THE UNIVERSITY OF CHICAGO

ELIMINATING THE CAPACITY VARIATION PENALTY FOR CLOUD RESOURCE MANAGEMENT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
CHAOJIE ZHANG

CHICAGO, ILLINOIS
MARCH 2023

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES ix

ACKNOWLEDGMENTS x

ABSTRACT xi

1 INTRODUCTION 1
1.1 The Cloud, Growth, and Ensuing Challenges 1
1.1.1 Rise of Cloud Computing 1
1.1.2 Data Centers as Large Power Consumers 2
1.1.3 Cloud Growth, Damage, and Ensuing Limits 3
1.2 The Opportunities of Variable Resource Capacity 5
1.2.1 Carbon Reduction 5
1.2.2 Grid Decarbonization and Renewable Generation 6
1.2.3 Power Cost Saving 7
1.2.4 Power Capacity Constraint 9
1.2.5 Motivating Example 10
1.3 Problem: Variable Capacity and Computing Productivity 12
1.4 Problem Summary 14
1.5 Thesis Statement 14
1.6 Thesis Project 15
1.7 Contribution and Thesis Organization 16

2 BACKGROUND 19
2.1 Power Grids and Decarbonization 19
2.2 Rapid growth of Datacenter Power Load 21
2.3 Batch and HPC Resource Management 22
2.4 Cloud Resource Management 24
2.5 Workloads 26

3 RELATED WORK 28
3.1 Resource Management and Scheduling with Unreliable Resources 28
3.2 Large-scale Power Management and Power Capping 29
3.3 Renewable Energy and Optimizing for Green Power Use 31
3.4 Coupling Resource Management with Power Grids 32
3.5 Managing Resource Revocation 33

4 PROBLEM AND APPROACH . . . 35
4.1 Problems . . . 35
4.1.1 Scheduling Problem Definition of Variable Capacity . . . 35
4.1.2 Challenges of Job Scheduling . . . 37
4.2 Scheduling Approach . . . 40
4.2.1 Characterizing Scheduling Performance . . . 40
4.2.2 Coping with Capacity loss and Preparing for Capacity Variation . . . 43
4.3 Summary . . . 45

5 UNDERSTANDING CAPACITY VARIATION . . . 46
5.1 Methodology . . . 46
5.1.1 Dimensions of Capacity Variation . . . 46
5.1.2 Job Scheduling . . . 47
5.1.3 Workloads . . . 47
5.1.4 Job Scheduling . . . 49
5.1.5 Systems . . . 51
5.1.6 Metrics . . . 52
5.2 Dimensions of Capacity . . . 53
5.2.1 Dynamic Range . . . 54
5.2.2 Variability Structure . . . 54
5.2.3 Change Frequency . . . 56
5.2.4 Summary . . . 59
5.3 Cloud Workload Drill Down . . . 59
5.3.1 Variation Ranges . . . 60
5.3.2 Job Dependencies . . . 62
5.3.3 Workload Mixes . . . 63
5.3.4 Drilldown Takeaways . . . 65
5.4 Real Variation Scenarios . . . 65
5.4.1 Variation from Price . . . 66
5.4.2 Variation from Carbon Emissions . . . 67
5.4.3 Variation from Stranded Power . . . 69
5.4.4 Scheduling Experiments on Real Variation Traces . . . 69
5.5 Summary . . . 72

6 COPING WITH CAPACITY LOSS . . . 73
6.1 Intelligent Termination Policy . . . 73
6.2 Foresight . . . 79
6.3 Case Study: A German Datacenter . . . 80
6.4 Summary . . . 84

7 A BROADER VIEW: PREPARING FOR CAPACITY VARIATION . . . 85
7.1 Uncertainty and Information Space . . . 86
7.1.1 Workload Information . . . 87
7.1.2 Capacity Information . . . 88

7.2 Scheduling Algorithms . . . 90
7.2.1 Other Techniques . . . 92
7.3 Evaluation . . . 93
7.3.1 Information Space . . . 93
7.3.2 Capacity Foresight . . . 94
7.3.3 Variation Range . . . 96
7.4 Summary . . . 99

8 SUMMARY AND FUTURE DIRECTIONS . . . 100
8.1 Summary . . . 100
8.2 Future work . . . 101
8.2.1 Different Types of Prediction/Foresight . . . 101
8.2.2 Flexible or Optional Workload . . . 102
8.2.3 Multi-datacenter Integration . . . 103
8.2.4 Grid Interaction . . . 105
8.2.5 Headroom Analysis and Seasonal Optimization . . . 106

REFERENCES . . . 108

LIST OF FIGURES

1.1 Explosions of Internet-scale applications with the rapid growth of demand during the global pandemic has accelerated the expansion of cloud data centers 2
1.2 ICT energy forecast[84] 4
1.3 Dynamic Carbon Content in Power Market and Volatile Grid Renewable Generation (right) give rise to variable power [72] – and thereby variable capacity for scheduler domains [122]. A collection of these domains/zones can be within a single data center or span several buildings at a single site. 6
1.4 A case study to compare a 20 MW data center (placed at bus 2) to large-scale storage in a 47 bus Southern California Edison (SCE) distribution network as a function of data center flexibility[143]. 7
1.5 Today's data centers assume resource capacity as a fixed quantity. Emerging approaches to exploit grid renewable energy and reduce carbon emissions give rise to variable power. 8
1.6 Resource capacity variation for a datacenter with a *fixed per-hour carbon budget* (left). As the power grid's generation mix varies, the data center's capacity varies. The resulting capacity is 10% greater for a fixed carbon budget (right). Example from the Germany electricity market on 13.03.2020[70]. 10
1.7 Monthly average quantity of resource capacity under variation for a datacenter with a *fixed per-hour carbon budget* in 2021 - 2022. As the power grid's generation mix varies, the data center's capacity varies, producing total capacity different across seasons and grids. The yearly opportunity is 1.6% - 19.8% greater for a fixed carbon budget. Example from MISO, ERCOT, and CAISO power grids over Aug 2021 - July 2022. 11
1.8 Variable Capacity Challenge Scheduler's Ability to Achieve High Goodput and Low Waiting Time vs. Fixed Capacity 12

4.1 Scheduler goodput for a variable capacity data center; increased dynamic range degrades goodput. 38
4.2 Job Scheduling in A Fixed Capacity Data Center vs. in A Variable Capacity Data Center 38
4.3 How capacity variation affects scheduler goodput is a key to reducing carbon emissions in our scheme. 40
4.4 A variety of publicly available workloads and the corresponding scheduling space we explore to understand scheduling performance in the face of variable resource capacity. 41
4.5 Space of resource variability (left) and dimensions illustrated on a variable capacity example (right). 42

5.1 Job parallelism and runtime in the workloads 51
5.2 HPC (Dedicated) Scheduler performance varying dynamic range and varying stepsize for the random walk. 53
5.3 Scheduling performance with random walk and random uniform resource variability structure, varying dynamic range. 55

5.4 Scheduling performance with random walk and random uniform resource variability structures, varying dynamic range. 56
5.5 Goodput versus change frequency (dynamic range 0.6: [0.4, 1.0]). 56
5.6 Goodput versus change frequency, varying dynamic range and structure of capacity variation. 57
5.7 Goodput versus change frequency, varying dynamic range and structure of capacity variation. 58
5.8 System Performance of Google's (Cloud) Borg TNG workloads using Dedicated Scheduling Model across 5 Borg Cells Varying Dynamic Range 60
5.9 System performance Varying Variation Ranges 61
5.10 System Performance of Traditional Scheduler of Independent Workloads and Same Workloads with Job Dependencies 62
5.11 System Performance of Traditional Scheduler Varying Workload Mixes of Long and Short Job Fraction (Note: Google workload long job fraction indicated by the red box) 64
5.12 Power price ($/MWh) (left) and resulting resource capacity for a 200-megawatt data center (right), using a constant cost purchase approach. Exemplar 24-hour day from MISO January 9, 2018, CIN.Markland grid node. 66
5.13 Carbon-emission rate (left) and resulting resource capacity at Constant Carbon purchase approach in the German power grid(December 2019, right). 67
5.14 Statistics of grids' historical carbon emission rate, variable capacity produced by constant carbon budget, and examplar variation traces (MW) 69
5.15 Stranded Power (curtailed and negative priced power) in 15-minute intervals for a node in the ERCOT power grid (left, each line is a different week), and the resulting resource capacity for a 200 megawatt datacenter for the week in December (right). 70
5.16 System Performance (Goodput) of Traditional Scheduler with Real Variation Traces, Comparing with Synthetic Random Walk of Range 0.2, 0.4, and 0.6 . . . 70

6.1 Coping with Capacity Loss Through Intelligent Termination Policies in A Variable Capacity Data Center 73
6.2 Job failure rate with resource capacity variation, varying dynamic range and variability structure. 74
6.3 Scheduling performance for various termination policies (random walk, dynamic range [0.4, 1.0], step size 0.05) 75
6.4 Goodput versus termination policy, varying dynamic ranges and structures . . . 76
6.5 Job failure rate versus termination policy, varying dynamic ranges and structures 77
6.6 Job terminations sorted by runtime. Workload distribution as reference. (Random walk, dynamic range 0.6:[0.4,1.0]) 78
6.7 Goodput versus advance warning (random walk, dynamic range 0.6: [0.4, 1.0]). . 79
6.8 Goodput versus warning time (foresight), varying dynamic range and structure. 80
6.9 Job failure rate versus warning time (foresight), varying dynamic range and structure. 81

6.10 24-hour resource capacity variation by Carbon-Emission-Aware approach for acquiring power of an exemplar day per month 82
6.11 Goodput for 12 exemplar days, comparing fixed and carbon-aware power consumption, various schedulers, Mira trace, and simulation. 82
6.12 Performance and Carbon emissions of a model German Datacenter 83
6.13 Resulting power cost of a model German Datacenter 83

7.1 Preparing for Capacity Variation Through Scheduling Schemes in A Variable Capacity Data Center 85
7.2 Variable capacity challenges scheduler's ability to achieve low waiting time . . . 86
7.3 System performance (Goodput, Job Failure Rate, Job Waiting Time) of scheduling schemes without foresight 93
7.4 System performance (Goodput, Job Failure Rate, Job Waiting Time) of scheduling schemes, coupled with foresight of 0 (None), 6, 24 hours 95
7.5 System performance (Goodput, Job Failure Rate, Job Waiting Time) of Scheduling Schemes Varying Variation Ranges 97

LIST OF TABLES

5.1 Key Statistics for Widely Used Public Workload Traces 49
5.2 Key Trace Statistics for Workload Used in Simulation 49

7.1 Source of Uncertainty and Information Space in Variable Capacity Data Center 87
7.2 Information Space and Corresponding Scheduling Algorithms Exploiting Information to Prepare for Capacity Variation 90
7.3 Tolerable degradation of goodput performance and the corresponding range of acceptable variation 98

ACKNOWLEDGMENTS

I would like to take this opportunity to express my deepest gratitude to my advisor Prof. Andrew A. Chien for guiding me patiently through my Ph.D journey. He guides and shapes me for critical thinking, open mindset, and rigorous research.

I am also thankful to my other committee memebers, Prof. Hank Hoffmann, Prof. Sanjay Krishnan, and Prof. Varun Gupta, for serving on my doctoral committee. They provided invaluable feedback on my work.

I am truely grateful for my friends and team members for their generous help on countless matters, inspiring collaborations, and moral support.

Finally, I want to thank my parents, Feimin Zhang and Qimei Cao, and my grandparents, Yanggao Zhang and Yayuan Xie, for their unconditional love.

ABSTRACT

Increasing power grid challenges due to rapid decarbonization and pressure for reduced carbon emissions and power cost compel data centers to operate with capacity varying in periods of hours or days, perhaps on a dynamic basis in concert with the use of renewable generation. With data centers exceeding 10% of load in many grids, the implied capacity variation may approach 50%. For today's computing, variable resource capacity is problematic, causing severe loss in throughput and corresponding resource efficiency.

Our approach is to create intelligent resource management for variable capacity resources. Traditional resource managers were built with the assumption of constant capacity, scheduling jobs that fail when capacity decreases, causing abrupt job failures and wasted resources. To understand scheduling performance under variable capacity, we define three key dimensions of variation that lead to performance loss. We use cloud and HPC production workloads and explore the multi-dimensional capacity change space, characterizing scheduler performance in resource efficiency, job failures, and waiting time. Moreover, to improve performance, we consider scheduling techniques to cope with capacity loss. We propose intelligent termination policies to minimize job failures and wasted resource efficiency. Then, we take a broader view to prepare for capacity variation altogether. We consider two dimensions of uncertainty in capacity and workload, exploring the corresponding information space that reduces uncertainty. We propose new scheduling techniques that exploit the information to prevent job failures and increase resource efficiency.

We evaluate traditional schedulers under varying resource capacities and using a diverse set of workloads, including one HPC and three cloud workloads. Results show that capacity variation can decrease goodput by up to 60%, incurring 15-40% job failures. Amongst variability dimensions, the results show that dynamic range, structures, and change frequency are all important; each in some cases producing 10 - 40% goodput losses. Drill down with Google cloud workloads shows that variable capacity can cause serious problems, including

up to 70% goodput loss, 20% job failures, and 15X increase in job wait time. Careful study of performance versus variability shows that avoiding major harm, such as goodput loss, requires a variation limit of <10% dynamic range. This prevents the cloud from significant temporal load shifting to reduce carbon emissions or power costs.

We designed and compared the performance of intelligent termination policies to cope with capacity loss considering a variety of workloads and variation traces. Our experimental results demonstrate that these new scheduling techniques achieve significant performance improvements under resource variability, with 10 - 66% goodput increase and 1.6 - 3x job failure reduction. Using job attributes and progress to minimize wasted computation produces 44% goodput increase on average and close to full reduction on job failures. Realistic examples show that with scheduling techniques, a typical data center can achieve benefits of up to 15% carbon emission reduction and 14% power cost savings by exploiting resource capacity variations. Then, we take a broader view and design new scheduling schemes that seek to prepare for variation with Google cloud workloads which represents a hard case. These new schedulers exploit a variety of potential information about workload and capacity variation to reduce uncertainty, increasing goodput by up to 180%, decreasing job failure rate by 5 - 15X, and job waiting time by 1.4 - 4X. Within the information space, runtime classification is critical. Exploiting this information, the *LongShort* algorithm can drastically improve the ability to support variation in capacity from <10 to 50% while maintaining performance. These results demonstrate the promising benefits of new scheduling schemes for capacity variations but require future validation with complex workload constraints.

While capacity variation poses serious challenges to conventional resource managers, our intelligent resource management shows significant improvement, eliminating the variation penalty and demonstrating promising benefits of future variable capacity data centers.

CHAPTER 1

INTRODUCTION

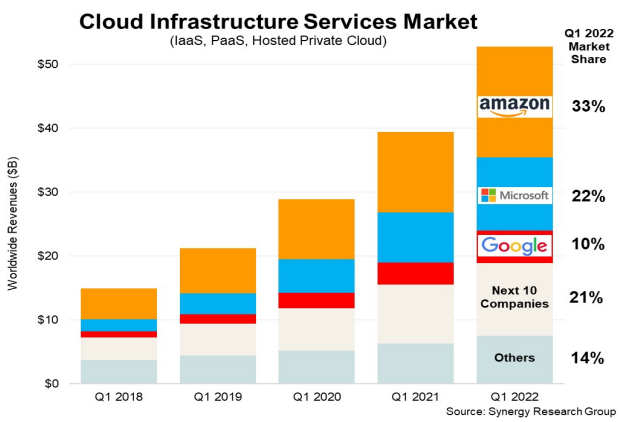
1.1 The Cloud, Growth, and Ensuing Challenges

1.1.1 Rise of Cloud Computing

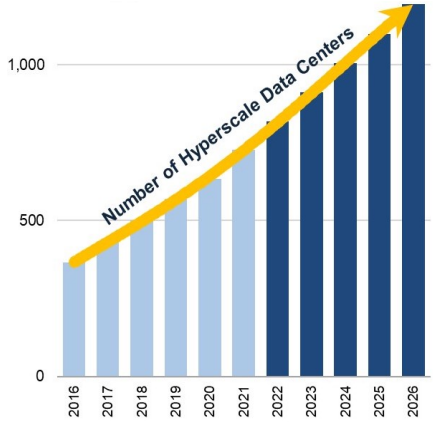
Cloud computing offers a platform for delivering elastic services over the Internet for a large number of users and with the use of hardware and software, thereby enabling scalability and resource-sharing. It offers many advantages over building out and managing a private infrastructure with instant use anytime anywhere, low upfront costs, and pay-as-you-go pricing. Since the incorporation of Salesforce and VMware in 1999 and the commercialization of Amazon EC2 in 2006[10], emerging cloud computing has become one of the most compelling paradigms and evolved to offer many branches of services, such as Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Large IT enterprises in play are dominating the cloud markets. Four companies own 67% of the world's $130 billion cloud market, and each has 60 or more data center locations. Overall, the eight largest providers control more than 80% of the market[3].

The explosion of data and data-driven workloads, such as big data analytics and query processing, has fast-tracked the growth of cloud capacity. With the increasing demand for machine learning applications and IoT services, this rapid growth has accelerated. With the global pandemic pushing companies to reshape their IT and application strategies radically by shifting to cloud services, this further prospers the growth of cloud computing. In addition to remote work, digital events such as video conferencing and streaming services create huge demand for cloud-based services. Infrastructure and platform cloud services grew 36% to $44 billion in the first quarter of 2022[22].

To meet the rapid growth of demand in cloud computing, IT companies strive to both increase the number of data centers all over the world and expand the size of data centers,



(a) Cloud revenue grows at 34% per year, and the top 3 providers account for 65% of total share[11]



(b) Number of hyperscale data centers projected to exceed 1,000 within three years[20]

Figure 1.1: Explosions of Internet-scale applications with the rapid growth of demand during the global pandemic has accelerated the expansion of cloud data centers

building hyperscale data centers which are significantly larger than traditional enterprise data centers in scale and performance. With a current pipeline of 314 new data centers under construction, the number of hyperscale data centers will exceed 1,000 in three years' time, and total capacity will double within less than four years[19, 12]. IDC projects worldwide spending on cloud infrastructure to have a CAGR of 12% till 2026, accounting for 67.9% of total infrastructure spend of compute and storage[13].

1.1.2 Data Centers as Large Power Consumers

The power requirements for cloud data centers have been growing rapidly (25-30%/year). It is no exaggeration to say that power is an important concern for large-scale cloud data centers.

The scale of cloud computing infrastructure, and its rapid growth reflects rapid commercial growth. Along with revenue growth, hyperscale cloud provider's corollary growth in power consumption rises from 10TWh in 2010 6.5x to 65 TWh per year in 2018 [97]. Based on a compound annual growth rate of 30%, [97] projects a potential rise to over 100

TWh/year in just the next few years. This growth can be seen in increasing numbers of data centers all over the world, which are causing increased power grid buildouts around the world. Within the data centers, sophisticated power management systems optimize how the power is used, according to metrics such as power usage effectiveness (PUE) [40]. Increasing the challenge of warehouse-scale computer design is the trend of increasing server power density[116]. Overall, hyperscale cloud data centers have become the fastest-growing consumer of electric power in many parts of the world. But worse than the cloud, power limits are constraining the scale of the world's largest supercomputers [47] and already define datacenter size. With the largest supercomputers approaching 50MW, datacenter complexes with multiple 40MW buildings, and aggregate loads of in excess of 200MW, with sites planned for 1GW, these are large power consumers indeed [38].

1.1.3 Cloud Growth, Damage, and Ensuing Limits

These power limits translate directly into limits on the amount of cloud computing that can be delivered. Compounding this, applications of computing drawn from every corner of commerce, society, science, and government [47, 152] are proliferating. The global cloud computing market is expanding at a rapid speed, with a projection to grow from $480.04 billion in 2022 to $1,712.44 billion by 2029, at a CAGR of 19.9%[4]. This growth, combined with climate concerns, has brought increasing attention to the cloud's power consumption and its environmental impact [85, 97, 72, 63]. Data centers consume 196 to 400 terawatt-hours (TWh) in 2020[1] and are forecast to reach 20.9% of projected electricity demand in Figure 1.2[84]. With the projected growth of energy consumption and expansion of hyperscale data centers, cloud computing already accounts for 2-4% of global greenhouse gas emissions with high growth rates[16].

Concerned about climate, governments around the world have adopted policies to reduce carbon emissions, increasing pressure to minimize data center power consumption and

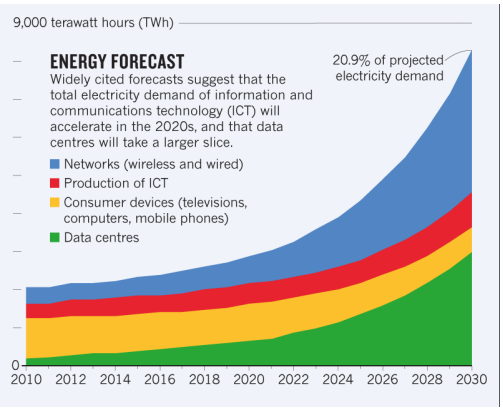


Figure 1.2: ICT energy forecast[84]

encouraging the use of renewable generation. For example, European Union goals include dramatic reduction of carbon emissions for the entire economy – 40% by 2030 and zero net by 2050[45]. In the United States, large states have adopted similar zero net carbon emissions goals for electric energy (California 2045) and for the entire economy (New York 2045) [98, 110, 105]. At the same time, wholesale electricity prices are skyrocketing in many countries, resulting in tripling in many markets in 2021-2022. This signals the continued energy crisis and large uncertainties of electricity forecast going forward[7], posing even more significant challenges to carbon reduction goals and power cost budgets of entities. These societal targets pose significant challenges for rapid hyperscale cloud growth (e.g. Amazon, Microsoft, Google, etc.) that is being accelerated further by exploding popularity of machine learning [79, 131]. In several areas of the United States, data centers already account for over 5% of the power load [120], and hyperscale power consumption growth is estimated at 20 to 40% CAGR. Several cloud computing providers have responded with aggressive goals to reduce carbon emissions – notably CEO's Satya Nadella of Microsoft, and Sundar Pichai of Google, and even Jeff Bezos of Amazon[33, 31, 32], but there is much work to be done in the face of rapid growth of the cloud.

Due to the significant amount of power footprint, cloud data centers now are facing serious power constraints that limit their long-term growth. With the expansion of hyperscale

data centers outrunning local utility grids, many countries and regions have to halt future constructions and projects simply because the local grid cannot keep up. Various locations, such as London, Ireland, and Singapore, are facing the same challenge that their electricity grids are at capacity and are having difficulties in building new infrastructures because of grid stability[21, 9].

1.2 The Opportunities of Variable Resource Capacity

These power limits make dynamic power management for cost, cooling, sharing, or simply to be a good citizen in a fluctuating or stressed power grid a source of variable capacity for data centers.

1.2.1 Carbon Reduction

Concern about the carbon footprint has led to significant public scrutiny from organizations such as Greenpeace [75], and a drive by many cloud providers to offset their carbon footprint (become carbon neutral), with Google, Facebook, Microsoft leading the charge in that area and recently Amazon agreeing to that as a long-term goal. In late 2018, Google raised the bar, adopting a goal beyond offsetting, matching its power consumption on an hourly basis, 24x7 over the entire year, with renewable energy in the same power grid[72]. The combination of the goals of extremely high energy use efficiency and reducing carbon footprint data centers lead to careful but aggressive large-scale power management. Recent studies suggest a growing trend of power management and sharing over large scheduling domains[122, 91, 149], and these large-scale power management creates dynamic power constraints as variable resource capacity, shown in Figure 1.3 (left).

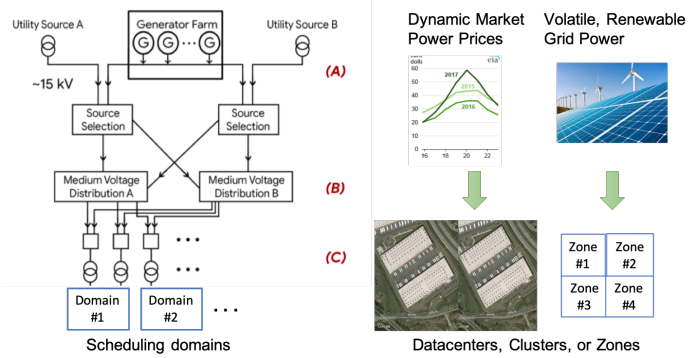


Figure 1.3: Dynamic Carbon Content in Power Market and Volatile Grid Renewable Generation (right) give rise to variable power [72] – and thereby variable capacity for scheduler domains [122]. A collection of these domains/zones can be within a single data center or span several buildings at a single site.

1.2.2 Grid Decarbonization and Renewable Generation

With ambitious goals to de-carbonize electric power generation in much of the world, power companies and grids have turned heavily to renewable sources such as solar and wind [76, 113]. The volatility of these resources, meaning their power is sometimes available and sometimes not, depends on the weather conditions and time of day. Because of this property, these resources are often characterized by a *capacity factor* such as 0.33 – the fraction of the nominal maximum generation that they provide over a full year.

Beyond this the challenge, renewables can be thought of as statistically available generators, so their power cannot be dispatched to be available when the grid loads need it – rather it just "is" available when the the wind is blowing or the sun is shining. The correlation and non-dispatchability of wind and solar generation result in diminishing benefits, "grid effective capacity factor" that diminishes with each additional unit of renewables added to the grid. This phenomena is well-documented and reflects a major challenge to high renewable fraction grids[144]. In fact, all of the high renewable fraction (RPS) grids are experiencing major challenges in these areas – at renewable fractions of 30-45%. So the drive to higher RPS is a significant challenge.

A critical solution is *adaptive loads*, which adjust their demand rapidly to match the

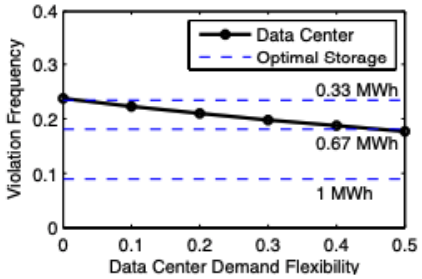


Figure 1.4: A case study to compare a 20 MW data center (placed at bus 2) to large-scale storage in a 47 bus Southern California Edison (SCE) distribution network as a function of data center flexibility[143].

available supply [67]. Such loads will be a staple of the future grid, because of their cost-effectiveness relative to energy storage. Dynamic power management of hyperscale data centers is an important potential adaptive load.

Existing studies show that a 20MW data center needs to provide 20% flexibility to be as effective as a 0.5MWh energy storage, and as much as 50% to be equivalent to 0.67MWh storage in terms of regulating and stabilizing the grid, shown in Figure 1.4[143]. This represents a sizable amount of variation range imposed on the data center.

Ambitious research has proposed new models of data center operation that synergize use and load with the grid, where ZCCloud represents a radical approach to operate with 100% dynamic range using zero-carbon power and low price [151, 55], or with the availability of local renewables [71, 80, 62] . These approaches all suggest that future data centers will have variable capacity, determined by external factors such as the general (grid-wide) or local (on-site) availability of renewable generated power.

1.2.3 Power Cost Saving

Even if carbon footprint is not compelling for all data center sites, power cost is always a concern. Global experience across dozens of major power grids has shown that increased renewable fractions (20%, 30%, 40%, etc.) produce growing swings in power pricing [144]. And,

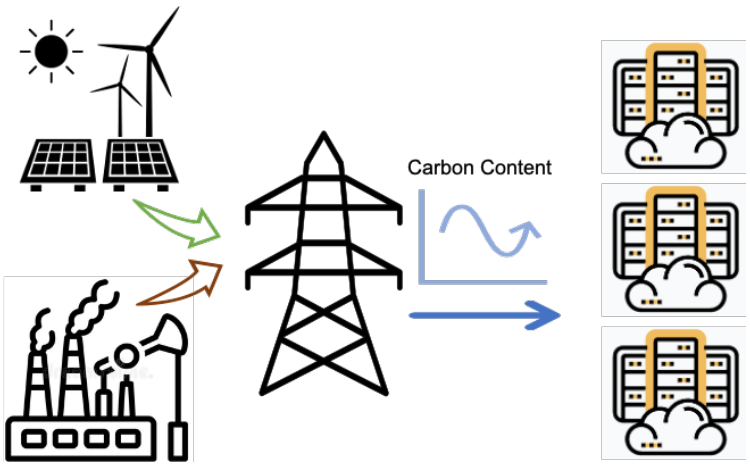


Figure 1.5: Today's data centers assume resource capacity as a fixed quantity. Emerging approaches to exploit grid renewable energy and reduce carbon emissions give rise to variable power.

when renewable generation coincides, supply can exceed demand, producing negative pricing and power curtailment (waste) in massive quantities [78, 76, 44, 55], leading to abundant opportunities to optimize data center power cost. This is a broad and perhaps unavoidable trend due to compounded uncertainties in forecasting these volatile generators and loads. Therefore, in addition to carbon footprint reduction, the goal to reduce power costs with fluctuating prices leads to complex local optimization and planning, driving variable capacity in data centers.

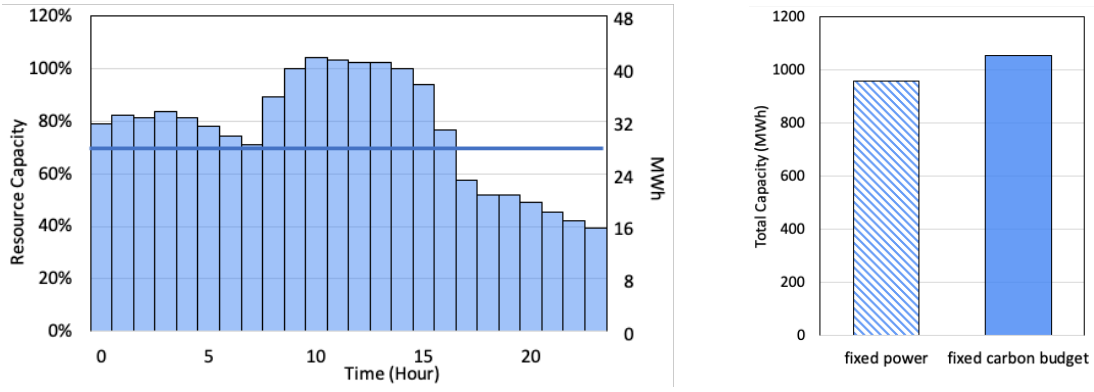
These varied scenarios suggest clusters, availability zones, scheduling domains, even entire data centers will have variable capacity, driven by external factors such as power allocation, market prices, or even general (grid-wide) or local (on-site) availability of renewable energy. This is the core motivation for the variable capacity resource scheduling problem. As shown in Figure 1.5, and external factor such as varying power creates variation in capability/capacity and the resource manager must effectively manage this varying capacity as it changes over time as in Figure 1.6a.

1.2.4 Power Capacity Constraint

In addition to benefits such as power cost savings, cloud data centers now are facing a major challenge of not being able to build new data centers and reaching a hard stop of the rapid growth brought out by power grids and governments. Ireland has recently imposed a de-facto moratorium on new data center construction as they have placed an unbearable strain on the power grid[9]. Major cloud enterprises, such as Microsoft and Google have to find alternative solutions as Ireland's grid operator has halted new constructions due to power constraints[17]. Likewise, Dominion Energy, the primary utility provider which serves close to 70 data centers as a global hub, can no longer guarantee the power demand due to overloaded transmission[5].

