Chapter 1

Demand Response for Computing Centers

Jeffrey S. Chase

Duke University

1.1 Introduction

The term *Demand Response (DR)* refers to policies or procedures to influence the timing or location of power demand in response to signals from the electricity supplier about energy production cost or availability. DR is an important element of "smart grid" initiatives to improve the reliability and efficiency of electrical power grids.

DR is a form of *demand-side management*, a term that refers to any means to manage the balance of electricity supply and demand in an electrical grid by influencing or modulating electricity demand, instead of or in addition to the conventional approach of modulating supply. DR is complementary to demand-side energy efficiency, another form of demand-side management.

Effective demand-side management can reduce environmental impact and operating cost for energy consumers. For example, energy-efficient computing, the primary focus of this book, influences demand by reducing the amount of energy consumed to perform a computational task. Advances in energy efficiency of computing centers reduce their operating costs and environmental

impact in an obvious and direct way: each unit of energy not consumed is one less unit to generate, transmit, and pay for. In particular, the "negawatts" saved by energy efficiency can substitute directly for megawatts produced by burning dirty and expensive fossil fuels [22].

DR offers similar benefits in an indirect way. In contrast to energy-efficient computing, the purpose of DR is not to reduce the amount of energy consumed for any given computing task. Rather, the purpose of DR is to reduce the cost for each unit of energy consumed by controlling when and where that unit is consumed, in order to consume it at a time and place with a low unit cost for energy. DR for computing centers involves scheduling and/or placement of computing loads in a way that considers the availability and cost of the electricity to run those loads. The cost metric may incorporate electricity prices, environmental impact, or other measures.

The role of DR in "green HPC" reflects a holistic view of computing and the electricity supply grid as an end-to-end system. In this holistic view, the ultimate measure of energy efficiency is the value of service delivered per unit of fuel consumed or pollution produced. The value derives from the benefit that the information technology service provides to its users (IT value). Effective DR can enhance energy efficiency on the supply side, even if it does not reduce the amount of electricity needed to produce a given unit of IT value. In particular, DR strategies can enhance end-to-end efficiency by shifting the electricity demand away from dirty electricity generators and onto clean energy, or by using energy opportunistically that might otherwise be wasted. DR strategies are also essential to functioning within supply constraints caused by power budgets [27, 25], brownout events [7], or intermittent generation [31, 29], e.g., local solar or wind power. Another form of DR is migrating workload in an Internet-scale service to exploit price disparities in regional electricity markets [26, 20].

One challenge of DR is that it often involves tradeoffs in the value of service produced. In general, making computing systems more energy-efficient enables them to produce the same IT value with less energy, and hence lower operating cost. In contrast, DR strategies entail some measurable reduction in service quality, and therefore may reduce IT value. For example, a DR strategy might incorporate admission control—the choice to deny or cancel a request for computing service during a period of high energy cost. A DR strategy might also defer or throttle a task, or migrate it to a remote provider; any of these choices could reduce the IT value by increasing response time. Another alternative is to reduce the demand for computing power by degrading result quality [2, 10].

Thus DR planning for computing facilities and data centers requires a careful consideration of the impact on IT value. In general, DR strategies are most suitable for what we might call *delay-tolerant computing*. For example, batch job workloads in HPC environments may be less sensitive to response time than interactive Web services or other data center applications.

Several intersecting trends suggest that effective DR will be an important

design goal for automated load management in computing centers that draw their electrical power from future smart grids. This chapter addresses the following questions:

- How does DR enhance energy efficiency on the supply side? Section 1.2 summarizes the role of DR in "greening" the electrical system to reduce fossil fuel use and carbon emissions.
- How does DR reduce electricity costs for facilities that can shift loads? Section 1.3 gives an overview of electricity pricing models and trends that increase the incentives for adaptive load control.
- Are computing facilities and data centers promising targets for DR strategies? Section 1.4 gives an overview of some factors and tradeoffs that determine their suitability and potential to employ DR.
- What factors influence the potential cost savings from DR in computing facilities? What impact does DR have on service quality? Section 1.5 develops a simple analytical model to understand the tradeoffs inherent in DR strategies for batch job scheduling. In particular, it illustrates the key factors that influence DR effectiveness in computing centers: facility load factor (utilization), surplus capacity, facility-scale energy proportionality, and electricity pricing factors.
- How do other changes to energy practices for computing facilities interact with DR? Section 1.6 discusses the impact of advances in facility-scale energy proportionality and dynamic pricing of cloud computing services.

1.2 Demand Response in the Emerging Smart Grid

DR is motivated by a need to balance electricity supply and demand at all levels of the power grid. Electrical grids have little or no energy storage capacity to use as a buffer, so supply must match demand at any point in time. If generation exceeds demand, then energy is wasted. If generation is insufficient to match demand, then outages may occur.

Electricity demand is highly dynamic. Fortunately, electrical demand over a region is predictable with sufficient accuracy and precision to enable a wide range of options for proactive management, including DR strategies. The installed base of electricity-consuming devices changes relatively slowly, and their usage patterns are generally driven by a few primary factors, such as weather, which can be predicted days or hours in advance.

As demand changes, suppliers must modulate generation to match the demand. DR offers a complementary response option: if demand exceeds supply, then reduce demand from selected electrical devices to match the current supply, instead of or in addition to increasing supply to meet the demand. DR offers a potential to improve end-to-end efficiency by avoiding reliance on high-cost generators, which are used primarily during periods of peak electricity demand (Section 1.2.1). DR is also an important tool to manage an electricity supply that is itself increasingly dynamic and difficult to modulate. For example, DR becomes more important as grids incorporate a larger share of fuel-free renewable electricity sources into the generation mix (Section 1.2.2).

1.2.1 Importance of Demand Response for Energy Efficiency

To satisfy dynamic demands, electrical suppliers maintain a mix of generating assets with various properties. As demand increases, suppliers dispatch their generating resources according to a plan that attempts to minimize their overall supply costs. Economic dispatch planning may be influenced by a range

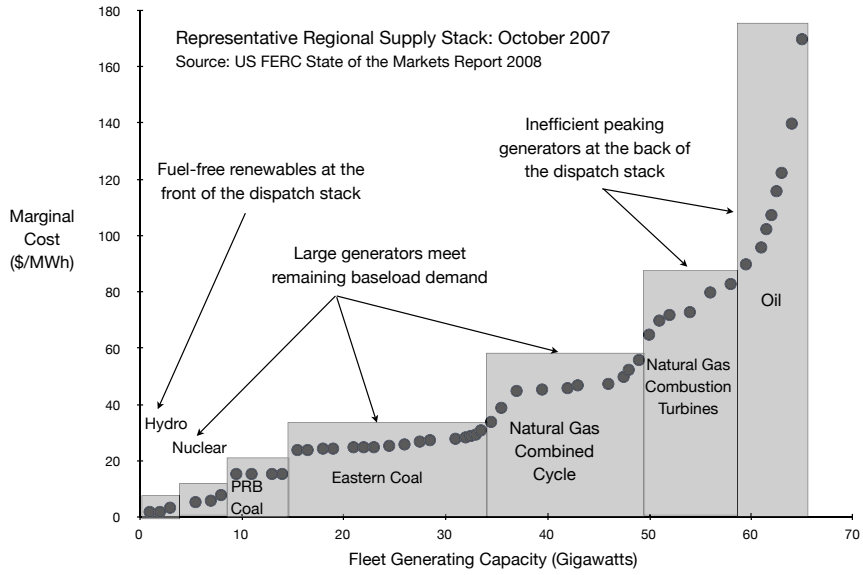

FIGURE 1.1: This figure illustrates a representative supply/dispatch stack for a US region with 65 GW of dispatchable generating capacity. The y-axis shows the marginal (e.g., fuel) cost of power from each generator, ordered by their priority in the dispatch stack. The least efficient generators have the highest marginal cost and are held in reserve for periods of high demand or constrained supply.

of factors, including predictions of how long the demand will last, and the cost of transmission from the candidate generating plant to the load.

