UNIVERSITY OF CALIFORNIA, SAN DIEGO

Energy and Cost Efficient Data Centers

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Baris Aksanli

Committee in charge:

Professor Tajana Simunic Rosing, Chair
Professor Sujit Dey
Professor Rajesh Gupta
Professor Ryan Kastner
Professor Dean Tullsen

2015

The dissertation of Baris Aksanli is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

To my parents, Yasemin and Mustafa,
my brother, Başar,
and Tarçın.

EPIGRAPH

But he did not understand the price. Mortals never do. They only see the prize.
Their hearts desire, their dream...
But the price of getting what you want is getting what you once wanted.
—Sandman (Neil Gaiman)

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables xi

Acknowledgments xiii

Vita xvi

Abstract of the Dissertation xix

Chapter 1 Introduction 1
1.1 Renewable Energy in Data Centers 4
1.2 Energy Efficient Network of Data Centers Connected with a Wide Area Network 5
1.3 Peak Power Aware Data Centers 7
1.4 Data Centers in the Grid 8
1.5 Thesis Contributions 9

Chapter 2 Renewable Energy in Data Centers 12
2.1 Related Work 13
2.1.1 Energy Prediction 13
2.1.2 Green Energy in Data Centers 14
2.2 Solar and Wind Energy Prediction 15
2.2.1 Solar Prediction Methodology 16
2.2.2 Wind Prediction Methodology 16
2.3 Green Energy Scheduling and Data Center Modeling . 17
2.3.1 Model Validation Using Experimental Testbed . 20
2.4 Results 21
2.5 Conclusion 24

Chapter 3 Renewable Energy in Wide Area Networks 25
3.1 Related Work 27
3.2 Data Center and Network Modeling 29
3.2.1 Data Center Model 29

3.2.2 Backbone Network 31
3.2.3 Simulation of Backbone Network with Data Centers 36
3.3 Results 36
3.4 Conclusion 42

Chapter 4 Energy Efficiency in Networks of Data Centers 43
4.1 Background and Related Work 44
4.2 Cost Minimization and Performance Maximization Algorithms 47
4.3 Methodology 51
4.4 Results 52
4.4.1 No Migration 52
4.4.2 Performance Maximization Using Migration . . 53
4.4.3 Cost Minimization Using Migration 54
4.4.4 Cost Min. Using a Green Energy Aware Network 57
4.5 Discussion 58
4.6 Conclusion 62

Chapter 5 Efficient Peak Power Shaving in Data Centers 63
5.1 Issues with the Existing Battery Placement Architectures 68
5.1.1 Centralized vs. Distributed Designs 68
5.1.2 Problems of the Distributed Design 71
5.1.3 DC Power Delivery in Data Centers 73
5.2 Detailed Battery Model 74
5.2.1 Battery Model Validation 77
5.3 Distributed Battery Control 78
5.4 Grid-tie Based Battery Placement 80
5.5 Methodology 84
5.5.1 Power Measurements and Workloads Run 85
5.5.2 Data center and Battery Simulation 87
5.5.3 Cost Models 88
5.6 Results 92
5.6.1 Accuracy of the Detailed Battery Model 92
5.6.2 Effects of Detailed Battery Model on Savings . . 95
5.6.3 Peak Shaving Efficiency of State-of-the-Art Designs 95
5.6.4 Performance of the Distributed Battery Control 99
5.6.5 Performance of the Grid-tie Design 103
5.7 Conclusion 107

Chapter 6 Data Centers & the Grid 109
6.1 Background 112
6.2 Data Centers Providing Regulation Services 114

6.2.1 Fixed Average Power 115
6.2.2 Varying Average Power 117
6.3 Evaluation 121
6.3.1 Methodology 121
6.3.2 Results: Fixed Average Power 122
6.3.3 Results: Varying Average Power 125
6.4 Conclusion 128

Chapter 7 Summary and Future Work 130
7.1 Thesis Summary 130
7.1.1 Renewable Energy in Data Center Systems . . . 131
7.1.2 Efficient Peak Power Shaving in Data Centers . 132
7.1.3 Data Centers in the Grid 133
7.2 Future Work Directions 134
7.2.1 Data Centers Causing Instabilities in the Grid . 134
7.2.2 Residential Energy Management 135

Bibliography 137

LIST OF FIGURES

Figure 1.1: Data center components [28] 2

Figure 2.1: System Architecture 18
Figure 2.2: Additional batch jobs running with predicted green energy . . 19
Figure 2.3: Average completion time of MapReduce jobs 23

Figure 3.1: Power curves for different network power schemes 33
Figure 3.2: Network Topology; squares = data centers, circles = routers . 33
Figure 3.3: Solar and wind energy availability 34
Figure 3.4: Green energy aware routing algorithm 35
Figure 3.5: MapReduce job completion time and power vs. bandwidth . . 39

Figure 4.1: Overview of the cost minimization algorithm 49
Figure 4.2: Daily brown and amortized green energy cost (¢/kWh) 52
Figure 4.3: Normalized performance maximization algorithm costs for data centers and network 54
Figure 4.4: Normalized cost minimization algorithm costs with different power tier levels and energy proportionality 56
Figure 4.5: Normalized total cost and utilization for cost min. with different power tier levels and network lease options using energy proportional servers 57
Figure 4.6: Comparison between SPR and GEAR energy consumption of routers and network profit with different energy proportionality schemes 58

Figure 5.1: Sample peak power shaving with batteries 64
Figure 5.2: Sample battery placement in data centers [73] 65
Figure 5.3: Different power delivery options with centralized and distributed battery placements 69
Figure 5.4: Peak power shaving comparison of centralized vs. distributed designs 70
Figure 5.5: Peak power shaving capabilities of the distributed design . . . 71
Figure 5.6: The maximum battery capacity with random battery selection 72
Figure 5.7: Effective capacity of 20Ah LA and LFP batteries 75
Figure 5.8: Cycle life of LA and LFP batteries rated at 20h [119][129] . . 76
Figure 5.9: Battery selection with communication based policies 81
Figure 5.10: Grid-tie based battery placement design 82
Figure 5.11: Data center workload mixture 86
Figure 5.12: DoD level variation 87
Figure 5.13: Avg. discharging current for the distributed design and grid-tie design over a 3 day period, with LFP batteries 88

Figure 5.14: Communication overhead analysis for the distributed control mechanisms 103
Figure 5.15: Communication overhead analysis for the grid-tie design . . . 107

Figure 6.1: Sample battery-based peak power shaving demonstration of a 21MW data center over 7 days 114
Figure 6.2: Regulation prices 123
Figure 6.3: Total savings result with CAISO and NYISO prices 124
Figure 6.4: Recharge shifting for NYSIO and CAISO 127

LIST OF TABLES

Table 2.1: Measured interference of MapReduce and Rubis 20
Table 2.2: Verification of simulation outputs 21
Table 2.3: Comparison of instantaneous and predicted green energy with different alternative energy sources 22
Table 2.4: Brown Energy for Inst. vs. Pred. Energy 23

Table 3.1: Inter-arrival and service time parameters 30
Table 3.2: Parameters and values used in the simulation 31
Table 3.3: Renewable energy availability in different locations 35
Table 3.4: Metrics and their definitions 37
Table 3.5: Baseline results: all bandwidth (AB), necessary bandwidth (NB) 38
Table 3.6: Summary of key results using green energy in wide area networks 40

Table 4.1: Summary and comparison of the related work 46
Table 4.2: Profit of network providers for performance maximization with different router energy proportionality schemes 55
Table 4.3: Profit of network providers for cost. min. with different router energy prop. and with server energy prop. 56
Table 4.4: Comparison of different policies with respect to total cost, MapReduce performance and green energy usage 60

Table 5.1: Group sizes, equivalent hierarchy level and the best beak power shaving performance for each group 73
Table 5.2: Battery model validation 77
Table 5.3: Group sizes in data center power delivery hierarchy 79
Table 5.4: Policies to control distributed batteries 80
Table 5.5: Comparison between the grid-tie design and the state-of-the-art (SoA) designs 85
Table 5.6: Workload parameters 86
Table 5.7: TCO/server breakdown for different designs. The components with different trends are highlighted 91
Table 5.8: Input parameters for the cost models 92
Table 5.9: Peak shaving capabilities of the square peak with centralized and distributed designs 94
Table 5.10: Battery lifetime estimation comparison 94
Table 5.11: CLR cost savings for distributed LA and LFP batteries 96
Table 5.12: TCO/server savings for distributed LA and LFP batteries . . . 96
Table 5.13: Centralized design peak shaving capabilities with different battery types. $P_{threshold}$ is set to 255W/server 97

Table 5.14: Peak shaving and battery recharging comparison of the distributed design with different battery types and AC vs. DC power options. $P_{threshold}$ is set to 255W/server 98
Table 5.15: Efficiency of centralized vs. distributed designs with different power equipment and delivery options 100
Table 5.16: Amount of energy shaved for a 10MW data center per week in MWh and % of power shaved compared to the peak 101
Table 5.17: Normalized average battery lifetime 102
Table 5.18: Peak shaving capabilities of the grid-tie design compared to the distributed design. $P_{threshold}$ is set to 255W per server 105
Table 5.19: Grid-tie vs. distributed design. EB=Extra Batteries, BL=Battery Lifetime, PS=Peak Shaving, ES=Extra Servers 105

Table 6.1: Best peak shaving percentages with different DoD levels 122
Table 6.2: Maximum PAR values for different DoD levels 123
Table 6.3: Error percentages if peak power costs are not considered . . . 125
Table 6.4: Regulation price analysis for battery discharge intervals 126
Table 6.5: Monthly savings using recharge shifting 126

ACKNOWLEDGMENTS

I would like to take this opportunity to thank many individuals who directly or indirectly supported me during my PhD process.

First, I want to thank my advisor Prof. Tajana Rosing for her guidance throughout my PhD. She has helped me understand what research truly is and constantly encouraged me to pursue my research studies. I feel extremely lucky and grateful as she let me be a part of her research team. I also want to thank my doctoral committee, Prof. Rajesh Gupta, Prof. Ryan Kastner, Prof. Dean Tullsen and Prof. Sujit Dey for their valuable feedback and contributions to my PhD.

I have to thank Inder Monga, who was my supervisor when I was an intern in Energy Sciences Network (Esnet) of Lawrence Berkeley National Laboratory. I am grateful to him for providing me such an opportunity and supporting me both during and after my internship. I also want to thank Eddie Pettis from Google, for his patience and time in answering my questions during our collaboration. It provided me invaluable experience. My research was made possible by funding from National Science Foundation (NSF) Project GreenLight, NSF ERC CIAN Grant 812072, NSF Flash Gordon, NSF IRNC TransLight/StarLight Grant 962997, Multiscale Systems Center (MuSyC), Terraswarm Research Center, UCSD Center for Networked Systems (CNS), Oracle, Microsoft, and Google. I thank them for their generous support.

My lab mates and colleagues helped me by their comments, discussions and guidance throughput my PhD. I sincerely thank all of them, either current or past, but owe special thanks to Jagannathan Venkatesh, Alper Sinan Akyurek, Christine Chan. I also want to thank my seniors Gaurav Dhiman, Vasileios Kontorinis and Richard Strong for their help and guidance during my early times of PhD.

I also want to thank my friends, Celal Ziftci, Efecan Poyraz, Furkan Can Kavasoglu, Doruk Beyter, Can Bal, and Ozgur Yigit Balkan who made the life in San Diego more enjoyable and memorable for me. I want to especially thank to Arda Kilinc for always being there as a friend and supporting me even though he was not in San Diego.

Last but most importantly, I want to thank my family, my parents, Yasemin

and Mustafa, and my brother Başar. They not only helped me go through my PhD, but have been always there to unconditionally support me. I can never fully explain the love I feel for them with words and will forever be grateful to them for everything they did for me. I am also grateful to have Tarçın, who has always been a very good and loyal friend to me.

Chapters 1 and 2 contain material from "Using datacenter simulation to evaluate green energy integration", by Baris Aksanli, Jagannathan Venkatesh and Tajana Simunic Rosing, which appears in IEEE Computer 45, September 2012 [19]. The dissertation author was the primary investigator and author of this paper.

Chapter 2 contains material from "Utilizing Green Energy Prediction to Schedule Mixed Batch and Service Jobs in Data Centers", by Baris Aksanli, Jagannathan Venkatesh, Liuyi Zhang and Tajana Simunic Rosing, which appears in ACM SIGOPS Operating Systems Review 45, no. 3, 2012 [20]. The dissertation author was the primary investigator and author of this paper.

Chapter 3 contains material from "Benefits of Green Energy and Proportionality in High Speed Wide Area Networks Connecting Data Centers", by Baris Aksanli , Tajana Rosing, and Inder Monga, which appears in Proceedings of Design Automation and Test in Europe (DATE), 2012 [17]. The dissertation author was the primary investigator and author of this paper.

Chapter 4 contains material from "A Comprehensive Approach to Reduce the Energy Cost of Network of Datacenters", by Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga, which appears in Proceedings of International Symposium on Computers and Communications (ISCC), 2013 [18]. The dissertation author was the primary investigator and author of this paper.

Chapter 4 contains material from "Renewable Energy Prediction for Improved Utilization and Efficiency in Datacenters and Backbone Networks", by Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga, which will appear in Computational Sustainability, Springer, 2015 [11]. The dissertation author was the primary investigator and author of this paper.

Chapter 5 contains material from "Distributed Battery Control for Peak Power Shaving in Data Centers", by Baris Aksanli, Tajana Rosing and Eddie

Pettis, which appears in Proceedings of International Green Computing Conference (IGCC), 2013 [16]. The dissertation author was the primary investigator and author of this paper.

Chapter 5 contains material from "Architecting Efficient Peak Power Shaving Using Batteries in Data Centers", by Baris Aksanli, Eddie Pettis, and Tajana Rosing, which appears in Proceedings of International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), 2013 [13]. The dissertation author was the primary investigator and author of this paper.

Chapter 6 contains material from "Providing Regulation Services and Managing Data Center Peak Power Budgets", by Baris Aksanli and Tajana Rosing, which appears in Proceedings of Design Automation and Test in Europe (DATE), 2014 [15]. The dissertation author was the primary investigator and author of this paper.

Chapter 7 contains material from "Optimal Battery Configuration in a Residential Home with Time-of-Use Pricing", by Baris Aksanli and Tajana Rosing, which appears in Proceedings of International Conference on Smart Grid Communications (SmartGridComm), 2013 [14]. The dissertation author was the primary investigator and author of this paper.

VITA

2010	B. S. in Mathematics, Bogazici University, Istanbul, Turkey
2010	B. S. in Computer Engineering, Bogazici University, Istanbul, Turkey
2010-2015	Graduate Student Researcher, University of California, San Diego, CA
2011	Intern, Energy Sciences Network (Esnet), Lawrence Berkeley National Laboratory, Berkeley, CA
2012	Intern, Datacenter Group (DCG), Intel Corporation, Hillsboro, OR
2012	M. S. in Computer Science, University of California, San Diego, CA
2013	Teaching Assistant, University of California, San Diego, CA
2015	Ph. D. in Computer Science, University of California, San Diego, CA

PUBLICATIONS

Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga. "Renewable Energy Prediction for Improved Utilization and Efficiency in Datacenters and Backbone Networks." *Computational Sustainability, Springer*. To appear in 2015.

Baris Aksanli, Alper S. Akyurek, Madhur Behl, Meghan Clark, Alexandre Donze, Prabal Dutta, Patrick Lazik, Mehdi Maasoumy, Rahul Mangharam, Truong X. Nghiem, Vasumathi Raman, Anthony Rowe, Alberto Sangiovanni-Vincentelli, Sanjit Seshia, Tajana Simunic Rosing, and Jagannathan Venkatesh. 2014. "Distributed control of a swarm of buildings connected to a smart grid: demo abstract." In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys '14)*. pp. 172-173. November 2014.

Baris Aksanli and Tajana Rosing, "Providing regulation services and managing data center peak power budgets." In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE '14)*. pp.1-4. March 2014.

Baris Aksanli and Tajana Rosing, "Optimal battery configuration in a residential home with time-of-use pricing." In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on* , pp.157-162, October 2013.

Baris Aksanli, Eddie Pettis, and Tajana Rosing. 2013. "Architecting Efficient Peak Power Shaving Using Batteries in Data Centers." In *Proceedings of the 2013 IEEE 21st International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS '13)*. pp. 242-253. August 2013.

Baris Aksanli, Jagannathan Venkatesh, Tajana Simunic Rosing, and Inder Monga, "A comprehensive approach to reduce the energy cost of network of datacenters." In *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pp.275-280, July 2013.

Jagannathan Venkatesh, Baris Aksanli, and Tajana Simunic Rosing, "Residential energy simulation and scheduling: A case study approach." In *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pp.161-166, July 2013.

Baris Aksanli, Tajana Rosing, and Eddie Pettis, "Distributed battery control for peak power shaving in datacenters." In *Proceedings of the International Green Computing Conference (IGCC)*, pp.1-8, June 2013.

Jagannathan Venkatesh, Baris Aksanli, Tajana Rosing, Jean-Claude Junqua, and Philippe Morin, "HomeSim: Comprehensive, Smart, Residential Energy Simulation and Scheduling." In *Proceedings of the International Green Computing Conference (IGCC)*, pp.1-8, June 2013.

Baris Aksanli, Jagannathan Venkatesh, and Tajana Simunic Rosing, "Using Datacenter Simulation to Evaluate Green Energy Integration." *Computer* , vol.45, no.9, pp.56-64, September 2012.

Vasileios Kontorinis, Liuyi Eric Zhang, Baris Aksanli, Jack Sampson, Houman Homayoun, Eddie Pettis, Dean M. Tullsen, and Tajana Simunic Rosing. "Managing distributed ups energy for effective power capping in data centers." In *Proceedings of the 39th Annual International Symposium on Computer Architecture (ISCA '12)*. pp. 488-499. June 2012.

Baris Aksanli, Tajana Simunic Rosing, and Inder Monga. "Benefits of green energy and proportionality in high speed wide area networks connecting data centers." In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE '12)*. pp. 175-180. March 2012.

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, no.3 pp. 53-57. January 2012.

Alper Sen, Baris Aksanli, and Murat Bozkurt. "Speeding up cycle based logic simulation using graphics processing units." *International Journal of Parallel Programming* vol.39 no.5 pp. 639-661. October 2011.

Alper Sen, Baris Aksanli, and Murat Bozkurt. "Using Graphics Processing Units for Logic Simulation of Electronic Designs." In *Microprocessor Test and Verification (MTV), 2010 11th International Workshop on*, pp. 73-76. IEEE, December 2010.

Alper Sen, Baris Aksanli, Murat Bozkurt, and Melih Mert. "Parallel cycle based logic simulation using graphics processing units." In *Parallel and Distributed Computing (ISPDC), 2010 Ninth International Symposium on*, pp. 71-78. IEEE, July 2010.

ABSTRACT OF THE DISSERTATION

Energy and Cost Efficient Data Centers

by

Baris Aksanli

Doctor of Philosophy in Computer Science

University of California, San Diego, 2015

Professor Tajana Simunic Rosing, Chair

Data centers need efficient energy management mechanisms to reduce their consumption, energy costs and the resulting negative grid and environmental effects. Many of the state of the art mechanisms come with performance overhead, which may lead to service level agreement violations and reduce the quality of service. This thesis proposes novel methods that meet quality of service targets while decreasing energy costs and peak power of data centers.

We leverage short term prediction of green energy as a part of our novel data center job scheduler to significantly increase the green energy efficiency and job throughput. We extend this analysis to distributed data centers connected with a backbone network. As a part of this work, we devise a green energy aware routing algorithm for the network, thus reducing its carbon footprint.

Consumption during peak periods is an important issue for data centers due to its high cost. Peak shaving allows data centers to increase their computational capacity without exceeding a given power budget. We leverage battery-based solutions because they incur no performance overhead. We first show that when using an idealized battery model, peak shaving benefits can be overestimated by 3.35x. We then present a distributed control mechanism for a more realistic battery system that achieves 10x lower communication overhead than the centralized solution. We also demonstrate a new battery placement architecture that outperforms existing designs with better peak shaving and battery lifetime, and doubles the savings.

Data centers are also good candidates for providing ancillary services in the power markets due to their large power consumption and flexibility. This thesis develops a framework that explores the feasibility of data center participation in these markets, focusing specifically on regulation services. We use a battery-based design to not only help by providing ancillary services, but to also limit peak power costs without any workload performance degradation.

Chapter 1

Introduction

Recent improvements in computer and network architectures have made internet-based applications and cloud computing systems popular. Some companies, such as Google, Amazon, Facebook, Microsoft, have multiple, geographically distributed data centers with thousands to millions of servers. One important problem of these huge computation-oriented structures is their energy consumption due to their significant demand. A recent study shows that the total energy consumption of all data centers in the world has increased by 56% from 2005 to 2010 [74]. As a result of this important issue, there is a large body of studies focusing on how to improve the energy efficiency in data centers. Even a small efficiency improvement translates into millions of dollars savings for large scale data centers. This topic will continue to be important in the future as the price of brown energy, the energy produced by non-renewable resources, rise due to additional taxes placed on carbon emissions [87]. Energy efficient solutions, ranging from utilizing green energy sources, such as solar and wind, to optimizing HW, SW and system design for energy efficiency, along with peak-power aware solutions will continue to be important.

The individual elements in a data center can be classified into two categories: IT and non-IT elements [28]. The former includes the components that do the computation whereas the latter maintain the functionality of the whole system. Servers comprise a large portion of the overall energy cost in IT component, with networking infrastructure, such as switches and routers, being a relatively

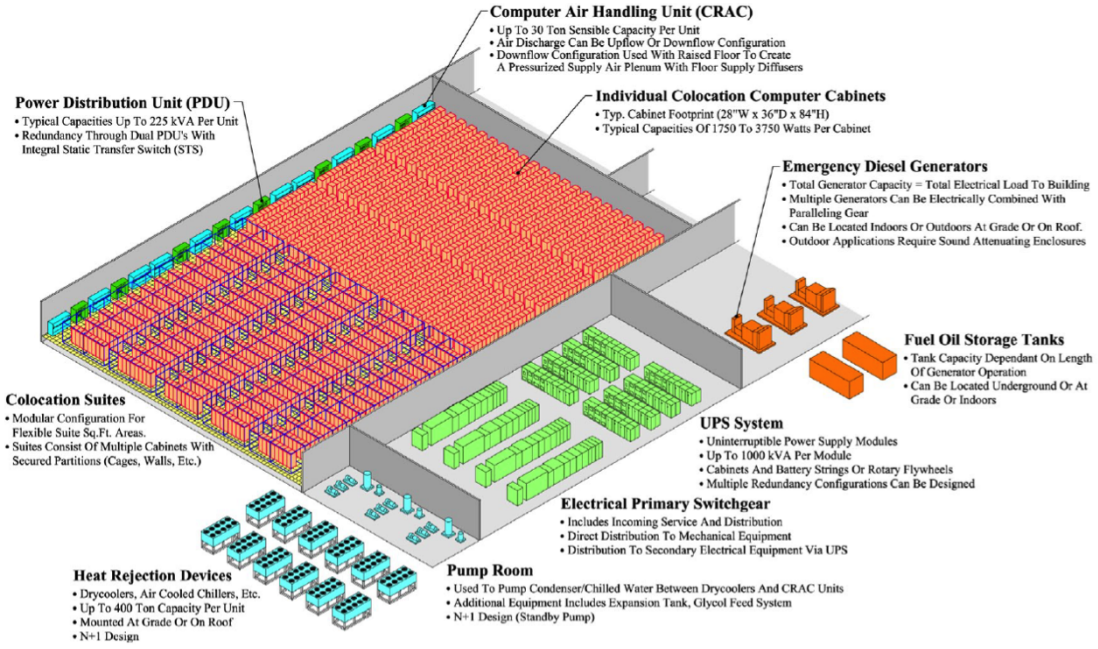


Figure 1.1: Data center components [28]

smaller component. There are a number of non-IT elements that contribute to the high energy costs of the data centers. Examples include power distribution units (PDU) which provide power to the IT elements, uninterruptible power supplies (UPS) used for emergency situations, and computer room air conditioners (CRAC) that keep the data center temperature within determined limits to ensure server reliability. Figure 1.1 shows all these components and how they interact.

In order to maintain the health of the electric grid and take advantage of the large energy consumption of data centers, energy providers charge data centers based on not only their energy use but also on their peak power level, which is measured as the highest power demand of the data center in a given billing period, e.g. a month. This is because the peak power level, especially for buildings that require considerable amount of power, determines the number of generators (i.e. power plants) that the energy providers need to activate at a given time. This activation can drastically impact the cost of energy production for energy providers. The cost of peak power is generally much higher than the cost of energy and can contribute up to 50% of the electricity bill [62] if data

centers do not carefully adjust their peak power levels. Data centers can apply traditional power management methods, such as dynamic voltage and frequency scaling (DVFS), to control their peak power levels at the expense of performance overhead. Recent studies propose new mechanisms that reduce the peak power of data centers without affecting the performance.

A power/energy management method that a data center uses can change its power demand significantly. Since data centers are, in nature, large buildings with high power demands, the fluctuations in their demand can automatically affect the dynamics of the electric grid. Thus, it is critical to study the relation between data centers and the grid. The recent work on this topic explores the opportunities for both data centers [10] and the utilities (energy providers) to exploit such a collaboration. Utilities generally allow their customers to help balance the energy supply and demand by creating ancillary services such as demand response, regulation services, spinning and non-spinning reserves. Each service has different properties, such as timing requirements, capacity allocation, etc. and thus, has separate compensation rates. Data centers can participate in this ancillary services market and make extra profits for their services. Although this seems as a mutually beneficial operation for data centers and the utilities, data centers should carefully allocate their resources for such operations to avoid an increase in their energy costs. This is because the contracts of power/energy usage and ancillary service participation are generally made separately .

This thesis proposes new energy and cost efficient solutions for data centers, and explores data center participation in regulation markets, one of the well-known ancillary services. It first analyzes how renewable energy can effectively be used in data centers and wide area networks connecting data centers. It then extends this analysis to a network of data centers where each data center is treated individually, but also modeled such that they can coordinate to increase the overall efficiency of the whole system. We continue with the peak power aware solutions to decrease the peak-to-average ratio of data centers in order to obtain savings. Our main focus on this part is to use batteries and using them as efficiently as possible. Lastly, we study data center - grid interaction, while focusing on regulation services. We

investigate how data centers can participate in these markets effectively.

Next, we introduce and classify different energy management mechanisms for data center systems. These mechanisms range from a single data center to a network of data centers connected with wide area networks. They include renewable energy usage, job migration to increase the efficiency of multiple data center systems, peak power shaving, and grid-connected mechanisms.

1.1 Renewable Energy in Data Centers

Data center energy needs are supplied mainly by non-renewable, or brown energy sources, which are increasingly expensive as a result of availability and the introduction of carbon emissions taxes [87]. Consequently, several data center operators have turned to renewable energy to offset the energy cost. The integration of renewable energy is complicated by the inherent variability of its output. Output inconsistency typically leads to inefficiency due to lack of availability or suboptimal proportioning, which carries an associated financial cost. These costs are mitigated in various ways: several data center owners, such as Emerson Networks, AISO.net, and Sonoma Mountain Data Center supplement their solar arrays with utility power, and other data center owners, such as Baronyx Corporation and Other World Corporation, have been forced to augment their input power with other forms of energy or through over-provisioning, respectively [59]. Previous investigation into the state of the art in data center green energy demonstrates that variability results in low utilization, on average 54%, of the available renewable energy [90].

Previous studies have investigated the several strategies to manage renewable energy as a part of data center operation. The work in [59] reduces the peak data center power with local renewable sources and power management algorithms. They investigate power capping, both of individual servers using dynamic frequency scaling, and of server pools by reducing the number of machines utilized in each pool. However, they have significant quality-of-service (QoS) violations when limiting peak power. The study in [90] explores brown energy capping in

data centers, motivated by carbon limits in cities such as Kyoto. The authors leverage distributed Internet services to schedule workloads based on electricity prices or green energy availability. By defining workload distribution as a local optimization problem, the authors demonstrated 35% lower brown energy consumption with a nominal (10%) hit on service level agreement (SLA) violations. The authors of [75] analyze the opportunities and problems of using supply-following loads to match green energy availability. When green energy is insufficient, workloads are terminated or suspended, restarting or resuming when availability returns. However, the results show very low green energy efficiency and a failure to meet required service-level guarantees. The above data center examples demonstrate the necessity of integrating renewable energy into the data centers, but do not address their highly variable nature, leading to severe under-utilization of these alternative energy resources.

1.2 Energy Efficient Network of Data Centers Connected with a Wide Area Network

Multiple data center networks offer great advantages in terms of both performance and energy. As each data center is located in a different location, their peak hours and electricity prices vary. The data center with the higher electricity price can send some of its workload to another data center with a lower price. Additionally, data centers generally contract a fixed price for a specific amount of energy, which is known as the tiered electricity price. This fixed price changes depending on the location, and it is beneficial to run jobs in a data center with a lower fixed price. However, the transfer should not increase the utilization so that the power consumption is more than the tier-price. Furthermore, the peak power costs with increased workloads can be high, omitting the savings obtained by relocating the workloads. The live migration of virtual machines over WAN has made this idea feasible, as it offers fast transmission without a serious performance hit [122].

Existing studies that are concerned with energy costs primarily propose "fol-

low the sun" or cheaper cost of brown energy strategies and generally neglect the cost of wide area networking (WAN) which is incurred for job migrations across the globe. The cost of WAN is relatively small when the data center operators own a WAN between their geographically distributed sites. Examples of work that leverages this idea include proposals where WANs are used to increase performance of the overall system via 1) reducing the electricity cost using only brown energy [33], 2) choosing the most suitable location for a new data center to be deployed [91], 3) minimizing the cost with different local brown energy markets [109], 4) migrating jobs to load balance the data centers in different locations [83], 5) capping the brown energy using utility green energy with different pricing in different locations [79]. However, their arguments are not applicable for large WAN costs and data centers that lease the network. Additionally, the large scale of these data center systems makes it hard to analyze them as a whole. Consequently, studies focusing on these networks model the data center parts only relevant to their problems and neglect the rest. This might result in significant inaccuracies when computing the associated savings.

WANs connect the geographically distributed data centers. Previous studies neglect the energy costs of the WANs, primarily because their energy consumption was considered insignificant compared to data centers. However, not all data center systems own their own network and might need to rent this service from a network provider. As the energy becomes more expensive or less available, these providers tend to charge their customers more to compensate for the high costs they have to pay. The network providers [54] also have Bandwidth-on-Demand services, especially for applications across multiple data centers [85]. The cost of these Bandwidth-on-Demand services can dramatically increase as WAN usage increases, since the network providers tend to charge their customers more with increased demand. They can also charge their customers based on the time of day when the WAN is used. For instance, the network might be more expensive in a peak hour compared to an idle period, aligning with electricity market prices [131]. Then, the increased cost of using the WAN turns into an important part of the data center costs.

1.3 Peak Power Aware Data Centers

Peak power costs are an important part of the data center utility bill. The data centers are charged based on the peak power level they achieve in a bill period, even though they operate at peak level very rarely. This phenomenon, thus, leads to high costs for data centers. Data centers can use peak power shaving methods to keep their peak power level below a predetermined power threshold, so that they can limit these costs. The peak power shaving methods can also allow data center providers to increase their computational capacity without exceeding a given power budget.

Data centers can leverage already existing energy management mechanisms to reduce their peak power level. These mechanisms include well-known dynamic voltage and frequency scaling based methods [50][88], virtual machine management (such as consolidation and resource management) [93], online job migration as described in the previous subsection [33][109]. Since these mechanisms rely on changing the computational resource dynamics, they may result in significant performance degradation. In contrast, recent work establishes that machines may repurpose energy from uninterruptible power supplies (UPSs) (i.e. batteries) to maintain power budgets during peak demand. The idea is to adjust the battery charge/discharge periods to make sure that the power threshold is not violated. Since these battery-based solutions do not interfere with workloads, they do not introduce any performance overhead. This is especially critical during the peak demand periods when several applications need fast response times simultaneously.

There are multiple approaches to use batteries for peak power shaving. The first approach is to use the existing batteries within the centralized UPS [62]. Nonetheless, this method is applicable to only short peaks because the UPS powers the entire data center. In addition to batteries, Wang et al. [125] analyze flywheels and ultra-capacitors for peak shaving, and identify which energy storage device might the most suitable for a given power demand curve. Kontorinis et al. [73] and Palasamudram et al. [97] propose overprovisioned distributed batteries to sustain longer peaks. This design leads to finer grained battery output control. But, batteries require high discharge current since each one powers an entire server.

High discharging current reduces both the effective battery capacity and the useful battery lifetime [45]. The models that these studies use cannot capture the negative effects of high discharge currents due to simplistic battery models and overestimate the battery lifetime. The distributed UPS implementations also require another layer of control to manage the distributed batteries at large scale. Since the number of batteries scales up with the number of servers in this design, the overhead of this battery management system can be serious, resulting in a slow response time to the peak power spikes and consequently a power budget violation.

1.4 Data Centers in the Grid

Recently, researchers have started to study the relations between data centers and the electric grid. They mainly model these interactions in the form of ancillary services and estimate the amount of savings data centers can obtain. These ancillary services include regulation services, demand response, voluntary load reduction and spinning and non-spinning reserves.

Out of the ancillary services, participating in regulation markets is the one that is most studied due to its higher return. But, this higher return requires fast responses from data centers' end. Chen et al. [37] use server-level DVFS to create the power consumption flexibility required to participate in regulation services. The data center first chooses which market it participates in, i.e. either hour or day ahead. It then reports the regulation capacity it can provide to the grid. It adjusts its power consumption based on the requests coming from the grid. These requests can demand either an increase or a decrease in consumption within the capacity agreed previously. These requests demanding power changes are fulfilled with DVFS.

Another well-known service that data centers can provide is demand response (DR). Ghamkhari et al. [55] analyze how data centers can participate in demand response with clever job scheduling. In another study, the authors analyze the potential of data centers for demand response participation [81]. Aikema et al. [10] study different types of ancillary services and show which one is more profitable

given the workload profile of the data center. The services they include in their study include regulation services, spinning and non-spinning reserves, voluntary load reduction and emergency DR. They conclude that the regulation service is the most profitable service for data center participation in ancillary services. They use different power management methods such as load shifting, DVFS, and job rescheduling to create the necessary flexibility in data center power consumption.

1.5 Thesis Contributions

This thesis focuses on energy supply and global efficiency of data centers along with their interactions with the grid. It shows methods to increase the energy efficiency of data centers using green energy, multiple data center systems where the overall efficiency can be improved with online job migration, battery-based peak power shaving solutions and how data centers can interact and collaborate with the electric grid. The following discussion demonstrates the contributions and the outlines of the rest of the thesis:

- It presents a new data center job scheduling methodology that uses green energy prediction to mitigate the variability issue of the green energy resources. We develop a data center model based on 200 Intel Nehalem servers. Our model uses the measured data obtained on a test bed of these servers that run a mixture of latency-critical service and throughput oriented batch jobs. This mixture enables our model to have a realistic data center environment. Our scheduler makes sure that the service jobs complete within their required response times and improves the batch job performance by executing additional batch tasks with available green energy. The results show that our predictive job scheduler increases green energy efficiency by 3x, the amount of work performed by green energy over brown energy by 1.6x and reduces the number of jobs terminated due to the lack of instantaneously available green energy by 7.7x. The predictive scheduler is described in chapter 2.