The reason behind the power constraint is data centers are viewed as large, steady power consumers. With the resilience and SLA requirements for cloud data centers, their power supply requires to be 24×7 available, posing severe stress to the grid during times of peak demand or in the event of failures. To account for peak use and guarantee power stability, regulators and grids need to conservatively plan for power capacity. One critical solution for data centers to address the problem is dynamically adjusting their power demand to meet power grids' needs. Large, flexible data center loads could not only reduce the stress on the grid but also become a great asset in stabilizing the grid and consuming excess power supply.

A group of studies has proposed peak shaving strategies for data centers to reduce demand during power grids' peak demand[35, 143]. These approaches demonstrate the feasibility of data center load shifting to stabilize the power grid. More promising, ERCOT has initiated Large Flexible Load Task Force (LFLTF) to allow large flexible loads, ranging from 20 - 75MW, such as crypto miners, into the grid connections as dispatchable loads[15]. These loads are introduced as a form of controllable resources, responding to the grid's demand and curtailment in a short time, adaptively adjusting their own power demand. ERCOT anticipates as much as 17GW of such large, flexible loads will interconnect with the grid by



(a) Hourly Capacity (b) Total Capacity

Figure 1.6: Resource capacity variation for a datacenter with a *fixed per-hour carbon budget* (left). As the power grid's generation mix varies, the data center's capacity varies. The resulting capacity is 10% greater for a fixed carbon budget (right). Example from the Germany electricity market on 13.03.2020[70].

2026[6], projecting a substantial existence in the future grid. These efforts demonstrate the opportunities for data centers to overcome the power capacity limits and continue to grow by offering flexibility to the power grid in times of stress, which manifests to the data centers as variable capacity.

1.2.5 Motivating Example

One promising reason for variable capacity data centers is to exploit the fluctuations in the power grid to reduce electricity costs and carbon footprint of data centers. Modern power grids include a complex mix of generators – wind, solar, hydro, as well as fossil-fuel and even nuclear. As load varies through the day or over the week, the power grid dynamically dispatches generators in an ever-changing mix to meet the current demand. While generally preference is given to renewables through economic dispatch because they have low incremental generation costs, they are not always available in sufficient quantity so carbon-emitting generators are used. This problem is much harder than most markets because power is what economists call a perishable resource – generation much be matched instantaneously with the load.

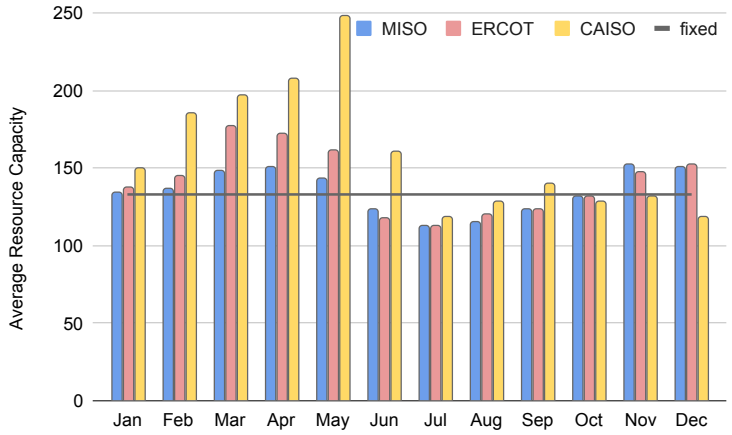
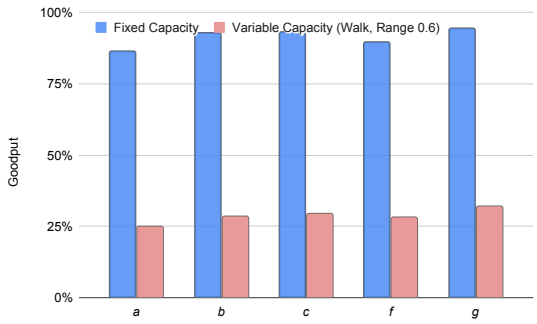


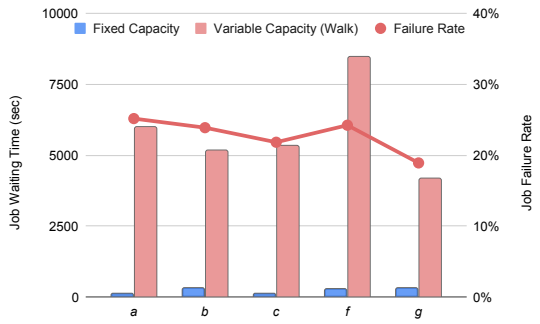
Figure 1.7: Monthly average quantity of resource capacity under variation for a datacenter with a *fixed per-hour carbon budget* in 2021 - 2022. As the power grid's generation mix varies, the data center's capacity varies, producing total capacity different across seasons and grids. The yearly opportunity is 1.6% - 19.8% greater for a fixed carbon budget. Example from MISO, ERCOT, and CAISO power grids over Aug 2021 - July 2022.

The net effect is that carbon-emissions content of power in most power grids varies widely with time. One example is illustrated in Figure 1.6a, where a fixed carbon emission budget for each hour, combined with the dynamic variation of power grid carbon content produces large variations in data center power available. The resulting variable power ranges from 16 to 42MW, compared to a constant power of 28MW. Not only does the level of power varies widely, but the carbon-emissions-based purchases also exploit this variation to buy a significantly larger quantity of power at the same level of carbon emissions. In Figure 1.6b, more than 10% capacity increase is observed over a 24-hour period. The example illustrates that following the carbon-emission content of the power market alone can produce resource capacities ranging from 40% to 105% within a single 24-hour period.

Now we expand the budget of fixed carbon emission to more grids and over one year, Figure 1.7 shows the average resource capacity achieved over the month of three major U.S. grids - MISO, ERCOT, and CAISO over the last year. The resulting total capacity of resource variations varies across months, seasons, and grids, but overall demonstrates a significantly larger quantity of power compared to a fixed capacity purchase with the same



(a) Goodput of Fixed vs. Variable Capacity



(b) Job Waiting Time of Fixed vs. Variable Capacity and Job Failure Rate under Variable Capacity

Figure 1.8: Variable Capacity Challenge Scheduler's Ability to Achieve High Goodput and Low Waiting Time vs. Fixed Capacity

amount of carbon emissions. Across three grids, capacity increases of 1.6% to 19.8% are observed over a one-year period, with an increase reaching as much as 56% over a one-month period. These examples show the great opportunities for adapting variable resource capacity and enlarging shifting capabilities in data centers. And as renewable generation increases, the carbon content in power grids will continue to exhibit increasingly larger differences, creating even greater benefits. Similarly, such resource capacity variation can arise from the dynamic power pricing of power markets, production quantity of local renewables, or intercluster power management that dynamically re-adjusts the power quantity available to each cluster.

1.3 Problem: Variable Capacity and Computing Productivity

However, variable resource capacity poses new scheduling challenges to the data center. Extensive research on job scheduling and resource management generally focuses on problems where the quantity of resources is fixed or constant. Changing resource capacity is a challenge for job schedulers and resource managers because of the uncertainty about future resource capacity. On one hand, this means that even if job runtime is known at start time, the

resources may not be available long enough to complete it. On the other hand, resources can increase rapidly, challenging the availability of workload to utilize them. Figure 1 portrays new data center scheduling and resource management paradigm where resource capacity varies over time, controlled by external factors such as the grid's carbon content, market power price, and inter-cluster power management.

To motivate the challenges of scheduling for variable capacity resources, consider Figure 1.8a, goodput (useful resource efficiency) achieved by a state-of-the-art scheduler, using one of the most widely used cloud workloads, comparing variable capacity with fixed capacity over 5 distinct clusters. The results show a large goodput decrease for all 5 clusters (varied workloads). Compared to traditional fixed capacity, the variation in capacity incurs a goodput degradation of 60 - 70%. The variation in this example is derived from typical power market variations as shown in Figure 1.6a[1]. Beyond goodput degradation, other negative effects include sharp increases in job failure rates, ranging from 50% to 70%, and large increases in job waiting times by up to 15x (see Figure 1.8b).

Walking through these simple illustrations, we see that there are certainly great opportunities to exploit carbon footprint and electricity cost benefits while maintaining or even increasing total available resources in data centers. However, such variations may induce significant negative impacts on the schedulers, and such impacts limit and prevent data centers from dynamically acquiring power and shaping the computation load accordingly. Therefore, characterization of the performance and effectively maintaining system performance under capacity variation is the key for the data center to enable or enlarge its shifting and load fluctuations capabilities to exploit power grid fluctuations.

In summary, the end of Dennard Scaling and the rapid growth of large-scale data centers advocate variable resource capacity to exploit carbon and cost benefits. Our approach is in line with these data center commitments and policy efforts to reduce carbon and power

1. This cluster's capacity varied by random walk with stepsize 0.15 and dynamic range of 0.6.

impact.

1.4 Problem Summary

We briefly describe the thesis problem in this section. Section 4.1 explains the problem in-depth with a detailed discussion.

As loads of cloud data centers grow to gigawatts, they have become dynamic entities, interacting with the power grid to optimize power cost, carbon emissions, grid stability, and computation. Such dynamism produces variable resource capacity, controlled by external factors. However, traditional data center resource managers have focused on time-invariant resource capacity and can suffer significant performance loss subject to such variability, limiting acceptable dynamic range. To enable larger benefits, new cloud needs new resource management techniques that can tolerate greater dynamic range of capacity variation, while maintaining good performance. We aim at creating intelligent resource management for variable resource capacity data centers to eliminate the variation penalty and exploit variation benefits.

1.5 Thesis Statement

We propose a new class of scheduling techniques that are robust in the presence of resource capacity variability. These techniques, exploiting the information space of uncertainty, can deliver high data center resource efficiency in variable capacity data centers for commercial public cloud workloads.

- **Variable capacity data centers** experience dynamic resource capacity, driven by external factors such as power allocation, power market prices, general (grid-wide) or local (on-site) availability of renewable energy, and intercluster power management.

These data centers experience multiple dimensions that shape resource variations, including variability structures, change frequency, and dynamic range.

- **Information Space of Uncertainty** explores dimensions of uncertainty including job runtime and resource capacity that lead to performance loss and includes varying amounts of information to help schedulers prepare for uncertainty.

- **Scheduling techniques** include intelligent termination policies to cope with capacity loss and variation-aware scheduling algorithms, effectively exploiting varying amounts of information about job runtime and resource capacity, to plan for capacity variations, with the goal of high data center utilization and no loss of QoS, increasing data center shifting capabilities.

1.6 Thesis Project

In this thesis, we propose intelligent resource management which exploits information of uncertainty for variable capacity resources to enable larger carbon emission and power cost benefits.

We define the scheduling problem of variable resource capacity and the multi-dimensional space of capacity variations. We define the space of variations with several key dimensions, dynamic range, variability structure, and change frequency. These correspond to a range of examples in the natural world (carbon content, power price, data center cooling, and more) that give rise to variation. We give examples covering a wide range of realistic scenarios and illustrate how varied and challenging they are. To study how well existing high-quality resource managers fare, we build a trace-driven job scheduling simulator and use four real large-scale workload traces, both cloud and HPC, covering a diverse range of synthetic and real variation traces. We explore the multi-dimensional capacity change space and characterize scheduler performance in resource efficiency, job failures, and waiting time under these

challenging scenarios. To drill down on the variation range of cloud data centers, we evaluate scheduler performance using Google cloud workloads varying workload mixes and structures with both synthetic and real variation traces.

To improve performance and cope with capacity loss, we propose intelligent termination policies that selectively terminate jobs to minimize goodput loss. We evaluate and compare these policies with foresight covering the whole problem space and with a case study to characterize improvements. Beyond coping with capacity loss, we identify two dimensions of uncertainty in capacity and workloads. We explore the information space, which reveals amounts of information about these dimensions, to reduce uncertainty and propose new scheduling schemes exploiting the information to prepare for capacity changes by optimizing job placements. We compare the performance of new scheduling schemes using Google cloud workloads with varying workload features and variation traces to demonstrate the generality of scheduling improvements.

1.7 Contribution and Thesis Organization

In this thesis project is a deeper understanding of the opportunities and challenges to make variable capacity resources useful. Specifically that the introduction of variation in capacity raises myriad problems in the effective use of data center resources. If useful coupling with external environments (eg. power grid, carbon emissions, power markets) is to be achieved, data centers must solve the problem of achieving high compute efficiency with variable resource capacity.

This problem's nature is dependent both on workload structure as well as the nature of variation. Variational studies show that realistic combinations face significant difficulties.

Exploration of a variety of termination policies as well as planning for change in capacity show promising benefits, but the properties of today's cloud workloads present a difficult challenge. We characterize how they limit any scheduling solutions, and thereby limit fea-

sible variation. Any greater solution will require changes in both workload and resource management systems. Specific contributions of the thesis include:

- Formal definition of the variable-capacity scheduling problem for data centers. This includes identifying the key dimensions of variation, and systematically characterizing their performance impact on current schedulers. This characterization covers multiple cloud and HPC workloads, and dimensions such as variation structure, range, frequency, as well as the impact of foresight.

- Empirical studies using real supercomputer center HPC workload traces and typical data center heterogeneity shows that capacity variation can significantly decrease goodput (15 - 60% with avg. 30%). Beyond such efficiency, job quality is also degraded, incurring job terminations (15 - 40% with avg. 25%). These studies show that several variability dimensions have the greatest negative impact – dynamic range, variability structure, and change frequency. Each of these dimensions could independently reduce goodput by 10 - 40%, with even greater losses in combination.

- Drill down on real cloud data centers with a range of workload traces and various cluster sizes reveals a dominant mode of VM usage with a large fraction of long or continuous running jobs, presenting an obstacle to resource flexibility. This property and inter-task dependencies further degrade performance, producing unacceptable goodput losses of 30-40%, and unconscionable job termination rate of 26%. These factors combine to produce a 60X increase in average waiting time. These impacts limit tolerable capacity variation to $<$10%, strictly limiting potential variation benefits.

- Scheduling techniques to cope with capacity loss by selectively terminating jobs using job attributes and progress to minimize wasted computation and improve resource efficiency. Evaluation of a range of workloads and variation ranges show that they are effective in mitigating performance degradation upon capacity loss, reducing job

failure rate by 2 - 5X, and enabling large increases in goodput by 44% on average.

- The framework of uncertainty in variable capacity data centers. This includes two dimensions of uncertainty in capacity and workloads which lead to performance loss, corresponding information space to reduce uncertainty, and new scheduling schemes which prepare for capacity changes by exploiting information to optimize job placements. This framework targets cloud data centers considering their workload properties.

- Experiments using real cloud workload traces show that with information, schedulers achieve significant improvements, increasing goodput (by 180%), decreasing job termination rate (by 5 - 15X), and job wait time (by 1.4 - 4X). Among information, job runtime classification is critical and enables scheduling algorithm to effectively achieves large load flexibility from <10% to 50% while maintaining performance.

The remainder of the thesis is organized as follows. Chapter 2 gives a brief background on recent advances in power grid decarbonizations, data center growth, resource management, and cloud workloads. In Chapter 3, we discuss the related research literature. In Chapter 4, we present the key research problem and our scheduling approach. Chapter 5 describes the scheduling performance under capacity variations with detailed multi-dimensional evaluation. In Chapter 6, we explain the two dimensions of uncertainty and the corresponding information space to reduce uncertainty. In Chapter 7, we describe our scheduling algorithms exploiting the information of uncertainty and evaluate their effectiveness. Finally, we summarize the thesis results and outline multiple future research directions in Chapter 8.

CHAPTER 2

BACKGROUND

2.1 Power Grids and Decarbonization

Growing concerns about carbon emissions and their long-term impacts on climate change have created a worldwide consensus on the transformation of power generation. Power grid decarbonization is one of the most fundamental pillars of the global effort to mitigate climate impacts. Globally, 30% carbon emissions have been contributed by coal-fired generation in 2018 and electricity has played an important role in achieving a carbon-neutral energy system by reducing carbon emissions.

Governments around the world have adopted policies to reduce carbon emissions, in particular a major promotion of renewable energy generation. For example, European Union goals include dramatic reductions of carbon emissions for the entire economy – 40% by 2030 and zero net by 2050 [45]. In the United States, large states have adopted similar zero net carbon emissions goals for electric energy (California 2045) and for the entire economy (New York 2045) [98, 110, 105]. These societal targets suggest that power grids need to project rapid growth in renewable electricity generation in response to these pledges and the goals of the Paris climate agreement[18].

In response to the actions from numerous governments to encourage the deployments and uses of renewable generation such as welfare, carbon taxes, and credits, the portion of renewable energy in the power grids has significantly boosted over the past years. The world's renewable energy power generation capacity has increased from 4,204 TWh in 2010 to 8,427 TWh in 2022, more than doubled. Wind and solar account for the major growth in renewable generation. In the US, together solar and wind comprised 5.2% of overall power generation in 2014. With an increase of 20 GW of the solar capacity forecast, EIA expects solar power to account for nearly half of new U.S. electric generating capacity in 2022[23].

EIA forecasts that wind and solar will provide 14% of U.S. electricity generation this year and 16% in 2023[14]. California has been a leader in setting Renewable Portfolio Standards (RPS), and requirements for power generation mix, reaching a 20% renewable mix in 2010. In September 2015, California adopted an RPS goal of 50% renewable by 2030. It surpassed RPS goals in 2020 with 34.5% of the state's electricity served by renewable sources such as solar and wind and a total of 59% came from renewable and zero-carbon sources in 2020. Other states across the midwest (included in the MISO power grid) have adopted a range of standards ranging from 25% (2015) in Illinois, 25-31% (2025) in Minnesota, and 55% (2017) in Vermont. Other large states include 50% by 2030 in New York, and 10GW by 2025 in Texas.

However, new challenges arose from the variability and uncertainty of renewable energy for electric grid operators which must continuously match variable supply with constantly changing demand. Unlike traditional generators such as fuel-based power plants whose electricity output is controllable, renewable generations often depend on variable factors that cannot be controlled. For example, wind and solar generation depend on sunlight and wind speed and such generation may come in bulk amounts or none at all. In the events of oversupply or congestion, the power grid market will reduce generation output, where plant generation is scaled back, through 1. economic curtailment, where price excess renewable electricity with zero or even negative prices, creating additional wastage from economic curtailment, 2. self-scheduled cuts, where reduce production from bids, 3. exceptional dispatch, when the ISO orders generators to turn down output. Such excess renewable capacity can be significant. For example, in 2015, the California ISO was forced to curtail more than 187,000 total megawatt-hours (MWh) of solar and wind generation. In 2022, that number rose to more than 1,504,000 MWh[2]. The total amount of curtailment has increased 8X from 2015 to 2022, suggesting that the challenges of variability and uncertainty drastically grow as power grids' decarbonization progresses to incorporate increased renewable genera-

tion such as wind and solar. Comparable waste and similar patterns of increase exist in other ISO's such as Eastern Region Coordinating District of Texas (ERCOT) and Mid-continent Independent System Operator (MISO), and many European countries such as Denmark, Germany, Ireland, and Italy.

These curtailment and uneconomic dispatch can be termed as stranded power. Various studies have focused on characterization of these stranded power in quantity and temporal structure[52, 54, 151, 88]. These works not only identify the significant quantity of stranded power in the grids now, but also observe a persistent phenomenon in future grids. These are the earliest studies that explore and characterize the dynamic variations (seasonal and time-of-day patterns) and the opportunities of stranded power in power grid decarbonization.

2.2 Rapid growth of Datacenter Power Load

Information and computing technologies (ICT) produced carbon emissions (8% of electric power in 2016) are growing most rapidly, and are projected to reach 13% by 2027 [64, 74]. Supercomputers and data centers are major elements of this consumption. With Dennard scaling long over [127, 46] power levels are growing rapidly: 15-petaflop systems exceeded 10 MW [106, 118] in 2017. Japan's newly announced 415 petaflop system exceeds 28 MW [49], and planned US DOE Exascale systems are expected to exceed 35 MW in 2021 [36, 107]. Centers often operate under power "caps", effective carbon footprint limits [41], or forced power reductions [89].

The rapid growth of cloud computing has produced a huge computing infrastructure with correspondingly large revenue – expanding at a rate of 20% per year to an estimated revenue of $331 billion in 2022[29]. Along with revenue growth, for hyperscale cloud provider's corollary growth in power consumption, rising from 10TWh in 2010 6.5x to 65 TWh in 2018 [97]. This article projects a potential rise to over 100 TWh in just the next few years. This rapid growth is manifest in increasing numbers of data centers all over the world, but also

in sophisticated power management systems that optimize how the power is used to deliver the most cloud computing by optimizing metrics such as power-use efficiency (PUE) [40]. Increasing the challenge of warehouse-scale computer design is the trend of increasing server power density[116]. These cloud data centers and supercomputers at extreme scales inspire radical approaches to scale data centers [134, 96, 53]. Overall, hyperscale cloud data centers have become the fastest-growing consumer of electric power in many parts of the world.

This large and growing power consumption produces several important problems. First, direct power cost, as well as implied costs in cooling and facilities is a growing problem. Second, the growing power use has significant associated environmental impacts [103, 104, 73] including growing carbon emissions.

Concern about the carbon footprint has led to significant public scrutiny from organizations such as Greenpeace [75], and a drive by many cloud providers to offset their carbon footprint (become carbon neutral), with Google, Facebook, Microsoft leading the charge in that area and recently Amazon agreeing to that as a long-term goal. In late 2018, Google raised the bar, adopting a goal beyond offsetting and adopting a goal of matching its power consumption on an hourly basis, 7x24 over the entire year, with renewable energy in the same power grid[72]. The combination of the goals of extreme, high energy use efficiency and reducing carbon footprint data centers lead to careful but aggressive large-scale power management. Recent studies suggest a growing trend of power management and sharing over large scheduling domains[122, 91, 149], and these large-scale power management creates dynamic power constraints as variable resource capacity.

2.3 Batch and HPC Resource Management

Resource management and job scheduling monitors and control resource usage, mapping jobs onto a set of machines, optimizing metrics such as makespan, job wait time, goodput, and resource utilization. While existing data centers deal with a great variety of workloads,

such as streaming and interactive jobs, batch workloads are an important workload. Many scientific computing resources serve large-scale computation as batch requests[37].

For example, the mainstream HPC cluster schedulers adopt the traditional batch job scheduler model and manage a long single queue. Users submit requests of job size in terms of a fixed amount of resources and job runtime estimates, generally overestimated, and the scheduler decides when and where to run the job request given the job waiting queue, current running workloads, and system availability based on different heuristics and policies. Common system-centric and job-centric optimization metrics include system utilization or throughput, job stretch, and job turnaround time. Slurm, Moab/Torque, and Cobalt are well-known and widely used[154, 133], offering high scalability and fault-tolerance in HPC environments. For example, Cobalt[133] is an open-source, component-based resource management package used on IBM Blue Gene systems. It uses utility function to prioritize jobs, which is similar to the popular policy scheduling mechanism from the Maui scheduler[83]. Its flexibility allows for easy modification and customization.

Conventional job scheduling algorithms include First Come First Serve (FCFS)[126], Shortest Job First (SJF)[77], Round-Robin (RR)[115], Min-Min[82], and Max-Min[43]. These heuristics aim at various goals, such as maximizing throughput, minimizing job stretch or response time, or maximizing fairness, and many studies have proposed hybrid algorithms to combine them[69, 66]. The most widely used scheduling policy is First Come First Serve (FCFS) which serves the job requests in order of their arrival time, which guarantees not only fast decision but also fairness[112]. Backfilling strategy, such as EASY backfilling, is widely adopted in addition to simple FCFS to enhance system utilization. Backfilling allows subsequent jobs in the queue to be moved and scheduled ahead if and only if they do not delay the existing requests[102].

A wide range of studies can be proposed to further improve various aspects of the data centers, such as increasing resource utilization, providing better support for heterogeneous

clusters, and special customization for hybrid workloads. A group of work focus on predicting the runtime of jobs than user runtime estimates to improve scheduling decisions[138, 128].

2.4 Cloud Resource Management

Modern cloud data centers are continually expanding their computing resources to meet growing needs for e-commerce, web search, social networking, enterprise IT, and big data analytics. With the enormous growth of services and as the complexity of applications multiply, cloud computing is widely-used across application domains and scenarios. Cloud computing allows data center infrastructure to be leased profitably to third parties and thus enable a pool of computing resources shared between applications and services that are accessed over the Internet[117]. Its growth has accelerated over the past years as it offers layers of abstraction to application designers and users without the careful maintenance and design of underlying infrastructure and resource management.

Cloud data centers, which are the providers of cloud services to cloud users, manage and allocate resources subject to performance guarantees, formally defined as Service-Level-Agreements (SLAs). Unlike traditional private clusters or supercomputers, cloud providers deploy virtualization software on physical machines for a couple of reasons. First, virtualization provides cloud data centers the flexibility to allocate arbitrary fractions of resources on physical machines on the fly to users. That is, a number of Virtual Machines (VMs) allocated on the same machines are isolated from each other and may run different operating systems, platforms, and applications. Therefore, virtualization provides flexibility, guarantees resource isolation to avoid contention and security problems, and reduces overhead for cloud data centers.

As cloud data centers serve requests for a wide range of applications and offer various features and services, they are exposed to a wide variety of workloads from both internal services and external customers. The first-party workloads comprise internal jobs for data

analytics, machine learning training, infrastructure management, and first-party services such as communication, gaming, and video streaming offered to third-party users. The third-party workloads contain various kinds of cloud services, such as Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) VMs, and even Software as a Service (SaaS) which cloud providers have very limited visibility into third-party uses[57].

These create both significant challenges and opportunities in terms of resource management and job scheduling in cloud data centers. On one hand, cloud resource managers need to carefully consider the complex variety and variability of a wide range of workloads and their resource requirements but also the SLA with which cloud providers have to comply. On the other hand, cloud data centers have other resource management objectives such as fault tolerance, load balance, and resource efficiency maximization for increased revenue. Some may consider energy use minimization in the new paradigm of green computing.

Quality-of-Service (QoS) of applications is one of the fundamental performance measures that cloud data centers are striving to improve performance to guarantee compliance with user requirements in the delivery of cloud resources. Some workloads, such as streaming services and user interactive applications, require fast allocation and low response time, whereas some batch workloads, such as big data analytics, may tolerate longer wait times. Therefore, as many of the users interpret their QoS as latency and response time, cloud data centers consider various infrastructures and techniques for maintaining an acceptable level of QoS while maximizing their revenue. For example, to prevent scenarios like overload where mandatory resource demands exceed the capacity of the cluster and jobs' SLA may not be satisfied, admission control mechanisms are generally implemented to handle these cases[148]. They direct the jobs whose SLA requirements can be satisfied in the current system to the job scheduling queue, waiting for the job scheduler to allocate resources and place. To control and guarantee low latency and response time of cloud workloads, cloud data centers generally introduce admission controls using quota systems or priority classes

to resource managers to guarantee the incoming workloads can be served within their SLA requirements[137, 8].

In addition, while traditional schedulers dedicate resources to jobs based on their requests, cloud computing systems generally use oversubscription to exploit the low utilization for better resource efficiency and greater revenue. That is, allocating a resource to multiple jobs, and depending on their statistical multiplexing to enable them to co-exist and achieve the expected performance[58, 122]. It is the fundamental mechanism that enables cloud data centers to substantially improve efficiency by over-committing resources multiple times. This approach achieves much greater loading of computing hardware – and thus greater efficiencies or revenues. Oversubscription exploits the fact that many jobs exhibit low average resource utilization – far less than requested[57, 141, 95]. Oversubscription schedulers can improve system throughput and resource utilization, as well as low latency for production jobs through statistical multiplexing of workloads. Of course, the level of oversubscription has to be carefully designed and tuned in case of unexpected spikes in usage[57]. These designs include but are not limited to complex characterization and prediction of resource utilization, cluster deployment size, server maintenance, and appropriate allocation size.

2.5 Workloads

While existing data centers deal with a great variety of workloads, such as streaming, interactive jobs, and more complex variants of workflows with dependent jobs, batch-scheduled workloads are an important workload. In commercial data centers and production environment, batch analytical computations and data processing are important growing loads [59, 57], consuming 65% to 90% of computing resources. For scientific applications, processors and memory resources are scheduled as dedicated, allowing fine performance control, and extreme scalability.

On one hand, HPC workloads are composed of applications with large-scale scientific

computation, such as Nuclear Physics, Astrophysics, and Climate Research. In HPC environments, users submit jobs with a fixed amount of resource requirements and job runtime estimates. These properties enforce an upper limit on a valid job's allocated computation resources and runtime, beyond which the job will be terminated. Because of the large parallelism these jobs exhibit, the runtime limit is generally constrained to less than 24 hours[121].

On the other hand, to provide high flexibility, availability, and scalability, cloud providers provision VMs for customers to run applications and maintain these resources in a running state to guarantee fast response and instant availability upon requests. To support a wide range of applications and also to maximize revenue by users' requested resources, cloud data centers support a large amount of continuously running jobs, coming from applications such as streaming services and VMs that are unnoticed and unterminated for a significant period, often termed as zombie VMs[86]. These long or continuously running jobs contribute to a significant portion of allocated resources in cloud data centers[137, 100].

One of the most widely used and publicly available industry cloud traces is the large-scale Google cluster workload trace, whose newest version is released in 2019 and contains historical information of eight Google Borg cells for the whole month of May 2019[137]. As pointed out in the studies, the Borg cluster trace exhibits an extremely heavy-tailed distribution where the top 1% of jobs consume over 99% of resources. That is, while a large fraction of jobs is short in runtime, a small number of the long-running jobs comprise most of the computation. This finding of heavy-tailed property is further supported by Borg 2011 data[119] and other large-scale batch workloads with different degrees of skewness[58].

While this general property of cloud workloads may give rise to specific flexibility to variable capacity data centers as many short jobs are more robust to variations as most of them can finish before the next capacity change arrives, but brings out challenges to the scheduler as it has to carefully determine placements for long-running jobs to avoid frequent disruptions from variations.

CHAPTER 3

RELATED WORK

Past research has explored the problem of resource management from different perspectives and optimization goals. Resource management in traditional data centers has looked at optimized simple heuristics and techniques like job preemptions to improve metrics like job wait time or deliver required Service-Level Objectives (SLOs). Some other resource management explores the space of energy consumption to either reduce brown energy consumption or maximize system performance under a power cap. Also, other studies optimize resource management with load changes like demand response by deferring or shuffling workloads.

We study resource management for both HPC and cloud scenarios responding to capacity changes that could arise from clusters, data centers, and site power management [122] or power grid dynamics [151, 87, 55].

