Energy

http://www.aimspress.com/journal/energy

AIMS Energy, 4(2): 300–312.
DOI: 10.3934/energy.2016.2.300
Received: 31 December 2015
Accepted: 25 February 2016
Published: 04 March 2016

Research article

Adaptive multi-tiered resource allocation policy for microgrids

Konstantinos Christidis *and **Michael Devetsikiotis**

Department of ECE, North Carolina State University, 890 Oval Dr, Raleigh, NC 27606, USA

* **Correspondence:** Email: kchrist@ncsu.edu; Tel: +1-919-513-7324.

Abstract: We consider a cluster of buildings within proximity that share a large-capacity battery for peak-shaving purposes, and draw power from the grid at a premium once they reach a certain threshold. Our goal is to identify a resource allocation policy that minimizes the amount of energy the cluster draws at a premium, while also ensuring fair access to all of its members. We introduce an adaptive policy that allows for maximum energy savings when the network load is low, and ensures fairness when the aggregate power level is high. We compare this adaptive policy with two standard resource allocation strategies with complementary advantages, and demonstrate through an extensive performance evaluation, that it combines the benefits of both. It is therefore suitable for a microgrid operator where equal weight is given to both cluster-wide cost minimization and fairness among all customers.

Keywords: microgrids; battery storage; peak shaving; resource pooling; resource allocation

1. Introduction

Utility tariffs that employ time-of-use pricing schemes or demand charges are becoming the norm for high-demand customers, in an effort to keep their demand under control ([1, 2, 3, 4]). On the customer side, an often used solution is to deploy a large capacity battery, charge it during off-peak hours when the electricity is cheap, and have it absorb part of the building's needs during the on-peak period of the day. Most of the time however, it is not economical for a single customer to deploy such a solution on their own, due to the high battery cost that outweighs the incurred energy cost savings.

Therefore, several buildings within proximity of each other may partner and collectively buy a single battery to serve the needs of the whole cluster. An energy consumption controller deployed in the cluster, along with smart meters with bidirectional connectivity in each building are used to ensure that the battery is (i) used optimally for the cluster, and (ii) shared fairly among its members.

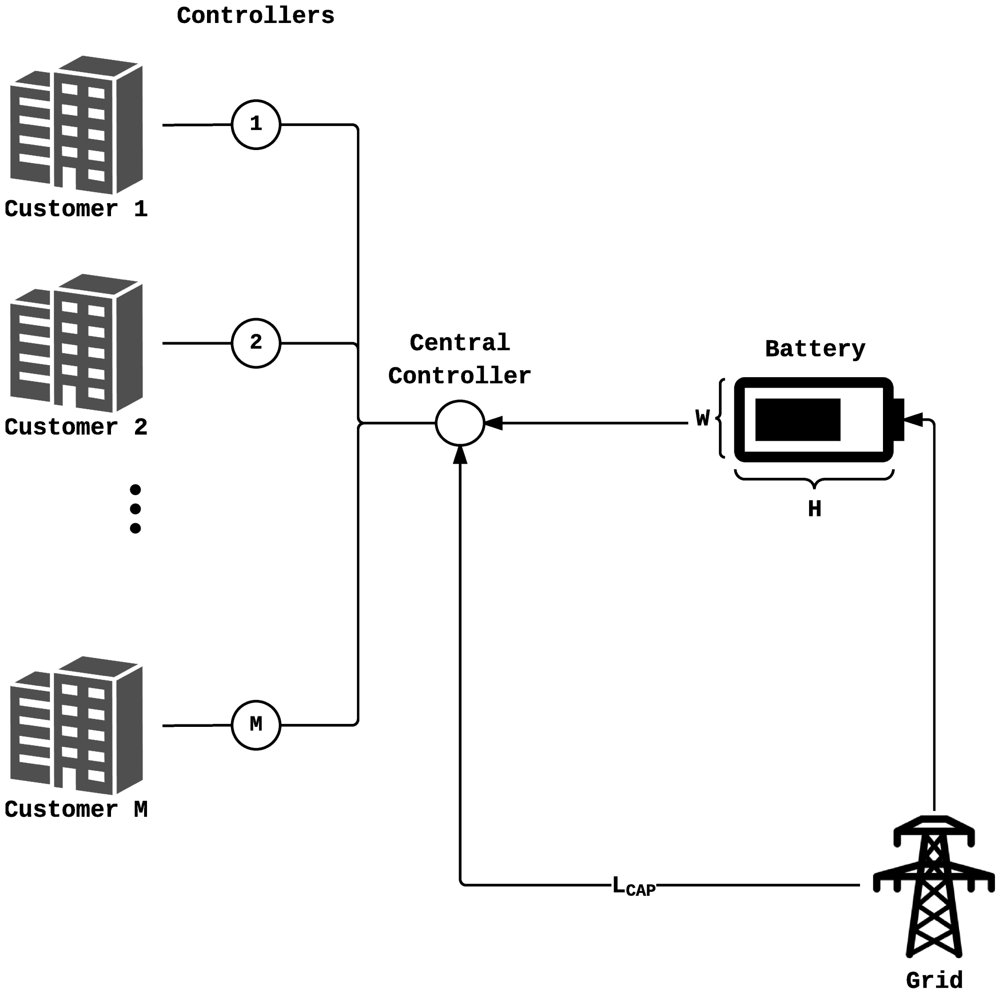

Figure 1. System diagram.

