Uncertainty and Decentralization: Two Themes in an Energy Transformation

Thesis by
Lucien D. Werner

In Partial Fulfillment of the Requirements for the
Degree of
Computing & Mathematical Sciences

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 25, 2023

© 2023

Lucien D. Werner
ORCID: 0000-0003-1613-1702

All rights reserved

ACKNOWLEDGEMENTS

The five and a half years I have spent at Caltech have been the most formative of my career. I am indebted to the professional and personal support I have received from the communities around me while on this journey.

First, I am grateful to my advisor Steven Low who has guided my graduate work and professional development with wisdom, creativity, dedication, and enthusiasm. His approaches to science, research, and entrepreneurship model what I aspire towards.

I am thankful to my co-advisor Adam Wierman for his impact on my research in large and small ways and for his tireless dedication to the growth of his students.

I am honored to be an alumnus of the the Netlab and RSRG groups. Although I didn't think I came to Caltech for the wonderful human beings I would meet here, being part of the kind, capable, and humble community that Steven and Adam have built have made me realize the true legacy of my time at Caltech.

Finally, I am thankful every day for my family Susan, Peter, Hélène, Mariel, Andrée, and Millicent, and for my partner Anuschka. To them I owe a large part of reaching this milestone in my life and I love them all deeply.

ABSTRACT

Over the last two decades, the rapidly decreasing units costs of solar, wind, and energy storage technologies have launched a fundamental transformation in how electric power is produced, distributed, and consumed. Proliferation of these technologies has effected a shift towards a more decentralized, flexible, and sustainable energy system that can meet the growing demand for energy while reducing greenhouse gas emissions from fossil fuels. The work in this thesis studies two principal themes in this transformation: uncertainty and decentralization.

Uncertainty is a key challenge in the modern grid resulting from the weather dependence of variable renewables and volatile loads like electric vehicles distributed throughout the grid. Electricity markets, whose function is to regulate the precise balance of supply and demand across the system, face a pressing need for dispatch mechanisms that account for uncertainty while providing participation incentives for generators and loads. We introduce a framework for multi-stage market dispatch and pricing under a general description of forecast uncertainty that enables system operators to explicitly incorporate uncertainty into market-clearing prices. In related work, we study mechanisms that guarantee feasibility of multi-interval dispatch under robust uncertainty and provide participation incentives for shiftable demand response in forward multi-interval markets.

The trend towards a more decentralized energy system stems from the inherent modularity of distributed energy resources (DERs), such as solar and storage, as well as the persistent growth in end-use loads. This evolution presents significant challenges to system operators who typically lack the tools and processes for managing a complex, distributed power system. To fill this gap, we introduce and implement a Microgrid Operating System (OS), a software platform for monitoring, modeling, and optimizing microgrids and distribution systems. The Microgrid OS is a central layer that links DER hardware, such as batteries, solar, and flexible loads, to energy applications like cost minimization, emissions reduction, and wholesale market participation. The core functions it provides are data acquisition and processing, system modeling and learning, and optimization & control. We present key modules of the Microgrid OS in the context of several implementation projects in microgrids, commercial buildings, and distribution networks.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Ognjen Stanojev et al. "Tractable Identification of Electric Distribution Networks". In: *arXiv preprint arXiv:2304.01615* (2023).
L.D.W helped with the conception of the project and participated in the prepration of the manuscript.

[2] Lucien D Werner et al. "Pricing uncertainty in stochastic multi-stage electricity markets". In: *Manuscript submitted to 2023 IEEE 64th Conference on Decision and Control (CDC)* (2023).
L.D.W conceived of the project, formulated and proved the theoretical results, implemented the simulation code, analyzed the results, and participated in the writing of the manuscript.

[3] Nicolas* Christianson et al. "Dispatch-aware planning for feasible power system operation". In: *Electric Power Systems Research* 212 (2022), p. 108597. DOI: 10.1016/j.epsr.2022.108597.
L.D.W conceived of the project, participated in designing the algorithm, participated in proving the theoretical results, implemented the simulation code, participated in analyzing the results, and participated in writing and revising the manuscript.

[4] Rupa Kurinchi-Vendhan et al. "Wisosuper: Benchmarking super-resolution methods on wind and solar data". In: *arXiv preprint arXiv:2109.08770* (2021).
L.D.W participated in the conception of the project and participated in the preparation of the manuscript.

[5] Syama Sundar* Rangapuram et al. "End-to-end learning of coherent probabilistic forecasts for hierarchical time series". In: *International Conference on Machine Learning* (2021), pp. 8832–8843.
L.D.W participated in the conception of the project, participated in the development of the algorithm, participated in the code implementation, participated in analyzing the results, and participated in preparing the manuscript and revisions.

[6] Lucien Werner, Adam Wierman, and Steven H Low. "Pricing flexibility of shiftable demand in electricity markets". In: *Proceedings of the Twelfth ACM International Conference on Future Energy Systems* (2021), pp. 1–14. DOI: 10.1145/3447555.3464847.
L.D.W. conceived the project, derived the theoretical results, implemented code for simulations, analyzed results, and wrote the manuscript.

[7] Tongxin Li, Lucien Werner, and Steven H Low. "Learning graphs from linear measurements: Fundamental trade-offs and applications". In: *IEEE Transactions on Signal and Information Processing over Networks* 6 (2020), pp. 163–178. DOI: 10.1109/TSIPN.2020.2975368.

L.D.W participated in the conception of the project, participated in the design of the main algorithm, implemented simulation code, prepared the results, and participated in the writing of the manuscript.

[8] Tongxin Li, Lucien Werner, and Steven H Low. "Learning Graph Parameters from Linear Measurements: Fundamental Trade-offs and Application to Electric Grids". In: *2019 IEEE 58th Conference on Decision and Control (CDC)* (2019), pp. 6554–6559. DOI: 10.1109/CDC40024.2019.9029949. L.D.W participated in the conception of the project, participated in the design of the main algorithm, implemented simulation code, prepared the results, and participated in the writing of the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Part I

Background and Introduction

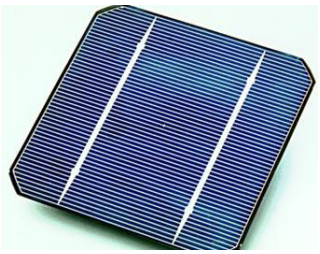
Chapter 1

INTRODUCTION

Over the last two decades, the rapidly decreasing units costs of solar, wind, and energy storage technologies have caused a fundamental transformation in how electric power is produced, distributed, and consumed. Their proliferation has effected a shift towards a more decentralized, flexible, and sustainable energy system that can meet the growing demand for energy while reducing greenhouse gas emissions from producing and combusting fossil fuels.

In addition to their low-carbon footprint, there are two key characteristics of solar, wind, and battery storage that distinguish them from the conventional generation technologies on the grid. First is that they are modular and scale from household size (10^2 W) to utility scale (10^9 W) with the same core subcomponents (e.g., solar cells, battery modules). This distinguishes renewable generation and storage technologies from traditional thermal generation types such as coal, nuclear, and gas plants, each of which is comprised of only a few large generating units (steam/gas turbine driving an electrical generator) with a characteristic scale of 10^8 W.

Second, the modularity of solar, wind, and battery storage enables a decentralized generation footprint, with renewable generation assets spread throughout the grid rather than concentrated at a handful of large facilities (Figure 1.2). This is opposed to the centralization of conventional thermal generating facilities into a small number of locations. There are fewer than 2,500 fossil fuel generating facilities in the US as



(a) 1.5W individual photovoltaic (PV) cell

(b) 250W solar panels on a home rooftop in the US [115]

(c) 2.2GW Bhadla Solar Park in Rajasthan, India [101]

Figure 1.1: Solar power generation size scales by using multiples of the same component PV cells, wired together in series and parallel and combined into panels and arrays.

of 2021 [128].

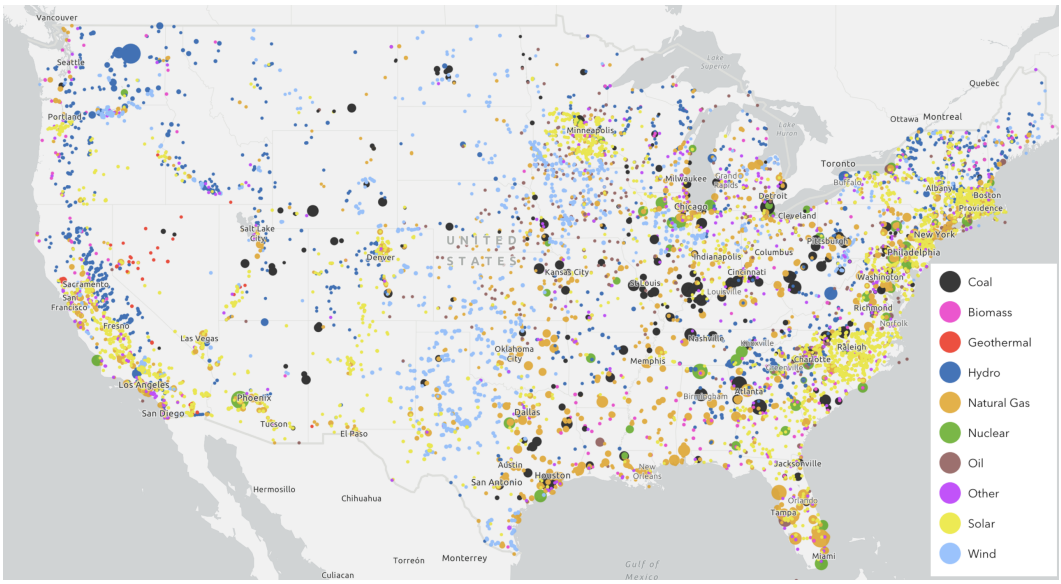


Figure 1.2: Map of generating plants in the US in 2023. Solar and wind are in yellow and blue circles, the size of which represents the size of the power plant [126]. The typical gas or coal power plant is between 500-1000MW, whereas the average size of a solar generator is 1-5MW.

In addition to the trends of modularity and proliferation of renewables, increased social consciousness around decarbonization has been driving electrification of end-use energy, particularly transportation and heating. This has resulted in increased variability and intensity of large loads like electric vehicle (EV) charging stations distributed throughout the grid (Figure 1.3).

Together, these changes in electricity supply and demand are raising significant engineering challenges for system operators, system planners, and end-use consumers of electrical energy. The research in this thesis addresses two of these challenges: 1) electricity market dispatch and pricing under uncertainty and 2) modeling and control of distributed energy resources (DERs).

1.1 Challenges facing a distributed, high-renewables energy system

The work in this thesis targets two principal trends in the energy system transformation currently underway: 1) increasing uncertainty and 2) decentralization of generation and load. Uncertainty arises from the weather-dependence of renewable energy, which causes the patterns of production from these sources to be spatially and temporally unpredictable. Decentralization, which is manifested in generation sources increasingly being located throughout the distribution grid, is being driven

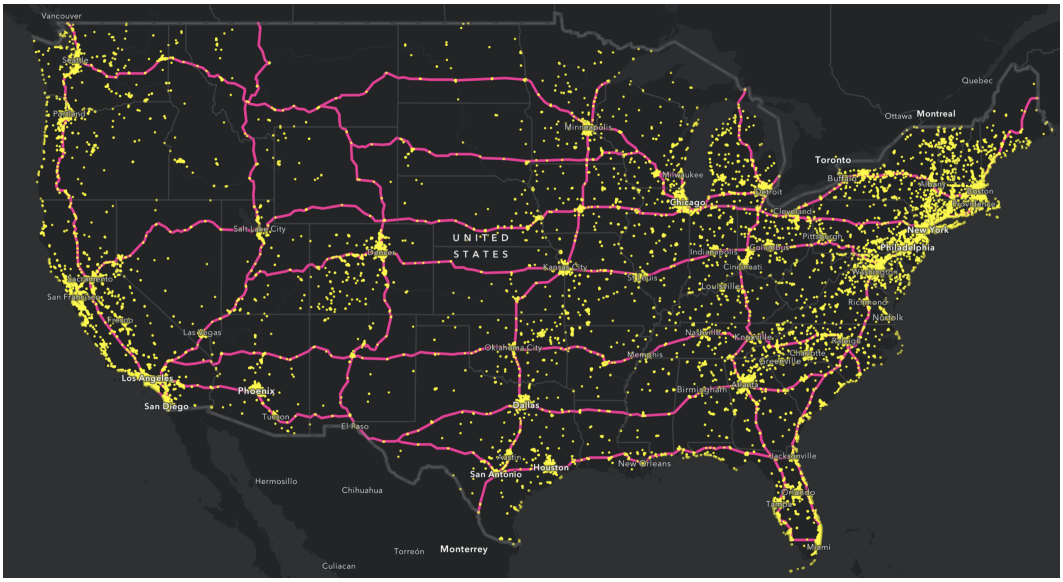


Figure 1.3: Map of EV charging stations in the US in 2022 [47]. DC fast chargers being deployed by Electrify America, a public EV charging network in the US, can draw up to 350kW which is over 100x the power draw of a single-family home (typically 1-2kW).

by accelerating electrification thermal (heating, cooling) and transportation (electric vehicles) end-use energy alongside distributed generation from renewables and storage. Both of these trends increasingly conflict with the control methodologies used to balance the grid today.

The first of these two themes is particularly relevant to system operators whose mandate is to manage the delivery of reliable electric power within their geographic territories. They must dispatch available supply at lowest cost to meet predicted demand, while maintaining a precise balance between these two quantities at all times. This problem is referred to as economic dispatch. Increasingly, it is crucial for system operators to quantify and manage the uncertainty in both net load and supply.

The second theme is of particular interest to end-use customers and distribution utilities. Customers increasingly care about reliability of their power supplies in the face of electrification of all energy use and extreme weather events that threaten the stability of the broader grid. In addition, societal consciousness about the importance of decarbonization is leading to proliferation of distributed zero-carbon power sources like rooftop solar and batteries. The propagation of distributed renewables creates challenges for distribution utilities such as the bi-directional flow of power, power quality degradation, overloading of lines and transformers,

and overall system stability. Utilities often do not have real-time visibility into their networks and therefore these issues are difficult to anticipate and mitigate until after the fact.

We now present the particular engineering and control consequences of uncertainty and decentralization that are the objects of study in this thesis.

Challenge #1: Dispatch mechanisms to handle uncertainty

Economic dispatch is the process of allocating available power generation resources to meet the demand for electricity while minimizing the total cost of generation. In a power system, there are usually multiple power generation sources available, such as thermal power plants, hydroelectric power plants, wind farms, and solar photovoltaic (PV) systems. Economic dispatch involves solving a mathematical optimization problem that minimizes the cost of producing electricity from each available source, as well as any constraints on the power generation, such as minimum and maximum output levels, ramp rates, and transmission limitations. The dispatch algorithm is typically run multiple times, to account for changes in electricity demand and the availability of power sources.

Uncertainty in the availability of power sources or in the demand for electricity can lead to suboptimal dispatch decisions, which can result in higher costs, increased emissions, or even blackouts. One challenge of economic dispatch under uncertainty is the need for accurate and timely information about uncertain parameters, such as weather forecasts for renewable production or electricity demand. Another challenge is the difficulty in balancing the trade-off between cost minimization and risk mitigation. For example, robust dispatch may be more expensive than deterministic dispatch but it provides greater protection against unexpected events. Additionally, some economic dispatch mechanisms, such as real-time dispatch or demand response, may require more flexible or responsive power systems, which can require additional investment in infrastructure and technology, such as fast-ramping resources like gas peaker plants or battery storage. Overall, the challenges of economic dispatch under uncertainty highlight the importance of developing robust and adaptive dispatch mechanisms that can handle a range of possible scenarios and ensure the reliability and efficiency of power systems.

The problem of uncertain economic dispatch is a currently a topic of particular interest both researchers and practitioners. Recently, research community has focused on applying techniques from robust and stochastic optimization to the dispatch prob-

lem. In practice, system operators have focused on methods such as multi-interval dispatch, multiple forward markets, and ancillary services such as operating reserves and ramp reserves.

Challenge #2: Pricing mechanisms to handle uncertainty

Related to the scheduling problem in economic dispatch is the dual problem of determining market clearing prices that support the system operator's desired dispatch. The standard theory of competitive equilibrium pricing introduced by Walras, Arrow, and Debreu is widely adopted in electricity markets around the world as a means of determining efficient clearing prices. [6] This theory was refined in the 1990s by Schweppe and Hogan to account for particular aspects of electricity systems such as transmission constraints. [118, 62, 66] More recently, price formation under uncertainty has received attention from the research community, particularly for multi-interval and multi-settlement markets [139, 57, 28, 94, 42].

However, in the existing literature, when the dispatch problem is posed as a stochastic or robust optimization problem, the price formation depends explicitly on the choice of primal dispatch procedure, which often ignores important implementations details that arise in practice. This constrains the flexibility and adaptability of many of the proposed pricing mechanisms. In contrast, the pricing being developed by system operators are often explicitly tailored to challenges in a particular market or geography and are not designed with broad market principles in mind. This also limits the generalizability of these approaches.

A related set of issues is the pricing of electricity market products (or ancillary services) beyond energy. Examples of these include capacity payments, demand response, operating reserves, and ramp reserves. Each of these services is related to the pricing of uncertainty at a particular time scale. However, understanding of how markets for the different services interact and how incentives align for participants in them is still nascent.

Challenge #3: DER control and coordination

As distributed energy resources (DERs) are becoming increasingly prevalent in modern power systems, their coordination and control are presenting significant challenges. One of the main challenges in DER coordination and control is the integration of different types of resources, such as solar, wind, energy storage, HVAC systems, flexible loads, electric vehicles. Coordinating the output of these resources requires sophisticated control algorithms to ensure that the system operates

efficiently and reliably. Another challenge is the variability of these resources, which can lead to fluctuations in power output and make it difficult to maintain grid stability (e.g., voltage fluctuations). Additionally, there is often a lack of visibility and control over DERs, as many of them are owned and operated by individual customers rather than utilities. This can make it challenging to coordinate and control DERs in a way that benefits the entire system. Finally, regulatory and market structures may not incentivize optimal coordination and control of DERs, which can impede their integration into the power system.

Aggregations of DERs, such as microgrids and virtual power plants (VPPs) have been the focus of increasing interest from end-use customers as well as system operators. Despite this interest, there is still comparatively little practical experience with microgrids. In contrast, microgrid control has been studied extensively by researchers; see for example [125]. These leaves a theory-practice gap that needs to be filled in order for DERs to scale widely in the energy system.

Challenge #4: Modeling DERs and microgrids

Microgrids are physical networks of DERs with complex topologies and varying levels of interconnectedness. Modeling the physical system accurately requires detailed information about the location, capacity, and characteristics of the different components, including generation and load sources, transformers, lines, switches, and breakers. Such information is often lacking or if it is available, it is contained in a diverse set of singleline diagrams or asset databases. Translating these varied sources into a single, coordinated model is a major challenge in practice which is not considered my much of the research work on microgrids.

Accurate modeling of distribution systems and microgrids also requires high-quality timeseries data on the state of system components. Examples include nodal voltages, line flows through conductors, power injections from generators and loads, and power quality metrics such as power factor, harmonic distortion and phase balance. This data can be difficult to obtain, particularly for smaller systems or for components that are owned and operated by individual customers or component operators. Thus, another major practical challenge that is also frequently overlooked by researchers is obtaining a cleaned, organized, and reliable timeseries data for microgrid systems.

These modeling difficulties leave a gap in the tools available for modeling micorgrids. At the network scale, there are the tools used by distribution system operators such as OpenDSS, Gridlab-D, Cyme, and Milsoft to model low-voltage distribution systems.

At the building and component scale, there are detailed component models of assets like transformers, breakers, HVAC systems, DERs, and buildings (e.g., Building Information Models [7]). However, there is a lack of modeling languages that integrate all of the components into a cohesive system model amenable to analysis, optimization, and real-time control.

Challenge #5: Market participation for DERs

Enabling market participation for DERs is seen as critical for their wide proliferation in energy systems. Market prices provide both coordinating signals during operation and financial incentives for investing in new capacity. In 2020, the Federal Energy Regulatory Commission (FERC) issued Order 2222 which mandated that system operators design and implement market mechanisms for DERs to participate in wholesale markets [22, 141]. Despite the regulatory mandate, progress towards DER markets has been uneven due in part to the following barriers.

First, in order to join the wholesale market, DERs often need to meet technical requirements for communication, control, and redundancy initially intended for large power plants that are too onerous given the small scale of most DERs. Second, utilities mandate that DERs go through a lengthy and complex interconnection process to connect to the grid and participate in the electricity market. The interconnection process almost always requires a manual review by a utility engineer and may also require extensive power flow studies to verify that the distribution grid has sufficient capacity to accommodate the DER. Third, regulatory barriers currently limit the market participation of DERs by preventing them from accessing certain markets or by imposing additional requirements and costs. Part of this is due to the reluctance of system operators to experiment with new planning or operating processes that could impact the high levels of reliability expected from the power system. Another part is due to the balkanized nature of utilities. In there US there are roughly 3,000 individual utilities, each with their own set of regulations and interconnection processes that apply within their service territory [127]. This heterogeneity makes it difficult to implement market participation mechanisms within DERs that scale easily across service territories.

In the next section, we specify our technical contributions that address important aspects of five challenges highlighted above and present the structure of the thesis.

1.2 Contributions of this Thesis

The technical contributions in this thesis are two-part. In the first part, consisting of Chapters 2-4, we study market mechanisms for efficient and incentive-compatible power system dispatch under uncertainty. We propose electricity market mechanisms based on stochastic and robust optimization frameworks that enable system operators to feasibly incorporate uncertainty into the dispatch procedure as well as to integrate the impact of this uncertainty into the market clearing prices. Incentive compatibility properties of the proposed mechanisms are established.

In the second part, consisting of Chapters 5-8, we present a software platform for monitoring, modeling, and controlling DERs and microgrids. Three key functionalities are exhibited in this platform, which we call the Microgrid Operating System (OS): 1) data collection (as implemented in the Meter Caltech project); 2) modeling and monitoring of microgrids and DERs (as implemented in the Digital Twin project); 3) optimization of microgrids and DERs (as implemented in collaborations with utilities and industry partners).

We now summarize the particular topics covered in each chapter.

Chapter 3: Feasible multi-interval economic dispatch under uncertainty

As introduced in Chapter 1, a rapid shift in the composition of the generation mix in power markets is creating several challenges for system operators. First, increasing renewable penetration from solar and wind is injecting variability and uncertainty into available power supply. Second, there is a lack of suitable market mechanisms tailored to the physical characteristics of DERs (such as energy storage) which are seeking to join markets in increasing numbers. Third, electrification of vehicle charging and thermal (heating/cooling) loads is impacting the shape and variability of the demand profile, leading to periods of high, sustained ramping.

These factors have a common theme of uncertainty, and system operators have been rapidly innovating on new market structures and dispatch procedures to handle it. These include multi-interval lookahead dispatch [63], ramping reserves [44], operating reserves [43], capacity markets [36], and multi-stage or intraday markets. Alongside, researchers have been investigating techniques from stochastic optimization to efficiently dispatch the market under uncertainty, including robust optimization [134, 94], chance-constrained optimization [99], scenario optimization [93], and distributionally robust optimization [42].

In this chapter of the thesis, we merge the practice of multi-interval dispatch with

the theoretical market mechanisms derived from robust optimization. In particular, we characterize the conditions under which the inflexible, deterministic dispatch procedures used in practice lead to infeasibility of the dispatch. To address these shortcomings, we present a dispatch mechanism that leverages the feasibility guarantees from robust optimization. We prove its feasibility and give bounds on its performance by analyzing upper and lower bounds on the competitive ratio of the algorithm. Finally, we show that our mechanism retains the excellent performance of the heuristic lookahead algorithms used in practice while still providing feasibility guarantees over the scheduling horizon.

Chapter 3: Pricing uncertainty in stochastic multi-stage electricity markets

Whereas the work in Chapter 3 studies the primal dispatch procedure, in this chapter we turn our attention to the dual pricing problem. In this problem, the system operator seeks to produce a set of market clearing prices that supports its desired primal dispatch. We focus particularly on the impact of uncertainty on the market clearing process.

Our work in this area draws on two main lines of inquiry. The first is dispatching and pricing multi-interval markets in the presence of intertemporal coupling constraints. This has been studied in [57], [27], [28], [139], [68], [64], among others. The second is dispatching and pricing using techniques from robust and stochastic optimization, particularly in [134], [94], and [42].

In contrast to the existing approaches on risk-based electricity pricing, we propose a pricing mechanism for multi-stage electricity markets that does not explicitly depend on the choice of dispatch procedure or optimization method. Instead, we introduce a reformulation of the multi-stage market clearing problem that accommodates any description of uncertainty. Our approach is applicable to a wide range of methodologies for the economic dispatch of power systems under uncertainty and includes as special cases multi-interval dispatch, multi-settlement markets, scenario-based dispatch, and chance-constrained dispatch policies. We prove that our pricing scheme provides both *ex-ante* and *ex-post* dispatch-following incentives by simultaneously supporting per-stage and ex-post competitive equilibria.

Chapter 4: Pricing demand-side flexibility

In contrast to the previous two chapters, which emphasize incentives for the supply side of electricity markets, the focus of this chapter is incentives for loads to offer their flexibility into markets. Enabling participation of demand-side flexibility is

key to both improving power system resilience as well as promoting the continued deployment of zero-carbon renewable generation sources. As such, it has been the subject of extensive study by system operators and researchers, with particular attention paid to mechanisms for load-shedding demand response [122], [23], [37], and [70].

Our work addresses the scenario of curtailment of near-zero-marginal-cost renewable resources during periods of oversupply, a particularly important cause of inefficient generation dispatch. Focusing on shiftable load in a multi-interval day-ahead economic dispatch setting, we show that incompatible incentives arise for demand using the standard shadow pricing approach. While the system's overall efficiency increases from dispatching flexible demand, the welfare of the demand side of the market can often decrease as a result of higher spot prices. We propose a mechanism to address this incentive issue. Specifically, by imposing a small number of additional constraints on the economic dispatch problem, we guarantee individual rationality for all market participants while simultaneously obtaining a more efficient dispatch. Our formulation leads to a natural definition of a uniform, time-varying flexibility price that can be offered to loads by the market operator to incentivize flexible bidding.

Chapter 5: An operating system for microgrids

The traditional operating model for the power system involves a system operator scheduling a dispatchable fleet of generators to meet forecast net demand. In particular, industrial, commercial and residential loads that are served by local utility companies form a passive system that is not actively observed and controlled. Rather, the aggregate demand at substations is forecast and the transmission system operator ensures sufficient supply is available to match this forecast. A reliable, low-cost, and low-carbon energy system requires a paradigm shift that converts these loads and connecting grids into an active system capable of closed-loop real-time response to random and frequent fluctuations in renewable generations and electricity prices.

In this chapter, we present the design and deployment of a software platform that extends the core functions (e.g., monitoring, modeling, optimization) present in control systems for high-voltage transmission grids to low-voltage grids (e.g., distribution systems, microgrids, and DERs). The platform, which we call the Microgrid Operating System (OS), is comprised of three layers: 1) data, 2) models/learning,

and 3) optimization/control. Each of these layers depends on the one below it, as illustrated in Figure 1.4, but each layer has intrinsic value to system planners and operators on its own.

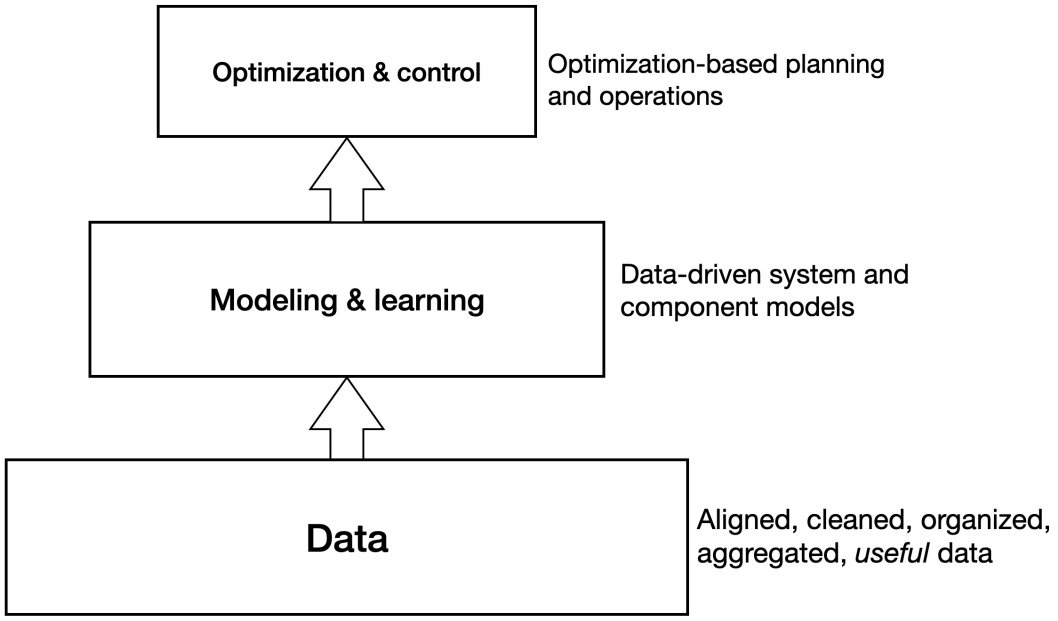


Figure 1.4: Layered architecture of the Microgrid OS. Reliable, organized, real-time data are the foundation for system models, which in turn are necessary for optimization-based control.

We extend the notion of layered thinking to the entire hardware-to-application stack for DERs and microgrids shown in Figure 1.5. The Microgrid OS is a software orchestration layer that links the hardware (e.g., batteries, PV, building loads, electric vehicle charging) with applications (e.g., resilience, peak shaving, carbon footprint reduction, energy market participation), allowing each layer to evolve independently from the others, as long as the interfaces between the layers are compatible. This paradigm is similar to those in the Internet and cloud computing that have facilitated rapid advances in cost, efficiency, and scale of these technologies.

We have developed and implemented several components of the Microgrid OS in the real-world testbed of the Caltech microgrid. The Caltech system is an ideal setting for understanding challenges in workplace decarbonization and for the testing of new technologies to overcome them. Despite Caltech's small size, its energy infrastructure is large and complex. The electricity, heating, and cooling needs correspond to those of a city of ~20,000 people, despite a campus population of ~5,000. The main load sources are ~120 buildings with anywhere from 30kW to

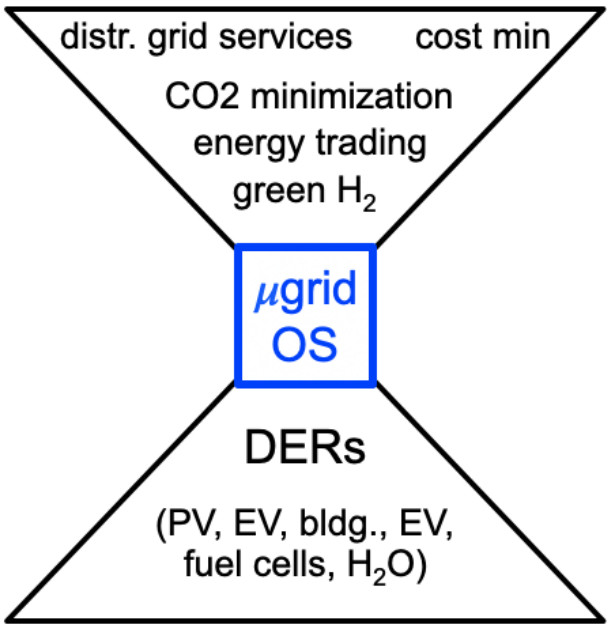


Figure 1.5: An hourglass architecture for DERs. The layered structure enables energy services to be defined without explicitly referencing the hardware that provides them. A software layer (Microgrid OS) is necessary to connect the hardware to the services.

500kW of load each. In recent years, Caltech produced nearly all of its ~100,000MW electricity consumption onsite (on an annual basis). The main generation resources on campus are a 12.5MW gas co-generation plant, 4MW of fuel cells (Bloom Energy), and 2MW of rooftop solar panels distributed across 15 arrays. When producing simultaneously, these assets can provide most of the 19MW peak campus demand. The generation and load sources are connected by a distribution network comprised on four substations, four interconnections with the Pasadena Water and Power (PWP) grid, and a radial network with voltage levels ranging from 208V up to 17kV. In addition to the electric distribution system, a district heating and cooling system supplies the campus buildings with chilled water and hot water (via steam and steam-to-water heat exchangers). Although the specific projects built and deployed in this thesis work were primarily developed within Caltech's system, the design philosophy and layered architecture of the Microgrid OS is transportable to any other distribution-level energy system, from single buildings up to utility distribution networks.

In the subsequent chapters of this thesis, we present particular projects at each of the main layers (data, models, optimization & control) of the Microgrid OS.

Chapter 6: Measuring and monitoring microgrids

Real-time, comprehensive, and reliable data is a prerequisite for effectively modeling and controlling systems of networked DERs. Despite its importance, a high-quality dataset is often a large gap in practical settings. Even where there is metering and monitoring infrastructure, as is the case with systems with building management systems (BMS) or SCADA systems, data is often not retained, organized, cleaned and made available for analysis. In the Meter Caltech component of the Microgrid OS project, we are systematically addressing the data gap by first deploying high-resolution smart meters on all building load and generation assets in the Caltech campus microgrid and then assembling the data collected by the meters into a comprehensive, granular database of electric power injections within the microgrid. This system provides a testbed for developing design, installation, and commissioning processes that can be scaled to other microgrids and networks.

The metering deployment is staged in three phases. Phase 0 consisted of the installation of 2 Egauge-brand smart meters with 15 current transformer sensors (CTs) each in the main load panels in the Annenberg Center building. The measurement quality and accuracy were validated and real-time data transmission pipeline between the meters and our Digital Twin monitoring platform was implemented. Phase 1 consisted of installing 20 additional meters in a self-contained subnetwork within the Caltech microgrid. These meters collect data from 13 building load switchboards, 2 PV arrays, 2 natural gas fuel cell arrays, and 2 interconnection points with the PWP grid. Phase 2, which is ongoing, extends the metering infrastructure to the entire Caltech campus.

The data collected from the meters consists of magnitudes and angles (with respect to an angle reference onboard each meter) of the three-phase bus voltages and line currents. These data are logged in a database every second and then used to compute derivative quantities such as per-phase power factor, real power, reactive power, and apparent power. In addition, we also collect 1-second-long voltage and current raw waveform snapshots every five seconds. Two types of measurements are derived from the waveforms. The first is the harmonic distortion of the current and voltage signals, which quantifies the deviation of the waveform from perfect sinusoids. This is an important metric for power quality which, when poor, can damage sensitive equipment. The second type is the phase offset of each waveform from each meter with respect to a common network reference signal. Using Network Time Protocol (NTP) and Precision Time Protocol (PTP) internet timing protocols, we

have been conducting experiments to validate the alignment of current and voltage phasors across meters in the network. Coherent synchrophasor data is typically only obtained in high-voltage transmission networks due to the expense of deploying phasor measurement units (PMUs). In contrast, our approach leverages inexpensive smart meters and IP timing protocols to replicate PMU functionality at a fraction of the standard cost. Synchrophasor data is necessary for state estimation algorithms which have been the subject of extensive study in distribution power systems. Our approach serves as a test case for providing synchrophasor data at relatively low cost, not just in microgrids but also in low-voltage distribution systems.

Chapter 7: Modeling microgrids

With the data collected from the Meter Caltech project as a foundation, we have designed, built, and deployed a software platform for modeling microgrids, called the Microgrid Digital Twin. The Digital Twin is a cyber-physical system consisting of multiple interconnecting sub-components: 1) a set of data collection interfaces to collect measurement data from smart meters and monitoring databases (e.g., BMS, weather data); 2) a processing pipeline to organize, clean, correct, and store ingested timeseries data; 3) a data integration module that fuses timeseries data with the database of physical assets contained in model of the network; 4) algorithms to learn statistical models of the data that can correct or augment the physical models; and 5) an interactive monitoring dashboard that integrates timeseries data and network models to make them useful to analysts, system planners, and operators.

Along with the software architecture design and implementation of the Digital Twin platform, the main technical contribution from this part of the thesis is a distribution system network modeling schema that can represent all physical components in the system and the connections between them. The schema incorporates information about each asset's labeling, physical properties, geographic location, hierarchical grouping, and connectivity. This schema builds upon standard tools for multi-phase distribution system modeling, such as OpenDSS, Cyme, and Gridlab-D, to include support for a wider range of components such as standard electrical distribution components (e.g., lines, transformers, buses, switches), protection devices (e.g., breakers, fuses), control devices (e.g., relays, variable frequency drives), meters, thermal assets (e.g., boilers, chillers, heat exchangers), and power generation sources(e.g., batteries, fuel cells, solar cells, inverters, emergency generators, and combustion turbines). This allows the Digital Twin to model and facilitate optimization of a broad range of energy systems at multiple scales, from individual

electrical load panels up to entire utility distribution networks. It also addresses a gap in the existing modeling tools for energy systems at the scale of microgrids that span component-level models (e.g., building information models) and system-level distribution network models (e.g., Cyme).

Chapter 8: Optimizing microgrids

Building on the data and modeling efforts described in the previous two chapters, this section of the thesis presents several projects on the optimal design and control of microgrids and distribution systems.

The first is a collaboration with Pasadena Water & Power (PWP) distribution utility where the placement and control of a fleet of energy storage devices was optimized in a distribution feeder correct over- and under-voltage conditions arising summertime peak load conditions. The principal challenges in this project were two-fold: 1) solving a non-convex optimal battery location problem; and 2) modeling power flow equations in a highly-unbalanced 4-wire distribution feeder with transformers. We showed that a relatively small 75kW/420kWh of energy storage, optimally distributed and controlled throughout a feeder with 2MW of peak load, could correct voltages at all ~800 load nodes to within ±5% of the nominal reference value. Our modeling validated the effectiveness of inexpensive smart control methods to address power quality issues, in contrast to the standard approach of expensive and disruptive transformer and conductor upgrades that utilities typically use in practice.

The second, ongoing project develops methods for optimal sizing of microgrids subject to physical, reliability, and financial constraints. Microgrids can serve many purposes depending on how they are designed and operated. Some microgrid operators, such as university campuses care primarily about reliability due to the high cost of power outages to operations. Other, emerging applications of microgrids take advantage of nascent opportunities to participate in wholesale electricity markets for energy and ancillary services. For both of these instances, we present a technique for optimizing the capital investment that operators must make in microgrid hardware assets (e.g., solar systems, batteries, generators, switchgear, and transformers) that incorporates physical constraints, operational constraints, financing cost, and ongoing cost/revenue streams.

Part II

Dispatching and pricing under uncertainty in electricity markets

A rapid shift in the composition of the generation mix in power markets is creating several challenges for system operators. First, increasing renewable penetration from solar and wind is injecting variability and uncertainty into available power supply. Second, there is a lack of suitable market mechanisms tailored to the physical characteristics of distributed energy resources (such as energy storage) which are seeking to join markets in increasing numbers. Third, electrification of vehicle charging and thermal (heating/cooling) loads is impacting the shape and variability of the demand profile, leading to periods of high, sustained ramping.

These factors have a common theme of uncertainty, and SOs have been rapidly innovating new dispatch procedures and market structures to handle it. These include multi-interval lookahead dispatch [63], ramping reserves [44], operating reserves [43], capacity markets [36], and multi-stage or intraday markets. Alongside, researchers have been investigating techniques from stochastic optimization to efficiently dispatch the market under uncertainty, including robust optimization [134, 94], chance-constrained optimization [99], scenario optimization [93], and distributionally robust optimization [42].

Uncertainty impacts the stability of pricing signals and can lead to market distortions such as out-of-merit dispatch, ramping shortages, and load shedding. Even with more advanced and accurate forecasts, SOs must still dispatch the system in a way that anticipates forecast uncertainty and the possibility of distribution shift over time. Pricing that incorporates characterization of uncertainty is necessary to fairly and efficiently compensate different resources for their contributions to a reliable power supply as well as to drive efficient investment in an optimal generation mix.

Some of the principal mechanisms SOs use in practice to manage uncertainty are multi-settlement dispatch and multi-interval scheduling. Multi-settlement dispatch makes use of a sequence of forward markets that culminate in a real-time when electricity is actually delivered. Examples of these forward markets are day-ahead markets, capacity markets, and intraday unit commitment markets. The common feature of these forward markets is that they only involve a financial settlement and have no immediate physical consequences on generator dispatch.[1] Multi-interval scheduling is a mechanism used in real-time when generator energy production must be physically dispatched. It involves optimizing the systems over a rolling lookahead horizon with predictions in order to more efficiently account for binding

[1]The unit commitment statuses of generators is decided in forward markets but the energy production is not determined until real-time.

intertemporal ramping and unit commitment constraints. Only the first interval of the horizon is physically and financially binding.

Multi-settlement systems and multi-interval dispatch interact in practice in various ways. Figure 1.6 illustrates a few of these interactions. The practice in North American ISOs is to have a single day-ahead forward market followed by successive rounds of multi-interval dispatch (e.g., every 5 or 15 minutes) that unfold in rolling fashion. In principle, an arbitrary number of forward market settlements is achievable and some markets such as the Electric Reliability Council of Texas (ERCOT) are exploring implementation of intraday forward markets that sit between the day-ahead market and real-time.

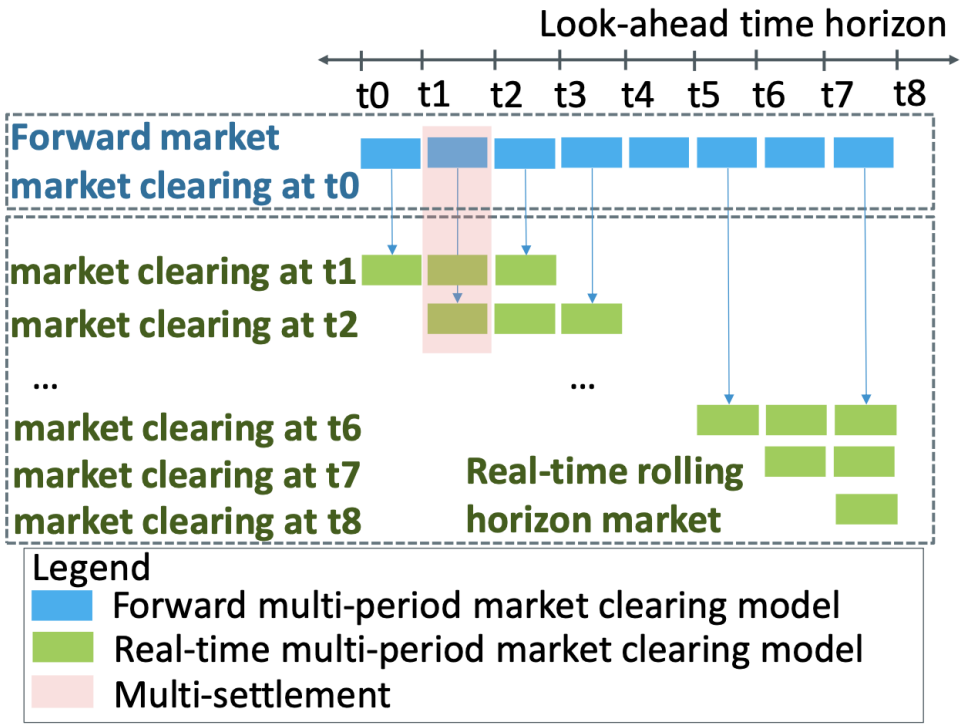


Figure 1.6: Schematic of multi-settlement, multi-interval economic dispatch [139]. The forward stage is a day-ahead financial market and the real-time stages are successive rounds of lookahead economic dispatch.

Multi-stage dispatch mechanisms are the focus in this part of the thesis. We study three particular aspects of them. The first (Chapter 2) is a method to retain feasibility of the multi-stage problem over the entire scheduling horizon under robust description of uncertainty. We establish the necessity of robust dispatch by proving fundamental limits on the feasibility of standard multi-interval dispatch even with perfect lookahead predictions. These limitations are addressed with a proposed

mechanism that combines robust policy-based dispatch with prediction-based multi-interval dispatch. We prove that the mechanism admits a feasibility guarantee to robust disturbances while retaining the excellent average-case guarantee of standard multi-interval dispatch.

The second (Chapter 3) is the derivation of market-clearing prices when the system operator is scheduling under uncertainty. The novelty of our approach is the formation of a price does not depend on the particular characterization of uncertainty. The requirement is simply that the dispatch optimization problem is tractable and produces and optimal dispatch and that its Jacobian with respect to the parameters for the optimal dispatch can be computed.

Finally, in Chapter 4, we study demand-side mechanisms to dispatch flexible load in the day-ahead multi-interval market. We show how the standard shadow price is not suitable for incentivizing load-side flexibility as the demand-side welfare can decrease when flexible loads are dispatched. We propose a mechanism to address this deficiency and prove that it provides rationality and incentive compatibility to flexible loads to bid into the market while preserving the incentive compatibility for generators.

Chapter 2

FEASIBLE MULTI-INTERVAL ECONOMIC DISPATCH UNDER UNCERTAINTY

In power systems with high penetrations of variable renewable energy production, sufficient flexible and dispatchable generation resources are necessary to ensure a stable energy supply. However, conventional dispatchable thermal generators are ramp-constrained, limiting how quickly they can modulate their production to accommodate large fluctuations in net demand. This poses a challenge for system operators on two fronts: *resource procurement* and *real-time generation scheduling*.

Resource procurement refers to the system operator's task of planning for sufficient available capacity and ramp for the system to meet uncertain net demand. Resource procurement takes place on longer timescales (e.g., years to day-ahead) and includes several problems familiar to power system operators including security-constrained unit commitment (SCUC), resource adequacy, and capacity planning. On shorter timescales (e.g., 5 to 15 mins), system operators must dispatch available generation resources efficiently to meet realized net demand. This is known as real-time economic dispatch (RTED).

Numerous methods have been devised in both of these domains to ensure robustness to uncertainty in net demand. For resource procurement problems, scenario-based optimization is common in practice, while other stochastic optimization techniques and robust optimization have been explored in the research community. For RTED, lookahead dispatch algorithms have been widely implemented by independent system operators (ISOs) in energy markets, and additional ancillary services such as flexible ramping products and load-following reserves have seen some adoption in markets with high demand variability.

The ultimate goal of both resource procurement and RTED is to deliver sufficient generation to meet realized demand while satisfying system constraints: that is, to guarantee feasibility of the dispatch in real time. However, a crucial challenge facing state-of-the-art methods today is that if resource procurement fails to account for the particular dispatch algorithm to be used, or if the RTED algorithm used does not appropriately consider procured resources (e.g., generation & ramp capacity) when

making decisions, then feasibility is not assured.[1] Moreover, merging the problems of resource procurement and feasible RTED algorithm synthesis, i.e., optimizing over both system specifications and RTED algorithms, is intractable for the class of general dispatch algorithms.

This motivates the goal of this work: *developing tractable methods for resource procurement and RTED that together yield provable guarantees of feasibility*.

Contributions

First, to motivate the need for a joint approach to resource procurement and RTED, we show in Section 2.2 that even on a single-bus system and with nearly full knowledge of future demand, offline feasibility over a set of demand trajectories is insufficient to guarantee the existence of *any* online dispatch algorithm that can feasibly meet those demand sequences.

Second, a practical joint algorithm for resource procurement and RTED is presented in Section 2.3. The first step is a robust optimization problem called *Dispatch-Aware Planning* (DAP), which determines adequate system capacity to ensure a feasible RTED algorithm exists. The second step is a dispatch algorithm called *Feasible Fixed Horizon Control (*FFHC*)* that minimally modifies the standard receding horizon control (RHC) algorithm to robustly use trusted predictions of demand.

Third, in Section 2.4 we give matching upper and lower bounds on the competitive ratio of *any* feasible online dispatch algorithm. These bounds imply identical bounds on FFHC.

Finally, we evaluate the proposed approach on a synthetic system derived from CAISO demand and generation data. We show that FFHC retains the excellent average-case performance of RHC but in cases where there are large demand fluctuations and the system is ramp constrained, RHC fails while FFHC remains feasible with minimal additional cost. The test cases presented in this work are designed to clearly demonstrate the feasibility properties of our approach. Although our algorithm is applicable to realistic problems like SCUC, large-scale simulations are not explored here and are the subject of future work.

[1] In systems with no inter-temporal coupling constraints (e.g., ramp or state-of-charge constraints), this feasibility mismatch does not arise. However, in practice ramp limits matter, i.e., they constrain the set of operating points reachable by the system.

Related work

This work bridges the online algorithms and power systems literatures. We briefly highlight some related work in each of these domains.

Online algorithms. RTED is an online decision-making problem characterized by a challenging combination of time-coupling and unknown time-varying constraints. None of the existing constrained online optimization literature, e.g., [9, 80, 79, 121], directly handles our setting.

The authors in [9] explore a related ramp-constrained online optimization problem, yet feasibility does not pose an issue due to the lack of unknown time-varying constraints. Recent work in online optimal control considers time-invariant [80] as well as time-varying and coupling constraints [79] on state and action. However, the feasibility guarantees depend on advance knowledge of the constraints.

The work in [121] comes closest to our setting. The authors optimize over affine policies to design algorithms for online optimization with switching costs and ramp limits that are robust to polytopic uncertainty in certain constraints. However, their approach does not consider the problem of guaranteeing feasibility, and their proposed algorithm is unable to fully exploit good predictions of near-term uncertainty.

Power systems. Resource procurement for system reliability and multi-interval economic dispatch are two key problems in power systems operation addressed by this work.

Our formulation of the resource procurement problem has broad applicability to several problems in power system reliability: in particular security-constrained unit commitment (SCUC) for day-ahead markets [15, 106, 136, 13, 133], resource adequacy [19, 117, 108], and capacity planning [138, 36, 38]. Most of the approaches in this literature do not consider behavior of causal RTED algorithms or generally involve scenario-based optimization [41, 10]. In practice, resource adequacy planning relies on regulatory standards and scenario-based studies and similarly ignore the behavior of the RTED algorithm.

Many ISOs have implemented multi-interval lookahead optimization for RTED, as it can better accommodate variability in forecasts for renewables and intertemporal constraints from conventional generation and storage [59, 137, 110, 44] Ancillary services such as flexible ramping products [131, 133, 54] and load-following reserves [43, 107, 97] have been studied and implemented in some markets. To our knowledge, all of the aforementioned proposals for multi-interval dispatch do not

provide provable guarantees for the feasibility of the lookahead optimization problem. There is prior work on utilizing affine policies to robustly dispatch reserves in the real-time market when ramp constraints are present [134, 135], but this work similarly does not explicitly consider the question of feasibility, and the affine policies utilized may be more conservative than the lookahead dispatch algorithms used by operators.

Our work is most closely related to research on adaptive robust unit commitment with causal affine real-time policies in [84, 85]. Here, robust policy-aware economic dispatch is combined with robust unit commitment, and an algorithmic framework for efficient computation of large-scale problems is proposed. Like [121], robustness comes at the expense of fully utilizing predictions. In contrast, we bring an online algorithms perspective to the problem of feasible RTED, focusing primarily on (a) designing feasible RTED algorithms that can fruitfully exploit predictions, and (b) characterizing the performance of feasible RTED algorithms in general.

Notation

$N \in \mathbb{Z}_+$ is the number of dispatchable generators and $T \in \mathbb{Z}_+$ is the length of the time horizon. We denote the ordered set of time intervals between indices a and b by $[a, b] \coloneqq \{a, \ldots, b\} \subset \mathbb{Z}_+$. The inequalities in (2.1) and subsequent optimization problems are element-wise.

2.1 Problem formulation

The problem of optimal power system planning and operation can be cast as a sequential optimization problem robust to uncertainty revealed prior to each stage.[2]

$$\min_{\mathbf{y}} \max_{d_1} \min_{\mathbf{x}_1} \cdots \max_{d_T} \min_{\mathbf{x}_T} \quad \bar{\mathbf{c}}^\top \mathbf{y} + \sum_{t=1}^{T} \mathbf{c}_t^\top \mathbf{x}_t \tag{2.1a}$$

$$\text{s.t.} \quad \mathbf{1}^\top \mathbf{x}_t = d_t \qquad \forall t \in [1, T] \tag{2.1b}$$

$$g_t(\mathbf{x}_t, \mathbf{y}) \leq \mathbf{0} \qquad \forall t \in [1, T] \tag{2.1c}$$

$$h_t(\mathbf{x}_{t-1}, \mathbf{x}_t) \leq \mathbf{0} \qquad \forall t \in [1, T] \tag{2.1d}$$

$$(d_1, \ldots, d_T) \in \mathcal{D} \tag{2.1e}$$

For concreteness, we limit our presentation to a single planning stage with decision variables $\mathbf{y} \in \mathbb{R}^K$ (e.g., generator capacities, ramp/line limits, unit com-

[2] For simplicity, in this work we assume that the only uncertainty is the demand, although uncertainty in generation (e.g., solar, wind) can also be accommodated.

mitments) followed by T generation dispatch stages, each with decision variables $\mathbf{x}_t \in \mathbb{R}^N, t = 1, \ldots, T$, where the initial operating point $\mathbf{x}_0$ is fixed. We assume the cost functions for planning variables $\bar{\mathbf{c}}$ and dispatches $\mathbf{c}_t$ are linear and known by the system operator. Constraint (2.1b) is the supply-demand balance constraint where d_t is the demand at time t. Constraints (2.1c) and (2.1d) are affine and represent capacity/planning constraints and intertemporal (ramp, state-of-charge) constraints respectively. We focus on a single-bus network with all dispatchable generators satisfying a net load trajectory $\mathbf{d} = (d_1, \ldots, d_T)$, which is contained in a bounded, known uncertainty set $\mathcal{D}$.[3] We assume that $\mathcal{D}$ is polytopic, i.e. it takes the form $\mathcal{D} = \{\mathbf{d} \in \mathbb{R}^T : \mathbf{E}\mathbf{d} \leq \mathbf{f}\}$ with parameters $\mathbf{E} \in \mathbb{R}^{L \times T}$ and $\mathbf{f} \in \mathbb{R}^L$ known to the system operator prior to solving the planning and dispatch problem.

Each dispatch stage depends on the planning decision as well as the previous dispatch. An example of a problem falling under this framework is SCUC followed by multi-interval real-time dispatch. However, this framework can be extended to include several planning stages in advance of dispatch, such as capacity planning and intraday unit commitment.

Planning problem

The goal of the planning problem is to determine a choice $\mathbf{y}^*$ of the planning decisions. Given that problem (2.1) is intractable due to the sequential min – max – min operators, an approach taken by power system operators in practice is to choose $\mathbf{y}^*$ by solving an "offline" form of the problem where the demand sequence $\mathbf{d} = (d_1, \ldots, d_T) \in \mathcal{D}$ (or a small number of scenarios) is assumed known in advance.

$$\min_{\substack{\mathbf{y} \in \mathbb{R}^K \\ \mathbf{x}_t \in \mathbb{R}^N}} \quad \bar{\mathbf{c}}^\top \mathbf{y} + \sum_{t=1}^{T} \mathbf{c}_t^\top \mathbf{x}_t \tag{2.2a}$$

$$\text{s.t.} \quad \mathbf{1}^\top \mathbf{x}_t = d_t \qquad \forall t \in [1, T] \tag{2.2b}$$

$$g_t(\mathbf{x}_t, \mathbf{y}) \leq \mathbf{0} \qquad \forall t \in [1, T] \tag{2.2c}$$

$$h_t(\mathbf{x}_{t-1}, \mathbf{x}_t) \leq \mathbf{0} \qquad \forall t \in [1, T] \tag{2.2d}$$

As written, (2.2) is a linear program; when $\mathbf{y}$ represents unit commitments, (2.2) becomes a MILP with the addition of integrality constraints on $\mathbf{y}$ (not shown).

While the resulting $\mathbf{y}^*$ from this offline optimization would be ex-post optimal, were the assumed demand sequence the true demand, this will not generally be

[3] Everything that follows can be extended to multi-bus setting with network constraints, as in [84, 85].

the case, as the planning problem is typically solved far in advance, when there is still uncertainty in future demand. To provide stronger guarantees in the face of demand uncertainty, system operators may wish for $\mathbf{y}^*$ to satisfy (2.2b) - (2.2d) *for any* $\mathbf{d} \in \mathcal{D}$. This motivates the following definition of offline feasibility. A planning decision $\mathbf{y}^*$ is *offline feasible* if and only if for all $\mathbf{d} \in \mathcal{D}$ there exists a dispatch sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$ satisfying the dispatch feasibility constraints (2.1b) - (2.1d).

Online dispatch problem

After the planning variables $\mathbf{y}^*$ are chosen, the task of the system operator is to determine real-time dispatches $\mathbf{x}_t$. They do so via an *online dispatch algorithm*: a sequence of functions $X_1, \ldots, X_T$, each of which maps a demand sequence to a dispatch for time t: $X_t : \mathcal{D} \to \mathbb{R}^N$. Crucially, the collection of functions $\{X_t\}_{t=1}^T$ must be *causal*, so the decision $X_t(\mathbf{d})$ at time t can only depend on information known to the system operator at time t. We will assume that the system operator knows the exact demand d_t at time t, and also has access to perfect predictions of demand $d_{t+1}, \ldots, d_{t+h}$ within a short lookahead window of length h. Thus $X_t(\mathbf{d})$ may only depend on demands through time $\min\{t + h, T\}$.

A desirable objective for an online dispatch algorithm is the satisfaction of dispatch feasibility constraints. This motivates the following definition of *online feasibility* of a dispatch algorithm as well as of a planning decision $\mathbf{y}^*$.

1. Given a fixed planning decision $\mathbf{y}^*$, a *feasible* online dispatch algorithm is a sequence of causal policies $\{X_t\}_{t=1}^T$ with the property that for any demand sequence $\mathbf{d} \in \mathcal{D}$, the produced decisions $X_1(\mathbf{d}), \ldots, X_T(\mathbf{d})$ satisfy the constraints (2.1b) - (2.1d).