3.1 Resource Management and Scheduling with Unreliable Resources

The broad literature on resource management typically assumes fixed resources [141, 140]. For example, some schedulers manage both data center placement (long-running processes) [141], and some consider latency-sensitive jobs [139, 140]. Others implement notions of priority [48, 125]. The general goal for these resource managers is to maximize system throughput (*goodput*) while maintaining quality. In addition, a great variety of resource management techniques have been proposed and adopted, including backfill[102], overcommitment[60, 141], and job preemption [48, 140]. A body of research studies has dealt with the addition and removal of resources or resource capacities varying with time. Some studies[151] consider the special case where the resource capacity during each time period is either fully available or wholly off. In the case of unpredictable failures of resources, many studies ex-

plored scheduling policies to assure secure grid job execution with unreliable resources. [129] considers risk-resilient strategies, preemptive, replication, and delay-tolerant, to provide security assurance under different risky conditions. [130] constructs statistical models to assess the reliability of resources based on prior performance and behavior and proposes algorithms that employ estimated reliability ratings of worker nodes for efficient task allocation. [155] proposes information models for unreliable resources, produced by high priority foreground load, by providing statistical information to allow users to cope with availability changes of volatile cloud resources. While these studies provide an invaluable backdrop and may be considered orthogonal to our scheduler study of capacity variations, we are not aware of any online scheduler studies exploring the management of highly variable resource capacity.

3.2 Large-scale Power Management and Power Capping

Production data centers have long adapted large-scale power management including power oversubscription and power capping at multiple levels of the power, hierarchy to improve power efficiency. Systems like Facebook's Dynamo[149] and IBM's CapMaestro[91] have focused on measuring and budgeting power at server or rack level. Dynamo[149] dynamically monitors power across the entire power hierarchy and makes coordinated control decisions that ensure safety and are performance-aware. CapMaestro[91] adapts global priority-aware power capping that accounts for power capacity at each level of the power distribution hierarchy and exploits server-level power capping. Flex[156] leverages workload properties to optimize rack-level placement for power while ensuring safety and quickly reduces the power draw by shutting down and power capping racks while respecting the workloads' requirements. Google recently published systems that do power management and shifting at a multi-megawatt scale, creating dynamic power constraints for schedulers[122]. It enables larger power-sharing domains, across tens of MW of equipment that improves power fungibility and reduces power stranding. Together with cluster schedulers that assign tasks with

different priorities to any node, it adapts a power capping service that is generator and job priority aware. These studies do not model schedulers, and interestingly suggest that many scheduling domains give better overall throughput – encouraging variable capacity models for scheduling. In academic studies, [61] optimizes power management and load scheduling in geo-distributed cloud data centers to minimize time-average eco-aware power cost while still ensuring Quality-of-Experience of user requests. It applies Lyapunov optimization theory to design an online control algorithm that decides the amount of power supply and scheduling plan for each data center. It views power supply quantity as controllable and exploited to optimize for other goals such as power cost and carbon footprint. We view power and resource capacity as uncontrolled change due to external factors such as power markets, renewable generation, and intercluster management.

Another body of research focuses directly on adapting data center power consumption to meet a power limit (power cap or emergency demand reduction). Power capping is often framed as – known, fixed caps – with variation in application behavior, and managing performance of a set of applications to stay within the caps [65, 109, 90]. Some work learns from previous power profiles or smart configuration selection to enable better resource management under power caps[142, 68]. [142] uses power profiles and job logs to estimate job power behavior and makes job allocation decisions by checking a window of jobs to make scheduling decisions to stay under the power budget while maximizing resource utilization. Such design is able to maintain less than 1% relative degradation when the power cap is set to 83% of the maximum. Of course, if excess resources are powered, switching idle servers off is a good idea, but not so simple. For example, [111] tries load concentration and dynamically turns on nodes to efficiently handle the load and off idle nodes to save energy. Results show that this method can reduce the total power consumption by as much as 86% while keeping performance degradation below 20%. Some explore algorithms like DVFS to enable fine-grained power control and tuning[147]. Other possibilities fall closer to our

study, deferring workloads[132] or exploiting energy storage [94, 34]. [132] incentives tenants to defer batch workloads subject to quality-of-service requirements to enable Emergency Demand Response. These studies typically exploit profiles of computation, power use, and even power markets. [94] dynamically deals with power mismatch through intelligent utilization of energy buffers and improves both energy efficiency by 40% and renewable energy utilization by 81%. However, these power constraints are normally fixed over a long period of time. On the other hand, we are interested in a scenario of constant variation, not just a catastrophic (and rare) emergency. Also, they require profiles about job speedup and energy usage to enable fine-grained power usage optimization. In contrast, our framing assumes a dynamic, uncontrolled change in power (resource capacity), due to external factors such as power markets, renewable generation, or perhaps the demand in the next data center building (unrelated applications or customers perhaps). We study a purer form of the problem, seeking to understand how to do resource management if capacity is reduced for *any* reason, not just the availability of power.

3.3 Renewable Energy and Optimizing for Green Power Use

One approach explores local management of workload and variable on-site renewable generation to reduce carbon emissions. These studies consider resource scheduling optimization of criteria such as green-power fraction, workload performance subject to cost, and grid power cost[71, 80, 42] in a system where there is a predictable, local source of renewable power (i.e. solar). GreenPar[80] proposes a scheduler in data centers partially powered by on-site generations of renewable energy. GreenPar increases the resource allocations to improve performance when green energy is available and reduces allocations to conserve brown energy subject to performance Service-Level-Agreement when renewable generation is insufficient. It utilizes information about job speedup, estimates of job runtime, and predictions of green energy production for intelligent allocation policy. [42] dynamically adapts the resource set

to the actual workload through shutdown policies to reduce brown energy consumption with local photovoltaic panels while considering various impacts, such as the cost of shutdown and wake-up (in terms of time and energy) and electric, thermal constraints imposed on the whole infrastructure. These scenarios are closely related to our study, but in many cases consider highly predictable, periodic generators such as solar photovoltaics. While useful local studies, our focus includes the properties of modern power grids in terms of both renewable mix, renewable mix as a function of time of day, week, and year. Our study assumes externally controlled variable capacity from complex power market carbon factors (generation, load, power markets) without any presumption of predictability and can be extended to general cases from other resources. The dynamics of pricing and power availability are much more complex in these environments [101, 28], which means dealing with more general variation increases the difficulty. [54, 52] characterizes the quantity and temporal structure of stranded power (curtailment and uneconomic dispatch) in the grids, highlighting the complex power availability and great combined opportunities for variable capacity data centers as power grids incorporate increased renewable generation.

3.4 Coupling Resource Management with Power Grids

A great variety of studies focuses on coupling resource management with power grids to exploit fluctuations and programs for power cost reduction and carbon footprint reduction. [92] considers data centers as dynamic loads and studies DC-grid coupling models to explore the impact on grid dispatch, power costs, and carbon emissions. Studies of various coupling approaches demonstrate that delegating load flexibility to the grid shows great grid benefits but creates rapid DC capacity variation and suggests a large dynamic range of as much as 0.6: [0.4, 1.0]. This study is in line with our study which aims to support a large dynamic range for data centers under resource capacity variations driven by external factors. Zero Carbon Cloud[151, 153] posits the creation of volatile data centers powered by stranded power

(wasted or negative-priced renewable power) zero-carbon footprint. These data centers can be operated without a power carbon footprint, and zero or low-priced power. Coupling with a traditional data center, ZCCloud system can reduce job wait times by more than 50%. The studies view data centers as wholly on or off subject to the intermittent availability of stranded power, which is due to a complex combination of variable generation, markets, etc. On the other hand, we explore variation in capacity less extreme, and structurally smoother, a much easier and more flexible resource management problem – but still unsolved. Other works explore the potential benefits of data center demand response by exploiting the flexibility of resource management[132, 157, 150]. [157] propose a strategy for a data center to provide regulation service reserves while providing Quality-of-Service guarantees of the jobs running in the data center. The proposed QoSG policy coordinates separate groups of servers to run different types of jobs. [150] reduce the peak power and demand charge of data centers by using partial execution. The study forms a workload scheduling partial execution problem subject to stringent SLAs on response quality into an integer problem and provides an optimal algorithm. These grid demand-response examples are related – but deal with rare circumstances (e.g. 4 hours a year). Our formulation of the variable capacity problem admits a rich, general externally imposed variation. It can vary at many time scales, with correlation or dependence across sites, and focuses on typical performance, but could perhaps include rare events.

3.5 Managing Resource Revocation

Several systems have done volatile resource management – early work on workstations and PC's in desktop grids [93, 51] to achieve high throughput on sequential jobs in the face of high rates of individual resource "failure" (revocation), and later work designed to exploit Amazon's Spot Instances[27] and Google Preemptible VM's[30]. There are a number of other scenarios where variable capacity is of interest for resource management. For example,

a dependent cloud (a meta-cloud that forms its resource pool from spare resources of others) typically experiences frequent capacity change. Users of Amazon's Spot Instances[27] and Google Preemptible VM's[30] face a related volatile resource utilization problem – how to make effective use of such resources. Other possible causes include partition software upgrades, compartmentalized security, etc. These systems employed statistical characterization [145, 146] to select appropriate resources, and preventative checkpointing to decrease application "failures" (preserve state across revocations) that have matured into commercial extensions which encapsulate the latency-insensitive, throughput model [26, 81]. These systems do not focus on which jobs to slow or terminate (in fact they generally don't have control over this). Further, most of these systems assume they can control available capacity (assuming cloud elasticity), and dynamically allocate and release resources based on availability to meet demand. However, all of these approaches do not address focus on parallel jobs, nor do solve the heart of the classical resource management problem – maximizing the utility of the given set of resources. Most of these systems are application-oriented, and deal with collections of single-node jobs. The capacity variation problem is large-scale resource-oriented, and formulated for a job scheduler managing a workload with complex mixes of co-run, run-before, and other kinds of task dependencies in the face of a rich set of service-level objectives (SLOs).

CHAPTER 4

PROBLEM AND APPROACH

In this chapter, we present the problem and our approach to understanding and improving scheduling performance under variable resource capacity. We formally define the variable resource capacity in the data center and show the challenges posed by job scheduling. We present three fundamental scheduling problems that arise in the face of variable resources. To understand and address these problems, we propose our scheduling approach and introduce three key components of our approach, understanding and characterizing the negative impact, methods to cope with capacity loss to mitigate negative impact, identifying the sources of impact and the resulting information space, and scheduling algorithms that consider various amount of information to prepare for capacity variation. The performance improvement and broad generality of our scheduling approaches demonstrate significant performance improvements in variable capacity data centers, enabling much larger shifting and power use fluctuation capabilities in data centers.

4.1 Problems

In this section, we formally define the data center with variable resource capacity and describe the challenges that may prevent the data center from effectively exploiting variable capacity and two key scheduling problems.

4.1.1 Scheduling Problem Definition of Variable Capacity

In a data center or cluster, let M denote the number of total machines, where each machine m has $r(m)$ resources. A traditional scheduler schedules a set of jobs J on M machines while optimizing one or various objectives. Each job $j \in J$ has submission time $s(j)$, resource requirement $r(j)$, and execution time $t(j)$. The data centers need to decide j_{mt}, which is the

decision variable of running job j on machine m at time t. In systems, such placements are subject to each machine's resource constraint and the total resource capacity constraint:

$$\begin{gathered}\forall t \in T, \forall m \in M, \sum_{j \in J} j_{mt} \times r(j) \leq r(m) \\ \text{subject to} \\ u_{mt} = 1 \iff \exists j \in J \; s.t. \; j_{mt} = 1 \\ \sum_{m \in M} u_{mt} \times r(m) \leq R(t)\end{gathered} \tag{4.1}$$

where the first constraint represents the individual resource constraint (CPU, memory) on each machine. The second constraint guarantees the total active number of machines within the data center capacity, where u_{mt} indicates whether a machine is active or not. In fact, in traditional data centers, the latter is implicitly fulfilled if the scheduler satisfies the first constraint. One of the most common optimization goals is to maximize the useful resource utilization (goodput) of the system. The resource utilization represents the percentage of compute resources allocated normalized by the entire resource capacity in the system over timespan T. It can be expressed as below:

$$max \; \frac{\sum_{t \in T} \sum_{m \in M} \sum_{j \in J} j_{mt} \times r(j)}{\sum_{m \in M} r(m) * T} \tag{4.2}$$

Based on the knowledge that current capacity will continue as M is constant, these schedulers make decisions that commit resources into the future. Because they have been designed to maximize goodput, they strive to fill as much of this capacity as possible with the information on job resource requirements and unknown or estimated job runtime.

However, in a data center with resource capacity variations, the available resource capacity is a function of time t, denoted as $R(t)$ where $R(t) \leq M$. Hence, all job placements are

now subject to a time-varying resource capacity constraint at each time slot t:

$$\forall t \in T, \forall m \in M, \sum_{j \in J} j_{mt} \times r(j) \leq r(m)$$

$$\text{subject to}$$

$$u_{mt} = 1 \iff \exists j \in J \; s.t. \; j_{mt} = 1 \tag{4.3}$$

$$\sum_{m \in M} u_{mt} \times r(m) \leq R(t)$$

$$\forall t \in T, R(t) \leq M$$

This constraint ensures that the total number of machines that have any amount of active running jobs does not exceed the current resource capacity $R(t)$.

4.1.2 Challenges of Job Scheduling

When resource capacity varies, even if the average capacity does not change, significant losses in system goodput (useful resource utilization based on total available resources) can result. In Figure 4.3, we present the resulting system goodput under dynamic capacity, even when a state-of-the-art scheduler [56] is used! As the dynamic range of variation increases from 0 to 0.6 (around an average capacity of 0.7), goodput decreases by 30%. Results are shown for capacity variability with random walk structure with stepsize of one-fourth the dynamic range.

What accounts for this degradation in goodput? Traditional schedulers assume a constant resource capacity of $R(t) = M$. Based on the assumption that current capacity will continue, these schedulers make decisions that commit resources into the future. Because they have been designed to maximize goodput, they strive to fill as much of this capacity as possible, shown in Figure 4.2a. However, the quantity of compute resources available $R(t)$ in variable capacity data centers can vary significantly and on short time scales compared to job runtime.

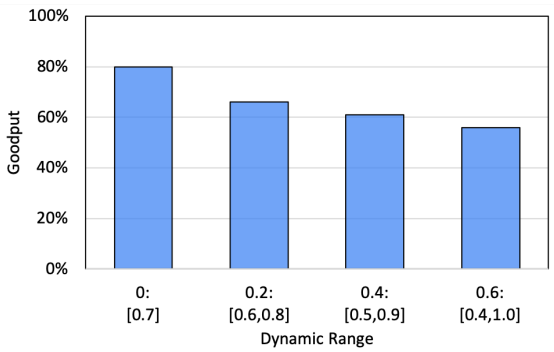
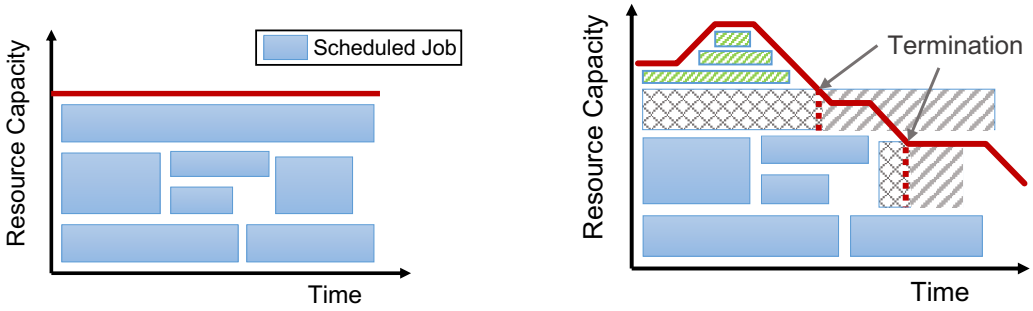


Figure 4.1: Scheduler goodput for a variable capacity data center; increased dynamic range degrades goodput.



(a) Fixed Capacity Data Center (b) Variable Capacity Data Center

Figure 4.2: Job Scheduling in A Fixed Capacity Data Center vs. in A Variable Capacity Data Center

On one hand, if resource capacity decreases, expressed as $R(t) < R(t-1)$, the schedule reflects an overestimate, and the resource capacity constraint in Equation 4.3 can be violated. This results in that jobs in the queue may have to wait longer before resources become available. Further, some running jobs may be terminated due to insufficient capacity to release the machine in order to enforce the time-varying resource capacity constraint (see gray in Figure 4.2b, showing the wasted computation and unfinished work in different patterns). Terminated jobs are put back into the queue, incurring further delays. They run from their beginning when rescheduled, so their runtime before termination is wasted. On the other hand, if resource capacity increases, the scheduler suddenly gains capacity, but may not make good use of it, as recent decisions could not take into account the greater capacity now

available, and may have misplaced jobs (see green in Figure 4.2b).

In this new world, key open research questions include:

- How do current schedulers respond to capacity variation?
  - What is the problem space of capacity variation? and how each contributes to scheduler performance?
  - What are the critical uncertainties in scheduling for variable capacity?
  - What are the dimensions of uncertainties and how do they affect scheduler performance?
- Can scheduler performance be improved in these challenging situations?
  - How to cope with adversity, when uncertainty resolves in a bad way. For example, when capacity drops precipitously. And what are the limits of adversity that can be tolerated with dramatic performance loss?
  - Can schedulers plan for uncertainty? and thereby improve performance?
  - What information might be available to reduce uncertainty? And how can it be exploited effectively?

The first question aims to understand the operation space of current schedulers under resource variability situations. Within this question covering the whole space of capacity variation, we can drill down on the problem space and explore how each dimension contributes to performance impact. With a general understanding of variable resource capacity, we also want to characterize the real variation examples and abstract them into generic problems. Finally, the last key questions focus on potential techniques to reduce performance degradation if there is any and thus expand the operation space for schedulers in order to enlarge variation benefits. So, answering our question depends on studying schedulers on variable capacity with various real production workloads and a range of resource variation

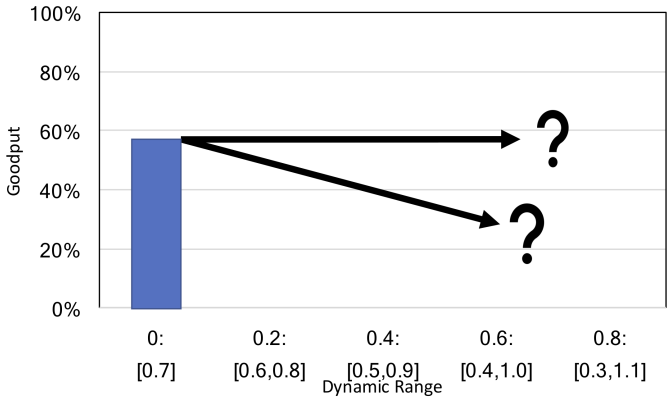


Figure 4.3: How capacity variation affects scheduler goodput is a key to reducing carbon emissions in our scheme.

scenarios and evaluation of scheduling algorithms. We present this question graphically in Figure 4.3.

4.2 Scheduling Approach

To address these problems, we present our scheduling approach to understanding the operation space of schedulers covering the problem space and improving the operation space by scheduling techniques in the face of capacity variations.

4.2.1 Characterizing Scheduling Performance

As variable resource capacity creates new scheduling challenges for data centers, we draw out the problem space of scheduling in variable capacity data centers by defining three key dimensions. Then we aim to understand the impact on scheduling performance within the problem space and resulting operation space. As conventional schedulers assume resource capacity is known and fixed going forward, capacity variations create uncertainty about capacity quantity. Therefore, whether a scheduler can effectively align job placements under resource variations remains an open question.

To characterize the challenge to conventional schedulers, we study and evaluate workloads

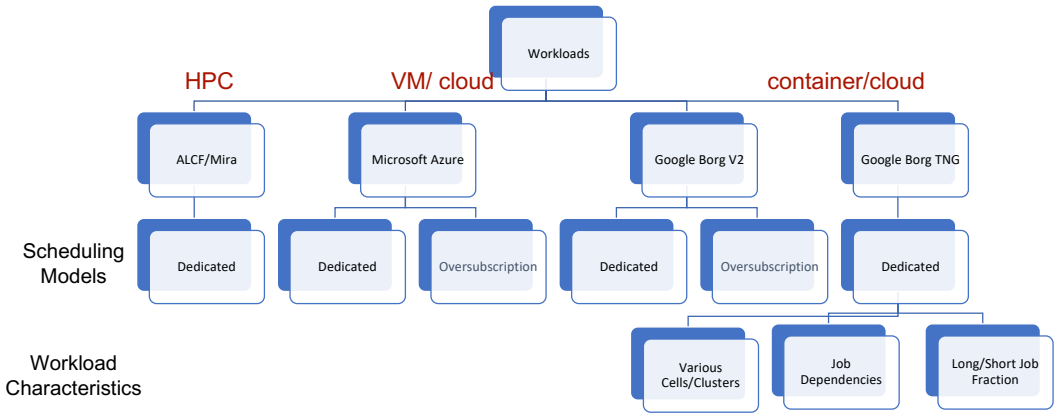
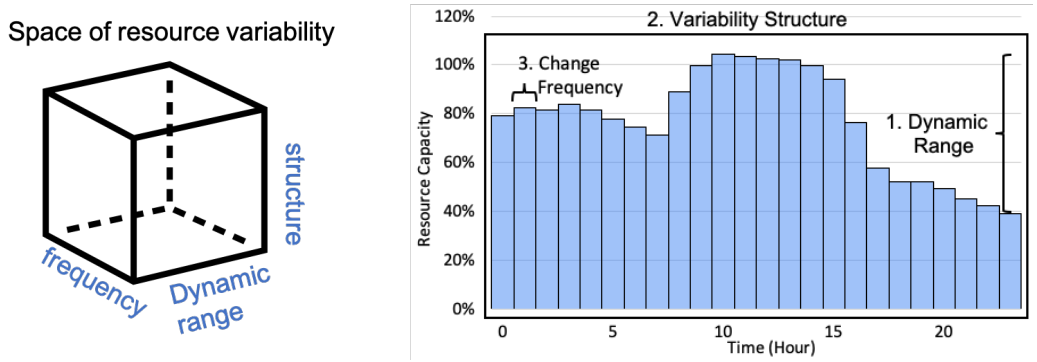


Figure 4.4: A variety of publicly available workloads and the corresponding scheduling space we explore to understand scheduling performance in the face of variable resource capacity.

and schedulers in the face of variable capacity under various scenarios and environments to cover a broad problem space. We carefully pick a variety of publicly available workloads. Figure 4.4 shows all the workloads we consider: ALCF/Mira, Azure, Borg V2, and Borg TNG workloads. These workloads are well-known exemplars of their respective environments and each of them corresponds to HPC, VM(cloud), and containerized cluster(cloud) workloads. To correspond to these workloads and their data center setting, we pick two distinct scheduling models, whose critical difference is whether the schedulers consider compute and memory capacity separately and if dedicated resources or oversubscription is practiced. For each workload combined with the corresponding scheduling model, we use a system model that varies the resource capacity available to the scheduler and evaluate performance.

Constant resource is a simple model; variable resources can have many different dimensions of variation. We consider the space of resource variations to be three-dimensional, as shown in Figure 4.5:

- Dynamic range: minimum to maximum capacity
- Variability Structure: random uniform, random walk



(a) Space of Resource Variability

(b) Modeled dimensions of capacity variation. Illustrated on a time-sequence of datacenter capacity from Figure 1.6a.

Figure 4.5: Space of resource variability (left) and dimensions illustrated on a variable capacity example (right).

- Change Frequency: frequency of capacity variation

These three dimensions represent the whole space of resource variability, illustrated in Figure 4.5a. The dynamic range captures the distance over which resource capacity varies – from a low to high watermark and back. It is the most foundational element of resource capacity change. Variability structure reflects how capacity is constrained to change from one time period to the next. Such constraints often reflect the realities of physical systems - inductance, momentum, inertia, and more – that prevent large instantaneous change. Change frequency reflects our choice to model time discretely – capacity varies only at time period boundaries – so change frequency reflects the size of those periods. In a real system, periods could be defined by external structures (power markets), data center physicals (cooling and power-sharing control systems), or other factors.

These dimensions are designed to cover a broad range of resource capacity variation scenarios, produced by various external factors. By defining the key dimensions of resource variability, we construct a generic multi-dimension space of resource variability.

Using these workloads and schedulers, covering a broad range of data centers and workloads we execute a set of scheduler experiments that explore this multi-dimensional capacity

variation space, characterizing scheduler performance. In effect, each experiment explores scheduler performance when actual resource capacity diverges from the scheduler's simple fixed estimate of stable resources. Our goal is to understand the capabilities or the operation space of state-of-the-art schedulers, where they can maintain robust performance and where they cannot. In addition to realistic workloads and scheduling models varying environments and use cases, to further drill down on understanding the operation space of conventional schedulers, we drill down on cloud workloads with a more complex and complete view. We consider Borg TNG workloads, the newly released Google cloud trace, which provides richer details and enables a focused study on various workload properties, such as job dependencies and runtime distribution. The whole space of workloads is depicted in Figure 4.4. By varying the workload properties, orthogonal to variable capacity, we explore the operation space and workload-specific limitations of traditional schedulers in the space of capacity variations.

We further collect a variety of real variation traces arising from various sources. We are to understand the dynamic variation behaviors and characteristics of data center capacity in real-world scenarios. Therefore, we characterize these traces and abstract them into generic problems, defined by three key dimensions. In addition, we evaluate the scheduling impact using these real variation traces, validating the results from synthetic traces. The broad and general characterization of scheduling performance impact under resource capacity variation lay the foundation for understanding the limits of future data centers to fully exploit carbon and power cost benefits. A detailed description of understanding capacity variation is in Chapter 5.

4.2.2 Coping with Capacity loss and Preparing for Capacity Variation

We propose various scheduling techniques to improve scheduling performance in the face of resource capacity variations. We consider two-prolonged approaches to address the scheduling challenges, coping with capacity loss and preparing for capacity variation.

First, as capacity varies over time, significant job failures can incur during capacity decreases resulting from the scheduler's incorrect assumptions about capacity, resulting in wasted computation and goodput loss. We characterize job failure rate in the problem space of variable capacity varying workloads and variation traces. Furthermore, to mitigate the negative impact of capacity decrease, we consider intelligent termination policies to cope with capacity loss. Intelligent termination policies, considering job features and progress, selectively terminate jobs to minimize wasted computation and improve resource efficiency. We evaluate and compare the intelligent termination policies to understand their effectiveness in improving performance and improved operation space upon capacity loss over a range of workloads and variation ranges. In addition, we consider using foresight of capacity variation, which is partial oracle information, to reduce uncertainty and enable the scheduler to cope with capacity loss. We evaluate workloads and variations varying the length of foresight to understand the usefulness of incremental foresight information. A detailed description of the scheduling approach to coping with capacity loss is in Chapter 6.

To take a broader view, we consider strategies to prepare for capacity variation, which proactively prepare for capacity increase and plan for capacity loss. As an extension of variation foresight, we present the dimensions of uncertainty that contribute to performance loss, job runtime, and capacity variation, and the information space of these two dimensions that may help reduce schedulers' uncertainty. We propose scheduling algorithms that exploit the information space to optimize job placement decisions to minimize wasted computation and maximize resource efficiency. We focus on workloads that may represent a challenging case for intelligent termination policies to demonstrate the additional improvements from the broader view of preparing for capacity variation. We empirically evaluate and compare these scheduling algorithms varying workload properties and variation ranges to demonstrate improved operation space and shifting flexibility. A detailed description of the scheduling approach to preparing for capacity variation is in Chapter 7.

4.3 Summary

This chapter explains the rise of variable resource capacity in data centers and lists a few fundamental scheduling problems. We then briefly explain our proposed scheduling approach. The approach contains two key components in addressing the scheduling problems: understanding the performance impact and problem space of capacity variation, identifying the sources of uncertainty and their information, and scheduling algorithms that exploit information to improve performance. The details of these two components are discussed in the following chapters.

CHAPTER 5

UNDERSTANDING CAPACITY VARIATION

In this chapter, we study the behavior of current schedulers under variable capacity to understand how such variation affects performance in data centers. This study provides the foundation, providing an understanding of the key dimensions of capacity variation that cause negative impact and the operation space of conventional schedulers. This systematic evaluation identifies the sources of uncertainty and corresponding information that scheduling algorithms need to exploit in the face of variable resource capacity. In Section 5.1, we first discuss the workloads, metrics, and schedulers we consider for a comprehensive study. Section 5.2 evaluates the whole space of capacity variations with various workloads and scheduling models to understand the individual and compound impact of variation dimensions. We further drill down on Borg TNG trace, the most recent and widely-used cloud workloads, looking into the space of workload characteristics to understand the operation space and specific limitations of cloud workloads in the face of capacity variation in Section 5.3. Section 5.4 includes the realistic scenarios that give rise to variable capacity and how to interpret these variations. Finally, Section 5.5 summarizes the chapter.

5.1 Methodology

We describe the dimensions of capacity variation, workloads, job schedulers, systems, and metrics used in experiments.

5.1.1 Dimensions of Capacity Variation

We consider three dimensions of variation, maintaining the average resource capacity constant in all cases. These dimensions, dynamic range, structure, and change frequency are illustrated in Figure 4.5b. We define them below.

Dynamic Range defines the distance between maximum and minimum capacity. We consider variation ranges of 0 (constant), 0.2, 0.4, and 0.6 as a fraction of maximum datacenter capacity. To normalize average capacity at 0.7, this produces dynamic ranges and intervals: 0: [0.7], 0.2: [0.6, 0.8], 0.4: [0.5, 0.9], and 0.6: [0.4, 1.0].

Structure defines how much the capacity can change between adjacent time periods. **Random Uniform:** Resource capacity in the next interval can be anywhere in the dynamic range and is drawn from a uniform distribution $\mathcal{U}([lbound, ubound])$, appropriate because power prices can be highly volatile. Or, **Random Walk:** Resource capacity in the next interval can change only by a maximum of *stepsize*, modeling some continuity from one interval to the next. Except where explicitly noted, we use stepsize of one-fourth of the dynamic range.

Change Frequency (Temporal Granularity) defines the frequency of resource capacity changes. Between changes, the capacity is constant. We vary the change frequency from 0.25 per hour (every 240 minutes) to 4 per hour (every 15 minutes).

5.1.2 Job Scheduling

The resource manager selects a job from the queue and, based on complex priority, selects the resources to run it on. A critical difference in approaches is whether the schedulers consider compute and memory capacity separately and if dedicated resources or oversubscription is practiced.

5.1.3 Workloads

We considered a variety of publicly available workloads (key statistics in Table 5.1). We picked the ALCF/Mira trace as the exemplar of large-scale HPC workloads as well as Azure and Borg V2 as node-sharing cloud workloads. Azure and Borg traces were chosen as the richest exemplars of data center cluster workloads. In addition to Borg V2 traces, we also

consider the Borg TNG workloads (became available in 2020) as the newly released traces provide richer details, more total quantity, and slightly changing distribution.

HPC workload We use a production trace from ALCF/Mira with a full range of jobs runtimes and parallelism[37]. Notably, some jobs exhibit parallelism as high as 49,152 nodes. The trace includes 6,571 jobs with scheduling events - job submission with requested runtime, nodes count, and timestamps of submission, start, and completion. Basic statistics are illustrated in Figure 5.1a (left). The x-axis is the job parallelism in log scale, the y-axis is the job runtime in hours, and the size of the markers represent the number of jobs sharing the same statistics. Job count is dominated by small to medium size and runtime. These characteristics are typical of both data center and high-performance workloads [119, 24]. Studies use a one-month subset as indicated in Table 5.2.