- It analyzes how renewable energy can improve the efficiency of wide area networks connecting multiple data center systems. The main target is to

increase the renewable energy integration to the networking systems without performance penalties for service and batch jobs running in the data center. We quantify the energy cost of data transfers over wide area networks and show that moving jobs may not always be feasible due to this cost. We design a green energy aware routing algorithm (GEAR) that ensures the quality of service requirements of the data center workloads are met and improves the energy efficiency by 10x. The details of GEAR are described in chapter 3.

- It uses green energy prediction in local renewable energy sites and varying brown energy prices to propose an online job migration algorithm among data centers to reduce the overall cost of energy. We uniquely consider network constraints such as availability, link capacity and transfer delay at the same time, i.e we model the impact of the network and create a more holistic multiple data center model. We investigate tiered power pricing, which penalize the data center for exceeding a certain level of energy consumption, along with WAN leasing costs/cost models, which leverage energy-aware routing. We also analyze the impact of new technologies in data center WAN, such as energy-proportional routing, green energy aware routing, and analyze leasing vs. owning the WAN. We observe that green energy prediction helps significantly increase the efficiency of energy usage and enables network provisioning in a more cost effective way. Similarly, we show that using a WAN to transfer workloads between data centers increases the performance of batch jobs up to 27% with our performance maximization algorithm, and decreases the cost of energy by 30% compared to no data migration with our cost minimization algorithm. We show the potential for green energy to go beyond simply cost reduction to improving performance as well. Our analysis of leasing WAN shows that network providers can increase profits by charging data center owners by bandwidth, but data centers can still benefit by using dynamic routing policies to decrease their energy costs. We additionally analyze server and router energy proportionality, demonstrating increases in both data center cost savings and network provider profits. This study is shown in chapter 4.

- It re-analyzes existing peak shaving designs using more realistic battery models and finds that the benefits of peak shaving may be overestimated by up to 3.35x with simplistic models. We address the battery coordination problem of the distributed battery placement designs by proposing a distributed battery control mechanism that achieves the battery lifetime and power shaving performance within 3.3% and 6% of the best centralized solution with only 10% of its communication overhead. The coordination is required to both reduce the communication overhead and to maximize the battery lifetime. We also present a new peak power shaving architecture that has the capability to provide "just enough" current to the data center, at a level that optimizes the individual battery lifetime. Our design places batteries centrally using grid-tie inverters to partially power loads. This new architecture has 78% longer battery lifetime and doubles the cost savings compared to the best existing distributed designs. Also, since the batteries are placed together, the communication overhead is reduced by 4x. The details of the distributed control mechanism and the new battery placement architecture are presented in chapter 5.

- It proposes a framework that analyzes the data center participation in the regulation markets while also considering the peak power objectives. Our framework consists of two cases corresponding to different peak power assumptions for a data center. We present multiple methods to address each case. It first analyzes if providing regulation services is reasonable and then computes the regulation capacity to maximize savings. We leverage data from different utility markets and show that for a 21MW data center, up to $480,000/year savings can be obtained, and 5.08% increase in data center profit percentage. We present our framework in chapter 6.

Chapter 1 contains material from "Using datacenter simulation to evaluate green energy integration", by Baris Aksanli, Jagannathan Venkatesh and Tajana Simunic Rosing, which appears in IEEE Computer 45, September 2012 [19]. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Renewable Energy in Data Centers

Green energy sources promise to mitigate the issues surrounding nonrenewable generation, but their output is very susceptible to environmental changes. This limits the use of green energy in time-sensitive applications. Prediction can reduce the uncertainty of the available resources, allowing end-users to scale demand with the predicted supply [120]. Data centers are a significant source of energy consumption with an estimated 2% global greenhouse gas emissions attributed to them [127]. However, the time-sensitive nature of their service-level workloads has precluded the use of green energy, as jobs might need to be stopped when the available green energy drops [75].

Data centers also have longer-running batch jobs (on the order of tens of minutes [71]) whose performance is measured in terms of throughput and job completion times instead of latency guarantees (e.g. web crawling, index update in search engines, web log analysis [121]). A number of computing frameworks have been developed to simplify the process of those jobs. Examples include MapReduce [43], Dryad [69], and Pregel [86]. The fault-tolerant nature of these frameworks mitigates source instability, allowing execution of a subset of the tasks in a job in order to scale with the available energy, as well as allowing re-execution of cancelled tasks that have been stopped due to a sudden lack of input energy.

Green energy prediction over short time intervals (tens of minutes) allevi-

ates these issues by scaling the workload to the expected available green energy, resulting in better maintenance of forward progress and allowing more tasks/jobs to continue their execution even if instantaneous green energy supply drops below the necessary amount. The system offsets the remainder of the immediate need with brown energy with the assurance that over the prediction interval the average green energy will ultimately be available. This allows a more efficient use of the available energy; reducing the amount of wasted green energy and the number of tasks/jobs that must be re-executed; and ultimately, increasing the overall throughput of the data center.

The contribution of this chapter is to develop a new data center job scheduling methodology that effectively leverages green energy prediction. We simulate a data center of 200 Intel Nehalem servers using measured data obtained on a small test bed of Nehalem servers that ran a mix of services (Rubis [112]) and batch jobs (MapReduce [65]). Our scheduler ensures that the required response time targets for services are met while minimizing the completion times and maximizing the number of MapReduce tasks run. We use green power data from a solar installation in San Diego [3], and wind power from National Renewable Energy Laboratory (NREL) [4] as our sources of green energy. Our results show maximum increase of 3x in green energy usage efficiency, a 1.6x increase in the amount of work performed by green energy over brown energy, and a 7.7x reduction in the number of jobs terminated due to the lack of instantaneously available green energy.

2.1 Related Work

2.1.1 Energy Prediction

Solar energy prediction is typically obtained with estimated weighted moving average (EWMA) models, because of its relative consistency and periodic patterns [67]. As long as the weather conditions remain consistent within a period, the prediction is accurate, but becomes inaccurate, with mean error well over 20%, with frequent weather changes. Recent work utilizing small-scale solar genera-

tion uses a weather-conditioned moving average (WCMA), taking into account the mean value across days and a measured factor of solar conditions in the present day relative to previous days [100]. While this work provides only a single future interval of prediction, it specifically addresses inconsistent conditions, with a mean error of under 10%.

Wind energy prediction can be separated into two major areas: time-series analysis of power data; and wind speed prediction and conversion into power. Kusiak et al. [76] present a comparison of several methodologies of time-series modeling of wind farms. The boosting-tree algorithm with both wind speed and power data performs well in their analysis, while the integrated model, a time-series analysis utilizing only wind speed measurements, performs poorly for calculating wind power, likely due to the cubic relationship between wind speed and power. Giebel et al. [57] focuse on the latter, describing a number of meteorological models including Numerical Weather Prediction (NWP), which forecasts atmospheric conditions over longer term. They use the resulting predictions to simulate the output of a wind farm providing accurate estimates for 3-6 hour time periods. However, this comes at the cost of needed a whole data center to calculate prediction. Sanchez et al. [114] suggest a statistical forecasting system that generates power curves (wind speed vs. wind power) for each turbine based on meteorological information and machine characteristics. They then utilize the power curves and available wind data for forecasting.

2.1.2 Green Energy in Data Centers

Green energy usage in a data center environment is a relatively new topic. Gmach et al. [58] augment a data center with PV and municipal solid waste based energy. However, since solid waste energy supply is constant over time, they do not address the problem of variability in renewable energy supply. Lee et al. [79] model an optimization problem which uses the market prices of brown and green energy to decide how much energy of each type should be bought in each interval. They do not make server level scheduling decisions based on the amount of green energy.

Stewart and Shen [118] analyze the energy requirement distributions of different requests and how to integrate green energy to the system. They state that the variable nature of green energy can be a problem, but do not propose solutions. Gmach et. al. [59] use wind and solar energy to cap the power usage of a data center environment. The paper addresses the problem of variability of green energy and overcomes this problem by adding extra energy storage. Krioukov et al. [75] use renewable energy for execution of MapReduce type jobs. They schedule MapReduce tasks with available green energy, but terminate them when the scheduler realizes that there is not enough green energy in subsequent intervals.

Our work, in contrast, uses prediction methods to estimate the amount of green energy in a given interval and utilizes that data to make decisions about scheduling policies of individual servers. We aim to increase the green energy usage efficiency by prediction as well as reduce the destructive impact of the variable nature of the green energy sources on batch job completion times. Additionally, unlike previous work, we include service jobs and batch jobs together in our model to obtain a more realistic system view, as data centers normally see both types of workloads.

2.2 Solar and Wind Energy Prediction

The focus of current work on large-scale green energy prediction is on medium to long-range time horizons lasting from hours to days. As such, the techniques are highly complex, requiring intensive data acquisition and analysis from using SCADA units [67] for solar energy to entire data centers [57] for NWP wind prediction models. Our prediction interval needs to be only as long as the workloads we desire to schedule, which is on the order of tens of minutes (our predictor uses 30 min). We chose this interval based on run-time experiments on the scalable, fault-tolerant Hadoop framework [65], which we use as our batch workload. Furthermore, as the response time constraints of services that run in data centers can be quite short (tens of ms), our job scheduler and predictor need to be fast. As a result, we designed solar and wind energy prediction models of lower

complexity and shorter time horizons.

2.2.1 Solar Prediction Methodology

We applied various time-series prediction algorithms described in the related work to the output data retrieved from a solar farm at the University of California, San Diego [3]. While most solar prediction algorithms are accurate when weather conditions are stable, exponential weighted moving average (EWMA) algorithms have 32.6% mean error in variable weather. We instead re-purpose the weather conditioned moving average (WCMA) algorithm [100], which was originally designed for wireless sensor networks to larger solar installations. WCMA takes into account the actual values from previous days and the current day's previous measurement samples. It averages the values for the predicted slot from previous days and scales it with a factor, which represents the correlation of the current day against the previous days. The details of this method can be found in [11]. It performs very well, with a mean error of 9.6% for 30 min prediction window even in artificially-created worst-case scenarios.

2.2.2 Wind Prediction Methodology

We develop a novel, low-overhead predictor that utilizes readily available data that has been shown to strongly correlate with wind energy prediction [76] wind speed and wind direction. Our algorithm produces weighted nearest-neighbor (NN) tables to generate wind power curves using available wind speed and direction data at each 30-minute interval. Weighted tables allow the algorithm to adapt to seasonal changes by weighting recent results highly, while the power curves offer flexibility, allowing the algorithm to be used with different wind farms. More details of this prediction method can be found in [20].

The algorithm has been tested against a wind farm installation over a year's worth of power output data provided by the NREL, and the meteorological data provided by the National Climatic Data Center (NCDC). The results show a mean error of 17.2% for a 30-minute prediction interval, equaling or outperforming the

time-series models described in [76], at much lower computational cost.

2.3 Green Energy Scheduling and Data Center Modeling

Our goal in this work is to evaluate the benefit of green energy prediction for increasing the data center job throughput while not sacrificing service jobs' response time constraints. To accomplish this we designed both predictive and instantaneous green energy based schedulers and compare them to the baseline of using only brown energy. The scheduler uses two separate job arrival queues as shown in Figure 2.1. One queue is for web services that have response time requirements (e.g. 90th percentile should be less than 150ms), and the other for batch jobs which are more concerned about throughput and job completion time. When a web services client request arrives, the controller allocates a server that has the smallest number of batch jobs running on it in order to reduce the interference effects between these two types of workloads. Additionally, we put a limit to the number of clients a host can serve to distribute the web-requests evenly among servers. This limit is determined by using current number of clients and total number of host machines. For simplicity, we assume that each server has at minimum one web services request queue, and one or more batch jobs slots to execute. Web services start execution whenever there are available computing resources (CPU and memory) to ensure their response time requirements are met whenever possible. Therefore, we guarantee that the system provides enough brown energy to maintain these service requests. In this work we use Rubis as representative of web services [112]. Based on our measurements and [47] we model the inter-arrival time of Rubis requests generated by a client using a log-normal distribution.

We use open source version of MapReduce, Hadoop [65], to represent batch jobs. Input data of any given job is split and processed by many map/reduce tasks distributed across a fixed number of map/reduce slots in a cluster as shown in Figure 2.1. If there are more tasks than the available slots, the tasks without slots are queued. If any task fails, the MapReduce scheduler starts a fresh copy the task.

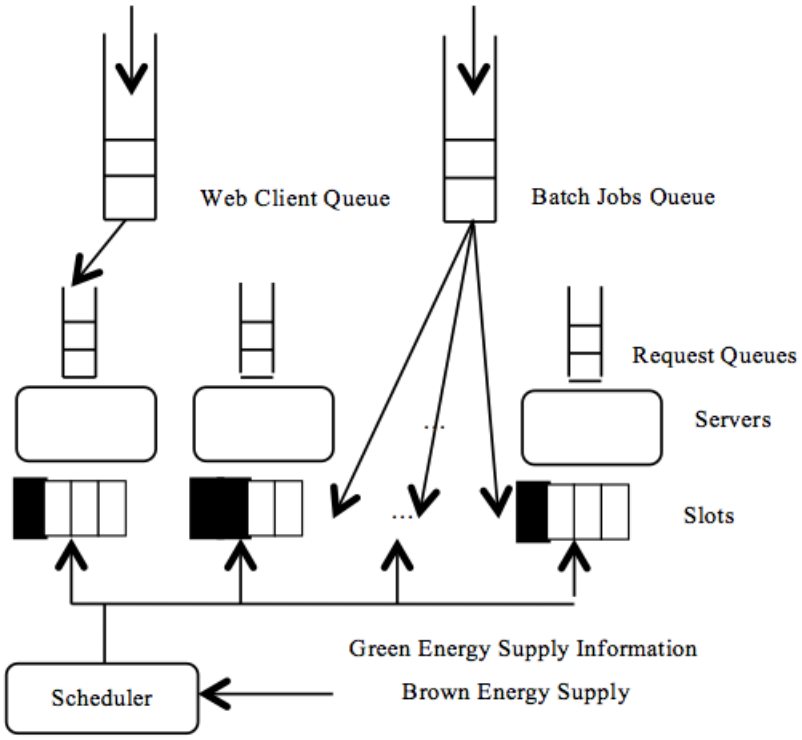


Figure 2.1: System Architecture

The arrival process of this type of jobs is modeled by a lognormal distribution, as demonstrated in [71]. The total number of servers given to a job depends on the energy availability and green energy scheduling algorithms. At each time instance, power consumption of servers is estimated using a linear model based on CPU utilization as in [46]. The overall data center energy cost is calculated using aggregate server power scaled by the power utilization efficiency ratio (PUE) to account for the impact of other sources of inefficiencies (e.g. cooling costs). We use our data center test bed measured average PUE value of 1.26.

Predictive green energy scheduler: Our green energy predictor uses a 30-min prediction interval, a duration that is longer than that of our run-time tests of MapReduce jobs to ensure enough energy is available to finish the tasks. The predictor provides the scheduler with an estimate of the next period's average green energy availability at the beginning of each batch job allocation interval. It then computes the number of batch job slots that can be used for the given amount

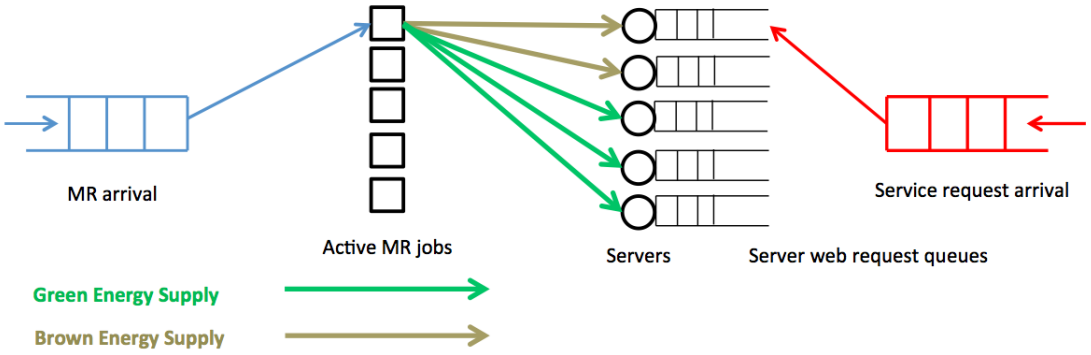


Figure 2.2: Additional batch jobs running with predicted green energy

of energy in that interval. When computing the number of extra slots the scheduler uses the average power/slot information we got from our measurements (see next subsection). If this number is greater than the current number of available slots, the remaining extra slots are distributed to the active MapReduce cluster, so that they can run more tasks in parallel. This process is shown in Figure 2.2. However, if this number is smaller, then the scheduler deallocates some jobs. Jobs that run more concurrent tasks than their base requirement have their slots reduced first. The tasks running in deallocated slots are either immediately terminated or restarted with green energy later on (jobs using more than needed slots), or continue but use brown energy instead. This decision is made depending on the number of concurrent tasks in a job. The energy consumed to run the terminated jobs in the previous interval is wasted. In the results section, we quantify this cost of incorrect energy prediction by using the green energy usage efficiency metric. The main benefit of a predictor is that the number of deallocated slots for batch jobs can be dramatically reduced, and the number of available slots increased.

Instantaneous green energy scheduler: We compare the impact of green energy prediction to the instantaneous use of green energy presented in [75]. To simplify evaluation we use the same algorithm as predictive scheduler, but with a 1min scheduling interval which reflects the instantaneous case well.

Table 2.1: Measured interference of MapReduce and Rubis

# cores MapReduce	1-4	5	6	7	8
Rubis QoS	0.047	0.08	0.1	0.4	0.93
MapReduce Performance	100%	94%	88%	83%	81%

2.3.1 Model Validation Using Experimental Testbed

We developed a discrete event-based simulation platform for scheduling a mix of service and batch jobs in a data center consisting of hundreds of servers. This enables us to evaluate the impact of using a combination of brown and green energy at scale. To ensure accuracy of our estimates, the parameters for our event-based simulator are obtained from measurements on Intel Nehalem [68] servers when running a mix of service (Rubis [112]) and batch workloads (MapReduce [121]) within Xen VMs. Rubis and MapReduce are run in separate VMs, with MapReduce run across 2 VMs, one utilizing 4 cores, and the other varying the number of cores occupied. Rubis is run with 9000 concurrent users.

Table 2.1 shows the measurements we obtained by scheduling an increasing number of MapReduce tasks on the same machine with service requests. We report a measure of normalized response time as Quality of Service (QoS) ratio, which is calculated using 90th percentile response time over the expected response time (for Rubis it is 150ms). We see that even in the worst case, where we allocate the maximum number of available cores to MapReduce jobs, normalized response time of Rubis, as measured by QoS ratio, is still less than 1. In addition, we see that the worst case performance impact on normalized MapReduce job completion times is maximum 20%. Mean measured service time of a single map or reduce task is around 10 minutes, though the maximum can be as high as 20 min, thus justifying our choice of 30min green energy prediction interval.

Given the measurements presented above, in our simulations we use 150ms as the target Rubis response time with 12ms service times for 1000 to 5000 clients representing different times in a day, 2min mean arrival time of MapReduce jobs [71] with average execution time of 10 min. To ensure that in our simulations we

Table 2.2: Verification of simulation outputs

	Measured	**Simulated**	**Error**
Avg. Power Consumption	246W	251W	3%
Rubis QoS Ratio	0.08	0.085	6%
Avg. MapReduce Comp. Time	112 min	121 min	8%

have at most 10% performance impact on MapReduce tasks, we use 5 slots per server. We compare simulation results using this setup to actual measurements on the Nehalem server. Table 2.2 shows that the average error is well below 10% for all quantities of interest, with power estimates having only 3% average error, while performance for services has only 6% and MapReduce completion times are within 8%.

2.4 Results

We use our discrete event-based simulation platform to schedule a mix of service (Rubis) and batch jobs (MapReduce) in typical data center container consisting of 200 Intel Nehalem servers. The overall duration of simulation is 4.5 days. Simulations are repeated until we obtain a statistically stable average.

Each server has a single web service queue that servers multiple clients. Incoming client requests are distributed over the servers evenly. The client arrival distribution is assumed to be exponential as in [89], while client requests are generated using a log-normal distribution with mean 100 ms and 15 ms as mean service time. MapReduce jobs arrive to the system with a mean of 2 min and each task has 10 min execution time on average. We use 5 MapReduce slots per host. Services QoS ratio in all of our simulations remains between 0.09 and 0.2, thus ensuring that web request response time requirements are never violated. The ratio gets closer to 1 when the number of web services clients exceeds 10000. The average queue length for web requests is 0.8 for 1000 clients and 5.5 for 5000 clients.

We use a number of metrics reported in Table 2.3 to compare our predictive scheduler (*Pred.*) with the state of the art instantaneous green energy usage (*Inst.*)

Table 2.3: Comparison of instantaneous and predicted green energy with different alternative energy sources

		GE Efficiency	***GE Job Ratio***	***% Incomplete Jobs***
Wind Energy	*Inst.*	30% ± 2.5%	35% ± 5%	10% ± 3.3%
	Pred.	90% ± 2%	50% ± 5%	1.3% ± 0.4%
Solar Energy	*Inst.*	60% ± 5%	28% ± 4%	8.6% ± 2.5%
	Pred.	93% ± 2%	45% ± 3%	2.4% ± 0.5%
Combined	*Inst.*	72% ± 5%	40% ± 4%	12% ± 2.5%
	Pred.	93% ± 3%	55% ± 5%	3% ± 0.5%

[75] when using only wind, only solar and combined two green energy sources. We define *GE Efficiency* as the ratio of the green energy doing useful work versus the total green energy available: $GE_{useful_work}/GE_{tot}$. Energy consumed by a task that is terminated before completion is not counted as a part of GE_{useful_work}. Green energy under-prediction is penalized by this metric. The percentage of jobs that are terminated as a result of the lack of green energy at the beginning of the scheduling interval, % incomplete jobs, is calculated relative to the overall number of jobs completed using green energy. This occurs when jobs launched with currently available green energy in a previous scheduling interval cannot be sustained due to the energy availability drop in the subsequent interval. Lastly, the efficiency of the system is in terms of green energy usage, *GE Job Ratio*, is defined as the total amount of work done with green energy, $Jobs_{GE}$, over the total work done in the system, $Jobs_{tot}$: $Jobs_{GE}/Jobs_{tot}$.

Table 2.3 shows that prediction improves green energy efficiency up to 3x relative to instantaneous energy. The main reason for this result is that the system has good quality information on green energy availability for a longer interval and hence can make better scheduling decisions. Therefore, less green energy is wasted and 5x fewer MapReduce tasks need to be terminated. Finally, our predictive scheduler increases the number of MapReduce tasks executed with green energy by 2x relative to the instantaneous approach as a result of more accurate energy provisioning. Figure 2.3 shows how the average completion time of MapReduce

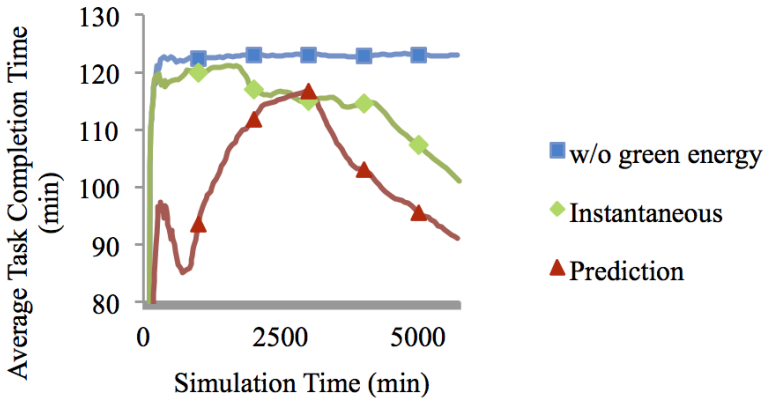


Figure 2.3: Average completion time of MapReduce jobs

Table 2.4: Brown Energy for Inst. vs. Pred. Energy

Total BE w/o GE	**Add. BE for Inst.**	**Add. BE for Pred.**
240 kWh	4.6 kWh	0.64 kWh

jobs changes over time as a function of the way green energy is used. The baseline case uses only brown energy to run a mix of services with just enough MapReduce jobs so that services response time constraints and performance requirements of MapReduce jobs (maximum 10% hit to completion times) are met. In this scenario, we create the MapReduce jobs at the same rate to highlight the green energy effect more clearly. Our green energy prediction scheduler decreases MapReduce task completion times on average by 20%. In contrast, instantaneous usage of green energy results in 12% higher average batch task completion times compared to prediction.

An alternate way to compare using predicted vs. instantaneous green energy schedulers is to supplement with brown energy whenever there is not enough green energy to complete batch jobs. In this way we ensure that all service jobs meet their response time requirements and all batch jobs complete, so none are terminated. The first column of Table 2.4 shows the amount of brown energy needed to run all the tasks in the absence of green energy. When we use green energy instantaneously and do not terminate any tasks when there is not enough green energy available,

we need extra 4.6 kWh of brown energy per data center container, but if we use our predictor, the extra brown energy needed is decreased by more than 7x to 0.64 kWh.

2.5 Conclusion

As the cost of brown energy is becoming a critical bottleneck in data center environments, the need for alternative energy sources is growing. In this section, we present a novel green energy predictor, along with a data center scheduling policy which uses prediction information to obtain better performance for batch jobs without significantly affecting the performance of latency sensitive web requests. We use a simulation platform to compare our predictive policy with instantaneous use of green energy. Our simulation platform has been verified by measurements on real systems, with maximum 8% error across all relevant metrics. Our results show that prediction leads to 3x better green energy usage and reduces the number of terminated tasks up to 7.7x compared to instantaneous green energy usage. The response time requirements of web requests stay well below the 90th%ile during all the experiments.

Chapter 2 contains material from "Using datacenter simulation to evaluate green energy integration", by Baris Aksanli, Jagannathan Venkatesh and Tajana Simunic Rosing, which appears in IEEE Computer 45, September 2012 [19]. The dissertation author was the primary investigator and author of this paper.

Chapter 2 contains material from "Utilizing Green Energy Prediction to Schedule Mixed Batch and Service Jobs in Data Centers", by Baris Aksanli, Jagannathan Venkatesh, Liuyi Zhang and Tajana Simunic Rosing, which appears in ACM SIGOPS Operating Systems Review 45, no. 3, 2012 [20]. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Renewable Energy in Wide Area Networks

The number of online services. such as search, social networks, online gaming and video streaming, has exploded. Due to data locality issues and the demand for fast response times, such services are usually distributed across geographically diverse set of data centers. This is clearly already the case for larger companies, such as Google and Facebook, but is also increasingly true of smaller companies who can leverage cloud offerings from companies such as Amazon [21]. This trend is also fueled by a dramatic increase in the usage of virtualization technology. For example, Amazon's EC2 allows load balancing between virtual machine instances [21].

Internet services usually have frontline service jobs and a background set of batch jobs that prepare data for the online services. For example, in order for eBay to be able to guarantee very low response times to their customer's requests, they need to have an updated and well indexed database of items, usually obtained by running batch jobs. Often, there are two classes of performance metrics used services response times, usually measured in 10s to 100s of milliseconds, and batch job throughput. Normally the service providers' goal is to ensure that service times are within specified bounds, while batch jobs are expected to progress at a reasonable rate.

In addition to performance, a key challenge in such distributed data centers

is the energy cost which includes the cost of computing and data transmission. A previous study [58] shows that as of 2007 at least 2% of the total carbon emission of the world comes from IT. World-wide power consumption due to IT has been growing, with more than 80% due to the way equipment is used [101][128]. The telecommunication infrastructure takes up to 40% of the total IT energy cost, and is expected to continue growing as demand for distributed data processing continues to rise [128].

One of the key trade-offs in the design of distributed services is how data center operators and network providers deliver the needed performance at minimum energy cost. While quite a bit of work has focused on energy optimization of data center computing, relatively little has been done for geographically distributed networks connecting the data centers. The overall electricity cost ("brown energy") of networking can be very high. For example, Telecom Italia is the second largest consumer of electricity in Italy [99]. One way to reduce these costs is to leverage green energy sources such as solar and wind. Intermittent green energy has been explored as a way to perform additional work in data centers [75] and to cap the peak power of a data center [59], but has not been leveraged to offset the cost of backbone networking.

An alternate way is to redesign network elements so that they consume less power. For simplicity purposes we model the total energy cost of backbone networking as a function of power consumption of routers and links. Typically the power cost of the links is a function of distance due to the need for signal amplification, while router power cost is largely fixed at the peak level as the primary objective of router design has been maximizing performance at all cost. As a result, routers dominate the backbone network's energy consumption [123]. Recently there have been a few publications studying how routers could be redesigned to be more energy proportional [31][51]. As the utilization levels of backbone networks tend to be low, around 30% [51], redesigning routers to be energy proportional and then enabling network to leverage this is important. Furthermore, routers are the primary network elements that ensure high speed connectivity between distributed data centers. Currently routes are typically determined statically by

using shortest-path algorithm. However, as routers become more energy proportional, and as their supply is complemented by using highly variable green energy, there will be a need for dynamic route adjustment depending on the current state of the load on particular connections, the performance needs of applications running in the data centers, and green energy availability.

In this chapter, we analyze the use of wide area network in a multiple data center system. The main goal is to improve the energy efficiency of the networking infrastructure, while ensuring service times and batch job throughput constraints are met for large scale distributed data center deployments. We also show that with increasing network speed, online job migration across multiple data centers becomes more feasible, increasing the total throughput. The main contributions of this chapter can be summarized as follows:

- We quantify the energy cost of a data transfer over the backbone network.
- We show that energy proportionality and green energy can make a dramatic difference to network energy efficiency.
- We design a novel green energy aware routing algorithm capable of ensuring quality of service needs are met while improving energy efficiency by 10x.
- We analyze the feasibility of online job migration with varying data center and backbone network properties, and show that the backbone network needs to have higher speed links than the conventional 10 Gbps ones.

3.1 Related Work

A number of projects have explored the idea of wide area job balancing for distributed data centers. A number of strategies have been developed to determine the best strategy for transferring data center jobs to locations where the electricity is cheaper [109][106][30]. This has been aided by the fast live VM migration that is possible with very short downtimes, on the order of a few seconds [122]. Green energy usage in data center systems is a very recent topic [82][83][59][58][75][79].

Work presented recently explored how to effectively leverage green energy availability to complement brown energy supply for data centers [82][83][75]. Green energy has been used to cap the peak power in the system [59]. Our work in the previous chapter studies green energy prediction as an effective way to dramatically increase the effective renewables utilization for data center operations in [20]. However, none of the projects that have looked at job balancing in distributed data centers consider the energy cost of the backbone networks while transport energy consumption can be significant for distributed cloud-based data center applications [26].

On aggregate, network service providers consume a lot of electricity, with Telecom Italia and British Telecom taking around 1% of nation's electricity [32]. This comes at a steep cost, with electricity costs reaching up to 50% of operating expenses for some providers. There has been quite a bit of research on energy efficient backbone networks. The first category includes shutting down idle network elements [51] and provisioning the network to identify the elements that can be shut down without affecting the connectivity [39][123][130]. Another way to increase network energy efficiency is to leverage the fact that line cards consume a large portion of the router power and by adjusting the number of active line cards the power consumption can be decreased significantly [36]. Additionally, dynamic software solutions such as energy aware routing [98][22] to select the energy efficient path and bandwidth adjusting to reduce the router power consumption [70] are used to improve network energy efficiency. Recent projects, like the GreenStar network, propose to experiment with using green energy to power zero-carbon data centers and migrating workloads over the network based on presence of renewable energy [95]. Another work uses brown and green energy together in a problem formulation to minimize carbon emissions [110]. However, energy aware policy they deploy and the green energy supply do not change and adapt over time.

In contrast to the related work, we focus on increasing the energy efficiency of the backbone network without shutting down any connected data centers, network devices or links connecting them. We showcase the effects of theoretical and practical proportionality in network elements on energy efficiency. We use dynamic

prediction of green energy availability to improve the reliability and decrease the carbon footprint of the network. In addition, we show that the design of dynamic routing policies leveraging green metrics effectively utilizes energy-efficiency of the routers and decreases the brown energy use significantly.

3.2 Data Center and Network Modeling

An effective strategy for managing backbone network energy costs, while at the same time ensuring that data center jobs meet their performance constraints, requires careful modeling of not only the network links, but also of the data centers and the servers within them. In this section, we present the models we use to represent data centers and the network elements. For the data center validation we use the methods illustrated in section 2.3.1, while for validating backbone network energy costs we leverage models of energy consumption of state-of-the-art backbone routers [123].

3.2.1 Data Center Model

Each data center container is modeled after the one we have on campus. It has 200 Intels Nehalem servers running Xen VM. The model of a single data center is the same as in section 2.3, where we run Rubis on our machines to model service-sensitive eBay-like workload [112] with 90th%ile of response times at 150ms and multiple MapReduce instances are run as batch jobs. A single MapReduce job consists of a number of tasks that are dispatched in parallel. The job is complete when all tasks finish. Although we have two types of jobs in a data center, we transfer only batch jobs between geographically distributed data centers, as service request sensitive tasks have very tight timing constraints, and often relay on fast local connections to ensure those constraints are met. We assume that data is replicated among the data centers automatically in order to ensure better reliability [56]. Thus, when a batch job is moved, relatively little data has to be moved with it.

Each data center scheduler uses two separate job arrival queues: web ser-

Table 3.1: Inter-arrival and service time parameters

Lognormal Distribution Parameters	α	μ
Rubis 1600 Clients Inter-arrival time (ms)	1.23	0.59
Rubis 3200 Clients Inter-arrival time (ms)	1.12	0.43
MapReduce Job Service time (sec)	1.44	5.24

vices (Rubis) and batch jobs (MapReduce). Service and batch job inter-arrival times are modeled using lognormal distribution based on our measurements of Rubis and MapReduce running on Nehalem servers and results of analysis presented in [71] and [47] respectively (see Table 3.1 for parameters). For simplicity, we assume that each server has at minimum one web services request queue, and one or more batch jobs slots to execute. Web services start execution whenever there are available computing resources (CPU and memory) to ensure their response time requirements are met whenever possible. Load balancing strategy described in [20] is used to distribute requests within data centers, as shown in section 2.3. Although data centers have the same number and type of servers, the request arrival rates are different for each of them representing varying demands based on location and the time of day. We leverage these differences for geographically distributed load balancing.