Although dispatch planning is complex, the dominating factor is a rank ordering of generators according to least marginal operating cost. The portfolio of generating assets is known as the *dispatch stack*, suggesting a relatively static order of dispatch from preferred plants that run continuously to higher-cost power plants that are used only when needed. Figure 1.1 illustrates a representative dispatch stack [8]. The plants dispatched last are the generators with the highest operating costs for fuel and emissions. These *stand-by* or *peaking* generators are used only when demands cannot be met from other sources. A 2007 Edison Electrical Institute report suggests that 20% of US generating capacity is used less than 10% of the time [5]. Only 42% of generation capacity is used 100% of the time (this base demand level is known as *baseload*).

Plants designed as peaking plants are intended to be used rarely, so they often lack efficiency features that would increase their capital cost. For example, a typical peaking power plant is a simple gas turbine that is significantly less efficient than combined-cycle gas plants that capture waste heat, as shown in Figure 1.1. The combined-cycle gas plants are cheaper to operate than simple gas turbines, but they are more expensive to build. The back of the dispatch stack also includes some of the dirtiest legacy plants.

Demand Response strategies can improve overall efficiency and reliability by limiting the peak and reducing the use of inefficient stand-by generation. For this reason, the US Energy Independence and Security Act of 2007 (EISA) mandates comprehensive planning and assessment of DR options for the electrical grid in the United States [9]. The 2009 US National Assessment of Demand Response Potential [18] suggests that DR strategies have potential to enable a 10-20% reduction in peak electricity demand relative to current projections, rendering 188 GW of reserve generating capacity unneeded in 2019. These reductions could allow earlier retirement of legacy assets, and free up resources and capital for investments in clean energy and energy efficiency.

1.2.2 The Role of Renewable Energy

Increasingly, the generating mix is being supplemented with the subclass of "renewable" assets that harvest natural energy flows such as wind and solar, rather than consuming fuel to generate power. Wind plants now make up almost half of new installed capacity in the US, and fuel-free renewables are the fastest-growing class of new capacity [8]. They have high capital cost relative to fossil fuel plants, but once installed they incur no costs for fuel or emissions.

Fuel-free renewables increase the importance of automated DR for two reasons. First, their near-zero operating cost places them at the front of the dispatch stack: by providing clean energy for free, they increase the relative (marginal, unburdened) cost of serving loads with fuel-driven generators. In

turn, this effect increases the relative benefit of damping the peak demand. Second, fuel-free power generators are intermittent, and it is not possible to control their output by modulating an input flow of fuel. These properties suggest that the burden of modulating the balance must shift to the demand side as they become more prevalent.

In principle, a computing facility under automated control can modulate its power demand at a fine time granularity to match a dynamic power budget. Researchers have begun to speculate how future DR strategies could play a role in accelerating deployment of renewables colocated with computing centers [29, 31]. These ideas are a first step to developing server backbone infrastructure that can continue to function, perhaps in a degraded mode, if access to fuel-generated power is disrupted.

Another relevant property of fuel-free renewables is that their capital cost is roughly linear with capacity even in small installations, thus they disrupt the economies of scale that motivated large, centralized generators in the past. Amory Lovins and other leading energy analyists have argued forcefully that this incremental scalability acts against inherent "diseconomies of scale" in centralized electricity generation and distribution [23]. Small-scale deployments distribute capital costs for generating assets, make use of the fragmented available space (e.g., rooftop solar), and reduce transmission costs and losses. They are also the building blocks of "smart microgrids" that can meet local power demands autonomously in the event that the supply of power from the grid backbone is disrupted [14, 13]. To encourage investment in distributed generation, some states have enacted *net metering* laws and *feed-in tariffs* that allow small private renewable energy systems to provide their surplus power to the grid for credit or payment.

These various factors should continue to drive the future power grid toward a larger number of distributed, smaller-scale, weakly controlled, intermittent power sources. In turn, that will add pressure on smart grid control software to balance the increasingly dynamic supply with the dynamic demand. This prospect suggests that DR will become an increasingly important element of integrated control strategies.

1.3 Electricity Pricing: A View to the Future

DR policy choices are driven by conditions in the power network, e.g., congestion, unanticipated demand, changes in supply output, or failure of assets for generating or transmitting electricity. Therefore, a DR strategy requires some stream of information about current or anticipated conditions in the power network. This information acts as a feedback signal from the electricity supplier to the consumer to modulate the consumer's demand.

The nature of the feedback signal is defined by the contract between the

electricity supplier and consumer. Some service contracts allow the supplier to modulate demand directly within certain bounds, in return for a lower tariff rate (Section 1.3.1). A more flexible feedback signal is a variable electricity price that reflects real-time supply and demand conditions (Section 1.3.2). Electricity contracts with hybrid forms of variable pricing are common in the electricity market today, reflecting various balances in the allocation of cost and risk among suppliers and consumers (Section 1.3.3). These contracts continue to evolve.

One premise of this chapter is that computing centers will have increasing exposure to variable pricing for power in the future, and will increasingly use DR as a tool to manage their costs and risks. For example, given an adaptive load control algorithm to curtail demand during price spikes, a consumer may lower its overall electricity costs by taking more of the supplier's price risk onto itself, in return for a lower average price.

It is also common for electricity contracts to include a charge for the customer's peak demand over a billing period, in addition to the energy usage charge. For example, a contract might specify a per-kW charge for the average demand over the 15-minute sampling interval with the maximum average demand among all sampling intervals in the billing period. For these contracts, DR strategies can also reduce charges by suppressing the demand peaks.

1.3.1 Dispatchable Demand Response

One simple form of DR contract is an *interruptible tariff*, which grants the supplier (a utility) a right to command the customer to reduce its demand according to prearranged terms. With *direct load control*, the utility issues direct commands to devices on the customer premises, e.g., to modulate systems for heating, cooling, pumping, or battery charging. Alternatively, the customer may simply agree to curtail load to a fixed level or by a fixed amount on command from the provider, but retain control over how to meet the target. Customers enter into these agreements in exchange for some payment or pricing incentive [28].

In these agreements, the utility manages the control algorithm to initiate the demand response in conjunction with capacity dispatch planning. In essence, the customer's DR commitment is a *dispatchable* resource on an equal footing with generating plants under the supplier's control. In 2008, the US Federal Energy Regulatory Commission issued several regulatory changes to treat dispatchable DR resources comparably to new generating capacity with respect to market function and dispatch planning [8].

Dispatchable DR agreements are most suitable when the DR policy choices made by the utility have negligible impact on the customer. In some electrical devices demand may be scheduled or shifted in time for short periods without impairing the function of the device. For example, consider a device that has a target running time over specific time intervals, such as a system for battery charging. A control algorithm can modulate the duty cycle over shorter time

intervals without missing the target. Other energy-hungry devices maintain a buffer against a leakage or drain rate: examples include pumping systems to maintain a water reservoir level, or thermal control in buildings, water heaters, or refrigeration. For these devices, modulating the duty cycle may cause the system to drift from a target objective, but this drift is acceptable within certain tolerances. These systems can be made more DR-tolerant by extending the buffer in some way, e.g., by increasing the size of the reservoir, or by adding insulation or thermal mass.

In contrast, DR for computing services involves managing service quality tradeoffs that may be dependent on the applications or load conditions within the center (Section 1.4). It is more suitable to arrangements that allow the center operator to control these tradeoffs. Even so, dispatchable DR arrangements are already present in the data center market. For example, some companies (e.g., *enernoc.com*) act as third-party Curtailment Service Providers to broker dispatchable demand reductions and mediate between data center operators and electrical utilities in managing peak loads.