Cloud workload We use the Microsoft Azure trace which includes VM submission, start and completion time, requested virtual cores and memory, and actual CPU utilization in 5-minute intervals[57]. As shown in Figure 5.1a (right), runlength of VMs also ranges from 5 minutes to more than 24 hours (summarized at 24 hours in the plot). The largest VMs are only 16 cores. Studies use a one-day subset to approximately match the number of jobs in the ALCF/Mira trace as indicated in Table 5.2. The Borg V2 trace contains job start times, end times, requested CPU and memory, and actual CPU and memory usage in 5-minute intervals. The job runlengths vary from a few seconds to much longer than 24 hours (summarized at 24 hours in the plot), with parallelism up to 0.5 GCU. Compared to Azure, the Borg trace has more small and short jobs, as well as a significant load from long-running jobs (see Figure 5.1a (right)).

Similar to the Borg V2 trace, Borg TNG workload traces include information about eight different Borg cells for the month of May 2019. In addition, it introduces the job dependencies information. As shown in Figure 5.1b, the runtime of jobs ranges from a few

seconds to much longer than 24 hours, with parallelism up to 0.5 GCU. The percentage of job dependencies in the workloads ranges from ∼0.1% to 12% across Borg cells. Since the number of jobs is massive, we focus on a single day from 5 borg cells, which include a total of ∼ 1.5 million submissions.

Workload	Mira	Azure [57]	Borg V2 [119]
Classification	HPC	Cloud (VM)	Cluster
Number of jobs	78,795	2, 013,767	47,351,173
Length (days)	365	30	29
Job runtime Avg, StDev (hrs)	1.7 & 3.0	51.8 & 169	1.84 & 21
Job parallelism Avg, StDev	1,975 & 4,100 nodes	2.6 & 2.4 cores	0.03 & 0.03 NCU[1]
Job memory Avg, StDev	31 & 65 TB	6 & 10 GB	0.03 & 0.02 (Normalized)[2]
Year of Trace	2014	2017	2011

Table 5.1: Key Statistics for Widely Used Public Workload Traces

5.1.4 Job Scheduling

The resource manager selects a job from the queue and, based on complex priority, selects the resources to run it on. One critical difference in scheduling approaches is whether they consider compute and memory capacity separately and whether dedicated resources or over-

1. The resource unit is rescaled by the largest GCU(Google Compute Unit) capacity of the machines in the traces

2. RAM measured in bytes, rescaled by maximum machine memory size in the traces

Workload	Mira'	Azure'	Borg'
Number of Jobs	6,571	442,784	204,749
Trace Length (days)	30	7	1
Job (Avg, StDev)			
Runtime (hrs)	1.8 & 2.9	3.6 & 8.8	0.6, 3
Parallelism	1,705 & 2,890 nodes	2.6 & 2.4 cores	0.03 & 0.02
Memory	27 & 46 TB	6 & 11 GB	0.02 & 0.02

Table 5.2: Key Trace Statistics for Workload Used in Simulation

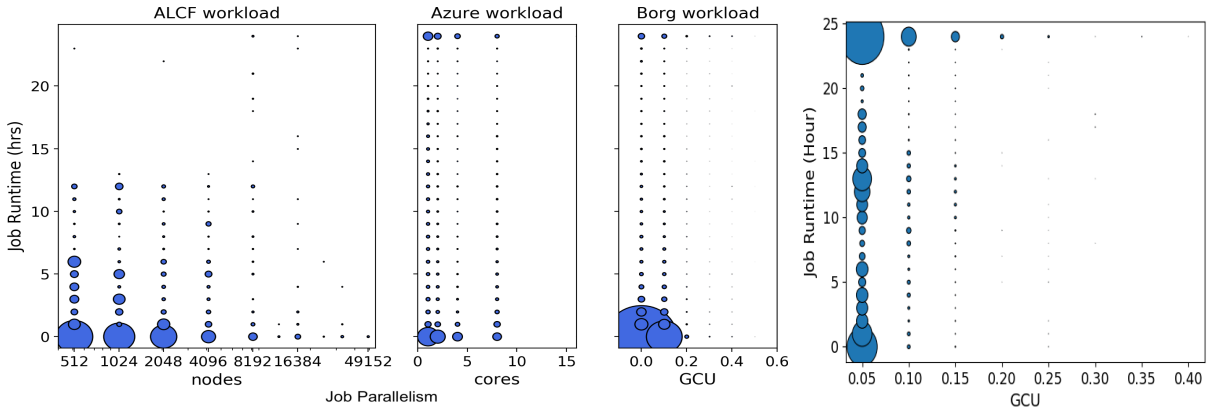
subscription is practiced.

Dedicated Resource Scheduling This model schedules (dedicates) the requested quantity of nodes, CPUs, and memory to the job. From the initiation time to completion/termination, these resources cannot be used by other jobs. Resource capacity reductions affect the number of CPUs (or nodes) available. For scientific applications, processor and memory resources are scheduled as dedicated, allowing fine performance control, and extreme scalability. We study a high-quality, mature commercial production HPC scheduler, Cobalt [56, 114] using HPC production workloads. Cobalt uses utility functions to calculate dynamic priority scores for each job. Then, it selects resources to run the highest-priority jobs. For the cloud, we use an FCFS policy on the same dedicated resource model to create the most comparable results. The FCFS policy is widely observed to give good results with the low-parallelism these commercial workloads exhibit and is widely-used[48, 139, 108, 83]. This model is also compatible with other types of batch execution, so to create the most comparable results, we use the dedicated resource model consistently for all of our studies.

Jobs are delay-tolerant, so the primary metric, *goodput*, depends on system throughput of jobs that complete successfully (are not terminated). Terminated (failed) jobs are re-queued to be scheduled again. In the cloud (dedicated) model, both CPU and memory limits are enforced.

Oversubscription Scheduling Among current cloud scheduling models, oversubscription scheduling is widely-adopted to increase system utilization. In this scheduling model, we continue to enforce the CPU and memory limits separately, and oversubscribe the CPU. The idea is to achieve higher resource utilization via statistical multiplexing, exploiting the gap between resource requests, max use, and typical use.

The Borg V2 trace defines the ratio of workload to resources, so we use it unchanged. The Azure trace does not define this ratio so we analyzed the trace and chose an oversubscription



(a) ALCF (left), Azure (middle), and Borg V2 (right) Workload (b) Borg TNG workload

Figure 5.1: Job parallelism and runtime in the workloads

of 125% of CPU resources (25% more virtual cores than physical cores), based on the job's 95th-percentile virtual core utilization instead of the requested amount.[3] This choice matches the CPU utilization reported for the Borg TNG traces. In the oversubscription model, jobs can be slowed down when resource capacity shrinks (lower CPU capacity) by the ratio of actual CPU usage to available capacity. Resource capacity reductions decrease the cores available; for the cloud workload, memory capacity is not a limiting factor. Jobs that are slowed down can *catch up* by claiming surplus CPU capacity in future 5-minute intervals based on actual CPU utilization from the trace.

5.1.5 Systems

HPC System The modeled resources are the Mira system, a 10-petaflops IBM Blue Gene/Q system, deployed at the Argonne Leadership Computing Facility, and at its inauguration, the world's 3rd fastest supercomputer. Mira contains 49,152 nodes (786,432 cores) and 760 TB memory [99].

3. This information may not be available in general, but this assumption produces an optimistic estimate.

Cloud System We use an Azure cloud cluster with 1,250 nodes (20,000 cores) and 160 TB of memory. This system is a close match in scale to the Mira system. We also model a Borg cluster with 630 nodes (336 GCU - Google-Compute-Unit) and 300 normalized bytes of memory. This system is sized to match the sampled Borg V2 trace used (Table 5.2). For Borg TNG workloads, we model a Borg cluster with 476 nodes (250 GCU - Google-Compute-Unit) and 300 normalized bytes of memory. This system is sized to match the sampled Borg V2 trace used (Table 2) at 90% resource utilization in a traditional data center operating at a fixed 70% of total capacity. This level of headroom and the actual resource utilization level in cloud data centers are validated by a series of studies of enterprise datacenters[137, 57].

For oversubscription, we use the same cloud traces and scale down the hardware to maintain a similar system load. Scaling for both average CPU utilization and the oversubscription rate produces an Azure cloud cluster with 200 nodes (3,200 cores) and 26 TB of memory. We also model a Borg V2 cluster with 630 nodes (336 GCU - Google-Compute-Unit) and 300 normalized bytes of memory. This system is sized to match the sampled Borg V2 trace used (Table 5.2).

5.1.6 Metrics

We use three metrics to quantify scheduler performance, Quality of Service (QoS), and user experience.

- *Goodput* measures the ability of the scheduler to utilize resources to complete jobs. We compute goodput as the ratio of node hours used by successfully completed jobs to the total available node hours. Losses include unscheduled resources and job runs that fail (terminate before completion).

- *SLO Miss Rate* measures the fraction of jobs that experience SLO violations. In the dedicated resource model, *Job Failure Rate* captures the fraction of job runs that fail to complete successfully. Compared to constant resource capacity, where this fraction is

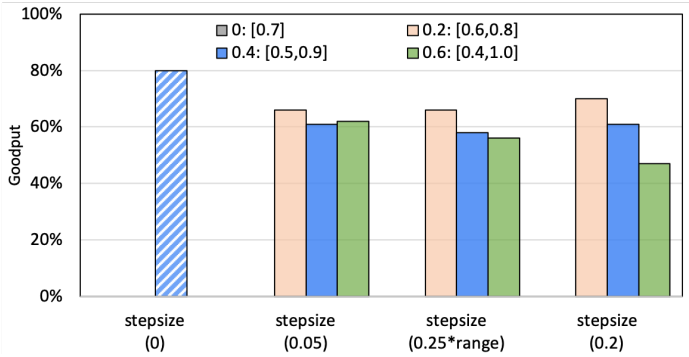


Figure 5.2: HPC (Dedicated) Scheduler performance varying dynamic range and varying stepsize for the random walk.

close to zero, the failure rate quantifies the disruptions from resource variability. In the oversubscription model, *Slowdown Rate* measures the fraction of jobs that experience a later completion time to the baseline. It quantifies the impact of both oversubscription and resource variability.

- *Job Waiting Time* represents the average interval between job arrival time in the queue and job start time (of its successful completion) and demonstrates measurement of user experience in job execution.

5.2 Dimensions of Capacity

We evaluate scheduler performance under varying resource capacities to understand how well they can manage variation and when it causes goodput loss. We explore variability dimensions of dynamic range, structure, and change frequency to understand how specific features of variable capacity affect performance. The worst of which, one might choose to avoid or perhaps engineer mitigation.

5.2.1 Dynamic Range

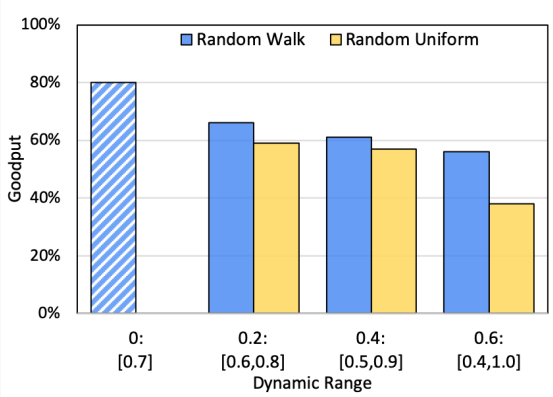
First, let's consider resource capacity variation with different dynamic ranges. In Figure 5.2, the patterned bar at the left is the scheduler performance (goodput) with the same average capacity but no capacity variation – the baseline. The groups of bars left to right reflect increasing dynamic range, all under random walk structure. Across the clusters, we vary random walk stepsize. The first and third bars have fixed stepsizes of 0.05 and 0.2 respectively. The center bar uses a stepsize scaled as one-fourth of the dynamic range.

As the dynamic range increases, the goodput declines for all stepsizes; with increased variability, scheduling performance degrades. By the time we reach the largest range, 0.6: [0.4,1.0], the goodput has declined by 25-45%. The largest dynamic range and largest stepsizes produce the worst performance. This performance loss is due to a dramatic increase in job failures which we will examine in greater detail later.

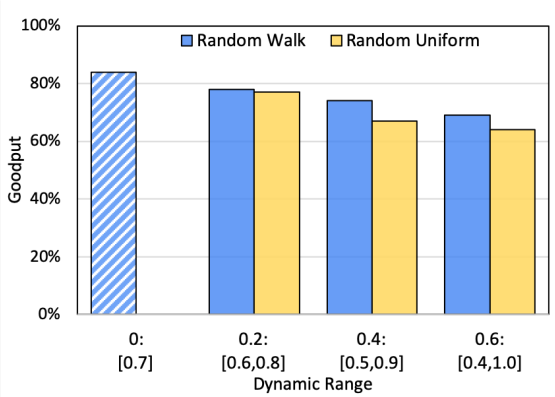
5.2.2 Variability Structure

We consider two variability structures, random walk and random uniform. In Figure 5.3, let's first look at the random walk case (blue), comparing it to no variation (patterned). Stepsizes are one-fourth of the dynamic range, and these HPC (dedicated) results were also presented in Figure 5.2. With increasing dynamic range, scheduler performance (goodput) degrades in both dedicated resource schedulers. Next, we compare random uniform (yellow in Figure 5.3). The dedicated resource schedulers experience goodput degradation as much as 35% worse (for a total degradation of 55%). This is because random uniform allows large jumps in capacity, disrupting the job schedule with terminations or wasted resources. For the dedicated scheduling models, we conclude variation structure can be as important as the dynamic range in degrading scheduler performance.

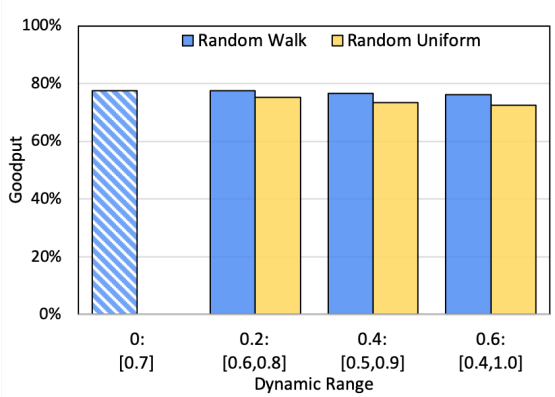
In Figure 5.4, we consider oversubscription. Both cloud workloads (oversubscription) exhibit little degradation for random walk. We believe this is because the safety margin



(a) HPC (Dedicated Resource)



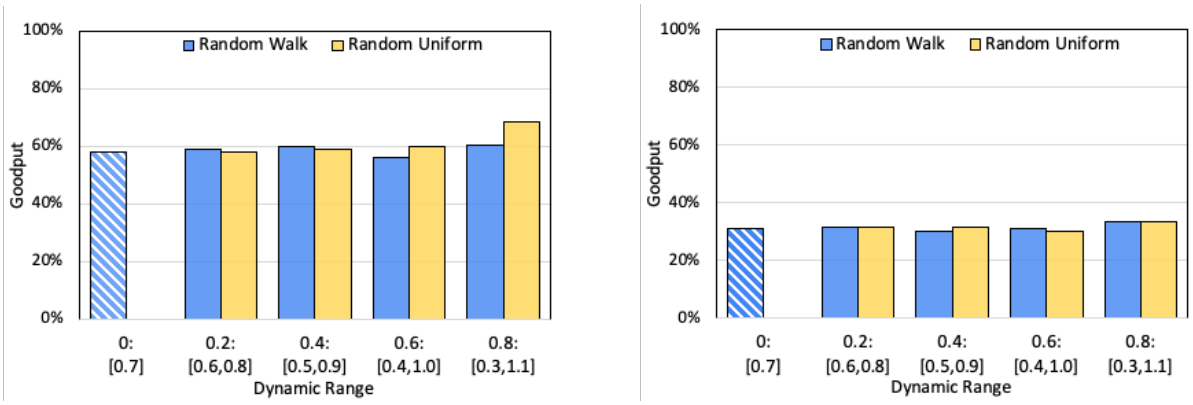
(b) Azure (Dedicated Resource)



(c) Borg (Dedicated Resource)

Figure 5.3: Scheduling performance with random walk and random uniform resource variability structure, varying dynamic range.

provided by statistical multiplexing allows much of the dynamic capacity change to be absorbed with little negative impact on goodput. More importantly, the oversubscription model adopts job slowdowns under capacity variations which can be viewed as a mitigation strategy than abrupt job terminations. The implications and impacts of job slowdowns need further study and validation. Random uniform also sees little goodput degradation. For Azure workload, in both random walk and random uniform cases, there is a significant goodput increase for large dynamic range cases (0.8:[0.3, 1.1]), whereas other dynamic ranges exhibit little goodput difference. It is because the 0.8 dynamic range allows large swings in capacity, producing time intervals with higher oversubscription levels during capacity decreases with job slowdowns. Moreover, random uniform shows even higher goodput because its structure



(a) Azure (oversubscription)

(b) Borg V2 (oversubscription)

Figure 5.4: Scheduling performance with random walk and random uniform resource variability structures, varying dynamic range.

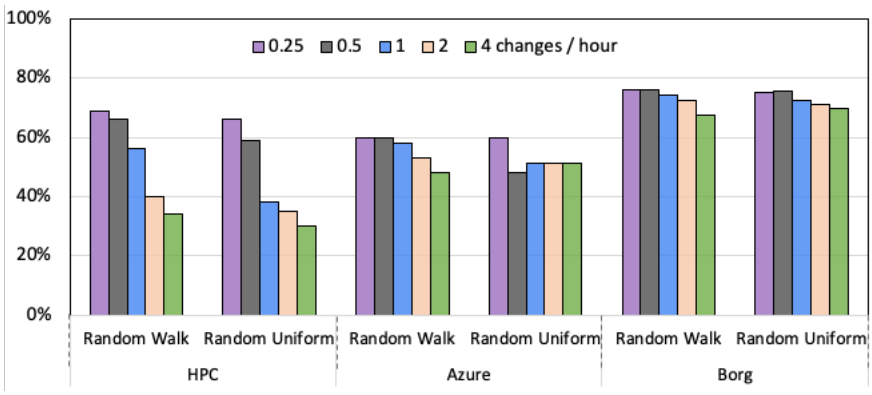


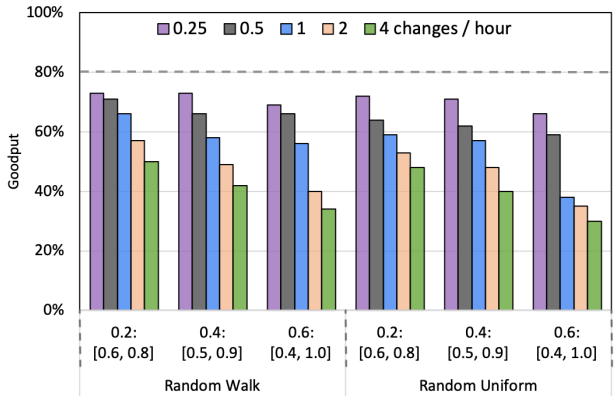
Figure 5.5: Goodput versus change frequency (dynamic range 0.6: [0.4, 1.0]).

produces more volatile capacity changes between time intervals. For the Borg workload, a similar trend is observed but less in quantity due to its lower CPU utilization.

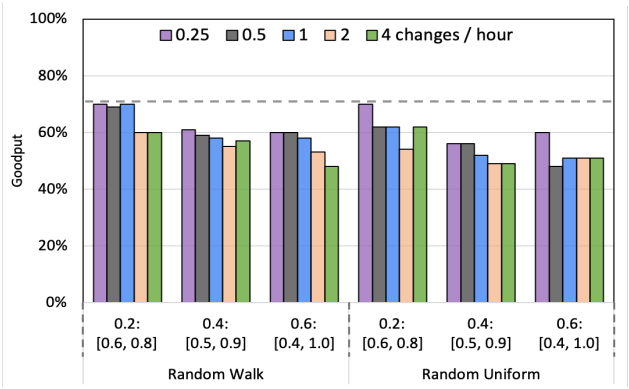
For cloud scheduling, we conclude that oversubscription scheduling is robust with various variation structures within a moderate dynamic range.

5.2.3 Change Frequency

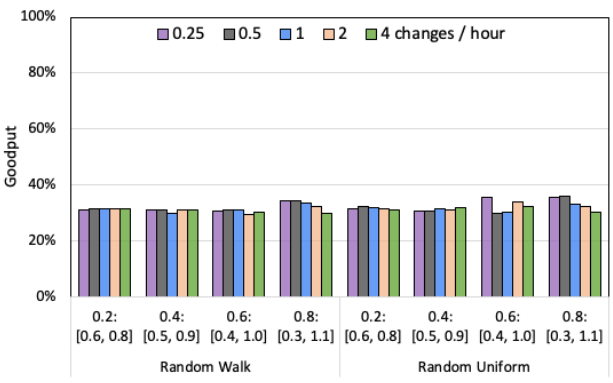
Change frequency is another dimension of capacity variation, so we start with a low rate (0.25 changes/hour), and increase to a high rate (4 changes/hour). Note that all prior experiments used a change frequency of 1 change/hour. We focus on a dynamic range of 0.6: [0.4, 1.0]



(a) HPC (Dedicated Resource)



(b) Azure (Dedicated Resource)

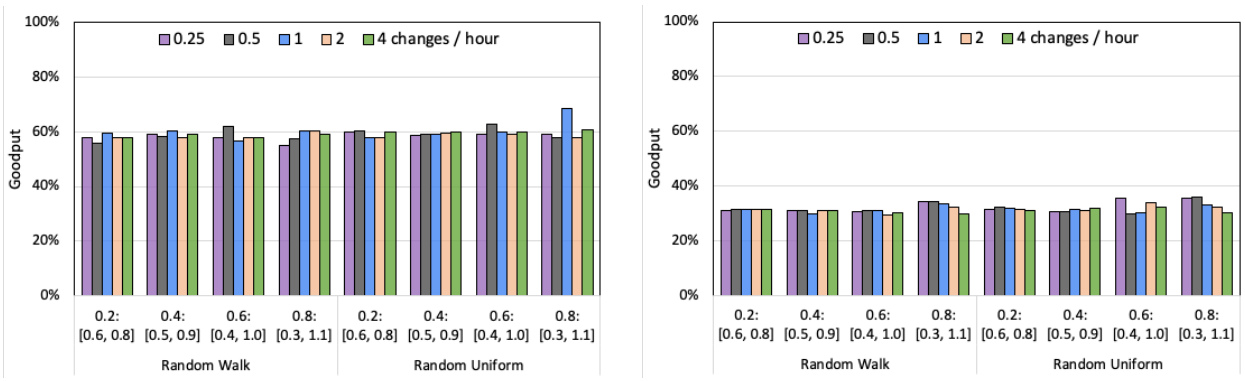


(c) Borg (Dedicated Resource)

Figure 5.6: Goodput versus change frequency, varying dynamic range and structure of capacity variation.

with stepsize of 0.15 first. In Figure 5.5, a significant goodput drop is observed across all structures and workloads as frequency increases. For HPC workload, goodput has fallen by as much as 50%. For Azure workload, higher change frequencies cause clear degradation in goodput (up to 30% overall, but 15% attributable to frequency); Borg V2 exhibits clear, but lesser degradation. These commercial workloads are less sensitive to resource variation because of their lower parallelism.

We combine change frequency with the other parameters (dynamic range and structure), putting it all together in Figure 5.6. With a very low change frequency of 0.25 changes/hour, performance approaches the fixed capacity case. The negative impact of increasing change frequency on goodput remains but is less extreme across all dynamic ranges. For HPC



(a) Azure (Oversubscription)

(b) Google (Oversubscription)

Figure 5.7: Goodput versus change frequency, varying dynamic range and structure of capacity variation.

(dedicated), as change frequency increases, there is a significant scheduler degradation for each increase, so that at 2 changes/hour, goodput has fallen by as much as 50%. For Azure (dedicated), higher change frequencies cause clear degradation in goodput (up to 30% overall, but 15% attributable to frequency); Borg V2 exhibits clear, but lesser degradation. These cloud (dedicated) workloads are less sensitive to resource variation because of their lower parallelism.

Both cloud workloads (oversubscription) have a different experience (see Figure 5.7). Goodput is close to the fixed resource case for all scenarios and shows no sensitivity to change frequency and dynamic range. Despite the overall stable performance varying all dimensions, we still see small variations of goodput between change frequency produced by more frequent, disruptive changes of resource capacity. While oversubscription scheduling design can absorb most of these disruptions, intelligent design and close monitoring are still recommended to avoid negative impacts. In general, beyond lower parallelism, the flexibility of statistical multiplexing provides greater tolerance of capacity variation.

Overall, the results suggest that for dedicated resource schedulers it would be productive to limit the change frequency of capacity variance as much as possible.

5.2.4 Summary

The thorough study of conventional schedulers on real HPC and cloud workloads shows that resource capacity variation can have a large impact on goodput, reducing it by up to 60%. Goodput in HPC (dedicated) and both cloud (dedicated) resource models are particularly sensitive to dynamic range, structure (and stepsize), and change frequency. In these systems, change frequencies of less than 0.25/hour (one per four hours) are desirable. In contrast, we find that cloud's oversubscription scheduling model is remarkably robust to capacity change – in almost all dimensions. While with a large dynamic range, goodput performance increases significantly, this may be an illusory benefit. It's likely that capacity troughs are eating into the safety margin that allows oversubscription without disrupting SLOs. But overall, cloud systems can tolerate wide dynamic ranges, up to 0.8 (80%) of full data center capacity without reduced performance. This is promising indeed for the variable capacity approach but requires further validation.

5.3 Cloud Workload Drill Down

In the previous section, results show that under the dedicated scheduling model, both cloud workloads experience much less degradation compared to HPC workloads. Moreover, performance is remarkably robust with the oversubscription scheduling model. With these positive results, we carefully revisit the impact of variable capacity on cloud workloads. These evaluations represent a simpler view of the cloud workloads as the utilization in Borg V2 workloads is maintained at a lower level, even more so under the oversubscription model. In addition, state-of-art cloud workloads exhibit a dominant mode of VM usage, where long or continuous running jobs contribute to a large fraction of total core hours, presenting a challenge to resource flexibility. Other factors including inter-task dependencies, job wait time measurements, and cross-cluster diversity are not yet taken into account as these are key properties and important SLA metrics in cloud services.

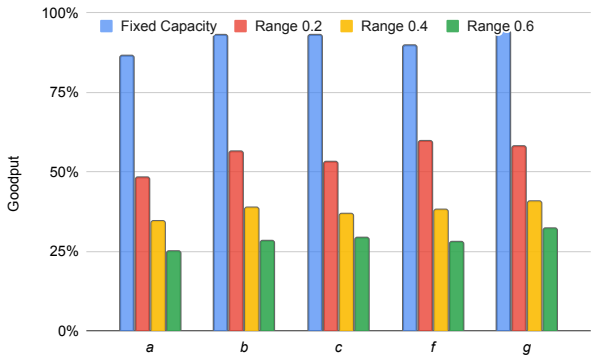
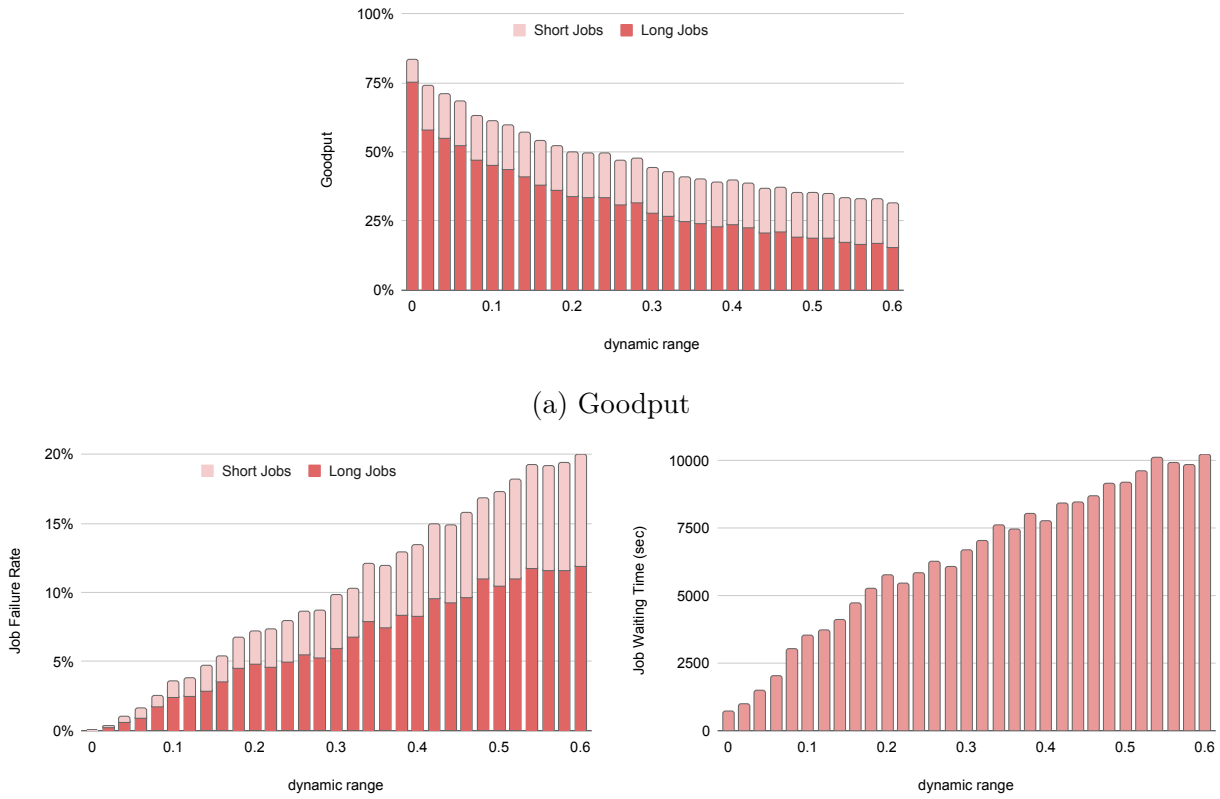


Figure 5.8: System Performance of Google's (Cloud) Borg TNG workloads using Dedicated Scheduling Model across 5 Borg Cells Varying Dynamic Range

To further drill down on if and how cloud workloads are negatively impacted under capacity variations and whether new scheduling techniques are necessary to improve cloud shifting capability, we revisit the scheduling problem of cloud workloads under variations, focusing on workload properties and key metrics to evaluate impacts. We use the Google Borg TNG trace released in 2020, which is the most recent large-scale cloud trace, with the dedicated resource scheduling model. It contains richer details, such as inter-task dependencies, and workloads covering eight Borg cells with diverse distributions and characteristics across cells, capturing the realistic complexity of cloud scheduling. Furthermore, it demonstrates a significant increase in allocated resources compared with Borg V2 workloads, reflecting a state-of-the-art utilization level in cloud data centers.

