Modeling and Control for Grid-interactive Efficient Data Centers

by

Yangyang Fu

B.S., PLA University of Science and Technology, 2012

M.S., Tongji University, 2016

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Civil, Environmental and Architectural Engineering

2020

This thesis entitled:
Modeling and Control for Grid-interactive Efficient Data Centers
written by Yangyang Fu
has been approved for the Department of Civil, Environmental and Architectural Engineering

Prof. Wangda Zuo

Prof. John Zhai

Prof. Kyri Baker

Dr. Jianming Lian

Dr. Sen Huang

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Fu, Yangyang (Ph.D., Architectural Engineering)

Modeling and Control for Grid-interactive Efficient Data Centers

Thesis directed by Prof. Wangda Zuo

The electrical grid is facing multiple challenges: increasing peak electricity demand, high penetration of variable renewable electricity generation, and transmission and distribution (T&D) infrastructure constraints. These challenges stress the electrical grid by making it more and more difficult to balance supply and demand for different time scales under the constraints of current infrastructure. Demand-side entities such as data centers with flexible electrical loads can also be utilized to serve the balancing, and their contributions can be as viable as supply-side counterparts.

Data centers have numerous opportunities to provide grid services considering their large capacities, flexible working environments and work loads, redundant design and operation, etc. Grid-interactive efficient data centers (GEDCs) have rich demand side resources, which have multi-scaled response time from milliseconds up to hours. The fast response resources such as servers can be used for providing frequency regulation (FR), as one of the ancillary services. Some electric markets in U.S., e.g., Pennsylvania New Jersey Maryland (PJM) Interconnection, allows demand sides to provide regulation services. However, due to the fact that either the programs or the markets are not mature especially from the data center's point of view, few of them are willing to participate. There are plenty of reasons, on top of which is the lack of a dynamic modeling and evaluation tool to assist the design and the control of GEDCs.

An equation-based object-oriented testbed for GEDCs are proposed and built in the Modelica environment. The proposed testbed considers different physical systems (thermal, electrical, and electromagnetic, etc.) with different time-scaled dynamics involved. End-to-end models include the computer servers, quality of service (QoS), uninterruptible power system, renewable energy resources such as solar panels, typical cooling system for data centers etc. The case studies show that this testbed can be used to perform various analysis, including detailed analysis of energy

efficiency and control performance in normal operation, as well as emergency operation.

The proposed testbed is then validated using measurement data from an actual data center located in Massachusetts. Based on the validated testbed, energy efficiency measures such as cooling system retrofitting, and control improvement are then proposed and evaluated. The resulted energy savings can be as much as 24.2% for the cooling system alone while data center room is still maintained in an acceptable thermal environment in terms of temperature and relative humidity.

To maximally harness the benefits from participating in energy market and regulation market, a real time optimization framework is proposed for GEDC without thermal energy storage systems (TESSs). This optimization framework first proposes a synergistic control strategy to enable FR service in GEDCs. The synergistic control strategy combines power management techniques at the server level with control of the chilled water supply temperature to track the regulation signal from the electrical market. A FR flexibility factor is also proposed to increase the IT capacity for FR. Then the performance of the control strategy is studied through numerical simulations using the proposed testbed. Simulation results show that with well-tuned control parameters, GEDCs can provide FR service in both regulation up and down. The performance of data centers in providing FR service is largely influenced by the regulation capacity bid, FR flexibility factor, workload condition, and cooling mode of the cooling system, and not significantly influenced by the time constant of chillers. The proposed synergistic control strategy can also provide an extra regulation capacity of 3% of the design power when chillers are activated, compared with a server-only control strategy. Finally, the proposed real-time multi-market optimization framework is investigated on the dynamic testbed with well-tuned parameters for FR service. Optimization results show that providing FR service over two days in January and July can save $23.6 and $115.8, respectively.

However, without storage systems, GEDCs are difficult to limit their demands that contribute to a large portion of the utility bill. To enable the demand limiting, GEDCs with TESSs are investigated. First, a synergistic control strategy for FR service by adjusting the chiller capacity, storage charging rate and IT server CPU frequency is developed. Then, a model predictive control framework is proposed to minimize the operational costs of GEDCs with TESSs from participating

in energy market and regulation market. Simulation results show that utilizing the TESS can not only reduce energy costs and demand charges but also harness FR revenues. The proposed multi-market optimization framework for the data center with TESS in two days can save the operational costs up to 8.8% ($1606.4) compared to the baseline data center with TESS. The savings are consisted of 0.2% ($38.7) of energy cost reduction, 6.5% ($1179.4) from demand reduction and 2.1% ($338.3) from regulation revenues. What's more, the proposed multi-market optimization framework for the data center with TESS can reduce the operation costs by $1793.2 in two days, saving 9.7% compared with a baseline data center without TESS.

Dedication

To all the people who have helped me become

To all the regrets and yearnings that keep me motivated

To this demanding but spiritually rewarding and interesting journey

Acknowledgements

The answer to the question "what I have lived for" is unfolding gradually over time to me in the past years, although finding the true answer is a lifetime journey. The three passions that governed Bertrand Russell are appreciated here, because they lead to my thinking about how to self-actualize, how to love and how to empathize.

The sought of knowledge brought and will keep bringing self-actualization to me. I would like to express my deepest thanks to my advisor Dr. Wangda Zuo for his constant guidance and support throughout this research work. Dr. Zuo has become not only an advisor but a mentor and a friend. I have learned from him a lot of things that I have been mentally blocking, and gradually walked out of my comfort zones. What's more, I would like to thank the rest of my thesis committee: Dr. John Zhai, Dr. Kyri Baker, Dr. Jianming Lian and Dr. Sen Huang, for their insightful comments, encouragement and understanding. Additionally, I owe my gratitude to all the co-authors that helped me publish the outcome of this research. Without them, I cannot imagine the delays of my graduation.

The sought of love relieves loneliness. I am grateful for being loved by my vast extended family. Their unconditional supports helped me get through those dark and cold days. I thank my fellow labmates for the simulating discussions, for the delightful working environment and for all the fun we have had. My sincere thanks also go to my circle of strong men and women. Dr. Mellisa Hoyer is the strongest women I have ever seen, which has motivated me for a long time. Thank Cory Mosiman for his friendship and support. Thank Daniel Sailer for blending me into American-styled life. Camping, baking, being loved by a dog, I cannot know how much more he still has.

Thank Justus and Simone for the remote calls from Europe. Thank Brenton Kreiger for leading me to the rocks. Thank Zhaozhong Chen for discovering my biceps. Thank Peilin Yang, Shiyao Liang and Xiaocheng Guo for witnessing me being a better and better rider on the snowboard. Thank Zac and Jillian for being such a nice couple to me. Seeing you two from falling into love to getting married is the happiest thing I ever had in the past three years. Thank Mohammad Dabbagh, Ammar Dehwah, and Archan Shah for always cheering me up. Thank Jiachen Mao from Massachusetts Institute of Technology for sharing great ideas. I also appreciated all the folks that have taught me how to behave.

I want to give my special thanks to all the people that helped me develop my sympathy and empathy in the past years. Although this is a lifetime practice, I will keep going and being devoted to.

Lastly, this research is based upon work supported by the U.S. Department of Energys Office of Energy Efficiency and Renewable Energy (EERE) under the Award Number DE-0007688.

Disclaimer

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1 Grid Balancing

1.1.1 U.S. Power Grid

A power grid, electrical grid or electric grid is an interconnected network for delivering electricity from producers to consumers as shown in Figure 1.1. There are four major portions of the electricity supply chain in a regional power grid: Generation, Transmission (high voltage power lines that move electricity over many miles), Distribution (low voltage power lines for local delivery of electricity) and Consumption (local consumers or loads).

- Generation. Generations have two types: centralized and decentralized. Centralized generation refers to large-scale power plant generation far from consumers, which includes coal, nuclear, natural gas, hydro, wind farms and large solar arrays. The electricity that power plants generate is delivered to customers over transmission and distribution power lines. Decentralized generation refers to local generation such as rooftop solar panels.

- Transmission and Distribution. Transmission and distribution refer to the different stages of carrying electricity over wires and poles from generators to consumers. Transmission lines move a large amount of power at a high voltage level, such as 100 kV and above. For most consumers such as homes, the power moving through transmission lines must be reduced to lower voltage levels by electricity distributors, and then delivered to consumers.

The distribution lines are usually at a voltage around 7-13 kV. Transformers are used to increase (step-up) or reduce (step-down) voltages in transmission and distribution grids.

- Consumption. The electricity is finally delivered and retailed to different consumers. Consumers are typically referred to industrial, commercial and residential consumers. Each of these consumers has different needs of energy services.

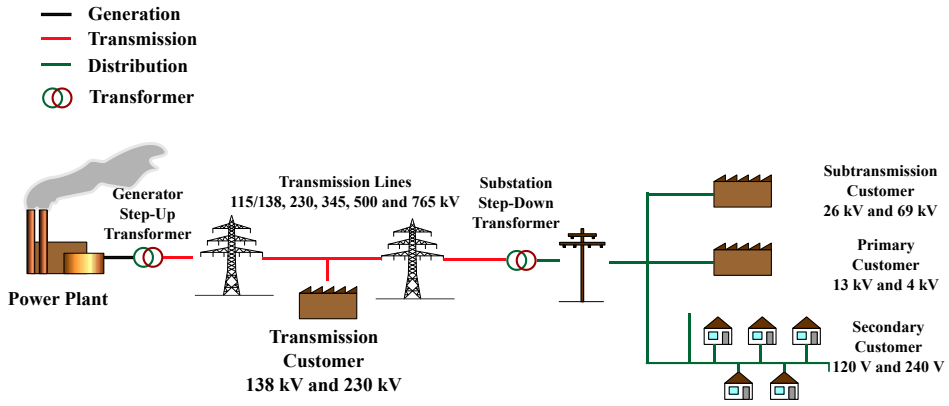


Figure 1.1: Electricity supply chain in a power grid

In the mid-1990s, the majority of electricity consumers in U.S. were served by an investor-owned, vertically-integrated monopoly utility that provided generation, transmission, distribution and billing simultaneously. In late the 1990s, regional electric grids in U.S. started a regulatory reform aimed at establishing a less-regulated and more market-based structure. The restructuring introduced competition in various utility functions, and separated the ownership of generation and distribution, which has led to systems where generation companies offer quantities of energy at various prices and distribution companies purchase quantities of energy to satisfy their demand. This exchange between generation companies and distribution companies is performed over wholesale markets, while purchases of electricity by consumers from distribution companies are transacted over retail markets. Because the reform was focused on generation, transmission and retailing, the local distribution lines are still considered a natural monopoly that would be subject to either regulation or municipal ownership. Meanwhile, for the restructured parts, wholesale markets greatly introduce competition on the supply side, which can eventually leads to the least-cost generation.

For the transmission service, nonpreferential third-parties, known in the U.S. as regional transmission organizations (RTOs) or independent system operators (ISOs) are created to be organized as non-profit companies and operate essentially as regulated entities; see Figure 1.2. In the U.S. these transmission companies do not own any transmission assets, but they can control access to those assets by virtue of approving the production schedules of the power plants within their regions, as well as operating real-time balancing markets. In each case, the decisions made by the RTO/ISO with regards to generation operations must satisfy the constraints of the transmission network and other reliability considerations.

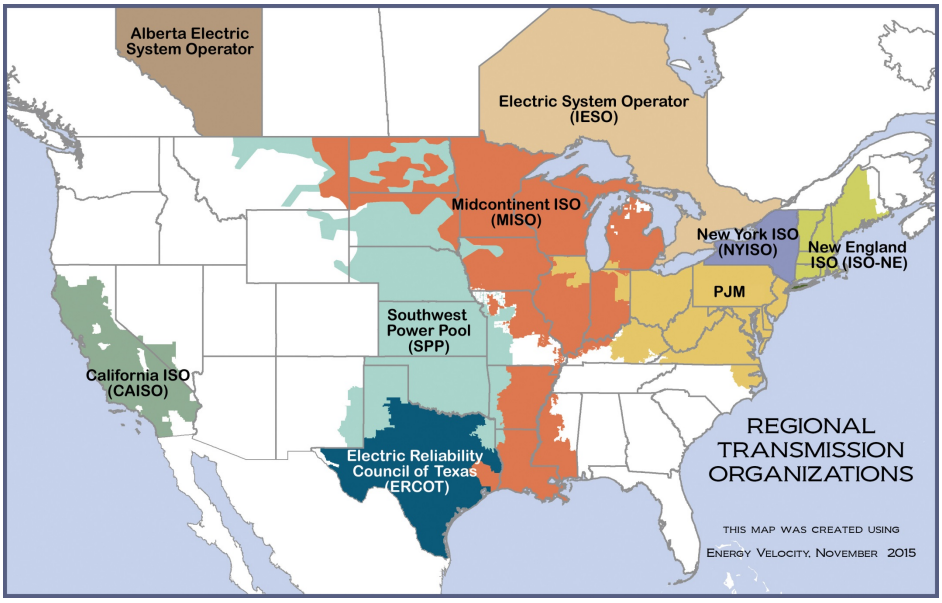


Figure 1.2: RTOs/ISOs in North America as of 2015

1.1.2 Balance Management

Balance management is a power system operation service to ensure the stabilization of the electrical grids through the continuous, real-time balancing of power demand and supply. At each point in time, the total generation must be equal to the total consumption to stabilize system

frequency. The imbalance in the electrical grid will deteriorate the power stability and quality, which may lead to the disconnection of system components, and ultimately, power blackouts.

Electricity supply and demand is balanced through self-supply, bilateral contracts or through ISO/RTO electricity auctions. Electric utilities not participating in an ISO or RTO must effectively balance their supply with customer demand by either relying on their own generation facilities (that is, self-supply), by purchasing electricity under bilateral contracts, or by shedding load.

Supply and demand in ISO/RTO regions is balanced through reverse auctions, in which the ISO/RTO matches bids for supply with bids for demand. Because RTOs/ISOs themselves do not own any physical assets, they must sign business contracts with generation companies to provide necessary services. The market mechanisms run by the RTO/ISO are used to procure generation supplies needed to maintain reliability. Once generation supplies are procured by the RTO/ISO, it can dispatch generation as needed to meet demand.

RTOs/ISOs run three types of electrical markets (capacity markets, energy markets and ancillary services markets) to manage the power grid over time scales ranging from cycles (one cycle = 1/60th of one second) to several years in advance of real-time dispatch, as shown in Figure 1.3.

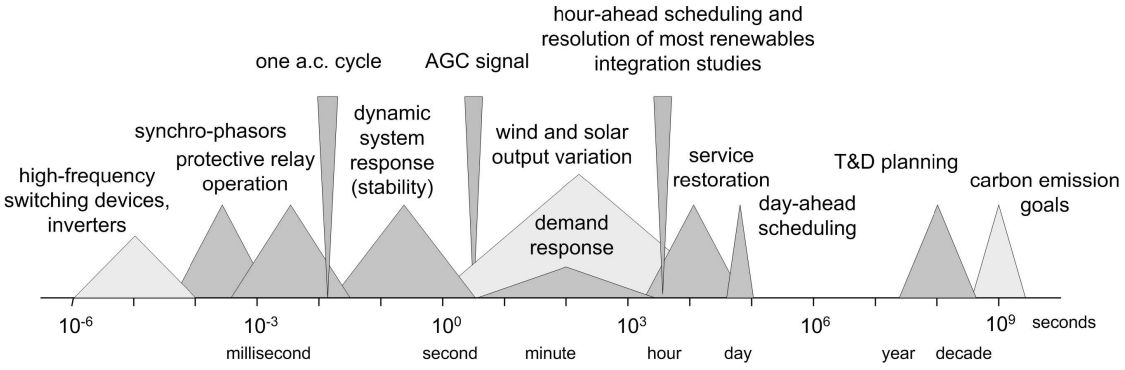


Figure 1.3: Schematic of time scales involved in power system planning and operations [126]

Capacity market is generally a forward market to ensure generation capacity online and ready to produce electricity at least one year ahead of time, by providing financial incentives for suppliers. The capacity market participants offer power supply resources into the market that provide supply or reduce demand. These resources include new and existing generators, upgrades for existing

generators, demand response (consumers reducing electricity use in exchange for payment), energy efficiency and transmission upgrades. When a participant offers these resources into the market, that participant is committed to increase supply or reduce demand by the amount they offered.

Energy market is used to ensure that enough generation capacity is online and able to produce energy on a day-ahead (24-hour ahead) to one-hour-ahead basis. Therefore, it usually consists of two markets: day-ahead market and real-time market. The "day-ahead market" is a forward market used to determine which generators are scheduled to operate during each hour of the following day and at what level of output, based on a projection of electricity demand in the following day. The "real-time market" is a spot market meaning that the product is procured for immediate delivery. It is used to adjust which generators are scheduled to run on an hour-ahead basis based on actual grid operating conditions.

Ancillary services market, to avoid potential instabilities caused by deviations between the planned and the actual load on a sub-hour basis, allows the RTO/ISO to maintain a portfolio of backup generation in case of unexpectedly high demand or if contingencies, such as generator outages, arise on the system. Ancillary services have many types, corresponding to the speed with which the backup generation needs to be dispatched. Take two important ancillary services (e.g., reserves and regulation) as an example. "Reserves" represent capacity that can be synchronized with the grid and brought to some operating level within 60, 30 or 15 minutes. "Regulation" represent capacity that can adjust its level of output within a few seconds in response to the fluctuations in the system frequency. The fluctuations are typically caused by slow response, inaccurate automatic generation control (AGC) and forced outages of power plants etc.

1.1.3 Challenges in a Modern Power Grid

One of the facts about power grids is that electric power must be generated and delivered in nearly the exact amount being consumed at any given moment in time when the grid-scale energy storage system is unavailable. Typically, grid operators need to dispatch vast amount of power from various sources to meet the demand variations from load valleys to load peaks every day. This

balancing process faces multiple challenges in recent years.

First, on the generation side, renewable energy resources (RES), such as wind and solar, are increasingly penetrated into the grid. The uncertainty of the generation by the RES is difficult to predict one hour ahead or day ahead, which pose challenges to economic dispatch. Also, the variations of generation by RES introduce more volatility to grid operations at higher levels of penetration.

Second, on the load side, the energy use are growing significantly over the past decades. Beside, the peak demand is also skyrocketing.

Third, the elements of the grid infrastructure (e.g., generators, transformers, or wires) are facing their physical constraints, which could be capital-intensive to upgrade. The infrastructure must be be sized to handle the peak demand at its location, because catastrophic damage to equipment can occur if the flow of power exceeds the rated capacity the element for too long. As the peak demand grows, the existing infrastructure are exposed to the danger of failure.

1.1.4 Buildings to Grid Services

The electrical grid is facing multiple challenges: increasing peak electricity demand, high penetration of variable renewable electricity generation, and transmission and distribution (T&D) infrastructure constraints. These challenges stress the electrical grid by making it more and more difficult to balance supply and demand for different time scales under the constraints of current infrastructure. Balance management services are largely provided by supply-side entities: integrated utilities, grid operators, and generators. However, demand-side entities such as buildings with flexible electrical loads can also be utilized to serve the balancing, and their contributions can be as viable as supply-side counterparts [97]. The motivation behind is that buildings can provide demand flexibility in response to price changes or direct signals.

Demand flexibility is the capability provided by on-site district energy resources (DERs) in buildings to reduce, shed, shift, modulate or generate electricity [97]. The flexibility has the potential to provide grid services that benefit the grid across the three major dispatchable categories:

capacity, energy and ancillary services. Building owners or occupants that deliver the grid services through dispatchable or non-dispatchable programs can be compensated through lower utility bills, or pre-negotiated payments while the grid also sees an immediate benefit. What's more, some of these grid services can also benefit the grid by avoiding or deferring T&D upgrades and associated capital cost which can increase utility customer rates.

Buildings can provide demand flexibility with five modes: efficiency, shedding, shifting, modulating, and generation. Some definitions from Ref. [97] are listed here and shown in Figure 1.4.

- Efficiency: the ongoing reduction in energy use while providing the same or improved level of building function.
- Shedding: the ability to reduce electricity use for a short time period and typically on short notice. Shedding is typically dispatched during peak demand periods and during emergencies.
- Shifting: the ability to change the timing of electricity use to minimize demand during peak periods or to take advantage of the cheapest electricity prices. A shift may lead to using more electricity during the cheapest time period and using thermal or battery storage at another time period when electricity prices increase.
- Modulating: the ability to balance power supply/demand or reactive power draw/supply autonomously (within seconds to sub-seconds) in response to a signal from the grid operator during the dispatch period.
- Generation: the ability to generate electricity for on site consumption and even dispatch electricity to the grid in response to a signal from the grid. Batteries can increase the ability of distributed generation to be dispatched as needed.

There are multiple grid services that buildings can provide through the demand flexibility. The grid services include generation, ancillary, and deliver services. Ref. [97] summarized potential grid services that can be provided by demand flexibility in buildings.

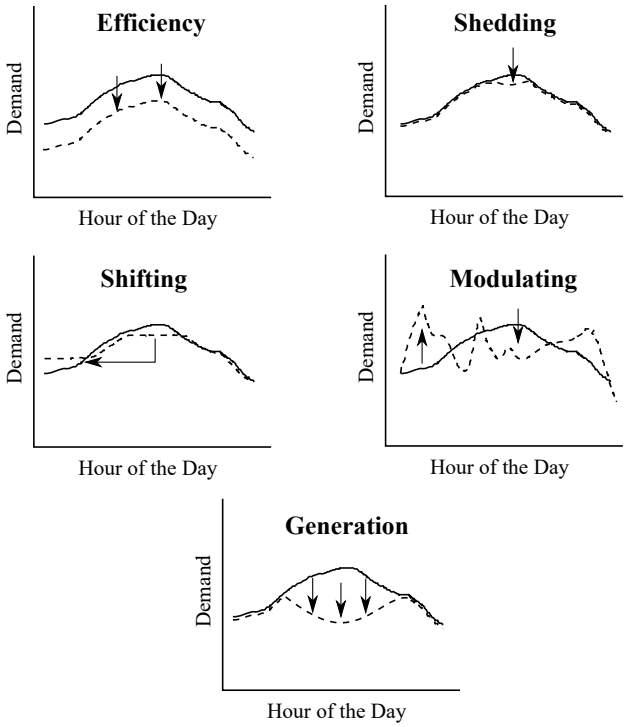


Figure 1.4: Building flexibility load curves (The solid line is the base load, and the dashed line is the resulting load)

Table 1.1: Grid services provided by demand flexibility in buildings

Grid Services		Strategies on Buildings' Side	Potential Demand Flexibility
Generation Services	Generation: Capacity	Buildings can defer generation capacity investment by reducing demand	Efficiency, Shedding, Generation
	Generation: Energy	Buildings can reduce the needs of generations by reducing demand	Efficiency, Shedding, Generation
Ancillary Services	Contingency Reservers	Buildings can provide a load reduction in the event of a system contingency	Shedding, Shifting
	Frequency Regulation	Buildings can modulate fast-responsing loads to track a power control signal	Modulating
	Ramping	Buildings can modulate fast-responsing loads to offset rapid changes in generation output	Modulating, Generation
Delivery Services	Non-Wires Solutions	Buildings can provide non-wires solutions in a variety of ways including: energy efficiency, demand response, distributed generation, voltage support, and energy storage	Efficiency, Shedding, Shifting, Generation
	Voltage Support	Multiple building technologies can provide limited voltage support including rooftop solar inverters and battery inverters	Modulating

1.1.5 Frequency Regulation

Maintaining the power system frequency at its nominal value (e.g. 60 Hz) is critical for power system stability and reliability. The real-time system frequency is influenced by the difference between the supply and the demand. If the supply is greater than the demand, the frequency will rise above its nominal value, and vice versa. Over-frequency and under-frequency events can damage electric power equipment, degrade the power quality, or even collapse the entire power grid. For example, when one generator is suddenly offline for some unexpected reasons, the system frequency will drop below its nominal value. It would be difficult for other offline generators to go back online immediately to compensate for the loss of generation. But it would be feasible to trigger online generators' protective actions. The under-frequency protection can trip the generator to operate at a slower speed. In this case, the imbalance of supply and demand is exacerbated and could lead to a system collapse.

Frequency regulation (FR) is the ancillary service that provides continuous, rapid, and automatic corrections for changes in electricity generation or use on a second-to-second basis in order to maintain the system frequency. Typically, FR resources are generators. FR uses certain amount of generators (e.g., about 1% of total generation) equipped with AGC and the information gathered through the supervisory control and data acquisition system to continuously track the demand variations. The frequency must be strictly maintained within a very narrow range in order to comply with the control performance standards and the balancing authority area control error limit reliability criteria. A real-time value, area control error (ACE), is used to quantify the energy balance at each moment. ACE is measured in MW and is calculated by the mismatch between the actual and scheduled electrical generation. The balancing authorities must monitor and carefully control ACE in a limited range in order to fulfill their obligations to the North American Electric Reliability Corporation (NERC) for system reliability. An automatic generated control signal based on ACE is then sent to the FR resources.

Besides generators, fast-ramping demand side resources (DSRs) in buildings can also provide

FR service to the grid by harnessing the demand flexibility provided by the modulating loads. Typical modulating loads on building side include energy storage systems such as flywheels, batteries and compressed-air energy system, electric boilers and heaters, and independent systems with variable frequency drivers (VFDs).

1.2 Grid-interactive Efficient Data Centers

Data centers are critical, energy-intensive infrastructures that support the fast growth of the information technology (IT) industry and the transformation of the economy at large. In 2010 data centers consumed about 1.1% to 1.5% of the total worldwide electricity and the number was about 1.7% to 2.2% for the U.S. [74]. The electricity usage will keep growing due to the surge in digital services associated with widespread Internet access. The electricity in data centers is mainly consumed by two parts: IT equipment (e.g., servers, storage, network, etc.) and infrastructure facilities (e.g., cooling system). The latter usually accounts for about half of the total energy consumption in a typical data center [125].

1.2.1 Typical Configurations in Data Centers

1.2.1.1 Cooling Systems

Many different cooling systems have been designed and operated for data centers. The data center room can be cooled by air, single- or two-phase liquid at room, rack or even chip level. However, the majority of existing data centers are cooled by air [39]. Thus, our current modeling efforts focus on air-cooled systems. The air-cooled systems supply to the data center room cold air, which is then drawn by the rack or servers [7]. The air can be cooled by chilled water systems or DX systems, which are introduced next.

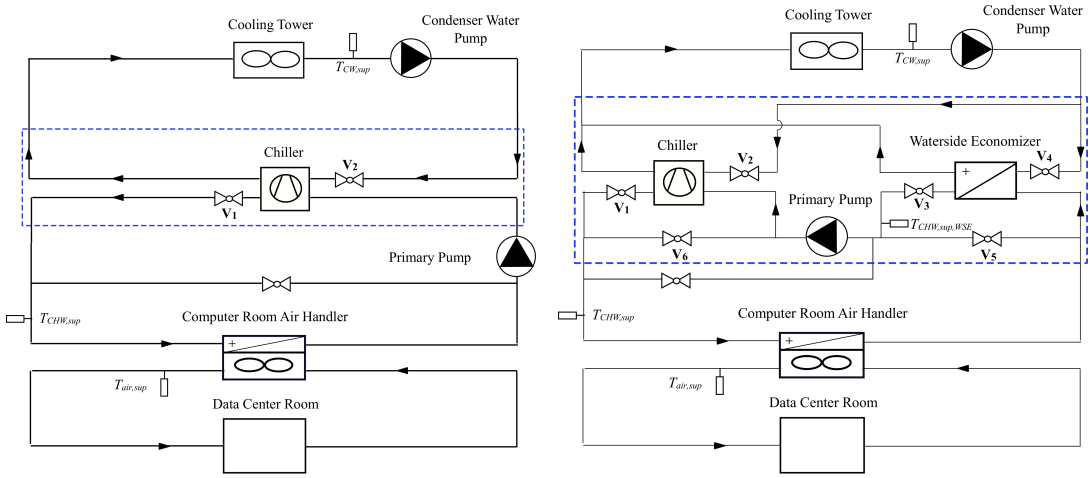
Chilled water systems are usually used for large data centers. A typical chilled water system includes chillers, Computer Room Air Handlers (CRAHs), pumps, and cooling towers, as shown in Figure 1.5(a). The heat generated in the data center room is first transferred to the chilled water

through the CRAHs, and the chillers then transfer the heat from the chilled water loop into the condenser water loop through a refrigeration system. The cooling towers at last reject the heat in the condenser water loop to the outdoor environment.

Economizers can also be installed to provide auxiliary cooling when outdoor conditions allow. The economizer can be installed on either the air side or the water side of chilled water system. If on the water side, it is typically called waterside economizer (WSE). WSE can be configured with a chiller in different ways [48]. For example, the WSE can be integrated with the chiller, meaning that the economizer can meet all or some of the load while the chiller meets the rest of the load, or non-integrated, meaning the economizer can only operate when it can meet the entire load.

A common configuration of a chiller plant with integrated WSE is shown in Figure 1.5(b). The WSE is located upstream of the chiller on the load side of the common leg. This configuration can guarantee that the WSE can meet the warmest return chilled water and maximize the number of hours when WSE can operate. The chiller plant with integrated WSEs can operate in three modes: Free Cooling (FC) mode when only the WSE is enabled for cooling, Partial Mechanical Cooling (PMC) mode when the chiller and WSE are both triggered, and Full Mechanical Cooling (FMC) mode when only the chiller is activated. When a particular cooling mode should be activated is determined by a cooling mode controller, as described in Section 3.2.3.1, and how to achieve the cooling mode is determined by the manipulation of the associated isolation valves V_1 to V_4, the chiller bypass valve (V_6) and the WSE bypass valve (V_5), as shown in Section 3.2.3.3.

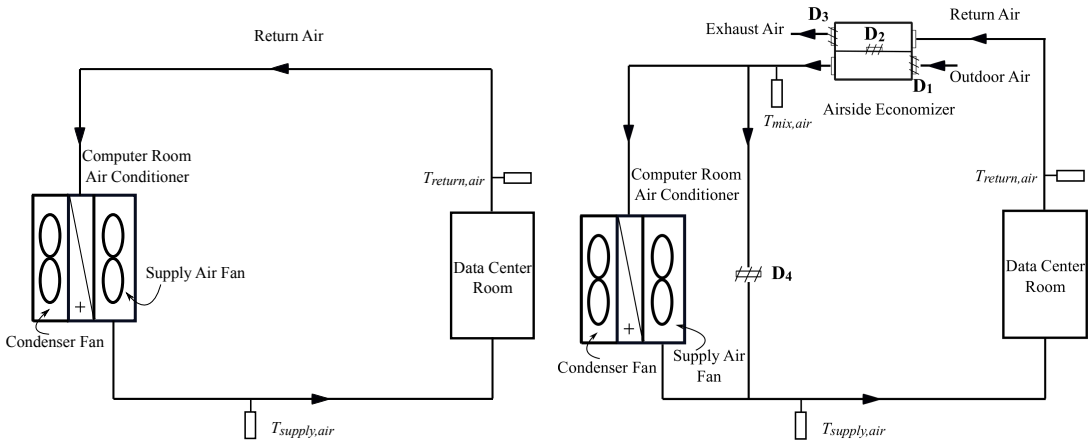
DX system is widely used in small data centers as a primary cooling system or as a backup system for the chilled water system. The major cooling source in a DX system are the Computer Room Air-Conditioner (CRAC) units. CRAC units typically have their own refrigeration system. They absorb heat from the data center room through evaporators, and reject heat to the outdoor environment (air-cooled CRAC) or a condenser water loop (water-cooled CRAC) through condensers. Based on the type of the condenser in the CRAC, the DX system is categorized into two classes: air-cooled or water-cooled. Figure 1.6(a) shows a schematic drawing for an air-cooled DX system, where the air-cooled CRAC is installed to cool the return air and send it back to the



(a) Without WSE (b) With an integrated WSE on the load side

Figure 1.5: Primary-only chilled water system

data center room by supply air fans. The heat extracted by the CRAC is then rejected to the outdoor environment through condenser fans. The DX system is usually installed together with air side economizers (ASEs), for example, as shown in Figure 1.6(b). The ASEs enable the system to efficiently use the cold outdoor air and reduce the operating time of CRACs.



(a) Without ASE (b) With an ASE

Figure 1.6: Air-cooled DX system

In Figure 1.6(b), with an ASE, the air-cooled DX system can also operate in the three modes as mentioned above. The only difference is that the mechanical cooling is provided by the CRACs in the air-cooled DX system instead of the chillers as in the chilled water system. Similarly, when to activate a particular cooling mode is determined by the cooling mode controller, and how to achieve the cooling mode is determined by the manipulation of the dampers such as D_1 to D_4 in Figure 1.6(b)[48].

1.2.1.2 Electrical Systems

Data centers are required to operate continuously. Electrical distribution systems in data centers are designed to power the IT equipment in a safe and reliable manner. One typical design is to power the data center with multiple sources. For example, data centers normally draw three-phase AC power from the power grid, and use diesel generators as backups during power grid failures. There is usually a time gap between the power grid failure and the activation of the

backup diesel generators, because the diesel generators usually need a short warmup time. To guarantee the safety of the data center during this time gap, the energy storage system, the so-called Uninterruptible Power Supply (UPS), is utilized to provide power. The UPS is typically sized to guarantee a 15-minute emergency power delivery for IT equipment of a fully-loaded data center and the critical equipment in the cooling system. The emergency operation of the UPS ends once the backup generators are brought online or the power grid recovers. The schematic drawing of a typical data center cooling and electrical system is shown in Figure 1.7 [116]. The fluid flow in the cooling system is denoted by solid lines, and the power flow in the electrical system is denoted by dashed lines.

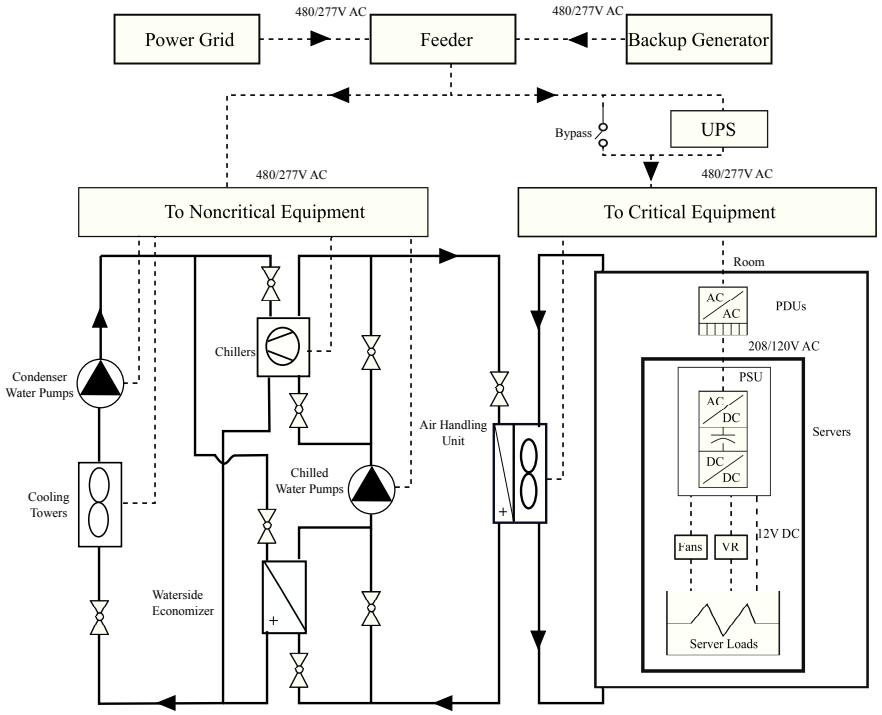


Figure 1.7: Schematic drawing of the cooling and electrical system in a data center

Note that in real data centers the electrical architecture has much more complexity and diversity. Based on their importance to keep the data center uninterrupted, all the equipment including IT and cooling equipment in a data center can be categorized into two types: critical and noncritical equipment. Critical equipment are indispensable to keep the data center functioning.

A typical design for critical equipment comprehends IT equipment and their supportive supply air fans in the CRAH. Noncritical equipment can be turned off for a short period of time without compromising the safety of the data center. They usually include all the other cooling equipment except CRAH supply air fans.

The data center is connected to the utility service and the backup generators at the building feeder. The incoming power is usually delivered to the data center building by a three-phase 480/277V AC system. During normal operation, the UPS is bypassed after it is fully charged. The power drawn by noncritical equipment enables the cold chilled water to be produced and delivered to the cooling coils in the CRAH. The supply air fans are critical equipment and powered to enable the heat transfer between the hot room air and the cold chilled water, and thus delivers the cold air to the servers. To supply the servers that usually are DC loads, the 480/277V AC power needs to be first transformed to a lower voltage such as 208/120V in the Power Distribution Units (PDUs). The 208/120V AC power is then converted to 12V DC power through the Power Supply Units (PSUs) in servers. The 12V DC power is finally delivered to the server loads and their auxiliary equipment such as fans. If the server requires power with a voltage less than 12V, then a Voltage Regulator (VR) should be used to decrease the voltage of the DC power. In emergency operation, before the backup generators are brought online, the UPS is only utilized to serve the critical equipment, and no power is delivered to the noncritical equipment.

1.2.2 Grid-interactive Efficient Data Centers

As a special type of buildings, data centers have numerous opportunities to provide grid services considering their large energy capacities, flexible working environments and work loads, redundant design and operation, etc. For example, the large energy usage in data centers makes them a suitable candidate for generation and delivery services, and the flexible working environment and work loads provides them the opportunities of ancillary services.

To advance the role buildings can play in energy system operations and planning, Department of Energy (DOE) in U.S. is developing a grid-interactive efficient building (GEB) strategy to

continuously integrate and optimize the DERs such as on-site generations and energy storage, for the benefit of the buildings owners, occupants, and the electric grid. The GEB is defined by DOE in 2019 as below [97].

> *A Grid-interactive Efficient Building is an energy efficient building with smart technologies characterized by the active use of DERs to optimize energy use for grid services, occupant needs and preferences, and cost reductions in a continuous and integrated way.*

The same definition can also be used for a grid-interactive efficient data center (GEDC), with a redefinition of "occupant" as IT equipment (e.g., servers), instead of human beings. The shifted meaning of occupant introduces specific constraints for the GEB strategy. The constraints include thermal environment and quality of service (QoS) delivered by the data centers.

GEDCs should be characterized by several features similar to that GEBs, as mentioned in Ref. [97]. First, they should be energy efficient. The efficiency can be achieved by optimizing the system design, installing high-performance site and support infrastructure, and utilizing optimal control during operation etc. Second, they should be connected. Besides the power supply, GEDCs are able to perform two-way communication with the grid by sending and receiving responsive signals. The GEDC strategy also responses to the occupants (e.g., servers) to guarantee their productivity. Third, they should be intelligent. GEDCs are able to support ubiquitous sensing and optimal control to manage DERs in ways that are beneficial to the grid, data center owners, and occupants. The optimality considers not only financial benefits but also business risk management in data centers. Finally, they should be flexible. GEDCs are capable of providing and harnessing demand flexibility.

1.2.3 Challenges in Approaching GEDCs

It is self-motivated for data centers to harness demand flexibility provided by energy efficiency. When it comes to other modes, for example, load modulation, data centers are barely willing to provide, which can be seen from one of the facts that data centers today are largely non-participants

in the demand response programs. The reasons for this are not mysterious. There are a number of significant challenges that lead to this unfortunate fact. Below, some of these biggest challenges are listed.

Firstly, decision-making tools. There is no mature tools to facilitate the decision-making process in terms of providing grid services from data centers. For example, it is difficult for data centers to harness their best portfolio of the grid services without a well-validated tool. The lack of tools exacerbates the situation where data centers are reluctant to provide grid services at the very beginning.

Secondly, regulation and market maturity. The opportunities for data center demand response participation may be limited by the immature market to simple, traditional smart pricing programs such as coincident peak pricing that it not well-suited for the risk tolerance of data centers [141].

Thirdly, risk management. Data centers are typically in the business of maximizing uptime and performance, and energy issues are certainly secondary to maintaining strong guarantees about these primary measures. However, participation in demand response programs always comes with some risk. This risk may be purely financial, e.g., in passive participation programs, or it may have the possibility of uptime/performance degradations, e.g., in active participation programs. As a result, risk management is a crucial issue for data center participation in demand response programs.

Forthly, right of control. An active debate within the demand and response field is that of who should have the control. For example, grid operators would like to have a guaranteed response when they ask for it; which leads to direct load control programs for which the grid sends a signal to a controller of the program participant. However, of course, such programs are inappropriate for data centers given the risk management issues.

Fifthly, multi-tenant issue. Many data centers are multi-tenant, i.e., they rent space to many different tenants. In such situations, the data center operator does not have control over the computing resources and so when a demand response signal is received, it cannot manage the

response directly and must find a way to encourage the tenants to respond appropriately.

Finally, real time operation. The systems in a GEDC (e.g., thermal, electrical and control systems) are interoperatable. A poweful control system is needed to coordinate and execute complex control strategies that adapt based on changing conditions over multiple time scales. The optimization techniques should be able to choose among different strategies and balance efficiency with flexibility and occupancy comfort. The data communication among systems needs to be cybersecure. All these needs complicate the design of such a control system, which requires a large amount of expertise.

1.3 Research Questions and Dissertation Structure

The above discussions indicate that approaching GEDCs still needs tremendous amount of efforts in both academia and industries, because each of the above-mentioned challenges could be a broad research topic. In this research, the focus is to mitigate the issues related to the decision-making tools and the operation of GEDCs. Although many studies have been conducted to improve energy efficiency in data centers, limited studies focused on harnessing the demand flexibility by shifting and shedding loads, and few have looked at the modulating loads. Therefore, in this study, we particularly want to answer the following questions:

(1) ***What is the suitable platform to assist the design and evaluation of GEDCs?***

Simulation is a cost-effective way to assist the design and evaluation of complex systems such as GEDCs. Most of current building energy simulation tools, for example, EnergyPlus, are designed for energy efficiency purposes. GEDCs involve different physical systems (thermal, electrical, and electromagnetic, etc.) with different time-scaled dynamics, which usually leads to a set of differential algebraic equations after mathematical modeling. Simulation is then conducted to numerically solve the mathematical equations in order to predict the system performance. To better represent the real systems, mathematical models with high-fidelity are necessary. To efficiently solve the high-fidelity models, robust and powerful

numerical solvers are required. Therefore, the complexity of the systems in GEDCs exposes special needs of the modeling and simulation platform.

To answer this question, Chapter 2 summarizes special needs for the modeling and simulation tools, and then reviewed several commonly used tools for building energy systems. A recommendation of the suitable modeling and simulation tool, Modelica, for GEDCs based on the review is finally proposed.

Chapter 3 develops a comprehensive testbed to help design and evaluate GEDCs. The proposed testbed considers different physical systems (thermal, electrical, and electromagnetic, etc.) with different time-scaled dynamics involved. End-to-end models include the computer servers, quality of service, uninterruptible power system, renewable energy resources such as solar panels, typical cooling system for data centers etc. The case studies shew that this testbed is able to perform various analysis, including detailed analysis of energy efficiency and control performance in normal operation, as well as emergency operation.

Chapter 4 validates the testbed using available measurement data from a data center located in Massachusetts. Based on the validated testbed, energy efficiency measures such as cooling system retrofitting, and control improvement were then proposed and evaluated.

(2) ***How to optimally operate GEDCs to harness the flexibility provided by modulating loads?***

Most of studies focus on energy efficiency improvement in data centers, and few focus on harnessing the demand flexibility provided by modulating loads. One of the major issues is planning the operation of DSRs to optimize the allocation of flexibility resources within operating limits in a way that is mutually economical for both building owners and electric grids. With a special focus on FR provided by modulating loads, this dissertation took two steps to investigate the control strategies to provide grid services and manage the operation risks.

First, in Chapter 5, this dissertation introduces a synergistic control strategy with a new

FR flexibility factor for GEDCs. The control strategy combines power management techniques at the server level with control of the chilled water supply temperature to track the regulation signal from the electrical market.

Second, a real-time multi-market optimization framework for GEDCs without storage systems is proposed in Chapter 5 to maximize their benefits from participating in both energy market and regulation market. The numerical case study investigates the optimal bids at each hour by considering the energy costs, demand costs and regulation revenues based on the proposed dynamic testbed of a virtual data center located in PJM.

(3) ***How to optimally operate GEDCs to harness the flexibility provided by shifting loads and modulating loads?***

Providing FR service might not be able to reduce power demands that contribute a large portion of the utility bill for data centers, because after providing FR service, the data center power will fluctuate around their baseline, and the sum of the fluctuations in a long run will neutralize over time, which might not change the power demands defined as the average power over a specific period. Therefore, storage systems especially thermal energy storage systems (TESSs) are proposed to reduce the data center power demands. The TESS can be considered as a shifting load because of its slow response. This research question aims to answer the question how to minimize the operation costs in GEDCs by adjusting shifting loads and modulating loads simultaneously.

Chapter 6 first proposes a synergistic control strategy to provide FR service in a GEDC with a TESS. The control strategy adjusts the chiller capacity, the storage charging rate and server CPU frequency together to respond to the FR signal from the electrical market. Then a multi-market optimization framework using model predictive control is introduced to minimize the operation costs and compared with different control schemes.

Lastly, Chapter 7 provides final conclusions, limitations and future work.

Note that the work contained in each chapter is largely associated with work already published or currently under review in scientific literature. These references are indicated at the beginning of each chapter.

Chapter 2

Literature Review

This chapter overviews the basics of GEDCs, including opportunities, available demand side resources, frequency regulation control strategies, PJM frequency regulation market rules and available modeling and simulation tools.

2.1 Opportunities in Data Centers

The concept of using data centers to provide demand response (DR) services stems from two critical challenges power grids are facing now. First, electric power grids need to balance generation/supply with increasing demand, partially due to increased data center use. Second, the increasing penetration of renewable energy generation in the grid has introduced more fluctuations in the power supply and thus further challenges the power grid management, especially as large-scale energy storage is not readily available.

Data centers are well-suited candidates to address these two grid-level challenges. The potential for data centers to provide DR encompasses several aspects:

Capacity Data centers represent very large loads for the grid. In 2010, data centers consumed about 1.1% to 1.5% of the total worldwide electricity and the number was about 1.7% to 2.2% for the U.S. [74]. The design load of an individual data center can be up to 50 MW or more [18]. Further, researches have shown that an optimized 30 MW data center is comparable to 7 MWh large-scale storage in providing DR service for the power grid [141]. One would potentially lose a huge storage capacity for power grid if data centers' large potential capacity for DR is not

utilized.

Flexibility Data centers can be considered as extremely flexible power loads for power grid. They can operate under a broad range of temperatures, which will result in a large range of power load. For example, American Society of Heating, Refrigerating, Air-Conditioning Engineers (ASHRAE) categorizes data centers into four types (A1-A4) based on their requirements of thermal environment. A Class A1 data center typically provides mission critical operations and requires tightly controlled thermal environment. ASHRAE suggests that the allowable supply air temperature in a Class A1 data centers should be within the range of 15 °C to 32 °C [4]. In addition, some data centers have delay-tolerant workloads, which can be shifted in time in response to electricity prices or other grid requests. The delay-tolerant workload is managed by the designs of novel hardware and algorithms that can adapt energy usage in proportion to the utilization of the computing system. Such designs include speed-scaling [29], power-capping [146, 27, 85], moving servers into and out of power saving mode [61], etc. Further, many internet-scale systems that depend on a number of geographically distributed data centers have geographical flexibility to distribute the workload to data centers at different locations [145, 52, 143].

Redundancy Data centers are designed to meet reliability standards to guarantee their uptime and performance [112]. Most data centers fall into the two high-availability classes defined by the Uptime Institute: Tier III (99.982% availability) and Tier IV (99.995% availability) [54]. Tier specifications address the number and nature of power and cooling distribution, required redundant components, and the ability to repair faults without interrupting IT load. Typical redundant equipment include power sources (e.g., backup generators, Uninterruptible Power Supply (UPS)), power delivery systems (e.g., transformers), chillers, pumps, Computer Room Air Handler units (CRAHs), etc.

Automation Nearly all sizable data centers (>1 MW IT load) have an Energy Management Control System (EMCS) that monitors and controls the cooling, electrical, and lighting systems [54]. The EMCS system often can provide limited flexibility in system operations to provide other services (e.g. DR).

2.2 Data Centers as Demand Side Resources

Ref. [54] summarized some potential demand side resources (DSRs) for data centers to participate in load-shedding and load-shifting. This paper adds additional resources that are capable of providing fast DR, shown in Table 2.1.

Site infrastructure, e.g., cooling systems, contribute to a variety of DSRs, in commercial buildings as well as in data centers, including chillers/Computer Room Air Conditioner units (CRACs), CRAHs/fans, temperature setpoints at equipment and system level, and more [69, 142, 92, 58].

Support loads, such as UPS and power delivery system, are unique resources for providing DR. Using a UPS as an on-site energy storage system can be ideal for DSR because of its capabilities to perform fast charging and discharging. Back-up generators powered by diesel or natural gas are usually configured to start in two to four seconds after a utility outage or voltage fluctuation or for greater than a 10% swing in voltage or frequency [54]. The traditional backup generators at data centers may not be environmentally friendly, in some cases even not meeting Environmental Protection Agency (EPA) emissions standards [4], which makes this form of response far from ideal. To fully utilize backup generators for in DR programs, it is necessary to reduce their emissions. In addition, a report from Lawrence Berkeley National Laboratory [54] mentions that redundant transformers could be powered down during a DR event to curtail their losses for shedding.

IT infrastructure, such as servers, storage, network etc., provide significant potential for data centers to participate in DR programs. For example, Servers are usually equipped with programmable power management mechanisms, and are capable of adjusting their power consumption using commands from certain interfaces. Fine-grained power management such as Dynamic Voltage/Frequency Scaling (DVFS) at the node level allows the processor to use a lower voltage at the cost of a slower clock frequency by offering high-resolution control [36]. Coarse-grained power management such as power capping at a low resolution and at a more aggregate level can limit the amount of electricity that servers can consume at any given time. Virtualization technologies consolidate and optimize servers, storage, and network devices in real time, reducing energy use by

enabling the optimal use of existing data center equipment as shown in [53]. Shutdown of servers and storage or job scheduling are also capable of providing load shedding in response to a DR event, usually by integrating with virtualization technologies. Load migration refers to temporarily shifting workloads from a system on one site to a system on another site. Migration between homogeneous platforms that have the same clusters requires less response time than that between heterogeneous platforms with different clusters.

Table 2.1: Demand side resources in a data center

Target	Strategy	Response Time	Ref
Chillers/CRACs	Adjust compressor frequency/speed	30 s $\sim$ 1 min	[69, 142]
	Reset chilled water supply temperature	$\sim$ 15 min	[114, 115]
	Activate/deactivate redundant chillers/CRACs	2 min $\sim$ 8 min	[53]
Supply air fan	Adjust frequency/speed	2 s $\sim$ 10 min	[92, 58, 59, 87, 127, 128, 10]
	Reset static pressure or flow rate	2 s $\sim$ 10 min	[91, 86, 87, 10]
	Activate/deactivate redundant CRAHs	N/A	N/A
Whole cooling system	Adjust supply/zone air temperature	5 min $\sim$ 50 min	[10, 148]
	Expand/shorten economization hours	5 min $\sim$ 20 min	[54]
UPS	Charge/discharge power	$\sim$ ms	N/A
	Utilize bypass techniques	N/A	[54]
Backup generator	Activate/deactivate backup generators	5 min $\sim$ 15 min	[4, 88]
Transformer	Shut down redundant transformers	N/A	[54]
IT equipment	Apply fine-grained power management	$\sim$ s	[82, 80, 26, 9, 132]
	Apply coarse-grained power management	$\sim$ s	[27, 9, 93]
	Utilize virtualization techniques	2 min $\sim$ 8 min	[53]
	Shut down servers/storage	2 min $\sim$ 8 min	[53]
	Schedule workloads	7 min $\sim$ 22 min	[53]
	Migrate workloads	2 min $\sim$ 175 min	[53]

2.3 Review of Frequency Regulation in Data Centers

Data centers have a rich pool of DSRs as shown in Table 3.1. FR service that requires fast responses from data centers can be individually or jointly provided by site infrastructure, support loads, and IT infrastructure. This section summarizes the state-of-art research of using data centers to provide FR.

2.3.1 Site Infrastructure

There are few studies in using data center site infrastructure (mainly cooling systems) to provide FR. However, the data center cooling systems are also commonly used in commercial buildings. Providing FR with commercial building cooling system has been well studied and the knowledge can be potentially applied to the data centers. Thus, this subsection mainly discusses the efforts of providing FR service by the cooling systems in commercial buildings.