We consider a group of customers that share a battery for peak-shaving purposes, and also sign a group contract with the utility where they effectively get a lower energy unit price in exchange for guaranteeing their power consumption won't exceed a certain threshold. The lower the threshold (the tighter the constraint), the lower the contract price. If the load exceeds the contract rate, the cluster is paying a premium for every unit above the threshold. Our goal is to minimize the amount of energy drawn at a premium.

When it comes to the utility, thanks to the group contract signed above, the cluster is treated as a single, atomic entity for billing purposes. However, when it comes to the group, each participant is responsible for their own usage and has to pay for it. This necessitates the need to monitor the intra-group demand on a per building/stakeholder level, and motivates us to examine the issue of *fair resource sharing*. We group the cluster customers into high- and low-demand tiers based on a stochastic characterization of their demand profiles. Our second, parallel goal then is to ensure that the delta between the overall unit prices that each customer is paying is as small as possible.

In order to achieve these goals, we examine two standard resource allocation policies with complementary benefits. We then introduce our own hybrid adaptive policy which allows us to find a sweet spot between the two earlier policies. We compare all three and perform sensitivity analyses to identify trends.

We begin with an overview of related and motivating work in Section 2. In Sections 3 and 4 we present our model formulation and resource-allocation policies respectively. We proceed with the performance evaluation of those policies in Section 5, and end with conclusions in Section 6.

2. Related work

Caron et al. [5] consider customers within a local distribution network that have access to the instantaneous total load on the grid, and introduce a distributed 'time/slackness' stochastic strategy that reduces the total cost for the customers and makes the overall load profile flatter. Koutsopoulos et al. [6] develop two online scheduling policies where demands are queued if the current load is higher than a threshold, and activated again either right before their deadline expires, or when there is enough residual energy. Using Jensen's inequality they also derive a lower bound on the average cost performance of all considered scheduling policies. Samadi et al. [7] consider a utility function for the customers and treat this as a welfare maximization problem, solved using a distributed pricing algorithm. Kim et al. [8] assume a time-varying price of electricity and a scheduler with statistical knowledge of future prices, and solve the resulting cost minimization problem by means of a Markov decision process. Alizadeh et al. [9] consider a neighborhood scheduler that queues the energy requests of 'deferrable' loads and optimizes the time at which they are served, so as to lower the overall cost and allow distributed energy resources to contribute. Fahrioglu et al. [10] look into how the utilities could design incentive-compatible contracts for effective demand management, using game theory principles.

Finally, we see parallels to our problem with the one examined by Goudarzi et al. [11] in multi-tier cloud computing systems, where the clients have Service Level Agreements (SLAs) and the system's profitability depends on the resource allocation that will allow these SLAs to be met.

3. Model formulation

In [12], the behavior of a power consumer is modeled as a k-state continuous time Markov Chain, since it is the random superposition of different appliances, themselves considered as multi-level on-off sources, according to [13]. We adopt the modeling assumptions presented in [6] because (i) they capture the bursty nature of arriving requests, as well as (ii) the fact that the chances of having large durations decrease fast, and (iii) they allow for mathematical tractability. We do not introduce any new assumptions.

In our model, a cluster (or 'network') comprised of M buildings generates power requests (or jobs). During a peak consumption period T, each building $n \in [1, M]$ issues requests (or jobs) according to a homogeneous Poisson process with an arrival rate λ_n. The time duration of each demand $\tau_{n,j}$ (where $j \in \mathcal{N}^+$ is a request counter), is exponentially distributed with parameter μ_n. The power requirement $p_{n,j}$ of each request is also considered exponentially distributed with parameter κ_n, and is measured in kW. Each customer request is then characterized by the tuple:

$$(\alpha_{n,j}, \tau_{n,j}, p_{n,j}) \tag{1}$$

where $\alpha_{n,j}$ is the interrarival time of requests in hours (hence exponentially distributed with parameter λ_n), $\tau_{n,j}$ is the duration of the request in hours, and $p_{n,j}$ is the instantaneous power consumption in kW, considered constant throughout $\tau_{n,j}$. Adopting the notation introduced in [5], the demand profile of each customer request is defined as:

$$d_{n,j}(t) := p_{n,j} \cdot \mathbb{1}_{\{\alpha_{n,j} \leq t \leq \alpha_{n,j} + \tau_{n,j}\}} \tag{2}$$

where $\mathbb{1}$ is the indicator function and $t \in [O, T]$. Likewise, the aggregate instantaneous load on the network is defined as:

$$L(t) := \sum_{n=1}^{M} \sum_{j} d_{n,j}(t) \tag{3}$$

The number of active requests on the network at any given time $t \in [0, T]$ is:

$$J(t) := \sum_{n=1}^{M} \sum_{j} \mathbb{1}_{\{\alpha_{n,j} \leq t \leq \alpha_{n,j} + \tau_{n,j}\}} \tag{4}$$

The cluster signs a 'tight guarantee' contract with the utility for a threshold of L_{CAP} kW. The unit cost for every kWh consumed when the power demand is equal to or less than L_{CAP} kW, is p_1. The integral over time of every kW of demand beyond L_{CAP} kW is charged at a premium of $p_2 = p_{BAT} \cdot p_1$ per kWh (i.e. p_{BAT} is a factor of the contract rate).