2. If $\mathbf{y}^*$ admits a feasible online dispatch algorithm, then $\mathbf{y}^*$ is said to be an *online feasible* planning decision.

A particular online dispatch algorithm that is widely used in practice is Receding Horizon Control (RHC), where at time t, dispatches are optimized over the h-step perfect lookahead horizon $[t, t + h]$. Only the first dispatch $\mathbf{x}_t^*$ is committed at each step of RHC; the remaining dispatches over the lookahead horizon are merely "advisory". This process is then repeated for the subsequent interval $t + 1$, and so on. However, RHC has a significant downside in that it is not necessarily feasible,

even if the planning decision $\mathbf{y}^*$ is online feasible. We demonstrate situations when RHC lacks feasibility in Section 2.5.

2.2 Offline feasibility does not imply online feasibility

Although we presented the planning and dispatch problems separately in the previous section, we show now why the common practice of solving for the planning variables $\mathbf{y}^*$ offline in a dispatch-unaware fashion can ultimately cause online dispatch infeasibility. We answer the following question: *For a particular demand uncertainty set, does offline feasibility of planning decisions* $\mathbf{y}^*$ *necessarily imply their online feasibility?*

We answer this question in the negative in the following theorem. This establishes that anything short of full knowledge of the demand sequence is insufficient for offline feasibility to imply online feasibility.

Theorem 1. *There exist choices of affine system constraints* $\{g_t\}$, $\{h_t\}$, *a polytopic demand uncertainty set* $\mathcal{D}$, *and fixed planning decisions* $\mathbf{y}^*$ *that are offline feasible, yet which are not online feasible if* $h < T - 1$.

In other words, this theorem states that offline feasibility does not imply online feasibility, even for online algorithms with arbitrary (but not full) knowledge of future demand. This fact necessitates performing resource procurement with explicit consideration of online feasibility.

Proof of Theorem 1. We construct an example system and an offline feasible demand uncertainty set $\mathcal{D}$ admitting no feasible online algorithm. Consider a 2-generator system with $\underline{\mathbf{x}} = (0, 0)$, $\overline{\mathbf{x}} = (2h + 4, 2h + 4)$ and $\overline{\Delta}_x = (2h + 4, 1)$, $\underline{\Delta}_x = (-2h - 4, -1)$, and initial operating point $\mathbf{x}_0 = (h + 2, h + 2)$, where $h \geq 0$. In a single time interval, the first generator can ramp to any operating point within its capacity bounds, while the second generator can only ramp up or down by 1 unit. Consider the demand uncertainty set

$$\mathcal{D} := \{\mathbf{d} \in \mathbb{R}^{h+3} : d_0, \ldots, d_{h+1} = 2h + 4, 0 \leq d_{h+2} \leq 4h + 8\}$$

and distinguish two demand sequences in this set:

$$\mathbf{d}^{(A)} = (2h + 4, \ldots, 2h + 4, 4h + 8)$$
$$\mathbf{d}^{(B)} = (2h + 4, \ldots, 2h + 4, 0)$$

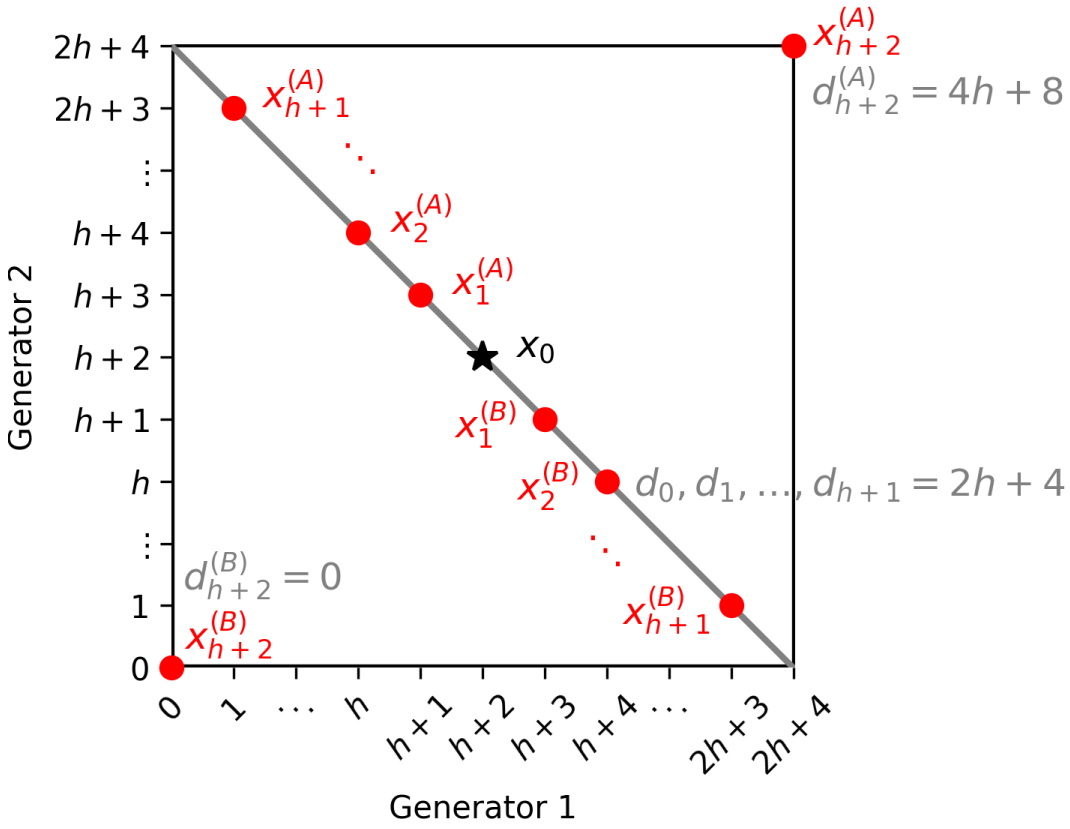


Figure 2.1: Illustration of the setting used in Theorem 1.

We illustrate the system, and the two demand sequences $\mathbf{d}^{(A)}$ and $\mathbf{d}^{(B)}$, in Figure 2.1: the outer box contains all of the operating points the system can realize while respecting capacity constraints and diagonal grey lines correspond to demands d_t, since an operating point along such a diagonal line corresponds to an aggregate generation of d_t.

Note that both $\mathbf{d}^{(A)}$ and $\mathbf{d}^{(B)}$ are offline feasible: for case (A), $\mathbf{x}^{(A)}$ as in Figure 2.1 with $x^{(A)}_t = (h+2-t, h+2+t)$ for $t \in [1, h+1]$ and $x^{(A)}_{h+2} = (2h+4, 2h+4)$ is feasible. For case (B), $x^{(B)}_t$ as in Figure 2.1 with $x^{(B)}_t = (h+2+t, h+2-t)$ for $t \in [1, h+1]$ and $x^{(B)}_{h+2} = (0,0)$ is feasible. Furthermore, it can be seen from the ramp constraints that $\mathbf{x}^{(A)}$ is the unique feasible dispatch meeting the demand sequence $\mathbf{d}^{(A)}$, and similarly $\mathbf{x}^{(B)}$ is the unique feasible dispatch meeting demand sequence $\mathbf{d}^{(B)}$. Finally note that any $\mathbf{d} \in \mathcal{D}$ is a convex combination of $\mathbf{d}^{(A)}$ and $\mathbf{d}^{(B)}$. By convexity, thus, $\mathcal{D}$ is offline feasible.

Now consider any online dispatch algorithm with lookahead h. At time $t = 1$, it must choose x_1 without seeing d_{h+2}; thus when choosing x_1 it does not know, for example, whether the full demand sequence is $\mathbf{d}^{(A)}$ or $\mathbf{d}^{(B)}$. Clearly if the algorithm

does not select $x_1 = x_1^{(A)}$, it will not be feasible if $\mathbf{d}^{(A)}$ is the full demand trajectory; likewise, if the algorithm does not select $x_1 = x_1^{(B)}$, it will not be feasible if $\mathbf{d}^{(B)}$ is the full demand trajectory. As $x_1^{(A)} \neq x_1^{(B)}$, the online algorithm cannot choose a decision at time 1 that allows it to remain feasible for both cases (A) and (B). Thus the algorithm is not feasible for $\mathcal{D}$. □

Remark 1. *The particular construction in the proof – namely, the dependence of generator capacities on h – is not crucial for the result. Other systems and offline feasible demand uncertainty sets can be constructed wherein both generator capacities and ramps are independent of h, yet for which no online dispatch algorithm with lookahead $h < T - 3$ is feasible.*

2.3 Joint algorithm for system planning and online dispatch

The counterexample in Theorem 1 establishes that the offline feasibility of a planning decision $\mathbf{y}^*$ for a particular demand uncertainty set does not imply its online feasibility. This motivates the joint approach in Algorithm 1, where we use affine policies to guarantee the existence of an online feasible $\mathbf{y}^*$ as well as an online feasible dispatch algorithm. Affine policies approximate online decision making during the planning stage (called *Dispatch-aware Planning (*DAP*)*), before passing the optimal planning variables to the RTED algorithm (called *Feasible Fixed Horizon Control (*FFHC*)*). Rather than using the (conservative) affine policies for determining actual dispatch schedules in real time, we subtly modify the standard RHC dispatch algorithm to include an affine-policy-based regularization term on the last decision of each subhorizon. This allows for online scheduling to exploit accurate short-term predictions without taking decisions that are too myopic. Other variants of fixed horizon control, like AFHC [82] or CHC [29], can be substituted for RHC in our algorithm at the expense of more burdensome notation.

Algorithm 1 Joint algorithm for planning and dispatch

1: **input:** Cost functions $\bar{\mathbf{c}}$, $\mathbf{c}_t$ and constraints g_t, h_t
2: Solve DAP problem (2.4)
3: Fix optimal planning variables $\mathbf{y}^*$
4: **for** $t = 1, \ldots, T$ **do**
5: $\quad$ d_{t+h} revealed
6: $\quad$ Solve FFHC problem (2.5) with $d_1, \ldots, d_{t+h}$ and $\mathbf{x}^*_{t-1}$ as parameters
7: $\quad$ **return** $\mathbf{x}^*_t$
8: **end for**

Dispatch-Aware Planning (DAP)

The dispatch-aware planning problem is defined in (2.4) below. The demand sequence $\mathbf{d} = (d_1, \ldots, d_T)$ resides in a known polytopic demand uncertainty set $\mathcal{D}$, and the linear planning/dispatch cost functions are known as well. Piecewise-linear cost functions can also be accommodated with additional notation.

The real-time scheduling policies $\{X_t(\cdot)\}_{t=1}^T$ are defined to be affine in the demand trajectory:

$$X_t(\mathbf{d}) := \mathbf{A}_t\mathbf{d} + \mathbf{b}_t \quad \forall t \in [1, T] \tag{2.3}$$

To optimize over the X_t is to optimize over the matrices $_t \in \mathbb{R}^{N\times T}$ and vectors $_t \in \mathbb{R}^N$. The X_t are causal, meaning $\mathbf{A}_t$ have 0's for all columns with index greater than t. This requirement can be enforced with entrywise constraints on the matrices. It is assumed that (2.4) has a feasible solution.

$$\min_{\substack{\mathbf{y} \\ X_1,\ldots,X_T}} \max_{\mathbf{d}\in\mathcal{D}} \quad \bar{\mathbf{c}}^\top\mathbf{y} + \sum_{t=1}^{T} \mathbf{c}_t^\top X_t(\mathbf{d}) \tag{2.4a}$$

$$\text{s.t.} \quad \mathbf{1}^\top X_t(\mathbf{d}) = d_t \qquad \forall t \in [1, T] \tag{2.4b}$$

$$g_t(X_t(\mathbf{d}), \mathbf{y}) \leq \mathbf{0} \qquad \forall t \in [1, T] \tag{2.4c}$$

$$h_t(X_{t-1}(\mathbf{d}), X_t(\mathbf{d})) \leq \mathbf{0} \quad \forall t \in [1, T] \tag{2.4d}$$

As convention we assume $X_0(\mathbf{d}) = \mathbf{x}_0$. Problem (2.4) is a linear program with semi-infinite constraints resulting from the "$\forall$" qualification on $\mathbf{d}$. Using strong duality of linear programs, (2.4) can be equivalently posed as a linear program with a finite number of additional variables and constraints [11]. The result is a tractable linear program that can be solved with off-the-shelf optimization solvers. However, depending on the length of the time horizon and the complexity of $\mathcal{D}$, (2.4) can be challenging to scale to large problem sizes. Although this scaling is *not* the focus of this chapter, strategies for improving scaling are discussed in Section 2.5.

After solving (2.4), the optimal planning variables $\mathbf{y}^*$ are fixed. It is then possible to schedule in real time using the optimal affine policies $X_1^*, \ldots, X_T^*$ applied to the real-time demand sequence. The resulting generation schedules are guaranteed to be feasible—that is, they satisfy (2.1b)-(2.1d) for any $\mathbf{d} \in \mathcal{D}$—but because of their robustness and inability to incorporate more refined demand predictions, the cost of the dispatch is likely to be quite conservative. In contrast, the algorithm we propose

next uses the policies to constrain online dispatches to an always-feasible region while still allowing accurate short-term predictions to be exploited.

Feasible Fixed Horizon Control (FFHC)

Economic dispatch in real-time (e.g., 5-min, 15-min) electricity markets is often a multi-interval optimization problem over an $(h+1)$-step horizon from which only the first dispatch decision is binding and the remaining are advisory. In the control literature this algorithm is referred to as receding horizon control (RHC) or model predictive control (MPC).

The version of fixed horizon control that we propose here, called *Feasible Fixed Horizon Control* (FFHC), is RHC with the addition of a robust affine constraint composed from the optimal policies from (2.4). FFHC is parameterized by the optimal solutions $\mathbf{y}^*, X_1^*, \ldots, X_T^*$ from (2.4).[4] FFHC has access to d_t and h perfect predictions of future demand $d_{t+1}, \ldots, d_{t+h}$, as well as the previously committed dispatch $\mathbf{x}_{t-1}^*$. The first decision $\mathbf{x}_t^*$ in the subhorizon of the following optimization problem determines the decision of FFHC in time t, denoted henceforth by FFHC(t).[5]

$$\begin{aligned}
\underset{\mathbf{x}_t,\ldots,\mathbf{x}_{t+h}}{\arg\min} \quad & \sum_{s=t}^{t+h} c_s^\top \mathbf{x}_s && && \text{(2.5a)}\\
\text{s.t.} \quad & \mathbf{1}^\top \mathbf{x}_s = d_s && \forall s \in [t, t{+}h] && \text{(2.5b)}\\
& g_s(\mathbf{x}_s, \mathbf{y}^*) \le \mathbf{0} && \forall s \in [t, t+h] && \text{(2.5c)}\\
& h_s(\mathbf{x}_s, \mathbf{x}_{s-1}, \mathbf{y}^*) \le \mathbf{0} && \forall s \in [t+1, t+h] && \text{(2.5d)}\\
& h_t(\mathbf{x}_t, \mathbf{x}_{t-1}^*, \mathbf{y}^*) \le \mathbf{0} && && \text{(2.5e)}\\
& h_{t+h+1}(X_{t+h+1}^*(\mathbf{d}), \mathbf{x}_{t+h}, \mathbf{y}^*) \le \mathbf{0} && \forall \mathbf{d} \in \mathcal{D}_{0:t+h} && \text{(2.5f)}
\end{aligned}$$

$\mathcal{D}_{0:t+h}$ is the restricted set of demand sequences in $\mathcal{D}$ that are possible given the already-revealed demand values from time 0 to $t+h$. In general for a pair of indices $r, s \in [0, T]$ and $r \le s$ and $\tilde{\mathbf{d}}_{r:s} = (\tilde{d}_r, \ldots, \tilde{d}_s)$ a subsequence of realized values, we define

$$\mathcal{D}_{r:s} := \{\mathbf{d} \in \mathcal{D} \mid d_t = \tilde{d}_t \ \forall t = r, \ldots, s\}.^6 \tag{2.6}$$

Clearly, $\mathcal{D}_{r:s} \subseteq \mathcal{D} \subseteq \mathbb{R}^{T+1}$.

[4] X_t^* refers to $(\mathbf{A}_t^*, \mathbf{b}_t^*)$.

[5] At $t = T - h$, the optimal solution of (2.5) determines FFHC's remaining dispatch decisions $\mathbf{x}_{T-h}^*, \ldots, \mathbf{x}_T^*$ because by that time, the entire demand sequence is known.

[6] The explicit dependence of $\mathcal{D}_{r:s}$ on $\tilde{\mathbf{d}}_{r:s}$ is suppressed in the notation for simplicity.

As was explored in Theorem 1, only enforcing constraints (2.5b) - (2.5e) does not always yield a feasible solution. Because the ramping constraint ties the previously committed decision $\mathbf{x}^*_{t-1}$ to all subsequent dispatches, a short-sighted dispatch early on could lead to infeasibility for a subsequent round. The addition of robust constraints (2.5f) on the last decision $\mathbf{x}_{t+h}$ ensures that earlier decisions are robust to future uncertainty. As in (2.4), the robust constraint in (2.5) can be transformed into auxiliary linear constraints on $\mathbf{x}_{t+h}$. Taking the optimal solutions of the first variables $\mathbf{x}^*_t$ from each subhorizon for $t = 1, \ldots, T$ gives the dispatch sequence from FFHC, which, as presented in the following theorem, is feasible.

Theorem 2. FFHC *is a feasible online dispatch algorithm. That is, for any* $\mathbf{d} \in \mathcal{D}$*,* $T - h$ *successive rounds of* FFHC(t) *produce a dispatch sequence* $\mathbf{x}^*_1, \ldots, \mathbf{x}^*_T$ *that satisfies* (2.1b)-(2.1d)*.*

Proof. The main idea of the proof is to match constraints between the offline robust capacity procurement problem and the online feasible dispatch problem.

Pick any $\mathbf{d} = (d_0, \ldots, d_T) \in \mathcal{D}$. We will inductively construct feasible solutions for each round of FFHC to satisfy demand $\mathbf{d}$ as its elements are revealed sequentially. We take the starting point $\mathbf{x}_0$ to be the one given by the optimal solution from the DAP synthesis problem; i.e., $\mathbf{x}_0 = c\mathbf{A}^*_0\mathbf{d} + \mathbf{b}^*_0$.[7] The solution sequence of FFHC(t) is $\mathbf{x}^*_t, \mathbf{x}^{(t)}_{t+1}, \ldots, \mathbf{x}^{(t)}_{t+h}$. We use the "$*$" notation to highlight that only the first decision of each subproblem is committed; the remaining decisions are merely advisory.

Start with $t = 1$. Let $\mathbf{d}_{1+h} := (d_0, \ldots, d_{1+h}, \tilde{d}_{2+h}, \ldots, \tilde{d}_T)$ be a sequence in $\mathcal{D}_{0:1+h}$ where $(d_0, \ldots, d_{1+h})$ is the demand subsequence known at time 1 and $(\tilde{d}_{2+h}, \ldots, \tilde{d}_T)$ is a sequence of ersatz values that fill out the remainder of the time horizon while satisfying the relevant constraints in $\mathcal{D}$.

We claim that sequence of dispatches

$$\begin{aligned} \mathbf{x}^*_1 &= \mathbf{A}^*_1\mathbf{d}_{1+h} + \mathbf{b}^*_1 \\ \mathbf{x}^{(1)}_2 &= \mathbf{A}^*_2\mathbf{d}_{1+h} + \mathbf{b}^*_2 \\ &\vdots \\ \mathbf{x}^{(1)}_{1+h} &= \mathbf{A}^*_{1+h}\mathbf{d}_{1+h} + \mathbf{b}^*_{1+h} \end{aligned} \tag{2.7}$$

is feasible for FFHC(1). This dispatch sequence is unique, despite the fact that demand values $d_{h+2}, \ldots, d_T$ are yet unknown, because of the causality constraints

[7] This assumption is not restrictive as $\mathbf{x}_0$ can be fixed if desired when solving the RAP synthesis problem.

on the $\mathbf{A}_t^*$ matrices. Since $\mathbf{d}_{1+h} \in \mathcal{D}$, (2.7) satisfies constraints (2.4b - 2.4d) for $t = 1, \ldots, h$ and these exactly match constraints (2.5b - 2.5e) in FFHC(1), then (2.7) is feasible for (2.5b - 2.5e) in FFHC(1). Constraint (2.5f) in FFHC(1) is also satisfied, because (2.5f) is just constraint (2.4d) for interval $h + 2$ restricted $\mathcal{D}_{0:1+h}$, which is a subset of $\mathcal{D}$ over which constraint (2.4d) is guaranteed to hold. Thus (2.7) is feasible for FFHC(1).

Next, assume there exists a feasible solution sequence for FFHC($t - 1$), namely $\mathbf{x}_{t-1}^*, \mathbf{x}_t^{(t-1)}, \ldots, \mathbf{x}_{t+h-1}^{(t-1)}$. Define $\mathbf{d}_{t+h} := (d_0, \ldots, d_{t+h}, \tilde{d}_{t+h+1}, \ldots, \tilde{d}_T)$ analogously to $\mathbf{d}_{1+h}$ from above. When rolling ahead to time t, an additional demand value d_{t+h} is revealed. Because we assume perfect forecasts, the first h demand values in demand sequence for time t coincide with the last h values in the demand sequence for time $t - 1$. At time t, the previous first decision $\mathbf{x}_{t-1}^*$ has been committed.

We claim that

$$\begin{aligned} \mathbf{x}_t^* &:= \mathbf{x}_t^{(t-1)} \\ \mathbf{x}_{t+1}^{(t)} &:= \mathbf{x}_{t+1}^{(t-1)} \\ &\vdots \\ \mathbf{x}_{t+h-1}^{(t)} &:= \mathbf{x}_{t+h-1}^{(t-1)} \\ \mathbf{x}_{t+h}^{(t)} &:= \mathbf{A}_{t+h}^* \mathbf{d}_{t+h} + \mathbf{b}_{t+h}^* \end{aligned} \tag{2.8}$$

is a feasible solution for FFHC(t). The sequence $\mathbf{x}_t^{(t-1)}, \ldots, \mathbf{x}_{t+h-1}^{(t-1)}$ satisfies constraints (2.5b - 2.5e) for $t, \ldots, t + h - 1$ in FFHC(t) which can be seen by matching identical constraints between the subproblems FFHC($t - 1$) and FFHC(t).

We now need to construct a feasible $\mathbf{x}_{t+h}^{(t)}$. Since $\mathbf{x}_{t+h-1}^{(t)} := \mathbf{x}_{t+h-1}^{(t-1)}$ satisfies constraint (2.5f) in FFHC($t - 1$) for any $\mathbf{d} \in \mathcal{D}_{0:t+h-1}$ and $\mathcal{D}_{0:t+h} \subseteq \mathcal{D}_{0:t+h-1}$, then $\mathbf{x}_{t+h}^{(t)} := \mathbf{A}_{t+h}^* \mathbf{d}_{t+h} + \mathbf{b}_{t+h}$ satisfies constraint (2.5e) for interval $t + h$ in FFHC(t).

It only remains to check that $\mathbf{x}_{t+h} = \mathbf{A}_{t+h}^* \mathbf{d}_{t+h} + \mathbf{b}_{t+h}$ satisfies constraint (2.5f) for FFHC(t). Note that $\mathcal{D}_{0:t+h} \subseteq \mathcal{D}$ and therefore any $\mathbf{d} \in \mathcal{D}_{0:t+h}$ satisfies the $\forall$ quantifier in constraint (2.5e). So

$$-\underline{\Delta}_x^* \leq \mathbf{A}_{t+h+1}^* \mathbf{d} + \mathbf{b}_{t+h+1}^* - \mathbf{A}_{t+h}^* \mathbf{d} - \mathbf{b}_{t+h}^* \leq \overline{\Delta}_x^*$$

holding $\forall \mathbf{d} \in \mathcal{D}_{0:t+h}$ implies

$$-\underline{\Delta}_x^* \leq \mathbf{A}_{t+h+1}^* \mathbf{d} + \mathbf{b}_{t+h+1}^* - \mathbf{x}_{t+h} \leq \overline{\Delta}_x^* \; \forall \mathbf{d} \in \mathcal{D}_{0:t+h}.$$

Thus (2.8) is feasible for FFHC(t). □

Remark 2. *The terminal constraint (2.5f) is only applied through* FFHC$(T-h-1)$. *Subsequent rounds of* FFHC *have no demand uncertainty.* FFHC*'s guaranteed feasibility distinguishes it from [121]. Moreover, the placement of constraint (2.5f) on the terminal decision enable* FFHC *to fully exploit all perfect predictions of demand, in contrast to work in [121, 84, 85].*

2.4 Upper and lower bounds on feasible online dispatch algorithms

We now turn to bounding the worst-case performance of the class of feasible dispatch algorithms, which contain our proposed FFHC as an instance. The exactly matching upper and lower bounds we obtain establish fundamental limits on the performance of algorithms for RTED. They also establish that, in general, feasibility of an online algorithm implies optimality. In other words, *feasibility is the best you can do*. Nonetheless, different algorithms can be distinguished in their average-case performance, as we examine in the experiments in Section 2.5.

We evaluate performance via the metric of competitive ratio, which has recently seen increasing use in the control and power systems communities [30, 26, 124]:

$$\mathrm{CR}_{\mathrm{ALG}} = \sup_{\mathbf{d}\in\mathcal{D}} \frac{\mathrm{Cost}_{\mathrm{ALG}}(\mathbf{d})}{\mathrm{Cost}_{\mathrm{OPT}}(\mathbf{d})} \tag{2.9}$$

A competitive ratio of 1 signifies optimal performance of the online algorithm, whereas a competitive ratio larger than 1 indicates suboptimal performance. We choose to focus on the competitive ratio because it is unitless and time-independent, thus facilitating fair comparison of algorithm performance across different problem instances and system parameters.[8]

In the following theorems we assume for clarity of exposition that $\mathcal{D} \subseteq \mathbb{R}^T_{\geq 0}$, dispatch variables $\mathbf{x}_t$ are always nonnegative, and costs are linear, positive, and potentially time-varying. However, performance bounds can be obtained in more general settings.

Theorem 3. *Suppose* ALG $:= \{X_t\}_{t=1}^T$ *is a feasible online dispatch algorithm for demand uncertainty set $\mathcal{D}$ on some arbitrary system, where costs are linear and time-varying, $\mathbf{c}_1, \ldots, \mathbf{c}_T \in \mathbb{R}^N_{>0}$. Then, the competitive ratio of* ALG *is bounded*

[8] Competitive difference upper and lower bounds can be obtained for FFHC, and more generally for arbitrary feasible online dispatch algorithms, that essentially match those in [9], with slight modifications due to the inclusion of supply-demand balance constraints in our setting. Further, [9] obtains upper and lower bounds on competitive difference matching up to a factor of 4, whereas our competitive ratio upper and lower bounds match exactly.

above as:

$$\mathrm{CR}_{\mathrm{ALG}} \leq \max_{s,t\in[1,T]} \frac{c_{\max,s}}{c_{\min,t}} \tag{2.10}$$

where $c_{\max,t} = \max_{i\in[1,N]} c_{i,t}$ *and* $c_{\min,t} = \min_{i\in[1,N]} c_{i,t}$.

Proof. Fix an arbitrary $\mathbf{d} \in \mathcal{D}$, and call $\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*$ an optimal solution of the offline dispatch problem (2.2) for this demand sequence. Feasibility of $\{X_t\}_{t=1}^T$ as well as the offline optimal dispatch trajectory implies that supply-demand balance holds for all t for each trajectory, i.e., $\mathbf{1}^\top \mathbf{x}_t^* = \mathbf{1}^\top X_t(\mathbf{d}) = d_t$ for all $t \in [1, T]$. We can bound the online algorithm's dispatch cost at time t above by $c_{\max,t} d_t$, and similarly we can bound the offline algorithm's dispatch cost at time t below by $c_{\min,t} d_t$. The competitive ratio is 1 by definition if all demands are 0. Otherwise, we obtain the following bound:

$$\begin{aligned}
\mathrm{CR}_{\mathrm{ALG}} &\leq \max_{\mathbf{d}\in\mathcal{D}} \frac{\sum_{s=1}^T c_{\max,s} d_s}{\sum_{t=1}^T c_{\min,t} d_t} \\
&\leq \max_{\mathbf{d}\in\mathcal{D}} \frac{\mathbf{1}^\top \mathbf{d} \cdot \max_{s\in[1,T]} c_{\max,s}}{\mathbf{1}^\top \mathbf{d} \cdot \min_{t\in[1,T]} c_{\max,t}} \\
&\leq \max_{s,t\in[1,T]} \frac{c_{\max,s}}{c_{\min,t}}.
\end{aligned}$$

□

Remark 3. *An even tighter, yet less interpretable, upper bound can be obtained on the competitive difference (CD) of any feasible online dispatch algorithm. Specifically, on a particular demand trajectory* $\mathbf{d} \in \mathcal{D}$*, the cost of any feasible online dispatch algorithm will be upper bounded by the "offline least-optimal" solution, i.e., the solution to (2.2) where the* min *has been replaced with a* max*, which will be the feasible dispatch with worst-case cost. Under the assumption of linear costs, this is yet another linear program, and thus* $\mathrm{CD}_{\{X_t\}_{t=1}^T}$ *can be bounded by the optimal value of the following linear program, which maximizes both over the difference between the offline least and most optimal, as well as over demand trajectories:*

$$\max_{\substack{\mathbf{x}_1,\ldots,\mathbf{x}_T\in\mathbb{R}^N \\ \mathbf{y}_1,\ldots,\mathbf{y}_T\in\mathbb{R}^N \\ \mathbf{d}\in\mathcal{D}}} \quad \sum_{t=1}^T \mathbf{c}_t^\top (\mathbf{x}_t - \mathbf{y}_t) \tag{2.11a}$$

$$\text{s.t.} \quad \mathbf{1}^\top \mathbf{x}_t = d_t,\ \mathbf{1}^\top \mathbf{y}_t = d_t \qquad \forall t \in [T] \tag{2.11b}$$

$$\underline{\mathbf{x}} \leq \mathbf{x}_t, \mathbf{y}_t \leq \overline{\mathbf{x}} \qquad \forall t \in [T] \tag{2.11c}$$

$$-\underline{\Delta}_x \leq \mathbf{x}_t - \mathbf{x}_{t-1} \leq \overline{\Delta}_x \qquad \forall t \in [T] \tag{2.11d}$$

$$-\underline{\Delta}_x \leq \mathbf{y}_t - \mathbf{y}_{t-1} \leq \overline{\Delta}_x \qquad \forall t \in [T] \tag{2.11e}$$

where $\mathbf{y}_0 = \mathbf{x}_0$.

In the following, we construct a system with linear, time-invariant costs and generators which have capacity and ramp constraints. We assume planning decisions $\mathbf{y}^*$ result in a certain generator having arbitrarily slow ramp limit ϵ. Thereby, in the regime where $\epsilon \to 0$, there is a demand trajectory on which any feasible dispatch algorithm must have competitive ratio arbitrarily close to the upper bound (2.10). This constitutes an exactly tight lower bound on the competitive ratio of any feasible dispatch algorithm.

Theorem 4. *Fix* $\epsilon \in (0, 1)$. *There exists a choice of system parameters* $\mathbf{y}^*(\epsilon)$ *with linear, time-invariant costs* $\mathbf{c} \in \mathbb{R}^N_{\geq 0}$ *and a polytopic demand uncertainty set* $\mathcal{D}$*, as well as a distinguished demand sequence* $\hat{\mathbf{d}} \in \mathcal{D}$*, such that for any feasible online dispatch algorithm* $\text{ALG} := \{X_t\}_{t=1}^T$*,*

$$\frac{\text{Cost}_{\text{ALG}}(\hat{\mathbf{d}})}{\text{Cost}_{\text{OPT}}(\hat{\mathbf{d}})} \geq \epsilon + (1-\epsilon)\frac{c_{\max}}{c_{\min}} \tag{2.12}$$

where $c_{\max} := \max_{i\in[1,N]} c_i$ *and* $c_{\min} := \min_{i\in[1,N]} c_i$.

Proof. Fix h to be some positive integer, independent of T. We construct a 2-generator system with costs $\mathbf{c} = (c_{\max}, c_{\min})$; capacity lower and upper bounds $\underline{\mathbf{x}} = (0, 0)$, $\overline{\mathbf{x}} = (2h, 2h)$; ramp lower and upper bounds $\overline{\Delta} = (\epsilon, 2-\epsilon)$ and $\underline{\Delta} = (-\epsilon, -2+\epsilon)$; and initial operating point $\mathbf{x}_0 = ((2-\epsilon)h, \epsilon h)$. We define the demand uncertainty set $\mathcal{D} \subset \mathbb{R}^T$ as follows:

$$\mathcal{D} = \{\mathbf{d} : d_0 = 2h, d_t \leq d_{t+1} \leq d_t + 2, d_t \leq 4h \,\forall t \in [0, T]\},$$

where we define $d_0 = \mathbf{1}^\top \mathbf{x}_0 = 2h$. Observe that $\mathcal{D}$ admits an online feasible algorithm: specifically, the online algorithm that chooses its operating point at time t in the set $\{\mathbf{x} : \mathbf{1}^\top \mathbf{x} = d_t, \mathbf{x} = \lambda \mathbf{x}_0 + (1-\lambda)\overline{\mathbf{x}}, \lambda \in [0, 1]\}$ is feasible for $\mathcal{D}$.

Now consider the specific demand trajectory $\hat{\mathbf{d}}$ with $\hat{d_t} = 2h$ for all $t \in [1, T]$. We claim that, for all times $t \in [1, T-2h]$, $X_t(\hat{\mathbf{d}})_1 \geq (2-2\epsilon)h$. We prove this by contradiction: suppose alternatively that $X_t(\hat{\mathbf{d}})_1 < (2-2\epsilon)h$ for some $t \in [1, T-2h]$. But consider another demand trajectory $\tilde{\mathbf{d}} \in \mathcal{D}$ defined by $\tilde{d_s} = 2h$ for $s \in [1, t+h]$ and $\tilde{d_s} = \min\{2h + 2(s-t-h), 4h\}$ for $s \in [t+h+1, T]$. As $\hat{\mathbf{d}}$ and $\tilde{\mathbf{d}}$ coincide in their entries through time $t+h$, causality dictates that $X_t(\hat{\mathbf{d}}) = X_t(\tilde{\mathbf{d}})$, so $X_t(\tilde{\mathbf{d}})_1 < (2-2\epsilon)h$ as well. But then the online algorithm cannot remain feasible for $\tilde{\mathbf{d}}$ for the rest of time: this is because feasibly meeting $\tilde{\mathbf{d}}$ for the rest of time,

and in particular remaining feasible for the sequence of demand increases beginning at time $t + h + 1$, requires $X_{t+h}(\tilde{\mathbf{d}}) = \mathbf{x}_0$ due to the ramp and capacity constraints. However, since $X_t(\tilde{\mathbf{d}})_1 < (2 - 2\epsilon)h$ and up-ramp on the first generator is bounded by ϵ, it is impossible for the online algorithm to reach $\mathbf{x}_0$ at time $t + h$. Thus the online algorithm cannot be feasible for $\mathcal{D}$, yielding a contradiction.

By the last paragraph's result, we know that for $t \in [1, T-2h]$, $X_t(\hat{\mathbf{d}})_1 \geq (2-2\epsilon)h$; as $X_t(\hat{\mathbf{d}})_1 = d_t - X_t(\hat{\mathbf{d}})_2$, it follows that the cost of the online algorithm on each of the first $T - 2h$ timesteps is lower bounded by $(2 - 2\epsilon)hc_{\max} + 2h\epsilon c_{\min}$. On each of the last $2h$ timesteps, we trivially lower bound the online algorithm cost by $2hc_{\min}$. Thus we obtain

$$\text{Cost}_{\text{ALG}}(\hat{\mathbf{d}}) \geq (T-2h)\left((2-2\epsilon)hc_{\max}+2h\epsilon c_{\min}\right)+(2h)^2 c_{\min} \tag{2.13}$$

Now we turn to providing an upper bound on $\text{Cost}_{\text{OPT}}(\hat{\mathbf{d}})$. Since the offline optimal knows all demands in advance, it will ramp maximally to transfer all generation onto the second, cheaper generator and will remain at this operating point for the rest of time. It will take $\frac{(2-\epsilon)h}{\epsilon}$ timesteps to ramp to the operating point $(0, 2h)$, since generator 1 has ramp limit ϵ and demand is constant through time. For each of these first $\frac{(2-\epsilon)h}{\epsilon}$ timesteps, we upper bound the offline optimal cost trivially by $2hc_{\max}$ in each step. Once the offline optimal reaches $(0, 2h)$, its cost in each step for the rest of time is exactly $2hc_{\min}$. We get:

$$\text{Cost}_{\text{OPT}}(\hat{\mathbf{d}}) \leq \frac{(2-\epsilon)h}{\epsilon}(2hc_{\max}) + \left(T-\frac{(2-\epsilon)h}{\epsilon}\right)(2hc_{\min}) \tag{2.14}$$

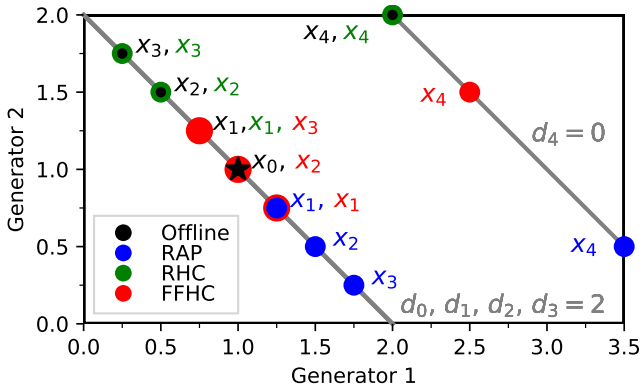
Forming the ratio of (2.13) with (2.14) and taking the limit as $T \to \infty$ yields the lower bound (2.12). □

2.5 Experiments

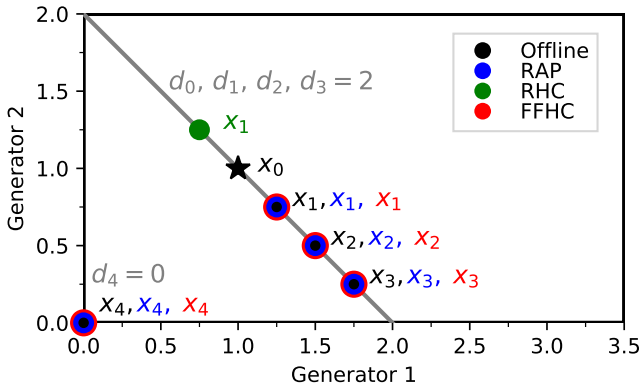
In this section we explore through simulations on simple systems how the proposed algorithm handles infeasibilities that otherwise arise when resource procurement is done in a dispatch-agnostic fashion. We also include a discussion about the scalability of the method to larger, more realistic power systems.

A two-generator case

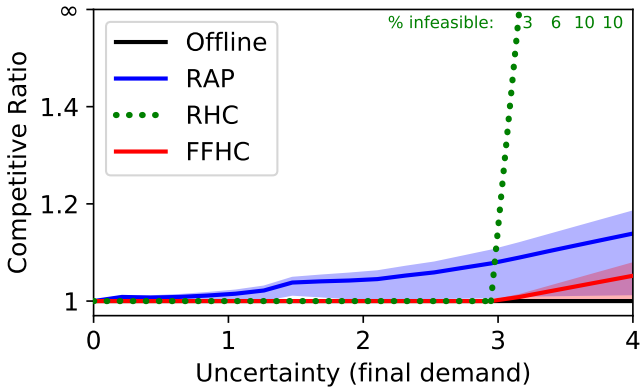
We use a two-generator case (same as the one presented in the proof of Theorem 1) to show how FFHC is able to compute a feasible dispatch when the standard RHC algorithm cannot.



(a) "Easy" case: $\mathbf{d}^{(A)}$



(b) "Hard" case: $\mathbf{d}^{(B)}$



(c) Competitive ratio vs. uncertainty

Figure 2.2: Counterexample from Theorem 1 revisited. Panel (a) shows a scenario where all algorithms are feasible. Panel (b) shows a scenario where RHC is unable to remain feasible, whereas FFHC remains feasible. Panel (c) illustrates the performance of the RHC, FFHC, and RAP against the offline optimal as the size of the uncertainty set grows. The dotted green line going to ∞ indicates that RHC becomes infeasible on some trajectories beginning at an uncertainty value of 3.

Setting $h = 2$ and $T = 4$, we define a demand uncertainty set $\mathcal{D} = \{(2, 2, 2, d) : d \in [1, 4]\}$. We run DAP on this system with costs $\mathbf{c} = (1, 3/4)$, $\bar{\mathbf{c}} = (10, 11)$, nominal max capacity $\overline{\mathbf{x}} = (2, 2)$, nominal ramp rates $\underline{\Delta}^{\text{nom}} = \overline{\Delta}^{\text{nom}} = (2, 1/4)$, and starting point $\mathbf{x}_0 = (1, 1)$. For offline planning and dispatch, the nominal max capacity of $(2, 2)$ is sufficient to satisfy all $\mathbf{d} \in \mathcal{D}$. DAP procures an additional 37.5% capacity and a proportional amount of ramp capacity on (lower cost) Generator 1.

Figures 2.2a and 2.2b show the performance of the algorithms RHC, FFHC, and RAP (affine policies synthesized in DAP), as well as the offline optimal, on the two demand sequences $\mathbf{d}^{(A)}$ and $\mathbf{d}^{(B)}$ distinguished in Theorem 1. $\mathbf{d}^{(A)}$ is an "easy" demand sequence: all algorithms are feasible, and FFHC takes more conservative (i.e., costlier) decisions than RHC, which immediately moves up to the top left of the capacity region to exploit the lower cost of Generator 2. On the other hand, $\mathbf{d}^{(B)}$ details the "hard" demand sequence , for which RHC is unable to remain feasible, since it mistakenly chooses to exploit the lower cost of Generator 2 production at $t = 1$, leaving it unable to meet d_4. FFHC is able to remain feasible in contrast.

Figure 2.2c compares the performance of the algorithms via competitive ratio as the demand uncertainty set is scaled. We parameterize the uncertainty set by u:

$$\mathcal{D}(u) := \{(2, 2, 2, d) : d \in [2 - u/2, 2 + u/2]\}$$

For $u \in [0, 4]$, we run DAP on $\mathcal{D}(u)$ to determine system parameters. We then sample trajectories from $\mathcal{D}(u)$ set using a hit-and-run sampler for polytopes [50], and compute the dispatch of each algorithm. The mean empirical competitive ratio of the trajectories along with upper/lower bounds (shaded) for each algorithm are shown in Figure 2.2c.

While the performance of RAP suffers in comparison to that of the offline optimal, both FFHC and RHC exactly match the offline performance for $u < 3$. For $u > 3$, RHC begins encountering infeasibility on some of the demand trajectories, and by $u = 4$ is infeasible for 10% of the sampled trajectories. Meanwhile, FFHC, just like RAP, always remains feasible, though its performance degrades slightly from that of the offline since its dispatches are influenced by the robust constraint.

Scenario based on CAISO load profile

The purpose of this example is to show the necessity of dispatch-aware planning to maintain feasibility under realistic net load profiles. For simplicity, we do not incorporate integer unit commitment variables and associated cost functions and therefore

the setting is not intended to represent the particular variety of unit commitment problems solved by system operators.

We consider the small power system shown in Figure 2.3, which has 1 GW of peak load and four generation sources: variable renewables (wind & solar), a fast-ramping gas turbine, a slow-ramping coal plant, and a transmission interconnection. This setup, while stylized, represents the scenario of a transmission-constrained zone within a larger grid where local infeasibilities could arise under high fluctuations of net demand and ramp shortages.

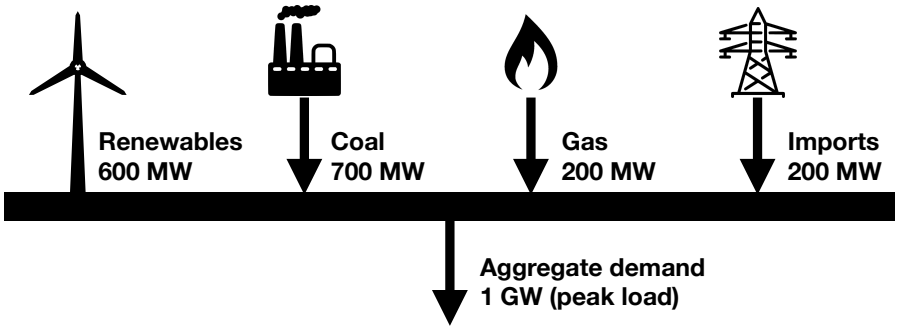


Figure 2.3: Power system for 4-generator case study. Capacities shown are peak values.

Details on capacity, ramp rates, and costs for the generation sources are given in Table 2.1.

Generation Type	Max Capacity (MW)	Ramp Rate (% cap./min)	Variable Cost ($/MWh)	Capacity Cost ($/MW)
Imports	200	±5	1.93	0
Gas	200	±2	2.56	1.08×10^6
Coal	700	±0.5	4.52	3.67×10^6
Renewables	600	Instantaneous	0	NA

Table 2.1: Parameters for generation sources. Costs are from [34, Table 1]. Max capacity values indicate the maximum available generation for each type. CAISO generation mix is used to derive a import cost [2]. Ramp rates are taken from reasonable ranges given in [71], [48].

The 24 hr nominal generation profile sampled at 15 min intervals is taken from CAISO's aggregate demand on Sept. 9, 2021 [21]. We subtract the variable renewable generation profile (from [21]) to get a nominal *net* demand curve (solid black line in Figure 2.4a), around which a demand uncertainty set is constructed. Trajectories are sampled uniformly from this set using a hit-and-run sampler [50].

Capacity costs are used in the DAP problem to determine an optimal robust generation mix (neither renewables nor imports are included in this step as they are considered already fixed).

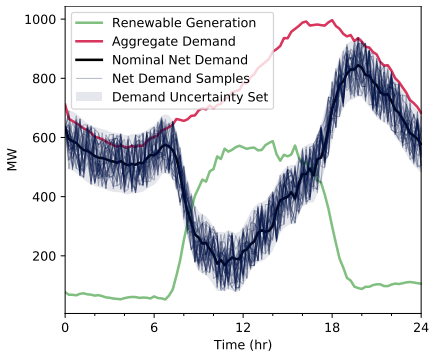
Due to the fast, sustained afternoon ramp event in the net load profile, the standard RHC dispatch runs into infeasibility for 29% of sampled demand trajectories. In contrast, FFHC is always feasible at little to no extra cost beyond that incurred by the offline optimal. Figure 2.4b shows the optimal solutions for a particular net demand trajectory. Prior to 20.5 hrs, both algorithms return identical solutions. After that, RHC becomes infeasible whereas FFHC does not. At the start of the ramp event, imports are already at their maximum and RHC chooses to myopically exploit the lower cost of the gas generator. In contrast, FFHC "pre-ramps" the slow coal generator and saves ramping capacity on the gas generator to accommodate later fluctuations. It is also notable that even though the planning variables (generator capacities and ramps) in this example are online feasible, RHC is still unable to remain feasible. Thus, the existence guarantee of an online feasible dispatch algorithm does not imply that even a good (on average) policy like RHC can produce a feasible sequence of dispatches.

Figure 2.4c shows that the feasibility guarantee of FFHC comes at a very minimal efficiency loss. When RHC is can stay feasible, both algorithms attain near-optimal cost with average empirical competitive ratios $CR_{RHC} = 1.0000$ and $CR_{FFHC} = 1.0002$. In comparison, the RAP algorithm, while robustly feasible, has a significantly higher average competitive ratio of 1.0934, indicating the value of using predictions (robustly!) in our approach. The robust resource procurement step (DAP) in this simulation procures 1066.6 MW of total generation capacity, which is 14% above peak demand of 938.8 MW. In ramp constrained power systems, additional capacity may be required to accommodate long high ramp events. DAP provides a way to directly optimize for this margin.

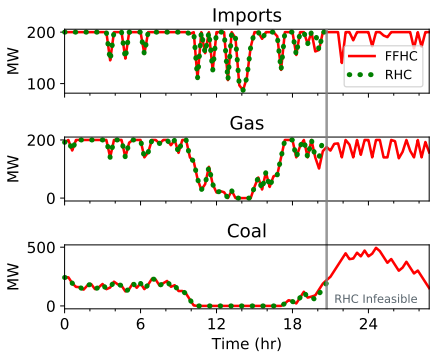
Discussion of algorithm scalability

DAP requires solving a robust linear program, a problem known to suffer from scalability issues. While scalability is not the focus of this work, in this section we discuss effective strategies for reducing the problem dimension and highlight relevant existing literature on this subject.[9]

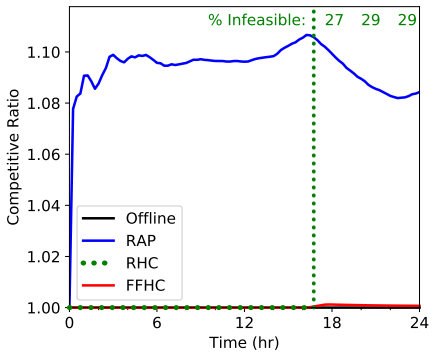
[9]We focus on DAP for scalability; the FFHC stage of our joint algorithm only includes a small robust constraint that does not appreciably affect computation.



(a) Aggregate demand, renewable generation, and net nominal demand, along with additional net demand trajectories from demand uncertainty set. U/L bounds are ±20% of nominal.



(b) Dispatch solutions for RHC vs. FFHC on a particular demand trajectory that is infeasible for RHC. Lookahead horizon is set at 1 hr for both.



(c) Average competitive ratio of each algorithm plotted vs. time.

Figure 2.4: Results from four-generator system using CAISO load and renewable generation profiles. Approximately 29% of 300 sampled demand trajectories are infeasible for RHC. The vertical dotted green line in panel (c) indicates at which time RHC first becomes infeasible and the values at the top of the frame display the percentage of infeasible trajectories (out of the total sampled) at regular time intervals.

Optimization problems for large N-generator power systems (e.g., SCUC, capacity planning) already present operators with a demanding computational task, with $O(NT)$ variables and $O(M)$ constraints where M can be as large as $O(N^2T)$ for mesh network topologies. We are concerned with the *additional* complexity our robust linear formulation adds to this baseline, which arises from 1) the robust description of the uncertainty set $\mathcal{D}$ and 2) expressiveness of the causal affine policy class.

For 1), we take the reasonable assumption that correlations between elements of the demand vector are limited to neighbors. This means that the number of constraints in $\mathcal{D}$ is $O(T)$, rather than $O(T^2)$ which would arise if full correlations were allowed. For 2) we observe that synthesized policies often only make use of a few previous demand steps, which we call memory m with $m = O(1)$. This allows us to limit the size of the affine policies to Nm variables, as opposed to NT for full-history policies. Using limited memory policies necessitates a careful reformulation of $\mathcal{D}$ and the problem constraints, but the downstream benefits for the size of the robust LP are significant, as the total number of constraints ultimately scales with the number of policy variables. Restricting policy memory also eliminates the $O(NT^2)$ causality constraints required for the full policies.

Table 2.2 summarizes the number of constraints and variables (in order sense) for each problem setting. Limited-memory policies allow for the multiplicative factor of T in both variables and constraints for the full-memory robust formulation to be reduced to a (tunable and small) constant factor m.

	Offline	Memory-T policies	Memory-m policies
Variables	$O(NT)$	$O(NT^2 + MT)$	$O(mNT + mM)$
Constraints	$O(M)$	$O(NT^2 + MT)$	$O(mM)$

Table 2.2: Comparison of number of variables and constraints for offline and two robust formulations.

After reducing the problem size in the proposed manner, the resulting problem may still be a large LP. We point the interested reader to the excellent discussion of this issue in [84, 85] where a constraint generation approach along with various other algorithmic tweaks allow for efficient solutions to large LP/MILP power system problems. All of the proposed methods therein are applicable to our setting.

2.6 Conclusion

In this work, we analyze properties of feasible online dispatch algorithms in general, and specifically propose a joint algorithm for resource procurement and RTED that exploits lookahead predictions for good performance while also guaranteeing feasibility. Our framework is applicable to several types of resource procurement problem including SCUC and resource adequacy, and is compatible with arbitrary fixed-horizon lookahead optimization problems. We further present exactly matching upper and lower bounds on the competitive ratio for the problem class of RTED. Finally, our computational results demonstrate that FFHC nearly matches the performance of the offline optimal while always remains feasible, which contrasts with the frequent infeasibility of RHC. Thus the proposed approach provides feasible RTED with nearly no loss of efficiency compared to the standard algorithm.

Future work includes applying this algorithmic framework to problems with energy storage and time-varying state-of-charge requirements, as well as designing incentive compatible prices for dispatches computed by feasible RTED algorithms.

Chapter 3

PRICING UNCERTAINTY IN STOCHASTIC MULTI-STAGE ELECTRICITY MARKETS

In this chapter, we propose a pricing scheme for multi-stage markets that is independent of the particular characterization of parameter uncertainty or the specific risk-aware method for optimizing dispatch decisions under parameter uncertainty. Our approach is distinguished from several recent works where the construction of the energy price intimately depends on the optimization paradigm (e.g., chance-constrained [42] or robust [94]). We show that our proposed price can be decomposed into components corresponding to the standard locational marginal price (LMP), intertemporal coupling, and uncertainty. Finally, we establish that this price clears the market under profit-maximizing assumptions on the participants and that it supports both *ex-ante* and *ex-post* dispatch-following incentives.

Related Work

Our work draws on two main lines of inquiry into electricity market mechanism design. The first is dispatching and pricing multi-interval markets in the presence of intertemporal coupling constraints. The second is dispatching and pricing using techniques from robust and stochastic optimization.

Pricing multi-period electricity markets In rolling-window real-time economic dispatch schemes, which are seeing wider deployment across system operators to mitigate uncertainty in real-time operation, distribution shift in predicted net demand can lead to lost opportunity cost and distorted truthful bidding incentives for generators. Several pricing mechanisms building on standard uniform pricing schemes have been proposed in recent years to mitigate the lack of dispatch-following incentives [63, 68, 139]. A more recent line of work [57, 27, 28] has proposed a non-uniform pricing scheme, Temporal Locational Marginal Pricing (TLMP), and has established a dual definition of dispatch-following incentives. Simultaneously satisfying a "partial equilibrium" (i.e., *ex ante* dispatch-following incentive in every stage) and a general equilibrium (i.e., *ex post*) forms the notion of "strong equilibrium," used in this work.

Our study is distinct from the prior literature on pricing multi-period markets as these

works do not incorporate uncertainty directly in the lookahead dispatch algorithms, but merely design prices to mitigate incentive misalignment as a result of inaccurate predictions and distribution shift. However, these lookahead algorithms might be infeasible [84, 32], necessitating our development of more general pricing schemes that can incorporate such robust constraints.

Pricing stochastic electricity markets There has been much recent interest in designing electricity markets incorporating robust or stochastic constraints to ensure reliable operation in the face of uncertainty. For example, such dispatch schemes include economic dispatch with robust constraints [94, 134], chance constraints [73, 92, 91], distributionally robust chance constraints [42, 109], and conditional value at risk constraints [99]. However, in the subset of these works that explicitly address the problem of designing price mechanisms for the stochastic dispatch problem, inconsistent notions of *ex ante* dispatch-following incentives are considered which leave open the need for out-of-market settlements to make up for lost opportunity cost. In contrast, our work proposes pricing mechanisms that can be applied to any formulation of stochastic or robust economic dispatch, and which ensure zero lost opportunity cost on the part of market participants by considering both *ex ante* and *ex post* dispatch-following incentives in the price specification.

3.1 Multi-stage dispatch under uncertainty

The day-ahead (DA) and real-time (RT) stages of electricity market clearing form a $T + 1$ stage sequential optimization problem, with coupling between the stages and uncertainty from load and renewables realized between each of the T stages. The first stage is the single-shot, DA optimization problem which determines a unit commitment and associated dispatch for the upcoming 24-hour time horizon. This dispatch, although not physically realized, may be financially settled. Subsequently, in real time, a receding-horizon multi-interval optimization is performed. The first interval from each of these T subproblems is financially binding. Between each of the subproblems, the SO utilizes updated forecasts of uncertain demand and renewable generation to improve the efficiency of the dispatch.

The stages of the sequential problem are temporally coupled in the manner depicted in Figure 3.1. The first (DA) stage couples to all of the subsequent stages because it fixes the unit commitment – and therefore the upper/lower generation bounds, ramp limits, etc. – in the T subsequent (RT) stages. Within the RT market, stages are coupled consecutively due to the form of ramping constraints and the battery

state-of-charge updates.

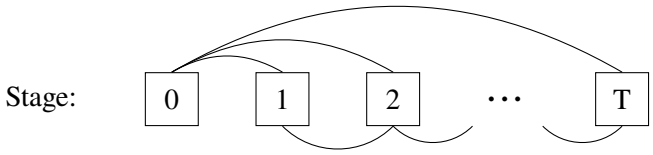


Figure 3.1: Coupling between $T + 1$ stages in DA + RT economic dispatch.

Since the $T + 1$ stages are solved and settled sequentially, we consider two groups of stages at a time: the period with no uncertainty, and the set of periods with remaining uncertainty. In the DA stage, the SO seeks to solve a stochastic optimization problem that fixes here-and-now decisions for the unit commitment while selecting policies for the wait-and-see decisions of RT stage 1. The purpose of the policies is to provide realization-dependent recourse in subsequent stages. However, in each of these stages, after uncertainty has been revealed, the multi-interval optimization is solved again for the next stage.

