX) [26]. The representation is particularly suited to data mining tasks like classification, search and clustering on time series. The basic idea is to represent time series by strings such as "aabcdab" by converting the values in each time interval into characters. Time series data collected from PMUs or smart meters can be represented efficiently as SAX strings. Clearly it is another form of data compression, but the real promise of SAX is the efficiency of data mining tasks for application like anomaly detection, using fast string processing [27].

3.2.3 Learning

A key enabler of the smart grid is accurate predictive capability, for example, prediction of load behavior for the next few hours or prediction of system stability for the next few seconds. In many cases such predictions are made by referring to some model that captures the behavior of the relevant sub-system. For example, load behavior over several hours is often modeled using neural networks, as in [28][29]. The predicted load is often a function of features such as past weather patterns, day of the week, time of the day and predicted weather pattern. More sophisticated models with many more variables will be required to capture the load-determining factors of the future smart grid. These factors will include demand response of energy consumers and the properties of renewable generation capacity at consumer sites.

These models are typically trained offline on large amounts of historical data to compute the right parameter values or structure of the model, and then applied to present conditions for prediction. This offline training or "learning" will need to extract and process some relevant subset of high volume historical data that may be stored in some distributed storage system. Like in the case of forensic analysis, the computational framework should be able to efficiently process this high volume distributed data even for these learning tasks. In some sense, even post-disturbance forensic analysis can be considered as a kind of learning, although not necessarily of some mathematical model parameters. Similar offline learning is required to train models for a variety of other application, like anomaly detection (Section 3.2.1) and weather based outage prediction [3].

Learning tasks may also be performed within real-time applications. An example is model-based state estimation as performed in [30] and [31], and discussed in detail in Chapter 4. Here, some model of the electrical grid is fit to the incoming electrical measurement data (e.g., synchrophasor data) giving values for the relevant state

variables (e.g., node voltage and current phasors) of the grid. Another example of such on-line learning is an adaptive neural network model that adjusts its parameters as real-time measurements become available [28].

These on-line learning tasks have very different requirements from the computational frameworks as compared to off-line learning discussed above. They often have to process data streaming in from the grid with very low computational latency so that the results can be effectively used in real-time, unlike off-line tasks where the issue is not real-time computation, but is high volume distributed data processing. In both cases, however, the current grid does not employ computational frameworks that can scale to the needs of the smart grid over the coming decades. The following section will discuss technologies that promise this desired scalability.

3.3 Data Processing Technologies

As highlighted in the beginning of the previous section, the applications discussed above will require a highly scalable data processing infrastructure to optimally exploit the large amounts of data from increasingly instrumented electricity grids. Three key capabilities that are critical for this scalability are:

1. Distributed and parallel processing of distributed high volume data
2. Low latency, parallel processing of real-time streaming data
3. Very fast access to subsets of high volume stored data

At the same time, the infrastructure should provide for easy development and deployment of new or enhanced applications and seamless integration of increasing number and variety of data sources and users. This section will discuss recently developed technologies from computer science and engineering that address each of these needs, respectively:

1. MapReduce
2. Stream computing
3. Hardware accelerated data processing

3.3.1 MapReduce: Massive distributed and parallel data processing

A simple calculation shows the need for massively parallel data processing to handle the high data volumes expected for future electricity grids. Consider a single data access channel connecting data storage drives with the processing nodes. Serial Advanced Technology Attachment (SATA) is the industry standard for commodity disk drive interfaces. The latest SATA specification (revision 3.0 [32]) released in May, 2009 targets a peak data bandwidth of 0.75 gigabyte/sec. Even at this peak speed, reading a 100 terabyte data set one a single data channel would require almost 38 hours. This would of course be an extremely naïve approach, but it highlights the clear need for parallel and distributed data access and processing to efficiently scale up to the petabyte scale.

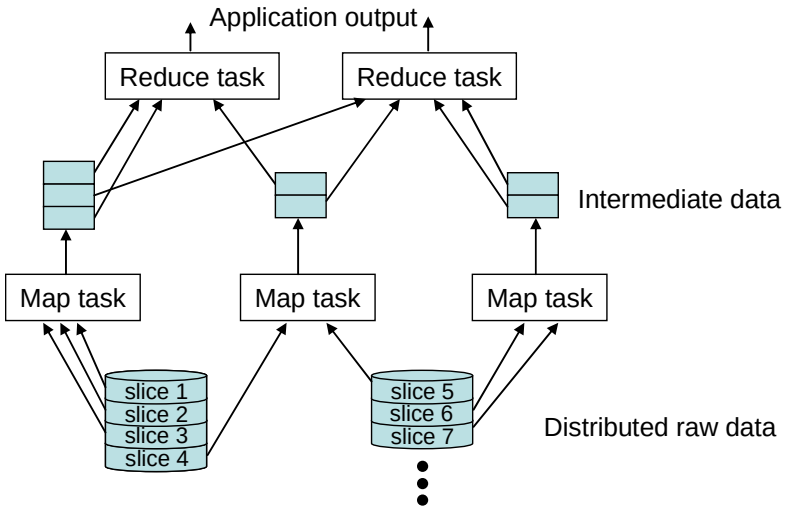


Figure 2: The MapReduce data processing model.

MapReduce [7] is a programming model for processing massive amounts of data in parallel that has found widespread use in internet-scale applications because of its ease of application development. A typical MapReduce implementation requires the programmer to write a "map" function and a "reduce" function, which together define the application program. This model is inspired by the map and reduce constructs in functional programming languages like Lisp [33]. The general flow of the model is as follows. Given a data set, perform the map() function on each record to compute (key, value) pairs for each record. In other words, the user-defined map() function is an application-specific mapping from the domain of data records to the domain of keys and values. The (key, value) pairs are grouped together by key and each group is processed by the user-defined reduce() function to compute corresponding outputs. Figure 2 outlines this data processing model.

Parallelism is achieved by running many instances of the map() function in parallel across multiple processing nodes, along with many instances of the reduce() function, which consume the results from the map stage. Data latencies can be kept low by dividing the stored data into slices and allocating each slice to a map() instance running on the local machine as far as possible. This scheme helps to reduce data transport across the network, keeping data latency low. Parallel programming is greatly simplified using this approach since the details of processor communication and system organization are typically performed by an underlying distributed file system and are hidden from the application developer. This combination of efficiency and ease of programming makes MapReduce particularly attractive for developing applications for a constantly evolving electricity grid infrastructure.

Many implementations of the MapReduce model have been developed, both proprietary [7] and in the public domain [9][8], and can efficiently support stored data volumes up to the petabyte range. One popular open source implementation is Hadoop. It enables

low cost, efficient computing over large clusters of commodity servers with high volumes of data stored in commodity disk drives and managed by the Hadoop Distributed File System (HDFS). Hadoop uses parallel I/O for fast read/write access times, and multi-core computing for maximally parallel processing of the data using the MapReduce model.

To exploit this programming model in any given application, the underlying algorithms often have to br re-architected to use the map()-reduce() function model. To address this issue, researchers have recently developed MapReduce versions of many widely used algorithms, and the list is continuously increasing. Of particular relevance to electricity grid applications are machine learning techniques, like neural networks, decision trees and clustering, that are used widely for forecasting and anomaly detection as discussed in Section 3.2. Many such algorithms have been implemented for MapReduce, as in [34]. A relevant example here is neural network training. One possible implementation of neural network training [34] can propagate training vectors through the network in the map stage. For n training vectors and p processors, each map() instance propagates on average n/p vectors through the network. The reduce() function can compute error gradients and back-propagate the errors in parallel to modify the neural network weights. With careful implementations, both the map and reduce stages run efficiently in parallel. A variety of other smart grid applications discussed in Section 3.2, that will need to process very high volumes of stored data, can be implemented in MapReduce to exploit its highly parallel, distributed processing infrastructure. It is particularly promising for high volume data processing tasks like post-disturbance analysis and offline model learning.

Some early steps in this direction have been taken in the power grid industry. Specifically, Hadoop was used recently by Tennessee Valley Authority (TVA) to analyze 15 terabytes of PMU data collected from 103 PMUs [35]. Their goal was to perform a forensic analysis of historical PMU data collected over the years to discover abnormal events in the power grid. The data collected was in the IEEE synchrophasor format C37.118-2005 [36] that consists of frequency and three phases of voltages and currents. Signal processing algorithms such as FFT and low/high pass filters were used to discover abnormal patterns and grid instabilities. Even so, 15 TB of data is much smaller than the petabyte range that is expected as grid instrumentation scales up over the coming years. It is worth noting here that in 2010 Google's highly optimized MapReduce implementation was being used to run more than 100,000 MapReduce jobs per day and handle datasets larger than 1 PB in size [7]. To achieve similar scalability for petabyte scale grid data, the data and the MapReduce implementation may benefit considerably from potential domain-specific customization, for instance, exploiting natural structure (e.g., time stamps) in the data for indexing and sharing tasks and processed data across applications.

3.3.2 Stream computing: Low latency stream data processing

Section 3.2 described several applications that would need to perform very low latency computations on real-time streaming data. It highlighted the need for a computational framework that can scale well with increasing large amounts of streaming data and increasingly complex and numerous applications running in parallel on this data. Similar challenges have been faced in other fields. An example is financial engineering, where split second investment decisions have to be made based on computations on large volumes of streaming data [37], often involving data analytics, pattern discovery and model training tasks similar to the case of smart grid applications. A popular solution emerging for these scenarios is stream computing.

Stream programming is typically done by creating a dataflow graph [38] of operators, which performs the computation required by the application. The inputs to the dataflow graph can be data from a variety of sources, such as internet sockets, spreadsheets, flat files, or relational databases, and may be consumed by one or more input operators in the graph. A synchronization primitive is an example of an operator which consumes data from multiple data streams and then outputs data only when data from each stream is read. This can be a useful operator for PMU data which can arrive at different times from different PDCs due to variable network delays and sampling times. Other relevant operators would be a fast Fourier transform (FFT) operator or a moving average operator. Each data element arriving at the input to an operator is typically treated as an event and the operator takes appropriate action when events occur at its input. Operators may be pipelined to perform in a sequential manner: output data from one operator is consumed by the next downstream operator. Operators may also perform in parallel if they do not have data dependencies: output from one operator may be consumed by two or more different operators working in parallel.

There operators are typically contained within containers called stream processing elements. For fast, parallel execution, the processing elements are automatically partitioned onto parallel processors and/or machines. The optimal partitioning depends on factors such as the amount and type of data streaming through different processing elements, the resource requirements for each of them and the dependencies between them. Hiding the details of parallel programming from the user greatly improves productivity and efficiency of streaming application deployment. The flexibility of input formats, the ease of developing and connecting the operators, and the automatic compilation onto parallel processors makes stream processing attractive.

There are many available implementations of streaming systems [39][40][41][10], S4 (Simple Scalable Stream System) [10] being noteworthy as an open source implementation. In S4, data elements are typically represented as (key, value) pairs. Processing elements take in such (key, value) pairs and emit new pairs. The value of the 'key' determines which processing elements will consume the data element. Because of

this similarity with MapReduce, S4 has also been described as "real-time MapReduce". Some aggressive implementations have developed new processors that are specifically designed to perform stream computing. Probably the most commonly seen example is a graphics processing unit, which has an architecture that is customized for performing graphics computations in a streaming manner: it contains deep chains of highly pipelined processing elements [42] that form a data flow graph. A more generic example better suited for a wider class of applications is the Imagine stream processor [43]. These processors provide the power and performance advantages of exploiting customized hardware instead of relying on general purpose servers that may have significant overhead.

Even though stream computing languages make application development quite easy, one may need to redesign traditional algorithms and applications to optimally use the stream processing flow. Reference [44] gives an overview of algorithmic advances and challenges in implementing efficient streaming versions of traditional algorithms. A specific example in this context is [45], where the authors implement decision tree learning for a streaming system. Real-time data will be available from increasingly numerous data sources across the grid. Stream computing frameworks hold the potential to enable scalable real-time applications that can extract information from this data to enable more complete situational awareness and very fast operational response and control.

3.3.3 Hardware accelerated data processing

Many data processing tasks need to access some subset of a massive data set. For example, to analyze trends in all the smart meter data from homes that experienced average daily temperature of more that 90°F during the last year, the first step is to extract the relevant records from all the recorded data (smart meter or other) for the entire year. Such data tasks are typically referred to as "queries" and performed using some database programming language, usually the Structured Query Language (SQL) [47].

Traditional database systems may not provide desired performance on these tasks as the volume of data increases toward the petabyte range. To address this challenge, the latest generation of database systems is beginning to employ hardware accelerators [11][48][49]. In particular, dedicated field-programmable gate arrays (FPGAs) and multi-core processors are programmed to perform atomic database query tasks, like filtering unwanted records and fields or combining records from different tables.

FPGAs are particularly well suited to these acceleration tasks because of the following reasons:

1. An FPGA provides a re-programmable array of basic circuit blocks called look-up tables (LUTs). The circuit blocks and the electrical connections between them can be re-programmed to implement any logic functionality. This enables a re-

optimization of the data access system if the data or application needs change over time. Such re-optimization allows the processing to occur at very high performance and very low power consumption levels.

2. A common issue when using general-purpose processors for parallel processing is that the parallel jobs have to compete for shared memory resources, which can enforce slow sequential behavior instead of the desired fast, parallel behavior. FPGAs now typically contain a reasonably large number of on-chip, distributed memory blocks that can be accessed in parallel. This enables many processes to operate in parallel without significant resource contention, breaking through the performance barriers imposed by sequential, von Neumann-style computation [50].

Two ways of integrating FPGAs in the data query flow have been proposed [11][48], as shown in Figure 3 and described next.

1. FPGA in the data path: Figure 3a shows an outline of this architecture. High volume data coming in from the data source (disks, network, etc) is first processed by an FPGA to reduce the amount of data significantly before sending it on to downstream general-purpose processor or memory. The data reduction may be achieved by filtering out the records and fields of the data that are not required by the query being performed. The goal is to filter data in the accelerators as fast as data can be read from the disk. This removes IO bottlenecks and allows for better memory utilization. Typically thousands of filtering streams are executing in parallel on the FPGA and performance is improved by 4-8×, and data reduction of > 90% is not uncommon [11]. More advanced data processing functions may also be incorporated in the FPGA that can further increase the performance and decrease the data volume that must be processed downstream [49].
2. FPGA as a co-processor: The approach here is to use the customized FPGA as an assistant for the general-purpose processor (CPU). The FPGA circuitry is optimized to perform certain atomic tasks that can be combined together to form SQL queries. These atomic tasks can be SQL functions like Join, Sort and GroupBy [47]. As Figure 3b shows, the CPU transfers these tasks over to the FPGA while it performs other tasks required to complete queries on the data. The figure is only conceptual and the actual implementation may have further enhancements, such as multiple FPGA engines that can stream processed data between each other before sending it back to the CPU or memory, or a network of processing nodes, each of which contains a CPU, FPGA and memory within it [48]. For a given application, these systems typically require a one-time optimization and compilation to program the FPGAs for optimum performance. One can envisage an online re-programming capability that re-programs the FPGA whenever there are changes in the application.

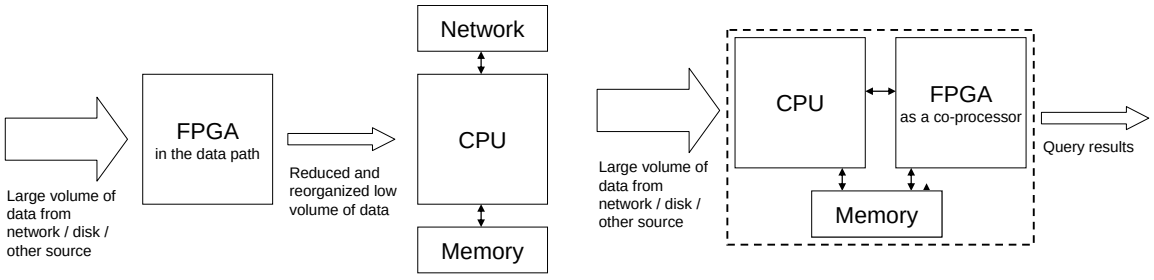


Figure 3: a) FPGA accelerator directly in the data path processes and reduces the data before sending it to the general purpose processor (CPU). b) FPGA accelerator as a co-processor is given specific high-volume data tasks by the CPU.

A similar approach that has also gained much research attention recently is to use graphics processing units (GPUs) to accelerate database query tasks [12][51]. GPUs have the advantage, like FPGAs, of supporting highly parallel processing on the chip because of highly pipelined and parallel processing elements. GPUs are able to provide higher processing power on the chip for the same power consumption in comparison to FPGA because they do not have to support the flexible field-programmability of FPGAs. On the other hand FPGAs provide the flexibility of optimizing the hardware for the specific application requirements, while GPUs are not optimized for data access and processing applications.

An important value that these hardware accelerators provide in the context of smart grid application is the possibility of accelerating pieces of these applications using FPGAs and/or GPUs. For instance, computational algorithms like wavelet transforms, nonlinear regression and simulation techniques, that form the core of many real-time situational awareness applications, can potentially be accelerated by exploiting the highly parallel computational fabric made available by GPUs and FPGAs. Similar ideas have been demonstrated in other areas of electrical and computer engineering, such as circuit simulation on GPUs [52] and speech recognition on FPGAs [53]. Indeed, gains have been demonstrated in very recent and early research [42] when using GPUs to accelerate transient stability simulation of large grids. Hence, these accelerators could be exploited both in the massive data-at-rest systems and within a stream computing framework to accelerate processing elements in the stream.

3.4 An Example Data Management Infrastructure for the Future Grid

Figure 4 shows an example data management infrastructure to illustrate how the discussed technologies may be deployed in a smart grid. Note that numerous different architectures are possible and this figure only shows one possibility for the purpose of illustration.

The architecture consists of the following main parts:

1. Data acquisition
2. Data communication
3. Data storage
4. Data processing

The data acquisition sub-system consists of the sensors and monitors: the example in Figure 4 shows data being acquired from PMUs, sensors and remote terminal units (RTUs) among others. The data communication sub-system may consist of data concentrators (e.g., PDCs or SCADA systems) and a highly reliable and extremely fast network that accumulates data from the data acquisition sub-system and transports it to receivers such as control centers and distributed mass storage systems. The key function of such data concentrators is to accumulate sensor/monitor data coming in from a region of the grid. In the case of PDCs, they may be organized in a flat or hierarchical level and may have local storage available to store the received data temporarily for a few days or weeks. Many parts of the acquisition and communication sub-systems may have local computation and storage to allow for localized and distributed real-time computation. For example, RTUs and PDCs in the future may have access to significant computing power to perform filtering and aggregation locally, to avoid loading the data network unnecessarily and to perform tasks that only need local grid information.

The data network as a whole has to satisfy some critical and special requirements. It should be able to chronologically synchronize data (e.g., PMU data) from different geographies, with different sampling rates, in different formats and with different incoming network latencies, based on the time stamps on the data packets. For PMU data, for instance, much of this synchronization may be done at the PDC level. Such networks will probably use some combination of internet protocols (IP) for communication, such as user datagram protocol (UDP) for data transfer and transmission control protocol (TCP) for control communications. UDP allows high rate communications, at the cost of dropped or lost data packets, while TCP allows highly reliable communication, at the cost of reduced bandwidth. Consequently, the network should be able to gracefully handle lost data without degrading the data bandwidth. Since there will probably be multiple consumers of the data from the network, each consuming different parts of the data, the network should be able to satisfy the heterogeneous needs such multiple consumers, such as different rates of receiving data. At the same time the network has to satisfy the very low data transport latency requirements of critical situational awareness applications, often in the range of 10s of ms. Detailed discussions on the expected properties of such networks can be found in [13]. Recent research has shown the feasibility of developing such networks using available technology, an example being the GridStat architecture described in [54].