Zhao et al. [148] proposed two methods of using HVAC systems in commercial buildings to provide FR: direct methods, such as adjusting static pressure setpoint, and indirect methods, such as adjusting zone air temperature setpoint. Many experimental studies focused on using the supply air fan to provide FR by changing static pressure in the air duct [91, 10], air flow rate setpoint [86, 87, 10] or frequency of the motor [92, 58, 59, 87, 127, 128, 10]. The response time can be as low as 2 s [92] by directly adjusting the VFD frequency. Some studies [114, 115] designed and evaluated a FR controller to adjust the CHWST setpoint for a chiller to track a FR signal.

Due to the specific nature of data centers, there are a few challenges and opportunities in adopting the outcome of the studies in commercial building to data centers. One is that these strategies are proposed for commercial buildings instead of data centers. Due to the high reliability requirement, data centers are required to have sufficient redundant capacity of the cooling equipment (CRAHs/CRACs, pumps, and even chillers), which is not necessary for commercial buildings. Therefore, there is a great potential for FR service using those redundant capacities in data centers.

In addition, most studies [114, 115] performed the evaluation of the FR service only on

isolated equipment, such as chillers, and their energy influence on the overall cooling system were separated. For instance, when regulation down is required, the chillers can raise their supply temperature setpoints to decrease power consumption. However, the supply air fans have to increase their power as a response to the increased chilled water temperature in the cooling coils, which counteracts the efforts of reducing the system power consumption. Thus, it is difficult to quantify the net benefits from electrical markets without considering the system as a whole.

2.3.2 IT Infrastructure

There are several techniques available to adjust the server power in order to limit data center power usage as mentioned in Table 3.1 [27, 80, 130, 93]. Some of the techniques (e.g. DVFS and dummy workload) can also be used to provide FR service because of their fast response.

Data center IT infrastructure is not always running at its full capacity. It is possible to use that unutilized capacity to provide FR service. A few studies focused on the FR service by using power management techniques such as DVFS [26, 25]. One recent publication investigated the possibility of providing FR service by introducing extra dummy workload to the servers [132]. The purpose of the dummy loads is to adjust the server utilization rates so that the server power can respond to external signals. However, there is still more spaces in using this resource to provide FR service which will be discussed in this paper.

2.3.3 Support Loads

Traditional usage of UPS in data centers is to serve as backup power. However, UPS can also be potentially used for peak shaving, power regulation, and assisting with renewable integration [95]. Studies showed that one could reduce operational costs by up to 30% by using UPS for peak demand shaving and FR service together [111, 110]. Others [73, 2] presented an architecture for distributed per-server UPS, which can participate in ancillary service markets without degrading the Quality of Service (QoS).

When considering degradation of the equipment, Chen et al. [28] concluded that it may

not be economical for batteries to participate in some markets. Analysis [95] suggestted that flow battery, a new type of electrochemical cell, allows for a fast response and placement flexibility, while conventional electrochemical energy storage technologies, such as Lead-Acid and Lithium-Ion, used for power backup, are less suitable for FR service.

There are also significant efforts underway within industry to improve the UPS design in order to enable FR service. In 2017, power management specialist Eaton launched the first pilot project of "UPS-as-a-Reserve" service [38], an initiative that enables data center owners to participate in regulation of the power grid while getting paid for their contribution. New battery technologies such as the Tesla Powerpack can charge or discharge instantly to provide FR service, voltage control, and provide spinning reserve services to the grid due to its low ramp time. However, such systems currently require large and expensive batteries to offer significant regulation capacity to the market.

2.3.4 Synergistic Strategies

Only a few of papers studied synergistic strategies that combine IT infrastructure and support loads for FR service. For example, Guruprasa et al. [56] developed a coupled data center and battery system, which allows data center to work in conjunction with a small battery to provide fast FR service. Some studies [83, 17] also considered the joint power management of a data center and plug-in electric vehicles for FR service.

It is worth to mention that the strategy that combines the IT infrastructure and site infrastructure (e.g. cooling systems) are not well studied yet because of several concerns towards manipulating the cooling system in data centers. On top of that might be the concern of cooling safety. For example, adjusting the room temperature might introduce hot spots in the racks. In addition, for an energy efficient data center whose power usage effectiveness is small, the power of the cooling system is relatively small compared with IT equipment. However, totally ignoring the capability of the cooling system might be a waste of existing resources since data center cooling energy still accounts for about $35 \sim 40$ % of the data center overall energy usage in the worldwide [66, 54]. In this dissertation, a synergistic control strategy that combines the operation of

the cooling system and IT equipment is proposed and the extra benefits of including the cooling system in the regulation control is also studied.

2.4 Overview of Regulation Market in PJM Territory

2.4.1 Traditional and Dynamic FR signals

FR is a service designed to maintain the frequency throughout the power grid system at something very close to its nominal value (e.g. 60 Hz in the United States). This is achieved by constantly and automatically balancing small fluctuations in supply and demand in real time. The service can be offered by FR resources such as generators on the supply side (which has traditionally been the case) or more recently, by DSRs on the demand side. Providing FR means FR resources are willing to increase or decrease their output (generation for generators, and consumption for DSRs) by following a control signal generated by the market operator (e.g. PJM). Details of PJM regulation services can be found in [89]. Here, only a few important and relevant features are discussed.

PJM separates FR resources into two groups: ramp-limited and capacity-limited. Typical ramp-limited resources include gas or coal-fired steam power plants which have large capacities but respond slowly to FR signals. Typical capacity-limited resources include batteries, flywheels, plug-in electric vehicles, and responsive loads which have small capacities but can respond quickly to FR signals. To fully utilize these two types of resources, PJM has developed two types of FR signals: the traditional regulation A signal (i.e., RegA) and the dynamic regulation D signal (i.e., RegD). Ramp-limited resources focus on following the slower moving RegA signal, and mostly get compensated for capacity. By contrast, capacity-limited resources can follow the faster moving RegD signal and get compensated mostly for performance.

2.4.2 Regulation Market Eligibility

Regulation offers may be submitted only for those resources electrically within the PJM RTO. To regulate, a resource must meet the following criteria:

- Generation and demand resources must be able to provide 0.1 MW of regulation capability in order to participate in the regulation market.
- Resources must be able to receive an AGC signal. A resources MW output must be telemetered to the PJM control center in a manner determined to be acceptable by PJM.
- New resources must pass an initial performance test (minimum 75% compliance required).
- Resources must demonstrate minimum performance standards. Regulating resources that have not met performance thresholds over a specified time period will be disqualified and must re-qualify to offer into the regulating market for applicable signal type (RegA or RegD). The disqualification threshold is based on the historic performance score. When the historic performance score falls below 40% by signal type, PJM will notify the resource owner and the resource will no longer be eligible to offer into the regulation market for the applicable signal type.
- Resources should give priority to the regulation signal by not allowing the sum of the regulating ramp rate and energy ramp rate to exceed the economic ramp rate. Only after a regulating resource has accounted for the regulation capability, may a generator use net of the economic base point and the regulation ramp rate to follow the energy signal.

2.4.3 Performance Calculation

In the PJM market, new resources aiming to enter the regulation market need to pass an initial test by obtaining at least 0.75 for a defined performance score. The initial test signals of RegA and RegD are available at [106]. The performance score is calculated as a composite score

of accuracy, delay and precision, which are shown below [89].

$$c_{sig,res} = \frac{COV(reg, res)}{\sigma_{reg}\sigma_{res}} \tag{2.1}$$

$$S_{accuracy} = \max_{\delta=0-5\ \min}(c_{reg,res(\delta)}) \tag{2.2}$$

$$S_{delay} = \left|\frac{5\ \min - \delta^*}{5\ \min}\right| \tag{2.3}$$

$$S_{precision} = 1 - \frac{1}{n}\sum\left|\frac{res - reg}{\overline{reg}}\right| \tag{2.4}$$

$$S_{performance} = \frac{S_{accuracy} + S_{delay} + S_{precision}}{3} \tag{2.5}$$

In the above equations, reg represents the regulation signal the DSRs receive from the electrical markets, and res represents the response signal the DSRs generate after control actions. c, COV and σ are the correlation coefficient, covariance, standard deviation of these two signals. In PJM, the response signal res is recalculated with a time shift δ ranging from 0 to 5 minutes in an increment of 10 seconds, which leads to 31 response signals $res(\delta)$. The accuracy score $S_{accuracy}$ is the maximum correlation coefficient c between reg and $res(\delta)$. The delay score S_{delay} is calculated based on the delay time δ^* when the maximum accuracy score is obtained using Eq. (2.3). The precision score $S_{precision}$ is defined as the relative difference between regulation signal and response signal, where n is the number of samples in the hour, and $\overline{reg}$ is the hourly average regulation signal. The final performance score $S_{performance}$ in that hour is calculated as the weighted average of the three individual scores.

2.4.4 Market Clearing Process

The detailed process and equations related to market clearing in PJM RTO can be referred to Ref. [89]. Here only a few summaries are listed.

The PJM Regulation Market provides PJM participants with a market-based system for the purchase and sale of the regulation ancillary service. Resource owners submit specific offers for regulation capability and regulation performance, and PJM utilizes these offers together with energy offers and resource schedules from the Markets Gateway System as input data to the Ancillary Ser-

vice Optimizer (ASO) which is an hour-ahead Market Clearing Engine (MCE). ASO optimizes the RTO dispatch profile and forecasts local marginal prices (LMPs) to determine hourly commitments of Regulation to meet the requirement.

Using the dispatch profile and forecasted LMPs, an opportunity cost, adjusted by the applicable performance score and benefits factor, is estimated for each resource that is eligible to provide regulation. The estimated opportunity cost for demand resources is zero. The adjusted lost opportunity cost is added to the adjusted regulation capability cost and the adjusted regulation performance cost to make the adjusted total regulation offer cost. The adjusted total regulation offer cost is then used to create the merit order price. All available regulating resources are then ranked in ascending order of their merit order prices, and the lowest cost set of resources necessary to simultaneously meet the PJM regulation requirement, PJM synchronized reserve requirement, PJM primary reserve requirement and provide energy in that hour is determined. The least cost set of regulation resources identified through this process are then committed. Prices for regulation are calculated simultaneously with energy and reserve every 5 minutes by the Locational Pricing Calculator. The highest merit order price associated with this lowest cost set of resources awarded regulation becomes the Regulation Market Clearing Price (RMCP). The Regulation Market Performance Clearing Price (RMPCP) is calculated as the highest adjusted performance offer from the set of cleared resources. The Regulation Market Capability Clearing Price (RMCCP)) is the difference between RMCP and RMPCP. These clearing prices are then used in market settlements to determine the credits awarded to providers and charges allocated to purchasers of the Regulation service.

2.5 Modeling and Simulation for GEDCs

2.5.1 Modeling and Simulation Needs

Modeling and simulation is an effective way to design and control GEDCs. The suitable modeling and simulation platform to serve the purpose of designing and controlling GEDCs should

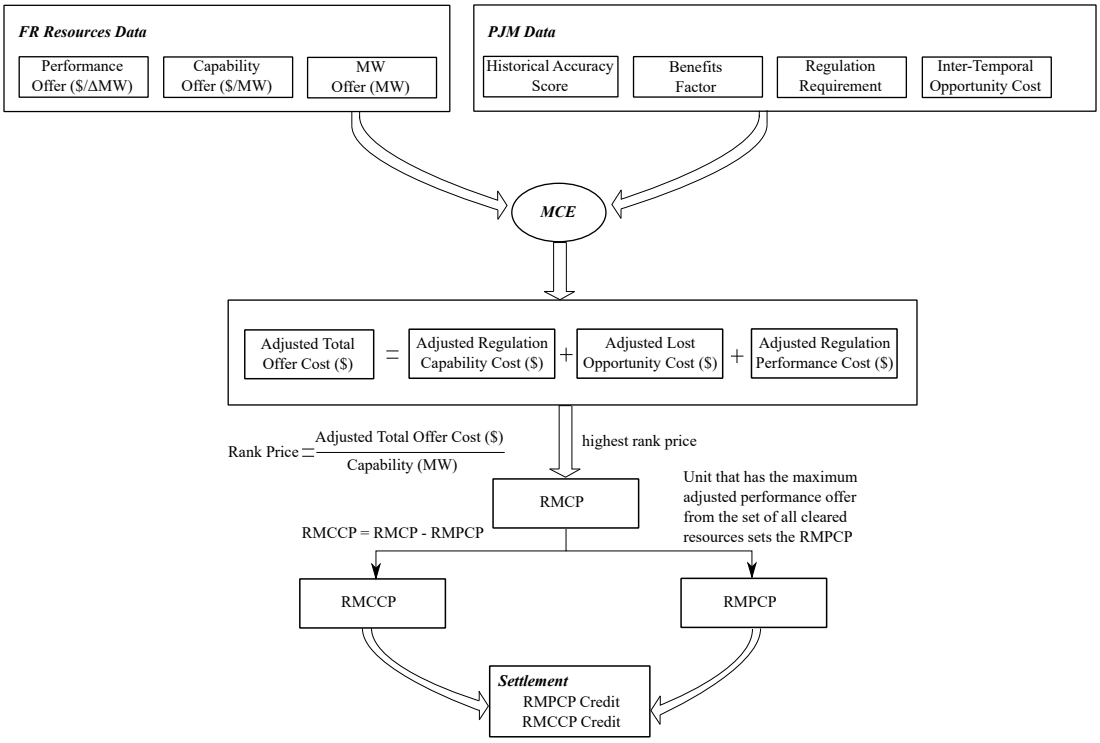


Figure 2.1: Regulation market clearing process in PJM

have the following features:

First, the platform should provide modeling flexibility. For design applications, fast and flexible prototyping can help designers and owners have a near-optimal design of the systems. For control evaluations, a GEDC can be optimized by examining different control strategies to find the best portfolio when providing grid services. The modeling flexibility in terms of design and control evaluations can be realized by two important features. One is modularity that can be provided by object-oriented programming languages such as C++ and Modelica. The modularity emphasizes compartmentalization and interrelation of independent, interchangeable modules in the platform. The other is conditional modeling. Conditional models involve discontinuities due to different user needs or physical phenomena (e.g., discrete control design). For example, to satisfy different needs of the design and the control, the ideal platform should be able to switch between steady-state and dynamic modeling.

Second, the platform should provide simulation flexibility. Simulation flexibility comes along with the modeling flexibility that can be utilized to create different use cases in the same platform such as energy analysis during design and control evaluations during operation. In the use case of energy analysis, building energy tools such as EnergyPlus approximate the HVAC equipment and controllers using steady-state models, resulting in algebraic equations. The only dynamics is from the building model. The resulting system model is not stiff, and explicit time integration algorithms are generally more efficient. In the use case of control evaluations, the modeling can lead to a hybrid numerical system that involves coupled continuous time, discrete time and discrete event dynamics. This requires solution methods with variable time steps and event handling. Therefore, the numerical solvers for simulation should also be adaptive to these different use cases so that the mathematical models could be solved efficiently.

Third, the platform should be able to provide flexibility for data synchronization with external tools. GEDCs, as complex engineered systems, requires expertise from different domains. The development of these systems may be concurrent and distributed, that is, divided between different teams or external suppliers, each in their own domain and each with their own tools.

Each participant develops a partial solution to a constituent system, that needs to be integrated with all the other partial solutions. The integration among different systems or tools is realized by synchronizing data among the different systems or tools. Data synchronization can facilitate both post-processing and co-simulation if required.

2.5.2 Review of Modeling and Simulation for GEDCs

Many researches have been conducted to perform the design and evaluation of the energy systems and control systems in GEDCs through computer modeling and simulation. This dissertation divides those researches into three categories based on the application purposes: energy efficiency, slow demand response (e.g, shedding and shifting) and fast demand response (e.g., frequency regulation and contingency reserve).

2.5.2.1 Data Center Energy Efficiency

Most of previous researches focus on improving energy efficiency in data centers, by considering energy efficiency techniques for the cooling system, or IT system or data center as a whole.

For the cooling system, researchers have proposed many techniques to provide efficient cooling by considering energy consumption or/and thermal environment in the data center room. Pan et al. [101] developed an energy simulation model for two office buildings with data centers in EnergyPlus [37] to evaluate potential retrofit energy savings. Kummert et al. [1] modeled and analyzed the system inertia of a data center cooling system using TRNSYS [44]. Kuei-Peng et al. [77] applied eQUEST developed with the DOE-2 [144] framework to explore the airside free cooling energy efficiency of data centers in 17 worldwide climate zones. Bash et al. [8] discuss a control algorithm for the CRAC units. The proposed control approach aims at minimizing the amount of heat removed by each CRAC unit, while enforcing the thermal constraints of the IT. Anderson et al. [3] consider a robust control approach to the control of the cooling system. Beitelmal et al. [11] developed a steady-state model for a centralized cooling system in a data center in EES [70] to enable the design and analysis of the heat removal processes. Computational

fluid dynamic (CFD) simulations are a widely used tool to simulate and predict the heat distribution in a data center. Such simulations take a very long time to execute and cannot be used in real-time control applications. In [121], Toulouse et al. discuss an innovative approach to CFD which is able to perform fast simulations for data centers.

For the IT system, researchers mainly focus on CPU power management [24, 105], network power management [96, 131] and storage power management [23, 55] for energy conservation. There are typically two approaches to numerically estimate the benefits of utilizing these techniques. One is through dedicated professional software. For example, in [55], the authors used the DiskSim simulator [19], augmented with a disk power model, to study the performance and power implications of Redundant Array of Independent/Inexpensive Disks configurations on the transaction processing workloads. In [21], CloudSim [22] was used to evaluate the development of dynamic resource provisioning and allocation algorithms that consider the synergy between various data center infrastructures, e.g., the hardware and power units. The other is to formulate and solve the energy efficiency problem as an optimization problem using commonly-used optimization engines. For example, Wang et al. [131] proposed a general framework for achieving energy efficiency in data center networks, where they modeled the energy-saving problem with a time-aware model and proved its NP-hardness.

There are also some researches focusing on the energy efficiency by considering coupled cooling system and IT system. Parolini et al. [103] considered data centers from a cyber-physical system perspective: a computational network representing cyber dynamics and a thermal network representing the physical dynamics. An coordinated control strategy that manages the IT and cooling system to achieve optimal performance with respect to both QoS and energy efficiency based on a discrete-time MPC approach was then proposed and evaluated. The in-house codes for this framework are not mentioned and released. Tran et al. [122] presented their modeling and simulation tool in Matlab/Simulink for data center energy simulation by considering IT system and cooling system together. Noguchi et al. [98] developed a model of the thermal behavior of a server rack in Modelica [41]; they defined a thermal circuit considering the conductance of server chassis,

cold and hot aisle to calculate each server temperature, and used an active controlled shutter to block-off cool air from a cold aisle when the temperature of the servers is less than a threshold, in this way the cooling power can be reduced by preventing excessive cooling of idle or shut-down servers. Fang et al. [42] provided a MPC-based control framework in Matlab, which divided the optimization into two subproblems, one for DVFS and IT workload assignment, the other to manage the cooling system operating point. They evaluated their results comparing the mathematical modeling of the thermal dynamics with a CFD simulation.

2.5.2.2 Data Center Demand Response

Researches related to the modeling of data center DR can be categorized into two types in terms of the way they deal with the cooling system. One is simplified (even ignored) cooling system models and the other is detailed cooling system models.

Most of the work on data center DR focused on the power management of the IT system (e.g., servers) and support loads (e.g., UPS and generator), which ignore or simplify the cooling system. The data center power is simply represented by a single variable or a set of simple equations in terms of workload schedules or CPU utilization, and the cooling power if considered is represented using a system parameter known as PUE. Because these studies are not focusing on dynamic representation of the IT system, in-house codes are usually used for the implementation. Luo et al. [90] proposed an IT workload management approach to process delay-tolerant jobs that have the same deadline of completion while maintaining their QoS. They formulated the mathematical scheduling problem to minimize the electricity costs of the data center IT system in response to real-time prices. Wang et al. [129] proposed a hierarchical DR framework for data centers to respond to real-time electricity prices. The data center energy system is simply modeled as a controllable parameter, which assumes the power can be perfectly adjustable. Bahrami et al. [6] considered the real-time pricing and modeled the data centers coupled decisions of utility company choices and workload scheduling as a many-to-one matching game with externalities. Li et al. [79] introduced a mix-integer-programming-based demand management solution for internet data centers consid-

ering the electricity market economic signals, which integrally optimizes the servers scheduling to accommodate the continuously received CPU-intense batch computing job requests. The data center power is modeled as a function of CPU utilization, and the optimization problem is solved in a C-based optimization engine, CPLEX [32]. Cupelli et al. [33] presented a Modelica-based modeling and simulation study to find the optimal sizing of data center UPS for FR service. The data center power is represented by a simple controllable variable.

There are limited researches using detailed cooling system models to investigate the data center DR. Cupelli et al. [34] modeled the site infrastructure and support loads system in data centers to evaluate the potential of ancillary services from the cooling system and generators. The detailed system is modeled using *PowerSystem* Modelica library [45] for the electrical system and *AixLib* Modelica library [94] for the thermal system. The detailed representation of the cooling system enables the evaluation of its dynamic response to external signals especially when chiller are manipulated to provide DR service. However, this platform did not model the IT system. Cupelli et al. [35] presented a Modelica framework for the optimal operation of data centers, leveraging their cooling system, delay-tolerant information technology workload and battery storage system for participating in demand response programs. This comprehensive platform, however, only considered the air loop for the cooling system, while the water loops (e.g., chilled water loop) are ignored.

2.5.2.3 Summary

Based on the previous review, several features in terms of the modeling and simulation of GEDCs can be summarized as follows.

First, many dedicated tools have been developed for different systems in data centers. For example, for the cooling energy prediction, typical building energy simulation tools such as EnergyPlus, and TRNSYS can be used. For the IT system, DiskSim can be used for the power evaluation of the storage system, and CloudSim can be used for the allocation of cloud services. These tools provide professional models merely for the dedicated system, and ignore other systems in the data

center.

Second, there are also a few researches focusing on energy efficiency and DR applications by considering both IT system and cooling system. Most of them utilized a simplified framework to couple multi-systems in a single platform such as Matlab. These simplified models cannot represent the system dynamics in the GEDCs, which might lead to false conclusions especially when the evaluation of various control strategies is needed.

Third, complex and dynamic multi-physical systems are usually represented in a Modelica-based environment. However, none of them are comprehensive enough to consider both cooling system and IT system. Also they are barely publicly accessible at this moment.

2.5.3 Comparison of Modeling and Simulation Tools

The previous subsection illustrates how the GEDC can be modeled in the literature, and finds out that data center can participate in DR programs by managing both IT system and cooling system. Because the cooling system usually has a larger time constant than the IT system, one way to simply the GEDC platform is to consider dynamic cooling system representation and steady-state IT system representation. While the IT system can be represented by simple algebraic equations as in Ref. [129], dynamic cooling system is usually represented by differential algebraic equations. Without available coupled modeling and simulation platform for GEDCs, this dissertation aims to extend existing dynamic cooling system modeling tools to support IT systems for GEDC applications. Therefore, this subsection gives a detailed comparison of the cooling system model tools and makes a recommendation for the final choice of GEDC platform development.

Commonly used building energy modeling and simulation tools that could be used directly or indirectly for GEDCs can be categorized into two types based on the programming language they are using. One type is based on imperative programming languages such as C/C++ and FORTRAN. Examples of such conventional building simulation programs include DOE-2 [14], TRNSYS [44], and EnergyPlus [37]. The other is based on equation-based programming languages such as Matlab/Simulink, and Modelica. Examples include Thermosys Toolbox [78] in Matlab/Simulink,

and Modelica Buildings Library [140] that can be executed in Modelica environment such as Dymola, OpenModelica, and JModelica.

2.5.3.1 Imperative Language-based Modeling and Simulation Tools

DOE-2 is widely used and accepted as a free building energy analysis program that can predict the energy use in buildings. DOE-2 solves the building envelope thermal dynamics with HVAC system operating performance sequentially due to its sequential software structure of Load-Systems-Plant-Economics [144]. Hour-by-hour evaluation makes it hard if not impossible to evaluate any control feedback loop. The HVAC equipment is calculated using semi-dynamic and steady-state models.

TRNSYS can be used to perform dynamic simulation of the behavior of a building system integrated with controls. Up through TRNSYS 17, two methods for solving the coupled system of algebraic and differential equations were provided: the successive substitution method and Powells method [71], where the latter is deactivated in TRNSYS 18. In successive substitution method, at each time step, the outputs of a given model are substituted for the inputs of the next model in the system. The performance of that next model is recomputed and its outputs are then substituted for the inputs of the next model. This process continues until all connected outputs have stopped changing (within error tolerance). So all the components are called at least once during this time step. This method does not solve nonlinear algebraic equations efficiently, and may not be able to find a solution if the equations are highly nonlinear.

EnergyPlus applies a simultaneous solution scheme to update the states of all the elements such as Zone, System, and Plant. For different elements, different solvers are applied to solve the equations [37]. For example, differential equations that are used to estimate zone air temperature can be solved using Euler formulas, 3rd-order BDF algorithm, and an analytic solution. To update the simultaneous solutions of Plant/System water loop, the secant method is used to predict updates to the plant leaving water conditions. Additionally, the Gauss-Seidel iterative method is used to reconcile supply and demand in the air loop and water loop. Most models in EnergyPlus are quasi-

steady energy balance equations used to predict the conditions present during each time step. Predictions for state variables, such as temperature, are averages over the time step.

2.5.3.2 Equation-based Modeling and Simulation Tools

EES is a general equation-solving program that numerically solves the DAE equations [70]. The built-in heat transfer library provides support for evaluation of the performance of the building system. Tarjan blocking algorithm [117] is first applied to break the large set of DAE equations into a number of smaller sets, which are easier to solve. For the differential equations, explicit method such as Euler Forward, and implicit method such as Euler Backward are available in EES. For algebraic systems, EES uses a variant of Newtons method to solve systems of non-linear algebraic equations. And Jacobian matrix needed in this iterative method is evaluated at each iteration. EES has several limitations [12, 13]: the available library for evaluating building system has limited component models, which means users may need to build their own models from scratch; also for a large HVAC system simulation, EES requires much longer computation time that other simulation software.

Some Matlab/Simulink tools like Thermosys [78] are also able to perform dynamic evaluation of building performance. Although these models share the same development environment, the developers of different tools have defined customized data structure for the models, which makes the integration of models from different tools time-consuming. As a mathematical programming tool, Matlab provides large amount of numerical methods for differential equation systems and algebraic equation systems. All these Matlab-based tools can benefit from the numerical solvers.

Modelica is an equation-based, and object-oriented modeling language [41]. Modelica Buildings Library, developed by Lawrence Berkeley National Laboratory, is an open-sourced, dedicated program for dynamic building and community energy performance simulation built on Modelica [140]. The Modelica source code is first translated into a so-called flat model. The flat model includes a set of equation declarations and functions, with all the object-oriented structure removed. Then the system of equations is transformed into a suitable form for the numerical solvers, and

manipulated using symbolic operations to decrease its complexity. Next, the system of equations is further arranged using the block lower triangular (BLT) decomposition that formulates the system into smaller equation blocks to solve sequentially. Finally, the procedural code (e.g., C++ code) is generated based on the previously computed BLT blocks [20]. Modelica-based modeling platforms provide some important features such as object-oriented, and acausal modeling, and support a rich library of numerical solvers for PDE, ODE, and DAE systems. The equation-based scheme also makes possible the nearly decoupled framework for model development: computer scientists can dedicate to developing and updating the numerical solvers, while the HVAC engineers can focus on high-quality physical models [81].

2.5.3.3 Pros and Cons

As mentioned above, two types of programs are developed in past few decades to perform building-level and community-level energy simulation, which also can be used for GEDC applications. One is based on imperative programming language, the other is based on equation-based modeling language. Most of these two types of software share a few common attributes: first, most of them are component-based modeling tools, whose hierarchical structure makes the modeling process much easier and convenient to the users; second, the large physical system is partitioned to many smaller sets, which are easier to solve simultaneously.

However, they are different in many other aspects.

(1) Most imperative language-based programs need specify the predefined sequence of the commands for the computer to perform calculations of assigned variables. For example, the DOE-2 needs solve Load-Systems-Plant-Economics sequentially to end the simulation. However, the equation-based programs usually do not need to predefine the equation orders. The compilers of the equation-based programs would use a partition algorithm such as Tarjan in Modelica and EES to automatically permute those equations into separate small blocks, and call numerical solvers to solve those small blocks.

(2) In most imperative language-based programs, models usually tightly couple physical equations, input/output routines with numerical solution methods, by making the numerical solution procedure part of the actual model equations [136, 138], such as the zone temperature model in Energy Plus, PID controller model in TRNSYS and Matlab/Simulink, and cooling load calculation in DOE-2. This intertwinement makes it difficult to extend these programs to support co-simulations with each other [107, 123] and effective optimization [136]. Also it is hard to apply these models in different user cases such as control design and verification, and coupled modeling of thermal and electrical system, because different numerical solvers might be required in different systems (stiff systems, non-stiff system and hybrid systems) [138]. So models for building energy systems and their numerical solution should be separated where possible. One solution is to use the equation-based programming language such as Modelica.

(3) In terms of evaluation of control strategies, many imperative-language-based tools have limited capability.

Some tools are difficult to capture the dynamic response of the control loop, because the controller is assumed to be ideal. For example, in EnergyPlus, the commonly used PI control loop is assumed to be ideal, i.e., there will be no overshoot and error bands [137]. EnergyPlus also ignores dead band or waiting time, which are frequently used in the discrete building control process such as the staging of cooling towers. Without representing the dynamic processes, it is impossible to evaluate the control performance.

Some dynamic tools may generate numerical artifacts. For example, Ref. [68] reported an unexpected fluctuation in the output of a TRNSYS PI controller model, which was used to control an air cooling system. They believed this fluctuation was actually a numerical artifact and is caused by the way that TRNSYS calculates the output of the PI controller. In TRNSYS, the PI controller model calculates a control action signal based on converged values of the controlled variables at the current time step but the calculated action is

not applied to the system until next time step. This time lagging can easily introduce instability, which will become numerical artifacts if the studied system rapidly changes. Such numerical artifacts may lead to a completely different control response compared to the real one, which makes the evaluation of the control strategies very difficult.

Most tools do not consider the pressure distribution in the fluid loop. For example, in EnergyPlus, the flow rates in both the duct and pipe networks are directly calculated according to the HVAC load. The pressure distribution in those networks are not considered. However, the pressure may be used as an input for building control systems especially in data centers (e.g. the head pressure control reported in Ref. [113]). Therefore tools like EnergyPlus might be challenging to model such a control system.

(4) To obtain reasonable trajectories of the state variables in a dynamic system, the time step should be employed carefully. For example, to capture the fast dynamics caused by a discrete event (e.g., an equipment switches on or off), the time step should be set small enough. But for other simulation periods when there is no event occurring, the time step can be set relatively large to accelerate the simulation without loss of accuracy. Therefore, to solve such a system, variable time steps should be used. Modelica and ESS both have solvers that support variable time steps. However, it is quite common that the time step is fixed and predefined in the imperative programming language-based building modeling tools. For example, EnergyPlus uses fixed time step (at least 1 minute) throughout the whole simulation, and DOE-2 requires a time step of 1 hour to perform the calculation.

2.6 Summary

This chapter provides literature reviews over a few topics including opportunities in developing a GEDC, demand side resources in a typical data center, frequency regulation control in data centers, ancillary services especially frequency regulation market in PJM, and modeling and simulation platform for GEDCs.

Section 2.1 reviewed the numerous opportunities in data centers in terms of providing grid services. By understanding these opportunities, data centers can be more motivated to participate in electric markets.

Section 2.2 summarized different demand side resources in a typical data center and the potential grid services they can provide.

Section 2.3 provided a detailed literature review with a special focus on how data centers can provide frequency regulation services to the grids.

Section 2.4 explained briefly the market mechanism of PJM territory in terms of frequency regulation service.

Section 2.5 reviewed the modeling and simulation platform to help design and evaluate GEDCs. First, the special modeling and simulation needs are identified. Then a brief literature review is conducted to understand current existing modeling and simulation framework for GEDCs. Lastly, current existing building energy modeling tools are compared, and a suitable candidate, Modelica, is identified to perform modeling and simulation for this dissertation.

In summary, to meet the modeling and simulation needs mentioned in Section 2.5.1, object-oriented equation-based modeling and simulation tools are identified as suitable candidates for the evaluation of design and control of GEDCs, on top of which, Modelica is the strongest candidate.

Chapter 3

Development of Numerical Testbed

Based on:

Yangyang Fu, Wangda Zuo, Michael Wetter, James W. VanGilder, Peilin Yang, "Equation-based object-oriented modeling and simulation of data center cooling systems", Energy and Buildings, Volume 198, 2019, Pages 503-519, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2019.06.037.

In this chapter, an open source data center package in the Modelica Buildings library to support modeling and simulation of cooling and control systems of data centers is developed. This package has been extended to support grid-interactive efficient data centers as well. The data center package contains major component models, such as server, Computer Room Air Handler, Computer Room Air Conditioner, models of different subsystem configurations such as chillers with differently configured waterside economizers, as well as templates for different systems. Two case studies are performed to investigate the performances of the cooling and electrical system under normal conditions and emergency situations such as a blackout: one is for a data center powered by conventional energy, and the other is for a data center powered by renewable energy. Simulation results show that the dynamic modeling and multi-domain simulation in the Modelica-based tool make it convenient for users to investigate not only the thermal performance but also the electrical performance.

3.1 Dynamic Modeling Platform

Data centers are critical, energy-intensive infrastructures that support the fast growth of the information technology (IT) industry and the transformation of the economy at large. In 2010 data centers consumed about 1.1% to 1.5% of the total worldwide electricity and the number was about 1.7% to 2.2% for the U.S. [74]. The energy in data centers is mainly consumed by two parts: IT equipment (e.g., servers, storage, network, etc.) and infrastructure facilities (e.g., cooling system). The latter usually accounts for about half of the total energy consumption in a typical data center [125].

Modeling and simulation is a cost-effective way to evaluate the design and operation of cooling systems. Different physical systems (thermal, electrical, and electromagnetic, etc.) with different time-scaled dynamics are involved in such systems. This usually leads to differential algebraic equations. Simulation is then conducted to numerically solve the mathematical equations in order to calculate the system performance.

Many tools have been developed in academia and industry to perform computer modeling and simulation of the energy systems in data centers. For example, eQuest [77], EnergyPlus [101, 57], TRNSYS [1], and some customized simulation tools such as Energy Modeling Protocol [109] have been used to study cooling systems with waterside economizers (WSEs) and airside economizers (ASEs) in data centers. Most of these traditional tools are based on imperative programming languages such as FORTRAN and C/C++. When implementing a physical model in these tools, model developers utilize their expertise to sort the physical equations in an order so that the unknowns (model outputs) in the equations can be solved based on given known variables (model inputs). The nonlinear equations are usually manipulated to be solved iteratively, and the differential equations are discretized to numerically approximate the state variables. Then, the model developers write the variable assignments into computer source codes. Other computer program procedures may be called in the source codes to calculate the input variables from a subsystem at each time step, which might request from the solver re-simulation of a subsystem iteratively [136].

The above-mentioned conventional simulation tools expose several disadvantages in terms of their modeling and simulation performance. First, in the imperative programming languages, mathematical equations are typically intertwined with numerical solvers. For example, a typical zone model is mathematically described as a first-order differential equation as shown in Eq. (3.1) where m is the zone air mass, $c_{p,a}$ is the specific heat capacity of the air, T_{zone} is the zone air temperature, t is time, and Q_i is the ith heat source in the zone.

$$mc_{p,a}\frac{dT_{zone}}{dt} = \sum_{i=1}^{n} Q_i(T_{zone}) \tag{3.1}$$

In EnergyPlus that is written in C++, the zone model is represented by discretizing the differential term on the left side over time. One of the discretization algorithm used for the zone model is the 3rd-order backward difference formula, which converts the differential equation into a set of algebraic equations as shown in Eq. (3.2).

$$mc_{p,a}\frac{\frac{11}{6}T^{t}_{zone} - 3T^{t-\delta t}_{zone} + \frac{3}{2}T^{t-2\delta t}_{zone} - \frac{1}{3}T^{t-3\delta t}_{zone}}{\delta t} = \sum_{i=1}^{n} Q_i(T^{t}_{zone}) \tag{3.2}$$

where δt is the length of the time step, and subscripts t,..., $t-3\delta t$ represent the time instance. The numerical method is integrated with the mathematical equations in the source codes, which leads to a program code that is hard to maintain. By accepting non-convergent solutions at intermediate time steps to the simulation results, the nested solver can also introduce numerical noises that can pose challenges to optimization programs [136]. Second, some platforms are not designed for evaluating the system dynamics and the semantics of their control have little in common with how actual control works [48]. For example, in EnergyPlus, the commonly used Proportional-Integral (PI) control loop is assumed to be ideal, i.e., there will be no overshoot. EnergyPlus also idealizes dead band or waiting time, which are frequently used in building controls. Moreover, many equipment models have built-in idealized control that requests flow rates, and flow rates are ideally distributed within a system rather than the results of friction-based flow distribution. This makes it difficult to model, test and verify actual control. Third, different numerical solvers for differential equations might be needed for different use cases [138]. However, most traditional tools

have predefined solvers in their physical component models. Forth, these tools are hard to support fast prototyping based on various user's needs. For example, the control logic for the WSE in DOE2.2 are predefined. Users are able to change the thresholds of particular conditions, but not the logic themselves. It is difficult for users to implement new logic. Fifth, these tools are difficult to perform multi-domain simulations. For example, to study the interactive performance of the thermal and the electrical system, one needs an external data synchronizer to couple these tools with electrical simulation tools as mentioned in Ref. [75].

Equation-based languages such as Modelica [41] can provide solutions to the above-mentioned issues. Modelica separates physical equations and numerical solvers wherever possible. The separation can mitigate the risks of intertwinement, and can fully take advantages of different expertise from different domains. For example, model developers can concentrate on how to develop efficient high-fidelity physical models, while computer engineers can focus on the development of robust numerical solvers. Also, the State Graph package [100] in the Modelica Standard Library can be used to perform discrete control which contains dead band or delay time. The rich library of numerical solvers in Modelica can be chosen for different systems and different use cases. Besides, Modelica models are convenient enough to be extended to support fast modeling and simulation. Furthermore, Modelica itself supports multi-domain simulation. Models from different domains are built in one single platform so that dedicated data synchronizer between different platforms is not required.

The Modelica Buildings library is designed to model and simulate the energy and control system at building and community level [140]. The Buildings library is free open-source, and has been demonstrated to have full capability to conduct energy efficiency analysis. Researchers have been active to utilize the Buildings library in a broad range of applications, such as dynamic modeling [40, 47, 46, 51], rapid prototyping of a district heating system [135], evaluation of feedback control [136], fault detection and diagnosis at the whole building level [76], optimal model-based control design and evaluation [63, 64], as well as coupled simulation between the cooling system and the detailed room model with fluid dynamics considered [119, 120]. However, this library so far

have very limited capability to support the design and operation of GEDCs. There are two main reasons. First, current models are insufficient to describe a complete energy system in the data center. For example, CRAC models and IT server models are missing. Second, existing models are not well constructed to meet the needs of fast modeling for design and control evaluations. For instance, to a model a chiller system with WSE, the users have to form the system level model by connecting the component models piece by piece. The lack of system templates extends the learning time of new users.

This study further extends the Modelica Buildings library to support fast modeling and simulation of energy systems for data center applications. As some of the data center cooling systems, such as chiller plants, are also commonly used by large commercial buildings and district cooling systems, the models developed by this study can also be used for those applications. This chapter first introduces typical air-cooled cooling systems in data centers such as chilled water system and direct expansion (DX) system. In Section 3.2, we give an introduction of the component models, subsystem models, cooling control models, and system models in the data center package. In Section 3.4 and 3.5, to demonstrate the capability of Modelica-based tools, we model and simulate an energy system under normal operation and emergency operation for a conventional and a renewable data center respectively. We then conclude the chapter in Section 3.6.

3.2 Mathematical Models

The data center package Buildings.Applications.DataCenters is built on Modelica Buildings library and Modelica Standard library as shown in Figure 3.1, and was publicly released in the Modelica Buildings library 5.0.0 (https://simulationresearch.lbl.gov/modelica/). It contains component models for the above-mentioned two typical air-cooled cooling systems in data centers. This package adopts the class hierarchy used by the Buildings library, and contains various reusable base classes. These base classes together with the inheritance and instantiation in the object-oriented modeling language Modelica facilitate fast model-based design of data center cooling systems. The following part introduces some of the key models for cooling components, subsystems, controls and

system templates.

3.2.1 Component Models

3.2.1.1 Group of Equipment

A group of chillers and pumps can be modeled on the base of existing chiller and pump model respectively. Take as an example the model of a parallel of chillers. The diagram for the ready-to-use model for the chiller parallel is shown in Figure 3.2. This model utilizes existing chiller and valve models in Modelica Buildings library and a filter model in Modelica Standard library.

$$(T_{c,o,k,j}, \dot{m}_{c,o,k,j}, \Delta P_{c,k,j}) = chiller(T_{c,i,k,j}, \dot{m}_{c,i,k,j}, u_j) \tag{3.3}$$

$$(T_{v,o,k,j}, \dot{m}_{v,o,k,j}, \Delta P_{v,k,j}) = valve(T_{v,i,k,j}, \dot{m}_{v,i,k,j}, u_{f,j}) \tag{3.4}$$

$$u_{f,j} = filter(u_j) \tag{3.5}$$

In the above equations, T, $\dot{m}$, and ΔP are temperature, mass flow rate and pressure losses respectively. For the subscripts, c,v and f are short for chiller, valve and filter, i and o refer to the inlet and the outlet, k represents the two sides of the equipment with 1 denoting evaporator side and 2 denoting condenser side, and j is the index of equipment ranging from 0 to N, where N is the design number of the chiller in the group. The functions $chiller$,$valve$, and $filter$ represent existing chiller, valve and filter model, respectively. The chiller model can be referred to [37], and the valve model can be referred to [5]. For the filter model, we use a second-order critical damping low pass filter to smooth the input control signal in order to represent the mechanical inertia in the equipment such as valves as introduced in [140]. Detailed equations are shown in Eq. (3.6) - (3.8). a is the coefficient of the state space equations for the first-order filter described in Eq. (3.7) and (3.8). f_{cut} is the cut-off frequency of the low pass filter, which passes signals with a frequency lower than f_{cut} and attenuates signals with a higher frequency. α is a frequency correction factor for different orders. Here we use 0.622 for the second order. u is the input signal, and u_f is the

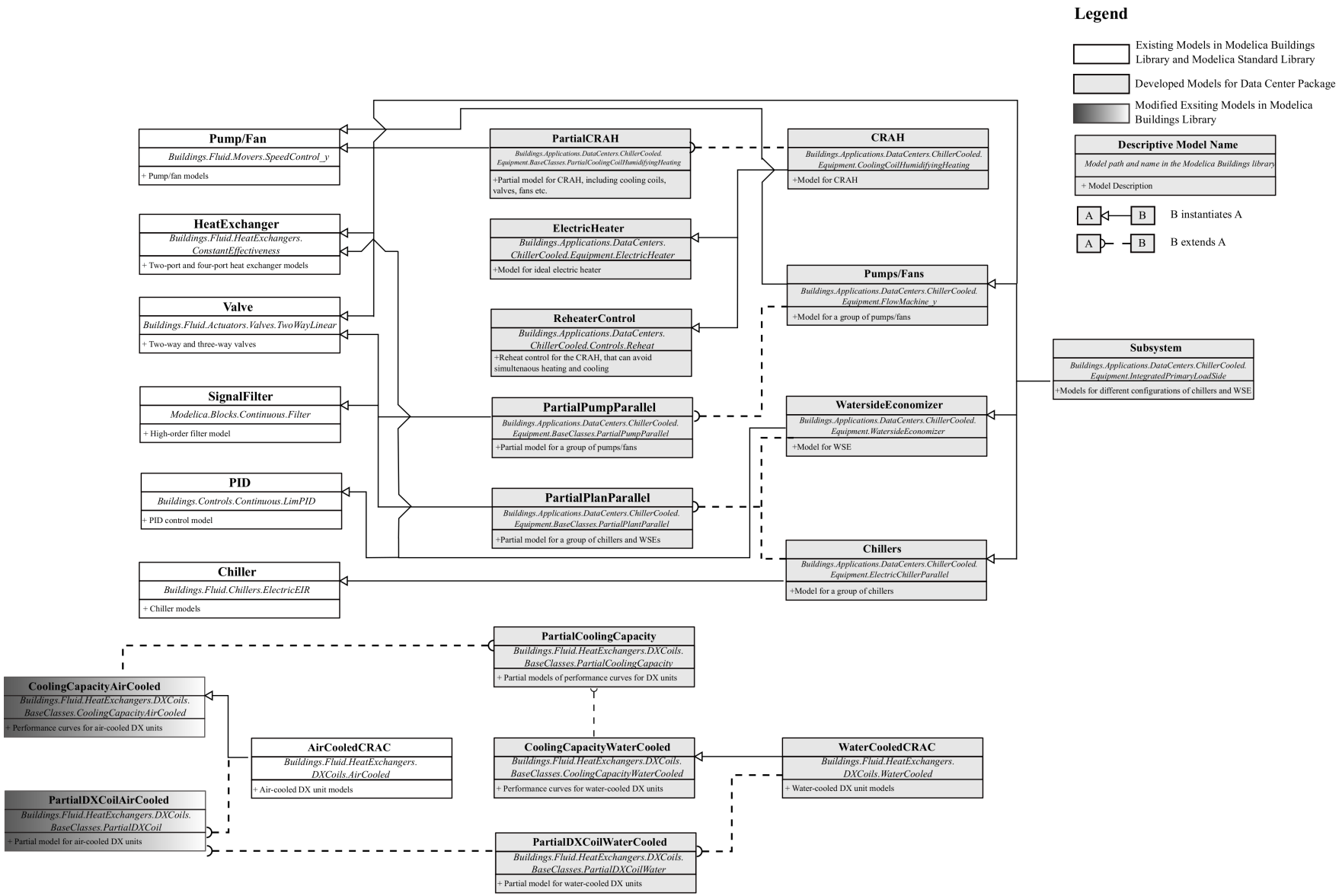


Figure 3.1: Class inheritance and instance diagram for part of the data center package

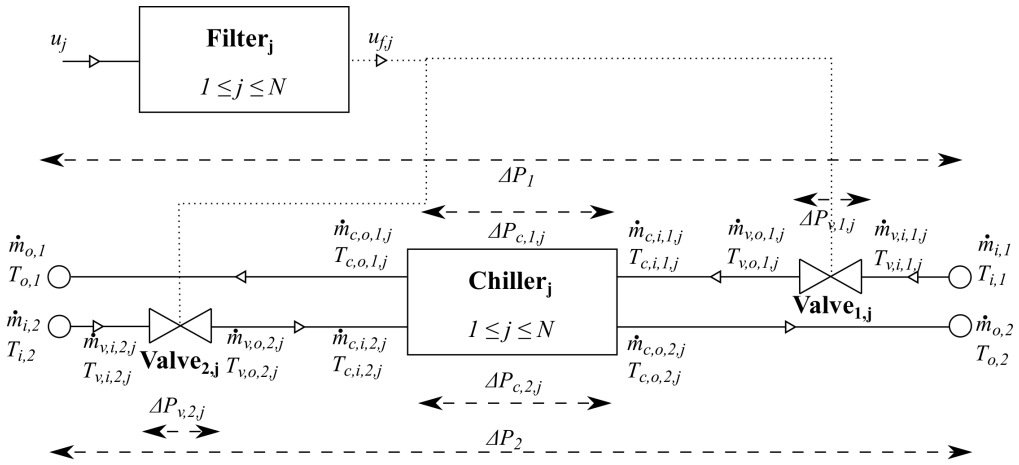


Figure 3.2: Diagram of the new chiller parallel model

output signal from the filter.

$$a = -\frac{2\pi f_{cut}}{\alpha} \tag{3.6}$$

$$\frac{dx}{dt} = a(x - u) \tag{3.7}$$

$$\frac{du_f}{dt} = a(u_f - x) \tag{3.8}$$

The group model is built by connecting the abovementioned individual component together in a way that the chillers are in parallel with each other, one valve is at the upstream of each side of each chiller, and each valve receives the signal of position from the filters. The mass flow rate at each component on each side such as $\dot{m}_{v,i,k,j}$ is calculated based on the pressure balance in the fluid network by combining Eq. (3.3) through Eq. (3.13). $\dot{m}_{i,k}$ and $T_{i,k}$ are the mass flow rate and temperature at the inlet of the group.

$$\dot{m}_{i,k} = \sum_{j=1}^{N} \dot{m}_{v,i,k,j} \tag{3.9}$$

$$\dot{m}_{c,i,k,j} = \dot{m}_{v,o,k,j} \tag{3.10}$$

$$T_{i,k} = T_{v,i,k,j} \tag{3.11}$$

$$T_{c,i,k,j} = T_{v,o,k,j} \tag{3.12}$$

$$\Delta P_{c,k,1} + \Delta P_{v,k,1} = ... = \Delta P_{c,k,j} + \Delta P_{v,k,j} \tag{3.13}$$

The outlet conditions of the grouped chillers such as mass flow rate $\dot{m}_{o,k}$ temperature $T_{o,k}$ and pressure loss ΔP_k are then obtained as follows:

$$\dot{m}_{o,k} = \sum_{j=1}^{N} \dot{m}_{c,o,k,j} \tag{3.14}$$

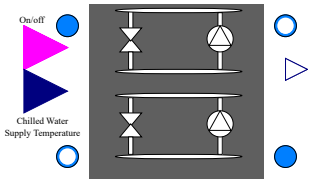
$$\dot{m}_{o,k} T_{o,k} = \sum_{j=0}^{N} \dot{m}_{c,o,k,j} T_{c,o,k,j} \tag{3.15}$$

$$\frac{1}{\Delta P_k} = \sum_{j=1}^{N} \frac{1}{\Delta P_{c,k,j} + \Delta P_{v,k,j}} \tag{3.16}$$

The implementation of the grouped chillers in Modelica benefits from the data structure *array*, which can vectorize the existing chiller model directly. The vectorized equipment model uses the same design parameters, but different performance curves if needed. The pseudo-code of vectorized chiller model *ElectricChillerParallel* is shown in Figure 3.3(a). First, a partial class of the electric chiller model *Fluid.Chillers.BaseClasses.PartialElectric* is instantiated through vectorization with a number n, which specifies the length of the chiller array. The keyword *replaceable* allows the model to be redeclared with a detailed chiller model later on. Line 3 specifies the medium used in the chillers. Line 4 defines the identical design parameters for the chillers with the keyword *each* in Modelica, such as the design capacity. Line 5 defines the performance curves of each chiller by assigning different curves from a performance curve array. The same instantiation method is also used to model a group of pumps. In addition, we add isolation valves in the vectorized models to avoid circulating flow among components. The implemented source code is packaged in a model and the graphic icon shown in Figure 3.3(b) is added to support graphical modeling.

```
1: replaceable Fluid.Chillers.BaseClasses.PartialElectric chillers[n]
2: constrainedby Fluid.Chillers.BaseClasses.PartialElectric(
3:       redeclare each final replaceable package Medium = Medium,
4:       each final parameters = parameters,
5:       final performanceCurve = performanceCurveArray)
6: "Chillers with identical design parameters but different performance curves"
```

(a) Pseduo code



(b) Modelica icon

Figure 3.3: Vectorized chiller model in Modelica

3.2.1.2 Waterside Economizer

The WSE model is built using a heat exchanger model with constant effectiveness, and a three-way valve model. The mathematical equations that define the distribution of fluid flow is similar to that in Section 3.2.1.1. Figure 3.4 shows the Modelica implementation. The three-way valve is on the chilled water side, and can be adjusted to control the chilled water supply temperature using a built-in PI controller. The three-way valve can be activated or deactivated based on users' needs for different control strategies. For example, in the FC mode, the mechanical

cooling is shut down, and only the WSE is activated to provide cooling. The chilled water supply temperature downstream of the WSE can be controlled at its set point by regulating the speed of the cooling tower fans or by modulating the three-way valve on the chilled water side in the WSE. The former control strategy requires deactivation of the three-valve, while the latter control needs to activate the three-way valve. The switch between activation and deactivation of the three-way valve is realized by setting a Boolean parameter *activateControl* to *True* or *False*.