The grid is equipped with a battery rated at W kW, a discharge duration of H hours at its nominal power rating, and a charging efficiency η [14]. At the beginning of T it is fully charged with a capacity of $C = W \cdot H$ kW.

Now, let $E(t)$ be the remainder of the cluster's aggregate instantaneous load beyond the contract cap of L_{CAP} kW, that is:

$$E(t) := max(0, L(t) - L_{CAP}) \tag{5}$$

Further, let L_{GC}, L_{BAT}, L_{GP} be the loads that are served by the grid at the contract rate, the battery, and the grid at the premium rate respectively. During peak hours, the cluster's power requests are served in the following order: when $L(t) \leq L_{CAP}$ all the demands are satisfied by the grid at the low, contract rate of p_1 per kWh. When $L_{CAP} < L(t) \leq L_{CAP} + W$:

$$L_{BAT} = E(t) \tag{6}$$

This is the demand served by the battery at a rate of p_2 per kWh. Finally, if $L(t) \geq L_{CAP} + W$, $E(t)$ is greater than the battery's nominal rate (W kW), so there will be a remainder that is served by the grid at a premium ($p_3 = p_{GP} \cdot p_1$).

Note that, in the policy that we consider in this work, the charge p_2 of drawing from the battery is higher than the contract rate p_1, even though the battery is charged during off-peak hours when the unit cost of electricity is cheaper. The difference is the self-imposed charge by the cluster in order to pay off the amortized procurement and investment battery costs. The cluster could also settle on a price-point p_2 that is less than p_1 (i.e. $p_2 < p_1$), depending on the discount rate, the amortization period, and the agreed upon payback period of the investment (book life). The former option constitutes a more aggresive policy, and is the one we focus on here. At any rate, the unit cost p_2 for drawing from the battery should be lower than that of drawing from the grid at a premium (p_3), otherwise one could just skip the battery in this setup altogether. We therefore have:

$$p_1 < p_2 < p_3 \tag{7}$$

Table 1. Notations.

Parameter	Description
T	Duration of peak period. Measured in hours.
M	Number of buildings in cluster.
n	Building identifier. $n \in [O, M]$.
λ_n	Interarrival rate for power requests of building n. (Rate of a homogeneous Poisson process.)
μ_n	Service rate for power requests of building n. (Rate of a homogeneous Poisson process.)
κ_n	Rate that determines the power level of the requests of building n. (Rate of a homogeneous Poisson process.)
j	Request identifier. $j \in \mathcal{N}^+$.
$\alpha_{n,j}$	Arrival instant of request j of building n. Measured in hours. $\alpha_{n,j} \in [O, T]$.
$\tau_{n,j}$	Time duration of request j of building n. Measured in hours. $\alpha_{n,j} \in [O, T]$.
$p_{n,j}$	Power level of request j of building n. Measured in kW. $p_{n,j} \in \mathcal{R}^+$.
W	Nominal power rating of battery. Measured in kW.
H	Discharge duration of battery at its nominal power rating W. Measured in hours.
η	Charging efficiency of battery.
C	Battery capacity. Measured in kWh.
RAF_n	Resource allocation factor for building n.
AUC_n	Average unit cost of energy for building n.
$d_{n,j}(t)$	Demand profile of request (n, j). Measured in kW.
$L(t)$	Aggregate instantaneous load on the cluster. Measure in kW.
$J(t)$	Number of active requests on the network at time $t \in [O, T]$.
L_{CAP}	Power threshold set in contract for discounted electricity. Measured in kW.
$E(t)$	Quantity by which $L(t)$ exceeds L_{CAP}. Measured in kW.
L_{GC}	Load served by the grid at the contract rate. Measured in kW.
L_{BAT}	Load served by the grid at the contract rate. Measured in kW.
L_{GP}	Load served by the grid at the premium rate. Measured in kW.
p_1	Unit cost for every kWh drawn from grid-contract.
p_2	Unit cost for every kWh drawn from the battery.
p_3	Unit cost for every kWh drawn from grid-premium.
p_{BAT}	Factor by which p_2 is greater than p_1.
p_{GP}	Factor by which p_3 is greater than p_1.

We are effectively dealing with three *servers* that we access in order of increasing cost: grid at the contract rate (**grid-contract**), **battery**, and grid at a premium (**grid-premium**). Each energy request from the buildings is serviced *immediately*, i.e. there is no deferral. As noted in [6], when each request is served upon arrival, $J(t)$ can be thought of as the occupation process of an $M/M/\infty$ service system; it is therefore a continuous time Markov chain whose steady-state probabilities can be derived from equilibrium equations [15]. All the adopted notations are summarized in Table 1.

