Datacenter Power Management in Smart Grids

Xue Liu
McGill University
xueliu@cs.mcgill.ca

Fanxin Kong
McGill University
fanxin.kong@mail.mcgill.ca

the essence of knowledge
Boston — Delft

Foundations and Trends® in Electronic Design Automation

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

X. Liu and F. Kong. *Datacenter Power Management in Smart Grids*. Foundations and Trends® in Electronic Design Automation, vol. 9, no. 1, pp. 1–98, 2015.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-793-8

Foundations and Trends® in Electronic Design Automation
Volume 9, Issue 1, 2015

Editorial Scope

Topics

Foundations and Trends® in Electronic Design Automation publishes survey and tutorial articles in the following topics:

- System level design
- Behavioral synthesis
- Logic design
- Verification
- Test
- Physical design
- Circuit level design
- Reconfigurable systems
- Analog design
- Embedded software and parallel programming
- Multicore, GPU, FPGA, and heterogeneous systems
- Distributed, networked embedded systems
- Real-time and cyberphysical systems

Information for Librarians

Foundations and Trends® in Electronic Design Automation, 2015, Volume 9, 4 issues. ISSN paper version 1551-3939. ISSN online version 1551-3947. Also available as a combined paper and online subscription.

Foundations and Trends® in Electronic Design Automation
Vol. 9, No. 1 (2015) 1–98

DOI: 10.1561/1000000038

Datacenter Power Management in Smart Grids

Xue Liu
McGill University
xueliu@cs.mcgill.ca

Fanxin Kong
McGill University
fanxin.kong@mail.mcgill.ca

Contents

Abstract

Cloud computing is a new computing paradigm and it is gaining wide popularity due to its benefits including reduced cost, ease of management, and increased reliability. In a cloud computing environment, companies or individuals offload their computing (hardware/software/data) to the cloud, which is supported by the computing infrastructure called datacenters. Datacenters consume large amounts of electricity to operate and bring enormous electricity bills to the operators. Associated emissions also cause significant negative impact to the environment. Meanwhile, a new kind of electrical grid, called the smart grid, is emerging. Smart grids enable two way communications between the power generators and the power consumers. Smart grid technology brings many salient features to help deliver power efficiently and reliably.

There are many efforts addressing either of the two tracks above. Different with them, we focus on cost-aware datacenter power management in presence of smart grids and review recent developments on this area. It involves understanding how a smart grid operates, where power goes in datacenters, and most importantly, how to reduce the power cost and/or negative environmental impact when operating datacenters. We first study new ideas of exploring spatial diversities provided by geographically distributed datacenters and show how to perform request routing. Then, we discuss the research that leverages temporal flexibilities given by delay-tolerant workload and present how to conduct workload scheduling. Thirdly, we study how to jointly optimize routing and scheduling by considering spatial diversities and temporal flexibilities together. These studies consider multiple features of smart grids, and develop different cost minimization approaches using techniques from the optimization, algorithmic, and feedback control fields. Moreover, we review studies incorporating the same solution framework with additional dimensions such as renewable and cooling.

X. Liu and F. Kong. *Datacenter Power Management in Smart Grids*. Foundations and Trends® in Electronic Design Automation, vol. 9, no. 1, pp. 1–98, 2015.

DOI: 10.1561/1000000038.

1

Introduction and Overview

Cloud computing is a new computing paradigm and it is gaining wide popularity due to the distinct benefits provided to the cloud users Armbrust et al. [2010]. Cloud computing enables users to have ubiquitous, convenient, and on-demand access to a shared pool of computing resources (e.g., servers, applications, and services). Cloud users no longer need massive capital investment in self-owned hardware and software or large expenses to operate them. They do not need to be concerned about over-provisioning or under-provisioning computing resources for their services. Instead, cloud users can pay for use of computing resources as needed and release them when not needed. This feature significantly simplifies users' resource management. Furthermore, the cloud keeps users' data safe and improves their service reliability.

The computing infrastructure supporting the cloud is called datacenters. The benefits for cloud users come with large-scale datacenter maintenance and management, which cause unacceptably heavy economic burden on cloud service providers. For example, a large datacenter hosts hundreds of thousands servers and requires megawatts of electricity Katz [2009]. This brings an electricity bill of millions of dollars annually to the operator. Moreover, associated carbon emissions from operating datacenters also cause significant negative impact to the environment. Organizations such as Google, Microsoft, Amazon, and many other cloud service providers are very concerned with their high power usages, electricity costs, and carbon emissions. Hence, an energy-efficient, low-cost, and environment-friendly datacenter operation is a key enabler of cloud computing.

In the mean time, a new technology for the next generation electrical grid, called the smart grid, has been emerging Farhangi [2010]. A smart grid uses both information and communication technology, and enables a two way communication between the power suppliers and the power consumers. Many salient features brought by a smart grid including self-healing, load adjustment, bi-directional energy flows, time-of-use pricing, demand response, etc., help deliver power more efficiently and reliably. Therefore, smart grids provide an excellent opportunity for better power management of datacenters. For example, cloud service providers could benefit from receiving real-time information of the electrical grid, such as electricity price and power availability. They can reduce their energy costs by dispatching workload to the datacenters where the electricity prices are lower and the power supplies are abundant. In this work, we review how to manage datacenter power consumption in the context of a smart grid, mainly focusing on utilizing features of time-of-use pricing and demand response to cut the datacenter electricity cost.