For simplicity we have a single controller that monitors and manages load of the data centers and the network. Each data center sends the available resource (CPU, memory etc.) profile to the controller every 30 min as MapReduce jobs typically take less than 30min to complete. Based on this information, the controller computes the average resource usage of the overall set of data centers. Then, starting from the center with least amount of extra resources, it balances the resources across the system. This process continues until the amount of available resources in each data center is more balanced under the constraint of available network bandwidth or a data center cannot find a task to transfer. The actual transfer of batch jobs is initiated by the controller once the re-balancing analysis completes. Data centers provide lists of candidate jobs, while the network computes the path and the available bandwidth of the path, depending on the routing policies used. Then

Table 3.2: Parameters and values used in the simulation

Parameter	Value	Parameter	Value
Mean Web Request Inter-arrival time	5ms	Number of data centers	5
Mean Web Request Service time	20ms	Number of routers	12
Service Request SLA	150ms	Idle Server Power	212.5W
Mean MapReduce Job Inter-arrival time	2min	Peak Server Power	312.5W
Mean MapReduce Task Service time	4 min	Idle Router Power	1381W
Average # tasks per MapReduce job	70	Peak Router Power	1781W
Average required throughput level per MR job	0.35	Number of line cards	10
Number of servers in a data center	200		

the controller computes the traffic matrix between data centers in terms of size of data (a function of the number of VMs) and initiates the transfers accordingly.

We compare the simulation results with a real experimental setup running a mix of Rubis and MapReduce workloads on a set of Intel Nehalem servers from our data center container, as shown in section 2.3.1. We also present the parameter values we use in our simulation in Tables 3.1 and 3.2.

3.2.2 Backbone Network

Our model is based on typical telecom network characteristics [132] consisting of routers, hubs, storage and computation elements, complemented with infrastructure PDUs, UPS, and air conditioners to keep them operational. Given that the large fraction of the overall network energy cost is due to routers, we

specifically focus on this aspect. In addition, routers maybe designed to be more energy proportional going forward, while optical links have a fixed energy cost that is a function of the distance between amplifiers [51]. Thus, in our analysis we neglect the link cost, as it is just a fixed offset to the overall energy consumed. The power consumption of the router can be estimated using a linear model [84] with bandwidth utilization ratio $0 < u < 1$, idle power, P_{idle} and peak power, P_{peak} as follows:

$$P = P_{idle} + u(P_{peak} - P_{idle}) \tag{3.1}$$

In current routers, P_{idle} is high, thus the energy consumption is not at all proportional to network load limiting the potential savings. However, there have already been a number of proposals on how routers can be made more energy proportional [31][51], ranging from turning off line cards that are not being used, to more complex circuits and system solutions. Figure 3.1 shows the power models of routers we use in our simulations. The non-proportional model represents measurements of an actual state-of-the-art router [123] that is capable of supporting four 100Gbps links concurrently. Its peak and idle power value are listed in Table 3.2. The step function proportional is the power curve we measured by removing line cards from the same router similar to on/off approach presented in [84]. Smooth proportionality model assumes techniques have been developed to "smooth out" the step proportional curve, while the ideal proportionality represents the best case linear proportionality.

In our simulations we model a subset of LBNL ESnets network topology as shown in Figure 3.2 [48]. We use 5 endpoints where data center containers reside (represented by squares) with 12 intermediate routers connected with all relevant connections (circles on Figure 3.2). Upon request for a larger backbone data transfer, the network identifies a path to carry the data between two endpoints of a transfer. State-of-the-art systems determine and configure that path statically by using shortest path algorithm.

ESnet dramatically improved on the state-of-the-art routing and bandwidth allocation algorithms by developing On-Demand Secure and Advance Reservation

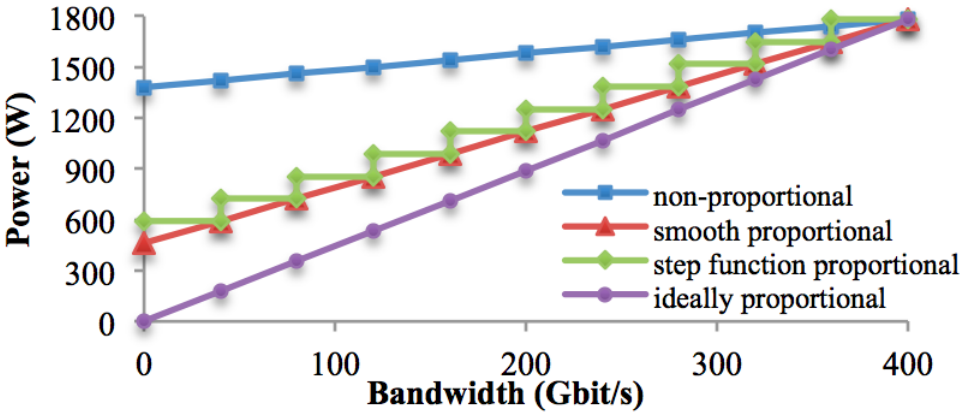


Figure 3.1: Power curves for different network power schemes

Figure 3.2: Network Topology; squares = data centers, circles = routers

System (OSCARS) [64]. OSCARS enables users to reserve dynamic virtual circuits (VC) by computing path online to construct VCs with required bandwidth. This solution works well in situations where the only goal is performance. However, the energy consumption is becoming another key constraint. As a result, an energy-aware dynamic routing algorithm is needed to identify and adjust the path during the transfer so both performance and energy constraints can be met in the most effective way.

To reduce the router's brown energy consumption, we assume that each routing site has its own green energy source, solar and/or wind. Table 3.3 lists the

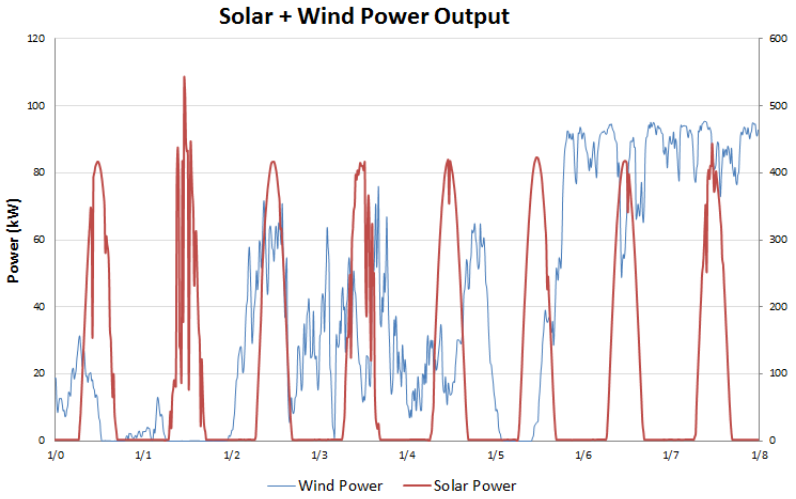


Figure 3.3: Solar and wind energy availability

types of renewable energy available in different locations. The renewable energy data, including location, and amounts generated over time has been provided by NREL [5][6]. We use a weather-conditioned moving average (WCMA) for solar and weighted nearest-neighbor (NN) table based algorithm for wind energy availability prediction over 30min intervals [20]. We also assume that green energy supply systems provide on average 80% of the energy need per router, 1.6 kW, where available.

In Figure 3.3, we show a subset of the green energy availability measurements. Solar data is gathered from the UCSD Microgrid and wind data is obtained from a wind farm in Lake Benton, MN, made available by the National Renewable Energy Laboratory. The representative outputs for the other various locations in our experiments (San Francisco, Chicago, etc.) are obtained by scaling and time-shifting the measured results from our available sources to published average availability data for the target areas [5][6].

To best leverage green energy availability, we design a novel green energy aware routing (GEAR) algorithm and compare it to shortest path routing (SPR) which is based on Dijkstra's algorithm [64]. GEAR selects the path capable of reserving the required bandwidth while ensuring it also has the lowest brown energy consumption. The algorithm in Figure 3.4 provides the overview of GEAR. GEAR

Location	Type	Location	Type	Location	Type
Chicago	Wind	New York	Wind	San Francisco	Solar+Wind
Atlanta	Solar	San Diego	Solar	Denver	-
Kansas	-	El Paso	Solar	Houston	Solar
Nashville	Wind	Cleveland	Wind	Washington DC	-

Table 3.3: Renewable energy availability in different locations

Algorithm I. GEAR

Inputs: Source, s; Destination, d; Paths, P; Required bandwidth, rb

Output: Path with lowest brown energy consumption

```
1.  n ← Number of paths between s and d
2.  be[1:n] ← Inf
3.  For i: 1 to n
4.      b ← bandwidth of P[i]
5.      If b >= rb
6.          be[i] ← 0
7.          For each router on P[i]
8.              be[i] += energy need of P[i] – green energy estimate P[i]
9.  index ← argmin be
10. Return P[index]
```

Figure 3.4: Green energy aware routing algorithm

analyzes brown energy need of each path with required bandwidth between a pair and selects the one with least brown energy need. The paths are pre-computed to avoid re-computation. We leverage the dynamic circuit construction capability of OSCARS to not only compute paths that best leverage green energy availability, but to also dynamically allocate those paths.

In addition to green energy aware routing, step proportional router design can be best leveraged by a new routing policy as well. In this case the additional bandwidth utilization might not always increase the power consumption of a router due to fairly coarse set of steps as shown in Figure 3.1. The network controller calculates how much extra power a path between two points would need and selects the path with the least extra power required. The algorithm is similar to GEAR, except that green energy usage in line 8 is set to zero.

3.2.3 Simulation of Backbone Network with Data Centers

We use a discrete event-based simulation platform that models the performance and energy cost of a large scale backbone network connecting geographically distributed data centers [20][19]. The simulator keeps track of each process in every data center. The main controller of the simulator is responsible for synchronizing both the data centers and the network. In our simulation, we set the load balance control interval to 30 min. This duration is appropriate given the typical length of batch jobs, and the fact that individual service requests are much shorter lasting. The load in each data center follows a day/night pattern appropriate for the particular location [39]. Power is estimated using models presented in section 2.1 for the data center, and using Figure 3.1 for power cost of routing. Renewable energy data has been obtained from NREL [5][6]. We do not quantify the power cost of supporting systems such as cooling as our goal is to compare the improvements to energy efficiency of backbone network as a function of changing availability of green energy and novel router designs. This could be easily accounted for by using a PUE factor.

3.3 Results

In the previous sections we describe the models we use for data centers and the backbone network, along with the simulator that we developed to evaluate the benefits of changing the design of routers, and leveraging green energy availability along the routes. The parameters that we use in simulation are shown in Tables 3.1 and 3.2. Each VM has a single job in it that is either service or MapReduce and is allocated 8GB, which is reasonable for current systems [21]. Predictor accuracy is 83% for wind and 90% for solar within the 30min rebalancing interval used by the overall system controller. Network is assumed to have 10% BW reserved for background data transfers in all our simulations, to account for the transfers other than data center load balancing. In all cases, except where otherwise noted, we assume 100Gbps backbone network links. The power profiles for various energy proportional router designs are given in Figure 3.1. We simulate four days. For our

Table 3.4: Metrics and their definitions

Metric	Definition
Network Related Metrics	
BW_{ave}	Average bandwidth per link in Gbps
$TotP_{ave}$	Average power consumption per router
$TotP_{max}$	Maximum power consumption per router
$BrownP_{ave}$	Average router "brown" power consumption
Energy Efficiency Metrics	
BrownE	% Brown energy used per router relative to total energy
BW_{ave}/BE	Ave. bandwidth util. efficiency per brown energy spent
$NetE_{eff}$	# MapReduce jobs completed per brown energy spent

analysis we define multiple metrics as shown in Table 3.4. In addition to traditional metrics, such as average bandwidth used and router power consumption, we also define two energy efficiency metrics. The first quantifies the increase in the number of batch jobs finished as a function of brown energy used, and the second evaluates how well bandwidth is utilized per brown watt consumed.

We first evaluate the job performance without distributed load balancing. In this case the batch job completion time is 22.8 min while service response time constraints are met. Next we analyze the benefits of leveraging the various types of network configurations for transferring jobs, ranging from baseline design that replicates the state of the art, to having a network populated with energy proportional routers that have green energy supply sources as well. The cases where there is no green energy use shortest path routing (SPR), while when green energy is available we compare SPR with our GEAR algorithm. We next provide the analysis of all these results.

Non-Proportional Routers: Data centers transfer batch job VMs to a remote center in order to reduce the computational burden and obtain higher performance for the waiting jobs. When transferring data, we use two different bandwidth allocation policies. The first one, *all-bandwidth policy*, allocates all the

Table 3.5: Baseline results: all bandwidth (AB), necessary bandwidth (NB)

Metric	AB	NB	Metric	AB	NB
Ave. MR job completion (min)	17.5	16.8	$TotP_{ave}$	85%	83%
Ave. MR task completion (min)	4.22	4.25	B_{ave}	66	48

available bandwidth of the links whenever a path is constructed. The second one, *necessary-bandwidth policy*, allocates just enough bandwidth to the path, so that the transfer time for data takes at most 100 sec through a 100Gbps path. The first policy yields faster data transfer, however it also saturates links. The second results in more network availability. Table 3.5 summarizes results for both policies. Using network to adaptively distribute batch jobs improves the job completion times by 30% while not changing the service's response times. Both AB and NB policies have comparable performance and power consumption as the dynamic power range of baseline network is very small. Bandwidth utilization is 1.5x lower for NB, which may enable additional data to be transferred as needed. As a result, our simulations show that with NB policy 34% of the tasks are executed in a remote center with 5% more tasks transferred than with AB.

In Figure 3.5 we explore the performance of the batch jobs and the average power consumption of a router with different bandwidth values available per link when utilizing the necessary bandwidth policy. Performance and power consumption do not change significantly between 50-100Gbps of the available network bandwidth. MapReduce job completion times approach the case where no load balancing is used as network bandwidth drops down to 10Gbps. This explains why today's load balancing is not done very often as most links are at 10Gbps. By increasing the number of servers by 2x and keeping the server load constant, we get better job performance. However, further increased of number of servers does not result in better performance as there is no need for extra resources with the fixed load rate. Increasing the server load 2x while keeping the server number fixed decreases performance by 7%. Further server load increase creates significant performance drop, 15%, with 10-20 Gbps bandwidth.

Energy Proportional Routers: We use three different proportionality

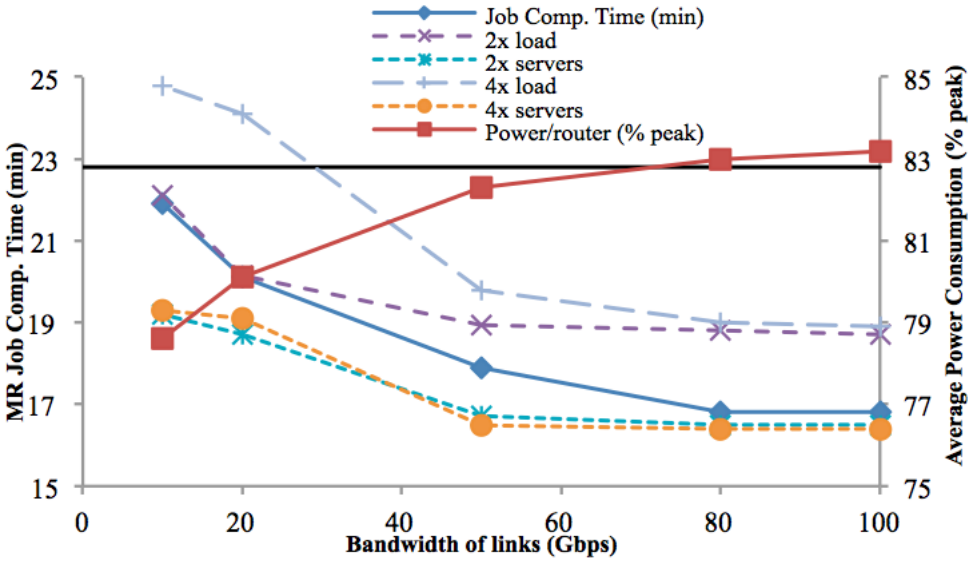


Figure 3.5: MapReduce job completion time and power vs. bandwidth

schemes as shown in Figure 3.1 ideal, smooth and step proportionality. Job completion time service times do not change significantly as compared to the non-proportional network case. Table 3.6 summarizes all results for the next subsections. All power numbers are reported as a percent of router peak power listed in Table 3.2. Bandwidth, BW_{ave}, is in Gbps. Looking at the columns corresponding to situations where no green energy is used, it is clear that NB allocation yields better power consumption for all proportionality schemes. Average power savings are around 70% if there is ideal, 50% for smooth and 35% for step function proportionality compared to non-proportional case. Network energy efficiency, $NetE_{eff}$, improves dramatically - by 3x, while bandwidth energy efficiency, BW_{ave}/BE, increases by at most 4x.

We also use the energy aware routing algorithm for step proportionality (described in Section 3.2.2) with NB policy in our simulation and obtain 48% of peak power per router on average. The dynamic policy results in 6% better power consumption compared to the state-of-the-art shortest path policy, but leads to 3% more transfer delay.

Green Energy & Non-Proportional Routers: For the next two scenarios we supplement the traditional grid (brown) power with green energy and

Table 3.6: Summary of key results using green energy in wide area networks

	w/o Green Energy								**w/ Green Energy**							
	Non-prop.		**Ideal Prop.**		**Smooth Prop.**		**Step Prop.**		**Non-prop.**		**Ideal Prop.**		**Smooth Prop.**		**Step Prop.**	
Policies	*AB*	*NB*	*AB*	*NB*	*AB*	*NB*	*AB*	*NB*	*SPR*	*GEAR*	*SPR*	*GEAR*	*SPR*	*GEAR*	*SPR*	*GEAR*
$TotP_{ave}\%$	85%	83%	33%	24%	51%	44%	61%	54%	83%	86%	24%	28%	45%	48%	53%	57%
$BrownP_{ave}\%$									62%	59%	8%	3%	15%	9%	20%	12%
$BrownE\%$	100%								75%	68%	33%	10%	33%	18%	38%	21%
$NetE_{eff}$	58	59	153	210	97	112	83	95	84	85	628	1675	358	559	251	419
BW_{ave}/BE	0.77	0.57	2	2	1.29	1.11	1.08	1.12	0.77	0.93	6	18	3.2	6.1	2.4	4.58
BW_{ave}	66	48	66	48	66	48	66	48	48	55	48	55	48	55	48	55

evaluate the benefits of green energy along with green-energy aware routing, and new router designs. Our goal is to reduce the brown energy consumption as much as possible by effectively leveraging renewable energy availability. Here we use our green energy aware routing (GEAR) algorithm. When there is a data transfer initiated between two data centers, GEAR chooses the path with the least brown energy needed, which may not be the shortest one. Thus, in Table 3.6 we compare GEAR to the shortest path routing (SPR) for all tests with green energy. The difference between SPR and GEAR routing algorithms when using green energy with non-proportional routers is minimal as non-energy proportional routers have very high idle power.

Green Supply & Energy Proportional Routers: We next combine GEAR with energy proportional router design. We do not implement any changes to GEAR specific to energy proportionality assumption as it chooses the path with smallest brown energy need regardless of the power curve used. The total (green + brown) power consumed by all networking elements with GEAR increases between 0.5- 5% compared to SPR depending on router design. However, GEAR compensates this increase by using more green energy, which results in lower brown energy usage. As a result, GEAR uses 7% less brown energy for non-energy proportional routers and 15% less for smooth proportional routers. The percentage of brown energy consumed when using GEAR, $BrownE$, drops dramatically with energy proportional hardware, dropping down to as low as 3% when ideal proportionality is assumed and as high as 12% with step proportionality.

Furthermore, GEAR has 2x better network energy efficiency, $NetE_{eff}$, and 2.3x better BW_{ave}/BE compared to SPR. Compared to non-proportional router design with no green energy usage, the improvement is 7x for $NetE_{eff}$ and 8x for BW_{ave}/BE with step proportionality, 10x for $NetE_{eff}$ and 11x for BW_{ave}/BE with smooth and 27x for $NetE_{eff}$ and 31x for BW_{ave}/BE with ideal proportionality. These dramatic improvements indicate that even relatively simple redesign of routers along with green energy availability and novel green-energy aware routing algorithm design can results in dramatic reductions in the operating expenses for backbone network operators.

3.4 Conclusion

High bandwidth and energy efficient backbone network design is critical for supporting large scale distributed data centers. In this chapter, we propose novel energy aware routing policies along with different energy proportionality schemes for network hardware. We use a simulation platform to compare our energy aware policies to state-of-the art routing policy with different power curves. Our results show that the network brown energy efficiency improves 10x with smooth proportionality and can be as high as 27x with ideal energy proportionality using energy aware policies.

Chapter 3 contains material from "Benefits of Green Energy and Proportionality in High Speed Wide Area Networks Connecting Data Centers", by Baris Aksanli , Tajana Rosing, and Inder Monga, which appears in Proceedings of Design Automation and Test in Europe (DATE), 2012 [17]. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Energy Efficiency in Networks of Data Centers

Multiple data center systems have been a widespread solution for companies in order to meet the constantly increasing need for computation, as shown in chapter 3. As the number of these huge buildings increases, their energy demand becomes a bigger problem due to elevated cost. Additionally, their energy needs are supplied mainly by non-renewable, or brown energy sources, which are increasingly expensive as a result of availability and the introduction of carbon emissions taxes [87]. We address this problem by efficiently integrating renewable energy into these systems as discussed in chapters 2 and 3. These multiple data center systems can also leverage temporal differences in workloads, energy prices and green energy availability, if applicable, by migrating workloads among each other. This online migration has become a feasible solution due to faster backbone networks, as shown in chapter 3.

This chapter expands the energy efficiency analysis of the previous chapter by focusing on data centers and the wide area networks (WAN) together. It proposes two online job migration (cost minimization and performance maximization) algorithms that use green energy prediction in local renewable energy sites and varying brown energy prices. We use the backbone network model described in chapter 3 to obtain a holistic multiple data center model. We investigate the impact of two aspects of data center operation typically overlooked in previous

studies: tiered power pricing, which penalize the data center for exceeding certain level of power restrictions, and WAN leasing costs/cost models, which leverage energy-aware routing.

4.1 Background and Related Work

Multi-data center networks offer advantages for improving both performance and energy. As each data center is in a different location, its peak hours and energy prices vary. A data center with high electricity prices may need to migrate work to another data center with a lower price, incurring some performance and power cost due to data migration. The live migration of virtual machines over high speed WAN has made this idea feasible, as it offers fast transmission with limited performance hit [122].

The study in [79] explores brown energy capping in data centers, motivated by carbon limits in cities such as Kyoto. The authors leverage distributed Internet services to schedule workloads based on electricity prices or green energy availability. Similarly, [33] optimizes for energy prices, to reduce overall energy consumption by distributing workloads to data centers with the lowest current energy prices. The insight is that renewable sources such as solar energy are actually cheapest during the day, when workloads are at the highest and utility sources are most expensive. Job migration is then modeled as an optimization problem, and the authors identify a local minimum energy cost among the available data centers that still meets deadlines.

Previous publications concerned with energy costs primarily propose a follow the sun cost management strategy [59][75][78][30][109][115] and generally neglect the cost of wide area networking (WAN) incurred by job migration between data centers. This assumption is reasonable for small data center networks that own the WAN and incur low network costs. Consequently, related work has WANs used to increase system performance via load balancing [83][77][108] or improve energy efficiency by migrating jobs [30][109][115]. However, these arguments are not applicable for large WAN costs and data centers that lease the network.For

example, WAN may be leased, with lease costs quantified per increment of data transferred, and thus might be too high to justify frequent migration of jobs between datacenters [106].

Data centers lease the WAN by agreeing to pay a certain price for a fixed bandwidth usage. However, as WAN usage increases, network owners [54] offer Bandwidth-on-Demand services, especially for data center applications [85]. Additionally, the WAN may take up to 40% of the total IT energy cost, and is expected to continue growing as demand for distributed data processing continues to rise [101] and as the server hardware becomes more energy efficient [9]. With the increasing importance of managing energy consumption in the network, WAN providers can charge users not just on the amount of bandwidth they use, but also the time of day when they use it. For example, using the network in a peak hour may be more expensive than when it is idle, reflecting electricity market prices [131]. Moreover, with the introduction of carbon taxes, WAN providers can also vary energy prices depending on the energy source. Consequently, data centers might be open to longer, less expensive paths on the network. For example, a data center may request a path that uses green energy to avoid paying extra carbon emission taxes, or a less-utilized path to avoid extra utilization costs. This chapter considers both the costs of geographically distributed data centers and the profits of the network provider. We model different network cost functions, along with the analysis of new technologies that would allow using more energy proportional routers in the future.

Furthermore, data centers often undergo a tiered power pricing scheme. The energy under a specific level may cost a fixed amount and this fixed price changes depending on the location, so it is beneficial to run jobs in a data center at a lower fixed price. Data migration should not increase the total power consumption to more than the amount specified by the specific tier level. Otherwise, extra power costs are calculated using higher prices, generally much higher than the fixed price.

Table 4.1 summarizes and compares the key state of the art contributions for managing distributed data centers in order to minimize an objective function, e.g. the overall cost of energy. Buchbinder et al. [33], Qureshi et al. [106] and Rao

Table 4.1: Summary and comparison of the related work

	Buchbinder 2011 [33]	**Qureshi 2009 [106]**	**Mohsenian-Rad 2010 [91]**	**Rao 2010 [109]**
Goal	Min. electricity bill	Min. electricity bill	Min. carbon footprint & job latency	Min. electricity bill
How	Move jobs where energy is cheaper	Move jobs where energy is cheaper	Migrate jobs to different locations depending on the goal	Move jobs where energy is cheaper
Workload	No specification	No specification	Service requests only	Service requests only
Perf. Constraints	X	X	Latency of service requests	Latency of service requests
Network Cost Model	Fixed cost per bandwidth	Fixed cost per bandwidth	X	X
Routing	X	Distance based routing	X	X
Green Energy	X	X	Local green energy, carbon tax	X
Network Delay	X	✓	✓	X

	Liu 2011 [83]	**Le 2010 [79]**	**Aksanli 2012 [20], [17]**
Goal	Min. brown energy use	Min. the total cost of energy	Max. batch job performance & Min. brown energy use
How	Move jobs to local green energy	Forward jobs to data centers	Move jobs where utilization is low
Workload	No specification	Different job types (not explicitly specified)	Mix of service and batch jobs
Perf. Constraints	X	SLA of service requests	Latency of service requests and throughput of batch jobs
Network Cost Model	X	X	X
Routing	X	X	Static routing vs. energy aware
Green Energy	Local Green Energy	Grid green energy carbon tax	Local green energy with prediction
Network Delay	X	X	✓

et al. [109] relocate jobs to where the energy is cheaper to minimize the energy cost. They do not model different energy types; perform detailed workload performance analysis and different routing options for both WAN providers and data centers. Le et al. [79] solves a similar problem including green energy in their model but they assume a centralized dispatcher and do not analyze network latency or cost. Liu et al. [83] and Mohsenian-Rad et al. [91] minimize the brown energy usage or carbon footprint. They either do not consider the variability of green energy or do not have a network model. In the previous chapter, we solve a load-balancing problem by modeling network properties, but do not consider energy costs. As we can see from this analysis, previous studies do not consider all the important aspects of multiple data center networks simultaneously in their models. This can lead to overestimated cost savings or overlooked performance implications due to not considering both the requirements of different types of applications and WAN characteristics.

In the next subsection, we present two online job migration solutions across data centers. First one reduces the total cost of energy by moving the computation to the locations with lower energy prices or additional renewable energy availability whereas the other one uses renewable energy to improve the overall job performance globally by matching renewable energy generation with computation. Our methods also bring a holistic approach by considering both data centers and WANs simultaneously.

4.2 Cost Minimization and Performance Maximization Algorithms

The goal of the cost minimization algorithm is to determine which workloads we need to transfer among different data centers during each interval to minimize the energy cost. The current algorithm assumes a centralized implementation for control for job migration decisions, though each data center generates its own workloads. We assume that green energy is generated and used locally, and is prioritized over brown energy to minimize the total cost, as green energy is a fixed,

amortized cost. Thus, we transfer workloads to data centers which have available capacity and extra green energy. Because of data centers' energy pricing scheme, energy in a particular location may have a fixed, low cost up to a specified amount of peak power capacity. After this level, energy becomes much more expensive. Therefore, our goals include maintaining utilization in data centers such that we do not increase the power consumption further than the power tier levels.

Figure 4.1 illustrates our cost minimization algorithm. The algorithm performs in discrete time steps of 30 minutes. Each interval begins with the calculation of the amount of energy required by each data center, incorporating the previous and incoming load rates. The former represents the active jobs at a given time, and the latter is determined by the statistical distributions of real applications. Each data center has its own workload distributions that represent different types of applications in a data center environment. The properties of these distributions are determined by applying statistical analysis on real data center traces, outlined in section 2.3.1. We estimate the green energy availability using prediction (section 2.2), obtain the current brown energy pricing, and check power restrictions. Based on the energy need and green energy availability, each data center determines if it has surplus green energy. The key assumption is that if brown energy has already been within the lower price region, it makes sense to use it for running jobs, while green energy can be used to both reduce power consumption and to run extra jobs which otherwise might not be scheduled.

Then workloads are transferred from the data centers with the highest need to those with the highest available green energy. The workload that can be transferred from a data center is determined by what is migrateable, while the workload that can be transferred to a particular data center is limited by the amount of additional green energy and WAN availability. This process continues until every data center is analyzed. If there are workloads remaining in any data centers at the end, the algorithm focuses on data centers with the cheapest brown energy cost. It moves workloads from the data centers with higher energy costs to those with the cheapest brown energy. The amount of data that can be transferred is limited by receiving datacenter's peak power constraints and tiered power levels.

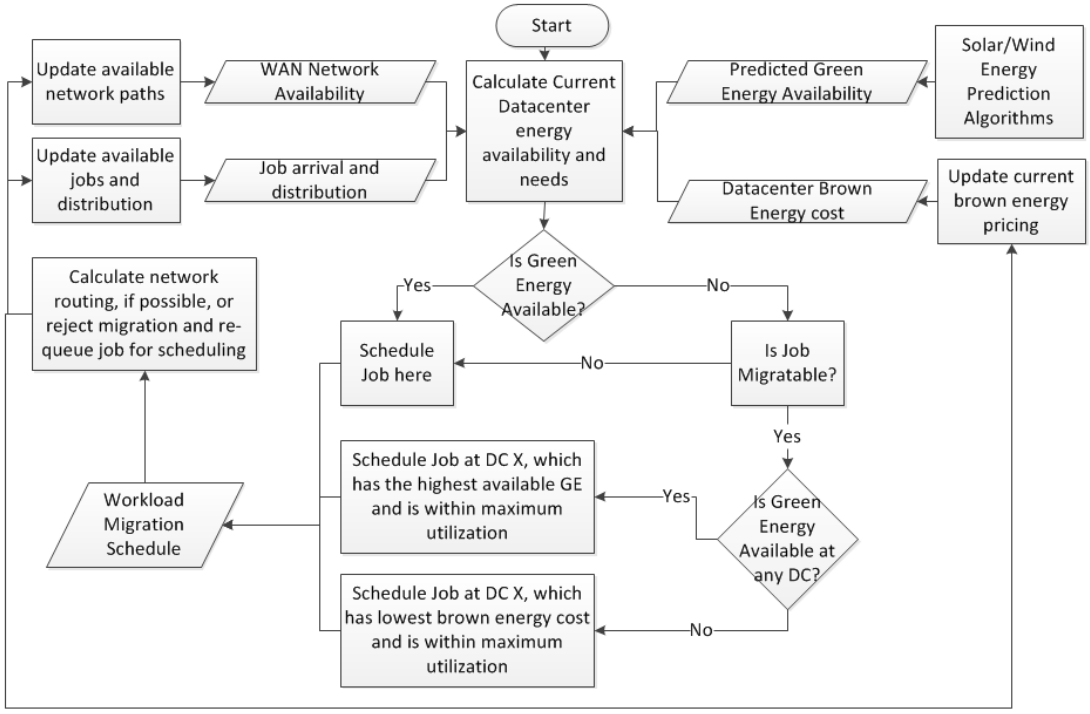


Figure 4.1: Overview of the cost minimization algorithm

If there are still unscheduled jobs remaining at the end of this process, they are scheduled in data centers where the market electricity prices are the lowest.

We can also modify this iterative part of our algorithm to maximize the performance of the workloads instead of minimizing the total cost of energy. In this case, we transfer the jobs that are actively waiting in the execution queue to data centers with excess green energy availability. The iterative process of the cost minimization algorithm is also valid here, but the migration depends only on green energy availability, i.e. jobs are not migrated to data centers with cheaper brown energy prices because extra brown energy would be required for these additional jobs. We denote this process as performance maximization as it runs additional jobs with surplus green energy.

At the end of this iterative process, we obtain a matrix representing workload transfers among data centers. This transfer matrix is then provided to the networking algorithm, which calculates the paths to be used and the amount of bandwidth that needed by each selected path. In our study, we analyze different

path selection algorithms, such as shortest path routing (SPR), green energy aware routing (GEAR), and network lease based routing. A detailed description of SPR and GEAR implementations is in section 3.2.2. Network lease based routing selects the path with the least per-bandwidth price in the case the WAN is leased. In our results, we analyze different network cost functions as well. If a selected path in the transfer matrix is unavailable due to network limitations, the job is rescheduled with a limitation on target data centers.

Our algorithm is similar to those proposed in related studies (section 4.1), but it minimizes the cost of energy more comprehensively. This is because it has a more complete view of data center energy costs, modeling both fixed energy costs under fixed amounts and variable, higher tier energy prices. This helps us to calculate the energy cost savings in a more accurate way. Secondly, it considers the side effects of the WAN, analyzing both the performance implications of different routing algorithms and additional leasing costs if necessary. This is key when multi-data center systems lease the WAN. Job migration may not be feasible for those systems if the cost of leasing the network is too high. Third, the green energy availability information is enhanced by using prediction which can provide information 30-minute ahead and thus help us allocate the workloads across multiple data centers in a more effective manner. Last but not the least; our algorithm is flexible in the sense that it can perform for both cost minimization and performance maximization purposes. It specifically shows that green energy can be used to maximize the performance rather than just minimizing the total cost of energy of geographically distributed multi-datacenter systems.

Note that the data center and WAN models, as well as the green energy prediction algorithms, used by our algorithms are taken from previous chapters: the data center model from sections 2.3 and 3.2.1, the backbone network model from section 3.2.2 and renewable energy prediction algorithms from 2.2.

4.3 Methodology

We use an event-based simulation framework to analyze and compare the results of our algorithms. The inputs to our simulator are derived from measurements performed on our data center container (sections 2.3 and 3.2.1) and data obtained from industrial deployments. This section first discusses how we construct the simulation environment including the data center loads, simulation parameters, green energy availability, and brown energy prices.

Data center load: The load models and mixtures for our experiments are the same as in chapter 3. More specifically, we use the data center model from section 3.2.1, backbone model from section 3.2.2). These models include both workload mixtures and power equations. The parameters of the simulation can be seen in Table 3.2. Note that we only migrate batch jobs due to the tight response time constraints of service jobs.

Green energy availability: The green energy availability, along with prediction, in different locations is the same as in section 3.2.2. Green energy availability in different locations are presented in Table 3.3.

Brown and green energy costs: Data centers contract power from utilities to obtain competitive prices for their expected loads. This can be seen as a tiered pricing scheme. If a data center exceeds the tiered amount in an interval, it is relegated to higher prices, sometimes even market prices. We obtain sample fixed pricing for the mid-west, the east and the west coasts [80]. Since market prices change over time, we use the California ISO [8] wholesale pricing database to obtain brown energy prices for various California locations, and time-shift and scale those values for the other locations based on published averages [2]. Figure 4.2 shows daily pricing values for brown energy in comparison to fixed costs. The straight lines correspond to fix, under-tier prices and the others show samples of variable, market prices which can be used to charge data centers that go over their tiered amounts.