4. Resource-allocation policies

We consider the following scheduling policies, all of them put into effect by a network-wide controller that regulates the power flows for every member of the cluster; see Figure 1 for a conceptual system diagram.

In the 'strict bounds' policy, we use the stochastic characteristics of the requests generated by each building, to define a factor proportional to its estimated power needs. We call this the **resource allocation factor** and it is defined as follows:

$$RAF_n = \frac{\frac{\lambda_n}{\mu_n \cdot \kappa_n}}{\sum\limits_n \frac{\lambda_n}{\mu_n \cdot \kappa_n}} \tag{8}$$

We then use this quantity to establish bounds for each building when drawing from the grid or the battery. Specifically, each building can draw up to $RAF_n \cdot L_{CAP}$ kW from the grid, and up to $RAF_n \cdot C$ kWh of the battery's initial charge at a maximum rate of W/M. Any demands above these thresholds are accommodated by the grid at a premium.

In the 'no bounds' policy, all customer jobs are queued according to a first-come, first-serve logic. The grid-contract server picks up jobs from the head of the queue until $L(t)$ reaches L_{CAP} kW, then the battery comes into play until it becomes full; finally, the grid-premium server picks up any outstanding requests.

The 'adaptive bounds' policy is a hybrid of the previous two; when the aggregate instantaneous load on the network is less than the grid contract cap L_{CAP}, access to the grid-contract resource is not constrained. If the about-to-be-admitted job's power level is such that the total load will exceed the cap, the controller switches to a strict bounds regime. If the overall demand $L(t)$ drops under L_{CAP} again, the 'no bounds' policy is put into effect again, and any throttling at the grid-contract level is ceased.

Note that contrary to most of the work presented in Section 2, we do not defer load requests; we attempt to identify a resource allocation policy that improves the cluster's welfare, while serving its requests without delays.

To that effect, we track two output performance measures. One, the amount of energy that the cluster draws from the grid at a premium; this is a quantity that we wish to minimize so as to decrease the utility charges. Two, the average unit cost of energy for each customer. We expect this to be higher for a building that cannot tap into the grid-contract server with the same frequency as another building. In a fair resource allocation regime, the relative difference (or standard deviation) of the average unit costs of all buildings should be minimal.

Our expectation is that the 'strict bounds' rule will be the fairest of all considered policies, since each customer has uncontended access to their own 'power' band in the grid-contract and battery servers. (Note that there is no point in assigning bounds for the grid-premium server, since its resources are effectively infinite.) This is conditional on our definition of the resource allocation factor, and its appropriateness as an index for each building's energy consumption throughout the peak period.

The 'no bounds' policy is expected to perform well in terms of peak shaving for the whole cluster, since it stacks the incoming jobs on top of each other, without leaving any unused resources in the lower-priced servers (grid-contract and battery).

Finally, we expect our 'adaptive bounds' policy to combine the attractive characteristics of both previous policies to some degree; the fairness of the 'strict bounds' policy, and the load compacting of the 'no bounds' one. The underlying algorithm behind our 'adaptive' policy is shown in Algorithm 1.

Algorithm 1 Adaptive algorithm

```
if $L(t) < L_{CAP}$ then
    current policy ← no bounds
    building quota ← $L_{CAP}$
else
    current policy ← strict bounds
    building quota (grid-contract) ← $L_{CAP} \cdot RAF_n$
    building quota (battery) ← $C \cdot RAF_n$ at $W/M$
```

5. Performance evaluation

All the results below are the averages of 100 replications with 95% confidence intervals.

We consider 10 buildings in our cluster; all of them share the same stochastic characteristics when it comes to the power requests they generate. On average, jobs arrive once per *minute*, last 50 *seconds*, and have a power level that we gradually increase in each batch of replications from 20 to 120 *kW* (this is the x-axis of Figure 2). For the battery we wish to pick a capacity that (a) prevents the cluster from reaching to the grid-premium server, when we're at the lower range of the cluster's capacity, and (b) does push the cluster toward grid-premium otherwise, so that we can investigate the interaction of the cluster with all three bands, as the aggregate power demands of the cluster increase. We therefore assume that a 1.5 *MWh* Lithium-ion battery with 90% round-trip efficiency [16] is shared by the cluster, i.e. $W = 500\ kW$, $H = 3\ hrs$, $C = 1500\ kWh$, $\eta = 0.9$. This is the rounded maximum capacity at which the cluster does not draw power from the grid-premium server

Figures 2a and 2b quantify how well each policy can utilize the available resources, as the load on the network increases. The goal is to minimize the cluster's exposure to the costly grid-premium server, so we want the energy drawn from it to be as small as possible.