1.1 Datacenter Overview

Cloud-oriented datacenters are usually built in independent buildings, which are dedicated for hosting computing devices. Each datacenter can take up to several hundred thousand square feet in size, and can have a peak power usage of tens of megawatts (MWs). A cloud service provider may have tens or even hundreds of such datacenters distributed geographically and connected by the Internet. This section presents an overview of such a large-scale system, from three aspects including major components and their functions, power usage, and datacenter distribution.

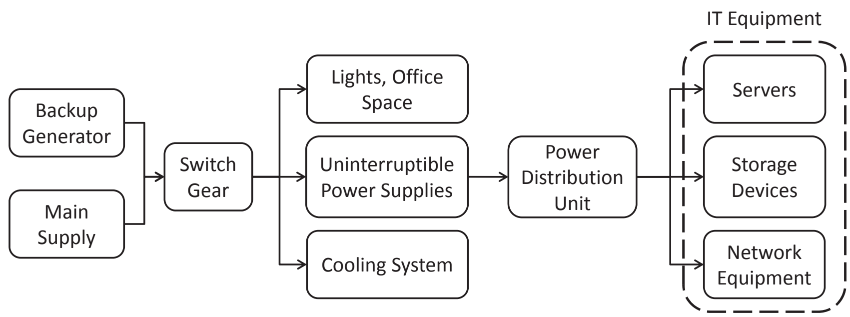

Figure 1.1: Components of a Datacenter.

1.1.1 Major Components

Fig. 1.1 shows typical components of a datacenter U.S. Environmental Protection Agency [2007]. There are three major components: IT equipment, cooling system, and power infrastructure. We list detailed description for each of the three major components as follows.

IT equipment. IT equipment includes servers for data processing, storage devices for data storage, and network equipment for data communications. They work together to support applications and services hosted in a datacenter. Servers are often mounted in rack cabinets and are interconnected by high-speed network equipment. Storage devices are usually placed alongside servers for lower access delay as well as easy management.

Cooling system. Cooling devices, usually referred to as computer room air conditioning (CRAC) units, extract the heat from IT equipment, and control the temperature and humidity in a datacenter. Typically, outside air is introduced into the top of a CRAC unit where it gets conditioned by passing through some coils containing chilled water. The chilled water is pumped from a chiller that cools down the returned hot water from CRAC units using mechanical refrigeration cycles or water-sider economizers Zhou et al. [2012], Liu et al. [2012]. The cooled air then passes IT equipment (primarily servers) through a raised floor plenum, and the fans pull the air into servers.

Power infrastructure. Different datacenters may have different

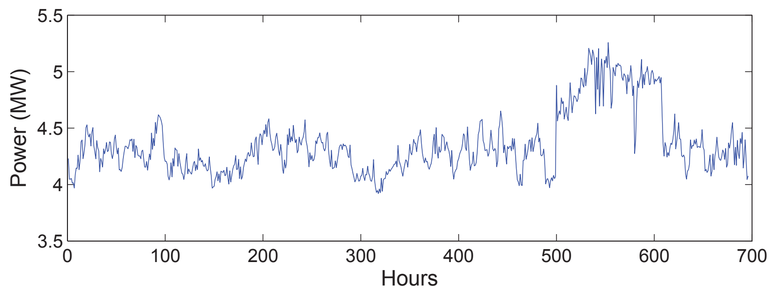

Figure 1.2: The figure is from Kong and Liu [2014], which illustrates the hourly power usage of a Google datacenter of 12,500 machines Goolge. All servers are assumed to be homogeneous. Each server has a peak power of 300W and idle power of 150W. Networking devices' power consumption equals 5% of the total peak power of all servers. Power Usage Effectiveness or PUE equals 1.5.

power infrastructure designs. We just give a typical case in the following. The power infrastructure generates and/or distributes power to cooling devices and IT equipment through a micro power grid that can integrate the electrical grid, local power supply (backup generator) and energy storage devices (UPS, uninterruptible power supplies). Electricity first flows to an UPS unit, which acts as a battery backup to prevent the IT equipment from experiencing power disruptions. The electricity is converted from AC to DC to charge the UPS batteries, and then reconverted from DC to AC before leaving the UPS. The electricity finally enters IT equipment through power distribution units (PDUs). Upon a power outage, UPS batteries keep powering up a datacenter using energy stored within them, until the backup power can start up or ramp up to match the datacenter's power load.