Local green energy costs are typically amortized over the lifetime of an installation, incorporating the capital and the maintenance costs. This is represented by a fixed offset to our cost model. We use data from [90] to obtain the capital

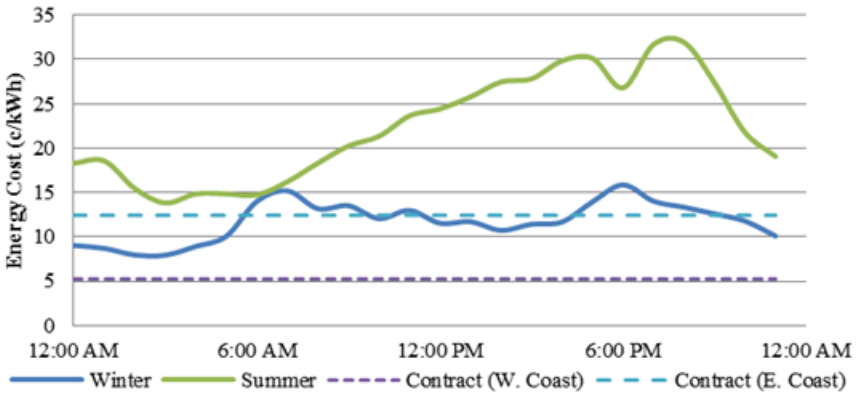


Figure 4.2: Daily brown and amortized green energy cost (¢/kWh)

and operational expenses of several solar and wind farms, amortized over their lifetimes, as representative solar and wind costs per interval.

4.4 Results

This section presents the simulation results for the base case of *no migration*, and the workload migration policies for *performance maximization* and *cost minimization*.

4.4.1 No Migration

In this scenario, each data center runs its own workload using only locally available green energy. This is the baseline for our comparisons, as it represents the nominal brown energy need and quantifies the performance of batch jobs without the overhead of migration. A power tier level accounts for 85% of data center's power needs, while the rest, when needed, is provided at variable market prices. We allow service and batch jobs to run on the same servers while ensuring that they meet quality of service (QoS) requirements (service job $QoS_{ratio} < 1$), and find that the average MapReduce job completion time is 22.8 min. Only 59% of the total green energy supply is consumed by data centers locally, motivating the distributed policies described previously.

4.4.2 Performance Maximization Using Migration

In this algorithm, we leverage migration to complete more batch jobs than previously possible. Data centers with high utilization transfer jobs to locations with low utilization or where there is excess green energy, effectively completing more work in the same amount of time.

Most MapReduce jobs (representative of batch jobs) complete within 30 min [20], which becomes the threshold for both the green energy prediction interval and the interval for checking data center utilization. At each interval, the controller retrieves the resource usage and green energy profiles of each data center and optimizes the system by initiating extra workloads in data centers with green energy availability while still meeting under-tier power constraints. It calculates the available transfer slots between each end-point pair, and selects the tasks to be executed remotely from each data center's active batch jobs. Once the tasks finish execution in a remote data center, the results are sent back to the original center. The key to this policy is that waiting tasks are migrated, as opposed to active tasks, resulting in more jobs executed overall (section 4.2).

Our simulation results show that the average completion time of MapReduce jobs is 16.8 min, 27% faster than the baseline, with no performance hit for service requests. Furthermore, since we are leveraging all available green energy for extra workloads, the percentage of green energy used is 85%, significantly higher than the baseline.

Figure 4.3 reports the total cost normalized against the no migration case with different tier levels specified as a percentage of the data center's peak power capacities and network lease options. Without tiered energy pricing (where all the consumption is charged using market prices), we demonstrate a 25% increase in the total energy cost. However, when we do include tiered energy pricing, we see more accurate results, with a cost increase of only 12% for a 70% level, and a total cost increase of 6% for an 85% level.

Since the WAN may not be owned by a data center, we also analyze the case where the network is leased. In this case, a bandwidth-dependent cost is incurred. Figure 4.3 shows the results of this analysis over different cost functions

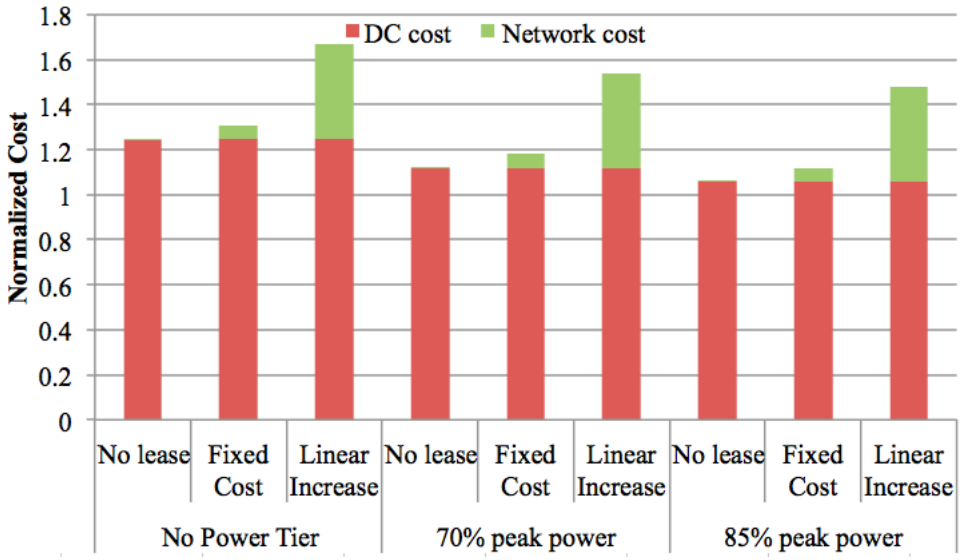


Figure 4.3: Normalized performance maximization algorithm costs for data centers and network

that net-work providers use. For linear increase (section 3.2.2), we see that the network cost can be up to 40% of the data center cost. This ratio increases with tiered energy pricing from $< 1\%$ to 25%, since this pricing scheme reduces data center power consumption and magnify the network cost.

For this policy, we also calculate the profit of network providers based on the energy costs associated with the WAN. Table 4.2 shows the profit normalized against fixed bandwidth cost and non-energy proportional routers. Energy proportionality of routers enables up to 37% more profit for network providers with ideal power curves and 20% with step proportionality WAN router power curve. We also observe that different power tier levels do not affect the savings of the network provider because the migration is based only on green energy availability in other locations.

4.4.3 Cost Minimization Using Migration

The main goal of the cost minimization policy is to maximize green energy usage and then leverage as much as possible inexpensive brown energy. Also, we show the impact of energy proportional servers to quantify the policy's benefit in

Table 4.2: Profit of network providers for performance maximization with different router energy proportionality schemes

Network Cost Function	**Profit**			
	Non-prop	*Step*	*Smooth*	*Ideal*
Fixed Cost	1x	1.2x	1.2x	1.4x
Linear Increase	4.5x	6.7x	6.8x	6.9x

future systems.

Unlike *performance maximization*, *cost minimization* does not transfer extra jobs, and thus, does not obtain any performance improvement. Furthermore, the overhead of network transfer decreases the performance of MapReduce jobs. We observe 23.8 min average job completion time for MapReduce jobs, 4.5% worse than the no migration case with green energy efficiency of 66%, a 7% improvement over no migration, with no performance overhead for service jobs.

In Figure 4.4, we show the impact of energy proportionality and tiered energy pricing to our model, normalized against the no migration case. We observe a 10% decrease in total cost when tiered energy pricing is incorporated into the model. Cost reduction grows to 15% when energy proportional servers are used. This shows the potential of cost minimization method in the future when servers become more energy proportional.

We also analyze how the total cost of data centers changes if the network is leased. Unlike the *performance maximization* policy, we prevent migration if the cost is higher than the potential savings. Figure 4.5 shows the results of this analysis, and additionally incorporates server energy proportionality. We use the same coefficients for the network cost functions as in the previous case. Neglecting the cost of network leasing can result in up to 15% error. The network costs are up to 17% of the data center cost, which is significantly less than results we saw with the performance maximization, where it is up to 40%. This is mainly because this policy sacrifices a potential increase in performance if the cost of a data transfer outweighs the cost savings. Figure 4.5 also shows how bandwidth utilization changes with different power tier levels and network lease options. First,

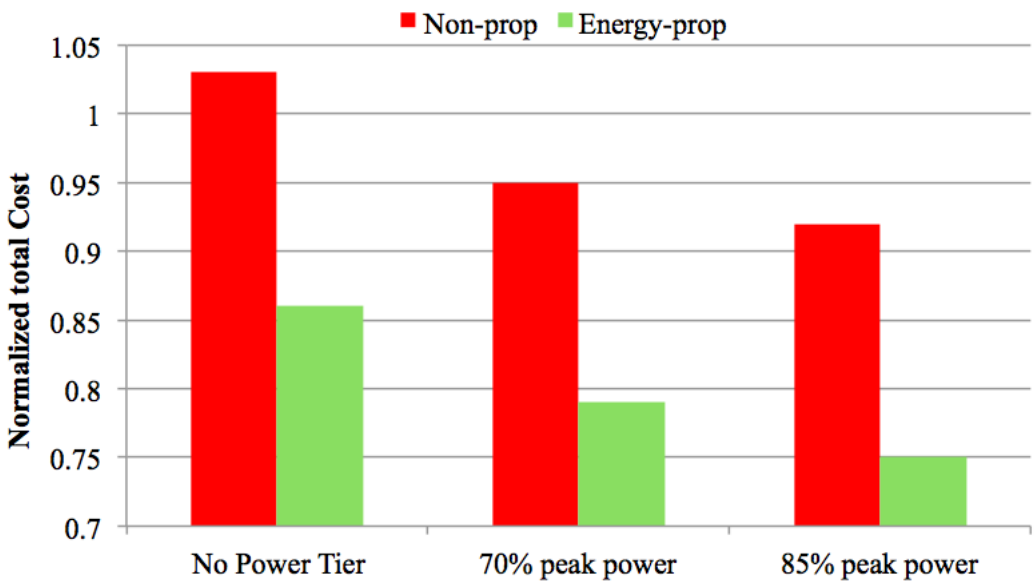


Figure 4.4: Normalized cost minimization algorithm costs with different power tier levels and energy proportionality

Table 4.3: Profit of network providers for cost. min. with different router energy prop. and with server energy prop.

Network Cost Function	**Profit**							
	Non-prop		*Step*		*Smooth*		*Ideal*	
	85%	*70%*	*85%*	*70%*	*85%*	*70%*	*85%*	*70%*
Fixed Cost	1x	1.2x	1.2x	1.4x	1.2x	1.4x	1.4x	1.6x
Linear Increase	2.2x	2.45x	3.26x	3.6x	3.4x	3.8x	3.5x	3.9x

as network costs become more dominant, bandwidth utilization decreases due to a growth in unfeasible data transfers. As a result, if the lease cost is not modeled, the average band-width utilization has up to 60% error. Introducing tiered power levels decreases network utilization because they create a more balanced energy cost scheme across data centers. Table 4.3 shows the normalized profit of the network providers. The cost minimization policy inherently limits network profits, since it only allows financially profitable transfers.

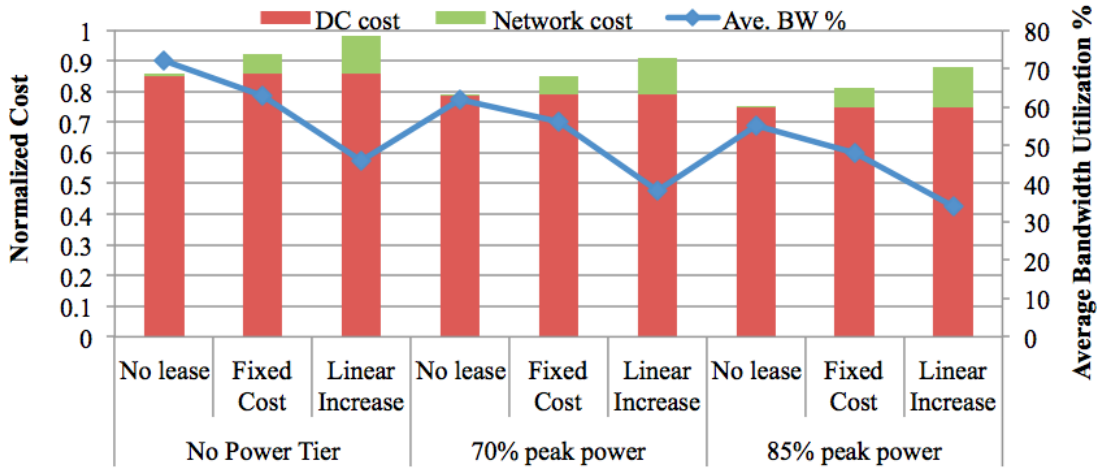


Figure 4.5: Normalized total cost and utilization for cost min. with different power tier levels and network lease options using energy proportional servers

4.4.4 Cost Min. Using a Green Energy Aware Network

We now investigate the cost minimization policy incorporating green energy aware routing (GEAR). Instead of simply selecting the shortest path between two data centers, GEAR chooses the path with the least brown energy need. As we only change the network routing policy for this scenario, data center cost values are similar compared to the previous case. An energy-aware network provides several benefits. Reducing brown energy costs of the WAN improves overall networking costs for both providers and data centers. It also provides a viable alternative for data centers, opting for cheaper green energy at the cost of GEAR's slightly increased network latency. Also, as network elements become more energy proportional in the future, we expect the energy savings obtained by GEAR to be more prominent.

Figure 4.6 compares SPR and GEAR in terms of router energy consumption and network provider profit, using fixed cost per bandwidth. GEAR with energy proportionality increases profits by 50% compared to the base case (non-proportional, SPR), and provides profit for all proportionality schemes. Without energy proportional routers, GEARs brown energy consumption is slightly lower than SPR (62% vs. 65% of SPR) with a 3% increase in network delay as a result

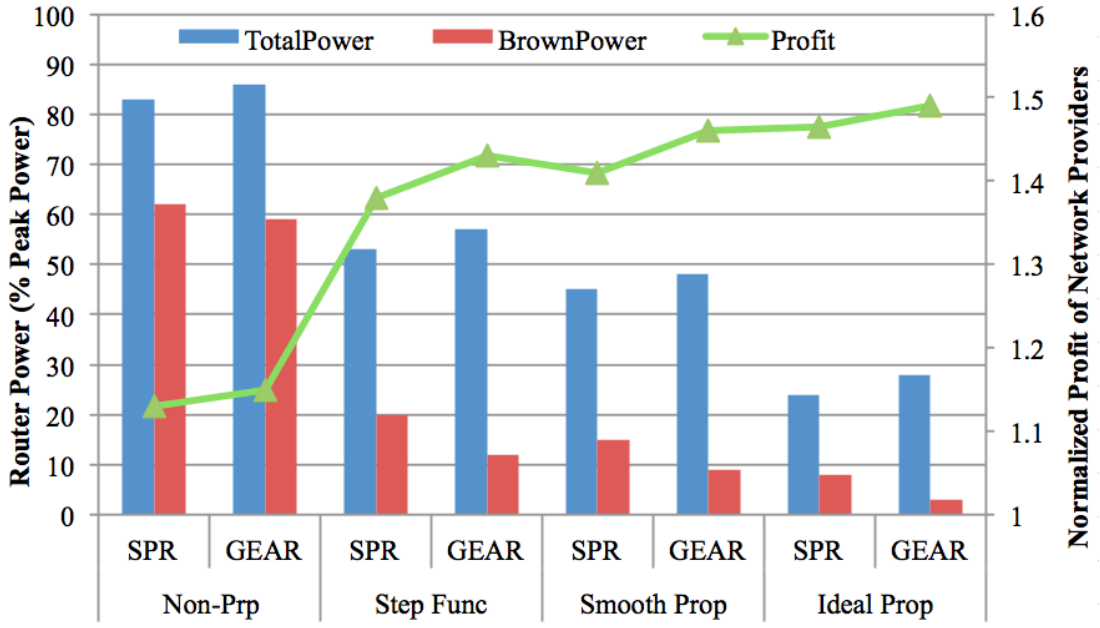


Figure 4.6: Comparison between SPR and GEAR energy consumption of routers and network profit with different energy proportionality schemes

of occasionally choosing a longer path, though with negligible overall effect on the job completion time.

4.5 Discussion

In this section, we first recap the most important results presented in chapters 3 and 4. We then compare our methodology with previous work, and explore the lessons learned with our analysis. Table 4.4 shows the comparison among the methods discussed in the previous sections. Our performance maximization algorithm uniquely leverages both workload and green energy differences across distributed data centers to maximize both throughput (27% improvement) and green energy efficiency (44% increase). We also demonstrate that the same variations in workloads and green energy can be leveraged for cost minimization, where our algorithm utilizes tiered energy pricing, and both migration and green energy aware routing. The results show up to 19% reduction in energy cost and 7% improvement in green energy usage while meeting QoS of latency sensitive applications,

and increasing job completion time of batch jobs by only 4%. Additionally, the comprehensive and novel aspects of our model provide a level of realistic simulation that previous models do not exhibit to make a complete analysis.

Green Energy Prediction and Workload Migration: Green energy prediction mitigates the inefficiency caused by the variability of renewable sources. We further improve inefficiency by matching our prediction horizon to the long-running batch jobs. The result is better decision making, and as the results indicate, up to 26% improvement in green energy efficiency. Previous work [91][80] only uses green energy as a method to reduce carbon footprint, and deploy workload migration to improve performance considering load balancing and resource availability [83]. In contrast, we show green energy can also be used to improve performance. We initially propose the idea in chapter 2 for a single data center, but now leverage prediction and availability across a network to run extra batch jobs in remote locations. We obtain 27% better batch job completion time compared to no migration with only a 6%-12% increase in total energy cost. Our work is the first to demonstrate the potential of green energy not only as a resource for environmental concerns, but also a means of performance improvement. While cost minimization precludes all potential migrations due to network costs, it still has 7% improvement in green energy usage.

WAN Ownership and Leasing: Related work assumes that WAN is part of the data center network, or applies static bandwidth costs. However, the WAN may be leased or owned, typically with bandwidth-dependent pricing. The work in this chapter considers such costs. Our first observation is that higher network cost reduces the bandwidth utilization. Secondly, despite increasing network costs with larger cost functions, data centers can obtain 2-19% cost savings by checking the financial feasibility of each potential migration. In contrast, when the data center owns the network, disregarding the initial WAN cost, it achieves up to 22% cost savings.

Tiered Energy Pricing: Previous studies on minimizing total energy cost, [33][109] use grid pricing as either fixed or variable with load. Others [59] attempt to limit data center peak load but do not consider how different power

Table 4.4: Comparison of different policies with respect to total cost, MapReduce performance and green energy usage

		Non-Energy Proportional Servers			**Energy Proportional Servers**		
Policy	**MapReduce Job Completion Time**	**Power Tier**	**Total Energy Cost**	**Green Energy Usage**	**Power Tier**	**Total Energy Cost**	**Green Energy Usage**
No Mig.	**22.8 min**	-	**1.22x**	**59%**	-	**0.99x**	**47%**
		85%	**1x**		85%	**0.85x**	
		70%	**1.10x**		70%	**0.92x**	
Perf. Max.	**16.8 min**	-	**1.25x**	**85%**	-	**1.03x**	**80%**
		85%	**1.06x**		85%	**1x**	
		70%	**1.12x**		70%	**1.05x**	
Cost Min.	**23.8 min**	-	**1.03x**	**66%**	-	**0.86x**	**60%**
		85%	**0.92x**		85%	**0.75x**	
		70%	**0.95x**		70%	**0.79x**	

levels can affect overall energy cost. Not modeling different cost regions for data center energy consumption may not be correct due to large power consumption of the data centers. We demonstrate that proposed improvements might be over-estimated by up to 20% when accurate pricing is taken into account. Both of our algorithms inherently attempt to remain below tiered power levels in order to avoid higher energy prices, and only exceed those limits when inevitable, i.e. when all data centers are over-provisioned. Consequently, while our algorithms' performance and cost benefits are tempered by the incorporation of tiered energy pricing, we can still show up to 15% cost savings.

Energy-Proportional Routing: We investigate the future of data center communication, analyzing the impact of energy proportionality of routers on network provider profit. We show that dynamic, green energy aware routing (GEAR) policies can improve energy efficiency by reducing brown energy consumption up to 65%. We quantify that energy proportionality can increase the profit of network providers up to 35% and 57% with fixed and linear policies, respectively. The difference in profit between an implementable proportionality scheme (i.e. step-function) and the ideal case is between 5-17% and decreases with increasing network lease costs. The key observation is that router energy-proportionality schemes can increase profits significantly if deployed, and that GEAR can decrease network brown energy use up to 3x with energy proportionality (chapter 3) with negligible performance impact.

Power-Proportional Computing for Future Systems: Current data center hardware is highly non-energy proportional, resulting in power-inefficient systems. There has been recent work [27] on designing energy-proportional elements. This chapter quantifies the benefits of this trend in both major aspects of a data center network: servers and network elements.It shows the benefit of optimizing the components individually and together into an ideal energy-proportional system, with up to 30% energy savings despite being limited by tiered energy pricing and network contracts. Table 4.4 quantifies both the impact of such systems, and the continued benefit of our algorithms in a power-proportional environment.

4.6 Conclusion

Energy efficiency and green energy usage in data centers and their networks has gained importance as their energy consumption, carbon emissions, and costs have increased dramatically. This chapter analyzes multiple data center systems and develops two online job migration algorithms. It also explores tiered energy pricing for data centers, network cost models and the costs of owning/leasing a data center WAN. Green energy variability is addressed by prediction algorithms. The performance maximization algorithm demonstrates the ability to leverage green energy to improve workload throughput, rather than simply reducing the operational costs. The chapter further explores the viability of the two algorithms in the face of emerging technologies in data center infrastructure, showing continued benefit of both the performance maximization and the cost minimization algorithms in the presence of energy proportional computing and communication.

Chapter 4 contains material from "A Comprehensive Approach to Reduce the Energy Cost of Network of Datacenters", by Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga, which appears in Proceedings of International Symposium on Computers and Communications (ISCC), 2013 [18]. The dissertation author was the primary investigator and author of this paper.

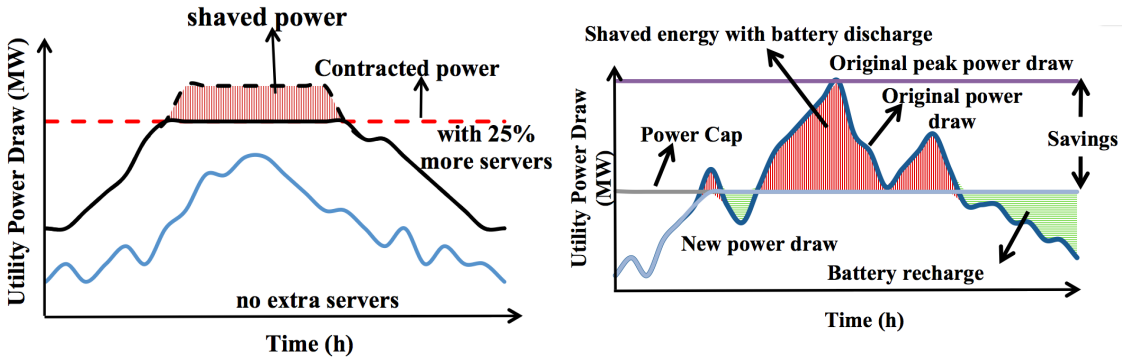
Chapter 4 contains material from "Renewable Energy Prediction for Improved Utilization and Efficiency in Datacenters and Backbone Networks", by Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga, which will appear in Computational Sustainability, Springer, 2015 [11]. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Efficient Peak Power Shaving in Data Centers

Warehouse-scale data centers consume several megawatts and require careful power provisioning to ensure that costly power infrastructure is utilized effectively [50]. The electricity bill of these data centers consists of two parts: 1) electricity cost, 2) peak power cost. This chapter focuses on the second part of the utility bill. The peak power cost is based on the maximum amount of power drawn by the data center during the bill period. The rate of this cost can be high, such as $12/kW [62]. Although data centers consume significant amount of power, they reach their peak capacities rarely [50]. This leads to increased peak-to-average power ratio, and thus, the peak power costs can contribute up to 50% of the utility bill [62]. If data centers can reduce their peak-to-average ratios, they may decrease their utility bill significantly.

Data centers often size their power infrastructure based on the expected peak power to avoid costly overages. The basic problem with power provisioning involves using as much power capacity as possible without exceeding a fixed power budget. Although individual machines may consume peak power, entire clusters of machines rarely operate at peak power simultaneously [50]. Several studies proposed peak shaving (capping) to increase power utilization [63][97], while maintaining power budgets and amortizing capital expenditures over more machines [73].



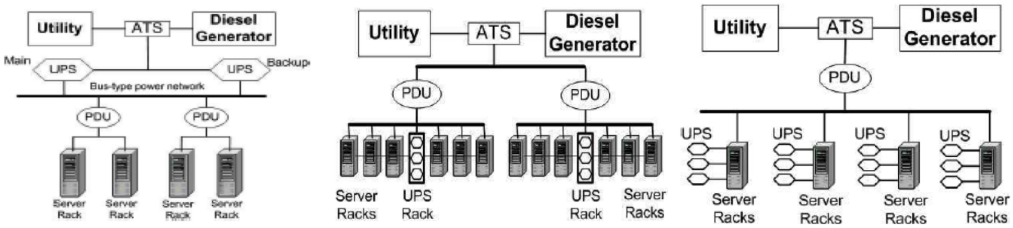
(a) with extra servers

(b) without extra servers

Figure 5.1: Sample peak power shaving with batteries

Although many mechanisms have been proposed for peak power shaving in data centers (such as dynamic voltage and frequency scaling (DVFS) [50][88], virtual machine power management [93], online job migration [33][109][131]), this chapter focuses on battery based peak power shaving because 1) Batteries already exist in data centers, 2) Batteries do not introduce the performance overhead associated with meeting the power budget. This is especially critical during the peak user demand. Battery-based peak shaving instead employs an uninterruptible power supply (UPS) to power machines.

Figure 5.1 illustrates two different strategies for using peak power shaving. The horizontal axes represent a 24-hour interval and the vertical axes show the aggregate power consumption. In Figure 5.1a, the dotted horizontal line denotes the contracted power for the data center. The lower curve indicates the power consumption of a nominal size data center without peak shaving. A significant amount of provisioned power is wasted during low activity periods, resulting in lower profit. The upper curve adds extra servers and handles oversubscribed power with peak shaving, so that the power utilization is higher. Peak shaving prevents the power consumption from exceeding the contracted energy costs shown by the shaded region. The dashed line illustrates how much power the data center would consume without peak shaving, which would then incur as much as 5x higher costs. Peak shaving increases the revenues by adding more machines to service more users and prevents utility-facing power consumption from exceeding the provisioned power



(a) Centralized batteries (most common)

(b) Rack level distributed batteries (Facebook)

(c) Server level distributed batteries (Google)

Figure 5.2: Sample battery placement in data centers [73]

with no performance cost.

Figure 5.1b uses peak shaving just to decrease the level of contracted power without increasing the number of servers. The upper horizontal line represents the original peak power demand and the lower one shows the power cap. The difference between the original power draw and the power cap corresponds to energy savings as the data center can contract for less power. If the power demand is greater than the power cap, the batteries provide energy. During low power demand, the batteries recharge to regain energy in preparation for the next peak.

The most common battery placement is centralized, shown in Figure 5.2a. If the data center uses a centralized UPS, the entire circuit is switched to battery until the batteries exhaust their capacity or the peak subsides. This technique is useful primarily with short pulses (a few minutes long) due to low battery capacity [62]. Recent trends in data centers focused on distributed UPS architectures, where individual machines [61] (shown in Figure 5.2c) or collection of racks [49] (shown in Figure 5.2b) have their own UPS. This architecture shaves power more effectively due to the finer granularity but only works for data centers willing to implement the non-standard power architecture [73].

The main disadvantage of a centralized UPS design is the double AC-DC-AC conversion, leading up to 35% energy loss. The distributed design can avoid this double conversion by taking batteries next to the servers. Recently, DC power distribution in data centers has been proposed as a solution to decrease the conversion losses. This chapter analyzes the conversion losses of these different designs

and quantify the effects the losses have on peak shaving capabilities.

Existing approaches discharge batteries in a "boolean" fashion: the entire data center power domain is fully disconnected from the utility power and supplied from the UPS. As a result, batteries discharge at much higher currents than rated, which lowers battery lifetime and raises the cost. This chapter first introduces an accurate battery lifetime model and shows that without an accurate model, savings estimations of the previous studies may not hold.

Distributed UPS design addresses this issue partly by providing the ability to discharge only a subset of batteries in a data center at a time and by using lithium iron phosphate (LFP) batteries which have both higher energy capacity and 5x more charge/discharge cycles than lead-acid (LA) batteries. The individual batteries are directly connected to servers, but still operate in boolean mode, leading to lowered battery lifetime and higher cost. Also, distributed batteries require coordination to provide the best performance. Palasamudram et al. [97] assume a centralized control mechanism and do not model the effects of coordination in their study. Kontorinis et al. [73] analyze the peak shaving performance of control mechanisms placed at different levels of power hierarchy. They conclude that the centralized controller for distributed batteries performs the best but do not comment on the feasibility of this centralized solution for a large scale system. In this chapter, we estimate that the response time of a centralized controller can take up to multiple seconds which may be too long to meet the power thresholds. We design a distributed battery control mechanisms that both address the latency problem of the centralized controllers and still provide near-optimal peak power shaving performance.

A key insight that we leverage in this chapter is that the ideal design should have the minimum management overhead of the centralized UPS with the capability to provide "just enough" current to the data center, at a level that optimizes the individual battery lifetime. This can be accomplished by an architecture where the batteries do not power an entire entity, e.g. server, data center, but are allowed meet the power demand partially. This chapter proposes a grid-tie inverter based architecture that has this property.

In summary, this chapter makes the following contributions.

1. We revisit the analyses for existing peak shaving designs using more realistic battery models and find that the benefits of peak shaving may be overestimated by up to 3.35x with simplistic models, resulting in unacceptably short peak power shaving times of only several minutes, for the centralized lead-acid UPS designs.

2. We present a distributed battery control mechanism that achieves the battery lifetime and power shaving performance within 3.3% and 6% of the best centralized solution with only 10% of its communication overhead. This power shaving enables 23MWh/week energy shaving or 8760 additional servers within the same power budget when scaled to a typical 10MW datacenter.

3. We propose a new peak power shaving architecture. We use a centralized UPS architecture using grid-tie inverters to partially power loads (in contrast to previous boolean discharge), so that the battery capacity decreases super-linearly with respect to discharge current [116], thus enabling the partial discharge architecture to overcome the efficiency problems associated with the state-of-the-art solutions. Our centralized grid-tie solution has 78% longer battery lifetime and doubles the cost savings compared to the best SoA distributed designs. Also, since the batteries are placed together, the communication overhead is reduced by 4x.

This chapter first outlines the issues with the existing designs and quantifies the relevant problems. It follows with a section demonstrating the battery model which is used to estimate the physical conditions of a battery. Afterwards, we first describe the distributed battery control solution and then present our grid-tie based battery placement architecture. The chapter continues with the methodology section where it explains the experimental methodology used, i.e. how we setup the experiments, along with any applicable parameters and cost models. The results are organized to show 1) the effects of accurate battery model, 2) performance of distributed battery control, 3) performance of the grid-tie based battery placement architecture. We finalize the chapter with a conclusion chapter.

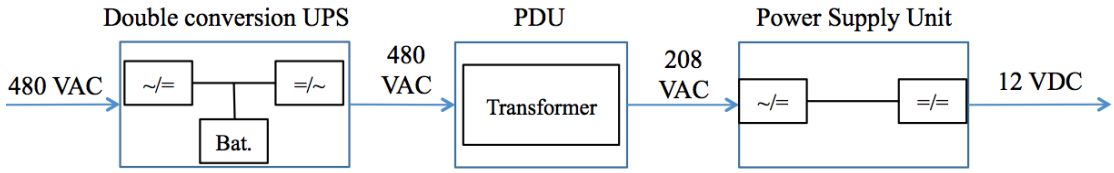
5.1 Issues with the Existing Battery Placement Architectures

5.1.1 Centralized vs. Distributed Designs

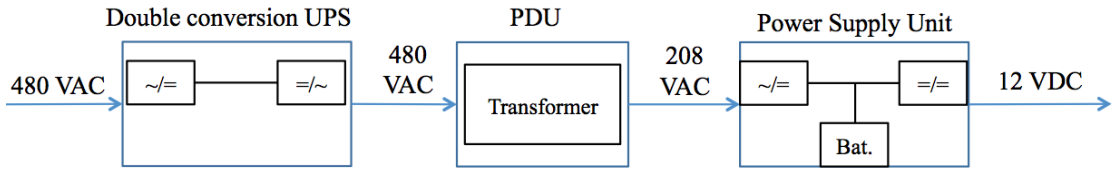
There are two main battery placement architectures: centralized and distributed. The centralized design uses batteries within the data center-level UPS and does not require additional power equipment or infrastructure. In Figure 5.3, we present different power delivery and battery placement options for data centers. We have PDU as the power distribution unit and UPS as the uninterruptible power supply, VAC and VDC as the voltage values using AC and DC power options, respectively. A common power delivery hierarchy for this design using AC distribution is shown in Figure 5.3a. When peak shaving occurs, the battery powers the entire data center, discharging the batteries at high rate. According to Peukert's Law, this drains battery capacity very quickly. Also, both the AC-DC-AC double conversion in UPS and the losses on the power delivery path result in up to 35% energy loss. These losses reduce both UPS efficiency and useful battery capacity.

The distributed design co-locates the servers and batteries and eliminates the DC-AC battery power conversion [73][97]. A sample design with is shown in Figure 5.3b. Each server may be switched to battery independently. This leads to finer grained control of the total battery output because only a fraction of the servers are operating on battery at any given time. Together, conversion efficiency and fine-grained control permit longer peak shaving than traditional centralized designs.