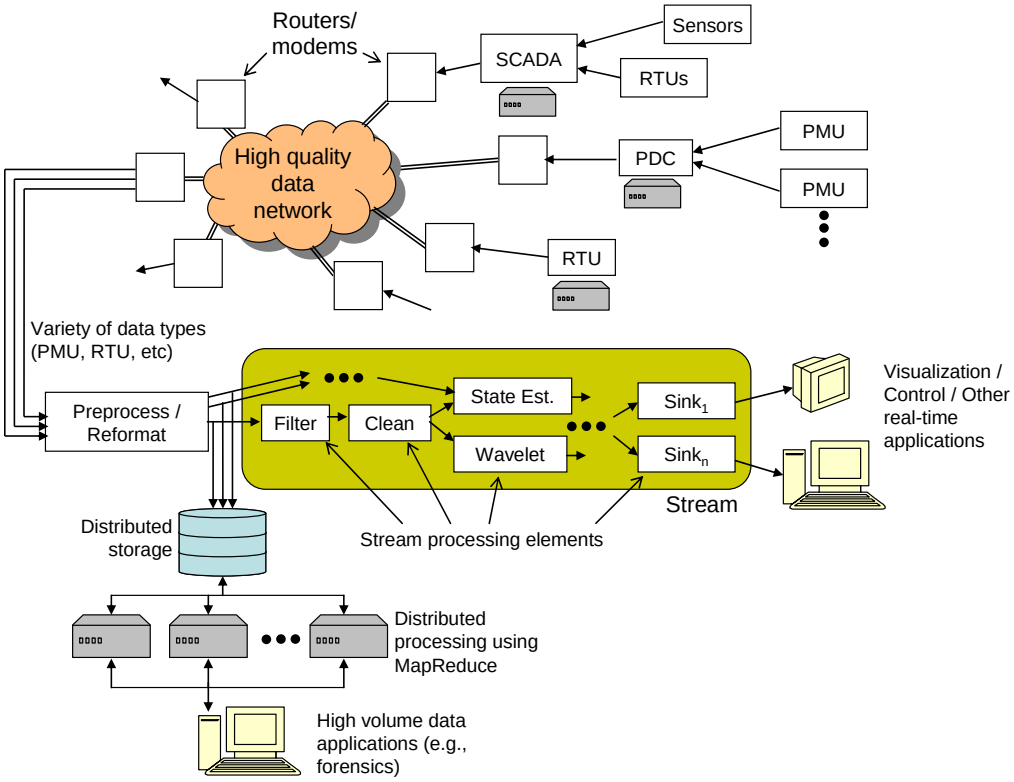


Figure 4: An example data processing infrastructure for the smart grid that uses stream computing for real-time applications and MapReduce for high data volume offline applications.

The data generated by the grid and by other sources relevant to the grid (such as weather data and grid asset records) will be stored in a storage system that will be distributed in geography, technology and ownership. Smart grid applications would require concurrent access to this data in many different modes, such as very frequent, relatively low volume for control applications, and infrequent and very high volume for forensic analysis. This wide variety in data generation, storage and use and the need for high reliability, concurrency and performance from the storage systems pose interesting engineering challenges. Fortunately, these challenges have been addressed by the information technology industry and distributed storage systems with high availability, reliability and performance have been developed. An example are internet-scale file systems, like the Hadoop distributed file system [9], which are designed to work with distributed, parallel computation frameworks like MapReduce. Given, the relevance of MapReduce to high data volume applications for the smart grid, these are doubly attractive. One drawback of typical internet-scale file systems is that they do not support non-MapReduce style of applications very well. To bridge this gap, researchers have recently proposed adaptations of more traditional storage cluster based, distributed file systems such as the Parallel Virtual File System [55] to support the highly distributed and parallel processing framework of MapReduce [56][57].

The final piece, and what has been the focus of this chapter, is the data processing sub-system. At the head of this sub-system, the data received from the network may need to

be preprocessed before further downstream storage or consumption, for example, to remove the UDP headers and/or to re-organize in a different format. Figure 4 shows an example of how stream computing and MapReduce might fit into this sub-system, and the overall data system. In the stream example, there are multiple stream processing elements, each performing some atomic task on the data flowing into it. For instance, the Filter element filters through only those records and fields in the data that are relevant to the downstream application. The Clean element gracefully handles missing data, resulting, for instance, from errors in the data network. The Wavelet element performs wavelet transforms on a sliding window of the data, for real-time anomaly detection as discussed in Section 3.2. The State Estimation element takes the same cleaned data to perform some state estimation task, as discussed in Chapter 4. Note that there may be multiple streams of data coming into the stream. For example, accurate state estimation will require not just synchrophasor data, but also other current grid status data such as relay status, data from RTUs and perhaps digital fault recorder data [58]. These other streams of data go through their own processing elements for pre-processing before being consumed by the State Estimation element. All streaming data is thus processed through the flow graph of stream processing elements and the results are finally presented by the Sink elements at the end of the stream, for use in tasks like visualization and control.

The MapReduce example shows that multiple distributed and heterogeneous processing units (computers and/or processors) are networked together to have access to the distributed historical data. Forensic and other high volume data applications running on a workstation are written in the MapReduce software framework and running. Each application runs as several parallel jobs on these processing units. The jobs are allocated to data chunks and processing units in a manner that reduces the amount of data movement across the storage network. To enable high performance, concurrent access to this high volume data by multiple applications, hardware-accelerated techniques discussed in Section 3.3.3 can prove very relevant.

Systems like the one described above may not be too far in the future since much of the underlying technology does exist. In fact, some early demonstrations may be feasible within a few years. The Pacific Northwest Smart Grid Demonstration Project [59] initiated in 2010 is the first multi-state effort to deploy multiple smart grid technologies across a large costumer and utility base. Indeed, some of the technologies discussed in this chapter are available to the demonstration project, for example stream computing and hardware-accelerated data access [60].

3.5 Conclusions

Future electricity grids will be much more instrumented than today's grids, given the increasing deployment of new monitoring devices like PMUs and smart meters. Consequently, they will have to deal with large amounts of high resolution and wide-area data from a variety of data sources, up to the petabyte range. This data will hold the

potential to substantially increase situational awareness, grid responsiveness and reliability, and optimality of grid operation, but only if it is efficiently processed by sophisticated applications to extract the relevant information and the data processing infrastructure possesses the following capabilities:

- low latency, yet complex processing of real-time data
- parallel processing of distributed data
- scalable to handle increasingly high volume and diversity of data
- support complex and diverse applications running concurrently
- ease of application development and deployment

This chapter discussed some of these applications and underlying algorithms, to identify two primary massive data processing modes: 1) fast, parallel and distributed processing of high volume historical data, and 2) fast, parallel and low latency processing of streaming data in real time. Similar challenges have been faced in other engineering domains, like internet systems and financial engineering and very promising technologies have been developed to address these challenges. Three technologies are particularly relevant in the context of the data processing needs of future electricity grids:

1. MapReduce for fast, parallel and distributed processing of high volume historical data
2. Stream computing for fast, parallel and low latency processing of streaming data in real time
3. Hardware acceleration for very fast, power-efficient data processing

Implementations and successful deployments of these technologies exist in other engineering domains. The electricity industry is well poised to exploit these technologies as data and infrastructure standards are being defined and the instrumentation infrastructure is being substantially expanded.

4 Large scale modeling and simulation of power grids

4.1 Introduction

The US power industry serves almost 300 million people, includes roughly 200,000 miles in transmission lines, and involves nearly 3,500 utility organizations. It is by far one of the largest machines built by mankind, and its reliable operation is of fundamental importance for the entire country's economy. Following the 2003 blackout, and the thorough investigations that followed it [3], the industry consensus was that among the main causes of that major and very costly failure, were (1) inadequate system understanding and (2) inadequate situational awareness. Both can be significantly improved through real-time, dynamic modeling, simulation and state estimation.

Presently, the state of real-time computations, in the power industry is, at an early stage of evolution regarding the use of computational capabilities. For example, the most frequently used on-line simulation tools are still the less accurate steady-state fast-decoupled power flow analysis and steady-state tracking state estimation tools, while the more accurate dynamic simulation and more computation intensive contingency analysis are typically performed offline.

Recently, however, the role of real-time, dynamic analysis is starting to change in the power industry. For example, researchers at the Pacific Northwest National Laboratory (PNNL) [1] have affirmed that most traditional power grid computations should be reformulated and converted to take advantage of the maturing high performance computing platforms, such as scalable multi-core shared memory architectures. They have shown that, through parallelization, significant speed-up and scalability can be achieved for both steady-state and dynamic state estimations, contingency analysis, and dynamic simulation. Not coincidentally, a very similar development is also taking place in Europe, where the European Commission is funding a four-year R&D project, called PEGASE [2]. It is implemented by a consortium composed of 21 partners including Transmission System Operators (TSOs), expert companies, and leading research centers in power system analysis and applied mathematics. The goal of PEGASE is to achieve real-time state estimation and dynamic simulation of the Pan-European wide interconnection for both operational and training use.

Both PNNL and PEGASE's efforts, although conceived on two different continents independently, identified a very similar set of computational bottlenecks for the reliable and efficient operation of continental dimension electricity grids. They include real-time and near real-time network wide dynamic simulation and state estimation, reliable, validated, static and dynamic models of complex network components (such as power electronics, FACTS, and HVDC devices), secure integration of renewable energy sources, and analysis of a large number of contingencies fast enough to provide timely options to system operators.

The computational challenges facing the power industry in the evolution to real-time dynamic analysis can be categorized along three major dimensions: (1) reduction of computation latency, (2) improvement in the computational scalability and capacity, and (3) improvement in the quality of computation results. The evolution of computational analysis along these three dimensions occurs by combining a multitude of application algorithmic and hardware architecture advances.

For example, today, power grid control relies widely on quasi-static state estimation based on data provided by the SCADA system, which delivers a snapshot of the grid status every few seconds. Depending on accuracy, state estimation latency ranges from seconds to minutes [1]. Since the effects from significant events, such as an equipment outage, can spread over hundreds of square miles in seconds [3], adequate reaction to

such events requires a significant reduction in latency of state estimation and other related analysis.

In the future, the wide-area instrumentation of the grid [4] will significantly enhance the visibility into the state of the grid and simultaneously greatly increase the scale of computational challenges. The available monitoring information is expected to increase by several orders of magnitude, a rate unprecedented for the power industry. If other industries are an indication, the availability of such large volume of data will only increase both the complexity of problems and the number of problem instances that need to be solved. For the power industry, this may lead to, for example, real-time contingency analysis and real-time transient stability analysis.

Another trend would be the transition from static to dynamic form of analysis. For example, current state estimations based on power flow analysis employ a quasi-static model formulation, but power grids are highly dynamic. In other industries, e.g. in automotives and semiconductors, every aspect of design is already subject to detailed virtual scrutiny through computer simulation. Achieving a similar degree of quality in computation analysis for the power industry will increase algorithmic complexity and computation cost significantly. It will require sophisticated techniques to create independent subtasks for efficient parallelization.

In the rest of this chapter, we will discuss in Section 4.2 a number of typical computational problems related to the power industry, and identify the common core numerical tasks involved therein. In Section 4.3 we will provide a detailed survey on these numerical techniques as applied in various scientific computing domains, and will formulate recommendations for the power industry to look into. We conclude this chapter in Section 4.4.

4.2 Computation Problems for the Power Industry

Simulation and state estimation are two major computation tasks used in the power grid operation. Each can be either static or dynamic. Power flow analysis, in Section 4.2.1, is the de facto static steady-state simulation of the power grid. Dynamic network simulation, in Section 4.2.2, involves formulating nonlinear differential equations to capture the dynamic nature of the grid. A special application of dynamic network simulation is to assess the stability of the grid in the event of disturbance, described in Section 4.2.3. Steady state and dynamic state estimations are described in Sections 4.2.4.

Almost orthogonal to these types of simulation is the contingency analysis that is closely related to the secure operation of the power grid. For example, the traditional method used in static security analysis involves solving power flow equations for each contingency; while the dynamic security assessment evaluates the transient performance of the system after each contingency disturbance such as short-circuit, loss of mechanical power, loss of electrical supply, and load fluctuations. The repeated assessment across a

large number of contingencies is highly time consuming and inadequate for real time applications. But from the computer science perspective, contingency analysis is an "embarrassingly parallel" problem because multiple contingency scenarios can be allocated to multiple processors and computed naturally in parallel. As shown in [1], parallel contingency analysis is one of the areas that could benefit tremendously from the increased computing power offered by high performance computing (HPC) platforms.

4.2.1 Power flow analysis

Power flow analysis is well understood and its models are relatively simple and easy to characterize. Quasi-static analysis of power networks requires the solution of large systems of nonlinear equations, which we will cover in more detail in Section 4.3.1.

Current power flow analysis typically solves the equations resulting from the modeling of transmission and distribution networks separately, and the formulation stops at the substation level where aggregated but crude load models are used. Future smart grid may require the extension of this formulation to a much larger scale (e.g. including parts of distribution network) to enable more accurate load modeling and to understand the impact of integration of renewable generation. The resulting systems are likely to be too large to be handled within a single multi-core system, and may require partitioning strategies to solve them on clusters or using some specialized hardware accelerators such as the SIMD accelerators [5].

4.2.2 Dynamic analysis

The dynamic analysis or transient simulation of power grids is performed by solving a system of differential algebraic equations (DAEs) under the given initial conditions. Dynamic network analysis is computationally challenging because the dynamic model of a power grid consists of a large set of DAEs that are typically sparse and stiff. Researchers at PNNL have reported that simulation of 30 seconds dynamics of the Western Interconnection require about 10 minutes computation time today on an optimized single-processor computer [1]. To dynamically simulate a large power Interconnection in real time would be even more challenging and require the application of high performance computing platforms.

A rough estimation on the computational effort can be done as follows. We assume that the real-time dynamic simulation will be performed in terms of dynamic phasors [6] and not at the full EMTP level. The expected time-step in this formulation will be on the order of 10-2s. An Eastern Interconnect Network-sized problem at around 30,000 buses can easily generate 106 state-variables. Furthermore, we estimate that we can reach a state solution at a cost of 104 operations per state variable and per time point. A rough estimate of the necessary computation power is therefore 1012 operations/second achievable only by advanced algorithms on HPC platforms.

Because of this high computational cost, dynamic simulation in the power industry is typically carried out only for a small number of interconnected power system models, and the computation is mainly performed offline. Today, the dynamic analysis of the interconnection-scale power systems is very rarely attempted, and the use such detailed dynamic analysis for real-time operational decision making is totally impractical.

Other industries also need to solve similar very large scale DAEs for dynamic simulation. One such an example is the Electronic Design Automation (EDA) industry where transient simulation of full-chip scale integrated circuits consisting of billions of electrical components (such as transistors and capacitors) is carried out on a daily basis for every complex chip design. Remarkable advances in speed, scalability, and capacity have been made by the EDA researchers in solving the large-scale system of DAEs. While these methods may not apply directly to the power industry, they could help spur further innovation, and enable real-time dynamic analysis of interconnection-scale power systems. We will cover in more detail the methods used by the EDA industry in Section 4.3.2.

Future direction for dynamic simulation is to achieve the capability of look-ahead dynamic simulation by enabling dynamic simulation on multiple processors. The goal is to significantly increase the speed of dynamic simulation so it can be performed faster than real-time so that system operators can look into the future to predict the dynamic status of the power grid. Together with Dynamic State Estimation, which will be discussed in Section 4.2.4.2, this will allow system operators to view the trajectory of the grid and even identify severe contingencies that have already begun but have not yet matured (such as voltage instability).

4.2.3 Stability analysis

The theory of stability analysis is a very mature and well-studied topic in dynamical systems [7] and applied to power systems [8] in particular. In today's power engineering practice, the dynamic stability problems consist of analyzing critical disturbances, predicting network reactions to these disturbances, and recommending the appropriate operating measures such as type of protection device, relay setting, and load shedding to mitigate the impact of these disturbances.

Efficient operation of the grid closer to its limits, especially in the presence of large intermittent generation sources, requires accurate dynamic modeling and stability margin evaluation. But because of the high computational complexity, and the lack of validated accurate models, stability analysis is seldom used in real-time operation.

Stability analysis is divided into small signal and transient stability analysis, each of which is discussed next.

4.2.3.1 Small signal stability analysis

Small-signal stability analysis determines the power system's ability to maintain proper operation and synchronism when subjected to "small" disturbances. A typical small-signal stability problem is the insufficient damping of system oscillations. From a mathematical point of view, a disturbance is considered to be small if the differential equations that describe the behavior of the power system can be linearized about the current operating point for the purpose of analysis.

Small-signal stability analysis reduces to eigen-analysis of the properly formulated linearization of the generally nonlinear system of differential equation describing the power network [8]. The computational cost of small-signal stability depends on the number of state variables needed to model the system. The cost is moderate for small problems but increases rapidly as the number of state variables grows.

Small-signal stability problems have two flavors: local and global. The local problems, called local plant mode oscillations, analyze rotor angle oscillations of a single generator, or a single plant against the rest of the power system, or the inter-machine or inter-plant mode oscillations between rotors of a few generators close to each other. Analysis of local small-signal stability problems requires the detailed representation of a small portion of the complete interconnected power system, while the rest of the system representation may be appropriately simplified by the use of simple models, such as system equivalents or model-order reduction techniques [9].

Global small-signal stability problems are caused by interactions among large groups of generators and have widespread effects. They involve oscillations of a group of generators in one area swinging against a group of generators in another area. Such oscillations are called inter-area mode oscillations. Analysis of inter-area oscillations in a large interconnected power system requires a detailed modeling of the entire system with hundreds of thousand state variables. For such large systems, special techniques have been developed that focus on evaluating a selected subset of eigenvalues associated with the complete system response [10]. Detailed discussion on this topic is given in Section 4.3.3.

In recent years, there has been significant progress in the development of parallel algorithms for large-scale eigenvalue problems [11][12][13] targeting high performance and hybrid computing platforms. A major research effort is needed for the adoption of these techniques to the power system stability problems in order to make possible faster or even real-time speed solutions for online monitoring.

4.2.3.2 Transient stability analysis

Not all power system stability problems fall into the small-signal category. The power system must maintain synchronous operation even in the presence of a large disturbance, such as the loss of generation or a transmission facility, or a significant change in load. The intermittent nature of renewable generation facilities is also prone to

cause instabilities. The system response to such phenomena consists of large excursions in power flows, voltages, currents, and angles. Therefore, the local linearization of the system differential equations employed by the small-signal stability analysis is no longer valid. For these situations, stability must be assessed through transient simulation methods by solving a set of nonlinear differential algebraic equations (DAEs).

Additionally, some of the analyzed events depart from the assumption of balanced operation of the three phases, and thus must use equations for all three phases. The method of symmetrical components [14] is typically used to formulate the unbalanced system equations.

The real time or faster than real time goal of transient power system stability analysis was contemplated in 1993 [15]. The simulation would simulate the power system model following an event and verify that it regains stability. It was using what was then state-of-the-art parallel DAE solver (Concurrent DASSL, DASPK). Those solvers have evolved and became the DASPK3.0 and DASPKADJOINT software packages [16].

An alternative way to analyze transient stability is by the so-called "direct methods" which are special forms of Lyapunov's second method [17][18]. These methods assess stability or calculate stability margins directly without actually performing transient simulation. The application of these methods to practical systems remains limited because of the difficulty of including detailed generator models into the energy function. Therefore, the best approach today seems to be a hybrid one where the computation of the energy function is incorporated in the transient simulation [19]. This way, transient simulations can also compute stability margins. It should be noted that, however, some significant progresses have been made on direct methods recently [68], and some even showed its applicability to solve a real power system's online transient stability problems.

Presently, it is clear that transient stability analysis is computationally difficult, and is thus rarely done in real-time for immediate operational purposes. However, with increased measurements from PMUs, massive computational power of HPC platforms, and new parallel algorithms, the stability analysis of the dynamic network is bound to move increasingly into the real-time operation of the power system, where it can deliver precise, predictive stability monitoring, a priori verification of operator actions, and ultimately participate in the automatic control and wide area protection.

4.2.4 State estimation

Given a network topology and measurements of observable network values, state estimation provides the current power grid state and is a central piece in grid control [20][21] [22] that drives other key grid operation functions such as contingency analysis, optimal power flow (OPF), economic dispatch, and automatic generation control (AGC).