1.1.2 Power Usage

The total power usage of a datacenter can be approximately defined as the sum of IT equipment power, cooling power, and power distribution losses. They are related via the power usage effectiveness (PUE) metric, which equals total power divided by IT equipment power. A lower PUE means a higher power efficiency, i.e., a larger portion of power is used for computing devices instead of supporting facilities. Modern datacenters usually have a PUE around 1.1 to 2. The power usage of

IT equipment consists of the aggregated power consumed by all servers (data storage included) and networking devices. Servers' power consumption can be estimated by a linear power model or dynamic voltage and frequency scaling (DVFS) power model (details will be discussed in Section 2). Networking devices' power consumption can be approximated as a constant offset and is in general less than 10% of the peak power of all servers in a datacenter Hamilton. Hence, if given a PUE and IT equipment power consumption, we can estimate the total power usage of a datacenter. For example, Fig. 1.2 (from Kong and Liu [2014]) shows power load estimation using a Google workload trace Goolge based on the linear power model in Fan et al. [2007]. The datacenter power load is highly variable but shows an approximate daily pattern of up-and-down. The first three weeks have a similar power load trace; while the last week experiences a load burst. A more detailed analysis on the Google workload trace can be found in Reiss et al. [2012]. A more accurate datacenter power usage estimation method relates PUE to the varying temperature. PUE is largely dependent on the scale of cooling power, and cooling power is closely tied to the temperature difference between the outside air and the insider air Liu et al. [2012]. For example, usually PUE for a cold weather will be smaller than that in a hot weather.

1.1.3 Datacenter Distribution

A cloud service provider may have tens or even hundreds of geo-distributed datacenters connected at Internet scale. Each datacenter is connected to many Internet Service Providers (ISPs) that are responsible for carrying traffic between service providers and thousands to millions of users. Datacenters are interconnected by the backbone network. Datacenters are distributed to different locations according to many factors, such as population, power availability, network proximity, and climate. The geo-distributed feature of datacenters brings a number of diversities in service response time, capital and operational costs, and carbon emissions. Hence, these diversities can be leveraged to optimize different kinds of performances for an Internet scale datacenter system. For example, user requests can be dispatched to the nearest datacenter (e.g., with the smallest round trip time (RTT)) for processing to reduce the response time perceived by users, or assigned to the datacenter location with the lowest temperature to reduce the datacenter cooling cost, or routed to the datacenter with the lowest electricity price to cut the datacenter operational cost.

1.2 Smart Grid Overview

Smart grids are the next generation power grids. Traditional electrical grids are used to carry power from several central power generators to a large number of customers. By contrast, a smart grid enables two-way energy and information flows to create an automated and distributed power delivery network. Smart grids have distinct features compared with traditional grids. For a comprehensive survey on smart grid systems, readers are referred to existing surveys, such as Fang et al. [2011]. In this work, we only describe features closely related to the datacenter power management.

Improved reliability. Smart grids employ advanced technologies for better self-monitoring and self-healing without manual intervention. With the help of real-time monitoring devices, a smart grid is able to reduce blackouts with minimum disruptions. It can automatically detect the problems, immediately respond to errors on power lines, and accurately isolate the error-prone links from the main power network. Reliable and quality power supply can significantly improve datacenter service availability.

Dynamic pricing. Communication and metering technologies of smart grids can inform electricity consumers (e.g., datacenters) via smart devices, when power demand is high in their regions. To motivate consumers to cut their load, the electricity price increases during high demand periods and decreases during low demand periods. Consequently, electricity consumers are stimulated to consume less during high demand periods as eventually they would see an economic gain by using energy at off-peak periods.

Enhanced sustainability. A smart grid is a key enabler for deep integration of renewable energy and distributed power generation. Smart grid technology allows for many distributed feed-in points and support bidirectional power flows. Furthermore, with the help of integrated monitoring and control, a smart grid can tackle the intermittancy and fluctuations of renewable energy, and can also maintain a consistent and stable power flow over the electrical grid. Hence, one electricity consumer can not only draw much cleaner power from the grid but also can install local power generators to become a power supplier.

Demand response. Generators and consumers can interact in a real-time manner with demand response support, adjusting demand to level out spikes. Advanced communication capabilities of a smart grid can provide consumers effective tools to receive incentive-based or emergency load reduction signals and to respond these signals accordingly. These capabilities not only eliminate the cost of adding reserve capacity for utilities but also cut the electricity bills for energy users who avoid coincidence with spikes. Moreover, consumers are even allowed to sell self-generated or stored energy back to the grid under policies such as net metering DSIRE.

1.3 Datacenter Power Management Overview

Along with the surging energy usage of datacenters, power management is becoming an important and active research area. An overview of challenges toward power management in datacenters is presented in Liu et al. [2009a]. Efforts such as the Climate Savers Computing Initiative (www.climatesaverscomputing.org) intended to help lower worldwide computer energy consumption by promoting widespread adoption of high efficiency power supplies and also by encouraging the use of power-saving features already present in users' equipment. Earlier research work on power management aims at reducing power consumption of servers within a single datacenter. Different hardware and software technologies have been proposed. In hardware level, the adoption of technologies, such as chip multiprocessing Barroso and Holzle [2007] and dynamic voltage and frequency scaling (DVFS) Horvath et al. [2007], has made more energy-efficient servers. In software level, the application of technologies, such as virtualization technologies Nathuji and Schwan [2007], Barham et al. [2003], dynamic power management Chase et al. [2001], Meisner et al. [2009], Ahmad and Vijaykumar [2010] and control theories Lefurgy et al. [2007], has further reduced servers' power consumption. In addition, there are new technologies developed to reduce the power of both servers and cooling device in a datacenter, such as Raghavendra et al. [2008], Liu et al. [2009b], Abbasi et al. [2012].