1.3.2 Variable Pricing

A more general alternative to drive DR strategies is to offer variable pricing that reflects varying supply costs through time to the customer. This approach gives less control to the utility, but it offers more flexibility to the customers to manage their own demand.

Variable pricing is a foundation of smart grid technologies. Wholesale electricity markets with dynamic pricing are currently operating in most regions of the United States. These competitive wholesale markets, administered by

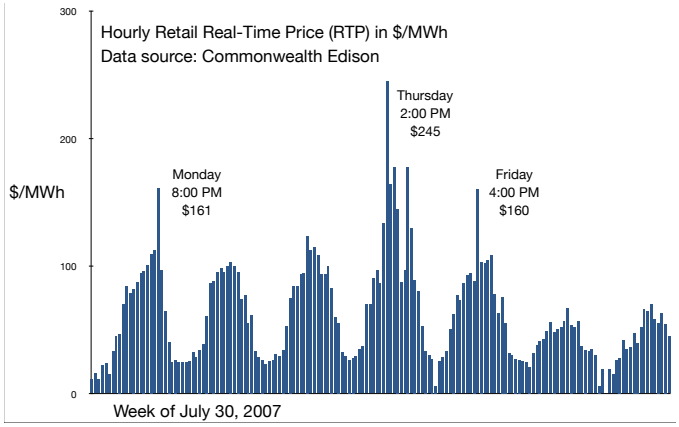

FIGURE 1.2: Real-time prices (RTP) are affected by diurnal activity cycles and hot weather. Peak demand and peak prices often occur on weekday afternoons. Prices during this hot summer week varied by an order of magnitude.

regional transmission organizations (RTOs) or independent system operators (ISOs), serve more than two-thirds of US electricity customers [34]. These markets use bidding protocols to set a dynamic price on electricity for delivery over specific time intervals within a given transmission region, e.g., on an hourly basis, or for spot intervals as short as 5-15 minutes.

While some very large computing centers may purchase electricity in the wholesale market, effective DR generally requires dynamic pricing in the retail markets where the vast majority of end users obtain their power. Retail pricing is decoupled from wholesale prices in most regional electricity markets in the United States; in 2009, penetration of dynamic (real-time) pricing at the retail level was still insignificant [9]. This decoupling is largely an artifact of older regulatory regimes that emphasized stable and predictable electricity pricing for consumers. The regulatory climate is changing to integrate more demand-side load management into the grid, including variable pricing schemes at the retail level [34].

To understand why, consider the effect of fixed-price regimes. Fixed-rate pricing is easy for customers and offers price stability, but providers bear the risk of price swings in the wholesale market. To ensure a profit, they must set the fixed-rate price at a sufficiently high level to balance this risk: the fixed price must be higher than the demand-weighted average of the wholesale price, or the retail supplier loses money. Thus the retail price must reflect not just the marginal cost of generation, but also the risk of supply constraints and price spikes in a dynamic wholesale market.

One straightforward variable pricing scheme is to pass the wholesale price directly to the consumer, e.g., by deriving the retail price from the wholesale price according to some preagreed function. This dynamic pricing is known as retail real-time pricing (RTP). Figure 1.2 shows retail prices from an RTP pilot in the State of Illinois: retail prices fluctuate by the hour according to market conditions, and customers are notified by SMS or e-mail before the end of the business day if prices will exceed some user-specified threshold at any time during the following day. Since RTP customers take the risk of price fluctuations in the wholesale market, they should see lower average prices. Although they are exposed to price spikes, they have an opportunity to reduce their costs by limiting their usage during high-price periods. Even at the residential level, price-responsive demand reductions have potential to damp wholesale price spikes [34], reducing costs for the market as a whole. The number of retail market suppliers offering RTP options to their customers increased by two-thirds between 2006 and 2008 [18].

One limiting factor for RTP and other forms of variable pricing is that they require advanced metering infrastructure (AMI) to monitor customer usage through time. Standard old-style electricity meters measure cumulative consumption, but do not record when the consumption occurred. This missing information is needed to bill the customer under a variable pricing regime. Metering devices that account usage through time had only about 5% penetration in US electricity markets in 2008 [18]. The US government has

provided various incentives for AMI deployment beginning with the Energy Policy Act of 2005 (EPAct).

1.3.3 Hybrid Pricing Models

Where variable pricing is available, various pricing and contract models have evolved that combine the stability of fixed pricing with the dynamic DR incentives of RTP, to varying degrees [5]. Variable pricing schemes and incentives may incorporate any of several common peak-pricing elements, or blend them to distribute costs and risks between the provider and consumer.

- *Time-of-use (TOU)* is a predictable form of variable pricing with fixed price levels over specific recurring time periods that are designated in advance according to a schedule. The price schedule may be a standard tariff for customers of a given class (e.g., residences), or a negotiated schedule tailored to specific customers and their demand levels. TOU pricing is already common for commercial and industrial (C&I) consumers in many regions of the US. Figure 1.3 shows the price schedule for a TOU tariff for light commercial customers of Pacific Gas & Electric during summer 2009. Basic TOU pricing reflects only those wholesale price variations that are anticipated at the time the schedule is set: the supplier bears the risk of any unexpected variation in the wholesale price, and must factor this risk into the TOU price levels.

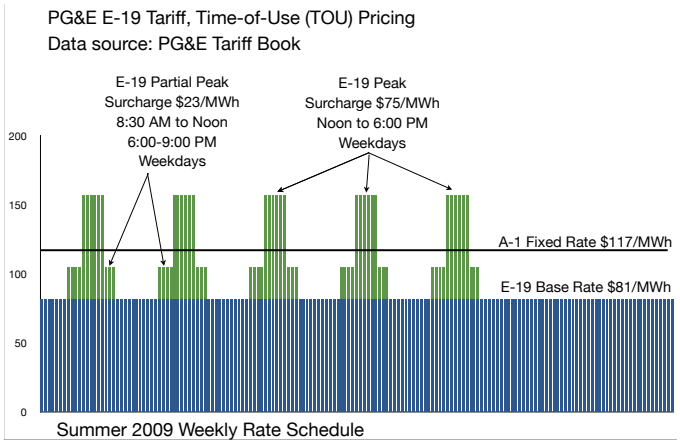

FIGURE 1.3: Electricity may be purchased on a tariff plan that varies prices according to the time of use (TOU) on a regular schedule. TOU pricing enables the customer to plan usage around scheduled surcharge periods that coincide with likely demand peaks.

- *Critical-peak (CPP)* pricing imposes a surcharge during intervals designated by the provider as "critical" due to unexpected supply constraints. CPP is more dynamic than pure TOU, but the supplier must call critical periods with a minimum advance notice, e.g., a day ahead or an hour ahead, and the contract may limit the number of CPP intervals and the CPP price levels. CPP pricing shifts more of the risk of critical periods to the customer, and so should reduce prices during non-critical periods.

- *Customer baseline load (CBL)* contracts specify a fixed price or schedule for a baseline demand level, and different pricing for demand that deviates from the CBL. For example *block and index* pricing is a forward futures contract for a block load at an agreed rate, with demand that deviates from the CBL charged or rebated at the dynamic price.

It is not yet clear how electricity pricing contracts will evolve and what forms they will take in the future. However, there is a clear shift toward more dynamic pricing coupled with incentives for customers that can modulate electricity demand in response to price signals from the electrical grid. The remainder of this chapter assumes that electricity contracts incorporate dynamic pricing for metered usage over specified intervals, and that the consumer controls how and when to modulate its electricity demand to respond to these price signals. To abstract from the pricing alternatives, we may suppose that the customer pays some base rate for electricity together with a surcharge over specific intervals, where both the amount of the surcharge and the surcharge intervals are agreed or announced in advance. It is possible that future models will include competitive bidding for electrical power by large customers, but we do not consider that case further. Pricing factors are discussed in more detail in Section 1.6.2.