5.3.1 Variation Ranges

We start by revisiting the performance evaluation under varying resource capacity dynamic ranges and stepsizes of each step of the random walk correspondingly with Borg TNG workloads. In Figure 5.8, the leftmost blue bars in each cluster are the scheduler performance (goodput) with no capacity variation – the baseline. Then each group of bars left to right reflects increasing dynamic range, all under random walk with the same pattern with stepsize scaled as one-fourth of the dynamic range. All cases have the same total capacity. As



(a) Goodput

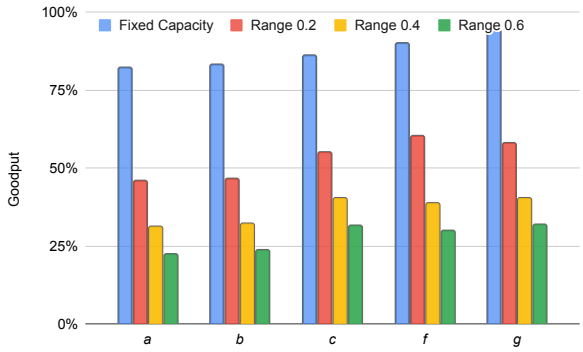
(b) Job failure rate

(c) Job waiting time

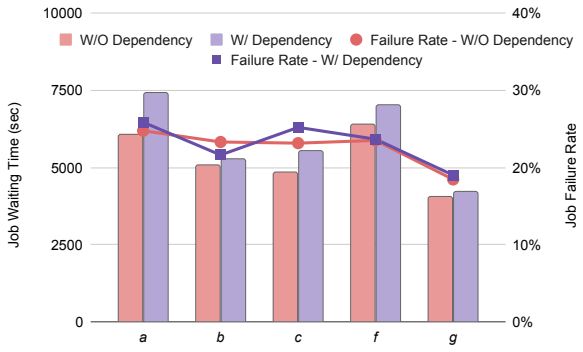
Figure 5.9: System performance Varying Variation Ranges

the dynamic range increases, the goodput degrades significantly, with approximately 40% further by each increase. This performance loss proves that in present-day cloud data centers, variable resource capacity still poses significant challenges to the scheduler, and such performance degradation grows drastically as the range of variations increases.

Moreover, we further drill down on the whole spectrum of variation ranges from 0 to 0.6 in small steps on Borg cell b in Figure 5.9. As the range of capacity increases, goodput shows a gradual drop, and the goodput curve becomes flattered towards larger variation ranges. Goodput demonstrates a heavy tail trend, with the first 5% variation range contributing to 15% goodput loss. Further drilling down to separate the total goodput into short jobs (in lighter red) and long jobs (in darker red), the total short jobs' goodput results remain stable while the goodput contributed by long jobs drastically decreases as the dynamic range



(a) Goodput vs. Variation Range, Jobs with Dependencies



(b) Job Waiting Time and Failure Rates, Comparing Independent Jobs vs. Jobs with Dependencies

Figure 5.10: System Performance of Traditional Scheduler of Independent Workloads and Same Workloads with Job Dependencies

increases. Job waiting time significantly increases with larger variations, and the first 5% variations incur 3X increase. The job failure rate also linearly increases, doubling every 0.1 step increase in variation range. Looking at long and short jobs, failure rate increases of long jobs over dynamic ranges notably outrun that of short jobs, demonstrating challenges on long-running jobs under variation. These results demonstrate why adapting data center load is hard for cloud providers. As major performance damages arise from the first 10% range of variation, even the smallest amount of variation can pose large scheduling challenges on cloud workloads, with serious goodput loss and user experience degradation.

5.3.2 Job Dependencies

As some cloud workloads have more complex structures than independent jobs, suggesting that at least a fraction of parallelism or job dependencies should be expected from distributed computing models such as MapReduce, Spark, and SQL. For example, among 8 Google Borg cells, the percentage of jobs with dependencies is widely different, from 0.1% to 12%. Other studies show jobs with dependencies compose 48.92% of batch jobs and 20% of all jobs in production clusters[135]. Job dependencies can further impact the scheduler's ability to

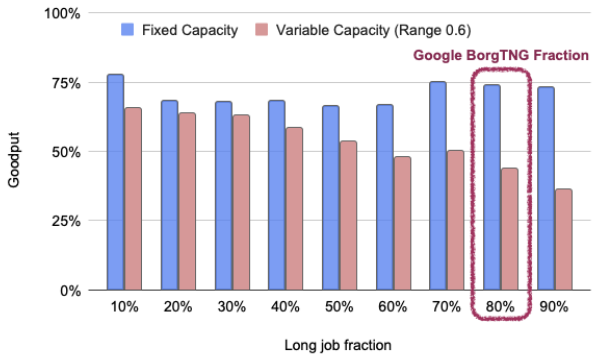
maintain system performance in the face of variable capacity. Scheduling delays in jobs that are on the critical path of workflows may result in limiting the scheduler's flexibility in scheduling waiting jobs and maximizing goodput. Moreover, terminations of these jobs may cause cascading job failures and further delays.

We compare performance on the same sampled workloads with and without job dependencies to evaluate the impact of workload structure. Figure 5.10a shows goodput varying dynamic ranges on the same workloads subject to job dependency constraints. Results show that while for some cells, the performance difference is not obvious, some cells exhibit further degradation in goodput from 10-15% when job dependencies are introduced into scheduling constraints. Figure 5.10b shows the job waiting time in bars (left y-axis) and job failure rates (right y-axis) in lines of workloads without and with job dependencies. It demonstrates that job failure rates of workloads with job dependencies under range 0.6 are similar to base workloads. However, job waiting time increases by 4 - 22%, due to the fact that the scheduler has decreased flexibility in scheduling existing job requests in the queue and job terminations may further delay-dependent job scheduling.

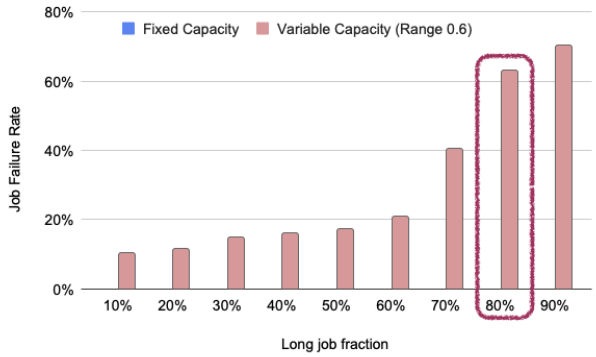
5.3.3 Workload Mixes

As discussed in Section 2.5, workloads across Borg cells are dominated by long jobs on core hours, with long jobs contributing to roughly 80% of total computation, following a Pareto distribution. Therefore, to understand what types of workload mixes can be supported well by the scheduler in the face of variation, we manually vary the core-hour fraction of long jobs in the workload mixes from 10% to 100% while keeping the total amount of computation constant. We measured the long job fraction in Google BorgTNG traces at 80%.

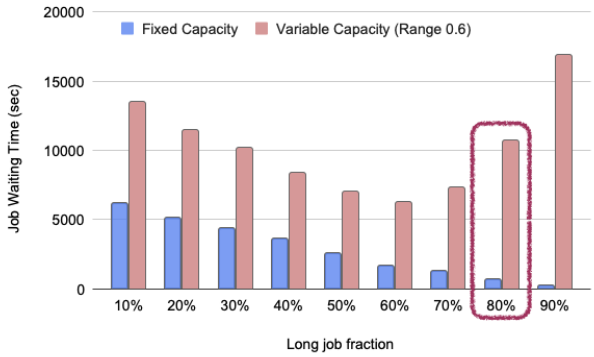
Figure 5.11 shows the performance results varying the fraction of long jobs on the x-axis. As we increase the long job fraction in the workloads, the goodput significantly decreases by more than 2X. The job failure rate increases as the long job fraction goes from 10% to



(a) Goodput



(b) Job failure rate



(c) Job waiting times

Figure 5.11: System Performance of Traditional Scheduler Varying Workload Mixes of Long and Short Job Fraction (Note: Google workload long job fraction indicated by the red box)

90%, reflecting the trend that long jobs are more prone to capacity changes. On the other hand, when there are mostly short jobs with 10-30% core-hours of long jobs, the workloads experience less degradation, due to a much smaller amount of job failures. For job waiting time in Figure 5.11c, the fixed capacity results show a continuous decrease. It is due to the decrease in the absolute number of total jobs as we hold the total computation constant while increasing the fraction of long jobs. A similar decrease in job waiting time is observed in the variable capacity results as the long job fraction increases. However, in the variable capacity scenario, as the long jobs become dominant in the total computation ($>$ 60% fraction), frequent job failures, reflected in Figure 5.11b, cause the jobs repeatedly being terminated and re-queued, and thus, drastically increases the job waiting time. These results suggest

that while a workload with mainly short jobs may tolerate capacity variations, general cloud workloads with a heavy-tailed distribution are not well-supported by traditional schedulers.

5.3.4 Drilldown Takeaways

Drilling down on real cloud workloads shows a dominant mode of VM usage. A substantial portion of long-running jobs, together with inter-task dependencies, present serious challenges to resource flexibility. These cloud-native workload characteristics produce severe goodput losses of 30-40%, job termination rate of 26%, and 60X increase in average waiting time under resource capacity variation, showing unacceptable performance degradation to the cloud data centers. These harms limit the acceptable dynamic range to <10%, constraining potential variation benefits. To enable larger variation benefits, new cloud needs new resource management techniques that can tolerate a greater dynamic range of capacity variation, while maintaining good performance.

5.4 Real Variation Scenarios

In the prior studies, we used synthetic variation, modeled on real variation, to systematically study the impacts of variation on scheduler performance and data center goodput. Here we use sources of variation from the real world directly, exploring how strategies such as price and carbon optimization based on varying power grid properties might implicate realistic variation properties and how they relate to synthetic traces we have extensively studied. Specifically, we focus on power prices, average carbon emissions/unit power, and stranded power. Optimization over these time-varying quantities is used to create variable resource capacity and derive variable resource traces from them. These traces are then used to evaluate scheduling systems. For each of these sources, we produce a set of sample traces with a duration of one year and a variety of temporal resolutions (5 minutes to hourly). These exemplar capacity traces are generated based on several simple policies, e.g. constant

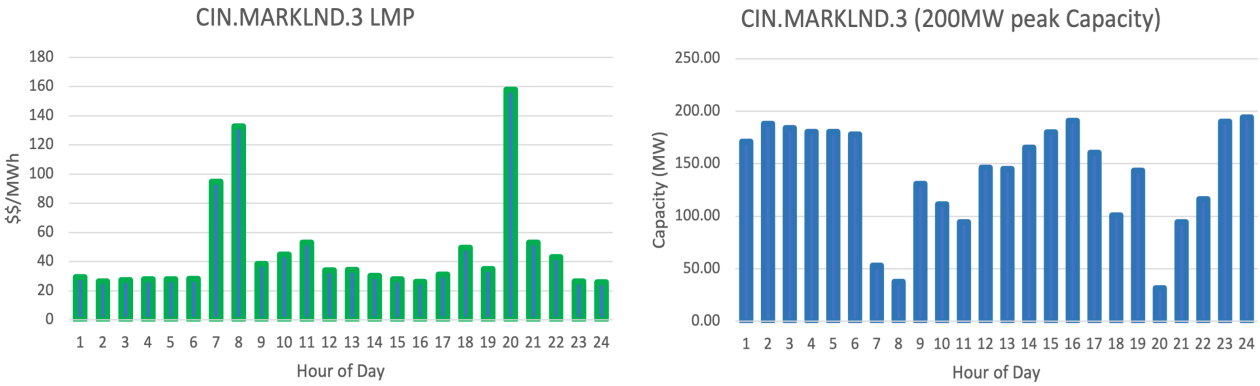
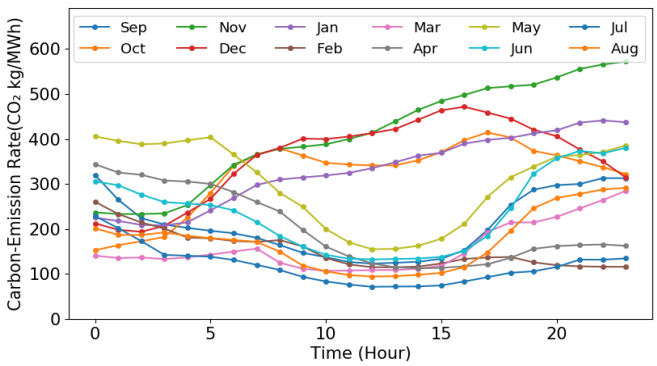


Figure 5.12: Power price ($/MWh) (left) and resulting resource capacity for a 200-megawatt data center (right), using a constant cost purchase approach. Exemplar 24-hour day from MISO January 9, 2018, CIN.Markland grid node.

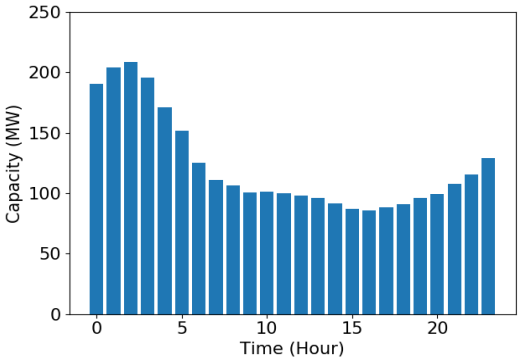
(hourly) carbon budget.

5.4.1 Variation from Price

In order to manage a supply cost (e.g. power), a common strategy is to constrain expenditures to a constant rate for an operating period. For example, in a fixed-capacity data center, the total power cost budget $Budget_{total}$ over a time period T is determined by $Budget_{total} = \sum_{t \in T} P_t \cdot Capacity$, where P_t is the power cost per unit capacity at time T. In a variable capacity data center with a constant price budget, the price budget $Budget_t$ at each time t is $\frac{Budget_{total}}{T}$. The capacity at time t is $Capacity_t = \frac{Budget_t}{P_t}$. In data centers or many types of machinery, this couples dynamic market price to resource capacity as illustrated in Figure 5.12, showing capacity variation of 5-fold [0.2, 1.0] or more. As the power prices can be volatile and spiky within minutes, variations produced by prices are similar to a random uniform during volatile price periods, dropping from 180MW to 52MW within a time interval. During these times, variation can be large over time periods as short as 5-minutes, and with very low (even negative) prices variable capacity may be limited by physical capacity.



(a) Carbon-emissions rate (mT/MWh) of one day per month from Aug 2019 - Jul 2020



(b) Resulting one-day resource capacity variation

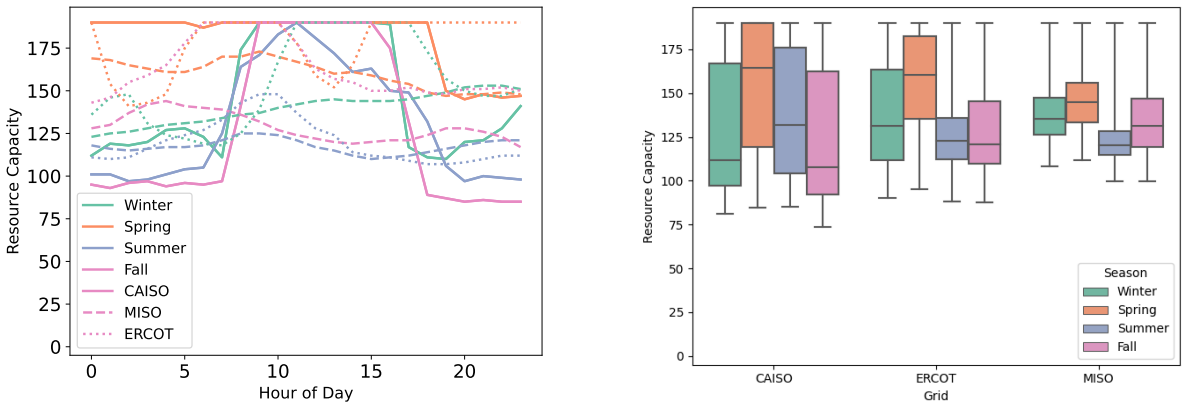
Figure 5.13: Carbon-emission rate (left) and resulting resource capacity at Constant Carbon purchase approach in the German power grid(December 2019, right).

5.4.2 Variation from Carbon Emissions

Concern is increasing about climate change, and thereby associated carbon emissions with power consumption. We derive capacity variation traces from power grids' historical carbon emissions to reflect the exploration of carbon opportunities. Carbon budgets must be managed against power grids with large fluctuations in carbon content. A basic strategy is a constant carbon budget for each time period. Similar to the constant power cost budget, the constant carbon budget approach keeps the carbon emission budget fixed at each time period. Figure 5.13a shows examples of carbon-emission rates in the German power grid for one examplar day per month from Aug 2019 - Jul 2020. As the grid has a high wind power penetration, the carbon emission rates follow a random walk structure of gradual changes. A resulting resource capacity variation at constant carbon purchase in Figure 5.13b shows a random walk of capacity variation, with a dynamic range [0.65, 1.05], comparable to a synthetic trace with a medium dynamic range of 0.4: [0.5, 0.9]. The step size varies from 0.01 to 0.2, where the former represents a much smaller change and the latter more challenging, representing the largest stepsize we evaluate, compared to a 0.1 stepsize from the synthetic traces of the same dynamic range.

Moreover, carbon emissions often vary not only daily, but also with patterns that differ by month of the year. For example, Figure 5.14a shows the resulting resource capacity of one exemplar day from three major power grids in the U.S. - CAISO, MISO, and ERCOT across four seasons in a year. Figure 5.14b depicts the yearly statistics of variable resource capacity. The lower and upper whiskers show the min and max resource capacity, and the lower and upper bound of the box shows the 75th, 25th-%tile. The midline of the box represents the median. It demonstrates that resource capacity varies over days and seasons and is widely distinct across the grids due to the renewable mix and weather. For example, CAISO, with a high solar power penetration, experiences rapid increases in generation during sunrise and rapid loss during sunset. This pattern of variable renewable generation produces capacity variation with a dynamic range of [0.4, 1.0], comparable with the largest range of synthetic traces. Throughout the rest of the day, the capacity is observed with a rather stable quantity, representing a simpler case than synthetic variations. MISO, which has wind-dominant renewable generation, exhibits variations with random walks. Its dynamic ranges are as large as [0.6, 1.0] but mostly within [0.65, 0.78], representing a scenario with smaller and simpler variations. ERCOT, with a mix of both solar and wind generation, demonstrates a structure and dynamic ranges which fall between CAISO and MISO.

Overall, across the grids, spring seasons provide more capacity opportunities, presenting larger quantities of total capacity, and summer seasons demonstrate less capacity also less variation (≤ 0.1), reflecting a combination effect from high demand and low renewable generation. Over one year period, a constant carbon approach produces up to 3% additional total capacity while achieving 9% carbon emission reductions, showing promising carbon benefits. Note that workload SLOs such as "catchup by end of the day" can have difficult interactions with the shape of variation curves.



(a) Realistic variation traces generated by constant carbon approach (MW)

(b) Statistics of variable capacity produced by constant carbon budget

Figure 5.14: Statistics of grids' historical carbon emission rate, variable capacity produced by constant carbon budget, and examplar variation traces (MW)

5.4.3 Variation from Stranded Power

A different approach to lower carbon emissions is stranded renewable power [55, 151], where excess renewable energy (power with zero-marginal carbon) can be used to power data centers intermittently. This excess case may be important for combatting climate [151, 152], and produces a nearly binary on-off resource capacity (Figure 5.15, ERCOT), while operating at zero carbon emissions, which can be viewed as an extremely volatile random walk of range 1.0: [0, 1.0] with stepsize 1.0. The graphs illustrate 15-minute intervals (high frequency of variation) and reflect variation over a week-long period. The power availability variation is day-to-day, week-to-week, and also by the season of the year.

5.4.4 Scheduling Experiments on Real Variation Traces

To validate our results of synthetic variation traces, we now look at the system performance using real variation traces, produced by the constant carbon approach from three major power grids in the U.S., shown earlier in Figure 5.14, to demonstrate the generality of scheduler performance under variable capacity. For each power grid, we focus on displaying

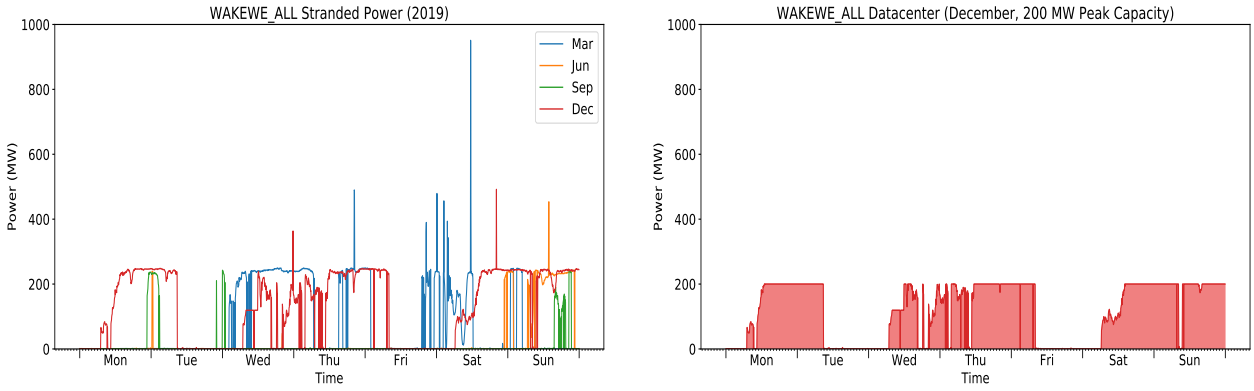


Figure 5.15: Stranded Power (curtailed and negative priced power) in 15-minute intervals for a node in the ERCOT power grid (left, each line is a different week), and the resulting resource capacity for a 200 megawatt datacenter for the week in December (right).

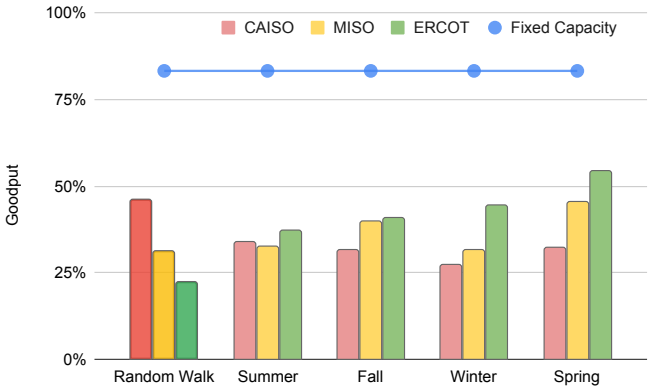


Figure 5.16: System Performance (Goodput) of Traditional Scheduler with Real Variation Traces, Comparing with Synthetic Random Walk of Range 0.2, 0.4, and 0.6

a representative single day, whose standard deviation of capacity variation is the 90%-tile value, for each season, reflecting the future grid with more fluctuations.

Figure 5.16 shows the goodput results of a variable capacity data center (same size as used with synthetic traces), following the variation of the examplar days across seasons and power grids, comparing with using synthetic traces of random walk (range 0.2, 0.4, and 0.6) in the leftmost cluster of bars. It shows the same level of goodput degradation as in Figure 5.10a due to capacity changes. Corresponding to the grid-level variation properties discussed in Section 5.4.2, goodput results under real variations show better performance than synthetic

traces with the same dynamic range by dealing with simpler, easier scenarios. For example, CAISO experiences more goodput degradation because its variation exhibits a large dynamic range. However, its performance still outperforms goodput from a synthetic random walk with 0.6: [0.4, 1.0] as the overall variation and stepsize are smaller, representing a simpler view. MISO and ERCOT, with random walks and smaller dynamic ranges, experience even less degradation.

For seasons like summer and winter where fewer opportunities are available due to high demand, the goodput is comparable with a walk with a range of 0.4. For spring and fall with more carbon emission fluctuations, the performance under realistic traces shows less degradation, around and even higher than results of range 0.2, because the variation curves are smoother and contain larger total quantities of resource capacity. These results reflect the opportunities from fluctuating renewables in the grids and variation benefits vary widely across seasons and grids. Overall, these real variation traces can be mapped to synthetic traces by varying three key dimensions of the problem space. In often cases, with the nature of renewables and hourly changing generation mixes, the real variation traces exhibit a smoother case of the variation space - a random walk structure with a 1hr change frequency with a 0.2-0.5 range with small step sizes.

Along with the performance validation, these results demonstrate that real variation traces from a constant carbon approach are a specific representation, maybe a simpler version, of the synthetic traces, which cover all the scenarios varying three key dimensions. Future power grids, as they drive to higher RPS goals, will experience increased volatility in power capacity due to high renewable fractions. We pick days whose standard deviation of capacity variation is 90%-tile for the season to project future grids with larger volatility. Although these results exhibit a simpler representation of the synthetic traces, a holistic view to take future power fluctuations into account using synthetic traces demonstrates realistic, challenging projections.

5.5 Summary

In this chapter, we presented our approach to understanding the performance impact under resource capacity variation. We defined the variation problem and three key dimensions of variation, dynamic range, structure, and frequency. Empirical evaluation of a diverse range of workloads, both cloud and HPC, shows that scheduling performance significantly degrades under capacity variation, with goodput loss of 15 - 60% and 30% on average. All dimensions of variations contribute to goodput loss, and each independently decreases goodput by 10 - 40%, suggesting careful control over these factors.

We further drill down on Google's cloud workloads with a range of clusters and sizes. Our study shows that cloud workloads have a dominant mode of VM usage, with a large fraction of long-running jobs and inter-task dependencies. These properties further degrade performance, incurring goodput losses of 30 - 40%, and job termination rate of 26%. These factors combine to produce a 60X increase in average waiting time. The negative impacts limit tolerable variation to <10% of the total load. These results all suggest that to enable larger benefits arising from dynamic capacities, such as carbon footprint and operating expenses, new data centers need new resource management techniques that can tolerate a greater dynamic range of capacity variation while maintaining good performance.

We show examples and empirical traces of sources that lead to resource capacity variation: power prices, carbon emission, and stranded power. We demonstrate that with key dimensions, these distinct scenarios can be characterized, abstracting it as a generic problem. Scheduling studies with real variation traces from carbon emissions further validate that these sources exhibit similar dynamic ranges and represent a specific, simpler case of the broad range of synthetic traces within the same range. Thus, they can be umbrellaed under the broader, systematic study of the whole problem space. As future grids trail to high-renewable power systems with accelerated decarbonization goals, more challenging cases in synthetic studies may represent a realistic projection to capture the increased volatility.

CHAPTER 6

COPING WITH CAPACITY LOSS

Results in Chapter 5 show that resource capacity variation can degrade scheduler performance (lower goodput). A key impact is increased job failures that arise when the scheduler assumes stable capacity, but capacity drops. Therefore, capacity decreases can cause a large number of job failures, resulting in wasted resource efficiency and damaged job experiences. We first characterize the resulting job failures from capacity loss. Then to cope with capacity loss, we propose and evaluate intelligent termination policies to minimize job failures and mitigate wasted work by selectively terminating jobs in Section 6.1, illustrated in Figure 6.1. Section 6.2 evaluates the effectiveness of using capacity foresight information to improve performance by varying the length of foresight. We conduct a case study of a hypothetical carbon-emission-aware datacenter in Germany to demonstrate the carbon, cost, and capacity benefits from variable capacity in Section 6.3. Finally, Section 6.4 summarizes the chapter.

6.1 Intelligent Termination Policy

We look at job failure rates of both HPC and cloud workloads varying dynamic range and variability structure to characterize the impact of capacity loss. Figure 6.2 shows that job

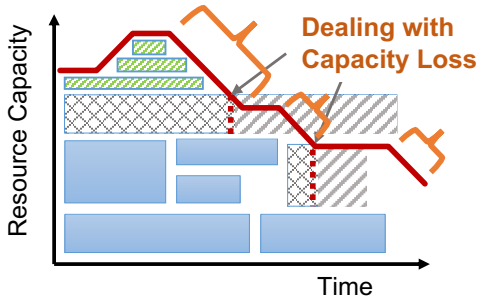


Figure 6.1: Coping with Capacity Loss Through Intelligent Termination Policies in A Variable Capacity Data Center

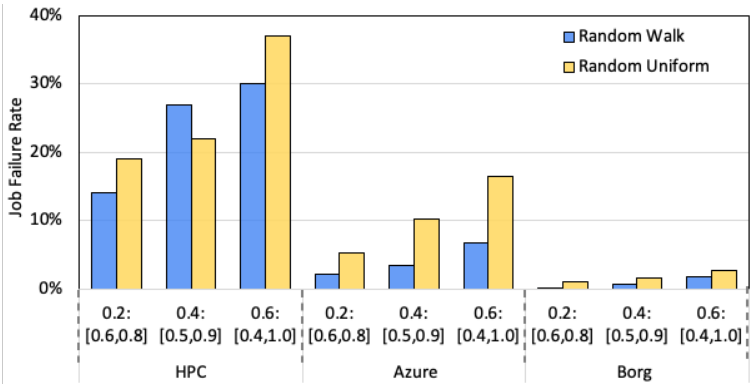


Figure 6.2: Job failure rate with resource capacity variation, varying dynamic range and variability structure.

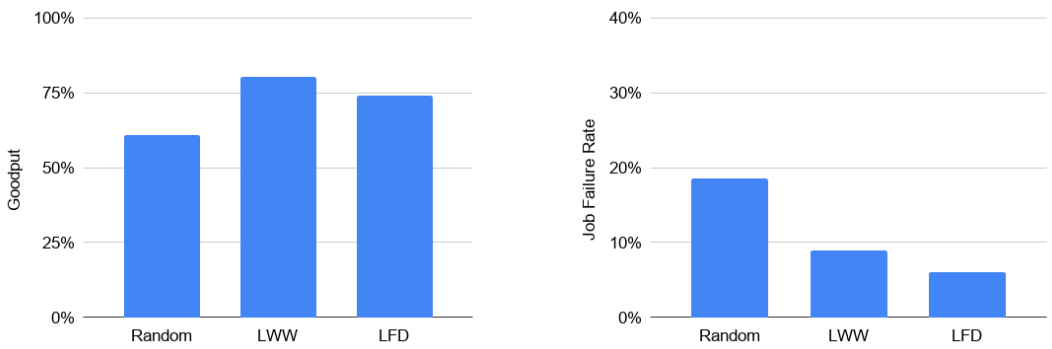
failure rates increase with the dynamic range and structure of resource capacity variation.[1] For HPC workload, the scheduler experiences 15% job failures with dynamic range [0.6, 0.8] and this rate rises to 30% for [0.4, 1.0] (random walk). For Random Uniform the rate is even worse, growing from 19% to 37%. Such high failure rates not only account for significant goodput losses, but they also produce a poor experience for applications. For Azure workload, the trend is similar, but the magnitude of degradation is much less, peaking at 6.7% for random walk and 16.5% for random uniform. For Borg V2 result, the failure rates are lower still, due to lower resource use.

When capacity decreases below the scheduled workload, jobs must be terminated (fail) until the total resources in use match the current capacity level. Because high job failure rates incur a large quantity of wasted computation, we explore how to best choose the jobs to be terminated with the goal of minimizing job failures and maximizing goodput and job experience considering job features and progress.