Notation

For each optimization interval indexed by $t \in \{0, \ldots, T\}$, each market participant $i \in \{1, \ldots, N\}$ has a dispatch vector $\mathbf{x}_{i,t} \in \mathbb{R}^{M_{i,t}}$ where $M_{i,t}$ is the dimension of the dispatch vector for i in stage t. The dispatch $\mathbf{x}_{i,t}$ includes all of the quantities associated with participant i in stage t. For conventional generators, this is just their power generation. For storage resources, it includes power generation and state-of-charge. We do not consider discrete variables, such as those needed for unit commitment, in our presentation here. They can be included without impacting our pricing or dispatch results, although the dispatch problem would need to be modified slightly as in [42, 102]. System states, such as nodal power injections, line flows, and voltage angles, can be written in terms of the individual dispatch variables $\mathbf{x}_{i,t}$ and are therefore not explicitly notated. For each t, we collect dispatch vectors into a single decision vector:

$$\mathbf{x}_t \coloneqq (\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{N,t}) \in \mathbb{R}^{M_t},$$

where $M_t \coloneqq \sum_i M_{i,t}$. Associated with each dispatch vector is a market price $\boldsymbol{\pi}_t \in \mathbb{R}^{M_t}$. The revenue (or payment) each participant receives over the entire horizon is $\boldsymbol{\pi}_t^\top \mathbf{x}_t$.

For each t we associate a random vector of uncertainty $\boldsymbol{\xi}_t \in \mathbb{R}^{P_t}$. Realizations of $\boldsymbol{\xi}_t$, denoted $\hat{\boldsymbol{\xi}}_t$, are obtained sequentially after the dispatch $\hat{\mathbf{x}}_{t-1}$ has been committed but

prior to computing $\mathbf{x}_t$. We also assume that the SO has access to a forecast $\boldsymbol{\theta}_t$ that represents their best knowledge at stage t about subsequent uncertainty $\boldsymbol{\xi}_{t+1}, \ldots, \boldsymbol{\xi}_T$. The composition of the forecast depends on what information is accessible. In the simplest case, $\boldsymbol{\theta}_t$ is just a point forecast of $\boldsymbol{\xi}_{t+1}, \ldots, \boldsymbol{\xi}_T$. When distributional information is available, $\boldsymbol{\theta}_t$ can be a set of parameters describing each forecast distribution and its support. Since stage 0 is the day-ahead/unit-commitment stage of the market clearing, which happens when no uncertainty has been realized, $\hat{\boldsymbol{\xi}}_0$ is defined to be a set of forecasts over the subsequent T RT intervals.

In the rest of the chapter, we denote by $\mathbf{a}_{\tau:t}$ the set of vectors $\{\mathbf{a}_j\}_{j=\tau}^{t}$. If $\tau > t$, we define this to be the empty set. For $\tau, t \in \mathbb{N}$ satisfying $\tau \leq t$, we define $[\tau, t] \coloneqq \{\tau, \tau + 1, \ldots, t\}$.

Ex-post Dispatch Problem and Prices

If the SO had perfect forecasts of uncertainty, it could solve the following optimization problem (3.1) for all time intervals simultaneously. This is a useful solution because it benchmarks the efficiency of dispatch algorithms and quantifies the impact of uncertainty.

Problem 1. *Given an uncertainty realization $\hat{\boldsymbol{\xi}}$, the ex-post dispatch problem for all $T + 1$ stages is:*

$$\min_{\mathbf{x}_0, \ldots, \mathbf{x}_T} \quad \sum_{t=0}^{T} \sum_{i=1}^{N} c_{i,t}(\mathbf{x}_{i,0:t}; \hat{\boldsymbol{\xi}}_{0:t}) \tag{3.1a}$$

$$\text{s.t.} \quad f_t(\mathbf{x}_t; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall t \tag{3.1b}$$

$$g_{i,t}(\mathbf{x}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall i,\ \forall t \tag{3.1c}$$

$$h_{i,t}(\mathbf{x}_{i,0:t}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall i,\ \forall t \tag{3.1d}$$

Our formulation contains three types of constraints: (3.1b) convex system-wide constraints f_t that couple decisions across market participants but *within* each stage (e.g., power balance, line flow limits, zonal constraints, reserve requirements); (3.1c) private constraints $g_{i,t}$ for participant i and stage t (e.g., generation limits, state-of-charge (SOC) limits); and (3.1d) private constraints $h_{i,t}$ for participant i coupling their decisions in stage t to all previous dispatches (e.g., ramping, storage SOC updates, unit commitment-dependent generation limits).

This formulation of economic dispatch incorporates linear power flow equations, network constraints, zonal constraints, reserve constraints, private constraints, and

intertemporal constraints for both conventional generators, flexible and inflexible loads, and storage.

Assumption 1. *We assume that for each i, t, functions $c_{i,t}$, f_t, $g_{i,t}$, and $h_{i,t}$ are convex w.r.t $\mathbf{x}_t$. We also assume that they are causal, in the sense that they possibly depend on any dispatches and uncertainty realized until time t. Finally, for non-triviality, we assume that Problem 1 has a feasible solution.*

If market dispatches $\mathbf{x}_0^*, \ldots, \mathbf{x}_T^*$ are generated by the optimal solution of Problem 1, then the market clearing price that supports a competitive equilibrium is just the dual multiplier associated with constraint (3.1b), cf. [139, 57].

Sequential Market Dispatch

In practice, solving Problem 1 is not a viable procedure for clearing the market due to the combination of uncertain inter-stage coupling constraints. Instead, SOs resort to solving a sequence of market-clearing optimization problems.[1] For each stage, updated forecasts of uncertainty are used as problem parameters, and advisory forward decisions are computed, but only the decision for the current stage is settled.

The market-clearing problem for stage t is presented in Problem 2, where the function $V_t : \mathbb{R}^{M_t} \to \mathbb{R}$ represents the forward cost of dispatch $\mathbf{x}_t$; we refer to this as the *forward value* or *cost-to-go* function. As with the functions in Problem 1, V_t may be parameterized by all uncertainty realized up to t, all previous dispatches, as well as forecasts of future uncertainty $\boldsymbol{\theta}_t$ that are available at time t:

$$V_t(\mathbf{x}_t; \hat{\mathbf{x}}_{i,0:t-1}, \hat{\boldsymbol{\xi}}_{0:t}, \boldsymbol{\theta}_t)$$

In service of simpler notation, we make this dependence on parameters implicit in the remainder of the manuscript and simply refer to $V_t(\mathbf{x}_t)$, except where an explicit reference to a particular parameter is necessary. In Section 3.1, we remark on how V_t is already incorporated in market dispatch problems in practice as well as on the theoretical benefits of abstracting the forward cost of decisions in this way.

Problem 2. *Let $\hat{\mathbf{x}}_{0:t-1}$ be the sequence of dispatches committed prior to stage t and $\hat{\boldsymbol{\xi}}_{0:t}$ the uncertainty realized through stage t. The sequential dispatch problem for*

[1] For example, this sequence could be the combination of a day-ahead forward market followed by real-time adjustment market clearings every 15 minutes.

interval t is:

$$\min_{\mathbf{x}_t} \sum_i c_{i,t}(\mathbf{x}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) + V_t(\mathbf{x}_t) \tag{3.2a}$$

$$\text{s.t.} \quad \boldsymbol{\lambda}_t \perp f_t(\mathbf{x}_t; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \tag{3.2b}$$

$$\boldsymbol{\mu}_{i,t} \perp g_{i,t}(\mathbf{x}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall i \tag{3.2c}$$

$$\boldsymbol{\eta}_{i,t} \perp h_{i,t}(\mathbf{x}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall i \tag{3.2d}$$

The dual multipliers associated with each set of constraints are indicated to the left of each constraint (and followed by "$\perp$"). When $V_t(\mathbf{x}_t)$ is convex with respect to $\mathbf{x}_t$, and the convexity conditions from Assumption 1 hold, then (3.2) is a convex optimization problem.

The following algorithm specifies how the system operator clears and settles the market over the multi-stage scheduling horizon. Note that at each stage, the SO requires a scheme for deciding the prices $\boldsymbol{\pi}_t^*$ (see below).

Algorithm 1.

1. *The SO generates a DA uncertainty forecast $\hat{\boldsymbol{\xi}}_0$ and solves Problem 3.2 for $t = 0$ to produce decisions $\mathbf{x}_0^*$ and prices $\boldsymbol{\pi}_0^*$.*
2. *For $t = 1, \ldots, T$:*
   a) *Nature realizes uncertainty $\hat{\boldsymbol{\xi}}_t$;*
   b) *The SO solves Problem 3.2 to produce dispatches $\mathbf{x}_t^*$ and prices $\boldsymbol{\pi}_t^*$;*
   c) *Each participant realizes dispatch $\hat{\mathbf{x}}_t := \mathbf{x}_t^*$ and settles with the SO i at $\boldsymbol{\pi}_{i,t}^{*\top} \hat{\mathbf{x}}_{i,t}$.*

Assumption 2. *Solving Problem 3.2 iteratively for $t = 0, \ldots, T$ produces a feasible sequence of dispatches. Note that such recursive feasibility is in general not guaranteed and may depend on the choice of V_t and $\boldsymbol{\theta}$; see [84, 32] for further consideration of these details.*

Specifying the cost-to-go function V_t

Depending on the parameterization of the uncertainty forecast $\boldsymbol{\theta}_t$ and the choice of the stochastic optimization model, the function V_t adopts different forms. We show below how several common stochastic paradigms fit into this framework. These encompass the multi-settlement and rolling-window optimization procedures (with

and without lookahead) used by SOs in practice as well as stochastic optimization formulations increasingly studied in the research literature.

Rolling dispatch without lookahead

This procedure is the traditional approach to dispatching the DA and RT markets, where each stage (or interval) is optimized without considering the forward consequences of the current dispatch. Thus, $\mathbf{x}_t$ is only coupled intertemporally to $\hat{\mathbf{x}}_{t-1}$ through the constraints (3.2e). In this case, $V_t := 0$ for all $t = 0, \ldots, T$. As this is convex, Problem 3.2 is therefore convex and tractable. Note however, that this choice of V_t may not necessarily lead to the feasibility of the entire dispatch sequence in practice, as is assumed Assumption 2.

Rolling dispatch with lookahead

To better handle uncertainty, SOs make use of forecasts and advisory decisions over a lookahead horizon of length $h > 1$. Exploiting lookahead predictions can increase the feasibility and ex-post optimality of the overall dispatch sequence since it allows for anticipating future ramp, unit commitment, and storage charge/discharge needs [137]. The forecast is a point forecast $\boldsymbol{\theta}_t = (\tilde{\boldsymbol{\xi}}_{t+1}, \ldots, \tilde{\boldsymbol{\xi}}_{t+h})$, available at time t, of the true uncertainties $\hat{\boldsymbol{\xi}}_{t+1}, \ldots, \hat{\boldsymbol{\xi}}_{t+h}$ to be realized.

$$
\begin{aligned}
V_t(\mathbf{x}_t; \boldsymbol{\theta}_t) := \\
&\min_{\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+h}} \sum_{\tau=t+1}^{t+h} \sum_{i=1}^{N} c_{i,\tau}(\mathbf{x}_{i,t+1:\tau}, \hat{\mathbf{x}}_{i,0:t}; \hat{\boldsymbol{\xi}}_{0:t}, \tilde{\boldsymbol{\xi}}_{t+1:\tau}) && && \text{(3.3a)} \\
&\text{s.t.} \quad f_\tau(\mathbf{x}_\tau; \hat{\boldsymbol{\xi}}_{0:t}, \tilde{\boldsymbol{\xi}}_{t+1:\tau}) \leq \mathbf{0} && \forall \tau && \text{(3.3b)} \\
&\qquad g_{i,\tau}(\mathbf{x}_{i,\tau}; \hat{\boldsymbol{\xi}}_{0:t}, \tilde{\boldsymbol{\xi}}_{t+1:\tau}) \leq \mathbf{0} && \forall i, \forall \tau && \text{(3.3c)} \\
&\qquad h_{i,\tau}(\mathbf{x}_{i,t+1:\tau}, \hat{\mathbf{x}}_{i,0:t}; \hat{\boldsymbol{\xi}}_{0:t}, \tilde{\boldsymbol{\xi}}_{t+1:\tau}) \leq \mathbf{0} && \forall i, \forall \tau && \text{(3.3d)}
\end{aligned}
$$

In the above, $\forall \tau$ means $\tau \in [t+1, t+h]$. By convention, if (3.3) is infeasible, $V_t = +\infty$.

Proposition 1. *$V_t(\mathbf{x}_t; \boldsymbol{\theta}_t)$ in (3.3) is convex in $\mathbf{x}_t$.*

For a proof of Proposition 1, see 5.6.1 in [18]. The result is a standard result in perturbation analysis for convex optimization problems.

Although V_t in (3.3) is convex, it is not possible to write down a closed-form solution in general. However, (3.3) can be incorporated into the formulation of

the problem (3.2), recovering the standard lookahead economic dispatch problem studied in [57, 139], which is a tractable convex optimization problem. Note that in a solution $\mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+h}$ to (3.2) with V_t defined as (3.3), only the first dispatch $\mathbf{x}_t$ is binding for the purposes of Algorithm 1. The remaining dispatches $\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+h}$ are advisory and are re-computed for each successive interval.

Chance-constrained optimization

Chance-constrained optimization has received increasing interest for its ability to optimize over decisions with constraints involving stochastic uncertainty [73, 42, 88]. The form of V_t presented next enables probabilistic guarantees on the feasibility of the advisory dispatch under a distributional assumption on uncertainty. At time t, we define p_t to be the distribution of future uncertainty $\boldsymbol{\xi}_{t+1:t+h}$ conditioned on all uncertainty realizations through time t. The forecast $\boldsymbol{\theta}_t$ collects parameters of this distribution or of the SO's best guess of this distribution. In this case, the risk-neutral chance-constrained lookahead value function is defined as follows:

$$V(\mathbf{x}_t; \boldsymbol{\theta}_t) :=$$

$$\min_{\mathbf{x}_{t+1},\ldots,\mathbf{x}_{t+h}} \quad \mathop{\mathbb{E}}_{\boldsymbol{\xi}_{t+1:t+h} \sim p_t} \left[\sum_{\tau=t+1}^{t+h} \sum_{i=1}^{N} c_{i,\tau}(\mathbf{x}_{i,\tau}; \boldsymbol{\xi}_\tau) \right] \tag{3.4a}$$

$$\text{s.t.} \quad \mathbb{P}_{\boldsymbol{\xi}_{t+1:\tau}}[f_\tau(\mathbf{x}_\tau; \hat{\boldsymbol{\xi}}_{0:t}, \boldsymbol{\xi}_{t+1:\tau}) \leq \mathbf{0}] \geq 1 - \varepsilon_\tau^f \qquad \forall \tau \tag{3.4b}$$

$$\mathbb{P}_{\boldsymbol{\xi}_{t+1:\tau}}[g_{i,\tau}(\mathbf{x}_{i,\tau}; \hat{\boldsymbol{\xi}}_{0:t}, \boldsymbol{\xi}_{t+1:\tau}) \leq \mathbf{0}] \geq 1 - \varepsilon_{i,\tau}^g \qquad \forall i, \forall \tau \tag{3.4c}$$

$$\mathbb{P}_{\boldsymbol{\xi}_{t+1:\tau}}[h_{i,\tau}(\mathbf{x}_{i,t+1:\tau}, \hat{\mathbf{x}}_{i,0:t}; \hat{\boldsymbol{\xi}}_{0:t}, \boldsymbol{\xi}_{t+1:\tau}) \leq \mathbf{0}] \geq 1 - \varepsilon_{i,\tau}^h \qquad \forall i, \forall \tau \tag{3.4d}$$

In the above, $\forall \tau$ means $\tau \in [t+1, t+h]$. By convention, if (3.3) is infeasible, $V_t = +\infty$. The hyperparameter ε's can be tuned by the SO to adjust the permissible probability of a constraint violation.

In general, (3.4) is intractable due to the difficulty in computing probabilities and expectations over arbitrary distributions p_t. In particular, the feasible set defined by the constraints may be nonconvex even if the constraint functions $f_\tau, g_{i,\tau}, h_{i,\tau}$ are convex. The structure of the constraints may also make the problem infeasible, e.g., a fixed advisory decision will generally be insufficient to guarantee feasibility under any demand realization, and uncertainty-dependent recourse will be necessary. However, by introducing suitable assumptions on the structure of the problem such as linearity of $c_{i,\tau}, f_\tau, g_{i,\tau}, h_{i,\tau}$, Gaussianity of p_t, separating joint chance constraints into individual chance constraints, and replacing advisory decisions with advisory

uncertainty-dependent affine policies, a tractable, convex counterpart to (3.4) can be formed. For details on such a transformation, we refer the reader to recent literature on chance-constrained optimization and economic dispatch [87, 73].

Other stochastic formulations

The procedure we have been following in Subsections 3.1 – 3.1 to formulate the sequential dispatch problem in the form (2) be applied to other stochastic optimization settings, including scenario-based optimization, robust optimization, and distributionally robust optimization, where there is an extensive literature on convex, tractable reformulations [109, 42, 93, 134, 32].

In fact, although all of these approaches to defining V_t rely on constructing a tractable optimization problem, this is not necessary for Problem 2. As long as V_t is convex and it is possible to obtain gradients of V_t for any input $\mathbf{x}_t$, then optimization (3.2) can be solved using gradient-based methods. And, as we will show in the next section, the price formation also depends only on being able to compute gradients of V_t for the market dispatch.

3.2 Pricing multi-stage uncertainty

In this section, we define the market clearing price and prove that it supports a competitive market clearing solution under *ex-ante* and *ex-post* definitions of dispatch-following incentives.

Model of market participation

In order to establish the properties of a competitive equilibrium, we first present the participant's model of market behavior. We assume that the agents are price-takers, in that they do not bid strategically to impact the price. Further, we assume that they optimize for the current stage of the optimization problem and do not price future decisions into the bid for the current interval. We express the agent's profit-maximizing behavior precisely through the following problem.

Problem 3. *Under a given price $\boldsymbol{\pi}_{i,t}$, agent i would self-schedule itself (denoted $\overline{\mathbf{x}}_{i,t}$) in interval t according to the following problem:*

$$\arg\max_{\overline{\mathbf{x}}_{i,t}} \quad \boldsymbol{\pi}_{i,t}^{\top}\overline{\mathbf{x}}_{i,t} - c_{i,t}(\overline{\mathbf{x}}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) \tag{3.5a}$$

$$\text{s.t.} \quad \overline{\boldsymbol{\mu}}_{i,t} \perp g_{i,t}(\overline{\mathbf{x}}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \tag{3.5b}$$

$$\overline{\boldsymbol{\eta}}_{i,t} \perp h_{i,t}(\overline{\mathbf{x}}_{i,t}; \hat{\mathbf{x}}_{i,0:t-1}, \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \tag{3.5c}$$

Equilibrium Concepts

We are interested in pricing mechanisms that the SO can implement to promote *dispatch-following incentives*. These incentives come in two varieties: *ex-ante*, which apply before uncertainty realization and dispatch, and *ex-post*, which apply after uncertainty has been realized and dispatches have been committed. Adopting terminology from [57, 27], we now present equilibrium notions that will encourage both ex-ante and ex-post dispatch following incentives.

Definition 1. *Let* $\mathbf{x}_0, \ldots, \mathbf{x}_T$ *be a dispatch sequence and* $\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_T$ *be a price sequence, and let* $\hat{\boldsymbol{\xi}}$ *be a realization of uncertainty. This pair of sequences supports a* ***general equilibrium*** *over the entire scheduling horizon* $t = 0, \ldots, T$ *if and only if the following conditions hold:*

1. *Market Clearing Condition. The dispatch sequence satisfies the system-wide constraints at all times:*

$$f_t(\mathbf{x}_t, \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall t \in [0, T]$$

2. *Incentive Compatibility. For each participant* i, $\mathbf{x}_{i,0}, \ldots, \mathbf{x}_{i,T}$ *is an optimal solution of the participant's* ex post *problem:*

$$\underset{\overline{\mathbf{x}}_{i,0}, \ldots, \overline{\mathbf{x}}_{i,T}}{\arg\max} \quad \sum_{t=0}^{T} \boldsymbol{\pi}_{i,t}{}^{\top} \overline{\mathbf{x}}_{i,t} - c_{i,t}(\overline{\mathbf{x}}_{i,t}, \overline{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) \tag{3.6a}$$

$$\text{s.t.} \quad \overline{\boldsymbol{\mu}}_{i,t} \perp g_{i,t}(\overline{\mathbf{x}}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall t \in [0, T] \tag{3.6b}$$

$$\overline{\boldsymbol{\eta}}_{i,t} \perp h_{i,t}(\overline{\mathbf{x}}_{i,t}, \overline{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0} \qquad \forall t \in [0, T] \tag{3.6c}$$

A dispatch and price sequence that supports a general equilibrium supports ex-post dispatch-following incentives. However, when the SO schedules in the presence of uncertainty, e.g. in the case of multi-interval lookahead or stochastic dispatch, a missing payments problem can arise due to distribution shift. The works [57, 27] discuss this issue extensively in the lookahead setting and further show how this missing payment problem arises even when there are perfect forecasts (but a truncated lookahead horizon). To address this, they introduce an additional notion of *partial equilibrium* at each dispatch stage which may be viewed as a condition on *ex-ante* dispatch-following incentives.

Definition 2. *Let $\mathbf{x}_t$ be the dispatch and $\boldsymbol{\pi}_t$ be the price from interval t, and let $\hat{\boldsymbol{\xi}}_{0:t}$ be a realization of uncertainty up through t. This pair supports a **partial equilibrium** for stage t if and only if the following conditions hold:*

1. *Market Clearing Condition:*

$$f_t(\mathbf{x}_t, \hat{\boldsymbol{\xi}}_{0:t}) \leq \mathbf{0}$$

2. *Incentive Compatibility: For each i, the subvector $\mathbf{x}_{i,t}$ of $\mathbf{x}_t$ is the optimal solution of (3.5) under price $\boldsymbol{\pi}_{i,t}$.*

The work in [57, 27] also adopts a dual notion of equilibrium that combines partial and general equilibrium.

Definition 3. *A dispatch sequence $\mathbf{x}_0, \ldots, \mathbf{x}_T$ and price sequence $\boldsymbol{\pi}_0, \ldots, \boldsymbol{\pi}_T$ support a **strong equilibrium** under sequentially realized uncertainty $\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_T$ if and only if they support both a general equilibrium and a partial equilibrium for each t.*

By employing this stronger notion of equilibrium, both ex-ante and ex-post incentive alignment can be guaranteed in the lookahead dispatch setting. We adopt this notion of strong equilibrium in our work to enable pricing that guarantees dispatch-following incentives in the case of general lookahead value function V_t, such as those in the case of stochastic optimization formulations of the market dispatch problem.

Pricing a strong equilibrium

In each interval, the market operator solves (2) to generate a dispatch for that interval for each participant $\mathbf{x}^*_{i,t}$ along with a price vector $\boldsymbol{\pi}^*_{i,t}$ defined as

$$\begin{aligned}\boldsymbol{\pi}^*_{i,t} := &\underbrace{-\mathsf{D}_{\mathbf{x}_{i,t}} f_t(\mathbf{x}^*_t; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\lambda}^*_t}_{\text{Locational marginal price}} \underbrace{- \nabla_{\mathbf{x}_{i,t}} V_t(\mathbf{x}^*_{i,t}; \boldsymbol{\theta}_t)}_{\text{Price of uncertainty}} \\ &\underbrace{-\mathsf{D}_{\mathbf{x}_{i,t}} h_{i,t}(\mathbf{x}^*_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\eta}^*_{i,t}}_{\text{Price of intertemporal coupling}}\end{aligned} \tag{3.7}$$

This price is defined in terms of optimal dual variables and derivatives of objective/constraint functions at the optimal point. The notation $\mathsf{D}_{\mathbf{x}_{i,t}} f_t(\mathbf{x}^*_t; \hat{\boldsymbol{\xi}}_t)$ represents the Jacobian of the function f_t with respect to variable $\mathbf{x}_{i,t}$ evaluated at $\mathbf{x}_{i,t} = \mathbf{x}^*_{i,t}$.

Our price admits a straightforward decomposition into several functional parts. The first component of the price is the standard locational marginal price (LMP). The

second component is a price on the intertemporal coupling between decisions. The price of ramping presented in [57] is a special case of this term; our formulation admits other intertemporal couplings, such as from storage state-of-charge [28]. The last term prices the cost of scheduling under uncertainty. The magnitude of this term is determined both by the particular choice of V_t as well as the quality of the uncertainty parameterization in $\boldsymbol{\theta}_t$. This price is discriminatory, in that each participant may see a different price. The necessity of such price discrimination when there are intertemporal coupling constraints on generators is proven in [57].

We now establish the equilibrium properties of this price. Given the prior convexity assumptions on $c_{i,t}$, f_t, $g_{i,t}$, and $h_{i,t}$, problems (3.5) and (3.6) are convex.

Theorem 5. *Fix a $t \in [0, T]$ and let $\mathbf{x}_t^*$ be the dispatch produced by the optimal solution of (2) and let $\boldsymbol{\pi}_t^*$ be the price as defined in (3.7) using optimal primal/dual variables from (2). This dispatch-price pair forms a partial equilibrium for interval t.*

Proof. For an interval t, we have realized uncertainty $\hat{\boldsymbol{\xi}}_t$ and a previous dispatch sequence $\hat{\mathbf{x}}_{0:t-1}$. Assume that problem (2) has been solved to optimality yielding optimal primal/dual solutions (not necessarily unique) $\mathbf{x}_t^*, \boldsymbol{\lambda}_t^*, \boldsymbol{\mu}_{i,t}^*, \boldsymbol{\eta}_{i,t}^* \; \forall i$.

The market clearing condition in Definition 2 is satisfied by primal feasibility of the optimal solution $\mathbf{x}_t^*$. Without loss of generality, the rest of the proof will be shown for a particular i. To show incentive compatibility, we write down the Lagrangian of (2) for a given t:

$$\begin{aligned}\mathcal{L}_t = &\sum_{i=1}^{N} c_{i,t}(\mathbf{x}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) + V_t(\mathbf{x}_t; \boldsymbol{\theta}_t) + \boldsymbol{\lambda}_t^\top f_t(\mathbf{x}_t; \hat{\boldsymbol{\xi}}_{0:t}) \\ &+ \sum_{i=1}^{N} \boldsymbol{\mu}_{i,t}^\top g_{i,t}(\mathbf{x}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) + \sum_{i=1}^{N} \boldsymbol{\eta}_{i,t}^\top h_{i,t}(\mathbf{x}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t})\end{aligned}$$

The stationarity conditions hold at optimality:

$$\begin{aligned}\mathbf{0} = &\nabla_{\mathbf{x}_{i,t}} c_{i,t}(\mathbf{x}_{i,t}^*, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) + \nabla_{\mathbf{x}_{i,t}} V(\mathbf{x}_t^*; \theta_t) \\ &+ \mathsf{D}_{\mathbf{x}_{i,t}} f_t(\mathbf{x}_t^*; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\lambda}_t^* + \mathsf{D}_{\mathbf{x}_{i,t}} g_{i,t}(\mathbf{x}_{i,t}^*; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\mu}_{i,t}^* \\ &+ \mathsf{D}_{\mathbf{x}_{i,t}} h_{i,t}(\mathbf{x}_{i,t}^*, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\eta}_{i,t}^*\end{aligned} \tag{3.8}$$

The argument uses the convex KKT theorem. We construct primal-dual solutions that satisfy the KKT optimality conditions (primal/dual feasibility, complementary slackness, and stationarity) of problem (3.5). Because (3.5) is convex, the

constructed primal-dual solution is also optimal. Define

$$\overline{\mathbf{x}}_{i,t} := \mathbf{x}^*_{i,t} \tag{3.9a}$$

$$\overline{\boldsymbol{\mu}}_{i,t} := \boldsymbol{\mu}^*_{i,t} \tag{3.9b}$$

$$\overline{\boldsymbol{\eta}}_{i,t} := \mathbf{0} \tag{3.9c}$$

$\overline{\mathbf{x}}_{i,t}$ satisfies primal feasibility of (3.5) because $\mathbf{x}^*_{i,t}$ is primal feasible for (3.2). $\overline{\boldsymbol{\mu}}_{i,t}$ and $\overline{\boldsymbol{\eta}}_{i,t}$ are dual feasible because both are non-negative by construction. Complementary slackness holds for $\overline{\boldsymbol{\mu}}_{i,t}$ because $\boldsymbol{\mu}^*_{i,t}$ is optimal for (2), and holds for $\overline{\boldsymbol{\eta}}_{i,t}$ trivially.

The Lagrangian of (3.5) is

$$\begin{aligned}\mathcal{L}_{i,t} = &-{\boldsymbol{\pi}^*_{i,t}}^\top \overline{\mathbf{x}}_{i,t} + c_{i,t}(\overline{\mathbf{x}}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t}) \\ &+ \overline{\boldsymbol{\mu}}_{i,t}^\top g_{i,t}(\overline{\mathbf{x}}_{i,t}; \hat{\boldsymbol{\xi}}_{0:t}) + \overline{\boldsymbol{\eta}}_{i,t}^\top h_{i,t}(\overline{\mathbf{x}}_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_{0:t})\end{aligned} \tag{3.10}$$

Now to check the stationarity condition,

$$\begin{aligned}\nabla_{\overline{\mathbf{x}}_{i,t}} \mathcal{L}_{i,t} &= -\boldsymbol{\pi}^*_{i,t} + \nabla_{\overline{\mathbf{x}}_{i,t}} c_{i,t}(\mathbf{x}^*_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_t) \\ &\quad + \mathsf{D}_{\overline{\mathbf{x}}_{i,t}} g_{i,t}(\mathbf{x}^*_{i,t}; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\mu}^*_{i,t} + \mathbf{0} \\ &= \mathsf{D}_{\mathbf{x}_{i,t}} f_t(\mathbf{x}^*_t; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\lambda}^*_t + \nabla_{\mathbf{x}_{i,t}} V(\mathbf{x}^*_t; \theta_t) \\ &\quad + \mathsf{D}_{\mathbf{x}_{i,t}} h_{i,t}(\mathbf{x}^*_{i,t}; \hat{\mathbf{x}}_{i,t-1}, \hat{\boldsymbol{\xi}}_t)^\top \boldsymbol{\eta}^*_{i,t} \\ &\quad + \nabla_{\overline{\mathbf{x}}_{i,t}} c_{i,t}(\mathbf{x}^*_{i,t}, \hat{\mathbf{x}}_{i,0:t-1}; \hat{\boldsymbol{\xi}}_t) \\ &\quad + \mathsf{D}_{\overline{\mathbf{x}}_{i,t}} g_{i,t}(\mathbf{x}^*_{i,t}; \hat{\boldsymbol{\xi}}_{0:t})^\top \boldsymbol{\mu}^*_{i,t} \\ &= \mathbf{0}\end{aligned}$$

where the first equality comes by from plugging (3.9) into (3.11) and the second equality comes from plugging in the price defined in (3.7). The third equality holds because the expression is identical to (3.8). □

Theorem 6. *The sequences of dispatches $\mathbf{x}^*_0, \ldots, \mathbf{x}^*_T$ and prices $\boldsymbol{\pi}^*_0, \ldots, \boldsymbol{\pi}^*_T$ produced by Algorithm 1 over the entire scheduling horizon form a general equilibrium.*

Proof. This result uses the same approach as the previous theorem. We construct a primal-dual solution for the individual participant's ex-post LOC problem (3.6) from the primal-dual variables computed over the scheduling horizon with Algorithm 1 and then show that this solution is optimal.

First, the market clearing condition is satisfied because constraints (3.2b) and 3.2b) hold for every t.

The Lagrangian of the individual participant's ex-post LOC problem is (3.6) is

$$\begin{aligned}\mathcal{L}_{i,t} = \sum_{t=0}^{T} & -\boldsymbol{\pi}_{i,t}^{*\top}\overline{\mathbf{x}}_{i,t} + c_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\boldsymbol{\xi}}_t) + \overline{\boldsymbol{\eta}}_{i,t}^{\top} g_{i,t}(\overline{\mathbf{x}}_{i,t};\hat{\boldsymbol{\xi}}_t) \\ & + \overline{\boldsymbol{\nu}}_{i,t}^{\top} h_{i,t}(\overline{\mathbf{x}}_{i,t},\overline{\mathbf{x}}_{i,t-1};\hat{\boldsymbol{\xi}}_t)\end{aligned} \tag{3.11}$$

Define

$$\overline{\mathbf{x}}_{i,t} := \mathbf{x}_{i,t}^{*} \qquad \forall t \tag{3.12a}$$

$$\overline{\boldsymbol{\eta}}_{i,t} := \boldsymbol{\eta}_{i,t}^{*} \qquad \forall t \tag{3.12b}$$

$$\overline{\boldsymbol{\nu}}_{i,t} := \mathbf{0} \qquad \forall t \tag{3.12c}$$

Primal/dual feasibility and complementary slackness follow from the same argument in Theorem 5.

The individual's stationarity condition, which must hold across the entire time horizon

$$\begin{aligned}\nabla_{\overline{\mathbf{x}}_{i,t}}\mathcal{L}_{i,t} &= \sum_{t=0}^{T} -\boldsymbol{\pi}_{i,t}^{*\top} + \nabla_{\overline{\mathbf{x}}_{i,t}} c_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\boldsymbol{\xi}}_t) \\ &\quad + \mathsf{D}_{\overline{\mathbf{x}}_{i,t}} g_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\boldsymbol{\xi}}_t)^{\top}\boldsymbol{\eta}_{i,t}^{*} \\ &\quad + \mathsf{D}_{\overline{\mathbf{x}}_{i,t}} h_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\mathbf{x}}_{i,t-1},\hat{\boldsymbol{\xi}}_t)^{\top}\boldsymbol{\nu}_{i,t}^{*} \\ &= \sum_{t=0}^{T} A_t^{(i)\top}\boldsymbol{\lambda}_t^{*} + \mathsf{D}_{\mathbf{x}_{i,t}} f_t(\mathbf{x}_t^{*};\hat{\boldsymbol{\xi}}_t)^{\top}\boldsymbol{\mu}_t^{*} + \nabla_{\mathbf{x}_{i,t}} V(\mathbf{x}_t^{*};\theta_t) \\ &\quad + \nabla_{\overline{\mathbf{x}}_{i,t}} c_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\boldsymbol{\xi}}_t) + \mathsf{D}_{\overline{\mathbf{x}}_{i,t}} g_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\boldsymbol{\xi}}_t)^{\top}\boldsymbol{\eta}_{i,t}^{*} \\ &\quad + \mathsf{D}_{\mathbf{x}_{i,t}} h_{i,t}(\mathbf{x}_{i,t}^{*};\hat{\mathbf{x}}_{i,t-1},\hat{\boldsymbol{\xi}}_t)^{\top}\boldsymbol{\nu}_{i,t}^{*} \\ &= \mathbf{0}.\end{aligned}$$

□

The last equality holds because equality (3.8) holds for each t. This reveals the motivation of the price construction in (3.7). Including a term for the intertemporal coupling constraints $h_{i,t}$ allows the pricing problem to decouple across intervals. Thus, the participant could leave the market after any interval and their lost opportunity cost would be 0.

The result in Theorem 6 shows that price (3.7) guarantees that each participant has zero lost opportunity cost at the end of the scheduling horizon. The intertemporal

coupling term compensates participants for any lost opportunity cost due to binding intertemporal constraints (e.g., ramping) whereas the uncertainty term compensates participants for any lost opportunity cost due to the system operator's uncertainty-aware scheduling procedure.

The following corollary holds immediately from Theorems 5 and 6:

Corollary 1. *The sequences of dispatches* $\mathbf{x}_0^*, \ldots, \mathbf{x}_T^*$ *and prices* $\boldsymbol{\pi}_0^*, \ldots, \boldsymbol{\pi}_T^*$ *produced by Algorithm 1 over the entire scheduling horizon support a strong equilibrium.*

A strong equilibrium is a desirable property of a market-clearing price because it provides dispatch-following incentives during each stage of scheduling horizon while also correcting the missing payment problem that arises ex-post.

3.3 Experiments

To explore the impact of uncertainty on dispatch efficiency and pricing, we present a simple test case with gas generators, renewables, and load. We consider a power system with a gas combined-cycle (C.C.) plant, a gas peaker plant, solar, wind, and load in a single bus network (no line constraints). The gas plants are ramp constrained whereas the renewables are not. Cost functions are linear and are parameterized by their marginal cost. All parameters for the generators are given in Table 3.1.

Generator	Pmin (MW)	Pmin (MW)	Ramp Rate (% Pmax/hour)	Cost ($/MWh)
Gas C.C.	350	550	25%	50
Gas Peaker	100	120	200%	70
Solar	0	250	NA	0
Wind	0	350	NA	0

Table 3.1: Generator parameters for test case. Ramp constraints on fossil generators limit the speed of their setpoint changes from hour to hour.

We obtained 24 hour load and renewable generation profiles from CAISO from Sep. 9, 2021 [21]. These include both forecast day-ahead trajectories and the actual, realized real-time trajectories, all of which were normalized to 1000MW peak demand. Sample realizations of the true trajectories were simulated by adding zero-mean Gaussian noise to the actual real-time trajectories

Algorithm 1 was implemented to clear the market in a rolling fashion. The dispatch horizon for a single run of the market is 24 hours, consisting of 289 individual stages: one DA dispatch and RT dispatches every 5 minutes.

The first stage ($t = 0$) is the DA unit commitment problem. Due to the non-zero minimum generation of the gas generators, unit commitment is needed to dispatch them efficiently. The unit commitment problem makes use of a 24-hour ahead hourly DA forecast, also obtained from the CAISO data. We fix the unit commitment statuses of the gas generators for each of the hours in the scheduling horizon at their optimal values.[2] The subsequent stages are RT stages and the market is cleared every 5 minutes. We assume that a perfect demand and renewable forecast is available for the current interval.

We implement the three mechanisms discussed in Section 3.1 for dispatching in RT. First is myopic scheduling, where only the current interval's cost and constraints are optimized but generator ramping constraints bind the current decision to the realized dispatch from the previous interval. This is a deterministic problem, as demand and renewable generation are assumed to be known, and does not account for the cost of future decisions in the scheduling horizon. Second is multi-interval lookahead scheduling with a 3-hour lookahead horizon (36 5-min intervals). A real-time point forecast for the lookahead horizon is computed by taking the point-wise mean of the real-time forecast samples. Third is a multi-interval chance-constrained lookahead problem, where the constraints for the advisory periods hold probabilistically and the objective function is the expected cost for the advisory periods. This formulation exploits a probabilistic (vs. point) forecast of uncertainty over the lookahead horizon.

Figure 3.2 shows the dispatch trajectories for each of the generators in the system under optimal *ex-post* scheduling. Note that due to its high cost relative to the other generators, the gas peaker is only active during the peak demand hours when the ramp needs of the system exceed available capacity.

Figure 3.3 presents the benefits of scheduling with lookahead and stochastic forward cost policies. When the forecast error is zero, including forward cost policies is more costly than myopic scheduling. This is due to the inherent conservatism and robustness to uncertainty that these policies provide. However, as uncertainty increases,

[2] In North American ISOs, there is often a financial settlement in the DA market. Although our formulations accommodate a financially settled DA market, we do not empirically analyze the DA market settlement in this work, as intertemporal coupling and uncertainty do not arise in the formation of the DA prices.

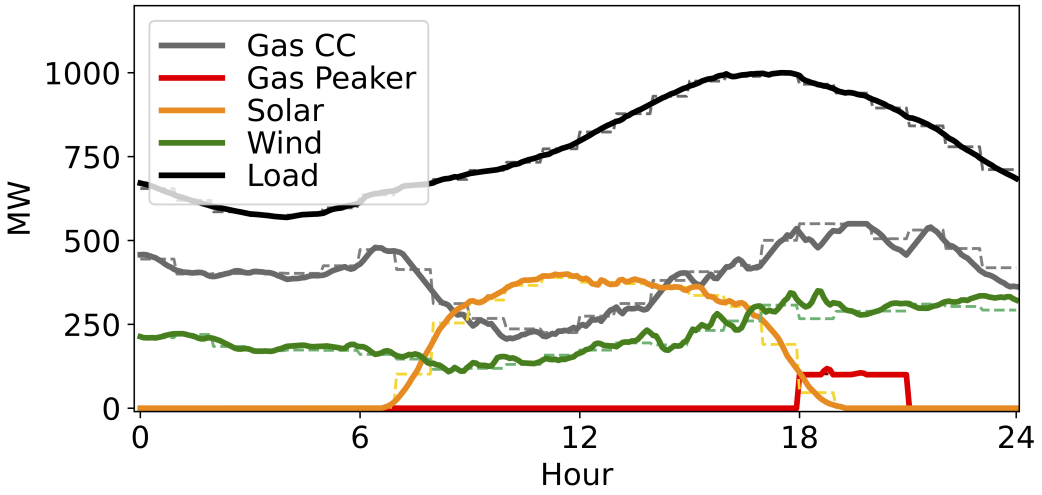


Figure 3.2: DA (dashed line) and optimal RT (solid line) dispatch trajectories for generators and load over a 24 hour scheduling horizon.

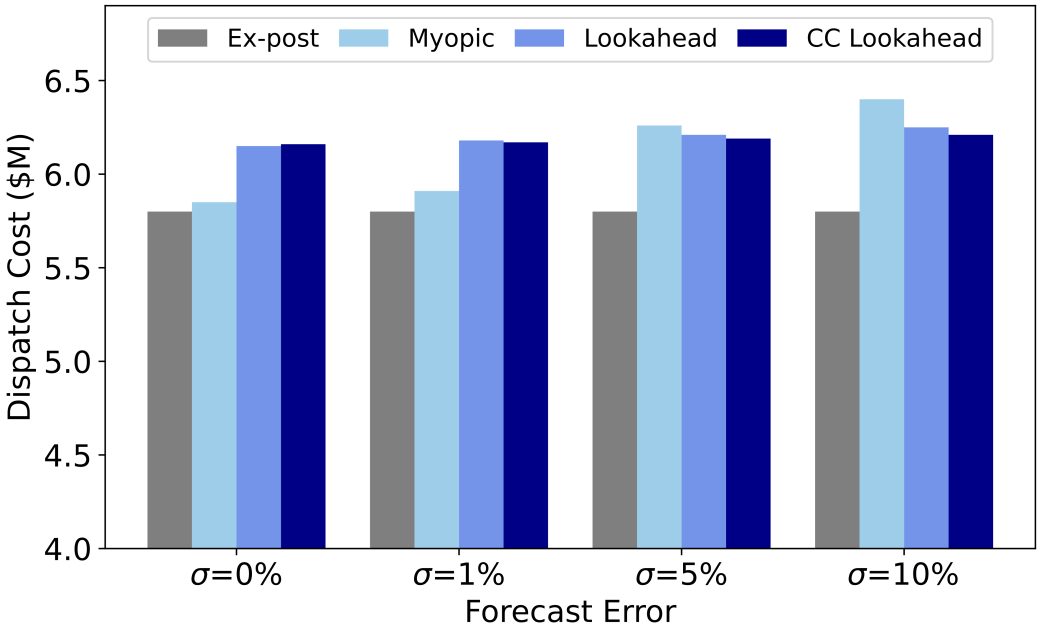


Figure 3.3: Total dispatch cost of the different pricing schemes under increasing forecast error. Forecast error is defined as the mean absolute percentage deviation from the true trajectory realization.

myopic scheduling becomes more costly than uncertainty-aware scheduling due to load shedding actions and sub-optimal dispatch of higher cost generators

Finally, we show how our proposed market clearing price (3.7) decomposes into its constituent components in Figure 3.4. The majority of the price relates to the uniform energy price, which is the shadow price of the power balance constraint. However,

for the ramp-constrained gas generator, there are additional terms that compensate them for the opportunity cost of the system operator's imperfect scheduling under uncertainty.

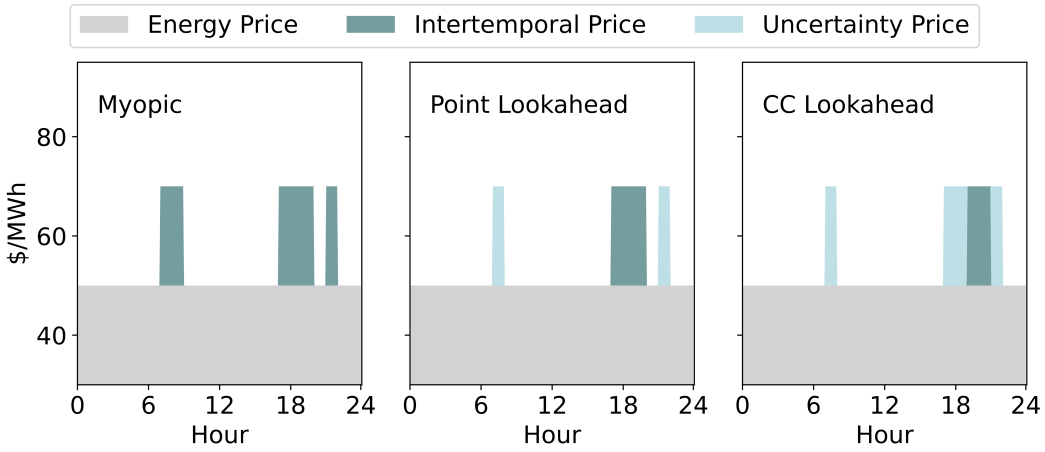


Figure 3.4: Price trajectory $\boldsymbol{\pi}^*_{i,t}$ for the gas combined-cycle generator under $\sigma = 10\%$ forecast uncertainty for different real-time forecast methodologies.

Under high forecast uncertainty, myopic scheduling results in having to dispatch the more expensive gas peaker during ramping events. Lookahead dispatch with point forecasts results dispatching the peaker less often for binding ramping constraints but more during other intervals due to the cost of uncertain dispatch. Chance-constrained lookahead dispatch is able to avoid most of the binding ramping constraints at the expense of more precautionary dispatches due to uncertainty.

3.4 Conclusion

In this chapter, we have presented a mechanism for pricing uncertainty in a multi-stage dispatch setting. Our mechanism builds on recent work in multi-interval pricing and stochastic optimization for economic dispatch. We show how several paradigms for stochastic economic dispatch, including multi-interval dispatch and chance-constrained dispatch, are accommodated in our formulation without needing to modify the price formation. Ongoing work includes examining the impact of our price on the system operator's merchandizing surplus, comparison of multi-settlement and single-settlement pricing methodologies, and empirical comparisons with other pricing methodologies, such as the standard LMP and the R-TLMP proposed in [57].

Acknowledgements: The authors gratefully acknowledge insightful conversations with Subhonmesh Bose (UIUC), Nathan Dahlin (UIUC), and Feng Zhao (ISO NE).

Chapter 4

PRICING FLEXIBILITY OF SHIFTABLE DEMAND IN ELECTRICITY MARKETS

In this chapter, we move to consider pricing mechanisms for load resources in markets. The traditional paradigm of generation following load is being transformed as variable, non-dispatchable resources like solar and wind are an ever increasing share of the generation mix. This creates situations, typically during midday, where there is an excess near-zero marginal cost renewable generation which must be curtailed in order to maintain supply-demand balance. While this scenario might have seemed far-fetched even a few years ago, it is already occurring in major markets. On October 11, 2020, renewables met more that 100% of the total demand in Southern Australia for several hours [111]. In the CAISO (California Independent System Operator) market, solar regularly provides more that 60% of total generation during the afternoon and reached an all-time peak of 80% in May 2019 [105]. Due to a generation queue dominated by renewables and a 100% zero-emission target for 2045, over-generation from renewables will become increasingly common in the California and other large markets [119].

Storage and demand response are two approaches for shifting surplus renewable generation from peak midday hours to periods of higher demand. In an influential 2017 report, researchers at Lawrence Berkeley National Lab analyzed opportunities for demand response and proposed the "Shape-Shift-Shed-Shimmy" taxonomy of flexible loads [74]. They argue that each type of load flexibility is applicable for a particular timescale and use case. Shift flexibility, where the total energy consumed over the time horizon (e.g., 24 hours) remains constant but can be shifted between time intervals, is identified as the form of demand response best suited to accommodate renewable over-generation. Sources of shiftable load include electric vehicle charging, commercial and residential HVAC, and non-time-sensitive industrial processes.

Mechanisms and incentives for offering demand response have been extensively studied but most often they focus on direct compensation for load shedding or peak shaving. Demand response programs implemented by ISOs (Independent System Operators) and utilities tend to be tailored to that same goal. Despite the

calls for attention, mechanisms for Shift flexibility in particular remain relatively understudied [74, 83]. The operational benefits of dispatching shiftable loads are clear to market operators, but as existing markets do not invite significant demand-side participation, from the consumer's point of view the advantages are less clear. This motivates the core questions of this work: *Is a flexible load better off offering its shiftable demand to the market operator than not? And, if not, can we redesign the market to encourage loads to offer shiftable demand to the marketplace?*

Contributions. The answers are No and Yes. We prove that there is incentive misalignment in traditional market designs where flexible loads may prefer not to expose their flexibility to the marketplace. To address this, we introduce a new mechanism where loads have incentives to offer flexibility and generator incentives remain aligned with the social welfare objective. More specifically, this work makes the following contributions.

First, we establish a market and utility model for analyzing shiftable demand. Ours is a variant of the multi-interval market, extensively studied with ramping inequality constraints [58, 140, 69, 65] where equality constraints are added to couple the demand consumption in all periods. Our framework for load utility is derived from the load utility model implied by the standard economic dispatch formulation.

Second, we identify a fundamental incentive incompatibility for loads offering flexible dispatch while being compensated with the standard electricity spot price. We show in Theorem 8 that even in very simple scenarios, loads are worse off under flexible dispatch, even as generators capture more profit and the efficiency of the dispatch solution improves. This counter-intuitive situation arises from the interplay between the time-coupling demand constraint and the power balance constraint that holds in each time interval.

Third, we propose a new multi-interval economic dispatch market that corrects the demand-side incentive incompatibility. Our mechanism preserves core features of the existing structure while making some novel changes: we add inequality constraints to constrain the demand allocation and clear the generation and demand sides of the market separately in a two-step procedure that ensures supply-demand balance and revenue adequacy. Loads that offer flexibility are compensated for deviating from their nominal baseline with a flexibility price, defined in Section 4.3, while inflexible loads continue to pay the baseline spot price for energy. Our main result, Theorem 10, proves that loads have incentives to offer flexibility under the new market design without disturbing dispatch-following incentives for generators.

Finally, in Section 4.4 we present a case study using generation data from CAISO. The case study highlights the importance of ensuring that shiftable loads have incentives to bid their flexibility into the marketplace. Our results show that curtailment of renewable generation can be eliminated, leading to a 15% reduction in the net generation costs.

Related work. This chapter builds on and contributes to three areas of the literature on electricity markets: (1) mechanisms for demand response, (2) multi-interval dispatch, and (3) incentive alignment in mechanism design.

Demand response has been extensively explored in both the academic literature and in practice. In both contexts, interest in demand response has mainly centered around rate-based demand reduction [122, 39, 70, 1] and incentive-based programs [23]. In the former category are time-of-use pricing [37], critical-peak pricing [61], and real-time pricing [5], all schemes which use a given price schedule to incentivize loads to consume energy during lower-cost periods. In the latter group are programs like direct load control [46] and emergency demand reduction [3] in which loads are given lump-sum or per-event payments by the system operator in exchange for the curtailment. Such programs are popular in practice since they lower demand and spot prices during peak load hours.

A drawback of many of these existing designs is that they tend to emphasize a particular variety of demand response—load shedding—and do not explicitly offer incentives for other types of flexible load. The demand response taxonomy in [74] identifies four major types of demand response, each requiring their own dispatch and incentive structures. A general mathematical formulation for optimal dispatch of flexible load is notably given in [83] but the formulation therein assumes knowledge of demand-side value functions. In practice these are very difficult to determine, partly for practical reasons (there are seldom opportunities for loads to reveal their price elasticity) and partly due to historical reasons (electricity has always been treated as an "on-demand" commodity) [72]. We are not aware of any works that formally analyzes incentives for offering shiftable demand—identified as the most significant potential source of demand-side flexibility in [74]—while also retaining the established economic dispatch market structure.

Another important theme in the demand response literature is strategic behavior by loads when reporting their baseline energy consumption. Because demand response almost always defined as a reduction from a baseline, there can be incentives for loads to inflate their baselines to give the appearance of a larger load reduction in

real-time. There are a number of works that analyzed incentives for misreporting and proposed mechanisms to discourage it [24, 132, 90, 31, 95]. While we retain the concept of a baseline in this work for convenience, our model is compatible with schemes to limit the incentives to misreport it, e.g., [96].

Multi-interval markets are of growing interest as a way to guarantee reliable electricity dispatch in the face of uncertain generation. Several substantial works have explored multi-interval market design including [58, 140, 69, 65]. The intertemporal constraints in all of these are limited to ramping limits, which only couple adjacent time periods. In contrast, along the lines of the model proposed in [83], our work considers equality constraints on demand consumption that couple all time periods together. This type of inter-temporal constraint introduces a particular incentive misalignment—a focus of this chapter.

More broadly, our work connects to the topic of mechanism design. Analysis of incentive and participation constraints in market mechanisms was pioneered by Hurwicz, Groves, and Ledyard, among others [56, 75]. The study of incentives in electricity markets has a rich history beginning with the seminal work of Schweppe [118] and has strongly influenced subsequent research on congestion pricing [62, 25] and non-convex pricing [55, 67, 81, 8]. In addition there has been research on market manipulation by generators, e.g., through market power and/or strategic curtailment of renewable generation. Some notable recent results in this direction include [116, 112, 14, 76]. While this body of work establishes a framework for analyzing electricity market incentives, it does so almost exclusively for the generation side of the market [89]. Efficiently dispatching demand-side resources to meet system needs requires similar evaluation of incentive structures.

4.1 Market model

We study an economic dispatch market for energy that the market operator (e.g., ISO/RTO) uses to calculate dispatch quantities and settlement prices. Our model is distinctive from the standard short-term setting in several important ways. First, we consider a multi-interval market with intertemporal equality constraints, which are necessary to model shiftable demand. This contrasts with an existing body of work on multi-interval markets with intertemporal inequality constraints. Second, we explicitly model and dispatch the demand side of the market. Typically demand is taken to be fixed with only generation being variable. Third, we evaluate the welfare of *both* generators and loads in our analysis of incentives. As loads are the

participants providing demand response flexibility, explicitly incorporating them into the social welfare formulation is crucial for quantifying the impacts of flexibility.

Market participants

The market has N generators, indexed by i, and operates over discrete time horizon of length T, indexed by t. The energy produced by generator i in interval t is denoted by $p_{i,t} \in \mathbb{R}$. We denote generators' production over the time horizon with the generation matrix $\mathbf{P} \in \mathbb{R}^{N \times T}$. It is sometimes convenient to refer to individual row/columns of this matrix. The t-th column, the market production vector for time t, is $\mathbf{p}_t = [p_{1,t}, \ldots, p_{N,t}]^\top \in \mathbb{R}^N$. Analogously, the i-th row, generator i's production across the entire time horizon, is denoted $\mathbf{p}_i = [p_{i,1}, \ldots, p_{i,T}]^\top \in \mathbb{R}^T$. Generator cost functions $c_{i,t}(p_{i,t}) : \mathbb{R} \to \mathbb{R}_+$ are assumed to be convex, monotonically increasing, sub-differentiable, and zero-crossing. This last property requires that $0 \in \mathbf{dom}(c_{i,t})$ and $c_{i,t}(0) = 0$. For convenience, we refer to the total cost function for each generator as

$$c_i(\mathbf{p}_i) = \sum_t c_{i,t}(p_{i,t}).$$

The market includes M demand participants, which we refer to as loads, indexed by j. Each load consumes a fixed amount of energy E_j over the T periods. We use $d_{j,t} \in \mathbb{R}$ to denote the the energy consumed by load j in interval t. Like with generators, we stack the $d_{j,t}$ into a demand matrix $\mathbf{D} \in \mathbb{R}^{M \times T}$. We refer to the t-th column with $\mathbf{d}_t = [d_{1,t}, \ldots, d_{M,t}]^\top \in \mathbb{R}^M$ and the j-th row with $\mathbf{d}_j = [d_{j,1}, \ldots, d_{j,T}]^\top \in \mathbb{R}^T$. Loads do not have preference functions that vary with consumption in each time interval. However, they do report a preferred baseline in each interval $d^0_{j,t} \in \mathbb{R}$. E_j is defined in terms of the cumulative baseline consumption of load j:

$$E_j = \sum_t d^0_{j,t}.$$

The use of a baseline is a common assumption in demand response (see e.g., [122]) that we retain here in order to provide a natural definition of flexibility Δ_j as the amount that the actual dispatch $\mathbf{d}_j$ deviates from load j's preferred baseline $\mathbf{d}^0_j$: $\Delta_j := \mathbf{d}_j - \mathbf{d}^0_j$.

Market mechanism

The structure of bids, the market clearing procedure, and the settlement structure are laid out in the following steps:

1. All participants (loads, generators) submit their bids. For loads, this takes the form of a triple
$$(E_j, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j) \in \mathbb{R}_+ \times \mathbb{R}_+^T \times \mathbb{R}_+^T$$
that consists of their energy requirement and lower/upper bounds on consumption in each time period.[1] For generators, the bid takes the form of a pair
$$(c_i, \overline{\mathbf{p}}_i) \in C \times \mathbb{R}_+^T$$
C is the set of all functions $c : \mathbb{R}_+^T \to \mathbb{R}$ that are convex, monotonically increasing, and contain the origin. Generators only submit their upper bounds on production; to avoid non-convexities arising from unit commitment, generation lower bounds are assumed to be 0.

2. The market operator collects bids and solves a market clearing optimization problem, defined in (4.1a) - (4.1e). Its solution provides an allocation of energy to each participant (the dispatch) and a unit price for energy in each time period.