While all of the above works address how to reduce datacenter energy consumption, this work focuses on the problem of how to reduce the datacenter energy cost. The difference between energy saving and cost reduction mainly stems from the dynamics of the electricity market. For example, to minimize energy consumption, workload should be allocated to the most energy-efficient datacenters as much as possible. However, these datacenters may not locate at the region with the cheapest electricity price, and thus this allocation method would increase energy cost. Furthermore, there are two reasons for considering energy cost. First, energy cost is one of the major concerns for datacenter operators, such as Google and Facebook. For example, Google consumed $2.60 \times 10^6 MWh$ electricity in 2010, amounting to hundreds of million dollars Shao et al. [2013]. Second, lowering datacenter energy expenditure also contributes to emission reduction. For example, the electricity price increases as power suppliers put more spinning reserves on-line to meet the rising power demand. These reserves not only have higher power generation costs, but also usually use carbon-intensive fuel such as coal and diesel. Hence, lowering energy cost makes datacenters use less power when the price is high, and thus reduce carbon emissions.

1.4 Road Map

Figure 1.3 depicts the road map of this work. In Section 2, we study ideas of exploring spatial diversities provided by geographically distributed datacenters and show how to carry out request routing to minimize energy cost. In this section, we first present the objectives and constraints of the energy cost minimization problem for geograph-

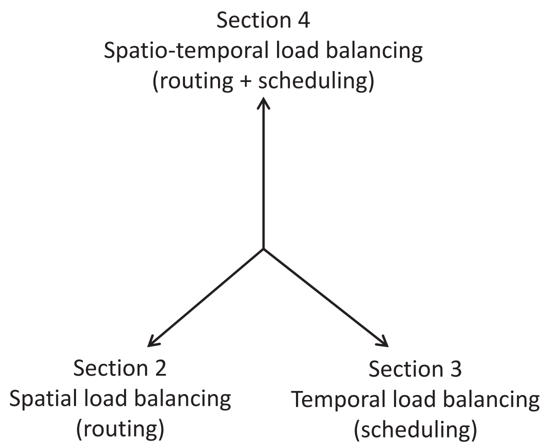

Figure 1.3: Road map of the work.

ically distributed datacenters. We then provide several solutions based on different pricing and power models using techniques from the optimization, algorithmic and feedback control fields. At the end of this section, we introduce several extensions of this model to deal with carbon emission reduction, renewable energy integration, and energy buffering. In Section 3, we provide the research that leverages temporal flexibilities given by delay-tolerant datacenter workload and present how to perform workload scheduling to save datacenter power cost. In Section 4, we study how to jointly optimize routing and scheduling by exploring both spatial diversities and temporal flexibilities. We first discuss the problem formulation and a cost minimizing solution, and then present several open problems for spatio-temporal load balancing. In Section 5, we conclude the work with summaries of main ideas discussed in previous sections.

References

Zahra Abbasi, Georgios Varsamopoulos, and Sandeep KS Gupta. Tacoma: Server and workload management in internet data centers considering cooling-computing power trade-off and energy proportionality. *ACM Transactions on Architecture and Code Optimization (TACO)*, 9(2):11, 2012.

Faraz Ahmad and TN Vijaykumar. Joint optimization of idle and cooling power in data centers while maintaining response time. In *ACM Sigplan Notices*, volume 45, pages 243–256. ACM, 2010.

Baris Aksanli, Jagannathan Venkatesh, Liuyi Zhang, and Tajana Rosing. Utilizing green energy prediction to schedule mixed batch and service jobs in data centers. *ACM SIGOPS Operating Systems Review*, 45(3):53–57, 2012.

Amazon. Amazon ec2 spot instance.

Michael Armbrust, Armando Fox, Rean Griffith, Anthony D Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, et al. A view of cloud computing. *Communications of the ACM*, 53 (4):50–58, 2010.

Ross Baldick, Ryan Grant, and Edward Kahn. Theory and application of linear supply function equilibrium in electricity markets. *Journal of Regulatory Economics*, 25(2):143–167, 2004.

Paul Barham, Boris Dragovic, Keir Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. Xen and the art of virtualization. *ACM SIGOPS Operating Systems Review*, 37(5):164–177, 2003.

Luiz André Barroso and Urs Holzle. The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007.

Mokhtar S Bazaraa, John J Jarvis, and Hanif D Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.

Dimitri P Bertsekas and John N Tsitsiklis. Parallel and distributed computation. 1989.

Paul Bogdan, Radu Marculescu, and Siddharth Jain. Dynamic power management for multidomain system-on-chip platforms: an optimal control approach. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 18(4):46, 2013.

Richard N Boisvert, Peter A Cappers, and Bernie Neenan. The benefits of customer participation in wholesale electricity markets. *The Electricity Journal*, 15(3):41–51, 2002.