We consider three policies:

- *Random*: Select a node randomly, terminate the associated job, and free its resources.
- *Least Wasted Work (LWW)*: Select the job whose termination wastes least work (small-

1. Here the job to be terminated is chosen randomly.



(a) Goodput (b) Job Failure Rate

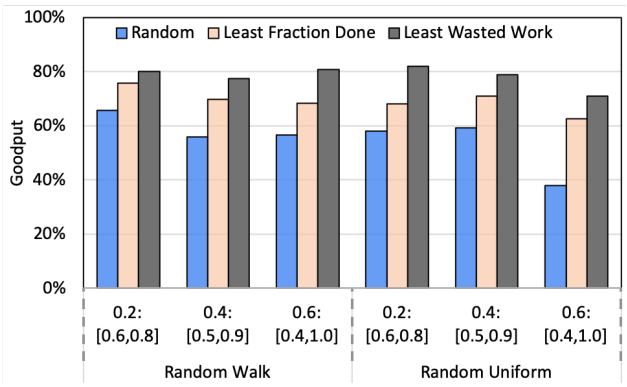
Figure 6.3: Scheduling performance for various termination policies (random walk, dynamic range [0.4, 1.0], step size 0.05)

est $nodes \times (t - start\ time)$, where t is the current time) and free its resources. This policy exploits the information of current progress (elapsed time) for each job, which is known and recorded in the system, to minimize wasted computation.

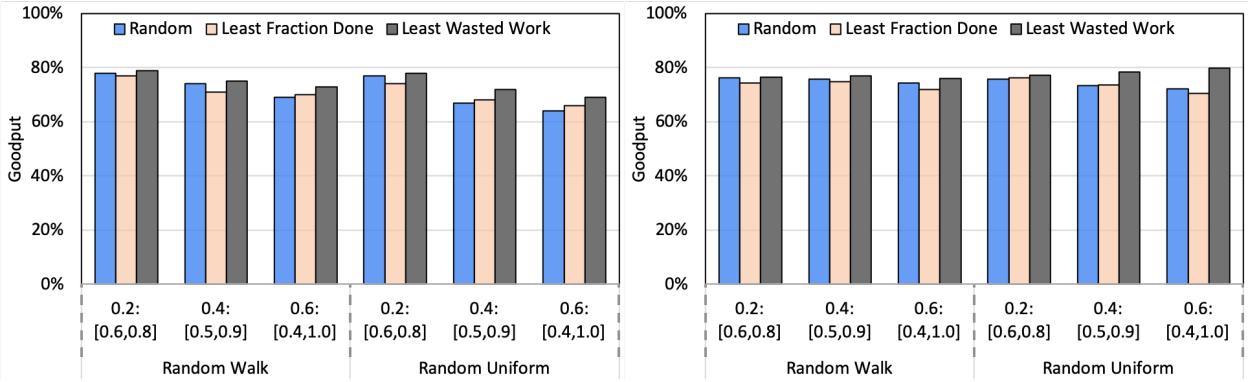
- *Least Fraction Done (LFD)*: Terminate the job which is least fraction completed (minimum $\frac{(t-start\ time)}{runtime_j}$, where t is the current time) and free its resources. This policy requires knowledge about job total runtime, in addition to elapsed runtime, to assess each job's progress to completion.

For each policy, we repeat the process of terminating jobs until the desired (lower) resource level is reached. For the HPC workloads, we use the requested runtime to compute LFD; for the commercial workloads we use the trace information for actual job length. However, in production, this information is not generally available. We compare the termination policies, using scheduler performance metrics of goodput and failure rate. In particular, we would like to improve scheduler performance under resource capacity variation, and further understand which of these termination policies work best.

We begin with a dynamic range of [0.4,1.0], as our experiments in Section 5.2.1 show significant scheduler performance degradation under these conditions. Both of our new policies, LWW and LFD perform much better than Random, increasing goodput and reducing



(a) HPC



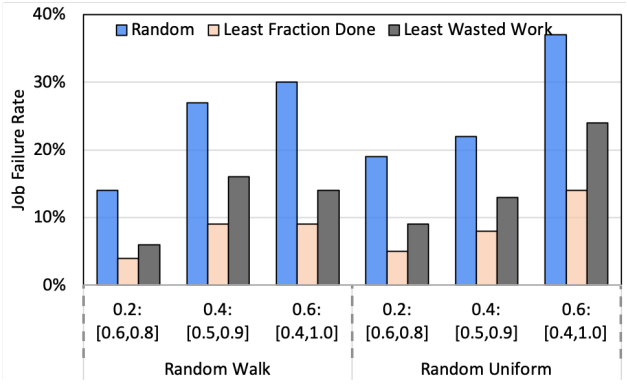
(b) Azure

(c) Borg

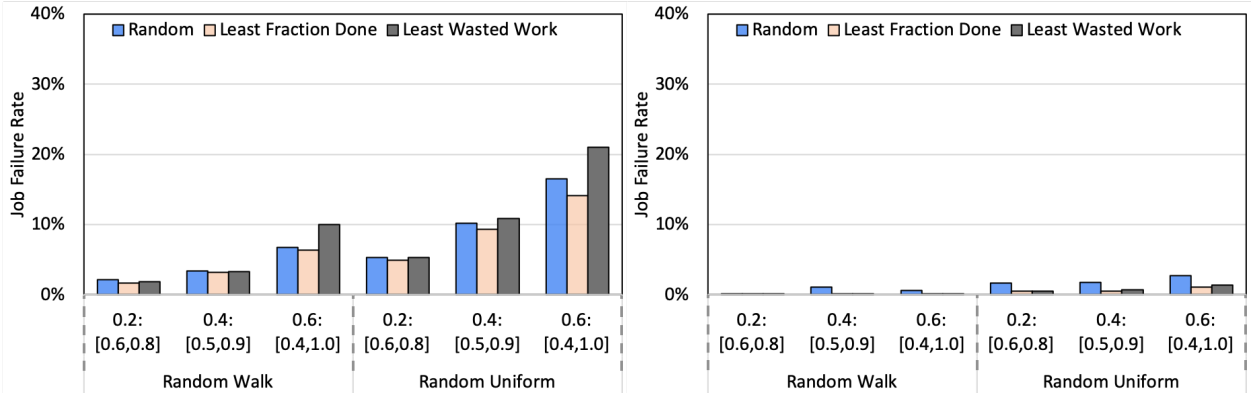
Figure 6.4: Goodput versus termination policy, varying dynamic ranges and structures

failure rate as shown in Figure 6.3. This experiment uses random walk with stepsize 0.05. The goodput shows 26% and 31% increases respectively, compared to Random. And the improvement in job failures is dramatic, improving by 2 to 3-fold. This is remarkable, given there is no advance warning. The improved result is comparable with goodput with stable capacity (80%). This result shows the promise of intelligent resource management to tolerate resource capacity variation.

Broadly, Figure 6.4 presents goodput results for a variety of dynamic ranges and variability structures. The results show that intelligent termination policies make a big difference. For HPC both intelligent termination algorithms improve performance, but the best performance is achieved with LWW (rightmost, gray). The goodput achieved by LWW approaches the stable resource capacity and is an average of 44% improvement over Random. For Azure



(a) HPC



(b) Azure

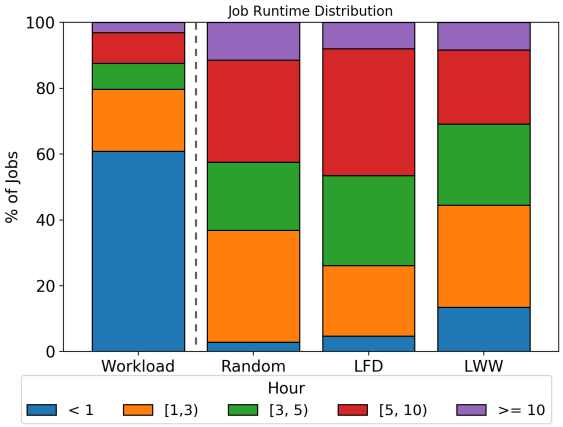
(c) Borg

Figure 6.5: Job failure rate versus termination policy, varying dynamic ranges and structures

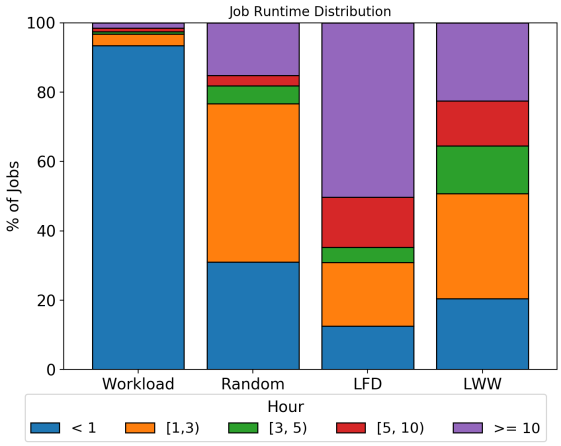
and Borg V2 workloads, the algorithm preference is similar, with LWW producing the highest goodput, but with smaller benefits.

Looking deeper, we observe how failure rates increase with dynamic range (leftmost, blue), and the dramatic reduction achieved by intelligent termination policies (see Figure 6.5). Job failures are reduced in all cases, and by as much as 3-4 fold. Interestingly, for all workloads, LFD reduces job terminations more, but LWW produces better goodput. This matches intuition as LWW terminates short and small jobs – and more of them, but wastes less work in general. On the other hand, LFD normalizes out job size and run length by calculating the fraction done. The Azure effects are smaller than HPC, and the Borg V2 effects are smaller still. Later, we examine the resulting fairness effects.

(a) HPC



(b) Azure



(c) Borg

Figure 6.6: Job terminations sorted by runtime. Workload distribution as reference. (Random walk, dynamic range 0.6:[0.4,1.0])

Intelligent Policies and Fairness Any intelligent policy can potentially affect different types of jobs in the workload differently. To examine this question, we plot the job termination distributions in Figure 6.6 with the overall workload distribution at the left of each graph for reference. Each bar represents one policy and the resulting job terminations distribution (by job length (runtime)).

All of the termination policies, including random, have a greater impact on longer and larger size jobs whose greater extent gives them a greater chance to interact with the temporal nature of capacity variation. Both LWW and LFD reflect these effects but recall that LWW has both higher termination rates and higher goodput. LWW achieves greater fairness by

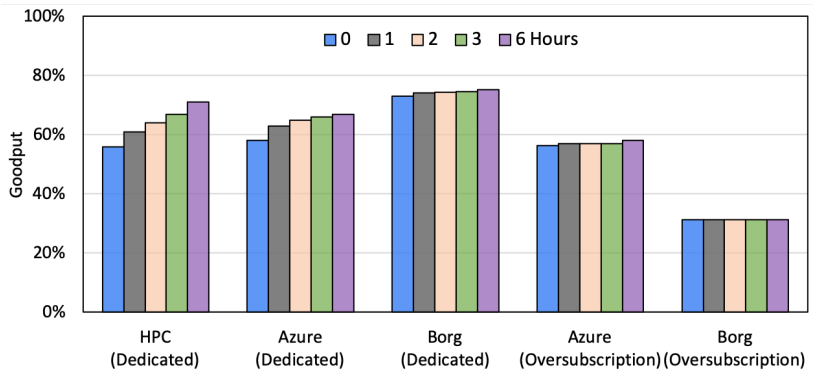


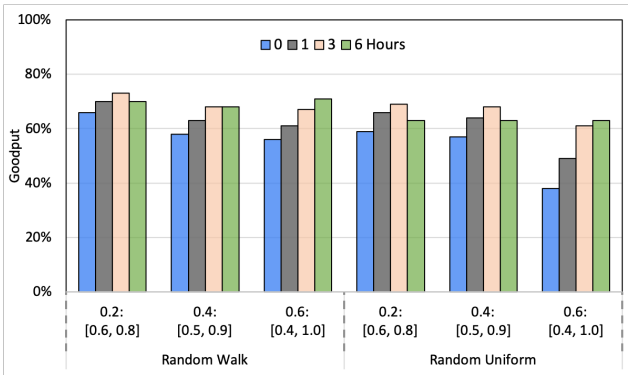
Figure 6.7: Goodput versus advance warning (random walk, dynamic range 0.6: [0.4, 1.0]).

runtime and size compared to LFD for all three of the dedicated workloads. Interestingly, in the Borg V2 workload, random is slightly fairer than both of the intelligent algorithms.

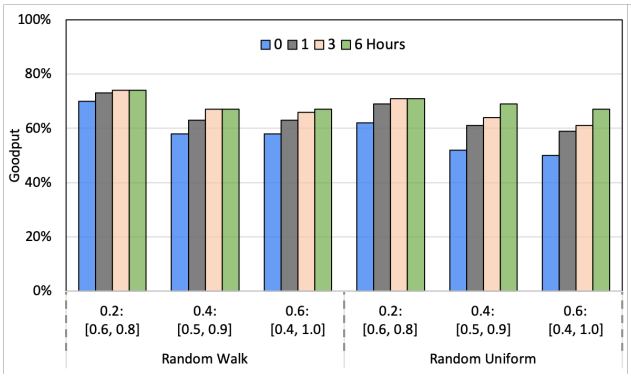
6.2 Foresight

With variable resource capacity, some forms of variation may be predictable or controllable – at a cost. We explore how well schedulers can exploit foresight of resource variability, adapting to an irregular resource projection. We give the scheduler foresight (a window of visibility) of zero (baseline) to 24 hours (longest job duration). The scheduler only makes commitments as long as the longest job, so 24 hours gives full knowledge. To fully exploit the oracle capacity information, we further exploit full information of job runtime so that the scheduler can fully align job placement with known capacity within the foresight window to avoid job failures. We first start with an example of dynamic range 0.6: [0.4,1.0] and the structure is random walk with stepsize of 0.15. In Figure 6.7, for all workloads, three to six hours of foresight is required to eliminate the variation penalty, but longer foresight has no further improvement (thus not shown).

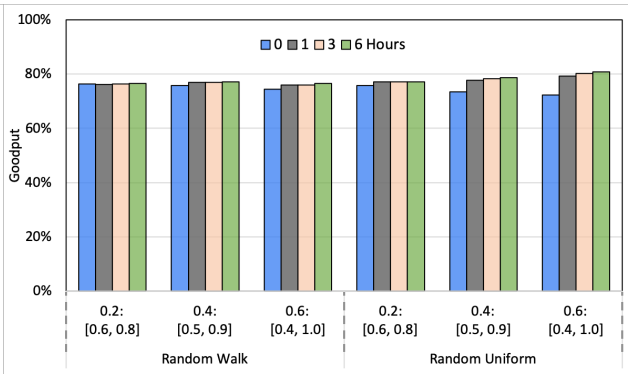
Putting it all together, we vary parameters of dynamic range, structure, and warning time, presenting the results in Figure 6.8. For HPC and random walk, there are strong benefits for increased advanced warning and pronounced benefits with larger dynamic ranges. For



(a) HPC



(b) Azure



(c) Borg

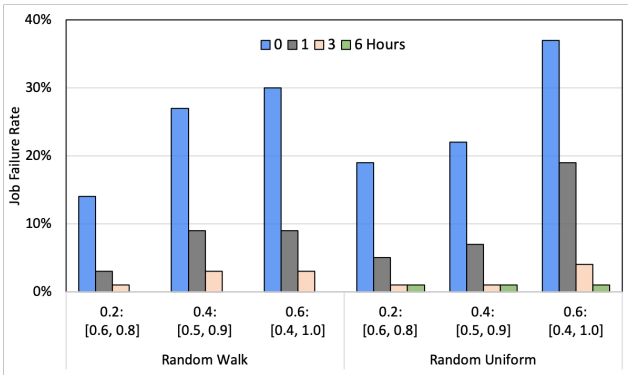
Figure 6.8: Goodput versus warning time (foresight), varying dynamic range and structure.

Azure, smaller benefits accrue for both random walk and random uniform. There are only small benefits for Borg V2.

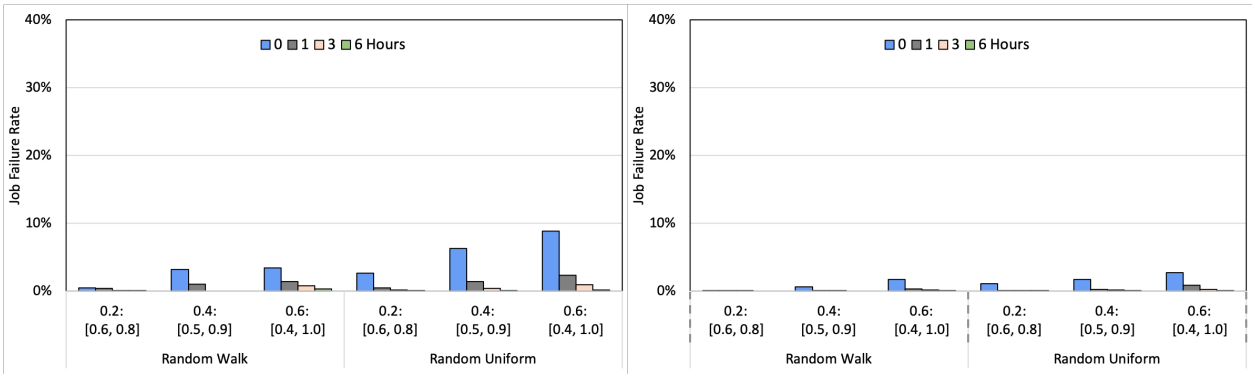
Next, we consider the impact of foresight on job failure rates (see Figure 6.9). In the HPC model and commercial workload models, the failure rate grows with the dynamic range. For all, increased advance warning reduces failure rates dramatically, though the failure rate starts much higher for HPC, and is lower for Azure, and Borg V2. Six hours are clearly enough to eliminate nearly all of the job failures.

6.3 Case Study: A German Datacenter

To illustrate the impact of carbon-based power acquisition and scheduling performance in a real-world scenario, we consider a hypothetical 40-megawatt data center operating in the



(a) HPC



(b) Azure

(c) Borg

Figure 6.9: Job failure rate versus warning time (foresight), varying dynamic range and structure.

German Power Market[70]. Because the power market varies every day and has a strong seasonal structure, we pick a set of exemplar days from the 12 most recent months (Sept 2019 - August 2020). When using constant carbon emissions per hour, they have power variation as shown in Figure 6.10. These twelve days have 24-hour capacity increases from 6% to 16% with an average of 11%.

We use the same HPC Mira workload, the corresponding system, and Cobalt HPC scheduler because evaluation results from Chapter 5 show that HPC workload can be more vulnerable to resource capacity changes than commercial workloads. We compare a traditional operating mode (fixed power), constant carbon emissions (carbon-emissions-aware), and then add foresight and then the scheduler enhancements, graphing goodput in Figure 6.11. Each cluster of bars depicts the results for a single exemplar day. Shifting from fixed to Carbon-

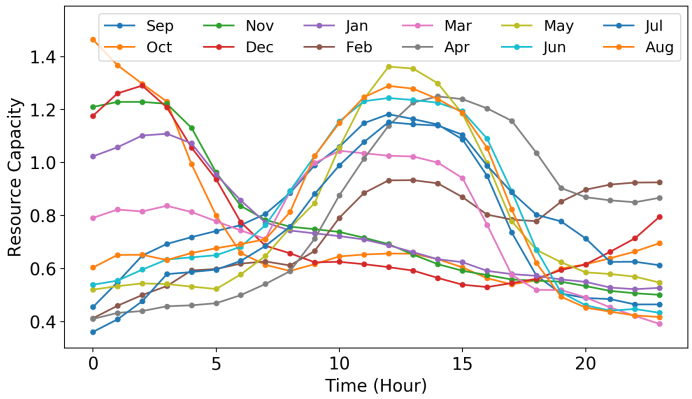


Figure 6.10: 24-hour resource capacity variation by Carbon-Emission-Aware approach for acquiring power of an exemplar day per month

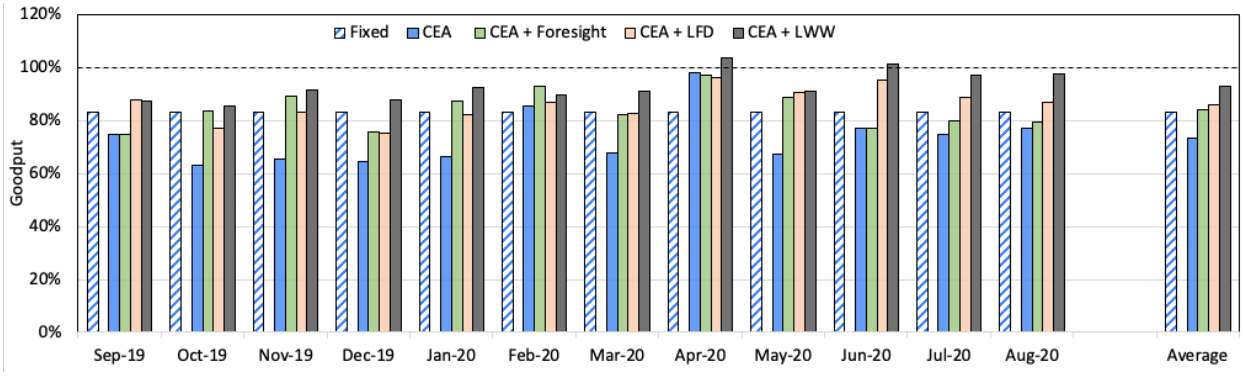
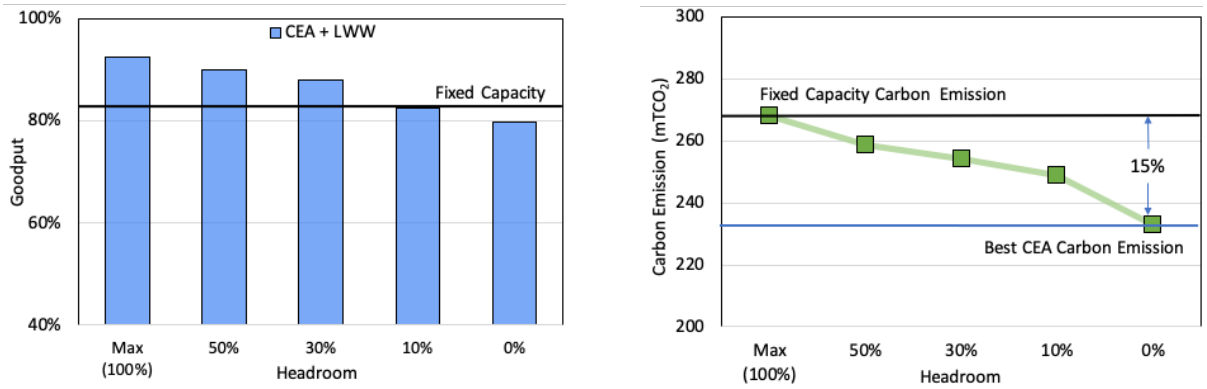


Figure 6.11: Goodput for 12 exemplar days, comparing fixed and carbon-aware power consumption, various schedulers, Mira trace, and simulation.

Emission-Aware(CEA) power acquisition produces a large drop in goodput as large at 24% on some days and 12% on average. Finally, we consider an alternate approach, modeling the use of advanced power market information, we give the scheduler 3 hours of foresight. This is a little optimistic, as power markets can be unpredictable. Next, we consider using the scheduler improvements identified (without foresight) in Section 6.1. Both LFD and LWW are productive, but LWW eliminates essentially all of this degradation. In fact, CEA+LWW is not only 11% better overall, but it also outperforms fixed on every one of the 12 days. Note that CEA+LWW actually exceeds the 100% line for the fixed power acquisition in two days! (this is correct, and reflects exploiting the headroom, which is the additional surge capacity in the data centers)



(a) Goodput (b) Carbon emissions

Figure 6.12: Performance and Carbon emissions of a model German Datacenter

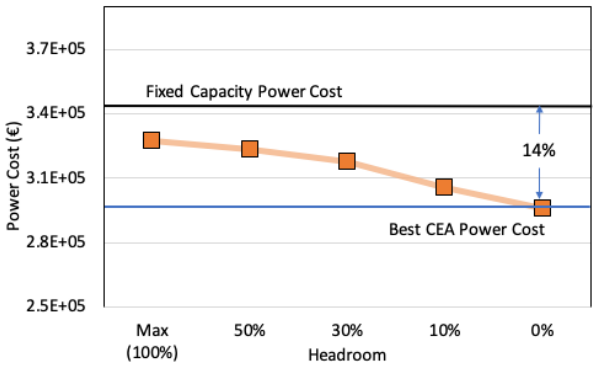


Figure 6.13: Resulting power cost of a model German Datacenter

Most data centers have power and cooling headroom[2] and hardware overprovisioning is increasingly popular [124, 123]. However, the quantity is of course limited. To assess these limits on performance and carbon emissions, we show the average goodput for our 12 exemplar days versus capacity headroom (see Figure 6.12). As data center headroom decreases, the potential benefit declines at 20%, it is effectively disappeared. However, CEA is still worthwhile, as carbon emissions are significantly reduced (by about 8-10%) at the same goodput, or by 15% with a 3% loss in goodput. Projects such as Zero-carbon Cloud and other lightweight and free-cooled datacenter projects have shown that this type of "headroom" can be often constructed for a fraction of the proportionate cost [152, 25].

2. Headroom is the ability to temporarily exceed these limits safely and can also exploit thermal inertia.

Finally, we calculated datacenter power cost, using hourly prices. The impact of the CEA power acquisition on total power cost is significant, with reductions that parallel carbon emission reductions reaching as large as 14%.

6.4 Summary

The study of traditional schedulers on real HPC and cloud workloads shows that capacity loss from variation can incur a large number of job terminations, imposing job failure rate from 15-37% on HPC workload and 1-16% on cloud workloads. To cope with capacity loss, we propose intelligent termination policies to reduce job failures and restore resource efficiency. Results show that they are effective in mitigating the impact of capacity loss from variation, increasing goodput by 10 - 66% and reducing job failures by 1.6 - 3X.

In addition, foresight information is powerful as it alone can drastically improve performance. Six hours of foresight can eliminate nearly all job failures, demonstrating promising opportunities to exploit information.

A case study of a German data center demonstrates the benefits of up to 15% carbon emission reduction and 14% power cost savings by effectively exploiting capacity variations while maintaining goodput.

CHAPTER 7

A BROADER VIEW: PREPARING FOR CAPACITY VARIATION

While Chapter 6 has shown promising results of intelligent termination policies, dealing with capacity loss can only minimize the job failures and wasted computation upon a capacity decrease in a reactive fashion and thus has a limit on how much it can mitigate. For example, in an extreme case where a data center is fully occupied with jobs of identical size, the opportunities to mitigate the negative impact in the event of capacity loss are very limited. Therefore, in this chapter, we take a broader view to consider strategies for preparing for capacity variation. That is, proactively preparing for capacity increase and planning for capacity decrease, shown in Figure 7.1. As evaluation in Section 6.2 demonstrates the powerful improvements of partial oracle information, we further explore information to improve performance. We present the dimensions of uncertainty that contribute to performance loss in the face of resource capacity variation and the information space of these two dimensions that may help reduce such uncertainty for the schedulers in Section 7.1. We propose scheduling algorithms that exploit such information to optimize job placement decisions preparing for capacity variation in Section 7.2. We empirically evaluate these scheduling algorithms using Borg TNG cloud workloads varying workload properties in Section 7.3 and summarize in Section 7.4

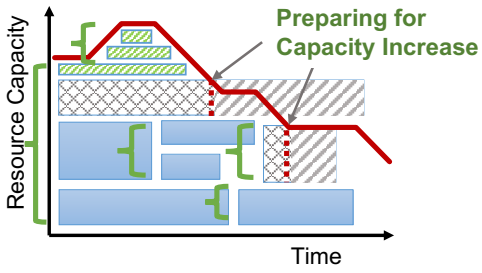


Figure 7.1: Preparing for Capacity Variation Through Scheduling Schemes in A Variable Capacity Data Center

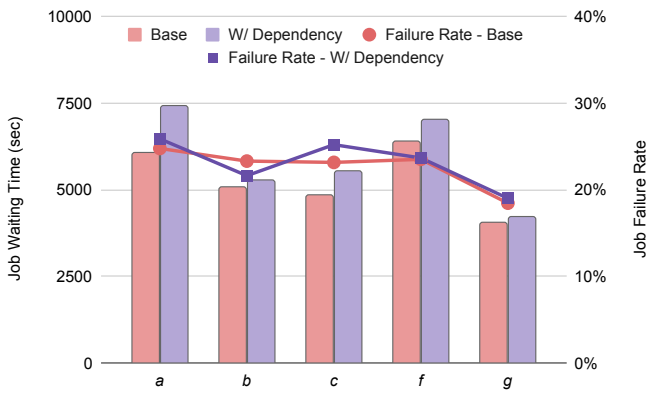


Figure 7.2: Variable capacity challenges scheduler's ability to achieve low waiting time

7.1 Uncertainty and Information Space

Compared with a traditional data center, the mismatch between resource capacity assumptions and actual variations, coupled with uncertainties of job runtimes, has limited job schedulers' capability to properly place jobs onto machines that will continue to be available until jobs' successful completion based on job run-length. To illustrate, the drill down of Google cloud workloads shows that the goodput significantly degrades due to abrupt job terminations. Figure 7.2 exhibits a significant increase in job failure rates, from 0%(fixed) to $> 20\%$, causing a large increase in job waiting time as jobs get re-queued and delayed.

Therefore, properties together can limit the types of workloads and structures that can be supported. Section 6.2 shows that with good job run length information and perfect capacity prediction, the scheduler can schedule jobs to avoid termination. To understand the effectiveness of information, we systematically organize the scheduler's information space based on two dimensions of uncertainty, job runtime, and resource capacity, and vary the amount of it. Conventional schedulers in fixed capacity systems fall outside this space, as indicated in Table 7.1. We explore the information space of workloads and resource capacity that can reduce the degree of uncertainty experienced by the scheduler. To explore the space, we look at them individually from no additional information to perfect knowledge.

		Resource Capacity			
		Fixed	Variable		
		-	Unknown	Bounded Range	Oracle/Prediction
Job Runtime	None	X	*X*		X
	Coarse, All jobs	-	*X*		X
	Exact, All jobs	-	*X*	*X*	X

Table 7.1: Source of Uncertainty and Information Space in Variable Capacity Data Center

7.1.1 Workload Information

We first look at workload information, as the runtime property of workloads impact scheduler performance in the face of variation. For example, longer-running jobs inherently experience more capacity changes and more potential decreases as they stay longer in the system. In addition, longer-running jobs account for a large amount of total computation and failures cause more wasted computation. These reasons make longer jobs more susceptible to capacity variation and shorter jobs more flexible in job placements. To identify this difference of jobs in the information space, we vary the amount of job runtime information to None, Job Runtime Classification, and Exact Job Runtime.

None Many cloud schedulers assume only job resource requirements, but no run duration information (and not duration limits) at scheduling [137, 50].

Job Runtime Classification Studies have shown cloud workloads have highly skewed runtimes. Simple partial information might consist of a classification, where each job is labeled {long, short} relative to a runtime threshold t. While knowing or predicting the exact runtime of jobs may be hard, many studies have demonstrated that given the skewness of cloud workloads, such simple classification is accurate[158, 39]. This provides runtime approximations of all jobs. Such classification allows the schedulers to separate the jobs into two bins based on the runtime and thus enables schedulers to make placement decisions to protect long-running jobs from frequent interruptions due to capacity changes.

Exact Job Runtime This represents the oracle information of all job runtimes. This oracle allows us to explore how well a scheduler could do, fully armed with certain job information. Exact job runtime enables the schedulers to assess risks for jobs based on runtime and make informed placement decisions. Various studies have demonstrated that by using exact or accurate job runtime information, a scheduler can significantly improve resource utilization and reduce job wait time under various scheduling algorithms, such as schedulers with Backfilling, Short-Job-First, and Gang. In practice, such information is not available for cloud workloads and is not even available in batch-scheduled supercomputers.