3. Generators are obligated to produce the dispatch quantities and are paid the unit price for whatever they produce. Loads must consume the dispatch quantities and must pay the unit price for whatever they consume. If any participant deviates from the dispatch schedule, the market operator has the ability to administratively penalize the violator, e.g., via large monetary penalties or exclusion from the market.[2]

The centerpiece of the market structure is the *market clearing optimization problem* in Step 2. We study a version of the economic dispatch problem used by ISOs, made distinctive in our case by the multi-interval setting and the inclusion of intertemporal equality constraints. For the sake of focusing our analysis on the impacts of these unique features, we do not consider unit commitment, start-up/no-load costs, and line congestion. We also ignore ramping constraints (i.e., intertemporal inequality constraints) for both loads and generators. As previously mentioned, these have been studied extensively on the generation side of market in [65, 140, 69, 58] among others. Finally, we consider a "single-shot" market-clearing procedure where dispatch quantities and prices are determined at the beginning of the dispatch horizon

[1] As $\mathbf{d}_j$ is assumed to be non-negative, all components of the bid are non-negative.

[2] This requirement reflects the auction design of most North American ISOs, see e.g., Section 2.1 in [33].

and adhered to through the remainder of it.[3] Relaxing these simplifying assumptions is discussed as future work in the Conclusion but we note here that incentive misalignment for loads arises even in the most straightforward setting of the problem.

The market clearing optimization problem is as follows:

$$\min_{\mathbf{p}_j, \mathbf{d}_j \ \forall i,j} \quad \sum_i c_i(\mathbf{p}_i) \tag{4.1a}$$

subject to

$$\lambda_t \ \perp \ \mathbf{1}^\top \mathbf{d}_t - \mathbf{1}^\top \mathbf{p}_t = 0 \qquad \forall t \tag{4.1b}$$

$$\rho_j \ \perp \ \mathbf{1}^\top \mathbf{d}^j = E_j \qquad \forall j \tag{4.1c}$$

$$\mu_i^-, \mu_i^+ \ \perp \ \mathbf{0} \le \mathbf{p}_i \le \overline{\mathbf{p}}_i \qquad \forall i \tag{4.1d}$$

$$\eta_j^-, \eta_j^+ \ \perp \ \underline{\mathbf{d}}_j \le \mathbf{d}_j \le \overline{\mathbf{d}}_j \qquad \forall j \tag{4.1e}$$

In the above, (4.1a) is the total generation cost; (4.1b) are the power balance constraints in each interval; (4.1c) enforces that each load's energy requirement E_j is met over the time horizon (these are the intertemporal equality constraints); and (4.1d) - (4.1e) ensure that the dispatch satisfies participants' minimum and maximum production/consumption limits.

Given an optimal solution to (4.1), the time-varying non-negative energy price π_t is defined for all t as

$$\pi_t := \lambda_t^* \tag{4.2}$$

where λ_t^* is the optimal dual variable for (4.1b).

By offering flexibility in the form of a box constraint on demand as in (4.1e), the efficiency of the dispatch is improved. This is expressed in the following theorem.

Theorem 7. *Let* OPT^0 *be the optimal value of problem (4.1a) - (4.1e) when* $\underline{d}_{j,t} = \overline{d}_{j,t} = d^0_{j,t}$ *for all* j *and* t*, assuming it exists. Let* OPT *be the optimal value of the problem where* $\underline{d}_{j,t} < d^0_{j,t} < \overline{d}_{j,t}$ *for at least one* j *or* t*. Then* $\text{OPT} \le \text{OPT}^0$.

Theorem 7 states that dispatching demand-side flexibility offers benefits to the market allocation in the form of lower cost (greater efficiency). A proof of this result follows immediately from the fact that relaxing constraint (4.1e) results in a large feasible set, which therefore gives a lower bound on the optimal value in

[3] The only assumption needed to support this is that the $\mathbf{d}_j^0$ are known at $t = 1$ and do not adjust over the course of the time horizon.

the case where the constraint is tight. The existence of OPT is guaranteed by the existence of OPT^0.

Notice that the formulation of economic dispatch in (4.1) reduces to the standard setting (T independent sequential economic dispatch problems) when $\underline{\mathbf{d}}_j = \mathbf{d}_j = \overline{\mathbf{d}}_j$. In this situation, constraints (4.1c) and (4.1e) are redundant and $\mathbf{d}_t$ in (4.1b) can be replaced by $\mathbf{d}_t^0$.

Utility models for generators and loads

An important goal of this work is to evaluate whether the market allocation, given by the optimal primal solution of (4.1), and the market-clearing price, given by the optimal dual variable of (4.1b), are aligned with the individual incentives. To study this question we need to introduce definitions of utility and the individual utility maximization problem for both loads and generators.

Let $\pi \in \mathbb{R}_+^T$ be the vector of market-clearing energy prices for the time horizon. We assume all agents are price takers and define the following utility models for generators and loads respectively.

Definition 4. *Let $\mathbf{p}_i \in \mathbb{R}^T$ be generator i's production vector and $\mathcal{P}_i$ be its private constraint set*

$$\mathcal{P}_i = \left\{ \mathbf{p}_i \in \mathbb{R}^T \;\middle|\; \mathbf{0} \leq \mathbf{p}_i \leq \overline{\mathbf{p}}_i \right\} \subseteq \mathbb{R}^T.$$

Generator i's utility is defined as

$$u_i\left(\mathbf{p}_i; \pi\right) := \pi^\top \mathbf{p}_i - c_i(\mathbf{p}_i). \tag{4.3}$$

We assume a generator acts rationally when facing the given price schedule π and therefore seeks to maximize its utility with

$$\begin{aligned} \arg\max_{\mathbf{p}_i} \quad & u_i\left(\mathbf{p}_i; \pi\right) \\ \text{s.t.} \quad & \mathbf{p}_i \in \mathcal{P}_i. \end{aligned} \tag{4.4}$$

In contrast to generators, loads do not have a cost function and are only constrained by a required amount of energy to be delivered over the time horizon, E_j. Instead we assume that there is a constant utility value $U_j \in \mathbb{R}_+$ that represents the value a load receives from having E_j satisfied. We assume that the load is indifferent to how energy is allocated across the intervals, as long as E_j is delivered and upper/lower consumption limits are respected.

Definition 5. *Let $\mathbf{d}_j \in \mathbb{R}^T$ be load j's consumption vector and $\mathcal{D}_j$ be its private constraint set,*

$$\mathcal{D}_j = \left\{ \mathbf{d}_j \in \mathbb{R}^T \;\middle|\; \underline{\mathbf{d}}_j \leq \mathbf{d}_j \leq \overline{\mathbf{d}}_j, \quad \mathbf{1}^\top \mathbf{d}_j = E_j \right\} \subseteq \mathbb{R}^T.$$

Load j's utility is defined as

$$u_j\left(\mathbf{d}_j; \pi\right) = U_j - \pi^\top \mathbf{d}_j. \tag{4.5}$$

We again assume each load acts rationally when facing the given price schedule π and therefore seeks to maximize its utility with

$$\begin{aligned} \arg\max_{\mathbf{d}_j} \quad & u_j(\mathbf{d}_j; \pi) \\ \text{s.t.} \quad & \mathbf{d}_j \in \mathcal{D}_j. \end{aligned} \tag{4.6}$$

A feature of this presentation of utility that deserves comment and justification is our representation of the positive "value" component of load utility with a constant U_j. This choice is made to align with the classical auction-based economic dispatch model in Section 4.1. Specifically, if we use utility functions (4.3) and (4.5) to construct the market's social welfare maximization problem subject to a shared market clearing constraint and private feasibility constraints, we get exactly the auction-based economic dispatch model (i.e., cost minimization) of the market described by (4.1). To see this, recall that the market's social welfare maximization problem is

$$\begin{aligned} \max_{\mathbf{p}_i, \mathbf{d}_j \; \forall i,j} \quad & \sum_i u_i(\mathbf{p}; \pi) + \sum_j u_j(\mathbf{d}_j; \pi) & \\ \text{s.t.} \quad & \mathbf{1}^\top \mathbf{p}_t = \mathbf{1}^\top \mathbf{d}_t & \forall t \\ & \mathbf{p}_i \in \mathcal{P}_i & \forall i \\ & \mathbf{d}_j \in \mathcal{D}_j & \forall j \end{aligned} \tag{4.7}$$

At an optimal point $\mathbf{p}_i^*, \mathbf{d}_j^* \; \forall i, j$ of (4.7),

$$\begin{aligned} & \mathbf{1}^\top \mathbf{p}_t^* - \mathbf{1}^\top \mathbf{d}_t^* = 0 \quad \forall t \\ \Rightarrow \; & \sum_i \mathbf{p}_i - \sum_j \mathbf{d}_j = \mathbf{0} \end{aligned}$$

Plugging in (4.3) and (4.5) into the objective function of (4.7), we get

$$\begin{aligned}\sum_i u_i(\mathbf{p};\pi) + \sum_j u_j(\mathbf{d}_j;\pi) &= \sum_i \pi^\top \mathbf{p}_i - c_i(\mathbf{p}_i) + \sum_j U_j - \pi^\top \mathbf{d}_j \\ &= \sum_j U_j - \sum_i c_i(\mathbf{p}_i) + \pi^\top \Big(\sum_i \mathbf{p}_i - \sum_j \mathbf{d}_j \Big) \\ &= \sum_j U_j - \sum_i c_i(\mathbf{p}_i)\end{aligned}$$

It is clear that the objective function of (4.7) differs from (4.1a) by only a constant factor and the constraint sets of the two problems are identical. Therefore, they have the same optimal solution (although the optimal value differs by a factor of $\sum_j U_j$). While the choice of U_j does not impact the optimal solution, intuitively, it should be a positive number, large enough so that $U_j - \pi^\top \mathbf{d}_j > 0$ for most realizations of π and $\mathbf{d}_j$. However this condition is not necessary for our analysis of prices and dispatch quantities.

4.2 Participation incentives

Our first set of results focuses on understanding the consequences of dispatching flexibility via the classical market formulation described in the previous section. We show in this section that, though dispatch-following incentives for generators remain intact, participation incentives for loads are misaligned and offering flexibility (i.e., $\underline{\mathbf{d}}_j < \overline{\mathbf{d}}_j$) to the market operator is not necessarily rational.

Participation incentives for loads

Participation constraints affect a rational agent's behavior. In particular, given a choice to enter into a market/mechanism or not, it is expected that a rational agent only does so if their utility is higher under participation than their best alternative. To put this precisely for the case of loads in our model, let $\mathbf{d}_j^0$ be the allocation a load receives outside of the flexibility mechanism (i.e., the load simply consumes its reported baseline). Let $\mathbf{d}'_j$ be the allocation a load receives by participating in the mechanism. j's participation constraint is satisfied if and only if $u_j(\mathbf{d}'_j) \geq u_j(\mathbf{d}_j^0)$.

Once a participant submits its bid to the market operator, it is obliged to obey the dispatch instruction that comes in return when the market is cleared. We show in the following theorem that, depending on the market outcome, loads can end up worse off by offering flexibility under the energy price in (4.2), despite the increase in efficiency that flexibility offers to the market as a whole (established in Theorem 7).

Theorem 8. *Assume the baseline solution $\mathbf{d}_j^0 \, \forall j$ is feasible for (4.1a) - (4.1e). Under the market dispatch (4.1a) - (4.1e) and energy price π_t given in (4.2), participation constraints for loads are are not guaranteed. That is, there exist choices of parameters $c_i(\cdot), \mathbf{d}_j^0, \overline{\mathbf{p}}_i, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j$ with $\underline{d}_{j,t} < \overline{d}_{j,t}$ for some j, t such that $u_j(\mathbf{d}'_j) < u_j(\mathbf{d}_j^0)$ for some j.*

Proof. Our proof takes the form of a counterexample. Consider a market environment with 2 time periods: $t = 1, 2$. There is a single load with demand given by $\mathbf{d} = [d_1, d_2]^\top$ and a single generator with generation given by $\mathbf{p} = [p_1, p_2]^\top$. The unit generation costs are $\mathbf{c} = [1, 2]^\top$ and the baseline demand is $\mathbf{d}^0 = [2, 2]^\top$. Thus $E = 4$. Generation is constrained by $\underline{\mathbf{p}} = [0, 0]^\top, \overline{\mathbf{p}} = [3, 3]^\top$, and demand by $\underline{\mathbf{d}} = \mathbf{d}^0(1 - \alpha) \leq \mathbf{d}^0(1 + \alpha) = \overline{\mathbf{d}}$, where $\alpha \in [0, 1]$. We parameterize the demand lower/upper bounds with the constant α to allow us to vary the offered flexibility between 0 ($\alpha = 0$) and its maximum ($\alpha = 1$).[4]

Market dispatch model (4.1a) - (4.1e) with these parameters gives the following optimization problem:

$$\begin{aligned} \min_{p,d} \quad & c^\top \mathbf{p} \\ \text{s.t.} \quad & \lambda \perp \mathbf{p} = \mathbf{d} \\ & \mathbf{1}^\top \mathbf{d} = 4 \\ & \mathbf{0} \leq \mathbf{p} \leq 3 \cdot \mathbf{1} \\ & 2(1 - \alpha) \cdot \mathbf{1} \leq \mathbf{d} \leq 2(1 + \alpha) \cdot \mathbf{1}. \end{aligned} \tag{4.8}$$

By (4.2) the energy price vector is $\pi = \lambda^*$. We assume the value constant for the load is 0 and take the optimal solution of (4.8) to be $(\mathbf{p}', \mathbf{d}', \lambda')$, we have the following form of load utility:

$$u(\mathbf{d}') = -\lambda'^\top \mathbf{d}'.$$

We solve (4.8) for $\alpha \in [0, 1]$ and compute demand utility $u(\mathbf{d}'; \alpha)$. Since α parameterizes the "amount" of flexibility demand offers, increasing values of α correspond to greater demand flexibility (looser bounds on min/max consumption in each interval). The results are shown in Figure 4.1.

[4] $\alpha = 1$ is the maximum because demand cannot be negative. The upper bound does not have the same restriction as the lower one but we stick to a single parameter here for simplicity.

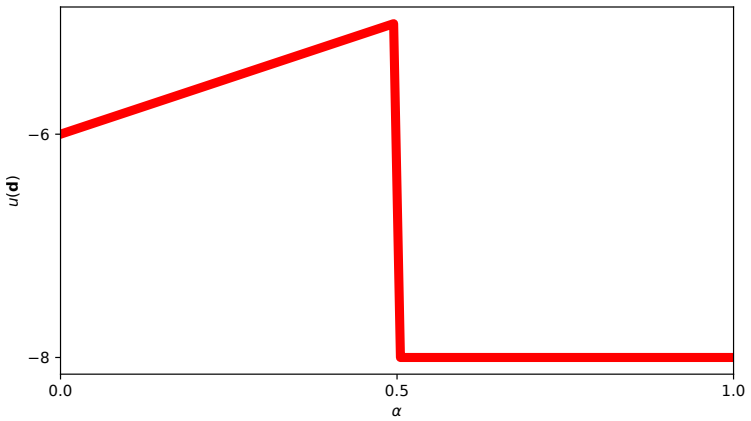


Figure 4.1: Demand utility vs. $\alpha \in [0, 1]$

Maximum consumer utility of -5 is reached as $\alpha \uparrow 0.5$ and $u(\mathbf{d}^0) \geq u(\mathbf{d})$ only for $\alpha < 0.5$. That is, the demand is worse off by offering for $\alpha \geq 0.5$ than none at all. In fact, if we had chosen the parameters differently (e.g., d^0 arbitrarily close to $\overline{p}$ in one interval), the demand participation constraint is violated for all $\alpha > 0$. □

Remark: *We retain the standard price-taking assumption in the above proof. With a single generator, this may be practically unrealistic. At the expense of greater complexity additional generators can be considered without changing the qualitative behavior we highlight. The purpose of the proof is to demonstrate that incentive violations arise even in the simplest of market settings.*

Analyzing which generation constraints in (4.1d) bind as α varies in the counterexample above gives insight into how misaligned incentives for loads come about.

Analogously to how a marginal generator $\hat{\imath}$ can be defined in the single-period economic dispatch, we define a marginal pair: generator *and* interval $(\hat{\imath}, \hat{t})$. If $\overline{p}_{\hat{\imath},\hat{t}} - \mathbf{1}^\top \mathbf{d}^0_{\hat{t}} \geq 0$, then flexible demand can shift to $\hat{t}$ from costlier intervals to take advantage of this excess supply without changing the price $\lambda_{\hat{t}}$. However, once the upper bound is exceeded for the marginal pair (i.e., $\mathbf{1}^\top \mathbf{d}^0_{\hat{t}} > \overline{p}_{\hat{\imath},\hat{t}}$), $\lambda^*_{\hat{t}}$ jumps up to the marginal cost of the next cheapest marginal pair. This surprising behavior occurs because time periods are coupled together through constraint (4.1c) Adding constraints on $\mathbf{d}_j$ that prevent this jump motivates the mechanism proposed in Section 4.3.

Participation incentives for generators

The previous section addresses the incentive misalignment for loads under the standard market structure. One may worry that a similar misalignment happens for

generators. In this section we show that there is no such issue on the generator side of the market, i.e., that utility-maximizing decisions of the generators exactly match the dispatch decision by the market operator.

Specifically, the following theorem states that the optimal solution of the market dispatch problem (4.1a) - (4.1e) provides dispatch following incentives to generators, provided we treat generators as pricetakers. Along the way, we also show that generators do not have negative profit (i.e., participation constraints are satisfied). Throughout, we make the standard assumption that (4.1a) - (4.1e) has a feasible point.

Theorem 9. *Let $\mathbf{p}_i^*, \mathbf{d}_j^* \; \forall i, j$ be the optimal primal solutions of (4.1a) - (4.1e). The energy prices are $\pi := \lambda^*$ where λ^* is the vector of optimal dual variables for constraint (4.1b). Then*

$$\begin{aligned} \mathbf{p}_i^* = \arg\max_{\mathbf{p}_i} \quad & u_i(\mathbf{p}_i; \pi) \\ s.t. \quad & \mathbf{p}_i \in \mathcal{P}_i \end{aligned}$$

Further, $u_i(\mathbf{p}_i^*) \geq 0$.

Proof. Start by forming the Lagrangian for (4.1a) - (4.1e):

$$\begin{aligned} \mathscr{L}(\mathbf{p}_i, \mathbf{d}_j; \lambda, \rho_j, \mu_i^{\pm}, \eta_j^{\pm}) = & \sum_i c_i(\mathbf{p}_i) \\ & + \sum_t \lambda_t(\mathbf{1}^\top \mathbf{d}_t - \mathbf{1}^\top \mathbf{p}_t) + \sum_j \rho_j \mathbf{1}^\top \mathbf{d}_j \\ & + \mu_i^{+\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i) - \mu_i^{-\top}(\mathbf{p}_i - \underline{\mathbf{p}}_i) \\ & + \eta_j^{+\top}(\mathbf{d}_j - \overline{\mathbf{d}}_j) - \eta_j^{-\top}(\mathbf{d}_j - \underline{\mathbf{d}}_j) \end{aligned}$$

We assume that (4.1a) - (4.1e) has a feasible point. Let $(\mathbf{p}_i^*, \mathbf{d}_j^*, \lambda^*, \rho_j^*, \mu_i^{\pm *}, \eta_i^{\pm *})$ denote its optimal solution, which exists because the c_i are continuous and the feasible set is compact. Compactness follows from constraint (4.1d) where it can be seen that all entries of $\mathbf{p}_i^*$ and $\mathbf{d}_j^*$ must be finite. Strong duality holds because all constraints are affine and the objective function is convex (see e.g., Prop. 5.3.1 in [12]).

Therefore the following KKT stationarity condition for holds for every i:

$$\frac{\partial \mathscr{L}}{\partial \mathbf{p}_i}(\mathbf{p}_i^*) = \nabla c_i(\mathbf{p}_i^*) - \lambda^* + \mu_i^{+*} - \mu_i^{-*} = \mathbf{0} \tag{4.9}$$

We compute the derivative of the λ term by noting that

$$\frac{\partial}{\partial p_{i,t}} \sum_t \lambda_t (\mathbf{1}^\top \mathbf{p}_t - \mathbf{1}^\top \mathbf{d}_t) = \lambda_t$$

Stacking this equation for each t into vector form gives

$$\frac{\partial}{\partial \mathbf{p}_i} \sum_t \lambda_t (\mathbf{1}^\top \mathbf{p}_t - \mathbf{1}^\top \mathbf{d}_t) = \lambda.$$

The price vector, as defined in (4.2), is $\pi = \lambda^*$. Thus (4.9) is

$$\nabla c_i(\mathbf{p}_i^*) - \pi + \mu_i^{+*} - \mu_i^{-*} = \mathbf{0} \tag{4.10}$$

Next, we rewrite (4.4) equivalently as a minimization of $-u_i$ over the same feasible set and take its Lagrangian.[5]

$$\mathscr{L}_i(\mathbf{p}_i; \mu_i^\pm) = c_i(\mathbf{p}_i) - \pi^\top \mathbf{p}_i + \mu_i^{+\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i) - \mu_i^{-\top}(\mathbf{p}_i - \overline{\mathbf{p}}_i)$$

The KKT stationarity condition is

$$\frac{\partial \mathscr{L}_i}{\partial \mathbf{p}_i} = \nabla c_i(\mathbf{p}_i) - \pi + \mu_i^+ - \mu_i^- = \mathbf{0} \tag{4.11}$$

It is clear that $\mathbf{p}_i = \mathbf{p}_i^*$, $\pi = \lambda^*$, $\mu_i^\pm = \mu_i^{\pm *}$ is a solution to (4.11) because $(\mathbf{p}_i^*, \lambda^*, \mu_i^{\pm *})$ satisfies (4.10).

Now we show that $\mathbf{p}_i^*$ satisfies participation constraints. Outside of the mechanism, the generator would produce $\mathbf{p}_i = \mathbf{0}$ with $u_i(\mathbf{0}) = 0$. This is because we assumed that $c_i(\mathbf{0}) = 0$. We need to show that 0 is a lower bound for $u_i(\mathbf{p}_i^*)$.

In (4.10), when $\mu_{i,t}^{+\,*} > 0$ then $\mu_{i,t}^{-\,*} = 0$ as only one of the lower/upper bounds can be attained at a time. But if $\mu_{i,t}^{-\,*} > 0$ then $p_{i,t}^* = 0$. When $\mu_{i,t}^{+\,*} = \mu_{i,t}^{-\,*} = 0$, then $p_{i,t}^* > 0$ and $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^*) = \pi_t$. Therefore $\pi_t p_{i,t}^* - c_{i,t}(p_{i,t}^*) = 0$. Finally, when $\mu_{i,t}^{+\,*} > 0$, then $\mu_{i,t}^{-\,*} = 0$ and $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p_{i,t}^*) < \pi_t$. So $\pi_t p_{i,t}^* - c_{i,t}(p_{i,t}^*) > 0$. In each of these three situations we get that $u_i(\mathbf{p}_i^*) \geq 0$.

□

Note that this theorem extends a well-known result for single-period economic dispatch to our multi-interval setting with equality constraints.

[5]Note that we use the same names for primal/dual variables in the individual problem as in the market dispatch problem. Although these variables *do not* represent the same quantities, we hope this is not cause for confusion.

4.3 Incentivizing flexibility

Section 4.2 highlights that loads have an incentive not to reveal their flexibility under the standard market design where only the energy price is used for settlement. This is problematic since exploiting the flexibility of loads is essential for system reliability, avoiding curtailment of renewable energy, and improving the economic efficiency of the dispatch. This section presents the main contribution of the chapter: a new market design that ensures both loads and generation have incentives that are aligned with the market operator's and, specifically, provides incentives for loads to reveal their flexibility to the market. First we introduce the market design and prove its incentive properties and following, in Section 4.4, we illustrate the market design using a case study.

A market design for flexibility

Our proposed design adopts a similar structure to the standard market while introducing three important components: (1) a small number of additional constraints on the demand allocation, (2) a time-varying price κ_t for flexibility, and (3) a two-stage market clearing scheme for the demand side of the market.

Before presenting the mechanism we must introduce some notation. First, let the constant

$$c_{\min} := \min_{i,t} \frac{\partial c_{i,t}}{\partial p_{i,t}}(p^0_{i,t})$$

be the smallest marginal cost (over all i and t) under the baseline allocation. Second, define $\mathcal{T} \subseteq \{1, \dots, T\}$ to be the subset of intervals for which it is true that $\frac{\partial c_{i,t}}{\partial p_{i,t}}(p^0_{i,t}) = c_{\min}$ for at least one $i \in \{1, \dots, N\}$. $\mathcal{T}^{\mathsf{c}}$ is the set of all intervals that do not meet this condition. Together, $\mathcal{T} \dot\cup \mathcal{T}^{\mathsf{c}} = \{1, \dots, T\}$.[6] In what follows, we assume that neither $\mathcal{T}$ and $\mathcal{T}^{\mathsf{c}}$ is empty. Third, for each $t \in \mathcal{T}$, define a generator index set

$$\mathcal{I}_t := \{i \mid \frac{\partial c_{i,t}}{\partial p_{i,t}}(p^0_{i,t}) = c_{\min}\} \subseteq \{1, \dots, N\}.$$

Fourth, define

$$P^{\text{cap}}_t := \sum_{i \in \mathcal{I}_t} p^{\text{cap}}_{i,t}$$

[6] In the real-world scenario of renewable curtailment, $c_{\min} = 0$ (since marginal cost of renewables is taken to be 0) and $\mathcal{T}$ is simply the set of intervals for which renewables are curtailed.

where

$$p_{i,t}^{\text{cap}} = \arg\max_{p \in \mathbb{R}} \; p$$
$$\text{s.t.} \quad \frac{\partial c_{i,t}}{\partial p_{i,t}}(p) = c_{\min}$$
$$p \leq \overline{p}_{i,t}$$

This (regrettably heavy) notation makes precise the amount of available *excess* capacity at the lowest price $c_{\min}$ in the baseline dispatch. Observe that when the c_i are linear, $p_{i,t}^{\text{cap}} = \overline{p}_{i,t}$ for $i \in \mathcal{I}_t$.

With this notation in hand, we summarize the structure of the market mechanism. Additional discussion of each step is provided following the exposition of the procedure.

1. Generators submit bids $(c_i, \overline{\mathbf{p}}_i)$ and loads submit bids $(\mathbf{d}_j^0, \underline{\mathbf{d}}_j, \overline{\mathbf{d}}_j)$ to the market operator.

2. Market operator collects bids, forms the market-clearing optimization problem (4.1a) - (4.1e) with the additional constraint $\mathbf{d}_j = \mathbf{d}_j^0$, and produces a baseline solution $(\mathbf{p}_i^0, \mathbf{d}_j^0, \pi^0)$ $\forall i, j$.

3. Market operator re-solves (4.1a) - (4.1e) with the addition of three new constraints:

$$\mathbf{1}^\top \mathbf{d}_t \leq P_t^{\text{cap}} \quad \forall t \in \mathcal{T} \tag{4.12a}$$
$$d_{j,t} \geq d_{j,t}^0 \quad \forall j, \; \forall t \in \mathcal{T} \tag{4.12b}$$
$$d_{j,t} \leq d_{j,t}^0 \quad \forall j, \; \forall t \in \mathcal{T}^{\mathsf{c}} \tag{4.12c}$$

   An interim solution and prices are computed: $(\tilde{\mathbf{p}}_i, \tilde{\mathbf{d}}_j, \tilde{\pi})$ $\forall i, j$.

4. The market operator defines a flexibility price

$$\kappa = \kappa(\mathbf{d}_j^0, \tilde{\mathbf{d}}_j, \pi^0, \tilde{\pi}) \in \mathbb{R}^T \tag{4.13}$$

   as a function of optimal solutions of the two market clearing problems. (We discuss the precise form of κ in Section 4.3.)

5. The market operator solves a demand dispatch problem (4.14a) - (4.14f), producing a final allocation for demand: $\mathbf{d}_j^*$ $\forall j$.

6. Generators are dispatched to produce energy $\tilde{\mathbf{p}}_i$ at price $\tilde{\pi}$. Loads are dispatched to consume energy $\mathbf{d}_j^*$ at price π^0 and contribute flexibility Δ_j^* compensated with price κ.

We now walk through the steps in more detail, beginning with Step 2. Step 2 establishes a baseline allocation that is used later in the procedure to ensure that participation constraints are satisfied.

In Step 3, additional inequalities (4.12a) - (4.12c) constrain the demand dispatch to a desirable region. (4.12a) enforces that the total demand does not exceed the total maximum capacity of the cheapest generator(s) in the interval—provided that there is spare capacity under the baseline solution. (4.12b) ensures that demand can only *increase* if there is excess cheapest generation in a period. (4.12c) guarantees that demand can only *decrease* during intervals where all of the cheapest generation is already dispatched. These additional linear inequalities only add $|\mathcal{T}|+T$ constraints to the market dispatch problem, which already has $(1 + M + N)T + M$ constraints. Due to the assumption that $\mathcal{T}$ and $\mathcal{T}^{\mathsf{c}}$ are non-empty, a solution to (4.12a) - (4.12c) exists: namely $\mathbf{D}^0$.

Step 4 defines a flexibility unit price κ. The definition of a flexibility price is central to our proposed mechanism. Rather than enforce a specific price function, here we present properties that a price of flexibility should satisfy. Later in Section 4.3 we provide examples that satisfy the given properties. Before introducing them, we first we define the concept of a *flexibility surplus*.

Definition 6. *The flexibility surplus* $S := \sum_t (\pi_t^0 - \tilde{\pi}_t)\mathbf{1}^\top \tilde{\mathbf{d}}_t$ *is the difference between the total demand-side energy payment if demand were paying baseline energy price* π_t^0 *and the total demand-side payment when demand pays the price,* $\tilde{\pi}_t$*. Because of Lemma 12 (below) and constraint (4.1c),* $S \geq 0$*. We interpret S as the improvement in welfare (over the baseline) of the demand side of the market as a whole when the dispatch* $\tilde{\mathbf{d}}_t$ *optimally utilizes flexibility.*

Now we establish properties that should be satisfied by a flexibility price κ:

- κ is uniform (each load faces the same κ)
- $\kappa_t \geq 0$ when $\sum_j \Delta_{j,t} \geq 0$ and $\kappa_t \leq 0$ when $\sum_j \Delta_{j,t} \leq 0$. This means that the payment for both up and down flexibility is non-negative, as at least some flexibility in both directions is necessary to dispatch shiftable demand.

- The sum of all flexibility payments equals the flexibility surplus: $\sum_t \kappa_t \mathbf{1}^\top (\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) = S$.

This last property is natural, as our scheme distributes the surplus arising from the increased economic efficiency of the flexibility dispatch to the loads that provide this flexibility. Another desirable property we seek when constructing κ_t is that its magnitude should reflect the value of flexibility to the system in interval t.

Step 5 maximizes social welfare for the demand side of the market given flexibility price κ and energy price π^0. In order to construct this welfare maximization problem, we need to update the definition of demand utility with a term that quantifies the benefit that comes from offering flexibility.

Definition 7. *Let $\mathbf{d}_j \in \mathbb{R}^T$ be load j's consumption vector and κ and π be the flexibility and energy price vectors, respectively. $\mathbf{d}_j^0$ is the load's reported baseline. Then demand utility is given by*

$$u_j^*(\mathbf{d}_j; \pi, \kappa, \mathbf{d}_j^0) = U_j - \pi^\top \mathbf{d}_j + \kappa^\top (\mathbf{d}_j - \mathbf{d}_j^0)$$

Next, we solve a demand allocation optimization where the total demand dispatch amount in each interval is fixed to be equal to the total interim demand dispatch from Step 3. This allows the settlement for the generation side of the market to remain unaffected by the redistribution on the demand side.

$$\max_{\mathbf{d}_j \ \forall j} \quad \sum_j u_j^*(\mathbf{d}_j) \tag{4.14a}$$

subject to

$$\mathbf{1}^\top \mathbf{d}_j = E_j \qquad \forall j \tag{4.14b}$$

$$\underline{\mathbf{d}}_j \le \mathbf{d}_j \le \overline{\mathbf{d}}_j \qquad \forall j \tag{4.14c}$$

$$d_{j,t} \ge d_{j,t}^0 \qquad \forall j,\ \forall t \in \mathcal{T} \tag{4.14d}$$

$$d_{j,t} \le d_{j,t}^0 \qquad \forall j,\ \forall t \in \mathcal{T}^\mathsf{c} \tag{4.14e}$$

$$\mathbf{1}^\top \mathbf{d}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t \qquad \forall t \tag{4.14f}$$

The optimal solution of the above problem ${\mathbf{d}_j}^*$ determines the actual consumption of load j over the horizon.

Finally, Step 6 settles the market with $(\tilde{\mathbf{p}}_i \ \forall i, \tilde{\pi})$ for generators and $(\mathbf{d}_j^* \ \forall j, \pi^0, \kappa)$ for loads. Load j *pays* ${\pi^0}^\top \mathbf{d}_j^*$ for energy because it is the price it *would* have payed in the baseline scenario. The load *receives* $\kappa^\top \Delta_j^*$ for deviating by Δ_j^* from its baseline.

Analyzing participation incentives

The following theorem establishes properties for both generator and load utility under the proposed market mechanism and settlement scheme. We show that incentives are aligned on both sides of the market.

Theorem 10. *Let* $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*)$ *be the energy and flexibility allocation from the market mechanism. Let* $(\tilde{\pi}, \pi^0, \kappa)$ *be the corresponding energy and flexibility prices. Then*

(i) $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*)$ *clears the market;*

(ii) $(\tilde{\mathbf{p}}_i, \mathbf{d}_j^*, \Delta_j^*, \tilde{\pi}, \pi^0, \kappa)$ *is revenue neutral for the market operator;*

(iii) $(\tilde{\mathbf{p}}_i, \tilde{\pi})$ *provides dispatch-following incentives for generators and satisfies their participation constraints;*

(iv) $(\mathbf{d}_j^*, \Delta_j^*, \pi^0, \kappa)$ *satisfies participation constraints for loads. Specifically, for each* j*,*

$$u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \geq u_j(\mathbf{d}_j^0; \pi^0);$$

(v) For j *for which* $\Delta_{j,t}^* = 0$ *for all* t *(no flexibility offered),*

$$u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \leq u_j(\tilde{\mathbf{d}}_j; \tilde{\pi});$$

For j *for which* $|\Delta_{j,t}^*| > 0$ *for some* t *(flexibility offered),*

$$\sum_j u_j^*(\mathbf{d}_j^*; \pi^0, \kappa, \mathbf{d}_j^0) \geq \sum_j u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).$$

Statement (v) is of particular importance and highlights that loads are better off offering flexibility than not: no load becomes worse off than at its baseline consumption but loads that do offer flexibility are (weakly) better off as a group than those that do not.

Before proving this theorem, we present and prove two lemmas. Let $\mathbf{d}^0$ and $\mathbf{p}^0$ be the optimal primal solutions to (4.1a) - (4.1e) in the baseline case (i.e., $\underline{\mathbf{d}} = \overline{\mathbf{d}}$), and let λ^0 be the associated optimal dual variable for (4.1b). In the presence of flexible demand, that is, $\underline{d}_{j,t} < \overline{d}_{j,t}$ for some t and j, we get the following result.

Lemma 11. *Let* $\tilde{\mathbf{d}}_j$ *for all* j *be optimal solutions of (4.1a) - (4.1e) with added constraints (4.12a - 4.12c). Then for all* j

$$u_j(\tilde{\mathbf{d}}_j) \geq u_j(\mathbf{d}_j^0).$$

The proof of this claim requires another technical lemma, which we state and prove below before returning to the proof of Lemma 11. To proceed, we associate dual variables $\beta \in \mathbb{R}^{|\mathcal{T}|}$, $\gamma^+ \in \mathbb{R}^{|\mathcal{T}|}$, $\gamma^- \in \mathbb{R}^{|\mathcal{T}^c|}$ with constraints (4.12a) - (4.12c) respectively.

Lemma 12. *$\lambda_t^0 \geq \tilde{\lambda}_t$ for $t \in \mathcal{T}^c$.*

Proof of Lemma 12. Let $(\tilde{\mathbf{P}}, \tilde{\mathbf{D}}, \tilde{\lambda}, \tilde{\rho}, \tilde{\mu}^{\pm}, \tilde{\eta}^{\pm}, \tilde{\beta}, \tilde{\gamma}^{\pm})$ be the optimal primal/dual solution of (4.1a) - (4.1e) with added constraints (4.12a) - (4.12c). The arguments for the existence of this solution and the existence of strong duality are the same as those given in the proof of Theorem 9 (see Appendix **??**).[7]

For each t, there is a set of marginal generators $\mathcal{N}_t \subseteq \{1, \ldots, n\}$. By its definition, a marginal unit produces strictly between its upper and lower bounds. Therefore, for $i \in \mathcal{N}_t$

$$\tilde{\mu}_{i,t}^+ = \tilde{\mu}_{i,t}^- = \mathbf{0}.$$

The KKT stationarity condition w.r.t. $\mathbf{p}_t$ is

$$\frac{\partial \mathscr{L}}{\partial \mathbf{p}_t}(\tilde{\mathbf{p}}_t) = \nabla c(\tilde{\mathbf{p}}_t) - \tilde{\lambda}_t \mathbf{1} + \tilde{\mu}_t^+ - \tilde{\mu}_t^- = \mathbf{0}. \tag{4.15}$$

From this equation we have that **for all** $i \in \mathcal{N}_t$,

$$\tilde{\lambda}_t = \frac{\partial c_{i,t}}{\partial p_{i,t}} \left(\tilde{p}_{i,t}\right). \tag{4.16}$$

That is, all marginal costs are equal in that time for the marginal units.

Next, we claim that no generator produces more than its baseline; that is $\tilde{p}_{i,t} \leq p_{i,t}^0$ for all $i \in \{1, \ldots, n\}$. In the baseline scenario there are three groups of generators: those producing at their upper bound, those producing at their lower bound, and the marginal units. Those already at their upper bound are unable to increase their production. Increasing the production of a unit at its lower bound would incur a higher cost than increasing production by the same amount for a marginal unit. In (4.16) we argued that all marginal units have the same marginal cost at the optimal point. Therefore they all would increase production or all decrease. Due to the convexity and monotonicity of the cost functions, a decrease in production would result in a lower value for λ_t by (4.16).

[7] We assumed previously that a baseline solution exists for (4.1a) - (4.1e). A feasible point for (4.1a - 4.1e) with added constraints (4.12a - 4.12c) is just this baseline solution.

This leaves two remaining possibilities: 1) at least one generator at its upper bound in the baseline case decreases its production or 2) a marginal unit decreases its production. In the first case, λ_t is unaffected since its value is determined by the cost function of a marginal unit. In the second case, due to the convexity and monotonicity of the cost functions, a decrease in production would result in a smaller value of λ_t by (4.16).

For all $t \in \mathcal{T}^{\mathsf{c}}$, $\tilde{d}_{j,t} \leq d^0_{j,t}$ for each j by constraints (4.12c). Thus we have $\mathbf{1}^\top \tilde{\mathbf{d}}_t \leq \mathbf{1}^\top \mathbf{d}^0_t$ for $t \in \mathcal{T}^{\mathsf{c}}$. By the power balance constraint (4.1b),

$$\mathbf{1}^\top \tilde{\mathbf{p}}_t \leq \mathbf{1}^\top \mathbf{p}^0_t.$$

From this we conclude that for at least one $i \in \mathcal{N}_t$, $p^0_{i,t} \geq \tilde{p}_{i,t}$. Since we take the cost functions to be convex and if $p^0_{i,t} \geq \tilde{p}_{i,t}\ \forall i$ as we showed just above, then

$$\frac{\partial c_{i,t}}{\partial p_{i,t}}(p^0_{i,t}) \geq \frac{\partial c_{i,t}}{\partial p_{i,t}}(\tilde{p}_{i,t}) \tag{4.17}$$

We conclude that

$$\lambda^0_t \geq \tilde{\lambda}_t.$$

□

Proof of Lemma 11. We take the aggregate marginal cost curve of (4.1a) - (4.1e) to be left continuous. This is equivalent to always picking the smallest value of the subgradient of $\sum_i c_i(\mathbf{p}_i)$ in the KKT condition for $\mathbf{p}_i$ when the subgradient is not unique.

For $t \in \mathcal{T}$, the following are true:

- $\tilde{d}_{j,t} \geq d^0_{j,t}$ for all j by primal feasibility from constraint (4.12b);
- $\lambda^0_t = \tilde{\lambda}_t = c_{\min}$ because of (4.12c). By definition of $\mathcal{T}$, $c_{\min}$ is always the marginal cost in $\mathcal{T}$. Note that this claim requires the assumption from the beginning of the proof. Otherwise, when constraint (4.12a) is tight, it could occur that $\tilde{\lambda}_t > \lambda^0_t$.

Analogously for $t \in \mathcal{T}^{\mathsf{c}}$ we have:

- $\tilde{d}_{j,t} \leq d^0_{j,t}$ for all j by primal feasibility from constraint (4.12c);

- $\tilde{\lambda}_t \leq \lambda_t^0$ by Lemma 12.

By definition of the price π, $\pi^0 \geq \tilde{\lambda}$ follows from the above. We have also established that $\tilde{d}_{j,t} \geq d_{j,t}^0$ only when $\tilde{\lambda}_t = \lambda_t^0 = c_{\min}$. Otherwise, $d_{j,t} \leq d_{j,t}^0$.

By definition of load utility in (4.5), we have that

$$\begin{aligned} u_j(\tilde{\mathbf{d}}_j) &= U_j - \sum_{t \in \mathcal{T}} \tilde{\pi}_t \tilde{d}_{j,t} - \sum_{t \in \mathcal{T}^c} \tilde{\pi}_t \tilde{d}_{j,t} \\ &= U_j - \sum_{t \in \mathcal{T}} c_{\min} \tilde{d}_{j,t} - \sum_{t \in \mathcal{T}^c} \tilde{\pi}_t \tilde{d}_{j,t} \\ &\geq U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \\ &= u_j(\mathbf{d}_j^0). \end{aligned}$$

□

Proof of Theorem 10. We prove each statement from the theorem in order, making use of the two technical lemmas just presented.

(i) First note that $\tilde{\mathbf{D}}$ is a feasible solution for (4.14b) - (4.14f). The optimal solution $\mathbf{D}^*$ of (4.14a) - (4.14f) satisfies $\mathbf{1}^\top \mathbf{d}_t^* = \mathbf{1}^\top \tilde{\mathbf{d}}_t$ for all t due to primal feasibility.[8] Similarly, the generation dispatch $\tilde{\mathbf{P}}$ satisfies $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t$ for all t by constraint (4.1b) and primal feasibility. Therefore $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \mathbf{d}_t^*$ for all t. Note that this is equivalent to $\sum_i \tilde{\mathbf{p}}_i = \sum_j \mathbf{d}_j^*$.

(ii) Total generation revenue is given by

$$\text{Rev}_{\text{gen}} = \sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{p}}_t = \sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{d}}_t$$

Total energy payments from demand are

$$\text{Pay}_{\text{demand}} = \sum_t \pi_t^0 \mathbf{1}^\top \mathbf{d}_t^* = \sum_t \pi_t^0 \mathbf{1}^\top \tilde{\mathbf{d}}_t$$

Total flexibility payments to loads are

$$\begin{aligned} \text{Rev}_{\text{flex}} &= \kappa^\top \sum_j \Delta_j^* \\ &= \sum_t \kappa_t \mathbf{1}^\top (\mathbf{d}_t^* - \mathbf{d}_t^0) \\ &= S \end{aligned}$$

[8] A feasible solution exists: observe that $\tilde{\mathbf{D}}$ satisfies constraints (4.14b - 4.14f). An optimal value is attained because the objective function is continuous and the feasible set is compact.

Revenue neutrality is the condition when

$$\text{Pay}_{\text{demand}} - \text{Rev}_{\text{gen}} = \text{Rev}_{\text{flex}}.$$

By the definition of S, this condition is satisfied.

(iii) The proof of this result follows exactly the one for Theorem 9. The KKT stationarity condition for $\mathbf{p}_i$ is unaffected by the addition of constraints (4.12a) - (4.12c).

(iv) By primal feasibility of $\mathbf{d}_j^*$ in (4.14a - 4.14f), $d_{j,t}^* \geq d_{j,t}^0$ for $t \in \mathcal{T}$ and $d_{j,t}^* \leq d_{j,t}^0$ for $t \in \mathcal{T}^{\mathsf{c}}$. In the proof of Lemma 11 (see Appendix **??**) we showed that $\pi_t^0 = \lambda_t^0 = c_{\min}$ for $t \in \mathcal{T}$ and $c_{\min} < \pi_t^0$ for all $t \in \mathcal{T}^{\mathsf{c}}$. As a consequence,

$$\sum_{t\in\mathcal{T}} \pi_t^0 d_{j,t}^* - \sum_{t\in\mathcal{T}} \pi_t^0 d_{j,t}^0 = c_{\min} \sum_{t\in\mathcal{T}} (d_{j,t}^* - d_{j,t}^0)$$

By primal feasibility of $\mathbf{d}_j^*$ and $\mathbf{d}_j^0$ from constraints (4.1c) and (4.14b) we have that

$$\begin{aligned}
&\sum_t d_{j,t}^* = E_j = \sum_t d_{j,t}^0 \\
\Rightarrow &\sum_{t\in\mathcal{T}} d_{j,t}^* + \sum_{t\in\mathcal{T}^{\mathsf{c}}} d_{j,t}^* = \sum_{t\in\mathcal{T}} d_{j,t}^0 + \sum_{t\in\mathcal{T}^{\mathsf{c}}} d_{j,t}^0 \\
\Rightarrow &\sum_{t\in\mathcal{T}} (d_{j,t}^* - d_{j,t}^0) = -\sum_{t\in\mathcal{T}^{\mathsf{c}}} (d_{j,t}^* - d_{j,t}^0)
\end{aligned}$$

Then

$$\begin{aligned}
c_{\min} \sum_{t\in\mathcal{T}} (d_{j,t}^* - d_{j,t}^0) &= -c_{\min} \sum_{t\in\mathcal{T}^{\mathsf{c}}} (d_{j,t}^* - d_{j,t}^0) \\
&\geq -\sum_{t\in\mathcal{T}^{\mathsf{c}}} \pi_t^0 (d_{j,t}^* - d_{j,t}^0)
\end{aligned}$$

By construction of κ, $\kappa_t \Delta_t^j \geq 0 \; \forall j, t$. Putting everything together,

$$\begin{aligned}
u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) &= U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* + \sum_t \kappa_t \Delta_{j,t}^* \\
&\geq U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* \\
&= U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^* - \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^* \\
&\quad + \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&\quad - \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&= U_j - \sum_{t \in \mathcal{T}} c_{\min} (d_{j,t}^* - d_{j,t}^0) - \sum_{t \in \mathcal{T}^c} \pi_t^0 (d_{j,t}^* - d_{j,t}^0) \\
&\quad - \left(\sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \right) \\
&\geq U_j - \sum_{t \in \mathcal{T}} c_{\min} d_{j,t}^0 + \sum_{t \in \mathcal{T}^c} \pi_t^0 d_{j,t}^0 \\
&= U_j - \sum_t \pi_t^0 d_t^{j,0} \\
&= u_j(\mathbf{d}^0; \pi^0)
\end{aligned}$$

(v) Let $\mathcal{M}_{\text{flex}} \subseteq \{1, \ldots, M\}$ be the index set of loads whose flexibility is dispatched. Complementarily, $\mathcal{M}_{\text{flex}}^c \subseteq \{1, \ldots, M\}$ is the index set of loads whose flexibility is *not* dispatched. For $j \in \mathcal{M}_{\text{flex}}^c$,

$$\begin{aligned}
u_j^*(\mathbf{d}_j^*; \pi^0) &= U_j - {\pi^0}^\top \mathbf{d}_j^* \\
&\leq U_j - \tilde{\pi}^\top \tilde{\mathbf{d}}_j \\
&= u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).
\end{aligned} \tag{4.18}$$

The inequality comes from the fact proved in Lemma 12.

Finally we show the second statement in (v):

$$\begin{aligned}
\sum_j u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) &= \sum_{j \in \mathcal{M}_{\text{flex}}} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) + \sum_{j \in \mathcal{M}_{\text{flex}}^{\mathsf{c}}} u_j^*(\mathbf{d}_j^*; \pi^0) \\
&= \sum_j U_j - \text{Pay}_{\text{demand}} + \text{Rev}_{\text{flex}} \\
&= \sum_j U_j - \text{Rev}_{\text{gen}} \\
&= \sum_j U_j - \sum_j \tilde{\pi}^\top \tilde{\mathbf{d}}_j \\
&= \sum_{j \in \mathcal{M}_{\text{flex}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}) + \sum_{j \in \mathcal{M}_{\text{flex}}^{\mathsf{c}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi})
\end{aligned}$$

From (4.18) we have that $\sum_{j \in \mathcal{M}_{\text{flex}}^{\mathsf{c}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}) \geq \sum_{j \in \mathcal{M}_{\text{flex}}^{\mathsf{c}}} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa)$. The sequence of equalities implies

$$\sum_{j \in \mathcal{M}_{\text{flex}}} u_j^*(\mathbf{d}_j^*; \pi^0, \kappa) \geq \sum_{j \in \mathcal{M}_{\text{flex}}} u_j(\tilde{\mathbf{d}}_j; \tilde{\pi}).$$

□

A price for flexibility

A core piece of our proposed market design is the flexibility price κ. How to properly compensate demand for flexibility is a challenging open question. Flexibility, as defined in this work, is a public good: in the interim, energy price-only settlement $(\tilde{\mathbf{P}}, \tilde{\mathbf{D}}, \tilde{\pi})$, even those loads who do not offer relaxed bounds on their consumption (i.e., offer flexibility to the market) still benefit from others that do by enjoying a lower price. To address this, our mechanism directly pays flexible loads that based on how much of the flexibility they offer is dispatched. We proceed in two stages: first we define a flexibility price κ that satisfies certain desirable properties (Step 4 in the mechanism); and second, we compute an allocation of energy and flexibility that maximizes individual utility while also respecting the previously-determined generation dispatch (Step 5 in the mechanism).

To this point, we have made the second stage concrete with (4.14a) - (4.14f) but we have not yet given specific examples of flexibility prices that satisfy the desirable properties of κ listed in Step 4 above. In this section we propose three different flexibility prices, commenting on their relative advantages. An interesting future research direction is to explore other forms of this price.

Optimization-based

Our first approachis to directly solve an optimization problem with the properties listed in Step 4 as constraints.

$$\begin{aligned} \min_{\kappa \in \mathbb{R}^T} \quad & f(\kappa) \\ \text{s.t.} \quad & S = \sum_t \kappa_t \mathbf{1}^\top (\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) \\ & \kappa_t \geq 0 \qquad\qquad \forall t \in \mathcal{T} \\ & \kappa_t \leq 0 \qquad\qquad \forall t \in \mathcal{T}^{\mathsf{c}} \end{aligned}$$

A benefit of this approach is that the choice of an objective function $f(\kappa)$ can be made in order to enforce desired structural properties. For example, setting $f(\kappa) = ||\kappa||_2$ yields a smooth price schedule. If prices that weight high-value time-periods more are desired, then one could have $f(\kappa) = ||\kappa||_1$.[9]

The adaptability of this formulation of κ is its main advantage. One potential disadvantage is that it does not yield a closed-form representation of κ in general, which could make the price difficult to interpret. The subsequent two designs we consider provide closed-form representations of κ.

Budget-balance

A contrasting formulation of κ is based on the market operator's budget balance condition:

$$\sum_t \tilde{\pi}_t \mathbf{1}^\top \tilde{\mathbf{p}}_t = \sum_t \pi_t^0 \mathbf{1}^\top \mathbf{d}_t^* - \sum_t \kappa_t \mathbf{1}^\top \left(\mathbf{d}_t^* - \mathbf{d}_t^0 \right) \tag{4.19}$$

This condition states that the total payments to generators equals the total energy payments from demand minus the total flexibility payments to loads. Noting that $\mathbf{1}^\top \tilde{\mathbf{p}}_t = \mathbf{1}^\top \tilde{\mathbf{d}}_t = \mathbf{1}^\top \mathbf{d}_t^*$ and enforcing equality for each t separately, we solve for (4.19) for κ_t to get

$$\kappa_t := \frac{\left(\pi_t^0 - \tilde{\pi}_t\right) \mathbf{1}^\top \mathbf{d}_t^*}{\mathbf{1}^\top \left(\mathbf{d}_t^* - \mathbf{d}_t^0 \right)}.$$

This form of κ_t satisfies our desired properties but it does have an important drawback. When $t \in \mathcal{T}$, $\tilde{\pi}_t = \pi_t^0 = c_{\min}$, which implies that $\kappa_t = 0$. So, κ_t is never strictly positive, which leads to only down-flexibility ($\Delta_t < 0$) being rewarded. Further, for

[9] If $f(\kappa)$ is a norm, then this formulation has the additional property that $\kappa_t = 0$ only if $\mathbf{1}^\top (\tilde{\mathbf{d}}_t - \mathbf{d}_t^0) = 0$. This means that intervals that do not dispatch flexibility will not have a non-zero price.

$t \in \mathcal{T}$, $\kappa_t = 0$ and so the prices do not capture the time-varying value of flexibility for those intervals. The following design avoids these disadvantages.

Flexibility surplus

Another closed-form version of κ_t can be defined using the flexibility surplus:

$$\kappa_t = \begin{cases} \frac{S}{2} \frac{P_t^{\text{cap}} - \mathbf{1}^\top \mathbf{d}_t^0}{\sum_t P_t^{\text{cap}} - \mathbf{1}^\top \mathbf{d}_t^0} \frac{1}{\mathbf{1}^\top \left(\tilde{\mathbf{d}}_t - \mathbf{d}_t^0\right)}, & t \in \mathcal{T} \\ \frac{S}{2} \frac{\tilde{\pi}_t}{\sum_t \tilde{\pi}_t} \frac{1}{\mathbf{1}^\top \left(\tilde{\mathbf{d}}_t^- \mathbf{d}_t^0\right)}, & t \in \mathcal{T}^{\mathsf{c}} \end{cases}$$

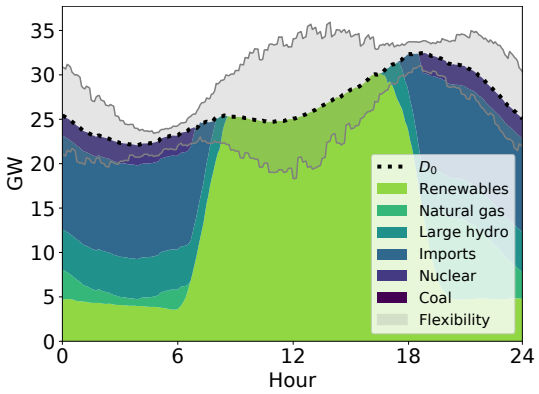
This form of κ_t is the product of three terms in both cases. The first, $\frac{S}{2}$, divides the total flexibility surplus evenly between up- and down-flexibility periods. The second term distributes that half-surplus amongst the time intervals. For $t \in \mathcal{T}$, an interval receives an amount proportional to its surplus (i.e., curtailed) lowest-cost generation. For $t \in \mathcal{T}^{\mathsf{c}}$, an interval receives the amount proportional to the interim price $\tilde{\pi}_t$in that time period. The third term divides by the total allocation of flexibility, as determined by the dispatch from (4.1a) - (4.1e) with (4.12a) - (4.12c).

Like the previous two flexibility prices, this κ satisfies all of the desired properties including budget balance. Its two-part specification reflects the different function flexibility has in $\mathcal{T}$ versus $\mathcal{T}^{\mathsf{c}}$. In $\mathcal{T}$, flexibility allows otherwise-curtailed low-cost generation to be dispatched. In $\mathcal{T}^{\mathsf{c}}$, flexibility allows for lesser amounts of more-costly generation to be dispatched. While this formulation addresses the zero-price shortcoming of the budget-balance formulation and has a closed form representation, it is vulnerable to volatility when $\mathbf{1}^\top \mathbf{d}_t^* - \mathbf{1}^\top \mathbf{d}_t^0$ is small.

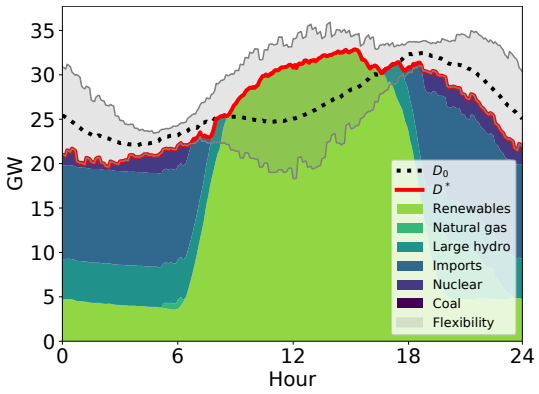
Comparing these three formulations, it is worth highlighting that, while a closed-form κ might be desirable for reasons convenience and interpretability, the optimized-based approach is more principled and adaptable. For this reason we choose to implement that version of the flexibility price in the case study in the next section.

4.4 Case study

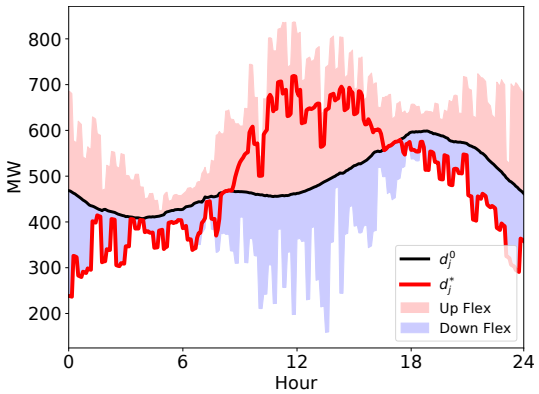
We conclude the chapter with a demonstration of the our new market dispatch of flexibility on a test case derived from the real-world CAISO market. Our numerical results show a significant increase in utility for loads when they allow their flexibility to be dispatched by the market operator, thus highlighting the value of redesigning the market to ensure participation incentives of loads are aligned.



(a) Baseline dispatch (demand does not offer flexibility)



(b) Flexible dispatch (demand offers flexibility)



(c) Individual load profile for $j = 29$: baseline (black) and flexible dispatch (red)

Figure 4.2: Comparison of the baseline market with the proposed market design in a CAISO case study.

