52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Aggregate Flexibility of a Collection of Loads[π]

Ashutosh Nayyar[1], Josh Taylor[2], Anand Subramanian[3], Kameshwar Poolla [3,4], and Pravin Varaiya[4]

***Abstract*— We consider a collection of flexible loads. Each load is modeled as requiring energy E on a service interval $[a, d]$ at a maximum rate of m. The collection is serviced by available generation $g(t)$ which must be allocated causally to the various tasks. Our objective is to characterize the *aggregate flexibility* offered by this collection. In the absence of rate limits, we offer necessary and sufficient conditions for the generation $g(t)$ to service the loads under causal scheduling without surplus or deficit. Our results show that the flexibility in the collection can be modeled as electricity storage. The capacity $Q(t)$ and maximum charge/discharge rate $m(t)$ of the equivalent storage can be computed in real time. Ex ante, these parameters must be estimated based on arrival/departure statistics and charging needs. Thus, the collection is equivalent a stochastic time-varying electricity storage. We next consider the case with charging rate limits. Here, we offer bounds on the capacity and rate of the equivalent electricity storage. We offer synthetic examples to illustrate our results.**

I. Introduction

California has set aggressive targets for deep integration of renewable energy, driven by climate change concerns and the promise of green-job creation. Energy production from these renewables is variable – it is not dispatchable, is intermittent, and is uncertain [17]. Variability is the most important obstacle to deep renewable integration.

The current approach to renewable integration is to absorb the attendant variability in operating reserves. This works at today's modest penetration levels. But it will not scale tomorrow, when we have to meet the aggressive renewable targets. Recent studies in California [3], [10] project that the spring time maximum up-regulation capacity needed to accommodate 33% renewable energy penetration will increase from 277 MW to 1,135 MW. Maximum load-following capacity requirements will increase from 2,292 MW to 4,423 MW. These large increases in necessary reserves are economically untenable, and diminish the net carbon benefit of renewables.

It is widely recognized that demand-side distributed energy resources (DERs) must play a key role in supplying zero-emissions regulation services that are necessary to enable deep renewable integration. These resources include thermostatically controlled loads (TCLs), electric vehicles (EVs), and strategic storage. Existing implementations and demonstration work focuses on dispatching DERs as individual deterministic resources, and on infrastructure (standards, sensors, and software platforms) such as open-ADR, s-MAP, to extract flexibility from DERs. However there is a critical gap to be filled if demand side resources are to make significant system-level impacts: we must offer a universal, portable model to describe the aggregated uncertain resource that facilitates its integration (forecasting, scheduling and dispatch) into system operator decisions.

A. Summary of Contributions

In this paper, we consider a collection of flexible loads. Each load is modeled as a task requiring energy E on a service interval $[a, d]$ at a maximum rate of m units of energy per unit time-interval. We first consider the case when the maximum rate m for each task is sufficiently large (that is, larger than the energy need of the task). For this case, we provide a necessary and sufficient condition for a generation profile to be adequate for meeting the energy needs of all tasks without surplus. This characterization allows us to quantify the instantaneous aggregate flexibility offered by the collection of loads in terms of the maximum energy that the collection can absorb or release at a time without creating a surplus or missing any task deadlines. Thus, we can model the collection of flexible loads as an *equivalent stochastic electricity storage*. We also characterize the minimum up and down reserves required for an inadequate generation profile to ensure that all tasks are serviced. For the case with charging rate limits, we provide necessary conditions for a generation profile to be adequate. We offer synthetic examples to illustrate our results.

B. Related Work

Coordinated aggregation of flexible loads is an active research area. There has been considerable work in developing and analyzing scheduling algorithms for controlling deferrable loads in response to renewable generation [5], [11], [18], [19], [20], [23]. Communication and control requirements to manage aggregate loads have been explored in [1], [14]. Coordinated aggregation policies and day-ahead bulk power purchase that maximize social welfare have been presented in [13]. Other demand response approaches

[1] Corresponding author, anayyar@berkeley.edu, Electrical Engineering and Computer Sciences, University of California, Berkeley.
[2]Electrical and Computer Engineering, University of Toronto, Canada.
[3]Mechanical Engineering, University of California, Berkeley.
[4]Electrical Engineering and Computer Sciences, University of California, Berkeley.
[π] Supported in part by EPRI and CERTS under sub-award 09-206; PSERC S-52; NSF under Grants 1135872, EECS-1129061, CPS-1239178, and CNS-1239274; the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore for the SinBerBEST Program; Robert Bosch LLC through its Bosch Energy Research Network funding program.

have been considered in [4], [6], [8], [9], [15], [21], [22]. In contrast to prior work, the focus of this paper is to quantify the instantaneous aggregate flexibility offered by the collection of loads in terms of charge and discharge capacity of a stochastic electricity storage.

The remainder of this paper is organized as follows. Section II contains the problem formulation where we model flexible loads, describe generation acquisition, and introduce scheduling policies. Our main results are found in Sections III and IV, where we characterize the aggregate flexibility of a collection of loads without and with rate limits respectively. Illustrative examples are offered in Section V, and conclusions are drawn in Section VI. All proof are collated in the Appendix.