1.4 Demand Response and Demand Elasticity for Computing

Computing centers—supercomputers, data centers, and other server ensembles—offer significant Demand Response potential for the following reasons:

- They are large power consumers and their share of electrical demand is growing. The analysis by Koomey [19] concludes that their electricity consumption grew at a rate of 16.7% per year worldwide between 2000 and 2005, and by up to 23% per year in some regions. The same paper estimates a growth rate of 12% per year worldwide between 2005 and 2010. A 2007 EPA study [11] projected that the US data center sector

would require 5 GW of new peak generating capacity over the 2007-2011 period under a baseline scenario.

- They have the means to modulate their power demand by controlling the flow of incoming jobs or requests to servers, or by suspending, resuming, or migrating work that is already in progress. Servers also have an increasingly rich array of platform-level power actuators under software control, which can select tradeoffs of power and performance, or cap the power budget at the granularity of individual servers or server ensembles such as chassis or racks [27]. Recent work has shown how to combine and extend these elements to modulate power usage for systems ranging from virtual machines [25] to "warehouse-sized" data centers [12].
- They increasingly run automated facility-wide policies to schedule and manage load. These policies can incorporate DR strategies to modulate power demand.
- Networking offers opportunities to shift computing loads and their electrical demand from one region to another. DR-aware load placement can address geographic imbalances of electrical supply and demand, even for interactive services that are sensitive to latency and intolerant of deferring work [26, 20].

The technical challenge for DR in computing is then to extend automated resource management policies to consider electricity cost as an optimization objective. These policies include scheduling, admission control, placement and request routing, and resource control.

Effective DR presumes that demand for electricity by a computing facility is elastic and price-responsive. DR strategies respond to higher prices by reducing service, typically substituting service at a later time or a different location. In general, DR for a computing facility compromises service quality by some observable measure. For example, if a DR strategy substitutes off-peak energy use for peak-period energy use, it reduces its demand by deferring work from a peak period to an off-peak period. As a result, any deferred tasks complete later. The degraded service quality is visible through standard measures of responsiveness, e.g., response time or stretch factor, even if the facility has sufficient future surplus capacity to defer work without compromising throughput (see Section 1.5).

A key difficulty is to balance electricity costs against other costs incurred by the candidate response options, e.g., costs to defer, deny, or migrate a computing task. The first challenge is to characterize and predict the impact on service quality. A distinct and perhaps more difficult challenge is to place a monetary value on the degraded service quality, so that its cost may be compared directly against the savings in the electric bill.

To make this more concrete, let us assume:

1. For a given schedule of activity, the facility consumes electricity over a sequence of discrete time intervals t: $electricity(t)$.

2. The facility incurs a cost for consuming electricity according to a function that varies with time, e.g., a base rate plus a variable surcharge: $rate(t)$.

3. For a given schedule of activity, the facility obtains some benefit (IT value) from the work that it completes in each timestep. Let us suppose that this benefit can be represented in a common currency to compare it directly with cost: $benefit(t)$.

The DR objective then is to determine a schedule of activity that maximizes the reward:

$$reward = \sum_t (benefit(t) - rate(t) \times electricity(t)) \tag{1.1}$$

Consider the common case of a computing facility that serves multiple workload components, e.g., jobs or virtual machines running on behalf of different contending users or groups. For example, cloud hosting centers, enterprise computing centers, and supercomputers execute tasks with a range of priority levels and urgency ranging from mission-critical to discretionary. Some workloads offer little opportunity for DR: for example, the IT value of urgent mission-critical tasks is likely to exceed any cost savings of deferring those tasks. Moreover, any new dynamic control incurs some risk of disrupting operations in unexpected ways. As another example, high-throughput computing environments cannot defer valuable work unless they maintain adequate reserve capacity to complete the work later. Section 1.6.1 discusses these practical issues in more detail.

In many cases, such as cloud data centers, the facility is itself a provider that receives revenue from customers according to various service agreements, which may include penalties for violating a service level objective (SLO). Any scheme for arbitrating resources assigns some relative value to the workloads, and uses them to prioritize relative measures to the contending tasks. The difficulty is in mapping these relative measures to an absolute value for the resources they run on, and the power they consume. That means quantifying the impact of policy decisions on service quality of each task, and the cost of that impact on each component of the workload, e.g., on each customer.

We can think of this challenge in terms of the contract that the facility presents to its customers. The contract may be explicit, as in a Service Level Agreement (SLA) between a provider and a customer, or it may be implicit in the definition of the service model for the system. In general, the contract imposes some performance constraint or Service Level Objective (SLO) on the facility. For example, the initial contract for Amazon's Elastic Compute Cloud (EC2) suggests that the provider will allocate to each EC2 instance (a virtual machine) all resources that it requests, up to a specified level encoded in the attributes of each instance type. This service model of a minimum *resource entitlement* (or share) is a defining characteristic of *proportional-share* scheduling systems. Alternatively, an SLO may specify constraints on

direct measures of application performance, such as bounds on a response time quantile or stretch factor.

If the facility's contract is defined exclusively by such constraints, then the facility is free to allocate any surplus resource as it sees fit, once it satisfies the constraints. In particular, the facility is free to allow surplus resources to idle at the discretion of a Demand Response strategy to reduce operating costs. For contracts that specify a penalty for violating the constraints, the DR strategy may choose to violate the constraints and pay the penalty if it is outweighed by other factors [16].

In practice, many computing centers are established by a community to serve its own needs, rather than operated for commercial profit with an explicit contract. Today, these systems typically operate on a "best possible" service model rather than a service constraint. For example, conventional proportional-share service models are defined to be *work-conserving*: any surplus resource is allocated to contending tasks in proportion to their shares, rather than maintained at the discretion of the provider. This means that the user of a proportional-share system has an opportunity to obtain any surplus resources for its own use, competing on a fair footing with other users. Conventional service models with this property are designed with the implicit assumption that the computing resource is a form of public good: although its use is exclusive, any surplus is free and open for use by the community. For example, the popular Condor job scheduler was orginally conceived as a system to "scavenge" these idle resources [21], which would otherwise be wasted.

DR motivates development of new service models that recognize that the surplus is not free. It is an open question how to design service models that allow the provider to balance the operating cost of surplus resources against the value of using them. In essence, the problem reduces to defining *utility functions* that place value ($benefit$) on service to applications. Several systems have experimented with utility-driven scheduling policies (e.g., [1, 16]), some for the explicit purpose of energy management [6, 7]. It is also intriguing to consider how applications themselves could manage these cost/benefit tradeoffs through *reflective control*, in which dynamic pricing for cloud service or power is exposed directly to advanced applications, which respond by modulating their functions and demands [2, 10].

1.5 Evaluating Demand Response: A Simple Model

Consider a system or facility at a single location, executing a workload. Deferring work during high-cost periods can reduce overall cost to run the workload, but it incurs a slowdown. Let us consider a simple model to illus-

trate the factors that influence the potential for cost savings from Demand Response, and the resulting slowdown.

This model focuses primarily on a specific example scenario for DR in computing centers: shifting of batch job workloads in time to minimize cost under a time-varying electricity price. The example scenario defers work to take advantage of lower prices in the future, and thus it presumes that workloads are *delay-tolerant* up to some bound. Batch job systems are an attractive target setting for DR because of their flexibility to schedule load levels through time, given the limited need for interactive response. However, the principles are relevant to other scenarios as well.

To simplify the analysis, suppose that the cost of electricity varies between two levels, a base price and a peak price, with some given regular period. Suppose further that the offered workload consists of a continuous stream of arriving jobs that drive the system at a constant load factor. Figure 1.4 illustrates this scenario. Section 1.6 relates these idealized assumptions to practice.