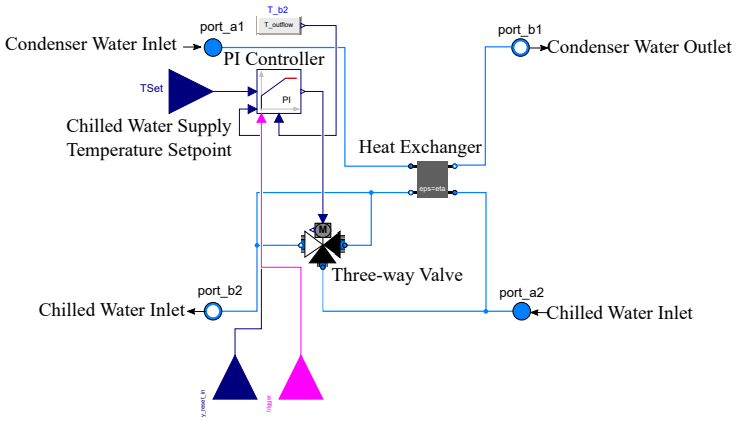


Figure 3.4: Waterside economizer model in Modelica

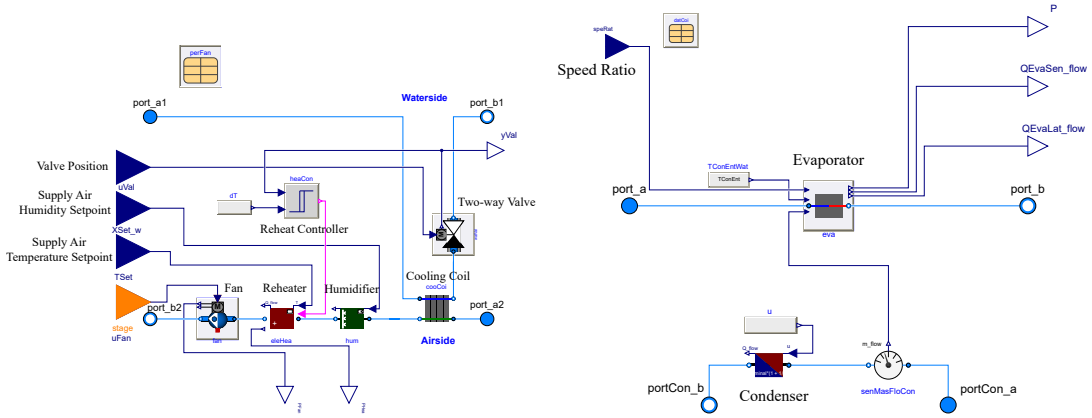
3.2.1.3 Computer Room Air Handler

As shown in Figure 3.5(a), the CRAH model named *CoolingCoilHumidifyingHeating* is built on the existing models of a cooling coil, a humidifier, a fan and a two-way valve on the water side of the cooling coil. An ideal electric reheater and an on/off controller with hysteresis is added in order to avoid simultaneous heating and cooling.

The electrical reheater is modeled as an ideal heat transfer process. The required heat flow $\dot{Q}_r$ to control the outlet temperature at its setpoint $T_{a,s}$ is shown in Eq. (3.17). And the power of the electrical heater P is then obtained in Eq. (3.18).

$$\dot{Q}_r = \max(0, \dot{m}_a C_{p,a}(T_{a,s} - T_{a,i})) \tag{3.17}$$

$$P = \eta \dot{Q}_r \tag{3.18}$$



(a) CRAH (b) CRAC

Figure 3.5: Modelica models for computer room cooling sources

The electrical reheater is activated only if the following two conditions are met simultaneously:

$$y_{w,v} \leq y_{w,v,sh} - \Delta y, \tag{3.19}$$

$$T_{a,o} - T_{a,s} \leq dT_{sh} - \Delta dT \tag{3.20}$$

where $y_{w,v}$ is the two-way valve position, $y_{w,v,sh}$ is a user-defined switching threshold, $T_{a,o}$ is the outlet air temperature, $T_{a,s}$ is the outlet air temperature setpoint, dT_{sh} is a user-defined temperature difference threshold, Δy and ΔdT are the control dead band for the valve position and temperature difference. These two conditions mean that the electric reheater is only activated when the valve on the waterside reaches its minimum position, and the outlet temperature is still lower than its setpoint.

3.2.1.4 Computer Room Air Conditioner

Both air-cooled and water-cooled CRAC models are available in Modelica Buildings library. The capacity of the refrigeration system in the CRAC is expressed using regression equations based on inlet temperature and flowrate on both the evaporator and condenser sides. The equations are based on Ref. [37]. The water-cooled CRAC model is shown in Figure 3.5(b), and detailed equations are illustrated as below.

This model uses the coil bypass factor (bf) to calculate sensible and latent heat thorough the cooling coils. The coil bypass factor at nominal conditions bf_0 can be calculated as:

$$bf_0 = \frac{h_{a,o,0} - h_{ADP,0}}{h_{a,i,0} - h_{ADP,0}} \tag{3.21}$$

where $h_{a,o}$ and $h_{a,i}$ are the enthalpy of air at the outlet and inlet respectively, h_{ADP} is the enthalpy of saturated air at the coil apparatus dew point, and subscript 0 is the nominal condition. The model uses modifiers to correct the nominal values in order to predict the off-design conditions. For example, Eq. (3.22) and (3.23) show the calculation of the available total cooling capacity $\dot{Q}_{t,a}$ and energy input ratio EIR from their corresponding nominal values $\dot{Q}_{t,a,0}$ and EIR_0 and the modifiers g_1 to g_6. r_a and r_w are air and water flow ratio as defined in Eq. (3.24) and (3.25), g_1 and g_4 are

biquadratic equations, and g_2, g_3, g_5, g_6 are polynomial equations. These modifier equations can be obtained from curve-fitting techniques.

$$\dot{Q}_{t,a} = \dot{Q}_{t,a,0} g_1(T_{a,wb,i}, T_{w,i}) g_2(r_a) g_3(r_w) \tag{3.22}$$

$$EIR = EIR_0 g_4(T_{a,wb,i}, T_{w,i}) g_5(r_a) g_6(r_w) \tag{3.23}$$

$$r_a = \frac{\dot{m}_a}{\dot{m}_{a,0}} \tag{3.24}$$

$$r_w = \frac{\dot{m}_w}{\dot{m}_{w,0}} \tag{3.25}$$

The sensible heat ratio (SHR) in the cooling coil is calculated using Eq. (3.26), where $h_{T_{a,i},w_{ADP}}$ is the enthalpy of the fictitious air with the same dry bulb temperature of the actual inlet air $T_{a,i}$ and the same humidity ratio of the saturated air at coil apparatus dew point condition w_{ADP}, $h_{a,i}$ is the enthalpy of the actual inlet air, and h_{ADP} can be calculated from Eq. (3.27).

$$SHR = \frac{h_{T_{a,i},w_{ADP}} - h_{ADP}}{h_{a,i} - h_{ADP}} \tag{3.26}$$

$$h_{ADP} = h_{a,i} - \frac{\dot{Q}_{t,a}}{m_a(1 - bf)} \tag{3.27}$$

The power consumption by the compressor P and heat rejection in the water-cooled condenser $Q_{t,w}$ can be calculated as

$$P = \dot{Q}_{t,a} EIR \tag{3.28}$$

$$\dot{Q}_{t,w} = \dot{Q}_{t,a}(1 + EIR) \tag{3.29}$$

The outlet conditions on the air side are then calculated based on $\dot{Q}_{t,a}$ and SHR as shown in following equations:

$$m_a \frac{dh_{a,o}}{dt} = \dot{m}_a(h_{a,i} - h_{a,o}) - \dot{Q}_{t,a} \tag{3.30}$$

$$m_a \frac{dw_{a,o}}{dt} = \dot{m}_a(w_{a,i} - w_{a,o}) - \frac{(1 - SHR)\dot{Q}_{t,a}}{h_{g,w}} \tag{3.31}$$

In the above two equations, $h_{g,w}$ is the latent heat of condensation of water, whose value is 2442 J/g at 25 °C, and m_a is the mass of the air volume in the evaporator. Here assuming that

the density of the air keeps constant, we can get

$$m_a = \tau_{a,0}\dot{m}_{a,0} \tag{3.32}$$

where $\tau_{a,0}$ is the design thermal time constant of the evaporator, and can be defined by users. The outlet temperature of the water-cooled condenser $T_{w,o}$ can then be obtained from Eq. (3.33). $C_{p,w}$ is the specific heat capacity of water, and m_w is the mass of the water volume of the condenser, that also can be calculated from user-defined time constant $\tau_{w,0}$ as shown in Eq. (3.34).

$$m_w C_{p,w}\frac{dT_{w,o}}{dt} = \dot{m}_w C_{p,w}(T_{w,i} - T_{w,o}) - \dot{Q}_{t,w} \tag{3.33}$$

$$m_w = \tau_{w,0}\dot{m}_{w,0} \tag{3.34}$$

3.2.1.5 Individual Server Power Model

Here we assume a data center with N_0 heterogeneous computer servers, and the server i has F_i different CPU frequencies $f_{i,1}$,..., f_{i,F_i}. Assuming $f_{i,1} \leq ... \leq f_{i,F_i}$, the set of the relative frequencies of the server i can be denoted by $\Phi_i = \{\frac{f_{i,1}}{f_{i,F_i}}, ..., \frac{f_{i,F_i}}{f_{i,F_i}}\}$. Thus the power of the individual server i can be calculated as:

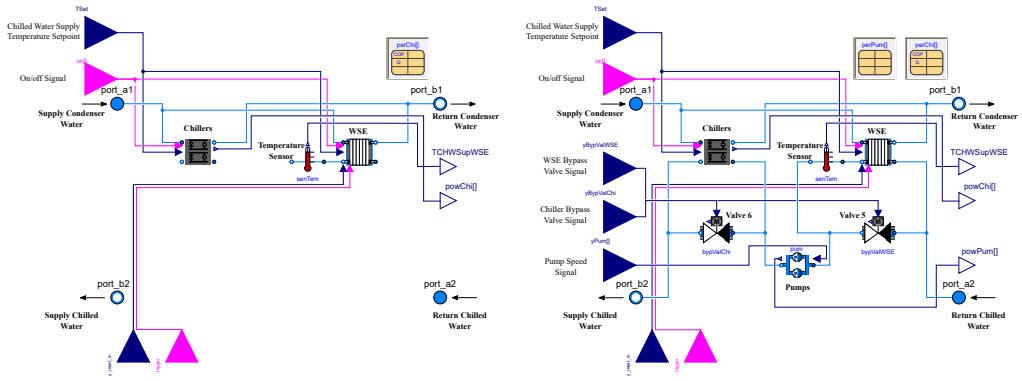
$$P_i = \alpha_i f_i(t)\lambda_i(t) + \beta_i, \tag{3.35}$$

where P_i, $f_i \in \Phi_i$, and λ_i are the power consumption, relative CPU frequency, and workload of server i, respectively. α_i and β_i are the coefficients that can be obtained through curve-fitting techniques.

3.2.2 Subsystem Models

Different subsystem models and their base classes that define the arrangement of chillers and WSEs are also built. The different subsystem models share the same base class as shown in Figure 3.6(a). The base class is built on a four-port fluid interface, representing the inlets and outlets for the chilled and condenser water, with instances of the chiller group model and the WSE group model. The connections of chillers and WSEs on the chilled water side are not declared in

the base class, because different subsystem models mentioned in Ref. [48] have different hydraulic configurations, such as chillers with integrated WSE on the load side, chillers with integrated WSE on the plant side, and chillers with nonintegrated WSE etc..



(a) Base class

(b) Chiller with integrated WSE on the load side

Figure 3.6: Modelica implementation of chillers and WSE subsystem

The different subsystems can then be modelled individually using a hierarchical approach. In this case, we first inherit the base class, then instantiate additional necessary equipment models, and finally add physical connections among components. For example, to model the subsystem where the WSE is integrated on the load side of a primary-only chilled water system (as shown in the dashed box of Figure 1.5(b)), we only need to extend the base class, add necessary instances such as bypass valves and pumps, expose model inputs and outputs, and finally connect them as in an actual system. Figure 3.6(b) shows the implementation of such a subsystem model based on the base class in Figure 3.6(a). On the left are the model inputs, including the on/off command, supply chilled water temperature setpoint, bypass valve position signal and pump speed signal from particular controllers. On the right are the model outputs, such as power from chillers and pumps.

This hierarchical modeling structure allows users to manage the complexity of large models, and to assemble system models as one would connect components in an actual system. This structure also facilitates debugging and verification of component models. For example, a lower-level model is first debugged and verified, and then instantiated in a higher-level model, which can

help identify modelling errors at the early stage of the model development.

3.2.3 Cooling Control Models

3.2.3.1 Cooling Mode Control

In both the chilled water and DX systems, a cooling mode control determines when to activate and deactivate the FC, PMC and FMC modes for the cooling system with economizers based on the system status and the environment. The cooling mode control is a supervisory control, and the output control signal will be taken as inputs by other equipment-level controllers as described in Section 3.2.3.2.

The cooling mode can be described as a finite-state machine. The cooling mode can transition from one state to another in response to some external inputs. For example, we can present a widely-used control strategy [113] for a chilled water system with integrated WSE using a state graph as shown in Figure 3.7(a). The chiller is switched on when

$$\Delta t_{chi,off} \geq \Delta t_1 \text{ and } T_{chw,sup,wse} > T_{chw,sup,set} + \Delta T_1 \text{ for } \Delta t_2, \tag{3.36}$$

and switched off if

$$\Delta t_{chi,on} \geq \Delta t_3 \text{ and } T_{chw,sup,wse} < T_{chw,sup,set} - \Delta T_2 \text{ for } \Delta t_4, \tag{3.37}$$

where $\Delta t_{chi,off}$ is the time of the chiller in Off status, $\Delta t_{chi,on}$ is the elapsed time since the chiller was switched on, Δt_1 to Δt_4 are time thresholds whose defaulted values are shown in Figure 3.7(a), $T_{chw,sup,wse}$ is the temperature of the supply chilled water downstream of the WSE, $T_{chw,sup,set}$ is the chilled water supply temperature set point, and ΔT_1 and ΔT_2 are the deadband temperature. The waiting time and dead band can prevent frequent short cycling.

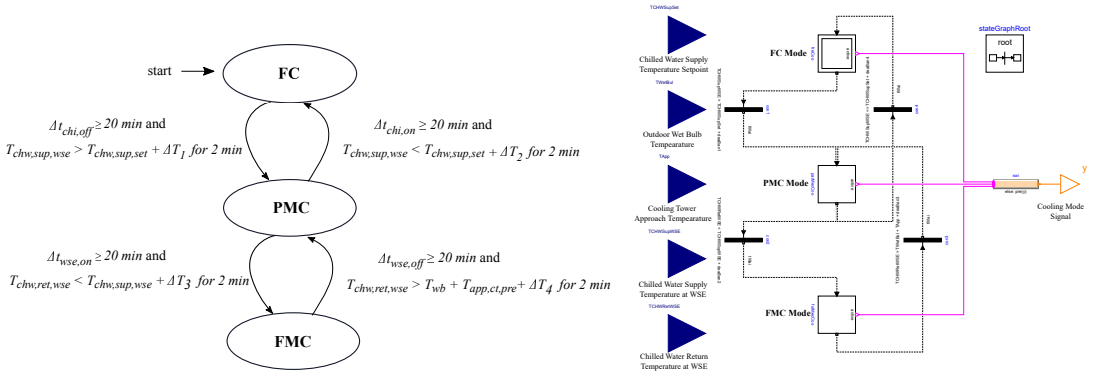
The WSE is enabled when

$$\Delta t_{wse,off} \geq \Delta t_7 \text{ and } T_{chw,ret,wse} > T_{wb} + T_{app,ct,pre} + \Delta T_4 \text{ for } \Delta t_8, \tag{3.38}$$

and is disabled when

$$\Delta t_{wse,on} \leq \Delta t_5 \text{ and } T_{chw,ret,wse} < T_{chw,sup,wse} + \Delta T_3 \text{ for } \Delta t_6, \tag{3.39}$$

where $\Delta t_{wse,off}$ is the elapsed time of the WSE in Off status, $\Delta t_{wse,on}$ is the elapsed time since the WSE was switched on, $T_{chw,ret,wse}$ is the temperature of the return chilled water upstream of the WSE, T_{wb} is wet bulb temperature of the outdoor air, $T_{app,ct,pre}$ is the predicted approach temperature of the cooling tower, ΔT_3 and ΔT_4 are the offset temperature, and Δt_5 to Δt_8 are time thresholds. In our application, we set $T_{app,ct,pre}$ as the nominal approach temperature in the cooling tower, although many other prediction algorithms can be used such as using a detailed cooling tower model [113] or engineering experience [118]. Figure 3.7(b) shows the Modelica implementation using the State Graph package.



(a) State graph (b) Modelica implementation

Figure 3.7: Cooling mode control for a chilled water system with integrated WSE

3.2.3.2 Equipment Control

Equipment-level control includes stage control and speed control of chillers, pumps and cooling towers. The stage control determines when and how many equipment are activated at a given time. The speed control regulates the speed of equipment such as cooling tower fans.

For the stage control of chillers, we implemented the following logic: If the cooling mode control outputs FC mode, then all the chillers should be commanded off. If the cooling mode control outputs PMC or FMC mode, at least one chiller should be active all the time. Then one additional chiller is commanded on when

$$Q_{ave} > Q_{up} + \Delta Q \text{ for } \Delta t_9, \tag{3.40}$$

and commanded off when

$$Q_{ave} < Q_{down} - \Delta Q \text{ for } \Delta t_{10}, \tag{3.41}$$

where Q_{ave} is the average cooling load in all the active chillers at the current time, Q_{up} and Q_{down} are the cooling load thresholds for staging up and down, respectively, and ΔQ is a deadband. The two conditions need to remain true for a predefined waiting time Δt_9 and Δt_{10}, respectively. The stage control was implemented in Modelica using the State Graph package.

For the speed control, here we take the cooling tower fans as an example. The cooling tower fan speed should be regulated differently in different cooling modes. One possible set of control logics is shown as follows.

- In the FC mode, the fan speed is controlled to maintain a predefined chilled water supply temperature downstream of the WSE.
- In the PMC mode, the fan speed is reset to 90%. Setting the speed to 100% can produce the condenser water as cold as possible and maximize the WSE output. However, with variable speed drives on the tower fans, changes from 90% to 100% do little to lower the condenser water temperature but increases the fan energy significantly [113].
- In the FMC mode, the fan speed is controlled to maintain the supply condenser water at its set point.

3.2.3.3 Valve Control

The transition among each cooling mode is achieved by manipulating the associated isolation valves and bypass valves. For example, in Figure 1.5(b), when the cooling system is in the FC mode, the isolation valves V_1 and V_2 in chillers, and V_5 for bypassing the WSE are closed. The isolation valves V_3 and V_4 in the WSE, and V_6 for bypassing the chillers are fully opened so that the chilled water can flow through the WSE, and then be delivered by the primary pumps to the CRAHs. In the PMC mode, V_1 and V_4 are fully open, and V_5 and V_6 are closed. In the FMC mode,

while V_3, V_4 and V_6 are closed, V_1, V_2 and V_5 are fully open to deliver the chilled water through the primary pumps, chillers, and then CRAHs.

3.2.4 System Templates

Templates for different systems are also provided. An example is shown in Figure 3.8, where the model of a primary-only chilled water system with an integrated WSE, its control system, boundary conditions, and post-processor is presented.

The boundary conditions for the cooling system are read from a weather data file. The cooling and control system are assembled by connecting the above-mentioned component and control models. The data center room model in the cooling system is simplified using a well-mixed volume, because the air flow management in the room is not the focus here. The cooling load is assumed to be constant during the simulation period. Post-processing provides an option to process the simulation results such as energy and control performances in the model. The simulation results of the system template model have been reported in Ref. [48].

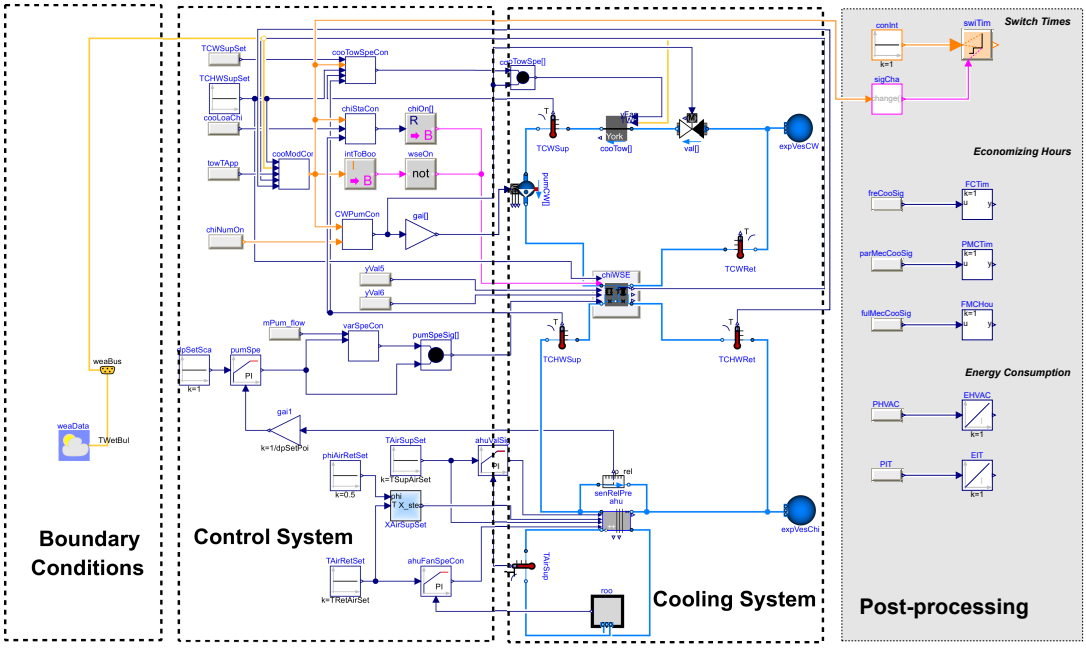


Figure 3.8: Modelica implementation of a primary-only chilled water system with an integrated WSE

3.3 Model Evaluation

Each component is verified in a simulation example, following the conventions of the Buildings library [140]. Taking advantages of the class hierarchy in the Buildings library and the object-oriented language Modelica, we built the data center package based on the base classes and ready-to-use component models. We validated the data center package using analytical verification and comparative testing. The former has also been used to validate all individual component models in the Buildings library. For example, the WSE model is validated by analytical verification, which compares its results with analytical solutions that are derived for certain steady-state boundary conditions. In addition, the CRAC model in Modelica is validated by comparing its simulation results with the same model in EnergyPlus.

3.4 Case Study 1: Conventional Data Center

This case study presents two scenarios to investigate the cooling system operation in a conventional data center located in Salem, Oregon, USA. This data center is powered by the power grid. The first scenario investigates the energy efficiency and control performance of the cooling system under normal operation (e.g. connected to grid), and the other one compares different operation strategies to explore the opportunities of effective operation of the cooling system under emergency situations (e.g. disconnected from grid and backup generators).

3.4.1 Description of Cooling and Electrical System

3.4.1.1 Cooling System

In the case study, the studied data center is cooled by a primary-only chilled water system with two identical chillers and one integrated WSE on the load side. The WSE is installed in parallel with chillers on the condenser water side. The design cooling load is 2,200 kW, which could be satisfied by two identical chillers in the FMC mode or one WSE in the FC mode. The number of variable-speed chilled water pumps, constant-speed condenser water pumps and variable-speed

cooling towers are equal to the number of chillers. One CRAH with one supply air fan delivers cold air to the room. The evaluation of the redundancy for the cooling system, such as a backup CRAH and redundant piping system, is not a purpose of this case study, therefore it is not considered here.

Dynamics in the cooling system models are represented using two methods: one is a lumped volume, parameterized by time constant or thermal mass, and the other is a signal filter that filters high frequency input signals. For example, the dynamics in the cooling coils are represented by a predefined time constant of 30s, and the thermal mass of the racks and the servers from Ref. [102] are added to the data center room model to calculate the thermal inertia inside the data center room. Furthermore, the dynamic behavior of the valves motors is represented by adding a second-order filter for the input position signals. For the control system, the room temperature is controlled at a set point of 25 °C by adjusting the fan speed in the CRAH. The supply air temperature is maintained at 18 °C by regulating the two-way valve on the waterside of the cooling coils. The chilled water supply temperature is set to be 6.5 °C under all cooling modes and load conditions.

3.4.1.2 Electrical System

The electrical system is modeled using the *Buildings.Electrical* package in the Modelica Buildings library. The electrical loads of the cooling equipment are represented by the balanced three-phase AC inductive resistances. The PDUs and PSUs are modeled by connecting existing converter and transformer models as in an actual physical system. The servers are represented as DC loads. The UPS is modeled as a battery storage that does not consider the voltage and thermal dynamics during the charging and discharging process in this case, and is sized based on the selected critical equipment (IT equipment and supply air fan in the CRAH). The charging and discharging of the battery are controlled by the following logic shown in Figure 3.9. The state of charge (SOC) of a battery is its available capacity expressed as a percentage of its rated capacity. Knowing the SOC gives the user an indication of how long a battery will continue to perform before it is depleted. As

it is not desired to deplete or overcharge the battery, the SOC of the battery should be kept within proper limits. Because the charging and discharging dynamics are not the purpose of this case study, we set the lower and upper bound of the SOC (SOC_l, and SOC_u) as 0 and 1, respectively. When connected to the utility or backup generators, the UPS is charged at a reference rate until being charged to SOC_u. When disconnected, the UPS discharges power to support critical equipment at a discharging rate of the minimum between the required rate $P_{dis,req}$ and the reference rate $P_{dis,ref}$. The potential voltage fluctuation during charging and discharging is not considered in this case.

```
1: if Connected then
2:     if SOC(t) < SOC_u then
3:         P_cha(t) ← P_cha,ref
4:     else
5:         P_cha(t) ← 0
6:     end if
7: else
8:     if SOC(t) < SOC_l then
9:         P_dis(t) ← 0
10:    else
11:        P_dis(t) ← min(P_dis,req, P_dis,ref)
12:    end if
13: end if
```

Figure 3.9: Pseudo codes of UPS charging and discharging control

3.4.2 Scenario 1: Normal Operation

Energy efficiency of the data center is considered as an important goal during normal operation. To quantify the energy efficiency, the Power Utilization Effectiveness $PUE = \frac{\int_{t_0}^{t_1} P_{feeder} dt}{\int_{t_0}^{t_1} P_{IT} dt}$ is used, where P_{feeder} is the total power that are delivered into the data center for the IT equipment and all their supporting infrastructure, P_{server} is the power used only by the servers, and t_0 and t_1 is the start and the end time for calculating the PUE. The closer the PUE is to 1, the more efficient the data center is. In this section, we investigate the different energy performances under different part load ratios (PLRs) of the cooling load in the data center room. The considered PLRs are

0.25, 0.50, 0.75 and 1.00, which represents the growing occupancy in the data center. The control settings for the cooling system in all PLRs are the same as the design condition as described in Section 3.4.1.1.

3.4.2.1 Simulation Results

Under design cooling load condition (PLR = 1.00), the breakdown of the annual electricity usage of the cooling system is shown in Figure 3.10. For the chilled water system with WSE, the economizing time, that is the period when the economizers are activated to pre-cool or fully cool the loads, is about 42% of the whole year (Figure 3.11). Because of the economizer operations, the fan in the CRAH is the major energy consumer, which takes up about 50.9% of total annual cooling energy. Chillers, pumps and cooling towers use 22.8%, 17.1% and 9.2%, respectively.

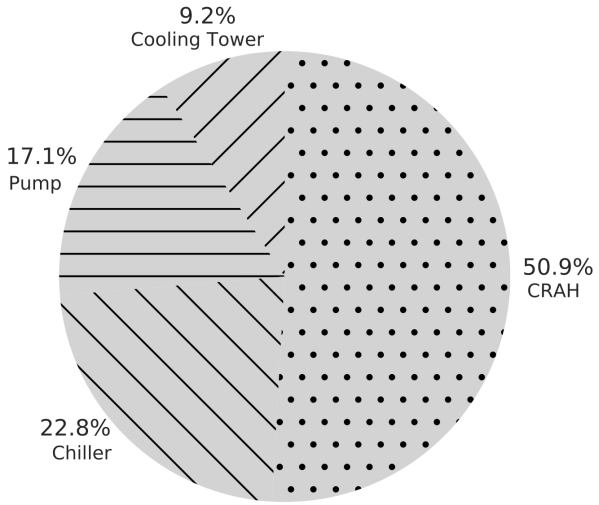
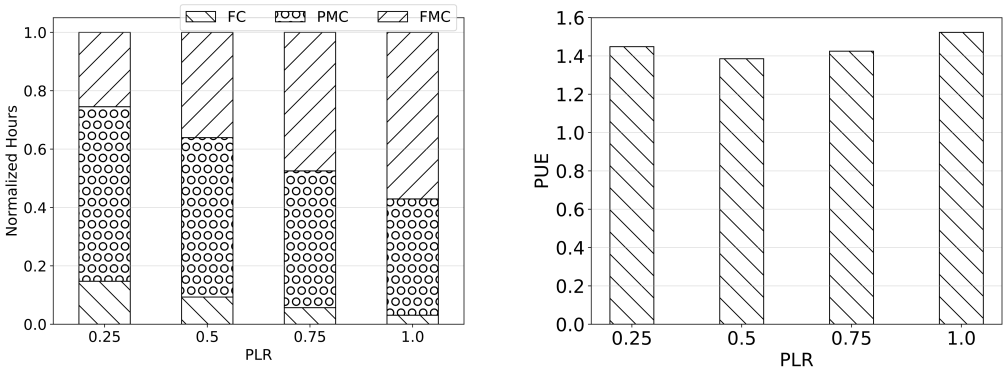


Figure 3.10: Breakdown of the electricity usage of the cooling system at PLR = 1.00

Figure 3.11(a) illustrates the normalized operation time of each cooling mode under different PLRs. As the PLR increases, the time when the cooling system stays in the FMC mode increases, and the time of the FC mode decreases. The time when the WSE is enabled decreases as the

PLR increases. The cooling mode controller described in Section 3.2.3.1 takes as inputs weather conditions and temperatures in the chilled water loop and the condenser water loop. Although the same weather file is used under different PLRs, the return chilled water temperatures are different. For example, the return chilled water temperature at PLR = 0.25 is higher than that at PLR = 1.00. Thus, condition (3.38) is faster to be triggered at PLR = 0.25 compared to PLR = 1.00, and hence the cooling system operates longer in the FC and the PMC mode at lower PLRs.

Figure 3.11(b) describes the relationship between PUEs and different PLRs. Among the four PLRs, the lowest PUE is 1.39 at PLR = 0.50, and hence maximum efficiency is achieved at part loads.



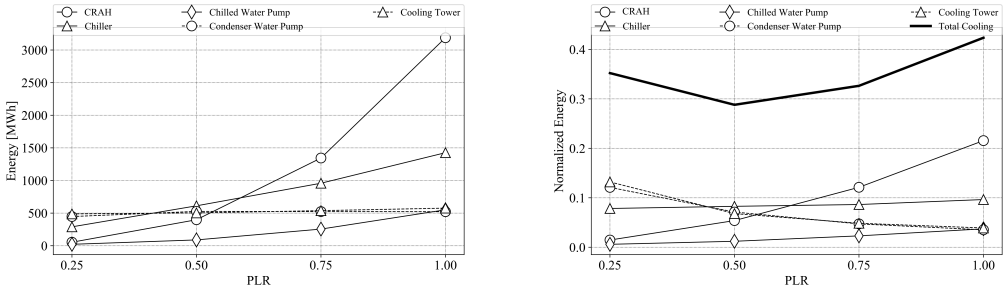
(a) Normalized Hours (b) PUE

Figure 3.11: Operational status at different PLRs

Figure 3.12(a) shows the detailed energy consumption by each cooling component at different PLRs. The energy usage by the CRAH has an approximately cubic relationship with the PLR. The reason is when the supply and room air temperature control setpoints are not changed as the PLR changes, the speed of the fan in the CRAH is linear to the PLR in the room, and the fan power has a cubic relationship with its speed. Similar profile can be observed in the chilled water pumps. The energy consumption of the chiller has a weak quadratic relationship with the PLR, which is determined by the performance curves of the chillers and the hours of the PMC and FMC modes. Since the condenser water pumps have constant speed, the energy consumption is almost constant

at PLR = 0.5, 0.75 and 1.00. A difference can be observed between PLR = 0.25 and other PLRs, because only one condenser water pump is activated during the FMC mode when the PLR is 0.25, while two condenser water pumps are needed at other larger PLRs during the FMC mode. As for the cooling towers, the annual energy increases as the PLR increases, although the relative increase, compared with the fan in the CRAH, is small. The major reason is that the cooling system runs at the PMC mode during the most time of the year especially when the PLR is low, and the speed of the cooling tower fans in the PMC mode is set to 90% all the time.

Figure 3.12(b) plots the normalized energy for each cooling component divided by the current cooling load. The energy efficiency of the cooling tower fans and condenser water pumps increases as the PLR increases, while the opposite trends happen in the CRAH fan, chillers and chilled water pumps. The cooling system efficiency as a whole by combing all the cooling equipment is highest at PLR = 0.50. At that PLR, to address 1 kW of cooling load, the cooling system needs about 0.29 kW electricity.



(a) Energy consumption (b) Normalized energy consumption

Figure 3.12: Detailed energy consumption in the cooling equipment

3.4.3 Scenario 2: Emergent Operation

The priority of emergency operation is to keep the data center safe. Thermally, safety means the heat generated by IT equipment can be removed timely to avoid the temperature of the IT equipment exceeding the maximum safety limit. Operation during emergency situations such as a blackout aims to maximize the use of the UPS by only powering critical equipment until backup

generators are online. Given the capacity of the UPS, the selection of critical and noncritical equipment has significant influence on the survival time. Typically, in a chilled water system with WSE, critical equipment are the IT equipment and the fan in the CRAH. However, there may be opportunities to cool the data center with the UPS by considering some other cooling equipment as critical equipment when the outdoor air is cold enough to activate the WSE, especially when the data center is operating under part load. This scenario studies the impact of the selection of critical equipment in a chilled water system with WSE during emergency mode on the thermal and electrical performance in a data center under different cooling load and outdoor conditions.

3.4.3.1 Problem Formulation

The selection of the critical equipment in order to provide a thermally reliable environment even during the emergent gap can be formed as an optimization problem shown in Eq. (3.42).

$$\begin{aligned} & min\ E(s) = \int_{t_1}^{t_2} P(s,t)dt \\ s.t. \quad & T_{room}(s,t) \leq T_{room,high} \\ & 0 < SOC(s,t) \leq 1, t \in [t_1, t_2) \\ & SOC(s,t) \geq 0, t = t_2 \end{aligned} \tag{3.42}$$

The optimization goal in this case study is to minimize the usage of the UPS power E during the gap by choosing the best operation strategy s, although other goals can also be considered. One constraint is that during the gap $t \in [t_1, t_2]$, the data center room temperature should not exceed a high limit $T_{room,high}$, because high temperature usually de-rates the power and IT equipment [4]. The other constraint is that the UPS must be able to support IT equipment during the gap, which means the SOC must be larger than zero before reaching the end of the gap.

In this study, we assume that the power grid fails at t_1=14:00, and the gap ends at t_2=14:15. The high temperature limit is set to 35 °C. The power $P(s,t)$ is simulated using the Modelica models, and is integrated over t_1 to t_2. Three strategies (s_1,s_2 and s_3) for the selection of critical equipment are considered in this study, as shown in Table 3.1. The optimization problem is solved

Table 3.1: Operation strategies for emergency situations

Index	Strategy
s_1	The IT equipment and fan in the CRAHs are powered by the UPS during the 15-minute gap.
s_2	The IT equipment and all the cooling equipment other than the chillers are powered by the UPS during the 15-minute gap.
s_3	When the SOC in the UPS is greater than 0.5, then activate s_2. Otherwise activate s_1.

by exhaustively simulating and comparing the three operation strategies. The same optimization problem is also formulated and solved for different PLRs under different cooling modes. The results are detailed in the next section.

3.4.3.2 Simulation Results

(1) FC mode

The recommended emergent operation strategies for different PLRs under the FC mode is summarized in Table 3.2. Detailed explanations are illustrated as follows.

Table 3.2: Recommended operation strategies for emergency situations in the FC mode

PLRs	s_1	s_2	s_3
0.25	x		
0.50		x	x
0.75		x	
1.00	x		

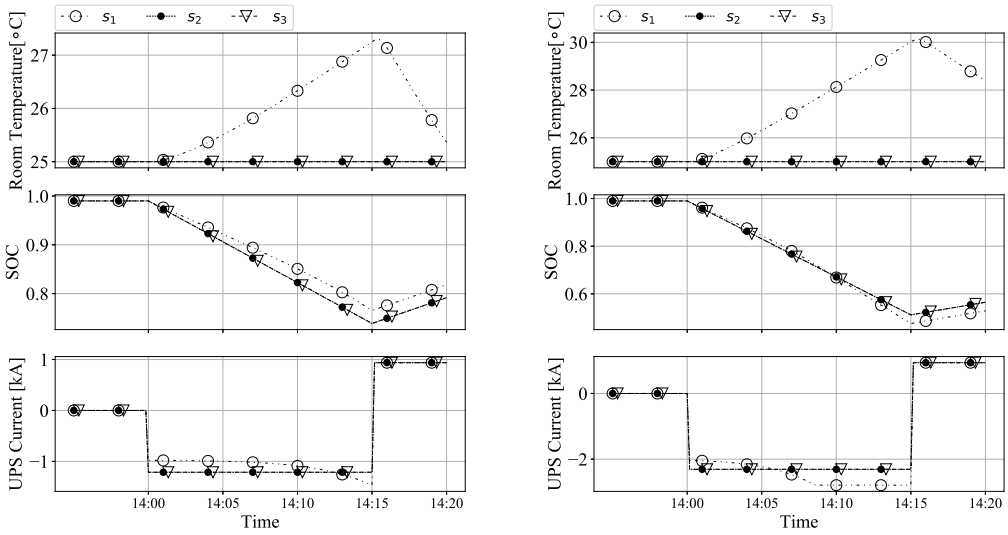
When the PLR is 0.25, s_1 performs better than s_2 and s_3 because it can keep the temperature within the high limit, and consumes the least energy from the UPS. The room temperature increases by 2.2 °C at the end of the power grid failure by utilizing s_1. The temperature rise is caused by the deactivation of cooling sources (chillers and economizers). On the contrary, the room temperature can be kept at 25 °C in s_2 and s_3 because the WSE is enabled to cool the room at the cost of more power drawn from the UPS. The SOC after 15 minutes in s_1 is 0.76, while that in s_2 and s_3 is 0.72.

The SOC in s_2 and s_3 is the same because the SOC in this case is greater than 0.5 all the time, which makes s_3 the same as s_2. Compared with s_1, s_2 and s_3 consume less fan energy, because in s_2 and s_3 the fan speed is kept at around 0.25, while in s_1 the fan speed ramps up to around 0.85 in the 15 minutes because the fan needs to deliver more air to reduce the room temperature. For the discharging current (negative) in the UPS, s_1 discharges slower than s_2 and s_3 at the beginning. However, as the fan needs more energy in s_1, the UPS need discharge faster. Although s_1 requires more fan energy, the increased fan energy is still less than the energy required by the activation of more cooling equipment in s_2 and s_3.

When the PLR is 0.5, s_2 and s_3 have the same performance, and perform better than s_1, because they can maintain a lower room temperature and consume less energy from the UPS, although they power more equipment. In s_2 and s_3, the fan speed is maintained at 0.5 as in the normal operation, while in s_3 the fan speed ramps up to 1 during the 15 minutes, which eventually leads to faster discharging and more energy consumption in the end.

When PLR is 0.75, s_2 performs better than s_1 and s_3, because it can lead to the lowest room temperature and consume the least energy. In s_1, the room temperature at the end of the gap increases to 33 °C. In s_2, the room temperature is still maintained at 25 °C. In s_3, the room temperature is kept at 25 °C before the first 10 minutes when the SOC of the UPS is greater than 0.5, and increases to 27 °C at the end of the failure. s_1 consumes more energy than s_2, because the fan in s_1 runs at the full speed during most of the gap. s_3 consumes slightly greater power than s_2 after the first 10 minutes because of the higher fan speed after deactivating the cooling equipment.

When PLR is 1.00, s_1 is the only strategy that can help the IT equipment survive 15 minutes, although the room temperature in the end reaches about 36 °C. For s_2, the room temperature is kept at 25 °C, but the UPS can only last about 13.5 minutes, which means the IT equipment has to be shut down for 1.5 minutes until the backup generators are on. Similarly, the UPS in s_3 can only last for 14.5 minutes. Though s_2 and s_3 can keep the data center room at a low temperature when data center is fully loaded, the reliability of the data center room is compromised.



(a) PLR = 0.25

(b) PLR = 0.50

(c) PLR = 0.75

(d) PLR = 1.00

Figure 3.13: Comparison of blackout in the FC mode at four PLRs for different strategies

(2) FMC Mode

During the FMC mode, the waterside economizer is activated for emergency operation. In the 15 minutes, the outdoor air dry bulb temperature is around 28.5 °C, and the wet bulb temperature is around 19.3 °C. The economizer can take some heat out when the condenser water temperature is lower than that in the chilled water loop. Simulation results show that s_1 is the best strategy in the FMC mode for all PLRs.

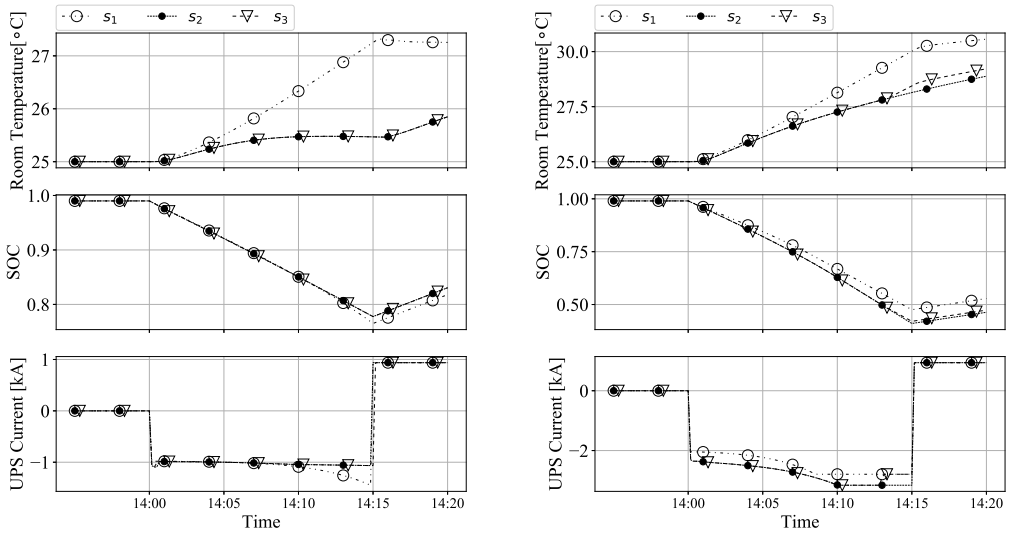
As shown in Figure 3.14, s_1 outperforms s_2 and s_3 when PLR = 0.25, 0.50 and 0.75, because it consumes the least energy while maintaining the room temperature within the high limit. When PLR = 1.00, s_1 is the only strategy that depletes the UPS after the gap.

3.5 Case Study 2: Renewable Data Center

Data centers with renewable energy become increasingly attractive, especially for small and medium ones as their installations are cheaper and smaller. Renewable data centers can provide not only decent economical savings, but new opportunities to improve reliability due to the additional generation. This case study aims to investigate how the renewable data center behave under emergent situations with a special focus on how it can extend the survival time of the UPS.

The studied system is the same as case study 1, except that the data center is connected to an additional PV system. To quantify the impact of the PV to the data center, we introduce a penetration factor r_{pv} that defines the ratio of the nominal power of the PV system to the nominal power of the data center. If r_{pv} is 0, then the data center cannot receive renewable power, and if r_{pv} is 1, the renewable energy can meet all the electrical demands in the data center at the nominal condition. To explore the survival time during emergency situation, we manually introduce a blackout at 12:00 pm, which lasts for 4 hours. The parameters r_{pv} and PLR are swept to show how the penetration of the PV system can affect the survival time of the UPS under different PLRs by using strategy s_2.

Figure 3.15 shows the lasting time of the UPS during FC and FMC mode for an emergency



(a) PLR = 0.25

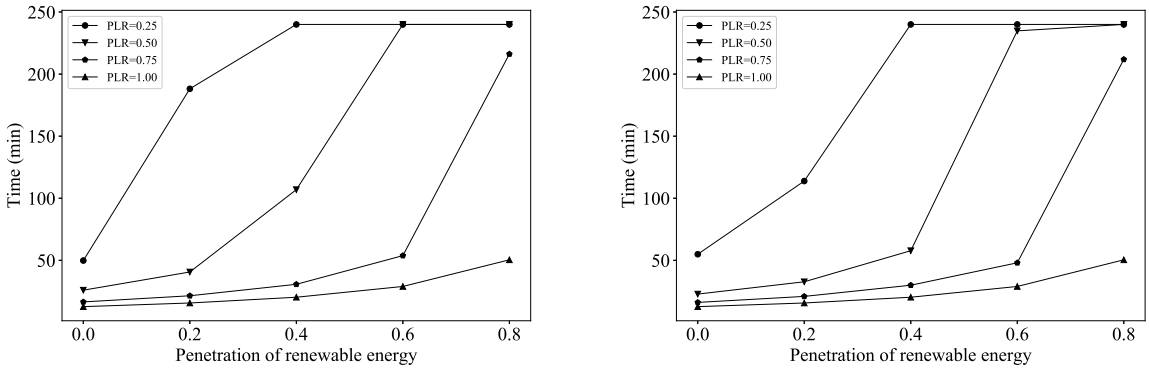
(b) PLR = 0.50

(c) PLR = 0.75

(d) PLR = 1.00

Figure 3.14: Comparison of blackout in the FMC mode at four PLRs for different strategies

situation. Generally, as the penetration of renewable energy increases, the lasting time increases as well because more power can be drawn from the renewable sources and less from the UPS. For example, as shown in Figure 3.15(b), when the data center is running at a PLR of 0.75 in PMC mode, the UPS can only last about 16 minutes at no penetration, and about 211 minutes at a penetration of 0.8. Note that the maximum lasting time is 240 minutes because we only simulate a blackout for 4 hours. Therefore, the penetration of renewable energy in a data center can extend the lasting time of the UPS, thus increasing the data center reliability. The extra time may also be utilized when there are emergent system maintenances needed to be performed onsite.



(a) FC mode

(b) FMC mode

Figure 3.15: Comparison of UPS lasting time in a renewable data center at different cooling modes, PV penetration factors, and PLRs

The relationship between the lasting time and penetration of renewable energy is nonlinear. The nonlinearity is caused by the difference between the penetration and the PLR. If the penetration is higher than the PLR, the lasting time will increase significantly because the generation of renewable energy can meet the most or entire electrical demand in the data center, which thus requires little power from the UPS.

Figure 3.16 compares the detailed SOC of the UPS and the current flows in two data centers with a penetration of 0 and 0.6. Without renewable resources, the UPS depletes after around 16 minutes. When the penetration of renewable resources increases to 0.6, the survival time can be extended to about 48 minutes. The SOC decreases non-smoothly when the PV system is penetrated,

because of the oscillations of the PV current and the UPS current. The sharp changes in the PV current are due to the movement of clouds.

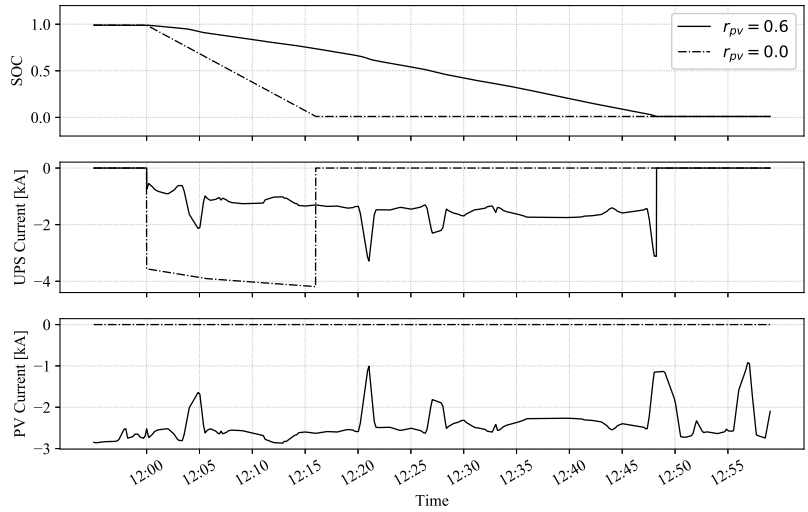


Figure 3.16: Comparison of detailed SOC and currents with and without PV

3.6 Summary

This chapter presents an open source, equation-based, object-oriented Modelica package for GEDCs. This package includes major cooling component models, control logic, subsystem models and system templates for both chilled water system and direct expansion system, which are designed to support rapid virtual prototyping. The case studies show that this package is able to perform various analysis, including detailed analysis of energy efficiency and control performance in normal operation, as well as emergency operation. By integrating with the electrical models, detailed electrical analysis is also supported for both conventional and renewable data centers.

Chapter 4

Validation of Numerical Testbed

Based on:

Yangyang Fu, Wangda Zuo, Michael Wetter, Jim W. VanGilder, Xu Han, David Plamondon, "Equation-based object-oriented modeling and simulation for data center cooling: A case study", Energy and Buildings, Volume 186, 2019, Pages 108-125, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2019.01.018.

In the previous chapter, an equation-based object-oriented modeling and simulation testbed is developed for GEDCs. This chapter aims to validate the developed testbed and provide new solutions to harness the demand flexibility of energy efficiency. The validation is conducted using measurement data from an actual data center located at the University of Massachusetts Medical School in Massachusetts, United States. This chapter is organized as follows: Section 4.1 gives a detailed description of the analyzed cooling and control systems, including the system configurations and different control strategies. Section 4.2 shows the management of the complex, large system model through a hierarchical modeling approach. The Modelica models are then calibrated using on-site measurement data in Section 4.3. In Section 4.4, we first identify several energy and control related issues in the baseline system through an annual simulation. Then we propose different energy efficiency measures (EEMs) to address the identified issues. A sequential search technique is applied to identify the combination of the most cost-effective EEMs in terms of energy savings and life cycle cost (LCC). After that, an optimization of the supply air temperature for the best EEMs is performed to evaluate the energy saving potentials. Conclusions are presented in Section 4.5.

4.1 System Description

The data center analyzed operates 24 hours per day, 365 days per year. The data center room has a floor area of 687 m^2 with a white space height of 3.35 m. The room contains 138 IT racks and 12 floor-mounted power distribution units. This case study only focuses on the cooling and control system, and the room-side air distribution management is not considered.

4.1.1 Cooling System

A primary-only chilled water system with airside economizers (ASEs) is used to provide cooling for the data center room, as shown in Figure 4.1. The size of detailed components is listed in Table 1. The current cooling load of the data center is about 316 kW. Two identical water-cooled chillers with a design coefficient of performance (COP) of 5.8 work in a Lead/Lag configuration to equalize their runtime. Each chiller has two variable-speed compressors. Two identical cooling towers with variable-speed fans eject the heat from the condenser water loop to the environment. Two chilled water pumps operate with variable speed drives, while two condenser water pumps work at a constant speed. Two Air Handler Units (AHUs) provide cool air to the data center white space. Each AHU consists of an array of 12 variable-speed supply air fans arranged in a parallel-flow configuration. The cool supply air is delivered to cold aisles through an underfloor plenum. The hot IT exhaust air is directed into open hot aisles, then enters a ceiling plenum, then mixing box, and finally returns to the AHUs. When the weather conditions allow, the ASEs are activated to mix the cold outdoor air and warm indoor air to provide precooling or free cooling. The activation and deactivation of ASEs are controlled by a cooling mode controller discussed in Section 4.1.2.1.