II. Problem Set-Up

A. Flexible Load Modeling

Consider a collection of flexible loads. For example, one can imagine a shopping mall where electric vehicles with various energy needs arrive and depart. A *load manager* is responsible for economically meeting the charging needs of the vehicles. Time is discrete with unit increments. Suppose u_t is the energy supplied to a load over the interval $[t, t+1)$. We will abuse notation and speak of u_t as the *power* supplied to the load at time t, in which case the sequence $u = (\cdots, u_t, u_{t+1}, \cdots)$ can be regarded as a power profile.

Definition 1: We regard each flexible load as a *task*. A task is parameterized by its total energy need E, arrival time a, departure time d, and maximum charge rate m. It is denoted by the quadruple (E, a, d, m). A task (E, a, d, m) can be serviced at times $t \in [a, d)$. If u_t is the power supplied to this task at time t, i.e. the energy supplied over the interval $[t, t+1)$, we can write

$$\sum_{a}^{d-1} u_t = E, \quad 0 \leq u_t \leq m. \tag{1}$$

Clearly, we require that the energy need of the task be less than the maximum energy it can receive, i.e. $E \leq m(d-a)$. We allow tasks to be *pre-emptive*, that is they can be serviced with interruptions. Scheduling problems for non-pre-emptive tasks reduce to bin packing problems which are NP hard [2].

Tasks are energy consumers. There are many power profiles that satisfy the constraints (1) and will service a task. Thus tasks have *flexibility* in the power profiles they can accept. The central objective of this paper is to characterize the power flexibility offered by a collection of tasks.

Definition 2: The *energy state* of the task (E, a, d, m) at time t is

$$e_t = E - \sum_{a}^{t-1} u_k,$$

where u_k is the power supplied at time k.

We will consider a collection of tasks

$$\mathbb{T} = \{(E^i, a^i, d^i, m^i) : i = 1 \cdots N\}$$

on a time horizon $H = [1, T)$, so all tasks have departed prior to time T. Subscripts index time and superscripts index tasks.

Definition 3: The set of *active tasks* at time t is

$$\mathbb{A}_t = \{i \ : \ a^i \leq t < d^i\}.$$

The *nominal aggregate load profile* at time t is

$$n_t = \sum_{i \in \mathbb{A}_t} \frac{E^i}{d^i - a^i}.$$

This is the power profile that results if each task is serviced at a *constant rate*. We can regard this nominal power profile as one that does not exploit load flexibility.

The arrivals, departures, and energy needs of tasks are random. Thus, we must regard n as a random process. We assume that when a task arrives, it announces its parameters (E, a, d, m) to the load manager.

B. Generation Modeling

The available generation g to service active tasks is drawn from (a) random renewable generation w such as distributed roof-top solar PVs, (b) bulk power b purchased in traditional forward markets, and (c) reserve generation r. We can write

$$g = w + b + r \tag{2}$$

We assume the renewable generation has zero marginal cost. The reserve generation can be positive or negative. It is scheduled in real-time and has two distinct interpretations:

(a) r *is bought.* Here, r is regarded as a regulation service procured by the load manager for real-time balancing of supply and demand. This could come from locally owned gas turbines and/or storage, or it could be purchased as an ancillary service from the grid. Here, load flexibility is used to minimize the cost of load-balancing.

(b) r *is sold.* Here, r is regarded as a regulation ancillary service sold by the load manager to the grid. Load flexibility is exploited to offer this service. We will argue that regulation services that flexible loads offer are fundamentally different than those supplied by traditional generator resources. Flexible loads cannot supply sustained biases in regulation power. These loads act essentially as storage.

Since w is random, g is a stochastic process. We assume g is *sized to meet the nominal aggregate load*, i.e.

$$\mathbb{E}g_t = \mathbb{E}n_t \tag{3}$$

where $\mathbb{E}$ denotes the expectation operator. This assumption is justified as follows. The best forecast of the procured (case (a)) or offered (case (b)) reserves on 5-minute blocks is $\mathbb{E}r = 0$. Bulk power is scheduled *ex ante* to supplement the expected renewable generation, i.e. $\mathbb{E}b = \mathbb{E}n - \mathbb{E}w$.

C. Scheduling Policies

The load manager allocates the available generation at time t to the active tasks using some scheduling policy. To enhance readability, we offer an informal definition:

Definition 4: A *scheduling policy* σ is an algorithm that allocates the available generation g to active tasks:

$$\sigma(g) = (u^1, \cdots, u^N)$$

such that $u_t^i = 0$ if $i \notin \mathbb{A}_t$, and

$$\sum_i u_t^i \leq g_t, \quad 0 \leq u_t^i \leq m^i, \quad u_t^i \leq \min(m^i, e_t^i)$$

Here, u_t^i is the power allocated to task i at time t.

Definition 5: The information state $\mathcal{I}_t$ at time t consists of

(a) task parameters $\left(E^i, a^i, d^i, m^i\right)$ for all current and past tasks, i.e. tasks with $a^i \leq t$,

(b) energy states e_t^i of for all current and past tasks,

(c) realized values $\{g_\tau : \tau \leq t\}$ of the available generation.