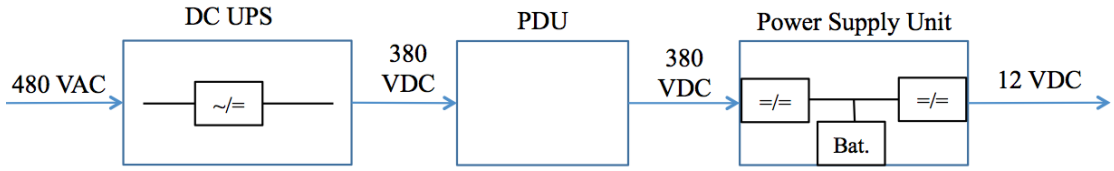
In Figure 5.4, we compare the power shaving capabilities of the state-of-the-art centralized and distributed designs during a fixed-magnitude spike in demand without considering conversion losses. We assume each server has a 20Ah LA battery in the distributed design because that is the maximum size that can fit in a rack [73]. The centralized design has equivalent aggregate capacity to the distributed batteries. In Figure 5.4a, the x-axis illustrates a range of peak server power values. We assume a provisioned power of 255W per server. This value



(a) State-of the art centralized power distribution



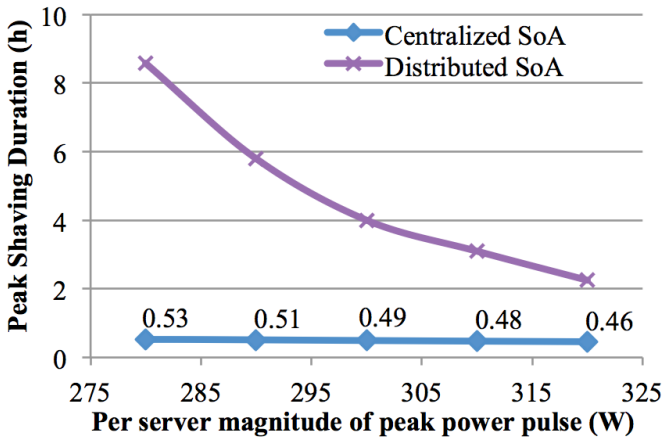
(b) AC power distribution with distributed batteries



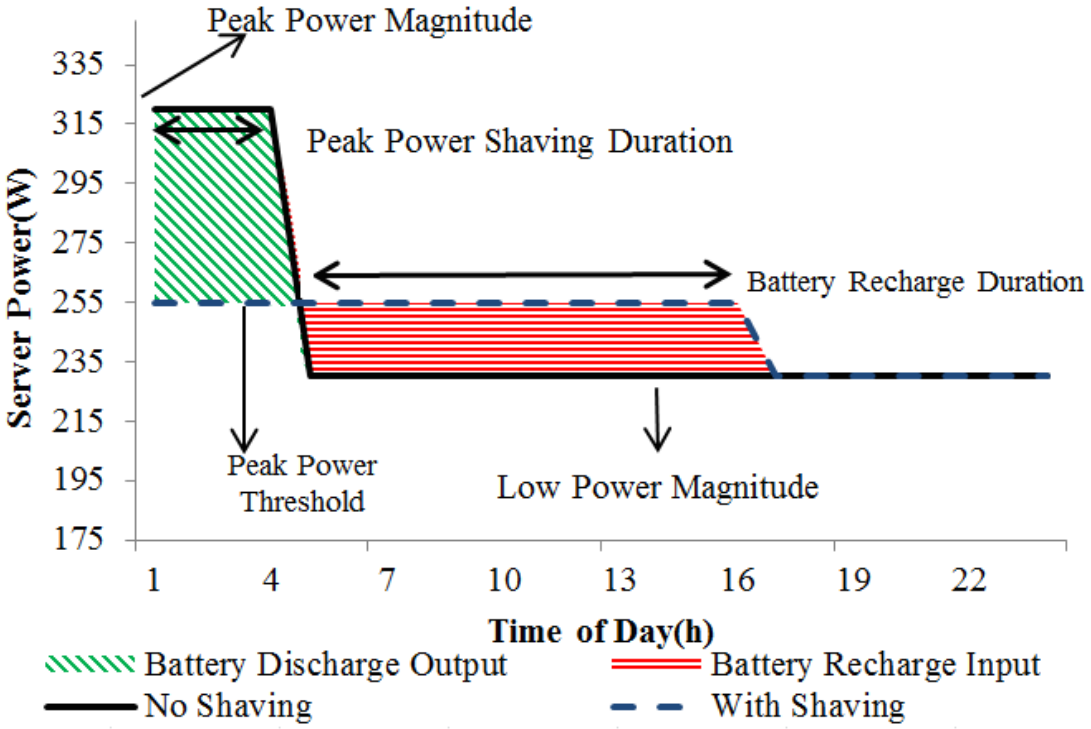
(c) DC power distribution with distributed batteries

Figure 5.3: Different power delivery options with centralized and distributed battery placements

limits the power consumption of the entire data center to $255 * (N_{servers})$, where $N_{servers}$ is the number of servers in the data center. The y-axis represents the peak shaving duration corresponding to different peak power spikes. We illustrate the fixed peak power magnitude and peak power threshold in Figure 5.4b. In this figure, the power curve of a data center consists of two long pulses: the peak pulse and the low pulse. The resulting power curve after peak shaving is mostly linear, having the value of the provisioned power. We define the duration batteries can sustain a specific peak pulse as the peak shaving duration. Figure 3-a has two curves showing the peak shaving durations for both centralized [62] and distributed [97] designs with different peak pulses. The former cannot scale its peak shaving duration for lower magnitude peaks, whereas the latter can throttle the battery energy. The latter reduces peak power even for higher peak spikes, outperforming the centralized design by 5x when shaving 25% above provisioned power.



(a) Peak shaving capabilities of different designs



(b) Illustration of fixed peak magnitude and peak shaving duration

Figure 5.4: Peak power shaving comparison of centralized vs. distributed designs

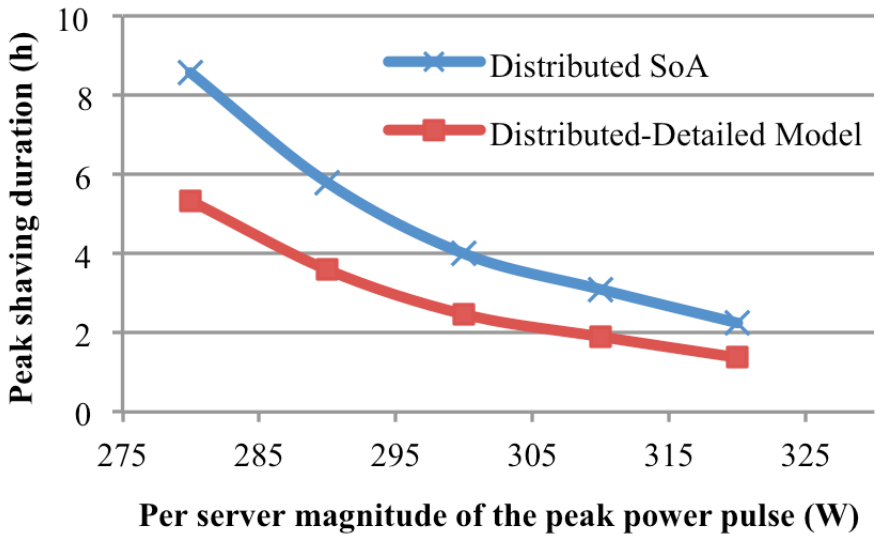


Figure 5.5: Peak power shaving capabilities of the distributed design

5.1.2 Problems of the Distributed Design

Even though the distributed design achieves finer grained control, each battery still needs to power the entire server with high discharge currents. The existing distributed architectures do not account for the negative effects of high discharging rate. Figure 5.5 shows the peak shaving capability of the distributed design with and without a detailed battery model. The figure setup is the same as in Figure 5.4a. The upper curve estimates peak shaving duration with a simplistic battery model and the lower curve uses a more exact model, both outlined in the next section. We see that the power shaving duration can be overestimated by up to 62% without a detailed battery model.

The ability to discharge batteries independently is crucial in the distributed design. However, since not all batteries are discharged at the same time, they may have very different discharge patterns, depending on server load. This variation results in capacity imbalances between them. Figure 5.6 represents this variation one and two years after the batteries are deployed when selecting batteries randomly each time battery power is needed. The outermost circle represents the nominal battery capacity. The innermost circle corresponds to the end of the battery's useful life. We consider a battery dead when it can use only 80% of its

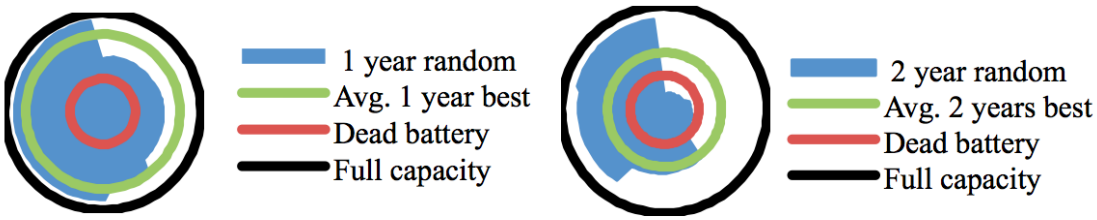


Figure 5.6: The maximum battery capacity with random battery selection

nominal capacity [103]. Each battery is denoted by a ray extending from the center. The length of the ray indicates the battery capacity. The line between the nominal and dead capacity indicates the ideal battery lifetime at each age. This graph illustrates that remaining battery capacities significantly deviate from the ideal. This deviation increases over time, resulting in early battery replacements, increasing the battery related costs. We may reduce this variation by selecting batteries more effectively. This requires coordination between the batteries, which may have delays on the order of seconds depending on network congestion. Large delays can lead to miscalculating the total available battery capacity, reducing the peak shaving.

Previous studies on distributed batteries [73][97] assume a centralized control mechanism to obtain the best peak shaving performance with them. Palasamudram et al. [97] do not actually model a controller but their solution depends on the coordination among all the batteries, implying centralized control. Kontorinis et al. [73] use controllers deployed at different levels of power hierarchy. Table 5.1 shows the different hierarchy levels used in that study and the corresponding number of batteries each controller needs to manage. Table I also shows the best peak shaving percentages obtained with each level of controller. Kontorinis et al. conclude that a centralized control mechanism is required to get the best performance of the distributed batteries. But, since the batteries are distributed to the servers, the centralized control mechanism needs to use the data center interconnect to manage the batteries. Kontorinis et al. do not analyze the effects of data center interconnect delays.

Table 5.1: Group sizes, equivalent hierarchy level and the best beak power shaving performance for each group

Hierarchy Level	Size of a Group	Best Peak Power Shaving
Server	1	10%
Rack	20	12%
PDU	200	16%
Cluster	1000	19%

5.1.3 DC Power Delivery in Data Centers

Currently, data centers distribute AC power because it is easy to deliver and transform. This requires multiple conversions in the power delivery hierarchy (Figure 5.3a and 5.3b), such as AC-DC-AC conversions in a centralized UPS and AC-DC conversion in the server power supply. These conversions reduce the efficiency of the centralized battery output and the distributed battery input. The former reduces the useful discharge time of the battery, and the latter leads to longer recharges.

In contrast, DC power distribution has been proposed to improve energy efficiency [104][53]. The AC utility input is converted to DC once within a centralized DC UPS. Delivery and transformation are handled using DC. The DC option aids UPS-based peak shaving because it eliminates multiple AC-DC conversions, and up to 35% energy loss on the power delivery path. Figure 5.3c shows a sample DC power distribution system with distributed batteries. This design can reduce the power distribution losses significantly compared to the AC distribution. We quantify these savings in the results section of this chapter. Despite its advantages, DC is not common, as it requires a new power infrastructure. It is a good option for new data centers but impractical for existing ones as the entire power distribution system must be redesigned.

5.2 Detailed Battery Model

Peak shaving using batteries needs accurate estimates of battery's physical behavior. This section demonstrates how we calculate the useful battery capacity over time and estimate its depth-of-discharge (DoD) along with its available capacity after recharging and discharging. The available battery capacity at a given time is defined as the state-of-charge (SoC) and reported as a percentage of the maximum capacity. State-of-Health (SoH) quantifies the maximum deliverable capacity of a battery over time as a percentage of its initial capacity.

There are several studies estimating battery SoC and SoH, especially for mobile devices, e.g. [29][107]. In this section, we combine a few models to both estimate the physical properties of the batteries and capture the negative effects of high discharging currents. Coulomb counting method presented in [94] describes the relation between DoD level and SoH. We take the model described in [45] to capture the effects of high discharge currents on SoH. We also include Peukert's law which states that the effective capacity of a battery decays exponentially with increasing discharging current [116]. The main benefit of this model is its simplicity and ability to easily leverage it in a large scale installation as it requires only voltage and current readings for all the calculations. We start describing our model by first calculating released capacity during a discharge event:

$$C_{released} = (\Delta t) I_{discharge} \tag{5.1}$$

where Δt is the length of the time interval and $I_{discharge}$ is the discharge current. Then, we compute the DoD as the released capacity over the effective capacity:

$$DoD = \frac{C_{released}}{C_{eff}} \tag{5.2}$$

$$C_{eff} = C_R \left(\frac{C_R}{I_{discharge} H}\right)^{k-1} \frac{SoH}{100} \tag{5.3}$$

where C_{eff} is the effective battery capacity when using $I_{discharge}$ as the discharging current and C_R is the rated capacity. We use H to denote rated

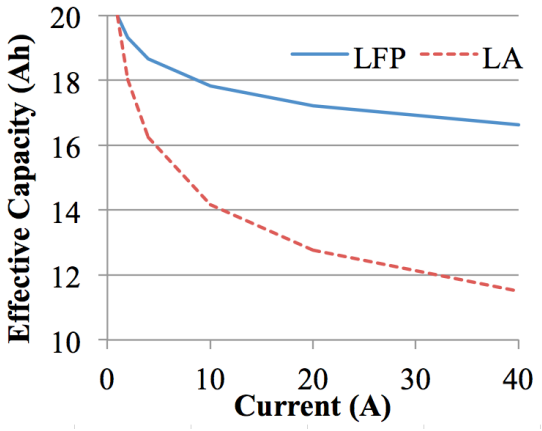


Figure 5.7: Effective capacity of 20Ah LA and LFP batteries

discharge time in terms of hours and obtain its value from the data sheets, which is generally 20 hours [116]. Peukert's exponent is shown by k, which changes depending on the battery type. For LA batteries, the typical value is around 1.15 whereas for LFP batteries it is 1.05 [66]. Effective capacity is also scaled with SoH value to reflect the capacity loss as the battery is used. The DoD is subtracted from the SoC at the end of each interval. When discharging ends, we save the total DoD value during that discharge period, DoD_{final} as $(100 - SoC)\%$.

Peukert's law states the effective capacity of a battery decreases with higher discharge current. Figure 5.7 shows this negative effect on 20Ah LA and LFP batteries. The horizontal and vertical axes show the effective battery capacity and discharging current respectively. The effective capacity of the LA battery decreases faster due to its greater nonlinear behavior, represented by a larger Peukert exponent. At 40A, corresponding to 2C rate for both of these batteries, the LA battery loses 42% of its nominal capacity, but the LFP battery loses only 15%.

We update the battery SoH after a complete charge/discharge cycle [94]. This update depends on the battery chemistry, determining Peukert's exponent, C_{eff} and DoD_{final}. The number of charge/discharge cycles decreases with deeper discharges, represented by a larger DoD_{final} value. We use a lookup table derived from effective capacity graphs provided in commonly available battery data sheets; similar to Figure 5.8 for each battery chemistry to define the effects of DoD_{final}.

In Figure 5.8, the horizontal axis shows the DoD level for charge/discharge

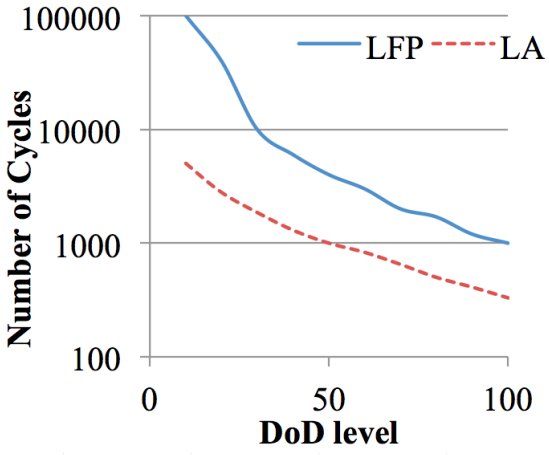


Figure 5.8: Cycle life of LA and LFP batteries rated at 20h [119][129]

at 20h discharge rate, which is defined as the current that drains the battery in 20h. The vertical axis is on a log scale and illustrates the number of cycles a battery can provide for a particular DoD level. As the battery is discharged deeper in each cycle, the available number of charge/discharge cycles decreases exponentially. LFP batteries provide 5x more cycle life compared to LA batteries in average.

We normalize the effect of one cycle with DoD_{final} value to calculate its impact on the battery lifetime. The battery lifetime is defined as the interval in which battery SoH is greater than a state of health value which determines when the battery is dead, SoH_{dead}. Battery manufacturers generally recommend 80% for this value [103][111], i.e. the battery is considered dead if the maximum capacity it can provide falls below 80% of its rated capacity. If the battery has $Cycles_{DoD_{final}}$ cycles with DoD_{final} value, the battery SoH is updated as:

$$SoH = SoH - (100 - SoH_{dead}) \frac{1}{Cycles_{DoD_{final}}} \frac{C_R}{C_{eff}} \tag{5.4}$$

This equation normalizes the effect of one cycle with DoD_{final} over the battery lifetime and penalizes high discharge currents.

Batteries generally include a battery management unit that both manages and monitors the voltage and the current of the battery. This unit makes it practical to use our model as it requires only voltage and current measurements of the

Table 5.2: Battery model validation

Battery	**Error**
$Li-ion_5$	$4.35\% \pm 2.05\%$
$Li-ion_6$	$5.83\% \pm 3.60\%$
$Li-ion_7$	$3.84\% \pm 2.75\%$

battery. In contrast, the simple battery model used in existing studies [63][97] does not calculate C_{eff}. They use nominal battery capacity, C_R, to compute DoD. This leads to up to 42% overestimated discharge duration and thus, overestimated peak shaving capabilities. They also assume the same battery capacity over its lifetime and do not consider the effects of decreasing SoH on C_{eff}, further increasing the errors.

5.2.1 Battery Model Validation

We validate our model using the battery data available from the NASA Ames Prognostics Data Repository [113]. The repository includes the measurements of 2Ah Li-ion batteries charging and discharging at different currents and temperatures. Each measurement has the complete charge/discharge profile of a single battery until the end-of-life condition. We check our model using the results of three batteries tested at room temperature. We compare the estimated SoH value of each selected battery using our battery model against the measurement points in the database. Table 5.2 shows that our model has 4.67% average error compared to the measurements.

Next, we introduce distributed battery control mechanisms for the distributed battery architecture and our centralized battery placement design that utilizes grid-tie inverters.

5.3 Distributed Battery Control

The distributed architecture permits finer grained control than centralized architectures because server batteries may be discharged independently. This process requires intelligent selection of batteries during each power peak. In section 5.1.2, we demonstrate that simple battery selection algorithms may distribute power load unevenly and induce high variations in battery SoH. This variations lead to premature battery replacements because capacity is reduced sooner than expected. Therefore, in this section we introduce a mechanism that monitors battery health and selects batteries in a way that minimizes this variation for the distributed battery design.

The distributed controller first estimates the number of batteries to discharge during each peak power pulse as follows:

$$N_{batteries} = \left\lceil \frac{P_{demand} - P_{threshold}}{V_{battery} I_{discharge}} \right\rceil \tag{5.5}$$

where $\lceil . \rceil$ is the ceiling function, P_{demand} is the peak power demand at a given time, $P_{threshold}$ is the peak power threshold to be maintained, $V_{battery}$ is the single battery voltage and $I_{discharge}$ is the single battery discharging current. We use 12V batteries [61] and set $I_{discharge}$ to 23A. Since the servers use the battery power without AC-DC conversion, the battery incurs no conversion losses in the server. In our experiments, the measured server peak power is 350W and power supply unit (PSU) efficiency is 80%. Therefore, the server actually uses 280W, which corresponds to 23A discharging current.

An ideal controller for the distributed design should poll every server to gather data on server power demand, battery SoC and SoH. This process requires message exchanges through the data center network. However, the controller becomes subject to communication delays between the thousands of servers and large background traffic. Previous work shows that the switch delay can increase by over 100x with excessive queuing in the switches [40].

Our method groups the batteries into multiple distributed controllers to address the communication complexity. Table 5.3 lists the possible group sizes

Table 5.3: Group sizes in data center power delivery hierarchy

Hierarchy Level	**Size of a group**
Server	1
Rack	20-50 [62]
PDU	200 [73]
Cluster	1000 [73]
Data center	Multiple clusters

and shows the corresponding level in the data center power hierarchy. The two extremes represent fully localized control, at each individual server, and the data center level, which is equivalent to fully centralized control. In between are rack level, PDU, which consists of approximately 10 racks, and cluster level, which is about the size of a typical data center container. We chose these hierarchy layers as they correspond to the typical organization found in the data center's power hierarchy.

Each level of the controller implements one of the policies shown in Table 5.4 to select a battery. *Random*, *Least-Recently-Used (LRU)* and *Max-SoH-local* policies make a local decision regarding which battery to use for peak power shaving from their immediate group. *Random* policy selects a random battery from available ones. *LRU*, also used in [73], always selects the next available battery from its local list. *Max-SoH-local* chooses the available battery with the greatest SoH value. We assume that the controllers do not know or predict the length of the upcoming peak power pulse. Hence, selecting the battery with the greatest SoH value is the best a controller can do because it minimizes the probability that the selected battery empties during the peak power pulse. These policies result in lower latency with smaller groups, but their knowledge about total power demand and battery status is limited.

We implement three other *Max-SoH* policies to address this problem. They are similar to *Max-SoH-local*, but controllers can communicate with other ones during a decision process. The *Max-SoH-global* policy represents a centralized controller and uses all data available in the system. Although this controller can

Table 5.4: Policies to control distributed batteries

Policy	Communication
Random	Local
LRU (Iterative) [73]	Local
Max-SoH-local	Local
Max-SoH-global	Global
Max-SoH-limited-communication	3 groups
Max-SoH-more-limited-communication	2 groups

make the best decision, it leads to large communication delays and becomes a single point of failure. *Max-SoH-limited* and *more-limited communication* policies are limited to two and one other groups. Each group's partners are assigned statically based on power and network infrastructure. We compare these policies with the local ones to demonstrate the trade-off between the communication overhead and power shaving, and battery lifetime performance.

Figure 5.9 shows the peak shaving and the battery selection process of a single group when communicating with others. The number of sharing groups depends on the policy. The controller first awaits power consumption and battery data from its sharing groups. It next computes the peak power that can be shaved by finding the number of batteries required and selects the batteries to use. Local batteries discharge immediately. Remote batteries require explicit signals to their controller. We use a timeout when waiting for the data from other groups to avoid problems, including miscalculating the total available battery capacity. The timeout may decrease the quality of selection since less data will be present.

5.4 Grid-tie Based Battery Placement

In previous sections, we show that previous designs do not capture the effects of a binary battery discharge, which requires high discharging current. Furthermore, the distributed design requires a centralized controller to obtain its peak performance. The performance of this controller highly depends on the data center

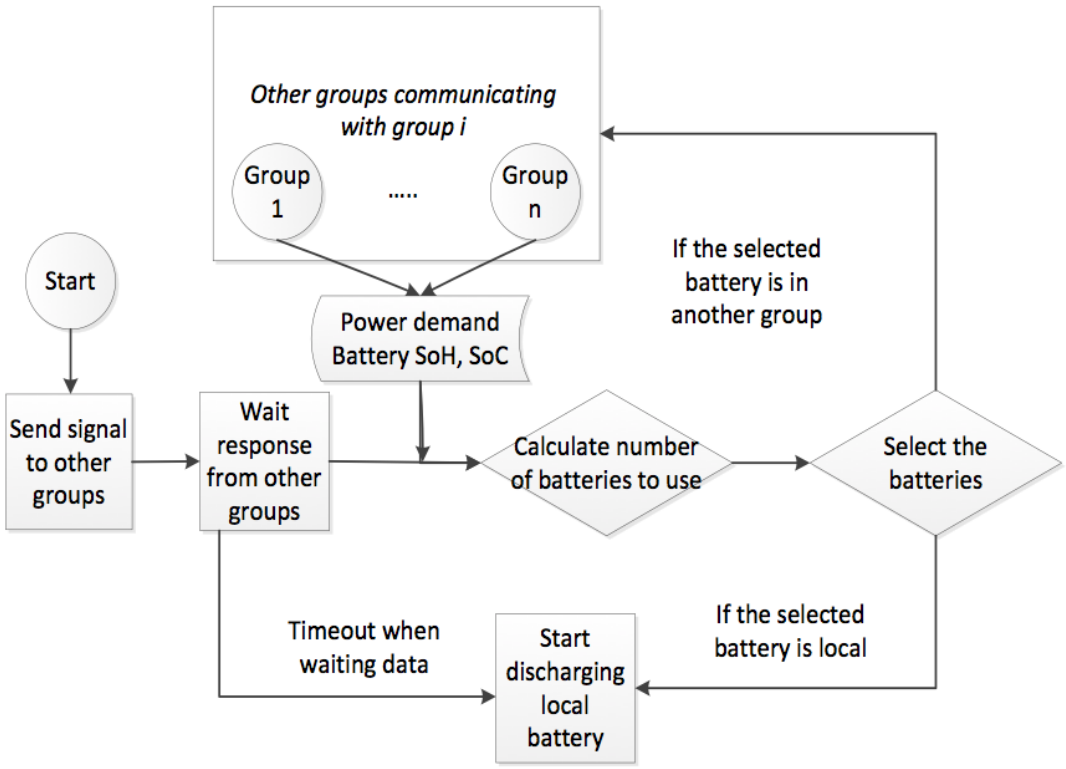


Figure 5.9: Battery selection with communication based policies

interconnect and can be negatively affected with increasing delay. Finer-grain control with the smaller batteries is a key to achieving good peak shaving results. This section presents our battery-based peak power shaving architecture. Our design places the batteries in a centralized location and connects their aggregate output to the utility grid with a grid-tie inverter. Our model outperforms the distributed design by exploiting the nonlinear nature of Peukert's Law, despite additional DC-AC conversion losses in the centralized UPS. It obtains improved battery lifetime and requires significantly less communication overhead than the state-of-the-art distributed designs.

Instead of decentralizing the batteries, we place them together and connect the batteries to the main grid using a grid-tie inverter. A grid-tie inverter allows any quantity of DC power to be converted into AC and fed into the grid in an efficient way [44][105]. Figure 5.10 presents the layout of our design. The integration of battery power to the system is controlled with two switches. When switch S_A

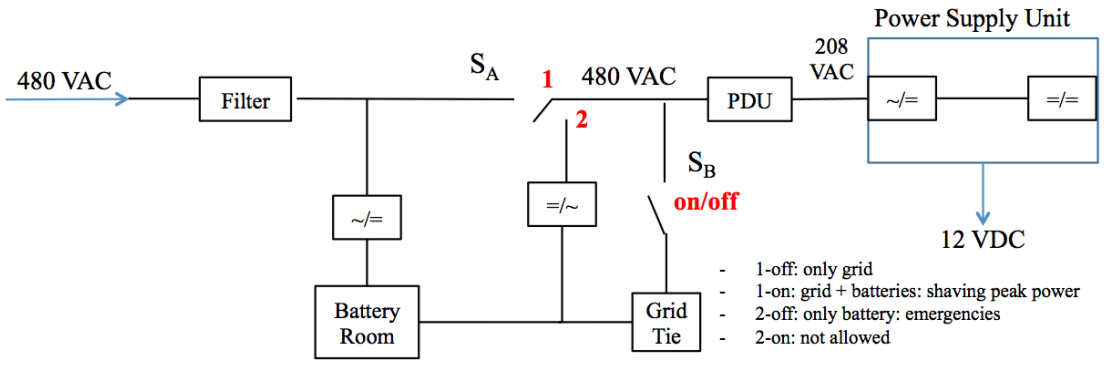


Figure 5.10: Grid-tie based battery placement design

is in mode "**1**", the data center operates normally, accepting any quantity of battery power output from the grid-tie, which is controlled by the switch S_B. If it is "**OFF**", the grid is the only power supplier, i.e. we are under the power threshold. If it is "**ON**", the batteries are active and shaving peak power. The batteries can be recharged directly by the grid through a rectifier. S_A is in mode "**2**" only in emergency cases, making sure that the only power supplier is the battery. The case where S_A is in mode "**2**" and S_B is "**ON**" is not allowed because it just combines the same battery output.

Even though the batteries are centralized, we still treat them as distributed and enable them to individually charge/discharge. The fact that grid-tie inverter allows any quantity of DC to be combined with AC makes it possible to adaptively select the discharge current of the batteries. Instead of using batteries with high current rates as in both state-of-the-art centralized and distributed designs, we can increase the number of batteries being discharged and scale down the current. In fact, this leads to finer grained control of the battery output compared to both existing designs. Furthermore, having more batteries used simultaneously with the same discharging current, we reduce the variation in battery discharge profiles. Decreasing both this variation and discharging current helps increase the battery lifetimes.

We place the batteries together and allow the discharging current to scale down instead of being in a binary mode. We have a set, Φ, of discharge currents, and we choose the smallest current from Φ that can sustain the peak demand with

the available batteries. Based on this current, we compute the number of batteries to use:

$$I_d = min_{I \in \Phi}\{I \mid IV_bN_b > (P_d - P_t) \ \ \& \ \ N_b \leq N_a\} \tag{5.6}$$

$$N_b = \left\lceil \frac{P_d - P_t}{V_bI} \right\rceil \tag{5.7}$$

where V_b is the voltage of a single battery, P_d is the peak power demand, P_t is the peak power threshold to sustain, N_a is the number of available batteries, N_b is the number of batteries required to discharge and I_d is the selected discharging current. These equations make sure that the minimum feasible discharging current is selected over all the selected batteries by ensuring the number of selected batteries is smaller than the number of available batteries. The set of available batteries include all the batteries having SoC greater than $100 - DoD_{goal}$, where DoD_{goal} is a predetermined value between 1 and 100 to better control the battery lifetime [73][97]. Larger DoD_{goal} values can shave bigger peak power pulses for a longer duration but they lead to shorter average battery lifetime values. We refer this process as the *discrete_current* policy.

This policy may select a subset of batteries to discharge. During battery selection, we choose the batteries available with the greatest SoH values. This minimizes the probability that a battery breaks down during discharging and it is the best a controller can do without any knowledge about the future power demand. The advantage of our architecture is that since the batteries are placed centrally they do not need to go through the data center network to coordinate for the battery selection process. They can use a dedicated network for this coordination. Thus, we can use a centralized controller with a much smaller expected latency.

Alternatively, we can use all the available batteries to discharge at the same current. We define the number of the available batteries, i.e. the ones with SoC greater than $100 - DoD_{goal}$, as N_a. The discharging current, I_d, becomes:

$$I_d = \left\lceil \frac{P_d - P_t}{V_bN_a} \right\rceil \tag{5.8}$$

where P_d is the power demand, P_t is the peak power threshold and V_b is the voltage of a single battery. Different than the previous policy, this policy does not let any battery to be idle during a peak power pulse, i.e. a battery is either drained or discharging. As a result, it does not have a predefined set of discharging currents and it selects the discharging current on-the-fly based on the number of available batteries. We refer this process as the *all_battery* policy. Since it discharges all the available batteries, there is no battery selection problem.

We use AC power delivery because it is most common in today's data centers and existing systems can apply our design without new infrastructure cost. Despite the power losses associated with the centralized placement, we still use it because of its simplicity and low maintenance requirements. We address this problem by adding 8% (see section 5.6.5 for more details) more batteries into our architecture and compensating the additional capacity cost with elevated battery life. Furthermore, our design can leverage a dedicated network to establish coordination among the batteries, instead of being dependent on the data center network, reducing the communication overhead.

We compare our grid-tie based design against SoA designs in Table 5.5 in terms of the key architectural challenges we describe in section 5.1. Our design leverages the useful properties of existing designs that are necessary to shave long peaks. We add the ability to adjust the discharging current adaptively and a detailed battery model to capture the effects of a high discharge current. Also, our design can facilitate the locality of the batteries by using a dedicated network to establish the communication, instead of using the data center network.

5.5 Methodology

This section demonstrates the methodology we use to evaluate distributed battery control and grid-tie based battery placement architecture. It outlines the experimental setup, such as power and workload models along with the simulation environment. Furthermore, it also presents two cost models that are used by different peak power shaving goals, 1) reducing the peak power level to obtain

Table 5.5: Comparison between the grid-tie design and the state-of-the-art (SoA) designs

	SoA [62] Centralized	**SoA [73][97] Distributed**	**Grid-tie Design**
Placement	Centralized	Distributed	Centralized
Selective Battery Discharge	X	✓	✓
Adaptive Current	X	X	✓
Battery Model	Simple	Simple	Detailed
Coordination Medium	N/A	Data center network	Dedicated network

savings, 2) adding additional servers within the original peak power budget to increase the revenue.

5.5.1 Power Measurements and Workloads Run

Same as in chapters 2, 3 and 4, we use measurements from our data center container on campus to estimate the overall power cost for a larger scale data center. Our container has 200 servers consisting of Nehalem, Xeon and Sun Fire servers running Xen VM. We run a mix of commonly used benchmarks to measure power and performance of service and batch jobs on our servers. We use RUBiS [112] to model service-sensitive eBay-like workload with 90th percentile of response times at 150ms, and Olio [23] to model social networking workloads with response times ranging from 100ms up to multiple seconds, depending on the type of request (e.g. text post vs. video upload). Multiple Hadoop [65] instances are run as batch jobs. We measure performance at 10ms sampling rate and obtain power at 60Hz.

The measurements are used to create an event-based simulator that embeds the power information and the workload characteristics to simulate a larger data center environment. We model each 8-core server with an M/M/8 queuing model, and a linear CPU utilization based power estimate commonly used by oth-

Table 5.6: Workload parameters

Workload	Average Time	
	Service	**Inter-arrival**
Search [88]	50ms	42ms
Social Networking [23]	1sec	445ms
MapReduce [38]	2min	3.3min

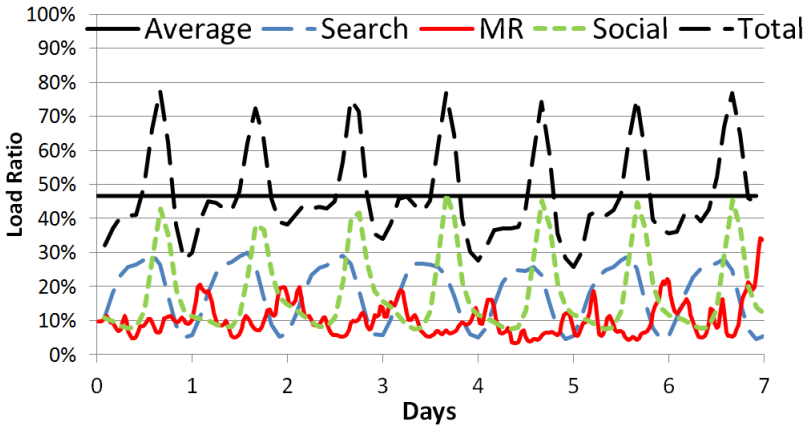


Figure 5.11: Data center workload mixture

ers [50][88]. In chapter 2, we show that the average simulation error is well below 10% for all quantities of interest.

To understand the benefits of peak power shaving, we model the typical user request load for a full data center. We use a year of publicly available traffic data of two Google products, Orkut and Search, as reported in Google Transparency Report [60]. A week's worth of workload combinations based on the waveform shown in Figure 3 of [38] where Social Networking and Search workloads represent service jobs, and MapReduce is for batch jobs. Table 5.6 shows the workload parameters, while Figure 5.11 compares each job's contribution to the total data center load. The maximum load ratio is around 80% with average of 45%.