Setup

Disaggregated demand-side data for bulk electricity markets is not readily available [52]. We therefore take existing publicly-available generation and aggregate load data from CAISO and simulate a demand side to to the market. Our simulations are implemented in Python and all optimization problems are solved with `CVXPY` [40, 4]. The simulations were run on a 2019 MacbookPro (2.8 GHz Quad-Core i7, 16GB RAM).

Throughout our experiments, we ran the single-shot market mechanism described in Section 4.3, which assumes an accurate demand forecast, and computed the flexibility price κ using the optimization-based formulation.

The test cases are constructed as follows. Generation time series data, disaggregated by resource type (e.g., renewable, hydro, coal), from July 2, 2020 is obtained from [20]. The data have observations every 5 minutes for 24 hours (288 total). At their peak, renewables (e.g., wind, solar, small hydro, biomass) account for approx. 60% of the net generation. We clean the data by removing trivial generation resources like batteries and negative values for solar generation at night (due to concentrating solar); the result is 6 generation resource types: renewables, natural gas, large hydro, nuclear, coal, and external imports from adjacent control areas. The aggregate demand D_0 profile is obtained from the resulting net generation. We scale up the entire renewable profile by 220% so that there is a set of intervals $\mathcal{T}$ where renewable generation alone exceeds aggregate demand and thus renewables must be curtailed. As we noted previously, this scenario is not (yet) the case in California but in other markets has already begun occurring [111].

We assume that conventional generation types and imports are dispatchable up and down without ramping limits, whereas renewables can only be curtailed. We also make the simplifying assumption that conventional generation can produce any amount from 0 to their upper limits, which are taken from the original data to be the maximum production at any point in the 24 hour window. Unit cost data in \$/MWh are the Variable O&M costs for 2020 from EIA's Annual Energy Outlook, see Table 1 in [35]. Unit costs for imports were assumed to be the average of costs for the other generation types present in our simulation.

The aggregate demand profile from the CAISO data (black dotted line in Figure 4.2a) is disaggregated proportionally into individual load profiles. These profiles are then perturbed with random noise to introduce temporal variability to the relative fraction of the aggregate each individual load consumes. The number of individual

loads m can be set arbitrarily and in our case study here, $m = 30$.[10] Half of these were designated inflexible loads and the other half to flexible loads. Centered around each of the individual flexible load profiles are upper and lower bound profiles for the consumption of each load in each time interval. These bounds are generated with a sinusoidal function which allows parametric scaling of flexibility by varying the amplitude and phase. We note here that despite not being able to access real-world load profiles, our load disaggregation scheme produces qualitatively similar results to the load shapes in [45]. The baseline load profile (black) and the flexibility range (grey) are shown for the market in aggregate in Figure 4.2a and for an individual load in Figure 4.2c.

Results

Figure 4.2 provides a detailed contrast between the traditional baseline market design, under which shiftable demands do not offer their flexibility, and the proposed design of this work, under which shiftable demands have incentives to expose their flexibility. The reduction of curtailment of renewable generation that results from shiftable demands is immediately clear from these figures.

In more detail, Figure 4.2a shows the generation dispatch as well as the baseline aggregate demand. The available aggregate flexibility is shown as a light grey overlay. Notice that renewables are curtailed between hours 8 and 17, as there is an excess supply available to meet the baseline aggregate demand D_0.

Figure 4.2a should be contrasted with Figure 4.2b, which shows the market dispatch (aggregate shown in red) when flexible demand is utilized. The flexibility upper/lower bounds (grey) and the baseline aggregate demand (dotted black) are superimposed for comparison. Load is dispatched *up* in periods with curtailed renewable (hours 8 - 17) and dispatched *down* during the remaining hours to compensate. In this simulation, for the hours when load is dispatched down, the lower bound on flexibility is often tight whereas the upper bound is not attained at any point over the time horizon. This highlights the point that both up- and down-flexibility are required in equal amounts due to the equality constraint for total demand over the time horizon (e,g., (4.1c), 4.14b). The limiting factor for shiftable loads to increase demand during the middle of the day (and therefore reduce renewable curtailment) could actually be their inability to reduce its demand at other times.

Rigure 4.2c drills deeper and considers the profile of an individual load. This

[10] Experiments with other values of m did not change results qualitatively.

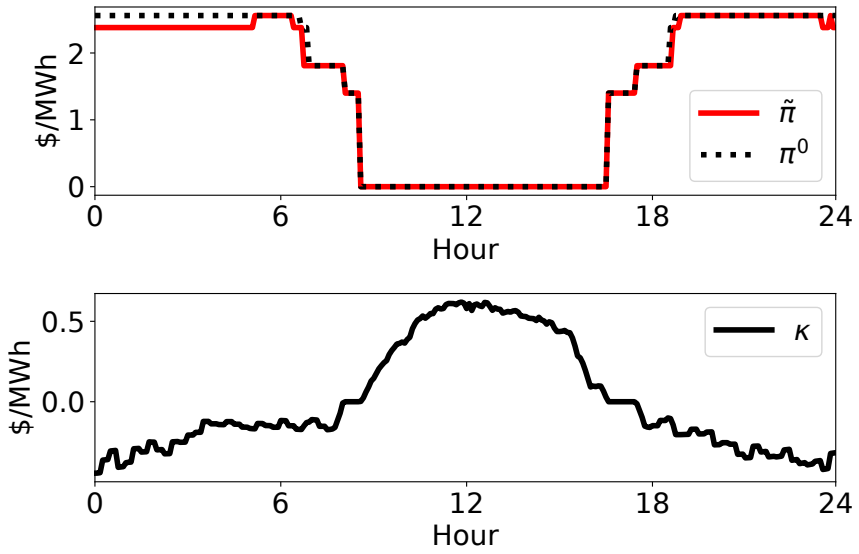


Figure 4.3: Illustration of energy prices $\tilde{\pi}$ and π^0 (top) and flexibility price κ (bottom) in the CAISO case study.

figure shows that feasibility of the flexible market dispatch for the load is indeed satisfied, as required by the constraints in (4.14b) - (4.14e). The black curve shows the baseline demand $\mathbf{d}_j^0$ for load $j = 29$ and the red curve shows the dispatch with flexibility $\mathbf{d}_j^*$. Both trajectories respect the upper and lower flexibility bounds $\overline{\mathbf{d}}_j, \underline{\mathbf{d}}_j$. Further, all loads (and therefore the aggregate load as well) change their dispatch under the flexibility dispatch allocation in the same direction (i.e., up or down) in each interval. This is due to constraints (4.12b) and (4.12c), without which the undesirable scenario where some loads increase and other simultaneously decrease their consumption could occur.

The case study also provides a concrete illustration of many of the properties of prices we proved previously. In particular, the top panel of Figure 4.3 shows this graphically that the baseline price π^0 is a lower bound for $\tilde{\pi}$, a property proven in Lemma 12. The lower panel of Figure 4.3 illustrates that the flexibility price κ satisfies its desired properties in that it is positive when up-flexibility is dispatched and negative when down-flexibility is dispatched. Its magnitude also reflects a time-varying value of flexibility; specifically, κ_t is most positive during the middle of the day when renewables have peak capacity and load should be dispatched up to utilize them and is most negative early and late in the day when expensive conventional generation dominates the generation mix and load should be dispatched down to reduce cost.

In Table 4.1 we quantify the economic value of the proposed market design as

Table 4.1: Total revenue, cost, and utility for generation and demand in CAISO case study. Amounts are in millions.

Amounts shown are in millions

	Baseline	With Flexibility	% change
Total Generation Revenue	\$11.91	\$10.58	-11.14
Total Generation Cost	\$6.39	\$5.39	-15.62
Total Generation Utility (profit)	\$5.52	\$5.19	-5.95
Total Demand Cost	\$11.91	\$10.90	-8.49
Total Demand Utility	\$-11.91	\$-10.58	+11.14
Total Flexibility Payment	\$0.00	\$0.32	—

compared to the baseline design by comparing market participant utility gains/losses between the two scenarios. The first observation from this table is that the demand side of the market increases its utility by 11% over the baseline while only needing to re-dispatch 10% of its total load. As flexibility is provided by the demand side of the of the market, our mechanism increases their utility to compensate.

The second observation is that each load individually is at least as well off under the flexibility mechanism as under the baseline scenario, but loads that offer flexibility are better off than those that do not. This can be seen by comparing total demand cost of \$10.90 to the total demand utility of -\$10.58. The difference in the magnitudes of these values is exactly the flexibility payment of \$0.32. Inflexible loads pay for energy but do not receive any benefit from the flexibility payment, which only goes to flexible loads.

Third is that generators are worse off under the flexibility mechanism due to a lower energy price $\tilde{\pi}$. Dispatching flexibility improves the overall efficiency (i.e., generation cost) of the dispatch but because the spot price decreases as well, that benefit is not captured by generators, instead going to the loads. From a generator's point of view, this is not desirable as it will lower their profits individually and collectively. However we remark that *any* improvement in market efficiency is likely to lower generator profits (for additional discussion of these see [72]). That does not mean that improvements in market efficiency ought to be avoided though. Rather, we take the view that incentives for improving system efficiency should be aligned with those of the market participants who actually provide the efficiency-improving service. In the setting we explore in this work, the deserving participants are flexible loads with shiftable demand.

4.5 Conclusion

This chapter focuses on a crucial and under-explored aspect of demand response markets: the incentives of loads with shiftable demand to expose flexibility to the market operator. We first show that relying on the energy spot price alone to compensate loads—as the standard market design does—leads to incentive misalignment: demand might end up worse off bidding flexibly than inflexibly. Our market mechanism addresses this shortcoming in two parts. The first constrains the total amount of flexibility that can be dispatched in each period, ensuring that costly generators cannot be dispatched. The second introduces a flexibility price and distributes the surplus that arises from the more efficient dispatch to loads that offer flexibility.

The flexibility price serves two useful purposes. One is to provide a time-varying signal to loads about the most profitable times to offer their flexibility to the market. A second value of the flexibility is to correct a free-rider problem that arises in an energy price-only market: flexibility is a public good, which means that all loads benefit from flexibility whether they contribute it themselves or not. In our mechanism, the flexibility payment, which is the product of flexibility price and flexibility dispatch, is only non-zero for flexible loads.

Importantly, our proposed mechanism has the same basic structure as the current economic dispatch market design, which provides a pathway to adoption. In this work though, our model sets aside several real-world electricity market features like startup costs, ramping constraints, line congestion, and rolling window market clearing. These undeniably impact market dispatch and are the focus of large portion of research on electricity market design. However they are typically evaluated without considering a responsive demand side of the market. In contrast, our focus here is on the mechanism for incorporating shiftable demand into the economic dispatch framework and analyzing the incentives that result. It will be important in future work to tease out how the above-mentioned generation-side characteristics interact with the demand-side structure in our model.

Finally, another important open problem motivated by our work relates to flexibility pricing. Our market design shows how to incorporate a flexibility price into the marketplace and proposes three potential designs for flexibility prices. The flexibility prices we introduce satisfy the minimal desired properties, but each have some drawbacks and thus a further exploration of the design of flexibility prices is an important research question. In particular, is there a stable and interpretable flexibility price, aligned both with individual and social welfare objectives, that incentivizes

loads to bid their flexibility into the market?

Part III

Microgrid Operating System (OS)

The second part of the thesis addresses the theme of decentralization by proposing an operating system for microgrids and distribution networks to help them meet the challenges of managing a distributed power resources. Realizing a 24/7 low-carbon, resilient, and reliable energy system calls for a paradigm shift towards active monitoring and real-time control in low-voltage distribution grids. Over a decade of research in the R&D community has resulted in significant advances in optimization and algorithms, providing ever more powerful tools for controlling and exploiting the latent flexibility present in distributed energy systems. However, two major obstacles to unlocking this potential in practical setting are 1) the lack of reliable real-time timeseries data and 2) the challenge of obtaining sufficiently detailed system-level models of the infrastructure. In this work, we present the design and implementation of a Microgrid Operating System (Microgrid OS) to bridge this gap in data and modeling.

The Microgrid OS is a software platform that provides the foundation for a collection of energy services applications such as demand response, carbon reduction, peak shaving, and state estimation. It does so by monitoring, modeling, and in some applications, controlling a network of heterogeneous hardware such as batteries, PV, EV chargers, and flexible load. In Chapter 5 we motivate the particular issues that DERs face in scaling and describe the three layers of the Microgrid OS architecture: data, models, and optimization. In the remaining chapters, we present implementation of Microgrid OS components undertaken thus far in a real-world testbed on the Caltech campus.

In Chapter 6 we present our practical experience designing and deploying a network of internet-connected smart electrical meters on the Caltech campus for the purpose of building a high-quality, granular energy dataset. In addition to collecting standard measurements like current, voltage, and power magnitudes, we also sample raw current and voltage waveforms to obtain phasors, with inter-meter synchronization accomplished with internet time protocols. Our low-cost synchrophasor implementation has the potential to yield synchrophasors for all nodal injections within a network, facilitating the use of data-driven topology and parameter estimation algorithms.

In Chapter 7 we share our work building a “Digital Twin” of the Caltech power system, which provides the foundation for downstream applications in system planning, control, and optimization. The Digital Twin models the complex and heterogeneous campus electrical network, which includes a variety of electrical, sensing, and com-

munication components. Our core contribution is a flexible, comprehensive network modeling schema that bridges capabilities of standard distribution system modeling tools (e.g., OpenDSS) and component/building-level models. The schema is applicable to any low-voltage multi-phase electrical network, from single-building microgrids up to distribution systems.

The Caltech Digital Twin includes: 1) topology and line parameters for a three-phase electrical network consisting of four high-voltage substations, two power plants, and approximately one hundred buildings, connected by thousands of buses, switches, lines, circuit breakers, and transformers; 2) real-time voltage, current, and power flow data from hundreds of meters in the network; 3) real-time generation data from on-site distributed generation sources including a gas co-gen plant, a steam turbine, fuel cells, and rooftop PV arrays; and 4) real-time temperature and HVAC data from buildings.

We discuss our software architecture for large-scale data collection, storage, analysis and visualization, with careful considerations to scalability and security. The back end of the software platform consists of data collection interfaces, data processing, and modeling modules. The front end consists of a browser-based visualization interface that allows users to interact with the network in real time and explore various scenarios at various levels of spatial and temporal granularity.

The data and modeling layers of the Microgrid OS provide a foundation for a collection of optimization applications, including energy cost reduction, power factor correction, and carbon emissions reduction. In Chapter 8 we detail several applications supported by the Microgrid OS related to power quality correction (volt-var control, power-factor correction), state estimation (topology and parameter estimation), and infrastructure planning (optimal storage placement and sizing). These applications were developed in collaboration with university campus facilities managers, distribution utilities, and industrial partners.

Chapter 5

DISTRIBUTED ENERGY RESOURCES (DERS) AND THEIR CHALLENGES

DER technologies facilitate decentralization because they enable the production of energy at the point of consumption, reducing reliance on centralized power plants and utility grids. DER technologies allow customers to generate their own energy from renewable sources such as solar panels, wind turbines, and geothermal systems.

By producing their own energy, individuals and communities can reduce their dependence on centralized power grids and potentially even sell excess energy back to the grid. This creates a more decentralized energy system, where power generation is distributed across multiple smaller sources rather than being controlled by a few large power plants. Moreover, DER technologies allow for greater control and customization over energy production and consumption, enabling individuals and communities to tailor their energy usage to their specific needs and preferences.

In the past 15 to 20 years, rapid growth of distributed energy resource (DERs) has been driven by advances in fundamental hardware, falling costs, and increasing demand for clean energy. The most crucial lever in these advances has been the swiftly (and sometimes exponentially) decreasing costs and production scale of critical building block technologies since 2008: photovoltaic modules, lithium ion batteries, electric vehicles, and wind turbines. Figure 5.1 shows exponential decreases in solar costs of the last three decades and and Figure 5.2 shows the 5x decreases in lithium-ion battery pack prices over the last decade.

One of the outcomes of these cost and scale trends is that wind and solar are the cheapest energy generation technologies as Figure 5.3 demonstrates. Renewables now compete favorably against combined-cycle gas in terms of cost of new capacity. In addition, renewables have additional revenue opportunities from production tax credits, power purchase agreements with large corporate buyers, and portfolio mandates from state and local governments.

In Figure 5.4 it is clear that this trend is only accelerating as interconnection queues in North American ISOs are now dominated by renewables and storage [114].

In addition to falling capital costs, these technologies have the common characteristic

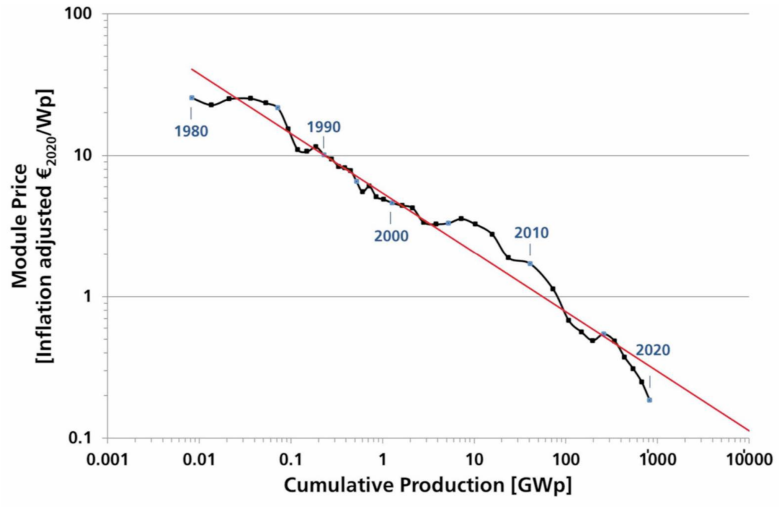


Figure 5.1: Per-watt solar module prices from 1980 - 2022 [51]. The exponential decrease in cost is sometimes referred to as Swanson's Law [49].

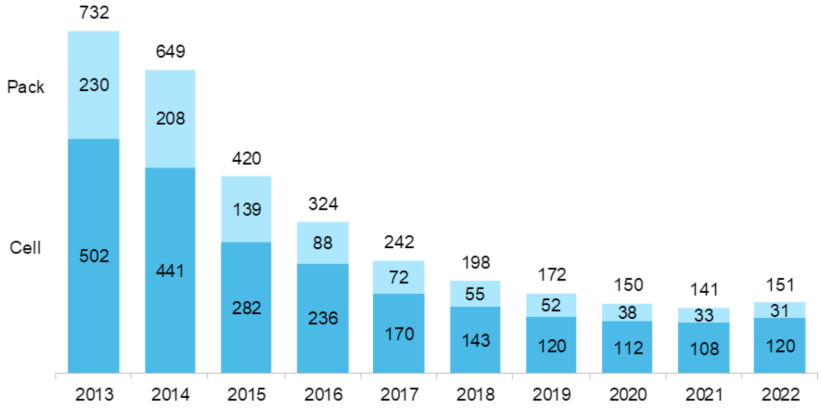


Figure 5.2: Average lithium-ion battery prices from 2013-2022 in $/kWh [17]. Both cell and pack costs have enjoyed similar magnitude decreases in unit costs.

in that they fundamentally facilitate decentralization. This contrasts with traditional generation technologies like gas, oil, nuclear, and coal which require large, megawatt or gigawatt scale to be logistically and financially viable. In contrast, solar generation plants can scale from sizes of hundreds of watts to hundreds of megawatts using the same fundamental components. There are similar dynamics for battery storage and wind. Despite this however, DERs do face significant challenges in scaling up further and faster due to the complexities of managing a large number of interconnected devices. DERs have transitioned to a regime where limits on their proliferation are determined by control and integration complexity rather than production cost.

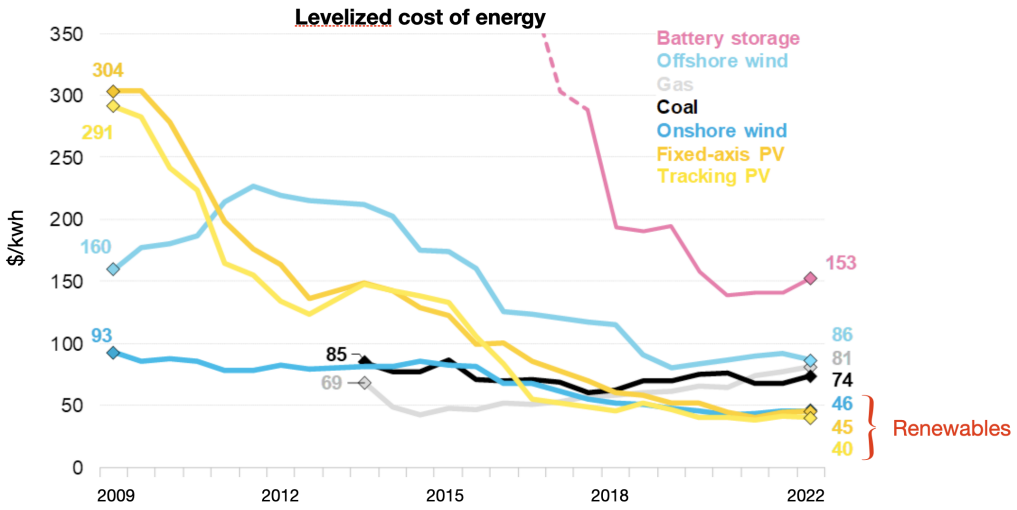


Figure 5.3: Levelized cost of energy (LCOE) in $/kWh for various generation technologies in the US, 2009-2022 [16]. Onshore wind and utility-scale solar have the lowest LCOE of any generator types since 2018.

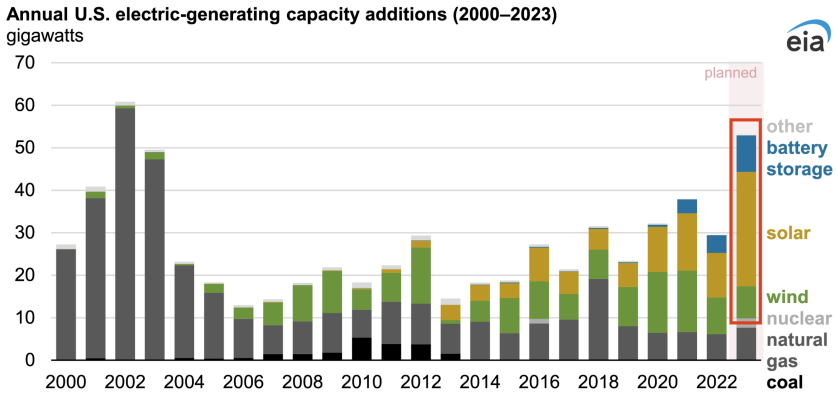


Figure 5.4: US generation capacity additions by generator type, 2000-2023 [129]. Passage of the Investment and Recovery Act in late 2022 has spurred significant additional capacity additions for renewable and storage in 2023.

5.1 Challenges for scaling DERs

We identify the following key challenges in scaling decentralized DERs: 1) interoperability of DER component technologies; 2) limited capacity in distribution networks to accommodate distributed generation; 3) limitations in current power system management practices; 4) lack of revenue opportunities in markets.

The first challenge of making DERs interoperable is mainly related to the diversity of DERs and the lack of standardization in their data, communication, and control protocols. Key difficulties are:

- Heterogeneity: DERs are manufactured by different vendors and targeted towards diverse end-use applications and customers.
- Communication: Standardized communication protocols for DERs are rapidly developing but still immature. DERs may use and support of various communication protocols depending on their manufacturer, making it difficult to integrate them into a single cohesive system.
- Security: DERs are vulnerable to cyber-attacks, and ensuring interoperability without compromising security can be challenging. Security measures must be built into the system to ensure that DERs are protected from cyber threats.
- Data: DERs generate large amounts of data, and managing this data can be a challenge. Data management systems must be in place to ensure that data is accurate, reliable, and accessible.

The second challenge of integrating large numbers of DERs into distribution grids arises because distribution systems were not historically engineered to support bi-directional power flow and numerous, independent power producing devices. Integrating large amounts of DERs into the distribution grid requires significant planning and coordination across various stakeholders, including utilities, regulators, and DER providers. Some key obstacles are:

- Power quality: DERs can cause power quality issues such as harmonic distortion, flicker, and voltage instability (both over and under voltage). These issues can cause disruptions to sensitive equipment and affect power quality.
- Capacity limitations: The existing infrastructure, such as lines, transformers, and power protection devices, may have capacity limitations that prevent the integration of additional DERs. Uncontrolled addition of DERs into distribution feeders can lead to equipment failures, outages, and increased cost from upgrading network infrastructure.
- Coordination and control: Integrating large amounts of DERs requires coordination and control of various elements of the power system, including distribution transformers, switchgear, and protection devices. Advanced control systems and algorithms must be put in place to ensure efficient and effective coordination.

- Regulation: Regulatory barriers such as interconnection standards, net metering policies, outdated grid access rules, and long wait times for interconnection hinders the integration of DERs.

The third challenge we identify for integrating large numbers of DERs are the policies and procedures that system operators (e.g., distribution utilities, facilities managers) use to plan and operate their systems. The basic problem is that power systems were not designed with a decentralized architecture in mind. This manifests itself in the following ways:

- Limited visibility: Traditional control systems have limited visibility into the performance of DERs, which can make it challenging to manage and control their output. Without real-time data on the performance of DERs, operators
- Reactive response: Many utilities utilize crude and reactive response measures to manage DERs, such as reducing output or disconnecting them from the grid. This can result in their inefficient use and limit their potential benefits to the grid.
- Lack of communication: Many utilities do not have effective means of real-time communication with DERs, which can hinder their effective coordination and control. Scalable and easily deployable communication systems are needed to help DERs seamlessly into the grid and that their performance is optimized.

The fourth challenge in scaling DERs are the lack of opportunities for generating revenue to cover fixed and operating costs and drive investment. In contrast to traditional bulk generators, DERs typically cannot participate in ISO-run energy or ancillary services markets. Although there is significant regulatory interest around this issue and utilities and ISOs are currently studying ways to integrate DERs, they still must justify their installation costs exclusively based on avoided cost metrics or fuzzy calculations about the value of reliability. Despite this current state of affairs, DERs have significant potential to contribute to markets for energy, demand response, ancillary services, capacity, and carbon.

- Energy Trading: DERs can participate in energy trading markets to sell excess energy back to the grid or to other participants. This can be achieved through

virtual power plants, which aggregate small-scale resources to act as a larger power plant.

- Demand Response: DERs can participate in demand response programs to reduce energy consumption during peak demand periods. This can be achieved through smart thermostats, energy storage systems, and other technologies.
- Ancillary Services: DERs can provide ancillary services such as frequency regulation, voltage stabilization, fast ramping services, and reactive power support to the grid. These services are necessary to maintain grid stability and reliability and the fast-response characteristics of DER inverters are very well suited to providing ancillary service.
- Capacity Markets: DERs can participate in capacity markets, where they can earn capacity credits for providing energy during peak load hours. This is particularly relevant to battery storage.
- Carbon Markets: DERs that generate clean energy or reduce emissions can participate in carbon markets and earn credits for their contributions to reducing greenhouse gas emissions.

These market opportunities are likely to grow as more DERs are deployed and as energy markets become more decentralized and flexible. However, there is an urgent need for scalable dispatch and control mechanisms to realize the market participation potential of DERs.

5.2 Layered architecture for DER control

Layered architecture is a technology design paradigm that separates different components of a system into distinct layers, each of which has a specific function and interacts with other layers in a well-defined manner. Each layer provides services to the layer above it and consumes services from the layer below it. Layered architectures provide several benefits for complex, interconnected technologies including:

- Modularity: Each layer can be developed and tested independently of the others, making it easier to maintain and update the system over time.
- Scalability: A layered architecture can be designed to scale horizontally, by adding more instances of a layer, or vertically, by adding more layers to the

stack. This allows the system to handle increasing amounts of data or traffic without sacrificing performance.

- Separation of Concerns: A layered architecture separates the concerns of a system into different layers, each with its own responsibilities. This promotes a cleaner design and makes it easier to debug and troubleshoot problems.
- Flexibility: Layers can be added, removed, or modified as needed, without affecting the other layers. This makes it easier to adapt the system to changing business requirements or technology trends.
- Security: A layered architecture can be designed to provide security at each layer with clearly-defined access points into each.

In the information technology (IT) sector, layered architectures have facilitated rapid scaling of innovations with examples that include TCP/IP protocols, open systems architecture (OSI), and virtualized or cloud computing stacks. In contrast, the electric power industry has a long history of vertical integration, which is the ownership and control of different stages in the production and distribution of electricity by a single company. The origins of this vertical structure can be traced back to the early days of electrification in the late 19th and early 20th centuries.

The early direct current (DC) power distribution systems championed by the Edison Illuminating Company, which was the first investor-owned electric utility, owned and operated the generating stations and the distribution lines to end customers [60]. Due to the low voltages inherent in this system design, it scaled poorly over longer transmission distances. The proliferation of alternating current (AC) transmission and distribution systems provided one of the first examples of how layered architectures enhanced the scalability of electric power. AC power can be easily and efficiently transformed from low to high voltages and back. This allows for power networks to be separated into distinct regions with their own operating characteristics and processes. Despite this innovation, the natural monopoly attribute of electric utilities meant that vertical integration of the power industry persisted through much of the 20th century.

In an attempt to drive more efficiency in the utility industry and reign in repeated rate increases, the U.S. government passed the Energy Policy Act of 1992, which led to the restructuring of the electric power industry and the separation of generation, transmission, and distribution functions. This resulted in the establishment of

competitive electricity markets, where generation companies could sell electricity to load serving entities (e.g., utility companies) as well as other market participants. These changes spurred investment in new, more efficiency generation technologies and opened the industry to competition [100].

Despite these advances, the power system is still primarily a relatively small (~10,000) set of bulk generators and load centers connected with a transmission network. Distribution utilities still manage their distribution networks in a centralized, top-down fashion. However, this operational paradigm is poorly suited to a power system comprised of millions of controllable generation and load devices. As these proliferate due to fast-paced improvements in battery storage, solar modules, electric vehicles, and thermal electrification, layered architectures have an important role to play in managing the ballooning complexity and stochasticity of the system.

The main conceptual contribution of this part of the thesis is a layered system architecture for DERs and microgrids. We will first present this architecture and then argue for how it addresses the four aforementioned challenges to scaling DERs.

The bottom-most layer of the DER stack consists of the hardware infrastructure. The top-most layer is a set of *applications* that the hardware can support. In between are control and communication layers that link the hardware to the applications. A schematic of this architecture is presented in Figure 5.5.

The hardware layer includes 1) generation resources such as solar PV and diesel generators, 2) storage such as electrochemical batteries and thermal storage (hot/cold water, sand), 3) loads like electric vehicles, heat pumps, and 4) network infrastructure like transformers, lines, breakers, and switchboards.

The communication layer consists of typical industrial automation protocols such as Modbus and BACNet, as well as new standards targeted towards DERs from organizations like SunSpec and OpenADR.

The application layer consists of standard objectives such as utility retail rate optimization (e.g., peak shaving, time-of-use, demand charge) and wholesale market participation (DA and RT energy, frequency regulation, ancillary services) as well as new market opportunities such as carbon footprint minimization, capacity services, and utility grid services (power factor correction, voltage correction).

Our main technical contribution in this part of the thesis is the design and implementation of a data, modeling, and control layer that links the application

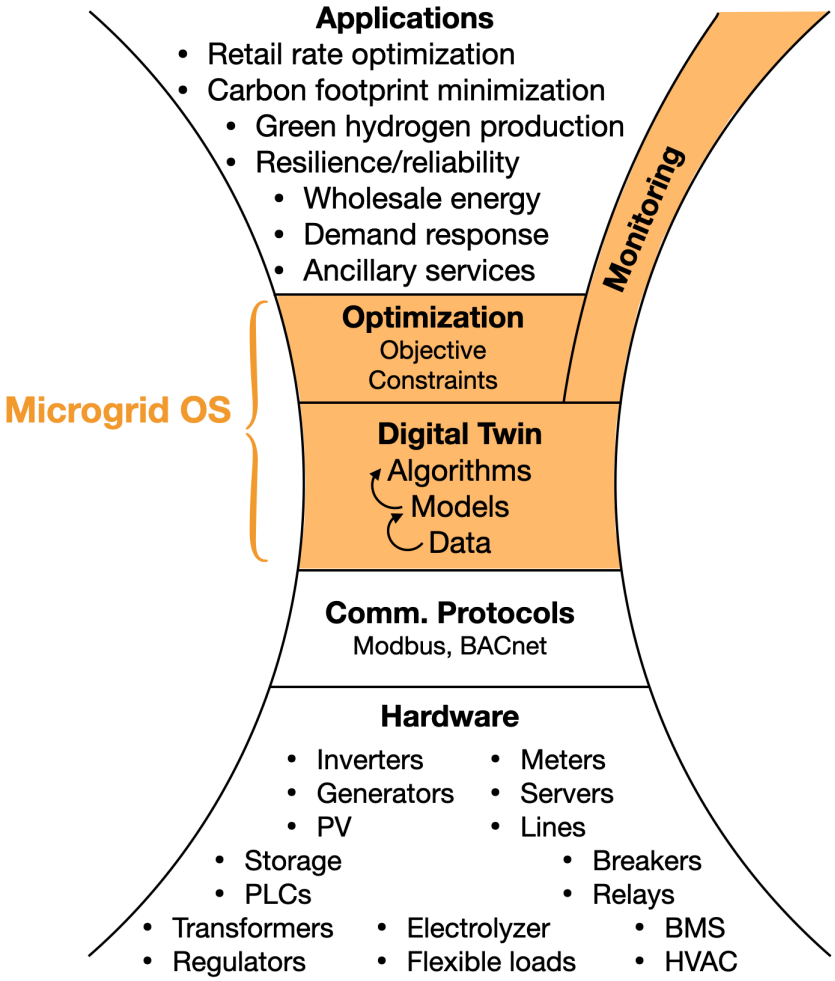


Figure 5.5: DA (dashed line) and optimal RT (solid line) dispatch trajectories for generators and load over a 24 hour scheduling horizon.

layer with the hardware layer in the DER hourglass. We call this layer the Microgrid OS.

Within the Microgrid OS layer, shown in Figure 5.6, there are three core strata: Data, Models & Learning, and Optimization & Control. Each of these strata serves as the foundation for the one above it.

The Data block in Figure 5.6 consists of:

- Hardware and software interfaces for data collection;
- Timeseries databases;
- Timeseries data cleaning, error detection, and aggregation;

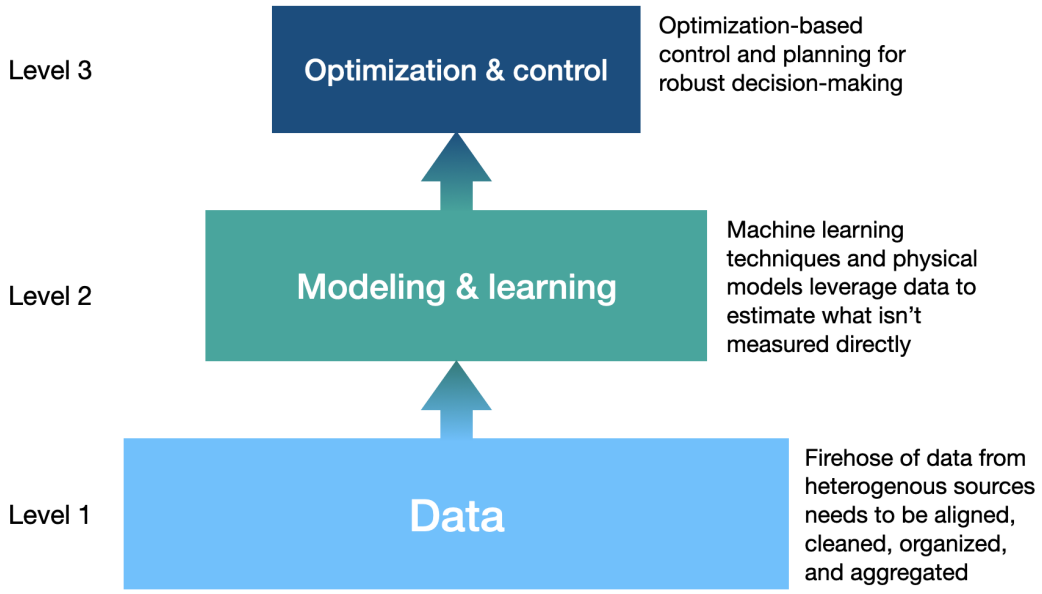


Figure 5.6: Three core functionalities of the Microgrid OS

- Network asset databases (including GIS data).

Examples of inputs into the Data block are interfaces with power meters, existing databases of timeseries data, and network assets (e.g., OpenDSS, Cyme) or singleline diagrams. The output of the Data block is an organized, consistent, relational dataset where timeseries are linked to specific quantities at specific locations/devices in the network model.

The Modeling & Learning block in Figure 5.6 consists of:

- Network and component model schema;
- Statistical timeseries models
- Data-derived network models
- Thermal building models
- End-user behavior models

Examples of the the models contained in the Models and Learning block are both first-principles models, such as the admittance matrices for power networks and heat transfer differential equations describing thermal systems, in addition to models learned from data, examples of which include statistical forecasts of timeseries

and predictive models for end-user behavior (e.g., EV charging arrival processes, building usage).

Finally, the Optimization and Control block leverages the models and data below to pose optimal control problems for planning and operating aggregations of DERs. In Chapter 8, we present several examples of such problems that we have explored in real-world systems, including optimal sizing of energy storage systems for reliability, microgrid co-design in electricity markets, and optimal battery placement and operation for voltage support in distribution networks.

The Microgrid OS and the Data-Models-Optimization paradigm in particular address the four challenges to scaling DERs presented at the outset of this part of the thesis. Interoperability of DERs is advanced by isolating interaction with hardware assets to a set of interfaces (i.e., APIs) which translate data and control signals into a common format within the Microgrid OS independent of the particular DER being interfaced with. The limited capacity of distribution networks to accommodate DERs is mitigated by the network models that are core to the Microgrid OS. These network models are prerequisites for accurate powerflow simulations and real-time state estimation algorithms, which allow planners to identify opportunities for DER expansion in their systems (e.g., hosting capacity) and enable operators to control DERs to respect physical constraints in the system. Limitations in current power system control practices often derive from a lack of visibility into the system in real time and incomplete knoweldge about the infrastructure that is being managed. By centralizing and coupling multiple different data types (e.g., timeseries data, assets databases, singleline diagrams) into a single, cohesive models, operators gain situational awareness over their systems and can take decisions in an informed, data-driven manner. Finally, the Microgrid OS provides the technical capabilities of communication and control for complex systems of DERs (e.g., campus microgrids) to be able to participate in utility and ISO market programs. Grid operators need to be able to both send dispatch signals to DERs (or aggregations of them) and receive telemetry validating that the resources produced or consumed as scheduled. The Microgrid OS provides a single portal for operators to communicate with and dispatch of DERs by translating grid signals into operational setpoints for each resource.

5.3 The Caltech Microgrid Testbed

The design principles for the Microgrid OS are general to any low-voltage distribution system, from single buildings up to large, multi-phase distribution systems with thousands of customer endpoints. However, the microgrid on the Caltech campus has presented an ideal testbed for understanding challenges in workplace decarbonization and for the testing of new technologies to overcome them. Despite Caltech's small size, its energy infrastructure is large and complex (see Figure 5.7. The electricity, heating, and cooling needs correspond to those of ~20,000 people in California. In 2020, Caltech consumed 108,934 MWh of electricity with a peak demand of about 19.2MW.In recent years, we have produced ~100% of our electricity onsite on an annual basis. The main generation resources on campus are a 12.5MW gas co-generation plant, 4MW of fuel cells (Bloom Energy), and 2MW of rooftop solar panels distributed across 15 arrays.

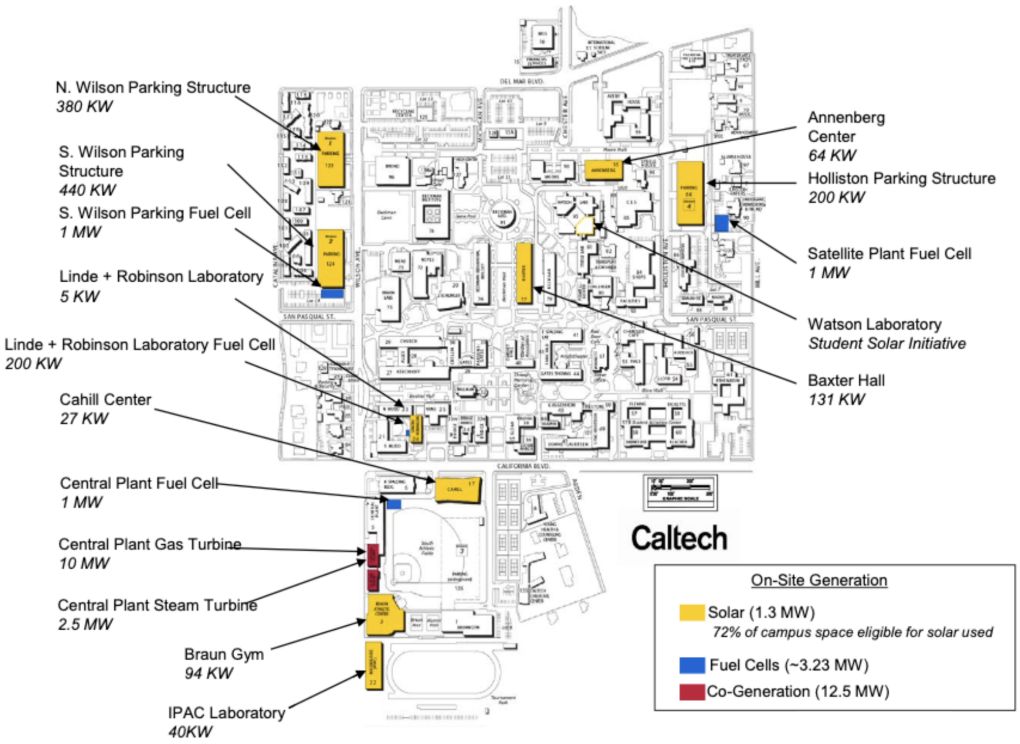


Figure 5.7: Map of Caltech Microgrid power generation assets. Source: 2020 Caltech Facilities Master Plan

The greenhouse gas (GHG) emissions have been rising steadily until Caltech's implementation of a Climate Action Plan, starting in 2008, to reduce its annual GHG

emissions. It has been tracking its yearly targets till about 2015. In 2020, Caltech emitted 64,517 mTCO2e against the Climate Action Goal of 51,000 mTCO2e. Of this emission, 92% is due to electric, heating and cooling loads on campus, including both Scope 1 and Scope 2 emissions

The Caltech campus has approximately 120 buildings connecting by a 2.4kV distribution network. Transformers in each building convert this voltage to 480V or 208Y/120V services for distribution to electrial panels within the buildings. There are four substations that connect to 3 17kV feeders from the utility grid. Within the Central Utility Plant on campus, there is also a 4.16kV service connect to large HVAC chillers and the cogen plant.

In the remaining chapters in this part of the thesis, we present further details of the Microgrid OS through a set of related implementation projects on the Caltech campus. These projects retain a focus on the data acquisition and system modeling components of the architecture, as these functions are a main obstacle to successful deployment of DERs at scale and to applying advanced control methodologies in practice.

Chapter 6

DATA: METER CALTECH

This chapter of the thesis presents a deployment project at Caltech to collect real-time, reliable, and granular data on the operation of the campus energy system. This project has unfolded in several phases. The first phase (Section 6.1) consisted of interfacing with and integrating the existing building management system (BMS) database that Caltech Facilities manages with the Microgrid OS. The second, ongoing phase involves (Sections 6.2-6.3) installing high-resolution smart meters on all of significant load and generation assets around campus, for the purpose of collecting both phase and magnitude data on electricity injections and line flows. The third phase (Section 6.4) involves building a network model for the electric microgrid, translating individual electric singleline diagrams and assorted asset databases into a coherent, system-wide network model that enables graph analysis and optimization.

6.1 Building Management System Data

Caltech has a building management system built on top of Niagara's Tridium framework. The system consists of a network of jaces, each of which aggregates a set of sensors and actuators connected serially. The two communication protocols used in the system are Modbus (both serial and TCP/IP) and BACNet. The data from the jaces is aggregated in a centralized "Histories" repository where each data register is recorded in a separate `.csv` timeseries file. In total, there are approximately 85,000 timerseries in the database with date timestamps ranging from 2015 to the present.

The baseline granularity of the data is 15 min average values, although a small number of sensors have higher/lower resolution. The datapoints we focused on corresponded to electric meters, hot/cold water meters, and building temperature sensors. The values recorded for electric meters are (typically):

- Phase-to-phase voltage RMS magnitudes (V)
- Phase current RMS magnitudes (A)
- Power factor averaged across 3 phases
- Total 3-phase real power (kW)

- Total 3-phase energy (kWh)
- Total 3-phase reactive power(kVar)
- Total 3-phase apparent power (kVA)

Notably the model of meter most prevalent in the Caltech network, Schneider Electric's PM800 series, does not compute the per-phase power factor. In addition, there is significant variation from meter to meter in which of the above quantities are included and how they are labeled.

The hot/cold water meters typically measure the following quantities:

- Supply temperature (°F)
- Return temperature (°F)
- Flow rate(gallons/min)

Temperature sensors collect air temperatures in °F inside of individual rooms or building zones.

The first challenge in making sense of the BMS data was identifying which series corresponded to which building and which specific physical elements they were measuring. Due to incomplete record keeping, it was not possible to do this thoroughly. Naming conventions are used inconsistently through the data and the jace-to-building mapping (which would at least allow one to physically locate a sensor within a particular building) is not one-to-one. Although we were able to identify the locations of most electrical meters by manually cross-checking with the single-line diagrams, there is no equivalent set of diagrams for the hot/cold water distribution loops and therefore we were only able to associate water meters, at best, with a particular building.

A second challenge in making use of the BMS data is its poor quality and completeness. Of the series present in the histories database, at least 25% have no data as shown in Figure 6.1. Missing chunks of data are common and data errors are frequent. These data errors stem from multiple causes including Modbus configurations issues, sensor configuration errors, and communication network outages. In addition, data is not well maintained in the BMS database. Figure 6.2 shows that at various points the size of the BMS database decreases, indicating mass data losses.

This makes it clear that while data may be collected, it is not retained over time for analysis.

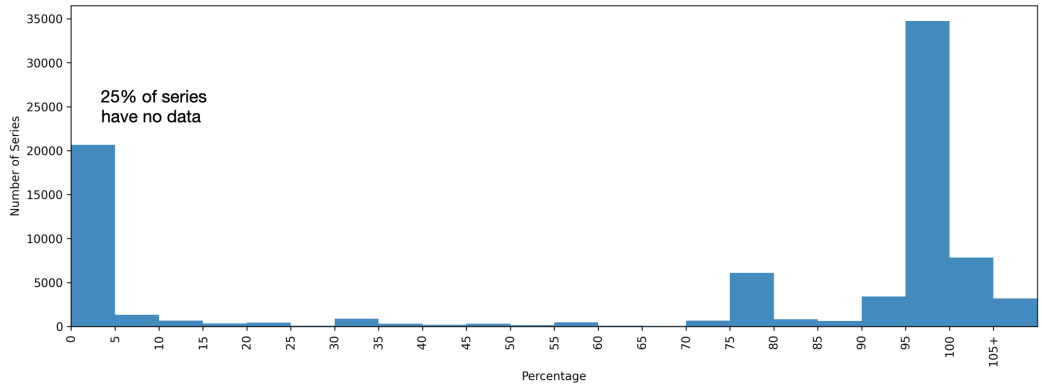


Figure 6.1: BMS timeseries with the percentage on the horizontal axis indicating the proportion of data present, based on a 15-minute time resolution. Horizontal axis values larger than 100% correspond to series with a sub 15-minute resolution.

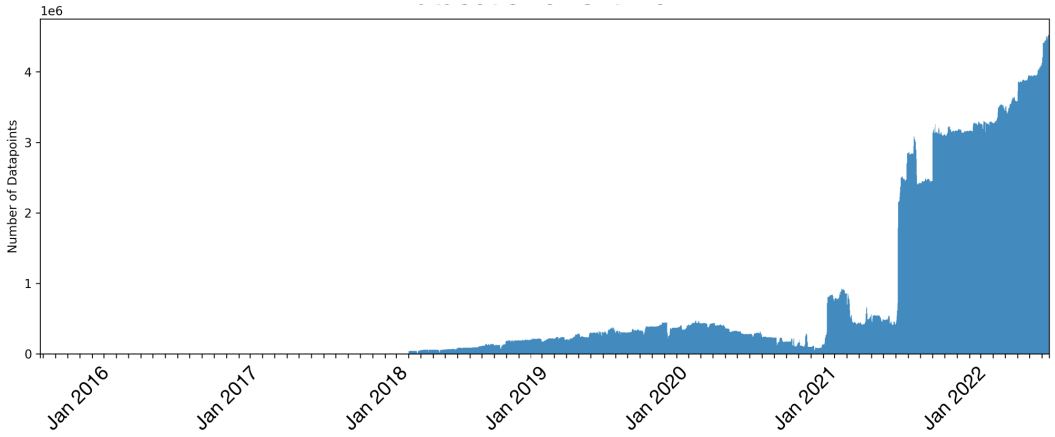
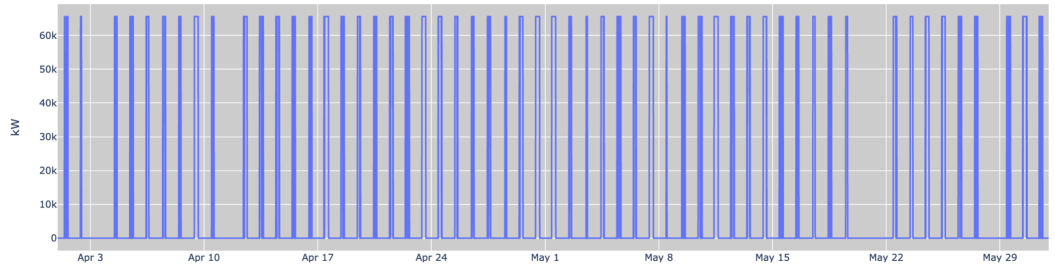


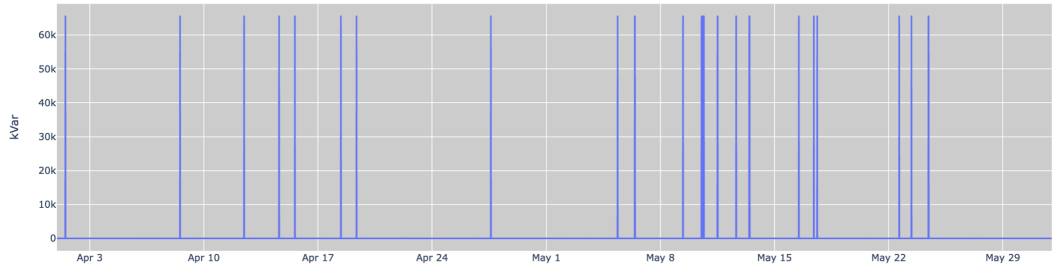
Figure 6.2: Number of timeseries in the BMS database vs. time. Sudden decreases in the data correspond to episodes where data was deleted or lost.

One common data corruption mode is shown in Figures 6.3a and 6.3b, where unphysical values of approximately 65,000 are interspersed with actual data points. We believe this is due to a register configuration error in the data processing pipeline but it was not feasible to locate and correct any of the errors of this type due to the complexity and inaccessibility of the configuration settings. In addition, the real and reactive power values shown in Figures 6.3(a) and 6.3(b) are not physically consistent with the apparent power magnitude in Figure 6.3(c), leading to a lack of confidence in the accuracy of the recorded values.[1]

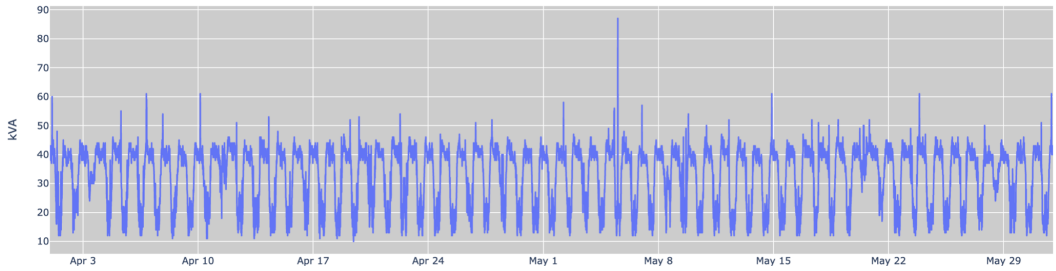
[1] In our conversations with Caltech Facilities analysts and engineers, skepticism about the validity



(a) kW



(b) kVar



(c) kVA

Figure 6.3: 5-min resolution kW, kVar, and kVA data from a meter in Annenberg Center, April - May 2021. Non-physical values for active and reactive power suggest sensor, communication, or database configuration errors within the BMS.

A third challenge, pertaining particularly to the electric meter data, is its crude resolution. In the time domain, data are averaged over 15 minutes. In the data domain, values are rounded to the nearest whole number for many registers (e.g., voltage) and are averages across phases (e.g., power factor). For the downstream tasks we seek to use this data for, such as power factor correction or phase balancing, the data does not have sufficiently granularity. Compounding this is the significant lag from real-time in updating the database. Data streams in from the sensors in the network unpredictably and there can be a lag of several hours from when a measurement is taken and when it is recorded in the database.

of BMS data is a common theme. They report not relying on BMS data at all, due to the frequency of obvious errors and gaps.

Despite these shortcomings, the BMS data are the most complete dataset available for energy use on campus. We implemented an API that would allow us to scrape new (as well as historical) data via HTTP get requests from the BMS database every 15 minutes. We implemented functionality to efficiently interleave incoming timeseries segments with our existing record, which was necessary given the multi-hour, unpredictable lag in BMS updates. This database provides the most complete record available of the thermal and electrical energy use on the Caltech campus. However, for the aforementioned reasons, dataset was not sufficiently granular, reliable, or complete to perform real-time monitoring and control. To address these shortcomings we embarked on a project, discussed in the next section, to install higher-resolution metering hardware and implement robust data processing pipeline to collect data from these meters in a consistent, organized manner.

6.2 Smart Meter Installation

In collaboration with Caltech Facilities Department, we have been deploying eGauge-brand smart meters to measure building load switchboards and DERs through the campus. The goal of the Meter Caltech project is to install meters on all of the building loads and distributed generation sources in the campus microgrid. This would correspond to approximately 120 - 150 individual meters.

The meter hardware we selected for the project is manufactured by eGauge Metering Systems. Their EG4015 series meter, shown in Figure 6.4, accomodates 15 current transformer (CT) inputs and can measure one set of 4-phase (ABCN) voltages up to 480V phase-to-phase. Measurements on all channels are recorded in the device's onboard memory every second. The onboard data can be accessed by a HTTP-based API over an Ethernet connection. Time stamping of measurements is implemented with Network Time Protocol (NTP). By working with Egauge engineers over the course of this project, we were able to introduce support for Precision Time Protocol (PTP) in the firmware for significantly enhanced precision in measurement alignment. (This will be discussed in further detail in Section 6.3.) The meter's oscilloscope function allows snapshots of waveforms on any channel to be recorded. The device's internal clock has a resolution of 1 microsecond and in practice, we observed that datapoints are sampled on each channel approximately every 400 microseconds, corresponding to a sampling frequency of 2.5kHz. This is sufficient to resolve harmonics up to 1.25kHz.

The installations of the meters themselves in switchboards called for careful ac-

Figure 6.4: Egauge meter Model EG4015 has inputs for 15 CTs and 4 voltage connections for 480V phase-to-phase.

counting during the installation process of what physical elements (i.e., conductors and buses) were being instrumented with CTs and voltage taps. One of the shortcomings in the BMS dataset that necessitated the smart metering deployment in the first place was poor accounting during the commissioning process of what physical components measurements actually corresponded to. For each placement of a CT or voltage tap, we tracked precisely which physical element in the network model (see Section 6.4) the sensor was being associated with. This detailed accounting during the installation phase was crucial to realizing the value of accurate measurements later on.

Given the magnitude of the metering deployment undertaking, the project is being implemented in three phases: (0) an initial test phase in a single building (Annenberg Center) to validate the suitability of the meter hardware and measurement quality; (1) a deployment of an additional 20 meters in the self-contained subnetwork connected to one of the campus substations; (2) complete metering of the remainder of campus injections.

Phase 0: Annenberg Center

Phase 0 of Meter Caltech consisted of installing two Egauge meters on the two main switchboards in Annenberg Center. The goals of this initial phase were 1) to understand the procedures required to install meters and CTs on building switchboards 2) to validate that the Egauge meters could record high-resolution, real-time electrical data 3) to confirm that we could communicate with the meters over the campus data network and stream data from the meters to our database server in real time.

As indicated in Figure 6.5, Annenberg Center is fed by two independent 2.4kV feeders from Substation 2. The medium-voltage feeders are transformed down to 480V L-L and terminate in two main switchboards, which are not connected by any tie circuit breaker as is commonly the case in other buildings. Main switchboard A (MSA) feeds the building load and main switchboard B (MSB) supplies power to a data center. There is a 56kW rooftop PV array that feeds the main busbar of MSA.

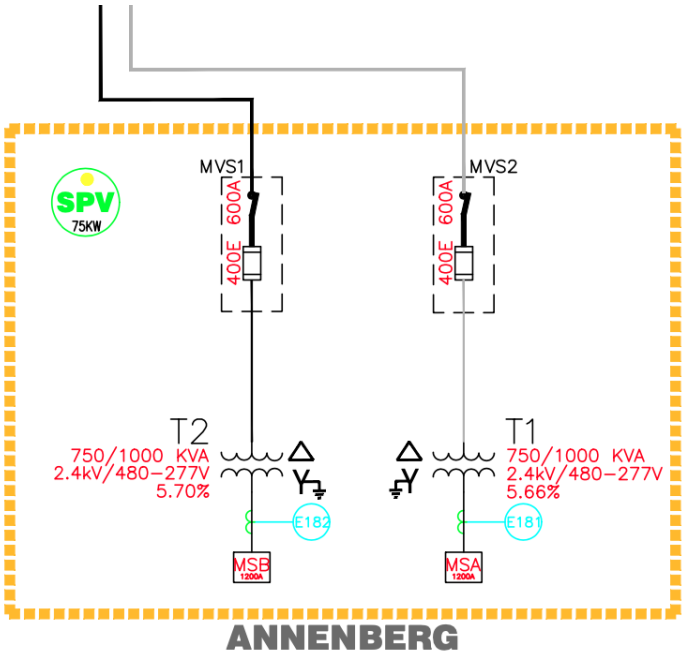


Figure 6.5: Annenberg Center single-line diagram. The solar PV (SPV) connects to MSA. The black and grey incoming feeders link the buildings to Substation 2.