We distinguish between two classes of scheduling policies. A *causal* scheduling policy allocates power at time t to active tasks using only the available information $\mathcal{I}_t$. A *non-causal* policy makes allocations u_t^i using oracle information – all past and future values of generation, and parameters of all past and future tasks. Of course, the load manager must implement a causal scheduling policy.

One of the simplest causal policies is Earliest Deadline First (EDF) which allocates available generation to the tasks in order of their deadlines. There are many other scheduling policies developed in the context of processor time allocation [2], [16] and other resource allocation problems.

When available generation g is allocated using a policy σ, we may have a surplus (power that could not be allocated to any task), and/or may suffer a shortfall (some tasks do not receive their energy demand, i.e. are not completed) at various times. In the event we have a surplus of power, we will require *down regulation*, and if we have a shortfall of power, we will need *up regulation*. Up/down regulation will be drawn from reserve generation r in real-time.

We now define two notions of generation adequacy.

Definition 6: The available generation g is called *exactly adequate* if there exists a *non-causal* scheduling policy that completes all the tasks without surplus. The available generation g is called *causally exactly adequate* if there exists a *causal* scheduling policy that completes all the tasks without surplus.

III. Without Rate Limits

We first characterize exact adequacy of generation for a collection of loads without rate limits. Using this characterization, we argue that the flexibility in the collection can be modeled as stochastic electricity storage. Finally, we show that this characterization allows us to determine minimum energy reserves (up and down) needed to complete all tasks.

We consider a collection of tasks $\mathbb{T} = \{(E^i, a^i, d^i, m^i) : i = 1 \cdots N\}$. In this section, we assume $m^i = \infty$, i.e. *the tasks do not have service rate limits*.

A. Characterizing Exact Adequacy

Theorem 1: Define the cumulative available energy as

$$G_t = \sum_{k=1}^{t} g_k.$$

Then, g is exactly adequate if and only if

$$\sum_{i:d^i \leq t+1} E^i \leq G_t \leq \sum_{i:a^i \leq t} E^i \quad \text{for all } t. \tag{4}$$

In the event g is exactly adequate, all tasks can be completed using the EDF scheduling policy. As this policy is causal, it follows that exact adequacy is equivalent to causal exact adequacy in the absence of rate limits.

Corollary 2: Suppose storage of energy capacity Q without charge/discharge rate limits is also available and has initial state of charge Q_1. Then g is exactly adequate (causally or non-causally) if and only if for all t we have

$$-Q_1 + \sum_{i:d^i \leq t+1} E^i \leq G_t \leq Q - Q_1 + \sum_{i:a^i \leq t} E^i \tag{5}$$

B. Selling Aggregate Flexibility as a Regulation Service

We first offer an alternate characterization of exact adequacy.

Theorem 3: Define the instantaneous and cumulative deviations between supply and nominal demand to be

$$v_t = g_t - n_t, \quad V_t = \sum_1^t v_k.$$

Then, g is exactly adequate if and only if

$$-x_t \leq V_t \leq y_t, \quad \text{for all } t \tag{6}$$

where

$$x_t = \sum_{i:a^i \leq t, d^i > t+1} E^i \frac{(t+1-a^i)}{(d^i - a^i)} \quad \text{and}$$

$$y_t = \sum_{i:a^i \leq t, d^i > t+1} E^i \frac{(d^i - t - 1)}{(d^i - a^i)}. \tag{7}$$

This result states that a generation profile is exactly adequate as long as the cumulative deviation process V_t is confined to the interval $[-x_t, y_t]$. It is important to note that x_t and y_t can be calculated causally, i.e. using only the information state $\mathcal{I}_t$. *We submit that this interval captures the aggregate flexibility offered by the collection of loads.* On the interval $[t, t+1)$, the collections of loads can forgo $(V_t + x_t)^+$ units

of energy while still ensuring exact adequacy. Similarly, the loads can collectively absorb $(y_t - V_t)^+$ units of energy from time t to $t+1$. We associate energy on a unit interval $[t, t+1)$ with power at time t. The metrics $(V_t + x_t)^+$ and $(y_t - V_t)^+$ therefore serve as the real-time up and down reserve power margins at time t offered by the collection of flexible loads.

Suppose the available generation is adequate. Theorem 3 implies that load flexibility offers a regulation ancillary service that can be sold by the load manager to the grid. This service is fundamentally different than regulation supplied by traditional generator resources. First, randomness in load arrivals, departures, and energy needs imply that the available regulation service is stochastic. Second, flexible loads cannot supply *sustained biases* in regulation power. The regulation service is therefore much more like electricity storage than traditional generation.