4.2.4.1 Static state estimation

Traditional state estimation only estimates static states, i.e. bus voltages and phase angles, and the problem can be formulated as a weighted maximum likelihood (WML) optimization problem. But such a formulation takes minutes to solve and cannot keep up with the pace of even the SCADA data flow that comes in about every four seconds.

In practice, the static state estimation as used in the Energy Management System (EMS) today implements the light-version of full-scale state estimation, called "tracking state estimation," which updates the states for the next instant of time with new set of measurement data obtained for that instant without fully running the static state estimation algorithm [20]-[28]. This allows continuous monitoring of the system without excessive usage of the computing resources.

The advent of PMUs led to a paradigm shift in the state estimation process, and with sufficient PMU coverage, quasi-static state estimation may not even be necessary. At present, PMU coverage is, however, far from sufficient for direct measurement of grid states. Nevertheless, accurate voltage and current phasor measurements from PMUs will enhance the accuracy of state estimators, and direct measurement of voltage phases at buses by PMUs further reduces the size of the matrices in state estimation models, thus improving both the convergence speed and precision of state estimation [29][30].

In the interim, systems with a few or no PMUs still need to rely on state estimation to calculate phase angles and determining the power flow for operational and market purposes. Simulations would need to be 100 times faster to match the real-time measurements from PMUs whose sampling rate is at least 30 times per second.

A particularly interesting recent work by PNNL [1] attempts to speed-up the static state-estimation using high performance computing (16 Cray MTA-2 parallel processors) and adapted algorithms (shared memory conjugate gradient) to perform state estimation of the 14000 bus Western Interconnection system. They achieved parallel speed-up about 10 times with 16 processors. Further algorithmic tuning demonstrates that real time static state estimation at the SCADA speed (about 4 seconds) is achievable in the immediate future with high performance computing platforms.

Future direction is to achieve real-time state estimation for continental-scale grid interconnection [1][2] by reducing solution time from minutes to seconds to even milliseconds to keep up with the SCADA measurement cycles (typically ~4 seconds) and PMU cycles (typically 30 samples per second). Parallelized state estimation would enable the timely update of power grid status as soon as new measurements come in and lead to significantly improved situational awareness.

4.2.4.2 Dynamic state estimations

Dynamic state estimation (or dynamic system tracking) is a natural evolution of static state estimation representing a qualitative improvement of visibility into the network state, which would enable real-time dynamic stability monitoring and dynamic contingency analysis.

The dynamic state estimator was introduced in [31] and a considerable amount of research has been performed to analyze its capabilities and to reduce its computational complexity. Reviews of developments in dynamic state estimation for power systems and hierarchical state estimation were presented in [32][33]. The dynamic state estimator assumes the existence of a predictor function (in general nonlinear) that, given the present state of the system, can produce the state of the system at the next time-point of interest.

Such a function would emerge naturally if the power system were modeled at the level of physically based dynamic differential equations. However, the majority of the dynamic state estimation literature does not assume the existence of such a detailed model. Instead, they impose a simplified parameterized predictor function (most often linear) and a dynamic state/parameter estimator based on the Kalman-Bucy [34] filter. The filter yields estimates of a state vector and the parameters of the function. It is also assumed that an initial value of the parameter vector is given along with its respective covariance matrix. In practical power grid interconnection, the idealized assumption of the Kalman filter theory, however, tends to break down. Significant research has been conducted to develop robust algorithms that can be used in real-life situations. Reference [35] presents an in-depth survey of these techniques.

A Pacific Northwest National Laboratory group, who performed the static state estimation on a parallel computing platform [1], has also performed a study for dynamic state estimation [36], and demonstrated the feasibility to track the 30 samples per second phasor measurement updates for a small-scale power system with 3 machines and 9 busses.

For an Eastern Interconnect size power system with thousands of busses, achieving dynamic state estimation with a similar latency as the currently used static state estimations poses a veritable computational challenge. The problem will certainly not be solvable on a shared memory computer system and will require the development of distributed computational techniques that can scale to thousands of processors.

4.3 Numerical Methods for Power Grid Computation

In Section 4.2, we surveyed the algorithms used for simulation and state estimation of power systems. At the core of these algorithms are a number of computational techniques that are commonly deployed. They are sparse linear solvers, non-linear optimization, solution of a large system of nonlinear equations, differential algebraic equations (DAEs), eigenanalysis of large systems, and dynamic system tracking.

Sparse linear solvers and non-linear optimization are surveyed in Chapter 2, Section 2.6 and Section 2.5, respectively. This section will review the state-of-the-art of the remaining computational techniques. Particular emphasis is placed on the scalable, parallel, high performance computing methods that are unavoidable for achieving real time and faster than real time performance.

4.3.1 Solving nonlinear system of equations

Numerous problems in power systems (such as power flow analysis or the inner loop of solving DAEs as shown in Figure 5) are formulated as systems of nonlinear equations that need to be solved efficiently and robustly.

The principal methods for solving nonlinear systems are variations of the Newton iteration [37]. The Newton iteration consists of repeatedly solving large sparse systems of linear equations (Ax=b) resulting from local linearization of the original problem, where A is the gradients or Jacobian matrix of the original nonlinear system. When started from a good initial guess, Newton methods exhibit quadratic convergence, i.e., the number of correct digits doubles in every iteration.

Of particular interest for high performance applications are the Newton-Krylov methods [38] where the linear system is solved by a Krylov subspace based iterative method. These methods can result in quadratic convergence of the nonlinear iteration, provided the linear systems have been solved sufficiently accurately. One difficulty, however, arises in iterating the linear solver to a low enough tolerance, a requirement that often necessitates an efficient and scalable preconditioner for the Krylov method. A preconditioner can be viewed as an approximate but very efficient solver for the linear system and may often reflect engineering intuition. Preconditioned Krylov solvers often lend themselves to high performance parallel implementation. An example of a scalable parallelizable Newton-Krylov solver is KINSOL [39] as developed by the Lawrence Livermore National Laboratory (LLNL). KINSOL is a solver for nonlinear algebraic systems based on Newton-Krylov solver technology. KINSOL employs the Inexact Newton method which uses GMRES [40] as linear iterative method. The package also includes a parallel implementation friendly band-block-diagonal preconditioner.

The available packages, such as KINSOL only provide the algorithmic framework which opens the way for specific applications to achieve very fast computational performance. The ultimate performance depends on the problem specific implementation of the components that need to be supplied by the user. In the case of a nonlinear solver, the bulk of the time is spent in (1) formulating the linearized system (updating matrix entries based on model evaluations), a highly parallelizable process, (2) applying the linearized operator, y=Ax, also easily paralellizable, and (3) applying the preconditioner, where the block-diagonal structure allows blocks to be processed in parallel. The ultimate performance will be determined by a careful trade-of in the choice of the

preconditioner, evaluation speed, and parallelizability, versus their capacity of reducing the total number of iterations. The quality of the preconditioner can be significantly improved by utilizing all the engineering insight that can be brought into the mathematical formulation.

Though popularly used, Newton methods suffer from two major shortcomings: low reliability and high computation cost. The former is related to convergence because the desired solution is not guaranteed unless the initial estimate is sufficiently close. Otherwise, the method may diverge, or converge to an unwanted solution. The robustness (i.e., convergence guarantee) can be achieved by modified Newton Methods, such as relaxed, or damped-Newton [41][42][43], or through the more expensive continuation methods [44] [45]. The latter is due to the computation of gradients or the Jacobian matrix, which has to be formed and factored at each iteration step. Computing gradients or Jacobian matrix is typically considered as an expensive operation, and numerous approximation techniques have been developed to alleviate the burden. One category is the quasi-Newton methods that avoid reevaluation and factorization of the Jacobian matrix at each iteration, but at the cost of sub-quadratic convergence speed. They either keep the Jacobian constant at the cost of the convergence speed or update it with rank-one or rank-two matrices built up from information from the previous iteration, a technique borrowed from the optimization community.

It is clear that computing the gradients for a set of nonlinear equations efficiently and accurately is important not only for solving nonlinear equation itself, but also for other optimization applications such as parameter tracking and model calibration where it requires to compute the derivatives of model outputs with respect to model parameters. Accurate gradients have a determinant effect on the convergence speed and the reliability of the algorithms. In this regard, Automatic Differentiation (AD) [46] is a technique that has not been fully exploited in the power grid computation areas.

AD is a method that can numerically evaluate the derivative of a function specified by a computer program as fast as the function evaluation itself, but it differs from both the symbolic differentiation and numerical differentiation. The idea behind AD is that any computer program implementing a vector function $y = F(x)$ can be decomposed into a sequence of elementary operations, each of which can be trivially differentiated. These elemental partial derivatives, when evaluated for a particular input argument, can be combined through chain rule to obtain the exact derivative for F. This process provides the ability to efficiently compute accurate derivatives that are "correct by construction." Because symbolic differentiation occurs only at the most basic level, AD avoids the computational problems inherent in complex symbolic computation.

4.3.2 Solving nonlinear differential-algebraic equations

Both dynamic (or transient) simulation and large-signal transient stability analysis require the solution of a large system of nonlinear differential-algebraic equations

(DAEs), constituted from the dynamic models of the network components through Kirchhoff laws.

Figure 5 summarizes the typical flow of solving a system of nonlinear DAEs. The solution uses a chosen numerical integration scheme along with suitable time discretization (time-stepping) to approximate the differential operator. Numerical integration using implicit methods will reduce the nonlinear DAE problem at each time step to a set of nonlinear system of equations, which needs to be solved using one of the methods, e.g., of the Newton-Krylov type, as discussed in the previous subsection. At the core of solving such set of nonlinear system of equations is a series of solutions to a system of linear equations.

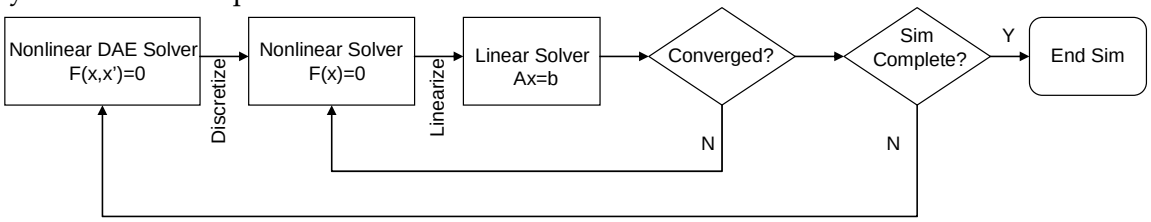


Figure 5: Nonlinear differential-algebraic equations (DAEs) solution flow.

One of the state-of-the-art general purpose DAE solvers is the IDA solver in the SUNDIALS package developed by the Lawrence Livermore National Lab [39]. The solver has a fully parallelizable version, and the simulation advances in time using a variable time step. At each step, the nonlinear system is solved through Modified or Inexact Newton methods, and the Jacobian is updated only when necessary for convergence. The linear systems are solved by one of five methods: two direct methods (dense or band) and three Krylov methods (GMRES, BiCGStab, or TFQMR). The direct methods allow either a user-supplied Jacobian or an internal approximation, while the Krylov methods include scaling and optional left preconditioning. Krylov methods with block diagonal preconditions are amenable to fast parallel implementation, as the blocks can be solved in parallel. The solver also implements sensitivity analysis.

An important technique for parallelizing DAE solvers is the Schur Complement Domain Decomposition Methods (DDM) [47]. Unlike PDEs that can rely on space domain decomposition, DAEs have no such geometrical information to explore in order to decompose the system. The connection between unknowns has to be explored explicitly with engineering insights so that the system can be decomposed into subsystems where the Schur complement domain decomposition methods can be applied.

Same as those general nonlinear system equation solvers, all existing general-purpose DAE packages only provide the shell of the simulator, and leave the task of achieving high performance to the specific application developers, who need to find the most effective way to implement those most CPU and memory consuming tasks based on their specific domain knowledge.

One exceptional example in adapting these general DAE solution techniques to solve a specific domain problem while achieving superior speed and scalability is the circuit simulation program developed in the Electronic Design Automation (EDA) area for integrated circuit (IC) design. The program is called SPICE (Simulation Program with Integrated Circuit Emphasis), which has become the golden standard of many distinct electrical circuit design problems. Circuit designers use SPICE to run many transient simulations on circuits with billions of circuit components (such as transistors and capacitors) on a daily basis.

For SPICE, the first step is to build the DAEs by parsing a netlist that describes the circuit to be simulated, where each electrical component has models recognizable by SPICE. The devices and their governing equations, both of which can be differential and/or regular algebraic equations, are formulated by SPICE in the form of a nonlinear DAE system. This set of equations determines the behavior of the electrical system. For power systems, similar DAE models can be derived except that the "netlists" for the power industry are not so much standardized and widely-accepted among different tool vendors as SPICE does.

Once the DAEs are formed, the solution to DAEs follows the similar flow as show in Figure 5. Below we will look at two specific approaches adopted by the EDA researchers in implementing SPICE to see how it achieves the desired speed, scalability, and capacity for simulating a full-scale chip design.

The first approach is to retain the high accuracy of simulation results while parallelizing every major computational task in simulation to achieve the desired speed and scalability. Typically for integrated circuit technologies, accuracy is utmost important. Therefore, both models and results need to maintain high accuracy during simulation. The performance is achieved by integrating both coarse-scale (multi-processor) and fine-scale (multi-thread) parallelism into every step of the entire simulation flow as shown in Figure 5. Careful choice of the composition of coarse-scale and fine-scale parallelism is proven to provide the best circuit simulation performance across a wide spectrum of parallel computing platforms, from high-end supercomputers, to large clusters, to multi-core desktops. One such exemplar implementation is the Xyce SPICE program [48] that was developed by the Sandia National Lab.

The second approach is to make a judicious tradeoff between accuracy and performance. The class of such SPICE simulators performing faster runs with relaxed accuracy constraints is called fast-SPICE. The motivation comes from two aspects: (1) higher level of chip integration necessitates the need to simulate larger IC designs well beyond the capacity limits of the conventional SPICE simulation, and (2) accuracy requirements for full-chip simulation does not need to be as stringent as those for more localized circuit blocks, allowing for a few percent errors as compared to full SPICE accuracy. This relaxed accuracy constraints on large-scale simulations has enabled EDA researchers to

develop a slew of Fast-SPICE simulation engines that have achieved 1,000 to 10,000 times speedup and similar scale capacity improvements. It is already a common practice in the EDA area to expect a commercial fast-SPICE tool performing simulation for circuits with hundreds of millions of devices in a matter of minutes [49][50].

Fast-SPICE techniques often use partitioning to decompose the large-scale electrical system in smaller scale subproblems to be solved more efficiently. They also perform matrix-based techniques to exploit parallelism and sparsity of the underlying system matrices. More importantly, they perform circuit-level, or component-level abstractions to achieve faster simulation for the entire system.

Since the underlying mathematics of SPICE is similar to that of the power system dynamic analysis, it makes sense for the power industry to look into the evolutionary path of SPICE in addressing those computational challenges, and see how the power industries transient simulation programs can be benefited from the success path of SPICE.

For example, power system analysis also displays similar features that make fast-SPICE like algorithms suitable for really large scale power system analysis. Approaching real-time simulation of power system networks of large size would require combining the state-of-the-art parallel DAE solvers, high performance computing platforms, and domain specific engineering insights for partitioning and parallelizing the problem. Techniques such as Model Order Reduction can be used for abstracting portions of the network that are not of detailed interest, and thus reduce the complexity of the network model. In other words, subsystems that do not require internal visibility can be abstracted with controlled loss of accuracy. The reduced models can contribute a significantly smaller (sometimes by orders of magnitude as was done in fast-SPICE) number of states to the system model. MOR techniques rely mainly on projection onto well chosen subspaces, such as Krylov subspaces for moment preservation, and balanced realization for H-infinity optimality. In order not to compromise the stability of the larger system simulation, the Order Reduction methods must preserve passivity. The literature on Model Order Reduction is vast and covers all engineering and scientific disciplines. We refer the reader to [51] for a power systems perspective on MOR, to [52] for the state of the art in linear system MOR techniques, and to [53] for extensions to nonlinear systems.

With such an attainable scale of speedup, power system operators in the future can quickly perform almost real-time analysis of the power system, enhance their situation awareness, and make correct decisions to operate the power system securely.

4.3.3 Eigenalaysis of large systems

As discussed earlier, small-signal stability analysis boils down to the eigenalaysis of large systems. We survey state-of-the-art eigenanalysis techniques in this subsection,

and the recommendation for the power industry is to adapt some of these techniques, particularly, computation of subsets of eigenvalues and/or eigenvectors, and adopt them to solve the large-scale small-signal stability problem in real time.

Several well-established global small-signal stability programs exist already in the power industry. For example, the AESOPS algorithm [54] uses a novel frequency response approach to calculate the eigenvalues associated with the rotor angle modes. The selective modal analysis (SMA) approach [55] computes eigenvalues associated with selected modes of interest by using special techniques to identify variables that are relevant to the selected modes, and then constructing a reduced-order model that involves only the relevant variables. The program for Eigen-value Analysis of Large Systems (PEALS) [56] uses two of these techniques: the AESOPS algorithm and the modified Arnoldi method. These two methods have been found to be efficient and reliable, and they complement each other in meeting the requirements of small-signal stability analysis of large complex power systems. But all of these analyses are currently performed off-line.

Eigenanalysis for large-scale power system stability analysis should rely on algorithms for the computation of subsets of eigenvalues and/or eigenvectors. A scalable code that performs such an analysis is ARPACK [57] which stands for ARnoldi PACKage. ARPACK software is capable of solving large scale Hermitian or non-Hermitian (standard and generalized) eigenvalue problems. It has been used for a wide range of applications. Parallel ARPACK (P_ARPACK) is provided as an extension to the current ARPACK library and is targeted for distributed memory message passing systems. The ARPACK software is based upon the Implicitly Restarted Arnoldi Method. Another solid package with similar functionality is SLEPc [58].

We give a brief explanation of the numerical methods underlying the computation of subsets of eigenvalues and/or eigenvectors. When the matrix is symmetric (Hermitian), the methods reduce to the Implicitly Restarted Lanczos methods (IRLM). Eigenvalue subset algorithms project the original matrix onto a well chosen low-dimensional subspace and approximate eigenvectors from that subspace. When the projection is done on a Krylov subspace, original eigenpairs can be approximated from the small projected matrix eigenpairs. The Lanczos or Arnoldi processes are used to produce bases for the Krylov subspace.

Implicit restart addresses the unfortunate aspect of the Lanczos or Arnoldi processes in that there is no way to determine in advance how many steps will be needed to determine the eigenvalues of interest to within a specified accuracy. The eigen information obtained through these processes is completely determined by the choice of the starting vector. Unless there is a very fortuitous choice of starting vector, the eigen information of interest probably will not appear before a very large number of iterations are performed. Clearly, it becomes intractable to maintain numerical orthogonality of

the projection subspaces. Extensive storage will be required, and repeatedly finding the eigen system of the projected matrix also becomes expensive. The need to control this cost has motivated the development of restarting schemes. Restarting means replacing the starting vector with an "improved" starting vector and then computing a new Arnoldi factorization with the new vector.

Implicit restart is an efficient and numerically stable method to extract interesting information from the large and sparse matrix, and to avoid numerical difficulties and storage problems normally associated with the standard approach. It does this by continually compressing the interesting information into a fixed-dimensional subspace. This is accomplished through the implicitly shifted QR mechanism. The algorithm is capable of computing a few eigenvalues with user-specified features, such as largest real part or largest magnitude. More detail can be found in [10].

An alternative class of methods for eigenvalue computation is the Jacobi-Davidson (JD) [59] method that addresses the drawbacks encountered in the Lanczos and Arnoldi methods. The latter two sometimes exhibit slow convergence and tend to have difficulties finding interior eigenvalues. This is cured by "shift-and-invert" variants of these methods, but these require a factorization of the shifted matrix, which may be very expensive or even not feasible. The JD method avoids this pitfall. The JD method belongs to the class of subspace methods, where approximated eigenvectors are sought in a subspace of the origin. Each iteration of these methods has two important phases: the subspace extraction phase and the subspace expansion phase. In the subspace extraction phase, a sensible approximated eigenpair is sought, of which the approximated vector is in the search space. In the subspace expansion phase, the search space is enlarged by adding a new basis vector to it in the hope of it leading to better approximated eigenpairs in the next extraction phase. One of the strengths of the JD method is that both phases may cooperate with each other to approximate those difficult-to-find interior eigenvalues. Reference [10] provides a good overview of the JD methods with pointers to advanced codes.