Figure 6.6 provides more detail on which breakers within the MSA and MSB panels we installed CTs. On each of the switchboards, we obtained a voltage reference from the three main busbars. We then measured the current flows in the main busbars with rope CTs on each phase and and additional four of 3-phase circuits feeding the main busbars with split-core CTs on each phase. This redundant sensor arrangement allowed us to validate and quantify the current measurement accuracy at the busbar node with Kirchhoff's Current Law.

CT capacities were sized to the ampacity of the circuit breaker for the feed being measured. 16" Rogowksi coil CTs manufactured by Accuenergy were used on the main busbars (see Figure 6.7) and 200-400A revenue-grade split-core CTs manufactured by Continental Control Systems were used on the other lines. The CTs guarantee measurement error of less than 1% for current magnitudes and less than

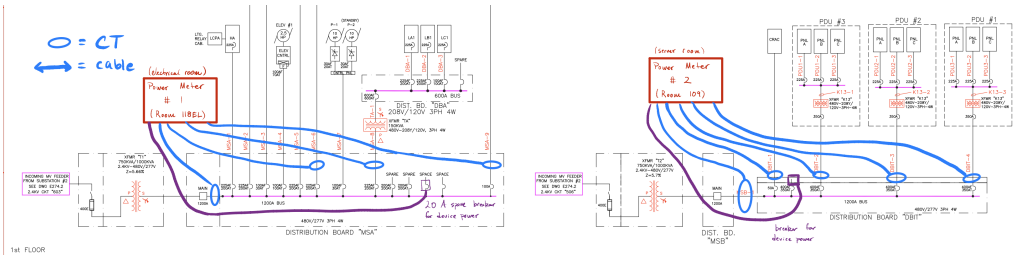


Figure 6.6: Schematic for installing meters on Annenberg Center MSA and MSB. CT connections are shown in blue and voltage connections are purple. The meters are protected by a small 20A breaker installed in a spare breaker slot in the switchboard or directly in the meter housing.

1° for phase angles.

Figure 6.7: 1600A Rogowksi coil CTs (orange loops) on the main lines feeding the busbars of MSB in Annenberg.

Voltage references were obtained either via the protection breaker in a spare slot in the switchboard or tapping directly into the copper busbars (Figure 6.8.

Each meter was enclosed in a polycarbonate case and connected to the switchboard via steel electrical conduit with separate pipes for high-voltage wiring and low-voltage wiring (Figure 6.9(a)). A small 20A breaker was installed inside the enclosure to protect the meter (Figure 6.9(b)). The meter itself is powered directly from the voltage taps and draws a negligible amount of power (max. $\sim$ 12W). Network connectivity was obtained via a Cat6e Ethernet connection directly to a Caltech campus network patch panel.

Following successful installation of the two meters in Phase 0 and initial validation

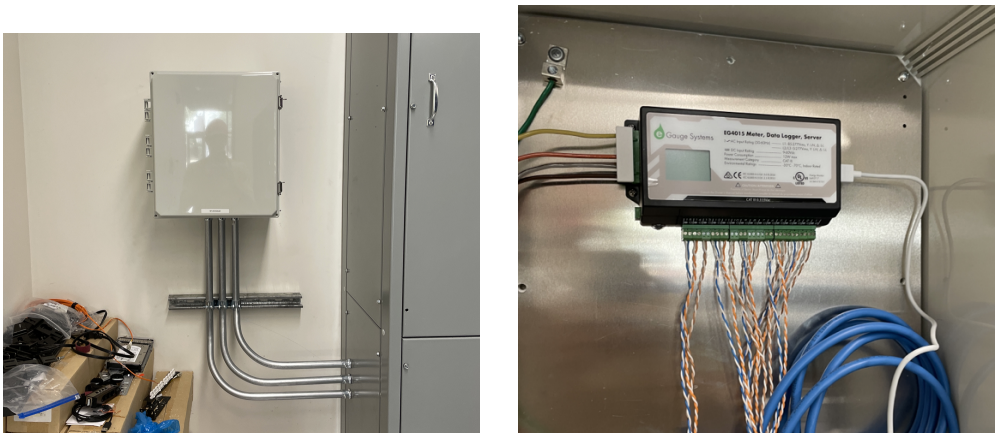
Figure 6.8: Installation of voltage taps via a 20A breaker in the main busbars of Annenberg Center main switchboard MSA.

of the data, we proceeded to Phase 1. The data collected in Phase 0 will be discussed in detail in Section 6.3.

Phase 1: Substation 3 Subnetwork

Phase 1 of the metering project consisted of installing an additional 20 eGauge meters on the sub-network of the Caltech campus microgrid serviced by Substation 3. This sub-network consists of 6 buildings, 2 solar arrays, 5 Bloom Energy natural gas fuel cells, 1 electrical substation, 1 EV charging array, and 3 grid interconnection points to the 17kV PWP distribution feeders. The buildings served by this substation are among the largest and most power-intensive on the campus and also had very poor data quality in the BMS, thus making this subnetwork a high-priority for metering. The redundant topologies of their switchgear (A and B main breakers connected by an open Tie breaker) often required that two meters be installed in every building, each with its own voltage reference. In addition, the injections from the PV and fuel cell DERs in the subnetwork are not measured and collected in the BMS, thus enabling the metering coverage from Phase 1 to fill gaps in the existing Caltech Facilities dataset.

The singleline diagram for Substation 3 sub-network is shown in Figure 6.10 with



(a) Meter enclosure (right) mounted on the wall and connected via electrical conduits to the main switchboard (left)

(b) Inside the eGauge meter enclosure. Power/voltage connections are on the left side of the device and CT sensor connections are on the underside. The enclosure chassis is grounded to the switchboard ground bar.

Figure 6.9: eGauge meter enclosure installed on the MSB main switchboard in Annenberg Center.

red stars showing the rough placement of the eGauge meters. To meter the medium-voltage 17kV lines in the substation itself, we used the existing control power transformers installed in the substation switchgear. Our strategy was to place 5A Continental Control Systems CTs on the existing current loops and to multiply our readings by the turns ratio of the already-installed control power transformers. We validated that this approach provided sufficient accuracy in Phase 0. We compared measurements from existing current loops of the CTs mounted within the Annenberg Center MSA switchboard with direct measurements from Rogowksi coil CTs on the main busbars. Errors between the two signals were no larger than sensor noise ($< 1\%$).

Due to the large number of meters that were installed in the Substation 3 network, we collaborated with Caltech campus information security (IMSS) to deploy a dedicated virtual local area network (VLAN) for the meters and our data storage server. This VLAN enables secure communication with the meters and protects our data collection infrastructure. As of the writing of this thesis, Phase 1 is complete with all 20 meters having been installed and commissioned.

Phase 2: Remainder of campus network

The remaining phase of Meter Caltech consists replicating what we did in Phases 0 & 1 in the remainder of the campus microgrid. This involves installing an additional

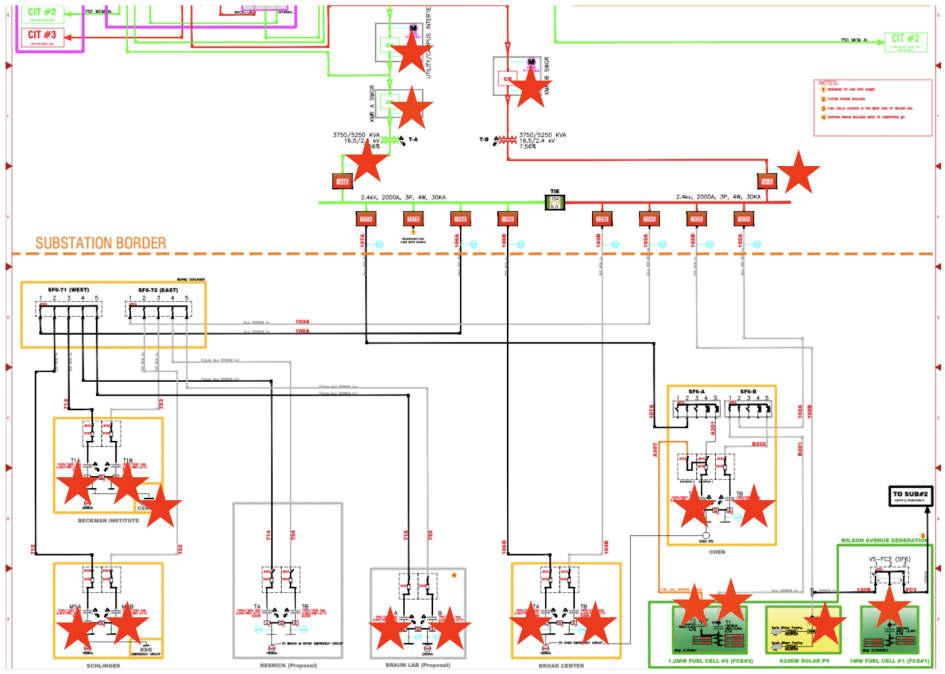


Figure 6.10: Singleline diagram of the Substation 3 sub-network with eGauge meter locations indicated by red stars.

100-120 meters. We will prioritize placing meters in the following locations:

- Remaining substations 1, 2 & 4 (including all PWP grid interconnection points);
- Central Plant and Satellite Plant generation assets and chilled water loads;
- Remaining Bloom Energy fuel cells (~2MW);
- Buildings and parking structures with rooftop PV arrays.

This prioritization allows us to fill urgent gaps in our timeseries dataset for the campus DERs and incoming grid power connections. Once these priorities are complete, which will require approximately 50 additional meters, we plan to complete the metering of the remaining building switchboards. For these building meters, we will explore using meters from different manufacturers (the Microgrid OS is hardware agnostic) and experiment with methods to make the installation and commissioning processes smoother and more scalable.

6.3 Smart Meter Data

The principal rationale for installing our own metering infrastructure was to obtain reliable, comprehensive electrical data for all of the power injection (generation and load) sources in the network. In particular, the quantities are collecting are:

- Per-phase current RMS magnitudes and phase angles;
- Phase-to-neutral (or phase-to-phase) voltage RMS magnitudes and phase angles;
- Per-phase power factor;
- Per-phase real and reactive power injections (where possible).

Note that the phase angles are all with respect to an onboard reference for each meter (typically the $L1$ channel, which corresponds to phase A voltage V_a).

An example of the total three-phase mains power data from Annenberg MSA is shown in Figure 6.11.

Figure 6.11: Total 3-phase power on Annenberg MSA switchboard. The upper green curve is the power generated by the rooftop PV array and red curve is the aggregate building load.

The waveform capture function of the eGauge smart meter allows current and voltage signals to be recorded in the time domain over short durations. An example of this data is shown in Figure 6.12. As would be expected, the voltage waveforms are corrupted by very little noise and maintain a consistent 120° offset between the phases. On the other hand, the current waveforms have significant distortion and fluctuate rapidly in response to changes in building load.

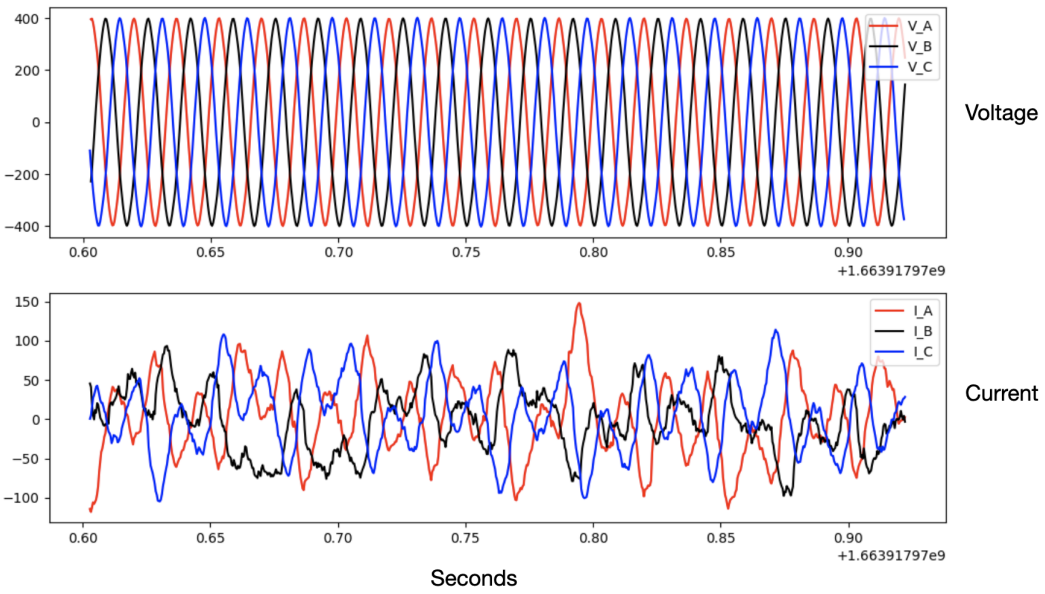


Figure 6.12: Waveform capture data for a 0.35 second window. The horizontal axis are Unix Epoch timestamps.

Using the eGauge Python API, we collect magnitudes and angles every second in our timeseries database as well as 1-second-long waveform captures every 5 seconds for all current and voltage channels.

Phasor data synchronization

One of our principal objectives in installing the eGauge meters was to implement them as a low-cost phasor measurement units (PMUs). This capability is not present in standard revenue power meters such as the Schneider Electric PM800 that is prevalent in the Caltech system. A PMU's main function is to record phasors in different locations in the network with respect to a common time reference (schematic in Figure 6.13). For high-quality PMUs, this tight time synchronization is accomplished using a GPS signal. In our case, we conducted experiments to understand whether standard internet timekeeping protocols Network Time Protocol (NTP) and Precision Time Protocol (PTP) could provide sufficient accuracy.

In order to limit measurement errors due to meter synchronization to less than 1° (or 0.28%), it is necessary to have measurement timestamps aligned with an accuracy of less than 0.1 milliseconds (ms). To validate that this was possible, we conducted the following experiment. We sent 150 repeated 0.25-second long capture requests every 1 second to the L3 channels(C-phase voltage) of two eGauge meters in the Chen Neuroscience building. The signals were recorded simultaneously on both

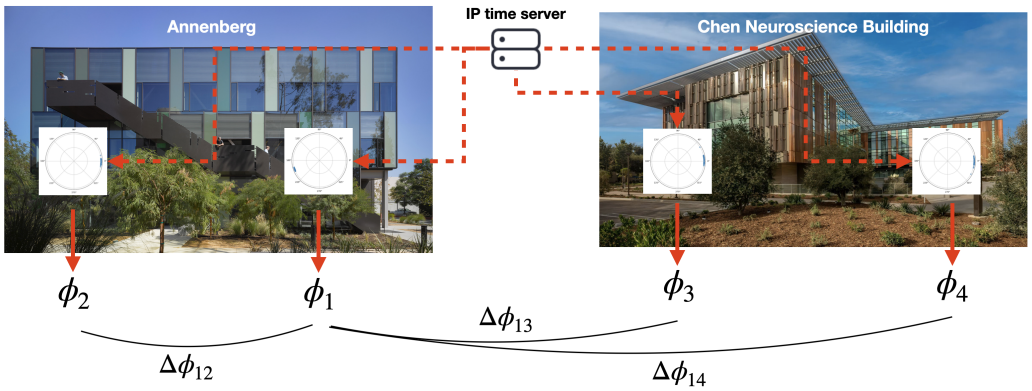


Figure 6.13: Schematic of phase offsets between meters in different locations within the electrical network, synchronized via network time servers.

meters and phase angles were calculated using the Fast Fourier Transform method. Given the stability of the voltage waveforms, we observed in separate experiments on the same device that the phase of any given waveform would drift negligibly over the period of several minutes. We then sought to observe the distribution of phase offsets between the signals on the different meters over the measurement window. To obtain sufficient measurement accuracy, it was necessary to have high-precision measurements over the 2.5-minute capture window in the form of small variance of the phase offset distribution. Our results are shown in Figure 6.14. The variance of the distribution we observe is 0.1 degrees, which suggests phase alignment at or below the tolerance required on average. This was both the case when timestamps were provided via PTP and when NTP was used (although in this case, the server polling interval needed to be quite small, otherwise the drift of the internal meter clocks would be unacceptably large). The initial synchronization experiments have been promising and we continue to conduct further tests to validate the reliability of the synchrophasor measurements for the other meters in the network and under various communication network loading conditions.

With sufficient time synchronization accuracy, a network of eGauges, such as the one we have installed in the Substation 3 subnetwork, produce a set of synchrophasors. This kind of data has many critical downstream use cases, such as harmonic disturbance detection and algorithmic power system state estimation (see for example the author's work in [123, 77]. Despite its utility, our dataset is the first of its kind (as far as the author is aware) due to the expense of traditional PMU hardware deployment. In typical use cases in low-voltage systems, a small number of PMUs (costing $10,000+ each) are deployed temporarily in network feeders to collect data

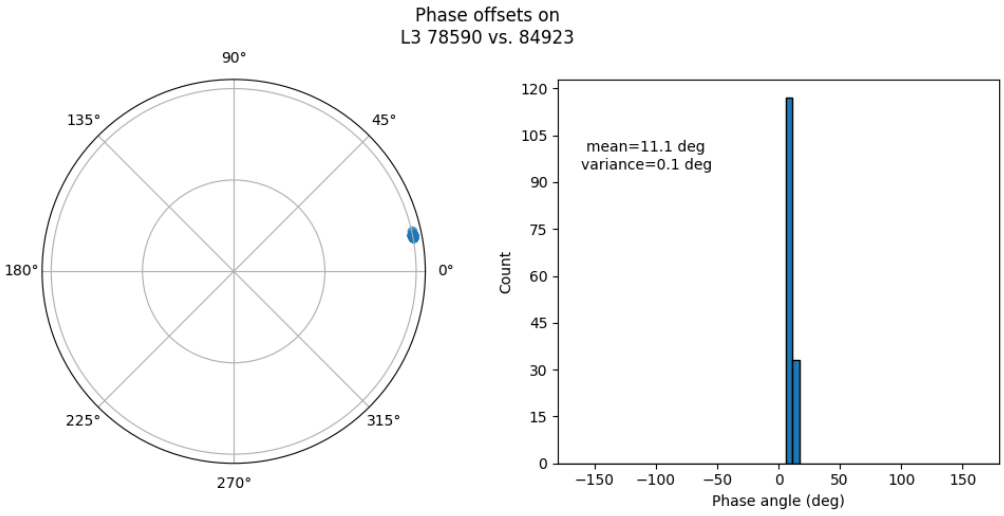


Figure 6.14: Distribution of phase offsets of the *C*-phase voltages of two meters in the Chen Neuroscience building. The meters are measuring voltages on different buses, leading to the average 11 degree phase difference between the signals. The small variance of the measurement distribution suggests accurate meter timestamp alignment.

on suspected power quality issues. In our deployment, data is collected continuously from our meters whose total installed cost is approximately $5,000 ($1500 for hardware, $3500 for installation). The dramatically lower cost of our hardware and the permanent installation of the meters on switchboards allow us to build a dataset that is novel for distribution networks.

6.4 Network Components and Connectivity

Along with the timeseries data discussed in the previous sections of this chapter, the other main data type in our models is network assets and their connectivity. The principal way of storing this data in practice is in set of electrical diagrams called single-line diagrams. A single-line diagram is a simplified network drawing that represents the electrical connections and components of a power system or electrical distribution system in a single line format. It is a basic representation of the system that uses symbols to represent the major components, such as generators, transformers, circuit breakers, switches, and loads, and their connections.

The purpose of a single-line diagram is to provide a clear, concise overview of the power system or distribution system, showing the flow of electrical power from the source to the loads, and the protection and control devices used to manage the flow of power. The diagram does not provide detailed information on the physical

layout or specific details of the components, but rather presents a simplified view of the system for ease of understanding. Thus the single-line diagram is both a form of asset database and and network configuration diagram. In practice, these diagrams are the most reliable and oftentimes only record of the electrical system configuration. Single-line diagrams are generated using CAD software but are then converted to pdf images or printed out and stored (see an example in Figure 6.10). Because of their data format, singleline diagrams are challenging to maintain as the network configuration evolves over time and therefore, the actual system state can be significantly different than what is contained in the diagrams.

The task for this part of the data pipeline was converting the single-line diagrams from images to computer-parsable network models which can be more easily maintained and updated. We followed several steps to complete this task. First was manual annotation of a set of single-line diagrams of the Caltech network. Second was defining a set of component models for the elements in the system. Third was translating the manual annotations into a structured element schema with consistent naming and tagging of elements. Finally, we discuss ongoing work on automated methods for parsing the single-line diagrams.

Manual annotation

Our single-line diagram dataset for the Caltech microgrid consists of seven separate diagrams that cover the four substations, Central and Satellite plants, and connecting medium voltage lines. The diagrams include all elements down to the main switchboards (480V or 208V level) in each building. We focused on two tasks for the annotations: 1) identifying the set of elements present in the system and 2) defining a name for each element. This process was highly tedious and took a team of 4 people several months to accomplish (see Figure 6.15 for an example of a single building). Lessons learned motivated us to pursue automated methods for parsing the diagrams.

Component models and model schema

Using our learning from the manual annotation process, as well as conventions from standard distribution system modeling packages such as OpenDSS [103], we defined a core set of power system components that spanned the following categories:

- Electrical system components (e.g., buses, lines, transformers, breakers, fuses)
- Generation and load sources (e.g., generators, PV, batteries, motors, loads)

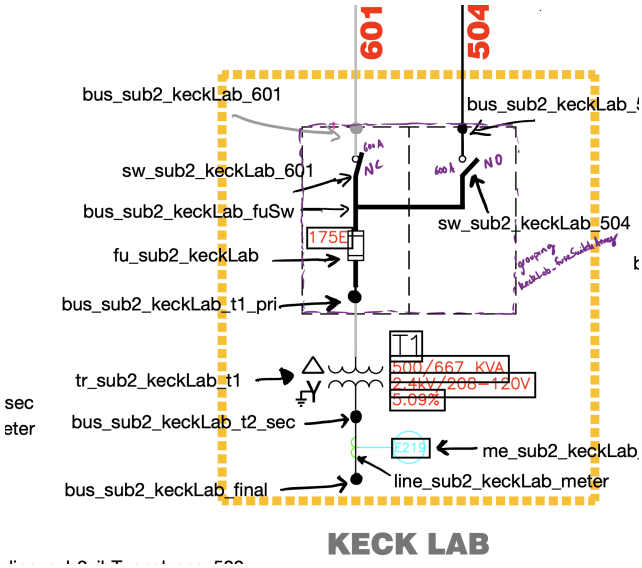


Figure 6.15: Annotated single-line diagram for Keck Lab

- Aggregations of components (e.g., switchboards, panels, buildings)

These components were formalized into a defined `json` object schema and all of the elements in the Caltech diagrams were manually written into a set of `json` files. Consistent naming and a formal schema allowed network structure and connectivity to be coded into the definition of the elements. For example, the from-bus and to-bus of a line element are linked by unique name strings to their bus objects which are defined elsewhere in the `json` file. The `json` schema is useful for formalizing the attributes of each component and introducing basic type checking and formatting constraints. An example of an object definition for a PV array is shown in Figure 6.16

Automated object detection

Due to the structured form of the single-line diagrams and the tedious, time-consuming process of manually annotating and converting them, we are currently exploring methods to accelerate the process using annotation tools and computer vision techniques.

One thrust of our approach is using computer vision methods such as template filtering and convolutional neural networks to perform object recognition on the model elements present in singleline diagrams. A second thrust is an human-in-the-loop annotation pipeline to correct errors in the automatically detected objects and to refine the model definition using expert input. These components interact with

```
"PV": {
    "type": "object",
    "additionalProperties": false,
    "required": [
        "name",
        "inverter",
        "nameplate_capacity"
    ],
    "properties": {
        "name": {
            "$ref": "#/$defs/types/noSpaceStringType"
        },
        "groups": {
            "$ref": "#/$defs/types/noSpaceStringArrayType"
        },
        "color": {
            "$ref": "#/$defs/types/colorType"
        },
        "nameplate_capacity": {
            "$ref": "#/$defs/types/powerType"
        },
        "dc_voltage": {
            "type": "number"
        },
        "geo_location": {
            "type": "string"
        },
        "inverter": {
            "$ref": "#/$defs/types/noSpaceStringType"
        },
        "info": {
            "type": "string"
        }
```

Figure 6.16: `json` schema object for a PV array.

each other to leverage the scalability of computer vision algorithms while retaining the flexibility of human annotation. Figure 6.17 shows a flowchart of the detection-annotation-model software pipeline currently being developed. The input into the pipeline is a single-line diagram image and the output is a set of model `json` files.

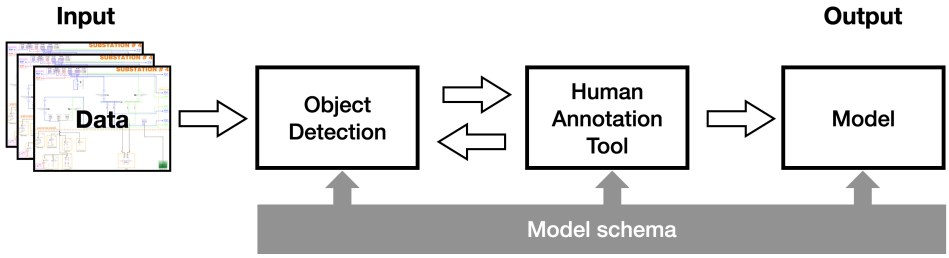


Figure 6.17: Flowchart for human-in-the-loop single-line diagram annotation pipeline.

Thus far, we have developed the annotation tool and are using it to build a labeled

dataset that can be used to train and evaluate object detection methods. However, just on its own, the annotation tool speeds up the human annotation process by greater than 20x, reducing the author's time to annotate a building's singleline diagram from 6 hours to 20 minutes. The interface we have developed is shown in Figure 6.18.

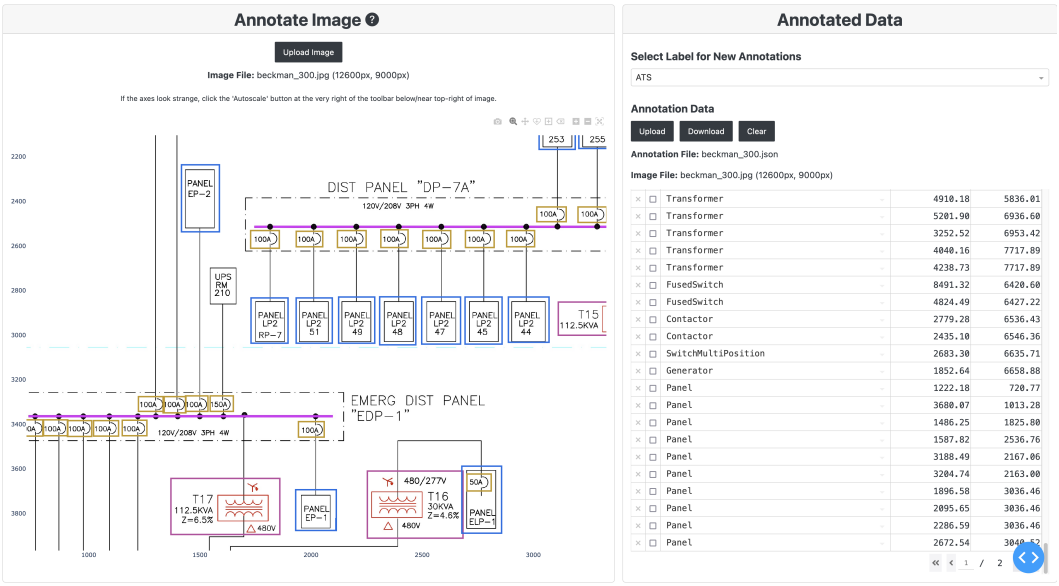


Figure 6.18: Human-in-the-loop annotation tool for single-line diagrams. The model schema is an input into the tool, limiting the available elements the annotator may select from to elements in the model.

By inserting algorithmic enhancements such as text recognition and object recognition into the annotation pipeline, we seek to reduce the need for human annotation to near-zero and provide a completely automated method for producing reliable, labeled computer-readable network models.

Chapter 7

MICROGRID DIGITAL TWIN

This chapter presents the software system we have designed and built to integrate timeseries and network data into a cohesive digital model of the energy system. Building on top of the timeseries and network model data discussed in Chapter 6, the Digital Twin provides a multi-scale model—from component to system level—of the energy system. Again, although the initial design and deployment of the Digital Twin was done in the Caltech system, the software is generic to any distribution system and we intend to deploy it to other campuses in the near future.

The inputs to the Digital Twin platform are data sources; specifically, timeseries and network model files. The output of the platform is a cohesive, integrated system model that users can interact with, visualize, and optimize. To achieve this, we have implemented and are continuing to develop the following key functional blocks:

1. Data interfaces;
2. Data processing modules
3. Data integration
4. Model learning
5. Visualization & interaction.

The high-level interactions between these blocks are illustrated in Figure 7.1.

The data interfaces allow any measurement device to stream data into the Digital Twin timeseries database. The data processing modules perform error detection, cleaning, and organizing of the ingested data. Data integration combines data from different sources (e.g., different types of meters, network models, GIS data) and combines them into a single, relational database with consistent data conventions. Model learning learns patterns or parameters of the existing data; examples include applications timeseries forecasting and line admittance estimation. Visualization is a user-facing interface that allows planners and operators to explore the real-time and historical states of the entire energy system.

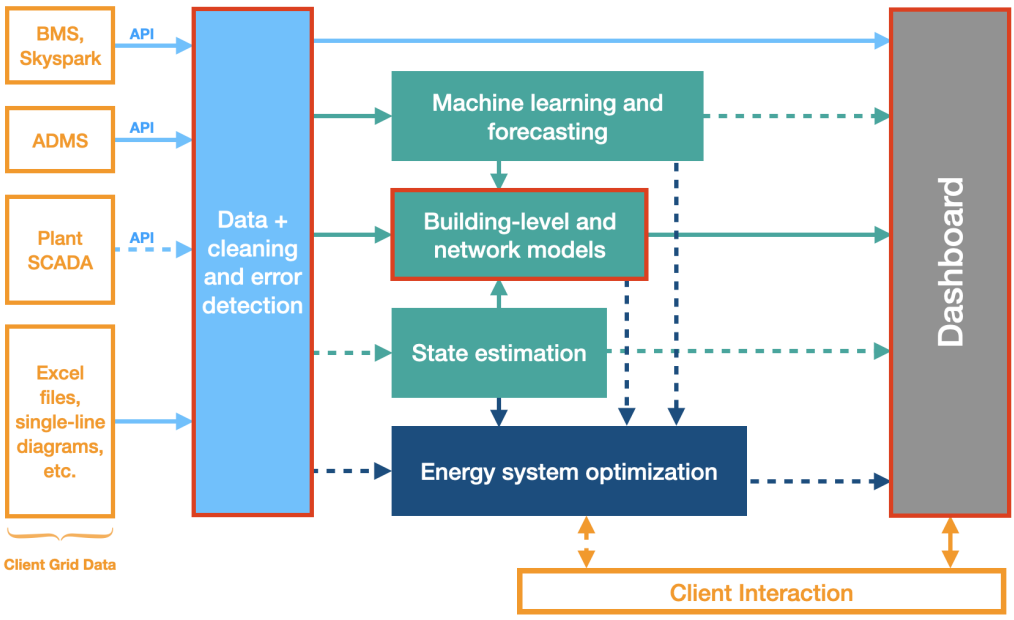


Figure 7.1: Digital Twin system architecture diagram

Implementation details of the Digital Twin are not the focus of this chapter. Nonetheless, implementing the system described here is a major software development effort. Figure 7.2 shows a block diagram of the compute, data storage, and and software modules that comprise our prototype production environment. We are leveraging virtualized computing (VMWare) for data processing and web hosting, an onsite network attached storage system (NAS) for timeseries data storage, and Amazon Web Services (AWS) S3 storage for archival data storage. Our codebase is mostly Python, with the front end dashboard based on the Dash and Plotly frameworks. Timeseries are stored in organized directories of `csv` and `parquet` files.

The Digital Twin platform is designed to be extensible and modular, with a base data model easily supporting the addition of new functionality. The work presented in this thesis does not represent an end point of its development and a team of software engineers and graduate students continues to develop it actively. In the rest of this chapter, we will discuss our work so far in each of the five functional blocks listed above.

7.1 Data interfaces

Data interfaces are a set of APIs that communicate with databases and hardware. Each of these APIs need to be customized to the particular requirements of the asset it interfaces. However, the output of each API, regardless of its input, produces data

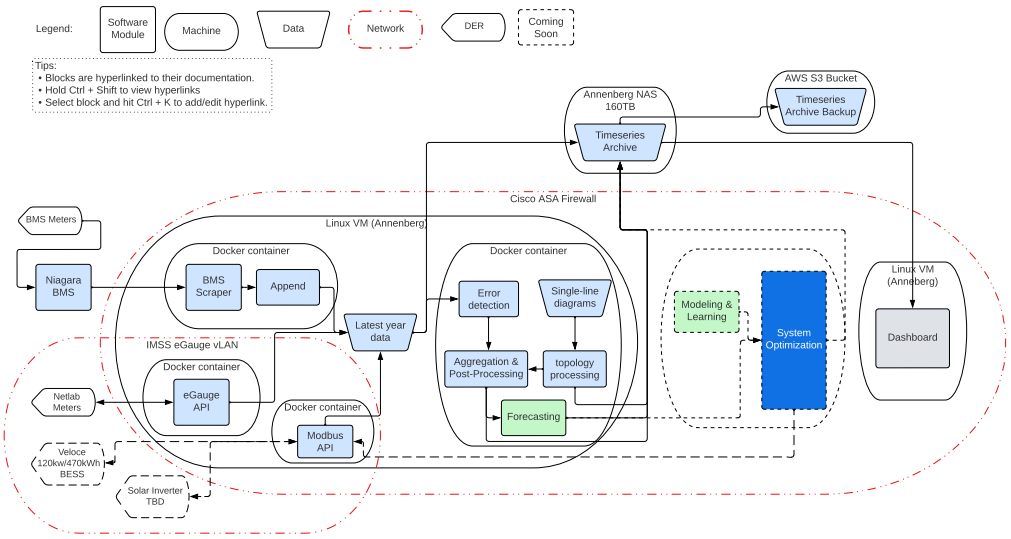


Figure 7.2: Digital Twin software system architecture block diagram.

in a consistent format internal to the Digital Twin.

We first implemented an HTTP-based API to collect time series data at regular 15-min intervals from the Caltech building management system Histories database (Niagara Iridium framework). Implementation challenges involved automated credential authentication and parsing of the native BMS data format.

We then implemented APIs for Egauge meters in a similar fashion, although rather than a single database interface, each smart meter required individual authentication and data pull requests. We are in the process of building similar APIs for other smart meter brands.

APIs for hardware control are also under development. To control DERs such as battery storage systems, we are designing and implementing Modbus TCP-based interfaces to be able to read register tables and write setpoints to them.

7.2 Data Processing

The data processing functionality of the Digital Twin ingests the raw timeseries data and subsequently cleans, transforms, and analyzes it make to make it suitable for monitoring, analysis, and control. This automated data correction pipeline is necessary to make the data from meters and single-line diagrams useful in real time. Some of the capabilities we have implemented are

- Data cleaning: identifying and correcting errors, outliers, or missing data

points in the timeseries data through interpolation and resampling;

- Pattern recognition: grouping similar timeseries data based on their characteristics, such as their patterns, trends, or seasonality;
- Anomaly detection: identifying unusual patterns, events, or outliers in the timeseries data that deviate from the normal behavior of the system;
- Data transformation: converting or rescaling timeseries data into useful formats. Examples include normalization, aggregation, and unit transformations;
- Physics-based checks: applying physical sanity checks on the data (e.g., power factor magnitudes must be less than 1 and must be consistent with real and reactive power magnitudes).

An example of our anomaly and pattern detection implementation is displayed in Figure 7.3. We use the autocorrelation of the signal to infer the presence of seasonal patterns and thresholds tuned on the data to identify likely outliers or missing data.

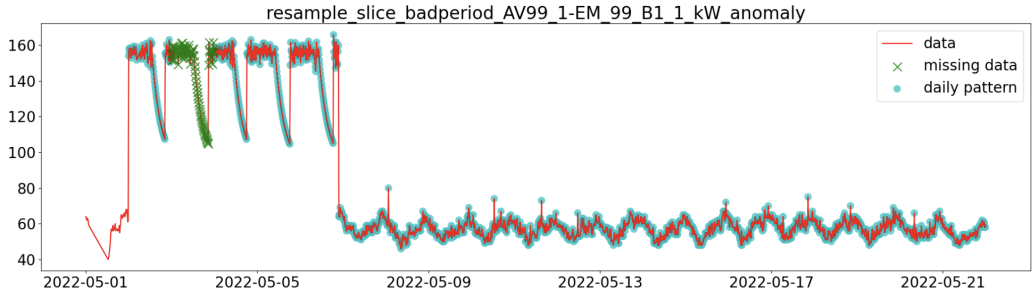


Figure 7.3: Detection of daily seasonality (blue) and outliers/missing data (green) in a power (kW) timeseries for a Caltech building.

7.3 Data integration

The motivating principle of the Digital Twin platform is to synthesize the heterogeneous data produced by and related to energy systems into a cohesive system model. One of the critical steps in accomplishing this is integrating data of diverse types from diverse sources. These types include

- Timeseries data (e.g., meter data, weather data);
- Geographic (GIS) location data;
- Electrical network diagram image files;

- Tabular asset databases (e.g., excel files).

Additional data sources that we have worked thus far are power system models in standard formats (e.g., OpenDSS, Cyme, Powerworld, MatPower).

The key to having all of these data sources "talk" to each other is to perform semantic indexing. Semantic indexing is the process of associated a single, internally consistent set of names (or keys) with all of the data objects. This key set can then be used to link data objects with each other in an automated fashion.

In our model, the hub of the semantic indexing system is the network model introduced in Section 6.4. This model leverages the grammar of `json` Schema to define element, data and object types and serves as the single source of truth of object naming. Any data source that joins the platform's database must first be specified and labeled in the network model schema. The use of a file schema enables easy application of units tests to each new element (to check correctness) and to the entire network model (to validate semantic consistency across elements). Further, it allows the model itself to be stored and distributed in a human-readable and widely-supported `json` format.

For an example, the timeseries data representing the power production of the PV array shown in Figure 6.16 would have a label linking it to the (unique) name of this PV array in the `json` files. Although seemingly trivial, implementing a consistent set of labels is the single most important step in data integration. In our real-world experience in power systems, poor or inconsistent labeling is the largest painpoint in building high-quality models.

The process of funneling raw data into a single set of semantic labels is non-trivial and its exact requirements will depend on the project or application. Defining a network model for a set of computerized, well-labeled databases is a straightforward, computer-based data parsing task. When the data are unorganized, not digitized, or poorly maintained, some degree of manual annotation is unavoidable.

For the Caltech microgrid deployment of the Digital Twin, we have combined manual and automated annotation to merge the following sets of data into our model:

- 15-minute electrical and thermal timeseries data from the BMS
- 15-minute historical timeseries data on DER generation from Excel files

- 1-second electrical timeseries data from eGauge smart meters (includes magnitudes and waveform/phasor data)
- GIS building data from Open Street Map [104]
- Manually annotated building and DER GIS data
- Paper and `pdf` single-line diagrams
- `csv` files of building-to-meter mappings and DER inventories

Based on the learning from this data integration process, we have begun implementing some human-in-the-loop (HIL) tools to speed up human annotation where necessary (see Section 6.4 for more detail). It is our observation that HIL tools can offer over 10-100x speedups in certain tasks that are not yet amenable to end-to-end automation (e.g., network diagram parsing). Further development of the Digital Twin will focus on developing more of these tool modules in the data interface and data processing layers. In particular, we will prioritize 1) tools to annotate and name the elements in the network model and 2) APIs to connect with DERs and sensors that support the Modbus protocol. The Modbus protocol is nearly universally supported in industrial automation and IoT applications and will allow the Digital Twin to scalably connect to a wide range of DERs.

7.4 Models from data

Building upon the preceding data processing and assimilation steps described thus far, we apply learning and inference techniques to derive additional structure and insight into the network model. The particular learning methods depend on the downstream applications. In this section, we discuss two particular applications that are the focus of ongoing research or software development: timeseries forecasting and topology/parameter estimation.

Example: timeseries forecasting

Time series forecasting is the process of predicting future values of a time series based on its past behavior and possibly a set of contemporaneous covariates. A time series is a sequence of data points recorded over time, typically at regular intervals, such as hourly, daily, weekly, or monthly. Timeseries forecasting is a core task in planning, operating, and analyzing energy systems with the main applications being load forecasting and renewable energy production forecasting. Applications that are market facing involve forecasting of electricity prices.

Timeseries forecasting methods generally fall into two categories: point forecasts and probabilistic forecasts. A point forecast is the forecaster's single best prediction of the variable of interest at some point in the future. In the probabilistic setting, the variable of interest is modeled as a random variable and a probabilistic forecast takes the form of a predictive distribution over future outcomes. The Digital Twin software architecture supports any forecasting algorithm that accommodates the input/output formats of the API.

Figure 7.4 displays our implementation of an autoregressive ARIMA point forecasting algorithm on the aggregate Caltech campus chilled water production. Using weather forecasts as covariates, we are able to forecast the series accurately up to seven days in advance.

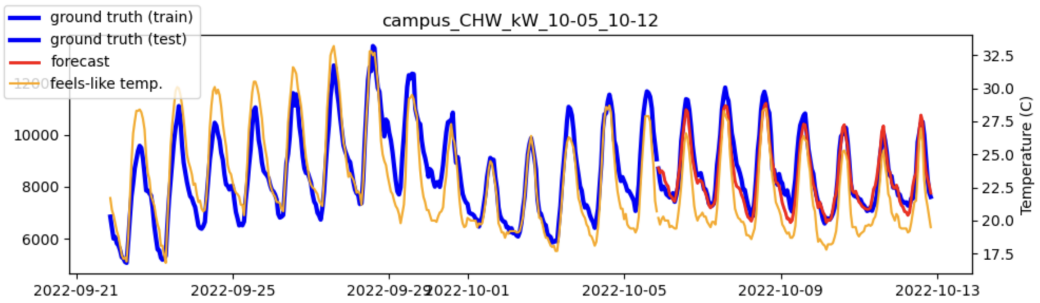


Figure 7.4: Point forecast of Caltech's campus chilled water load. An ARIMA variant, SARIMAX, uses weather forecasts as covariates and accounts for strong daily seasonality in the timeseries.

Probabilistic methods for energy forecasting are the objects of increasing interest in the research literature and are being called for by operators who are in need of quantifying uncertainty in forecasts of key system parameters. The author's recent work on hierarchical probabilistic forecasting using deep autoregressive forecasting methods established the state-of-art performance in this task, which is particularly suited to the structured setting of energy forecasting [113]. We are currently working to integrate probabilistic forecasts into the Digital Twin data processing pipeline.

Example: topology and parameter state estimation

Even when singleline diagrams of the electrical system are available, the difficultly in maintaining them over time typically results in the electrical topology being quite different in reality from what the diagrams suggest. Research in the last decade on the problem of topology and parameter estimation from data has resulted in

numerous algorithms for recovering the network state from current, voltage, and power state measurements.[1]

We are developing several algorithms to perform network state estimation using the data collected by the Digital twin, focusing on approaches for sample efficiency [77, 78] and estimation accuracy [123]. The data necessary to support such algorithms typically consists of synchrophasors for all power injection sources (we collect this data from eGauge meters; see Section 6.3) in the connected network. The inferred topology can be used to identify errors in switch and breaker statuses in the network model (e.g., Figure 7.5 as well as estimate unknown 3-phase line impedances.

7.5 Visualization

We have implemented a dashboard interface for the Digital Twin to allow users to interact and explore data and models. The dashboard is a critical tool for visualizing the timeseries data and for making the data accessible/downloadable for further detailed analysis. It also makes the connection between timeseries data and the network model explicit by allowing users to click on components in the singleline diagrams and visualize the timeseries data associated with them (e.g., the energy generation timeseries from a PV array).

The dashboard is also a platform that supports human-based editing of the network model. We have found that clear, comprehensive visualization drastically speeds up the model review process for expert annotators. Therefore, to meet this need, we immediately render whatever functionality that is implemented in the back end (e.g., anomaly detection, forecasting, network modeling) through the dashboard.

[1] Despite the algorithmic progress, few of these algorithms have actually been validated on real-world settings due to the challenges of acquiring synchrophasor datasets.

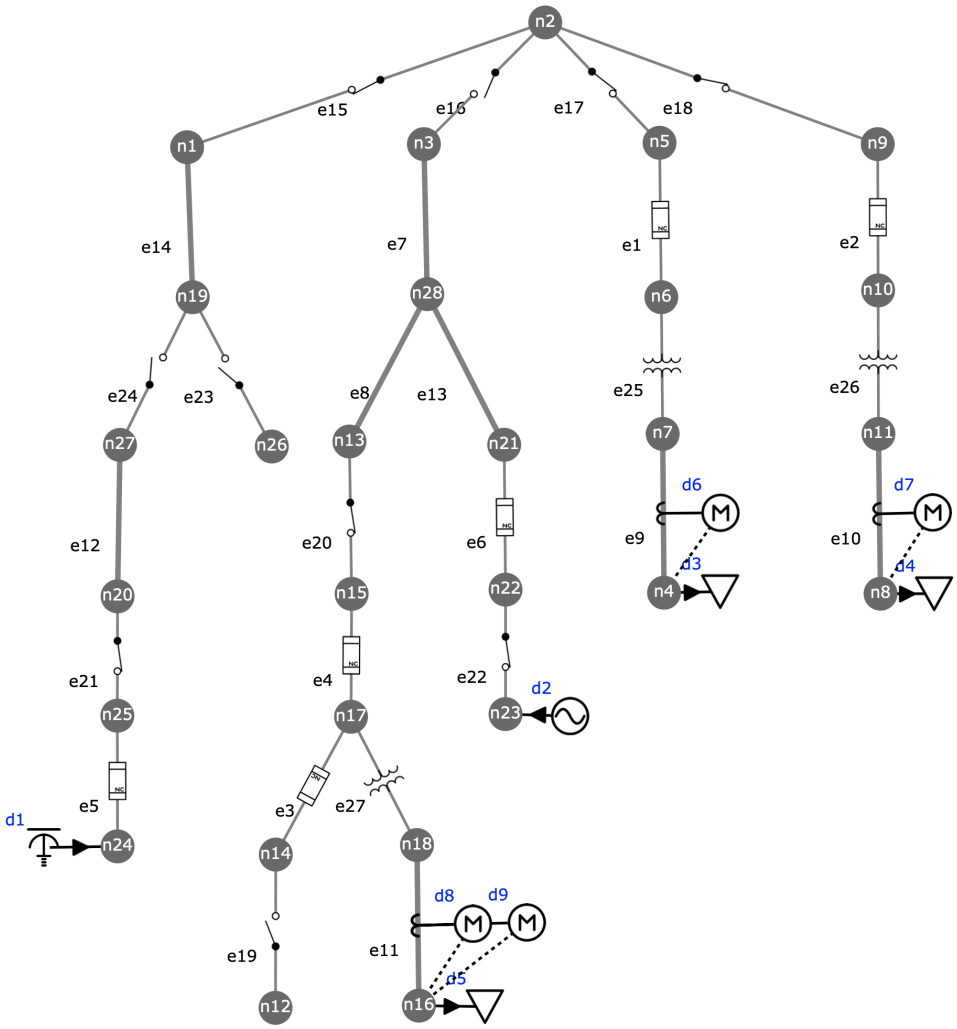


Figure 7.5: Computer-parsable singleline diagram generated from the network schema for the Kerckhoff Building at Caltech. Switch and breaker positions are taken from singleline sources but may not represent actual positions. State estimation algorithms can be used to validate or correct their statuses.

Figure 7.6: Caltech Digital Twin dashboard system overview

Chapter 8

OPTIMIZATION AND CONTROL OF DERS

In this final chapter of the thesis, we present three projects related to the optimization and control of microgrids and distribution grids. These projects are examples of energy services applications that sit on top of the data and model layers in the Microgrid OS in Figure 5.5.

In Section 8.1, we introduce a methodology for optimizing the infrastructure investment for a grid-connected microgrid participating in wholesale energy markets. We incorporate both physical and financial constraints (e.g., debt financing, market constraints) into the problem formulation to optimize the payback period for the microgrid site.

In Section 8.2, we show how to optimally size a battery system to provide resilience (as well as cost and emissions reductions) to a commercial building with rooftop PV. We optimize the cost of the battery system to meet the desired level of reliability and outage duration for the building's critical loads.

In Section 8.3, we detail a novel approach to voltage control in a large distribution feeder using distributed energy storage assets. We solve an optimal battery location problem while accounting for multi-phase power flow equation constraints, physical constraints on voltage and line flows, and physical constraints on the locations of the batteries. It is shown that controlling the charge and discharge of the fleet of optimally placed batteries during peak load conditions can steer the voltage deviations at customer nodes to be within nominal ranges.

8.1 Microgrid co-design in electricity markets

The purpose of this section is to model the finances of the installation and operation of a zero-carbon microgrid consisting of solar (PV), battery energy storage, and commercial building load. The goal will be to model both microgrid and market operation of the system and to optimize the installed capacities of PV and battery, as well as the market operations over the life of the system.

Assumptions

We make the following assumptions about the microgrid model and the market participation:

- Solar will be placed on the roof of an existing commercial building(s) and that there is sufficient space on site for the batteries. The property acquisition cost and/or recurring least cost will not be included in investment or operational costs.
- We assume that building load, solar, and battery injections feed into the same bus in a main switchboard. This basically means that there is a single-bus network with no losses. Powerflow equations reduce to power balance.
- The battery cannot charge from the grid.
- When the power produced onsite by the microgrid is insufficient to meet building load, power can be drawn from the utility grid at the volumetric retail rate.
- Market participation and consumption from the utility grid cannot happen simultaneously.

Notation

See Table 8.1.

Vector quantities denoted in boldface (e.g., $\mathbf{x}$). The t-th energy of $\mathbf{x}$ is denoted x_t. In some of the formulations we need to introduce auxiliary variables. Auxiliary binary variables are denoted with z_t and auxiliary continuous variables are denoted with v_t. Appropriate superscripts will be added to these variables to distinguish them.

Capacity and operational optimization

We will go through the constraints for the system (i.e., powerflow), the PV constraints, the battery constraints, and the balance-of-system (BOS) constraints.

Constraints

The units of the $\mathbf{u}$ and $\mathbf{x}$ power injection variables are given as watts (W). This number represents the *average* power injection over interval t. Care must be taken when converting from these variables to quantities with units of energy (Wh) such as $\mathbf{u}^{\text{soc}}$ that the correct conversion is done (will depend on the length of the interval). When

Variable	**Domain**	**Unit**	**Description**
|=|=|=|=| t	$[1, \ldots, T]$	hour	index of hourly interval
i	$[1, \ldots, N]$	year	index of year
T	$\mathbb{Z}_+$		number of hours in the simulation. T is not necessarily $8760 \times N$
N	$\mathbb{Z}_+$		number of years in the simulation (e.g., 25)
$\mathbf{u}^{\text{batt}}$	$\mathbb{R}^T$	MWac	Battery charge (−) or discharge (+). t-th entry is u_t^{batt}
$\mathbf{u}^{\text{pv}}$	$\mathbb{R}_+^T$	MWac	PV production t-th entry is u_t^{pv}
$\mathbf{u}^{\text{load}}$	$\mathbb{R}_+^T$	MWac	Building load t-th entry is u_t^{load}
$\mathbf{u}^{\text{soc}}$	$\mathbb{R}_+^T$	MWh	Battery state of charge t-th entry is u_t^{soc}
$\mathbf{x}^{\text{DA}}$	$\mathbb{R}^T$	MWh	Energy sold to day-ahead market
$\mathbf{x}^{\text{RT}}$	$\mathbb{R}^T$	MWh	Energy sold to real-time market
$\mathbf{x}^{\text{PPA}}$	$\mathbb{R}^T$	MWh	Energy sold to PPA offtaker
$\mathbf{x}^{\text{DSP}}$	$\mathbb{R}^T$	MWh	Energy procured from distribution service provider (DSP)
$\mathbf{x}^{\text{REC}}$	$\mathbb{R}^T$	MWhg	Quantity of renewable energy sold as a renewable energy credit (REC)
$\boldsymbol{\beta}$	$\mathbb{R}_+^T$	MWac	Random variable representing the building load trajectory
$\boldsymbol{\alpha}$	$[0, 1]^T$	MWac	Random variable representing the solar production as a fraction of its nameplate capacity
r^{xfmr}	$[0, 1]$		Transformer loss (1 = no loss)
r^{ITC}	$[0, 1]$		fraction of upfront capital cost eligible for investment tax credit
$r^{\text{tax}}_{\text{sales}}$	$[0, 1]$		Sales tax rate on capital investment
$r^{\text{tax}}_{\text{state,income}}$	$[0, 1]$		State tax rate on income
$r^{\text{tax}}_{\text{federal,income}}$	$[0, 1]$		Federal tax rate on income
r^{debt}	$[0, 1]$		Fraction of investment costs serviced by debt

Table 8.1: Definition of variables and parameters. c_t and x_t are variables. All other symbols are parameters.

we enforce power balance, for example, equality of average power injections over each interval is being enforced, as opposed to equality of instantaneous injections.

Power balance The powerflow equations consist of the power balance at the point of common coupling (PCC) with the grid.

$$r^{\text{xfmr}}(u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}}) = x_t^{\text{RT}} + x_t^{\text{PPA}} - x_t^{\text{DSP}} \qquad \forall t \qquad (8.1a)$$

All of the quantities in (8.1) are positive. The signs preceding each term are determined by the convention illustrated in Figure 8.1.[1] $r^{\text{xfmr}} \in [0, 1]$ is the loss associated with transformer stepup.

The x^{RT}, x^{PPA}, and x^{DSP} quantities cannot be sold simultaneously (they must each represent a "unique" Wh). However, x^{REC} can be sold unbundled for the same Wh that is sold in the power markets. We can theoretically sell as many RECs as there are Whs produced by the solar array. Since we enforce that the battery only charged from the PV and not from the grid, we can simply sum up the microgrid u quantities to get the clean power exported to the grid.

$$\max\{0, u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}}\} = x_t^{\text{REC}} \qquad \forall t$$

The max expression is necessary because sometimes the microgrid imports energy from the distribution grid to feed building load (although not to charge the battery as enforced by a complementarity constraint below). Unfortunately this constraint is non-convex (equality constraint with ReLu). We will propose a lifted mixed-integer linear equivalent.[2]

There are two cases.

Case 1: $u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}} \leq 0 \Rightarrow x_t^{\text{REC}} = 0$**.** Introduce a binary variable $z_t^{\text{rec},1} \in \{0, 1\}$ and take $z_t^{\text{rec},1} = 1 \Rightarrow x_t^{\text{REC}} = 0$. The set of constraints that enforces this case is

$$0 \leq x_t^{\text{REC}} \leq M(1 - z_t^{\text{rec},1}) \qquad \forall t \qquad (8.2a)$$

$$u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}} \leq M(1 - z_t^{\text{rec},1}) \qquad \forall t \qquad (8.2b)$$

The value of M will be chosen so that it doesn't constrain the values of x^{REC} or the u variables when $z_t^{\text{rec},1} = 0$.

Case 2: $u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}} > 0 \Rightarrow x_t^{\text{REC}} = u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}}$**.** Introduce a binary variable $z_t^{\text{rec},2} \in \{0, 1\}$ and take $z_t^{\text{rec},2} = 1 \Rightarrow x_t^{\text{REC}} = u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}}$. The set of

[1] In general, the "u" variables represent quantities within the microgrid and the "x" variables represent quantities on the grid side.

[2] Equivalent in the sense that any u and x satisfying the max formulation will also satisfy the lifted version, and vice versa.

constraints that enforces this case is

$$-M(1-z_t^{\text{rec},2}) \le x_t^{\text{REC}} - u_t^{\text{batt}} - u_t^{\text{pv}} + u_t^{\text{load}} \le M(1-z_t^{\text{rec},2}) \qquad \forall t \qquad (8.3a)$$

$$u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}} \ge -M(1-z_t^{\text{rec},2}) \qquad \forall t \qquad (8.3b)$$

Finally we want one case or the other so we enforce

$$z_t^{\text{rec},1} + z_t^{\text{rec},2} = 1 \qquad \forall t \qquad (8.4a)$$

$$z_t^{\text{rec},1}, z_t^{\text{rec},2} \in \{0,1\} \qquad \forall t \qquad (8.4b)$$

$M \ge 0$ is a non-negative constant that should be chosen so that it doesn't additionally constrain the varaibles when the case is not active. In principle, this value of M could be different for each of the four constraints above. However, a sufficient condition for this to happen is that

$$M \ge \max\{|x_t^{\text{REC}}|, |u_t^{\text{batt}} + u_t^{\text{pv}} - u_t^{\text{load}}|\} \quad \forall t$$

Setting $M = \text{Y}^{\text{BOS}}$ satisfies this.

A final note: these constraints are a minimum set of requirements to constrain x_t^{REC}. Additional constraints can be added to accommodate more complicated rules; e.g., bundled REC constraints. Also note that this formulation relies on on non grid charging for the battery. If grid charging is allowed, then microgrid "net" needs to be changed.