4.1.2 Control System

The control system is composed of a system-level cooling mode control and an equipment-level control with various controllers, as shown in Figure 4.2. The solid lines show the hierarchical

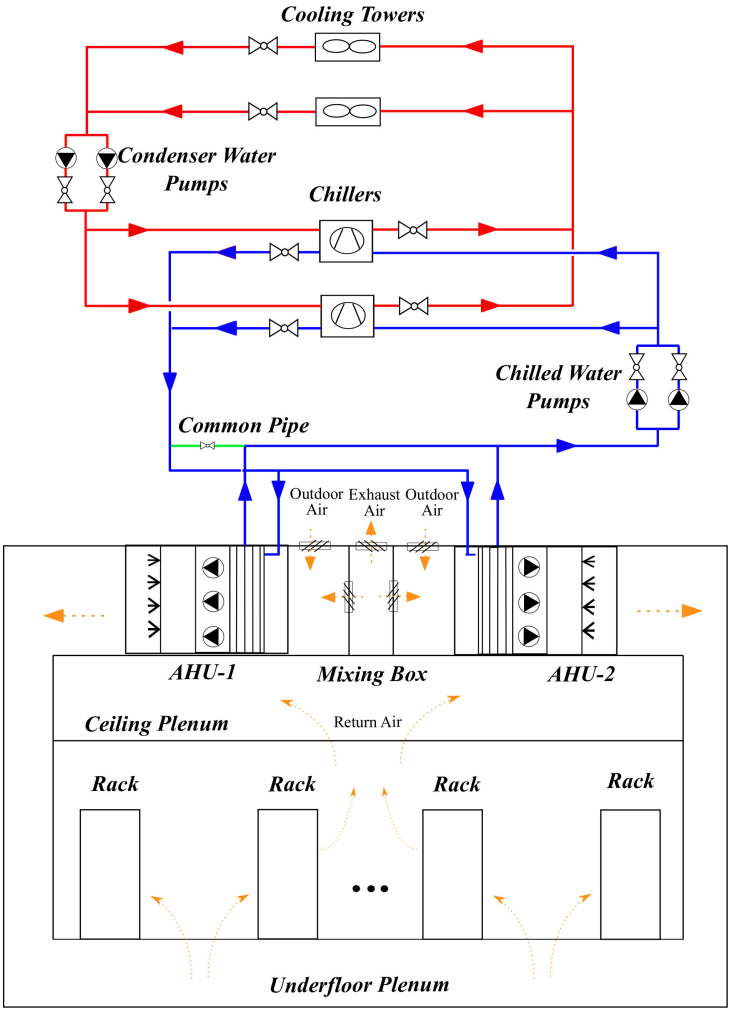


Figure 4.1: Schematic drawing of the cooling system in the data center

Table 4.1: Nominal information of components in the cooling system

Equipment	Qty.	Nominal Equipment Information		Unit	Value
AHU	2	Cooling Coil	Air Flowrate	m^3/s	39.2
			Cooling Capacity	kW	696
			Sensible Heat Ratio	-	0.99
			Water Flowrate	kg/s	0.025
		Heating Coil	Qty.	-	4
			Power	kW	31.2
		Steam Humidifier	Qty.	-	4
			Capacity	kg/s	0.019
		Fan	Qty.	-	12
			Head	Pa	622
			Power	kW	3.42
			Flowrate	m3/s	3.26
Chiller	2	Nominal Capacity		kW	774
		Design COP		-	5.8
		Evaporator	Flowrate	m^3/s	0.028
			Design Outlet Temperature	°C	10
		Condenser	Flowrate	m^3/s	0.026
			Design Inlet Temperature	°C	29.4
		Compressor	Number	-	2
			Speed Type	-	Variable Speed
			Power	kW	67
Chiller Water Pump	2	Head		mH2O	41
		Power		kW	12
		Flowrate		m^3/s	0.028
		Speed Type		-	Variable Speed
Condenser Water Pump	2	Head		mH2O	29.5
		Power		kW	8
		Flowrate		m^3/s	0.026
		Speed Type		-	Constant Speed
Cooling Tower	2	Nominal Capacity		kW	893
		Design Approach Temperature		°C	4.4
		Number of Cells		-	1
		Number of Fans		-	1
		Fan Speed Type		-	Variable Speed

relationship between different controls. The dashed arrows describe the actual control signal flow between different controls. Based on the operational status and outdoor air conditions, the cooling mode controller selects a particular cooling source from the three available choices: chillers only, ASEs only, or both chillers and ASEs. The signal from the cooling mode controller is then sent to the equipment-level controllers to determine the appropriate operating point of individual equipment.

4.1.2.1 System-level Control

The chilled water system with ASEs can operate in three cooling modes to provide cooling for the data center: (1) Free Cooling (FC) mode, where only ASEs are activated; (2) Partial Mechanical Cooling (PMC) mode, where chillers and ASEs work simultaneously; and (3) Fully Mechanical Cooling (FMC) mode, where only chillers are utilized. As the cooling system has to operate 24 hours per day, 365 days per year, the system off state is not considered. The staging among the different cooling modes is controlled by prescribed transition conditions, which is described by a state graph shown in Figure 4.3.

The transition between FC and PMC mode is determined by air temperature setpoint in the underfloor plenum $T_{floor,set}$ and outdoor air conditions, such as dry bulb temperature, $T_{OA,db}$, and dew point temperature $T_{OA,dp}$.The cooling system switches from FC to PMC mode, when

$$T_{OA,db} > T_{floor,set} + \Delta T_1 \; and \; T_{OA,dp} > T_{OA,dp,low} + \Delta T_2 \tag{4.1}$$

and from PMC to FC mode when

$$T_{OA,db} < T_{floor,set} - \Delta T_1 \; or \; T_{OA,dp} < T_{OA,dp,low} - \Delta T_2 \tag{4.2}$$

where $T_{OA,dp,low}$ is the low cutoff limit for $T_{OA,dp}$, and ΔT_1 and ΔT_2 are temperature dead band settings.

The transition between PMC and FMC mode is governed by $T_{OA,db}$, $T_{OA,dp}$, and data center return air temperature $T_{RA,db}$. The cooling system switches from PMC to FMC mode when the

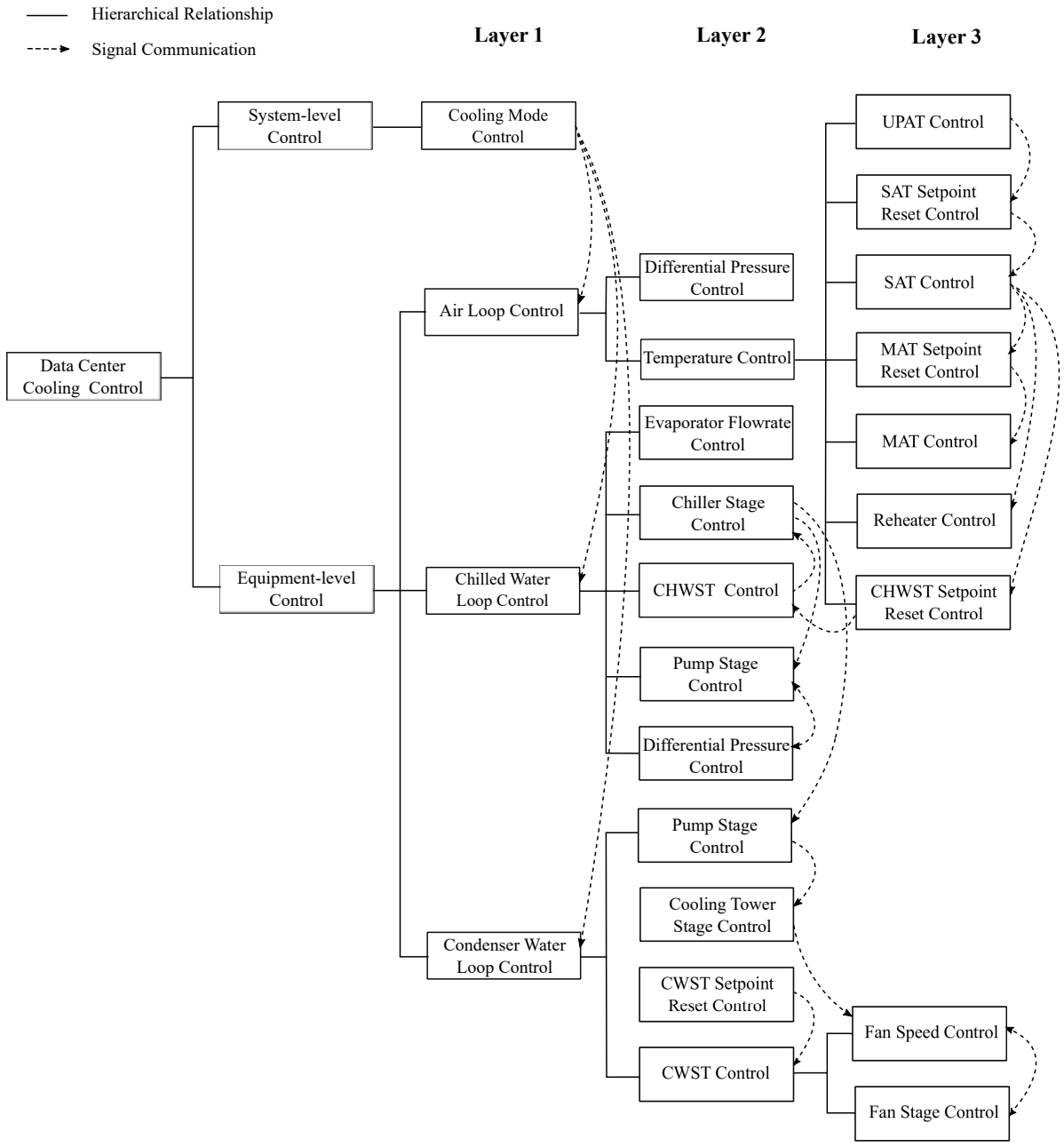


Figure 4.2: Structure of the data center cooling control

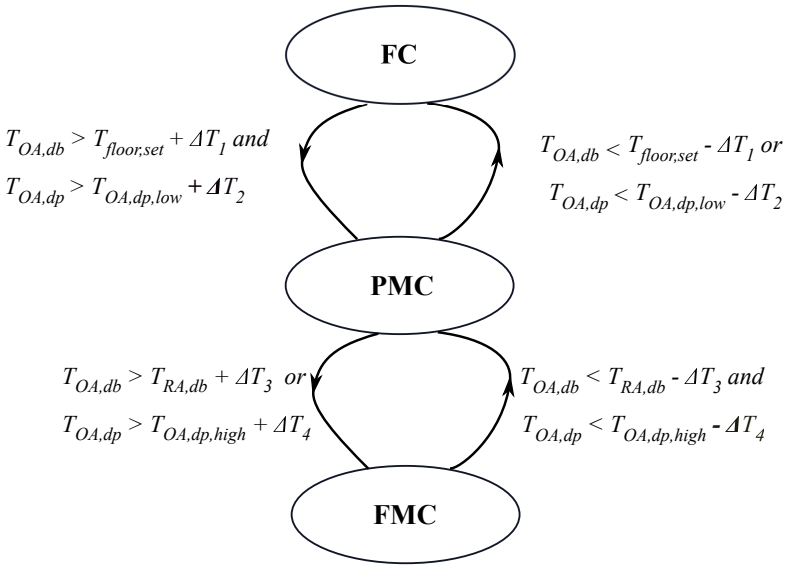


Figure 4.3: State graph of the cooling mode controller

following conditions are triggered:

$$T_{OA,db} > T_{RA,db} + \Delta T_3 \; or \; T_{OA,dp} > T_{OA,dp,high} + \Delta T_4 \tag{4.3}$$

and from FMC to PMC mode, when the following conditions are met:

$$T_{OA,db} < T_{RA,db} - \Delta T_3 \; and \; T_{OA,dp} < T_{OA,dp,high} - \Delta T_4 \tag{4.4}$$

where $T_{OA,dp,high}$ is the high cutoff limit for $T_{OA,dp}$, and ΔT_3 and ΔT_4 are temperature dead band settings. The $T_{OA,dp}$, $T_{OA,db}$, and $T_{RA,db}$ are read from measured data. The $T_{floor,set}$, $T_{OA,dp,high}$ and $T_{OA,dp,low}$ are set to 22.2 °C, 12.75 °C, and 11.65 °C, respectively. The dead bands ΔT_1 and ΔT_3 are set to 1.1 °C, and ΔT_2 and ΔT_4 are 0.55 °C. To prevent short-cycling, all the conditions must remain true for 2 minutes before switching to next state.

4.1.2.2 Equipment-level Control

As shown in Figure 4.2, the equipment-level control consists of multiple layers with complicated interactions among different controllers. Layer 1 is designed to coordinate the operation of the three major fluid loops of the cooling system: air, chilled water, and condenser water. Each loop has multiple groups of different controls in Layer 2. For instance, the air loop has two groups of controls. One is to control the differential pressure in the underfloor plenum to ensure that a reasonable amount of air passes through the perforated tiles to the data center room. The other is designed for the temperature control. Some groups in Layer 2 also have multiple controllers (Layer 3) dedicated to different control objectives. For example, the condenser water supply temperature (CWST) control in the condenser water loop consists of controls for cooling tower fan staging and fan speed. The details are explained in a top-down approach from Layer 1 to Layer 3 as follows.

Air Loop Control

Air loop control includes the control for the underfloor plenum and AHUs. The average static pressure in the underfloor plenum is controlled at a setpoint of 12.4 Pa by modulating the AHU fan

speed. The AHUs run all the time. The fans in each AHU are equipped with variable frequency drives and they are controlled to run at the same speed.

The temperature control in the air loop determines the supply air temperature (SAT) setpoint for AHUs, mixed air temperature (MAT) setpoint, outdoor air damper position, chilled water supply temperature (CHWST) setpoint, and control signals for the reheaters in the AHUs. The control strategies and interactions are schematically shown in Figure 4.4. The underfloor plenum air temperature (UPAT) is maintained at its setpoint $T_{floor,set}$ = 22.2 °C by resetting the SAT setpoint for AHUs in a range from 15.6 °C to 23.3 °C using:

$$f = \begin{cases} y_{ref,1}, & u \leq u_{ref,1} \\ (u - u_{ref,1})\frac{y_{ref,2} - y_ref,1}{u_{ref,2} - u_{ref,1}} + y_{ref,1}, & u_{ref,1} < u < u_{ref,2} \\ y_{ref,2}, & u \geq u_{ref,1} \end{cases} \tag{4.5}$$

where u and y are input and output signals respectively. The $u_{ref,1}$, $u_{ref,2}$, $y_{ref,1}$, and $y_{ref,2}$ are predefined reference values. In this case, u is the output of a proportional-integral-derivative (PID) controller (PID-1) and $u_{ref,1} = 0$, $u_{ref,2} = 1$, $y_{ref,1}$ = 23.3 °C, and $y_{ref,2}$ = 15.6 °C. It is worth mentions that (4.5) is also used by other controllers in Figure 4.4 but with different reference values for both input and output signals.

Using the reset SAT setpoint and measured SAT, two PID controllers (PID-2 and PID-3) are adopted to control the SAT for AHU-1 and AHU-2, respectively. The output signal y_2 and y_3 from the two PID controllers, ranging from 0 to 1, are then used in different control strategies under different cooling modes.

- In the FC mode, the SAT is maintained at its setpoint by adjusting the MAT setpoint. The maximum of the output signals y_2 and y_3 is used to reset the MAT setpoint within a range of 14.4 °C to 25.3 °C through (4.5). The MAT is then maintained at its setpoint by adjusting the outdoor air dampers through a PID controller (PID-4).

- In the PMC and FMC modes, the system will either reset the CHWST setpoint or activate

reheaters to maintain the SAT. To reset the CHWST setpoint, the output signals y_2 and y_3 are mapped to the CHWST setpoint within the range of 7.8 °C to 12.2 °C. The minimum of the mapped setpoints $CHWST_{set,1}$ and $CHWST_{set,2}$ is then sent to the chillers as the CHWST setpoint. For the reheaters, y_2 and y_3 are mapped to a control signal ranging from 0 to 1 in order to adjust the power of reheaters in AHU-1 and AHU-2, respectively. Take AHU-1 as an example. The reference values in the CHWST setpoint reset control are set to $u_{ref,1} = 0.4$, $u_{ref,2} = 1$, $y_{ref,1} = 12.2$ °C and $y_{ref,2} = 7.8$ °C. In the reheater control, they are set to $u_{ref,1} = 0$, $u_{ref,2} = 0.4$, $y_{ref,1} = 1$ and $y_{ref,2} = 0$. When the output signal y_2 of PID-2 is less than 0.4, the CHWST setpoint reset control is deactivated, and the reheater control is activated. Reverse actions are triggered when y_2 is greater than 0.4.

Chilled Water Loop Control

Chilled water loop control is composed of controls for the chillers and chilled water pumps. At the current cooling load, only one chiller is needed when FMC or PMC mode is activated. The chilled water pumps are set up to run one pump per chiller. The speed of the chilled water pumps is modulated by a PI controller to maintain a constant pressure difference of 206 kPa between the inlet and the outlet of the chiller evaporators. The bypass valve in the common leg is regulated by a PI controller to maintain a constant flowrate through the evaporators.

Condenser Water Loop Control

Condenser water loop control includes controls for the condenser water pumps and cooling towers. The condenser water pumps and cooling towers are staged based on the number of operating chillers: one condenser water pump and cooling tower is commanded on if one chiller is required. The CWST setpoint is reset from 21.1 °C to 29.4 °C as the outdoor air wet bulb temperature increases from 17.2 °C to 25.6 °C using the mapping algorithm in (4.5).

The speed and number of operating fans in cooling towers are manipulated to control the CWST at its setpoint. The fan speed is adjusted by a PI controller to reduce the difference between the CWST and the setpoint, and the number of working fans is determined as follows:

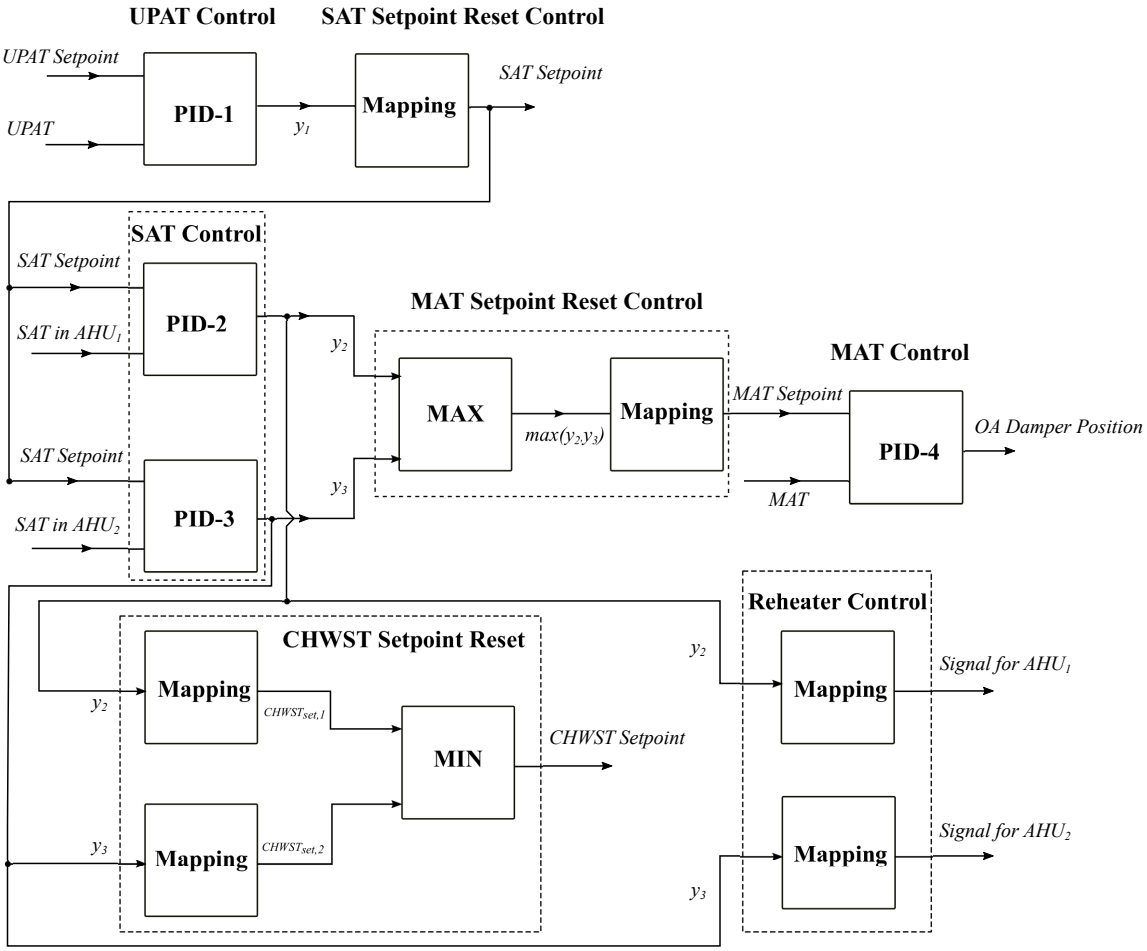


Figure 4.4: Temperature control for the air loop

One additional fan is switched on if

$$T_{cws} > T_{cws,set} + \Delta T \; and \; SP_{fan} > SP_{high} + \Delta SP \tag{4.6}$$

and switched off if

$$T_{cws} < T_{cws,set} - \Delta T \; or \; SP_{fan} < SP_{low} - \Delta SP \tag{4.7}$$

where T_{cws} is condenser water supply temperature, $T_{cws,set}$ is condenser water supply temperature setpoint, ΔT is temperature dead band, SP_fan is cooling tower fan speed, SP_{high}, SP_{low} are the high and low threshold of the fan speed,, and ΔSP is the fan speed dead band. In this case study, ΔT is set to 1 °C, SP_{high} is set to 0.8, SP_{low} is set to 0.4, and ΔSP is set to 0.1. To prevent short-cycling, the conditions described in (4.6) an (4.7) need to remain true for 5 minutes before the control actions are triggered.

4.2 Modelica Models

The cooling and control systems are modeled using the Modelica language. The Modelica Buildings library version 5.0.0 provides models for data center cooling systems [48]. The following sections illustrate how to perform modeling and simulation by taking advantage of object-oriented, equation-based modeling. We first introduce the implementation of equipment models and control system models, and then demonstrate the system model by integrating the models of equipment and control.

4.2.1 Component Models for the Cooling System

Most components of the cooling system are modeled directly using the existing models in the Modelica Buildings library. The pipes and ducts are modeled using *Buildings.Fluid.FixedResistances.PressureDrop*, which is a flow resistance with a fixed flow coefficient. The cooling tower is modeled using *Buildings.Fluid.HeatExchangers.CoolingTowers.YorkCalc*, which uses a polynomial to predict the approach temperature for the cooling tower at off-design conditions. The performance of chiller

compressor is predicted using the DOE-2 electrical chiller model [65], which consists of 3 performance curves: $CAPFT$ a curve that represents available cooling capacity as a function of the evaporator and condenser temperature, $EIRFT$ a curve that represents the full load efficiency as function of the evaporator and condenser temperature, and $EIRFPLR$ a curve that represents the efficiency as a function of the part-load ratio. The head and power of the pumps/fans are represented as a quadratic equation in terms of the flowrate. Detailed curves are shown in Table 4.2, where T_{wb} is the wet bulb temperature, T_{ran} is the range temperature defined as the temperature difference between the supply and return condenser water, r is the water to air mass flowrate ratio, T_{chws} is the chilled water supply temperature, T_{cws} is the condenser water supply temperature, PLR is the part load ratio, Q is the flowrate of water or air, H is the head of the pump/fan, P is the power of the pump/fan. a, b, c, d, e, f are the coefficients that needed to be calibrated for each model.

For component models that are not included in the Buildings library, we constructed the models based on Modelica standard library (Version 3.2.2 Build 3) and Buildings library (Version 5.0.0). For example, the Buildings library has no model for a chiller that is equipped with two variable-speed compressors. However, such a dual-compressor chiller model can be built quickly by utilizing the existing chiller models in the Buildings library. As shown in Figure 4.5, we instantiated an electric chiller object model (*Buildings.Fluid.Chillers.ElectricEIR*) twice to represent two variable-speed compressors, denoted as *Compressor 1* and *Compressor 2*, respectively. Each compressor has its own performance curves to calculate the off-design performance.

A stage control is also included to activate the compressors based on load conditions and other control commands. When the chiller is commanded on, one compressor will be turned on immediately. The second compressor is staged on if

$$T_{chws} > T_{chws,set} + \Delta T \; and \; \Delta t_{off} > \Delta t_{thr} \tag{4.8}$$

Table 4.2: Performance curves for major cooling equipment

Model	Performance Curves
Cooling Tower	$T_{app} = a_1 + a_2T_{wb} + a_3T_{wb}^2 + a_4T_{ran} + a_5T_{ran}T_{wb} + a_6T_{wb}^2T_{ran} + a_7T_{ran}^2 + a_8T_{wb}T_{ran}^2 + a_9T_{wb}^2T_{ran}^2 + a_10r + a_{11}T_{wb}r + a_{12}T_{wb}^2r + a_{13}T_{ran}r + a_{14}T_{wb}T_{ran}r + a_{15}T_{wb}^2T_{ran}r + a_{16}T_{ran}^2r + a_{17}T_{wb}T_{ran}^2r + a_{18}T_{wb}^2T_{ran}^2r + a_{19}r^2 + a_{20}T_{wb}r^2 + a_{21}T_{wb}^2r^2 + a_{22}T_{ran}r^2 + a_{23}T_{wb}T_{ran}r^2 + a_{24}T_{wb}^2T_{ran}r^2 + a_{25}T_{ran}^2r^2 + a_{26}T_{wb}T_{ran}^2r^2 + a_{27}T_{wb}^2T_{ran}^2r^2$
Chiller	$CAPFT = b_1 + b_2T_{chws} + b_3T_{chws}^2 + b_4T_{cws} + b_5T_{cws}^2 + b_6T_{chws}T_{cws}$ $EIRFT = c_1 + c_2T_{chws} + c_3T_{chws}^2 + c_4T_{cws} + c_5T_{cws}^2 + c_6T_{chws}T_{cws}$ $EIRFPLR = d_1 + d_2PLR + d_3PLR^2$
Pump/Fan	$H = e_1 + e_2Q + e_3Q^2$ $EIRFPLR = d_1 + d_2PLR + d_3PLR^2$

and staged off if

$$T_{chws} < T_{chws,set} - \Delta T \; and \; \Delta t_{on} > \Delta t_{thr} \tag{4.9}$$

where T_{chws} is the chilled water supply temperature, $T_{chws,set}$ is the chilled water supply temperature setpoint, ΔT is a temperature dead band of 1 °C, Δt_{off} is the elapsed time since the compressor was off, Δt_{on} is the passing of time after the compressor was commanded on last time, and Δt_{thr} is the time threshold (e.g. 20 minutes in this case) to prevent short cycling of compressors.

4.2.2 Control System Model

4.2.2.1 Cooling Mode Control

Figure 4.6 shows the Modelica implementation of the cooling mode control described in Figure 4.3. On the left are the connectors for the control input signals expressed as real numbers, including $T_{floor,set}$, $T_{OA,db}$, $T_{OA,dp}$, and $T_{RA,db}$. In the middle is the state graph implemented using the Modelica state graph package. On the right are signal conversions from Boolean to Integer signals, followed by an Integer connector, which outputs the control signal of the cooling modes.

There are three states in the cooling mode controller, indicated by the squared block icons in the middle of Figure 4.6. The states are FC, PMC, and FMC mode. The initial state is set to FMC mode when simulation starts. The transitions between the states are represented by the horizontal

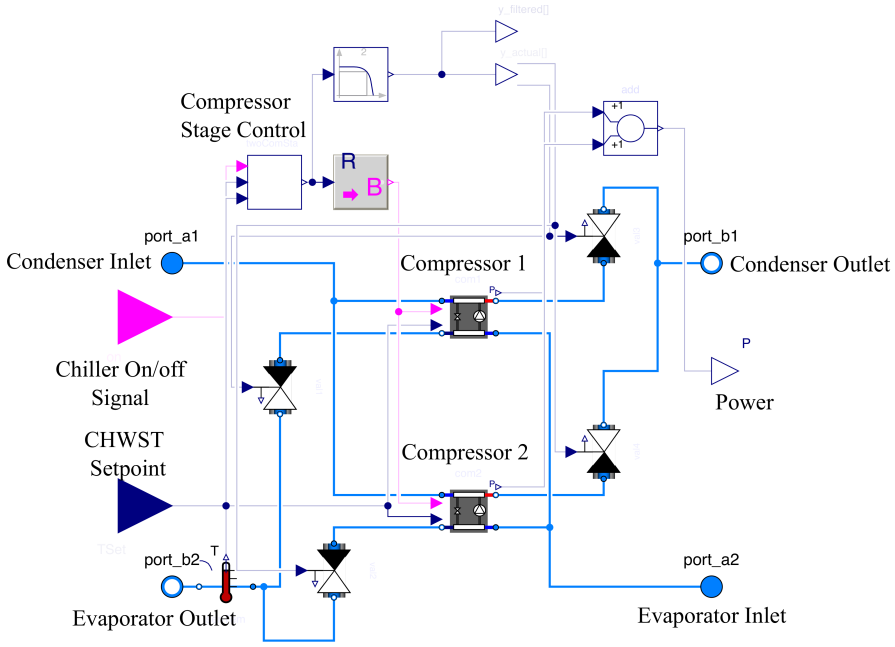


Figure 4.5: Diagram of Modelica model for a chiller with two variable speed compressors

black bars, and each transition has exactly one preceding state and one succeeding state.

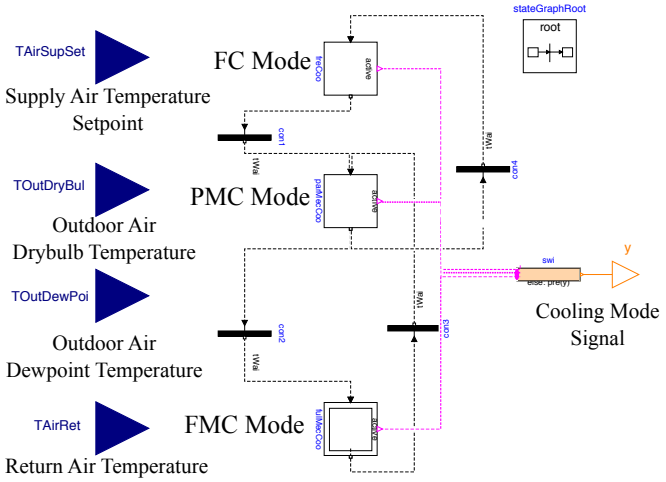


Figure 4.6: Diagram of Modelica models for the cooling mode control

4.2.2.2 Equipment-Level Controls

The hierarchy of the equipment-level controls described in Figure 4.2 is implemented in Modelica by a bottom-up approach. We first declare the equipment controls at Layer 3 as Modelica classes. Then we instantiate these classes and encapsulate their instances layer by layer (from Layer 3 to Layer 1) to formulate the control models for the different loops in Layer 1.

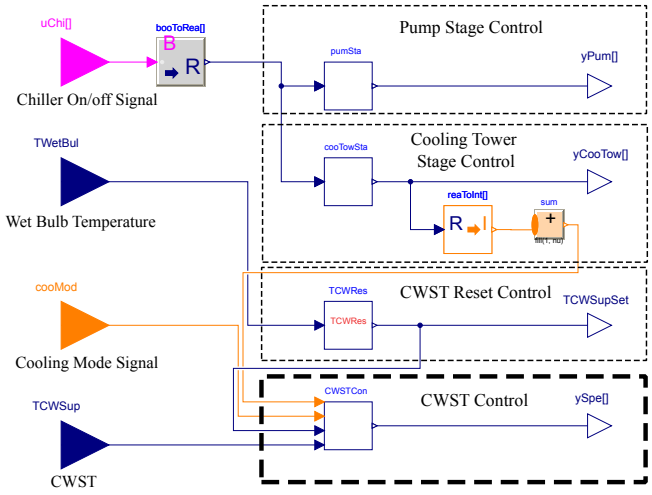
Figure 4.7 shows a part of the hierarchical models of the condenser water loop control. Each icon encapsulates a model that may encapsulate other models. Figure 4.7(a) shows the model for the condenser water loop control in Layer 1, including pump stage control, cooling tower stage control, CWST setpoint reset control, and CWST control, as indicated in the dash boxes. Figure 4.7(b) shows the implementation of CWST control in Layer 2. The CWST control model further encapsulates the instantiations of fan stage control model and the fan speed control model in Layer 3. For instance, Figure 4.7(c) demonstrates the implementation of the fan speed control model. Taking advantages of object inheritance, and instantiation in the object-oriented Modelica, this hierarchical modeling structure allows users to manage the complexity of large models, and to assemble system models as one would connect components in an actual system. This structure also

facilitates debugging and verification of component models. For example, a lower-level model is first debugged and verified, and then instantiated in a higher-level model, which can help identify modelling errors at the early stage of the model development.

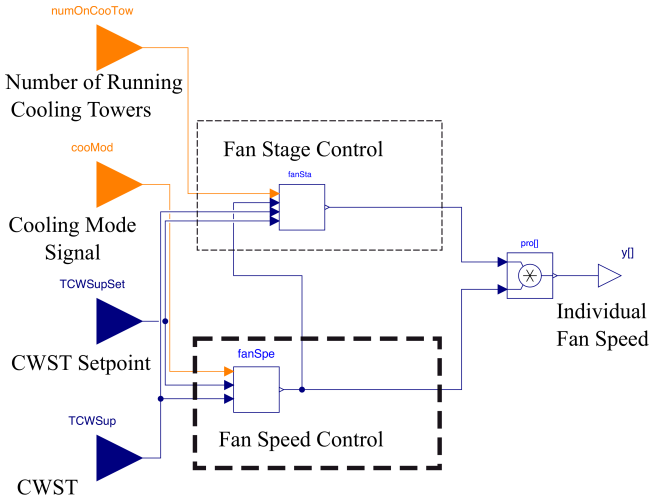
4.2.3 System Models

After implementing the necessary equipment and control models, this section introduces the system model which combines both the physical plant and control system. Several important assumptions are made to simplify the system level model. First, the cooling load in the data center room consists of only heat generation from IT equipment. The heat transfer from envelope and lighting are not considered. Second, the cooling load in the data center is assumed to be constant and equally distributed into two zones (158 kW in each zone). Each zone is cooled by a separate AHU. This assumption is reasonable because the measured heat transfer and power for the two AHUs are almost identical. Third, each zone is modeled as an ideally-mixed volume by assuming the air in the data center room to be completely mixed, because our focus here is the backend cooling system instead of air flow distribution in the room. Forth, the underfloor plenum is modeled using a lumped resistance model instead of a detailed air flow model. Last, communication and computation in the control system are assumed to be instantaneous.

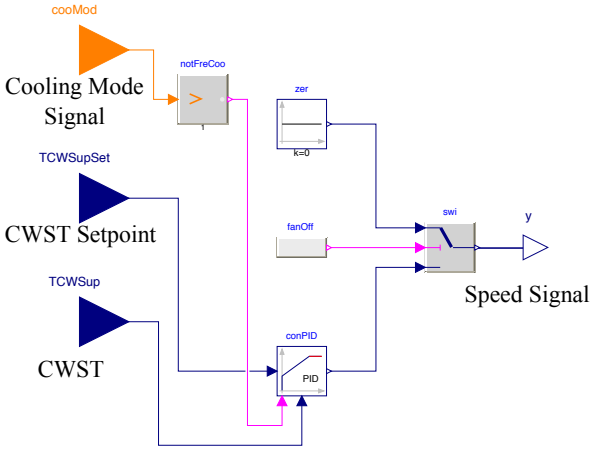
As shown in Figure 4.8, the integrated system model is composed of two parts: the cooling system and the control system. The cooling system is shown at the right side, where the red solid lines represent the condenser water loop, the blue solid lines represent the chilled water loop, and the yellow lines represent the air loop. The controls are displayed on the left side of Figure 4.8, and include the cooling mode control, and equipment-level controls such as the condenser water loop control, the chilled water loop control, and the air loop control. Each component model that formulates the system model is verified in a simulation example following the procedure used by the Modelica Buildings library [140]. We validated the customized models using analytical verification, which has also been used to validate all individual component models in the Modelica Buildings library. For example, the model for the chiller with two variable-speed compressors is verified by



(a) Condenser Water Loop Control



(b) Fan Control



(c) Fan Speed Control

Figure 4.7: Detailed energy consumption in the cooling equipment

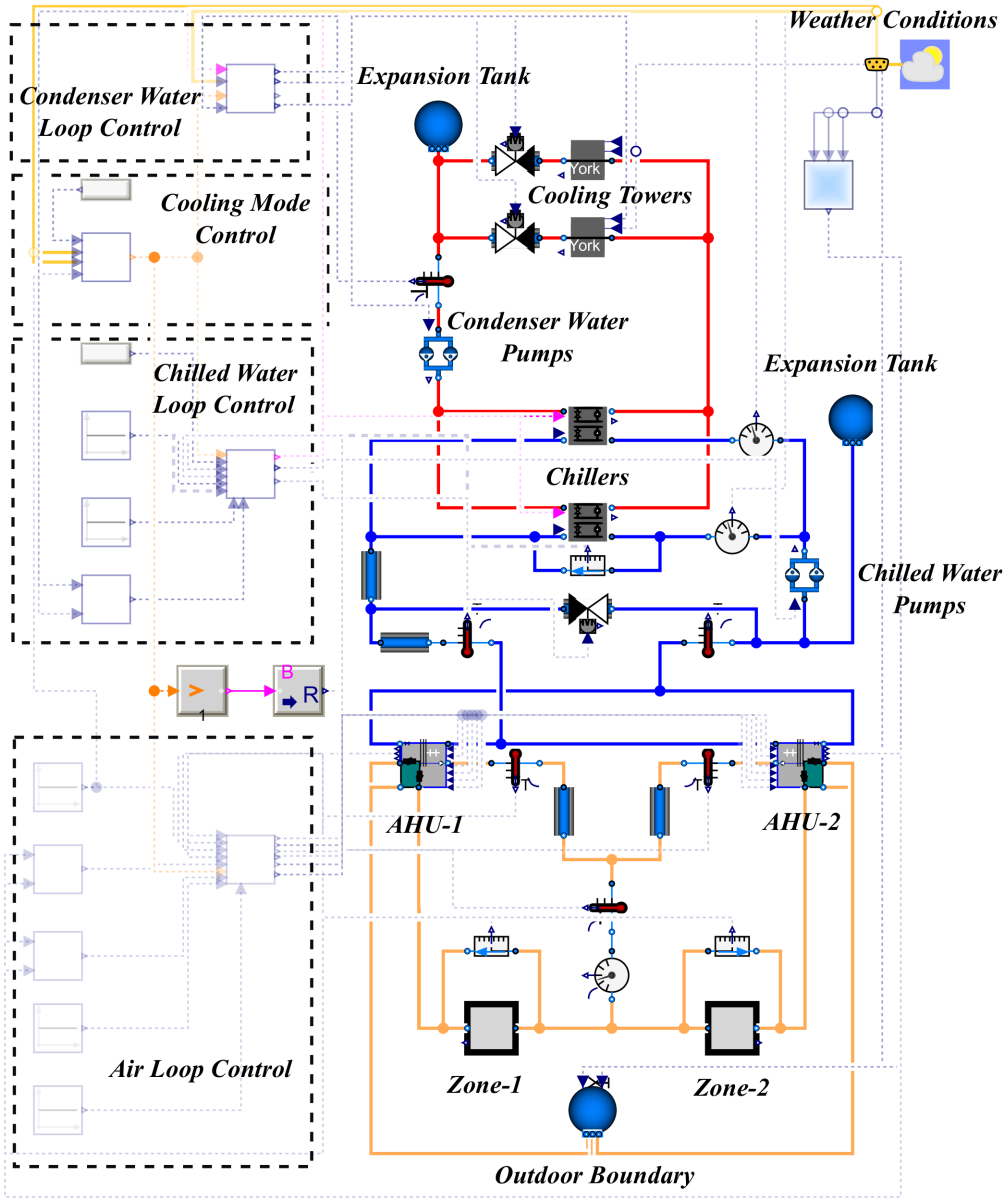


Figure 4.8: Implementation of the system-level model in Modelica

comparing its simulation results with the analytical solutions that are derived for certain steady-state or transient boundary conditions.

4.3 Model Calibration

To evaluate the performance of potential retrofit solutions, we need to establish a baseline model which will predict the performance of the data center cooling system. Figure 4.9 describes a general procedure to calibrate the baseline model. To automate the calibration process, we formulated it as an optimization problem. The objective of the calibration is to minimize the difference between the model output and the corresponding measurement. The difference is defined by the Normalized Mean Bias Error (NMBE), e_{NMBE}. The formulation of the optimization problem is shown in (4.10).

$$\begin{aligned} \min \quad & e_{NMBE} = \frac{\int_{t_0}^{t_0+\Delta t} |f(p, in, t) - M(t)| dt}{\int_{t_0}^{t_0+\Delta t} |M(t)| dt} \\ \text{s.t.} \quad & p_{lb} \leq p \leq p_{ub} \end{aligned} \tag{4.10}$$

where f is the calibrated model, which are shown in Table 4.3. M is the corresponding measurement, in are the inputs for the Modelica models, which can be obtained from measurement data, t_0 and Δt are initial time and length of the calibration period, p are the adjustable parameters of the model, and p_{lb} and p_{ub} are the lower and upper bounds of the parameters p, respectively. The optimization problem is solved using the Particle Swarm Optimization algorithm [67] in GenOpt.

To calibrate the baseline model, the abovementioned optimization problem was established and solved for each cooling equipment using measurement data from October 3 to November 3 in 2017. The measurement data were sampled at a 5-minute interval and divided into two sets: the first 80% were used for calibrating the models, and the remaining 20% were used to evaluate the calibrated models. Table 4.3 shows the calibration problems and their results for different equipment. The AHUs are calibrated by adjusting nominal UA values ($UA_{nominal}$) to predict the outlet temperatures on both air and water side. For the chiller, since the chiller needs two

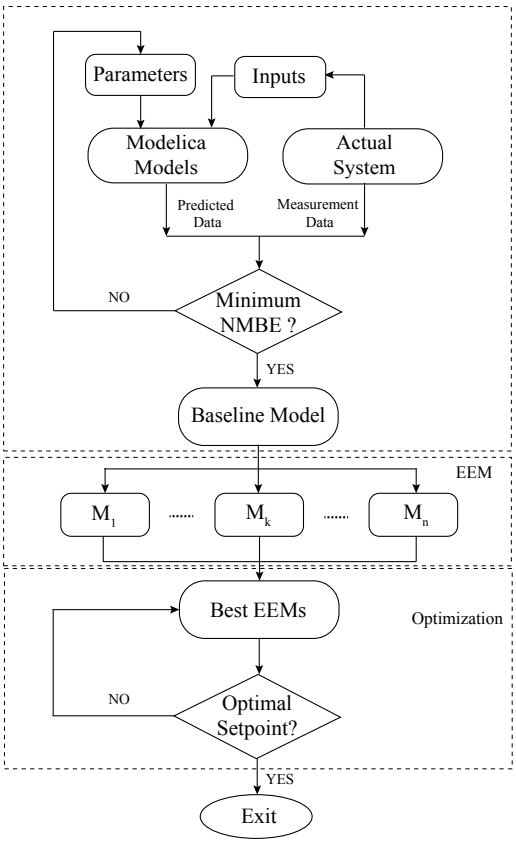


Figure 4.9: Flowchart of the baseline model calibration

compressors to run at the same time at current cooling load level and we only have the measurement data for the chiller as a whole, we assume the two compressors have the same performance during the calibration. Because the condenser water pump runs at constant speed and the measured power is almost constant, we only calibrated the nominal power ($P_{nominal}$) and nominal head ($H_{nominal}$) instead of the performance curves listed in Table 4.3.

Table 4.3: Calibration and evaluation results

Model		Input	Range	Adjusted Parameters	Calibrated Results	Relative Errors Calibration	Evaluation
AHU-1		$T_{wat,in}$	[8, 12.1] °C	$UA_{nominal}$	28.6 kW/K	7%	3%
		$T_{air,in}$	[24.2, 28.1] °C				
AHU-2		m_{wat}	[12.5, 14.5] kg/s	$UA_{nominal}$	30.2 kW/K	8%	3%
		m_{air}	[30.3, 37.7] kg/s				
Chiller	Compressor 1 / Compressor 2	T_{chws}	[8, 12.1] °C	b_i	[1.002148,3.300191E-02,3.741670E-04,-5.925358E-03,-2.599267E-05,-2.172126E-04]	8%	6%
		T_{cws}	[21.2, 25.8] °C	c_i	[4.475957E-01,-1.054652E-02,7.126870E-04,1.158632E-02,5.151510E-04,-9.831355E-04]		
		PLR	[0.35, 0.56]	d_i	[2.519108E-01,2.756914E-01,4.725826E-01]		
Cooling Tower		T_{wb}	[1.8, 22.0] °C	a_i	[-2.3982747E-01,-3.6702405E-02,1.590029E-03,1.15951251E-01,-1.6564918E-02,3.22868E-04,-3.726566E-03,3.84672E-04,-8.94952E-06,1.898438674,-8.1176766E-02,9.73283E-04,1.120285767,-1.1128052E-02,-4.79369E-04,-1.699013E-02,3.24994E-05,1.81282E-05,-4.3584417E-02,-1.518778E-03,1.66684E-04,-6.0704364E-02,2.121175E-03,2.57478E-05,-2.285692E-03,5.7106E-06,1.01121E-06]	4%	4%
		T_{ran}	[3.3, 7.8] °C				
		r	[0.9, 1.8]				
Chilled Water Pump		Q	[0.02, 0.03] m^3/s	e_i	[7.66E05, 1.74E07, 8.58E06]	2%	3%
				f_i	[9.20, 4.21E05, 6.82E06]		
Cooling Tower Fan		SP	[0.5, 1]	e_i	N/A	3%	3%
				f_i	[0,0,0,7.23789E03]		
Condenser Water Pump		N/A	N/A	$P_{nominal}$	8.623E03 W	1%	1%
				$H_{nominal}$	2.89288E05 Pa		

Relative errors between measurement and prediction are within 8% for all component models during calibration and evaluation. A detailed comparison for the cooling tower is shown in Figure 4.10. The system electricity consumption error during the calibration and evaluation period is obtained as 5% and 6% after the calibration of the component models.

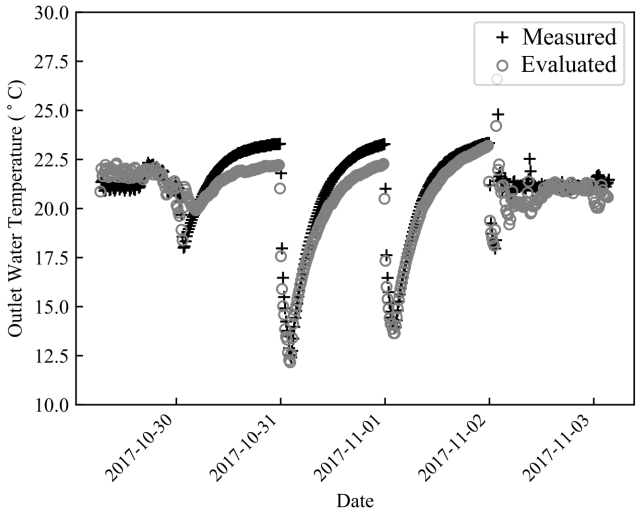


Figure 4.10: Evaluation results for cooling towers

Although only limited measurement data are utilized for calibration, it is sufficient enough to predict for other seasons. The reasons are listed as follows. First, at the early October, the outdoor air temperature is high enough for the system to run at mechanical cooling (PMC and FMC) mode, while at the late October the system can work at FC mode. The measurement data might be in a broad range for calibration. Second, when the mechanical cooling (PMC and FMC) is activated, the performance of chiller is calibrated using the inlet condenser water temperature (T_{cws}), the outlet evaporator temperature (T_{chws}). Because both temperatures are controlled within a limited range during the whole year, and the measurement in October covers most of the range (see Table 4.3), it is acceptable to use measurement in October to calibrate chillers. Similar situation happens for pumps and cooling coils in AHUs. For the cooling tower, although the wet bulb temperature in October is much higher than that in February, but it has little influence on the performance of the

cooling tower, because in cold days, the cooling towers are shut off.

4.4 Retrofit Solutions

We first performed an annual numerical simulation of the baseline system. Based on the analysis of the energy and control performance, we identified several control related issues and proposed corresponding design solutions to improve the control performance. The new design solutions were then examined using the system models. Lastly, we conducted an optimization to further improve the energy saving potentials of the proposed solutions. All the simulations were performed using the local weather data in 2017.

4.4.1 Baseline System

The simulation of the baseline system with calibrated equipment models and control settings shows that the cooling system works in FC, PMC, and FMC modes for 6310, 17, and 2433 hours in 2017, respectively. We identified three potential improvements in terms of control and energy efficiency.

First, the cooling coils are degraded possibly because of fouling. The calculated overall thermal conductance according to the measured data is $UA = 28.6$ kW/K for the AHU-1, which is only 37% of the design value ($UA = 77.4$ kW/K) calculated from the manufacture data.

Second, the simulation results show that the cooling system generally operates either in FMC mode or FC mode. However, it rarely works in PMC mode, where the return air is pre-cooled by cold outdoor air. The cooling mode control are defined in (4.1) to (4.4). However, the difference between $T_{OA,dp,low}$ in (4.2) and $T_{OA,dp,high}$ in (4.3) is only 1.1 °C, which is small. As a result, the system will be able to stay in PMC mode only if $T_{OA,dp}$ is between 12.2 °C and 13.3 °C. Otherwise, it will move to FMC or FC mode.

Third, there is simultaneous heating and cooling in the AHUs, which is caused by the control of the SAT and the chilled water flow rate as described in Section 4.1.2.2. In FMC mode, it is possible that the SAT is lower than the set point (e.g. the control output signal y_2 from PID-2 is

less than 0.4). This will activate the reheaters in AHU-1, but the CHWST will remain the same because the chiller CHWST reset control can only be activated when y_2 is larger than 0.4. As a result, the air is overcooled by the chilled water and then heated by reheaters in AHU-1. Besides the SAT control, lacking flow rate control for the chilled water in AHUs also contributes to the simultaneous heating and cooling problem. As described in Section 4.1.2.2, the speed of the primary pumps is modulated to maintain a fixed pressure difference of 206 kPa between the inlet and outlet of the evaporators, and the bypass valve is adjusted to guarantee that the chilled water through the evaporators is 0.03 m^3/s all the time. Without direct control of the chilled water through the cooling coils, it can lead to a persistent oversupply or undersupply of chilled water to the cooling coils. The oversupply of chilled water can over-cool the supply air.

4.4.2 Energy Efficiency Measures

To address the energy inefficiencies identified above, we propose the following three energy efficiency measures (EEMs) for the cooling and control systems. The rest of this paper uses M to represent the system with the corresponding EEMs implemented.

4.4.2.1 M_1: Clean Cooling Coils

In system M_1, we propose cleaning the fouled cooling coils on both water and air sides. We assume that the UA value of the cleaned cooling coils can be the same as the design nominal UA value. However, the simulation results show that cleaning the cooling coil alone actually results in 76% more energy consumption than the baseline results because improving the cooling efficiency makes the existing over-cooling problem even worse. After being cleaned, the heat transfer effectiveness of the cooling coils increases, which means under the same CHWST, the cleaned cooling coils cool the supply air to a lower temperature than the fouled cooling coils. As a result, the AHUs with clean cooling coils need additional reheat energy to maintain the same SAT when the water flowrate through the cooling coils is not regulated. For example, when $T_{floor,set}$ is reset to 23.3 °C, the CHWST can be as high as 12.2 °C. The fouled cooling coils can cool the supply air to

around 21.0 °C, and we just need to reheat it to 23.3 °C. However, the clean cooling coils can cool the supply air to around 18.0 °C, and additional energy is expended to bring the SAT to 23.3 °C. Furthermore, the energy of chillers and cooling towers in M_1 increases because the heat generated by the reheaters increases the thermal load of the chillers and hence cooling towers.

4.4.2.2 M_2: Improve Cooling Mode Control

To increase the operating time of PMC mode, we need to make it easier to move from FMC to PMC and more difficult to move back. To achieve this goal, we propose to increase the $T_{OA,dp,high}$ by setting it to 15 °C in system M_2. This temperature is a high $T_{OA,dp}$ recommended by ASHRAE for the data center equipment environment [4]. All the other settings remain the same as the baseline system.

The annual simulation show that M_2 can save 9.0% of cooling energy compared to the baseline system because the improved cooling mode controller allows the cooling system to operate less in FMC mode, and more in FC and PMC modes. The detailed explanation is as follows:

- The higher cutoff limit of $T_{OA,dp}$ (15 °C) reduces the operational time of FMC mode in M_2 (from 2,433 hours to 1,632 hours). Due to a higher $T_{OA,dp}$, the system can stay in PMC mode longer. As a result, M_2 operates in PMC mode for 188 hours in the whole year, compared to 16 hours for the baseline system. As the chillers only need to address part of the cooling load in PMC mode, they consume less energy than in FMC mode.