4.3.4 Techniques for system tracking

As discussed in previous section, dynamic state estimation (or dynamic system tracking) is a natural evolution of static state estimation to improve the quality of result, and it would then enable real-time dynamic stability monitoring and dynamic contingency analysis. We review those techniques that are promising to achieve dynamic system tracking with a similar latency as the currently used static state estimations. Recommended techniques that are of particular interests and worthy to look into are the particle filters [60], where the accuracy of the state estimates as a function of ensemble size is an important research question.

The classical Kalman filter provides a complete and rigorous solution for state estimation of linear systems under Gaussian noise. For the more general practical

applications, a wide range of suboptimal methods have been developed. The estimation problem for nonlinear, non-Gaussian cases is complicated by the fact that the distribution of the state is not completely characterized by the second moment, as is the case with Gaussian distributions. In particular, the probability density of the state evolves according to the Fokker-Planck partial differential equation.

The extended Kalman filter (XKF) [61] proceeds by adopting the formulae of the classical Kalman filter with the Jacobian of the dynamics matrix playing the role of the linear dynamics matrix. This approach thus mimics the classical Kalman filter by propagating the error covariance in the linear case. Although XKF estimation is effective in many practical cases, the method fails to account for the fully nonlinear dynamics in propagating the error covariance, which, in turn, fails to represent the error probability density.

A broader category of filters is known as particle filters [60]. Unlike XKF estimation, particle filters do not use the Jacobian of the dynamics. The starting point for particle filters is choosing as a set of sample points, i.e., an ensemble of state estimates, which captures the initial probability distribution of the state. These sample points are then propagated through the true nonlinear system and the probability density function of the actual state is approximated by the weighted ensemble of the estimates.

For high dimensional problems, particle filters are an attractive alternative to the classical Kalman filter methods. The basic idea is to replace the expected values by suitable Monte-Carlo sample averages. An appealing aspect of these algorithms is that, unlike the Markov chain approximation method or any other standard discretization schemes, one does not need to fix a grid to approximate the dynamics. Particle methods are very flexible and easy to implement; and they are ideally suited for a parallel computing architecture.

Although the ensemble scheme is easy to implement, it suffers from severe degeneracy problems, especially in high dimensions. The main difficulty is that, after a few time steps, all the weights tend to concentrate on a very few particles, which drastically reduces the effectiveness of sample size. A common remedy for this paucity of significant particles is to occasionally re-sample in order to rejuvenate the particle cloud. Various re-sampling methods exist. One of them is the occasional resampling technique, whose main idea is to periodically resample with replacement from the discrete distribution approximating the filter to obtain a uniform distribution of weights. Undesirably, resampling adds extra noise to the approximation scheme. Hence it is important not to resample too frequently.

Examples of particle filters include the unscented Kalman filter [62], the central difference Kalman filter [63], and the ensemble Kalman filter (EnKF) [64]-[66]. The former two types of filters chose their sample points deterministically, and the number

of sample points required is of the same order as the dimension of the system. In contrast, the number of samples required in the EnKF is heuristically determined, and its error statistics are predicted by using a Monte Carlo or ensemble integration to solve the Fokker-Plank equation. While one would expect that a large ensemble would be needed to obtain useful estimates, the literature on EnKF suggests that an ensemble of size 50 to 100 is often adequate for systems with thousands of states. The EnKF estimation has been widely used in weather forecasting, where a large number of measurements are available with the models of extremely high order and nonlinear, and the initial states of high uncertainty.

An example of a large scale implementation of a nonlinear distributed Kalman filter to estimate the state of a sparsely connected, large-scale, multi-dimensional dynamical system monitored by a network of sensors is presented in [67]. Local Kalman filters are implemented on smaller dimensional subsystems, obtained by spatially decomposing the large-scale system. The (approximated) centralized Riccati and Lyapunov equations are computed iteratively with only local communication and low-order computation by a distributed iterate collapse inversion (DICI) algorithm.

4.4 Conclusion

To realize the true potential of the future smart grid requires system operators to run a slew of robust, high-quality simulation and analysis tools in real time on much larger scale power system models, just like many other industries such as aviation, automotives, semiconductors, finance, and meteorology.

In this chapter, we have reviewed a number of core computation problems that are frequently encountered in the power industry today. The tradeoff between speed and solution quality made in the existing tools renders them inadequate to handle the computation challenges of the future smart grid. For example, the online simulation tools such as power flow analysis and state estimation all suffer from low accuracy and may not reflect power system reality, while other high quality tools such as dynamic transient simulation and stability analysis have limited capacity and take too much computation time, thus they are only used in offline analysis environment.

The recommendation for the power industry is to accelerate the adoption state-of-the-art high performance computing techniques to enhance existing power system simulation capabilities, and reduce computation latency, improve scalability, and quality. The goal is to enable real-time simulation of large-scale power systems, both statically and dynamically, to support power system operators' online decision-making.

We have also surveyed a number of key computational techniques and their latest developments in computation speed and solution quality. We believe that many computation technologies developed in other domains have matured enough to be readily adapted by the power industry to address its computational challenges. The key

is to have the right synergy between computation technologies and power domain expertise.

5 Conclusion and Recommendation

The future smart grid will be a complex, heterogeneous system of systems, orders of magnitude more complex than the current power grid with a growing and unprecedented need for solutions to address the scale of data that are being generated and to react to that data in real-time.
Current IT infrastructure and software tools used by power grid operators are, however, suffering from the following limitations and deficiencies:

- The inability to include stochastic uncertainty in fuel prices, renewables, or weather for decision-making
- The inability to collect, manage, and mine the high-volume data emerging from the grid to gain insights into the operational status of the grid
- The limited computational capabilities and leverage of multi-core machines to improve the solution quality, reduce response latency, and increase the analysis and optimization scalability
- An inherent lack of code/tools integration that would enable utilities and system operators to optimize across different operational domains
- The lack of automation and autonomy of grid operation – as current software was design with a human in the loop and to support human decision making

Therefore, considerable innovations are required to transform the current power system into a 21st century smart grid.

At the core of secure operations of future smart grid are system operators, for example, Independent System Operators (ISOs) and Regional Transmission Operators (RTOs), who monitor and control the physical power grid. The system operators leverage analysis and optimization tools such as security-constrained unit commitment and economic dispatch during their decision-making process. All these computing tools, however, will require a major makeover to meet the dynamic challenges of the future smart grid and perform these analyses not only 100 times faster than existing practices but also more accurately. Stochastic analysis will be an essential part of these tools to accurately reflect the intermittent nature of renewable energy resources and uncertainty in fuel prices and loads.

To effectively take advantage of the benefits offered by the wide-area instrumentation of the physical grid, it requires a Massive Data Set Management (MDSM) system that can aggregate and analyze the petabyte-scale data sets from PMUs, IEDs, SCADA, smart meters, smart appliances, and other sensors (weather, wind, solar) to deduce useful

information. The MDSM system will then supply the distilled information as needed for improved grid modeling, simulation, and operation. Although this has been traditionally a weak part of the grid operation, technologies developed from the computer science areas, such as MapReduce to process distributed massive data-at-rest, stream computing to process fast data-in-motion, and hardware-accelerated data warehousing, are all at a stage that they can be effectively adapted and applied to the massive data processing tasks required by smart grid applications.

Additionally, the traditional power systems analyses like power flow, state estimation (static and dynamic), and transient stability will need to be 10-100x faster to enable real-time solutions.

Addressing these fundamental issues will require real-time solutions of large-scale networks and systems, derived from a plethora of incoming streaming data and sound mathematical models. These solutions will greatly enhance situational awareness, reduce the need for additional sensors, and allow faster than real-time offline analysis, experiments, and insights without impacting the operational systems, i.e., through the availability of a "virtual power grid" as shown in Figure 6. Each major physical component of the grid will be represented by a mathematical model, and each model will consist of the combined network differential equations and algebraic equations that approximate the behavior of the component. The model parameters will be calibrated and refined based on the physical grid data, which is provisioned by the Massive Data Set Management (MDSM) system. The calibration and refinement can occur even in real-time by continuously adapting the models through feedback of the streaming real-time measurements of the power grid from the MDSM system. Having such a system will enable system operators to accurately simulate the real-time phenomena that are happening at the physical power grid by leveraging the virtual grid but without impacting the physical grid. It will also enable operators to replay historical outage or cascading events at any time quickly.

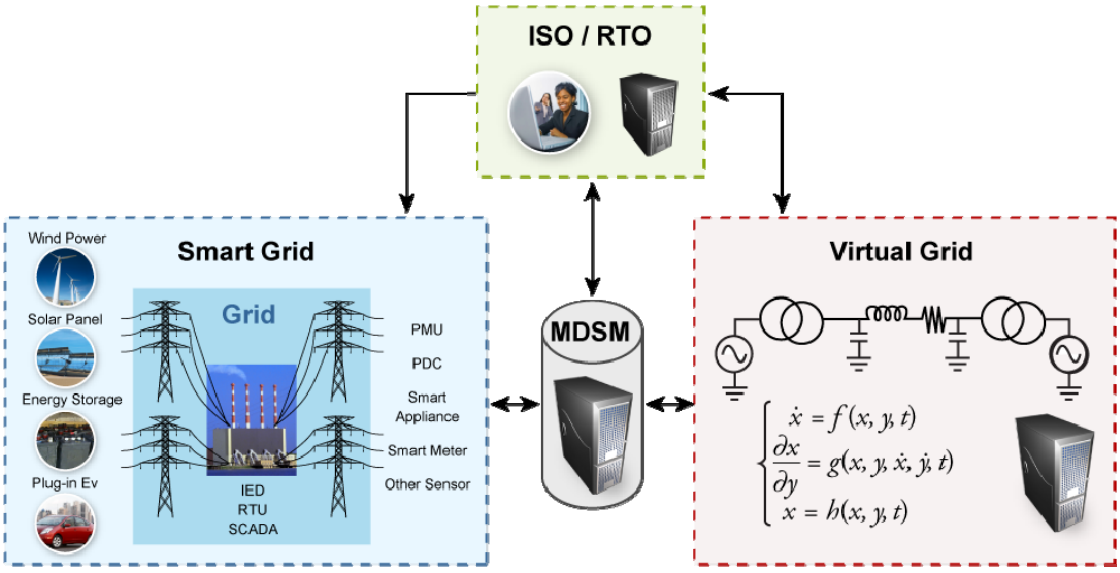


Figure 6: The role of virtual power grid in the operation of a future smart grid. MDSM provides the data analysis and interface between the physical and virtual grid.

Such vastly improved situational awareness, combined with massive computing power in control centers, will make possible a suite of applications that are not possible today. They range from sophisticated networks monitoring for stability, security, and reliability, wide area protection, prevention of cascading failures, ability to pre-test by detailed dynamic simulation of all management and control operations, to real-time contingency planning. They will evolve in the long term all the way to almost full automatic control of the grid, relegating human intervention to only the most sensitive and complex operations. Fully realizing these features of such a virtual grid will require a hybrid-computing platform that supports petaflop-scale computation capability, high-speed network communication, and very-likely domain-specific acceleration.

The smart grid will be the next generation of the world's already most complex machine and it will operate closer to its transfer limits. This will require significant improvements in simulation and modeling capabilities throughout the grid operation. To add these new capabilities, we will need to overcome daunting computational challenges, e.g. adding stochastic analysis to SCUD and SCED without degrading performance, using appropriate computational frameworks for stored and streaming data analysis, parallelizable and scalable solutions of linear system of equations, nonlinear system of equation, and nonlinear differential algebraic equations. The solutions will be obtained by reaching out and working with other disciplines like, Mathematics, Computer Science, and Electrical Engineering.

6 References

Chapter 2 References

[1] T.S. Dillon, K.W. Edwin, H.D. Kochs, and R.J, Taud, "Integer programming approach to the problem of optimal unit commitment with probabilistic reserve determination", IEEE Transactions Power Systems, vol. PAS-97 (1978), pp. 2154-2166.

[2] N.P. Padhy, "Unit commitment - A bibliographical survey", IEEE Transactions on Power Systems, 19 (2004), pp. 1196-2005.

[3] G. B. Sheble, G. N. Fahd, "Unit commitment - Literature synopsis", IEEE Transactions on Power Systems, 9 (1994), pp.128-135.

[4] W. L. Snyder Jr., H. D. Powell, Jr, and J. C. Rayburn, "Dynamic programming approach to unit commitment", IEEE Transactions Power Systems, 2 (1987), p.339-347.

[5] S. Takriti and J. Birge, "A stochastic model for the unit commitment problem," IEEE Transactions on Power Systems, 11 (1996), pp. 1497-1508.

[6] J. Carpentier, "Contribution to the economic dispatch problem", Bull Soc. France Elect., 8 (1962), pp. 431-447.

[7] G. Torres and V. Quintana, "An interior-point method for nonlinear optimal power flow using voltage rectangular coordinates", IEEE Transactions on Power Systems, 13 (1998), pp. 1211-1218.

[8] J. Lavaei and S. Low, "Convexification of optimal power flow problem", in Forty-Eighth Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, Oct. 2010.

[9] O. Alsac and B. Stott, "Optimal load flow with steady state security", IEEE Transactions on PAS-93, 3 (1974), pp. 745-751.

[10] F. Capitanescu, M. Glavic, D. Ernst, and L. Wehenkel, "Contingency filtering techniques for preventive security-constrained optimal power flow", IEEE Transactions on Power Systems, 22 (2007), pp. 1690-1697.

[11] W. Qiu, A. J. Flueck, and F. Tu, "A new parallel algorithm for security-constrained optimal power flow with a nonlinear interior point method", in IEEE Power Engineering Society General Meeting, 2005, pp. 2422-2428.

[12] A. Schellenberg, W. Rosehart, and J. Aguado, "Cumulant-based stochastic nonlinear programming for variance constrained voltage stability analysis of power systems", IEEE Transactions on Power Systems, 21 (2006), pp. 579 – 585.

[13] FERC 2006, Federal Energy Regulatory Commission, "Security-Constrained Economic Dispatch: Definition, Practices, Issues and Recommendations – A Report to Congress Regarding the Recommendations of Regional Joint Boards For the Study of Economic Dispatch Pursuant to Section 223 of the Federal Power Act as Added by Section 1298 of the Energy Policy Act of 2005," July 31, 2006, http://www.ferc.gov/industries/electric/indus-act/joint-boards/final-cong-rpt.pdf

[14] R. H. Kerr, J. L. Scheidt, A. J. Fontana Jr., and J. K. Wiley, "Unit commitment", IEEE Transactions on PAS-85, 5 (1966), pp. 417-421.
[15] A. Sousa and G. Torres, "Globally convergent optimal power flow by trust-region interior-point methods", in Power Tech, 2007 IEEE Lausanne, 2007, pp. 1386-1391.
[16] J. R. Birge, and S. Takriti, "Using integer programming to refine Lagrangian-based unit commitment solutions," IEEE Transactions on Power Systems, 15 (2000), pp. 151-156.
[17] Z. Quyang, and S. M. Shahidehpour, "Short-term unit commitment expert system", Electric Power Systems Research, 20 (1990), pp. 1-13.
[18] S. Saneifard, N. R. Prasad, and H. A. Smolleck, "A fuzzy logic approach to unit commitment", IEEE Transactions on Power Systems, 12 (1997), pp.988-995.
[19] A. H. Mantawy, Y. L. Abdel-Magid, and S. Z. Selim, "Integrating genetic algorithms, tabu search and simulated annealing for the unit commitment problem," IEEE Transactions Power Systems, 14 (1999), pp. 829-836.
[20] S. J. Huang, "Enhancement of hydroelectric generation scheduling using ant colony system based optimization approaches", IEEE Transactions on Energy Conservation, 16 (2001), pp. 296-301.
[21] A. J. Wood and B. F. Wollenburg, "Power Generation Operation and Control", John Wiley and Sons, New York, 1996.
[22] R. Mukerji, 2010. "NYISO Day-ahead unit commitment design," Technical Conference on Unit Commitment Software, Federal Energy Regulatory Commission, June 2, 2010.
[23] A. Ott, "Unit commitment in PJM", Technical Conference on Unit Commitment Software, Federal Energy Regulatory Commission, June 2, 2010.
[24] M. Rothleder, "Unit commitment in the CAISO market", Technical Conference on Unit Commitment Software, Federal Energy Regulatory Commission, June 2, 2010.
[25] Energy Policy Act. EPAct 2005, p. 1138.
[26] J. D. Glover, M. S. Sarma, and T. J. Overbye, "Power Systems Analysis and Design", Thomson Learning, Toronto, 2008.
[27] H. Glavitsch and R. Bacher, "Optimal power flow algorithms", in Volume 41 of Analysis and Control System Techniques for Electric Power Systems. ACADEMIC Press Inc., 1991.
[28] E. Acha, C.R. Fuerte-Esquivel, H. Ambriz-Perez, and C. Angeles Camacho, eds., "FACTS: Modelling and Simulation in Power Networks", John Wiley & Sons, 2005.
[29] D. Gan, R. Thomas, and R. Zimmerman, "Stability-constrained optimal power flow", IEEE Transactions on Power Systems, 15 (2000), pp. 535-540.
[30] Q. Jiang and G. Geng, "A reduced-space interior point method for transient stability constrained optimal power flow", IEEE Transactions on Power Systems, 25 (2010), pp. 1232-1240.
[31] R. A. Jabr, "Radial distribution load flow using conic programming", IEEE Transactions on Power Systems, 21 (2006), pp. 1458–1459.
[32] R. A. Jabr, "Optimal power flow using an extended conic quadratic formulation", IEEE Transactions on Power Systems, 23 (2008), pp. 1000-1008.

[33] D. W. Wells, “Method for economic secure loading of a power systems”, in Proceedings of IEEE, 115 (1968), pp. 606-614.
[34] G. C. Contaxis, C. Delkis, and G. Korres, “Decoupled optimal power flow using linear or quadratic programming”, IEEE Transactions on Power Systems, PWRS-1 (1986), pp. 1-7.
[35] R. A. M. van Amerongen, “Optimal power flow solved with sequential reduced quadratic programming”, Electrical Engineering, 71 (1988), pp. 213-219.
[36] W. Min and L. Shengsong, “A trust-region interior-point algorithm for optimal power flow problems”, International Journal of Electrical Power & Energy Systems, 27 (2005), pp. 293–300.
[37] A. Sousa and G. Torres, “Robust optimal power flow solution using trust region and interior-point methods”, IEEE Transactions on Power Systems, (to appear).
[38] D. I. Sun, B. Ashley, B. Brewer, A. Hughes, and W. F. Tinney, “Optimal power flow by Newton approach,” IEEE Transactions on power Apparatus and systems, Vol. PAS-103, 10 (1984), pp. 2864-2875.
[39] J. Santos and G. da Costa, “Optimal power flow solution by Newton’s method applied to an augmented Lagrangian function”, Generation, Transmission and Distribution, IEE Proceedings, 142 (1995), pp. 33-36.
[40] E. C. Baptista, E. A. Belati, and G. R. da Costa, “Logarithmic barrier-augmented Lagrangian function to the optimal power flow problem”, International Journal of Electrical Power & Energy Systems, 27 (2005), pp. 528-532.
[41] H. Wang, C. E. Murillo-Sánchez, R. D. Zimmerman, and R. J. Thomas, "On computational issues of market-based optimal power flow", IEEE Transactions on Power Systems, 22 (2007), pp. 1185-1193.
[42] R. A. Jabr, “A primal-dual interior-point method to solve the optimal power flow dispatching problem”, Optimization and Engineering, 4 (2003), pp. 309–336.
[43] F. Capitanescu, M. Glavic, D. Ernst, and L. Wehenkel, “Interior-point based algorithms for the solution of optimal power flow problems”, Electric Power Systems Research, 7 (2007), pp. 508-517.
[44] H. D. Chiang, B. Wang, and Q. Y. Jiang, “Applications of trust-tech methodology in optimal power flow of power systems”, in Optimization in the Energy Industry, J. Kallrath, P. M. Pardalos, S. Rebennack, and M. Scheidt, eds., Energy Systems, Springer Berlin Heidelberg, 2009, pp. 297-318.
[45] Q. Jiang, G. Geng, C. Guo, and Y. Cao, “An efficient implementation of automatic differentiation in interior point optimal power flow”, IEEE Transactions on Power Systems, 25 (2010), pp. 147-155.
[46] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, “An algorithm for nonlinear optimization using linear programming and equality constrained subproblems”, Mathematical Programming, 100 (2004), pp. 27-48.
[47] O. Alsac, J. Bright, M. Prais, and B. Stott, “Further developments in LP-based optimal power flow”, IEEE Transactions on Power Systems, 5 (1990), pp. 697-711.