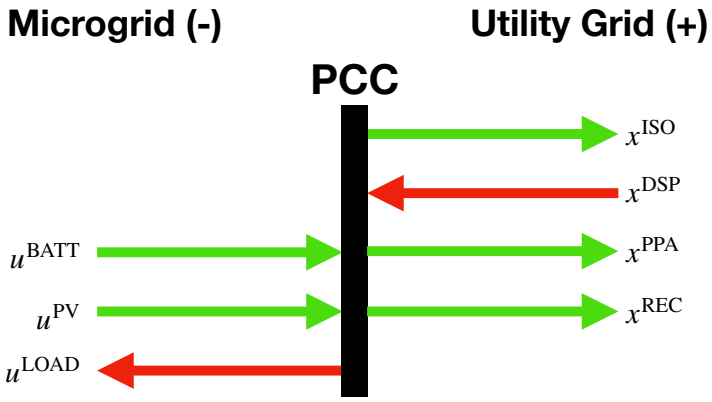


Figure 8.1: Positive sign indicates export to grid, negative sign indicates import from grid.

Building Load Building load is assumed to follow a known-in-advance timeseries $\boldsymbol{\beta} \in \mathbb{R}_+^T$. This means building load is non-controllable. These constraints can be expanded to allow for a controllable load trajectory.

$$u_t^{\text{load}} = \beta_t \qquad \forall t \tag{8.5a}$$

PV

The PV production is curtailable and can vary between 0 and the maximum available power. The maximum available power is given by the DC nameplate capacity y^{PV} scaled by a parameter timeseries $\boldsymbol{\alpha} \in [0, 1]^T$ which represents the fraction of AC PV power produced by the PV inverter in each time interval.[3] The $\boldsymbol{\alpha}$ parameter combines all losses from max theoretical DC power production (in W) to actual AC injection by the inverter (in W), including DC-AC conversion losses, DC losses (in wiring and panels), soiling, shading, clouds, etc. The DC nameplate capacity is constrained by the parameter Y^{PV}, the site's max physical capacity.[4]

$$0 \leq u_t^{\text{pv}} \leq y^{\text{PV}} \alpha_t \qquad \forall t \tag{8.6a}$$

$$0 \leq y^{\text{PV}} \leq \text{Y}^{\text{PV}} \tag{8.6b}$$

Battery

Standard battery constraints with integer variables z_t^{c}, z_t^{d} to prevent simultaneous charge/discharge. We define an extra variable for the total battery injection u_t^{batt} which is positive when discharging and negative when charging.

$$u_t^{\text{batt}} = u_t^{\text{d}} - u_t^{\text{c}} \qquad \forall t \tag{8.7a}$$

Given that we seek to simultaneously optimize capacity as well as charging operation,

[3] For this optimization, we take the powerfactor to be 1, so that all power produced by the inverter is real. At a later point, we could introduce q, for example if we want to offer grid services to the utility.

[4] I assume that the DC nameplate is a parameter of each panel that scales with its physical area (m^2).

the first two of the following constraints present bilinear integer constraints.

$$
\begin{aligned}
&0 \le u_t^{\mathrm{c}} \le z_t^{\mathrm{c}} y^{\mathrm{BATT}} && \forall t \\
&0 \le u_t^{\mathrm{d}} \le z_t^{\mathrm{d}} y^{\mathrm{BATT}} && \forall t \\
&0 \le y^{\mathrm{BATT}} \le \mathbf{Y}^{\mathrm{batt}} && \\
&z_t^{\mathrm{c}} + z_t^{\mathrm{d}} \le 1 && \forall t \\
&z_t^{\mathrm{c}}, z_t^{\mathrm{d}} \in \{0, 1\} && \forall t
\end{aligned}
$$

These can be replaced by the following lifted constraint set, with additional variables $v_t^{\mathrm{c}}, v_t^{\mathrm{d}} \in \mathbb{R}$ introduced. The formulation is given in vectorized notation.

$$
\begin{aligned}
&0 \le u_t^{\mathrm{c}} \le v_t^{\mathrm{c}} && \forall t && \text{(8.8a)} \\
&0 \le u_t^{\mathrm{d}} \le v_t^{\mathrm{d}} && \forall t && \text{(8.8b)} \\
&v_t^{\mathrm{c}} \le Y^{\mathrm{BATT}} z_t^{\mathrm{c}} && \forall t && \text{(8.8c)} \\
&v_t^{\mathrm{d}} \le Y^{\mathrm{BATT}} z_t^{\mathrm{d}} && \forall t && \text{(8.8d)} \\
&Y^{\mathrm{BATT}}(z_t^{\mathrm{c}} - 1) \le v_t^{\mathrm{c}} - y^{\mathrm{BATT}} \le 0 && \forall t && \text{(8.8e)} \\
&Y^{\mathrm{BATT}}(z_t^{\mathrm{d}} - 1) \le v_t^{\mathrm{d}} - y^{\mathrm{BATT}} \le 0 && \forall t && \text{(8.8f)} \\
&z_t^{\mathrm{c}} + z_t^{\mathrm{d}} \le 1 && \forall t && \text{(8.8g)} \\
&z_t^{\mathrm{c}}, z_t^{\mathrm{d}} \in \{0, 1\} && \forall t && \text{(8.8h)}
\end{aligned}
$$

Note that in the original formulation, $z_t^{\mathrm{c}} = 0 \Rightarrow u_t^{\mathrm{c}} = 0$. In (8.8), $z_t^{\mathrm{c}} = 0 \Rightarrow v_t^{\mathrm{c}} = 0 \Rightarrow u_t^{\mathrm{c}} = 0$ from (8.8a),(8.8c). (8.8e) implies the 3rd constraint from the original formulation. In the original formulation, $z_t^{\mathrm{c}} = 1 \Rightarrow 0 \le u_t^{\mathrm{c}} \le y^{\mathrm{BESS}}$. In (8.8), $z_t^{\mathrm{c}} = 1 \Rightarrow v_t^{\mathrm{c}} = y^{\mathrm{BATT}}$. With this equivalence, (8.8c) enforces the 3rd constraint in the original formulation and (8.8a) recovers the 1st.

The remaining battery constraints enforce the SOC update and SOC constraints. It is assumed that the units of SOC are Watt-hours (Wh) and the duration of each charge/discharge action is 1 hour. If the interval is not one hour, then a scaling term will need to be introduced into (8.9a).

$$x_t^{\text{SOC}} = x_{t-1}^{\text{SOC}} + \eta_c u_t^{\text{c}} - \frac{1}{\eta_d} u_t^{\text{d}} \qquad \forall t \tag{8.9a}$$

$$\text{SOC}_{\min} y^{\text{SOC}} \leq x_t^{\text{SOC}} \leq \text{SOC}_{\max} y^{\text{SOC}} \qquad \forall t \tag{8.9b}$$

$$x_0^{\text{SOC}} = \text{SOC}_0 \tag{8.9c}$$

$$x_T^{\text{SOC}} = \text{SOC}_f \tag{8.9d}$$

$$0 \leq y^{\text{SOC}} \leq Y^{\text{SOC}} \tag{8.9e}$$

Note that there is an additional variable x_0^{SOC} introduced for convenience. The SOC variable x_t^{SOC} is the state of charge *at the end* of interval t, after the charge/discharge action has occurred; x_{t-1}^{SOC} is the SOC at the beginning of interval t. It is in units of energy to allow us to optimize both the power y^{BATT} and energy y^{SOC} capacity of the battery. If we desire to fix the E-P ratio of the battery, then add a constraint $y^{\text{SOC}} = R y^{\text{BATT}}$.

We might also like to enforce that the battery does not charge from grid power and only uses solar to charge. This means that we need to enforce the complementarity constraint

$$u_t^{\text{c}} x_t^{\text{DSP}} = 0 \quad \forall t$$

Since we already have integer variables defined for both battery charging (8.8a) and grid power (8.11d), we simply add a constraint

$$z_t^{\text{c}} + z_t^{\text{DSP}} \leq 1 \quad \forall t \tag{8.10a}$$

This complementarity constraint is sufficient to enforce no grid charging because we assume that $x^{\text{RT}} \geq 0$.

Market dispatch

The surplus power generated by the microgrid, beyond what is needed to serve building load, can be sold into markets. We consider three scenarios

1. RT market dispatch only
2. DA-RT market dispatch
3. PPA pricing

In the 3rd scenario, all power is dispatched and paid a fixed $ per MWh price. Simultaneously to all of these scenarios, we consider it possible that the surplus power can earn REC revenues (value stacking). When selling power into any of these markets, we enforce that power *cannot* be simultaneously bought from the utility.

RT dispatch only The DA commitment is always zero in this scenario and the excess produced by the microgrid is scheduled and sold in to the RT ISO market. When the RT price is positive, it will always be profitable to sell; when the price is negative, it will be unprofitable. Therefore we allow for curtailment to 0 of the RT commitment.

DA-RT dispatch only In this scenario, we allow a financial DA commitment to be made and settled in advance at the DA price p^{DA}. A RT adjustment is allowed and the final, physical power provided to the grid in RT is x^{RT}. The settlment rule is

$$p_t^{\mathrm{DA}} x_t^{\mathrm{DA}} + p_t^{\mathrm{RT}} (x_t^{\mathrm{RT}} - x_t^{\mathrm{DA}})$$

x_t^{DA} must be physical, in that if there is no RT adjustment, the x_t^{DA} commitment is physically realizable. Therefore, we must enforce $0 \leq x_t^{\mathrm{DA}} \leq y^{\mathrm{BOS}}$. However, it not necessary to enforce any of the complementarity constraints.

PPA In this scenario, all excess power is sold at the PPA price.

We wish to enforce that the microgrid (1) can only *sell* excess power[5] at the LMP into the wholesale markets and (2) can only *buy* power at retail rates from the utility. Furthermore, we wish to enforce that the microgrid cannot do (1) and (2) simultaneously (during the same interval). This can be accomplished with complementarity constraints.

$$x_t^{\mathrm{RT}} x_t^{\mathrm{DSP}} = 0 \qquad \forall t$$
$$x_t^{\mathrm{PPA}} x_t^{\mathrm{DSP}} = 0 \qquad \forall t$$

The total amount of power that can be injected (positive or negative) through the PCC (point of common coupling) with the distribution service provider's (DSP) grid is physically capped by y^{BOS}.[6] We also enforce for accounting purposes, none of x^{RT}, x^{DSP}, x^{PPA} can individually exceed y^{BOS}.

[5] "Excess" defined as $\max\{0, u^{\mathrm{batt}} + u^{\mathrm{pv}} - u^{\mathrm{load}}\}$.

[6] Capping the total injection by y^{BOS} simultaneously controls both the grid-side injection $|x_t^{\mathrm{RT}} + x_t^{\mathrm{PPA}} - x_t^{\mathrm{DSP}}|$ and the microgrid injection $|u^{\mathrm{batt}} + u^{\mathrm{pv}} - u^{\mathrm{load}}|$ through the power balance constraint (8.1).

The original set of market dispatch constraints with the complementarity constraints is

$$
\begin{aligned}
& 0 \le y^{\mathrm{BOS}} \le \mathrm{Y}^{\mathrm{BOS}} && \\
& |x_t^{\mathrm{RT}} + x_t^{\mathrm{PPA}} - x_t^{\mathrm{DSP}}| \le y^{\mathrm{BOS}} && \forall t \\
& 0 \le x_t^{\mathrm{RT}} \le y^{\mathrm{BOS}} && \forall t \\
& 0 \le x_t^{\mathrm{DA}} \le y^{\mathrm{BOS}} && \forall t \\
& 0 \le x_t^{\mathrm{DSP}} \le y^{\mathrm{BOS}} && \forall t \\
& 0 \le x_t^{\mathrm{PPA}} \le y^{\mathrm{BOS}} && \forall t \\
& x_t^{\mathrm{RT}} x_t^{\mathrm{DSP}} = 0 && \forall t \\
& x_t^{\mathrm{PPA}} x_t^{\mathrm{DSP}} = 0 && \forall t
\end{aligned}
$$

First, $|x_t^{\mathrm{RT}} - x_t^{\mathrm{DSP}}| \le y^{\mathrm{BOS}}$ is redundant and can be replaced by $x_t^{\mathrm{RT}} + x_t^{\mathrm{PPA}} \le y^{\mathrm{BOS}}$. The typical procedure is used to turn the bilinear equality constraint into an equivalent set of mixed-integer linear (MIL) constraints by introducing auxiliary integer variables z_t^{ISO}, z_t^{DSP}, and z_t^{PPA}.

$$
\begin{aligned}
& 0 \le y^{\mathrm{BOS}} \le \mathrm{Y}^{\mathrm{BOS}} && \\
& x_t^{\mathrm{RT}} + x_t^{\mathrm{PPA}} \le y^{\mathrm{BOS}} && \forall t \\
& 0 \le x_t^{\mathrm{RT}} \le y^{\mathrm{BOS}} z_t^{\mathrm{ISO}} && \forall t \\
& 0 \le x_t^{\mathrm{DSP}} \le y^{\mathrm{BOS}} z_t^{\mathrm{DSP}} && \forall t \\
& 0 \le x_t^{\mathrm{PPA}} \le y^{\mathrm{BOS}} z_t^{\mathrm{PPA}} && \forall t \\
& z_t^{\mathrm{ISO}} + z_t^{\mathrm{DSP}} \le 1 && \forall t \\
& z_t^{\mathrm{PPA}} + z_t^{\mathrm{DSP}} \le 1 && \forall t \\
& 0 \le x_t^{\mathrm{DA}} \le y^{\mathrm{BOS}} && \forall t \\
& z_t^{\mathrm{ISO}}, z_t^{\mathrm{DSP}}, z_t^{\mathrm{PPA}} \in \{0, 1\} && \forall t
\end{aligned}
$$

This reformulation brings up a similar problem as with the battery constraints, where a mixed-integer bilinear term in the RHSs of the second and third constraints has been introduced. We handle it in a similar way by introducing auxiliary variables

$v_t^{\text{ISO}}, v_t^{\text{DSP}}, v_t^{\text{PPA}} \in \mathbb{R}$.

$$0 \leq y^{\text{BOS}} \leq \text{Y}^{\text{BOS}} \tag{8.11a}$$
$$x_t^{\text{RT}} + x_t^{\text{PPA}} \leq y^{\text{BOS}} \quad \forall t \tag{8.11b}$$
$$0 \leq x_t^{\text{RT}} \leq v_t^{\text{ISO}} \quad \forall t \tag{8.11c}$$
$$0 \leq x_t^{\text{DSP}} \leq v_t^{\text{DSP}} \quad \forall t \tag{8.11d}$$
$$0 \leq x_t^{\text{PPA}} \leq v_t^{\text{PPA}} \quad \forall t \tag{8.11e}$$
$$v_t^{\text{ISO}} \leq \text{Y}^{\text{BOS}} z_t^{\text{ISO}} \quad \forall t \tag{8.11f}$$
$$v_t^{\text{DSP}} \leq \text{Y}^{\text{BOS}} z_t^{\text{DSP}} \quad \forall t \tag{8.11g}$$
$$v_t^{\text{PPA}} \leq \text{Y}^{\text{BOS}} z_t^{\text{PPA}} \quad \forall t \tag{8.11h}$$
$$\text{Y}^{\text{BOS}}(z_t^{\text{ISO}} - 1) \leq v_t^{\text{ISO}} - y^{\text{BOS}} \leq 0 \quad \forall t \tag{8.11i}$$
$$\text{Y}^{\text{BOS}}(z_t^{\text{DSP}} - 1) \leq v_t^{\text{DSP}} - y^{\text{BOS}} \leq 0 \quad \forall t \tag{8.11j}$$
$$\text{Y}^{\text{BOS}}(z_t^{\text{PPA}} - 1) \leq v_t^{\text{PPA}} - y^{\text{BOS}} \leq 0 \quad \forall t \tag{8.11k}$$
$$z_t^{\text{ISO}} + z_t^{\text{DSP}} \leq 1 \quad \forall t \tag{8.11l}$$
$$z_t^{\text{PPA}} + z_t^{\text{DSP}} \leq 1 \quad \forall t \tag{8.11m}$$
$$0 \leq x_t^{\text{DA}} \leq y^{\text{BOS}} \quad \forall t \tag{8.11n}$$
$$z_t^{\text{ISO}}, z_t^{\text{DSP}}, z_t^{\text{PPA}} \in \{0, 1\} \quad \forall t \tag{8.11o}$$

Objective function

The objective function of the optimization problem will account for the *present value* of all fixed and variable costs associated with the microgrid investment.[7] We seek to optimize the total net present value P of the project over its entire term.

The items to be considered as part of this profit/loss are:

- Capacity cost for solar
- Capacity cost for battery (for both power and energy)
- Capacity cost for balance-of-system (BOS) equipment including utility transformer, switchgear (ATS), telemetry. This can be thought of as the cost o the grid connection.
- O&M for solar, battery, and BOS

[7] A minor caveat to this is that we won't consider the fixed, non-capacity dependent costs (e.g., property expense) in the objective function, since they won't influence the optimal solution. Such costs will be included in the full cash flow calculation however.

- Degradation cost for battery (get from SAM) We won't consider cell degradation variable cost for now. Rather, just assume a replacement time period and have that be capacity-scaled expense baked into F^{BATT}.
- Retail energy cost (can include variable and demand charge)
- Revenue from DART markets
- Revenue from RECs. RECs are currently levied in units of $ per MWh on renewable energy. Eligible power for RECs is anything produced by the solar array, which is potentially stored in the battery.
- Revenue from a power purchase agreement (PPA)
- Taxes, incentives, depreciation schedules, cost of financing, and discount rates.

Define $f(\mathbf{y})$ to be the revenue-cost associated with capacity-scaling investments and $g(\mathbf{x})$ be the revenue-cost associated with variable, ongoing production. The basic difference is the same as that between fixed and variable costs. Although items like O&M cost, debt interest payments are ongoing, they are all scaled by the initial capital investment and therefore truly are fixed costs. Items like income tax, production tax credit (PTC), and market revenues are part of the variable costs because they depend on how much power the microgrid produces.

NPV P is then defined

$$P = f(\mathbf{y}) + g(\mathbf{x}) \tag{8.12}$$

Components of fixed costs

- Installation cost (direct capital cost, indirect capital cost, sales tax). The components of direct cost include module and inverter costs, balance of system equipment (conduit, cables), and installer labor and overhead. SAM also adds a contigency fee to this (e.g., 4%) but not sure if this is needed. Components of indirect cost include permitting and environmental studies, engineering overhead, and grid interconnection fees. Unlike SAM, we do not include land cost.
- O&M cost (fixed annual, annual by capacity, battery cell replacement cost)
- Debt cost

- Investment tax credit (ITC)
- Utility connection charge

Components of variable costs

- REC revenues
- ISO revenues
- PPA revenues
- Utility volumetric charge
- Utility demand charge
- Production tax credit (PTC)
- Income taxes assessed on all income less interest and depreciation. For this reason, we need to calculate the depreciation schedule, which then becomes a fixed parameter in the optimization.

Next, we define the cost functions explicitly.

Capacity costs

There are two categories of costs that scale with the installed capacities y^{PV}, y^{BATT}, and y^{BOS}, namely capital (or investment) cost and fixed, annualized costs (e.g., O&M). We do not consider fixed annual costs that do not vary with capacity because they do not affect the optimization solution (however they do affect the cashflow). We assume that both kinds of costs are unit costs, in $ per unit of capacity.

The upfront unit capital cost (in year 0) is scaled by the cost of capital. We show in Section 8.1 that the present value of financing is a multiplicative scaling term A. The cost of servicing debt **is also** future discounted because the principal and interest payments will be paid for with future case. However, when calculating the interest and principal payments scheuldes, future discounting should not be considered, because they are computed in nominal year-0 dollars.

Define

$$\mathbf{c} = \begin{bmatrix} C^{\text{pv}} \\ C^{\text{batt}} \\ C^{\text{bos}} \end{bmatrix}, \quad f_i = \begin{bmatrix} F_i^{\text{pv}} \\ F_i^{\text{batt}} \\ F_i^{\text{bos}} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y^{\text{PV}} \\ y^{\text{BATT}} \\ y^{\text{BOS}} \end{bmatrix}$$

The total year-0 capital cost is given by

$$f_1(\mathbf{y}) := \mathbf{c}^\top \mathbf{y}(1 + r^{\text{debt}} A - \frac{r^{\text{ITC}}}{1 + r^{\text{NPV}}}) \tag{8.13}$$

The investment tax credit r^{ITC} is applied to the upfront (year 0) investment cost and since we apply the tax credit as a positive cash flow (rather than an offset of a tax bill otherwise due) in year 1, we discount its NPV by one year at rate r^{NPV}. $r^{\text{debt}} \in [0, 1]$ is the fraction of the upfront capital cost serviced by debt.

The O&M portion of the fixed cost is discounted using the NPV formulation with a give rate r^{NPV}.[8] The total present value (in year 0) of the ongoing fixed costs is

$$f_2(\mathbf{y}) := \sum_{i=1}^{N} \frac{1}{(1 + r^{\text{NPV}})^i} (F_i^{\text{PV}} y^{\text{PV}} + F_i^{\text{BATT}} y^{\text{BATT}} + F_i^{\text{BOS}} y^{\text{BOS}}) \tag{8.14}$$

Then the total capacity cost with sales and income taxes applied is

$$f(\mathbf{y}) := -f_1(\mathbf{y})(1 + r^{\text{tax}}_{\text{sales}}) - f_2(\mathbf{y})(1 + r^{\text{tax}}_{\text{income}}) \tag{8.15}$$

Variable costs and revenues

Variable costs and revenues scale with the amount of power produced by the microgrid. Taxed future net earnings (revenues minus costs) are discounted by r^{NPV} and have taxes applied.

The variable costs consist of a volumetric charge from utility for energy plus a demand charge. For simplicity, we will assume that the demand charge is applied over the entire year, although in reality the billing period is probably shorter.

$$g_1(\mathbf{x}) = \sum_{t=1}^{T} p_t^{\text{DSP}} x_t^{\text{DSP}} + p^{\text{DC}} \max_t \{x_t^{\text{DSP}}\}$$

The revenues come from three sources: ISO markets, PPA, and RECs. The unit of variable cost/revenue is $ per MWh.[9] The production tax credit $p_t^{\text{PTC}} \in \mathbb{R}_+$ is a $-per-MWh credit for all energy produced and sold into ISO and PPA markets. Since the battery is charged by solar, any energy export from the microgrid is eligible for the PTC. We apply the REC price to the net of solar production, building load,

[8] We assume that inflation is baked into this discount rate (i.e., nominal discount rate).

[9] If the interval is anything but an hour, there will need to be a scaling term introduced into g to convert between MW and MWh.

and battery charging. This is valid because we enforce that the battery can only be charged by solar. Other more complicated constraints can be implemented in (8.1b).

$$g_2(\mathbf{x}) = \sum_{t=1}^{T} p_t^{\mathrm{RT}} x_t^{\mathrm{RT}} + (p_t^{\mathrm{DA}} - p_t^{\mathrm{RT}}) x_t^{\mathrm{DA}} + p_t^{\mathrm{PPA}} x_t^{\mathrm{PPA}} + \frac{p_t^{\mathrm{PTC}}}{1 - r_{\mathrm{income}}^{\mathrm{tax}}} (x_t^{\mathrm{RT}} + x_t^{\mathrm{PPA}}) + p_t^{\mathrm{REC}} x_t^{\mathrm{REC}}$$

We scale the PTC by $1 - r_{\mathrm{income}}^{\mathrm{tax}}$ because it will be multiplied by the same in the next expression and the PTC should be counted as a straight positive, untaxed cashflow.

The convention is that revenues are positive and costs are negative. $r_{\mathrm{income}}^{\mathrm{tax}}$ is the combined federal and state income tax rate.

$$g(\mathbf{x}) := \sum_{i=1}^{N} \frac{1 - r_{\mathrm{income}}^{\mathrm{tax}}}{(1 + r^{\mathrm{NPV}})^i} \left(-g_1(\mathbf{x}) + g_2(\mathbf{x})\right) \tag{8.16}$$

The tax rate is applied to both positive and negative values of $g(\mathbf{x})$ because a negative profit incurs a positive tax benefit, which is treated as a positive cashflow for that year.

Debt and interest

The objective of this section is to show that the cost of financing debt can be included as a parameter of the optimization problem. This may not be obvious due to the non-linear compounding of interest and potentially complicated debt repayment schedules.

For this section, let $C \in \mathbb{R}_+$ be the unit cost of investment quantity $y \in \mathbb{R}_+$, and let $r^{\mathrm{debt}} \in [0, 1]$ be the fraction of investment cost serviced by debt. Without financing ($r^{\mathrm{debt}} = 1$), the present value (in year 0) of total investment cost is

$$\mathrm{TC} = Cy$$

In the financed case, the amount $Cy(1 - r^{\mathrm{debt}})$ is paid upfront in year 0, and a loan for the rest Cyr^{debt} is taken out and repaid over N years (in actual dollars) plus interest, giving a total cost of

$$\mathrm{TC} = Cy(1 - r^{\mathrm{debt}}) + L + I$$

L and I are the *present values* of the principal of the loan of Cyr^{debt} and interest paid on it respectively.[10]

[10] Interest needs to be computed on Cyr^{debt}, not L!

The question is whether we can write $L + I = yA$ for some A, which may only depend on some constants like discount rate, loan interest rate, etc. This will allows the total cost to be expressed as

$$\text{TC} = y \underbrace{(C(1 - r^{\text{debt}}) + A)}_{:=\tilde{C}}$$

In other words, the form of the investment cost term stays the same, but $\tilde{C}$ replaces C. This means that the objective function of the optimization problem remains linear in the decision variable y.

First, some notation is presented. Let B_0 be the quantity of the loan (i.e., Cyr^{debt}) and B_i be the remaining principal balance at the beginning of year i.[11] Let r^{int} be the interest rate, and N be the loan term. Let b_i be the principal payment and s_i be the interest payment due in period i. By definition, the sum of principal payments is B_0.

$$B_0 = \sum_{i=1}^{N} b_i$$

and the sum of interest payments is

$$I = \sum_{i=1}^{N} s_i$$

The interest payment in period i is

$$s_i = r^{\text{int}} B_i$$

and remaining balance at the beginning of period i is

$$B_i = \begin{cases} B_{i-1} - b_i, & i = 2, \ldots, N \\ B_0, & i = 1 \end{cases}$$

We will now show that L and I can be computed in closed form for two commonly used repayment schedules: even total payments and even principal payments.

[11] Nothing is paid down in year 0, so $B_1 = B_0$.

Even total payments In this amortization scheme, the total of principal and interest payments in each period (i.e., one year) is a constant. The relation that holds between two successive time periods is

$$\begin{aligned} & b_i + s_i = b_{i-1} + s_{i-1} \\ \Leftrightarrow \quad & b_i + r^{\text{int}} B_i = b_{i-1} + r^{\text{int}} B_{i-1} \\ \Leftrightarrow \quad & b_i + r^{\text{int}}(B_{i-1} - b_{i-1}) = b_{i-1} + {}_{i-1} \\ \Leftrightarrow \quad & b_i = b_{i-1}(1 + r^{\text{int}}) \end{aligned}$$

Each principal payment can therefore be written in terms of the last b_N which allows us to solve back for each b_i:

$$\begin{aligned} & b_i = \frac{b_N}{(1 + r^{\text{int}})^{N-i}} \\ \Rightarrow \quad & B_0 = b_N \sum_{i=1}^{N} \frac{1}{(1 + r^{\text{int}})^{N-i}} \\ \Leftrightarrow \quad & B_0 = \frac{b_N}{(1 + r^{\text{int}})^{N}} \sum_{i=1}^{N} (1 + r^{\text{int}})^{i} \\ \Leftrightarrow \quad & B_0 = \frac{b_N}{(1 + r^{\text{int}})^{N}} \left(\frac{(1 + r^{\text{int}})^{N+1} - (1 + r^{\text{int}})}{r^{\text{int}}} \right) \\ \Rightarrow \quad & b_N = B_0 \frac{r^{\text{int}}(1 + r^{\text{int}})^{N}}{(1 + r^{\text{int}})^{N+1} - (1 + r^{\text{int}})} \\ \Rightarrow \quad & b_i = B_0 \frac{r^{\text{int}}(1 + r^{\text{int}})^{i}}{(1 + r^{\text{int}})^{N+1} - (1 + r^{\text{int}})} \end{aligned}$$

To get the present value of the principal L, each b_i needs to be discounted by the appropriate multiple of r^{NPV}:

$$L = \sum_{i=1}^{N} \frac{b_i}{(1 + r^{\text{NPV}})^{i}} = B_0 \underbrace{\frac{r^{\text{int}}}{(1 + r^{\text{int}})^{N+1} - (1 + r^{\text{int}})} \sum_{i=1}^{N} \frac{(1 + r^{\text{int}})^{i}}{(1 + r^{\text{NPV}})^{i}}}_{:=\text{Nasty}_1}$$

Next, calculate the annual interest payment s_i due.

$$
\begin{aligned}
s_i &= r^{\text{int}} B_i \\
&= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_{i-1} - b_i, & i = 2 \ldots, N \end{cases} \\
&= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 - \sum_{k=1}^{i-1} b_k, & i = 2 \ldots, N \end{cases} \\
&= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 \left(1 - \frac{r^{\text{int}}}{(1+r^{\text{int}})^{N+1} - (1+r^{\text{int}})} \sum_{k=1}^{i-1} (1 + r^{\text{int}})^k \right), & i = 2 \ldots, N \end{cases} \\
&= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 \left(1 - \frac{r^{\text{int}}}{(1+r^{\text{int}})^{N+1} - (1+r^{\text{int}})} \frac{(1+r^{\text{int}})^i - (1+r^{\text{int}})}{r^{\text{int}}} \right), & i = 2 \ldots, N \end{cases} \\
&= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 \left(1 - \frac{(1+r^{\text{int}})^i - (1+r^{\text{int}})}{(1+r^{\text{int}})^{N+1} - (1+r^{\text{int}})} \right), & i = 2 \ldots, N \end{cases}
\end{aligned}
$$

The yearly s_i is discounted by the appropriate factor of r^{NPV} and summed to get I.

$$
\begin{aligned}
I &= \sum_{i=1}^{N} \frac{s_i}{(1 + r^{\text{NPV}})^i} \\
&= B_0 \underbrace{r^{\text{int}} \left(\frac{1}{1 + r^{\text{NPV}}} + \sum_{i=2}^{N} \frac{1}{(1 + r^{\text{NPV}})^i} \left(1 - \frac{(1 + r^{\text{int}})^i - (1 + r^{\text{int}})}{(1 + r^{\text{int}})^{N+1} - (1 + r^{\text{int}})} \right) \right)}_{:=\text{Nasty}_2}
\end{aligned}
$$

Therefore, we can write

$$L + I = B_0 \times \left(\text{Nasty}_1 + \text{Nasty}_2\right)$$

To get the expression for $\tilde{C}$ when $B_0 = Cyr^{\text{debt}}$, we have

$$\tilde{C} = C + Cr^{\text{debt}}(\text{Nasty}_1 + \text{Nasty}_2)$$

Even principal payments In this amortization scheme, the principal payments in each period (i.e., one year) are constant. Therefore

$$b_i = \frac{B_0}{N}$$

L is then

$$L = \frac{B_0}{N} \sum_{i=1}^{N} \frac{1}{(1 + r^{\text{NPV}})^i} = B_0 \underbrace{\frac{1 - \frac{1}{N(1+r^{\text{NPV}})^N}}{r^{\text{NPV}}}}_{\text{Nasty}_3}$$

The annual interest payment is

$$\begin{aligned} s_i &= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 - \sum_{k=1}^{i-1} b_k, & i = 2 \ldots, N \end{cases} \\ &= r^{\text{int}} \begin{cases} B_0, & i = 1 \\ B_0 \left(1 - \frac{i-1}{N}\right), & i = 2 \ldots, N \end{cases} \end{aligned}$$

Discounting each s_i by appropriate factor gives the present value of the interest

$$\begin{aligned} I &= \sum_{i=1}^{N} \frac{s_i}{(1 + r^{\text{NPV}})^i} \\ &= B_0 \underbrace{\left(\frac{1}{1 + r^{\text{NPV}}} + \sum_{i=2}^{N} \frac{1 - \frac{i-1}{N}}{(1 + r^{\text{NPV}})^i} \right)}_{:=\text{Nasty}_4} \end{aligned}$$

Therefore, here we can also write

$$L + I = B_0 \times (\text{Nasty}_3 + \text{Nasty}_4)$$

To get the expression for $\tilde{C}$ when $B_0 = Cyr^{\text{debt}}$, we have

$$\tilde{C} = C + Cr^{\text{debt}}(\text{Nasty}_3 + \text{Nasty}_4)$$

Taxes and incentives

There are taxes on investments (fixed costs), which include sales and property taxes, and taxes on cashflows (variable costs/revenues), which include state and federal income tax. The ITC and PTC are accounted for as positive cashflows (in year 0 for ITC) and ongoing for PTC.

The tax basis for investment taxes includes (taken from SAM Installation Costs module)

- Sales tax on total direct costs (excludes permitting, land, engineering overhead, and grid interconnection costs) in year 0
- Sales tax on total installed cost

The tax basis for income taxes includes

- Any revenue from markets (DART, REC, PPA)
- Operating costs (fixed and variable)
- Depreciation expense
- Interest payments on debt

Taxable income does not include principal payments on debt or investment/production tax credits. For now, we assume that the tax basis for federal and state taxes is the same. We do not include the adjustment to the taxable cashflow that arises from depreciation because it does not affect the optimal solution. The depreciation shield in each year is solely a consequence of the depreciation schedule and the total amount of depreciation will not change depending on how it is depreciated.

Optimization problem

Combining constraints and objective function from the previous sections, our capacity-operational optimization problem is

$$\max_{\mathbf{y},\mathbf{x}_t} \quad P \tag{8.17a}$$

$$\text{s.t.} \quad (8.1) - (8.11) \tag{8.17b}$$

Note that some integer and auxiliary variables are not explicitly included here as optimization variables. So, really what is meant is that there exist such variables along with $\mathbf{y}, \mathbf{x}_t$ that satisfy (8.1) - (8.11). This optimization problem is a mixed-integer linear program.

Cash flow

We have optimized most of the ingredients of the cash flow in the objective function of the optimization. However there are some details about how taxes and depreciation are handled that are currently outside the scope of the optimization.

One key point about how we compute cash flow is that we do not separately consider value of inflation. Rather, it is baked into the nominal discount rate r^{NPV}. To determine the real discount rate r^{discount}, we can use the following relation

$$1 + r^{\text{NPV}} = (1 + r^{\text{discount}})(1 + r^{\text{inflation}})$$

We should also discount the future value of the combined principal and interest payments because these will be cash flows.

The approach should be to calculate the entire cashflow of the project in nominal year-0 dollars and then discount that timeseries.

We will use the following steps to generate cashflow timeseries for the project term.

1. Run optimization problem to generate "optimal" **y**, **x**, and **u**. The **x** solution gives a prospective operational schedule over the project term.

2. Compute investment cost, which is the total installed cost times the sales tax (assume 5% rate).

3. Use operational fixed cost and variable cost/revenue values to generate annual cashflows from operations. This will necessitate combining aspects of **y** and **x**. The result of this step should be the EBITDA.

4. Compute the depreciation basis and schedule. From SAM, "The depreciable basis is the sum of the total installed cost and total construction financing cost, less the sum of investment-based incentives and 50% of any investment tax credits." Sum depreciation expense with EBITDA to get profit before taxes.

5. Compute state and federal taxes. The tax basis consists of EBITDA, depreciation expense, interest payments (but not principal). We assume that state taxes are computed first and whatever the state tax expense is needs to be considered a positive cash flow for the purpose of computing the federal tax. This is taken to be the only difference between federal and state tax bases.

6. Apply tax credits. The ITC is assumed to be a positive cashflow applied in year 1. The PTC will be a applied in real dollar amounts to the MWh of production in each year.

7. Get net cash flow for each year and discount appropriately.

Simulation design

This section contains a few comments on the implementation of this optimization problem. For the most part, it follows the formulation in these notes.

One key parameter to pick is T. For computational reasons, we cannot optimize every hour of every year for 25 years ahead. What seems to be possible is to pick $T = 8760$ which is one year. What's not clear is how we model load growth, degradation in system components (PV degradation, inverter efficiency loss), weather changes etc.

The demand charge is implemented by introducing an auxiliary variable x^{DC} along with the constraints

$$x^{\text{DC}} \geq x_t^{\text{DSP}} \quad \forall t$$

The demand charge is then $p^{\text{DC}} x^{\text{DC}}$.

Sample results

We apply our methodology to case study in the Texas ERCOT market. We simulate the market operation of a microgrid over a 25-year time horizon at a node near Houston. The microgrid is located in an industrial park composed of several warehouses with capacity for significant solar. The buildings draw a small amount of load. It is assumed that there is 2MW interconnection limit and space at the site for a maximum of 5MW of rooftop solar. The objective of the optimization problem is to determine the rooftop PV and energy storage capacities that should be installed to minimize the payback period of the infrastructure. In addition, we add a nonnegative profit constraint to ensure that the project is net profitable.

Cost data is taken from the National Renewable Energy Lab (NREL) Annual Technology Baseline for commercial solar and storage [98] and market data are acquired from the ERCOT energy price database.

Our results show a payback period of 9 years when the microgrid's onsite generation and storage are able to offset a majority of building load and gain additional revenue by trading the excess power into the day-ahead and real-time energy markets.

The results from this optimization were used as a basis for an $80M grant application to the US Department of Energy's Grid Resilience and Innovation Partnerships (GRIP) program for the installation of a similar microgrid in ERCOT. Our partners in the project application were a large commerical real estate developer which would serve as the site host and provide financing and an energy trading firm which would handle the interaction with the ERCOT spot market.

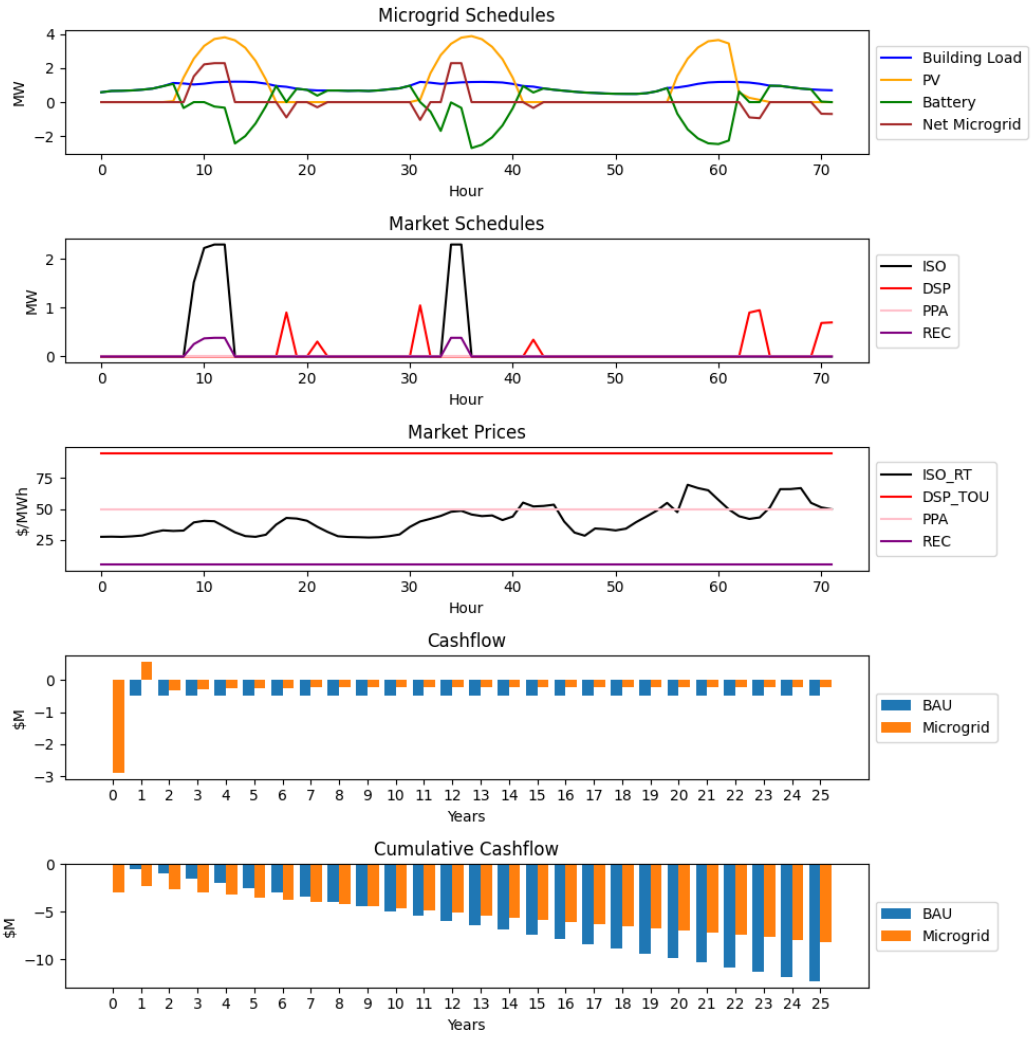


Figure 8.2: Microgrid operational and financial optimal solutions. The first panel shows the generation/consumption trajectories for building load and microgrid DERs. The second panel shows the market schedules for the net injection of the microgrid at the interconnection point. The third panel displays the market prices for DART markets and utility rates. The fourth panel shows the annual discounted cash flows for the project. The fifth panel dispalys the cumulative discounted cash flows for the project.

8.2 Resiliency planning for BESS

In this section, we explore a planning application of microgrid optimization. We consider a system operator that seeks to install a battery energy storage system (BESS) at a building with existing rooftop solar, so that solar generation, battery injections, building load, and grid power are aggregated at the building's main switchboard. These components for a building-level microgrid which can provide

benefits in 1) resiliency to outages and load shedding, 2) carbon emissions reduction, and 3) energy cost reduction. Our objective is to optimally size the battery with respect to cost to provide a given level of resiliency. In addition, we will quantify the ancillary benefits the battery would provide in terms of energy cost reduction and carbon emissions reduction.

For an expository case study, we focus on Baxter Hall on the Caltech campus, which has 115kW of rooftop PV already installed an average building load of ~100kW. We assume that the the building load and PV production can be measured but not controlled. We also assume that the output from the BESS inverter with unity powre factor can be controlled continuously throughout its charge-discharge range. A schematic of the micorgrid is shown in Figure 8.3. It is expected that the power capacity capacity of the battery should be roughly proportionate to the building's load and that the energy capacity of the battery should depend on the length of the outage.

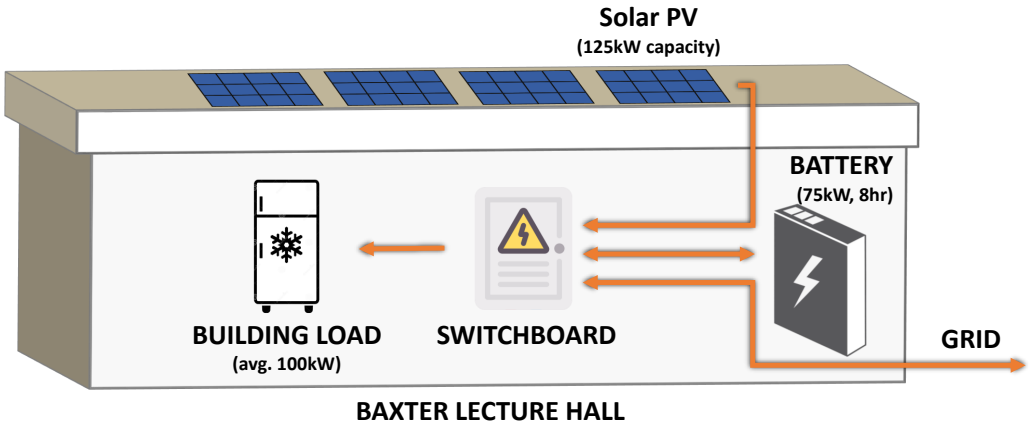


Figure 8.3: Schematic for a PV-BESS-Load building microgrid at Baxter Hall on the Caltech campus

Note that although the methodology used in this case study is fully generalizable, the precise outcomes presented are illustrative only and actual benefits will vary depending on the building size, load profile, and available PV.

Defining resilience

To quantify resilience and the impact a BESS has on it, we adopt the methodology in [130] to triage load into three tiers: Tier 1 loads are mission critical and must be served 100% of the time. Tier 2 loads are critical but can be curtailed if necessary to preserve Tier 1 loads. However, over the duration of the outage, 80% of the Tier

2 load should be met. Tier 3 loads are discretionary and can be curtailed but should be satisfied on average 25% of the time.

Each organization can specify what proportion of its load belongs each tier. Define R_1, R_2, and R_3 as the proportion of Tier 1,2 & 3 load, respectively. It must hold that

$$R_{1,t} + R_{2,t} + R_{3,t} = 1 \quad \forall t = 1, \ldots, T$$

We consider a nominal load profile $\mathbf{d} \in \mathbb{R}^T$ where T is the number of periods in a prospective outage. Given resiliency levels R_i, the following defines the load profiles for each tier of load:

$$\mathbf{d}_i := R_i \mathbf{d}, \quad i \in \{1, 2, 3\}$$

If desired, time-varying resiliency levels $R_{i,t}$ can be defined as well.

A 100% resiliency level means that the available power supplying the building (sum of grid, BESS, and PV), denoted by $\mathbf{u}_t \in \mathbb{R}^T$, satisfies

$$u_t \geq \sum_i d_{i,t} \; \forall t = 1, \ldots, T.$$

A 0% resiliency level means that $u_t = 0 \; \forall t$. In order to meet the criteria in [130], it must at least be true that

$$u_t \geq d_{1,t} \; \forall t = 1, \ldots, T.$$

By scaling R_2 from 0 to $1 - R_1$, overall resilience levels from R_1 to 100% can be calibrated. Note that according to [130], the resilience of Tier 1 load guaranteed *in each interval* but the resilience of Tier 2 load is only guaranteed *on average*.

For an example of how the resilience of a system can be defined, let $R_1 = 0.2$ and $R_2 = 0.3$. Then $R_3 = 0.5$. In this example, a "20% resilient" system would serve all of Tier 1 load for the duration of the outage but none of Tier 2 & 3 load. A "40% resilient" system would meet all of the Tier 1 load and at least 2/3 of the *cumulative* Tier 2 load during the outage. However, there may be periods where less than 2/3 of Tier 2 load is being served and others where more than 2/3 is being served. An "80% resilient" system meets all of Tier 1 load, all of Tier 2 load, and 3/5 of Tier 3 load on average. However, there may still be periods where not all of Tier 2 load is satisfied, but if there were to be the case, then there would conversely be periods where more than all of Tier 2 load would need to be satisfied.

Quantifying the resiliency benefit

A principal benefit of placing a BESS in a building is that it can be controlled to satisfy building load in the event of an outage in the Caltech or utility distribution systems.[12] The resiliency a BESS would afford a building depends on its energy capacity and the size of critical building load. The combination of rooftop PV and battery—when controlled optimally—can provide significant reliability benefits for the building.

For the experiments here and all that follow in this section, we consider the same Baxter Hall scenario introduced above, with a building with 90.3kW of load on average and a peak load of 257.5kW during the week of August 22-29, 2022. The minimum load during this week was 41.1kW.[13] The 115kW nameplate PV array on the roof, producing 5600kWh over the entire weak and ~110kW at peak every day. PV generation was derived from the solar radiation profiles for Pasadena during that same week [120].

In these simulations, we assume that a 75kW/600kWh (8hr) battery has been installed so that it injects power into the main building switchboard (see Figure 8.3) and supporting load when there is no grid power. We then vary the outage duration from 1 to 24 hours and calculate how much load shed must occur if the battery is being discharged. These results are presented in Table 8.2. During short outages of less than 1 hour, the battery can support nearly all of the nominal building load. For all outages up to 24 hours in length that begin both at 12am and 12pm, the battery is able to provide power for both Tier 1 and Tier 2 loads, with a large portion of the Tier 3 load being covered for short outages.[14]

Optimally sizing a battery system for resilience

Next, we pose the question: to guarantee a given level of resilience (e.g., 60%), how big must a battery system be? The answer will depend on the load profile, available building-connected PV, and the definitions of Tier 1,2 &3 loads.

We address this question by solving an optimization problem (8.18). The objective is to minimize the cost of buying a battery while still meeting reliability criteria. These criteria are encoded as constraints on the optimal solution. Because load

[12] A battery would serve a similar purpose as the fuel cells on the Caltech campus or the emergency diesel generators in many high-priority buildings.

[13] For the month of August 2022, average load was 87.4kW and peak load was 292.9kW, so this week is representative of summertime load conditions.

[14] Assumes the 10%,15%,75% breakdown of load into Tiers 1,2,3 presented in [130].

Outage Scenarios	Loadshed without battery	Loadshed with battery
<1 hour outage @ 12am	100%	0%
2 hour outage @ 12am	100%	12%
4 hour outage @ 12am	100%	56%
12 hour outage @ 12am	70%	58%
24 hour outage @ 12am	72%	67%
<1 hour outage @ 12pm	55%	0%
2 hour outage @ 12pm	56%	24%
4 hour outage @ 12pm	64%	49%
12 hour outage @ 12pm	75%	65%
24 hour outage @ 12pm	71%	65%

Table 8.2: Comparison of building load that is shed in various outage scenarios. "100%" means all of building load is shed during outage; "0%" means that no load is shed.

management and battery scheduling during the outage is necessary to realize the resiliency benefit of the battery, we include a set of constraints on the operation of the battery and the battery-solar-building microgrid.

$$\max_{U^{\max}, E^{\max}} \quad c_{\mathrm{u}} U^{\max} + c_{\mathrm{e}} E^{\max} + \lambda ||\mathbf{u} - \mathbf{d}_1 - \mathbf{d}_2||_2 \tag{8.18a}$$

$$\text{s.t.} \quad u_t = s_t + u_t^{\mathrm{d}} - u_t^{\mathrm{c}} \qquad \forall t \tag{8.18b}$$

$$0 \le u_t^{\mathrm{c}} \le z_t^{\mathrm{c}} U^{\max} \qquad \forall t \tag{8.18c}$$

$$0 \le u_t^{\mathrm{c}} \le z_t^{\mathrm{d}} U^{\max} \qquad \forall t \tag{8.18d}$$

$$z_t^{\mathrm{c}} + z_t^{\mathrm{d}} \le 1 \qquad \forall t \tag{8.18e}$$

$$z_t^{\mathrm{c}}, z_t^{\mathrm{d}} \in \{0, 1\} \qquad \forall t \tag{8.18f}$$

$$x_{t+1} = x_t + \eta_c u_t^{\mathrm{c}} - \frac{1}{\eta_d} u_t^{\mathrm{d}} \qquad \forall t \tag{8.18g}$$

$$\mathrm{SOC}^{\min} E^{\max} \le x_t \le \mathrm{SOC}^{\max} E^{\max} \qquad \forall t \tag{8.18h}$$

$$x_0 = \mathrm{SOC}_0 \tag{8.18i}$$

$$u_t \ge d_{1,t} \qquad \forall t \tag{8.18j}$$

$$\sum_t u_t \ge 0.8 \sum_t d_{2,t} + 0.25 \sum_t d_{3,t} \tag{8.18k}$$

$$1 \le \frac{E^{\max}}{U^{\max}} \le 10 \tag{8.18l}$$

The objective function (8.18a) is the sum of energy and power capacity costs and a regularization term, weighted by scalar $\lambda \in \mathbb{R}$, on the net load curtailment. This

term promotes the smoothness and uniqueness of the curtailed Tier 2 load profile. (Without it, the charge/discharge of the battery can be very noisy.) Constraint (8.18b) ensures power balance during an outage (all building load must be met by a combination of PV power, denoted s_t, and BESS power, denoted by the difference $u_t^{\mathrm{d}} - u_t^{\mathrm{c}}$). Constraints (8.18c-i) enforce operational constraints on the battery, with variables z_t^{c} and z_t^{d} introduced to enforce complementarity of charge/discharge actions u_t^{c} and u_t^{d}. Constraints (8.18j) and (8.18k) encode the resiliency requirements from the methodology in [130]. Constraint (8.18l) ensures that the energy-to-power ratio stays within reasaonble ranges (between a 1 hour and 10 hour battery). Optimization problem (8.18)is a mixed-integer second-order cone program, efficiently solvable in practice with a MISOC solver such as Gurobi.

Table 8.3 summarizes the remaining parameters in the case study that appear in (8.18). We adopt the default partition of nominal load into tiers given in [130].

Parameter	**Value**	**Description**
c_{u}	\$965.83/kW	Cost of battery power capacity[15]
c_{e}	\$237.64/kWh	Cost of battery energy capacity[16]
η_c	0.95	Charging efficiency of modern li-ion battery system
η_d	0.95	Discharging efficiency of modern li-ion battery system
SOC_0	1.0	Battery state-of-charge when outage begins (assume full)
R_1	10%	Tier 1 load, cannot be curtailed
R_2	15%	Tier 2 load, can be curtailed up to 20%
R_3	75%	Tier 3 load, discretionary load
λ	0.01	Smoothing parameter to enforce uniqueness

Table 8.3: Parameters for optimization problem (8.18)

The optimal solutions of optimization problem (8.18) are displayed in teh pliots in Figure 8.4.

The first panel shows the load and solar profiles for the week of Aug 22-29, 2022 without any battery injections. The load is disaggregated into tiers, with the Tier 1 load being a constant 10% of the average load (rather than varying with time). Thus only the priority and discretionary loads in Tiers 2 and 3 fluctuate with time.

The second panel of Figure 8.4 shows the combined solar and battery net generation profile in relation to the same load profiles. Optimal charge/discharge of the battery over a two day outage results in no curtailment of Tier 1 loads, only 20% curtailment of Tier 2 loads, and 75% curtailment of Tier 3 loads.

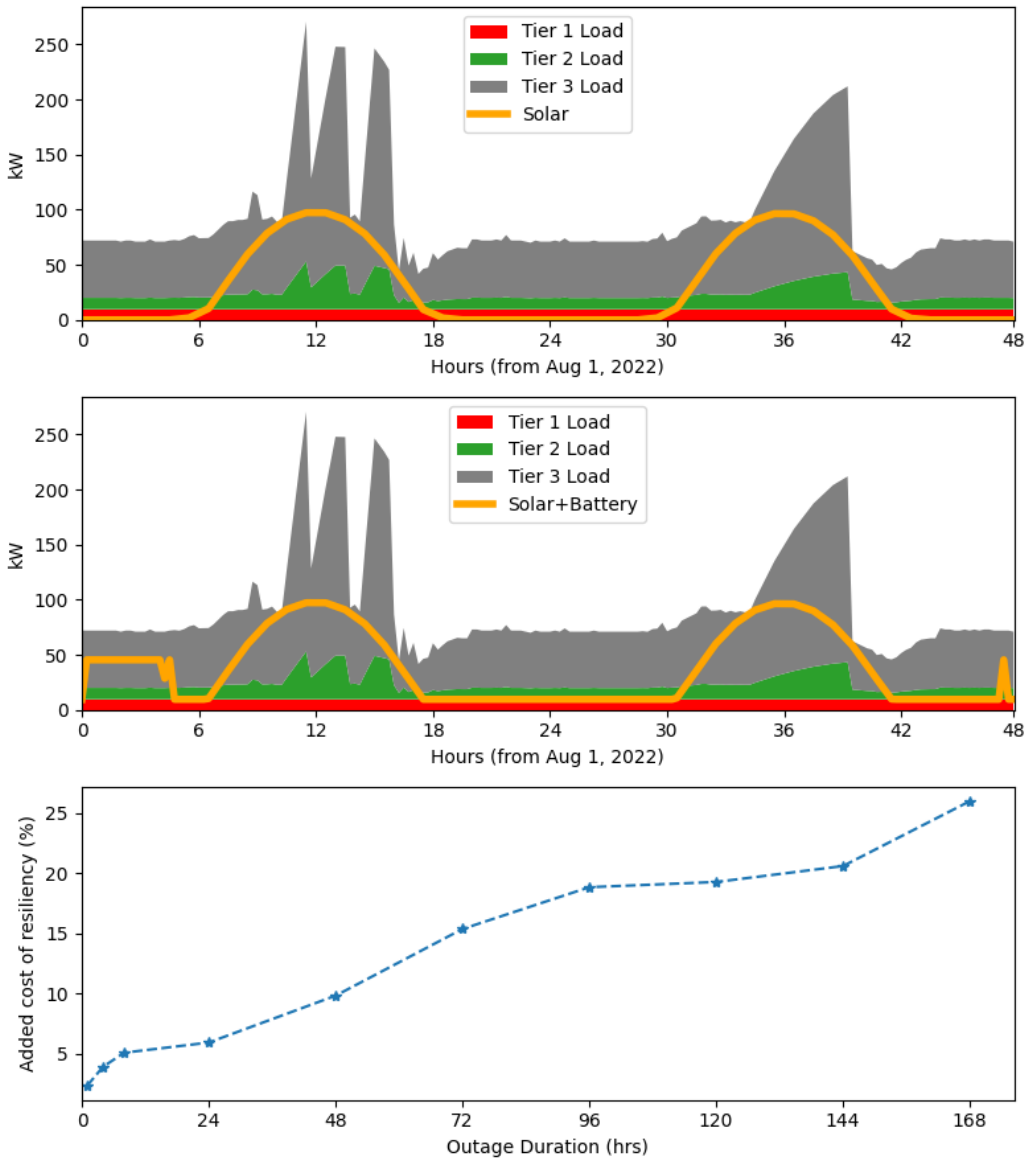


Figure 8.4: Simulation of battery and load microgrid operation at Baxter Hall over 48 hours in August 2022. The first panel shows nominal operation with significant curtailment of critical load. The second panel shows optimal discharge of the battery to meet critical and priority loads when PV generation is not available. The last panel shows the added cost of resilience that a battery incurs over the building's energy cost baseline (and over the ~15-year lifespan of the battery).