In order to participate in ancillary service markets, the load manager must *ex ante* declare the level of service that the flexible loads can supply at delivery time. This can be modeled as an equivalent *stochastic electricity storage* with V_t being the state of the storage, $-x_t$ the discharge capacity, and y_t the charge capacity. Note that x, y are random processes. We can bound the capacities of the equivalent stochastic storage *ex ante* in terms of their expected values. We define

$$\begin{aligned} \Phi_1 &= \mathbb{E}[(V_t + x_t)^+] \geq (\mathbb{E}[V_t] + \mathbb{E}[x_t])^+ \\ \Phi_2 &= \mathbb{E}[(y_t - V_t)^+] \geq (\mathbb{E}[y_t] - \mathbb{E}[V_t])^+. \end{aligned}$$

If the generation is sized to meet the nominal load, that is, $E[g_t] = E[n_t]$, then $E[V_t] = 0$. Therefore, the long-term average capacities offered are

$$\Phi_1 \geq \mathbb{E}[x_t], \quad \Phi_2 \geq \mathbb{E}[y_t].$$

C. Examples

1) Deterministic Arrivals: Consider a collection of identical loads with deterministic, uniform arrival rate $1/\lambda$, and constant service window $d^i - a^i = N$ and no (effective) rate limits. The i^{th} task is parameterized by $(E^i = E, a^i = i\lambda, d^i = i\lambda + N, m^i = \infty)$. Under this model, at any time (beyond an initial transient period), we have

$$x_t \approx \frac{E(N-1)}{2\lambda}, \quad y_t \approx \frac{E(N-1)}{2\lambda}.$$

The constant $\frac{E(N-1)}{2\lambda}$ is the expected reserve capacity that the resource manager can offer in the forward reserve market.

2) Bernoulli Arrivals: Consider the above example with the arrival process being Bernoulli, i.e. at each time a new task arrives with probability $1/\lambda$. A task arriving at a^i has the parameters $(E, a^i, d^i = a^i + N, m^i = \infty)$. Then the expected number of active tasks at any time is N/λ and

$$\mathbb{E}[x_t] \approx \frac{E(N-1)}{2\lambda}, \quad \mathbb{E}[y_t] \approx \frac{E(N-1)}{2\lambda} \tag{8}$$

3) EVs Parking Garage: Consider a garage that supports EV charging. Suppose that, on average, 50 EVs arrive per hour, park for h hrs, and are charged at the nominal rate of 3 KW. Assume the arrival process is Bernoulli. Using (8), we calculate that the aggregate flexibility offered by these EVs can be treated as stochastic storage with capacity $\pm 75h^2$ KWh.

D. Buying Minimum Energy Reserves

We now consider a different use case. Suppose the available generation is not exactly adequate. The load manager must at various times purchase up-regulation (when there is a shortfall), and down-regulation when there is a surplus. These must be procured in real-time. Load flexibility offers a degree a freedom in procurement of up/down regulation services. We are interested in the minimum energy reserves necessary to ensure that all loads are serviced.

Let n denote the nominal aggregate power profile. If we do not exploit load flexibility, the surplus generation at time t is simply $(g_t - n_t)^+$. Similarly, the shortfall in generation is $(n_t - g_t)^+$. Load flexibility allows for reduced surplus/shortfall, allowing the load manager to procure reserves at lower cost. Theorem 3 allows us to characterize the minimum energy reserve process that is necessary to service the loads. For simplicity, we consider the situation where $x_t = y_t = c$. From Theorem 3, if $-c \leq V_t \leq c$, there is no surplus/shortfall. The minimum energy reserve process is the smallest modification of g needed to confine the cumulative deviation V in the sleeve $[-c, c]$. More precisely, we have

Theorem 4: Define the new process $\tilde{V}_t$ with $\tilde{V}_0 = 0$ and

$$\tilde{V}_{t+1} = \begin{cases} c & \text{if } \tilde{V}_t + v_t > c \\ \tilde{V}_t + v_t & \text{if } -c \leq \tilde{V}_t + v_t \leq c \\ -c & \text{if } -c > \tilde{V}_t + v_t, \end{cases} \tag{9}$$

The minimum energy up and down reserve process needed to service all loads is

$$\begin{aligned} \eta_t^{\text{down}} &= (\tilde{V}_t + v_t - c)^+ \quad \text{(down reserves)} \\ \eta_t^{\text{up}} &= (-c - \tilde{V}_t - v_t)^+ \quad \text{(up reserves)} \end{aligned}$$

We note that these required reserves can be computed causally in real-time using only the information state $\mathcal{I}_t$. While we have determined the minimum *energy* reserves, it is straightforward to compute the minimum *cost* reserves if we are given energy prices for up/down regulation.

IV. With rate constraints

We now consider the case where tasks have rate limits. We offer necessary conditions for exact adequacy and illustrative examples.