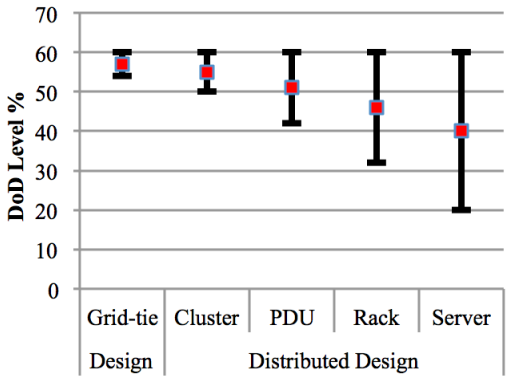
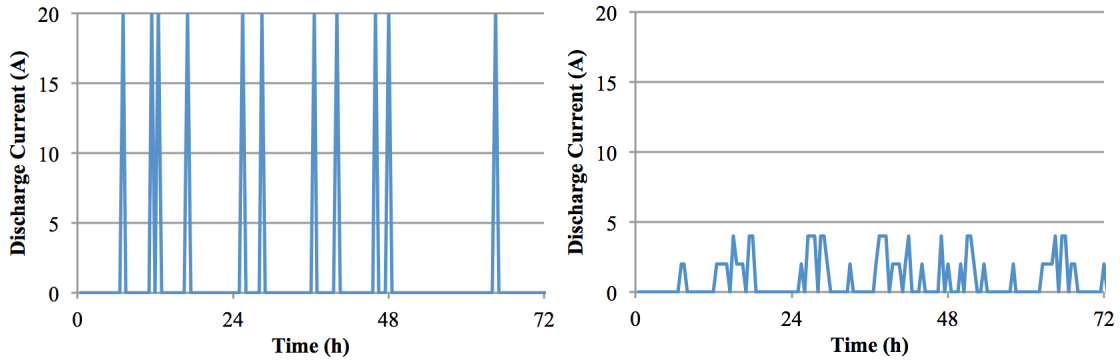


Figure 5.12: DoD level variation

5.5.2 Data center and Battery Simulation

We limit our data center simulation to a week because it is not computationally feasible over long periods due to fine event granularity. We extract the power profile of the data center as well as the charge/discharge profile of the batteries in the given time-frame and scale these profiles appropriately for longer time intervals. We refer to this process as data center workload simulation. The main goal of this pre-processing is to analyze the required DoD level and discharging current profiles for the batteries.

Figure 5.12 shows the DoD level variation of the grid-tie design and the distributed design with different level controllers over a week using LFP batteries when DoD_{goal} is set to 60%. Both designs shave 15% of the peak power. The grid-tie architecture is more consistent, followed by high level distributed controllers. In these cases, the batteries use all the available capacity, because the battery power is distributed evenly across the batteries. In contrast, the DoD value is distributed between 20% and 60% approximately uniformly with a server level controller since individual server power profiles vary and there is no coordination between them. In Figure 5.13, we present the average discharging current profile of the distributed and grid-tie design over a 3 day period from the same experiment described above. The grid-tie design reduces the discharging current significantly without affecting the amount of peak power shaved, and thus can decrease the negative effects of high discharging current.



(a) Distributed design

(b) Grid-tie design

Figure 5.13: Avg. discharging current for the distributed design and grid-tie design over a 3 day period, with LFP batteries

We include both LFP and LA batteries in our study and assume that the battery capacity per server is 40Ah and 20Ah respectively with 12V nominal voltage. These capacity values are the maximum that can fit into a 2U server [73]. Each battery is allowed to discharge up to DoD_{goal}. We change the DoD_{goal} to see how it impacts both average battery lifetime and peak power level that can be sustained. Our battery model estimates the SoC and SoH of each battery. After analyzing short-term battery usage profiles, we use our battery model and simulate only charge/discharge cycles of the batteries. We run the simulation for several years of simulation time to estimate the battery lifetime. We consider a battery dead when its SoH goes below 80% [103][111]. We refer to this process as battery simulation.

5.5.3 Cost Models

This section presents the cost models to quantify the benefits of the peak power shaving. For each different model, we show the domains they are applicable to, how they are calculated and specifically focus on how the battery cost affects the overall cost.

Co-location Rental (CLR) Cost Model

Co-location providers rent their data center equipment and space to retail customers. This applies to companies that require a data center-like system but do not want to build their own. A well-known example for a co-location renter is content delivery networks (CDNs) [97]. These renters make long-term power contracts with co-location providers and pay based on their provisioned power, instead of their actual consumed power. As a result, decreasing their peak power consumption immediately translates to savings (Figure 5.1b). Palasamudram et al. [97] target this domain for their distributed battery-based peak shaving design and calculate the total cost as:

$$Cost_{total} = c_p PP_{total} + \frac{c_b}{L} BC_{total} \tag{5.9}$$

where c_p is the unit power price, PP_{total} is the total provisioned power, c_b is the unit battery price, BC_{total} is the total battery capacity and L is the expected battery lifetime. Then, they calculate the savings as:

$$Savings = 100 \times \frac{Cost_{total}(nobatteries) - Cost_{total}(batteries)}{Cost_{total}(batteries)} \tag{5.10}$$

where $Cost_{total}(batteries)$ and $Cost_{total}(nobatteries)$ represent total cost with and without batteries, respectively. When calculating the total cost without the batteries, we can just neglect the battery related parts of Equation 5.9. The main purpose of peak shaving in this case is to reduce the provisioned power level so that the co-location renters can contract for less power.

Total Cost of Ownership (TCO) Model

There are several companies that own their data centers, where they still make power contracts based on their peak power consumption to reduce their cost of energy. However, they achieve this peak value rarely and under-utilize the provisioned power. A solution to this is to add more servers to the data center, which improves the power utilization but also increases the peak power level. A

peak shaving mechanism can ensure that the provisioned power level is not violated with additional servers. In this case, the provisioned power level does not decrease but both the provisioned power and the data center equipment can be used to host more servers and thus TCO/server reduces. Also, assuming that each server brings a constant amount of revenue, the total profit increases [73]. This also shows that the savings is directly proportional to TCO/server reduction.

This analysis is made by collecting the depreciation and opex data from the APCs commercial TCO calculator [24]. This model computes the TCO/server by dividing it into multiple parts, calculating each part separately and analyzing how each part changes with more servers within the same power budget. Table 5.7 summarizes the different components of TCO and shows the TCO breakdown for different designs. More servers decrease the TCO/server and increase the profit obtained from a server. We compare the TCO/server of each battery placement design in our study with the TCO/server of a data center which does not use batteries for peak shaving (base model). The part we are interested in TCO partitioning is the UPS depreciation, accounting for the battery costs. If the associated UPS depreciation cost is high, we can obtain negative savings compared to the base model. Some reasons for high UPS depreciation include short average battery lifetime (requires frequent replacements) or using an inappropriate battery type for peak shaving (low energy density, short service time, etc.). Table 5.8 lists the input values for both this model and CLR model.

Our grid-tie design requires more power distribution infrastructure than the distributed design because we keep transmitting power throughout the data center, even if the power is not drawn from the utility. For example, a 10MW data center may have 1MW worth of additional servers due to peak shaving. In our case, the extra power is provided from the UPS through the data center power infrastructure to the servers. In the distributed case, this extra power is not provided through the data center power infrastructure. Although all the servers are connected to the main power infrastructure, during a peak pulse some of them may disconnect themselves from the main power infrastructure and get power locally from the on-board UPS. Therefore, the provisioned power infrastructure is sufficient. This

Table 5.7: TCO/server breakdown for different designs. The components with different trends are highlighted

TCO Component	**w/o peak shaving**	**Distributed Design Breakdown**			**Grid-tie Design**	
		TCO/server trend with more servers	**Battery Model**		**TCO/server trend with more servers**	**Breakdown**
			Simple	**Detailed**		
Facility space depreciation	$3.40	Decreasing	$2.74		Decreasing	$2.72
UPS depreciation	$0.13	Constant	$1.67	$5.00	Constant	$3.33
Power infrastructure depreciation	$5.94	Decreasing	$4.79	$4.79	Constant	$5.94
Cooling infrastructure depreciation	$2.46	Decreasing	$1.98		Decreasing	$1.96
Racks, monitoring, installation	$8.97	Decreasing	$7.23		Decreasing	$7.17
Data center opex	$7.49	Decreasing	$6.04		Decreasing	$5.99
Server depreciation	$31.25	Constant	$31.25		Constant	$31.25
Server opex	$1.56	Constant	$1.56		Constant	$1.56
PUE overhead	$1.94	Constant	$1.94		Constant	$1.94
Utility monthly energy cost	$8.71	Constant	$8.71		Constant	$8.71
Utility monthly power cost	$4.20	Decreasing	$3.39		Decreasing	$3.36
Total	$76.04	Decreasing	$71.30	$74.63	Decreasing	$73.94

Table 5.8: Input parameters for the cost models

Input	LA value	LFP value
Battery unit price - rated with 20h	$2/Ah [92]	$5/Ah [25]
Per server capacity	20 Ah [73]	40 Ah [73]
Peukert's exponent	1.15 [66]	1.05 [66]
Battery nominal voltage	12V [61]	
Data center depreciation time	10 years [28]	
Server depreciation time	4 years [28]	
Utility energy price	4.7¢/kWh [35]	
Utility power price	12$/kW [62]	

means that our approach has constant power infrastructure depreciation, whereas the distributed design decreases this depreciation with more servers. But, our design does not require a custom PSU or power distribution, as opposed to the DC architecture. This makes it practical for the existing data centers. The additional peak shaving opportunities with the grid-tie design outweigh the additional infrastructure costs.

5.6 Results

This section first presents the accuracy results of the detailed battery model and how it affects the savings obtained by battery-based peak power shaving using state-of-the-art designs. It then shows the effectiveness of the distributed battery control mechanism presented in section 5.3 and the grid-tie based battery placement design shown in section 5.4.

5.6.1 Accuracy of the Detailed Battery Model

We start our evaluation by comparing the power capping capabilities of the state-of-the-art (SoA) battery placement designs with both LA and LFP batteries. The SoA centralized design adjusts the battery capacity to handle only emergency

cases, which last only a few minutes. We assume that this design has a 3200 Ah LA battery as proposed in [61][73] to support a single data center container. In distributed case, each server has a dedicated 20Ah LA or 40Ah LFP battery, the maximum possible given their volume, same as in [73]. These battery capacities are adjusted to match previous work. In Table 5.9, we compute how long the batteries can shave a fixed average peak power pulse per server with specified magnitude where the data center power cap is defined at 255W/server. We first apply the simplistic battery model used by recent existing studies. This model accounts only for the total battery capacity and ignores the effects of high discharge currents and nonlinear behavior of different battery types [97][62]. Table 5.9 shows that the centralized design can shave a peak for only 7 minutes whereas the distributed design can successfully shave peaks of over 3 and 6 hours with LA and LFP batteries, respectively.

Next, we use the detailed battery model presented in section 5.2 to account for the battery type and the negative effects of high discharging currents. The peak power shaving amount can be overestimated by 133% in the centralized design. The discharging current in the distributed design is still high, but the rate of the discharging current is lower relative to total battery capacity. This results in error of 64% for LA batteries, and 14% for LFP. LFP's error rate is up to 4.5x lower than the LA's because of its more linear discharge behavior. However, the error, a result of an inaccurate model and interaction with physical devices the batteries, is still significant to affect peak power shaving decisions, such as determining battery design or the total needed capacity.

We use this detailed battery model with our long term battery simulation to estimate the average lifetime of an LA and LFP battery when shaving peak power. Table 5.10 compares our long-term battery lifetime estimates with previous work. Neglecting the effects of high current results in high error: as much as 210% and 240% longer battery lifetime estimates leading to severely underestimated battery costs and overstated cost savings due to peak shaving.

Table 5.9: Peak shaving capabilities of the square peak with centralized and distributed designs

Peak power/ server (W) - shaving (%)	**Centralized - 3200Ah total capacity LA**			**Distributed - 20Ah/server LA - 40Ah/server LFP**					
	Power Capping Durtaion (min)		**Error (%)**	**Power Capping Duration (min)**				**Error (%)**	
	Model			**Simple Model**		**Detailed Model**		**LA**	**LFP**
	Simple	**Detailed**		**LA**	**LFP**	**LA**	**LFP**		
300 - 15%	8	4	100%	240	480	148	423	62%	13%
310 - 17%	7	3	133%	186	372	114	327	63%	14%
320 - 20%	7	3	133%	135	270	82	237	64%	14%

Table 5.10: Battery lifetime estimation comparison

	LA	**LFP**
Low current rated estimations	3 years	10 years
Detailed model estimations	1.4 years	4.1 years

5.6.2 Effects of Detailed Battery Model on Savings

Inaccuracies in battery lifetime estimation may lead to underestimated battery costs and overestimated cost savings. Table 5.11 and 5.12 show the CLR and TCO/server savings, for both LA and LFP batteries with varying battery lifetime values. We obtain 9.5% and 19% peak power capping with distributed LA and LFP batteries, respectively. This peak shaving also enables 10.5% and 24% extra servers to be deployed within the same power budget when using the TCO model. We first use inexpensive batteries rated at low currents. In this case, CLR savings are 2.8% and 6.4%; TCO/server savings are 0.9% and 1.86% for LA and LFP batteries, respectively. If we do not capture the effects of high discharge current, these savings are 6.4% and 15.15% for CLR model and 2.65% and 6.24% for the TCO model. The savings are overestimated by up to 2.94x and 3.35x for LA and LFP.

Next, we use batteries with larger rated currents: 10h, 5h and 1h [39], that are also more expensive: 8, 10, 12$/Ah for LFP and 3, 4, 5$/Ah for LA. The average LFP lifetime increases to 5, 6 and 8 years and 2, 2.5 and 3 years for LA. Table 5.11 shows that CLR cost savings become 3.2%, 2.7%, 1% for LA and 5.5%, 1.9%, -1.8% for LFP batteries with 10h, 5h and 1h rated batteries. Similarly, Table 5.12 shows that TCO/server savings become 1.2%, 0.94%, 0.28% for LA and 1.42%, -0.33% and -2.08% for LFP with 10h, 5h and 1h rated batteries respectively. Although the battery lifetime values are closer to the low current rated estimates, higher battery price overshadows the savings obtained by fewer battery replacements.

5.6.3 Peak Shaving Efficiency of State-of-the-Art Designs

We continue our evaluation by comparing the peak shaving capabilities of SoA battery placement designs. We also include our battery model to account for the high discharge currents. Most data centers use a centralized LA battery, powering the entire data center when it is active and not over=provisioned for peak shaving. The capacity of this battery is adjusted to handle only emergency cases, which last a few minutes. We assume that this design has 3200 Ah worth of

Table 5.11: CLR cost savings for distributed LA and LFP batteries

LA - distributed design					**LFP - distributed design**				
Cost	**$2/Ah**	**$3/Ah**	**$4/Ah**	**$5/Ah**	**Cost**	**$5/Ah**	**$8/Ah**	**$10/Ah**	**$12/Ah**
Lifetime	**CLR savings (%)**				**Lifetime**	**CLR savings (%)**			
1	0.9	-3.6	-8.1	-12.7	1	-25.3	<-50	<-50	<-50
2	5.5	3.2	0.9	-1.3	2	-2.7	-16.3	-25.3	-34
3	6.4	4.6	2.7	1	3	1.8	-9.4	-17	-24
4	8.2	5.9	4.6	3.2	4	6.4	-1.8	-7.2	-13
7	Not possible				7	11	5.5	1.9	-1.8
10					10	15.5	12.7	11	9.1

Table 5.12: TCO/server savings for distributed LA and LFP batteries

LA - distributed design					**LFP - distributed design**				
Cost	**$2/Ah**	**$3/Ah**	**$4/Ah**	**$5/Ah**	**Cost**	**$5/Ah**	**$8/Ah**	**$10/Ah**	**$12/Ah**
Lifetime	**TCO/server savings (%)**				**Lifetime**	**CLR savings (%)**			
1	0.02	-2.08	-4.18	-6.27	1	-13.48	-26.63	-35.4	-44.17
2	2.21	1.2	0.2	-0.81	2	-2.52	-9.1	-13.48	-17.87
3	2.65	1.86	1.07	0.28	3	-0.33	-5.59	-9.1	-12.61
4	3.09	2.52	1.95	1.38	4	1.86	-2.08	-4.71	-7.35
7	Not possible				7	4.05	1.42	-0.33	-2.08
10					10	6.24	4.93	4.05	3.18

Table 5.13: Centralized design peak shaving capabilities with different battery types. $P_{threshold}$ is set to 255W/server

Peak Power per Server (W) - Shaving(%)	Power Shaving Duration (min)		
	Centralized - LA not scaled	Centralized - LA scaled	Centralized - LFP scaled
300 - 15%	3.8	24.2	70.5
310 - 17.5%	3.7	23.3	68.1
320 - 20.3%	3.5	22.5	65.8

LA batteries [61][73]. Then, we compute the amount of time a battery can shave a fixed peak pulse and how long it takes to fully recharge it during low demand. Table 5.13 shows that the centralized design shaves the peak power for only 3-4 minutes when $P_{threshold}$ is set to 255W per server. It cannot sustain long peaks and needs to apply other policies such as DVFS which have performance overhead.

To address this problem, we increase the capacity of the centralized battery by 5x and obtain 6x longer peak shaving. LA batteries have large volume, so the capacity cannot be scaled significantly. The increase in peak shaving duration is more than 5x because the stress on discharging current rate decreases non-linearly as a result of Peukert's law [116]. In contrast, the recharging duration increases almost linearly with scaling capacity. The peak shaving duration, despite increased capacity, is still not sufficient enough to sustain peaks lasting hours. Another option is to use LFP batteries with more energy density and less nonlinear battery behavior. This can scale up the capacity further. We use a total capacity of 40K Ah [73] and get up to 70 minutes of peak shaving at high cost. The peak shaving benefits are insufficient to compensate for high battery costs. This analysis shows that centralized battery design is not a good option for peak shaving when the battery powers the entire data center in boolean fashion as in the state-of-the-art work.

The distributed design allows battery power to be controlled in finer granularity by selectively discharging only a subset of all the batteries. We analyze the power shaving duration of distributed LFP and LA batteries in Table 5.14. The

Table 5.14: Peak shaving and battery recharging comparison of the distributed design with different battery types and AC vs. DC power options. $P_{threshold}$ is set to 255W/server

Peak Power per Server (W) - Shaving(%)	**Power Shaving Duration (min)**		
	Distributed - LA with AC	**Distributed - LFP with AC**	**Distributed - LFP with DC**
300 - 15%	192.9	552.2	552.2
310 - 17.5%	157.1	451.1	451.1
320 - 20.3%	132.3	381.1	381.1
Low Power per Server (W)	**Recharging Duration (h)**		
	Distributed - LA with AC	**Distributed - LFP with AC**	**Distributed - LFP with DC**
220	8.5	17	14.8
210	6.6	13.2	11.6
200	5.4	10.8	9.4

size of each LA and LFP battery is set to be 20Ah and 40Ah, respectively. These are the maximum capacities that can fit in a 2U server [73]. Although the LFP capacity is more than LA by 2x, it shaves a given peak for 3x longer because LFP battery behavior is less nonlinear at high current, proving to be a better fit for the distributed design. But, recharging all the batteries back to back takes more time for LFP due to its larger capacity. Since batteries can selectively discharge, this is not much of an issue.

Another important key challenge is to reduce the conversion losses that impact the effective battery input/output. The distributed design puts the batteries next to the servers and increases the effective battery capacity compared to the centralized design. DC power delivery can be used to further eliminate the conversion losses on the power path, reducing the input power required to recharge the battery. We show the best and common efficiency values for the power infrastructure of both AC and DC options in Table 5.15. It also shows the amount of energy wasted to recharge the batteries and battery output wasted before going

into the servers.

The centralized design does not waste a lot of grid power but the battery output loss is 15%, which further reduces its peak shaving duration. We see that the distributed DC design obtains the best efficiency by having the smallest total conversion losses. The AC counterpart provides similar battery output power but it wastes the grid input 3x more than the DC design and results in longer discharges. Table 5.14 also shows the comparison between AC and DC distributed options in terms of effective discharge and recharge durations. Discharging capabilities are the same but the DC design takes 14% shorter time to fully recharge, which makes it a safer option as batteries get ready for the next peak earlier. Although the DC option is more energy efficient, it is an unfeasible option for existing data centers because its high cost to replace the power infrastructure.

5.6.4 Performance of the Distributed Battery Control

We next evaluate the performance of the communication based distributed controllers, which increase the overall battery lifetime by balancing the power demand across the batteries. We use 1000 40Ah LFP batteries with configurations shown in Table 5.3, with policies described in Table 5.4. Tables 5.16 and 5.17 summarize the comparison between different policies and group sizes in terms of peak shaving and average battery lifetime. To calculate the best peak power shaving for each configuration we first use the workload distribution shown in Figure 5.11 to create the power profile of the data center over a week. We initially set a power cap, e.g. 280W/server, and reduce it in each simulation experiment until we cannot guarantee that cap. We then compute the power shaving percentage with the amount of power shaved over the peak.

Table 5.16 shows energy savings per week due to various peak power shaving strategies scaled to a data center of peak capacity 10MW, along with peak power shaving percentages for each configuration based on the smallest power cap we can guarantee. Googles 10MW, 45 container data center, with 40000 servers [61] is an example of such a deployment. The best peak power shaving can be achieved with a centralized controller as much as 19% of the peak power of the entire data center,

Table 5.15: Efficiency of centralized vs. distributed designs with different power equipment and delivery options

Unit	**Efficiency Common**	**Efficiency Best**	**Design**	**% Battery energy wasted before providing server power**		**% Grid energy wasted before charging the batteries**	
				Common	**Best**	**Common**	**Best**
Centralized double conversion UPS	85% [117]	90% [53]	Distributed w/ AC power	5%	2%	35%	23%
AC distribution PDU	98% [53]						
Server AC PSU	75% [41]	90% [53]	Distributed w/ DC power	2%	2%	12%	8%
DC UPS	92% [107][117]	95% [53]					
DC distribution PDU	99% [53]						
Server DC PSU	92% [53]		SoA Centralized	38%	15%	5%	5%
Filter + Rectifier	95% [53]	97% [102]					
Grid-tie inverter	95% [96]						

Table 5.16: Amount of energy shaved for a 10MW data center per week in MWh and % of power shaved compared to the peak

Policies	**Data Center Partitioning**				
	1 container	**5 PDUs**	**10 PDUs**	**50 Racks**	**1000 servers**
Local	30 (19%)	14.3 (16%)	11.2 (15%)	4.8 (12%)	2.5 (10%)
Max-SoH-global	30 (19%)	30 (19%)	30 (19%)	30 (19%)	30 (19%)
Max-SoH-lim.comm.	30 (19%)	23.1 (18%)	14.3 (16%)	6.6 (13%)	2.5 (10%)
Max-SoH-more-lim.comm.	30 (19%)	18.1 (17%)	11.2 (15%)	4.8 (12%)	2.5 (10%)

equivalent to 30MWh/week of the 10MW data center, or 9380 more servers with no additional peak power cost. Although we have the same total battery capacity in all of the configurations, the power shaving capability decreases significantly with lower level controllers because of their limited knowledge of the total power demand. They shave up to 50% less power and 92% less energy compared to the best solution. In contrast, we observe that our PDU level controllers with communication can shave 18% of the peak power and 23MWh energy, within 6% and 23% of the centralized solution.

Table 5.17 shows the average battery lifetime, normalized to the case with the individual server level controllers. Local policies perform poorly regardless of their battery selection algorithm as they are unaware of batteries in other groups. Changing the group size does not affect performance of the local policies, except for *Max-SoH*, which reduces to *Max-SoH-global* when there is only one group. The centralized controller gives the best results, performing 2x better than the local policies by processing the data from all the batteries. The performance of policies with limited communication depends on the group size and communication span. Increasing span with 5 PDU level controllers using limited communication by one

Table 5.17: Normalized average battery lifetime

Policies	**Data Center Partitioning**				
	1 container	**5 PDUs**	**10 PDUs**	**50 Racks**	**1000 servers**
Local	1.02	1.03	1.04	1.04	1.00
LRU	1.07	1.07	1.07	1.07	1.00
Max-SoH-local	1.97	1.07	1.07	1.07	1.00
Max-SoH-global	1.97	1.97	1.97	1.97	1.97
Max-SoH-lim. comm.	1.97	1.91	1.76	1.77	1.73
Max-SoH-more-lim. comm.	1.97	1.59	1.59	1.51	1.48

group results in up to 20% longer battery lifetime, within 3.3% of the centralized solution. Thus, our distributed controllers well approximate the performance of the centralized controller in terms of both power shaving and battery lifetime, showing that intelligent control and good characterization of data center's physical infrastructure can dramatically improve the overall system efficiency.

Communication Overhead Analysis

In this architecture, each group controller polls the servers in its group using the data center network to collect server power consumption and battery statistics. The controller then delivers the battery selection decision to the servers. Intra-rack communication is extremely fast, but relaying messages through multiple switches introduces far more delays. Assuming a common a fat-tree topology [72], we model the links in the network with 10 Gbps capacity, which can transmit a 1K package at 1us. We evaluate an ideal network, without queuing delay, a network with normal level congestion where a single message transmission delay in a switch is 50us and a network with a high level congestion reaching 350us delay [40]. In this experiment, container level models global communication.

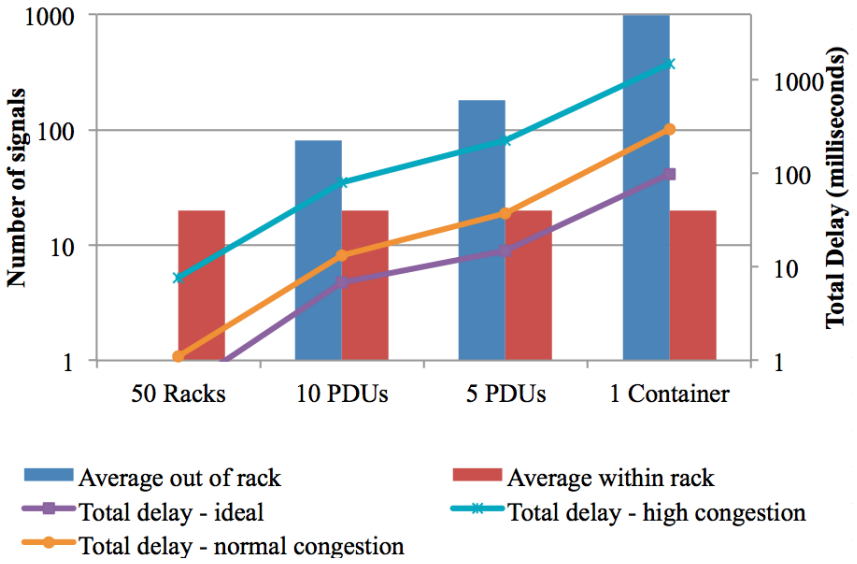


Figure 5.14: Communication overhead analysis for the distributed control mechanisms

Figure 5.14 shows the results of the communication analysis. The vertical axes are on a log scale. The total delay increases exponentially with higher level controllers because of the increasing number of out-of-rack communication signals, going over several hops. Rack level controller gives the best results with only tens of ms total delay even in the presence of high congestion. However, it has 32% less power shaving and 11% shorter battery lifetime compared to the centralized solution. In contrast, the container level controller may have seconds of delay, 100x more than the rack level in high congestion. With 5 PDU controllers there is a 10x decrease in the total communication delay relative to the global solution while being within 6% and 3.3% of the centralized controller in terms of peak power shaving and battery lifetime. Clearly this is a great replacement for the centralized control for peak power shaving with batteries.

5.6.5 Performance of the Grid-tie Design

We compare our grid-tie design with existing designs in terms of energy efficiency, average battery lifetime, cost savings, and communication overhead. As we place the batteries in a centralized location, we still lose 15% of battery

output because of the conversion losses (see Table 5.15). However, batteries are used at lower discharge current and have higher effective battery capacity. This reduces the effects of the conversion losses. Instead of 15% performance difference, we get an average of 8% performance loss compared to the distributed design as shown in Table 5.18. We compensate for this performance loss by adding 8% more battery capacity, which is feasible because we are not limited by rack size as in the distributed design.

Table 5.19 shows the power shaving statistics of our grid-tie design and the distributed design. We analyze our design with and without additional battery capacity as well as with *all_battery* and *discrete_current* policies (see section 5.4). The average battery lifetime does not change with additional battery capacity, but the *all_battery* policy results in longer average battery lifetime. The average battery lifetime estimates are 5.4 and 2.2 years for LFP and LA respectively using the *discrete_current* policy. We obtain 6.4 and 2.5 years with the *all_battery* policy. The battery lifetime values are 60% and 78% higher compared to the distributed design for LFP and LA batteries respectively since the discharging current can be scaled down with our design so that the negative impact on the battery lifetime is minimized. The *all_battery* policy scales down the discharging current more by using all available batteries and thus performs better than the *discrete_current* policy.

Our grid-tie design with 8% larger capacity obtains similar peak shaving performance compared to the distributed design. It compensates the increased battery costs with longer battery lifetime. Our design achieves up to 11% and 5.5% savings for LFP and LA batteries when renting from co-location providers. These savings are 70% and 100% higher than the distributed design. Similarly, we obtain up to 2.77% and 1.87% TCO/server savings using LFP and LA respectively. These TCO/server savings correspond to up to $75K/month for a 10MW data center [24]. The TCO savings are 48% and 107% higher than the savings of the distributed design.

Table 5.18: Peak shaving capabilities of the grid-tie design compared to the distributed design. $P_{threshold}$ is set to 255W per server

Peak Power Per Server (W) Shaving %	**Peak Power Shaving Duration (Min)**	
	Distributed - LFP	**Grid-tie - LFP**
300 - 15%	552	516
310 - 17.5%	451	418
320 - 20.3%	381	351

Table 5.19: Grid-tie vs. distributed design. EB=Extra Batteries, BL=Battery Lifetime, PS=Peak Shaving, ES=Extra Servers

Design - Policy	**EB**	**LFP**					**LA**				
		BL	**PS**	**ES**	**CLR Savings**	**TCO/server Savings**	**BL**	**PS**	**ES**	**CLR Savings**	**TCO/server Savings**
Grid-tie	8%	6.4	20%	25%	11%	2.77%	2.5	9.9%	11%	5.5%	1.87%
all_battery	0%	yrs	16%	19%	7.7%	1.36%	yrs	7%	8%	2.8%	1.14%
Grid-tie	8%	5.4	20%	25%	9.4%	2.77%	2.2	9.9%	11%	4.9%	1.44%
discrete_current	0%	yrs	16%	19%	6.1%	1.36%	yrs	7%	8%	2.2%	0.42%
Distributed - N/A	N/A	4 yrs	19%	24%	6.4%	1.86%	1.4 yrs	9.5%	10.5%	2.7%	0.9%

Communication Overhead Analysis

The distributed design requires a centralized controller to get the best peak shaving performance [97][73]. Since the batteries are distributed to the servers, this controller communicates with the batteries through the data center interconnect. High network usage leads to large signal delays to/from batteries. This can affect the performance of the controller negatively by increasing the response time to a peak pulse or transmitting outdated battery and server load information. The distributed design can also use multiple controllers placed at different levels of power hierarchy, as shown in previous subsection. A decentralized control mechanism may, however, significantly reduce the peak shaving capabilities. Our design can isolate itself from the data center interconnect, achieving fast communication even with high network congestion.

Figure 5.15 compares the total delay of our grid-tie design during a discharge process with that of different controllers deployed in distributed design. We analyze the worst-case scenario where the controller needs to poll each battery. The left vertical axis is on a log scale and shows the communication delay whereas the right vertical axis presents the peak shaving percentage achieved by each configuration. The data center interconnect assumptions are the same as in the previous subsection. In this experiment, cluster level corresponds to centralized communication for the distributed design. The low-level controllers have less total delay compared to our design in the ideal network case, but as the network congestion increases, our design performs better, except for the rack level controller, which has 60% less peak shaving performance than our design. Our design has similar peak shaving performance (1% better) compared to the centralized control in distributed design. But, even in the ideal case of network, our design has around 20 ms total delay compared to 100 ms of the centralized control for the distributed design. Even in this case, we obtain 4x less communication overhead, and this difference increases exponentially as the network delay ramps up.

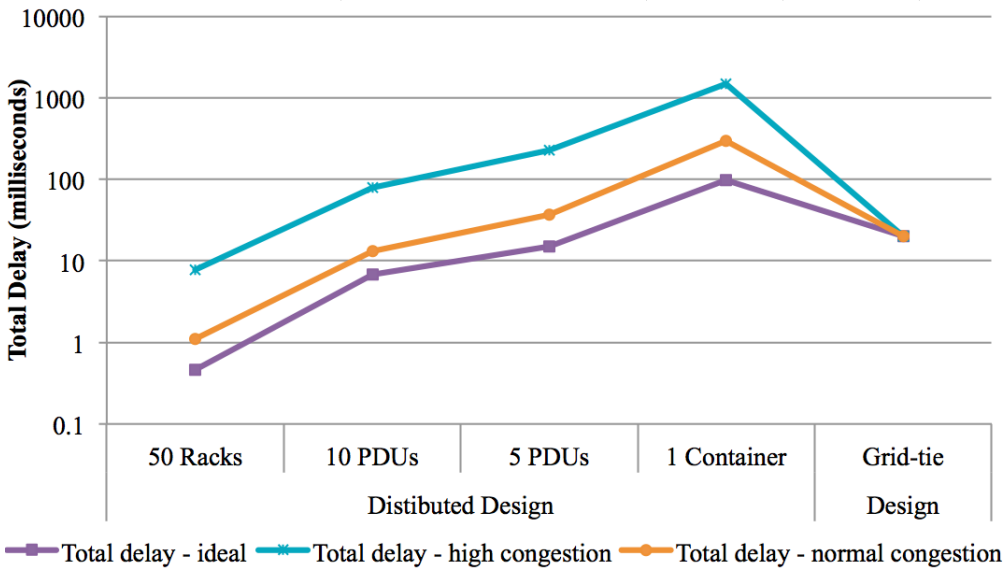


Figure 5.15: Communication overhead analysis for the grid-tie design

5.7 Conclusion

Peak power shaving with batteries in data centers has gained significant importance because of its ease of applicability and great performance. In this chapter, we first identify the issues with the existing designs and address the key challenges of architecting a cost and energy efficient battery-based peak shaving design. We first use a detailed battery model to capture the negative effects of high discharging currents. Our results indicates that not having a detailed battery model overestimates the battery lifetime up to 2.44x and leads to 3.35x error in cost saving estimates. Second, we propose a distributed control mechanism to manage the physical properties of the batteries. Our mechanism removes the single point of failure of the traditional centralized control and reduces its communication overhead by 10x while being within 6% and 3.3% of its peak power shaving and battery lifetime, respectively. This power shaving leads to 23.1 MWh energy shaving when scaled to a typical 10MW data center. Third, we introduce a new grid-tie based design which preserves the advantages of the existing designs, such as individual control of the batteries, and eliminates the key drawbacks, such as adaptively selecting the discharge current. It can use a fast, dedicated network to

coordinate the batteries, reducing the communication overhead by 4x compared to the distributed design. Our design achieves up to 78% longer battery lifetime and doubles the savings compared to the state-of-the-art designs. This chapter illustrates the benefits of correctly modeling and tracking the physical phenomena (batteries). Thus, designing an appropriate infrastructure to manage the batteries is critical for obtaining great results.