The model considers a single recurring interval of this schedule, with parameters normalized to the length of the interval, the base energy price, and the system's peak power draw, as illustrated in Figure 1.5. Four key factors characterize the potential cost savings and resulting slowdown of DR:

- *Price variability.* Deferring work reduces cost only when it costs less to do the work later. The simplified pricing model consists of a constant *base price* representing a floor on the price of electricity, with a variable *surcharge* y, normalized to the base price, that captures additional costs due to congestion during peak periods, or other factors. See Figure 1.5(a). The goal of the DR strategy is to schedule work to avoid these surcharges, subject to various constraints. Higher surcharge rates increase the potential savings from a DR strategy.

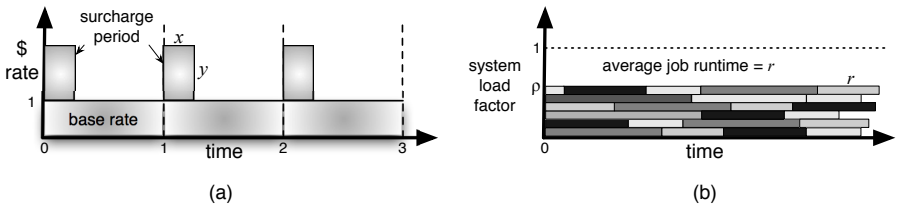

FIGURE 1.4: A simple scenario for the analytical model to illustrate demand response factors. Electrical power is charged at a base rate, with a surcharge of y times the base rate for critical periods of x time units of each interval on a regular schedule. The system's offered load is an idealized job mix that drives the system at a constant load factor ρ with no queuing. The average job execution time is r.

- *Surplus capacity.* The system can defer work only if it has spare capacity to run the work later. Without this surplus capacity, deferring work causes monotonically increasing backlogs and slowdowns for later jobs. We characterize the load level of the system as a utilization or *load factor* ρ as a share of its peak capacity to do work: $0 \leq \rho \leq 1$.[1] See Figure 1.4(b) and Figure 1.5(b). DR is an option only when the system is not saturated: average $\rho < 1$.

- *Surcharge time.* A DR strategy defers work from periods of high surcharge to periods of lower (or zero) surcharge. The opportunity for benefit depends in part on the share x of each interval constituting the *surcharge period* during which surcharges apply. See Figure 1.5(a) and (b).

- *Energy proportionality.* Deferring work can reduce cost only if the system draws less power when it is doing less work. The system power draw is a function of its instantaneous utilization or load factor ρ: $power(\rho)$. Suppose that the system draws a base power i when it is idle, where i is given as a share of the system's peak power. Then $power(\rho)$ ranges between i and the peak, normalized as 1. The energy proportionality of the system can be characterized by its *dynamic range* $1-i$ [3]. A dynamic range of 100% ($i = 0$) corresponds to a fully energy-proportional system.

[1] The load factor ρ may be viewed as a measure of IT asset efficiency, since it represents the utilization of installed capacity of IT assets [17]. It is analagous to (but distinct from) the *load factor* as the term is used in the electricity sector: it is the ratio of average power (or output of work or electricity) to the peak power (or capacity to do work or generate electricity).

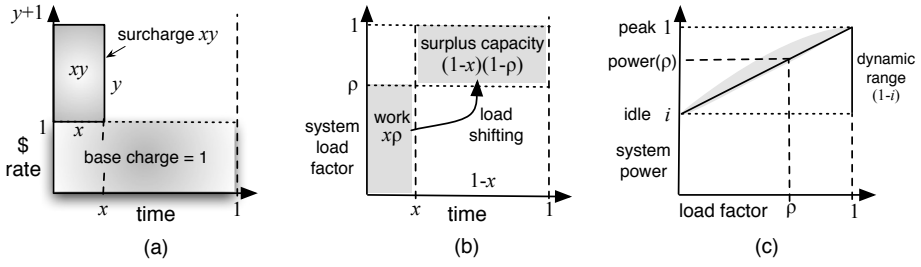

FIGURE 1.5: For the idealized scenario in Figure 1.4, the potential cost savings from demand response is determined by the magnitude (y) and period (x) of on-peak electricity surcharges, the system load factor ρ and off-peak surplus capacity, and the system's energy proportionality. We use a linear approximation of power as a function of load: proportionality is characterized by the *dynamic range* of power consumption from idle (i) to peak, which is the slope of the line.

In this model, the total electricity cost to run the system at full power for one recurring interval is $1 + xy$. Consider the case without DR, in which the system runs at a constant load level ρ. It executes work $x\rho$ during each surcharge period. If the system is perfectly energy-proportional ($i = 0$), then the base cost for energy during the interval is ρ, and it incurs a surcharge of $xy\rho$, for a total per-interval energy cost of $\rho(1 + xy)$.

If the system is not fully energy-proportional, then it is necessary to estimate the amount of power the system can save by shifting some or all of its load over the surcharge period. We consider an idealized model of energy proportionality in which power is linear with load factor ρ: $power(\rho) = i+\rho(1-i)$. The dynamic range is the slope of the line. See Figure 1.5(c). For example, if a system consumes 60% of its peak power even while idling in its lowest-power state ($i = 0.6$), then its power varies across 40% $(1 - i)$ of its range as ρ ranges from 0 to 1, and the slope of the line is the dynamic range $1 - i = 0.4$. The linear model of energy proportionality was used in early work on energy management for server ensembles [7]; it is also suggested by the recent paper on energy-proportional systems by Barroso and Holzle. This idealized model roughly approximates to the behavior of current-generation servers, but it also applies at facility scale [32, 12] (see Section 1.6.3). By this linear model, if the system runs at utilization ρ then we approximate its power draw as $i + \rho(1 - i)$; thus the cost to run the system at utilization ρ for one interval is $(i + \rho(1 - i))(1 + xy)$.

Now consider a DR strategy in this idealized setting. If the DR strategy can defer the $x\rho$ work to a subinterval in which no surcharge applies, then it can idle to consume less power during the surcharge period. During each interval, the system has surplus capacity $(1 - \rho)(1 - x)$ to complete deferred work without incurring a surcharge for the work, and without impacting other work scheduled during the interval. To stay idle when surcharges apply, the system must shift $x\rho$ work onto this surplus capacity. Refer to Figure 1.5(b). It is easy to see that the balance condition reduces to $\rho = 1 - x$. If $\rho > 1 - x$, then the DR strategy lacks sufficient capacity to idle during surcharge times: it must run some work even when surcharges apply to avoid creating a backlog. On the other hand, if $\rho \leq 1 - x$, then the DR strategy can idle during the surcharge period. In general, the DR strategy can minimize its costs by shifting $MIN(x\rho, (1-x)(1-\rho))$ work, and incurs a surcharge for the unshifted residual.