A. Characterizing Exact Adequacy

Theorem 5: If g is exactly adequate (causally or non-causally), then

$$\sum_{i:d^i \le T} E^i = G_T \tag{10}$$
$$L_t \le G_t \le U_t \quad \text{for all } t < T,$$

where

$$\begin{aligned} L_t &= \sum_{i:d^i \le t+1} E^i + \sum_{i:a^i \le t, d^i > t+1} (E^i - (d^i - t - 1)m^i)^+ \\ U_t &= \sum_{i:a^i \le t} E^i - \sum_{i:a^i \le t, d^i > t+1} (E^i - (t + 1 - a^i)m^i)^+. \end{aligned} \tag{11}$$

Comparing Theorems 1 and 5, it is evident that the rate limits tighten the lower and upper bounds on cumulative generation by

$$\sum_{i:a^i \le t, d^i > t+1} \max\{0, E^i - (d^i - t - 1)m^i\}$$

and

$$\sum_{i:a^i \le t, d^i > t+1} \max\{0, E^i - (t + 1 - a^i)m^i\}$$

respectively. It should be noted that while Theorem 1 provides necessary and sufficient conditions for exact adequacy, the conditions of Theorem 5 are only necessary.

Corollary 6: Define the instantaneous and cumulative mismatches between supply and nominal demand to be

$$v_t = g_t - n_t, \quad V_t = \sum_1^t v_k.$$

If g is exactly adequate, then $V_T = 0$ and

$$-x_t^* \le V_t \le y_t^*, \quad \text{for all } t \le T \tag{12}$$

where

$$\begin{aligned} x_t^* = &\sum_{i:a^i \le t, d^i > t+1} E^i \frac{(t + 1 - a^i)}{(d^i - a^i)} \\ &- \sum_{i:a^i \le t, d^i > t+1} \max\{0, E^i - (d^i - t - 1)m^i\} \end{aligned} \tag{13}$$

and

$$\begin{aligned} y_t^* = &\sum_{i:a^i \le t, d^i > t+1} E^i \frac{(d^i - t - 1)}{(d^i - a^i)} \\ &- \sum_{i:a^i \le t, d^i > t+1} \max\{0, E^i - (t + 1 - a^i)m^i\} \end{aligned} \tag{14}$$

B. Examples

1) Deterministic arrivals: Consider a collection of identical loads with deterministic, uniform arrival rate $1/\lambda$, deferrability window $d^i - a^i = N$ The ith task is characterized by $(E^i = E, a^i = i\lambda, d^i = i\lambda + N, m^i)$. Let $m^i = E/k$, $1 \le k \le N$. Then,

$$x_t \approx y_t \approx \frac{E(N - k)}{2\lambda}.$$

2) Bernoulli Arrivals: Consider the case where the arrival process is Bernoulli, i.e. at each time a new task arrives with probability $1/\lambda$. A task arriving at a^i has the parameters $(E, a^i, d^i = a^i + N, m^i)$. Let $m^i = E/k$, $1 \le k \le N$. Then,

$$\mathbb{E}[x_t] \approx \mathbb{E}[y_t] \approx \frac{E(N - k)}{2\lambda}.$$

V. Empirical Analysis

We now use synthetic examples to illustrate our results. We are primarily interested in exploring the effect of rate limits. Through synthetic test cases, we explore (a) the impact of the rate limits on load-balancing requirements, i.e. the sufficiency gap in Theorem 5, and (b) the performance of our minimum energy reserve procurement in the case with rate limits.

A. Adequacy with rate limits

In the rate limited case, satisfying the conditions of Theorem 5 do not assure generation adequacy. In this study, we explore how inadequate profiles satisfying these conditions are. Specifically, we calculate the total unsatisfied energy requirement when servicing a collection of tasks with a generation profile that satisfies (13) and (14).

We perform these calculations on test cases where 2 identical tasks arrive every hour for 100 hrs. Each task has an energy need of $E = 10$ kWh and over a constant service interval $d - a = 4$ hrs. We randomly generated power profiles to service these tasks. Each power profile is supplemented by additional up/down reserves so that the available generation g satisfies (13) and (14). The available generation is then allocated to tasks using the least-laxity first (LLF) scheduling algorithm [2], [16]. All results are averaged over 100 test cases.

In Figure 1, the unmet energy demand of the collection of tasks (as a percentage of the total energy demand) is plotted as the maximum service rate is varied. We observe that for generation profiles that satisfy the conditions of Corollary 6, a very small fraction ($< 1\%$) of the energy demand is unmet.

A detailed study of these test cases shows that:

(a) The maximum percentage of unmet tasks over all test cases was 20%. Even here, at least 85% of each tasks' energy requirements were met.
(b) The maximum unmet energy requirement over all test cases for any task was 50%. For this generation profile, 97% of all tasks were completed.

While not conclusive, these results suggest that statistical multiplexing mutes the problem of rate constraints, and that the exact adequacy characterization of Theorem 5 is, in practice, almost necessary and sufficient.

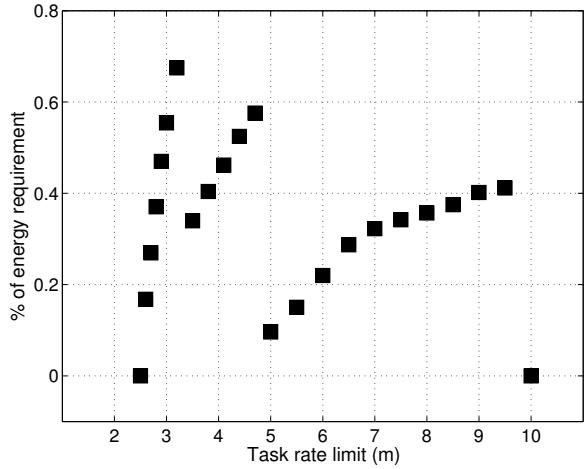

Fig. 1: Percentage of total task energy requirement unsatisfied after procuring reserves based on deferrability limits as a function of task rate limit.