Ulrich Brenner. A faster polynomial algorithm for the unbalanced hitchcock transportation problem. *Operations Research Letters*, 36(4):408–413, 2008.

Niv Buchbinder, Navendu Jain, and Ishai Menache. Online job-migration for reducing the electricity bill in the cloud. In *NETWORKING*, pages 172–185. Springer, 2011.

Michael R Bussieck and Armin Pruessner. Mixed-integer nonlinear programming. *SIAG/OPT Newsletter: Views & News*, 14(1):19–22, 2003.

Eduardo F Camacho and Carlos Bordons Alba. *Model predictive control*. Springer, 2013.

Niklas Carlsson and Martin Arlitt. Towards more effective utilization of computer systems. In *ACM/SPEC International Conference on Performance Engineering (ICPE)*, pages 235–246. ACM, 2011.

Jeffrey S Chase, Darrell C Anderson, Prachi N Thakar, Amin M Vahdat, and Ronald P Doyle. Managing energy and server resources in hosting centers. In *ACM SIGOPS Operating Systems Review*, volume 35, pages 103–116. ACM, 2001.

Changbing Chen, Bingsheng He, and Xueyan Tang. Green-aware workload scheduling in geographically distributed data centers. In *IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 82–89. IEEE, 2012.

Benoît Colson, Patrice Marcotte, and Gilles Savard. Bilevel programming: A survey. *4OR*, 3(2):87–107, 2005.

Wei Deng, Fangming Liu, Hai Jin, Chuan Wu, and Xue Liu. Multigreen: cost-minimizing multi-source datacenter power supply with online control. In *Proceedings of the fourth international conference on Future energy systems*, pages 149–160. ACM, 2013.

Javier Diaz, Camelia Muñoz-Caro, and Alfonso Niño. A survey of parallel programming models and tools in the multi and many-core era. *IEEE Transactions on Parallel and Distributed Systems*, 23(8):1369–1386, 2012.

Chuansheng Dong, Fanxin Kong, Xue Liu, and Haibo Zeng. Green power analysis for geographical load balancing based datacenters. In *International Green Computing Conference (IGCC)*, pages 1–8. IEEE, 2013.

Joseph Doyle, Donal O'Mahony, and Robert Shorten. Server selection for carbon emission control. In *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, pages 1–6. ACM, 2011.

DSIRE. Database of state incentives for reneables & efficiency. *http://www.dsireusa.org/*.

Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power provisioning for a warehouse-sized computer. *ACM SIGARCH Computer Architecture News*, 35(2):13–23, 2007.

Xi Fang, Satyajayant Misra, Guoliang Xue, and Dejun Yang. Smart grid-the new and improved power grid: a survey. *IEEE Communications Surveys & Tutorials*, 14(4):944–980, 2011.

Hassan Farhangi. The path of the smart grid. *IEEE Power and Energy Magazine*, 8(1):18–28, 2010.

Federal Energy Regulatory Commission. Electric power markets. *http://www.ferc.gov/market-oversight/mkt-electric/overview.asp*.

Roger Fletcher and Sven Leyffer. Solving mixed integer nonlinear programs by outer approximation. *Mathematical programming*, 66(1-3):327–349, 1994.

Peter Xiang Gao, Andrew R Curtis, Bernard Wong, and Srinivasan Keshav. It's not easy being green. *ACM SIGCOMM Computer Communication Review*, 42(4):211–222, 2012.

Yan Gao, Zheng Zeng, Xue Liu, and PR Kumar. The answer is blowing in the wind: Analysis of powering internet data centers with wind energy. In *Proceedings IEEE INFOCOM (Mini Conference)*. IEEE, 2013.

Reinaldo C Garcia, Javier Contreras, Marco Van Akkeren, and João Batista C Garcia. A GARCH forecasting model to predict day-ahead electricity prices. *IEEE Transactions on Power Systems*, 20(2):867–874, 2005.

Siddharth Garg, Diana Marculescu, and Radu Marculescu. Custom feedback control: enabling truly scalable on-chip power management for mpsocs. In *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, pages 425–430. IEEE, 2010.

Arthur M Geoffrion. Generalized benders decomposition. *Journal of optimization theory and applications*, 10(4):237–260, 1972.

Mahdi Ghamkhari and Hamed Mohsenian-Rad. Energy and performance management of green data centers: A profit maximization approach. *IEEE Transactions on Smart Grid*, 4(2):1017–1025, 2013.

Íñigo Goiri, Ryan Beauchea, Kien Le, Thu D Nguyen, Md E Haque, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. Greenslot: scheduling energy consumption in green datacenters. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*, page 20. ACM, 2011.

Íñigo Goiri, Kien Le, Thu D Nguyen, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. Greenhadoop: leveraging green energy in data-processing frameworks. In *Proceedings of the 7th ACM european conference on Computer Systems*, pages 57–70. ACM, 2012.

Íñigo Goiri, William Katsak, Kien Le, Thu D Nguyen, and Ricardo Bianchini. Parasol and greenswitch: managing datacenters powered by renewable energy. In *Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems*, pages 51–64. ACM, 2013.

Andrew V Goldberg. An efficient implementation of a scaling minimum-cost flow algorithm. *Journal of algorithms*, 22(1):1–29, 1997.