In that regard, as Figure 2a shows, the 'strict bounds' policy performs the worst, and the 'no bounds' policy the best, due to its ability to "stack" the jobs on top of each other and fully use the resources of the grid-contract server, before moving to the more expensive battery server, and then to the most expensive grid-premium server. The 'adaptive bounds' policy lies in between those two. In the beginning, it stays close to the performance of the 'no bounds' policy; the instantaneous load on the network is such that the policy can accommodate within the grid-contract server. As the load increases, the number of times the adaptive policy switches over to the strict policy increases; this is when the distance between the 'no bounds' and the 'adaptive' curve begins to grow in Figure 2a. Finally, the load becomes so large that the adaptive policy effectively degenerates to the strict one; this corresponds to the rightmost side of the figure.

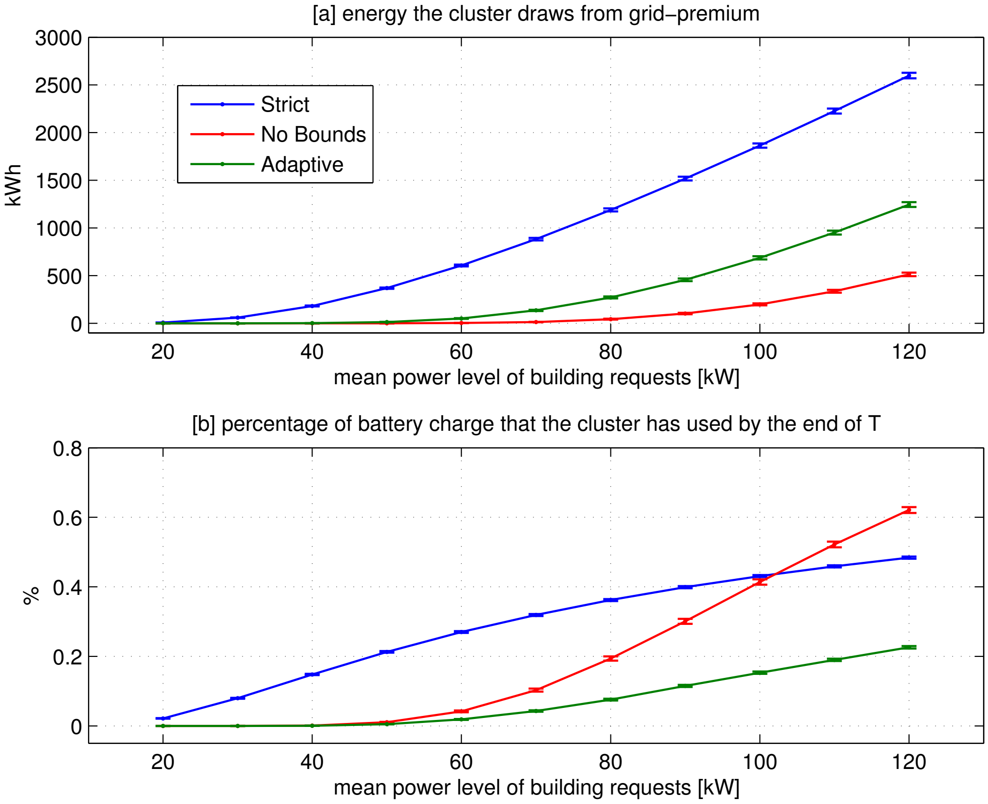

Figure 2. Utilization of premium-priced resources as the network load increases. (a) Top: Energy the cluster draws from grid-premium., (b) Bottom: Percentage of the initial battery charge C that the cluster has used by the end of T.

The adaptive policy is overall closer to the performance of the 'no bounds' policy than the strict one, i.e. it does not just perform as the average of these two. Given that the 'no bounds' policy here is the optimal one, this result is highly desirable. Notice how all three strategies begin at the same starting point; when the demand is low enough that access to the grid-premium server is minimized, all policies perform the same when it comes to absorbing the cluster's demand via the grid-contract and battery servers.

Figure 2b complements Figure 2a by showing how the battery server is used. Beginning with the leftmost side of the figure, we see that the strict rule is the first one to access the battery server; as soon as each building hits the ceiling of their allocated band in grid-contract, it moves on the battery server, regardless of any unused resources in the rest of grid-contract. On the contrary, both the 'no bounds' and adaptive policies defer access to the battery server until grid-contract is full. The results on the

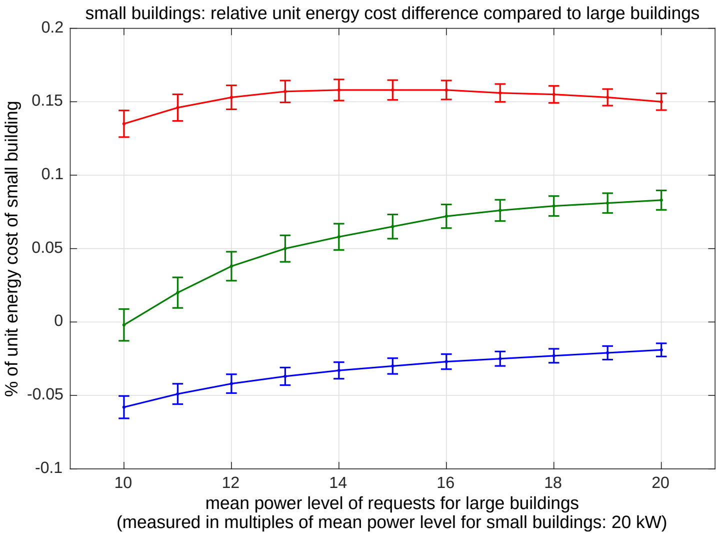

Figure 3. Small buildings: relative unit energy cost difference compared to large buildings. The units in the x-axis are measured in multiples of 20 kW, the mean power level of small buildings.