Now consider the impact of the DR strategy on energy cost. We can determine an upper bound on the energy cost savings from DR as follows. Suppose that the system has sufficient surplus capacity to shift all of the work $x\rho$ out of the surcharge period, i.e., $\rho \leq 1 - x$. If the system is perfectly energy-proportional, then it draws zero power while idling ($i = 0$) during the surcharge period, so it can eliminate the surcharge and pay only the base cost ρ for each interval, instead of the cost with surcharge of $\rho(1 + xy)$. Dividing through by ρ, we have the *idealized savings of DR in an energy-proportional system*, measured as a percentage of energy cost:

$$1 - \frac{1}{1+xy} \tag{1.2}$$

In practice, the system is not perfectly energy-proportional, and consumes some power i even when it is not doing work. This effect reduces the potential savings: a DR strategy can reduce the surcharge incurred, but cannot eliminate it. Consider the case where the highest savings occurs: the balance point ($\rho = 1-x$), where the system idles during each surcharge period and otherwise runs at full power and maximum efficiency. Figure 1.7 depicts this scenario. The system incurs a charge of $ix(1+y)$ while idling during each surcharge period: it consumes power i at a cost rate of $(1+y)$ for time x. The off-surcharge cost is again just the base cost ρ: peak power 1 for time $1-x$ at the base rate 1. Thus the *best-case idealized savings of DR under the linear power model*, measured as a percentage of energy cost, becomes:

$$1 - \frac{\rho + ix(1+y)}{(i+\rho(1-i))(1+xy)} \tag{1.3}$$

Figure 1.6 summarizes the interaction of these factors. The figure shows normalized absolute cost savings: how much of the surcharge $xy\rho$ can be avoided. There is no cost to save if $\rho \to 0$, and no opportunity to shift load if $\rho \to 1$. In other cases the cost and potential savings is linear with ρ: the potential savings grows linearly as the system becomes busier and incurs higher costs, but declines linearly when the system has too much load to allow it to idle during the surcharge period. These two lines bound a triangle defining the potential savings. Whatever amount of work is shifted, systems that are more energy-proportional (lower i) save more from shifting that work: thus systems that are not perfectly energy-proportional ($i > 0$) obtain savings given by a point in the interior of the triangle, rather than on an upper edge. The peak savings for a perfectly energy-proportional system is given by a point on the parabola $yx(1-x)$: for any given x, the peak savings and top vertex of the triangle occurs at the point on the parabola where $\rho = 1-x$. Thus the savings of DR is zero if $x \to 0$ (surcharges never apply) or if $x \to 1$ (surcharges always apply). For any point in the triangle, the magnitude of the savings grows linearly with y: savings is unbounded as y increases.

This cost savings from DR comes at the price of a slowdown as work is deferred to avoid surcharges. The system incurs the maximum average slowdown if it idles whenever surcharges apply: $\rho \leq 1-x$. For example, consider again the balance point $\rho = 1-x$ depicted in Figure 1.7. If each job requires r units (intervals) of running time to complete, then under the DR strategy it receives $1-x$ units of service in each interval, and requires $r/(1-x) = r/\rho$ intervals to complete. The additional residence time of each job drives the load factor to 1 when the system is active during non-surcharge periods. The average throughput is unchanged.

This ideal case establishes an upper bound on the slowdown from DR: the stretch factor $1/\rho$. If jobs vary in their runtime around a mean of r, then $1/\rho$

is the average stretch factor: some jobs are slowed less and some are slowed more. In the worst case, a job arrives at the start of a surcharge period, and does not quite complete before the next surcharge period: $r = 1 - x$ (plus ϵ). The job completes in time $1 + x$ instead of time $1 - x$, and the *worst-case stretch factor for short jobs* is:

$$\frac{1+x}{1-x} \tag{1.4}$$

The worst-case stretch factor grows without bound as $x \to 1$. However, the worst case applies only to the shortest jobs. The maximum runtime of a job subject to this worst case is $r = 1 - x$, and $r \to 0$ as $x \to 1$.

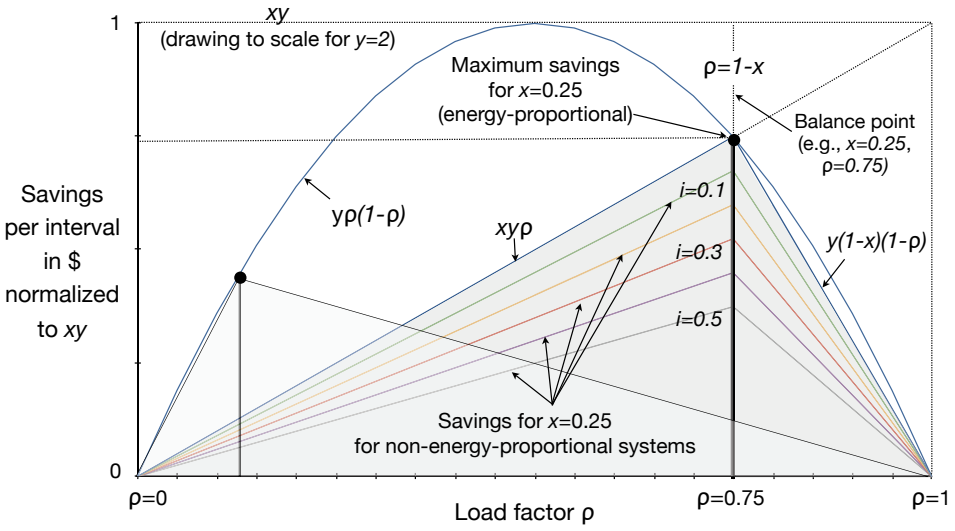

FIGURE 1.6: For any given surcharge time x and surcharge y, the savings is given by a shaded triangle. The system can shift all work out of the surcharge period if $\rho <= 1 - x$: savings grows linearly with the load factor ρ. At higher load factors $\rho > 1 - x$, the system does not have sufficient spare capacity to idle during the surcharge period: savings declines linearly with the spare capacity. For any x and ρ, the savings is always proportional to y: higher y values make the triangle taller. For any given x, y, and ρ, imperfect energy proportionality limits the savings: higher i values make the triangle shorter. Savings approaches zero as $x \to 0$, $x \to 1$, $\rho \to 0$, $\rho \to 1$, or $y \to 0$. The figure is drawn for $x = 0.25$ and $y = 2$.

1.6 Demand Response in Practice

The analytical model is useful to illustrate the key factors that influence effectiveness of a DR strategy. A realistic scenario is likely to be more complicated in several key respects:

- Both the price curves and job properties are more dynamic, and often are not known with certainty in advance. For example, a strategy that defers work may expose itself to risk that it will face an unexpected backlog or incur higher costs later.

- The model presumes that the system has the flexibility to suspend, slow, or migrate jobs as needed to implement the strategy, with zero cost. In practice, a DR strategy may have a limited set of actuators, and it must account for their costs.

- The model presumes that the system is unconstrained by the need to manage varying levels of parallelism in jobs. It does not preclude parallel jobs, but it presumes that it can reach any target utilization level by running some subset of its ready jobs. In practice, certain combinations may be infeasible due to the varying resource requirements of jobs.

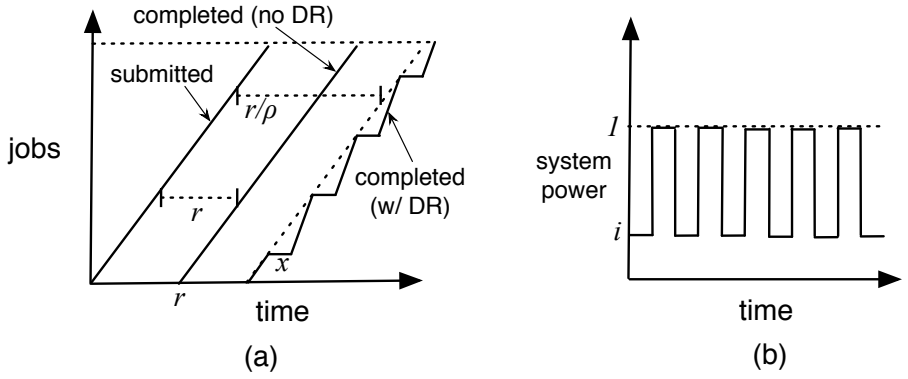

FIGURE 1.7: Job throughput and slowdown (a) and system power (b) under a simple illustrative Demand Response scenario. Each job completes after exactly r units of running time. The system has just enough surplus capacity to idle during surcharge periods, and otherwise runs at full power: $\rho = 1 - x$. Average throughput is not affected, but jobs incur an average stretch factor of $1/\rho$.

It is an open problem to develop DR strategies that can manage these factors in practical online scenarios. The benefits of practical DR strategies will approach those derived from the model, although they may be modestly less. It is also important to consider realistic values of the parameters to estimate what these benefits might be in practice.