[48] F. Bouffard, F. Galiana, and J. M. Arroyo, "Umbrella contingencies in security-constrained optimal power flow", in 15th Power Systems Computation Conference (PSCC 05), Liege, Belgium, Aug 2005.
[49] D. Ernst, D. Ruiz-Vega, M. Pavella, P. M. Hirsch, and D. Sobajic, "A unified approach to transient stability contingency filtering, ranking an assessment", IEEE Transactions on Power Systems, 16 (2001), pp. 435-443.
[50] A. Monticelli, M. V. F. Pereira, and S. Granville, "Security-constrained optimal power flow with post-contingency corrective rescheduling", IEEE Transactions on Power Systems, 2 (1987), pp. 175-180.
[51] Y. Li and J. D. McCalley, "Decomposed SCOPF for improving efficiency", IEEE Transactions on Power Systems, 24 (2009), pp. 494-495.
[52] A. M. Geoffrion, "Generalized Benders decomposition", Journal of Optimization Theory and Applications, 10 (1972), pp. 237-260.
[53] F. Capitanescu and L. Wehenkel, "A new iterative approach to the corrective security-constrained optimal power flow problem", IEEE Transactions on Power Systems, 23 (2008), pp. 1533-1541.
[54] Y. Li, "Decision making under uncertainty in power system using Benders decomposition", PhD thesis, Iowa State University, Iowa, 2008.
[55] S. M. Hadfield, "On the LU Factorization of Sequences of Identically Structured Sparse Matrices within a Distributed Memory Environment", PhD thesis, University of Florida, Gainsville, FL, 1994.
[56] J. M. T. Alves, C. L. T. Borges, and A. L. O. Filho, "Distributed security constrained optimal power flow integrated to a DSM based energy management system for real time power systems security control", in VECPAR'06: Proceedings of the 7th international conference on High performance computing for computational science, Berlin, Heidelberg, 2007, Springer-Verlag, pp. 131-144.
[57] M. Rodrigues, O. R. Saavedra, and A. Monticelli, "Asynchronous programming model for the concurrent solution of the security constrained optimal power flow problem", IEEE Transactions on Power Systems, 9 (1994), pp. 2021-2027.
[58] N. Karmarkar, "A new polynomial time algorithm for linear programming", Combinatorica, 4 (1984), pp. 373-395.
[59] M. H. Wright, "The interior-point revolution in optimization: history, recent developments, and lasting consequences", Bull. Amer. Math. Soc., 4 (2005), pp. 39-56
[60] R. Fletcher and S. Leyffer, "Nonlinear programming without a penalty function", Mathematical Programming, 91 (2002), pp. 239-269.
[61] R. Fletcher, S. Leyffer, and Ph. L. Toint, "On the global convergence of a filter-SQP algorithm", SIAM Journal on Optimization, 13 (2002), pp. 44-59.
[62] A. Waechter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming", Mathematical Programming, 106 (2006), pp. 25-57.
[63] A. Ben-Tal, A. Nemirovski, "Lectures on Modern Convex Optimization", MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2001.

[64] B. Borchers and J. G. Young, "Implementation of a primal-dual method for SDP on a shared memory parallel architecture", Computational Optimization and Applications, 37 (2007), pp. 355-369.
[65] J.F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones", Optimization Methods and Software 11-12 (1999) 625-653, http://sedumi.ie.lehigh.edu/
[66] K. C. Toh, M. J. Todd, and R. H. Tutuncu, "On the implementation and usage of SDPT3 - a Matlab software package for semidefinite-quadratic-linear programming, version 4.0", Handbook of semidefinite, cone and polynomial optimization: theory, algorithms, software and applications, M. Anjos and J.B. Lasserre eds., June 2010, http://www.math.nus.edu.sg/~mattohkc/sdpt3.html
[67] S. J. Benson, Y. Ye, and X. Zhang, "Solving large-scale sparse semidefinite programs for combinatorial optimization", SIAM Journal on Optimization, 10(2) (2000), pp. 443-461, http://www.mcs.anl.gov/hs/software/DSDP
[68] M. Yamashita, K. Fujisawa and M. Kojima, "Implementation and Evaluation of SDPA 6.0 (SemiDefinite Programming Algorithm 6.0)", Optimization Methods and Software, 18 (2003), pp. 491-505, http://sdpa.indsys.chuo-u.ac.jp/sdpa/
[69] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming", SIAM Journal on Optimization, 10 (2000), pp. 673-696.
[70] C. Helmberg and K. C. Kiwiel, "A spectral bundle method with bounds", Mathematical Programming, 93(2) (2002), pp. 173-194.
[71] J. Nocedal and S. J. Wright, "Numerical Optimization", Springer, New York, 2006.
[72] C. M. Chin and R. Fletcher, "On the global convergence of an SLP-filter algorithm that takes EQP steps", Mathematical Programming, 96 (2003), pp. 161-177.
[73] N. Grudinin, "Reactive power optimization using successive quadratic programming method", IEEE Transactions on Power Systems, 13 (1998), pp. 1219-1225.
[74] P. E. Gill, W. Murray, and M. A. Saunders, "SNOPT: an SQP algorithm for large-scale constrained optimization", SIAM Review, 47 (2005), pp. 99-131.
[75] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, "Trust-Region Methods", SIAM, Philadelphia, 2000.
[76] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds", SIAM Journal on Numerical Analysis, 28 (1991), pp. 545-572.
[77] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, "A globally convergent Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds", Mathematics of Computation, 28 (1997), pp. 261-288.
[78] R. Andreani, E. Birgin, J. Martinez, and M. Schuverdt, "On augmented Lagrangian methods with general lower-level constraints", SIAM Journal on Optimization, 18 (2007), pp. 1286-1309.
[79] R. Andreani, E. Birgin, J. Martinez, and M. Schuverdt, "Augmented Lagrangian methods under the constant positive linear dependence constraint qualification", Mathematical Programming, 111 (2008), pp. 5-32.

[80] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, (2010). LANCELOT webpage: http://www.cse.scitech.ac.uk/nag/lancelot/lancelot.shtml
[81] S. Leyffer and A. Mahajan, "Software for nonlinearly constrained optimization", Preprint ANL/MCS-P1768-0610, June 2010.
[82] H. Mittelmann (2010), "Decision Tree for Optimization Software", http://plato.asu.edu/bench.html
[83] S. Leyffer, "Integrating SQP and branch-and-bound for mixed integer nonlinear programming", Computational Optimization and Applications, 18 (2001), pp. 295-309.
[84] A. R. Conn, K. Scheinberg, and L. N. Vicente, "Introduction to Derivative-Free Optimization", SIAM, Philadelphia, PA, USA, 2009.
[85] A. Griewank and G. F. Corliss, eds., "Automatic Differentiation of Algorithms: Theory, Implementation, and Application", SIAM, Philadelphia, PA, 1991.
[86] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing", Science, 220 (1983), pp. 671–680
[87] M. Goldberg and T. Spencer, "A new parallel algorithm for the maximal independent set problem", SIAM Journal on Computing, 18 (1989), pp. 419-427.
[88] J. A. Nelder and R. Mead, "A simplex method for function minimization", The Computer Journal, 8 (1965), pp. 308-313.
[89] K. I. M. Mckinnon, "Convergence of the Nelder-Mead simplex method to a nonstationary point", SIAM Journal on Optimization, 9 (1999), pp. 148-158.
[90] Y. del Valle, G. K. Venayagamoorthy, S. Mohagheghi, J. C. Hernandez, and R. G. Harley, "Particle swarm optimization: Basic concepts, variants and applications in power systems", IEEE Transactions on Evolutionary Computation, 12 (2008), pp. 171-195.
[91] L. F. Wang, and C. Singh, "Multi-objective stochastic power dispatch through a modified particle swarm optimization algorithm", in Proc. IEEE Swarm Intell. Symp., 2006, pp. 128–135.
[92] A. George and J. W.-H. Liu. "Computer Solution of Large Sparse Positive Definite Systems", Prentice-Hall, NJ, 1981.
[93] I. S. Duff, A. M. Erisman, and J. K. Reid, "Direct Methods for Sparse Matrices", Oxford University Press, Oxford, UK, 1990.
[94] J. W. Demmel, M. T. Heath, and H. A. van der Vorst, "Parallel numerical linear algebra", Acta Numerica (1993), pp. 111-197.
[95] A. Gupta, G. Karypis, and V. Kumar, "Highly scalable parallel algorithms for sparse matrix factorization", IEEE Transactions on Parallel and Distributed Systems, 8 (1997), pp. 502-520.
[96] A. Gupta and V. Kumar, "A scalable parallel algorithm for sparse matrix factorization", Technical Report 94-19, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1994.
[97] J. W. H. Liu, "The multifrontal method for sparse matrix solution: Theory and practice", SIAM Review, 34 (1992), pp. 82-109.

[98] A. Gupta, S. Koric, and T. George, "Sparse matrix factorization on massively parallel computers" In SC09 Proceedings.
[99] J. W. Demmel, J. R. Gilbert, and X. S. Li, "An asynchronous parallel supernodal algorithm for sparse Gaussian elimination", SIAM Journal on Matrix Analysis and Applications, 20 (1999), pp. 915-952.
[100] P. R. Amestoy, I. S. Duff, J. Koster, and J. Y. L'Excellent, "A fully asynchronous multifrontal solver using distributed dynamic scheduling", SIAM Journal on Matrix Analysis and Applications, 23(1) (2001), pp. 15-41.
[101] P. R. Amestoy, I. S. Duff, and J. Y. L'Excellent, "Multifrontal parallel distributed symmetric and unsymmetric solvers", Computational Methods in Applied Mechanical Engineering, 184 (2000), pp. 501-520.
[102] A. Gupta, "Improved symbolic and numerical factorization algorithms for unsymmetric sparse matrices", SIAM Journal on Matrix Analysis and Applications, 24 (2002), pp. 529-552.
[103] A. Gupta, "A shared- and distributed-memory parallel general sparse direct solver", Applicable Algebra in Engineering, Communication, and Computing, 18 (2007), pp. 263-277.
[104] Y. Saad, "Iterative Methods for Sparse Linear Systems", 2nd edition. SIAM, 2003.
[105] M. Benzi, "Preconditioning techniques for large linear systems: A survey", Journal of Computational Physics, 182(2002), pp. 418-477.
[106] D. Hysom and A. Pothen, "A scalable parallel algorithm for incomplete factor preconditioning", SIAM Journal on Scientific Computing, 22 (2000), pp. 2194-2215.
[107] G. Karypis and V. Kumar, "Parallel threshold-based ILU factorization", Technical Report TR 96- 061, Department of Computer Science, University of Minnesota, 1996.
[108] M. J. Grote and T. Huckle, "Parallel preconditioning with sparse approximate inverses", SIAM Journal on Scientific Computing, 18(1996), pp. 838-853.
[109] E. Chow, "A priori sparsity patterns for parallel sparse approximate inverse preconditioners", SIAM Journal on Scientific Computing, 21 (2000), pp. 1804-1822.
[110] E. Chow, "Parallel implementation and practical use of sparse approximate inverse preconditioners with a priori sparsity patterns", International Journal of High Performance Computing Applications, 15 (2001), pp. 56–74.
[111] E. Chow, R.D. Falgout, J.J. Hu, R.S. Tuminaro, and U.M. Yang, "A survey of parallelization techniques for multigrid solvers", In Michael A Heroux, Padma Raghavan, and Horst D. Simon, editors, Parallel Processing for Scientific Computing, SIAM 2006.
[112] G. Karypis and V. Kumar, "ParMETIS: Parallel graph partitioning and sparse matrix ordering library", Technical Report TR 97-060, Department of Computer Science, University of Minnesota, 1997.
[113] G. Karypis and V. Kumar, "Parallel algorithms for multilevel graph partitioning and sparse matrix ordering", Journal of Parallel and Distributed Computing, 48 (1998), pp. 71-95.

[114] D. Bozdag, A. H. Gebremedhin, F. Manne, E. G. Boman, and U. V. Catalyurek, "A framework for scalable greedy coloring on distributed memory parallel computers", Journal of Parallel and Distributed Computing, 68(4) (2008), pp. 515–535.
[115] Z. Hu, X. Wang, and G. Taylor, "Stochastic optimal reactive power dispatch: Formulation and solution method", International Journal of Electrical Power & Energy Systems, 32 (2010), pp. 615-621.
[116] N. Alguacil and A. Conejo, "Multiperiod optimal power flow using Benders decomposition", IEEE Transactions on Power Systems, 15 (2000), pp. 196-201.
[117] J. F. Benders, "Partitioning procedures for solving mixed-variables programming problems", Numerische Mathematik, 4 (1962), pp. 238-252.
[118] V. Quintana, G. Torres, and J. Medina-Palomo, "Interior-point methods and their applications to power systems: a classification of publications and software codes", IEEE Transactions on Power Systems, 15 (2000), pp. 170-176.
[119] X. Ding, X. Wang, and Y. H. Song, "Interior point cutting plane method for optimal power flow, IMA Journal of Management Mathematics", 15 (2004), pp. 355-368.
[120] F. Capitanescu and L. Wehenkel, "Improving the statement of the corrective security-constrained optimal power flow problem", IEEE Transactions on Power Systems, 22 (2007), pp. 887-889.
[121] L. T. Biegler, P. Bonami, A. R. Conn, G. Cornuejols, I. E. Grossmann, C. D. Laird, J. Lee, A. Lodi, F. Margot, N. Sawaya, and A. Waechter , "An algorithmic framework for convex mixed integer nonlinear programs", Special issue of Discrete Optimization in memory of George B. Dantzig (1914-2005),Guest Editors: E. Balas, A. Hoffman, and T. McCormick, Discrete Optimization, 5 (2008), pp. 186-204.
[122] B. H. Kim and R. Baldick, "A comparison of distributed optimal power flow algorithms", IEEE Transactions on Power Systems, 15 (2000), pp. 599-604.
[123] M. Zhang and Y. Guan, "Two-stage robust unit commitment problem." Optimization online, 2009.
[124] H.B. Gooi, D.P. Mendes, K.R.W. Bell, D.S. Kirschen, "Optimal scheduling of spinning resrve," IEEE Transactions on Power Systems, 14 (1999), 1485-1492.

Chapter 3 References

[1] NERC, "Real-time application of synchrophasors for improving reliability", August 15, 2010.
[2] Federal Energy Regulatory Commission, "Assessment of demand response and advanced metering," Staff Report, Dec 2008.
[3] H. Li, L. Treinish, J. Hosking, "A statistical model for risk management of electric outage forecasts", IBM Journal of Res. & Dev. 54(3), May-June 2010, pp. 8:1-8:11.
[4] T. Niimura, "Forecasting techniques for deregulated electricity market prices – extended survey," Power Systems Conf. and Expo., 2006.

[5] U.S.-Canada Power System Outage Task Force , "Final report on the August 14, 2003 blackout in the United States and Canada: Causes and Recommendations," April 5, 2004.
[6] Electric Power Research Institute, "Smart grid program overview," 2010.
[7] J. Dean, S. Ghemawat, "MapReduce: A flexible data processing tool", Communications of the ACM, 53(1), 2010.
[8] H. Liu, D. Orban, "MapReduce: a MapReduce implementation on top of a cloud operating system."
[9] T. White, "Hadoop, The Definitive Guide," O'Reilly Press, 2009.
[10] L. Neumeyer, B. Robbins, A. Nair, A. Kesari, "S4: Distributed stream computing platform", International Workshop on Knowledge Discovery Using Cloud and Distributed Computing Platforms, KDCloud 2010.
[11] Netezza White Paper, "The Netezza data appliance architecture: a platform for high performance data warehousing and analytics". Netezza Corp., Marlborough, MA, 2009.
[12] P. Bakkum and K. Skadron. "Accelerating SQL database operations on a GPU with CUDA," 3rd Workshop on General-Purpose Computation on Graphics Processing Units, Mar 2010.
[13] EPRI, "Transmission fast simulation and modeling (T-FSM) – architectural requirements," 2005.
[14] NASPI, "Synchrophasor technology roadmap," Mar 13, 2009.
[15] J. Bank, O. Omitaomu, Y. Liu, "Visualization and classification of power system frequency data streams", 2009 IEEE Conf on Data Mining Workshops, pp.650-655.
[16] I. Monedero, C. Leon, J. Ropero, A. Garcia, J. Elena, J. Montano, "Classification of electrical disturbances in real time using neural networks", IEEE Transactions on Power Delivery, 2(3), July 2007, pp. 1288-1295.
[17] S. Rovnyak, S. Kretsinger, J. Thorp, D. Brown, "Decision trees for real-time transient stability prediction", IEEE Transactions on Power systems, 9(3), August 1994, pp.1417-1426.
[18] C. DeMarco, "Situational awareness: singular value methods for PMU data interpretation", PSERC Seminar, March 2010.
[19] A. Tiranuchit, R. Thomas, "A posturing strategy against voltage instabilities in electric power systems", IEEE Transactions on Power Systems, 3(1), 1988.
[20] G. Liu, V. Venkatasubramanian, "Oscillation monitoring from ambient PMU measurements by frequency domain decomposition", IEEE International Symposium of Circuits and Systems, 2008, pp.2821-2824.
[21] X. Liu, Y. Li, Z. Liu, Z. Huang, Y. Miao, Q. Jun, Y. Jiang, W. Chen, "A novel fast transient stability prediction method based on PMU", Power & Energy Society General Meeting, 2009, pp.1-5.
[22] NERC Steering Group, "Technical analysis of the August 14, 2003 blackout", July 13, 2004.
[23] J. Minkel, "The 2003 Northeast blackout – five years later", Scientific American, August 13, 2008.