rightmost end will also be interpreted in conjunction with our observations on Figure 3 below. Given that access to grid-premium is inevitable due to the high aggregate load, the fact that the ‘no bounds’ policy uses up most of the battery charge is a desirable quality; the more it draws from the battery, the longer it delays its transition to the premium band. The adaptive policy uses the battery less than the strict rule, but this is because it utilizes grid-contract in a far more optimal manner.

In Figure 3 we are switching our focus to fairness. Let us consider 9 large buildings and 1 small one in our 10-building cluster. A building falls into the ‘large’ tier if its average power level is at least an order of magnitude larger than a building from the ‘small’ tier. In this example, the small building generates requests with an average power level of 20 kW. For the large buildings in the cluster, this number begins at 200 kW, or 10 times the level of a small building, and eventually grows to 400 kW (this is the x-axis of Figure 3). The mean interarrival and service time rates remain fixed for all buildings at $\frac{1}{60 \text{ seconds}}$ and $\frac{1}{50 \text{ seconds}}$ as before. (Note that tiers could have also been drawn by changing how often the buildings generate requests, or how long their requests need to be serviced for, i.e. by modifying the λ or the μ parameter in the (λ, μ, κ) tuple and keeping all other values fixed.)

Going back to the considered setup, this is a network where the requests coming from large buildings dominate the network when it comes to using its resources, and as we increase their power level, this phenomenon becomes even more prevalent. It is therefore a good stress test for the fairness performance of all policies. We evaluate fairness as follows: at the end of each run, we multiply the amount of energy each building drew from each of the three servers by their unit cost. We add up these products, and divide the result by the overall energy usage of the building:

$$AUC_n := \frac{p_1 \cdot \int_0^T L_{GC,n}(t)dt + p_2 \cdot \int_0^T L_{BAT,n}(t)dt + p_3 \cdot \int_0^T L_{GP,n}(t)dt}{\sum_j p_{n,j} \cdot t_{n,j}} \tag{9}$$

This gives us the average unit cost of energy (notated as AUC in the formula above) for each building. In a fair policy, the standard deviation between the unit costs of all buildings should be as small as possible. In our case, since we are dealing with 9 large buildings which have similar unit costs due to their identical stochastic characteristics, we focus on the relative difference between the two tiers as a less biased metric of fairness.

$$\frac{AUC_{\text{small}} - AUC_{\text{large}}}{AUC_{\text{small}}} \tag{10}$$

As with the standard deviation, we wish to keep that difference as small as possible.

Examining Figure 3 we observe that, as the network becomes saturated with power requests, all policies eventually settle to a steady state, but the levels at which they settle are different. The strict strategy performs the best, with the relative difference remaining within 5% throughout the entire range of scenarios considered. The adaptive policy is slightly worse, due to its operation at times as 'no bounds'; it settles at a relative difference in unit costs of around 7%. However, it constitutes a significant improvement over 'no bounds', improving its fairness by a factor of two.

For the adaptive policy, also observe that under heavy load, the unit cost for small buildings is slightly higher than that of large buildings. This happens because in the adaptive regime, when access to grid-contract happens on a first-come, first-serve basis (i.e. when it is effectively operating as 'no bounds') there will be instances that this server's resources are used entirely by the dominant traffic generator on the network, that is, the 'large building' tier. Such occurrences bring the tier's average unit cost down. This is not possible in the 'strict' regime since the large buildings are always limited to their own allocated bands and cannot capture grid-contract in its entirety; this is why the strict policy practically tends to zero.

Finally, we examine how sensitive the policies are to changes in the access costs of the premium-priced servers (battery and grid-premium), and whether they pass along this cost difference to the large and small customers in the cluster in a *proportional manner*. This is always tied directly to the aggregate load on the cluster; performing this sensitivity analysis on a different operating point will bring about different results. However, for a given operating point, this evaluation allows us to identify how each policy utilizes the premium-priced resources (battery and grid-premium) and whether it passes along the cost increases fairly (i.e. does the relative unit cost difference decrease?).

In our considered setup, the 'large building' tier now generates requests with a fixed average power level of 300 kW. All other building-related parameters remain the same as before. This is a heavily loaded network that keeps all servers busy.