1.6.1 Load Factor and Capacity Provisioning

The model estimates the cost reductions possible from DR at a given offered load and a given system capacity: the load factor ρ is the ratio of load to capacity. The model shows that DR can reduce costs if $\rho < 1$, i.e., the system has surplus capacity.

Recent studies of industry data centers suggest that they have substantial surplus capacity. A recent McKinsey study suggests that many industry data centers with mixed workloads are overprovisioned well beyond their need to handle expected load surges [17]. The study suggests that in many cases structural overprovisioning emerges from organizational factors rather than technical considerations. It argues that a primary goal of data center efficiency efforts should be to reduce capacity to match the load; as a benchmark for these efficiency efforts, it introduces an efficiency metric called CADE that is linear with capacity utilization. Overprovisioned systems have high potential for DR, but this benefit has limits: for example, in the model, there can be no advantage to overprovisioning so that ρ is below the balance point $\rho = 1 - x$. Steps taken to improve capacity utilization (higher ρ) , e.g., through server consolidation using virtual machines, do not reduce the potential for DR until they reach this level. Also, there is no energy cost to maintain surplus standby capacity if it is powered off when not in use (see Section 1.6.3).

Another recent study of a well-managed interactive Web service showed that servers spent a large majority of their time with CPU utilizations below 50% [3]. However, Web data centers have limited opportunity to use their surplus capacity for DR savings by deferring requests. First, most Web activity is interactive, and so is relatively inelastic: requests are short and it would disrupt users to defer them. Second, Web service request loads tend to be highly dynamic; studies tend to show regular diurnal request load peaks on weekday afternoons, and flash crowds may also occur. Both electrical grids and Web data centers are provisioned with surplus capacity to handle these peaks. Unfortunately, the peaks often coincide: peak demands on the power grid also tend to occur on weekday afternoons, e.g., when demand is driven by air-conditioning systems. Despite these limitations, recent work has shown that distributed Web services have substantial opportunity to reduce electricity costs by routing requests to take arbitrage regional disparities in electricity prices, even given the interactive response constraints [26, 20].

Batch job systems may also have bursty job arrivals, but batch jobs can often be deferred without disrupting users. This makes batch systems more attractive candidates for DR, but it also means they may tend to run at higher

load factor ρ. Because response time is less crucial, batch systems have less need for surplus capacity to handle peak loads. These systems tend to be provisioned to sustain the throughput needed to serve a target average load.

In a mission-critical computing center that runs at full utilization, $\rho = 1$, DR offers no cost savings without compromising throughput. However, the center can drive ρ down by investing in surplus capacity. Adding surplus capacity improves average response time to users; with DR, it can also reduce operating costs.

Considering the grid and computing center together as an end-to-end system reveals that investments in computing capacity can be compared directly to investments in peaking generation capacity. For example, suppose a large center runs at full capacity ($\rho = 1$) to serve a given job load, and that the power to run the center is drawn from a grid that experiences a demand spike for one hour of each day. If the center has 25 racks at 40kW each, then adding an additional rack permits idling the entire data center during the peak hour without impacting throughput: $\rho = 0.96$. The center delays jobs during the idle period, but offers better service for the rest of the day. Idling the center during the demand spike eliminates the need for one megawatt of peaking generation capacity and any fuel that it consumes.

1.6.2 Price Variability

Prices vary within different locations or regions, according to demand, proximity to generating capacity, and the availability and cost of transmission. In 2008, a year of unstable fuel prices, electricity spot prices in the US fluctuated between \$40/MWh and \$160/MWh. These levels are representative of marginal provider costs (as given in Figure 1.1), but reflect congestion and pricing factors as well [8].

Unusual market conditions occasionally drive real-time market prices well above or below the marginal provider costs; the highest prices exceed the lowest prices by an order of magnitude [5], but may spike above that level in extreme cases. For example, price spikes to \$8000/MWh have occurred during extreme weather events (Northeast US in summer 1999). In California in 2000-2001, electricity suppliers drove wholesale prices to regulatory cap levels (\$1000/MWh). However, it is dangerous to infer too much about price variability from these extreme events in freshly deregulated markets. Indeed, one motivation for DR is that it reduces the market power of suppliers to drive extreme price spikes by withdrawing supply, as apparently occurred in California during the 2000-2001 crisis [30, 4].

Figure 1.2 and Figure 1.3 are more likely to be representative of pricing conditions encountered in practice. For the E-19 tariff in Figure 1.3, the DR model parameters are $x = 0.25$ and $y = 0.92$ if we consider only the on-peak periods. The maximum savings is 18%. Considering both on-peak and partial-peak periods, the parameters are $x = 0.54$ and $y = 0.58$; in this case y is the time-weighted average surcharge for on-peak and partial-peak periods. The

maximum savings from DR is 24%, but it can be obtained only if the center runs at less than half of its capacity: $\rho = 1 - x = 0.46$. PG&E's residential A-6 tariff for the same season had a higher on-peak surcharge ($y = 1.55$) and a potential DR savings of 28% for a facility that is loaded at an average 75% of capacity. In the RTP example in Figure 1.2, the top 5% of pricing intervals averages \$149/MWh, and the 95% average price is \$52. Taking \$52 as the base rate, the average normalized surcharge is $y = 1.86$ for $x = 0.05$. A center paying these prices could save 8% even if it is loaded at an average 95% of capacity. Taking the top 2% of pricing intervals as the surcharge period, $y = 2.25$, and the maximum savings is 4.3%.

It is important to note that customers can often lower their average prices by accepting the higher risk of volatility that comes with variable pricing. Thus DR may be viewed as a risk-control measure with an indirect benefit of lowering electricity prices during normal operation, while limiting exposure to the resulting price spikes. For example, the A-6 tariff mentioned above offers a discount of 41% on the base price of electricity (11 hours per day), and a 27% average discount on electricity outside of peak periods. The customer obtains these benefits by accepting the surcharge for expected peak periods. A DR strategy can avoid the surcharges if it can defer electricity usage during these surcharge periods.

1.6.3 Energy Proportionality at Facility Scale

The model illustrates the importance of energy proportionality for DR savings. In essence, energy proportionality captures the degree to which a system can reduce its power draw by shedding load. The idealized model presumes that the system power is linear with instantaneous facility utilization or load factor ρ. Refer to Figure 1.5(c) and the discussion in Section 1.5.

We can quantify energy proportionality at the granularity of servers or other individual components, or at the granularity of ensembles or an entire facility. For example, recent results from SPECpower benchmark indicate that server systems are increasingly energy-proportional, primarily as a result of advances in CPUs and power supplies. Servers with dynamic ranges of 70% of higher are common. However, their power profiles increasingly deviate from the linear model, which tends to underestimate their power draw at CPU utilization levels that are low but non-zero. Also, for data-intensive workloads, the energy costs of memory, storage, and I/O may dominate CPU activity [33], and these costs tend to be less energy-proportional than CPUs.

In server ensembles, further improvements are possible by concentrating load on a minimal subset of servers and stepping down surplus servers to a low-power state (e.g., [7]). This technique can be combined with various approaches to active server scaling at the platform level, such as dynamic voltage scaling. Several commercial products and services offer support for energy-proportional ensembles using these techniques. Recent studies suggest that server ensembles can approach full energy proportionality with active

management [32, 12]. Related techniques have been applied in storage ensembles, with some success (e.g., [35]). There has also been some recent attention to energy-proportional networking for data centers [15].

A large share of power in computer centers and data centers feeds ancillary equipment including cooling and power distribution, rather than servers. One measure of their relative impact is the ratio of total power to power for servers and other IT equipment—the ratio known as Power Usage Effectiveness or PUE. Recent studies have estimated a typical PUE value of 2.0 [19], suggesting that about half of the energy in today's data centers goes to servers. The EPA target for state-of-the-art data centers is a PUE of 1.2 in 2011 [11]. Google reports PUE levels for Google-designed data centers on a quarterly basis, and has succeeded in meeting a PUE of 1.2 in 2010. Active server management pushes PUE up, making efficient power distribution and cooling more important.