[24] A. J. Conejo, M. A. Plazas, R. Espínola, A. B. Molina, "Day-ahead electricity price forecasting using the wavelet transform and ARIMA models," IEEE Transactions on Power Systems, 20(2), May 2005.
[25] D. Shasha, Y. Zhu, "High Performance Discovery in Time Series," Springer, 2004.
[26] J. Lin, E. Keogh, L. Wei, S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series", Data Mining and Knowledge Discovery Journal, 2007, pp. 107-144.
[27] A. Aho, M. Corasick, "Efficient string Matching: an aid to bibliographic search", CACM 18(6), 1975, pp.333-340.
[28] A. Khotanzad, R. Afkhami-Rohani, T-L. Lu, A. Abaye, M. Davis, D. Maratukulam, " ANNSTLF – a neural-network-based electric load forecasting system, IEEE Trans on Neural Networks, Vol 8, No 4, July 1997.
[29] X. Guo, Z. Chen, H. Ge, Y. Liang, "Short term load forecasting using neural networks with principal component analysis", 3rd Intl Conf on Machine Learning and Cybernetics, Aug. 2004.
[30] H. Zhao, "A new state estimation model of utilizing PMU measurements", International Conference on Power System Technology, 2006, pp.1-5.
[31] A. P. S. Meliopoulos, G. J. Cokkinides, F. Galvan, B. Fardanesh, P. Myrda, "Advances in the SuperCalibrator concept – practical implementations," Hawaii Intl. Conf. on System Sciences, 2007.
[32] Serial ATA International Organization, "SATA-IO Releases SATA Revision 3.0 Specification". Press release, May 27, 2009, http://www.sata-io.org/documents/SATA-Revision-3.0-Press-Release-FINAL-052609.pdf.
[33] P. Siebel, "Practical Common Lisp," Apress, Apr 2005.
[34] C. Chu, S. Kim, Y. Lin, Y. Yu, G. Bradski, A. Ng, "Map-Reduce for machine learning on multicore", Neural Information Processing System NIPS 2006, pp.281-288.
[35] C. Bisciglia, "The smart grid: Hadoop at the Tennessee Valley Authority (TVA)", Jeun 2, 2009, http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/.
[36] K. Martin, D. Hamai, M. Adamiak, et.al, "Exploring the IEEE standard C37.1-2005" synchrophasors for power systems", IEEE Transactions on Power Delivery, 3(4), Oct. 2008, pp.1805-1811.
[37] H. Wu, B. Salzberg, D. Zhang, "Online event-driven subsequence matching over financial data streams," ACM SIGMOD Intl. Conf. on Management of Data, 2004.
[38] B. Gedik, H. Andrade, K.L. Wu, P. Yu, M. Doo,"SPADE: The System S declarative stream processing engine", SIGMOD 2008, pp. 1123-1133.
[39] C. Ballard, et. al, "IBM InfoSphere Streams, harnessing data in motion", Sep. 2010.
[40] M. Stonebreaker, U. Cetintemel, "One size fits all: an idea whose time has come and gone", Intl. Conf. on Data Engineering, ICDE 2005.
[41] D. Talia, P. Trunfio, "Systems and techniques for distributed and stream data mining", CoreGrid Technical Report, TR-0045, July 4, 2006.

[42] V. Jalili-Marandi, V. Dinavahi, "SIMD-based large-scale transient stability simulation on the graphics processing unit," IEEE Transactions on Power Systems, 25(3), Aug. 2010.
[43] J. H. Ahn, W. J. Dally, B. Khailany, U. J. Kapasi, A. Das. Evaluating the imagine stream architecture," Annual International Symposium on Computer Architecture, June 2004.
[44] S. Muthukrishnan, "Data streams: algorithms and applications," Now Publishers, 2005.
[45] G. Hulten, L. Spencer, P. Domingos, "Mining time changing data streams", ACM Conf. on Knowledge Discovery and Data Mining, 2001.
[46] R. Reddi, J. Hazra, D. Seetharam, A. Sinha, "Stream computing based transient stability prediction for power grids", 2011.
[47] D. D. Chamberlin, R. F. Boyce, "SEQUEL: A structured English query language," ACM SIGFIDET Workshop on Data Description, Access and Control, 1974.
[48] T. C. Scofield, H. A. Delmerico, V. Chaudhary, G. Valente, "XtremeData dbX: An FPGA-based data warehousing appliance," Computing in Science and Engineering, 12(4), Jul/Aug 2010, pp.66-73.
[49] R. Mueller, J. Teubner, G. Alonso, "Streams on wires – a query compiler for FPGAs", Intl. Conf. on Very Large Data Bases, 2009.
[50] J. Backus, "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs," Communications of the ACM, 21(8), Aug, 1978.
[51] P. B. Volk, D. Habich, W. Lehner, "GPU-based speculative query processing for database operations". First Intl. Workshop on Accelerating Data Management Systems Using Modern Processor and Storage Architectures (ADMS'10), Sep 2010.
[52] Z. Feng and P. Li, "Multigrid on GPU: tackling power grid analysis on parallel SIMT platforms," in Proc. of IEEE/ACM Intl. Conf. on Computer-Aided Design, Nov 2008.
[53] E. C. Lin, K. Yu, R. A. Rutenbar and T. Chen, "Moving speech recognition from software to silicon: the In Silico Vox project," Proc. Intl. Conf. on Spoken Language Processing (InterSpeech2006), Sep 2006.
[54] D. E. Bakken, C. H. Hauser, H. Gjermundrød, A. Bose, "Towards more flexible and robust data delivery for monitoring and control of the electric power grid," Technical Report EECS-GS-009, Elec. Engg. and Comp. Sc., Washington State University, May 2007.
[55] Parallel Virtual File System, http://www.pvfs.org
[56] W. Tantisiriroj, S. Patil, and G. Gibson, "Crossing the chasm: sneaking a parallel file system into Hadoop," SC08 Petascale Data Storage Workshop, 2008.
[57] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar, M. Shah, and R. Tewari, "Cloud analytics: do we really need to reinvent the storage stack?," HotCloud'09 - 1st USENIX Workshop on Hot Topics in Cloud Computing, June 2009.

[58] A. Monticelli, "Electric power system state estimation," Proc. of the IEEE, 88(2), Feb 2000.
[59] Pacific Northwest Smart Grid Demonstration Project, http://www.pnwsmartgrid.org.
[60] L. Hall, "Pacific Northwest Smart Grid Demonstration Project," Presentation to NWESS, Feb 18, 2010.

Chapter 4 References

[1] Zhenyu Huang and Nieplocha, J., "Transforming power grid operations via High Performance Computing," Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE, Pittsburgh, PA, 2008
[2] http://www.fp7-pegase.eu
[3] U.S.-Canada Power System Outage Task Force , "Final report on the August 14, 2003 blackout in the United States and Canada: Causes and Recommendations," April 5, 2004.
[4] October 7-8, 2009, NASPI Work Group Meeting in Chattanooga, Tennessee, http://www.naspi.org/meetings/archive/meetingarchive.stm
[5] Z. Feng and P. Li, "Multigrid on GPU: tackling power grid analysis on parallel SIMT platforms," in Proc. of IEEE/ACM Intl. Conf. on Computer-Aided Design, Nov 2008.
[6] S. Sanders, J. Noworolski, X. Liu, and G. Verghese, "Generalized averaging method for power conversion circuits," Power Electronics, IEEE Transactions on, vol. 6, no. 2, pp. 251–259, 1991.
[7] V. I. Arnold, "Mathematical methods of classical mechanics." Springer-Verlag, 1982. ISBN 0-387-96890-3.
[8] Prabha Kundur, "Power System Stability and control", The EPRI Power System Engineering Series, McGraw-Hill, Inc., 1994.
[9] N. Singh, R. Prasad, H. O. Gupta, "Reduction of power system model using Balanced Realization, Routh and Pade approximation methods," Int. J. Model. Simul. Pages 57-63, 2008, ACTA Press.
[10] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, Henk van der Vorst, "Templates for the Solution of Algebraic Eigenvalue Problems," Philadephia, PA, ISBN 9780898719581.
[11] I.J.Bush, A.G.Sunderland, and G. Pringle, "Using Scalable Eigensolvers on HPCx: Case Studies," HPCx Technical Report HPCxTR0510, 2005.
[12] D. Antonelli, and C. Vomel, "PDSYEVR. ScaLAPACK's parallel MRRR algorithm for the symmetric eigenvalue problem," Lapack working note 168, (2005).
[13] A.G.Sunderland and E. Breitmoser, "An Overview of Eigensolvers for HPCx," HPCx Technical Report HPCxTR0312, 2003.
[14] Charles L. Fortescue, "Method of Symmetrical Co-Ordinates Applied to the Solution of Polyphase Networks." Presented at the 34th annual convention of the

AIEE (American Institute of Electrical Engineers) in Atlantic City, N.J. on 28 July 1918.

[15] Anupindi, K.; Skjellum, A.; Coddington, P.; Fox, G.; "Parallel differential-algebraic equation solvers for Power System Transient Stability Analysis," Northeast Parallel Archit. Center, Syracuse Univ., NY Proceedings of the Scalable Parallel Libraries Conference, 6-8 Oct 1993 Page(s): 240 - 244

[16] Shengtai Li, Linda Petzold, "Adjoint sensitivity analysis for time-dependent partial differential equations with adaptive mesh refinement," Journal of Computational Physics, Volume 198, Issue 1, 20 July 2004, Academic Press Professional, Inc. San Diego, CA, USA

[17] P.C. Magnusson, "Transient Energy Method of Calculating Stability," AIEE Trans, Vol 66 pp. 747-755, 1947.

[18] A.A Fouad, Vijay Vittal, "Power Systems Transient Stability Analysis using the Transient Energy Function Method," Perntice Hall, 1992.

[19] Tang, C.K., Graham, C.E., El-Kady, M., Alden, R.T.H., "Transient stability index from conventional time domain simulation," Power Systems, IEEE Transactions on, Aug 1994, Vol. 9 Issue 3, Page(s): 1524 – 1530

[20] F. C. Schweppe, J. Wildes, and D. Rom, "Power system static state estimation," Power Syst. Eng. Group, M.I.T. Rep. 10, Nov. 1968

[21] F. C. Schweppe and J. Wildes, "Power system static-state estimation, part I: Exact model," IEEE Trans. Pourer App. Syst., vol. PAS-89, pp. 120- 125, Jan. 1970.

[22] F. C. Schweppe and D. B. Rom, "Power system static-state estimation, part 11: Approximate model," IEEE Trans. Power Apparatus and Systems, Vol. PAS-89, pp. 125-130, Jan. 1970.

[23] D. M. Falco, P. A. Cooke, A.Brameller, "Power System Tracking State Estimation And Bad Data Processing", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-101, No. 2 February 1982

[24] E. Handschin, "Real-Time Data Processing Using State Estimation in Electric Power Systems," in Real-Time Control of Electric Power Systems, E. Handschin, Ed., Elsevier, Amsterdam, 1972.

[25] R. D. Massiello and F. C. Schweppe, "A Tracking Static State Estimator" , IEEE Trans. PAS, March/April 1971, pp 1025 -1033

[26] E. Handschin, F. C. Schweppe, J. Kohlas, A. Fiechter, "Bad Data Analysis for Power System State Estimation", IEEE Trans. Power App. Syst., vol. P AS-94, pp.329-337, Mar/Apr.1975.

[27] W. W. Kotiuga, "Development of a Least Absolute Value Power System Tracking State Estimator", IEEE Transactions on Power Apparatus and Systems, Vol. PAS-104, No. 5, May 1985.

[28] S. C. Tripathy, S. Chohan, "On Line Tracking State-Estimation in Power Systems", IE (I) Journal-EL, 1992.

[29] F. Chen, X. Han, Z. Pan, L. Han, "State Estimation Model and Algorithm Including PMU," DRPT2008 April, 2008, Nanjing, China, pp.1097-1102.

[30] H. Zhao, "A New State Estimation Model of Utilizing PMU Measurements", 2006 International Conference on Power System Technology, pp.1-5.
[31] Atif. S. Debs and Robert. E. Larson, "A Dynamic Estimator for tracking the state of a Power System," IEEE Transactions on Power Apparatus and Systems, Vol. PAS 89, NO.7, September-October, 1970.
[32] P. Rousseaux, Th. Van Cutsem, and T. E. Dy Liacco, "Whither dynamic state estimation," Int. J. Elect. Power, vol. 12, no. 2, pp. 105–116, Apr. 1990
[33] I. W. Slutsker, S. Mokhtari, and K. A. Clements, "Real time recursive parameter estimation in energy management sys- tems," IEEE Trans. Power Syst., vol. 11, pp. 1393–1399, Aug. 1998.
[34] R. S. Bucy, "Lectures on Discrete Time Filtering," Berlin, Germany: Springer-Verlag, 1994.
[35] Jain, A., Shivakumar, N.R., "Power system tracking and dynamic state estimation", Power Systems Conference and Exposition, 2009. PSCE '09. IEEE/PES, 2009 Page 1 – 8.
[36] Zhenyu Huang, Kevin Schneider, and Jarek Nieplocha. "Feasibility Studies of Applying Kalman Filter Techniques to Power System Dynamic State Estimation", in: Proceedings of the 8th International Power Engineering Conference, Singapore, 3-6 December 2007.
[37] J E Dennis, R B Schnabel, "Numerical methods for unconstrained optimization and nonlinear equations", Prentice-Hall, Englewood Cliffs, N. J
[38] Pueyo, A.,and Zingg, D.W.,"Efficient Newton–Krylov Solver for Aerodynamic Computations," AIAA Journal, Vol. 36, No. 11, 1998, pp. 1991– 1997.
[39] Hindmarsh, A. C., et al., "SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers," ACM Trans. Math. Software, 31:363-396, 2005
[40] Saad, Y., and Schultz, M. H., "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems," SIAM Journal on Scientific and Statistical Computing, Vol. 7, No. 3, 1986, pp. 856– 869.
[41] Dennis, J. E., "On the convergence of Broyden's method for nonlinear systems of equations," Mathematics of Computation, Vol. 25, pp. 559-567 (1971).
[42] Dennis, J. E. and More, J. J., "A characterization of superlinear convergence and its application to quasi-Newton method," Mathematics of Computation, Vol. 28, pp. 549-560 (1974).
[43] Dennis, J. E. and More, J. J., "Quasi-Newton methods, motivation and theory," SIAM Review, Vol. 19, pp. 46-89 (1977).
[44] Allgower, E. L. and Georg, K., Numerical Continuation Methods: An Introduction, Springer, New York (1990).
[45] Deuflhard, P., "Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms," Springer, New York (2004).
[46] Corliss, G., Faure, C., Griewank, A., Hascoet, L., and Naumann, U., Eds. 2002. "Automatic, Differentiation of Algorithms – From Simulation to Optimization." Springer, New York.

[47] David Guibert and Damien Tromeur-Dervout, "A Schur Complement Method for DAE/ODE Systems in Multi-Domain Mechanical Design," Domain Decomposition Methods in Science and Engineering XVII, Lecture Notes in Computational Science and Engineering, 2008, Volume 60, III, 535-541
[48] Xyce Parallel Circuit Simulator, http://xyce.sandia.gov
[49] Synopsys, "NanoSim: High Performance Full-Chip Simulation," http://www.synopsys.com/Tools/Verification/AMSVerification/CircuitSimulation/Pages/NanoSim.aspx
[50] Cadence, "Virtuoso UltraSim Full-Chip Simulator," http://www.cadence.com/products/cic/UltraSim_fullchip/pages/default.aspx
[51] Dimitrios Chaniotis, and M. A. Pai, "Model Reduction in Power Systems Using Krylov Subspace Methods," IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 20, NO. 2, MAY 2005
[52] Z. Mahmood, T. Moselhy, L. Daniel, "Passive Reduced Order Modeling of Multiport Interconnects via Semidefinite Programming", IEEE Conference on Design Automation and Test in Europe (DATE), March 2010.
[53] B. N. Bond, L. Daniel, "Stable Macromodels for Nonlinear Descriptor Systems through Piecewise-Linear Approximation and Projection," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 28, no. 10, p. 1467-1480, October 2009.
[54] Byerly, R.T., Bennon, R.J., Sherman, D.E., "Eigenvalue Analysis of Synchronizing Power Flow Oscillations, Power Apparatus and Systems," IEEE Transactions on, Jan. 1982, Vol. PAS-101 Issue:1 page(s): 235 – 243.
[55] Sancha, J.L., Perez-Arriaga, I.J., "Selective modal analysis of power system oscillatory instability," Power Systems, IEEE Transactions on, May 1988 Vol. 3 Issue:2 page(s): 429 – 438.
[56] Sauer, P.W., Rajagopalan, C., Pai, M.A., "An explanation and generalization of the AESOPS and PEALS algorithms [power system models]," Power Systems, IEEE Transactions on, Feb 1991, Vol. 6, Issue: 1, page(s): 293 – 299.
[57] R.B Lehoucq, D.C. Sorensen, C. Yang, "ARPACK Users' Guide: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods (Software, Environments and Tools)," Society for Industrial & Applied Mathematics, U.S. (30 April 1998)
[58] VICENTE HERNANDEZ, JOSE E. ROMAN, and VICENTE VIDAL, "SLEPc: A Scalable and Flexible Toolkit for the Solution of Eigenvalue Problems", ACM Transactions on Mathematical Software, Vol. 31, No. 3, September 2005.
[59] Gerard L. G. Sleijpen, Henk A. Van der Vorst, "A Jacobi–Davidson Iteration Method for Linear Eigenvalue Problems," SIAM REVIEW, Society for Industrial and Applied Mathematics Vol. 42, No. 2, pp. 267–293
[60] F. E. Daum and J. Huang, "The Curse of Dimensionality for Particle Filters," Proc. IEEE Conf. Aero., vol. 4, pp. 1979-1993, 2003.
[61] A. Gelb, "Applied Optimal Estimation," The M.I.T Press, 1974.

[62] E. A. Wan and R. van der Merwe, "The Unscented Kalman Filter for Nonlinear Estimation," Proc. The IEEE AS-SPCC Symposium, 2000.
[63] K. Ito and K. Xiong, "Gaussian Filters for Nonlinear Filtering," IEEE Trans. Auto. Cont., vol. 45, pp. 910927, 2000.
[64] E. Kalnay, "Atmospheric modeling, data assimilation and predictability," Cambridge University Press, 2003.
[65] G. Evensen, "Advanced Data Assimilation for Strongly Nonlinear Dynamics," Monthly Weather Review, vol. 125, pp. 1342-1354, 1997.
[66] G. Evensen, "Sequential Data Assimilation for Nonlinear Dynamics: The Ensemble Kalman Filter," In Ocean Forecasting: Conceptual basis and applications, edited by N. Pinardi and J. D. Woods, Springer-Verlag Berlin Heidelberg, 2002.
[67] Usman A. Khan and José M. F. Moura, "Distributing the Kalman Filter for Large-Scale Systems," IEEE Transactions on Signal Processing, Vol. 56, No. 10, October 2008.
[68] Hsiao-Dong Chiang, "Direct Methods for Stability Analysis of Electric Power Systems: Theoretical Foundation, BCU Methodologies, and Applications," Wiley, ISBN-10: 0470484403, December 2010.

Discussant Narrative

Framework for Large-Scale Modeling and Simulation of Electricity Systems for Planning, Monitoring, and Secure Operations of Next-generation Electricity Grids

Loren Toole
Los Alamos National Laboratory

The authors present a well-researched, comprehensive vision of future smart grid M&S including online and offline applications and their ontology; real-time speeds, data volume and scalability are given special attention; recommendations for methods development based on a critique of state-of-art are "grand challenge" in nature.

There are four chapters in this paper, 1) introduction, 2) security constrained unit commitment and economic dispatch with stochastic analysis, 3) massive data processing, and 4) large scale M&S of power grids. I mainly address chapter 3.

My first observation will focus on a possible differentiation of need not suggested by the authors. It is important to understand the relative value of each M&S level versus specific system level needs. This discussion will address the common perception that all levels of Smart Grid will demand similar M&S, which I contend is not the case. This issue has significant implications also for computational scale and speed, which is another subjective dimension that we should explore if time is available.

My second observation addresses the authors' claim that power flow solution accuracy, speed and complexity requires parallel AC solvers. For problems involving grid security at bulk transmission/generation level, there is a significant need to improve solution techniques by using parallel direct methods. For problems involving grid security at distribution or sub transmission level the likelihood of non convergence is reduced. A key determinant could be topology, the underlying structure of nodes and links; this can be broadly classified by looped or radial topology. Parallelization will allow the extent of infeasible solution space to be more quickly defined, but it will probably not result in more robust power flow solutions which overcome topology induced failures.

My third observation is intended to highlight somewhat related but conflicting Smart Grid M&S needs i.e. the potential to operate in a "fuzzy" decision space. I offer an example quoted by the authors that includes both static and dynamic aspects of grid response following a contingency. Operational requirements such as NERC A (least severe) to NERC D (most severe) level events can serve as guidance to further organize M&S requirements with more sophistication regarding the risk of not performing actionable decisions, quickly and precisely.

My fourth observation is perhaps the most open ended of all posed. I am a proponent of not increasing the complexity of M&S for utility applications unless indicated by need but with insight into the full scope of issues involved. The authors' claim that stochastic analysis is required for renewable integration needs more consideration. This is in fact a less demanding issue below specified levels of renewable penetration; the typical utility rule is below 20-30% capacity penetration. While it is likely that some anecdotal evidence will be presented pro and con in terms of the need for stochastic modeling, I am hoping this will mainly stimulate the discussion, not resolve the issue.