- More importantly, M_2 works in FC mode for 6,938 hours, which is 628 hours more than the baseline system. By increasing the time staying in PMC mode, it also increases the possibility of switching from PMC mode to FC mode. For example, as shown in Figure 4.11, the baseline system works in FMC mode all the time from July 12 to July 15, but M_2 can work in FC mode for almost two days in the same period. The $T_{OA,db}$ is lower than $T_{floor,set}$ as well as $T_{RA,db}$ during most of the time. At the beginning of July 12, $T_{OA,dp}$ is higher than the high dew point temperature cutoff limit in both the baseline system ($T_{OA,dp,high,Baseline}$)

and M_2 ($T_{OA,dp,high,M_2}$), thus FMC mode is activated in both systems. However, as $T_{OA,dp}$ continues decreasing to $T_{OA,dp,high,M_2}$, M_2 can operate in PMC mode, while the baseline system still works in FMC mode. When M_2 works in PMC mode, (4.1) is easily triggered because $T_{OA,db}$ is lower than $T_{floor,set}$, which switches the system from PMC to FC mode. Therefore, with a higher $T_{OA,dp,high}$, it is easier for the cooling system to switch from FMC to PMC mode, and then to FC mode.

However, increasing $T_{OA,dp,high}$ might pose challenges for the humidity control of the underfloor air. Figure 4.12 shows the box plot of the hourly relative humidity (RH) in both M_2 and the baseline system. The central rectangular box spans the first quartile to the third quartile, and the dashed line inside displays the median. The lower and upper whiskers represent the 0.1 percentile and the 99.9 percentile, respectively, which means there is only 0.1% of the data between the minimum and the lower whisker, and 99.9% of data between the minimum and the upper whisker. In the baseline system, the RH is within the boundary preferred by the operators. However, M_2 exceeds the upper bound since it introduces more humid outdoor air in the data center room for free cooling, and the cooling coil in the AHUs has a very limited capacity for dehumidification with a design sensible heat ratio of 0.99.

4.4.2.3 M_3: Improve SAT Control

To mitigate the problems of simultaneously heating and cooling in the AHUs, we propose adding a two-way valve on the waterside of cooling coils to regulate to maintain the SAT in PMC and FMC modes in system M_3. Instead of maintaining a constant differential pressure between the inlet and outlet of the evaporators, M_3 adjusts the speed of the primary pumps to maintain a fixed differential pressure of 83 kPa between the inlet and the outlet of the cooling coils. The bypass valve in the common leg is adjusted to achieve a minimum flow rate through evaporators. The above proposal will regulate the amount of chilled water passing through the cooling coils to avoid an oversupply of chilled water and reduce simultaneous heating and cooling in AHUs.

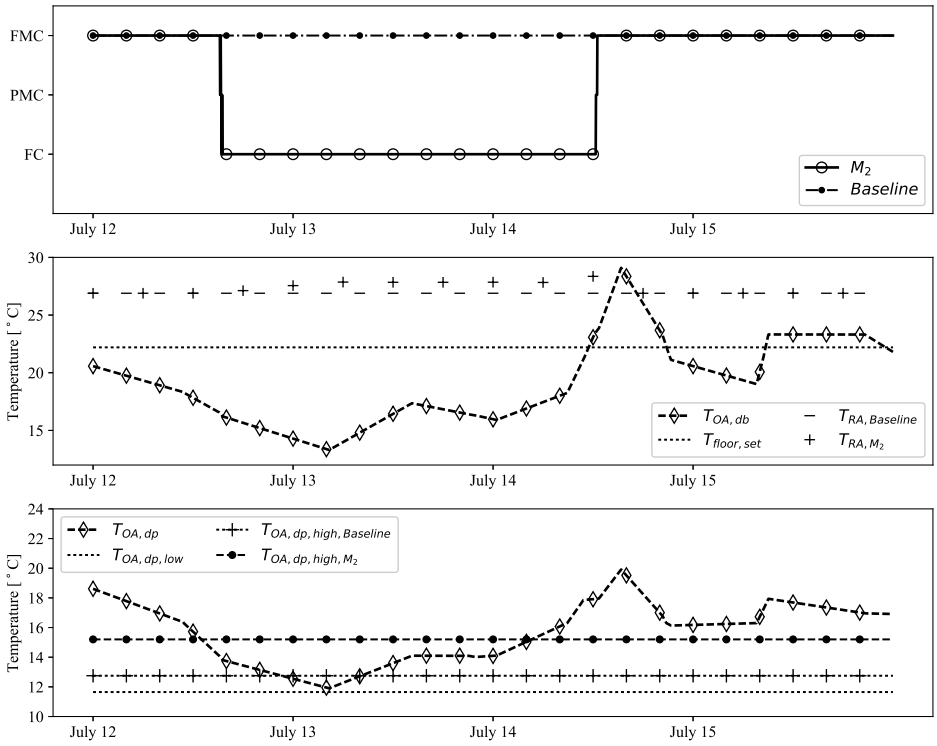


Figure 4.11: Comparison of the system details in M_2 and the baseline system

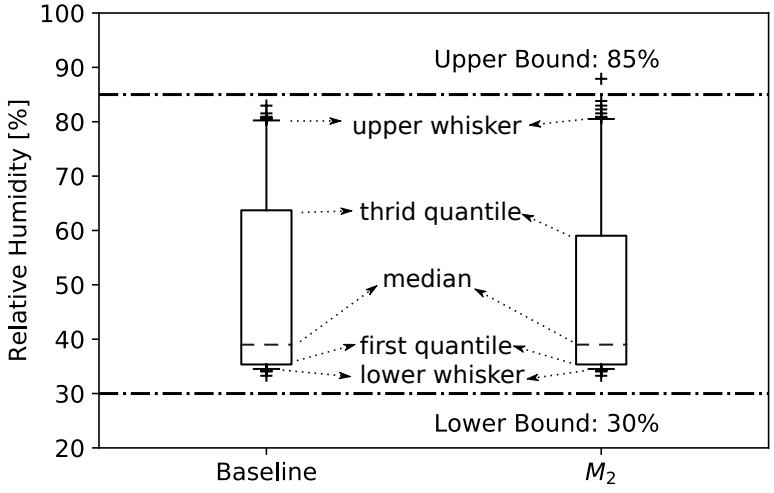


Figure 4.12: RH of underfloor air in M_2 and the baseline system

Simulation results show that 9.4% of cooling energy can be saved in M_3. Most of the savings are from reheaters. Figure 4.13 compares the operational status of M_3 and the baseline system during July 12 to July 15. Although the systems operate in the same cooling mode, to track the same SAT setpoint, the reheaters in the baseline system need to be activated, while those in M_3 are completely deactivated (see the bottom figure in Figure 4.13).

4.4.3 Sequential Search for EEMs

The process of designing and retrofitting real buildings often involves choosing among discrete options, for example, different EEMs. To propose the best EEMs for a building retrofit, engineers need to explore and search the design space of possible EEMs. Genetic Algorithms (GAs) are most commonly used for building energy optimizations [134]. Others seek to develop the Pareto Frontier the set of cost-optimal solutions over a range of energy savings [133, 43]. However, GAs are typically for optimization problems with large amount of decision variables, such as a large amount of different EEMs. Because we only proposed three EEMs in this case, the Sequential Search Technique [30, 62] is utilized to find the most cost-effective retrofit solutions.

4.4.3.1 Sequential Search Technique

The basic principle of Sequential Search technique is schematically shown in Figure 4.14. All the proposed EEMs are simulated individually. These simulations comprise an initial iteration of the optimization process. As illustrated in Figure 4.14, the most cost-effective option (points with steepest slope compared with the optimal design in previous iteration), based on simulation results and energy-related costs, is chosen as the baseline point for the next iteration. The chosen EEM is then removed from future evaluations by the search. Remaining EEMs are simulated in the presence of this new baseline point and the iterative process repeats. The method can provide intermediate optimal points, that is, the minimum cost designs at various levels of energy savings, which enables the engineers to make their choice when they are interested in intermediate solutions, rather than a global optimum [30, 62].

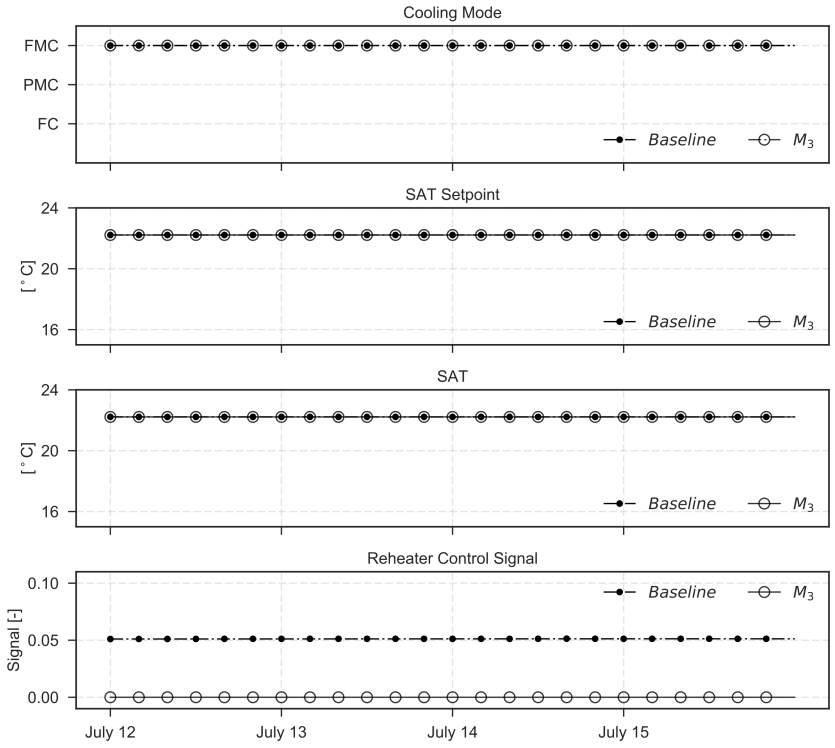


Figure 4.13: Cooling and heating in the AHUs. Simultaneous heating and cooling is avoided in M_3

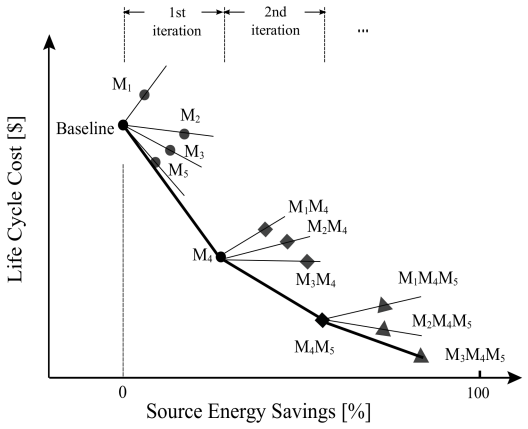


Figure 4.14: Illustration of Sequential Search Technique

4.4.3.2 Results of Sequential Search Technique

We performed a sequential search technique among the abovementioned three EEMs considering energy savings and LCC. For energy savings, only electricity consumed by the cooling system was taken into account. The LCC is calculated using Eq.(4.11). The N is the life cycle, and the r_d is the real discount rate. The $C_{i,n}$, $C_{o,n}$, and $C_{m,n}$ are capital cost, operation cost and maintenance cost in year n, respectively. In this study, we set N to 40 years, and r_d to 0.02. Capital costs for different measures were estimated by experienced engineers of the analyzed data center, and operation costs were calculated using a variable basic service charge offered by the utility company in Massachusetts [124], which is also shown in Table 4.4. Maintenance cost for each measure was set to 0 in this study.

$$LCC = \sum_{n=1}^{N} \frac{C_{i,n} + C_{o,n} + C_{m,n}}{(1 + r_d)^n} \tag{4.11}$$

The results of Sequential Search are documented in Table 4.5. The calculated LCCs are listed in the $6th$ column, and the energy savings compared with the baseline point of each iteration (optimal solution in previous iteration) are listed in the $7th$ column. The slopes between the baseline point in each iteration and the evaluated EEMs are shown in the last column, where N/A means the slope is not calculated because the EEM cannot save energy. In Iteration 1, a single EEM is compared with the baseline system. M_2 is identified as the most cost-effective solution and then serves as the baseline point for Iteration 2. The simulation results show that the combination of M_2 and M_3 is better than that of M_1 and M_2, because M_1 cannot save energy at all. In the third iteration, compared with the new baseline point M_2M_3, the only combination of $M_1M_2M_3$ has no advantage in terms of both energy savings and LCC.

Sequential search among the proposed EEMs shows that M_2 is suggested if one EEM is adopted, and M_2M_3 is suggested if two EEMs are considered. The combination $M_1M_2M_3$ cannot further reduce energy and LCC compared with M_2M_3, and hence is not an effective retrofit option.

Table 4.4: Utility rates used in the studied data center

Month	Price ($/kWh)	Month	Price ($/kWh)	Month	Price ($/kWh)
January	0.10759	May	0.06823	September	0.08515
February	0.10632	June	0.08505	October	0.08360
March	0.08565	July	0.08993	November	0.08908
April	0.07226	August	0.08752	December	0.10415

Table 4.5: Sequential search process for the proposed EEMs

Iteration #	EEM	Energy (MWh)	Annual Operation Cost ($)	Initial Cost ($)	LCC	Energy Savings (%)	Slope
1	*Baseline*	446	39,121	0	1,070,174	0	0
	M_1	787	68,574	1,000	1,876,875	-76.5	N/A
	M_2	406	35,574	100	973,244	8.97	-1,080,768
	M_3	404	35,458	110,037	1,080,008	9.42	104,426
2	*Baseline*	406	35,574	100	973,244	0	0
	M_1M_2	656	57,208	1,100	1,566,052	-61.6	N/A
	M_2M_3	358	31,481	110,137	971,315	11.8	-16,316
3	*Baseline*	358	31,481	110,137	971,315	0	0
	$M_1M_2M_3$	358	31,523	111,137	973,464	0	N/A

4.4.4 Optimal Underfloor Plenum Air Temperature Setpoint

To further investigate the energy saving potentials, we proposed to optimize the cooling system by adjusting the control setpoint of the UPAT in addition to the searched EEMs in Section 4.4.2. The following section describes the setup and results of the optimization.

4.4.4.1 Optimization Problem Setup

The optimization problem is formulated as:

$$\begin{aligned} \min \quad & \sum_{con} E_{con}(T_{floor,set}) \\ \text{s.t.} \quad & T_{floor,set,l} \leq T_{floor,set} \leq T_{floor,set,u} \end{aligned} \tag{4.12}$$

where E is the energy consumption, and the subscript con represents different electricity consumers in the cooling system, including chillers, pumps, cooling towers, and AHUs. The energy consumed by the server fans are considered constant in the optimization problem. The UPAT setpoint $T_{floor,set}$ is chosen as the only design variable in this case study. The subscript l and u are the lower and upper bound. Here we assume that the rack inlet temperature is the same as the UPAT. Thus, we can set $T_{floor,set,l} = 18$ °C and $T_{floor,set,u} = 27$ °C based on ASHRAEs recommended range for rack inlet temperatures [4]. The optimization problem is then solved using exhaustive search or parametric analysis with an increment of 0.1 °C for the design variable.

4.4.4.2 Optimization Results

We performed the optimization on three systems: the baseline system, M_2, and M_2M_3. The relationship between the $T_{floor,set}$ and the annual energy consumption for the three systems is shown in Figure 4.15. The systems with optimal $T_{floor,set}$ are denoted as $Baseline_{opt}$, $M_{2,opt}$, and $M_2M_{3,opt}$ respectively. The results show that the optimal $T_{floor,set}$ of $Baseline_{opt}$, and $M_{2,opt}$ is 25.1 °C, and that of $M_2M_{3,opt}$ is 27 °C. The additional energy savings by optimizing the UPAT are around 20 - 25 MWh for all three cases. As a result, the combined energy savings for $Baseline_{opt}$, $M_{2,opt}$, and $M_2M_{3,opt}$ are 4.5%, 14.6% and 24.2% compared with the baseline system, respectively.

Figure 4.15 shows that starting from $T_{floor,set} = 18$ °C, the energy consumption reduces when increasing $T_{floor,set}$. However, for the baseline system and M_2, when $T_{floor,set}$ reaches around 25.1 °C, the annual energy consumption starts to increase. The reason is that when $T_{floor,set}$ is higher than 25.1 °C, the increased reheat energy is larger than the savings from the chillers and their associated pumps. For example, for the baseline system (Figure 4.16), when $T_{floor,set}$ increases from 25.1 °C to 27 °C, the reheat energy in one year increases from 20.4 MWh to 99.0 MWh, but the energy consumed by the chillers only decreases from 89.5 MWh to 87.9 MWh. In M_2M_3, the reheaters consume no energy at all even when $T_{floor,set}$ increases because reheating in the AHUs is avoided.

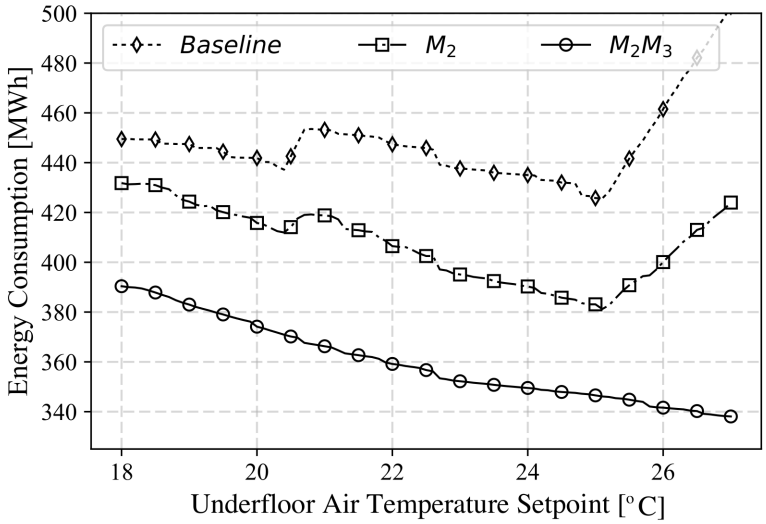


Figure 4.15: Relationship between $T_{floor,set}$ and annual energy consumption for different systems

It is worth mentioning that the equipment-level control strategies have significant influence on the design space of the above optimization problem. For example, Figure 4.15 shows that with the baseline reheat control, the baseline system and M_2 have local optima around 20.5 °C. The local optima are caused by the activation of reheaters in the AHUs. In M_2M_3, the annual energy consumption monotonously decreases as $T_{floor,set}$ increases, because the reheaters are off for the entire range of $T_{floor,set}$.

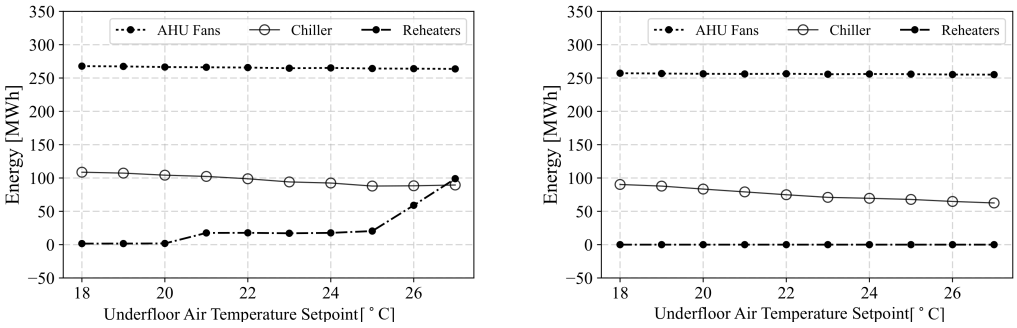


Figure 4.16: Energy impact of $T_{floor,set}$ on major cooling equipment: (left) baseline system; (right) M_2M_3

Figure 4.17 shows the operating time under different UPATs during the whole year in the baseline system and three optimal systems. The UPATs are controlled at their setpoints with a tolerance of 1.5 °C for about 96% of the year in the baseline system, and about 99% in all three optimized systems.

Figure 4.18 compares the RH in the underfloor plenum for a whole year in a box plot with whiskers of 0.1 and 99.9 percentile. The RH in all four systems is within the preferred range. The median RHs in the $Baseline_{opt}$, $M_{2,opt}$, and $M_2M_{3,opt}$ are lower than that in the baseline system, because the RH decreases as the dry bulb temperature increases if the dew point temperature is the same. Taking the baseline system and $Baseline_{opt}$ for example, the only difference is that the $Baseline_{opt}$ utilizes $T_{floor,set}$=25.1 °C instead of 22.2 °C in the baseline system. Because these two systems have the same $T_{OA,dp}$ thresholds in the cooling mode controller, and their cooling coils have limited capacity of dehumidification, we can assume these two systems have the same dew point temperature in the underfloor plenum for most time. However, UPAT in $M_{2,opt}$ is controlled at 25.1 °C, about 3 °C higher than the baseline system. Therefore, the median RH in the $Baseline_{opt}$ is lower than the baseline system.

To understand when the energy savings are achieved in the three optimal systems, we show detailed analysis in Figure 4.19. The energy savings in $Baseline_{opt}$, $M_{2,opt}$, and $M_2M_{3,opt}$ mostly take place during transition (e.g. October - November) and summer seasons (e.g. September).

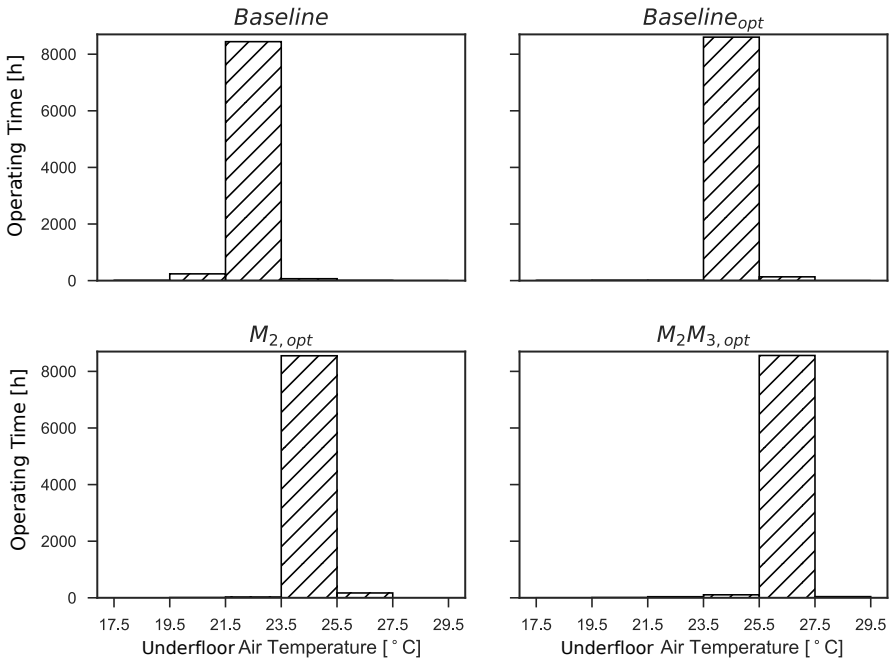


Figure 4.17: Operating time under different UPATs for four different systems during a year

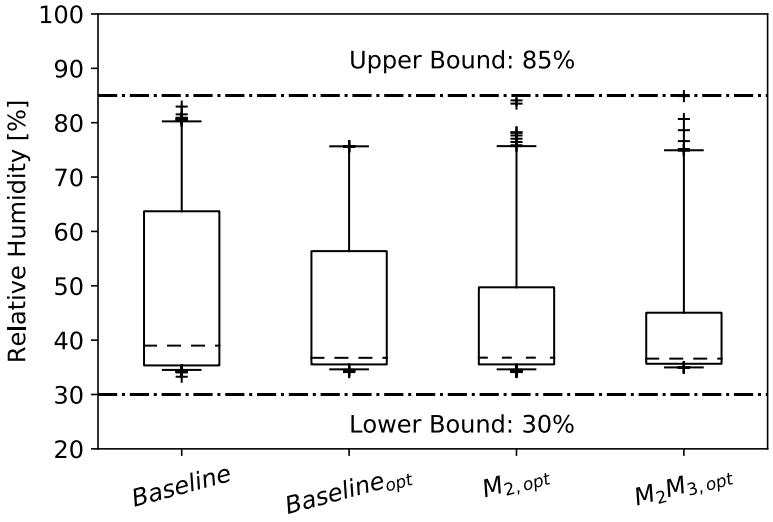


Figure 4.18: Box plot of the underfloor RH in different systems

There are barely energy savings from winter because of free cooling. The maximum daily energy saving in all cases is about 2000 kWh, where all the mechanical cooling is totally deactivated.

To understand where the energy savings are from, we break down the savings in $Baseline_{opt}$ as an example shown in Figure 4.20. For a single day, the cooling tower fans can save up to 150 kWh, and all the pumps together can save up to 480 kWh. The largest saving is from deactivation of the chiller, which can reduce energy by about 900 kWh in a day. For the AHUs, due to the increase of $T_{floor,set}$, the fans need to operate at a higher speed to deliver more air to the data center room in order to keep the room at the setpoint. Therefore, during a winter day, the AHU fans in the $Baseline_{opt}$ can consume 10 kWh more energy than the baseline system due to the increase of the fan speed. For the transition and summer seasons, the AHUs in the $Baseline_{opt}$ can save up to 500 kWh if the reheaters are deactivated, but can also consume about 500 kWh more energy if the reheaters are triggered due to the control in Section 4.1.2.2.

4.5 Summary

In the present study, an equation-based dynamic modeling and simulation approach is performed to evaluate energy and control performance, to develop EEMs, and to optimize the operation in a medium-size data center located in Massachusetts, United States. The baseline cooling and control systems is built in Modelica and calibrated using measurement data. Three individual EEMs related to energy and dynamic control performance are proposed: (M_1) cleaning the cooling coils in the AHUs; (M_2) increasing high cutoff limit of outdoor air dew point temperature in cooling mode controller; (M_3) improving AHU controls to avoid simultaneous heating and cooling.

The intermediate cost-effective retrofit solutions among the proposed EEMs are then identified through the Sequential Search technique. If only one EEM is adopted when budget is limited, M_2 is suggested due to its low initial cost and considerable energy savings. If two EEMs are considered, M_2 and M_3 together can save up to 19.7% cooling energy. Adopting three EEMs simultaneously is not recommended in this case, because it is hardly a cost effective retrofit.

In the end, by optimizing the underfloor air temperature setpoint in addition to the proposed

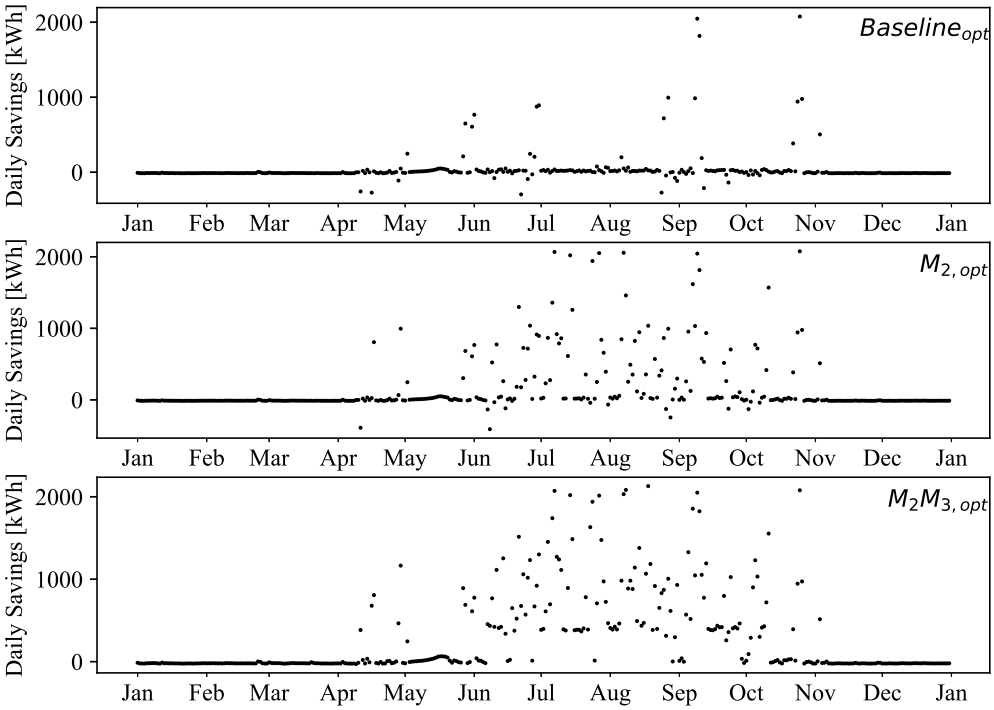


Figure 4.19: Daily energy savings for different systems

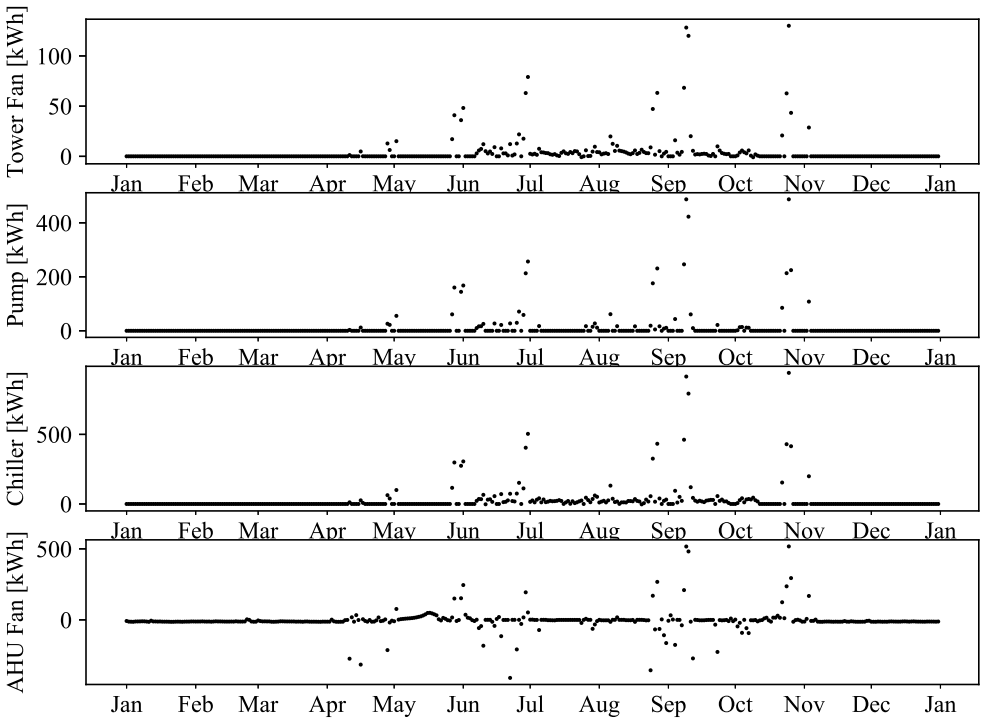


Figure 4.20: Daily energy savings in $Baseline_{opt}$

EEMs, the data center can potentially save as much as 24.2% energy for the cooling system with M_2 and M_3 adopted. The optimal settings for the cooling system can maintain the data center room in an acceptable thermal environment in terms of temperature and relative humidity. The energy savings for all the optimized systems come from summer and transition seasons.

This case study also demonstrates two important features in Modelica-based tools. One is the capability of complexity management through hierarchical modeling, which supports fast modeling of various user cases. The other is the ability to evaluate discrete control involving delay time and dead band, which are commonly used in the control of the cooling system.

Chapter 5

Multi-market Optimization for GEDCs without Thermal Energy Storage Systems

Based on:

Yangyang Fu, Xu Han, Kyri Baker, Wangda Zuo, Jim W. VanGilder, "Assessments of Data Centers for Provision of Frequency Regulation", under review in Energy Conversion and Management.

Yangyang Fu, Wangda Zuo, Kyri Baker, "Multi-market Optimization of a Data Center without Storage Systems", submitted to The American Modelica Conference 2020.

This chapter proposes a synergistic control strategy for data center frequency regulation, which combines power management techniques at the server level with control of the chilled water supply temperature to track the regulation signal from the electrical market. A frequency regulation flexibility factor is also proposed to increase the IT capacity for frequency regulation. The performance of the control strategy is studied through numerical simulations using an equation-based object-oriented Modelica platform designed for data centers. Simulation results show that with well-tuned control parameters, data centers can provide frequency FR service in both regulation up and down. The performance of data centers in providing FR service is largely influenced by the regulation capacity bid, frequency regulation flexibility factor, workload condition, and cooling mode of the cooling system, and not significantly influenced by the time constant of chillers. In addition, compared with a server-only control strategy, the proposed synergistic control strategy can provide an extra regulation capacity of 3% of the design power when chillers are activated.

A real-time multi-market optimization framework for the data center without storage systems

is also proposed to maximize their benefits from participating in both energy market and regulation market. Simulation results shows that providing frequency regulation service over two days in January and July can save $24.8 and $123.6, respectively.

5.1 Server Aggregator

One straightforward way to control the data center power at its reference signal to provide regulation service is to find the optimal frequency for each individual server at each time step. This eventually leads to a large scale optimization problem considering the number of servers in a medium or large data center.

An aggregated server model described in Ref [82] is adopted here. This model can output the real-time power and service response time based on CPU frequency, workload arrival rate, and number of active servers.

$$P_{servers}(t) = \lambda(t) \sum_{0}^{r} b_i f_{agg}(t)^i + \sum_{0}^{s} c_j N_{act}(t)^j, 0 \leq i \leq r, 0 \leq j \leq s \tag{5.1}$$

where b_i, c_j are constant coefficients that can be obtained from curve fitting techniques, $\lambda(t)$ is the workload, f_{agg} is the aggregated relative frequency, ranging from 0 to 1, and N_{act} is the active number of servers at current time. f_{agg} and N_{act} can be optimally determined in order to minimize cost.

Here we use the average response time to quantify the service quality of a data center. The workloads are modeled as GI/G/m queues, which assumes a general distribution with independent arrival times and a general distribution of service times. The total time that a job spends in the queuing system is known as response time. The response time usually consists of two parts: waiting time, that is, the time that a job spends in a queue waiting to be serviced; and service time, that is, the time that a job needs to be executed. The average response time model is adopted from [16]. Details are shown as follows.

$$\mu(t) = k f_{agg}(t) \tag{5.2}$$

$$t_s = \frac{1}{\mu(t)} \tag{5.3}$$

$$\rho(t) = \frac{\lambda(t)}{N_{act}(t)\mu(t)}, 0 \leq \rho(t) \leq 1 \tag{5.4}$$

$$P_m = \begin{cases} \frac{\rho(t)^{N_{act}(t)} + \rho(t)}{2}, & \rho(t) \geq 0.7 \\ \rho(t)^{\frac{N_{act}(t)+1}{2}}, & \rho(t) < 0.7 \end{cases} \tag{5.5}$$

$$t_w = \frac{C_A^2 + C_B^2}{2N_{act}(t)} \frac{P_m}{\mu(t)(1-\rho(t))} \tag{5.6}$$

$$t_r = t_s + t_w \tag{5.7}$$

In the above equations, μ is the mean service rate, k is a constant parameter, assuming the service rate is proportional to the frequency, ρ is the average utilization of the server, representing the fraction of occupied time, P_m is approximated probability that an arriving job is queued, C_A and C_B are constant coefficients reflecting the type of data centers.

5.2 Proposed Synergistic Control Strategy

In this section, we propose a synergistic control strategy for data centers to provide FR service, which is evaluated at a whole system level in the Section 5.3. This strategy is composed of four major parts. The first one is **Baseline Routine**, which predicts the baseline power usage when the data center provides no FR. The second one is **Bidding Capacity**, which is the capacity bid that the data center submits to the electrical market. The third one is **Server Power Management**, where an aggregator is adopted to represent the aggregated performance of servers in the data center. The clock frequency of the aggregator can be directly changed by a Proportional-Integral-Derivative (PID) controller in order to follow the regulation signal. Based on that, the desired frequencies for individual servers will be determined by a set of predefined assignment rules and

then be propagated to all servers. The forth one is **Cooling Power Management**, which adjusts the CHWST setpoint to respond to the regulation signal.

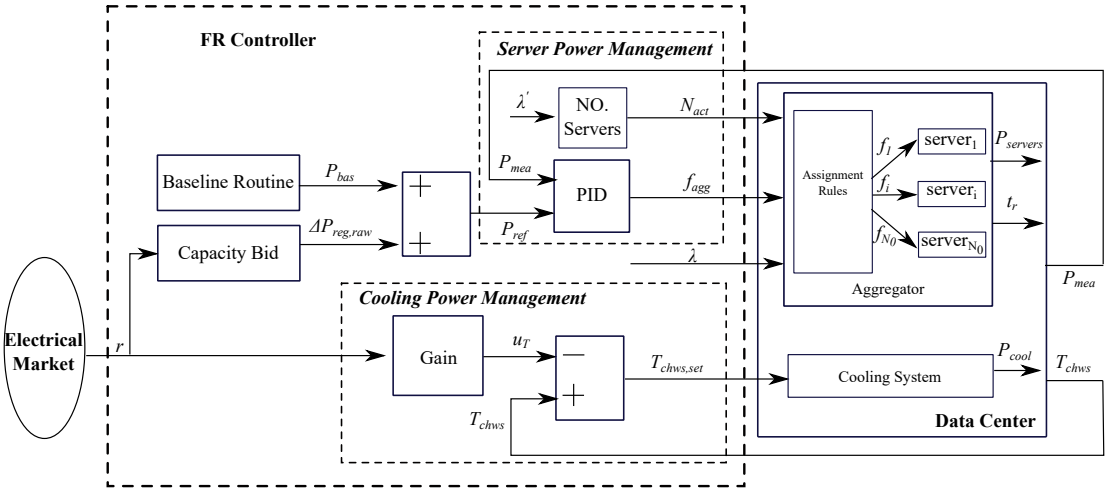


Figure 5.1: Synergistic control strategy for FR

Figure 5.1 shows the workflow of the proposed synergistic control strategy. The **Baseline Routine** outputs the prediction of the overall power profile for the data center P_{bas} when no FR service is provided. In this paper, the prediction is performed using detailed energy models, although many other methods such as machine learning techniques can also be used. The detailed energy models and baseline settings can be referred to Section 5.3.1. The **Bidding Capacity** is a module that can calculate the optimal capacity bid for the data center at each time step, and output raw regulation power $\Delta P_{reg,Raw}$ based on the optimal capacity bid and received regulation signal r from the electrical market. In this paper, we assume the capacity bid C_{reg} is known, since finding the optimal bid is not the focus here. Then, the reference power P_{ref} for the data center to track is the summation of the predicted baseline power P_{bas} together with the raw regulation power $\Delta P_{reg,Raw}$.

The **Server Power Management** first determines the number of required active servers in the aggregator N_{act} based on the predicted workload $\lambda^{'}$ in the next time step (e.g., one hour ahead). Then a closed-loop control using a PID controller is utilized to minimize the error between the measured total power usage P_{mea} and the reference power P_{ref} by adjusting the aggregated

frequency of the server aggregator. Meanwhile, the **Cooling Power Management** applies an open-loop control to adjust the cooling system power usage by resetting the CHWST setpoint in response to the received regulation signal r.

The server aggregator receives the aggregated frequency f_{agg} and the required number of active servers N_{act} from the FR controller. Assuming there are N_0 number of servers in the data center, the server aggregator then calculates the CPU frequency f_i for an individual server i based on predefined assignment rules. The cooling system receives CHWST setpoint from the FR controller. Both the IT system and the cooling system respond in such a way that their total power P_{mea} is adjusted to track the reference power P_{ref}.

For the aggregator, there are several assignment rules to control the individual server's frequency [82, 132]. We can also represent the aggregated server power $P_{servers}$ of all servers under an assignment rule using a simplified model [82] and this approach is adopted by this paper and detailed in Section 5.2.1. For the FR controller, more details are described in the rest of this section.

5.2.1 Server Power Management

The servers in the data center are represented by an aggregator, which is characterized by the active number of servers N_{act}, and the aggregated frequency f_{agg} as shown in Section 5.1. Base on these two parameters, the aggregator can output the total power of the servers $P_{servers}$ and the average service response time t_r. The **Server Power Management** is used to determine N_{act} and f_{agg} at each time step based on the normalized raw regulation signal received from the electrical market, r, ranging from -1 to 1, and incoming actual workload λ.

5.2.1.1 Reference Power

The reference power P_{reg} is calculated as

$$\Delta P_{reg,raw}(t) = r(t)C_{reg} \tag{5.8}$$

$$P_{ref}(t) = P_{bas}(t) + \Delta P_{reg,raw}(t) \tag{5.9}$$

where $\Delta P_{reg,raw}$ is the raw power signal and C_{reg} is the regulation capacity that the data center bids in the market.

5.2.1.2 Number of Active Servers

The number of servers in a data center needs to satisfy the following condition in order to ensure the stability of the IT service. This condition means that the service capability $N_{act}(t)\mu(t)$ in the data center should be greater than the workload) $\lambda(t)$:

$$N_{act}(t)\mu(t) > \lambda(t), \tag{5.10}$$

where $\mu(t)$ is the actual service rate, which denotes the number of requests that a single server can process every second. The service rate is typically proportional to the server's CPU frequency, as defined in Eq. (5.2) [82, 132].

Under design conditions, to guarantee reliability, a scaling factor γ as defined in Eq. (5.11) is utilized here to describe the design redundancy of the servers [82]. The γ is set to greater than 1. If $\gamma = 1$, it means all the CPU clock frequencies need to set at the maximum level just to serve the average workload, which limits the potential of FR. The γ is defined as

$$\gamma = \frac{\mu_0 N_0}{\lambda_0}, \tag{5.11}$$

where μ_0 is the nominal service rate of a single server, N_0 is the nominal number of servers in a data center room, and λ_0 is the nominal workload to be served by the data center.

When using a server aggregator model as described in Section 5.1, the γ can then be rewritten as:

$$\gamma = \frac{kN_0}{\lambda_0} = \frac{kN_{act}(t)}{\lambda^{'}(t)}, \tag{5.12}$$

where k is a constant parameter, assuming the service rate is proportional to the aggregated frequency, N_{act} is the number of active servers at current time step, and $\lambda^{'}$ is the predicted workload, here we use the mean workload of the current time step as the prediction.

The number of active servers is calculated at an interval of 1 hour because the servers have relatively long wakeup time. The detailed formula is shown in Eq. (5.13):

$$N_{act}(t) = \lceil \frac{\gamma \lambda^{'}(t)}{k} \rceil, \tag{5.13}$$

where the operator $\lceil x \rceil$ is the ceiling function that gives the least integer greater or equal to x.

However, when aiming to provide FR service, Eq. (5.13) cannot fully exploit the design redundancy introduced by γ. Here we propose to revise it by adding a FR flexibility factor β during the operation:

$$N_{act}(t) = \lceil \beta \frac{\gamma \lambda^{'}(t)}{k} \rceil, N_{act}(t) \in [0, N_0]. \tag{5.14}$$

The greater β is, the more servers are activated for a specific workload. The influence of β on the FR service performance will be investigated in Section 5.3.

5.2.1.3 Aggregated Frequency Control

A PID controller is used to follow the reference power P_{ref} by directly changing f_{agg} at an interval of 4 s.

$$f_{agg}(t) = K_p e(t) + K_i \int_0^t e(x)dx + K_d \frac{\mathrm{d}e(t)}{\mathrm{d}t}, f_{agg}(t) \in [f_{min}, f_{max}] \tag{5.15}$$

$$e(t) = P_{ref}(t) - P_{mea}(t) \tag{5.16}$$

In the above equations, K_p, K_i, and K_d denote the coefficients for the term P, I and D, respectively. e is the control errors between P_{ref} and P_{mea}. The maximum aggregated frequency is 1, while the minimum frequency varies based on the number of active servers due to the constraints of QoS. Details on how to determine f_{min} are described in Section 5.2.1.4.

5.2.1.4 Minimum Aggregate Frequency

The same approach in Ref. [82] is used here to find the minimum allowable aggregated frequency. Using a service response time model shown in Eq. (5.2) and Eq. (5.7), we know that the

response time of the servers depends on the aggregated frequency. If the frequency is low, then it takes relatively long time for the servers to respond to the arrival workload, which means the QoS of the data center is compromised. To enable FR and guarantee the QoS, the aggregated frequency should meet a minimum value. From Eq. (5.2) and Eq. (5.10), we can get

$$f_{agg}(t) > \frac{\lambda(t)}{kN_{act}(t)} \tag{5.17}$$

Combining Eq. (5.17) and Eq. (5.14), we can obtain a lower bound for the aggregated frequency:

$$f_{agg}(t) \geq \frac{\lambda(t)}{\beta\gamma\lambda'(t)} \tag{5.18}$$

To ensure the QoS while providing FR, the response time should satisfy:

$$t_r \leq t_u \tag{5.19}$$

where t_u is the upper response time limit of the data center.

In Eq. (5.7), the service time $t_s = \frac{1}{\mu(t)}$ accounts for the majority of the response time [82]. A necessary condition to guarantee the response time constraint is that

$$t_s = \frac{1}{\mu(t)} \leq t_u \tag{5.20}$$

By substituting Eq. (5.2) into Eq. (5.20), we can get another lower limit of the aggregated frequency:

$$f_{agg}(t) > \frac{1}{kt_u} \tag{5.21}$$

Combining Eq. (5.18) and (5.21), we can define f_{min} as:

$$f_{min} = \max(\frac{\lambda(t)}{\beta\gamma\lambda'(t)}, \frac{1}{kt_u}) \tag{5.22}$$

5.2.2 Cooling Power Management

The cooling system power is managed by resetting the CHWST. The regulation signal from the electrical market is directly used to change the CHWST setpoint $T_{chws,set}$ by Eq. (5.23).

$$T_{chws,set}(t) = T_{chws}(t) - r(t)\Delta T, \tag{5.23}$$

where T_{chws} is the CHWST at current time step, ΔT is the user defined regulation range for the temperature, and varies based on the design supply temperature range of chillers. Here we set $\Delta T = 2$ °C. The negative sign at the right term means when regulation up is needed, the temperature setpoint should be reduced, and vice versa.

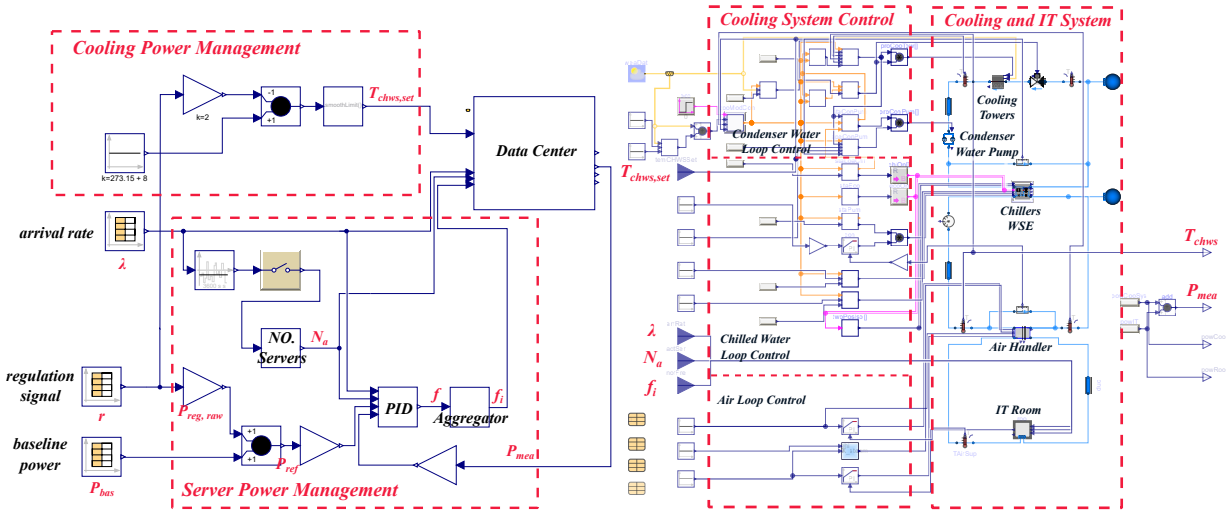
5.3 Parametric Study

The purpose of this case study is to investigate the performance of the proposed control strategy for tests with RegA and RegD signals. Via the case study, we try to understand how the FR service performance can be affected by some important factors, such as regulation capacity bid, FR flexibility factor, thermal response time of the chiller, workload condition, and cooling mode of the cooling system.

5.3.1 Simulation Setup

The considered data center is located in Chicago, which is in ASHRAE Climate Zone 5A (Cool Humid) and within the PJM market territory. The configuration of the cooling system is shown in Figure 1.7. The number of servers in the data center is 8000. The design factor γ is set to 1.5 [82]. The total nominal electrical load is 2680 kW, with a design power usage effectiveness of 1.35. The calibrated coefficients for Eq. (5.1) are $b_0 = 0.0154$, $b_1 = 1.5837$, $b_2 = 0.1373$, $c_0 = -22.3540$ and $c_1 = 121.0212$ using the method mentioned in Ref. [82], with a mean absolute percentage error of 3.6%. When not providing FR, the server aggregator operates at a frequency of 0.8, and the CHWST setpoint is set to 8 °C. A typical workload from a web service data center is normalized and used here as shown in Figure 5.3 [82]. The test signal for 2019 is downloaded from the PJM homepage as shown in Figure 5.4 [106]. To guarantee the QoS of the data center, the maximum average response time is set to 6 ms. The simulation is performed in a Modelica-based environment as shown in Figure 5.2 using models in [48, 49, 50].

This paper adopted the cooling system control in a previous study [47]. The cooling tower fan speed is controlled to satisfy the requirement of temperature setpoint under the maximum fan



(a) FR Controller (b) System Model

Figure 5.2: Modelica implementation of the studied data center for FR service

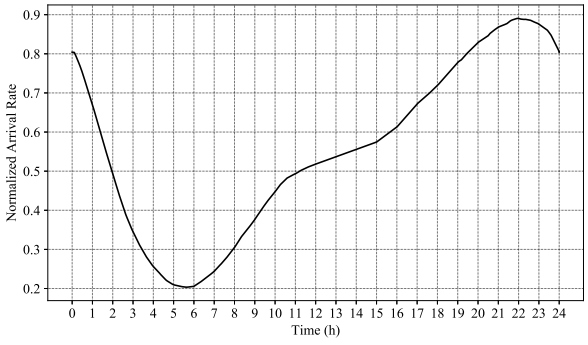
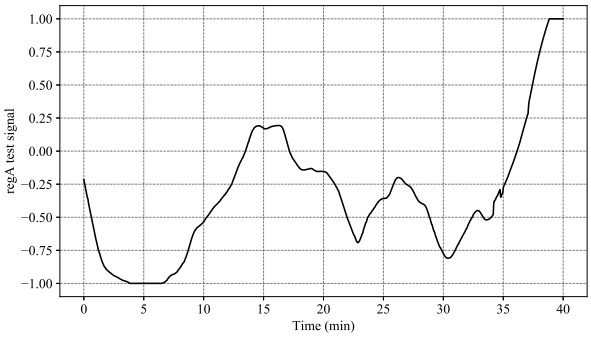
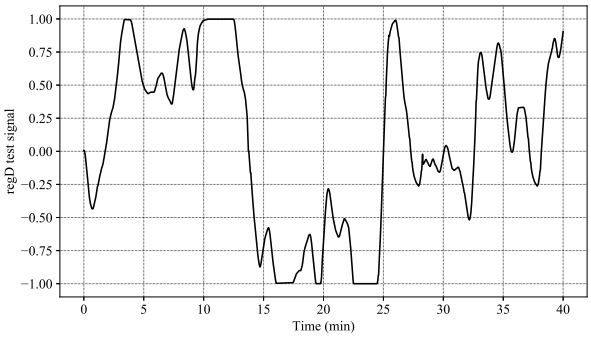


Figure 5.3: Normalized daily workload



(a) RegA



(b) RegD

Figure 5.4: Raw test signal from the PJM market

speed. In FC mode, the fan speed is controlled to maintain a predefined CHWST at the downstream of the economizer, and not exceed the predefined maximum fan speed that is 90% of the normal speed. In PMC and FMC modes, the fan speed is controlled to maintain the supply condenser water at its setpoint. The FR performance will be influenced by this local control in the cooling system because the CHWST setpoint is adjusted in the proposed strategy.