7.1.2 Capacity Information

Variable capacity data center gives rise to the uncertainty of resource capacity. In a traditional data center, a scheduler knows the perfect information of resource capacity into the infinite horizon, which is the fixed data center capacity. In contrast, in a variable capacity data center, the scheduler does not know what capacity will be available in 1 hour, 3 hours, or 10 hours into the future. With capacity variation, the scheduler must act in the face of uncertainty, sometimes making the wrong assumptions and causing future misalignment of workload and resources. Therefore, we consider three types of resource capacity information that might be available to reduce uncertainty and its impact – None, Bounded Range, and Foresight.

None As a baseline, we consider a data center whose capacity is dynamically managed to optimize power cost or carbon content. In this case, no information is available to the scheduler for future capacity, so the schedulers must manage the risk of capacity decreases.

Bounded Range This illustrates the additional information on resource capacity where the lower and upper bound of the variation are known. If the dynamic range of variation is known, the scheduler can assess where the current capacity lies in the variation space and

estimate if the capacity is likely to increase or decrease in the future intervals to be more aggressive or conservation and change placement strategies accordingly to either capture more opportunities or avoid job failures. The scheduler would also have information on the number of machines that are always available despite capacity changes based on the lower bound.

Foresight Near-term capacity decreases can cause significant harm, and increases can be difficult to exploit. We consider a model where precise future capacity is known for some horizon – 0, 1, 2, .. up to 24 hours – into the future. The shorter periods could be achieved with a time series predictor and the latter corresponds to a power-grid day-ahead plan. Since foresight, unlike scheduling algorithms, is a limited horizon of future variation, it can be combined with any scheduling scheme to further improve performance by varying the length of foresight.

With foresight, the scheduler can further optimize the placements for all jobs that have runtime shorter than the horizon to eliminate job terminations and maximize goodput. That is, with the visibility of the variation curve within a window, the scheduler is able to pack the resources with short jobs within the curve as tight as possible to maximize goodput and minimize job wait time without terminations. If the foresight is longer than the runtime of any jobs, it is equivalent to full knowledge as jobs can be scheduled to avoid unknown future capacity decrease. However, if it is short, the job scheduler may still schedule jobs more than the resource capacity available in the future. And for jobs that have longer runtime, since the scheduler does not have information beyond the horizon, it needs to fall back to the existing strategy.

		Resource Capacity			
		Fixed	Variable		
		Oracle	Unknown	Bounded Range	Oracle
Job Runtime	None	*Conventional*	*Random*		
	Coarse, All jobs	|	*LongShort*		
	Exact, All jobs	|	*JobSize-Order*	*TwoMode*	*Oracle*

Table 7.2: Information Space and Corresponding Scheduling Algorithms Exploiting Information to Prepare for Capacity Variation

7.2 Scheduling Algorithms

To minimize job failures and improve resource efficiency, we consider scheduling algorithms to prepare for capacity variation. That is, scheduling algorithms to make informed placement decisions to not only avoid terminations during future capacity decreases as much as possible but also account for potential capacity increases to maximize utilization. We propose scheduling algorithms that exploit the information space of job runtime and variable capacity, demonstrated in Table 7.2.

Baseline It has no additional information on workloads or variable capacity, representing a class of traditional scheduling schemes that are variation oblivious. It uses a first-come-first-serve (FCFS) scheduling policy and applies a first-fit placement policy to all jobs. For a job Job_i, the algorithm places it onto the first machine M_k which both has availability $A_k = True$ and unused resource capacity $R_k - \sum_{j \in J} RR_j \cdot P_{jk} \geq RR_i$ It is widely observed to give good results with low parallelism and is widely used. Upon resource capacity decreases, as some running jobs may have to be terminated (fail) due to capacity changes, the scheduler randomly selects a machine M_r, terminates the associated jobs $\{j \in J | P_{jr} = 1\}$, and relinquish its resources. It repeats the process until the desired (lower) resource level is reached, $\sum_{m \in M} A_m = C_t$. The terminated jobs will be put back into the waiting queue to be rescheduled and all intermediate computation will be lost.

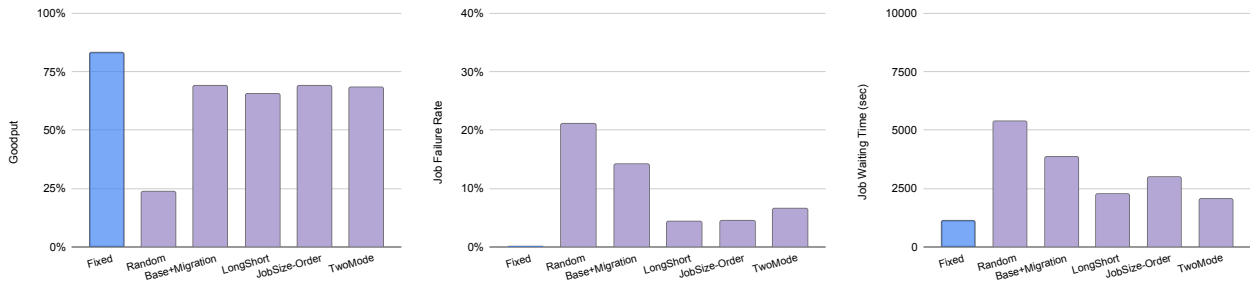
LongShort We consider a scheme where a scheduler exploits the information of job runtime classification to reduce job terminations of long jobs. LongShort uses the job runtime classification information to separate jobs into two groups: long and short jobs. To match these, it then separates the resources into two groups, where one group of resources is stable and another group represents the resources that may be periodically unavailable due to variations. LongShort makes placement decisions for jobs within its group of resources to reduce long-running job terminations from capacity decreases. If a scheduler knows a job J_i is long, as its runtime exceeds the time threshold t, which means the job is more likely to be terminated as it encounters more capacity volatility due to its runtime, the scheduler places the long-running job in the stable group of resources $M_k \in M_{stable}$ by first-fit policy where M_k is always available and satisfies the resource constraint. Otherwise, if the job is short that it is more likely to finish before the next capacity fluctuation, the scheduler places it in the other group $M_{unstable}$ by a reversed first-fit policy to improve resource utilization.

JobSize-Order This algorithm further improves resource utilization by grouping jobs with similar runtime and placing them on the same machines by exploiting exact job runtime information. Moreover, it exploits job runtime to further avoid terminations and minimize wasted work of both long and short jobs by placing longer jobs on safer machines. It orders the jobs waiting to be scheduled based on job runtime in descending order and reversely orders the resources based on the likelihood to be unavailable in the future. Based on such order, the scheduler starts scheduling jobs in a Longest Job First fashion and places the jobs on machines, in descending order of availability, by a first-fit placement policy. Therefore, the longer jobs in the queue, even in the short job group, will be placed earlier and on machines with higher availability to further reduce job terminations and wasted goodput. By prioritizing job runtime in scheduling, the algorithm improves resource utilization.

TwoMode By knowing the dynamic range of capacity variation and estimating if the capacity will likely increase or decrease, the scheduler makes the decision to switch between two modes: optimistic in scheduling to capture resource opportunities or pessimistic to reduce job terminations using job runtime information. If the scheduler discovers that the current capacity is relatively low in the space and will likely increase in the future, it will order jobs based on runtime in descending order, onto machines starting from which have the highest availability, similar to the JobSize-Order algorithm. Otherwise, if the scheduler knows the current capacity is higher than average and may encounter a capacity decrease, to avoid job terminations, it starts scheduling from the shortest jobs, as a Shortest-Job-First policy, on these servers starting from which have the lowest availability (ascending order on availability). Such optimistic scheduling aims to minimize potential computation waste while capturing resource capacity before it is gone as short jobs may successfully complete before the resources are made unavailable.

7.2.1 Other Techniques

Migration One technique that is widely adapted in previous work in the face of adaptive loads and data center failures to mitigate performance degradation is migration capability. As one may suggest using such a common mitigation strategy to solve the scheduling problem under variable capacity, we consider enabling migration capability in a data center to minimize potential job terminations. We use migration as an optimal strategy for coping with capacity loss (in Chapter 6) to compare with scheduling algorithms preparing for capacity variations. While it adapts the same job scheduling mechanism, upon resource capacity decreases, running jobs will be immediately migrated from resources to be unavailable to available ones to avoid terminations if there are enough free resources. Once running jobs are completely migrated or available resources are completely filled, it will terminate all running jobs that no longer fit into currently available servers, ordering based on least



(a) Goodput (b) Job failure rate (c) Job waiting time

Figure 7.3: System performance (Goodput, Job Failure Rate, Job Waiting Time) of scheduling schemes without foresight

wasted computation, and then continue with scheduling arriving jobs. To understand the effectiveness and limitation of this technique, it does not consider the migration overhead and therefore eliminates the potential impact from suboptimal job placements by the scheduler (fragmentation, runtime alignment, etc.). Thus, this migration technique can be viewed as a job-centric, ideal version of Least Wasted Work (an intelligent termination policy from Chapter 6 to cope with capacity loss).

7.3 Evaluation

We combine the foresight information with scheduling schemes to quantify the improvements a scheduler can achieve with different amounts of information and scheduling schemes that exploit it. Figure 7.3 shows the goodput, job failure rate, and job waiting time on the y-axis comparing four scheduling schemes, varying amounts of information with the largest variation range of 0.6.

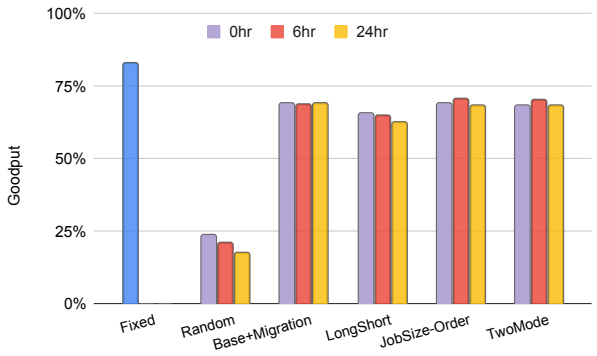
7.3.1 Information Space

In the case of 0hr foresight, all scheduling approaches with additional information improve performance significantly compared to *B*ase. Base with Migration capability demonstrates the best achievable performance improvement from a traditional scheduler, achieving 80%

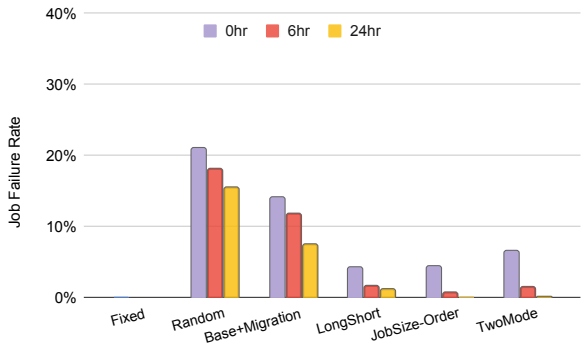
of fixed capacity goodput. However, Base with Migration does not plan for uncertainty and strives to fill as much resource capacity as possible, causing job terminations in addition to migrations due to capacity overestimation. It reflects up to 10% job failure rates with 4% migrations rate, with 3X job waiting time compared to fixed capacity, as a result. *L*ongShort algorithm, which uses the least amount of additional information – job runtime classification, shows 5X job failure reductions as it effectively reduces most terminations and rescheduling from long-running jobs, resulting in 3X goodput increase and 2.4X job waiting time reduction compared to the Base. *J*obSize-Order algorithm further improves LongShort by exploiting exact job runtime to improve reliability and maximize goodput based on exact runtime ordering, showing 5% incremental goodput increase on top of *L*ongShort results. However, as both *L*ongShort and *J*obSize-Order prioritize long jobs over short jobs in scheduling to maximize goodput, the job waiting time improvements are limited by both long-running job placements and short job terminations, showing $\sim$ 3X compared to fixed capacity. Finally, *T*woMode knows not only the exact job runtime but also the dynamic range of variation. As it switches between pessimistic and optimistic scheduling, it changes scheduling priority between long and short jobs and does not limit the placement of long jobs. Therefore, it exhibits the best results in goodput and job waiting time, with 2.6X reductions compared to Base, demonstrating the improvements achieved by obtaining additional information.

7.3.2 Capacity Foresight

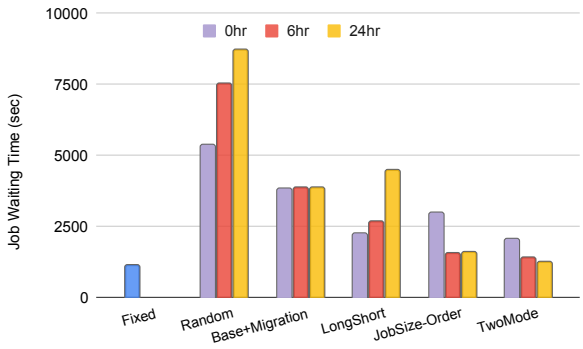
Now we combine these scheduling algorithms with the additional amount of capacity foresight, varying from 0hr to 24hrs, to understand the incremental benefits in Figure 7.4. The *B*ase scheduler, despite increasing capacity foresight, cannot effectively exploit to align jobs with capacity as it has no job runtime information. Conservatively scheduling all jobs on capacity lower bound within the foresight window causes large resource waste, producing no improvements. Similar to the Base, Random with Migration has no job runtime information



(a) Goodput



(b) Job failure rate



(c) Job waiting time

Figure 7.4: System performance (Goodput, Job Failure Rate, Job Waiting Time) of scheduling schemes, coupled with foresight of 0 (None), 6, 24 hours

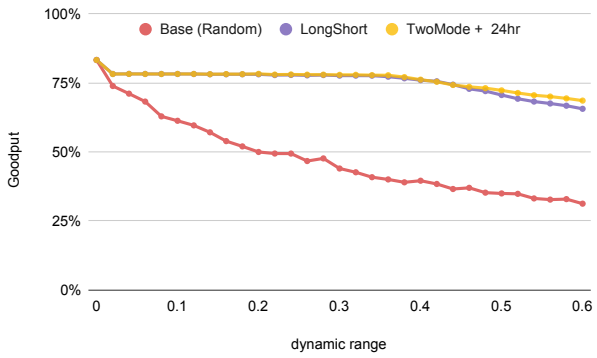
and thus cannot eliminate job terminations and migrations with foresight. As migration may incur additional overhead and constraints, these results demonstrate that simply enabling migration capability on a traditional scheduler or providing capacity foresight does not solve the problem of variable capacity. *L*ong-Short uses capacity foresight to plan for capacity changes by prioritizing placements on safe resources based on the future and further reduces job failure rate by 3X. However, only knowing the classification of job runtimes limits its ability to fully align jobs with variations because it can only assume each job to be the maximum possible runtime of its classification for scheduling. Therefore, it does not gain additional benefits in goodput and job waiting time due to rough estimation and conservative scheduling. On the other hand, *JobSize-Order* is able to fully exploit foresight information

to optimize placements according to exact job runtime. Such improvement is observed as the length of foresight increases, and with 24hr foresight, it achieves 90% of fixed capacity goodput performance, $\sim 0\%$ failures with 24hr foresight, 1.4X job waiting time compared to fixed capacity is because some long-running jobs have to wait longer for stable resources to be free. Finally, with full knowledge of job runtime, dynamic range, and capacity foresight, *T*woMode achieves 90% of fixed capacity goodput performance, 0.09% job failures, and a comparable job waiting time. It demonstrates the performance achievable with complete information in two dimensions of uncertainty despite a large range of variation. Overall, it is clear that even 24-hour foresight is less helpful in general than more intelligent scheduling.

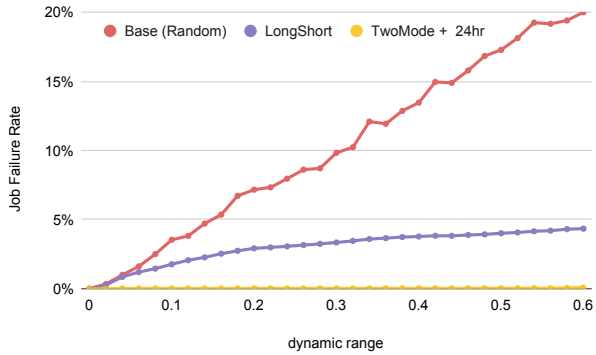
In summary, with additional information on job runtime and resource capacity, all scheduling algorithms effectively improve scheduling performance, restoring most of the degradation under a large dynamic range of variation. Foresight (oracle) of resource capacity can be useful but only if it can be coupled with additional job runtime information to allow the scheduler to exploit. If the scheduler is full-armed with information within the space, it achieves the best goodput with no performance degradation. On the other hand, runtime classification is critical information to improve scheduling. Exploiting only this information, *L*ongShort algorithm effectively obtains 95% of the full scheduling improvements.

7.3.3 Variation Range

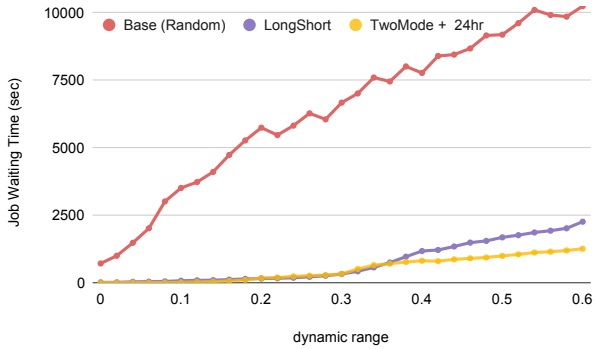
To understand whether scheduling improvements can enable more shifting ability in cloud data centers, now we expand the performance evaluation of scheduling schemes from the largest variation range to the whole spectrum from 0 to 0.6 in small steps of 0.02. Figure 7.5 displays the performance of *L*ongShort, which requires the least amount of additional information while achieving 95% of full improvements, and TwoMode with 24hr foresight, which exploits full knowledge and demonstrates the best performance across the board, comparing with Random as the baseline. The goodput drastically decreases as the range of



(a) Goodput



(b) Job failure rate



(c) Job waiting time

Figure 7.5: System performance (Goodput, Job Failure Rate, Job Waiting Time) of Scheduling Schemes Varying Variation Ranges

variation increases in the baseline. The degradation of goodput and acceptable variations are summarized in Table 7.3. Such degradation in performance suggests a very limited range of 6%, as the first 5% capacity variation contributes to 15% goodput loss and job failures double every 0.1 step increase of variation range. However, two scheduling schemes significantly improve performance across the full range, showing more than 2X improvements. Both scheduling schemes effectively maintain goodput, showing support for variation in capacity up to 50% with 7% goodput losses. Job failure rates of both schemes are controlled below 5%, and TwoMode + 24hr foresight shows nearly 0%. Job waiting time is maintained under 300 seconds for $<$ range 0.3 (43% variation) and under 100 seconds for $<$ range 0.2 (29% variation), demonstrating the ability of both schemes to maintain SLOs while

Goodput	Acceptable Variation		
Tolerable Degradation	Base (Random)	LongShort	TwoMode + 24hr foresight
5%	0%	1%	1%
7%	0%	49%	51%
10%	0%	60%	60%
12%	3%	63%	66%
15%	6%	69%	74%
17%	6%	74%	83%
20%	9%	83%	91%

Table 7.3: Tolerable degradation of goodput performance and the corresponding range of acceptable variation

drastically increasing shifting capabilities. Table 7.3 further shows that under the same tolerable degradation, *L*ongShort achieves $> 90\%$ acceptable variation of *T*woMode with 24hr foresight.

While *T*woMode + 24hour Foresight represents the ideal performance with the most amount of information, the results of *L*ongShort demonstrate the promising improvements achievable by a scheduling scheme with a small amount of critical information - job runtime classification. With scheduling schemes that exploit additional information, schedulers can effectively support cloud workloads under a 3 - 5X larger range of variation, demonstrating the achievable shifting ability of cloud data centers with scheduling efforts. Surprisingly, foresight (even 24-hr) is less helpful in general than more intelligent scheduling. As promising as these improvements are, one caveat for adopting variation-aware scheduling schemes in a variable capacity cloud data center is to consider job-level catch-up or latency constraints or priorities. We used a simple workload model assuming all jobs are delay-tolerant with the same priority. We discuss future directions tackling workload constraints and complexity in Section 8.2.2.

7.4 Summary

Variable resource capacity poses significant challenges to traditional schedulers as variation causes a large number of terminations, degrading goodput and increasing job waiting time. While techniques to cope with capacity loss are effective, cloud workloads represent a harder case. Uncertainties of workloads and variation together can limit the types of workloads and structures that can be supported. Therefore, to further improve performance, we propose the framework of uncertainty for cloud workloads. This includes two dimensions of uncertainty - capacity and job runtime, corresponding information space to reduce uncertainty, and new scheduling schemes which prepare for capacity changes by exploiting information to optimize job placements.

Experiments using Borg TNG cloud workloads show that these scheduling schemes can effectively improve performance, increasing goodput by up to 180%, decreasing job termination rate by 5 - 15X, and decreasing job waiting time by 1.4 - 4X, demonstrating the importance of information in preparing for capacity variation. Capacity foresight requires coupling with job runtime information and is less helpful in general than more intelligent scheduling. Among the information, job runtime classification is critical. Exploiting this information alone, *LongShort* scheduling algorithm achieves >90% full benefits and significantly improves tolerable variation range from <10% to 50% while maintaining performance. These results demonstrate the promising benefits of exploiting the information space under capacity variations but require validation on workloads with complex constraints and priorities.

CHAPTER 8

SUMMARY AND FUTURE DIRECTIONS

8.1 Summary

This thesis presents intelligent resource management for variable capacity data centers to eliminate variation penalties and enlarge variation benefits.

Variable resource capacity arises from the increasing pressure of reducing carbon emissions and power costs in data centers. For today's computing, variable resource capacity is problematic, causing severe loss in throughput and corresponding resource efficiency. To effectively achieve external benefits, data centers must support variable resource capacity while maintaining high compute efficiency. To evaluate and characterize the performance impact of capacity variation, we define the variable resource capacity problem, including three key dimensions of variation, dynamic range, structure, and change frequency, and characterizing scheduling impact. The empirical study of real cloud and HPC workloads shows that capacity variation can significantly degrade goodput by 15 - 60%, causing 15 - 40% job failures. These studies show that all variation dimensions have negative impacts, and each could independently reduce goodput by 10 - 40%, with even greater losses in combination. A drill down on production Borg TNG workloads reveals a dominant mode of VM usage. Coupling with other workload properties, such as inter-task dependencies, performance further decreases, producing goodput losses of 30 - 40% and up to 26% job failures. These harms strictly limit the cloud's shifting flexibility to $< 10\%$. These results all suggest future data center requires new resource management techniques that can tolerate greater dynamic range of capacity variation, while maintaining good performance.

To improve performance, we consider intelligent termination policies to cope with capacity loss by selectively terminating jobs to reduce failures and minimize wasted computation. Evaluation of a range of real workloads and variations show significant improvements on

HPC workloads, reducing job failures by 2 - 5x and increasing goodput by 44% on average.

While cloud workloads represent a hard case with low parallelism and complex workload properties, we take a broader view to proposing scheduling techniques preparing for capacity variation altogether. We consider two dimensions of uncertainty in resource capacity and workload which contribute to performance degradation, exploring the information space to reduce uncertainty. Scheduling algorithms that exploit the information to minimize job failures and increase resource efficiency achieve significant improvements. Experiments show that algorithms increase the goodput by 130%, decreasing job termination rate by 10X and job waiting time by 2X. Capacity foresight is less helpful in general than more intelligent scheduling and requires coupling with job runtime information to be useful. On the other hand, job runtime classification is critical information. Exploiting it alone can effectively enable up to 50% load flexibility while maintaining performance (7% goodput degradation). These results demonstrate great opportunities for new scheduling techniques under capacity variations but require validation on workloads with complex constraints and priorities.

8.2 Future work

We outline a few promising research directions for future exploration in variable capacity data centers based on our study and findings.

8.2.1 Different Types of Prediction/Foresight

Our intelligent resource management explores the information space of resource capacity and job runtime to reduce uncertainty exposed to the scheduler. The results demonstrate the effectiveness of exploiting information and the promising opportunities to use prediction or foresight to reduce scheduling uncertainty. While our study only exact capacity forecasts with a finite horizon, predictions can and often in many cases take many other forms, which might not be as perfect and straightforward.

What other forms of prediction/information are also useful? This requires a careful search for realistic examples of predictions. One interesting direction is to look into the time-series prediction models, which may offer bounds and probability of variations in the future. Another realistic scenario is to consider either conventional forecasts, such as day-ahead forecasts from power grids, or weather forecasts considering local renewables. Evaluating these different forms of information can demonstrate other interesting opportunities provided by realistic information.

How accurate do these predictions need to be and how does it impact the scheduling performance? We consider a horizon with perfect knowledge to demonstrate the scheduler capability with full-armed capacity information. In many cases, such capacity information can be predicted, but with a margin. For example, consider fuzzy predictions, such as some percentage of inaccuracy or capacity information with some margins. First, to characterize the impact of fuzziness, we can conduct sensitivity tests on the scheduling performance. Using a range of workloads and variation scenarios, we start by injecting inaccurate predictions into the foresight information provided to the scheduler, varying the amount of inaccuracy, false categorizations - false-positive/false-negative in capacity increase, and the degree of inaccuracy in the total information. Similarly, for capacity prediction with some margins, evaluations can include varying the amount of margin to the scheduler. Finally, after identifying which part of the fuzziness in the problem space impacts the scheduling's ability to reduce uncertainty and cope with variation, we propose new scheduling techniques that can either reason about when the predictions seem to be inaccurate or make decisions with a safety margin to prepare for incorrect capacity information.

8.2.2 Flexible or Optional Workload

While our studies show that workloads can be significantly impacted by capacity variation due to a high job failure rate, we use a simple workload model assuming all jobs can be

dropped and rerun. In reality, many cloud data centers consider classes or categories. For example, Borg considers several classes of jobs, and each is associated with different priorities in scheduling and running. Azure considers categories of jobs to be delay-tolerant and delay-sensitive. These groups show a more complex separation of workloads and can significantly impact our scheduling choices as some workloads are more important or sensitive to potential job terminations and need to be prevented. We first start by understanding and characterizing how much the delay-sensitive workloads a scheduler can support under capacity variation by varying the fraction of delay-sensitive jobs in the workload mixes. We increase the delay-sensitive fraction in small steps and see if and where performance drops. Furthermore, to understand if our proposed scheduling techniques can well support different mixes while maintaining performance, we evaluate and compare the new scheduling schemes with varying delay-sensitive job fractions. These performance results can provide insights into a realistic mix of delay-tolerant and delay-sensitive workloads and provide an understanding for cloud providers to incentivize or offer mechanisms for users to provide more flexibility.

8.2.3 Multi-datacenter Integration

Our intelligent resource management demonstrates the feasibility of variable capacity data centers and promising carbon and power cost benefits. The success of intelligent resource management encourages ongoing research on the continuous enhancement of resource management approaches to support more challenging scenarios and to further improve system performance. We believe that multi-data center coordination with flexible load shifting is a promising next step for scheduling research in variable capacity data centers.

The multi-datacenter integration is a natural enhancement to the current single data center study. *What are the relationships of resource variability across data centers? and does that affect the scheduler's operation spaces?* For multi-data center scenarios, the difference

is that one data center can shift workloads elsewhere besides termination or slowdown while facing resource deficit. Therefore, understanding the performance of load shifting enabled multi-datacenters under resource variability is important.

There are two important aspects in multi-data center scenarios: the relationship between data centers' variability patterns, and the amount of workload a data center can shift at any time. For example, the capacity variations between two data centers, depending on the locations, time, and sources of variation, can be independent, correlated, and counter-varying. Ideally, a counter-varying relationship represents the best possible results and may achieve close to fixed capacity results because workload can always be shifted to free resources in other data centers. On the other hand, correlated variability will degrade performance more because other data centers will likely experience resource decreases when the data center faces a resource deficit.

How should the shiftable load be constrained? and how do these constraints impact operation spaces or performance in various scenarios? For the impact of shiftable loads, defining a fraction of the total workloads as shiftable load and varying the fraction from small to large may show increasing performance. As more shiftable loads allow more flexibility in scheduling, it is important to understand how would load shifting decisions impact performance. For example, coupling with intelligent resource management may minimize the amount of load that needs to be shifted while achieving benefits. Another key question would be, how would intelligent resource management in multi-datacenter scenarios be different from the single data center?

This extension leads to a more flexible and coordinated view as many cloud providers support regional management and shifting. To perform a solid study with a deep understanding of the scheduling performance within the space, a multi-datacenter simulator with a global scheduler and coordination capability is necessary.

8.2.4 Grid Interaction

Our study of intelligent resource management which enables large dynamic variation while maintaining performance gives the data center more flexibility to dynamically shift loads and thus provides more opportunities for data centers to interact with the grids. Demand response programs as an emerging research area gain attention from datacenter to reduce electricity bills. These programs all serve as ways for data centers to interact and help stabilize the grids in exchange for monetary benefits. While data centers may choose to actively participate in markets as ancillary service markets and regulation services, it is very challenging due to their complexity and impact on performance. On the other hand, a large group of works explores data centers participating in voluntary programs to individually manage data center load through the use of pricing signals because of their flexibility. However, these programs either have little benefits to incentivize data centers or the load requirements are ad-hoc and arbitrary.

Data Center Flexibility *What flexibility a data center should provide to the grid?* The ideal strategy for a data center is to provide as much of its tolerable range without compromising performance as possible to maximize the benefits received from the grid. First, the data center needs to evaluate the trade-off between potential degradation of system performance if any, and the cost and carbon benefits from the larger variation range. It requires comprehensive characterization of performance across the variation ranges covering all possible scenarios and constraints and accurate predictions of future data center load. Another interesting direction is, considering if the grids can provide information on variations in return to help data centers prepare for load shifting, data centers can exploit such information to further optimize resource management strategies and explore what information might be useful.

Grid Coordination *How do the grids make use of and coordinate across entities with different flexibility?* Existing research studies have shown that grid coordination benefits individual data centers to reduce operational carbon emissions and reduce grid disruptions through a cooperative scheme. Unlike the existing demand response programs that occur only occasionally, we enable new and prominent data center flexibility exposed to the grid through intelligent resource management. With these variation ranges, an interesting direction for grid optimization is to consider how to utilize and coordinate such flexibility, as they may be different in time scale and absolute ranges and how should the grid price them or incentivize the data centers to provide an accurate and large dynamic range.

8.2.5 Headroom Analysis and Seasonal Optimization

Our experiments show that in a variable capacity data center, it is important and beneficial to have a nontrivial amount of headroom capacity. This capacity gives intelligent resource managers the ability not only to "catch up" jobs waiting in queue for better performance but also to effectively exploit the capacity opportunities during low-cost or low-carbon time in power grids. While having as much as possible headroom is important for performance, more complex problems associated with cost, quantity, and location are not as simple.

How much headroom capacity is useful and cost-wise beneficial? The ultimate goal of the data center with or without headroom capacity is to maximize goodput per TCO, where TCO can be roughly viewed as two separate components, CapEx (capital expenses) and OpEx (operational expenses). Therefore, to consider the cost associated with headroom capacity, one has to consider the CapEx, such as building cost, landfill, additional maintenance, etc. In addition, the data center also has to consider the electricity and the carbon associated with headroom. Our results demonstrate significant improvement and promising opportunities with enough headroom. Considering the associated cost and resulting benefits, data centers can weigh these two components and explore the incremental benefits of adding each unit of

headroom.