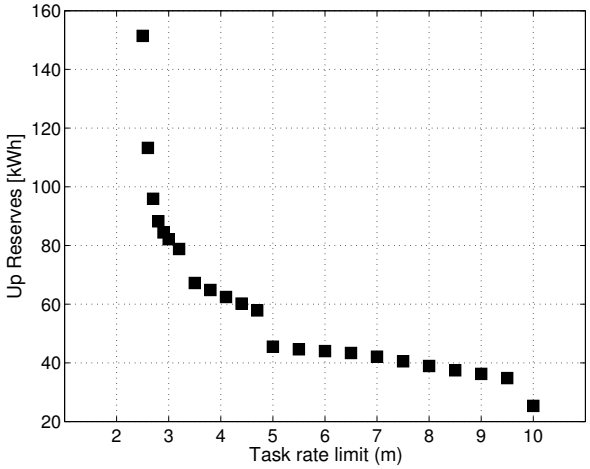

Fig. 2: Up reserves required to meet load requirements as a function of task rate limit.

B. Impact of rate limits on reserve requirements

We now illustrate the ability of these metrics to quantify the benefit afforded by increased deferrability on reserve requirements. Using the same test case parameters described in Section V-A, we quantify the impact of changing the task rate limit on the amount of up (to cover deficits) and down (to cover surpluses) reserves required for load-balancing. In each test case, we compute the up and down reserves according to deviations of the cumulative mismatch process beyond the described bounds. We also include any unsatisfied task requirements in up reserve calculations.

Figures 2 and 3 show the average amounts of up and down reserves required to complete the tasks. The rate limit ranges from 2.5 (no flexibility) to 10 (full flexibility). These figures demonstrate the benefit of load flexibility on reserve requirements. As rate limit increase, we see dramatic reductions in both up and down reserve requirements. These reductions are most significant at lower rate limits indicating that modest load flexibility is sufficient to realize the majority of load-balancing cost reductions.

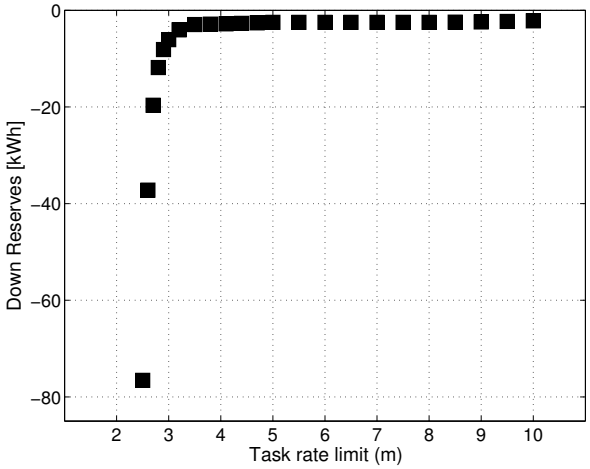

Fig. 3: Down reserves required to meet load requirements as a function of task rate limit.

VI. Conclusions

In this paper we addressed the problem of quantifying the aggregate flexibility offered by a collection of flexible loads. We modeled each load as a task requiring energy E on a service interval $[a, d]$ at a maximum rate of m. For the case without rate limits, we provided a necessary and sufficient condition for a generation profile to be adequate for meeting the energy needs of all tasks without surplus. This characterization allowed us to quantify the instantaneous aggregate flexibility offered by the collection of loads in terms of the maximum energy that the collection can absorb or release at a time without creating a surplus or missing any task deadlines. We also characterized the minimum up and down reserves required with an inadequate generation profile. For the case with charging rate limits, we provided necessary conditions for a generation profile to be adequate.

We view this work as an initial step towards the goal of finding a simple, universal stochastic storage model for aggregate demand flexibility. Such a model would allow the system operator to incorporate demand flexibility in its scheduling and dispatch decisions without requiring detailed information about the needs and constraints of individual loads. We believe such models are necessary if demand side resources are to make significant system-level impacts.

APPENDIX

A. Proof of Theorem 1

Suppose g is exactly adequate. Therefore there exists some scheduling policy σ that completes all tasks without surplus. Fix t. The cumulative generative G_t must exceed the total energy needs of all departed tasks, i.e.

$$G_t \geq \sum_{i:d^i \leq t+1} E^i$$

Also, G_t cannot exceed the total energy needs of all active and past tasks (else we would incur a surplus). Thus

$$G_t \leq \sum_{i:a^i \leq t} E^i$$

establishing (4).

Now suppose (4) holds at all times t. We show that earliest deadline first (EDF) scheduling complete all the tasks. With EDF, we allocate energy to an active task i only if the energy needs of all active tasks with departure time less than d^i have been met. We proceed inductively.