5.3.2 Simulation Scenarios

We swept the following parameters (Table 5.1) to investigate the FR performance in the studied data center. The regulation capacity is chosen as 5%, 10%, and 15% of the design electrical load, which is 134 kW, 268 kW and 402 kW, respectively. The FR flexibility factor β is set to 0.9, 1.0, and 1.1, to investigate the influence of the operational redundancy on the performance of regulation. The chiller's thermal response time τ is set to 5 min, 10 min and 15 min, which can reflect different types of chillers as referred in [114, 115]. As shown in Figure 5.3, three different workloads are analyzed and compared: light, medium and heavy, which happen during 5:00-6:00, 12:00-13:00 and 22:00-23:00, respectively. The simulation is conducted for a cold day when the cooling system is in FC mode, and a hot day when the cooling system is in FMC mode. Both RegA and RegD are evaluated in this case study.

Table 5.1: Swept parameters for FR service

Parameters	Values	Units	Comments
C_{reg}	134, 268, 402	kW	regulation capacity
β	0.9, 1.0, 1.1	–	FR flexibility factor
τ	5, 10, 15	min	time constant of the chiller's response to the CHWST setpoint
workload	light, medium, heavy	–	requested IT service
cooling mode	FC, FMC	–	cooling mode that determines the activation and deactivation of different cooling sources

5.3.3 Results and Discussion

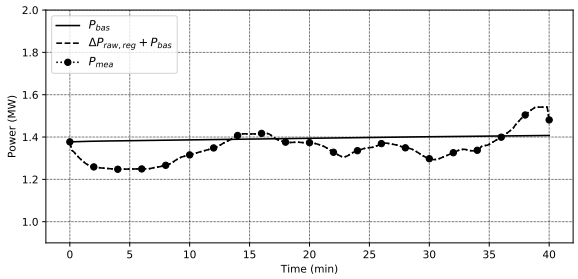
Using the above settings, numerical simulations were performed and the results are presented in the following subsections.

5.3.3.1 Regulation Capacity Bid

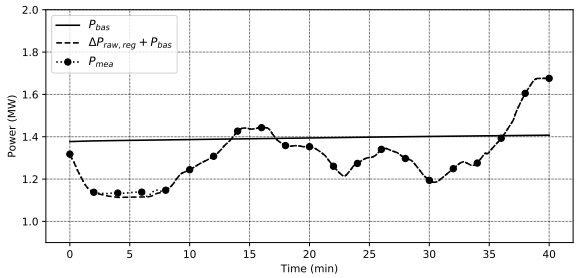
The regulation capacity bid C_{reg} has a major influence on the FR service performance in PJM market, especially when regulation down is required. The larger the bidding regulation capacity is, the worse the service performance is. For example, in Table A.5, the performance score defined by Eq. (2.5) decreases from around 0.98 to around 0.89 as C_{reg} increases from 5% to 15% at $\beta = 1.1$.

The decrease is due to insufficient regulation capacity as shown in Figure 5.5. When the bidding capacity is 5%, the system can generally track the reference signal in an accurate way. Both regulation down and regulation up requests can be met. When the bidding capacity is 10% and 15%, the regulation performance is mainly influenced by regulation down because the maximum regulation down capacity is achieved at the minimum aggregated frequency required by the constraints of QoS. Because there are sufficient number of active servers, regulation up can be met for both bids (10% and 15%). If the bid is further increased to a relatively large value, with specific amount of active severs operating at their maximum frequency, the data center cannot meet the regulation up either. Figure 5.5 also shows that the data center provides asymmetric regulation up and regulation down capacity. It is easier to provide regulation up because of the design redundancy introduced by γ. Regulation down is constrained by the QoS.

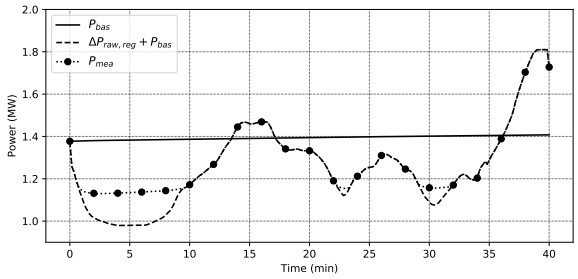
Figure 5.6 shows the average response time of the servers whiling providing FR service. When the bid is small (e.g., 5%), the average response time is within 6 ms as required. However, when the bid increases, the proposed strategy cannot strictly constraint the average response time to be within 6 ms. The violations happen only when the system cannot provide sufficient regulation down capacity. When regulation down is required, the servers work at a low frequency constrained by the minimum frequency as defined in Eq. (5.20). However, because Eq. (5.20) is not a sufficient



(a) $C_{reg} = 5\%$

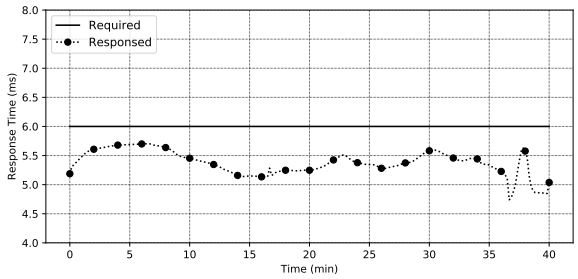


(b) $C_{reg} = 10\%$

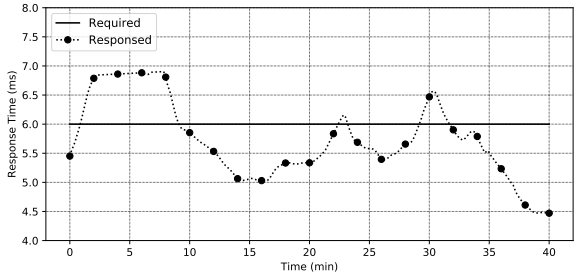


(c) $C_{reg} = 15\%$

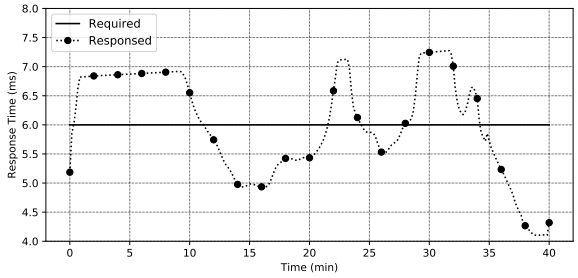
Figure 5.5: Detailed signal tracking for RegA test at $\beta = 1.1$, $\tau = 5$, medium load, FMC mode and different bids



(a) $C_{reg} = 5\%$



(b) $C_{reg} = 10\%$



(c) $C_{reg} = 15\%$

Figure 5.6: Detailed response time for RegA test at $\beta = 1.1$, $\tau = 5$, medium load, FMC mode and different bids

condition, it happens that the system would violate the constraints when it cannot provide enough regulation down capacity.

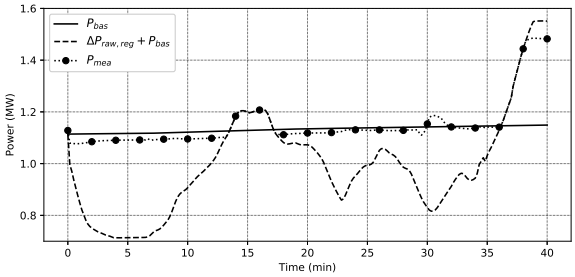
5.3.3.2 Flexibility Factor

The larger β is, the more servers are activated for a specific workload. When more servers are activated, based on Eq. (5.1), more base power related to the amount of active servers is needed, which can increase the regulation up capacity. In addition, more active servers means a smaller minimum aggregated frequency as shown in Eq. (5.22), which can decrease the power usage for the frequency-related term as shown in Eq. (5.1). The tradeoff of power consumption between the decreased frequency and increased number of active servers determines if regulation down capacity can be increased or decreased.

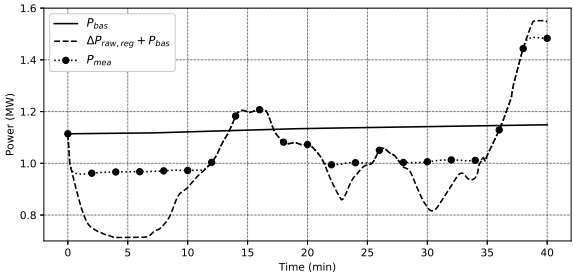
Generally, the proposed β can significantly improve the FR performance for almost all the scenarios when it is appropriately tuned. Figure 5.7 shows detailed signal tracking performance for a RegA test with a medium load in FC mode. The regulation down capacity can be increased as β increases, which then significantly improves the performance score as shown in Table A.2.

For most scenarios, when β is less than 1, there is a significant degradation of the FR service compared to $\beta = 1.0$. For example, in Table A.3, the performance score is only around 0.55 when β is 0.9. The reason can be seen in Figure 5.7(a), which shows that the system can barely provide regulation down service at a small β. This inability is due to the large allowable minimum aggregated frequency for FR service that is illustrated in Figure 5.8. The minimum aggregate frequency decreases from about 0.80 to 0.67 as β increases from 0.9 to 1.1. When β is 0.9, and regulation down service is required, the server aggregator can only work at its minimum frequency (around 0.80), which leads to a similar power consumption compared with the baseline. When β is 1.1, because the minimum frequency is decreased to 0.67, more regulation down capacity can be provided.

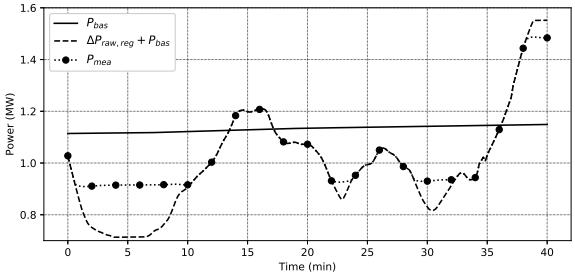
When the bidding regulation capacity C_{reg} is small (e.g. 134 kW) at medium and heavy workload, the performance of FR service is almost the same for β=1.0 and β=1.1 (e.g., Table A.2



(a) $\beta = 0.9$

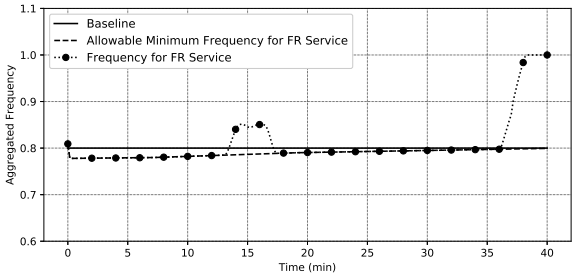


(b) $\beta = 1.0$

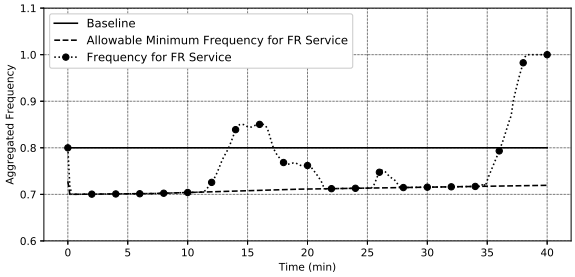


(c) $\beta = 1.1$

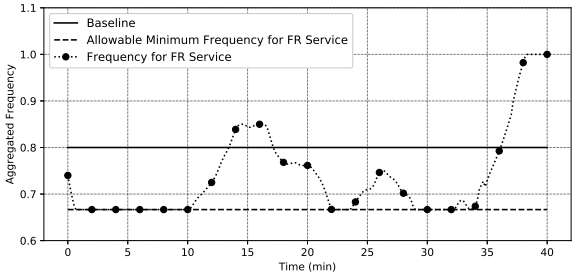
Figure 5.7: Detailed signal tracking for RegA test at $C_{reg} = 15\%$, $\tau = 5$, medium load, FC mode and different β



(a) $\beta = 0.9$



(b) $\beta = 1.0$



(c) $\beta = 1.1$

Figure 5.8: Controlled frequency for RegA test at $C_{reg} = 15\%$, $\tau = 5$, medium load, FC mode and different β

and Table A.3). The reason is when C_{reg} is small, the aggregated frequency only needs be somewhere between the minimum frequency and the maximum frequency, which can track the reference power so well that both regulation down and regulation up can be met.

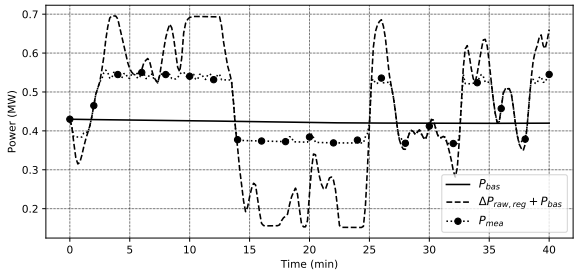
5.3.3.3 Other Parameters

Thermal time constant of the chillers: It has little influence on the FR service using the proposed synergistic control strategy as shown in Table A.1 to Table A.12. In FC mode, the chillers are off, therefore little influence can be observed. In FMC mode, the chillers are activated to provide cooling. Although they are slow-response resources, the delays can be compensated by the fast-response resource, e.g., servers that act like a battery system.

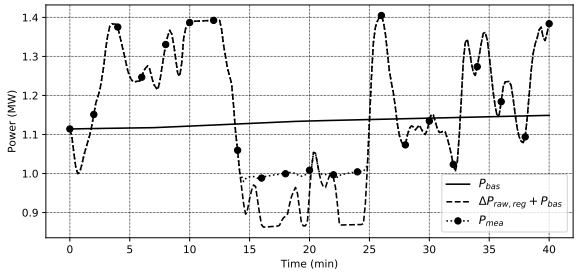
Workload: The larger the workload is, with the same amount of bidding capacity, the better the regulation performance is. For example, Figure 5.9 compares the detailed power signal tracking for a RegD test under different workloads but with the same C_{reg}, β, τ and cooling mode. With a larger workload, more servers are activated based on Eq. (5.14), which can subsequently provide more regulation up and regulation down, thus improving the regulation performance.

Cooling mode: The proposed strategy for both RegA and RegD test can provide better performance in FMC mode than FC mode. For example, by comparing Table A.5 and Table A.2, we can find out that the performance score is about 0.90 in FMC mode and about 0.83 in FC mode when bidding 268 kW with a β of 1.0. The reason is that when the data center works in FMC mode, where chillers are activated to provide cooling, the proposed strategy can provide more regulation down capacity as shown in Figure 5.10.

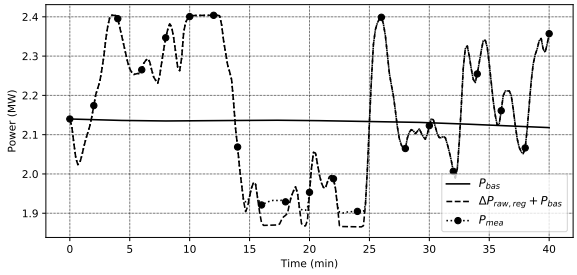
In summary, the simulation results show that the data center using the proposed control strategy is able to participate in the PJM regulation market as a new resource when the parameters are well-tuned. The regulation performance is largely influenced by the bidding regulation capacity C_{reg}, FR flexibility factor β, workload condition and cooling mode of the cooling system, and minimally influenced by the thermal time constant of chillers τ.



(a) Light load

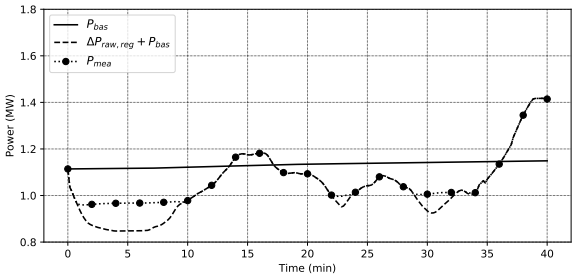


(b) Medium load

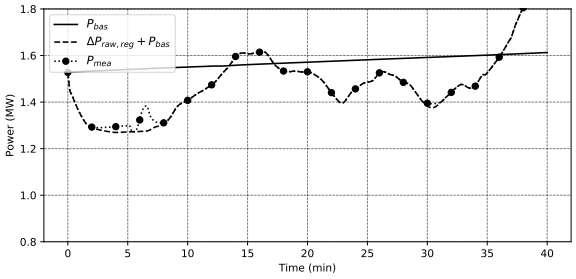


(c) Heavy load

Figure 5.9: Detailed signal tracking for RegD test at $C_{reg} = 10\%$, $\beta = 1.0$, $\tau = 10$, FC mode, and different workloads



(a) FC mode



(b) FMC mode

Figure 5.10: Detailed signal tracking for the RegA tests at $C_{reg} = 10\%$, $\beta = 1.0$, $\tau = 10$, medium load, and different cooling modes

5.4 Synergistic Control versus Servers-only Control

As stated in the literature review, most of current researches focus on servers-only strategies, where only the CPU frequency of the IT servers is adjusted to respond to the regulation signal. Few studies consider synergistic control strategies, especially the cooling system and the servers, for FR service in data centers. This section aims to numerically investigate the benefits of including cooling system in the FR control strategy. Two strategies are compared in this section. s_1 - the servers-only control strategy that only adjusts the aggregator frequency to respond to regulation signal; s_2 - the proposed synergistic control strategy that adjusts the aggregator frequency and reset CHWST simultaneously to respond to regulation signal. The strategy that only resets CHWST is not considered here, because of its relatively slow response and small capacity. The comparison of s_1 and s_2 focuses on the different maximum regulation capacities for each strategy under FMC mode and FC mode.

The regulation capacity is identified as the maximum symmetric power range that the data center can operate within. Data centers have a nonlinear baseline since cooling system operation changes intra-hour and hour-to-hour in response to varying weather, equipment staging and workloads. Similarly, potential regulation capacity varies throughout the day as a function, for example, of weather, workloads, and how the cooling system respond to the workloads and weather etc. In this section, we use a simulation-based environment to determine the regulation capacity for each time step. Here we set the time step to 1 h, because typically in PJM regulation market, the regulation capacity bid can be updated on an hour basis.

The regulation capacity can be found using a model perturbation method as introduced in [104]. The model perturbation method uses mathematical models to study the relationship between the data center system power response and changes in the control inputs (e.g., f_{agg} in s_1) at each time step. Figure 5.11 illustrates the process of determining the regulation capacity of the data center using s_1. For example, at 12:00, the model is assumed to be tracking a baseline aggregated frequency, and it intends to determine the regulation capacity for the current hour

(12:00-13:00). The control input f_{agg} is adjusted from the baseline by simulating 0.02 increments between a minimum aggregated frequency (calculated from Eq. (5.22) to a maximum aggregated frequency of 1. Simulation power responses are then compared with the baseline power to determine the regulation up capacity and regulation down capacity. The symmetric regulation capacity of the whole system is then determined by the minimum of the regulation up and regulation down capacity. The same approach is also used for s_2, which adjusts an additional control input CHWST with an increment of 0.5 °C between an lower limit of 6 °C and 10 °C.

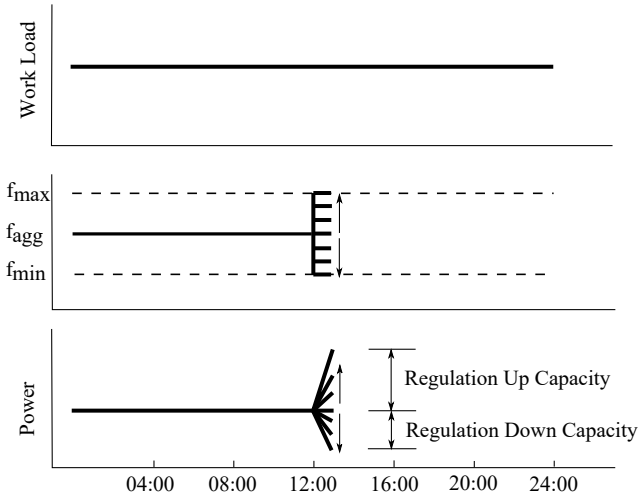
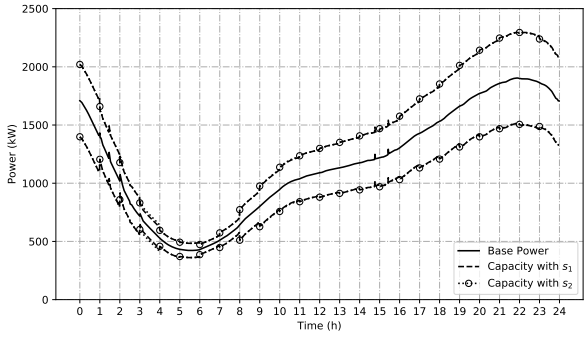
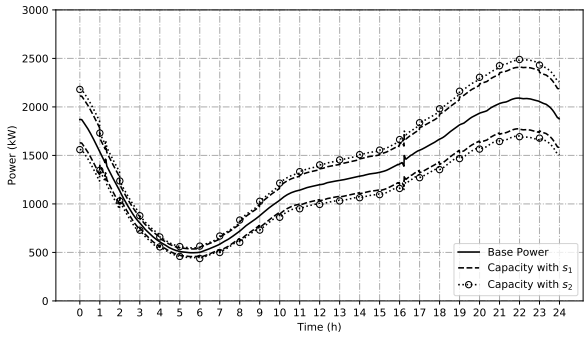


Figure 5.11: Illustration of model perturbation method to find the regulation capacity

Figure 5.12 shows the regulation capacity for the data center operating in FC mode and FMC mode with $\beta = 1.1$, $\tau = 5$ min and RegA signal. In FC mode, the data center can provide almost same regulation capacity using s_1 and s_2. To achieve regulation down capacity, servers work at the allowable minimum frequency in both strategies, thus consume the same amount of power for a given workload. However, the CHWST is reset to 10 °C in s_2 instead of 8 °C in s_1. To address the same cooling load, the higher the CHWST is, the more power is consumed by chilled water pumps and CRAH fans to maintain the same supply air temperature and room temperature, and the less power can be consumed by cooling tower fans because of the CHWST control logic during FC mode described in Section 5.3.1. The constant-speed condenser water pumps consume the same amount of power in the two strategies. The tradeoff among the power changes of cooling towers, pumps,



(a) FC mode



(b) FMC mode

Figure 5.12: Comparison of regulation capacities at different cooling modes

and CRAH fans determines how much more or less capacity s_2 can provide compared with s_1.

In FMC mode, s_2 can provide 18 - 76 kW more regulation capacity than s_1. The minimum regulation capacity for both strategies occurs at around 4am-7am when the workload is the lightest. For example, the capacity is about 34 kW (1.2% of the nominal data center power) for s_2 and 16 kW (0.6% of the nominal data center power) for s_1 at 4am. The maximum regulation capacity is at around 9pm-11pm when the workload is the heaviest. The capacity at 10pm by s_2 is about 460 kW (17% of the data center nominal power), and that by s_1 is about 384 kW (14% of the data center nominal power). In FMC mode, s_2 outperforms s_1 in terms of regulation capacity because increasing CHWST can reduce the power consumption of the chillers. The tradeoff power changes of chillers, CRAH fans, pumps and cooling towers determines that increasing CHWST in the proposed range can reduce the cooling system power consumption, thus increase the regulation down capacity. Note that the CHWST range should be case-specific in a different cooling system based on the curve between the system efficiency and CHWST.

In summary, the proposed synergistic control strategy can better extract the extra potentials of the cooling system together with servers to provide FR service for power grids, especially when chillers are activated. Without chillers, the cooling system consumes a small amount of power, and resetting CHWST can only provide an insignificant regulation capacity.

5.5 Multi-market Optimization

5.5.1 Real-time Optimization Framework

A real-time optimization framework is applied for optimizing the operation of the data center without thermal storage system in the presence of real-time (or day-ahead) energy prices, peak demand charges, and frequency regulation revenue. For each optimization time step, the overall

objective can be described as:

$$\begin{aligned} \min \quad & J(C_{reg}) = E_{cos} + D_{pen} - R_{rev} \\ s.t. \quad & 0 \leq C_{reg}(t) \leq C_{reg,max}(t) \\ & t_r(t) \leq t_{r,u} \\ & s(t) \geq s_l \end{aligned} \tag{5.24}$$

where C_{reg} is the regulation capacity bid at each hour, $C_{reg,max}$ is the maximum capacity the data center can provide for regulation, t_r is the response time of the data center service, $t_{r,u}$ is the allowable upper limit of the response time, s is the regulation performance score defined by PJM, and s_l is the lowest allowable performance score by PJM to participate in regulation market. The cost function J has three terms: energy cost E_{cos}, demand penalty D_{cos} and regulation revenue R_{rev}. The energy cost is calculated by Eq. (5.25).

$$E_{cos} = \int_{t}^{t+\Delta t} p_{em}(t) P_{DC}(t) dt \tag{5.25}$$

where p_{em} is the real-time price signals for energy use at time t, P_{DC} is the total power consumption for the data center at time t. The calculation period starts from time t and ends at $t + \Delta t$, where Δt is the optimization step, and is set to 1 hour in this study.

The electric demand during the current optimization horizon is penalized by the demand price p_{dm} as shown in Eq. (5.26).

$$D_{pen} = p_{dm} \cdot \max((P_{dm} - P_{dm,lim}), 0) \tag{5.26}$$

where p_{dm} is the demand price, P_{dm} is the power demand calculated as the average power for each 30-min interval, and $P_{dm,lim}$ is the limit of required demand. This function means if the demand in current step exceeds a predefined demand value, then the optimization cost function is penalized by the demand difference. Otherwise, no penalization is applied. Note that p_{dm} and P_{dm} are both utility specific, and may vary from this definition.

The revenues from regulation service is computed as follows [104, 132].

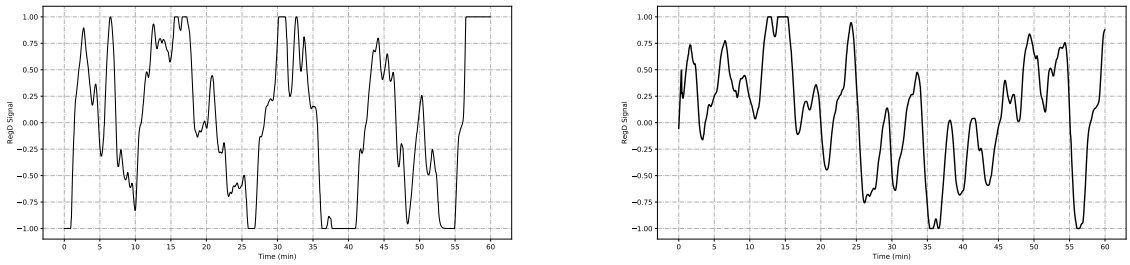
$$R_{rev} = \int_{t}^{t+\Delta t} p_{rm}(t) C_{reg}(t) dt \tag{5.27}$$

where p_{rm} is the real-time price signal from the regulation market, and $C_{reg}(t)$ is the regulation capacity bid for each time step.

The price signals such as p_{em} and p_{rm} need to be predicted one optimization step ahead, e.g. 1 hour in this study. Many researches have been conducted for this purpose. In this paper, historical prices of these two electrical markets are used, which means the hourly ahead prices are assumed to be perfectly predicted. The demand limit $P_{dm,lim}$ can also be predefined by the data center operators based on historical operation conditions. The maximum regulation capacity is calculated using the model perturbation method mentioned in Section 5.4.

5.5.2 Case Study

A data center as shown in Figure 5.2 is used to investigate the benefits from participating in different electrical markets. The data center is considered as a price taker only. This case study investigates the maximum benefits that data centers can obtain from both the real-time energy market and the regulation market in PJM. For the regulation service, only dynamic regulation is studied here, because its price is usually much higher than traditional regulation.



(a) January (b) July

Figure 5.13: Example of one-hour historical RegD signal of the PJM market in January and July

5.5.2.1 Case Description

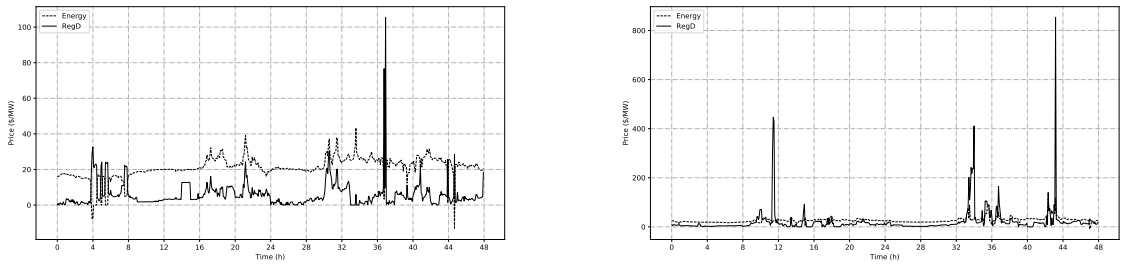
The data center is located in Chicago, which is in ASHRAE Climate Zone 5A and within the PJM market territory. For the cooling system, there are two chillers and one integrated waterside economizer providing cooling to the data center room. This cooling system can operate in three modes: Free Cooling (FC) mode when only the WSE is enabled for cooling, Partial Mechanical Cooling (PMC) mode when the chiller and WSE are both triggered, and Full Mechanical Cooling (FMC) mode when only the chiller is activated. There are also two cooling towers, two constant-speed condenser water pumps, two variable-speed chilled water pumps, and one variable speed fan. The cooling system and its control are modelled using an open-source equation-based Modelica environment [48, 47, 49, 50].

For the IT system, the design number of servers is 8000. The design factor γ is set to 1.5 [82]. The total nominal electrical load is about 2700 kW. The calibrated coefficients for Eq. (5.1) are $b_0 = 0.0154$, $b_1 = 1.5837$, $b_2 = 0.1373$, $c_0 = -22.3540$ and $c_1 = 121.0212$ using the method mentioned in Ref. [82]. When not providing FR, the server aggregator operates at a frequency of 0.8 with a regulation flexibility factor of 1.0, and the chilled water temperature setpoint is set to 8 °C.

For the multi-market optimization, all the settings are the same as the baseline except an additional frequency regulation controller as designed in Section 5.2 is used to provide regulation service for the grids by adjusting the CPU frequency and chilled water temperature setpoint. The regulation flexibility factor is set to 1.1 when providing regulation services. The QoS when providing regulation services is guaranteed by constraining the average response time of the data center service to 6 ms. The lower limit of the performance score in PJM to disqualify a regulation resource is 0.4 [89]. Here we set it to a higher value, 0.9. The real-time optimization is performed at a one-hour interval for 2 days in both January (when cooling system operates at FC mode) and July (when cooling system operates at FMC mode).

The price signals of the real-time energy market and the regulation market over the opti-

mization period is plotted as shown in Figure 5.14. A real-time web service in Wikipedia [132] is used as the workload arrival profile, which is shown in Figure 5.15.



(a) January (b) July

Figure 5.14: Two-day historical real-time prices of PJM energy market

5.5.2.2 Results and Discussions

Table 5.2 compares the total cost of the data center in terms of baseline operation and multi-market optimization. The baseline system is denoted as *BL*, and the multi-market optimization is denoted as *BL+MM*. In both January and July, the data center without energy storage systems, using the proposed optimization framework, can benefit from participating in both energy market and regulation market. In the two days considered, *BL+MM* can save $123.6 in July, while the saving is $24.8 in January.

The savings mainly come from the revenues in the regulation market, and the cost for energy use and demand charge are almost the same in the *BL* and *BL+MM*. Because the sum of the RegD signal over a long time period (e.g. 1 hour) is almost 0, providing regulation service in the *BL+MM* leads to the similar energy use, thus similar energy cost compared with the *BL* where no regulation service is provided. By utilizing the demand cost defined in Eq. (6.14), the data center can provide regulation service without increasing monthly demand, thus no extra demand charge would be added to utility bills. The revenue from July is much higher than that in January because the price for dynamic regulation (RegD) resources is higher in July. As shown in Figure 5.14, during the studied two days, the average price from regulation market in July is about 21 $/MW, while that

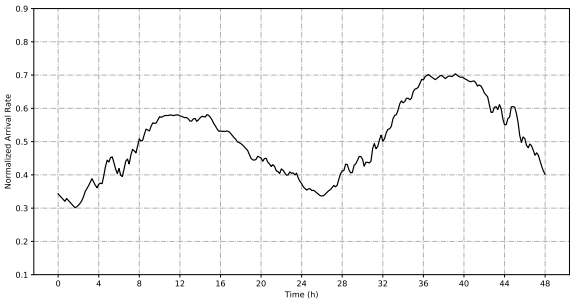
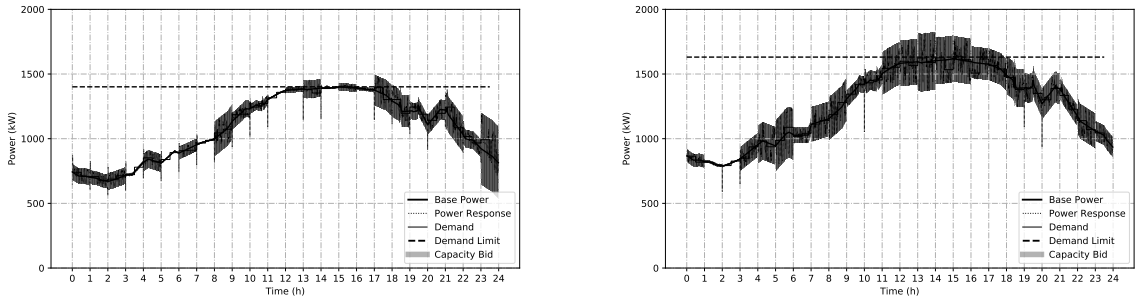


Figure 5.15: Two-day historical arrival rates in the data center

Table 5.2: Multi-market optimization of GEDCs without TESS

Costs	January		July	
	BL	*BL+MM*	*BL*	*BL+MM*
Energy Cost ($)	1043.3	1042.9	1591.4	1590.6
Demand Cost ($)	10459.3	10457.5	12063.9	12062.8
Regulation Revenue ($)		-22.6		-121.7
Total Cost ($)	11502.6	11477.8	13655.3	13531.7

in January is only about 5.8 $/MW.



(a) January (b) July

Figure 5.16: Optimal hourly FR capacity bids

Figure 5.16 shows the hourly capacity bids in from hour 0 to hour 24 in both January and July. The demand for each 30 minutes is denoted as the thin solid line. The demand limit used for demand cost as shown in Eq. (6.14) is denoted as the dashed line. The optimal capacity bid at each hour is denoted as the shaded area. At non-peak hours (e.g., 3:00 - 6:00), the optimal bid is mainly influenced by the price from energy market, price from regulation market and detailed shape of RegD signal. Because the demand is lower than the demand limit, the tradeoff between the energy cost and revenues from regulation market determines the optimal bid. The energy cost is highly influenced by the energy use, which is determined by the detailed shape of the RegD signal. If the sum of the RegD signal is larger than 0, then more energy would be consumed when providing frequency regulation service, thus the energy cost would increase. Although the energy cost increases in this case, the data center can get revenues from regulation market. If the sum of the RegD signal is no large than 0, then at that hour, the data center can bid at their maximum capacity.

At peak hours (e.g., 12:00 - 16:00), the optimal bid is mostly influenced by the demand limit and the RegD signal. Figure 5.16 shows that at these hours, the bid is small so that the demand cannot exceed the required demand limit to avoid demand penalty. At 13:00, the bid is about 69 kW, but it is only about 5 kW at 14:00. The difference is caused by the detailed shapes of

the RegD signals in these two hours. At 13:00, the sum of the RegD signal in first 30 minutes is slightly greater than 0, but in the second 30 minutes it is much smaller than 0. This means that regulation capacity bid in this hour can increase the demand in the first 30 minutes, but the demand in the second 30 minutes can be decreased compared with the same time in the baseline system. Therefore, at this hour, the data center can bid a large capacity as long as the demand in the first 30 minutes will not exceed the demand limit. The same situation happens at 14:00 but with a large sum of RegD signal at the first 30 minutes. Also because the power at 14:00 is much closer to the demand limit, the data center can only bid a small capacity at this hour.

5.6 Summary

Section 5.2 proposed a synergistic control strategy for data centers to provide FR service in an electric market. A FR flexibility factor which has been shown to improve regulation performance scores was also introduced.

Section 5.3 investigated the important factors that can affect the FR performance in the PJM market through a case study that sweeps different parameters with different values at a whole-system level. Simulation results showed that the data center can provide a regulation capacity as large as 17% of its nominal power. Furthermore, the performance of the data center providing FR service for both RegA and RegD using the proposed strategy is largely influenced by the bidding regulation capacity C_{reg}, FR flexibility factor β, workload condition and cooling mode of the cooling system, and minimally influenced by the time constant of chillers τ. The main findings of the paper can be summarized as follows:

- The larger the bidding regulation capacity is, the worse the FR service performance is. The performance is degraded mainly because of the insufficient regulation down capacity, which is determined by the minimum aggregated frequency and the number of active servers.
- The proposed flexibility factor β can improve the FR service when tuned appropriately according to the coefficients of the aggregator model. In this paper, the preferable lower

limit is 1.0, and the upper limit should be decided by considering the power tradeoff between more activated servers and increased minimum aggregated frequency.

- Thermal time constant of the chillers τ has minimal influence on the provided FR service, because the fast-response resources (e.g., servers) can compensate the delays caused by the slow-response resources (e.g., chillers).

- The larger the workload is, with the same amount of bidding capacity, the better the regulation performance is.

- The system can provide better FR service in FMC mode than FC mode, because larger regulation down capacity can be provided in the former.

Section 5.4 compared the FR capacity the data center can provide with different control strategies. The proposed synergistic control strategy can provide an extra regulation (76 kW) of 3% of the design power in the data center during FMC mode compared with a server-only control strategy, while in FC mode these two strategies have almost the same regulation capacity. The capacity difference is mainly caused by the tradeoff of power changes among different cooling equipment in different operational conditions.

Section 5.5 developed a real-time multi-market optimization framework for the data center without storage systems to maximize their benefits from participating in both energy market and regulation market. Then, a case study was conducted to numerically investigate the optimal bids at each hour by considering the energy cost, demand costs and regulation revenues using a virtual data center located in PJM. Simulation results shows that using the proposed multi-market optimization framework can minimize the operational cost. Compared with the baseline system, providing frequency regulation service over the considered two days can save \$24.8 in January and \$123.6 in July.

Chapter 6

Multi-market Optimization for GEDCs with Thermal Energy Storage Systems

Based on:

Yangyang Fu, Wangda Zuo, Xu Han, Kyri Baker, "Optimizing the Participation of Data Centers with Thermal Energy Storage System in Energy and Ancillary Service Market", in preparation.

Previous chapter provides a multi-market optimization framework for GEDCs without energy storage systems. One major disadvantage is that the optimization framework cannot limit the power demand when providing frequency regulation service because the studied GEDC have no energy storage system that can be utilized to shift loads.

This chapter introduces a TESS into the GEDC, and provides a synergistic control strategy to enable the provision of FR service by using the cooling system with TESS and IT equipment and to reduce the power demand. A multi-market optimization framework using model predictive control schemes is also proposed to minimize the operation costs in GEDCs. Simulation results show that utilizing the TESS can not only reduce energy costs and demand charges but also harness FR revenues. The data center with TESS can obtain $ 182 (89.5%) more FR revenues compared to that without TESS in two days. What's more, the proposed multi-market optimization framework for the data center with TESS can reduce the operation costs by $1793.2 in two days, saving 9.7% compared with a baseline data center without TESS. The economic baseline can help the system save $1186.2 in the energy and the regulation market in two days, saving 6.6% compared to a data center without an economic baseline.

6.1 Thermal Energy Storage System

Energy storage systems has been continuously proposed to reduce the peak power in data centers. During a power valley, the energy storage system can be charged so that the power drawn from the power grid can be maintained approximately constant, and during a power peak, the energy storage system is discharged so that the electrical demand can be met.

There are two types of energy storage system in general: electrical energy storage system (EESS) and thermal energy storage system (TESS). The EESS, such as uninterruptible power supply (UPS) batteries, are commonly used in data centers. Many researches have focused on shaving the server-side power demand using the EESS. However, these batteries have several limitations to serve the power shaving purpose. First, batteries in data centers have limited capacities. The design purpose of batteries in data centers is to serve critical equipment (e.g., IT equipment, air fans etc.) for a short time during an emergency situation, such as an power outage. They are typically designed to provide power for 5-15 minutes. Second, the performance of batteries deteriorates very fast, especially when fast charging and discharging are required. Third, batteries are currently expensive, requiring high capital costs and recycling costs. Therefore, many researchers have shifted their attentions to find an alternative of the EESS.

TESS, because of their cheap costs, has gained increasing attentions for power shaving in data centers. Many industrial data centers, such as a Google data center in Taiwan, have adopted TESS to avoid high operation costs during on-peak periods [72]. There are also some researches on utilizing TESS for power management in data centers. Zheng et al. proposed a TESS to shave the data center power profile, which can lead to a significant amount of capital and operation cost savings [149]. Zhang et al. designed and evaluated a cooling strategy that exploits thermal and electrical energy storage techniques to cut the electricity bill for data center cooling, without causing servers in a data center to overheat [147]. Oro et al. overviewd the energy saving potentials of direct air cooling and TESS in data centers, and concluded that when using TESS in combination with an off-peak electricity tariff the operational cooling cost can be drastically reduced [99].

However, none of them considered the potential of using TESS to provide ancillary service such as frequency regulation (FR) to the power grid. There are multiple reasons, on top of which may be the concerns of the large thermal time constant (e.g, hours) of TESS compared with the time scale (e.g., seconds) required by FR.

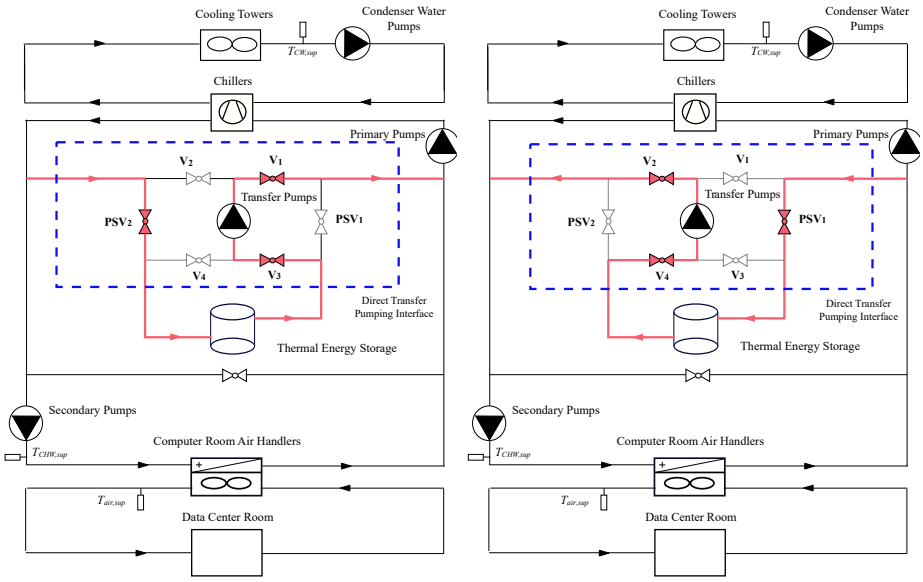
6.1.1 Stratified Chilled Water Storage System

Stratified chilled-water storage systems have been successfully applied in hospitals, schools, industrial facilities, campus and district cooling systems, power generation (combustion turbine inlet air cooling), emergency cooling systems, and others. A typical inside-the-plant connection scheme for a stratified sensible storage system serving a primary/secondary chilled water system is shown in Figure 6.1(a) [5]. The primary pumps operate at a constant speed, while the secondary pumps operate at various speeds. Besides the traditional configuration of a primary-secondary pump system, a direct transfer pumping interface is installed between the TESS and the chillers in order to allow chilled water from the TESS to be pumped directly into the chilled water supply system while return water flows from the system to the tank. The direct transfer pump interface is composed of several pumps and valves. Valves V_1 to V_4 are shutoff valves to control the direction of the fluid flow. Pressure-sustaining valves (PSVs) PSV_1 and PSV_2 are modulating valves that typically are used to control the flow rate of the passing fluid. Transfer pumps provide pressurized flow to match with the system level pressure.

6.1.2 Operating Modes

The stratified chilled-water storage system for a data center usually can operate in four operating modes. Different control strategies have been proposed to switch the cooling system among these different operating modes.

- M_1: charging storage while meeting loads. Under this mode, chillers are activated to remove heat from storage and the data center room.



(a) Charging (b) Discharging

Figure 6.1: Chilled water system with a TEES

- M_2: meeting loads from storage only. Under this mode, the cooling system discharges storage to meet cooling loads without operating chillers.
- M_3: meeting loads from storage and chillers. Under this mode, the cooling system is required to discharge storage and operate chillers to meet the cooling loads simultaneously.
- M_4: meeting loads from chillers only. Under this mode, only chillers are activated to meet the cooling loads, and there is no fluid flow to or from storage.

The charging and discharging of a stratified chilled-water storage is realized by operating the direct transfer pump interface as shown in Table 6.1 and Figure 6.1. The charging and discharging process can be characterized by charging rate u_s as detailed in Section 6.2.1. The charging rate describes the "cooling energy" that being added or removed from the storage in every second. A positive value means charging, and a negative value means discharging.

Table 6.1: Operations of charging and discharging storage

Actions	V_1	V_2	V_3	V_4	PSV_1	PSV_2	Pumps
Charging	open	closed	open	closed	closed	modulated	modulated
Discharging	closed	open	closed	open	modulated	closed	modulated
Idle	open	closed	closed	closed	closed	closed	closed

6.2 Models and Control

The section describes the mathematical models and control strategies that are used in this study. Here only list TESS-related models and control, while the other detailed chilled water system models and control can be referred to Chapter 3.

6.2.1 Stratified Tank Model

Here a concept named "state-of-charge (SoC)" denoted as x in Eq. (6.1) is introduced to describe the level of charge of a chilled water stratified TES tank to its nominal capacity. The SoC is 1 when the tank is fully charged, and 0 when the tank is fully depleted. SoC is difficult to measure

but it can be estimated from the tank temperatures. In this section, the average tank temperature $T_{s,avg}$ among different stratified volumes are used to estimate the current SoC as shown in Eq. (6.1).

$$x(t) = \frac{T_{st,0} - T_{s,avg}(t)}{T_{st,0} - T_{sb,0}}, \tag{6.1}$$

where $T_{st,0}$ and $T_{st,0}$ are the return temperature at the top of the tank and the supply temperature at the bottom of the tank under the nominal condition, respectively.

The charging rate u_s during the charging and discharging process is measured as:

$$u_s(t) = \dot{m}_{sb} C_{p,w} (T_{st} - T_{sb}), \tag{6.2}$$

where the u_s is the charging rate. When the u_s is positive, the tank is charged with cold water, and when the u_s is negative, the tank is discharged to provide cooling. The $\dot{m}_{sb}$ is the mass flowrate at the bottom of the tank. During charging process, the cold water flows into the tank, and the $\dot{m}_{sb}$ is set to a positive value. During discharging process, the cold water flows out of the tank, and the $\dot{m}_{sb}$ is set to a negative value. The T_{st} is the temperature of the water flowing into or out of the top of the tank. The T_{sb} is the temperature of the water flowing into or out of the bottom of the tank.

6.2.2 Operating Mode Control

The cooling system can switch among different operating modes in real time based on two controllable inputs: storage charging rate setpoint $u_{s,set}$ and chiller capacity setpoint $u_{c,set}$. The staging conditions for entering each mode is listed in Table 6.2. The detailed strategies that determines the schedules for $u_{s,set}$ and $u_{c,set}$ are described in Section 6.2.3 and Section 6.2.4, and the detailed operating strategies for the cooling system to track $u_{s,set}$ and $u_{c,set}$ are discussed in Section 6.2.5. Each staging condition has a delay time Δt that introduces extra delays of operations to avoid frequently switching.

In the above table, $u_{c,min}$ and $u_{c,max}$ are the minimum and maximum chiller capacity setpoint. Users can define these two bounds based on the requirements of the chiller manufacture. $u_{s,min}$ and

Table 6.2: Staging conditions among different operating modes

Operating Modes	Staging Conditions
M_1	$u_{c,min} \leq u_{c,set} \leq u_{c,max}$ for Δt, and $0 < u_{s,set} \leq u_{s,max}0$ for Δt
M_2	$0 \leq u_{c,set} < u_{c,min}$ for Δt, and $u_{s,min} \leq u_{s,set} < 0$ for Δt
M_3	$u_{c,min} \leq u_{c,set} \leq u_{c,max}$ for Δt, and $u_{s,min} \leq u_{s,set} < 0$ for Δt
M_4	$u_{c,min} \leq u_{c,set} \leq u_{c,max}$ for Δt, and $u_{s,set} = 0$ for Δt

$u_{s,max}$ are the maximum discharging and charging rate of the tank, which is physically constrained by the system design and operation, such as transfer pump and temperatures in the tank.

6.2.3 Load-shifting Control Strategy

There are few researches focusing on the rule-based control of chilled water storage system in data centers for load shifting purposes. Here a typical storage-priority control strategy designed for commercial buildings in [60] is adopted to provide baseline storage control. The $u_{s,set}$ and $u_{c,set}$ are determined at an interval of one hour. During off-peak periods, the chillers are operated at their full capacity until the tank is full or charged to the SoC with which the cumulative cooling load during the next on-peak period can be met without operating the chillers. Thus short-term cooling load prediction is required. When the storage is full or at the desired level, the chillers operate the capacity necessary to meeting the cooling load. During on-peak periods, the chillers operate at a constant capacity C_0 for the entire on-peak period so that the TESS is just depleted at the end of the on-peak period.

For the chillers, their capacities at each time step can be determined as shown in Eq. (6.3).

$$u_{c,set}(t) = \begin{cases} u_{c,max}(t), & t \text{ is off-peak} \\ C_0 - \dot{Q}^{'}(t), & t \text{ is on-peak} \end{cases} \tag{6.3}$$

The maximum chiller capacity $u_{c,max}$ can be defined in the following equations:

$$u_{c,max}(t) = \min\{C_{chi}, \dot{Q}^{'}(t) + (\dot{m}_{pri} - \dot{m}^{'}_{sec})C_p(T_{s,1} - T_{sb,0})\} \tag{6.4}$$

$$\dot{m}^{'}_{sec} = \frac{\dot{Q}^{'}(t)}{C_p(T_{st,0} - T_{sb,0})} \tag{6.5}$$

where C_{chi} is the nominal chiller capacity, $\dot{Q}^{'}(t)$ is the predicted hourly mean cooling load, $\dot{m}_{pri}$ is the mass flow rate of the primary pumps, $\dot{m}^{'}_{sec}$ is the predicted mass flow rate of the secondary pumps, and C_p is the specific heat of the fluid. $T_{s,1}$ is the current temperature of the first segment of the tank. Eq. (6.4) means the maximum chiller capacity should be the minimum of nominal chiller capacity and the possible maximum cooling load of the system.

The constant chiller capacity C_0 is determined only for on-peak periods and only calculated at the first hour of each on-peak period. Details can be seen as follows.

$$C_0 = \begin{cases} \frac{1}{N}\{\sum_{i=0}^{N-1} \dot{Q}^{'}(t+i) - x(t)\frac{C_{sto}}{dt}\}, & \nu = 0 \\ \frac{\nu}{1-(1-\nu)^N}\{\sum_{i=0}^{N-1}(1-\nu)^{N-1-i}\dot{Q}^{'}(t+i) - (1-\nu)^N x(t)\frac{C_{sto}}{dt}\}, & \nu > 0 \end{cases} \tag{6.6}$$

where i is the hour index at the on-peak periods, and $i = 0$ means the first hour of the current on-peak period. N is the number of hours at each on-peak period. S_{cap} is the nominal storage capacity. dt is the time step used in this controller, and here is set to 1 hour. ν is the storage loss, and $\nu = 0$ means a lossless storage system. Eq. (6.6) means that the chillers should be able to meet the extra cooling loads that cannot be addressed by the remaining storage energy.