Where should the headroom capacity be placed and how to coordinate the use of headroom capacity? In addition, to justify the need and the amount of headroom to be placed on-site, multi-regional data centers can further optimize the benefits by a global view of the headroom capacity and its placement. Power grids have distinct profiles of renewable generations, supply and demand, and price structures, and the large difference over both long and short periods of time give rise to headroom optimization by its placement. For example, CAISO has a large fraction of solar power which means the daily carbon emission and power prices vary with the daylight patterns. Therefore, the long-term operational cost and maybe even capital cost can be widely-different across regions or power grids. In addition, the ability to productively utilize the headroom capacity and the associated incremental benefits can also differ across data centers, depending not only on the data center load but also on the global resource managers. In summary, a global operator needs to carefully take into consideration and orchestrate these dimensions to maximize the effectiveness of headroom capacity and data center revenue.

REFERENCES

[1] The amount of data center energy use - akcp monitoring. https://www.akcp.com/blog/the-real-amount-of-energy-a-data-center-use/#:~:text=In%202020%2C%20the%20data%20center,require%20104%20TWh%20in%202020. (Accessed on 12/18/2022).

[2] California iso - managing oversupply. http://www.caiso.com/informed/Pages/ManagingOversupply.aspx. (Accessed on 12/18/2022).

[3] Chart: Amazon, microsoft & google dominate cloud market | statista. https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/. (Accessed on 12/18/2022).

[4] Cloud computing market size, share amp; covid-19 impact analysis, 2022-2029.

[5] Dominion energy admits it can't meet data center power demands in virginia - dcd. https://www.datacenterdynamics.com/en/news/dominion-energy-admits-it-cant-meet-data-center-power-demands-in-virginia/. (Accessed on 12/07/2022).

[6] Ercot, market participants, and bitcoin miners get acquainted at second task force meeting. https://www.emergingenergyinsights.com/2022/04/ercot-market-participants-and-bitcoin-miners-become-acquainted-at-second-task-force-meeting/#:~:text=ERCOT%20anticipates%20that%20as%20much,in%20north%20and%20west%20Texas. (Accessed on 12/18/2022).

[7] Executive summary – electricity market report - july 2022 – analysis - iea. https://www.iea.org/reports/electricity-market-report-july-2022/executive-summary. (Accessed on 12/18/2022).

[8] Google-flex-jsspp.pdf. https://jsspp.org/papers18/Google-Flex-JSSPP.pdf. (Accessed on 12/18/2022).

[9] government-statement-on-the-role-of-data-centres-in-irelands-enterprise-strategy.pdf. https://enterprise.gov.ie/en/publications/publication-files/government-statement-on-the-role-of-data-centres-in-irelands-enterprise-strategy.pdf. (Accessed on 12/18/2022).

[10] The history of cloud computing - solved. https://solved.scality.com/solved/the-history-of-cloud-computing/#:~:text=The%20term%20%E2%80%9Ccloud%20computing%E2%80%9D%20itself,in%20academic%20work%20before%20that. (Accessed on 12/18/2022).

[11] Huge cloud market still growing at 34% per year; amazon, microsoft & google now account for 65% of the total | synergy research group. https://www.srgresearch.com/articles/huge-cloud-market-is-still-growing-at-34-per-year-amazon-microsoft-and-google-now-account-for-65-of-all-cloud-revenues. (Accessed on 12/07/2022).

[12] Hyperscale data center capacity doubles in under four years; the us still accounts for half | synergy research group. https://www.srgresearch.com/articles/as-hyperscale-data-center-capacity-doubles-in-under-four-years-the-us-still-accounts-for-half-of-the-total. (Accessed on 12/18/2022).

[13] Idc tracker finds spending on compute and storage cloud infrastructure increased strongly across most regions in the second quarter of 2022. https://www.idc.com/getdoc.jsp?containerId=prUS49732022#:~:text=Long%20term%2C%20IDC%20predicts%20spending,compute%20and%20storage%20infrastructure%20spend. (Accessed on 12/18/2022).

[14] Increased solar capacity drives u.s. renewable energy growth through 2023. https://www.eia.gov/pressroom/radio/transcript/steo_renewable_electricity_06072022.pdf. (Accessed on 12/18/2022).

[15] Large flexible load task force (lfltf). https://www.ercot.com/committees/tac/lfltf. (Accessed on 12/18/2022).

[16] Measuring greenhouse gas emissions in data centres: the environmental impact of cloud computing | insights & sustainability | climatiq. https://www.climatiq.io/blog/measure-greenhouse-gas-emissions-carbon-data-centres-cloud-computing. (Accessed on 12/18/2022).

[17] Microsoft and amazon reportedly halt plans to build data centers in ireland | engadget. https://www.engadget.com/microsoft-aws-data-centers-ireland-power-constraints-183054626.html. (Accessed on 12/07/2022).

[18] The paris agreement | united nations. https://www.un.org/en/climatechange/paris-agreement. (Accessed on 12/18/2022).

[19] Pipeline of over 300 new hyperscale data centers drives healthy growth forecasts | synergy research group. https://www.srgresearch.com/articles/pipeline-of-over-300-new-hyperscale-data-centers-drives-healthy-growth-forecasts. (Accessed on 12/18/2022).

[20] Pipeline of over 300 new hyperscale data centers drives healthy growth forecasts | synergy research group. https://www.srgresearch.com/articles/pipeline-of-over-300-new-hyperscale-data-centers-drives-healthy-growth-forecasts. (Accessed on 12/07/2022).

[21] Power crunch: Dublin grapples with cloud growth, utility constraints | data center frontier. https://www.datacenterfrontier.com/featured/article/11427205/power-crunch-dublin-grapples-with-cloud-growth-utility-constraints. (Accessed on 12/07/2022).

[22] Public cloud ecosystem quarterly revenues leap 26% to $126 billion in q1 | synergy research group. https://www.srgresearch.com/articles/public-cloud-ecos

ystem-quarterly-revenues-leap-26-to-126-billion-in-q1. (Accessed on 12/07/2022).

[23] U.s. energy information administration - eia - independent statistics and analysis. https://www.eia.gov/todayinenergy/detail.php?id=50818#:~:text=Solar.,by%2021.5%20GW%20in%202022. (Accessed on 12/18/2022).

[24] Alibaba production cluster data. https://github.com/alibaba/clusterdata, 2018.

[25] Lancium. https://www.lancium.com, 2018. A startup company, building low-cost, free-cooled datacenters to exploit excess renewable power.

[26] Amazon spot fleet. https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html, 2019.

[27] Amazon spot instance, 2019. https://aws.amazon.com/ec2/spot/.

[28] California independent system operator (caiso), 2019. One of several US regional wholesale power markets.

[29] Gartner forecasts on worldwide public cloud revenue. https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020, 2019.

[30] Google preemptible VMs. https://cloud.google.com/preemptible-vms/, 2019.

[31] Google aims to run on carbon-free energy by 2030. www.cnbc.com, September 2020.

[32] Jeff bezos pledges $10 billion to fight climate change, planet's 'biggest threat'. www.npr.org, February 2020.

[33] Microsoft makes 'carbon negative' pledge. www.bbc.com, January 2020.

[34] Baris Aksanli and Tajana Rosing. Providing regulation services and managing data center peak power budgets. In *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–4. IEEE, 2014.

[35] Ilari Alaperä, Samuli Honkapuro, and Janne Paananen. Data centers as a source of dynamic flexibility in smart girds. *Applied Energy*, 229:69–79, 2018.

[36] ALCF. Aurora. http://aurora.alcf.anl.gov/, 2015.

[37] William Allcock, Paul Rich, Yuping Fan, and Zhiling Lan. Experience and practice of batch scheduling on leadership supercomputers at Argonne. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 1–24. Springer, 2017.

[38] Mike Allen. And the title of the largest data center in the world and largest data center in us goes to... *Datacenters.com*, 2017. https://www.datacenters.com/news/and-the-title-of-the-largest-data-center-in-the-world-and-largest-data-center-in.

[39] Pradeep Ambati, Noman Bashir, David Irwin, and Prashant Shenoy. Good things come to those who wait: Optimizing job waiting in the cloud. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 229–242, 2021.

[40] Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. The datacenter as a computer: Designing warehouse-scale machines. *Synthesis Lectures on Computer Architecture*, 13(3):i–189, 2018.

[41] Natalie Bates, Girish Ghatikar, Ghaleb Abdulla, Gregory A Koenig, Sridutt Bhalachandra, Mehdi Sheikhalishahi, Tapasya Patki, Barry Rountree, and Stephen Poole. The electrical grid and supercomputing center: An investigative analysis of emerging opportunities and challenges. In *3rd D-A-CH Conference on Energy Informatics*. Springer, 2014.

[42] Anne Benoit, Laurent Lefèvre, Anne-Cécile Orgerie, and Issam Raïs. Reducing the energy consumption of large-scale computing systems through combined shutdown policies with multiple constraints. *The International Journal of High Performance Computing Applications*, 32(1):176–188, 2018.

[43] Upendra Bhoi, Purvi N Ramanuj, et al. Enhanced max-min task scheduling algorithm in cloud computing. *International Journal of Application or Innovation in Engineering and Management (IJAIEM)*, 2(4):259–264, 2013.

[44] Lori Bird, Jaquelin Cochran, and Xi Wang. Wind and Solar Energy Curtailment: Experience and Practices in the United States. Technical report, NREL, March 2014.

[45] Bloomberg. European union aims to be first carbon neutral major economy by 2050. *Fortune*, November 2018.

[46] Shekhar Borkar and Andrew A. Chien. The future of microprocessors. *Commun. ACM*, 54, May 2011.

[47] Katherine Bourzac. Supercomputing poised for a massive speed boost. *Nature*, November 2017. `https://www.nature.com/articles/d41586-017-07523-y`.

[48] Eric Boutin, Jaliya Ekanayake, Wei Lin, Bing Shi, Jingren Zhou, Zhengping Qian, Ming Wu, and Lidong Zhou. Apollo: scalable and coordinated scheduling for cloud-scale computing. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI'14)*, pages 285–300, 2014.

[49] John Boyd. Japan's fugaku supercomputer completes first-ever sweep of high-performance benchmarks. *IEEE Spectrum*, June 2020. `https://spectrum.ieee.org/tech-talk/computing/hardware/japans-fugaku-supercomputer-is-first-in-the-world-to-simultaneously-top-all-high-performance-benchmarks`.

[50] Brendan Burns, Brian Grant, David Oppenheimer, Eric Brewer, and John Wilkes. Borg, omega, and kubernetes. *Communications of the ACM*, 59(5):50–57, April 2016.

[51] Andrew Chien, Brad Calder, Stephen Elbert, and Karan Bhatia. Entropia: architecture and performance of an enterprise desktop grid system. *Journal of Parallel and Distributed Computing*, 63(5):597–610, 2003.

[52] Andrew A Chien. Characterizing opportunity power in the california independent system operator (caiso) in years 2015-2017. Technical report, Technical Report TR-2018-07. University of Chicago, 2018.

[53] Andrew A Chien, Richard Wolski, and Fan Yang. The zero-carbon cloud: High-value, dispatchable demand for renewable power generators. *The Electricity Journal*, pages 110–118, 2015.

[54] Andrew A Chien, Fan Yang, and Chaojie Zhang. Characterizing curtailed and uneconomic renewable power in the mid-continent independent system operator. *arXiv preprint arXiv:1702.05403*, 2016.

[55] Andrew A. Chien, Fan Yang, and Chaojie Zhang. Characterizing curtailed and uneconomic renewable power in the mid-continent independent system operator. *AIMS Energy*, 6(2):376–401, December 2018.

[56] Cobalt: Component based lightweight toolkit. `https://github.com/ido/cobalt`, 2019. IBM Commercial Job scheduler for Mira, JuQueen, and other Blue Gene systems.

[57] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 153–167. ACM, 2017.

[58] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP '17, pages 153–167, 2017.

[59] Carlo Curino, Subru Krishnan, Konstantinos Karanasos, Sriram Rao, Giovanni Matteo Fumarola, Botong Huang, Kishore Chaliparambil, Arun Suresh, Young Chen, Solom Heddaya, et al. Hydra: a federated resource manager for data-center scale analytics. In *NSDI*, pages 177–192, 2019.

[60] Mehiar Dabbagh, Bechir Hamdaoui, Mohsen Guizani, and Ammar Rayes. Toward energy-efficient cloud computing: Prediction, consolidation, and overcommitment. *IEEE network*, 29(2):56–61, 2015.

[61] Xiang Deng, Di Wu, Junfeng Shen, and Jian He. Eco-aware online power management and load scheduling for green cloud datacenters. *IEEE Systems Journal*, 10(1):78–87, 2014.

[62] Onur Derin and Alberto Ferrante. Scheduling energy consumption with local renewable micro-generation and dynamic electricity prices. In *First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods, and Tools*, volume 12, pages 1–6, 2010.

[63] Emily Dreyfuss. How google keeps is power hungry operations carbon neutral. *Wired*, December 2018.

[64] HS Dunn. The carbon footprint of ICTs. *Global Information Society Watch*, 2010.

[65] D. A. Ellsworth, A. D. Malony, B. Rountree, and M. Schulz. Dynamic power sharing for higher job throughput. In *SC '15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2015.

[66] Samir Elmougy, Shahenda Sarhan, and Manar Joundy. A novel hybrid of shortest job first and round robin with dynamic variable quantum time task scheduling technique. *Journal of Cloud computing*, 6(1):1–12, 2017.

[67] E3: Energy and Environmental Economics. Investigating a higher renewables portfolio standard for california. Technical report, Report for the California Public Utilities Commissions, 2014.

[68] Maja Etinski, Julita Corbalan, Jesus Labarta, and Mateo Valero. Parallel job scheduling for power constrained hpc systems. *Parallel Computing*, 38(12):615–630, 2012.

[69] Kobra Etminani and M Naghibzadeh. A min-min max-min selective algorihtm for grid task scheduling. In *2007 3rd IEEE/IFIP International Conference in Central Asia on Internet*, pages 1–7. IEEE, 2007.

[70] Electricity market data: Generations, prices, power, 2020. `https://www.smard.de`.

[71] Íñigo Goiri, William Katsak, Kien Le, Thu D Nguyen, and Ricardo Bianchini. Parasol and greenswitch: Managing datacenters powered by renewable energy. In *ACM SIGPLAN Notices*, volume 48, pages 51–64. ACM, 2013.

[72] Google. White paper: Moving toward 24x7 carbon-free energy at Google data centers: Progress and insights. Technical report, Google, October 2018.

[73] Albert Gore, Davis Guggenheim, Laurie David, Lawrence Bender, Scott Z Burns, Jeff Skoll, Leslie Chilcott, Bob Richman, Jay Cassidy, and Dan Swietlik. *An Inconvenient Truth*. Paramount, 2007.

[74] Greenpeace. Clicking clean: A guide to building a green internet, 2015 edition, 2015. `http://www.greenpeace.org/usa/wp-content/uploads/legacy/Global/usa/planet3/PDFs/2015ClickingClean.pdf`.

[75] Greenpeace. Clicking clean: Who's winning the race to build a Green Internet, 2017 edition, 2017. `http://www.greenpeace.org/`.

[76] GWEC. Global wind report: Annual market update. Technical report, Global Wind Energy Council, 2016. Documents curtailment around the world.

[77] Maryam Hamayun and Hira Khurshid. An optimized shortest job first scheduling algorithm for cpu scheduling. *J. Appl. Environ. Biol. Sci*, 5(12):42–46, 2015.

[78] Siqi Han. The wind is wasted in china. `https://www.wilsoncenter.org/`, 2015.

[79] Karen Hao. Training a single AI model can emit as much carbon as five cars in their lifetimes. *Technology Review*, June 2019.

[80] Md E Haque, IŽigo Goiri, Ricardo Bianchini, and Thu D Nguyen. Greenpar: Scheduling parallel high performance applications in green datacenters. In *Proceedings of the 29th ACM on International Conference on Supercomputing*, pages 217–227. ACM, 2015.

[81] Aaron Harlap, Andrew Chung, Alexey Tumanov, Gregory R Ganger, and Phillip B Gibbons. Tributary: spot-dancing for elastic services with latency slos. In *2018 USENIX Annual Technical Conference (ATC'18)*, pages 1–14, 2018.

[82] XiaoShan He, XianHe Sun, and Gregor Von Laszewski. Qos guided min-min heuristic for grid task scheduling. *Journal of computer science and technology*, 18(4):442–451, 2003.

[83] David Jackson, Quinn Snell, and Mark Clement. Core algorithms of the maui scheduler. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 87–102. Springer, 2001.

[84] Nicola Jones. How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722):163–167, 2018.

[85] Nicola Jones. How to stop data centres from gobbling up the world's electricity. *Nature*, September 2018.

[86] Arijit Khan, Xifeng Yan, Shu Tao, and Nikos Anerousis. Workload characterization and prediction in the cloud: A multiple time series approach. In *2012 IEEE Network Operations and Management Symposium*, pages 1287–1294. IEEE, 2012.

[87] Kibaek Kim, Fan Yang, Victor Zavala, and Andrew A. Chien. Data centers as dispatchable loads to harness stranded power. *IEEE Transactions on Sustainable Energy*, 2016.

[88] Kibaek Kim, Fan Yang, Victor M Zavala, and Andrew A Chien. Data centers as dispatchable loads to harness stranded power. *IEEE Transactions on Sustainable Energy*, 8(1):208–218, 2016.

[89] Osamu Kimura and Ken-ichiro Nishio. Saving Electricity in a Hurry: A Japanese Experience after the Great East Japan Earthquake in 2011. *ACEEE Summer Study on Energy Efficiency in Industry*, 2013.

[90] Kien Le, Ricardo Bianchini, Jingru Zhang, Yogesh Jaluria, Jiandong Meng, and Thu D. Nguyen. Reducing electricity cost through virtual machine placement in high performance computing clouds. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 22:1–22:12, New York, NY, USA, 2011. ACM.

[91] Yang Li, Charles R Lefurgy, Karthick Rajamani, Malcolm S Allen-Ware, Guillermo J Silva, Daniel D Heimsoth, Saugata Ghose, and Onur Mutlu. A scalable priority-aware approach to managing data center server power. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 701–714. IEEE, 2019.

[92] Liuzixuan Lin, Victor M Zavala, and Andrew A Chien. Evaluating coupling models for cloud datacenters and power grids. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*, pages 171–184, 2021.

[93] Michel J Litzkow, Miron Livny, and Matt W Mutka. Condor-a hunter of idle workstations. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1987.

[94] Longjun Liu, Chao Li, Hongbin Sun, Yang Hu, Juncheng Gu, Tao Li, Jingmin Xin, and Nanning Zheng. Heb: deploying and managing hybrid energy buffers for improving datacenter efficiency and economy. *ACM SIGARCH Computer Architecture News*, 43(3):463–475, 2016.

[95] Qixiao Liu and Zhibin Yu. The elasticity and plasticity in semi-containerized co-locating cloud workload: A view from alibaba trace. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 347–360, 2018.

[96] John Markoff. Microsoft Plumbs Ocean's Depths to Test Underwater Data Center. *The New York Times*, 2016.

[97] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. Recalibrating datacenter energy use estimates. *Science*, February 2020.

[98] C Megerian and J Panzar. Gov. Brown signs climate change bill to spur renewable energy, efficiency standards. Los Angeles Times, 9 2015.

[99] Mira-ALCF. MIRA: A 10-Petaflop, 4 MW IBM Supercomputing at Argonne. `https://www.alcf.anl.gov/mira`, 2019.

[100] Asit K Mishra, Joseph L Hellerstein, Walfredo Cirne, and Chita R Das. Towards characterizing cloud backend workloads: insights from google compute clusters. *ACM SIGMETRICS Performance Evaluation Review*, 37(4):34–41, 2010.

[101] The mid-continent independent system operator (miso), 2019. One of several US regional wholesale power markets.

[102] Ahuva W. Mu'alem and Dror G. Feitelson. Utilization, predictability, workloads, and user runtime estimates in scheduling the ibm sp2 with backfilling. *IEEE transactions on parallel and distributed systems*, 12(6):529–543, 2001.

[103] NCSA. Blue waters. `https://en.wikipedia.org/wiki/Blue_Waters`, June 2010.

[104] NERSC. Computational and Theory Research Facility. `https://www.nersc.gov/assets/Uploads/CRT-for-NUG.pdf`, February 2015.

[105] New York State Energy Planning Board. The energy to lead: 2015 New York state energy plan, 2015. `http://energyplan.ny.gov/Plans/2015.aspx`.

[106] NUDT. TIANHE-2. `http://www.top500.org/system/177999`, June 2015.

[107] OLCF. Summit. `https://www.olcf.ornl.gov/summit/`, 2015.

[108] Kay Ousterhout, Patrick Wendell, Matei Zaharia, and Ion Stoica. Sparrow: distributed, low latency scheduling. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 69–84, 2013.

[109] Tirthak Patel and Devesh Tiwari. Perq: Fair and efficient power management of power-constrained large-scale computing systems. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '19, page 171–182, New York, NY, USA, 2019. Association for Computing Machinery.

[110] Ivan Penn and Inyoung Kang. California today: A move to mandate 100% carbon-free electricity. *New York Times*, August 2018.

[111] Eduardo Pinheiro, Ricardo Bianchini, Enrique V Carrera, and Taliver Heath. Load balancing and unbalancing for power and performance in cluster-based systems. 2001.

[112] Chris N Potts and Mikhail Y Kovalyov. Scheduling with batching: A review. *European journal of operational research*, 120(2):228–249, 2000.

[113] DOE Wind Program. Wind vision: A new era for wind power in the united states. Technical report, DOE National Renewable Energy Laboratory, `http://energy.gov/eere/wind/wind-vision`, May 2015.

[114] Qsim: an event-driven scheduling simulator for cobalt. `https://trac.mcs.anl.gov/projects/cobalt/wiki/qsim`, 2013.

[115] Ishwari Singh Rajput and Deepa Gupta. A priority based round robin cpu scheduling algorithm for real time systems. *International Journal of Innovations in Engineering and Technology*, 1(3):1–11, 2012.

[116] Parthasarathy Ranganathan, Phil Leech, David Irwin, and Jeffrey Chase. Ensemble-level power management for dense blade servers. *ACM SIGARCH computer architecture news*, 34(2):66–77, 2006.

[117] Aaqib Rashid and Amit Chaturvedi. Cloud computing characteristics and services: a brief review. *International Journal of Computer Sciences and Engineering*, 7(2):421–426, 2019.

[118] John Rath. Blue Waters: Awesome Power, Awesome Efficiency. `http://www.datacenterknowledge.com/archives/2010/06/24/blue-waters-awesome-power-awesome-efficiency/`, June 2010. 15 MW power.

[119] Charles Reiss, John Wilkes, and Joseph L Hellerstein. Google cluster-usage traces: format+ schema. *Google Inc., White Paper*, pages 1–14, 2011.

[120] Greenpeace Reports. Clicking clean virginia. online, February 2019.

[121] Gonzalo P Rodrigo, Per-Olov Östberg, Erik Elmroth, Katie Antypas, Richard Gerber, and Lavanya Ramakrishnan. Towards understanding job heterogeneity in hpc: A nersc case study. In *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pages 521–526. IEEE, 2016.

[122] Varun Sakalkar, Vasileios Kontorinis, David Landhuis, Shaohong Li, Darren De Ronde, Thomas Blooming, Anand Ramesh, James Kennedy, Christopher Malone, Jimmy Clidaras, et al. Data center power oversubscription with a medium voltage power plane and priority-aware capping. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 497–511, 2020.

[123] R. Sakamoto, T. Cao, M. Kondo, K. Inoue, M. Ueda, T. Patki, D. Ellsworth, B. Rountree, and M. Schulz. Production hardware overprovisioning: Real-world performance optimization using an extensible power-aware resource management framework. In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 957–966, 2017.

[124] R. Sakamoto, T. Patki, T. Cao, M. Kondo, K. Inoue, M. Ueda, D. Ellsworth, B. Rountree, and M. Schulz. Analyzing resource trade-offs in hardware overprovisioned supercomputers. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 526–535, 2018.

[125] Malte Schwarzkopf, Andy Konwinski, Michael Abd-El-Malek, and John Wilkes. Omega: flexible, scalable schedulers for large compute clusters. 2013.

[126] Uwe Schwiegeishohn and Ramin Yahyapour. Improving first-come-first-serve job scheduling by gang scheduling. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 180–198. Springer, 1998.

[127] John Shalf, Sudip Dosanjh, and John Morrison. Exascale computing technology challenges. In *VECPAR 2010*, pages 1–25. 2011.

[128] Warren Smith, Valerie Taylor, and Ian Foster. Using run-time predictions to estimate queue wait times and improve scheduler performance. In *Workshop on Job scheduling strategies for Parallel Processing*, pages 202–219. Springer, 1999.

[129] Shanshan Song, Kai Hwang, and Yu-Kwong Kwok. Risk-resilient heuristics and genetic algorithms for security-assured grid job scheduling. *IEEE Transactions on Computers*, 55(6):703–719, 2006.

[130] Jason Sonnek, Abhishek Chandra, and Jon Weissman. Adaptive reputation-based scheduling on unreliable distributed infrastructures. *IEEE Transactions on Parallel and Distributed Systems*, 18(11):1551–1564, 2007.

[131] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.

[132] Qihang Sun, Shaolei Ren, Chuan Wu, and Zongpeng Li. An online incentive mechanism for emergency demand response in geo-distributed colocation data centers. In *Proceedings of the Seventh International Conference on Future Energy Systems*, page 3. ACM, 2016.

[133] Wei Tang, Zhiling Lan, Narayan Desai, and Daniel Buettner. Fault-aware, utility-based job scheduling on blue, gene/p systems. In *2009 IEEE International Conference on Cluster Computing and Workshops*, pages 1–10. IEEE, 2009.

[134] Daniel Terdiman. Is Google building a hulking floating data center in SF Bay? *CNET*, 2013. `http://www.cnet.com/news/is-google-building-a-hulking-floating-data-center-in-sf-bay/`.

[135] Huangshi Tian, Yunchuan Zheng, and Wei Wang. Characterizing and synthesizing task dependencies of data-parallel jobs in alibaba cloud. In *Proceedings of the ACM Symposium on Cloud Computing*, pages 139–151, 2019.

[136] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E. Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *Proceedings of the Fifteenth European Conference on Computer Systems*, 2020.

[137] Muhammad Tirmazi, Adam Barker, Nan Deng, Md Ehtesam Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. Borg: the next generation. In *EuroSys'20*, Heraklion, Crete, 2020.

[138] Dan Tsafrir, Yoav Etsion, and Dror G Feitelson. Backfilling using system-generated predictions rather than user runtime estimates. *IEEE Transactions on Parallel and Distributed Systems*, 18(6):789–803, 2007.

[139] Alexey Tumanov, Angela Jiang, Jun Woo Park, Michael A Kozuch, and Gregory R Ganger. Jamaisvu: Robust scheduling with auto-estimated job runtimes. Technical report, Technical Report CMU-PDL-16-104. Carnegie Mellon University, 2016.

[140] Alexey Tumanov, Timothy Zhu, Jun Woo Park, Michael A Kozuch, Mor Harchol-Balter, and Gregory R Ganger. Tetrisched: global rescheduling with adaptive plan-ahead in dynamic heterogeneous clusters. In *Proceedings of the Eleventh European Conference on Computer Systems*, page 35. ACM, 2016.

[141] Abhishek Verma, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Large-scale cluster management at google with borg. In *Proceedings of the Tenth European Conference on Computer Systems*, page 18. ACM, 2015.

[142] Sean Wallace, Xu Yang, Venkatram Vishwanath, William E Allcock, Susan Coghlan, Michael E Papka, and Zhiling Lan. A data driven scheduling approach for power management on hpc systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, page 56. IEEE Press, 2016.

[143] Adam Wierman, Zhenhua Liu, Iris Liu, and Hamed Mohsenian-Rad. Opportunities and challenges for data center demand response. In *International Green Computing Conference*, pages 1–10. IEEE, 2014.

[144] Ryan H. Wiser, Andrew D. Mills, Joachim Seel, Todd Levin, and Audun Botterud. Impacts of variable renewable energy on bulk power system assets, pricing, and costs. Technical Report LBNL-2001082, 11/2017 2017. A link to a webinar recorded on December 13, 2017 can be found at `https://youtu.be/EMrFAklQnPI`.

[145] Rich Wolski and John Brevik. Providing statistical reliability guarantees in the aws spot tier. In *Proceedings of the 24th High Performance Computing Symposium*, pages 1–9, 2016.

[146] Rich Wolski, John Brevik, Ryan Chard, and Kyle Chard. Probabilistic guarantees of execution duration for amazon spot instances. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2017.

[147] Chia-Ming Wu, Ruay-Shiung Chang, and Hsin-Yu Chan. A green energy-efficient scheduling algorithm using the dvfs technique for cloud datacenters. *Future Generation Computer Systems*, 37:141–147, 2014.

[148] Linlin Wu, Saurabh Kumar Garg, and Rajkumar Buyya. Sla-based admission control for a software-as-a-service provider in cloud computing environments. *Journal of Computer and System Sciences*, 78(5):1280–1299, 2012.

[149] Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. Dynamo: facebook's data center-wide

power management system. *ACM SIGARCH Computer Architecture News*, 44(3):469–480, 2016.

[150] Hong Xu and Baochun Li. Reducing electricity demand charge for data centers with partial execution. In *Proceedings of the 5th international conference on Future energy systems*, pages 51–61, 2014.

[151] Fan Yang and Andrew A Chien. Zccloud: Exploring wasted green power for high-performance computing. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 1051–1060. IEEE, 2016.

[152] Fan Yang and Andrew A. Chien. Large-scale and extreme-scale computing with stranded green power: Opportunities and costs. *IEEE Transactions on Parallel and Distributed Systems*, 29(5), December 2017.

[153] Fan Yang and Andrew A Chien. Large-scale and extreme-scale computing with stranded green power: Opportunities and costs. *IEEE Transactions on Parallel and Distributed Systems*, 29(5):1103–1116, 2017.

[154] Andy B Yoo, Morris A Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 44–60. Springer, 2003.

[155] Chaojie Zhang, Varun Gupta, and Andrew A Chien. Information models: Creating and preserving value in volatile cloud resources. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, pages 45–55. IEEE, 2019.

[156] Chaojie Zhang, Alok Gautam Kumbhare, Ioannis Manousakis, Deli Zhang, Pulkit A Misra, Rod Assis, Kyle Woolcock, Nithish Mahalingam, Brijesh Warrier, David Gauthier, et al. Flex: High-availability datacenters with zero reserved power. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 319–332. IEEE, 2021.

[157] Yijia Zhang, Ioannis Ch Paschalidis, and Ayse K Coskun. Data center participation in demand response programs with quality-of-service guarantees. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, pages 285–302, 2019.

[158] Salah Zrigui, Raphael Y de Camargo, Arnaud Legrand, and Denis Trystram. Improving the performance of batch schedulers using online job runtime classification. *Journal of Parallel and Distributed Computing*, 164:83–95, 2022.