Chapter 5 contains material from "Distributed Battery Control for Peak Power Shaving in Data Centers", by Baris Aksanli, Tajana Rosing and Eddie Pettis, which appears in Proceedings of International Green Computing Conference (IGCC), 2013 [16]. The dissertation author was the primary investigator and author of this paper.

Chapter 5 contains material from "Architecting Efficient Peak Power Shaving Using Batteries in Data Centers", by Baris Aksanli, Eddie Pettis, and Tajana Rosing, which appears in Proceedings of International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS), 2013 [13]. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Data Centers & the Grid

The conventional, centralized electric grid has been evolving into a more distributed structure with the increasing penetration of renewable energy sources, wide usage of electric vehicles and increasing number of large-scale smart buildings. These changes are making it harder for the grid to preserve its stability. The grid has to maintain its supply/demand balance at any time to avoid voltage/frequency deviations, which may harm the stability of the grid and hence prevent the grid operations from being performed normally [12]. Utility companies employ ancillary services, that help preserve the stability. These services include demand-response (DR), spinning and non-spinning reserves and regulation services. DR is designed to motivate customers to voluntarily reduce their energy consumption at times of high overall demand or when the system reliability is endangered. Utilities usually increase the price of electricity to incentivize their customers. Spinning and non-spinning reserves provide electricity when the grid unexpectedly needs more power on a very short notice. They include explicit contracts with electricity providers and thus, need to promptly answer to the notifications from the utility. Regulation service is used to correct short-term changes in electricity use that affect the the power system stability. This chapter focuses on contract-based regulation services, which are used to balance the demand and supply. For example, at a time when there is high renewable energy generation, the utility might need higher power demand from consumers. Similarly, it might request users to reduce their consumption when the power generation is at a premium. When needed, the

utility issues fine-grained command signals to the contracted loads, which then adjust their power consumption accordingly.

Traditionally, utilities accept only generator sources to provide regulation services. Recently, they have also allowed non-generator sources to participate [34]. These sources should have large consumption and some power flexibility to allow adjustments. Power consumption of data centers is growing rapidly, up to 100MW per individual site [42]. The data center's ability to adjust the power consumption at run time by employing techniques such as dynamic voltage-frequency scaling (DVFS), virtual machine (VM) migration and peak power shaving make them a good choice for regulation services. Peak power shaving decreases the peak power level a data center achieves over a time, thus reduces its contribution to the monthly utility bill.

While both DVFS and VM migration have some performance overhead, some recently proposed battery-based peak shaving techniques [62][73][13] are capable of reducing the power consumption at no cost to performance. Significant savings, of up to \$75K/month for a 10MW data center, can be obtained when using batteries for peak power shaving. However, none of the existing battery-based peak power shaving designs consider the feasibility of using the energy storage in data centers to participate in the regulation markets. This is one key contribution of this chapter.

There are a few recent studies that investigate the data center participation in the ancillary services market [10][37][55][126]. The savings a data center can obtain depend on the type of the ancillary service provided and how much capacity can be allocated towards providing that service. Wang et al. [126] model a distributed set of data centers and explore DR. They use VM migration among data centers to create flexibility in power consumption and send VMs to locations with lower energy costs. Ghamkhari et al. [55] analyze the savings of a single data center participating in voluntary load reduction. They increase the waiting times of the workloads to reduce the consumption when necessary. Chen et al. [37] specifically focuses on regulation service for data centers and explore the ability of data centers to provide this service. They use dynamic power capping techniques

to follow the signals of the utility and provide the service at the expense of performance degradation. Aikema et al. [10] analyze a number of different ancillary services for data centers, including DR, spinning reserves and regulation. However, their analysis is also based on slowing down the workload. They assume that the nominal data center consumption can be all considered as the regulation capacity, which can result in dramatic performance penalties.

Existing methods result in performance degradation, which is a serious concern for response-time critical workloads. They also do not consider peak power costs. When providing regulation services, the utility is given the prerogative to demand a change in the power consumption of the data center by as much as a maximum amount specified in the contract over the given interval. This amount, if not properly handled, may raise the peak power level of the data center, and increase the utility bill. The data center should adjust its average power demand and the regulation capacity to be allocated to ensure that the peak power costs do not eliminate the savings from providing regulation services.

In this chapter, we propose a framework that analyzes the data center participation in the regulation markets while also considering the peak power objectives. We use a battery-based peak power shaving design to avoid performance penalty to workloads. We study two most common battery types, lead-acid (LA) and lithium iron phosphate (LFP), as in chapter 5. Our framework consists of two cases corresponding to different peak power assumptions for a data center. We present a method for each case, which first analyzes if providing regulation services is feasible and if so, shows how the regulation capacity should be adjusted to maximize savings. Normal peak shaving takes place to limit peak power costs if the data center chooses not to participate. We leverage the data from NY-ISO and CAISO markets to demonstrate the effectiveness of our framework. Our results show that for a 21MW data center, up to $480,000/year savings can be obtained using our methods, corresponding to 1280 more servers operating, and 5.08% increase in data center profit percentage. Also, if peak power costs are not considered when providing regulation services, its savings can be overestimated as high as 385%.

This chapter first gives background information on how participation in regulation markets takes place. It analyzes how data centers can participate in these markets in an effective way. The participation is analyzed in different cases to consider all possible scenarios. It then presents results of the proposed framework using realistic utility market dynamics.

6.1 Background

Providing regulation services requires the agreement of both the regulation service provider, in our case data centers, and the independent system operator (ISO), which provides the electricity to end-users. The data center determines the capacity of the regulation service it is willing to provide based on the price bid of the utility. Determining this service can take place in real-time, hour-ahead or day-ahead markets, which all have different pricing schemes and requirements in terms of how long the service should be provided.

In a given service interval, the regulation service provider should determine its average power demand, P_{ave}, and the regulation capacity, C_{reg}, it can provision in that interval, and give this information to the utility. By giving this information, the regulation service provider agrees that the utility can issue fine grained signals that can change the power consumption of the service provider to any value within the interval $[P_{ave} - C_{reg}, P_{ave} + C_{reg}]$. Within this interval, the average power demand of the service provider is P_{ave} [37]. For data centers, P_{ave} depends on the resource utilization and it is typically around 50% [28] and C_{reg} changes based on the load flexibility [37]. Some previous studies, e.g. [10], incorrectly assume that the power demand of a data center providing regulation service does not change within the interval in which the service is provided. When power demand is set to a value in $[P_{ave} - C_{reg}, P_{ave}]$ due to the utility feedback, one of the power shaving methods can be used. For example, DVFS-based solutions [88] reduce the peak at the expense of lower performance. Alternatively, batteries can be used to shave power [62][73][13] to avoid performance impact at cost of additional hardware. Utility may also demand the power be increased from P_{ave} to a value in

$[P_{ave}, P_{ave} + C_{reg}]$ to balance its larger energy supply and smaller demand. In this case there may be a conflict between a need to keep peak power under a predefined threshold due to electricity pricing vs. the need to respond to utility controls as a part of the regulation services contract.

In this chapter, we assume that the data center already uses peak power shaving methods to reduce the related costs in the monthly electricity bill. We assume that the cost of electricity (¢/kWh) is constant for a given day. To avoid the performance degradation of traditional power shaving methods, we leverage the battery-based peak power shaving method described in section 5.4. This method allows for very fine-grained battery output control, leads to smaller discharging currents and longer battery lifetime, and has smaller communication overhead compared to the other designs. Fast response times are needed to ensure that the data center receives the best prices for the regulation services it provides [37]. We do not interfere with any of the jobs running and instead control the battery output to track the differences between the actual and the targeted power demand.

As shown in chapter 5, in battery based peak power shaving, the data center first determines a fixed peak power threshold, P_{th}, and then discharges the batteries when the actual demand is over that threshold and recharges the batteries during lower demand. The physical properties of the batteries influence the choice of this threshold. Chapter 5 shows that if the batteries are discharged deeply or with high discharging currents, their expected lifetime decreases. Hence, first a fixed limit for battery depth-of-discharge (DoD) is found that is economically feasible, then the peak power limit is estimated based on this DoD limit. Typical DoD limit values are 20-40% for LA batteries and 60% for LFP batteries [73].

Figure 6.1 shows the power shaving process for a data center with 50,000 servers (21MW peak capacity) over 7 days, using 40Ah/server LFP batteries. The y-axis shows the power consumption in MWs and the x-axis corresponds to time in hours. The straight line is the nominal power demand of the data center, while the dashed line shows the best case power demand of the data center from the utility with 60% DoD goal. The dashed line shaves 20.5% of peak power compared to the original peak, shown by the upper horizontal line. In this particular case,

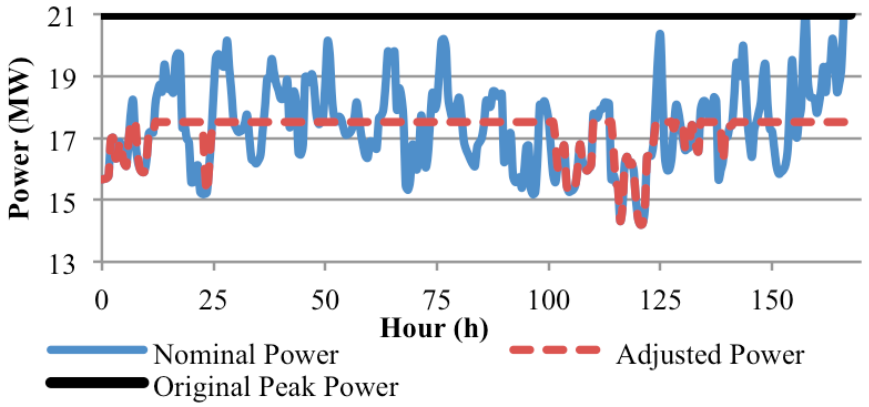


Figure 6.1: Sample battery-based peak power shaving demonstration of a 21MW data center over 7 days

it is not possible to lower this threshold without allowing batteries to discharge deeper. The difference between the original peak power level and the adjusted peak power level corresponds to savings. Peak power level can be further reduced but the battery DoD limit has to be increased. In that case, the expected battery lifetime reduces and the battery replacement costs become larger than the peak power shaving savings.

6.2 Data Centers Providing Regulation Services

This section gives a detailed feasibility analysis for the participation of data centers in a regulation market. Data centers that leverage batteries to shave peak power can respond to utility commands for regulation services by changing the battery charge and discharge intervals, thus requiring only minor changes to the already implemented battery control system. We analyze two different types of data center operation where both peak power shaving and regulation service controls are present. The first case does not alter the average data center power demand, but increases the peak power threshold to match the allotted regulation capacity. When participating in the regulation markets, we specifically show how the decision mechanism should be designed to consider peak power costs. The other solution does not change the fixed power threshold but create flexibility in

data center power consumption by adjusting the access to the batteries. This is the key point not to degrade workload performance. We show that this is essential to obtaining savings in a conservative market, like CAISO, where peak power costs are higher than the regulation prices.

6.2.1 Fixed Average Power

Careful control of data center batteries can ensure that the data center can keep its average power equal to the peak power threshold, $P_{ave} = P_{th}$. If the regulation capacity is C_{reg}, then the data center power consumption can be any value in the interval $[P_{ave} - C_{reg}, P_{ave} + C_{reg}]$ so the peak power of the data center is set to $P_{th} + C_{reg}$, instead of P_{th}. The savings from regulation services need to be larger than the difference between the original and the elevated peak power cost. The peak power cost is charged with the largest consumption over a month, and thus, it increases with the maximum regulation capacity. Then the condition becomes: $C_{reg}c_r \geq (P_{th} + C_{reg_{max}})c_{pp} - P_{th}c_{pp}$, where c_r is the hourly regulation price in \$/MW, $C_{reg_{max}}$ is the maximum regulation capacity over a month in MW, c_{pp} is the monthly peak power cost in \$/MW, which is around \$12,000/MW. The above constraint is defined for an hourly interval, while the peak power cost is charged on the monthly basis. We assume that the data center provides regulation services in each interval over a month and average the monthly peak power costs over all these intervals to obtain a lower bound of the average regulation price that guarantees cost savings:

$$c_{r_{ave}} \geq \frac{c_{pp}}{30 \times 24} \frac{C_{reg_{max}}}{c_{reg_{ave}}} \tag{6.1}$$

where $c_{r_{ave}}$ is the average monthly regulation price in \$/MW and $C_{reg_{ave}}$ is the average regulation capacity that can be provisioned over a month in MW. The ratio $C_{reg_{max}}/c_{reg_{ave}}$ is the main determining factor of this lower bound. It ranges between 2-5 depending on the peak power threshold, battery type and DoD limits. The lower bound obtained of equation 6.1 can be unreachable in some markets, e.g. CAISO [1], but feasible in others, e.g. NYISO [7].

Equation 6.1 assumes that the regulation service should be provided in each interval over the month with a price larger than $c_{r_{ave}}$ to make up for the extra peak power cost. A solution to this problem is to limit the $C_{reg_{max}}/C_{reg_{ave}}$ ratio, thus limiting the maximum regulation capacity so that the possibility of reaching a very high peak power level is eliminated. In this case, the condition in equation 6.1 becomes:

$$c_{r_{ave}} \geq \frac{c_p p}{30 \times 24} PAR \tag{6.2}$$

where PAR (peak-to-average ratio) is the fixed $C_{reg_{max}}/C_{reg_{ave}}$ ratio and the maximum regulation capacity is limited by $C_{reg_{ave}}PAR$,which prevents the peak power from exceeding a predefined limit.

After setting a lower bound on the regulation price, the next step is to set the regulation capacity, C_{reg}. The upper bound of the power flexibility interval, $P_{th} + C_{reg}$, cannot be greater than the nominal data center power demand in that interval, P_{nom}, when the batteries are discharging. Note that at this point, if data center does not provide regulation services, the batteries discharge to reduce the data center power demand from utility from P_{nom} to P_{th}. This adjusted consumption can be increased by stopping some batteries discharging, up to P_{nom}. In addition, the lower bound, $P_{th} - C_{reg}$ cannot be smaller than the actual demand when the batteries are charging. Since the actual demand averages to P_{th} in an interval, we can determine the regulation capacity in each interval, t, as:

$$C_{reg}(t) = \begin{cases} min(P_{nom}(t) - P_{th}, C_{reg_{max}}), & \text{if} \quad P_{nom}(t) < P_{th} \\ min(P_{th} - P_{nom}(t), P_{peak} - P_{th}, C_{reg_{max}}), & \text{otherwise} \end{cases} \tag{6.3}$$

Equation 6.3 ensures that the actual power demand stays within the acceptable range and does not exceed $C_{reg_{max}}$. Since the average power demand from the utility stays the same as in the case where peak shaving is used without regulation services, the expected battery lifetime does not change. While in this example the peak power cost is a limiting factor for regulation services., in the following cases we We next explore the possibility of providing regulation services without changing the original peak power level.

6.2.2 Varying Average Power

We analyze this option in two parts. First, we focus on intervals with batteries discharging. Since the goal now is to provide regulation services without increasing the original data center peak power, so the upper limit of the regulation interval, $P_{ave} + C_{reg}$, should not exceed the peak power threshold, P_{th}. Thus, the data center should reduce its average power demand further than P_{th} when providing regulation services. Although the peak power cost of the data center does not increase, the batteries may need to discharge deeper than the allowed DoD limit to create the power flexibility required by the utility to provide regulation services, thus decreasing the expected battery lifetime and increasing the battery costs. We use the battery model described in section 5.2 to estimate the battery costs. It models the effect of each charge/discharge cycle on the battery lifetime, based on the DoD level and the discharging current in that cycle, and calculates the cost of each cycle. This property allows us to get the cost difference of using a battery with different DoD levels in a single cycle.

We first start our analysis by proving that in a discharging interval it is not possible to reduce the average power consumption further than the original best peak power threshold without violating the DoD limit requirement for some batteries for only that interval. We assume that the total battery capacity consists of a collection of smaller batteries.

Proof: Assume that we have the best peak level, P_{best} with the DoD level, D, where $0 \leq D \leq 100$. Suppose that we can lower P_{best} in the interval t_1,where $P_{nom}(t_1) \leq P_{best}$ without discharging any battery further than D in t_1. Let us denote the new peak power level in t_1 as $P_{best_{t1}}$, such that $P_{best_{t1}} < P_{best}$.Since we do not discharge any battery further than D, there must be enough energy in some batteries to provide the energy difference, $E_{diff} = (P_{best} - P_{best_{t1}})|t_1|$, where $|t_1|$ is the length of the interval t_1. But, this energy should be restored back to the batteries later to ensure that the only DoD violation happens in t_1. There are two situations we need to focus:

1. There is a collection of intervals $\{t_i\}$ where $\sum_{t_i}(P_{best} - P_{adj}(t_i))|t_i| \geq E_{diff}$ where $P_{adj}(t_i) < P_{best}$ and $t_i > t_1$ for all i. Here P_{adj} denotes the power

demand after using batteries, i.e. the dashed line in Figure 6.1. This makes sure that after the interval t_1, there is enough energy slack in a collection of some recharging intervals to store back E_{diff} so that it can safely be distributed over all discharging intervals, rather than to be used only in t_1. Thus, P_{best} is actually not the best peak power level with D, which is a contradiction.

2. There is not any collection of intervals satisfying the above condition. Then, since we need to charge the batteries back with E_{diff} to make sure that the only DoD violation would be in t_1, there has to be an interval, $t_2 > t_1$ in which recharging the batteries violates the original peak power threshold, i.e. $P_{adj}(t_2) > P_{best}$. So, we cannot sustain the original peak, P_{best}. ■

We need to investigate the tradeoff between increasing the DoD limit in an interval to provide regulation service and its savings. We use t for the interval in which we analyze the feasibility of providing regulation services. Since we do not want to increase the peak power threshold, P_{th}, the average power the data center reports to the utility in the interval t, $P_{ave}(t)$, should be smaller than P_{th}. Thus,C_{reg}, can be at most $P_{th} - P_{ave}(t)$ and the savings become $(P_{th} - P_{ave}(t))c_r(t)$ where $c_r(t)$ is the regulation price in the interval t.

Some batteries may need to discharge further than the fixed DoD limit, D, to account for the additional power demand in the interval t, $P_{th} - P_{ave}(t)$. We distribute this additional demand to all batteries to minimize the extra DoD and to limit their discharging current. We obtain the extra DoD in the interval t, D_{extra}, as:

$$\frac{P_{th} - P_{ave}(t)}{NVC_{eff}} \times |t| \times 100 \tag{6.4}$$

where N is the number of batteries, V is the single battery voltage and C_{eff} is the effective battery capacity which is based on the discharging current and Peukert exponent reflecting the physical properties of the battery, as described in 5.2. We distribute the required battery power to all batteries and discharge them with the same current each time, and thus, assume that their expected lifetime is

the same. Then, we calculate the cost of discharging all batteries up to DoD value ($D + D_{extra}$, rather than D, in one cycle, as:

$$Cost_{bat_{extra}} = NC_R c_{bat}\left(\frac{1}{lt(D + D_{extra}, I_{d_{extra}})} - \frac{1}{lt(D, I_d)}\right) \tag{6.5}$$

where c_{bat} is the unit battery cost in \$/Ah, $I_{d_{extra}}$ is the single battery discharging current when providing regulation and I_d is the original single battery discharging current in the interval t. In equation 6.5, the crucial part is the function $lt(D, I)$ which calculates the expected battery lifetime (in cycles) when the battery is used with D depth of discharge limit and I discharging current. This function considers type-specific battery properties and penalizes higher DoD values and discharging currents to reflect their negative effects on battery lifetime. More details of $lt()$ can be found in equation 5.4. In equation 6.5, we compute the cost of using all batteries in one cycle with $(D + D_{extra}, I_{d_{extra}})$ and (D, I_d) and take the difference. Then, the main optimization goal becomes:

$$\begin{aligned} & max(P_{th} - P_{ave}(t))c_r(t) - Cost_{bat_{extra}} \\ \text{s.t.} \quad & P_{ave}(t) < P_{th} \\ & 0 < D_{extra} < 100 - D \end{aligned} \tag{6.6}$$

which maximizes the savings of regulation services. The constraints in the other two equations ensure that $P_{ave}(t)$ does not violate peak power limits and does not require that the energy battery provides is more than the total battery capacity. In this case, the regulation price is limited by battery characteristics and fixed DoD limits as they are the main inputs of battery lifetime calculation. In the results section of this chapter, we show that current regulation prices in NYISO and CAISO markets cannot compensate for the increased battery costs due to the larger DoD.

Next, we focus on the intervals where the data center nominal power is less than the peak power threshold. In hours between 0-3, 18-25 and 63-70 shown in Figure 6.1 the difference between the nominal power and the peak power threshold can provide the flexibility required by regulation services. We schedule the bat-

tery recharge events to create this flexibility and still stay within the peak power threshold. Thus, we select:

$$C_{reg} = \frac{P_{th} - P_{nom}(t)}{2}$$
$$P_{ave} = \frac{P_{th} + P_{nom}(t)}{2} \tag{6.7}$$

where $P_{nom}(t)$ is the actual data center power demand in interval t. However, the data center cannot provide regulation services during all the intervals the batteries recharge or are idle because it has to ensure that batteries have enough energy stored before being discharged again. Thus, we need to determine which subset of intervals where $P_{nom}(t) < P_{th}$ should be selected to provide regulation services:

$$\max_{I} \sum_{t_i \in I} \frac{P_{th} - P_{nom}(t_i)}{2} c_r(t_i)$$
$$I \subseteq \{t | P_{adj}(t) < P_{th}\} \tag{6.8}$$

where $P_{adj}(t)$ is the adjusted power in interval t. The main goal is to select a set of intervals that maximize the regulation savings. These intervals should be a subset of the intervals with adjusted power less than the peak power threshold to ensure that we do not work with the discharging intervals and only use intervals with some flexibility. We shift the recharge events among these intervals, intervals with high regulation price provides regulation services and the others make sure that batteries have enough energy before the next peak event. Also, a recharge interval should not be shifted forward beyond a discharge interval because that recharge event might be necessary to prevent the peak in the relevant discharge interval.

This part considers only with battery recharging intervals. Thus it does not put any pressure on the battery DoD limits. Computation of regulation capacity does not allow any peak power violations. However, its applicability might be limited as the data center needs to select a subset of recharging intervals fitting

the above restrictions. The availability of those intervals depends highly on the fixed DoD limit. In the results section, we demonstrate that this method can obtain savings even in a price conservative market as it does not increase peak power costs or battery costs.

6.3 Evaluation

In this section, we evaluate the two methods described in the previous section. The *Fixed Average Power* method does not change the average data center power demand when providing regulation, but increases the peak power level. It can be seen as a representative of the previous studies [10][37] as it does not change the average power consumption in an interval, but it differs strictly from them as it considers peak power management. *Varying Average Power* method does not modify the peak power level but instead change the average power using the battery charge and discharge events. A data center can use each method separately, and decide which one is more cost-efficient based on its power demand profile, peak power goals, the battery type used and the power market it participates in.

Next, we first describe our experimental setup, including data center and battery settings. Then, we present the effectiveness of *Fixed* and *Varying Average Power* methods for data center participation in regulation markets.

6.3.1 Methodology

We model a large data center with 50,000 Sun Fire servers, each at 175W idle and 350W peak power, and use a linear, CPU-utilization based function to compute the power consumption of a single server as in chapter 5. We also consider non-server power consumption with the power usage efficiency (PUE) metric and use 1.2 for this value, which corresponds to an energy efficient data center [28]. We use the same workload mixture introduced in Table 5.6 and Figure 5.11, and scale the workload information for 50,000 servers. Then, we use an event-based simulation platform with this workload information to extract the total data center power demand profile.

Table 6.1: Best peak shaving percentages with different DoD levels

DoD value	**20%**	**40%**	**60%**	**80%**
Best peak shaving %	15.4%	18.3%	20.6%	21.4%
Battery type	LA		LFP	

We use battery-based peak power shaving to reduce the cost of peak power in the monthly utility bill. The battery types we use in our work are LA and LFP batteries, with capacity 40Ah per server as in chapter 5. Another reason we use multiple types of batteries is that they have different levels of optimum DoD levels for peak power shaving and in our evaluation, we show how different levels of DoD levels affect the efficiency of regulation services. Table 6.1 shows the peak power shaving percentage of the total data center peak power, obtained with different DoD levels and appropriate battery types for each DoD level, using the workload information from Figure 5.11. Peak power prices and battery properties are taken from Table 5.8.

We target the day-ahead option in the regulation services market as it has higher prices than the other ones. Data centers can estimate the expected load for the day ahead, which can allow them to provision their resources as a function of the regulation services in the day-ahead market [37]. We use pricing from NYISO and CAISO to show the importance of the market data center participates in. We get the NYISO numbers from previous studies [10] and CAISO ones from their database [1]. Figure 6.2 shows the daily regulation prices in our evaluation.

6.3.2 Results: Fixed Average Power

In this section, we evaluate our first method that does not change the average power demand of the data center, but instead increases the peak power level to match the regulation capacity provided. However, these previous studies do not account for peak power costs, which can affect the regulation capacity and in turn, total savings. We estimate the total savings for different DoD levels and PAR values with both NYISO and CAISO prices. The maximum PAR values

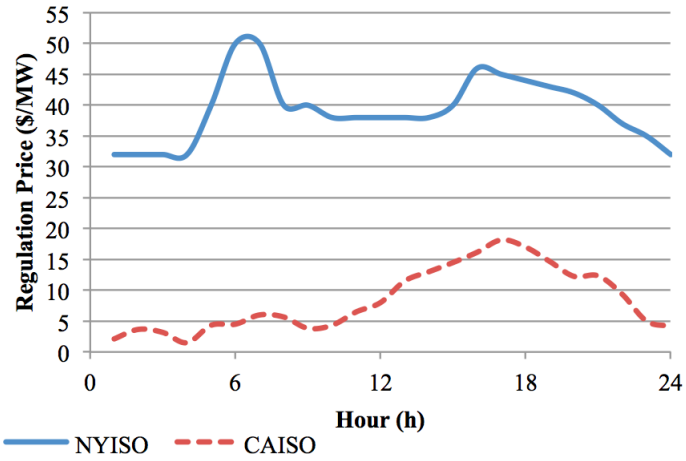


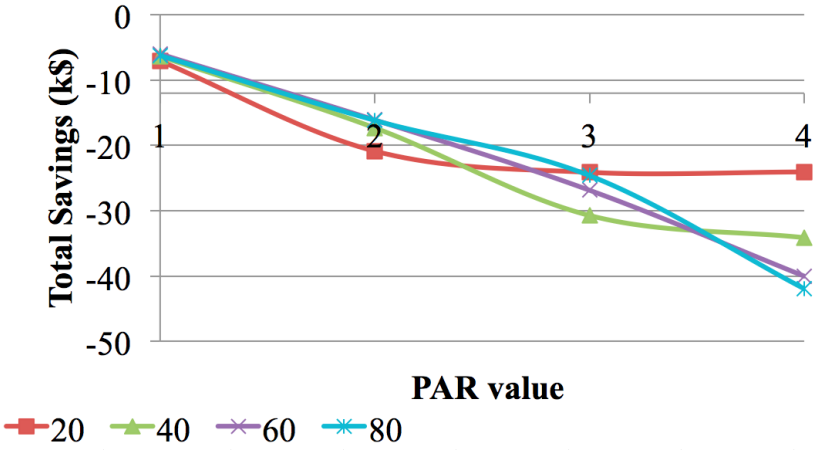
Figure 6.2: Regulation prices

Table 6.2: Maximum PAR values for different DoD levels

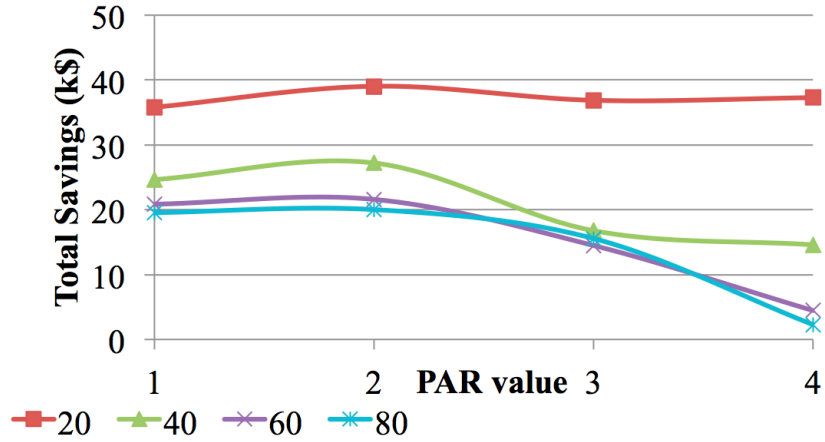
DoD value	**20%**	**40%**	**60%**	**80%**
PAR value	2.2	3.3	4	4.1

that can be obtained with varying DoD limits are listed in Table 6.2.

Figure 6.3 shows the total savings of providing regulation services in CAISO and NYISO markets. In both graphs, the x-axis corresponds to changing PAR values and the y-axis shows the total savings in a month in thousand dollars. The savings are calculated as the difference between the profit from providing regulation services and the cost of increased peak power level. We select the regulation capacity using different PAR values based on equation 6.3. Since the average power consumption does not change compared to the no-regulation case, there is no extra battery cost. We observe that the data center can obtain savings for any PAR value in the NYISO market, whereas the CAISO prices do not lead to any savings. The best PAR value is 2 in the NYISO market whereas it is 1 in the CAISO market because limiting the peak power increase limits the additional peak power costs and increases the savings. The best DoD value is 20% for the NYISO and 60% for the CAISO markets. The maximum savings are $40,000 with NYISO pricing, corresponding to 5% savings overall the electricity bill. These savings can also be used to accommodate $1280 more servers within the same



(a) with CAISO prices



(b) with NYISO prices

Figure 6.3: Total savings result with CAISO and NYISO prices

energy budget. The savings are always negative with CAISO pricing, which means that the data center should not participate in the regulation market. This analysis shows the necessity of investigating all the options, such as battery DoD level, the market pricing dynamics and the maximum regulation capacity limit, when providing regulation services along with meeting peak power shaving goals.

Table 6.3 shows the average error in savings if peak power costs are not considered with different PAR values as was done in previous studies. We see that the error is smaller with lower PAR values since PAR value limits the maximum regulation capacity and its effects on the peak power cost. The error is up to 385%

Table 6.3: Error percentages if peak power costs are not considered

	Error Percentage			
PAR value	**1**	**2**	**3**	**4**
NYISO	37%	52%	66%	79%
CAISO	182%	257%	324%	385

for the CAISO market and 80% for the NYISO one. It is much higher in CAISO market as peak power costs are much larger than possible regulation savings due to low prices.

6.3.3 Results: Varying Average Power

We first focus on the intervals where the batteries are discharging and assume that the original peak power levels obtained with a given battery configuration should not be increased to avoid high peak power costs. Thus, the flexibility interval in which regulation services can be provided should be created under the original best peak power threshold. As a result, batteries discharge further than their allowed DoD limit to create this flexibility range. The best peak shaving with LA and LFP batteries are obtained with 40% and 60% DoD at 17.2MW and 16.7MW respectively for our data center with 50,000 servers. The expected battery lifetime values are 2.5 and 6.4 years for LA and LFP batteries respectively.

Table 6.4 shows how much the minimum regulation price should be for a given regulation capacity to compensate for the increased cost of using batteries with deeper discharges. We change the amount of regulation capacity to be provided and calculate the extra DoD level required by the batteries for each case. This extra DoD leads to a higher cycle cost. We compute it by estimating the battery lifetime if the battery is used with the extra DoD in each cycle. Lastly, we calculate the minimum regulation prices in $/MW that makes up for the increased cycle cost. Table 6.4 shows that the required minimum regulation prices are much higher than the actual prices (both NYISO and CAISO) for both LA and LFP batteries, and the minimum prices increase with increased regulation capacity. Thus,

Table 6.4: Regulation price analysis for battery discharge intervals

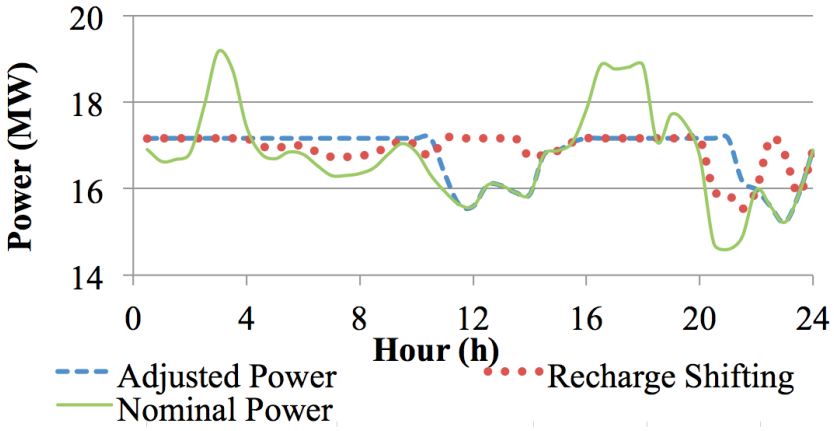
		Battery Life (yrs)		Min. Reg. Price ($/W)	
C_{reg} (MW)	D_{extra}	LA	LFP	LA	LFP
0.5	2.08	2.4	6.2	183	138
1	4.17	2.3	6	191	143
1.5	6.25	2.2	5.8	199	148

Table 6.5: Monthly savings using recharge shifting

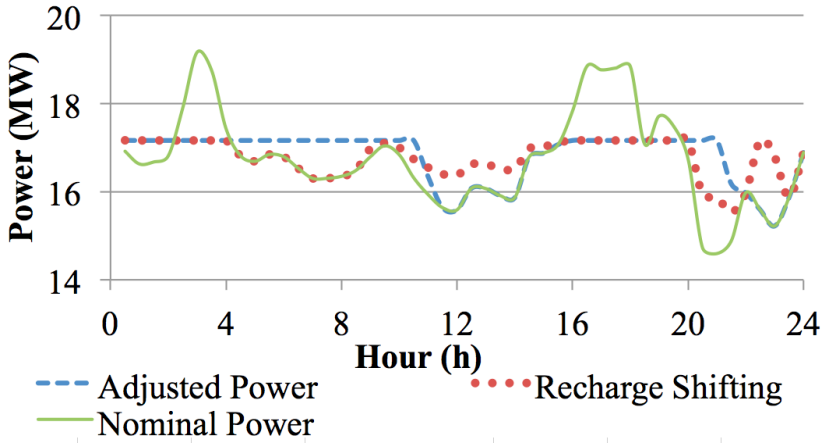
DoD	NYISO Savings ($)	CAISO Savings ($)
40	33628	11312
60	14600	5132

we can conclude that this method is not feasible for data centers with peak power shaving and battery lifetime limitations when providing regulation services with the current market prices. This method can become feasible with lower battery prices or less nonlinear battery behavior with higher DoD. Additionally, emergency DR events, with high prices as much as $500/MW [10], might be a good target to compensate for high battery costs.