In Figure 4a, we are gradually increasing the cost parameter p_{GP} that applies when accessing grid-premium (this is the x-axis of Figure 4a). This corresponds to the scenario where the cluster is willing to sign a contract with tighter guarantees, and thus a higher premium if these guarantees are not honored. (The reward is cheaper access to grid-contract.) In Figure 4b, we keep the cost of accessing grid-premium fixed, and instead gradually increase the cost of drawing from the battery (while keeping it below the grid-premium unit cost at all times). This could correspond to the case when the cluster increases the self-imposed fee of accessing the battery in an attempt to pay off the battery installation costs sooner, or invest in more capacity.

We observe (Figure 4a) that the relative difference between unit costs of small and large customers gradually increases in the 'strict' policy. This happens because the large customers are, in relative terms, more exposed to the grid-premium band than the small customer. (Remember that access to

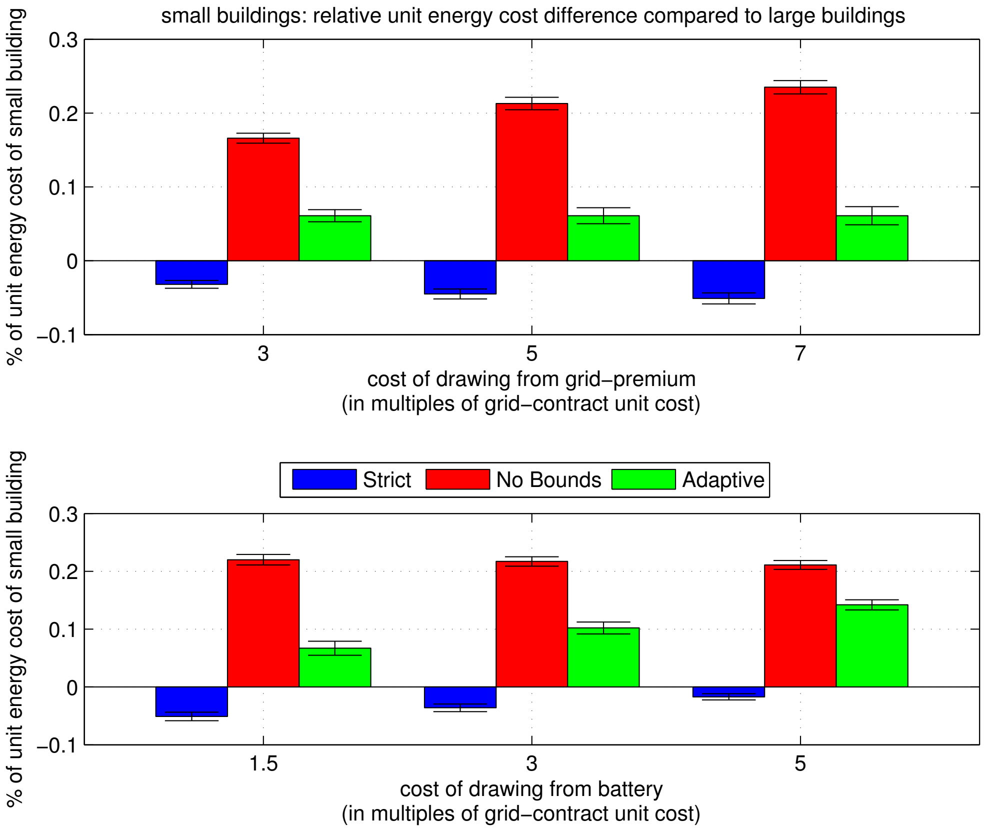

Figure 4. Small buildings: relative unit energy cost difference compared to large buildings. (a) Top: p_{GP} **ranges from** 3 **to** 7. p_{BAT} **remains fixed at** 1.5. **The strict policy drew** 910 kWh **from the battery and** 8,952 kWh **from grid-premium. Those numbers are** 1500/6047 **for the 'no bounds' policy, and** 643/7,303 **for the adaptive one, (b) Bottom:** p_{BAT} **ranges from** 1.5 **to** 5. p_{GP} **remains fixed at** 7. **The strict policy drew** 910 kWh **from the battery and** 8,952 kWh **from grid-premium. Those numbers are** 1500/6047 **for the 'no bounds' policy, and** 643/7,303 **for the adaptive one.**

the grid-premium server is unrestricted in all policies.) Therefore, the average unit cost of the 'large-building' tier increases more than that of the small building. Conversely, and following the same logic, the relative difference decreases when the battery costs increase (Figure 4b); for the power level that we picked for the small building, the percentage of its jobs served by the battery server compared to grid-premium is higher than that of the large buildings, where a sizeable portion is served at grid-premium. A similar argument can be made for the increase that the adaptive rule demonstrates in Figure 4b; the percentage of small building jobs served at the battery is higher than that of large buildings jobs (a considerable portion of which is now served at grid-contract).