For the storage, the charging rate can be scheduled as shown in Eq. (6.7), which means that the charging rate of the storage should be constrained by their maximum physical bounds and the difference between the chiller capacity and the current cooling load.

$$u_{s,set}(t) = \begin{cases} \min\{u_{c,set}(t) - \dot{Q}^{'}(t), u_{s,max}(t)\}, & t \text{ is off-peak} \\ \max\{u_{c,set}(t) - \dot{Q}^{'}(t), u_{s,min}(t)\}, & t \text{ is on-peak} \end{cases} \tag{6.7}$$

The maximum discharging rate $u_{s,min}$ and the maximum charging rate $u_{s,max}$ can be determined using Eq. (6.8) and Eq. (6.9).

$$u_{s,min}(t) = \max\{-\dot{Q}^{'}(t), u_{s,x(t+1)=x_{min}}\} \tag{6.8}$$

$$u_{s,max}(t) = \min\{u_{c,max}(t) - \dot{Q}^{'}(t), u_{s,x(t+1)=x_{max}}\} \tag{6.9}$$

where $u_{s,x(t+1)=x_{min}}$ means the charging rate that leads to the minimum SoC of the storage at next time step, and $u_{s,x(t+1)=x_{max}}$ means the charging rate that leads to the maximum SoC of the storage at next time step. Detailed calculations can be referred to Eq. (6.10) and Eq. (6.11). These two charging rates can guarantee that the tank is maintained at the desired level.

$$u_{s,x(t+1)=x_{min}} = -\frac{x(t) - x_{min}}{dt} S_{cap} \tag{6.10}$$

$$u_{s,x(t+1)=x_{max}} = -\frac{x(t) - x_{max}}{dt} S_{cap} \tag{6.11}$$

6.2.4 Frequency Regulation Control Strategy

A similar synergistic control strategy as proposed in Section 5.2 for the provision of FR service is also used for the data center with a thermal storage system. Details of this strategy is shown in Figure 6.2. The **Economic Baseline**, **Cooling Power Management** and **Capacity Bid** are illustrated in this section while the other parts are the same as described in Section 5.2.

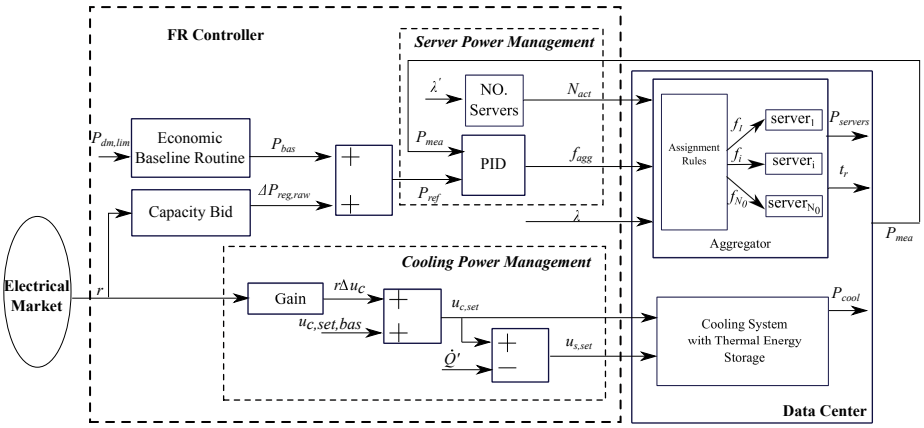


Figure 6.2: Synergistic control strategy for FR service in a data center with a TESS

6.2.4.1 Economic Baseline

The **Economic Baseline** calculates an economic baseline point for the GEDC based on a user-defined demand limit $P_{dm,lim}$ at each time step. The economic baseline point is then used for

the GEDC to provide FR service. The baseline point depends on how the storage is controlled when not providing FR service. Available control strategies include chiller-priority control , and storage-priority control as in Section 6.2.3 etc. However, those rule-based control strategies might not be able to provide optimal control of storage to minimize operation cost. This dissertation proposes a model predictive control (MPC) scheme to find the most economic baseline power profile with limited demand. The MPC controller optimizes the control signal chiller capacity $u_{c,set}$ over a finite prediction horizon PH, and implements the optimal signal to a control horizon CH. Details are shown as follows.

$$\begin{aligned} \min \quad & J(u_{c,set}(t)) = E_{cos}(t) + D_{pen}(t) \\ s.t. \quad & x_{min} \leq x(t) \leq x_{max} \\ & T_{roo,min} \leq T_{roo}(t) \leq T_{roo,max} \\ & u_{c,min,eco} \leq u_{c,set} \leq u_{c,max,eco} \end{aligned} \tag{6.12}$$

where x_{min} and x_{max} are the required minimum and maximum SoCs of the chilled water tank. The T_{roo} is the data center room temperature, and it should be within a range from the lower limit $T_{roo,min}$ and the upper limit $T_{roo,max}$. The $u_{c,min,eco}$ and $u_{c,max,eco}$ are the minimum and maximum controllable chiller capacity for the economic baseline.

The cost function J have two parts: energy cost E_{cos} and demand penalty D_{pen}. The energy cost is calculated by Eq. (6.13).

$$E_{coc} = \int_{t}^{t+PH} p_{em}(t) P_{dc}(t) dt \tag{6.13}$$

where p_{em} is the predicted energy price at time t, P_{dc} is the predicted total power consumption for the data center at time t. The calculation period starts from time t and ends at $t+PH$, where PH is the prediction horizon. The electric demand during the current prediction horizon is penalized by the demand price p_{dm} as shown in Eq. (6.14).

$$D_{pen} = p_{dm} \max_{t,t+PH}((P_{dm} - P_{dm,lim}), 0) \tag{6.14}$$

where p_{dm} is the demand price, P_{dm} is the power demand calculated as the average power for each 30-min interval, and $P_{dm,lim}$ is the limit of required demand. This function means if the demand in current optimization step exceeds a predefined demand limit, then the optimization cost function is penalized by the cost of the demand differences. Otherwise, no penalization is applied. Note that p_{dm} and P_{dm} are both utility specific, and may vary from this definition.

6.2.4.2 Cooling Power Management

The **Cooling Power Management** modulates the power of the cooling system to respond to the FR signal. For the cooling system with a thermal storage system, the chiller capacity is controlled to respond to the regulation signal from the electric market as shown in Eq. (6.15).

$$u_{c,set}(t) = u_{c,set,bas} + r\Delta u_c(t), \tag{6.15}$$

where $u_{c,set}$ is the chiller capacity updated on a 4 s basis, and $u_{c,set,bas}$ is a predefined base chiller capacity that can be calculated to minimize the data center operation cost on a hour basis. The r is the raw regulation signal from the electrical market, ranging from -1 to 1. The $\Delta u_c(t)$ is the regulation range of the chiller capacity, and it could be estimated from Eq. (6.16).

$$\Delta u_c(t) = \min(u_{c,fr,max} - u_{c,set,bas}, u_{c,set,bas} - u_{c,fr,min}), \tag{6.16}$$

where $u_{c,fr,min}$ and $u_{c,fr,max}$ are the minimum and maximum controllable chiller capacity for FR service, respectively.

The charging rate of the storage is then determined as shown in Eq. (6.17). When $u_{c,set}$ is greater than $\dot{Q}^{'}$, then the storage is required to charge at a rate of $u_{s,set}$. When $u_{c,set}$ is less than $\dot{Q}^{'}$, then the storage is required to discharge at a rate of $u_{s,set}$.

$$u_{s,set}(t) = u_{c,set}(t) - \dot{Q}^{'}(t) \tag{6.17}$$

6.2.4.3 Frequency Regulation Capacity Bid

At each hour, the regulation capacity bid is determined using the predicted work load using the following equations. For the server aggregator, the power consumption as defined in Eq. (5.1) increases as the aggregated frequency f_{agg} increases. Therefore, the minimum and maximum power consumption of the servers can be obtained at the minimum frequency f_{min} and the maximum frequency f_{max}, respectively given the predicted workload.

$$P_{agg,min} = P_{agg}(f_{min}) \tag{6.18}$$

$$P_{agg,max} = P_{agg}(f_{max}) \tag{6.19}$$

The baseline server power consumption when the aggregator operates a frequency of f_{bas} can be calculated as

$$P_{agg,bas} = P_{agg}(f_{bas}) \tag{6.20}$$

The regulation up and down capacity of the aggregator then can be estimated as

$$C_{agg,reg,up} = P_{agg,max} - P_{agg,bas} \tag{6.21}$$

$$C_{agg,reg,do} = P_{agg,bas} - P_{agg,min} \tag{6.22}$$

The symmetric regulation capacity of the aggregator is then obtained as

$$C_{agg,reg} = min(C_{agg,reg,up}, C_{agg,reg,do}) \tag{6.23}$$

For the cooling system with a thermal storage system, the regulation capacity is related to the chiller capacity. Here we use a simple method to determine the available regulation range of chillers as shown in Eq. (6.24), and the other cooling equipment are ignored. The $u_{c,fr,max}$ and $u_{c,fr,min}$ are the maximum and minimum chiller capacity for FR service, which can be defined by users. COP is the nominal coefficient of performance of the chiller.

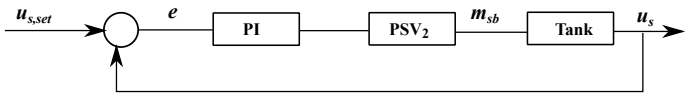
$$C_{chi,reg} = min(\frac{u_{c,fr,max} - u_{c,set,bas}}{COP}, \frac{u_{c,set,bas} - u_{c,fr,min}}{COP}) \tag{6.24}$$

Therefore, the total regulation capacity bid can be estimated as

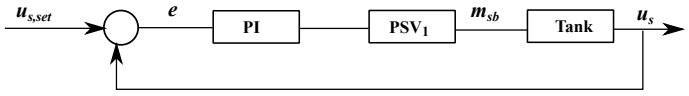
$$C_{reg} = C_{agg,reg} + C_{coo,reg} \tag{6.25}$$

6.2.5 Charging/Discharging Operating Strategy

The charging and discharging control of the chilled water system with chilled water storage is realized using some local control loops as illustrated in the following. Figure 6.3 describes the control diagram of the charging and discharging of the tank. The pressure-sustaining valves are modulated to regulate the water flow so that the charging and discharging rate u_s can track their references $u_{s,set}$. During charging process, a controller (e.g., PI) is used to track the errors e between the actual charging rate u_s and the reference charging rate $u_{s,set}$ by adjusting the valve position of the PSV_2. The changes of the valve position lead to the changes of water flowing into the tank, which eventually lead to the changes of actual charging rate. The same control loop is also used for discharging process but with modulating the PSV_1 instead.



(a) Charging



(b) Discharging

Figure 6.3: Diagram for tank charging/discharging control

For chillers, their cooling capacity at each time step can be controlled by adjusting the temperature setpoint of the chilled water leaving the evaporator $T_{eva,lvg,set}$ as illustrated in Figure 6.4. Here assume that the chillers have constant chilled water flow rate. When the cooling capacity is required to increase, then the $T_{eva,lvg,set}$ should be decreased.

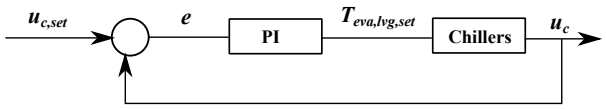


Figure 6.4: Diagram for chiller capacity control

6.3 Multi-market Optimization Framework

In this section, an optimization framework is proposed to enable the GEDC to maximize their benefits from both energy market and regulation market. Figure 6.5 shows the general process of the proposed optimization framework. The optimization framework uses a MPC to find the optimal control signal for each control horizon. The MPC controller consists of three major parts: **Predictor**, **Model** and **Optimizer**. The **Predictor** predicts the future signals over the prediction horizon, such as price signals from electrical markets and cooling load in the data center room. In this dissertation, historical prices and cooling load are used, which assumes perfect predictions. The **Optimizer** provides optimization algorithms to solve the optimization problem formed in the MPC controller. Here we use an open-source simulation-based optimization engine GenOpt [139]. The **Model** contains mathematical models to represent the studied data center cooling and IT systems, a FR controller that determines the economic baseline and estimates the possible maximum FR capacity based on the control input $\boldsymbol{u}_{c,set,bas}$, and a module that calculates the cost function.

The MPC controller outputs the optimal control signal $u^*_{c,set,bas}$ for the next control horizon, which is evaluated in a Modelica-based dynamic testbed. The system states $\boldsymbol{S}(u^*_{c,set,bas})$ after receiving the optimal control signal are then propagated to the MPC controller to initialize the framework for next time step. The MPC model in the MPC controller in this dissertation is the same as the models in the Modelica testbed, which means the MPC model can perfectly represent the "real" system.

The MPC cost function is formulated as follows. For each optimization time step, the optimal

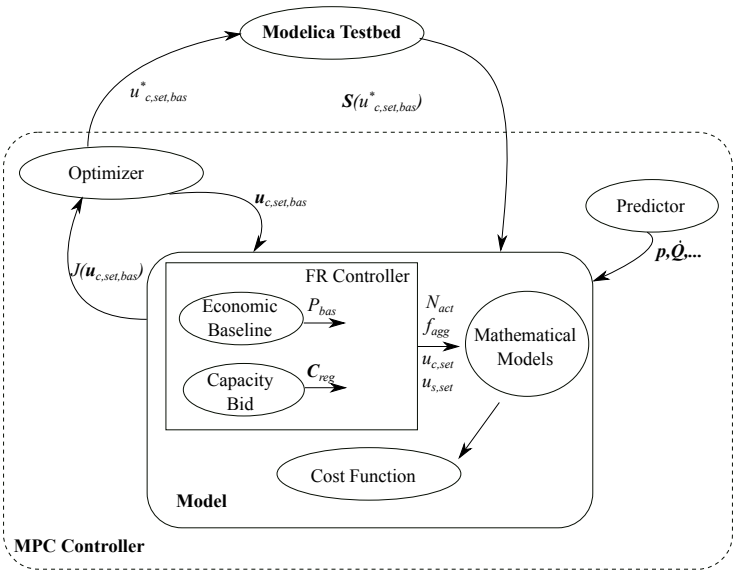


Figure 6.5: Diagram of multi-market optimization framework

setpoint $u^*_{c,set,bas}$ can be calculated as:

$$\begin{aligned} \underset{u_{c,set,bas}}{\text{argmin}} \quad & J(u_{c,set,bas}(t)) = E_{cos}(t) + D_{pen}(t) - R_{rev}(t) \\ s.t. \quad & x_{min} \leq x(t) \leq x_{max} \\ & T_{roo,min} \leq T_{roo}(t) \leq T_{roo,max} \\ & s(t) \geq s_l \\ & u_{c,min,fr} \leq u_{c,set,bas} \leq u_{c,max,fr} \end{aligned} \tag{6.26}$$

where the s is the regulation performance score defined by PJM, and s_l is the lowest allowable performance score by PJM to participate in regulation market. The $u_{c,min,fr}$ and $u_{c,max,fr}$ are the minimum and maximum controllable chiller capacity for the FR service. The cost function J has three terms: energy cost E_{cos}, demand penalty D_{pen} and regulation revenue R_{rev}. The former two terms have the same definitions as in Eq. (6.13) and Eq. (6.14). The revenues from regulation service is computed as follows.

$$R_{rev} = \int_t^{t+PH} p_{rm}(t) C_{reg}(t) dt \tag{6.27}$$

where p_{rm} is the predicted price signal from the regulation market, and C_{reg} is the regulation capacity bid for each control horizon.

6.4 Case Study

This case study aims to investigate the performance of the proposed multi-market optimization framework for a data center with a TESS. The data center is considered as a price taker only. The real-time energy market and the regulation market in PJM are investigated here. For the regulation service, only dynamic regulation is studied here, because its price is usually much higher than traditional regulation.

This case study considered six scenarios as listed here:

- *BL*: this scenario serves as the baseline of the GEDC without TESS as described in Section 5.3 in Chapter 5. This baseline system provides no FR services.

- *BL+MM*: this scenario applies the multi-market optimization framework proposed in Section 5.5 to the baseline system *BL*.

- *BL+LS*: this scenario introduces a stratified chilled water tank in parallel to the chillers as shown in Figure 6.1. The storage-priority control strategy is utilized to control the tank to provide load shifting. The on-peak period is defined as 11:00 - 18:00 because during this period the energy price is usually higher than other periods.

- *BL+ILS*: this scenario improves the storage-priority control strategy by defining the on-peak period as 8:00 - 18:00 to consider the energy price peaks and the actual workload peaks in the data center. Note that this is not a generic improvement that can be applied to any data centers.

- *BL+LS+MM*: this scenario applies to the *BL+LS* the proposed multi-market optimization framework in Section 6.3 but with a baseline for FR service calculated from the *BL+LS*.

- *BL+OPLS+MM*: this scenario applies the proposed multi-market optimization framework in Section 6.3 to the studied GEDC. The baseline power is calculated as described in Section 6.2.4.1. This scenario can minimize the operation costs from energy market and regulation market with required power demands.

6.4.1 System Description

The schematic drawing of the studied system is shown in Figure 6.1. For the IT system, the design number of servers is 8000. The design factor γ is set to 1.5 [82]. The total nominal electrical load is 2680 kW. The calibrated coefficients for Eq. (5.1) are $b_0 = 0.0154$, $b_1 = 1.5837$, $b_2 = 0.1373$, $c_0 = -22.3540$ and $c_1 = 121.0212$ using the method mentioned in Ref. [82]. For the cooling system, to simply the implementation, there is only one chiller and one storage tank for

cooling. The nominal chiller capacity C_{chi} is 2000 kW with a design COP of 5.8, and the thermal time constant of the evaporator is set to 5 min. The tank is sized to fully address 4-hour nominal cooling load in the data center using Eq. (6.28). The prices from the energy market and regulation market is plotted in Figure 6.6. A two-day historical workload in the data center is plotted in Figure 6.7. The demand charge rate is set to 7.48 \$/kW.

$$C_{sto} = C_{chi} * 4 * 3600 \tag{6.28}$$

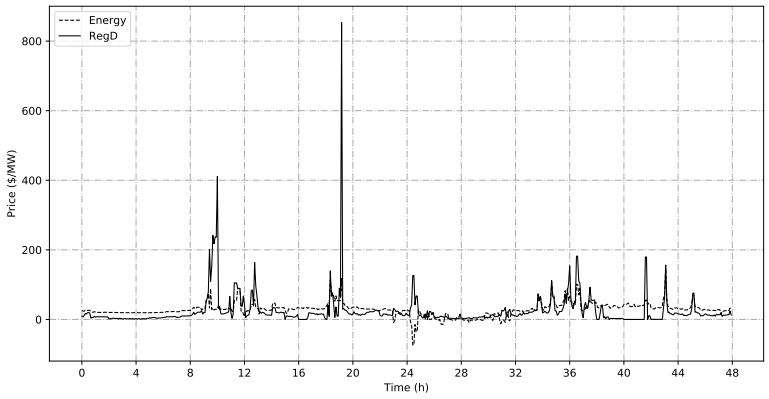


Figure 6.6: Two-day historical real-time prices of PJM markets

The baseline system using only the storage-priority control described in Section 6.2.3 is simulated to compare with the multi-market optimization framework. The on-peak period is defined as 11:00 - 18:00 and the other hours are in the off-peak period. C_0 in Eq. (6.6) is calculated for a lossless tank.

When providing FR service, the proposed optimization framework is used. The regulation flexibility factor is set to 1.1. The demand limit is set to 1990 kW, a lower value than the baseline demand 2046 kW in BL. The QoS is guaranteed by constraining the average response time of the data center service to 6 ms. The lower limit of the performance score in PJM to disqualify a regulation resource is 0.4 [89]. The minimum and maximum SoC of the tank is set to 0.05 and 0.95, respectively. The room temperature is bounded within 25 ± 3 °C. The control input $u_{c,set,bas}$

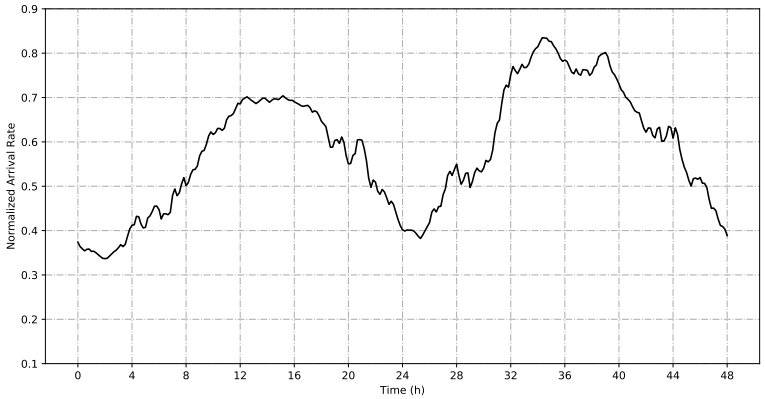


Figure 6.7: Two-day historical arrival rates in the data center

and $u_{c,set}$ are constrained within $[0, C_{chi}]$ for simplification.

When applying the proposed optimization framework, the prediction horizon in the **Economic Baseline** is set to 12 hours, and that for the multi-market optimization is set to 7 hours. The control horizon in all scenarios is set to 1 hour.

6.4.2 Chiller Capacity and Storage Charging Rate Control

The control performance of the tank charging and discharging rate control for load shifting and FR service is investigated here. Here we use the simulation results from *BL+LS* and *BL+LS+MM* as an example as shown in Figure 6.8 and Figure 6.9, respectively.

When not providing FR service, the trajectory of the control signals shown in Figure 6.8 can match their setpoints in a generally good way, especially when the system operates at mode M_3. At mode M_3, the storage provides the majority of the required cooling to the data center room, while the chiller only accounts for a small portion. Thus the chiller can track the setpoint that is within the bounds of the chiller's physical available capacity. At mode M_1 when the storage is charged with a maximum chiller capacity set to its nominal capacity, the control setpoint cannot be tracked very well. The reason is that the maximum chiller capacity $u_{c,max}$ as calculated in Eq. (6.3) and Eq. (6.4) tends to be overestimated. For the chiller, the available capacity is a function of different variables such as temperatures at the condenser and the evaporator. Merely using the nominal capacity as the available capacity cannot reflect the actual available capacity. For the storage, the charging rate setpoint equals to the difference between the cooling load and the chiller capacity setpoint at current time step as shown in Eq. (6.7). There, the storage charging rate setpoint is also overestimated when the chiller capacity setpoint is overestimated when charging is required. A better capacity estimation method can be used here to improve the prediction of the chiller capacity setpoint when charging the storage is required.

When providing FR service, the chiller capacity setpoint is reset to respond to the regulation signal as shown in Eq. (6.15) at a 4-second interval. The chiller capacity setpoint varies from a minimum of 0 to a maximum of design capacity. Therefore, during most of time, the chiller

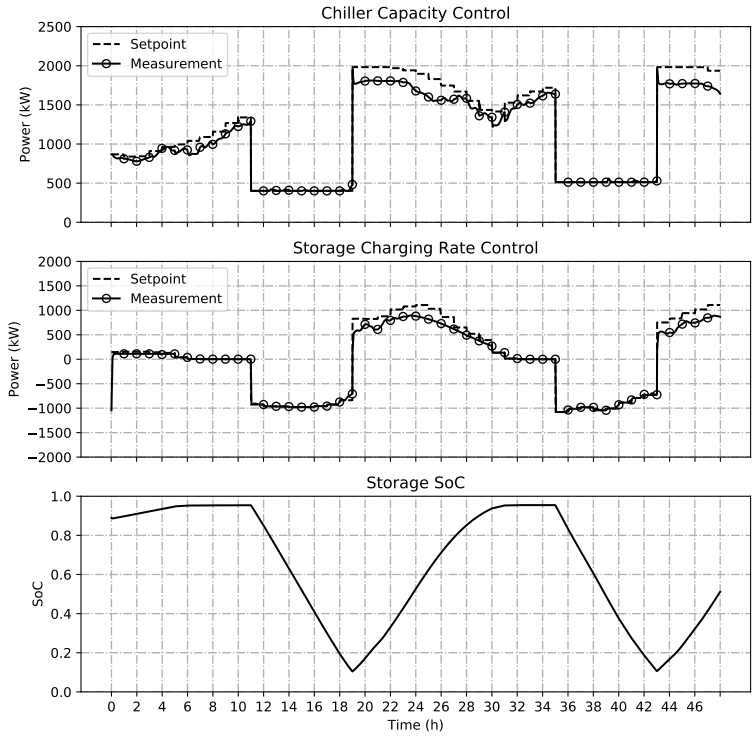


Figure 6.8: Control performance in *BL+LS*

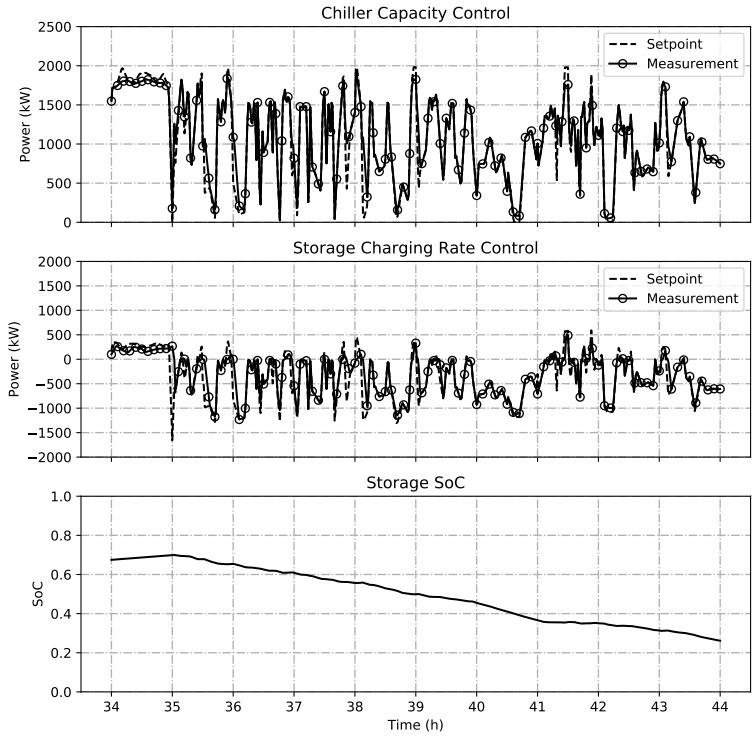


Figure 6.9: Control performance in *BL+LS+MM*

capacity can be controlled at the setpoint as shown in Figure 6.9. The similar control signal tracking performance can also be observed in storage charging rate control. The mismatches between the setpoint and the measurement are due to the delays in the thermal system and the limited physical capacities. The delays are caused by the thermal delays in the cooling equipment and the control delays in the controllers (i.e., as introduced in Table 6.2). The limited capacity may be caused by the physical sizes and operational constraints.

6.4.3 Frequency Regulation Control

This section investigates the control performance of the proposed FR control strategy as defined in Section 6.2.4, including the discussion of the FR signal tracking performance, hourly control inputs and hourly FR capacity bids.

Take *BL+OPLS+MM* as an example, during the simulated two days, the minimum hourly regulation performance score is 0.75, the maximum is 0.98, and the hourly average is 0.93. Therefore, the proposed synergistic control strategy can work very well to track the FR signal from the electrical market. The minimum score happens at hour 39, whose tracking performance is illustrated in Figure 6.10. The mismatch is caused by the overestimation of the regulation capacity bid, which leads to insufficient regulation down.

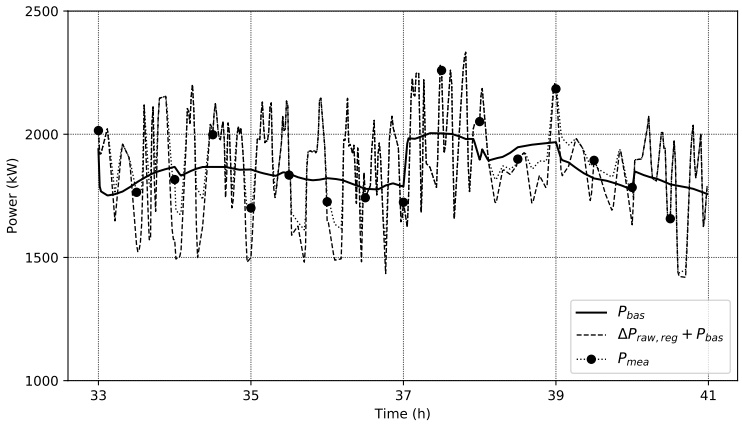


Figure 6.10: FR signal tracking in *BL+OPLS+MM*

Figure 6.11 illustrates the detailed control signals and performance at each hour during the simulated two days. The top figure shows the hourly control signal of the base chiller capacity setpoint $u_{c,set,bas}$. The control signal is about 1000 kW, half of the nominal capacity of the chiller, when the system power is far from its demand limit, e.g., hour 0 to hour 30. When the system approaches its peak power, the control signal is significantly increased. The reason is that when the system operates at a small load with a power demand far less than the limit, setting $u_{c,set,bas}$ to around 1000 kW can provide the maximum FR capacity based on Eq. (6.24) without violating the demand limit, thus gain the maximum FR revenues. When the system reaches its demand limit, the FR capacity bid should be as small as possible to avoid violating the demand limit as shown in the bottom figure. For example, the FR capacity bid around hour 37 is so well controlled that the demand in that hour is almost at the limit.

The GEDC with TESS can provide significantly larger FR capacity than that without TESS. Without TESS, the capacity bids only vary from 5 kW to 223 kW as shown in Figure 5.16(b). However, with TESS, the capacity bids can vary from 242 kW at hour 2 to 378 kW at hour 34 as shown in Figure 6.11. The reason is that with storage installed, the system can bid their maximum FR capacity into the market every hour, while the system without storage can only bid a capacity between 0 and their maximum FR capacity to balance the FR revenues and the potential increase of energy costs.

Although the chiller capacity and storage charging rate are not perfectly controlled, the proposed strategy can still be able to provide FR service when the bids are reasonable. The reason is that the fast-responding resources such as servers can compensate the delays caused by slow-responding resources such as the cooling system.

Figure 6.12 shows the detailed trajectories of some other important variables such as room temperature, aggregated frequency f_{agg}, the number of active servers N_a and storage SoC. The data center room temperature T_{room} is well controlled within the bounds 25 ± 3 °C. The fluctuations are caused by constantly resetting the chilled water supply temperature to control the chiller capacity at its setpoint. This strategy has been experimentally proven to be feasible to track dynamic FR

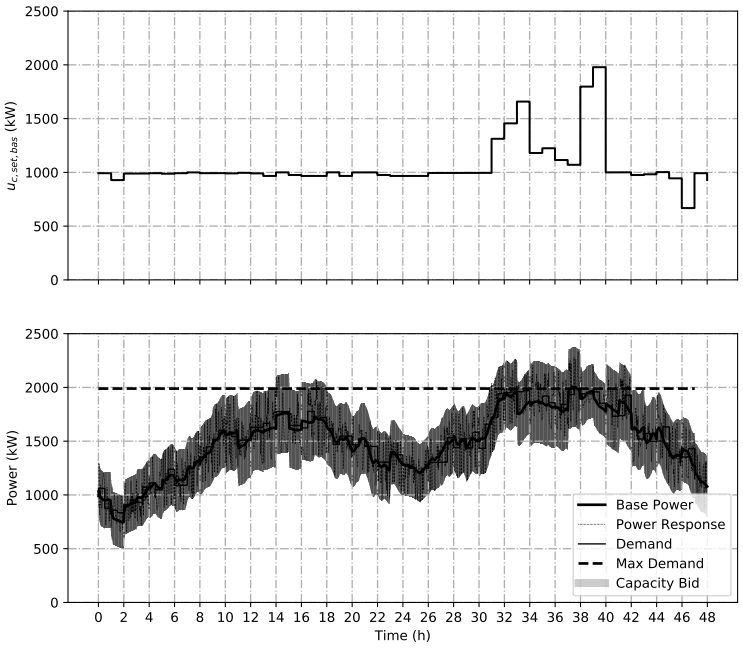


Figure 6.11: FR capacity in *BL+OPLS+MM*

signal in Ref. [114]. The allowed minimum aggregated frequency f_{min} varies as the work load and the number of active servers N_a change. The aggregated frequency f_{agg} is well controlled to fluctuate within the minimum and the maximum frequency. The average response time T_r of the data center service is controlled within 6ms. The number of active servers N_a varies every hour as recommended in Ref. [82] based on the changes of the incoming work load in the data center.

6.4.4 Multi-market Optimization Results

Table 6.3 shows the operation costs of two days in different scenarios. The energy consumption in all scenarios are similar, but the total operation costs are so different.

Table 6.3: Comparison of two-day operation costs

Items	*BL*	*BL+MM*	*BL+LS*	*BL+ILS*	*BL+LS+MM*	*BL+OPLS+MM*
Energy (MWh)	72.1	72.3	72.5	72.3	72.9	72.0
Energy Cost ($)	2246.8	2246.1	2212.3	2209.0	2197.5	2173.6
Demand (kW)	2168	2168	2148	1990	2145	1990
Demand Cost ($)	16216.9	16216.9	16064.6	14885.2	16044.6	14885.2
FR Cost ($)	0	-203.4	0	0	-385.4	-388.3
Total Cost ($)	18463.7	18259.6	18276.9	17094.2	17856.7	16670.5

6.4.4.1 TESS in Multi-markets

With TESS, the data center can reduce its energy costs and demand costs when not providing FR service. The *BL+LS* costs $34.5 (1.5%) less than *BL* for energy, $152.3 (1%) less for demand charges, and thus $186.8 (1%) less for total costs. The cost reductions of energy and demand with TESS are due to the control strategy that utilizes chillers as little as possible during on-peak periods, which can then reduce the demand and avoid the high energy prices during on-peaks.

The TESS can also help the system to harness more FR revenues besides energy and demand reductions. When providing FR service, *BL+LS+MM* costs $48.5 less than *BL+MM* for energy, and $172.3 less for demand charges. Although the energy and demand reductions are limited, together with fast-responding resources such as servers, *BL+LS+MM* can provide $182 (89.5%)

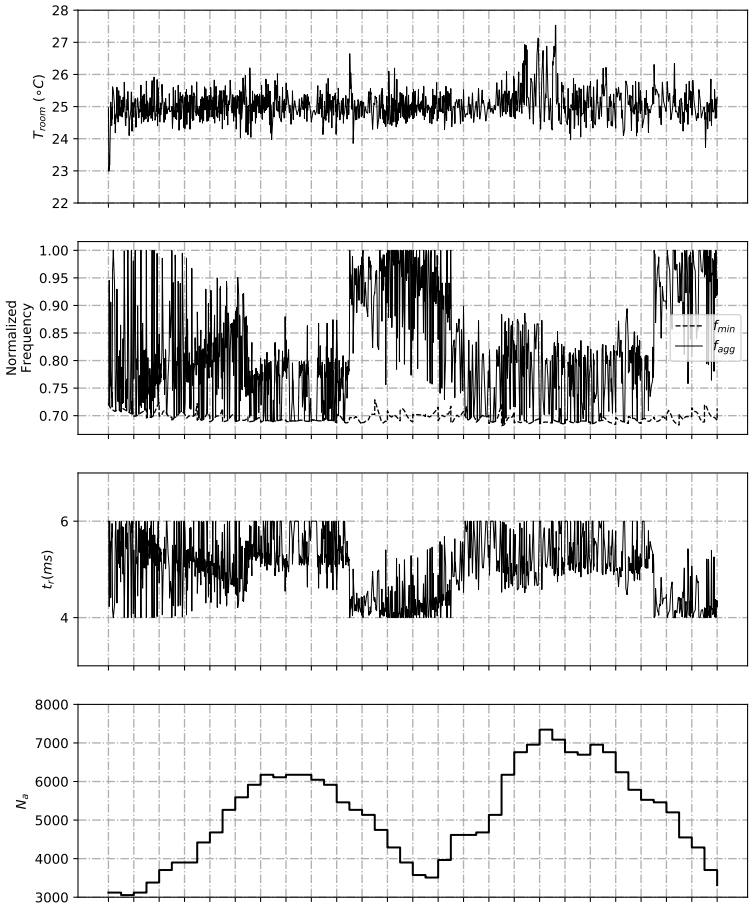


Figure 6.12: Details in *BL+OPLS+MM*

more FR revenues. The super gain of FR revenues in *BL+LS+MM* benefits from the larger FR capacity bids with TESS installed as described in Section 6.4.3.

The system with TESS can gain a significant amount of FR revenues by participating in energy market and regulation market without compromising energy and demand costs. For example, *BL+LS+MM* can harness $385.4 in the regulation market but with similar energy costs and demand costs as in *BL+LS*. The slight differences in energy costs and demand costs in *BL+LS* and *BL+LS+MM* are caused by the FR signal profiles as discussed in Section 5.5.2.2.

However, with the conventional control strategy in Section 6.2.3, the storage is not maximally utilized for energy efficiency and FR service. For example, the demand in *BL+LS+MM* is only slightly reduced compared with *BL*, *BL+MM* and *BL+LS*. The reason is that when providing FR service, the system needs to follow a baseline power profile. Because the FR signals in a long run sum to zero, the energy and the demand use when participating regulation market significantly depends on the submitted baseline profile, and it should have very similar energy use and demand compared with the baseline profile. The proposed multi-market optimization framework in Section 6.3 introduces a demand limiting strategy to calculate the economic baseline. The demand limiting strategy can significantly reduce the power demand, thus reduce the utility bill.

The *BL+OPLS+MM* can significantly reduce the operation costs by 9.7% in two days, saving about $1793.2 compared with *BL*. The demand reduces from 2168 kW to 1990 kW, saving $1332, which contributes the majority of the cost reductions. The proposed optimization framework can also harness $388.3 from the regulation market, and reduce $73.2 of energy costs. These savings benefit from the maximum utilization of TESS in energy market and regulation market.

6.4.4.2 Economic Baseline in Regulation Market

The *BL+OPLS+MM* can save $1186.2 (6.6%) of total costs, compared with *BL+LS+MM*. The savings are from the proposed economic baseline in Section 6.2.4.1. Here we denote the economic baseline as *BL+OPLS*. Figure 6.13 compares the control signals of the three baseline profiles *BL+LS*, *BL+ILS* and *BL+OPLS*.

Comparing with *BL+LS*, *BL+ILS* improves the storage-priority control strategy to consider the coincident peaks of energy prices and workloads in data centers. The improvement leads to a small energy cost reduction of $3 but with a significant amount of demand cost reduction of $1179.4. The demand cost reduction is due to the expansion of the on-peak period of energy prices and workloads. However, this improvement might not be generic enough for industrial application. There are two major reasons. First, even for the same data center, the energy prices and the workloads might have different peaks on different days, which might require different settings of the on-peak and off-peak periods. Second, different types of data centers have different occurrences of load peaks. For example, internet data centers might have a peak at noon, while research data centers might have a peak at midnight. Each type of data center should define their own on-peak and off-peak periods accordingly. The above-mentioned limits can be addressed by the proposed MPC-based economic baseline control.

BL+OPLS can eliminate the needs of coincident peaks of energy prices and workloads to some extend due to the proposed economic baseline. Using the rule-based control strategy, *BL+LS* and *BL+ILS* start to charge the tank at hour 19 with the maximum chiller capacity. However, *BL+OPLS* starts to charge the tank at hour 23 to avoid the high energy prices during hour 18 to hour 23. The tank is fully charged to its maximum SoC (i.e., 0.95) in *BL+LS*, but it is only charged to about 0.87 in *BL+OPLS*. The reason is that the data center in these two days operates at its part load, and thus the tank does not need to be fully charged to meet the loads on the next on-peak. Without charging the tank to its maximum level, about 0.9 MWh energy can be saved in *BL+OPLS* in two days.

Note that, at hour 24, the energy price from the energy market is negative, which means the data center should use as much power as possible to save costs. However, the chiller at hour 24 is not operating at its full capacity to charge the tank. The reason is that the tank reaches its maximum charging rates constrained by its physical sizes, which results in a total cooling load together with the heat generated by IT servers less than the chiller maximum capacity. Therefore, the chiller at this hour is not required to operate at the full capacity.

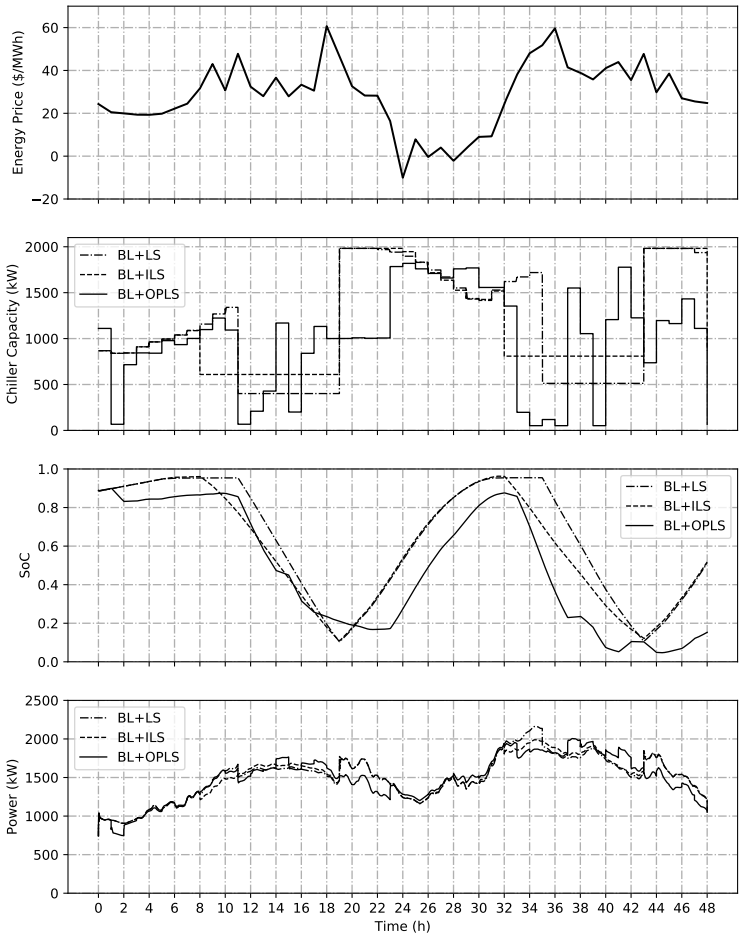


Figure 6.13: Control comparison of *BL+LS*, *BL+ILS* and *BL+OPLS*

In summary, the proposed optimization framework can provide a more economic baseline and thus reduce the total operation costs by reducing energy costs, limiting power demand, and harnessing FR revenues.

6.5 Summary

This chapter provides a multi-market optimization framework for GEDCs with TESS. The proposed framework includes two hierarchical model predictive control structures: one is for the economic baseline, which provides an efficient baseline power profile with limited demand, and the other is for multi-market operation, which provides the optimal control signal for the chiller and the storage in order to minimizing the operation costs. Detailed contributions are listed as follows.

Section 6.1 introduces some basics of the studied chilled water system with a stratified chilled water tank, including the system configurations, and general operating modes in terms of charging and discharging control.

Section 6.2 provides detailed mathematical models and control strategies for load shifting and FR service using the TESS. First, a tank model that can calculate the state-of-charge and the charging rate based measurable variables such as temperatures and flowrates is developed. Then, the staging control strategy that switches the cooling system from different operating modes is described. Next, a conventional storage-priority control strategy is introduced to provide load shifting service for the data center. Forth, to enable FR service using the TESS in the data center, a new synergistic control strategy is proposed. Lastly, the detailed operating strategy to track the chiller capacity setpoint and the storage charging/discharging rate setpoint is introduced.

Section 6.3 proposes a multi-market optimization framework for the data centers with TESS to minimize their operation costs. This framework utilizes model predictive control schemes to find the optimal chiller capacity setpoint so that the system can operate at the least cost.

Section 6.4 investigates the performance of the proposed optimization framework numerically. Simulation results show that utilizing the TESS can not only reduce energy costs and demand charges but also harness FR revenues. The data center with TESS can obtain $ 182 (89.5%) more

FR revenues compared to that without TESS in two days. What's more, the proposed multi-market optimization framework for the data center with TESS can reduce the operation costs by $1793.2 in two days, saving 9.7% compared with a baseline data center without TESS. The economic baseline can help the system save $1186.2 in the energy and the regulation market in two days, saving 6.6% compared to a data center without an economic baseline.

Chapter 7

Conclusions, Limitations and Future Work

7.1 Concluding Remarks

(1) ***What is the suitable platform to assist the design and evaluation of GEDCs?***

This question addresses the needs of simulation tools to model, control and simulate GEDCs. This simulation tool should be able to represent the interaction between cooling system components, system controllers, and IT system components over the short timescales associated with ancillary service provision. Modelica was used to create a dynamic model of the system, which was simulated to analyze the system dynamics associated with providing regulation ancillary services under various load conditions.

The most significant original contributions of this research are three-folds:

1) To meet the various modeling and simulation needs of GEDCs, Modelica is well-suited for the study of dynamic systems and controls.

2) A comprehensive testbed based on Modelica is developed to assist the design and operation of GEDCs. The testbed considers different physical systems (thermal, electrical, and electromagnetic, etc.) with different time-scaled dynamics involved. End-to-end models include the computer servers, quality of service, uninterruptible power supply, renewable energy resources such as solar panels, typical cooling system for data centers etc.

3) The validated testbed has been proven to be able to perform dynamic control evaluations, besides detailed analysis of energy efficiency.

(2) ***How to optimally operate GEDCs to harness the flexibility provided by modulating loads?***

This question addresses the use of modulating loads to gain benefits from electrical markets. The strategy in this dissertation is to provide FR service to the grid and get compensated from the energy and regulation market simultaneously.

Many original contributions have come out from this research as illustrated in Chapter 5:

1) A synergistic control strategy with a new FR flexibility factor for GEDCs is introduced to enable the provision of FR service using modulating loads in both the cooling system and the IT system. With well-tuned parameters, this control strategy is able to track the FR signal and maintain a high regulation performance score.

2) The important factors that influence the FR performance when using the proposed synergistic control strategy are identified, including the regulation capacity bid, frequency regulation flexibility factor, workload condition, and cooling mode of the cooling system. The time constant of chillers has minimal influence on the FR performance when fast-responding resources such as servers are present.

3) A model-based perturbation method is utilized to determine the maximum FR capacity GEDCs can provide at each time step based on weather conditions, workloads, etc. This method finds out that the proposed synergistic control strategy can provide an extra regulation capacity (76 kW) of 3% of the design power in the data center when chillers are activated compared with a server-only control strategy.

4) A real-time multi-market optimization framework for GEDCs without storage systems is developed to maximize their benefits from participating in both energy market and regulation market by finding the optimal bids at each hour. Simulation results shows

that providing frequency regulation service over two days in January and July can save $23.6 and $115.8, respectively.

(3) ***How to optimally operate GEDCs to harness the flexibility provided by shifting loads and modulating loads?***

This question addresses the use of shifting loads and modulating loads together to gain benefits from electrical markets. Because providing FR service may not be able to control the electrical demand in GEDCs, a thermal energy storage system is proposed. The GEDCs cooled by chillers and thermal storage can then be controlled to provide FR service while reducing the system demands.

The following specific contributions have come out of this research:

1) A synergistic control strategy is proposed to enable FR service using thermal storage systems by adjusting the chiller capacity, storage charging/discharging rate and sever CPU frequency.

2) A multi-market optimization framework with demand limiting is developed to minimize the operation costs of the GEDCs with TESSs. Simulation results show that utilizing the TESS can not only reduce energy costs and demand charges but also harness FR revenues. The data center with TESS can obtain $ 182 (89.5%) more FR revenues compared to that without TESS in two days. What's more, the proposed multi-market optimization framework for the data center with TESS can reduce the operation costs by $1793.2 in two days, saving 9.7% compared with a baseline data center without TESS. The economic baseline can help the system save $1186.2 in the energy and the regulation market in two days, saving 6.6% compared to a data center without an economic baseline.

7.2 Limitations

This section provides the limitations of the methods developed in this dissertation. The limitations are based on the objective of this research, which is to explore the opportunities and methods of enabling GEDCs. With this research moving forward to implementation, methods presented in this dissertation must be improved for field tests with real-time control. The major limitations of the current method are listed in the following:

(1) The multi-market optimization frameworks assume perfect predictions. The proposed multi-market optimization frameworks for GEDCs with and without thermal energy storage systems intend to use calibrated simulation models to represent actual GEDCs and the market operations. This requires the GEDC operation status and electrical market operations to be known hours ahead. For example, the multi-market optimization framework needs to know the energy prices and regulation prices in the next hour so that it can calculate the optimal control signal to minimize the operation cost in advance. However, in the optimization framework, the MPC controller uses the same mathematical models as the Modelica testbed, which represents perfect predictions of GEDC operation status. Besides, the energy prices and regulation prices from the electrical market are assumed to be known in advance by using historical data. The workload in GEDCs uses historical profiles from web-based data centers. What's more, the RegD signal is assumed known ahead of time. The perfect predictions of prices, workload, and RegD signal result in possible maximum benefits of the proposed framework. When considering real implementation, the benefits calculated from this dissertation can be overestimated.

Thus, possible improvement toward real implementation should include online robust predictors for prices of the electrical markets, data center incoming workload, regulation signal, and data center operation states. The uncertainties introduced by the predictors should be controlled within a desired level. Although previous research has provided some methods to predict the above-mentioned signals, the accuracy of the methods for online control remains

unknown. In addition, incorporating those predictors into the MPC would be cumbersome and time-consuming to deal with. Therefore, a predictor that can balance the accuracy and the calculation speed for online implementation seems an important step.

(2) The FR control strategy is designed for limited applications. First, the GEDC in this dissertation is assumed to provide delay-tolerant noncritical services, such as web-based services and delay-tolerant computational workload in universities. This type of data centers allow the degradation of the QoS to some extent, which provides extra spaces for energy efficiency and ancillary services. The same control strategy especially for the IT system might not be used in mission-critical data centers.

Second, the proposed FR control strategy especially for the IT servers only works when the data center operates at part load. The introduced design redundancy and FR flexibility factor exploit the unutilized space for FR service. However, when the data center is not so well-designed that it approaches its full capacity very fast, the proposed FR control strategy might not work as desired.

Third, adjusting the chiller capacity in a TESS system for FR service may lead to frequent staging of chillers if where are multiple chillers designed to provide cooling. The FR control strategy in Section 6.2.4 ignores the potential frequent staging of chillers. Therefore, it might be more suitable for a system with only one chiller.

(3) The uncertainties of the numerical testbed and optimization framework in this dissertation are not fully understood. The uncertainties come from two parts: modeling uncertainties and measurement uncertainties.

A Modelica-based dynamic simulation testbed is developed and validated in Chapter 3 and Chapter 4. However, the validation is not fully conducted for all the components and systems simulated in this dissertation due to limited access to the measurement data. The validation is only conducted for an academic data center, which uses chillers and airside economizers to provide cooling for almost constant workload in the room. For the cooling

system with waterside economizers used in Chapter 5, there are unknown uncertainties of the system operation status if waterside economizer models are not validated. What's more, the proposed optimization framework is tested only on this numerical testbed. No field test is available at this moment due to the above-mentioned theoretical limitations and lack of practical resources. Therefore, the performance of the proposed optimization framework remains uncertain when it can be implemented on a real GEDC.

The uncertainties introduced by measurement sensors such as temperature sensors are not investigated when designing the FR controller. These measurement uncertainties can significantly influence the performance of model-based control strategies [108]. Consequences include unstable control actions or non-optimal setpoints in optimal control. Therefore, impact of the measurement uncertainties on the control should be considered and evaluated before being implemented on site.

(4) Before the electrical market is cleared, FR resources need to submit their bids, including regulation capacity offer (MW), and performance offer ($/ΔMW). The performance offer ($/ΔMW) is the price that the FR resources bid into the market. After collecting the data from all the FR resources, the electrical market will be cleared by setting the market prices to highest ranked price. Therefore, the performance offer provided by GEDCs can potentially influence the market clearing prices. However, this dissertation only considers GEDCs as the price-takers, where they only respond to price signals to schedule energy consumption and provision of FR service. As the demand-side continues to provide more ancillary services, it will be important to accurately determine the costs associated with their service provision, so that market clearing prices represent the true cost of providing service.

7.3 Future Work

Future work associated with the above-mentioned limitations is proposed as follows:

(1) For the proposed optimization framework, practical predictors should be developed, including simplified models for the cooling system and the IT system, energy market price predictors, server workload predictor, and regulation signal predictor etc. The major improvement should be finding suitable predictors that can balance the needs of model accuracy and computation time for online applications.

The cooling system and the IT system for the optimization framework should be simplified so that real-time optimization can be feasible. The relationship of the power profile between the cooling system and the IT system can be simplified for online optimization purposes. The idea is to estimate the cooling energy from the predicted IT energy, which could be quantified by PUE. The PUE can be estimated as a system response to different inputs such as weather conditions, workload, room temperature, supply air temperature etc. The estimation can be performed using curve-fitting techniques or machine learning techniques.

For the prices and server workload, predictors using machine learning techniques can be utilized. This involves determining the influential features (e.g., date and time, weather conditions, etc.), finding the optimal estimators (e.g., artificial neural network, support vector machine, etc.), online model calibration schemes etc. An example for energy price prediction can be found in [31].