Goolge. Clusterdata2011_1. *https://code.google.com/p/googleclusterdata/wiki/ClusterData2011_1*.

Sriram Govindan, Anand Sivasubramaniam, and Bhuvan Urgaonkar. Benefits and limitations of tapping into stored energy for datacenters. In *38th Annual International Symposium on Computer Architecture (ISCA)*, pages 341–351. IEEE, 2011.

Sriram Govindan, Di Wang, Anand Sivasubramaniam, and Bhuvan Urgaonkar. Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters. In *Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 75–86. ACM, 2012.

Richard J Green and David M Newbery. Competition in the british electricity spot market. *Journal of political economy*, pages 929–953, 1992.

Yuanxiong Guo and Yuguang Fang. Electricity cost saving strategy in data centers by using energy storage. *IEEE Transactions on Parallel and Distributed Systems*, 24(6):1149–1160, 2013.

Yuanxiong Guo, Zongrui Ding, Yuguang Fang, and Dapeng Wu. Cutting down electricity cost in internet data centers by using energy storage. In *IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–5. IEEE, 2011.

James Hamilton. James hamilton's blog-energy proportional datacenter. *http://perspectives.mvdirona.com/*.

Jin Heo, Dan Henriksson, Xue Liu, and Tarek Abdelzaher. Integrating adaptive components: An emerging challenge in performance-adaptive systems and a server farm case-study. In *28th IEEE International Real-Time Systems Symposium (RTSS)*, pages 227–238. IEEE, 2007.

Tibor Horvath and Kevin Skadron. Multi-mode energy management for multi-tier server clusters. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*, pages 270–279. ACM, 2008.

Tibor Horvath, Tarek Abdelzaher, Kevin Skadron, and Xue Liu. Dynamic voltage scaling in multitier web servers with end-to-end delay control. *IEEE Transactions on Computers*, 56(4):444–458, 2007.

Randy H Katz. Tech titans building boom. *Spectrum, IEEE*, 46(2):40–54, 2009.

Enver Kayaaslan, B Barla Cambazoglu, Roi Blanco, Flavio P Junqueira, and Cevdet Aykanat. Energy-price-driven query processing in multi-center web search engines. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 983–992. ACM, 2011.

Daniel S Kirschen. Demand-side view of electricity markets. *IEEE Transactions on Power Systems*, 18(2):520–527, 2003.

Leonard Kleinrock. *Theory, volume 1, Queueing systems*. Wiley-interscience, 1975.

Fanxin Kong and Xue Liu. A survey on green-energy-aware power management for datacenters. *Accepted to ACM Computing Surveys*, 2014.

Fanxin Kong, Wang Yi, and Qingxu Deng. Energy-efficient scheduling of real-time tasks on cluster-based multicores. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011*, pages 1–6. IEEE, 2011.

Fanxin Kong, Xue Liu, and Lei Rao. Optimal energy source selection and capacity planning for green datacenters. In *The 2014 ACM international conference on Measurement and modeling of computer systems (Poster)*, pages 575–576. ACM, 2014.

Andrew Krioukov, Christoph Goebel, Sara Alspaugh, Yanpei Chen, David E Culler, and Randy H Katz. Integrating renewable energy using data analytics systems: Challenges and opportunities. *IEEE Data Eng. Bull.*, 34 (1):3–11, 2011.

Andrew Krioukov, Sara Alspaugh, Prashanth Mohan, Stephen Dawson-Haggerty, David E Culler, and Randy H Katz. Design and evaluation of an energy agile computing cluster. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2012-13*, 2012.

Rupa Krishnan, Harsha V Madhyastha, Sridhar Srinivasan, Sushant Jain, Arvind Krishnamurthy, Thomas Anderson, and Jie Gao. Moving beyond end-to-end path information to optimize cdn performance. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 190–201. ACM, 2009.

Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations research*, 14(4):699–719, 1966.

Kien Le, Ricardo Bianchini, Margaret Martonosi, and T Nguyen. Cost-and energy-aware load distribution across data centers. *Proceedings of HotPower*, pages 1–5, 2009a.

Kien Le, Ozlem Bilgir, Ricardo Bianchini, Margaret Martonosi, and Thu D Nguyen. Managing the cost, energy consumption, and carbon footprint of internet services. *Technical Report, Dept. of Computer Science, Rutgers University*, 2009b.

Kien Le, Ricardo Bianchini, Thu D Nguyen, Ozlem Bilgir, and Margaret Martonosi. Capping the brown energy consumption of internet services at low cost. In *International Green Computing Conference (IGCC)*, pages 3–14. IEEE, 2010.

Kien Le, Ricardo Bianchini, Jingru Zhang, Yogesh Jaluria, Jiandong Meng, and Thu D Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*, page 22. ACM, 2011.

Charles Lefurgy, Xiaorui Wang, and Malcolm Ware. Server-level power control. In *Fourth International Conference on Autonomic Computing*, pages 4–4. IEEE, 2007.

Jie Li, Zuyi Li, Kui Ren, and Xue Liu. Towards optimal electric demand management for internet data centers. *IEEE Transactions on Smart Grid*, 3(1):183–192, 2012.

Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. Dynamic right-sizing for power-proportional data centers. In *Proceedings IEEE INFOCOM*, pages 1098–1106. IEEE, 2011a.

Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. Online dynamic capacity provisioning in data centers. In *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1159–1163. IEEE, 2011b.

Minghong Lin, Zhenhua Liu, Adam Wierman, and Lachlan LH Andrew. Online algorithms for geographical load balancing. In *International Green Computing Conference (IGCC)*, pages 1–10. IEEE, 2012.

Jie Liu, Feng Zhao, Xue Liu, and Wenbo He. Challenges towards elastic power management in internet data centers. In *29th IEEE International Conference on Distributed Computing Systems Workshops*, pages 65–72. IEEE, 2009a.

Liang Liu, Hao Wang, Xue Liu, Xing Jin, Wen Bo He, Qing Bo Wang, and Ying Chen. Greencloud: a new architecture for green data center. In *Proceedings of the 6th international conference industry session on Autonomic computing and communications industry session*, pages 29–38. ACM, 2009b.

Shuo Liu, Shaolei Ren, Gang Quan, Ming Zhao, and Shangping Ren. Profit aware load balancing for distributed cloud data centers. In *IEEE 27th International Symposium on Parallel & Distributed Processing (IPDPS)*, pages 611–622. IEEE, 2013.

Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H Low, and Lachlan LH Andrew. Geographical load balancing with renewables. *ACM SIGMETRICS Performance Evaluation Review*, 39(3):62–66, 2011a.

Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H Low, and Lachlan LH Andrew. Greening geographical load balancing. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 233–244. ACM, 2011b.

Zhenhua Liu, Yuan Chen, Cullen Bash, Adam Wierman, Daniel Gmach, Zhikui Wang, Manish Marwah, and Chris Hyser. Renewable and cooling aware workload management for sustainable data centers. In *ACM SIGMETRICS Performance Evaluation Review*, volume 40, pages 175–186. ACM, 2012.

Tan Lu and Minghua Chen. Simple and effective dynamic provisioning for power-proportional data centers. In *46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2012.

Jianying Luo, Lei Rao, and Xue Liu. Data center energy cost minimization: a spatio-temporal scheduling approach. In *2013 Proceedings IEEE INFOCOM (Mini Conference)*, pages 340–344. IEEE, 2013.

Jianying Luo, Lei Rao, and Xue Liu. Temporal load balancing with service delay guarantees for data center energy cost optimization. *IEEE Transactions on Parallel and Distributed Systems*, 25(3):775–784, 2014.

Vimal Mathew, Ramesh K Sitaraman, and Prashant J Shenoy. Reducing energy costs in internet-scale distributed systems using load shifting. In *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS)*, pages 1–8, 2014.

David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

David Meisner, Brian T Gold, and Thomas F Wenisch. Powernap: eliminating server idle power. In *ACM Sigplan Notices*, volume 44, pages 205–216. ACM, 2009.

K Meng, ZY Dong, and KP Wong. Self-adaptive radial basis function neural network for short-term electricity price forecasting. *IET Generation, Transmission & Distribution*, 3(4):325–335, 2009.

Ripal Nathuji and Karsten Schwan. Virtualpower: coordinated power management in virtualized enterprise systems. *ACM SIGOPS Operating Systems Review*, 41(6):265–278, 2007.

Michael J Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

TS Eugene Ng and Hui Zhang. Predicting internet network distance with coordinates-based approaches. In *Proceedings IEEE INFOCOM*, volume 1, pages 170–179. IEEE, 2002.

James B Orlin. A faster strongly polynomial minimum cost flow algorithm. *Operations research*, 41(2):338–350, 1993.

Darshan S Palasamudram, Ramesh K Sitaraman, Bhuvan Urgaonkar, and Rahul Urgaonkar. Using batteries to reduce the power costs of internet-scale distributed networks. In *Proceedings of the Third ACM Symposium on Cloud Computing*, page 11. ACM, 2012.

John G Proakis. *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001.

Ignacio Quesada and Ignacio E Grossmann. An lp/nlp based branch and bound algorithm for convex minlp optimization problems. *Computers & chemical engineering*, 16(10):937–947, 1992.

Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the electric bill for internet-scale systems. *ACM SIGCOMM Computer Communication Review*, 39(4):123–134, 2009.

Ramya Raghavendra, Parthasarathy Ranganathan, Vanish Talwar, Zhikui Wang, and Xiaoyun Zhu. No power struggles: Coordinated multi-level power management for the data center. In *ACM SIGARCH Computer Architecture News*, volume 36, pages 48–59. ACM, 2008.

Ashikur Rahman, Xue Liu, and Fanxin Kong. A survey on geographic load balancing based data center power management in the smart grid environment. *IEEE Communications Surveys & Tutorials*, 16(1), 2014.

Lei Rao, Xue Liu, Marija Ilic, and Jie Liu. MEC-IDC: joint load balancing and power control for distributed internet data centers. In *Proceedings of the 1st ACM/IEEE International Conference on Cyber-Physical Systems*, pages 188–197. ACM, 2010a.