I offer two questions for discussion, based on my first and second observations:
What system needs for distribution, sub transmission and transmission-level Smart Grid will change the relative importance (or inclusion) of these M&S factors? (1) failure of N-k contingency analysis; (2) stochastic (variations); (3) solve larger instances of UC and ED.
Will parallel direct M&S methods ensure consistent improvement of solution robustness using AC power flow solvers? Suggest examples derived by current non parallel methods which represent exceptions.

Recorder Summary

Framework for Large-Scale Modeling and Simulation of Electricity Systems for Planning, Monitoring, and Secure Operations of Next-generation Electricity Grids

Jeff Dagle
Pacific Northwest National Laboratory

Authors' response to discussant presentation:

- Stochastic analysis is not only useful for renewable integration or demand response, but also useful for managing pricing signal uncertainties.
- Stochastic analysis may also need to consider the temporal and spatial correlations among loads.
- R&D is needed for model calibration and stochastic optimization. Industry should determine the priority of the problems that are most relevant to them, and researchers should not dictate this. The goal of the paper is to propose a number of challenging research problems.
- Some of those challenges are interrelated. For example, one reason for the need of N-k contingency analysis is related to stochastic optimization that may push the power system operating closer to its limit – need more refined analysis. Also, need to rank those N-k contingences to avoid high complexity.
- Most smart grid investment today is installing instrumentation to enhance the capability to gather a lot of data, but the key challenge going forward is making use of that data.
- Encouraged looking at data in two ways: modeless-based approach vs. model-based approaches. The former looks at data in absence of a model, good for visualization, fast reaction or control; the latter is more elaborate: for example, model augmented with measurements (like state estimation) – existing state estimation will become obsolete, would enable on-line stability analysis; this all becomes possible with data plus increased computational power.
- Massive data-notion of machine learning, data models as streaming data. Scenario: learn from data, feed this back into the models in real-time to calibrate the models.
- Wavelets (analyze the data at certain frequencies, good for data compression).

General Discussion

- Reformulating historical problem formulations to better leverage existing and emerging computational capabilities is a good idea, but the specifics of how to do that are still unknown. The concepts presented in this paper represent a grand challenge.
- We should develop a set of test cases so that people outside the power industry can get their teeth into solving significant and relevant problems. For example, optimization experts can put their talents into solving the right problems. In addition to test cases, we need data. A robust dialog needs to exist between academia and industry. DOE has a role to help solve this problem.
- The traditional differences between the mathematical community and the engineering community are breaking down. For example, in the area of preconditioning. To achieve parallelism, we need to rely on the decomposition of the problem. To be effective this has to take into account the physical characteristics. Algorithms can't be developed in a vacuum, we need better problem-specific insight; robustness depending on portioning the problem, different for nonlinear. Sometimes approximations can be made that might be less than linearizing the problem. Impossible to generically formulate the problem.
- There needs to be greater clarity on capabilities and limitations associated with the state-of-the-art tools currently available to industry for these applications, some of which might be proprietary. The state-of-the-art in industry is not state-of-the-art in academia.
- While there are some similarities with the SPICE-like solution method, there are also some significant differences that may require tweaking.
- You can get significant different answers between DC and AC optimal power flow solutions, it is not simply the ease at optimizing the problem but voltage constraints and other issues can impact the final solution.
- Unit commitment without constraints is not complete, and the stochastic approach needs to include voltage stability considerations.
- With dynamic state estimation, if you are using the data from the system, you can only estimate parameters for the linear model. The nonlinear parameters can only be verified with a large disturbance that tests the parameters.
- There is a need for research to guarantee convergence for non-linear problems.
- There is a need for solving non-convex optimization of dispatch.
- There is a need to develop distributed database (vs. partitioned database) with non-consistent data.
- There is a need to compare linear with non-linear estimates.
- There is a need for optimization of reactive power dispatch.
- Feedback control is fundamentally important. Need to understand uncertainty including uncertain time delays.
- The solutions presented in this paper need consideration of how to implement in a deregulated market.

Appendix 1: Workshop Agenda

Appendix 1

Computational Needs for the Next Generation Electric Grid

Workshop Agenda

The Statler Hotel
130 Statler Drive
Cornell University
Ithaca, NY

Monday, April 18, 2011

6:30 – 8PM: Reception, Statler Hotel, Rowe Room – 2nd floor

Tuesday, April 19, 2011:

Statler Amphitheater

7:30: Continental Breakfast – Amphitheater Foyer – 1st floor

8:00: **Welcome**

Lance Collins, Cornell Dean of Engineering

8:15: **DOE Research Vision and Implementation Plans, Charge to the Group, and Expected Outcomes**

Gilbert Bindewald, Acting Deputy Assistant Secretary for Permitting, Siting, and Analysis, Office of Electricity Delivery and Energy Reliability, US Department of Energy

8:45 **Introductions and Review of Paper Session Format**

Bob Thomas, Cornell University

9:00: **Paper 1: Running Smart Grid Control Software on Cloud Computing Architectures**

Presenter/Co-authors: *Kenneth Birman*, Lakshmi Ganesh, and Robbert van Renessee, Cornell University

Paper Discussant: James Nutaro, Oak Ridge National Laboratory
Session Facilitator: Joe Eto, Lawrence Berkeley National Laboratory
Session Recorder: Ghaleb Abdulla, Lawrence Livermore National Laboratory

10:30: Break

11:00: **Paper 2: Coupled Optimization Models for Planning and Operation of Power Systems on Multiple Scales**

Presenter/Author: *Michael Ferris*, University of Wisconsin

Paper Discussant: Ali Pinar, Sandia National Laboratory
Session Facilitator: Bob Thomas, Cornell University
Session Recorder: Alejandro Dominguez-Garcia, University of Illinois at Urbana-Champaign

12:30: Lunch – Statler Hotel, Taylor Room, 2nd floor

1:30: **Paper 3: Mode Reconfiguration, Hybrid Mode Estimation, and Risk-bounded Optimization for the Next Generation Electric Grid**

Presenter/Co-authors: *Andreas Hofmann* and Brian Williams, MIT Computer Science and Artificial Intelligence Laboratory

Paper Discussant: Bernard Lesieutre, University of Wisconsin
Session Facilitator: Joe Eto, Lawrence Berkeley National Laboratory
Session Recorder: Hyung-Seon Oh, National Renewable Energy Laboratory

3:00: Break

3:30: **Paper 4: Model-based Integration Technology for Next Generation Electric Grid Simulations**

Presenter/Co-authors: *Janos Sztipanovits*, Graham Hemingway, Vanderbilt University; Anjan Bose and Anurag Stivastava, Washington State University

Paper Discussant: Henry Huang, Pacific Northwest National Laboratory
Session Facilitator: Bob Thomas, Cornell University
Session Recorder: Victor Zavala, Argonne National Laboratory

5:00: Review of plans for tomorrow - Adjourn

6:30: Reception – Johnson Museum of Art – Lobby
The Johnson Museum of Art is a short walk from the meeting site.
If it's raining transportation will be provided.

7:00: Dinner – Johnson Museum of Art – 6th Floor

Wednesday, April 20, 2011

Statler Amphitheater

8:00 Continental Breakfast – Amphitheater Foyer – 1st floor

8:30: **Paper 5: Research Needs in Multi-Dimensional, Multi-Scale Modeling and Algorithms for Next Generation Electricity Grids**

Presenter/Author: Santiago Grijalva, Georgia Institute of Technology

Paper Discussant: Roman Samulyak, Brookhaven National Laboratory
Session Facilitator: Joe Eto, Lawrence Berkeley National Laboratory
Session Recorder: Sven Leyffer, Argonne National Laboratory

10:00: Break

10:30: **Paper 6: Long Term Resource Planning for Electric Power Systems Under Uncertainty**

Presenter/Co-authors: *Sarah M. Ryan* and James D. McCalley, Iowa State University; David L. Woodruff, University of California Davis

Paper Discussant: Jason Stamp, Sandia National Laboratory
Session Facilitator: Bob Thomas, Cornell University
Session Recorder: Chao Yang, Lawrence Berkeley National Laboratory

12:00: Lunch – Statler Hotel, Taylor Room, 2nd floor

1:00: **Paper 7: Framework for Large-scale Modeling and Simulation of Electricity Systems for Planning, Monitoring, and Secure Operations of Next- generation Electricity Grids**

Presenter/Co-authors: *Jinjun Xiong*, Emrah Acar, Bhavna Agrawal, Andrew R. Conn, Gary Ditlow, Peter Feldmann, Ulrich Finkler, Brian Gaucher, Anshul Gupta, Fook-Luen Heng, Jayant R Kalagnanam Dr. Ali Koc, David Kung, Dr. Dung Phan, Dr. Amith Singhee, Dr. Basil Smith, IBM

Paper Discussant: Loren Toole, Los Alamos National Laboratory
Session Facilitator: Joe Eto, Lawrence Berkeley National Laboratory
Session Recorder: Jeff Dagle, Pacific Northwest National Laboratory

2:30: Break

3:00: **Wrap-up session to include 5-minute presentations**
Session Facilitator: Bob Thomas, Cornell University

Anyone attending the Workshop is invited to make a presentation on any topic they feel is even remotely related to the theme of the Workshop. Anyone interested in making a 5-minute presentation should make himself or herself known to Bob Thomas anytime before the end of dinner on Tuesday, April, 19, 2011.

4:30 DOE Closing Thoughts and Next Steps

Gilbert Bindewald, Acting Deputy Assistant Secretary for Permitting, Siting, and Analysis, Office of Electricity Delivery and Energy Reliability

5:00: Adjourn

Appendix 2: Written Comments Submitted after the Workshop

Appendix 2

Written Comments Submitted After the Workshop on Computational Needs for the Next Generation Electric Grid

Advancing Effectiveness of Optimization Used in the Next Generation Electric Grid

Prof. Christine Shoemaker
Cornell University

There was a considerable discussion of optimization methods in the workshop at Cornell. Given the complexity of the Electric Grid and the enormous number of decisions that need to be made, optimization methods will be an essential tool in converting a powerful computational system into a powerful smart grid.

The description of existing applications of optimization methods to operating the grid presented in the workshop implies that there is a lot of room for improvement both in the algorithms and in the breath of their implementation. For example, electric grid problems tend to be stochastic, adaptive, nonlinear (and possibly multimodal), with a large number of decision variables and many time steps in a dynamic optimization problem. In addition, answers need to be computed quickly and sometimes extremely quickly. None of the optimization methods discussed in the conference (and probably none in existence) can compute the best answers extremely quickly for stochastic, nonlinear, high dimensional, dynamic optimization problems with many time steps. The algorithms can be improved by developing new methods and more effective ways to combine existing algorithms to achieve the goals.

Part of the difficulty is that there are many levels and different types of electric grid problems that could be helped by optimization. For example, decision making problems characteristics differ over each of many different decision periods ranging from fractions of a second to decades. To overcome this problem there are two approaches which can support each other: a) identify the various classes of problems that need solutions and b)

improve optimization methods for those classes. For example, the characteristics of a class include i) how quickly a decision needs to be computed, ii) the importance of random factors and the ability to adapt quickly to them, iii) mathematical characteristics like the absence or presence of multiple local minima or discontinuities, and iv) the number of time steps into the future that should be incorporated into the decision.

A way to support the effective use of optimization is to generate and release to the public some standard test problems that represent the important classes of electric grid problems. For example, for each of the major classes of problems, there could be a several examples –one of which is easier to solve that some of the others. The optimization problems could be defined by mathematical equations or a simulation model so that the objective functions, decision variables, and constraints can be incorporated into the optimization code. The existence of these test problems would encourage optimization groups to try to figure out ways to solve the problems with more accuracy and higher wallclock speed (e.g. with parallelism in a cloud or supercomputer) and with appropriate incorporation of random and nonlinear factors.

The difficulty right now is that the optimization community at large does not have an easy way of understanding what kinds of problems are of interest to the electric grid community since even articulating the problem requires knowledge of the workings of the complex electric grid. The optimization community is always looking for test problems by which they can compare their algorithms performance on challenging publicly available problems.

There are of course interdisciplinary teams who are familiar with simulation of power grids and optimization, but this represents only a small fraction of the researchers and types of algorithms that could potentially contribute to most successfully exploiting the power of the new computational resources for the Next Generation Electric Grid. By creating clear classes of important problems and then providing equations or simulation code for each class of problem, the most efficient use of existing optimization methods and identification of ways to improve existing methods could be greatly accelerated. This would draw new people and techniques into the area of power grid optimization and enable existing groups to more efficiently focus their efforts.

Of course DOE should also fund development of optimization approaches to enhance the effectiveness of electric grids, but the creation of the clearly defined classes of problems and some publically available problems presented mathematically or as simulation code will expand the number of researchers who are able to effectively respond to these future funding opportunities and the identification of classes of problems and test problems can be expected to advance the field more quickly.

Appendix 3: Participant Biographies

Appendix 3

Computational Needs for the Next Generation Electric Grid

Workshop Participants

Ghaleb Abdulla, *Recorder*, Scientist/Tech Lead, Lawrence Livermore National Laboratory, abdulla1@llnl.gov

Ghaleb Abdulla is a computer scientist and a technical lead in the informatics group at the Center for Applied Scientific Computing. He has a diverse research background and his previous projects cover the areas of network traffic characterization and modeling, large scale scientific data management, spatial data management, large scale graph representation and querying, data centric workflows, and data mining. His latest projects are in the areas of data mining for scientific experimental data and Energy Informatics. He is a frequent panelist for the DOE office of science and NSF in the data sciences area and he served on the program committee for several related conferences. He is the editor of a special issue of the applied optics journal in the area of "applied data mining and machine learning algorithms for optics applications."

Khaled Abdul-Rahman, *Guest* - Director, Power Systems Technology Development, California ISO, kabdul@caiso.com

Khaled Abdul-Rahman received his BS (86), and MS (90) degrees in E.E from Kuwait University, his Ph.D. in Electrical Engineering from Illinois Institute of Technology (IIT), Chicago, IL in 1993. Dr. Abdul-Rahman is currently the Director of Power System Technology Development (PSTD) Group at the CAISO. Current responsibilities include overseeing the technical implementation of California's integrated forward and real-time markets technology projects, as well as the integration of the Voltage and Dynamic Stability Analysis functions to the market applications. Dr. Abdul-Rahman is a key member for developing and executing CAISO's smart grid technology vision and strategy. Dr. Abdul-Rahman has been closely involved with various different types of entities in the electric industry including academic institutions, utilities, independent system operators, power systems software vendors, Database vendor, and many engineering consulting firms. Dr. Abdul-Rahman has many IEEE transactions publications and other conference papers.

Gil Bindewald, *Organizer*, Deputy Assistant Secretary for Permitting, Siting and Analysis, Department of Energy, Office of Electricity Delivery and Energy Reliability, Gilbert.Bindewald@hq.doe.gov
Gil Bindewald is Acting Deputy Assistant Secretary for Permitting, Siting, and Analysis within the U.S. Department of Energy's (DOE) Office of Electricity Delivery and Energy Reliability. He leads the Department's efforts to provide, on a national-level, unbiased technical assistance and analysis to facilitate electricity infrastructure investment needed to deliver clean, affordable, and reliable electricity to customers. He also leads activities in the areas of Advanced Computation & Grid Modeling, Renewables Integration, and Workforce Development. Before coming to DOE, he worked as a project manager with Westinghouse Electric Corporation, and on the technical staff of Massachusetts Institute of Technology's Lincoln Laboratory.

Ken Birman, *Author*, Professor, Cornell University, ken@cs.cornell.edu
Ken Birman is the Rao Professor of Computer Science here at Cornell University, and has worked on problems involving high assurance computing for most of his career. The Isis system he developed in the 1990's played key roles in the New York Stock Exchange, French Air Traffic Control System and US Navy AEGIS, and the virtual synchrony model it supports is used in many modern cloud computing settings. Ken was the 2010 winner of the IEEE Kanai Award, is a Fellow of the ACM, and he'll be talking about the ways that high assurance, cloud computing and control of the future power grid will need to converge if we're going to achieve some of the ambitious goals that workshops like this one are laying out.

Anjan Bose, *Author*, Regents Professor, Washington State University, bose@wsu.edu
Anjan Bose is Regents Professor in the School of EECS at Washington State University. He has worked in the area of electric power engineering for over 40 years in industry, government and academia. He is a Member of the National Academy of Engineering and a Fellow of the Institute of Electrical and Electronic Engineers.

Hsiao-Dong Chiang, *Guest* - Professor, Cornell University, chiang@ece.cornell.edu
Hsiao-Dong Chiang is a Professor in the School of Electrical and Computer Engineering, Cornell University and is Founder of Bigwood Systems, Inc., Ithaca, NY 14850. His research and development interest lies nonlinear systems theory, computation, and practical applications to electric grids analysis, operation and on-line assessment and controls. His expertise includes power system stability and control, security assessments and enhancements, distribution network analysis and automation, and global optimization techniques and their engineering applications.

Jeff Dagle, *Reporter*, Chief Electrical Engineer, Pacific Northwest National Laboratory, jeff.dagle@pnl.gov
Jeff Dagle has been an electrical engineer at the Pacific Northwest National Laboratory for 22 years. He received his BS and MS degrees from Washington State University.

Currently he manages several projects in the areas of transmission reliability and security, including the North American SynchroPhasor Initiative (NASPI).

Munther Dahleh, *Guest* - Professor, MIT, dahleh@mit.edu
Munther Dahleh is a professor of electrical engineering and computer science at MIT. He is currently the acting director of the laboratory for information and decision systems. Dr. Dahleh is interested in problems at the interface of robust control, filtering, information theory, and computation which include control problems with communication constraints and distributed mobile agents with local decision capabilities. He is also interested in various problems in network science including distributed computation over noisy network as well as information propagation over complex engineering and social networks.

Gary Ditlow, *Author*, Research Staff Member, IBM, ditlow@mac.com
Gary Ditlow is a Research Staff Member at the IBM TJ Watson Research Center. He is currently working on Smart Grid data mining, analytics, and simulation. He has worked in many VLSI design areas including phase change memories, array design, logic synthesis, statistical timing, array logic minimization, circuit tuning, circuit analysis, chip placement, sparse matrix solvers, database machines, and data mining to optimize chip profit. He has received three Outstanding Technical Achievement Awards, one Research Award and 18 issued patents. He has a BSEE from the University of Maryland and an MS in Computer Science from RPI.

Alejandro Dominguez-Garcia, *Reporter*, University of Illinois at Urbana-Champaign, aledan@illinois.edu
Alejandro D. Dominguez-Garcia is an Assistant Professor in the Electrical and Computer Engineering Department at the University of Illinois, Urbana, where he is affiliated with the Power and Energy Systems area. His research interests lie at the interface of system reliability theory and control, with special emphasis on applications to electric power systems and power electronics. Dr. Dominguez-Garcia received the Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, in 2007 and the degree of Electrical Engineer from the University of Oviedo (Spain) in 2001. Dr. Dominguez-Garcia received the NSF CAREER Award in 2010.

Joe Eto, *Organizer*, Staff Scientist, Lawrence Berkeley National Laboratory, jheto@lbl.gov
Joseph H. Eto is a staff scientist at the Lawrence Berkeley National Laboratory where he manages a national lab – university – industry R&D partnership called the Consortium for Electric Reliability Technology Solutions (CERTS) http://certs.lbl.gov. Joe has been involved in the preparation of every major Department of Energy study on electricity since 2000, including the Power Outage Study Team (2000); the National Transmission Grid Study (2002); the US-Canada Final Report on the August 14, 2003 Blackout; the

DOE National Electric Transmission Congestion Study (2006 and 2009). In 2011, he completed a major study for the Federal Energy Regulatory Commission on the use of frequency response metrics to assess the reliability impacts of increased variable renewable generation. Joe has authored over 150 publications on electricity policy, electricity reliability, transmission planning, cost-allocation, demand response, distributed energy resources, utility integrated resource planning, demand-side management, and building energy-efficiency technologies and markets.