Notice that for the adaptive policy, as Figure 4a shows, a cost increase in grid-premium leaves the relative unit cost difference between small and large buildings unaffected. Part of the large building jobs that were served in grid-premium under the strict regime, are now served in grid-contract (in those time instances when the instantaneous total load falls below grid-contract threshold). This leaves the same percentage of 'grid-premium'-served jobs for both types of buildings, so the cost increase affects them both in equal measure. This is, again, a side-effect of the way the adaptive policy utilizes the lower-priced servers (when operating under a 'no bounds' rule), and its advantage over the strict strategy. We note that in both scenarios, the adaptive policy remains a fairer policy than the 'no bounds' one, in consistence with what we saw in the previous section (Figure 3).

The analysis above focuses on (a) the relative ease of access that each type of customer (small/large) has to the grid-contract server, and (b) to the effect the price increases on premium-priced servers have across policies, both by means of the relative unit energy cost difference of small buildings compared to big buildings. Future extensions of this work may also wish to consider the frequency with which each customer accesses the battery when assessing fairness, as extensive battery usage leads to more charging cycles and eventually a degradation of the battery's life.

6. Conclusions

As we have demonstrated, our adaptive policy combines the benefits of both the strict and the 'no bounds' policy. In our view, and in light-load conditions, fairness is not an issue since the power resource is not congested; what matters is to avoid premium charges since every request can be accommodated by the low-priced grid-contract server.

The adaptive policy operates similarly to the 'no bounds' rule, thus allowing a cluster to utilize its own resources (grid-contract and battery) optimally, and to avoid premium charges from the utility. On the other hand, when the network is heavily loaded with power requests, access to the grid-premium server is unavoidable, so ensuring fairness –in the sense of a similar average unit cost for all buildings– is a key matter. For that reason, our adaptive rule evolves into a strict regime, enforcing each customer to limit their requests in each server to their own properly-sized band. The size of these bands is calculated using the resource allocation factor formula that we introduced (Eq. 8).

Because the adaptive policy smartly alternates between these two regimes when it most makes sense, it performs better than an average of the two base policies would. We have shown via sensitivity analyses that premium energy usage is closer to the 'no bounds' policy (the optimal in that front) than to the strict one. Similarly, its fairness performance is closer to the 'strict' strategy than to the 'no bounds' one. It is therefore an optimal combination of the two policies, suitable for a microgrid operator where equal weight is given to both cluster-wide cost minimization and fairness among all customers.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. Time-of-Use. Pacific Gas and Electric Company, 2015. Available from: http://www.pge.com/touintro/.
2. Eyer J, Corey G (2010) Energy storage for the electricity grid: Benefits and market potential assessment guide. *Sandia National Laboratories*: 69-73.

3. Demand Charges 101. Stem, 2015. Available from: http://www.stem.com/resources/learning.
4. Public Utilities Code Section 748 Report to the Governon and Legislature on Actions to Limit Utility Cost and Rate Increases. California Public Utilities Commission, 2012. Available from: http://www.cpuc.ca.gov/NR/rdonlyres/339C0DD6-0298-4BC7-AAD9-A27779AA43D4/0/2012SB695ReporttoGovernorandLegislatureFinalv2.pdf.
5. Caron S, Kesidis G (2010) Incentive-based energy consumption scheduling algorithms for the smart grid. *First IEEE International Conference on Smart Grid Communications (SmartGridComm)* IEEE, 391-396.
6. Koutsopoulos I, Tassiulas L (2012) Optimal control policies for power demand scheduling in the smart grid. *IEEE Journal on Selected Areas in Communications* 30: 1049-1060.
7. Samadi P, Mohsenian-Rad AH, Schober R, et al. (2010) Optimal real-time pricing algorithm based on utility maximization for smart grid. *First IEEE International Conference on Smart Grid Communications (SmartGridComm)*IEEE: 415-420.
8. Kim T, Poor H (2011) Scheduling power consumption with price uncertainty. *IEEE Transactions on Smart Grid* 2: 529-527.
9. Alizadeh M, Scaglione A, Thomas R (2012) From packet to power switching: Digital direct load scheduling. *IEEE Journal on Selected Areas in Communications* 30: 1027-1036.
10. Fahrioglu M, Alvarado F (2000) Designing incentive compatible contracts for effective demand management. *IEEE Transactions on Power Systems* 15: 1255-1260.
11. Goudarzi H, Pedram M (2011) Multi-dimensional SLA-based resource allocation for multi-tier cloud computing systems. *IEEE International Conference on Cloud Computing (CLOUD)* IEEE: 324-331.
12. Ardakanian O, Keshan S, Rosenberg C (2012) On the use of teletraffic theory in power distribution systems. *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet* ACM, 21.
13. Richardson I, Thomson M, Infield D, et al. (2010) Domestic electricity use: A high-resolution energy demand model. *Energy and Buildings* 42: 1878-1887.
14. Ghiassi-Farrokhfal Y, Keshav S, Rosenberg C (2015) Toward a realistic performance analysis of storage systems in smart grids. *IEEE Transactions on Smart Grid* 6: 402-410.
15. Bertsekas D, Gallager R, Humblet P (1992) Data networks, Volume 2, New Jersey: Prentice-Hall International.
16. Ferreira H, Garde R, Fulli G, et al. (2013) Characterisation of electrical energy storage technologies. *Energy* 53: 288-298.