Next, we investigate intervals when the batteries are idle or recharging. We use the framework described in section 6.2.2 to obtain the intervals in which regulation can be provided. Table 6.5 shows the monthly savings for both 40% and 60% DoD levels, corresponding to the best peak shaving configurations for LA and LFP batteries respectively, in both NYISO and CAISO markets. NYISO savings are almost 3x higher than CAISO due to the higher regulation prices. An important observation is that lower DoD limits lead to more than 2x in savings, because there is more flexibility in the recharging intervals with a lower DoD limit. This result does not have additional peak power costs as the peak power level does not change. The advantage of this method is that it does not increase the original peak power threshold (unlike *Fixed Average Power method*) and does not put extra burden on the batteries (unlike the first part of this method). Also, it can obtain



(a) NYISO



(b) CAISO

Figure 6.4: Recharge shifting for NYSIO and CAISO

savings in a more conservative market, i.e. CAISO, where *Fixed Average Power* method cannot.

The core of this method is shifting the time of battery recharge. Figure 6.4 shows the recharge shifting for NYISO (6.4a) and CAISO (6.4b) prices for a sample day where the DoD limit is set to 40%. In both graphs x-axis shows the time in hours and y-axis is the power consumption in MW. The straight line stands for the nominal consumption, the dashed line is the adjusted power consumption using batteries without recharge shifting and dotted line represents the consumption with recharge shifting. For both graphs, recharge shifting is visible between hours 4-15 and 20-24. The way recharge occurs is different because of

the pricing schemes. In 4-15 in the NYISO market, the data center first decreases its adjusted power consumption until t=11.5hrs to create flexibility for regulation services. Then in 11.5-15hrs, it increases the consumption to complete the battery recharge. In contrast, when using CAISO prices, it stops battery recharging completely in 4-8.5hrs and then provide regulation services in 8.5-15.5hrs by recharging the batteries. By shifting the recharge periods, the data center tries to provide regulation services in the intervals with higher regulation prices. The data center has the same recharge shifting events in both NYISO and CAISO markets in 20-24hrs due to same pricing trend in both markets.

In this section, we introduced two complementary methods for data centers with peak power budgets to participate in regulation services markets. *Fixed Average Power* method keeps the original data center power demand the same and thus may violate the peak power thresholds when providing regulation services. *Varying Average Power* method creates the required flexibility in power demand strictly under the peak power threshold at the expense of more aggressive battery usage. The feasibility of these methods depends on the peak power thresholds, battery properties and the regulation market dynamics. We calculate the savings from these methods on top of already existing peak power shaving savings. We observe that in markets with high regulation prices *Fixed Average Power* method brings more savings whereas in price-conservative markets *Varying Average Power* method is more beneficial. The advantages of our methods are that they consider peak power budgets and do not lead to any workload performance degradation.

6.4 Conclusion

Electricity utilities present different options for both generator and non-generator sources in the ancillary services market. Data centers can provide regulation services in these markets with their high energy consumption and ability to create flexibility in their demand profiles, and obtain additional savings. Existing studies all degrade the workload performance to provide regulation services and do not consider peak power costs which can affect the regulation capacity and re-

sult in up to 385% overestimation in savings. Our solution adopts a battery-based peak shaving method and has no performance impact on the workloads. It shows two options to provide regulation services considering the peak power costs. We see that increasing the peak power limits may be feasible in markets with high regulation prices but in a more conservative market, the best practice is to provide regulation services only in intervals when the batteries are recharging. Our methods can obtain $480,000/year savings, which can accommodate 1280 more servers and increase the profit percentage by 5.08%.

Chapter 6 contains material from "Providing Regulation Services and Managing Data Center Peak Power Budgets", by Baris Aksanli and Tajana Rosing, which appears in Proceedings of Design Automation and Test in Europe (DATE), 2014 [15]. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Summary and Future Work

The number of data centers has been increasing over the last decade to meet the rapidly-exploding computation demand. These warehouse-scale, compute-oriented buildings host millions of servers globally, and constantly require high amount of energy to maintain their operability. This high demand, in turn, translates into elevated electricity bills, which have become one of the largest components of data center operational expenses. Additionally, increased cost of fossil-based, brown energy and carbon emission penalties have forced data centers to search for alternative energy sources. As a result, it has become extremely important for data centers to be energy efficient, which also corresponds to cost savings.

7.1 Thesis Summary

This thesis proposes approaches to energy efficiency problem of data centers from multiple dimensions that are complementary to each other. The proposed methods can be applied either individually or together and they can dramatically lower the utility bill, corresponding to millions of dollars of savings for large scale data centers. The thesis first proposes mechanisms to efficiently integrate renewable energy to both data centers and the wide are networks connecting multiple data centers. It makes use of a predictive approach that helps data centers reduce their carbon footprint without any performance hits caused by the highly variable nature of green energy sources. Second, it presents holistic cost minimization and

performance maximization approaches for multiple data center systems which make use of online job migration made possible by recent technological improvements. These approaches both increase the renewable energy penetration and decrease the overall energy cost significantly by identifying and modeling the key aspects of multiple data center systems. The thesis also targets to reduce the peak power level of data centers to decrease the utility bill. It presents battery-based peak power aware solutions that minimize the deployment costs by optimizing for battery lifetime and enable data centers to effectively reduce their peak power levels without an impact on job performance. Lastly, the thesis demonstrates how data centers and the electric grid can collaborate in a mutually beneficial way, where the data center can increase its profits and the utilities can maintain the health of the electric grid. Next, we show how we increase the efficiency of data centers with renewable energy, peak power management and collaborating with the electric grid in detail. Finally, we provide some future research ideas.

7.1.1 Renewable Energy in Data Center Systems

Green energy usage in data centers systems has gained importance as their energy consumption, carbon emissions, and costs have increased dramatically. Existing studies focus on using immediately available green energy to supplement the non-renewable, or brown energy at the cost of canceling and rescheduling jobs whenever the green energy availability is too low. This thesis first proposes an adaptive data center job scheduler which utilizes short term prediction of solar and wind energy production. This enables the data center to scale the number of jobs to the expected energy availability, thus reducing the number of cancelled jobs and improving green energy usage efficiency as compared to just utilizing the immediately available green energy.

This thesis also investigates the importance of wide area networks to improve the job performance in data center systems. We not only quantify the performance benefits of leveraging the network to run more jobs, but also analyze its energy impact. We compare the benefits of redesigning routers to be more energy efficient to those obtained by leveraging locally available green energy as

a complement to the brown energy supply. We design novel green energy aware routing policies for wide area traffic and compare to state-of-the-art shortest path routing algorithm. Our analysis indicate that using energy proportional routers powered in part by green energy along with our new routing algorithm results in significant improvement in per router energy efficiency with increased batch job throughput due.

Previous work leverages geographically separated data centers by migrating workloads over WAN, leveraging demand and price differences. However, the work neglects several key cost and energy contributions: the financial network, and consequently, data migration costs, focusing solely on latency and quality of service costs. Additionally, these studies assume a simpler, and ultimately inaccurate, model for data center energy costs. This thesis explores tiered energy pricing for data centers, network cost models and the costs of owning/leasing a data center WAN. We develop algorithms for energy management, focusing on 1) performance maximization, and 2) cost minimization. With the performance maximization algorithm, we demonstrate the ability to leverage green energy to actually improve workload throughput, rather than simply reducing the operational costs. We further explore the viability of our new algorithms in the face of emerging technologies in data center infrastructure, showing continued benefit of both the performance maximization and the cost minimization algorithms in the presence of energy proportional computing and communication.

7.1.2 Efficient Peak Power Shaving in Data Centers

Peak power shaving allows data center providers to keep their power demand under a predetermined threshold. This operation may either directly correspond to savings due to reduced peak power level or increase the computational capacity without exceeding a given power budget. Recent studies show that data centers can leverage the stored energy in batteries (or other energy storage devices) to achieve lower peak power levels. Battery-based peak power shaving is extremely useful since it does not interfere with workloads, resulting in no performance overhead. The battery placement designs can vary across data centers. The most well-known

designs are the traditional centralized design and the distributed design where the total battery capacity is distributed across the servers or the racks.

This thesis first focuses on the distributed battery design and proposes a novel distributed battery control design that has no performance impact, reduces the peak power needs, and accurately estimates and maximizes the battery lifetime. We demonstrate that models which do not take into account physical characteristics of batteries can overestimate their lifetime and in turn, their savings. In contrast, our design closely approximates the best centralized solution with an order of magnitude smaller communication overhead. The thesis then demonstrates an architecture where batteries provide only a fraction of the data center power, exploiting nonlinear battery capacity properties to achieve longer battery life and longer peak shaving durations. This architecture demonstrates that a centralized UPS with partial discharge sufficiently reduces the cost so that double power conversion losses are not a limiting factor, thus contradicting the recent trends in warehouse-scale distributed UPS design. Our architecture almost doubles the battery lifetime with increased cost savings and significantly reduced communication overhead due to central battery placement.

7.1.3 Data Centers in the Grid

Utilities have been using ancillary services to keep the electric grid safe and operational. These services include regulation services, spinning and non-spinning reserves, demand response, voluntary load reduction, etc. Utilities employ these services mostly to eliminate the supply/demand imbalances. Traditionally, utilities only allowed generator sources to participate in these ancillary services, such as power plants, solar farms, wind turbines, etc. Recently, they have allowed non-generator sources participation in these services as well. Data centers are good candidates for participating in the ancillary services market due to their large power consumption and flexibility that can be created with various energy/power management mechanisms.

This thesis focuses on one such ancillary service for data centers, regulation services. This is because the possible return of regulation providing is higher than

the other ancillary services, but it also requires more prompt response from data centers to utility signals. On the other hand, the regulation participation contracts are kept separate than the normal energy and peak power costs. Therefore, such a participation by a data center should carefully be analyzed for savings. This thesis develops a framework that explores the feasibility of data center participation in the regulation services markets. It uses a battery-based design that can not only help with providing ancillary services, but can also limit peak power costs without any workload performance degradation. The proposed framework considers energy costs, peak power costs, and regulation market dynamics simultaneously and computes the regulation capacity that the data center should allocate for regulation service providing. Our results indicate that significant amount of savings is possible with careful regulation capacity bidding.

7.2 Future Work Directions

7.2.1 Data Centers Causing Instabilities in the Grid

With the integration of highly distributed renewable energy sources and large-scale smart buildings, the electricity grid becomes more prone to experience instabilities due to unexpected fluctuations in energy consumption. As we show in this thesis, data centers are good candidates to participate in the ancillary services to help utilities maintain the operational environment of the electric grid. This is because data centers are a type of smart building because of their innate automation and the fact that their loads can be significantly controlled. Existing studies focusing on this relation between data centers and the grid assume that the data center can help the grid via ancillary services but do not consider how data centers can lead to imbalances in the grid.

Due to their significant power demand, data centers may not only lead to unstable regions in the grid circuit but also threaten other buildings in their surroundings as well. To prevent this, data centers can tune their power management techniques to account for possible instability events they may cause. However, these instabilities also depend on the other buildings in the neighborhood. There-

fore, a data center power control that considers grid instabilities requires two-way communication between the data center and the utility. The utility constantly monitors the power consumption of each building in the grid and based on these values, it can anticipate an instability event along with its major cause. Then, it has to respond to these instabilities and this can have severe impacts on a data center, loss of power (leading to loss of service or violating SLAs), increasing operational costs, etc. Thus, a data center can communicate with the utility to minimize the instability events that it causes by adjusting its power consumption. This adjustment may require a completely new power management mechanism or a combination of existing methods.

7.2.2 Residential Energy Management

The focus of building energy consumption research has been on commercial and industrial sectors, as they constitute a majority of energy consumption. However, residential energy consumption constitutes 38% of the total energy consumption in the US, with millions of individual customers [52]. The technological improvements in the smart grid domain, such as smart metering, different types of sensors (motion, occupancy), etc. enable residential energy consumption to be monitored and tracked more effectively. This monitoring inevitably leads to smarter control mechanism for the residential domain, including load shifting, peak shaving, voltage regulation, energy arbitrage, etc. A good example of smart residential control mechanisms is load shifting where the house demand is classified as deferrable and non-deferrable, and the non-deferrable part is rescheduled based on energy availability or cheaper energy prices [124].

Most of the control mechanisms mentioned above has a similar counterpart in data center systems. Therefore, we can apply the data center power management mechanisms to residential domain. However, residential houses require appropriate automation techniques before these methods can be applied. For example, load shifting with deferrable residential workloads needs the appliances in a house to be automatically controlled. Similarly, peak power shaving and/or voltage regulation with heating, ventilation and air conditioning (HVAC) units requires HVAC unit

to be programmable and remotely controlled.

We will start our residential energy management research with cost-efficient integration of energy storage devices into houses. Residential energy consumption shows significant diurnal patterns that can be leveraged by energy storage devices. Batteries can store energy from either local renewable sources or from the grid when the electricity is cheaper, and provide it when the prices are higher. As we show in chapter 5, battery performance and lifetime depends highly on how these chemical devices are used. We initially develop a framework that considers the physical properties of batteries, tests the feasibility of a battery deployment and finds the best battery types and configurations for a particular residential configuration [14]. Next step is to validate the outcomes our framework through simulations that are informed by measurements, and show how much savings can be obtained by using batteries in a residential house.

Other data center power management methods can also be mapped into the residential domain, such as load shifting, peak power management, etc. Furthermore, single house analysis can be extended to a neighborhood with several houses, where energy allocation becomes a more complex problem due to the heterogeneous nature of different houses.

Chapter 7 contains material from "Optimal Battery Configuration in a Residential Home with Time-of-Use Pricing", by Baris Aksanli and Tajana Rosing, which appears in Proceedings of International Conference on Smart Grid Communications (SmartGridComm), 2013 [14]. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] Caiso. http://oasis.caiso.com.

[2] Electric power monthly. http://www.eia.gov/electricity/monthly.

[3] Energy recommerce. http://www.mypvdata.com.

[4] National renewable energy laboratory. http://www.nrel.gov.

[5] National renewable energy laboratory: Solar resources. http://www.nrel.gov/gis/solar.html.

[6] National renewable energy laboratory: Wind resource. http://www.nrel.gov/rredc/wind_resource.html.

[7] Nyiso. http://www.nyiso.com.

[8] California iso, retrieved from oasis. http://oasis.caiso.com, 2012.

[9] Dennis Abts, Michael R Marty, Philip M Wells, Peter Klausler, and Hong Liu. Energy proportional datacenter networks. In *ACM SIGARCH Computer Architecture News*, volume 38, pages 338–347. ACM, 2010.

[10] David Aikema, Rob Simmonds, and Hamidreza Zareipour. Data centres in the ancillary services market. In *Green Computing Conference (IGCC), 2012 International*, pages 1–10. IEEE, 2012.

[11] B. Aksanli, J. Venkatesh, T. Rosing, and I. Monga. Renewable energy prediction for improved utilization and efficiency in datacenters and backbone networks. In Jörg Lässig, Kristian Kersting, and Katharina Morik, editors, *Computational Sustainability*. Springer, 2015.

[12] Baris Aksanli, Alper S. Akyurek, Madhur Behl, Meghan Clark, Alexandre Donzé, Prabal Dutta, Patrick Lazik, Mehdi Maasoumy, Rahul Mangharam, Truong X. Nghiem, Vasumathi Raman, Anthony Rowe, Alberto Sangiovanni-Vincentelli, Sanjit Seshia, Tajana Simunic Rosing, and Jagannathan Venkatesh. Distributed control of a swarm of buildings connected

to a smart grid: demo abstract. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 172–173. ACM, 2014.

[13] Baris Aksanli, Eddie Pettis, and Tajana Rosing. Architecting efficient peak power shaving using batteries in data centers. In *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2013 IEEE 21st International Symposium on*, pages 242–253. IEEE, 2013.

[14] Baris Aksanli and Tajana Rosing. Optimal battery configuration in a residential home with time-of-use pricing. In *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pages 157–162. IEEE, 2013.

[15] Baris Aksanli and Tajana Rosing. Providing regulation services and managing data center peak power budgets. In *Proceedings of the conference on Design, Automation & Test in Europe*, page 143. European Design and Automation Association, 2014.

[16] Baris Aksanli, Tajana Rosing, and Eddie Pettis. Distributed battery control for peak power shaving in datacenters. In *Green Computing Conference (IGCC), 2013 International*, pages 1–8. IEEE, 2013.

[17] Baris Aksanli, Tajana Simunic Rosing, and Inder Monga. Benefits of green energy and proportionality in high speed wide area networks connecting data centers. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 175–180. EDA Consortium, 2012.

[18] Baris Aksanli, Jagannathan Venkatesh, Tajana Rosing, and Inder Monga. A comprehensive approach to reduce the energy cost of network of datacenters. In *Computers and Communications (ISCC), 2013 IEEE Symposium on*, pages 000275–000280. IEEE, 2013.

[19] Baris Aksanli, Jagannathan Venkatesh, and Tajana Šimunić Rosing. Using datacenter simulation to evaluate green energy integration. *Computer*, 45(9):0056–64, 2012.

[20] 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.

[21] Amazon. Amazon ec2. http://aws.amazon.com/ec2.

[22] Matthew Andrews, Antonio Fernández Anta, Lisa Zhang, and Wenbo Zhao. Routing and scheduling for energy and delay minimization in the powerdown model. *Networks*, 61(3):226–237, 2013.

[23] Apache. http://incubator.apache.org/olio/.

[24] APC. Infrastruxure total cost of ownership, infrastructure cost report. http://www.apc.com/tools/isx/tco, 2008.

[25] A.P.P.Corp. Portable power product design, assemble and quality control. http://www.batteryspace.com/lifepo4cellspacks.aspx.

[26] Jayant Baliga, Robert W.A. Ayre, Kerry Hinton, and Rodney S. Tucker. Green cloud computing: Balancing energy in processing, storage, and transport. *Proceedings of the IEEE*, 99(1):149–167, 2011.

[27] Luiz André Barroso and Urs Hölzle. The case for energy-proportional computing. *IEEE computer*, 40(12):33–37, 2007.

[28] Luiz André Barroso and Urs Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture*, 4(1):1–108, 2009.

[29] Luca Benini, Giuliano Castelli, Alberto Macii, Enrico Macii, Massimo Poncino, and Riccardo Scarsi. Discrete-time battery models for system-level low-power design. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 9(5):630–640, 2001.

[30] Bernd Bergler, Christopher Preschern, Andreas Reiter, and Stefan Kraxberger. Cost-effective routing for a greener internet. In *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, pages 276–283. IEEE Computer Society, 2010.

[31] Aruna Prem Bianzino, Luca Chiaraviglio, and Marco Mellia. Grida: A green distributed algorithm for backbone networks. In *Online Conference on Green Communications (GreenCom), 2011 IEEE*, pages 113–119. IEEE, 2011.

[32] Raffaele Bolla, Roberto Bruschi, Franco Davoli, and Flavio Cucchietti. Energy efficiency in the future internet: a survey of existing approaches and trends in energy-aware fixed network infrastructures. *Communications Surveys & Tutorials, IEEE*, 13(2):223–244, 2011.

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

[34] CAISO. Non-generator sources. http://www.caiso.com/informed/Pages/StakeholderProcesses/CompletedStakeholderProcesses/NonGenerator ResourcesAncillaryServicesMarket.aspx.

[35] D. E. Carolinas. Utility bill tariff. http://www.duke-energy.com/pdfs/scscheduleopt.pdf, 2009.

[36] Joseph Chabarek, Joel Sommers, Paul Barford, Cristian Estan, David Tsiang, and Steve Wright. Power awareness in network design and routing. In *INFOCOM 2008. The 27th Conference on Computer Communications. IEEE*. IEEE, 2008.

[37] Hao Chen, Can Hankendi, Michael C Caramanis, and Ayse K Coskun. Dynamic server power capping for enabling data center participation in power markets. In *Proceedings of the International Conference on Computer-Aided Design*, pages 122–129. IEEE Press, 2013.

[38] Yanpei Chen, Archana Ganapathi, Rean Griffith, and Randy Katz. The case for evaluating mapreduce performance using workload suites. In *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2011 IEEE 19th International Symposium on*, pages 390–399. IEEE, 2011.

[39] Luca Chiaraviglio, Marco Mellia, and Fabio Neri. Energy-aware backbone networks: a case study. In *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*, pages 1–5. IEEE, 2009.

[40] Cisco. Priority ow control: Build reliable layer 2 infrastructure. http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9670/white_paper_c11-542809.pdf.

[41] C.S Computing. www.climatesaverscomputing.org/resources/certification.

[42] Gary Cook and J Van Horn. How dirty is your data: A look at the energy choices that power cloud computing. *Greenpeace (April 2011)*, 2011.

[43] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[44] Nan Deng, Christopher Stewart, and Jing Li. Concentrating renewable energy in grid-tied datacenters. In *Sustainable Systems and Technology (ISSST), 2011 IEEE International Symposium on*, pages 1–6. IEEE, 2011.

[45] Stephen Drouilhet and Bertrand L Johnson. A battery life prediction method for hybrid power applications. In *AIAA Aerospace Sciences Meeting and Exhibit*, 1997.

[46] Dimitris Economou, Suzanne Rivoire, Christos Kozyrakis, and Partha Ranganathan. Full-system power analysis and modeling for server environments. International Symposium on Computer Architecture-IEEE, 2006.

[47] Deniz Ersoz, Mazin S Yousif, and Chita R Das. Characterizing network traffic in a cluster-based, multi-tier data center. In *Distributed Computing Systems, 2007. ICDCS'07. 27th International Conference on*, pages 59–59. IEEE, 2007.

[48] Esnet. Network topology. http://www.es.net/network/network-maps.

[49] Facebook. Hacking conventional computing infrastructure. http://opencompute.org/, 2011.

[50] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power provisioning for a warehouse-sized computer. In *ACM SIGARCH Computer Architecture News*, volume 35, pages 13–23. ACM, 2007.

[51] Will Fisher, Martin Suchara, and Jennifer Rexford. Greening backbone networks: reducing energy consumption by shutting off cables in bundled links. In *Proceedings of the first ACM SIGCOMM workshop on Green networking*, pages 29–34. ACM, 2010.

[52] Center for Climate and Energy Solutions. Energy and technology. http://www.c2es.org/category/topic/energy-technology, 2011.

[53] Brian Fortenbery, Ecos Consulting EPRI, and William Tschudi. Dc power for improved data center efficiency. 2008.

[54] M. Fratto. http://www.networkcomputing.com/data-center/229503323, 2009.

[55] Mahdi Ghamkhari and Hamed Mohsenian-Rad. Data centers to offer ancillary services. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, pages 436–441. IEEE, 2012.

[56] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *ACM SIGOPS Operating Systems Review*, volume 37, pages 29–43. ACM, 2003.

[57] Gregor Giebel, Richard Brownsword, George Kariniotakis, Michael Denhard, and Caroline Draxl. The state-of-the-art in short-term prediction of wind power: A literature overview. Technical report, ANEMOS. plus, 2011.

[58] Daniel Gmach, Yuan Chen, Amip Shah, Jerry Rolia, Cullen Bash, Tom Christian, and Ratnesh Sharma. Profiling sustainability of data centers. In *Sustainable Systems and Technology (ISSST), 2010 IEEE International Symposium on*, pages 1–6. IEEE, 2010.

[59] Daniel Gmach, Jerry Rolia, Cullen Bash, Yuan Chen, Tom Christian, Amip Shah, Ratnesh Sharma, and Zhikui Wang. Capacity planning and power management to exploit sustainable energy. In *Network and Service Management (CNSM), 2010 International Conference on*, pages 96–103. IEEE, 2010.

[60] Google. Google transparency report. http://www.google.com/transparencyreport/traffic.

[61] Google. Google summit. http://www.google.com/corporate/datacenter/events/dc-summit-2009.html, 2009.

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

[63] Sriram Govindan, Di Wang, Anand Sivasubramaniam, and Bhuvan Urgaonkar. Leveraging stored energy for handling power emergencies in aggressively provisioned datacenters. In *ACM SIGPLAN Notices*, volume 47, pages 75–86. ACM, 2012.

[64] Chin Guok. A user driven dynamic circuit network implementation. *Lawrence Berkeley National Laboratory*, 2009.

[65] Hadoop. http://hadoop.apache.org.

[66] F. Harvey. Table with peukert's exponent for different battery models. http://www.electricmotorsport.com/store/ems_ev_parts_batteries.php.

[67] Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004.

[68] Intel. Intel microarchitecture nehalem. http://www.intel.com/technology/architecture-silicon/next-gen.

[69] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, and Dennis Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *ACM SIGOPS Operating Systems Review*, volume 41, pages 59–72. ACM, 2007.

[70] Krishna Kant. Power control of high speed network interconnects in data centers. In *INFOCOM Workshops 2009, IEEE*, pages 1–6. IEEE, 2009.

[71] S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan. An analysis of traces from a production mapreduce cluster. In *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*, pages 94–103, May 2010.

[72] Dzmitry Kliazovich, Pascal Bouvry, and Samee Ullah Khan. Greencloud: a packet-level simulator of energy-aware cloud computing data centers. *The Journal of Supercomputing*, 62(3):1263–1283, 2012.

[73] Vasileios Kontorinis, Liuyi Eric Zhang, Baris Aksanli, Jack Sampson, Houman Homayoun, Eddie Pettis, Dean M Tullsen, and T Simunic Rosing. Managing distributed ups energy for effective power capping in data centers. In *Computer Architecture (ISCA), 2012 39th Annual International Symposium on*, pages 488–499. IEEE, 2012.

[74] Jonathan Koomey. Growth in data center electricity use 2005 to 2010. *A report by Analytical Press, completed at the request of The New York Times*, 2011.

[75] 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.

[76] Andrew Kusiak, Haiyang Zheng, and Zhe Song. Short-term prediction of wind farm power: a data mining approach. *Energy Conversion, IEEE Transactions on*, 24(1):125–136, 2009.

[77] Mahendra Kutare, Greg Eisenhauer, Chengwei Wang, Karsten Schwan, Vanish Talwar, and Matthew Wolf. Monalytics: online monitoring and analytics for managing large scale data centers. In *Proceedings of the 7th international conference on Autonomic computing*, pages 141–150. ACM, 2010.

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

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

[80] Kien Le, Ozlem Bilgir, Ricardo Bianchini, Margaret Martonosi, and Thu D Nguyen. Managing the cost, energy consumption, and carbon footprint of internet services. In *ACM SIGMETRICS Performance Evaluation Review*, volume 38, pages 357–358. ACM, 2010.

[81] Adam Wierman Zhenhua Liu Iris Liu and Hamed Mohsenian-Rad. Opportunities and challenges for data center demand response.

[82] 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, 2011.

[83] 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, 2011.

[84] Priya Mahadevan, Puneet Sharma, Sujata Banerjee, and Parthasarathy Ranganathan. A power benchmarking framework for network devices. In *NETWORKING 2009*, pages 795–808. Springer, 2009.

[85] Ajay Mahimkar, Angela Chiu, Robert Doverspike, Mark D Feuer, Peter Magill, Emmanuil Mavrogiorgis, Jorge Pastor, Sheryl L Woodward, and Jennifer Yates. Bandwidth on demand for inter-data center communication. In *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, page 24. ACM, 2011.

[86] Grzegorz Malewicz, Matthew H Austern, Aart JC Bik, James C Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 135–146. ACM, 2010.

[87] Jennifer Mankoff, Robin Kravets, and Eli Blevis. Some computer science issues in creating a sustainable world. *iEEE Computer*, 41(8):102–105, 2008.

[88] David Meisner, Christopher M Sadler, Luiz André Barroso, Wolf-Dietrich Weber, and Thomas F Wenisch. Power management of online data-intensive services. In *Computer Architecture (ISCA), 2011 38th Annual International Symposium on*, pages 319–330. IEEE, 2011.

[89] David Meisner and Thomas F Wenisch. Stochastic queuing simulation for data center workloads. In *Exascale Evaluation and Research Techniques Workshop*, 2010.

[90] R. Miller. Green data centers. data center knowledge. http://www.datacenterknowledge.com/archives/category/infrastructure /green-data-centers/, 2011.

[91] Amir-Hamed Mohsenian-Rad and Alberto Leon-Garcia. Energy-information transmission tradeoff in green cloud computing. *Carbon*, 100:200, 2010.

[92] E. motor sport. Ev construction, thundersky batteries. http://www.electricmotorsport.com/store/ems_ev_parts_batteries.php.

[93] Ripal Nathuji, Karsten Schwan, Ankit Somani, and Yogendra Joshi. Vpm tokens: virtual machine-aware power budgeting in datacenters. *Cluster computing*, 12(2):189–203, 2009.

[94] Kong Soon Ng, Chin-Sien Moo, Yi-Ping Chen, and Yao-Ching Hsieh. Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries. *Applied energy*, 86(9):1506–1511, 2009.

[95] Kim Khoa Nguyen, Mohamed Cheriet, Mathieu Lemay, Bill St Arnaud, Victor Reijs, Andrew Mackarel, Pau Minoves, Alin Pastrama, and Ward Van Heddeghem. *Renewable energy provisioning for ICT services in a future internet*. Springer, 2011.

[96] US Department of Energy. Grid-tie energy efficiency. www1.eere.energy.gov/solar/review_meeting/pdfs/prm2010_apollo.pdf, 2010.

[97] 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.

[98] Henrik Petander. Energy-aware network selection using traffic estimation. In *Proceedings of the 1st ACM workshop on Mobile internet through cellular networks*, pages 55–60. ACM, 2009.

[99] S. Pileri. Energy and communication: Engine of the human progress. INTELEC 2007 keynote talk, Sep 2007.

[100] J Recas Piorno, Carlo Bergonzini, David Atienza, and T Simunic Rosing. Prediction and management in energy harvested wireless sensor nodes. In *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on*, pages 6–10. IEEE, 2009.

[101] Global Action Plan. An inefficient truth. *Global Action Plan Report, http://globalactionplan. org. uk*, 2007.

[102] Emerson Network Power. Rectifier energy efficiency. http://www.emersonnetworkpower.com/en-US/Brands/EnergySystems/Pages/ensys_eSureRectifiers.aspx.

[103] PowerSonic. Technical manual of la batteries. http://www.power-sonic.com/technical.php.

[104] Annabelle Pratt, Pavan Kumar, and Tomm V Aldridge. Evaluation of 400v dc distribution in telco and data centers to improve energy efficiency. In *Telecommunications Energy Conference, 2007. INTELEC 2007. 29th International*, pages 32–39. IEEE, 2007.

[105] Hao Qian, Jianhui Zhang, Jih-Sheng Lai, and Wensong Yu. A high-efficiency grid-tie battery energy storage system. *Power Electronics, IEEE Transactions on*, 26(3):886–896, 2011.

[106] 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.

[107] Daler Rakhmatov, Sarma Vrudhula, and Deborah A Wallach. Battery lifetime prediction for energy-aware computing. In *Proceedings of the 2002 international symposium on Low power electronics and design*, pages 154–159. ACM, 2002.

[108] 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, 2010.

[109] 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 *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE, 2010.

[110] Sergio Ricciardi, Davide Careglio, Francesco Palmieri, Ugo Fiore, Germán Santos-Boada, and Josep Solé-Pareta. Energy-aware rwa for wdm networks with dual power sources. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[111] Peng Rong and Massoud Pedram. An analytical model for predicting the remaining battery capacity of lithium-ion batteries. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 14(5):441–451, 2006.

[112] RUBiS. http://rubis.ow2.org.

[113] B Saha and K Goebel. Battery data set, nasa ames prognostics data repository, 2007.

[114] Ismael Sanchez. Short-term prediction of wind energy production. *International Journal of Forecasting*, 22(1):43–56, 2006.

[115] Ananth Narayan Sankaranarayanan, Somsubhra Sharangi, and Alexandra Fedorova. Global cost diversity aware dispatch algorithm for heterogeneous

data centers. In *ACM SIGSOFT Software Engineering Notes*, volume 36, pages 289–294. ACM, 2011.

[116] SmartGauge. Peukert's law equation and its explanation. http://www.smartgauge.co.uk/peukert.html, 2011.

[117] Energy Star. Uninterruptible power supply energy efficiency values. www.energystar.gov/index.cfm?c=specs.uninterruptible_power_supplies.

[118] 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.

[119] Maciej Swierczynski, Remus Teodorescu, and Pedro Rodríguez Cortés. Lifetime investigations of a lithium iron phosphate (lfp) battery system connected to a wind turbine for forecast improvement and output power gradient reduction. 2008.

[120] J. Taneja, D. Culler, and P. Dutta. Towards cooperative grids: Sensor/actuator networks for renewables integration. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 531–536, Oct 2010.

[121] Chunqiang Tang, Sunjit Tara, R Chang, and Chun Zhang. Black-box performance control for high-volume non-interactive systems. In *Proceedings of USENIX ATC*, 2009.

[122] Franco Travostino, Paul Daspit, Leon Gommans, Chetan Jog, Cees De Laat, Joe Mambretti, Inder Monga, Bas Van Oudenaarde, Satish Raghunath, and Phil Yonghui Wang. Seamless live migration of virtual machines over the man/wan. *Future Generation Computer Systems*, 22(8):901–907, 2006.

[123] R Tucker, Jayant Baliga, Robert Ayre, Kerry Hinton, and W Sorin. Energy consumption in ip networks. In *ECOC Symposium on Green ICT*, 2008.

[124] Jagannathan Venkatesh, Baris Aksanli, Jean-Claude Junqua, Philippe Morin, and T Simunic Rosing. Homesim: Comprehensive, smart, residential electrical energy simulation and scheduling. In *Green Computing Conference (IGCC), 2013 International*, pages 1–8. IEEE, 2013.

[125] Di Wang, Chuangang Ren, Anand Sivasubramaniam, Bhuvan Urgaonkar, and Hosam Fathy. Energy storage in datacenters: what, where, and how much? *ACM SIGMETRICS Performance Evaluation Review*, 40(1):187–198, 2012.

[126] Rui Wang, Nagarajan Kandasamy, Chika Nwankpa, and David R Kaeli. Datacenters as controllable load resources in the electricity market. In *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*, pages 176–185. IEEE, 2013.

[127] Brian J. Watson, Amip J. Shah, Manish Marwah, Cullen E. Bash, Ratnesh K. Sharma, Christopher E. Hoover, Tom W. Christian, and Chandrakant D. Patel. Integrated design and management of a sustainable data center. In *ASME 2009 InterPACK Conference,*, pages 635–644, July 2009.

[128] Molly Webb. Smart 2020: Enabling the low carbon economy in the information age. *The Climate Group. London*, 1(1):1–1, 2008.

[129] Windsun. Lead-acid batteries: Lifetime vs depth of discharge. http://www.windsun.com/Batteries/Battery_FAQ.htm, 2009.

[130] Ming Xia, Massimo Tornatore, Yi Zhang, Pulak Chowdhury, Charles Martel, and Biswanath Mukherjee. Greening the optical backbone network: A traffic engineering approach. In *Communications (ICC), 2010 IEEE International Conference on*, pages 1–5. IEEE, 2010.

[131] Yanwei Zhang, Yefu Wang, and Xiaorui Wang. Capping the electricity cost of cloud-scale data centers with impacts on power markets. In *Proceedings of the 20th international symposium on High performance distributed computing*, pages 271–272. ACM, 2011.

[132] Yi Zhang, Pulak Chowdhury, Massimo Tornatore, and Biswanath Mukherjee. Energy efficiency in telecom optical networks. *Communications Surveys & Tutorials, IEEE*, 12(4):441–458, 2010.