Peter Feldmann, *Author*, IBM, feldmann@us.ibm.com

Michael C. Ferris, *Author*, Professor, University of Wisconsin-Madison, ferris@cs.wisc.edu

Michael C. Ferris is Professor of Computer Sciences and Industrial and Systems Engineering at the University of Wisconsin, Madison, USA, having received his PhD from the University of Cambridge, England in 1989.Dr. Ferris' research is concerned with algorithmic and interface development for large scale problems in mathematical programming, including links to the GAMS and AMPL modeling languages, and general purpose software such as PATH, NLPEC and EMP. He has worked on several applications of both optimization and complementarity, including cancer treatment plan development, radiation therapy, video-on-demand data delivery, economic and traffic equilibria, structural and mechanical engineering. Ferris received the Beale-Orchard-Hays prize from the Mathematical Programming Society and is a past recipient of a NSF Presidential Young Investigator Award, and a Guggenheim Fellowship; he serves as co-editor of Mathematical Programming, and is on the editorial boards of SIAM Journal on Optimization, Transactions of Mathematical Software, and Optimization Methods and Software.

Rajit Gadh, *Guest* - Professor, UCLA - Smart Grid Energy Research Center, gadh@ucla.edu

Rajit Gadh is a Professor at the Henry Samueli School of Engineering and Applied Science at UCLA, and the Founding Director of the UCLA Smart Grid Energy Research Center - http://smartgrid.ucla.edu. Dr. Gadh's research interests include Smart Grid Architectures, Smart wireless communications, sense and control for Demand Response, Micro Grids and Electric Vehicle Integration, Mobile Multimedia, Wireless and RFID Middleware, and RFID/Wireless sensors for Tracking Assets. He has over 150 papers in journals, conferences and technical magazines, and, 3 patents granted. He has a Doctorate degree from Carnegie Mellon University (CMU), a Masters from Cornell University and a Bachelors degree from IIT Kanpur, and he has taught as a visiting researcher at UC Berkeley, has been a Assistant, Associate and Full Professor at University of Wisconsin-Madison, and was a visiting researcher at Stanford University for a year.

James Gallagher, Guest, NYISO

Lakshmi Ganesh, *Author*, PhD Student, Cornell University, lakshmi@cs.cornell.edu
Lakshmi Ganish is a PhD student at Cornell University, working with Dr. Ken Birman, and Dr. Hakim Weatherspoon on data center power management techniques. She has worked on low-power storage design, effective and safe power-oversubscription techniques, and fast inter-data center communication protocols, among other topics. She expects to graduate this year, and will be in the market for a research position in the area of power-aware computing.

Avi Gopstein, *Guest - DOE*, Senior Advisor, Department of Energy, avi.gopstein@science.doe.gov
Avi Gopstein is the senior energy technology advisor to Under Secretary Koonin at the Department of Energy. In this role he focuses on strategic R&D challenges across all DOE energy technology programs, including modeling and simulation of applied energy systems, risks and opportunities for technology development, and the interaction of engineered and natural systems at the scale of energy. Prior to joining the Administration Mr. Gopstein worked on diverse engineering and systems challenges ranging from integrated defense systems and spacecraft dynamics, to biotechnology and the interaction of the human body with engineered systems.

Paul Gribik, *Guest* - Sr. Director, Market Development and Analysis, Widwest ISO, pgribik@midwestiso.org
Paul Gribik is Senior Director of Market Development and Analysis at the Midwest ISO. Before joining MISO, Paul was a consultant with PA Consulting Group, Perot Systems Corporation, and Mykytyn Consulting Group. He has worked on electricity market restructuring issues in the Midwest, New England and California. Paul has a Ph.D. in Operations Research, a M.S. in Industrial Administration and a B.S. in Electrical Engineering from Carnegie-Mellon University.

Santiago Grijalva, *Author*, Associate Professor, Georgia Institute of Technology, sgrijalva@ece.gatech.edu
Santiago Grijalva is an Associate Professor of Electrical and Computer Engineering and the Director of the Advanced Computational Electricity Systems (ACES) Laboratory at The Georgia Institute of Technology. His areas of research are: real-time control, power system informatics, and sustainable electricity systems. Dr. Grijalva's industry experience prior to joining Georgia Tech was in EMS systems, and as Software Architect for PowerWorld Corporation. Dr. Grijalva graduate degrees in Electrical Engineering are from the University of Illinois at Urbana-Champaign.

Graham Hemingway, *Author*, Vanderbilt University, graham.hemingway@vanderbilt.edu
Graham Hemingway received his Ph.D. in Computer Science from Vanderbilt University. He currently is a Research Scientist at the Institute for Software Integrated System (ISIS). His research interests include modeling, simulation and analysis of safety-critical cyber-physical systems, distributed and heterogeneous simulation, and model-based collaboration and development platforms.

Ian Hiskens, *Guest* - Professor, University of Michigan, hiskens@umich.edu
Ian A. Hiskens is the Vennema Professor of Engineering in the Department of Electrical Engineering and Computer Science at the University of Michigan in Ann Arbor. He has held prior appointments in the Queensland electricity supply industry (for ten years), and various universities in Australia and the USA. Dr Hiskens' research addresses issues that arise in the design and operation of power systems, using concepts from nonlinear systems, optimization and networks. His recent activities have focused largely on integration of renewable generation and controllable load.

Andreas G. Hofmann, *Author*, Research Scientist, MIT Computer Science and Artificial Intelligence Lab, hofma@csail.mit.edu/shasen@mit.edu
Dr. Andreas Hofmann is a research scientist at MIT, where his research focuses on model-based autonomous systems, robotics, and optimal planning and control. He is also Vice-president of Autonomous Solutions at Vecna Technologies, where he leads development of autonomous robotic systems. Dr. Hofmann was previously a co-founder of Gensym Corp., a company specializing in automatic, supervisory-level control software for the chemical process and manufacturing industries. He has a bachelor's degree in Electrical Engineering from MIT, a master's in Electrical Engineering from Rensselaer Polytechnic Institute, and a Ph.D. in computer science from MIT.

Zhenyu (Henry) Huang, *Discussant*, Staff Engineer, Pacific Northwest National Laboratory, zhenyu.huang@pnl.gov
Henry Huang is currently a Staff Engineer with the Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington, USA. He is recipient of the 2008 Pacific Northwest National Laboratory Ronald L. Brodzinski's Award for Early Career Exceptional Achievement and the 2009 IEEE Power and Energy Society Outstanding Young Engineer Award. Dr. Huang has over 90 peer-reviewed publications. His research interests include high performance computing, phasor technology, and power system stability and simulation.

Mark Johnson, *Guest - DOE*, mark.johnson2@hq.doe.gov

Barbara Kenny, *Guest* - Program Director, National Science Foundation, bkenny@nsf.gov
Dr. Barbara Kenny is a Program Director at the National Science Foundation in the Engineering Education and Centers Division. She works primarily on the Engineering Research Center program, where she is the Energy, Sustainability and Infrastructure Cluster Leader. She has oversight responsibility for centers in diverse fields including nanotechnology, biomedical, robotics and power systems. Prior to coming to the National Science Foundation in 2006, she worked at the NASA Glenn Research Center in aerospace power systems, and she started her career as a U.S. Air Force officer in facilities engineering.

Alexandra Landsberg, *Guest* - Program Manager, Department of Energy, SC/ASCR, sandy.landsberg@science.doe.gov
Sandy Landsberg works for the Department of Energy, Office of Science, in the Office of Advanced Scientific Computing Research where she is an Applied Mathematics program manager. The Applied Mathematics program supports basic research and develops mathematical models, algorithms and numerical software for enabling predictive scientific simulations of DOE-relevant complex systems.

Philippe Larochelle, *Guest - DOE*, Postdoctoral Researcher, ARPA-E, philippe.larochelle@hq.doe.gov
Phil Larochelle is a postdoctoral researcher at the Oak Ridge Institute for Science and Education working as a support contractor to the Department of Energy's Advanced Research Projects Agency for Energy (ARPA-E). He works primarily on electricity storage and renewable energy integration onto the grid with program directors Mark Johnson and Rajeev Ram. Prior to ARPA-E, Phil was an alternative power and energy storage analyst Lux Research Inc. and senior energy analyst at the Alliance for Climate Protection. He received a Ph.D. in Physics from Harvard University and B.Sc. in Physics from M.I.T.

Bernard Lesieutre, *Discussant*, Professor, University of Wisconsin-Madison, lesieutre@engr.wisc.edu
Bernie Lesieutre is a Professor of Electrical Engineering at the University of Wisconsin-Madison and holds a research position at Lawrence Berkeley National Laboratory. Prior to LBNL and UW-Madison, he served on the faculty at MIT, and has held visiting faculty positions at Cornell and Caltech. His primary research involves the modeling and analysis of electric power systems for the purposes of assessing and improving reliability.

Dr. Sven Leyffer, *Reporter*, Argonne National Laboratory, leyffer@mcs.anl.gov
Sven Leyffer is a computational mathematician in the Mathematics and Computer Science Division at Argonne National Laboratory. Sven is a Senior Fellow of the Computation Institute at the University of Chicago, and an Adjunct Professor at

Northwestern University. Sven is the current SIAM Vice President for Programs, and serves on the editorial board of Computational Optimization and Applications, and Computational Management Science. He is an editor-in-chief of Mathematical Methods of Operations Research and the area editor for nonlinear optimization of Mathematical Programming Computation. In addition, Sven edits the SIAG/OPT Views-and-News. Together with Roger Fletcher and Philippe L. Toint, Sven was awarded the Lagrange prize in optimization in 2006. In 2009, Sven became a SIAM Fellow.

Eugene Litvinov, *Guest* - Sr. Director, ISO New England, elitvinov@iso-ne.com
Eugene Litvinov is the Senior Director of Business Architecture and Technology at ISO New England. He is responsible for advanced System and Markets solutions, Smart Grid and Technology strategy, is a lead of the Research and Development activities at ISO New England, is a chairman of the Power System Engineering Research Center (PSERC) industry advisory board, and an editor of the IEEE Transactions on Power Systems. He has more than 35 years of professional experience in the area of Power System modeling, analysis and operation; Electricity Markets design, implementation and operation; Information Technology. Dr. Litvinov holds BS, MS and PhD in Electrical Engineering.

James Momoh, *Guest – Professor* - Howard University, jmomoh@howard.edu
James Momoh is the former Department Chair and current professor of Electrical and Computer Engineering at Howard University, Director of the Center for Energy Systems and Control (CESaC), and former Director of ECE department at NSF. His research area is focused on optimization applications, power flow, distribution automation, computational intelligence applications, and smart grid. He is also the author of several papers and books including a best-seller, 2nd edition of Electrical Power System Applications of Optimization by CRC Press. Finally, he is a fellow of IEEE.

James Nutaro, *Discussant*, Research staff, Oak Ridge National Laboratory, nutarojj@ornl.gov
Jim Nutaro is part of the research staff at Oak Ridge National Laboratory in the Computational Sciences and Engineering Division and an adjunct faculty member in the Electrical Engineering and Computer Science department at the University of Tennessee. His work concentrates on methods and technology for modeling and simulation of engineered systems. Dr. Nutaro has recently published "Building Software for Simulation" and is area editor for the journals Simulation: the Transactions of the Society for Computer Simulation International and the ACM Transactions on Modeling and Computer Simulation.

HyungSeon Oh, *Reporter*, Power System Engineeer, National Renewable Energy Laboratory, HyungSeon.Oh@nrel.gov
HyungSeon Oh is an engineer III at the National Renewable Energy Laboratory. He received his Ph.D. in electrical and computer engineering from Cornell University. His

specialization includes power system planning, renewable energy, smart grids, storage devices, and nonlinear dynamics.

Ali Pinar, *Discussant*, Sandia National Laboratories, apinar@sandia.gov
Ali Pinar received his PhD in computer science from U. Illinois. After graduation, he joined LBNL, where he started working with power systems experts on vulnerability analysis of the power grid. He is now at Sandia National Labs, where he keeps working on various applications of networks, including the power system. His research is on graph algorithms, optimization, and parallel and scientific computing.

Sarah Ryan, *Author*, Professor, Iowa State University, smryan@iastate.edu
Sarah M. Ryan is Professor and Director of Graduate Education in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. Her research examines the planning and operation of energy, manufacturing and service systems under uncertainty, including stochastic optimization in bulk energy transportation networks, robust optimization of generation expansion planning, and asset management in electric transmission systems. Her work on capacity expansion and resource allocation was initiated under an NSF Faculty Early Career Development (CAREER) award. Professor Ryan earned the B.S. in systems engineering from The University of Virginia and the M.S. and Ph.D. in industrial and operations engineering from The University of Michigan.

Roman Samulyak, *Discussant*, Brookhaven National Lab / Stony Brook University, rosamu@bnl.gov
Roman Samulyak holds a joint appointment as professor at the Department of Applied Mathematics and Statistics, Stony Brook University, and scientist at Computational Science Center, Brookhaven National Laboratory. He received his PhD in Applied Mathematics from New Jersey Institute of Technology. R. Samulyak conducts research in mathematical modeling, numerical algorithms and supercomputer simulations of complex physics processes occurring in energy research applications. Recent examples include targets for future particle accelerators, multiphase magnetohydrodynamics, pellet fueling of thermonuclear fusion devices, alternative approaches to nuclear fusion, and optimization methods for electric power grids.

William Schulze, *Guest* - Robinson Professor, Cornell University, wds3@cornell.edu
William Schulze is the Robinson Professor of Applied Economics and Public Policy in the Dyson School at Cornell University. His research interests include behavioral, experimental and environmental economics, electricity markets, and optimal operation and investment in large scale power systems, both in terms of economic efficiency and environmental impacts.

Christine Shoemaker, *Guest – Professor* - Cornell University, cas12@cornell.edu
Christine Shoemaker is the Joseph P. Ripley Professor of Engineering at Cornell University where she is affiliated with the School of Civil and Environmental Engineering and School of Operations Research and Information Engineering. She received her PhD in Mathematics supervised by Richard Bellman in the area of optimal control and dynamic programming with applications to the environment. Her computational research currently focuses on parallel and serial algorithms for global optimization of computationally expensive simulation models and her applications are on environmental and hydrologic problems including optimization of hydropower generation and carbon sequestration. She has won several awards including Distinguished Member ASCE, Fellow of INFORMS, Fellow of AGU (Hydrology), and a Humboldt Research Prize from Germany.

Paul M. Sotkiewicz, *Guest - Industry*, PJM Interconnection, sotkip@pjm.com
Paul M. Sotkiewicz, Ph.D., is chief economist in the Market Services Division at the PJM Interconnection, provides analysis and advice with respect to PJM's market design and market performance including demand response mechanisms, intermittent and renewable resource integration, market power mitigation strategies, capacity markets and the potential effects of climate change and other environmental policies on PJM's markets.

Anurag Srivastava, *Author*, Assistant Professor, Washington State University, asrivast@eecs.wsu.edu
Anurag K. Srivastava received his Ph.D. from Illinois Institute of Technology, Chicago and working as Assistant Professor at Washington State University since August 2010. His research interests include power system operation and control, power system modeling and real time simulation. Dr. Srivastava is a senior member of IEEE and author of more than 75 publications.

Jason Stamp, *Discussant*, Sandia National Laboratories, jestamp@sandia.gov
Jason Stamp is a Principal Member of the Technical Staff in the Energy Systems Analysis department at Sandia National Laboratories in Albuquerque, New Mexico. His research areas include cyber security for control systems, and the development of energy microgrids for military applications (he is currently the lead engineer of the SPIDERS project for the Department of Defense). He received a Bachelor of Science degree in Electrical Engineering from Rose-Hulman Institute of Technology in Terre Haute, Indiana in 1995, and his Ph.D. in Electrical Engineering from Clemson University in 1998.

Janos Sztipanovits, *Author*, Professor, Vanderbilt University, janos.sztipanovits@vanderbilt.edu
Janos Sztipanovits is currently the E. Bronson Ingram Distinguished Professor of Engineering at Vanderbilt University. He is founding director of the Institute for

Software Integrated Systems (ISIS). His research areas are at the intersection of systems and computer science and engineering. His current research interest includes the foundation and applications of Model-Integrated Computing for the design of Cyber Physical Systems.

Bob Thomas, *Organizer*, Professor Emeritus - Cornell University, rjt1@cornell.edu
Bob Thomas is currently a Professor Emeritus, Cornell University where he has been a member of the faculty of the School of Electrical and Computer Engineering since 1973. He has had assignments in Washington DC that include the first Program Director for the Power Systems Program at the National Science Foundation and the U.S. Department of Energy as a Senior Advisor and an inaugural member of the U.S. Department of Energy Secretary's Electricity Advisory Committee (EAC). He is the founding Director of the NSF Power Systems Engineering Research Center (PSerc). He has received 5 teaching awards and the IEEE Centennial and Millennium medals. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, ASEE and a Life Fellow of the IEEE.

Travis Togo, *Guest* - Operations Research Analyst, Bonneville Power Administration, txtogo@bpa.gov
Travis Togo received his undergraduate degree at The University of Montana and graduate degree at Louisiana State University in mathematics. He is a senior operations research analyst in the short-term operations planning organization at Bonneville Power Administration. Currently he leads an effort to improve BPA's modeling capability.

Loren Toole, *Discussant*, Senior R&D Engineer, Los Alamos National Security LLC, ltoole@lanl.gov
G. Loren Toole serves as a Senior R&D Engineer at Los Alamos National Laboratory. His engineering experience extends over 30 years, including over 10 years an electric utility planner. His current research interest is focused on application of optimization methods to grid expansion analysis. He holds degrees in Electrical Engineering from the Georgia Institute of Technology and serves as an Adjunct Professor, University of Missouri College of Engineering (Columbia).

Robbert Van Renesse, *Author*, Principal Research Scientist, Cornell University, rvr@cs.cornell.edu
Robbert van Renesse is Principal Research Scientist in the Department of Computer Science at Cornell University in Ithaca, NY. He is interested in distributed systems, particularly in their fault tolerance and scalability aspects. Van Renesse is an ACM Fellow.

David Woodruff, *Author*, Professor, University of California Davis, dlwoodruff@ucdavis.edu
David L. Woodruff is Professor at the Graduate School of Management at the University of California at Davis. His research concerns computational aspects of multi-stage

optimization under uncertainty, which includes solution algorithms, problem representation and modeling language support. He has worked on applications in operations, logistics, and has been involved recently in a number of applications in energy.

Jinjun Xiong, Author, IBM, jinjun@us.ibm.com
Jinjun Xiong is a research staff member at IBM Thomas J. Watson Research Center. His research background was in the area of electronic circuit design automation with focus on modeling and simulation. He is a member of the IBM's smart energy research team, working on the Pacific Northwest Smart Grid Demonstration Project, a DoE Smart Grid Demonstration project led by Battelle.

Dr. Chao Yang, *Reporter*, Lawrence Berkeley National Laboratory, cyang@lbl.gov
Chao Yang is a staff scientist in the computational research division at Lawrence Berkeley National Lab. He has a Ph.D in computational science and engineering from Rice University (1998). His research interests are in large-scale scientific computing and numerical algorithms with applications in material science, chemistry, biology and power systems.

Victor Zavala, *Reporter*, Assistant Computational Mathematician, Argonne National Laboratory, vzavala@msc.anl.gov
Victor Zavala is a computational mathematician at Argonne National Laboratory. He holds a PhD degree in Chemical Engineering from Carnegie Mellon University. His research activities include optimization and control algorithms for building energy management, electricity markets, and power plants. He is has served as a consultant for industrial companies such as General Electric, BuildingIQ, and ExxonMobil.

Ray Zimmerman, *Guest* - Senior Research Associate, Cornell University, rz10@cornell.edu
Ray Zimmerman is a Senior Research Associate at Cornell University and the principal developer of the widely used power systems simulation tool called MATPOWER. His research interests include optimization problems and tools related to the electric power industry and its changing and growing computational needs. Recent work has focused on tools that can properly allocate and value resources under uncertainty, including assets such as storage and flexible loads and public goods such as system reliability.