In recent years, energy-proportional cooling has received more attention. For example, temperature-aware workload placement helps reduce cooling demands for ensembles running below full capacity [24]. Other "smart cooling" techniques modulate fan speeds, compressor duty cycles, and other mechanical systems. A recent study suggests that combining these techniques with active server management can yield facility-level energy proportionality roughly following the linear model with dynamic ranges of 70% to 80% [32].

Bibliography

[1] Alvin Auyoung, Laura Grit, Janet Wiener, and John Wilkes. Service contracts and aggregate utility functions. In *Proceedings of the IEEE Symposium on High Performance Distributed Computing*, pages 119–131, 2006.

[2] Woongki Baek and Trishul M. Chilimbi. Green: a framework for supporting energy-conscious programming using controlled approximation. In *Proceedings of the 2010 ACM SIGPLAN conference on Programming language design and implementation*, PLDI '10, pages 198–209, New York, NY, USA, 2010. ACM.

[3] Luiz Andre Barroso and Urs Holzle. The case for energy-proportional computing. *Computer*, 40:33–37, 2007.

[4] Severin Borenstein. The trouble with electricity markets: Understanding california's restructuring disaster. *Journal of Economic Perspectives*, 16(1):191–211, 2002.

[5] Steven Brathwait, Dan Hansen, and Michael O'Sheasy. Retail electricity pricing and rate design in evolving markets, July 2007.

[6] Michael Cardosa, Madhukar R. Korupolu, and Aameek Singh. Shares and Utilities based Power Consolidation in Virtualized Server Environments. In *Proceedings of the 11th IFIP/IEEE international conference on Symposium on Integrated Network Management*, June 2009.

[7] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle. Managing Energy and Server Resources in Hosting Centers. In *Proceedings of the 18th ACM Symposium on Operating System Principles (SOSP)*, pages 103–116, October 2001.

[8] Federal Energy Regulatory Commission. State of the Markets Report, August 2009.

[9] Federal Energy Regulatory Commission. National Action Plan on Demand Response (Draft), March 2010.

[10] Azbayer Demberel, Jeffrey Chase, and Shivnath Babu. Reflective Control for an Elastic Cloud Application: An Automated Experiment Workbench.

In *Proceedings of the First Workshop on Hot Topics in Cloud Computing (HotCloud)*, June 2009.

[11] United States Environmental Protection Agency (EPA). Report to Congress on Server and Data Center Energy Efficiency, Public Law 109-431, August 2007.

[12] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power Provisioning for a Warehouse-sized Computer. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*, June 2007.

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

[14] M.M. He, E.M. Reutzel, Xiaofan Jiang, R.H. Katz, S.R. Sanders, D.E. Culler, and K. Lutz. An architecture for local energy generation, distribution, and sharing. In *Energy 2030 Conference, 2008. ENERGY 2008. IEEE*, pages 1 –6, November 2008.

[15] Brandon Heller, Srini Seetharaman, Priya Mahadevan, Yiannis Yiakoumis, Puneet Sharma, Sujata Banerjee, and Nick McKeown. Elastictree: saving energy in data center networks. In *Proceedings of the 7th USENIX Conference on Networked System Design and Implementation*, NSDI'10, pages 17–17, Berkeley, CA, USA, 2010. USENIX Association.

[16] David Irwin, Laura Grit, and Jeff Chase. Balancing risk and reward in a market-based task service. In *Proceedings of the Thirteenth International Symposium on High Performance Distributed Computing (HPDC-13)*, June 2004.

[17] James M. Kaplan, William Forrest, and Noah Kindler. Revolutionizing data center energy efficiency, July 2008.

[18] David Kathan, Caroline Daly, Jignasa Gadani, Diane Gruenke, Eric Icart, Ryan Irwin, Carey Martinez, Kendra Pace, John Rogers, Christina Switzer, Carol White, and Dean Wight. Assessment of Demand Response and Advanced Metering, September 2009.

[19] Jonathan G. Koomey. Worldwide electricity used in data centers. *Environmental Research Letters*, September 2008.

[20] Kien Le, Ozlem Bilgir, Ricardo Bianchini, Margaret Martonosi, and Thu D. Nguyen. Managing the cost, energy consumption, and carbon footprint of internet services. In *Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, SIGMETRICS '10, pages 357–358, New York, NY, USA, 2010. ACM.

[21] M. Litzkow, M. Livny, and M. Mutka. Condor - A Hunter of Idle Workstations. In *Proceedings of the 8th International Conference on Distributed Computing Systems*, pages 104–111, 1988.

[22] Amory Lovins. The negawatt revolution. *Across the Board, The Conference Board Magazine*, 27(9), September 1990.

[23] Amory B. Lovins, E. Kyle Datta, Thomas Feiler, Karl R. Rabago, Joel N. Swisher, Andre Lehmann, and Ken Wicker. *Small is Profitable: The Hidden Economic Benefits of Making Electrical Resources the Right Size*. Rocky Mountain Institute, 2002.

[24] Justin Moore, Jeff Chase, Parthasarathy Ranganathan, and Ratnesh Sharma. Making Scheduling "Cool": Temperature-Aware Workload Placement in Data Centers. In *Proceedings of the 2005 USENIX Annual Technical Conference*, pages 61–74, April 2005.

[25] Ripal Nathuji and Karsten Schwan. VirtualPower: Coordinated Power Management in Virtualized Enterprise Systems. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, October 2007.

[26] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the Electric Bill for Internet-Scale Systems. In *Proceedings of the ACM SIGCOMM Conference*, October 2009.

[27] Partha Ranganathan, P. Leech, David Irwin, and Jeffrey Chase. Ensemble-level power management for dense blade servers. In *33rd International Symposium on Computer Architecture (ISCA)*, June 2006.

[28] Federal Energy Regulatory Commission Staff Report. National Assessment of Demand Response Potential, June 2009.

[29] Navin Sharma, Sean Barker, David Irwin, and Prashant Shenoy. Blink: Supply-side Power Management in Data Centers. In *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Newport Beach, California, March 2011.

[30] Kathleen Spees and Lester B. Lave. Demand Response and Electricity Market Efficiency. *The Electricity Journal*, 20, April 2007.

[31] Christopher Stewart and Kai Shen. Some Joules Are More Precious Than Others: Managing Renewable Energy in the Datacenter. In *Proceedings of the Workshop on Power-Aware Computing and Systems (HotPower)*, October 2009.

[32] Niraj Tolia, Zhikui Wang, Manish Marwah, Cullen Bash, Parthasarathy Ranganathan, and Xiaoyun Zhu. Delivering Energy Proportionality with

Non Energy-Proportional Systems — Optimizing the Ensemble. In *Proceedings of the Workshop on Power-Aware Computing and Systems (HotPower)*, October 2009.

[33] Dimitris Tsirogiannis, Stavros Harizopoulos, and Mehul A. Shah. Analyzing the energy efficiency of a database server. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD '10, pages 231–242, New York, NY, USA, 2010. ACM.

[34] Jon Wellinghoff, David L. Morenoff, James Pederson, and Mary Elizabeth Tighe. Creating Regulatory Structures for Robust Demand Response Participation in Organized Wholesale Electric Markets. In *ACEEE Summer Study on Building Efficiency*, August 2008.

[35] Qingbo Zhu, Zhifeng Chen, Lin Tan, Yuanyuan Zhou, Kimberly Keeton, and John Wilkes. Hibernator: helping disk arrays sleep through the winter. In *Proceedings of the Twentieth ACM Symposium on Operating Systems Principles*, SOSP '05, pages 177–190, New York, NY, USA, 2005. ACM.