The third panel of Figure 8.4 shows the results of an ablation study on the cost premium of reliability, as provided by a BESS system. To Compute the annual baseline energy cost of the building, we use the Pasadena Water & Power's large commercial time-of-use volumetric rate for energy ($/kWh) and apply a demand charge to the single 15-min monthly peak ($/kW). We divide the demand charge

by 2 to account for the fact that the August peak is likely to be higher than during the rest of the year, thus yielding a more realistic estimate of annual energy cost. We then scale the battery size (and cost) so that it can meet progressively longer outage durations (shown on the horizontal axis). This battery cost is then divided by the asset lifespan cost of supplying energy building, yielding the additional "cost of reliability" shown on the vertical axis. Remarkably, resilience to outages of up to 48 hours can be had for as little as +10% of the cost of supplying the building's energy.

Our calculations show that total *energy* provided by solar could meet 65% of aggregate building load over the week. However, the solar profile does not match the building's profile, which is where a battery could provide significant resilience value for critical building loads.

Cost and emissions reduction

Given that grid outages are rare (despite their potentially significant impact), the BESS can be controlled to provide decarbonization and cost benefits to the campus while simultaneously maintaining its primary resiliency function. For decarbonization, load shifting is necessary to minimize consumption during periods of high-carbon intensity. For cost efficiency, peak shaving of building load in response to time-of-use schedules and demand charge peaks can reduce energy cost significantly.

In situations where there is a tradeoff between resiliency, decarbonization, and cost reduction, multiple objectives can be co-optimized to yield the greatest *net* benefit. The relative weighting of each objective can be set to tailor specific application scenarios. In this scenario, the complementary timing of the optimal battery charge/discharge schedules suggest that resiliency, emissions, and cost reduction benefits can all be value stacked without sacrificing any one particularly.

Figure 8.5 shows a 14% reduction in the weekly energy cost for the building, which is achieved by discharging during peak demand. These savings are realized with brief and infrequent discharges (see panel 2). Figure 8.6 shows a modest 5% reduction is weekly CO2 emissions for the building, which is achieved by cycling the battery daily to shift load from high-carbon intensity periods to low-carbon intensity periods on the grid. Given that peak load shaving opportunities come during the day and decarbonization load shifting opportunities occur at night, these two objectives are compatible and their benefits are additive.

Operating the Baxter Hall 75kW/600kWh BESS over a 10-year span, its net CO2

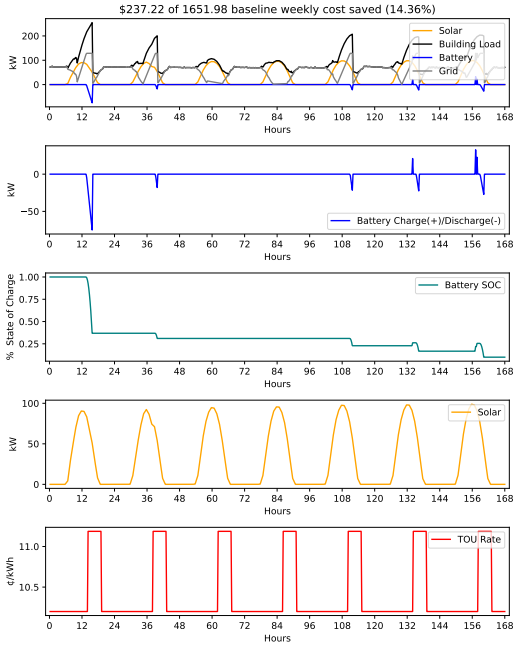


Figure 8.5: Battery charge/discharge schedules and state-of-charge trajectories optimized for time-of-use energy cost minimization.

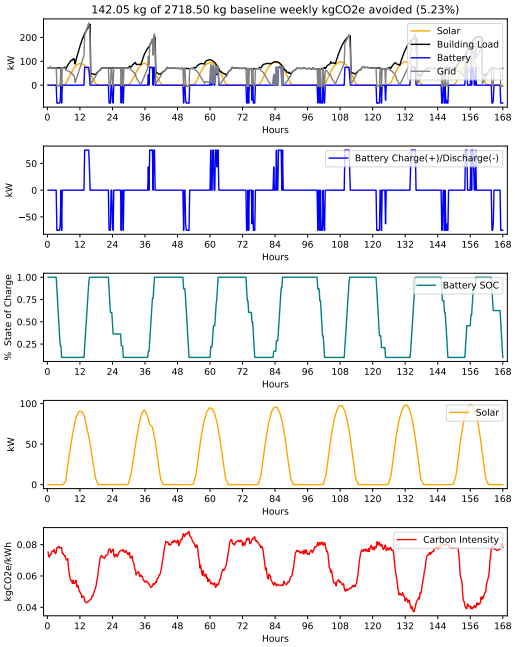


Figure 8.6: Battery charge/discharge schedules and state-of-charge trajectories optimized for carbon emissions reduction.

reduction would be

$$(20 \text{ kgCO2e/day}) \cdot (365 \text{ days/year}) \cdot 10 \text{ years} \approx 73,000 \text{ kgCO2e} \tag{8.19}$$

which corresponds to 0.61 kgCO2e avoided per dollar of investment. This figure will almost certainly increase in the future, as higher solar PV penetration leads to a more aggressive duck curve and greater difference in carbon intensity between day and night.

The net energy cost reduction over the same 10 year span would be

$$(3 \text{ \$/day}) \cdot (365 \text{ days/year}) \cdot 10 \text{ years} \approx \$120,000 \tag{8.20}$$

which roughly breaks even with the initial investment in the bESS. Remarkably, the return on investment can be driven higher—up to a factor of 5x over a 10-year period—if we choose a smaller battery (e.g., 20kWh capacity), due to the peak shaving events being infrequent and short. However, this comes at a tradeoff of lesser resiliency benefits.

BESS testbed at Caltech

Based on the findings in the experiments above, we are collaborating with Caltech Facilities to deploy a 80kW/4hr BESS system at a Caltech building in Fall 2023. Although the resiliency, decarbonization, and cost reduction benefits that the BESS will provide will be modest at this small scale, the pilot project offers a low-risk investment to validate and quantify these benefits in a real-world test bed. If this pilot is successful, it will provide a template for scaling up to additional Caltech buildings and to a scale sufficient (likely >2MW) to positively impact reliability of the entire Caltech campus. In addition, the hardware testbed will give us the opportunity to deploy and refine hardware-software interfaces discussed in Section 7.1. While such interfaces are straightforward in concept, their implementation is invovled and will provide important lessons for scalable deployment in the future.

Conclusions

Based on the simulations and scenarios explored in this section, our main finding is that a small BESS system deployed at a building with critical load and rooftop solar offers significant resiliency benefits to critical loads. In addition, we observe that:

- Even a small-sized battery (~10kWh) can be extremely efficient at peak shaving, reducing demand charge. The 10-year net cost saving can be as much as 5x the initial investment;
- The cost-optimal battery size (kW) and duration (kWh) depend on the building location, the load profile, and the main objective (resiliency, decarbonization, cost reduction);
- Despite the fact that we presented the Baxter Hall, with rooftop PV, as the case study, buildings without rooftop PV likely provide even greater cost and emission reductions due to the comparatively greater time-shifting impact of the BESS.

8.3 Voltage control in a distribution feeder

Increasing penetration of distributed renewable generation and accelerating electrification of vehicles and heating/cooling loads have the potential to erode power quality in distribution feeders. For example, voltage magnitudes at customer endpoints far away from the substation can often drop below minimum acceptable values (±5% per-unit (p.u.) during summer peak load hours. Correcting these issues traditionally

requires upgrades to physical infrastructure like conductors and transformers so that larger loads at the network edge can be accommodated at nominal voltage levels. Pasadena Water & Power (PWP) is exploring a new approach that places a fleet of small batteries (5kW/28kWh) strategically throughout the grid and control them during peak loading conditions to prevent voltage violations.

Project objectives. In this project, we collaborate with PWP to study the following questions:

- At which nodes should the batteries be placed to optimally impact customer endpoint voltage?
- How should the charging/discharging processes of the batteries be scheduled to support voltage during peak load conditions?

The answers to these questions depend on load pattern as well as control objectives and constraints. For example, objectives may include minimizing voltage deviations from their nominal values or capital investment. Constraints may include voltage limits, line limits, budget, availability of state information, and limited communication and control capabilities.

Method

In collaboration with PWP, we have developed a mathematical model of the PWP feeder and a systematic method to answer this type of questions. The core of this method is the formulation of a general-purpose multi-phase optimal power flow (OPF) problem that takes into account of energy use in the feeder as well as various control objectives and constraints. We have used this method to determine optimal battery placement assuming power injections from batteries are fully controllable. Extensions of this work address control strategies for the batteries that account for the limited observability and communication capabilities of the inverters.

Implementing distributed voltage control in real distribution networks requires tackling several modeling challenges. First is simulating unbalanced multi-phase power flow. We adopt the linearized 3-phase power flow model for which distribution line losses are assumed to be relatively small ($< 10\%$) and voltages phasors are approximately balanced. The second is incorporating power losses due to transformers which, despite being ubiquitous in real-world distribution grids, are often ignored in powerflow studies. The feeder under consideration in this project has three tiers

coupled by a combination of single- and multi-phase transformers. We model the electrical properties of multi-phase transformers (in both YY and ΔY configurations) using admittance matrices derived from equivalent circuit representations. Finally, limited observability of the network below the substation requires simulation of realistic loading patterns.

The goal is to place a limited number of batteries at strategic locations in the distribution network and then charge/discharge them to correct voltage deviations. We formulate the battery location problem as a mixed-integer linear program with a binary variable associated with the placement decision for each node in the network. We impose a constraint on the total battery capacity in the system and model limits on battery charge/discharge rates and states of charge. Network constraints include linearized multi-phase powerflow equations and voltage bounds at all nodes. Power demand from customers is modeled by disaggregating time series of the net injection at the substation according to defined end-user load profiles. The objective of the optimization is to minimize the sum of the deviations of voltage magnitudes from nominal values at customer endpoints. We solve this multi-interval network optimization problem to optimality to obtain the battery placement locations and corresponding charge/discharge schedules. The current charge/discharge schedules assume full communication and control capability of the batteries.

Technical details

We first introduce our model and some basic definitions.

Preliminaries and notation

A distribution network is composed of buses and lines connecting these buses. It is usually multi-phase and radial. A distribution network is a directed graph $\mathsf{G} := (\mathcal{N}, \mathcal{E})$ Let $\mathcal{N} := \{0, 1, \ldots, n\}$ denote a set of buses in a distribution network where 0 represents the substation bus and let $\mathcal{N}_+ := \{1, \ldots, n\}$. The substation bus in the network often has a fixed voltage V_0. The set of lines $\mathcal{E}$ contains ordered pairs (i, j) if there is a line $i \to j$ between buses $i \in \mathcal{N}$ and $j \in \mathcal{N}$. We write $i \sim j$ if either $i \to j$ or $j \to i$. We consider a three-phase network and let a, b, c denote the three phases and let $\Phi_i \subseteq \{a, b, c\}$ denote the phases of bus $i \in \mathcal{N}$ and $\Phi_{ij} \subseteq \{a, b, c\}$ denote the phases of line $i \sim j$. For each bus $i \in \mathcal{N}$, we let V_i^ϕ be its phase-ϕ complex voltage for any $\phi \in \Phi_i$ and define $V_i := [V_i^\phi]_{\phi \in \Phi_i}$. Moreover, for each bus $i \in \mathcal{N}$, let I_i^ϕ denote the current of phase $\phi \in \Phi_i$ for bus i and define

$I_i \coloneqq [I_i^\phi]_{\phi\in\Phi_i}$ and let s_i^ϕ denote the complex power injection of phase $\phi \in \Phi_i$ of bus i and define $s_i \coloneqq [s_i^\phi]_{\phi\in\Phi_i}$. For each line $i \sim j$, we let I_{ij}^ϕ denote the current of phase $\phi \in \Phi_{ij}$ from bus i to j and define $I_{ij} \coloneqq [I_{ij}^\phi]_{\phi\in\Phi_i}$. Finally, we denote by z_{ij} the phase impedance matrix and we assume it is full-rank and define $y_{ij} \coloneqq {z_{ij}}^{-1}$. Our goal is to control voltages in the distribution network G during fixed time horizons $\mathcal{T} \coloneqq \{1, \ldots, T\}$. In the remainder of this article, we use $V_i(t), I_i(t)$ and $s_i(t)$ to represent the voltage, current and power injection for bus i at time $t \in \mathcal{T}$.

Table 8.4 below summarizes the notation.

Table 8.4: Nomenclature and Notation.

Indices and Sets			
i, j	Indices of buses	$\mathcal{T}$	Set of time horizons
$\mathcal{N}$	Set of buses	$\mathcal{E}$	Set of lines
Φ_i	Set of phases of bus i	Φ_{ij}	Set of phases of line $i \sim j$
Parameters			
V_0^{ref}	Fixed substation voltage		
V_i^ϕ	Voltage of phase-ϕ for bus i	V_i	Voltages of all phases for bus i
I_i^ϕ	Current of phase-ϕ for bus i	I_i	Current of all phases for bus i
I_{ij}^ϕ	Current of phase-ϕ for line $i \sim j$	I_{ij}	Current of all phases for line $i \sim j$
s_i^ϕ	Power injection of phase-ϕ for bus i	s_i	Power injection of all phases for bus i
z_{ij}	Phase impedance matrix	γ, β	Battery charging/discharging rate limits
		$\underline{V}_i^\phi, \overline{V}_i^\phi$	Voltage bounds for bus i
Decision Variables			
P	Battery locations		

Battery placement optimization

Our goal is to solve the following battery placement optimization problem based on the bus injection model (BIM) [86] subject to given battery dynamics and voltage

regulation constraints:

$$\min_{P} \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{N}_+} \mathcal{L}\left(V_i(t)\right) \tag{8.21a}$$

$$\text{subject to:} \tag{8.21b}$$

$$\text{(BIM)} \quad s_i(t) = \sum_{j:i\sim j} \operatorname{diag}\left[V_i(t)^{\Phi_{ij}} \left(V_i(t)^{\Phi_{ij}} - V_j(t)^{\Phi_{ij}}\right)^H y_{ij}^H\right],\ i \in \mathcal{N}, t \in \mathcal{T} \tag{8.21c}$$

$$\text{(Battery Dynamics)} \quad s_i(t) = s_i(t-1) + P_i e_i(t-1),\ i \in \mathcal{N}_+, t \in \mathcal{T} \tag{8.21d}$$

$$e_i(t) \in [\gamma, \beta],\ i \in \mathcal{N}_+, t \in \mathcal{T} \tag{8.21e}$$

$$\sum_{i\in\mathcal{N}_+} P_i = \ell,\ P_i \in \{0,1\},\ i \in \mathcal{N}_+ \tag{8.21f}$$

$$\text{(Voltage Regulation)} \qquad V_0 = V_0^{\text{ref}}, \overline{V}_i^{\phi} \le \left|V_i^{\phi}(t)\right| \le \overline{V}_i^{\phi},\ i \in \mathcal{N}_+, \phi \in \Phi_i \tag{8.21g}$$

where $\ell > 0$ denote the number of batteries to be allocated and $P = (P_1, \ldots, P_n)$ and $\mathcal{L} : \mathbb{C}^3 \to \mathbb{R}_+$ denotes some loss function for voltage regulation.

Linearized branch Flow model

Optimization (8.21) is not computationally efficient to solve. Therefore, we follow the LPF proposed in [53] and consider the following linearized optimization.

$$\min_{P} \sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{N}_+} \mathcal{L}\left(V_i(t)\right) \tag{8.22a}$$

$$\text{subject to:} \tag{8.22b}$$

$$\text{(LPF with Transformers)} \qquad s_0 = -\sum_{k\in\mathcal{N}_+} s_k^{\Phi_0} \tag{8.22c}$$

$$\Lambda_{ij} = -\sum_{k\in\text{Down}(j)} s_k^{\Phi_{ij}} \tag{8.22d}$$

$$S_{ij} = \gamma^{\Phi_{ij}} \operatorname{diag}(\Lambda_{ij}) \tag{8.22e}$$

$$v_i = v_0^{\Phi_i} - \sum_{(k,l)\in\mathcal{P}_i} \left[S_{kl} z_{kl}^H + z_{kl} S_{kl}^H\right]\ i \in \mathcal{N}_+ \tag{8.22f}$$

$$\text{(Battery Dynamics)} \quad s_i(t) = s_i(t-1) + P_i e_i(t-1),\ i \in \mathcal{N}_+, t \in \mathcal{T} \tag{8.22g}$$

$$e_i(t) \in [\gamma, \beta],\ i \in \mathcal{N}_+, t \in \mathcal{T} \tag{8.22h}$$

$$\sum_{i\in\mathcal{N}_+} P_i = \ell,\ P_i \in \{0,1\},\ i \in \mathcal{N}_+ \tag{8.22i}$$

$$\text{(Voltage Regulation)} \qquad V_0 = V_0^{\text{ref}}, \overline{V}_i^{\phi} \le \left|V_i^{\phi}(t)\right| \le \overline{V}_i^{\phi},\ i \in \mathcal{N}_+, \phi \in \Phi_i \tag{8.22j}$$

where $\mathcal{P}_j$ denotes the path from bus 0 to bus j and Down(j) denotes the downstream of j for $j \in \mathcal{N}_+$ and $v_i = V_i V_i^H$ for all $i \in \mathcal{N}$.

Optimal battery placement results

The modeling and simulation were done using a feeder from the PWP network with approximately 800 buses and 600 customers, which draw ~2MW in peak aggregate load. Currently, deviations from nominal voltages as large as 13% p.u. are observed for some customers at some times (see Figure 8.7).

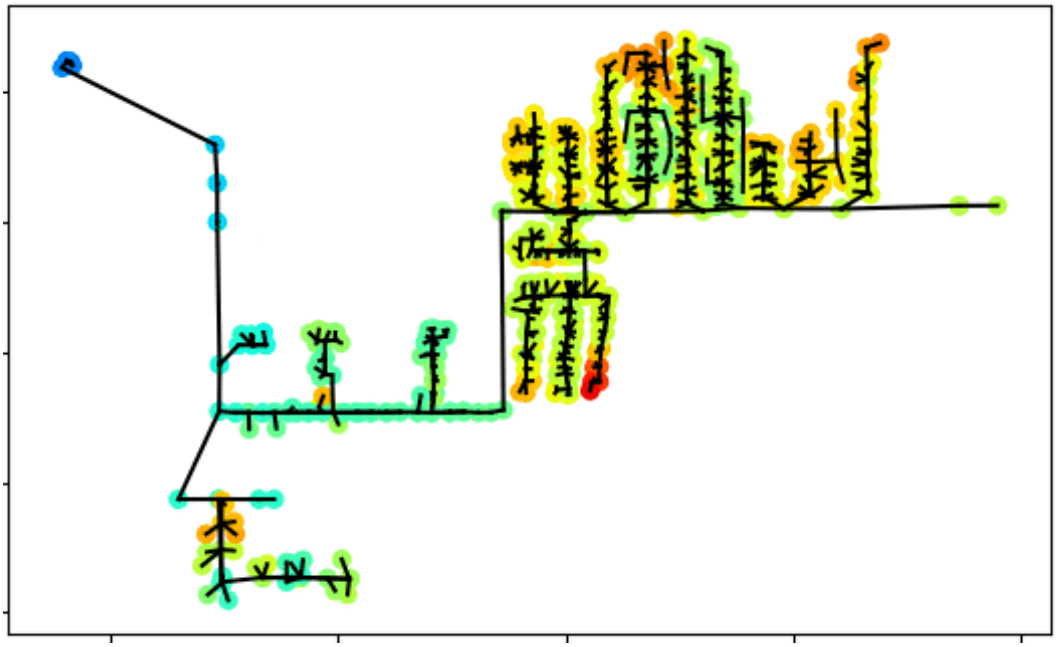


Figure 8.7: Voltage magnitudes in the PWP feeder. Blue indicates overvoltage and red indicates undervoltage. Green is approximately 1 p.u. A significant number of nodes in the feeder experience undervoltage while the substation voltages are above the nominal range.

The optimal battery placement results from optimization problem (8.21) are shown in Figure 8.8. Candidate locations were constrained to be in the 240V regions of the network due the 240V nominal voltage of the battery inverters. In addition, some power poles were not suitable for mounting batteries due to physical constraints. These placement conditions were incorporated into (8.21) as an extra set of constraints on the integer placement variables.

Our powerflow simulations show that a collection of 15 batteries, each with a capacity of 27.5kWh and discharge rate of 5kW, discharged over 4 midday peak-load hours can correct nodal voltages to be within the required ±5% p.u. tolerance (see Figure 8.9).

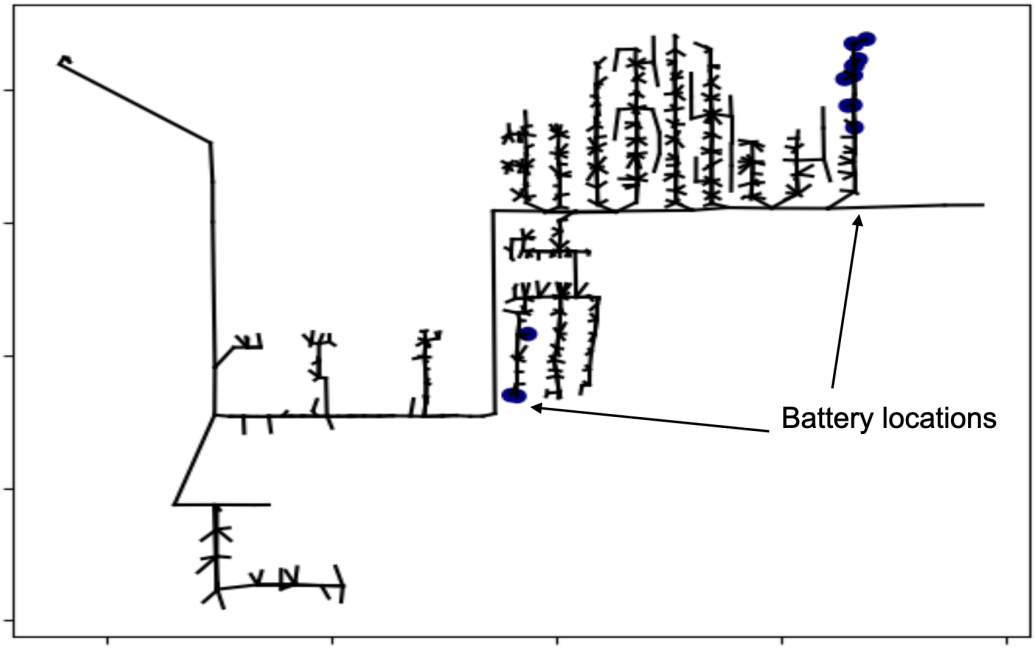


Figure 8.8: Dark blue circles show the optimal battery placement locations. The selected nodes are typically at the end of long or over-loaded lateral lines in the feeder.

Ongoing work focuses on designing controllers for the batteries, which face communication constraints and allow for limited customization of their on-board power electronics. Such approaches include optimizing open-loop time-of-use settings based on historical and projected network load and designing centralized feedback controllers that are robust to communication lags and noisy or infrequent network state measurements.

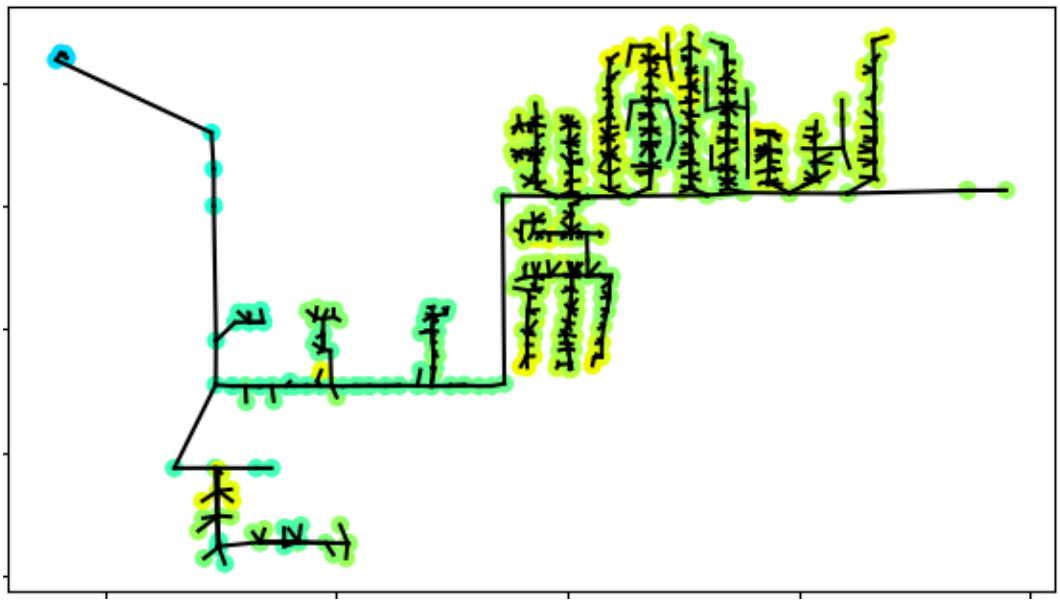


Figure 8.9: Voltage magnitudes in the PWP feeder. Blue indicates overvoltage and red indicates undervoltage. Green is approximately 1 p.u. All nodal voltages are corrected to within ±5% p.u. when batteries are discharged optimally at peak feeder load.

BIBLIOGRAPHY

[1] *2019 Assessment of Demand Response and Advanced Metering*. Dec. 2019. URL: https://www.ferc.gov/sites/default/files/2020-04/DR-AM-Report2019_2.pdf.

[2] *2021 Summer Loads and Resource Assessment*. Tech. rep. CAISO, May 2021. URL: http://www.caiso.com/Documents/2021-Summer-Loads-and-Resources-Assessment.pdf.

[3] HA Aalami, M Parsa Moghaddam, and GR Yousefi. "Demand response modeling considering interruptible/curtailable loads and capacity market programs". In: *Applied Energy* 87.1 (2010), pp. 243–250.

[4] Akshay Agrawal et al. "A rewriting system for convex optimization problems". In: *Journal of Control and Decision* 5.1 (2018), pp. 42–60.

[5] Hunt Allcott. "Real time pricing and electricity markets". In: *Harvard University* 7 (2009).

[6] Kenneth J Arrow and Gerard Debreu. "Existence of an equilibrium for a competitive economy". In: *Econometrica: Journal of the Econometric Society* (1954), pp. 265–290.

[7] Salman Azhar. "Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry". In: *Leadership and management in engineering* 11.3 (2011), pp. 241–252.

[8] Navid Azizan et al. "Optimal pricing in markets with nonconvex costs". In: *Operations Research* 68.2 (2020), pp. 480–496.

[9] Masoud Badiei, Na Li, and Adam Wierman. "Online convex optimization with ramp constraints". en. In: *2015 54th IEEE Conference on Decision and Control (CDC)*. Osaka: IEEE, Dec. 2015, pp. 6730–6736. ISBN: 978-1-4799-7886-1. DOI: 10.1109/CDC.2015.7403279. URL: http://ieeexplore.ieee.org/document/7403279/ (visited on 04/29/2021).

[10] Kyri Baker, Gabriela Hug, and Xin Li. "Energy Storage Sizing Taking Into Account Forecast Uncertainties and Receding Horizon Operation". In: *IEEE Transactions on Sustainable Energy* 8.1 (Jan. 2017), pp. 331–340. ISSN: 1949-3037. DOI: 10.1109/TSTE.2016.2599074.

[11] A. Ben-Tal and A. Nemirovski. "Robust solutions of uncertain linear programs". en. In: *Operations Research Letters* 25.1 (Aug. 1999), pp. 1–13. ISSN: 0167-6377. DOI: 10.1016/S0167-6377(99)00016-4. URL: https://www.sciencedirect.com/science/article/pii/S0167637799000164 (visited on 10/03/2021).

[12] Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

[13] Dimitris Bertsimas et al. “Adaptive Robust Optimization for the Security Constrained Unit Commitment Problem”. In: *IEEE Transactions on Power Systems* 28.1 (Feb. 2013). Conference Name: IEEE Transactions on Power Systems, pp. 52–63. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2012.2205021.

[14] Suzhi Bi and Ying Jun Zhang. “False-data injection attack to control real-time price in electricity market”. In: *2013 IEEE Global Communications Conference (GLOBECOM)*. 2013, pp. 772–777.

[15] Roy Billinton and Mahmud Fotuhi-Firuzabad. “A reliability framework for generating unit commitment”. en. In: *Electric Power Systems Research* 56.1 (Oct. 2000), pp. 81–88. ISSN: 0378-7796. DOI: 10.1016/S0378-7796(00)00104-8. URL: https://www.sciencedirect.com/science/article/pii/S0378779600001048 (visited on 09/30/2021).

[16] BloombergNEF. *Cost of New Renewables Temporarily Rises as Inflation Starts to Bite*. https://about.bnef.com/blog/cost-of-new-renewables-temporarily-rises-as-inflation-starts-to-bite/. Accessed: May 20, 2023.

[17] BloombergNEF. *Lithium-ion Battery Pack Prices Rise for First Time to an Average of $151/kWh*. https://about.bnef.com/blog/lithium-ion-battery-pack-prices-rise-for-first-time-to-an-average-of-151-kwh/. Accessed: May 20, 2023. 2022.

[18] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[19] James Bushnell. “Electricity Resource Adequacy: Matching Policies and Goals”. en. In: *The Electricity Journal* 18.8 (Oct. 2005), pp. 11–21. ISSN: 1040-6190. DOI: 10.1016/j.tej.2005.08.005. URL: https://www.sciencedirect.com/science/article/pii/S104061900500103X (visited on 09/30/2021).

[20] CAISO. *California ISO Open Access Same-time Information System (OASIS)*. 2020. URL: http://oasis.caiso.com/mrioasis/logon.do.

[21] *California ISO - Today’s Outlook*. Accessed on 2021-09-10. URL: http://www.caiso.com/TodaysOutlook/Pages/default.aspx (visited on 09/10/2021).

[22] Craig Cano. “Ferc order no. 2222: A new day for distributed energy resources”. In: *Federal Energy Regulatory Commission* (2020).

[23] Yanxin Chai et al. “Incentive-based demand response model for maximizing benefits of electricity retailers”. In: *Journal of Modern Power Systems and Clean Energy* 7.6 (2019), pp. 1644–1650.

[24] Hung-po Chao. "Demand response in wholesale electricity markets: the choice of customer baseline". In: *Journal of Regulatory Economics* 39.1 (2011), pp. 68–88.

[25] Hung-po Chao et al. "Flow-based transmission rights and congestion management". In: *The Electricity Journal* 13.8 (2000), pp. 38–58.

[26] Vincent Chau, Shengzhong Feng, and Nguyen Kim Thang. "Competitive Algorithms for Demand Response Management in Smart Grid". en. In: *LATIN 2018: Theoretical Informatics*. Ed. by Michael A. Bender, Martín Farach-Colton, and Miguel A. Mosteiro. Vol. 10807. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 303–316. ISBN: 978-3-319-77403-9 978-3-319-77404-6. DOI: 10.1007/978-3-319-77404-6_23. URL: http://link.springer.com/10.1007/978-3-319-77404-6_23 (visited on 10/04/2021).

[27] Cong Chen, Ye Guo, and Lang Tong. "Pricing Multi-Interval Dispatch under Uncertainty Part II: Generalization and Performance". In: *arXiv:1912.13469 [cs, eess]* (Dec. 2020). arXiv: 1912.13469. URL: http://arxiv.org/abs/1912.13469 (visited on 12/30/2020).

[28] Cong Chen, Lang Tong, and Ye Guo. "Pricing Energy Storage in Real-time Market". In: *2021 IEEE Power & Energy Society General Meeting (PESGM)*. ISSN: 1944-9933. July 2021, pp. 1–5. DOI: 10.1109/PESGM46819.2021.9638013.

[29] Niangjun Chen et al. "Using Predictions in Online Optimization: Looking Forward with an Eye on the Past". en. In: *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. Antibes Juan-les-Pins France: ACM, June 2016, pp. 193–206. ISBN: 978-1-4503-4266-7. DOI: 10.1145/2896377.2901464. URL: https://dl.acm.org/doi/10.1145/2896377.2901464 (visited on 04/25/2021).

[30] Shiyao Chen, Yuting Ji, and Lang Tong. "Large scale charging of Electric Vehicles". In: *2012 IEEE Power and Energy Society General Meeting*. ISSN: 1944-9925. July 2012, pp. 1–9. DOI: 10.1109/PESGM.2012.6345497.

[31] Yongbao Chen et al. "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings". In: *Applied Energy* 195 (2017), pp. 659–670.

[32] Nicolas Christianson et al. "Dispatch-aware planning for feasible power system operation". en. In: *Electric Power Systems Research* 212 (Nov. 2022), p. 108597. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2022.108597. URL: https://www.sciencedirect.com/science/article/pii/S0378779622006733 (visited on 02/11/2023).

[33] Antonio J Conejo and Ramteen Sioshansi. “Rethinking restructured electricity market design: Lessons learned and future needs”. In: *International Journal of Electrical Power & Energy Systems* 98 (2018), pp. 520–530.

[34] *Cost and Performance Characteristics of New Generating Technologies, Annual Energy Outlook 2021*. Tech. rep. Energy Information Administration (EIA), Feb. 2021. URL: https://www.eia.gov/outlooks/aeo/assumptions/pdf/table_8.2.pdf.

[35] *Cost and Performance Characteristics of New Generating Technologies, Annual Energy Outlook 2021*. Tech. rep. Energy Information Administration (EIA), Feb. 2021.

[36] Peter Cramton, Axel Ockenfels, and Steven Stoft. “Capacity Market Fundamentals”. en. In: *Economics of Energy & Environmental Policy* 2.2 (Apr. 2013). ISSN: 21605882. DOI: 10.5547/2160-5890.2.2.2. URL: http://www.iaee.org/en/publications/eeeparticle.aspx?id=46 (visited on 12/06/2021).

[37] S Datchanamoorthy et al. “Optimal time-of-use pricing for residential load control”. In: *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2011, pp. 375–380.

[38] Shahab Dehghan, Nima Amjady, and Antonio J. Conejo. “Reliability-Constrained Robust Power System Expansion Planning”. In: *IEEE Transactions on Power Systems* 31.3 (May 2016), pp. 2383–2392. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2015.2464274.

[39] Ruilong Deng et al. “A survey on demand response in smart grids: Mathematical models and approaches”. In: *IEEE Transactions on Industrial Informatics* 11.3 (2015), pp. 570–582.

[40] Steven Diamond and Stephen Boyd. “CVXPY: A Python-embedded modeling language for convex optimization”. In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[41] Krishnamurthy Dvijotham, Misha Chertkov, and Scott Backhaus. “Storage Sizing and Placement through Operational and Uncertainty-Aware Simulations”. In: *2014 47th Hawaii International Conference on System Sciences*. ISSN: 1530-1605. Jan. 2014, pp. 2408–2416. DOI: 10.1109/HICSS.2014.302.

[42] Yury Dvorkin. “A Chance-Constrained Stochastic Electricity Market”. In: *IEEE Transactions on Power Systems* 35.4 (July 2020). Conference Name: IEEE Transactions on Power Systems, pp. 2993–3003. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2019.2961231.

[43] Eirk Ela, Michael Milligan, and Brendan Kirby. *Operating Reserves and Variable Generation*. TP-5500-51978. NREL, Aug. 2011. URL: https://www.nrel.gov/docs/fy11osti/51978.pdf.

[44] Erik Ela and Mark O'Malley. "Scheduling and Pricing for Expected Ramp Capability in Real-Time Power Markets". In: *IEEE Transactions on Power Systems* 31.3 (May 2016). Conference Name: IEEE Transactions on Power Systems, pp. 1681–1691. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2015.2461535.

[45] *End-Use Load Profiles for the U.S. Building Stock*. URL: https://www.energy.gov/eere/buildings/downloads/end-use-load-profiles-us-building-stock.

[46] Torgeir Ericson. "Direct load control of residential water heaters". In: *Energy Policy* 37.9 (2009), pp. 3502–3512.

[47] ESRI. *ESRI EV Charging Stations*. https://links.esri.com/EVChargingStations. Accessed: May 23, 2023.

[48] Andreas Feldmüller. *Flexibility of coal and gas fired power plants*. Paris, Sept. 2017. URL: https://www.cleanenergyministerial.org/sites/default/files/2018-01/Andreas%20Feldmueller%20Siemens.pdf.

[49] Financial Times. *Swanson's Law provides green ray of sunshine for PV*. https://www.ft.com/content/d9f9f1b4-a3f0-11e5-873f-68411a84f346. Accessed: June 8, 2023.

[50] Francesc Font-Clos. *Hit and Run*. https://github.com/fontclos/hitandrun. 2021.

[51] Fraunhofer Institute for Solar Energy Systems (ISE). *Photovoltaics Report*. https://www.ise.fraunhofer.de/content/dam/ise/de/documents/publications/studies/Photovoltaics-Report.pdf. 2023.

[52] Natalie Mims Frick et al. "End-Use Load Profiles for the US Building Stock: Market Needs, Use Cases, and Data Gaps". In: (2019).

[53] Lingwen Gan and Steven H Low. "Convex relaxations and linear approximation for optimal power flow in multiphase radial networks". In: *2014 Power Systems Computation Conference*. IEEE. 2014, pp. 1–9.

[54] Mohammad Ghaljehei and Mojdeh Khorsand. "Day-ahead Operational Scheduling with Enhanced Flexible Ramping Product: Design and Analysis". In: *IEEE Transactions on Power Systems* (2021). Conference Name: IEEE Transactions on Power Systems, pp. 1–1. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2021.3110712.

[55] Paul R Gribik, William W Hogan, Susan L Pope, et al. "Market-clearing electricity prices and energy uplift". In: *Cambridge, MA* (2007).

[56] Theodore Groves and John O Ledyard. *Incentive compatibility ten years later*. Tech. rep. Discussion paper, 1985.

[57] Ye Guo, Cong Chen, and Lang Tong. "Pricing Multi-Interval Dispatch under Uncertainty Part I: Dispatch-Following Incentives". In: *arXiv:1911.05784 [cs, eess]* (Dec. 2020). arXiv: 1911.05784. URL: http://arxiv.org/abs/1911.05784 (visited on 12/30/2020).

[58] Ye Guo, Cong Chen, and Lang Tong. "Pricing multi-interval dispatch under uncertainty part I: Dispatch-following incentives". In: *IEEE Transactions on Power Systems* (2021).

[59] X.S. Han, H.B. Gooi, and D.S. Kirschen. "Dynamic economic dispatch: feasible and optimal solutions". In: *IEEE Transactions on Power Systems* 16.1 (Feb. 2001). Conference Name: IEEE Transactions on Power Systems, pp. 22–28. ISSN: 1558-0679. DOI: 10.1109/59.910777.

[60] M. Hasanuzzaman, R. Saidur, and N. A. Rahim. "Energy savings in the combustion based process heating in industrial sector". In: *IEEE Access* 1 (May 2013), pp. 475–485. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2013.2260814.

[61] Karen Herter. "Residential implementation of critical-peak pricing of electricity". In: *Energy policy* 35.4 (2007), pp. 2121–2130.

[62] William W Hogan. "Contract networks for electric power transmission". In: *Journal of regulatory economics* 4.3 (1992), pp. 211–242.

[63] William W Hogan. "ELECTRICITY MARKET DESIGN: MULTI-INTERVAL PRICING MODELS". en. In: (). Accessed 2023-03-23. URL: https://scholar.harvard.edu/files/whogan/files/hogan_hepg_multi_period_062220.pdf.

[64] William W Hogan. "Electricity Market Design: Multi-Interval Pricing Models". en. In: (2020), p. 40. URL: http://lmpmarketdesign.com/papers/hogan_hepg_multi_period_062220.pdf.

[65] William W Hogan. *Electricity Market Design: Multi-Interval Pricing Models*. 2020.

[66] William W Hogan. "Electricity market restructuring: reforms of reforms". In: *Journal of Regulatory Economics* 21.1 (2002), pp. 103–132.

[67] William W Hogan and Brendan J Ring. "On minimum-uplift pricing for electricity markets". In: *Electricity Policy Group* (2003), pp. 1–30.

[68] Bowen Hua et al. "Pricing in Multi-Interval Real-Time Markets". en. In: *IEEE Transactions on Power Systems* 34.4 (July 2019), pp. 2696–2705. ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2019.2891541. URL: https://ieeexplore.ieee.org/document/8606259/ (visited on 12/30/2020).

[69] Bowen Hua et al. "Pricing in multi-interval real-time markets". In: *IEEE Transactions on Power systems* 34.4 (2019), pp. 2696–2705.

[70] A Rezaee Jordehi. "Optimisation of demand response in electric power systems, a review". In: *Renewable and sustainable energy reviews* 103 (2019), pp. 308–319.

[71] B. Kirby and M. Milligan. *Method and Case Study for Estimating the Ramping Capability of a Control Area or Balancing Authority and Implications for Moderate or High Wind Penetration: Preprint*. English. Tech. rep. NREL/CP-500-38153. National Renewable Energy Lab., Golden, CO (US), May 2005. URL: https://www.osti.gov/biblio/15016292 (visited on 10/03/2021).

[72] Daniel S Kirschen. "Demand-side view of electricity markets". In: *IEEE Transactions on power systems* 18.2 (2003), pp. 520–527.

[73] Xiaolong Kuang et al. "Pricing Chance Constraints in Electricity Markets". In: *IEEE Transactions on Power Systems* 33.4 (July 2018). Conference Name: IEEE Transactions on Power Systems, pp. 4634–4636. ISSN: 1558-0679. DOI: 10.1109/TEC.2018.2821922.

[74] LBNL. "2025 California Demand Response Potential Study". In: (2017).

[75] John O Ledyard. "Incentive compatibility". In: *Allocation, Information and Markets*. Springer, 1989, pp. 141–151.

[76] Yen-Yu Lee et al. "New indices of market power in transmission-constrained electricity markets". In: *IEEE Transactions on Power Systems* 26.2 (2010), pp. 681–689.

[77] Tongxin Li, Lucien Werner, and Steven H Low. "Learning Graph Parameters from Linear Measurements: Fundamental Trade-offs and Application to Electric Grids". In: *2019 IEEE 58th Conference on Decision and Control (CDC)* (2019), pp. 6554–6559. DOI: 10.1109/CDC40024.2019.9029949.

[78] Tongxin Li, Lucien Werner, and Steven H Low. "Learning graphs from linear measurements: Fundamental trade-offs and applications". In: *IEEE Transactions on Signal and Information Processing over Networks* 6 (2020), pp. 163–178. DOI: 10.1109/TSIPN.2020.2975368.

[79] Tongxin Li et al. "Information Aggregation for Constrained Online Control". In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 5.2 (June 2021), 18:1–18:35. DOI: 10.1145/3460085. URL: https://doi.org/10.1145/3460085 (visited on 09/23/2021).

[80] Yingying Li, Subhro Das, and Na Li. "Online Optimal Control with Affine Constraints". In: *arXiv:2010.04891 [cs, eess, math]* (Oct. 2020). arXiv: 2010.04891. URL: http://arxiv.org/abs/2010.04891 (visited on 04/20/2021).

[81] George Liberopoulos and Panagiotis Andrianesis. "Critical review of pricing schemes in markets with non-convex costs". In: *Operations Research* 64.1 (2016), pp. 17–31.

[82] Minghong Lin et al. “Online algorithms for geographical load balancing”. In: *2012 International Green Computing Conference (IGCC)*. June 2012, pp. 1–10. DOI: 10.1109/IGCC.2012.6322266.

[83] Yanchao Liu, Jesse T Holzer, and Michael C Ferris. “Extending the bidding format to promote demand response”. In: *Energy Policy* 86 (2015), pp. 82–92.

[84] Alvaro Lorca et al. “Multistage Adaptive Robust Optimization for the Unit Commitment Problem”. In: *Operations Research* 64.1 (Feb. 2016), pp. 32–51. ISSN: 0030364X. DOI: 10.1287/opre.2015.1456. (Visited on 10/03/2021).

[85] Álvaro Lorca and Xu Andy Sun. “Multistage Robust Unit Commitment With Dynamic Uncertainty Sets and Energy Storage”. In: *IEEE Transactions on Power Systems* 32.3 (May 2017), pp. 1678–1688. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2016.2593422.

[86] Steven H Low. “Convex relaxation of optimal power flow—Part I: Formulations and equivalence”. In: *IEEE Transactions on Control of Network Systems* 1.1 (2014), pp. 15–27.

[87] Miles Lubin, Yury Dvorkin, and Scott Backhaus. “A Robust Approach to Chance Constrained Optimal Power Flow With Renewable Generation”. In: *IEEE Transactions on Power Systems* 31.5 (Sept. 2016). Conference Name: IEEE Transactions on Power Systems, pp. 3840–3849. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2015.2499753.

[88] James Luedtke and Shabbir Ahmed. “A Sample Approximation Approach for Optimization with Probabilistic Constraints”. en. In: *SIAM Journal on Optimization* 19.2 (Jan. 2008), pp. 674–699. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/070702928. URL: http://epubs.siam.org/doi/10.1137/070702928 (visited on 04/07/2021).

[89] Peter B Luh et al. “Payment cost minimization auction for deregulated electricity markets using surrogate optimization”. In: *IEEE Transactions on Power systems* 21.2 (2006), pp. 568–578.

[90] Johanna L Mathieu, Duncan S Callaway, and Sila Kiliccote. “Examining uncertainty in demand response baseline models and variability in automated responses to dynamic pricing”. In: *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE. 2011, pp. 4332–4339.

[91] Robert Mieth and Yury Dvorkin. “Distribution Electricity Pricing Under Uncertainty”. In: *IEEE Transactions on Power Systems* 35.3 (May 2020). Conference Name: IEEE Transactions on Power Systems, pp. 2325–2338. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2019.2954971.

[92] Robert Mieth, Jip Kim, and Yury Dvorkin. "Risk- and variance-aware electricity pricing". en. In: *Electric Power Systems Research* 189 (Dec. 2020), p. 106804. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2020.106804. URL: https://www.sciencedirect.com/science/article/pii/S0378779620306076 (visited on 03/24/2023).

[93] Mohammad Sadegh Modarresi et al. "Scenario-Based Economic Dispatch With Tunable Risk Levels in High-Renewable Power Systems". In: *IEEE Transactions on Power Systems* 34.6 (Nov. 2019). Conference Name: IEEE Transactions on Power Systems, pp. 5103–5114. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2018.2874464.

[94] Daniel Munoz-Alvarez et al. "Piecewise affine dispatch policies for economic dispatch under uncertainty". en. In: *2014 IEEE PES General Meeting | Conference & Exposition*. National Harbor, MD, USA: IEEE, July 2014, pp. 1–5. ISBN: 978-1-4799-6415-4. DOI: 10.1109/PESGM.2014.6939369. URL: http://ieeexplore.ieee.org/document/6939369/ (visited on 01/19/2022).

[95] Deepan Muthirayan et al. "A minimal incentive-based demand response program with self reported baseline mechanism". In: *IEEE Transactions on Smart Grid* 11.3 (2019), pp. 2195–2207.

[96] Deepan Muthirayan et al. "Mechanism design for demand response programs". In: *IEEE Transactions on Smart Grid* 11.1 (2019), pp. 61–73.

[97] Aramazd Muzhikyan, Amro M. Farid, and Kamal Youcef-Toumi. "Relative merits of load following reserves & energy storage market integration towards power system imbalances". en. In: *International Journal of Electrical Power & Energy Systems* 74 (Jan. 2016), pp. 222–229. ISSN: 0142-0615. DOI: 10.1016/j.ijepes.2015.07.013. URL: https://www.sciencedirect.com/science/article/pii/S0142061515002999 (visited on 09/30/2021).

[98] National Renewable Energy Laboratory. *2022 Annual Technology Baseline (ATB) for Electricity*. Available at: https://atb.nrel.gov/electricity/2022/index. 2022.

[99] Mariola Ndrio, Avinash N. Madavan, and Subhonmesh Bose. "Pricing Conditional Value at Risk-Sensitive Economic Dispatch". In: *2021 IEEE Power & Energy Society General Meeting (PESGM)*. ISSN: 1944-9933. July 2021, pp. 01–05. DOI: 10.1109/PESGM46819.2021.9637845.

[100] Richard G. Newell, Daniel Raimi, and Daniel Shawhan. *US Electricity Markets 101*. 2022. URL: https://www.rff.org/publications/explainers/us-electricity-markets-101/.

[101] James O'Malley. *India is harnessing renewable energy through the world's biggest solar farm. Here's how it happened.* 2022. URL: https://www.

businessinsider.com/india-harnessing-renewable-energy-through-worlds-biggest-solar-farm-2022-11.

[102] Richard P. O'Neill et al. "Efficient market-clearing prices in markets with nonconvexities". en. In: *European Journal of Operational Research* 164.1 (July 2005), pp. 269–285. ISSN: 03772217. DOI: 10.1016/j.ejor.2003.12.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0377221703009196 (visited on 03/30/2023).

[103] *OpenDSS: Open Source Distribution System Simulator*. http://smartgrid.epri.com/SimulationTool.aspx. Accessed: May 2, 2023. 2023.

[104] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org. 2017.

[105] California Independent System Operator. *California ISO records banner year for renewables integration*. URL: http://www.caiso.com/Documents/CaliforniaISORecordsBannerYearforRenewablesIntegration.pdf.

[106] U.A. Ozturk, M. Mazumdar, and B.A. Norman. "A solution to the stochastic unit commitment problem using chance constrained programming". In: *IEEE Transactions on Power Systems* 19.3 (Aug. 2004), pp. 1589–1598. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2004.831651.

[107] Nikolaos G. Paterakis et al. "Load-Following Reserves Procurement Considering Flexible Demand-Side Resources Under High Wind Power Penetration". In: *IEEE Transactions on Power Systems* 30.3 (May 2015), pp. 1337–1350. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2014.2347242.

[108] Johannes P. Pfeifenberger and Kathleen Spees. *Resource Adequacy Requirements: Reliability and Economic Implications*. Tech. rep. The Brattle Group, Sept. 2013. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.641.5126&rep=rep1&type=pdf.

[109] Bala Kameshwar Poolla et al. "Wasserstein Distributionally Robust Look-Ahead Economic Dispatch". In: *IEEE Transactions on Power Systems* 36.3 (May 2021). arXiv:2003.04874 [cs, eess, math], pp. 2010–2022. ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2020.3034488. URL: http://arxiv.org/abs/2003.04874 (visited on 02/10/2023).

[110] James E. Price and Mark Rothleder. "Recognition of extended dispatch horizons in California's energy markets". In: *2011 IEEE Power and Energy Society General Meeting*. ISSN: 1944-9925. July 2011, pp. 1–5. DOI: 10.1109/PES.2011.6039484.

[111] "Quarterly Energy Dynamics Q4 2020". In: (Jan. 2021). URL: %22https://aemo.com.au/-/media/files/major-publications/qed/2020/qed-q4-2020.pdf?la=en%5C&hash=BE12D7DEBE790078B5E7C047FB7BD0BB.

[112] Sepehr Ramyar, Andrew Lu Liu, and Yihsu Chen. "Power market model in presence of strategic prosumers". In: *IEEE Transactions on Power Systems* 35.2 (2019), pp. 898–908.

[113] Syama Sundar* Rangapuram et al. "End-to-end learning of coherent probabilistic forecasts for hierarchical time series". In: *International Conference on Machine Learning* (2021), pp. 8832–8843.

[114] *Renewable Power Generation Costs in 2021*. July 2022. URL: https://www.irena.org/publications/2022/Jul/Renewable-Power-Generation-Costs-in-2021.

[115] David Roberts. *Utilities fighting against rooftop solar are only hastening their own doom*. 2017. URL: https://www.vox.com/energy-and-environment/2017/7/7/15927250/utilities-rooftop-solar-batteries.

[116] Navid Azizan Ruhi et al. "Opportunities for price manipulation by aggregators in electricity markets". In: *IEEE Transactions on Smart Grid* 9.6 (2017), pp. 5687–5698.

[117] Pablo A. Ruiz and George Gross. "Short-Term Resource Adequacy in Electricity Market Design". In: *IEEE Transactions on Power Systems* 23.3 (Aug. 2008), pp. 916–926. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2008.926094.

[118] Fred C Schweppe et al. *Spot pricing of electricity*. Springer Science & Business Media, 2013.

[119] California State Senate. *100 Percent Clean Energy (SB100)*. Sept. 2017. URL: https://focus.senate.ca.gov/sb100.

[120] Manajit Sengupta et al. "The National Solar Radiation Data Base (NSRDB)". In: *Renewable and Sustainable Energy Reviews* 89.June (2018), pp. 51–60. DOI: 10.1016/j.rser.2018.02.014.

[121] Ming Shi, Xiaojun Lin, and Sonia Fahmy. "Competitive Online Convex Optimization With Switching Costs and Ramp Constraints". In: *IEEE/ACM Transactions on Networking* 29.2 (Apr. 2021). Conference Name: IEEE/ACM Transactions on Networking, pp. 876–889. ISSN: 1558-2566. DOI: 10.1109/TNET.2021.3053910.

[122] Pierluigi Siano. "Demand response and smart grids—A survey". In: *Renewable and sustainable energy reviews* 30 (2014), pp. 461–478.

[123] Ognjen Stanojev et al. "Tractable Identification of Electric Distribution Networks". In: *arXiv preprint arXiv:2304.01615* (2023).

[124] Bo Sun et al. "ORC: An Online Competitive Algorithm for Recommendation and Charging Schedule in Electric Vehicle Charging Network". en. In: *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*. Virtual Event Australia: ACM, June 2020, pp. 144–155. ISBN: 978-1-4503-8009-6. DOI: 10.1145/3396851.3397727. URL: https://dl.acm.org/doi/10.1145/3396851.3397727 (visited on 10/04/2021).

[125] Hongbin Sun et al. "Review of challenges and research opportunities for voltage control in smart grids". In: *IEEE Transactions on Power Systems* 34.4 (2019), pp. 2790–2801.

[126] Synapse Energy Economics, Inc. *Interactive Map of U.S. Power Plants*. https://www.synapse-energy.com/tools/interactive-map-us-power-plants. Accessed: May 23, 2023.

[127] U.S. Energy Information Administration. *Annual Electric Power Industry Report*. https://www.eia.gov/electricity/annual/pdf/epa.pdf. 2019.

[128] U.S. Energy Information Administration. *Preliminary Monthly Electric Generator Inventory (based on Form EIA-860M as a supplement to Form EIA-860)*. https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=epmt_1_01. 2021.

[129] U.S. Energy Information Administration. *U.S. Energy Information Administration (EIA) - Today in Energy*. https://www.eia.gov/todayinenergy/detail.php?id=55719. Accessed: May 20, 2023.

[130] *VOR123: Value-of-resilience, Unleashing Community Microgrids to deliver cost-effective resilience benefits to businesses, municipalities, and communities*. Tech. rep. Clean Coalition, 2022. URL: https://clean-coalition.org/disaster-resilience/#adder.

[131] Beibei Wang and Benjamin F. Hobbs. "A flexible ramping product: Can it help real-time dispatch markets approach the stochastic dispatch ideal?" en. In: *Electric Power Systems Research* 109 (Apr. 2014), pp. 128–140. ISSN: 0378-7796. DOI: 10.1016/j.epsr.2013.12.009. URL: https://www.sciencedirect.com/science/article/pii/S0378779613003386 (visited on 09/26/2021).

[132] Fei Wang et al. "Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description". In: *IEEE Transactions on Smart Grid* 9.6 (2018), pp. 6972–6985.

[133] Qin Wang and Bri-Mathias Hodge. "Enhancing Power System Operational Flexibility With Flexible Ramping Products: A Review". In: *IEEE Transactions on Industrial Informatics* 13.4 (Aug. 2017), pp. 1652–1664. ISSN: 1941-0050. DOI: 10.1109/TII.2016.2637879.

[134] J. Warrington et al. “Robust reserve operation in power systems using affine policies”. In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. ISSN: 0743-1546. Dec. 2012, pp. 1111–1117. DOI: 10.1109/CDC.2012.6425913.

[135] Joseph Warrington et al. “Rolling Unit Commitment and Dispatch With Multi-Stage Recourse Policies for Heterogeneous Devices”. In: *IEEE Transactions on Power Systems* 31.1 (Jan. 2016), pp. 187–197. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2015.2391233.

[136] Lei Wu, Mohammad Shahidehpour, and Tao Li. “Stochastic Security-Constrained Unit Commitment”. In: *IEEE Transactions on Power Systems* 22.2 (May 2007), pp. 800–811. ISSN: 1558-0679. DOI: 10.1109/TPWRS.2007.894843.

[137] Le Xie and Marija D. Ilic. “Model predictive economic/environmental dispatch of power systems with intermittent resources”. In: *2009 IEEE Power & Energy Society General Meeting*. ISSN: 1932-5517. July 2009, pp. 1–6. DOI: 10.1109/PES.2009.5275940.

[138] M. Xu and X. Zhuan. “Optimal planning for wind power capacity in an electric power system”. en. In: *Renewable Energy* 53 (May 2013), pp. 280–286. ISSN: 0960-1481. DOI: 10.1016/j.renene.2012.11.015. URL: https://www.sciencedirect.com/science/article/pii/S0960148112007185 (visited on 09/30/2021).

[139] Jinye Zhao, Tongxin Zheng, and Eugene Litvinov. “A Multi-Period Market Design for Markets With Intertemporal Constraints”. en. In: *IEEE Transactions on Power Systems* 35.4 (July 2020), pp. 3015–3025. ISSN: 0885-8950, 1558-0679. DOI: 10.1109/TPWRS.2019.2963022. URL: https://ieeexplore.ieee.org/document/8945389/ (visited on 12/30/2020).

[140] Jinye Zhao, Tongxin Zheng, and Eugene Litvinov. “A multi-period market design for markets with intertemporal constraints”. In: *IEEE Transactions on Power Systems* 35.4 (2019), pp. 3015–3025.

[141] Ella Zhou, David Hurlbut, and Kaifeng Xu. *A Primer on FERC Order No. 2222: Insights for International Power Systems*. Tech. rep. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2021.