(a) *Base case:* At time $t = 1$, EDF will satisfy all tasks with departure time 2 and use any remaining energy to satisfy other active tasks. (4) for $t = 1$ ensures that there is no surplus or deficit at time $t = 1$.

(b) *Inductive step:* Assume there is no surplus/deficit up to time $t = n$. We show that there is no surplus/deficit by time $t = n + 1$. The amount of energy demand that must be satisfied by the cumulative generation until time $t = n + 1$ is $\sum_{i:d^i \leq (n+1+1)} E^i$. Of this total amount, the amount that must be supplied in the interval $[n, n + 1)$ is $\max(\sum_{i:d^i \leq (n+2)} E^i - G_n, 0)$. To ensure no deficit between times $t = n$ and $t = n + 1$, we need

$$g_{n+1} \geq \max\left(\sum_{i:d^i \leq (n+2)} E^i - G_n, 0\right)$$

which is true if and only if

$$G_{n+1} \geq \sum_{i:d^i \leq (n+2)} E^i \tag{15}$$

(15) is true because (4) is assumed to hold for all t.

The total energy demand of active tasks at time $n + 1$ is $\sum_{i:d^i > n+1, a^i \leq n+1} E^i$. Some of this energy demand was satisfied by the generation till time n. Since, there is no surplus or shortfall till time n, the amount of energy demand of tasks active at time $n+1$ that was satisfied by past generation is $(G_n - \sum_{i:d^i \leq n+1} E^i)$. Therefore, the *unsatisfied* energy demand of tasks active at $n + 1$ is

$$\sum_{i:d^i > n+1, a^i \leq n+1} E^i - (G_n - \sum_{i:d^i \leq n+1} E^i).$$

To ensure that there is no surplus at $t = n+1$, g_{n+1} must be less than the unsatisfied energy demand of active tasks at $n + 1$. That is,

$$g_{n+1} \leq \sum_{i:d^i > n+1, a^i \leq n+1} E^i - (G_n - \sum_{i:d^i \leq n+1} E^i)$$
$$\Leftrightarrow G_{n+1} \leq \sum_{i:a^i \leq n+1} E^i,$$

which we know is true because (4) is assumed to hold for all t. Thus, there is no surplus/deficit by time $t = n + 1$.

This completes the induction, establishing that g is exactly adequate.

B. Proof of Corollary 2

We encode energy storage as an energy surplus Q_1 added to energy balance at $t = 1$ paired with a vehicle s with $E^s = Q$, $a^s = 1$, and $d^s = \infty$. Substituting these values into Theorem 1, we obtain

$$\sum_{i:d^i \leq t+1} E^i \leq G_t + Q_1 \leq Q + \sum_{i:a^i \leq t} E^i \quad \text{for all } t, \tag{16}$$

which establishes the claim.

C. Proof of Theorem 3

By definition,

$$V_t = G_t - \sum_{k=1}^{t} n_k \tag{17}$$

Observe that

$$\sum_{k=1}^{t} n_k = \sum_{k=1}^{t} \left(\sum_{i \in \mathbb{A}_k} \frac{E^i}{(d^i - a^i)} \right)$$
$$= \sum_{i:d^i \leq t+1} E^i + \sum_{i:a^i \leq t, d^i > t+1} E^i \frac{(t + 1 - a^i)}{(d^i - a^i)} \tag{18}$$
$$= \sum_{i:a^i \leq t} E^i - \sum_{i:a^i \leq t, d^i > t+1} E^i \frac{(d^i - t - 1)}{(d^i - a^i)} \tag{19}$$

Simple substitution of (17) - (19) into the statement of Theorem 1 yields the result.

D. Proof of Theorem 4

Note that as long as $-c \leq V_t \leq c$, then $\{\tilde{V}_t\}$ and $\{V_t\}$ are the same process and there is no surplus or deficit. If the first surplus/deficit occurs at $t + 1$, the amount of energy to be discarded is $(V_t + v_t - c)^+ = (\tilde{V}_t + v_t - c)^+$. Further, $\{\tilde{V}_{t+1}\}$ is the "corrected" value obtained from $\{V_{t+1}\}$ by subtracting the amount of discarded energy. Similar arguments hold for deficits.

E. Proof of Theorem 5

Suppose g is exactly adequate generation with rate constraints. Then, g must also be exactly adequate without rate constraints. Therefore, from Theorem 1, we have

$$\sum_{i:d^i \leq t+1} E^i \leq G_t \leq \sum_{i:a^i \leq t} E^i \quad \text{for all } t \leq T. \tag{20}$$

Any task active at time t with a departure time $d^i > t+1$ can receive at most $(d^i - t - 1)m^i$ units of energy on $[t+1, d^i)$. As a result, the cumulative generation until time t must supply at least $\max\{0, E^i - (d^i - t - 1)m^i\}$ to any task active at time t. Thus,

$$\sum_{i:d^i \leq t+1} E^i + \sum_{i \in \mathbb{A}_t} \max\{0, E^i - (d^i - t - 1)m^i\} \quad \leq G_t.$$

Also, any task active at time t with arrival time $a^i \leq t$ could have received at most

$$\begin{aligned} &\min\{E^i, (t + 1 - a^i)m^i\} \\ &= E^i - \max\{0, E^i - (t + 1 - a^i)m^i\} \end{aligned} \tag{21}$$

units of energy on $[a^i, t + 1)$. Thus,

$$G_t \leq \sum_{i:a^i \leq t} E^i - \sum_{i \in \mathbb{A}_t} \max\{0, E^i - (t + 1 - a^i)m^i\}.$$

REFERENCES

[1] M. Alizadeh, A. Scaglione, R.J. Thomas, and D. Callaway, "Information infrastructure for cellular load management in green power delivery systems," *IEEE SmartGridComm Conf.*, 2011.