Lei Rao, Xue Liu, Le Xie, and Wenyu Liu. Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment. In *Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010b.

Lei Rao, Xue Liu, Le Xie, and Zhan Pang. Hedging against uncertainty: A tale of internet data center operations under smart grid environment. *IEEE Transactions on Smart Grid*, 2(3):555–563, 2011.

Lei Rao, Xue Liu, Marija D Ilic, and Jie Liu. Distributed coordination of internet data centers under multiregional electricity markets. *Proceedings of the IEEE*, 100(1):269–282, 2012a.

Lei Rao, Xue Liu, Le Xie, and Wenyu Liu. Coordinated energy cost management of distributed internet data centers in smart grid. *IEEE Transactions on Smart Grid*, 3(1):50–58, 2012b.

Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Towards understanding heterogeneous clouds at scale: Google trace analysis. *Intel Science and Technology Center for Cloud Computing, Tech. Rep*, 2012.

Shaolei Ren, Yuxiong He, and Fei Xu. Provably-efficient job scheduling for energy and fairness in geographically distributed data centers. In *IEEE 32nd International Conference on Distributed Computing Systems (ICDCS)*, pages 22–31. IEEE, 2012.

Huajie Shao, Lei Rao, Zhi Wang, X Liu, and K Ren. Optimal load balancing and energy cost management for internet data centers in deregulated electricity markets. *IEEE Transactions on Parallel and Distributed Systems*, 2013.

Petter Skantze, Marija Ilic, and Jeffrey Chapman. Stochastic modeling of electric power prices in a multi-market environment. In *IEEE Power Engineering Society Winter Meeting*, volume 2, pages 1109–1114. IEEE, 2000.

Solarbuzz. Solar electricity prices. *http://solarbuzz.com/*.

J Michael Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Christopher Stewart and Kai Shen. Some joules are more precious than others: Managing renewable energy in the datacenter. In *Proceedings of the Workshop on Power Aware Computing and Systems*, 2009.

Goran Strbac, Anser Shakoor, Mary Black, Danny Pudjianto, and Thomas Bopp. Impact of wind generation on the operation and development of the uk electricity systems. *Electric Power Systems Research*, 77(9):1214–1227, 2007.

Jinlong Tu, Lian Lu, Minghua Chen, and Ramesh K Sitaraman. Dynamic provisioning in next-generation data centers with on-site power production. In *Proceedings of the fourth international conference on Future energy systems (e-Energy)*, pages 137–148. ACM, 2013.

Jonathan Turner. New directions in communications (or which way to the information age?). *IEEE communications Magazine*, 24(10):8–15, 1986.

Rahul Urgaonkar, Bhuvan Urgaonkar, Michael J Neely, and Anand Sivasubramaniam. Optimal power cost management using stored energy in data centers. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 221–232. ACM, 2011.

U.S. Environmental Protection Agency. Report to congress on server and data center energy efficiency. *http://hightech.lbl.gov/documents/data_centers/epa-datacenters.pdf*, 2007.

Luis N Vicente and Paul H Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global optimization*, 5(3):291–306, 1994.

Cormac Walsh. *Maximal Effective Bandwidth Under "leaky Bucket" Constraints*. Hewlett-Packard Laboratories, 2001.

Peijian Wang, Lei Rao, Xue Liu, and Yong Qi. Dynamic power management of distributed internet data centers in smart grid environment. In *IEEE Global Telecommunications Conference (GLOBECOM)*, pages 1–5. IEEE, 2011.

Peijian Wang, Lei Rao, Xue Liu, and Yong Qi. D-pro: dynamic data center operations with demand-responsive electricity prices in smart grid. *IEEE Transactions on Smart Grid*, 3(4):1743–1754, 2012.

Dan Xu and Xin Liu. Geographic trough filling for internet datacenters. In *Proceedings IEEE INFOCOM*, pages 2881–2885. IEEE, 2012.

Hong Xu and Baochun Li. Joint request mapping and response routing for geo-distributed cloud services. In *Proceedings IEEE INFOCOM*. IEEE, 2013.

Jianguo Yao, Xue Liu, Wenbo He, and Ashikur Rahman. Dynamic control of electricity cost with power demand smoothing and peak shaving for distributed internet data centers. In *IEEE 32nd International Conference on Distributed Computing Systems (ICDCS)*, pages 416–424. IEEE, 2012a.

Yuan Yao, Longbo Huang, Abhishek Sharma, Leana Golubchik, and Michael Neely. Data centers power reduction: A two time scale approach for delay tolerant workloads. In *2012 Proceedings IEEE INFOCOM*, pages 1431–1439. IEEE, 2012b.

Yanwei Zhang, Yefu Wang, and Xiaorui Wang. Greenware: greening cloud-scale data centers to maximize the use of renewable energy. In *Middleware*, pages 143–164. Springer, 2011.

Rongliang Zhou, Zhikui Wang, Alan McReynolds, Cullen E Bash, Thomas W Christian, and Rocky Shih. Optimization and control of cooling microgrids for data centers. In *13th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 338–343. IEEE, 2012.