For the regulation signal, especially RegD signal, the prediction is difficult. In reality, RegD signal is the high-pass filtered product of area control error (ACE). Thus, if RegD signal is assumed predictable, that means ACE is assumed predictable. Although previous research has provided some methods to predict ACE with simplified interconnected power system models. The accuracy of the methods for online supervisory control remains unknown. New methods should be developed for the online control purposes.

(2) There are few improvements that can be performed currently to extend the proposed FR control strategies to mission-critical GEDCs, and full-capacity GEDCs. However, it is feasible to address the possible fact that cooling equipment might be frequently switched

due to the changing of chiller capacity for FR controls in Chapter 6. One method is to design a staging controller for the chillers. Detailed logic for a system with multiple identical chillers might be illustrated as follows. One chiller should be staged on if

$$\frac{u_{c,set}}{N_{chi}} \geq u_{on} \text{ for 20 minutes,} \tag{7.1}$$

and one chiller should be staged off if

$$\frac{u_{c,set}}{N_{chi}} \leq u_{off} \text{ for 20 minutes,} \tag{7.2}$$

where N_{chi} is the design number of chillers, u_{on} and u_{off} are the staging on and off thresholds for chillers. The delay time can be determined from chiller manufactures to ensure equipment safety. Note the large delay time will cause large delays of the response of the cooling system to the FR signal. However, the delays can be compensated by fast responding resources such as IT servers. Therefore, the performance of FR service might not be degraded if the IT servers have enough regulation capacity.

Another method is to include constraints for the number of staging operations during the interested time period. For example, if the constraint is added to Eq. (6.26), the optimization algorithm will find an optimal control signal for the chiller capacity that can avoid switching too many times in a given time slot.

(3) The uncertainties in the proposed testbed and optimization framework can be comprehended and evaluated in order to improve the control. The modeling uncertainties can be quantified by calibrating all the component models and the system model as a whole. The measurement uncertainties of different sensors have different distributions. The distributions can be investigated using methods mentioned in [84]. These distributions then give important information on how to design robust rules to improve the control system.

(4) To calculate the performance offer, GEDCs need to know their opportunities costs to provide FR service. Per Order 755 rule 2), all regional transmission organizations (RTO) and

independent system operators (ISO) in the U.S. include opportunity costs in their ancillary service market marginal clearing prices and are either calculated explicitly or calculated implicitly from unit-commitment optimization shadow prices, which represent the cost of specific constraint violations in optimization processes. PJM defines generator opportunity costs as shown in Eq. (7.3), where $\pi_{g,t}$ is the generator opportunity cost for a given time t, $x^*_{g,t}$ is the economic maximum dispatch setpoint without considering ancillary service provision, $x^{AS}_{g,t}$ is the dispatch setpoint including the provision of an ancillary service, $p^E_{g,t}$ is the forecasted locational marginal price at the generator node, and $c(x)$ is the generator offer curve into the wholesale energy market.

$$\pi_{g,t} = \int_{x^{AS}_{g,t}}^{x^*_{g,t}} [p^E_{g,t} - c(x)]dx \tag{7.3}$$

For the demand side FR resources such as GEDCs, the opportunity costs are not defined by PJM. But a similar formula as the generator can be established for providing ancillary services by considering capacity opportunity costs and performance opportunity costs as demonstrated in [15].

Bibliography

[1] Aayush Agrawal, Mayank Khichar, and Sanjeev Jain. Transient simulation of wet cooling strategies for a data center in worldwide climate zones. Energy and Buildings, 127:352–359, 2016.

[2] 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.

[3] Michael Anderson, Michael Buehner, Peter Young, Douglas Hittle, Charles Anderson, Jilin Tu, and David Hodgson. Mimo robust control for hvac systems. IEEE Transactions on Control Systems Technology, 16(3):475–483, 2008.

[4] ASHRAE. Thermal guidelines for data processing environments. 4 edition, 2015.

[5] ASHRAE. Hvac systems and equipment. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, GA, pages 1–10, 2016.

[6] Shahab Bahrami, Vincent WS Wong, and Jianwei Huang. Data center demand response in deregulated electricity markets. IEEE Transactions on Smart Grid, 2018.

[7] C B Bash, C D Patel, and R K Sharma. Dynamic thermal management of air cooled data centers. In Thermal and Thermomechanical Proceedings 10th Intersociety Conference on Phenomena in Electronics Systems, 2006. ITHERM 2006., pages 445–452, may 2006.

[8] CB Bash, Chandrakant D Patel, and Ratnesh K Sharma. Dynamic thermal management of air cooled data centers. In Thermal and Thermomechanical Proceedings 10th Intersociety Conference on Phenomena in Electronics Systems, 2006. ITHERM 2006., pages 8–pp. IEEE, 2006.

[9] Natalie Bates, Girish Ghatikar, Ghaleb Abdulla, Gregory A Koenig, Sridutt Bhalachandra, Mehdi Sheikhalishahi, Tapasya Patki, Barry Rountree, and Stephen Poole. Electrical grid and supercomputing centers: An investigative analysis of emerging opportunities and challenges. Informatik-Spektrum, 38(2):111–127, 2015.

[10] Ian Beil, Ian Hiskens, and Scott Backhaus. Frequency regulation from commercial building hvac demand response. Proceedings of the IEEE, 104(4):745–757, 2016.

[11] Monem H Beitelmal and Chandrakant D Patel. Model-based approach for optimizing a data center centralized cooling system. Hewlett-Packard (HP) Lab Technical Report, 2006.

[12] Stéphane Bertagnolio and Jean Lebrun. Simulation of a building and its hvac system with an equation solver: application to benchmarking. In Building Simulation, pages 234–250. Springer, 2008.

[13] Stéphane Bertagnolio, Gabrielle Masy, J Lebrun, and P André. Building and hvac system simulation with the help of an engineering equation solver. Proceedings of SimBuild, 3(1):53–60, 2008.

[14] B Birdsall, WF Buhl, KL Ellington, AE Erdem, and FC Winkelmann. Overview of the doe-2 building energy analysis program. Report LBL-19735m rev. w, Lawrence Berkeley Laboratory, Berkeley, CA, 1990.

[15] D. H. Blum, T. Zakula, and L. K. Norford. Opportunity cost quantification for ancillary services provided by heating, ventilating, and air-conditioning systems. IEEE Transactions on Smart Grid, 8(3):1264–1273, May 2017.

[16] Gunter Bolch, Stefan Greiner, Hermann De Meer, and Kishor S Trivedi. Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. John Wiley & Sons, 2006.

[17] Marco Brocanelli, Sen Li, Xiaorui Wang, and Wei Zhang. Joint management of data centers and electric vehicles for maximized regulation profits. In 2013 International Green Computing Conference Proceedings, pages 1–10. IEEE, 2013.

[18] Richard Brown et al. Report to congress on server and data center energy efficiency: Public law 109-431. 2008.

[19] John S Bucy, Gregory R Ganger, et al. The DiskSim simulation environment version 3.0 reference manual. School of Computer Science, Carnegie Mellon University, 2003.

[20] Peter Bunus and Peter Fritzson. Automated static analysis of equation-based components. Simulation, 80(7-8):321–345, 2004.

[21] Rajkumar Buyya, Anton Beloglazov, and Jemal Abawajy. Energy-efficient management of data center resources for cloud computing: A vision. Architectural Elements, and Open Challenges, 2010.

[22] Rodrigo N Calheiros, Rajiv Ranjan, Anton Beloglazov, César AF De Rose, and Rajkumar Buyya. Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and experience, 41(1):23–50, 2011.

[23] Enrique V Carrera, Eduardo Pinheiro, and Ricardo Bianchini. Conserving disk energy in network servers. In Proceedings of the 17th annual international conference on Supercomputing, pages 86–97, 2003.

[24] Jeffrey S Chase, Darrell C Anderson, Prachi N Thakar, Amin M Vahdat, and Ronald P Doyle. Managing energy and server resources in hosting centers. ACM SIGOPS operating systems review, 35(5):103–116, 2001.

[25] Hao Chen, Michael C Caramanis, and Ayse K Coskun. The data center as a grid load stabilizer. In 2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC), pages 105–112. IEEE, 2014.

[26] Hao Chen, Ayse K Coskun, and Michael C Caramanis. Real-time power control of data centers for providing regulation service. In 52nd IEEE Conference on Decision and Control, pages 4314–4321. IEEE, 2013.

[27] Hao Chen, Can Hankendi, Michael C Caramanis, and Ayse K Coskun. Dynamic server power capping for enabling data center participation in power markets. In 2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pages 122–129. IEEE, 2013.

[28] Hao Chen, Zhenhua Liu, Ayse K Coskun, and Adam Wierman. Optimizing energy storage participation in emerging power markets. In 2015 Sixth International Green and Sustainable Computing Conference (IGSC), pages 1–6. IEEE, 2015.

[29] Lijun Chen and Na Li. On the interaction between load balancing and speed scaling. IEEE Journal on Selected Areas in Communications, 33(12):2567–2578, 2015.

[30] Craig Christensen, Greg Barker, and Scott Horowitz. A sequential search technique for identifying optimal building designs on the path to zero net energy. In Proceedings of the Solar Conference, pages 877–882. American Solar Energy Society; American Institute Of Architects, 2004.

[31] Antonio J Conejo, Miguel A Plazas, Rosa Espinola, and Ana B Molina. Day-ahead electricity price forecasting using the wavelet transform and arima models. IEEE transactions on power systems, 20(2):1035–1042, 2005.

[32] IBM ILOG Cplex. V12. 1: Users manual for cplex. International Business Machines Corporation, 46(53):157, 2009.

[33] Lisette Cupelli, Nikhil Barve, and Antonello Monti. Optimal sizing of data center battery energy storage system for provision of frequency containment reserve. In IECON 2017-43rd Annual Conference of the IEEE Industrial Electronics Society, pages 7185–7190. IEEE, 2017.

[34] Lisette Cupelli, Pooyan Jahangiri, Antonello Monti, and Dirk Müller. Ancillary services from data center hvac systems and back-up generator sets. In IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, pages 5579–5584. IEEE, 2016.

[35] Lisette Cupelli, Thomas Schütz, Pooyan Jahangiri, Marcus Fuchs, Antonello Monti, and Dirk Müller. Data center control strategy for participation in demand response programs. IEEE Transactions on Industrial Informatics, 14(11):5087–5099, 2018.

[36] Howard David, Eugene Gorbatov, Ulf R Hanebutte, Rahul Khanna, and Christian Le. Rapl: memory power estimation and capping. In 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), pages 189–194. IEEE, 2010.

[37] U.S. DoE. Energyplus engineering reference. The reference to energyplus calculations, 2016.

[38] Eaton. Eaton launches industry first ups-as-a-reserve service to support the power grid in frequency containment reserve. http://powerquality.eaton.com/emea/about-us/news-events/2017/pr031017.asp?act=smtc&id=&key=&Quest_user_id=&leadg_Q_QRequired=&site=&menu=&cx=98&x=16&y=8. Accessed: 2019-09-30.

[39] Khosrow Ebrahimi, Gerard F Jones, and Amy S Fleischer. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. Renewable and Sustainable Energy Reviews, 31:622–638, 2014.

[40] Bryan Eisenhower, Kazimir Gasljevic, and Igor Mezi. Control-Oriented Dynamic Modeling and Calibration of a Campus Theater Using Modelica. Proceedings of SimBuild, 5(1):112–119, 2012.

[41] Hilding Elmqvist, Sven Erik Mattsson, and Martin Otter. Modelica: The new object-oriented modeling language. In 12th European Simulation Multiconference, Manchester, UK, 1998.

[42] Qiu Fang, Jun Wang, Qi Gong, and Mengxuan Song. Thermal-aware energy management of an hpc data center via two-time-scale control. IEEE Transactions on Industrial Informatics, 13(5):2260–2269, 2017.

[43] Fan Feng, Yangyang Fu, Jin Hou, and Peng Xu. Optimizing the topologies of heating, ventilation, and air-conditioning water systems in supertall buildings: A pilot study. Science and Technology for the Built Environment, 24(4):371–381, 2018.

[44] Gilles Fraisse, Christelle Viardot, Olivier Lafabrie, and Gilbert Achard. Development of a simplified and accurate building model based on electrical analogy. Energy and buildings, 34(10):1017–1031, 2002.

[45] Rüdiger Franke and Hansjürg Wiesmann. Flexible modeling of electrical power systems–the modelica powersystems library. In Proceedings of the 10 th International Modelica Conference; March 10-12; 2014; Lund; Sweden, pages 515–522. Linköping University Electronic Press, 2014.

[46] Yangyang Fu, Sen Huang, Draguna Vrabie, and Wangda Zuo. Coupling Power System Dynamics and Building Dynamics to Enabling Building-to-Grid Integration. In Proceedings of the 13th International Modelica Conference, Regensburg, Germany, March 4–6, 2019, number 157. Linköping University Electronic Press, 2019.

[47] Yangyang Fu, Xing Lu, and Wangda Zuo. Modelica Models for the Control Evaluations of Chilled Water System with Waterside Economizer. In Proceedings of the 13th International Modelica Conference, Regensburg, Germany, March 4–6, 2019, number 157. Linköping University Electronic Press, 2019.

[48] Yangyang Fu, Michael Wetter, and Wangda Zuo. Modelica models for data center cooling systems. In 2018 Building Performance Analysis Conference and SimBuild, Chicago, Illinois, United States of America, 2018.

[49] Yangyang Fu, Wangda Zuo, Michael Wetter, James VanGilder, Xu Han, and David Plamondon. Equation-based object-oriented modeling and simulation for data center cooling: A case study. accepted by Energy and Buildings, 2019.

[50] Yangyang Fu, Wangda Zuo, Michael Wetter, James W. VanGilder, and Peilin Yang. Equation-based object-oriented modeling and simulation of data center cooling systems. Energy and Buildings, 198:503 – 519, 2019.

[51] Yangyang Fu, Wangda Zuo, Michael Wetter, Jim W. VanGilder, Xu Han, and David Plamondon. Equation-Based object-oriented modeling and simulation for data center Cooling: A case study. Energy and Buildings, 186:108–125, 2019.

[52] Venkata Ganti and Girish Ghatikar. Smart grid as a driver for energy-intensive industries: a data center case study. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2012.

[53] Girish Ghatikar, Venkata Ganti, Nance Matson, and Mary Ann Piette. Demand response opportunities and enabling technologies for data centers: Findings from field studies. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2012.

[54] Girish Ghatikar, Mary Ann Piette, Sydny Fujita, Aimee McKane, Junqiao Han Dudley, Anthony Radspieler, KC Mares, and Dave Shroyer. Demand response and open automated demand response opportunities for data centers. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2009.

[55] S. Gurumurthi, Jianyong Zhang, A. Sivasubramaniam, M. Kandemir, H. Franke, N. Vijaykrishnan, and M. J. Irwin. Interplay of energy and performance for disk arrays running transaction processing workloads. In 2003 IEEE International Symposium on Performance Analysis of Systems and Software. ISPASS 2003., pages 123–132, March 2003.

[56] Ranjini Guruprasad, Prakash Murali, Dilip Krishnaswamy, and Shivkumar Kalyanaraman. Coupling a small battery with a datacenter for frequency regulation. In 2017 IEEE Power & Energy Society General Meeting, pages 1–5. IEEE, 2017.

[57] Sang-Woo Ham and Jae-Weon Jeong. Impact of aisle containment on energy performance of a data center when using an integrated water-side economizer. Applied Thermal Engineering, 105:372–384, 2016.

[58] He Hao, Anupama Kowli, Yashen Lin, Prabir Barooah, and Sean Meyn. Ancillary service for the grid via control of commercial building hvac systems. In 2013 American control conference, pages 467–472. IEEE, 2013.

[59] He Hao, Yashen Lin, Anupama S Kowli, Prabir Barooah, and Sean Meyn. Ancillary service to the grid through control of fans in commercial building hvac systems. IEEE Transactions on smart grid, 5(4):2066–2074, 2014.

[60] Gregor P Henze, Robert H Dodier, and Moncef Krarti. Development of a predictive optimal controller for thermal energy storage systems. HVAC&R Research, 3(3):233–264, 1997.

[61] Jin Heo, Praveen Jayachandran, Insik Shin, Dong Wang, Tarek Abdelzaher, and Xue Liu. Optituner: On performance composition and server farm energy minimization application. IEEE Transactions on Parallel and Distributed Systems, 22(11):1871–1878, 2011.

[62] Scott Horowitz, Craig Christensen, Michael Brandemuehl, and Moncef Krarti. Enhanced sequential search methodology for identifying cost-optimal building pathways. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2008.

[63] Sen Huang, Wangda Zuo, and Michael D Sohn. Amelioration of the cooling load based chiller sequencing control. Applied Energy, 168:204–215, 2016.

[64] Sen Huang, Wangda Zuo, and Michael D Sohn. Improved cooling tower control of legacy chiller plants by optimizing the condenser water set point. Building and Environment, 111:33–46, 2017.

[65] Mark Hydeman and Kenneth L Gillespie. Tools and techniques to calibrate electric chiller component models. ASHRAE transactions, 108(1):733–741, 2002.

[66] Uptime Institute. Annual data center survey results. Report, Uptime Institute, 2019.

[67] James Kennedy. Particle swarm optimization. Encyclopedia of machine learning, pages 760–766, 2010.

[68] Donghun Kim, Wangda Zuo, James E Braun, and Michael Wetter. Comparisons of building system modeling approaches for control system design. In Proceedings of the 13th International Conference of the International Building Performance Simulation Association (Building Simulation 2013), pages 3267–3274. Citeseer, 2013.

[69] Young-Jin Kim, Leslie K Norford, and James L Kirtley. Modeling and analysis of a variable speed heat pump for frequency regulation through direct load control. IEEE Transactions on Power Systems, 30(1):397–408, 2015.

[70] SA Klein and FL Alvarado. Ees: Engineering equation solver, user manual. F-chart software, 2008.

[71] Sanford A Klein. Trnsys-a transient system simulation program. University of Wisconsin-Madison, Engineering Experiment Station Report, pages 38–12, 1988.

[72] Data Center Knowledge. Google embraces thermal storage in taiwan. `https://www.datacenterknowledge.com/archives/2012/04/03/google-embraces-thermal-storage-in-taiwan`. Accessed: May 6, 2019.

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

[74] Jonathan Koomey. Growth in data center electricity use 2005 to 2010. A report by Analytical Press, completed at the request of The New York Times, 9, 2011.

[75] V V G Krishnan, Y Zhang, K Kaur, A Hahn, A Srivastava, and S Sindhu. Cyber-Security Analysis of Transactive Energy Systems. In 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T D), pages 1–9, apr 2018.

[76] Dongkyu Lee, Byoungdoo Lee, and Jin Woo Shin. Fault detection and diagnosis with modelica language using deep belief network. In Proceedings of the 11th International Modelica Conference, Versailles, France, September 21-23, 2015, number 118, pages 615–623. Linköping University Electronic Press, 2015.

[77] Kuei-Peng Lee and Hsiang-Lun Chen. Analysis of energy saving potential of air-side free cooling for data centers in worldwide climate zones. Energy and Buildings, 64:103–112, 2013.

[78] Bin Li, Richard Otten, Vikas Chandan, William F Mohs, Jeff Berge, and Andrew G Alleyne. Optimal on–off control of refrigerated transport systems. Control Engineering Practice, 18(12):1406–1417, 2010.

[79] Jie Li, Zhen Bao, and Zuyi Li. Modeling demand response capability by internet data centers processing batch computing jobs. IEEE Transactions on Smart Grid, 6(2):737–747, 2014.

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

[81] Pengfei Li, Yaoyu Li, John E Seem, Hongtao Qiao, Xiao Li, and Jon Winkler. Recent advances in dynamic modeling of hvac equipment. part 2: Modelica-based modeling. HVAC&R Research, 20(1):150–161, 2014.

[82] Sen Li, Marco Brocanelli, Wei Zhang, and Xiaorui Wang. Data center power control for frequency regulation. In 2013 IEEE Power & Energy Society General Meeting, pages 1–5. IEEE, 2013.

[83] Sen Li, Marco Brocanelli, Wei Zhang, and Xiaorui Wang. Integrated power management of data centers and electric vehicles for energy and regulation market participation. IEEE Transactions on Smart Grid, 5(5):2283–2294, 2014.

[84] Zhengwei Li, Gongsheng Huang, and Yongjun Sun. Stochastic chiller sequencing control. Energy and Buildings, 84:203–213, 2014.

[85] Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. Dynamic right-sizing for power-proportional data centers. IEEE/ACM Transactions on Networking (TON), 21(5):1378–1391, 2013.

[86] Yashen Lin, Prabir Barooah, Sean Meyn, and Timothy Middelkoop. Demand side frequency regulation from commercial building hvac systems: An experimental study. In 2015 American Control Conference (ACC), pages 3019–3024. IEEE, 2015.

[87] Yashen Lin, Prabir Barooah, Sean Meyn, and Timothy Middelkoop. Experimental evaluation of frequency regulation from commercial building hvac systems. IEEE Transactions on Smart Grid, 6(2):776–783, 2015.

[88] Zhenhua Liu, Adam Wierman, Yuan Chen, Benjamin Razon, and Niangjun Chen. Data center demand response: Avoiding the coincident peak via workload shifting and local generation. Performance Evaluation, 70(10):770–791, 2013.

[89] PJM Interconnection LLC. Pjm manual 11: Energy & ancillary services market operations, 2019.

[90] Jianying Luo, Lei Rao, and Xue Liu. Temporal load balancing with service delay guarantees for data center energy cost optimization. IEEE Transactions on Parallel and Distributed Systems, 25(3):775–784, 2013.

[91] Mehdi Maasoumy, Jorge Ortiz, David Culler, and Alberto Sangiovanni-Vincentelli. Flexibility of commercial building hvac fan as ancillary service for smart grid. arXiv preprint arXiv:1311.6094, 2013.

[92] Jason MacDonald, Sila Kiliccote, Jim Boch, Jonathan Chen, and Robert Nawy. Commercial building loads providing ancillary services in pjm. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.

[93] Josiah McClurg. Fast Demand Response with Datacenter Loads: A Green Dimension of Big Data. PhD thesis, University of Iowa, 2017.

[94] D Müller, M Lauster, A Constantin, M Fuchs, and P Remmen. Aixlib-an open-source modelica library within the iea-ebc annex 60 framework. BauSIM, 20162016:3–9, 2016.

[95] Iyswarya Narayanan, Di Wang, Anand Sivasubramaniam, Hosam K Fathy, Sean James, et al. Evaluating energy storage for a multitude of uses in the datacenter. In 2017 IEEE International Symposium on Workload Characterization (IISWC), pages 12–21. IEEE, 2017.

[96] Sergiu Nedevschi, Lucian Popa, Gianluca Iannaccone, Sylvia Ratnasamy, and David Wetherall. Reducing network energy consumption via sleeping and rate-adaptation. In NsDI, volume 8, pages 323–336, 2008.

[97] Monica Neukomm, Valerie Nubbe, and Robert Fares. Grid-interactive efficient buildings. Technical report, US Department of Energy, Navigant, 2019.

[98] Tomomichi Noguchi and Hiroaki Nishi. Active controlled shutter for effective cooling of servers in data center. In IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society, pages 1668–1673. IEEE, 2015.

[99] Eduard Oró, Victor Depoorter, Noah Pflugradt, and Jaume Salom. Overview of direct air free cooling and thermal energy storage potential energy savings in data centres. Applied thermal engineering, 85:100–110, 2015.

[100] Martin Otter, Karl-Erik Årzén, and Isolde Dressler. StateGraph-a Modelica library for hierarchical state machines. Modelica 2005 proceedings, 2005.

[101] Yiqun Pan, Rongxin Yin, and Zhizhong Huang. Energy modeling of two office buildings with data center for green building design. Energy and Buildings, 40(7):1145–1152, 2008.

[102] Zachary M Pardey, Dustin W Demetriou, Hamza Salih Erden, James W VanGilder, H Ezzat Khalifa, and Roger R Schmidt. Proposal for standard compact server model for transient data center simulations. ASHRAE Transactions, 121(1):413–422, 2015.

[103] Luca Parolini, Bruno Sinopoli, Bruce H Krogh, and Zhikui Wang. A cyber–physical systems approach to data center modeling and control for energy efficiency. Proceedings of the IEEE, 100(1):254–268, 2011.

[104] Gregory S Pavlak, Gregor P Henze, and Vincent J Cushing. Optimizing commercial building participation in energy and ancillary service markets. Energy and Buildings, 81:115–126, 2014.

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

[106] PJM. Ancillary services. `https://www.pjm.com/markets-and-operations/ancillary-services.aspx`. Accessed: May 6, 2019.

[107] M Trcka Radosevic, JLM Hensen, and AJ Th M Wijsman. Distributed building performance simulationa novel approach to overcome legacy code limitations. HVAC&R Research, 12(S1):621–640, 2006.

[108] Kui Shan, Shengwei Wang, Fu Xiao, and Yongjun Sun. Sensitivity and uncertainty analysis of cooling water control strategies. HVAC&R Research, 19(4):435–443, 2013.

[109] Arman Shehabi, Srirupa Ganguly, Kim Traber, Hillary Price, Arpad Horvath, William W Nazaroff, and Ashok J Gadgil. Energy implications of economizer use in california data centers. Technical report, Lawrence Berkeley National Laboratory, 2008.

[110] Yuanyuan Shi, Bolun Xu, Di Wang, and Baosen Zhang. Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains. IEEE Transactions on Power Systems, 33(3):2882–2894, 2018.

[111] Yuanyuan Shi, Bolun Xu, Baosen Zhang, and Di Wang. Leveraging energy storage to optimize data center electricity cost in emerging power markets. In Proceedings of the Seventh International Conference on Future Energy Systems, page 18. ACM, 2016.

[112] Telecommunication Industry Association. Standards, Technology Dept, and American National Standards Institute. Telecommunications Infrastructure Standard for Data Centers. Telecommunication Industry Association, 2005.

[113] Jeff Stein. Waterside Economizing in Data Centers: Design and Control Considerations. ASHRAE Transactions, 115(2), 2009.

[114] Leo Su and Leslie K Norford. Demonstration of hvac chiller control for power grid frequency regulation-part 1: Controller development and experimental results. Science and Technology for the Built Environment, 21(8):1134–1142, 2015.

[115] Leo Su and Leslie K Norford. Demonstration of hvac chiller control for power grid frequency regulation-part 2: Discussion of results and considerations for broader deployment. Science and Technology for the Built Environment, 21(8):1143–1153, 2015.

[116] Mengshu Sun, Yuankun Xue, Paul Bogdan, Jian Tang, Yanzhi Wang, and Xue Lin. Hierarchical and hybrid energy storage devices in data centers: Architecture, control and provisioning. PloS one, 13(1):e0191450, 2018.

[117] Robert Endre Tarjan. Data structures and network algorithms, volume 44. Siam, 1983.

[118] Steven T Taylor. How to design & control waterside economizers. ASHRAE Journal, 56(6):30–36, 2014.

[119] Wei Tian, Yangyang Fu, Qiujian Wang, Thomas Alonso Sevilla, and Wangda Zuo. Optimization on Thermostat Location in an Office Room Using the Coupled Simulation Platform in Modelica Buildings Library: A Pilot Study. In 2018 COBEE conference, 2018.

[120] Wei Tian, Thomas Alonso Sevilla, Wangda Zuo, and Michael D Sohn. Coupling fast fluid dynamics and multizone airflow models in Modelica Buildings library to simulate the dynamics of HVAC systems. Building and Environment, 122:269–286, 2017.

[121] Michael M Toulouse, Guislain Doljac, Van P Carey, and Cullen Bash. Exploration of a potential-flow-based compact model of air-flow transport in data centers. In ASME 2009 international mechanical engineering congress and exposition, pages 41–50. American Society of Mechanical Engineers Digital Collection, 2009.

[122] Van Giang Tran, Vincent Debusschere, and Seddik Bacha. Data center energy consumption simulator from the servers to their cooling system. In 2013 IEEE Grenoble Conference, pages 1–6, June 2013.

[123] Marija Trcka, Michael Wetter, and Jan Hensen. Comparison of co-simulation approaches for building and hvac/r system simulation. In Proceedings of the International IBPSA Conference, Beijing, China, 2007.

[124] Unitil Corporation. Energy for business, 2017. [Online; accessed 30-December-2017].

[125] Ward Van Heddeghem, Sofie Lambert, Bart Lannoo, Didier Colle, Mario Pickavet, and Piet Demeester. Trends in worldwide ICT electricity consumption from 2007 to 2012. Computer Communications, 50:64–76, 2014.

[126] Alexandra Von Meier. Integration of renewable generation in california: Coordination challenges in time and space. In 11th International Conference on Electrical Power Quality and Utilisation, pages 1–6. IEEE, 2011.

[127] Evangelos Vrettos, Emre C Kara, Jason MacDonald, Göran Andersson, and Duncan S Callaway. Experimental demonstration of frequency regulation by commercial buildings-part i: Modeling and hierarchical control design. IEEE Transactions on Smart Grid, 9(4):3213–3223, 2018.

[128] Evangelos Vrettos, Emre C Kara, Jason MacDonald, Göran Andersson, and Duncan S Callaway. Experimental demonstration of frequency regulation by commercial buildings-part ii: results and performance evaluation. IEEE Transactions on Smart Grid, 9(4):3224–3234, 2018.

[129] Cheng Wang, Bhuvan Urgaonkar, Qian Wang, and George Kesidis. A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing. In 2014 IEEE 22nd International Symposium on Modelling, Analysis & Simulation of Computer and Telecommunication Systems, pages 305–314. IEEE, 2014.

[130] Cheng Wang, Bhuvan Urgaonkar, Qian Wang, George Kesidis, and Anand Sivasubramaniam. Data center power cost optimization via workload modulation. In Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, pages 260–263. IEEE Computer Society, 2013.

[131] Lin Wang, Fa Zhang, Jordi Arjona Aroca, Athanasios V Vasilakos, Kai Zheng, Chenying Hou, Dan Li, and Zhiyong Liu. Greendcn: A general framework for achieving energy efficiency in data center networks. IEEE Journal on Selected Areas in Communications, 32(1):4–15, 2013.

[132] Wei Wang, Amirali Abdolrashidi, Nanpeng Yu, and Daniel Wong. Frequency regulation service provision in data center with computational flexibility. Applied Energy, 251:113304, 2019.

[133] Weimin Wang, Radu Zmeureanu, and Hugues Rivard. Applying multi-objective genetic algorithms in green building design optimization. Building and environment, 40(11):1512–1525, 2005.

[134] Michael Wetter. Simulation-based building energy optimization. PhD thesis, 2004.

[135] Michael Wetter. A Modelica-based model library for building energy and control systems. In Proceedings of the 11th IBPSA Conference, Glasgow, Scotland, pages 652–659, 2009.

[136] Michael Wetter. Modelica-based modelling and simulation to support research and development in building energy and control systems. Journal of Building Performance Simulation, 2(2):143–161, 2009.

[137] Michael Wetter. A view on future building system modeling and simulation. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2011.

[138] Michael Wetter, Marco Bonvini, and Thierry S Nouidui. Equation-based languages A new paradigm for building energy modeling, simulation and optimization. Energy and Buildings, 117:290–300, 2016.

[139] Michael Wetter et al. Genopt-a generic optimization program. In Seventh International IBPSA Conference, Rio de Janeiro, pages 601–608, 2001.

[140] Michael Wetter, Wangda Zuo, Thierry S Nouidui, and Xiufeng Pang. Modelica buildings library. Journal of Building Performance Simulation, 7(4):253–270, 2014.

[141] 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.

[142] Xue Xue, Shengwei Wang, Chengchu Yan, and Borui Cui. A fast chiller power demand response control strategy for buildings connected to smart grid. Applied Energy, 137:77–87, 2015.

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

[144] Don A York and Charlene C Cappiello. Doe-2 engineers manual (version 2. 1a). Technical report, Lawrence Berkeley Lab., CA (USA); Los Alamos National Lab., NM (USA), 1981.

[145] Linquan Zhang, Shaolei Ren, Chuan Wu, and Zongpeng Li. A truthful incentive mechanism for emergency demand response in colocation data centers. In 2015 IEEE Conference on Computer Communications (INFOCOM), pages 2632–2640. IEEE, 2015.

[146] Qi Zhang, Mohamed Faten Zhani, Shuo Zhang, Quanyan Zhu, Raouf Boutaba, and Joseph L Hellerstein. Dynamic energy-aware capacity provisioning for cloud computing environments. In Proceedings of the 9th international conference on Autonomic computing, pages 145–154. ACM, 2012.

[147] Yanwei Zhang, Yefu Wang, and Xiaorui Wang. Testore: Exploiting thermal and energy storage to cut the electricity bill for datacenter cooling. In Proceedings of the 8th International Conference on Network and Service Management, pages 19–27. International Federation for Information Processing, 2012.

[148] Peng Zhao, Gregor P Henze, Sandro Plamp, and Vincent J Cushing. Evaluation of commercial building hvac systems as frequency regulation providers. Energy and Buildings, 67:225–235, 2013.

[149] Wenli Zheng, Kai Ma, and Xiaorui Wang. Exploiting thermal energy storage to reduce data center capital and operating expenses. In 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA), pages 132–141. IEEE, 2014.

Appendix A

Performance Scores for Different Scenarios

Table A.1: Performance for RegA test signal during light load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.6536	0.5749	0.3861	0.5382
		10	0.6509	0.5857	0.3853	0.5406
		15	0.6362	0.6125	0.3864	0.5450
	1.0	5	0.8506	0.7382	0.6439	0.7442
		10	0.8475	0.7292	0.6435	0.7401
		15	0.8479	0.7224	0.6456	0.7386
	1.1	5	0.9149	0.8150	0.7614	0.8304
		10	0.9148	0.8049	0.7592	0.8263
		15	0.9129	0.8210	0.7598	0.8312
268.0	0.9	5	0.6542	0.5815	0.3322	0.5226
		10	0.6425	0.6286	0.3318	0.5343
		15	0.6500	0.5618	0.3315	0.5144
	1.0	5	0.7231	0.6886	0.4702	0.6273
		10	0.7008	0.5936	0.4703	0.5882
		15	0.7037	0.5885	0.4723	0.5882
	1.1	5	0.7967	0.7068	0.5455	0.6830
		10	0.8020	0.7382	0.5433	0.6945
		15	0.8040	0.7410	0.5415	0.6955
402.0	0.9	5	0.6353	0.5606	0.3187	0.5048
		10	0.6445	0.5724	0.3173	0.5114
		15	0.6498	0.5331	0.3187	0.5005
	1.0	5	0.6717	0.6169	0.4120	0.5669
		10	0.6770	0.6035	0.4109	0.5638
		15	0.6864	0.5817	0.4112	0.5597
	1.1	5	0.7124	0.6144	0.4651	0.5973
		10	0.6964	0.6435	0.4647	0.6015
		15	0.7136	0.6354	0.4632	0.6041

Table A.2: Performance for RegA test signal during medium load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8258	0.4914	0.4627	0.5933
		10	0.8280	0.4957	0.4604	0.5947
		15	0.8280	0.4954	0.4603	0.5946
	1.0	5	0.9987	1.0000	0.9469	0.9819
		10	0.9987	1.0000	0.9475	0.9821
		15	0.9987	1.0000	0.9472	0.9820
	1.1	5	1.0000	1.0000	0.9475	0.9825
		10	1.0000	1.0000	0.9474	0.9825
		15	1.0000	1.0000	0.9473	0.9824
268.0	0.9	5	0.8168	0.4972	0.4084	0.5741
		10	0.8170	0.4938	0.4082	0.5730
		15	0.8145	0.4897	0.4088	0.5710
	1.0	5	0.9063	0.8229	0.7694	0.8329
		10	0.9061	0.8210	0.7699	0.8323
		15	0.9063	0.8199	0.7701	0.8321
	1.1	5	0.9774	0.9371	0.8784	0.9310
		10	0.9786	0.9342	0.8787	0.9305
		15	0.9776	0.9375	0.8788	0.9313
402.0	0.9	5	0.8121	0.4988	0.3819	0.5642
		10	0.8120	0.4988	0.3819	0.5642
		15	0.8119	0.5039	0.3817	0.5658
	1.0	5	0.8392	0.7132	0.6391	0.7305
		10	0.8404	0.7126	0.6390	0.7307
		15	0.8392	0.7135	0.6391	0.7306
	1.1	5	0.9600	0.8056	0.7428	0.8361
		10	0.9602	0.8051	0.7406	0.8353
		15	0.9600	0.8051	0.7407	0.8353

Table A.3: Performance for RegA test signal during heavy load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8103	0.4501	0.3822	0.5476
		10	0.8100	0.4499	0.3822	0.5474
		15	0.8098	0.4499	0.3819	0.5472
	1.0	5	1.0000	1.0000	0.9474	0.9825
		10	1.0000	1.0000	0.9492	0.9831
		15	1.0000	1.0000	0.9494	0.9831
	1.1	5	1.0000	1.0000	0.9471	0.9824
		10	1.0000	1.0000	0.9473	0.9824
		15	1.0000	1.0000	0.9475	0.9825
268.0	0.9	5	0.8286	0.4501	0.3706	0.5498
		10	0.8283	0.4503	0.3707	0.5497
		15	0.8286	0.4499	0.3707	0.5497
	1.0	5	0.9847	0.9508	0.9126	0.9494
		10	0.9846	0.9508	0.9126	0.9493
		15	0.9846	0.9511	0.9107	0.9488
	1.1	5	1.0000	1.0000	0.9471	0.9824
		10	1.0000	1.0000	0.9470	0.9823
		15	1.0000	1.0000	0.9490	0.9830
402.0	0.9	5	0.8366	0.4526	0.3665	0.5519
		10	0.8366	0.4529	0.3663	0.5520
		15	0.8364	0.4526	0.3668	0.5520
	1.0	5	0.9714	0.8850	0.8017	0.8860
		10	0.9714	0.8875	0.7995	0.8861
		15	0.9714	0.8875	0.7998	0.8862
	1.1	5	0.9983	0.9989	0.9454	0.9809
		10	0.9983	0.9989	0.9468	0.9813
		15	0.9983	0.9989	0.9455	0.9809

Table A.4: Performance for RegA test signal during light load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9869	0.9431	0.4293	0.7864
		10	0.9879	0.9351	0.4346	0.7859
		15	0.9869	0.9404	0.4310	0.7861
	1.0	5	0.9703	0.8682	0.7080	0.8488
		10	0.9726	0.8761	0.7062	0.8516
		15	0.9712	0.8467	0.7053	0.8411
	1.1	5	0.9879	0.9051	0.8160	0.9030
		10	0.9877	0.9011	0.8160	0.9016
		15	0.9870	0.9021	0.8135	0.9009
268.0	0.9	5	0.9788	0.9468	0.3557	0.7604
		10	0.9802	0.9424	0.3556	0.7594
		15	0.9806	0.9432	0.3550	0.7596
	1.0	5	0.9735	0.9187	0.5078	0.8000
		10	0.9722	0.8940	0.5084	0.7915
		15	0.9672	0.9146	0.5074	0.7964
	1.1	5	0.9409	0.9104	0.5747	0.8087
		10	0.9418	0.9064	0.5728	0.8070
		15	0.9431	0.8875	0.5740	0.8015
402.0	0.9	5	0.9751	0.9463	0.3318	0.7511
		10	0.9759	0.9426	0.3310	0.7498
		15	0.9767	0.9424	0.3307	0.7499
	1.0	5	0.9758	0.9144	0.4329	0.7744
		10	0.9754	0.9149	0.4325	0.7743
		15	0.9763	0.9164	0.4330	0.7752
	1.1	5	0.9694	0.9211	0.4809	0.7905
		10	0.9694	0.9188	0.4810	0.7897
		15	0.9690	0.9165	0.4813	0.7890

Table A.5: Performance for RegA test signal during medium load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9035	0.9015	0.4903	0.7651
		10	0.9254	0.9336	0.4839	0.7810
		15	0.9278	0.9296	0.5109	0.7894
	1.0	5	1.0000	1.0000	0.9446	0.9815
		10	1.0000	1.0000	0.9477	0.9826
		15	1.0000	1.0000	0.9481	0.9827
	1.1	5	1.0000	1.0000	0.9452	0.9817
		10	1.0000	1.0000	0.9493	0.9831
		15	1.0000	1.0000	0.9459	0.9820
268.0	0.9	5	0.9030	0.9086	0.4157	0.7424
		10	0.9330	0.9421	0.4208	0.7653
		15	0.9339	0.9556	0.4242	0.7712
	1.0	5	0.9370	0.9243	0.8400	0.9004
		10	0.9461	0.9282	0.8449	0.9064
		15	0.9516	0.9339	0.8431	0.9096
	1.1	5	0.9767	0.9861	0.9362	0.9663
		10	0.9778	0.9856	0.9341	0.9658
		15	0.9788	0.9857	0.9330	0.9658
402.0	0.9	5	0.9024	0.9081	0.3954	0.7353
		10	0.9285	0.9493	0.3994	0.7591
		15	0.9215	0.9256	0.4011	0.7494
	1.0	5	0.9045	0.8418	0.6906	0.8123
		10	0.9164	0.8286	0.6961	0.8137
		15	0.9206	0.8311	0.6940	0.8152
	1.1	5	0.9475	0.9043	0.8103	0.8874
		10	0.9569	0.9114	0.8139	0.8941
		15	0.9643	0.9172	0.8129	0.8981

Table A.6: Performance for RegA test signal during heavy load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8226	0.9308	0.3526	0.7020
		10	0.9061	0.9661	0.3405	0.7376
		15	0.9356	0.9846	0.3597	0.7600
	1.0	5	1.0000	1.0000	0.9480	0.9827
		10	1.0000	1.0000	0.9479	0.9826
		15	1.0000	1.0000	0.9459	0.9820
	1.1	5	1.0000	1.0000	0.9473	0.9824
		10	1.0000	1.0000	0.9495	0.9832
		15	1.0000	1.0000	0.9487	0.9829
268.0	0.9	5	0.8225	0.9246	0.3456	0.6976
		10	0.9050	0.9607	0.3426	0.7361
		15	0.9314	0.9747	0.3575	0.7545
	1.0	5	1.0000	1.0000	0.9465	0.9822
		10	1.0000	1.0000	0.9502	0.9834
		15	1.0000	1.0000	0.9445	0.9815
	1.1	5	1.0000	1.0000	0.9460	0.9820
		10	1.0000	1.0000	0.9484	0.9828
		15	1.0000	1.0000	0.9482	0.9827
402.0	0.9	5	0.8395	0.9537	0.3465	0.7132
		10	0.9150	0.9637	0.3538	0.7442
		15	0.9367	0.9686	0.3602	0.7552
	1.0	5	0.9444	0.9624	0.8906	0.9324
		10	0.9621	0.9733	0.9047	0.9467
		15	0.9744	0.9792	0.9087	0.9541
	1.1	5	1.0000	1.0000	0.9483	0.9828
		10	1.0000	1.0000	0.9479	0.9826
		15	1.0000	1.0000	0.9498	0.9833

Table A.7: Performance for RegD test signal during light load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9041	0.8353	0.5876	0.7757
		10	0.9006	0.8325	0.5867	0.7733
		15	0.9045	0.8354	0.5882	0.7760
	1.0	5	0.9432	0.8467	0.6744	0.8214
		10	0.9445	0.8471	0.6740	0.8219
		15	0.9445	0.8468	0.6748	0.8220
	1.1	5	0.9466	0.8669	0.7405	0.8514
		10	0.9465	0.8628	0.7399	0.8497
		15	0.9460	0.8708	0.7399	0.8523
268.0	0.9	5	0.7792	0.8156	0.4327	0.6758
		10	0.7777	0.8129	0.4328	0.6745
		15	0.7800	0.8047	0.4331	0.6726
	1.0	5	0.8004	0.8172	0.4823	0.7000
		10	0.8011	0.8168	0.4819	0.6999
		15	0.7981	0.8125	0.4811	0.6973
	1.1	5	0.8021	0.8263	0.5043	0.7109
		10	0.8011	0.8210	0.5049	0.7090
		15	0.8032	0.8149	0.5032	0.7071
402.0	0.9	5	0.7524	0.7978	0.3622	0.6375
		10	0.7506	0.7792	0.3636	0.6311
		15	0.7532	0.7931	0.3633	0.6365
	1.0	5	0.7657	0.7957	0.3979	0.6531
		10	0.7663	0.7894	0.3974	0.6511
		15	0.7696	0.7933	0.3980	0.6537
	1.1	5	0.7709	0.7946	0.4135	0.6597
		10	0.7736	0.7958	0.4131	0.6608
		15	0.7737	0.7978	0.4135	0.6617

Table A.8: Performance for RegD test signal during medium load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9175	0.8179	0.6286	0.7880
		10	0.9175	0.8179	0.6286	0.7880
		15	0.9176	0.8179	0.6286	0.7880
	1.0	5	1.0000	1.0000	0.8634	0.9545
		10	1.0000	1.0000	0.8636	0.9545
		15	1.0000	1.0000	0.8636	0.9545
	1.1	5	1.0000	1.0000	0.8644	0.9548
		10	1.0000	1.0000	0.8644	0.9548
		15	1.0000	1.0000	0.8644	0.9548
268.0	0.9	5	0.9185	0.8232	0.6132	0.7850
		10	0.9185	0.8232	0.6132	0.7850
		15	0.9185	0.8232	0.6132	0.7849
	1.0	5	0.9530	0.8500	0.7476	0.8502
		10	0.9531	0.8500	0.7477	0.8503
		15	0.9530	0.8500	0.7480	0.8503
	1.1	5	0.9861	1.0000	0.8222	0.9361
		10	0.9861	1.0000	0.8218	0.9360
		15	0.9860	1.0000	0.8221	0.9360
402.0	0.9	5	0.9174	0.8272	0.5819	0.7755
		10	0.9176	0.8272	0.5819	0.7756
		15	0.9175	0.8272	0.5820	0.7756
	1.0	5	0.9471	0.8417	0.6680	0.8189
		10	0.9473	0.8417	0.6680	0.8190
		15	0.9472	0.8417	0.6680	0.8190
	1.1	5	0.9488	0.8500	0.7232	0.8407
		10	0.9488	0.8500	0.7232	0.8407
		15	0.9489	0.8500	0.7232	0.8407

Table A.9: Performance for RegD test signal during heavy load in FC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.8850	0.7857	0.6005	0.7571
		10	0.8850	0.7857	0.6005	0.7571
		15	0.8852	0.7857	0.6006	0.7572
	1.0	5	1.0000	1.0000	0.8635	0.9545
		10	1.0000	1.0000	0.8636	0.9545
		15	1.0000	1.0000	0.8635	0.9545
	1.1	5	1.0000	1.0000	0.8635	0.9545
		10	1.0000	1.0000	0.8635	0.9545
		15	1.0000	1.0000	0.8636	0.9545
268.0	0.9	5	0.9050	0.7911	0.5984	0.7648
		10	0.9050	0.7911	0.5984	0.7648
		15	0.9050	0.7911	0.5983	0.7648
	1.0	5	0.9885	1.0000	0.8334	0.9406
		10	0.9884	1.0000	0.8333	0.9406
		15	0.9885	1.0000	0.8334	0.9406
	1.1	5	1.0000	1.0000	0.8639	0.9546
		10	1.0000	1.0000	0.8639	0.9546
		15	1.0000	1.0000	0.8639	0.9546
402.0	0.9	5	0.9129	0.7988	0.5957	0.7691
		10	0.9131	0.7989	0.5956	0.7692
		15	0.9131	0.7989	0.5956	0.7692
	1.0	5	0.9536	0.8625	0.7621	0.8594
		10	0.9536	0.8625	0.7617	0.8593
		15	0.9537	0.8625	0.7619	0.8594
	1.1	5	0.9988	1.0000	0.8585	0.9524
		10	0.9988	1.0000	0.8583	0.9524
		15	0.9988	1.0000	0.8584	0.9524

Table A.10: Performance for RegD test signal during light load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9530	0.9940	0.6341	0.8604
		10	0.9550	0.9935	0.6305	0.8597
		15	0.9547	0.9943	0.6338	0.8610
	1.0	5	0.9615	0.9064	0.7398	0.8692
		10	0.9612	0.9444	0.7379	0.8812
		15	0.9620	0.9522	0.7370	0.8837
	1.1	5	0.9667	0.9947	0.7949	0.9188
		10	0.9667	0.9989	0.7922	0.9193
		15	0.9654	0.9989	0.7924	0.9189
268.0	0.9	5	0.8972	0.9628	0.4784	0.7795
		10	0.8991	0.9824	0.4790	0.7868
		15	0.8906	0.9760	0.4779	0.7815
	1.0	5	0.9003	0.9304	0.5275	0.7861
		10	0.9021	0.9390	0.5274	0.7895
		15	0.9017	0.9425	0.5271	0.7904
	1.1	5	0.9015	0.9276	0.5478	0.7923
		10	0.9016	0.9562	0.5461	0.8013
		15	0.9013	0.9667	0.5468	0.8049
402.0	0.9	5	0.9080	0.9599	0.3887	0.7522
		10	0.9122	0.9608	0.3886	0.7539
		15	0.9133	0.9843	0.3899	0.7625
	1.0	5	0.9042	0.9196	0.4242	0.7493
		10	0.9083	0.9254	0.4226	0.7521
		15	0.9084	0.9326	0.4237	0.7549
	1.1	5	0.9001	0.9050	0.4371	0.7474
		10	0.9050	0.9247	0.4369	0.7556
		15	0.8520	0.9325	0.4377	0.7408

Table A.11: Performance for RegD test signal during medium load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9615	0.9992	0.6875	0.8827
		10	0.9726	0.9994	0.6906	0.8875
		15	0.9769	0.9988	0.6891	0.8883
	1.0	5	1.0000	1.0000	0.8642	0.9547
		10	1.0000	1.0000	0.8648	0.9549
		15	1.0000	1.0000	0.8649	0.9550
	1.1	5	1.0000	1.0000	0.8648	0.9549
		10	1.0000	1.0000	0.8648	0.9549
		15	1.0000	1.0000	0.8652	0.9551
268.0	0.9	5	0.9491	0.9986	0.6364	0.8614
		10	0.9599	0.9983	0.6375	0.8653
		15	0.9649	0.9983	0.6390	0.8674
	1.0	5	0.9792	0.9999	0.8067	0.9286
		10	0.9801	0.9997	0.8059	0.9286
		15	0.9790	0.9996	0.8071	0.9286
	1.1	5	0.9991	1.0000	0.8589	0.9527
		10	0.9991	1.0000	0.8559	0.9517
		15	0.9992	1.0000	0.8595	0.9529
402.0	0.9	5	0.9436	0.9978	0.6226	0.8547
		10	0.9568	0.9979	0.6252	0.8600
		15	0.9596	0.9974	0.6247	0.8605
	1.0	5	0.9597	0.9657	0.7267	0.8840
		10	0.9617	0.9739	0.7278	0.8878
		15	0.9643	0.9738	0.7262	0.8881
	1.1	5	0.9716	0.9997	0.7952	0.9222
		10	0.9731	0.9992	0.7934	0.9219
		15	0.9722	0.9992	0.7940	0.9218

Table A.12: Performance for RegD test signal during heavy load in FMC mode

C_{reg}	β	τ	accuracy	delay	precision	performance
134.0	0.9	5	0.9579	0.9999	0.7003	0.8860
		10	0.9808	1.0000	0.7674	0.9161
		15	0.9894	1.0000	0.7814	0.9236
	1.0	5	1.0000	1.0000	0.8657	0.9552
		10	1.0000	1.0000	0.8644	0.9548
		15	1.0000	1.0000	0.8657	0.9552
	1.1	5	1.0000	1.0000	0.8663	0.9554
		10	1.0000	1.0000	0.8671	0.9557
		15	1.0000	1.0000	0.8659	0.9553
268.0	0.9	5	0.9479	0.9996	0.6357	0.8611
		10	0.9717	0.9990	0.6548	0.8752
		15	0.9817	0.9990	0.6587	0.8798
	1.0	5	1.0000	1.0000	0.8645	0.9548
		10	1.0000	1.0000	0.8650	0.9550
		15	1.0000	1.0000	0.8642	0.9547
	1.1	5	1.0000	1.0000	0.8656	0.9552
		10	1.0000	1.0000	0.8650	0.9550
		15	1.0000	1.0000	0.8644	0.9548
402.0	0.9	5	0.9439	0.9988	0.6164	0.8530
		10	0.9665	0.9967	0.6272	0.8634
		15	0.9762	0.9985	0.6332	0.8693
	1.0	5	0.9955	1.0000	0.8430	0.9461
		10	0.9967	1.0000	0.8503	0.9490
		15	0.9971	1.0000	0.8498	0.9490
	1.1	5	1.0000	1.0000	0.8636	0.9545
		10	1.0000	1.0000	0.8645	0.9548
		15	1.0000	1.0000	0.8650	0.9550