[2] S. Baruah, and J. Goossens, "Scheduling real-time tasks: Algorithms and Complexity," in *Handbook of Scheduling: Algorithms, Models and Performance Analysis,* J.Y-T. Leung, Ed. Boca Raton, FL: CRC Press, 2004, chpt. 28.

[3] California Independent System Operator (CAISO), "Integration of renewable resources: Operational requirements and Generation fleet capability at 20% RPS," Technical Report, August 2010.

[4] D.S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conversion and Management*, 50(5): 1389-1400, May 2009.

[5] S. Chen, T. He, and L. Tong, "Optimal deadline scheduling with commitment", *Allerton Conf. on Commun., Control, and Computing*, 2011.

[6] J. L. Mathieu, M. Kamgarpour, J. Lygeros, and D. S. Callaway, "Energy Arbitrage with Thermostatically Controlled Loads", *European Control Conference*, 2013.

[7] M.L. Dertouzos, and A.K. Mok, "Multiprocessor online scheduling of hard-real-time tasks," *IEEE Trans. Softw. Eng.*, 15(12): 1497-1506, December 1989.

[8] D.J. Hammerstrom, J. Brous, D.P. Chassin, et al., "Pacific Northwest GridWise testbed demonstration projects," Tech. Rep. PNNL-17167, Pacific Northwest National Laboratory, October 2007.

[9] G. Heffner, C. Goldman, B. Kirby, and M. Kintner-Meyer, "Loads providing ancillary services: Review of international experience," Technical Report LBNL-62701, Lawrence Berkeley National Laboratory, May 2007.

[10] U. Helman, "Resource and transmission planning to achieve a 33% RPS in California - ISO Modeling tools and planning framework," *FERC Technical Conference on Planning Models and Software*, 2010.

[11] G. Hug-Glanzmann, "Coordination of intermittent generation with storage, demand control and conventional energy sources", *IREP 2010 Bulk Power System Dynamics and Control - VIII*, 2010.

[12] M. Ilic, L. Xie, and J.Y. Joo, "Efficient coordination of wind power and price-responsive demand", *IEEE Trans. Power Syst.*, 26(4): 1885-1893, November 2011.

[13] L. Jiang, and S. Low, "Multi-period optimal procurement and demand responses in the presence of uncertain supply," *50th Conf. on Decision and Control*, December 2011.

[14] J.-Y. Joo and M. D. Ilic, "A multi-layered adaptive load management (ALM) system: Information exchange between market participants for efficient and reliable energy use," *IEEE Power and Energy Society Transmission and Distribution Conf. and Exposition*, New Orleans, 2010.

[15] T.F. Lee, M.Y. Cho, Y.C. Hsiao, P.J. Chao, and F.M. Fang, "Optimization and implementation of a load control scheduler using relaxed dynamic programming for large air conditioner loads," *IEEE Trans. Power Syst.*, 23(2): 691-702, May 2008.

[16] C.L. Liu, and J.W. Layland, "Scheduling algorithms for multiprogamming in a hard-real-time environment," *Journal of the ACM*, 20(1): 46-61, January 1973.

[17] North American Electric Reliability Corporation (NERC),"Accommodating high levels of variable generation," White Paper, April 2009.

[18] A. Papavasiliou, and S. Oren, "Supplying renewable energy to deferrable loads: Algorithms and economic analysis," *IEEE Power & Energy Society General Meeting*, 2010.

[19] M. Roozbehani, M. Ohannessian, D. Materassi, and M. A. Dahleh, "Load-Shifting under Perfect and Partial Information: Models, Robust Policies, and Economic Value," *Operations Research*, 2012 (Submitted).

[20] A. Subramanian et al., "Real-Time Scheduling of Deferrable Electric Loads," *Proceedings of the Amer. Control Conf.*, Montreal, 2012.

[21] J.A. Short, D.G. Infield, and L.L. Freris, "Stabilization of grid frequency through dynamic demand control," *IEEE Trans. Power Syst.*, 22(3): 1284-1293, August 2007.

[22] K. Spees and L.B. Lave, "Demand response and electricity market efficiency," *The Electricity Journal*, 20(3): 69-85, April 2007.

[23] K. Trangbaek, M. Petersen, J. D. Bendtsen, and J. Stoustrup, "Exact power constraints in smart grid control," IEEE Conference on Decision and Control and European Control Conference, Orlando, Florida, December 2011.