Minimizing Cost Sharing Among Residential Electric Customers with Solar and Storage

Michaela Levine#1, Amanda Sun#2, Zhenhua Zhang#3

#*Civil and Environmental Engineering, Stanford University*
450 Serra Mall, Stanford, CA 94305, USA
1mblevine@stanford.edu
2asun3@stanford.edu
3zhenhua@stanford.edu

Abstract — **This study introduces distributed energy resources (DERs) to the optimal aggregation and pricing scheme developed by Patel et al. (2013) to group residential electric customers according to their cost-to-serve in wholesale electricity markets [1]. The bisection algorithm used in Patel et al. (2013) was applied to several use cases of distributed solar and/or storage considering various ownership structures and deployment strategies. Utility management of solar and/or storage resources results in a lower total cost of procurement than the deployment of the resource at the customer level for demand charge management and energy arbitrage. The average cost to serve customer groups of all sizes decreases significantly when customers have access to both solar and storage compared to either solar or storage alone.**

I. INTRODUCTION

Traditionally, utilities purchase electricity in bulk from wholesale markets on behalf of all of their customers in aggregate and charge their customers a flat rate for energy consumption. While this structure provides several advantages to the customer as the utility takes on the burden of purchasing power and forecasting demand, which is difficult to do with certainty on an individual household basis, it does result in cost sharing amongst a utility's customers. Customers that consume electricity primarily in off-peak and lower price times have to pay a higher rate than they would if the utility purchased wholesale electricity just to serve them to compensate for the utility's purchases at a higher price to serve other customers who consume at peak times. Thus, customers that are cheaper to serve subsidize those customers that are more expensive to serve through cost sharing.

While time-of-use (TOU) rates have been introduced as a means to reduce sharing by offering time-varying retail rates, consumers have been found to be more responsive to the average price of power [4]. Reference [1] demonstrated an optimal aggregation and pricing scheme in which residential customers are grouped based on their cost to serve and a given level of tolerated uncertainty in load forecasting. The scheme aggregates customers into groups where a single flat rate could be offered to all customers in the group to minimize cost sharing. This aggregation can be used by any load serving entity (LSE) or utility that purchases electricity from wholesale markets. The cost to serve customers was modeled as the cost to procure power in a two-stage electricity market in this scheme. The bisection algorithm developed the authors to partition customers results in stable groups such that no individual customer can reduce its cost by moving to a different group. This aggregation and pricing scheme could be useful to LSEs in deregulated markets—where customers can be offered a variety of rate plans—to reduce turnover by providing customers meaningfully cheaper and stable rate plans.

This study introduces distributed energy resources (DERs) to the optimal aggregation and pricing method developed by [1]. Behind-the-meter (BTM) and community solar are proliferating, changing the net load profiles of utilities' customers and the cost of procurement of power in wholesale markets. As the cost of battery energy storage falls, the deployment of both BTM and utility-scale batteries will change load curve dynamics as well. We examine how the pricing and grouping dynamics change in the scheme developed by [1] for several common use-cases of DERs: customer-sited solar and/or storage, community solar, and utility-owned solar and/or storage. We also consider the impact of the deployment of DERs to the total cost of procurement to the utility.

II. LITERATURE REVIEW

Several methods have been introduced to address the issue of cost-sharing. As discussed, [1] uses a bisection algorithm to aggregate a LSE's customers into groups to minimize both their per unit cost of electricity on the wholesale market and load forecast error. Given a predetermined group size M**,** the algorithm converges to an optimal vector of households that give the minimum possible cost to serve. The authors then determine how best to determine M, as larger groups will have a higher average cost to serve due the inclusion of customers whose consumption aligns more closely with peaks in electricity prices, but will also have lower forecasting errors. The authors use an autoregressive moving average (ARMA) model to predict customer loads and the coefficient of variance (CV) as the metric to calculate forecast error. They show that LSEs can set the maximum CV they are willing to tolerate and determine the minimum group size *M* such that the CV of all groups of that size does not exceed that tolerance. The LSE can then run the bisection algorithm multiple times to sort all customer into groups and determine the cost to serve each group.

Though [1] uses a bisection algorithm for optimal pricing schemes, that is not the only way to segment the customer population. Reference [2] introduces a K-means algorithm for customer clustering, applied to three methods for load curve classification: classical K-means, modified K-means (in which the effect of outliers is removed by just considering densely situated points within each cluster), and hierarchical K-means. Reference [3] also applies K-means clustering to group customers together. The authors use encoding to match residential customer load shapes to those in a dictionary, which was populated by applying K-means clustering to the dataset to get representative load shapes. The value of K was determined by then finding the number of profiles needed so that matching the load profiles to the dictionary profiles did not produce a squared error of a significant size. While these methods focus on typical load shapes in grouping customers, our analysis groups customers based on cost, which is a function of both variations in load shapes and market prices.

Standard economic models predict that customers respond to marginal price rather than average price. However, for a nonlinear price schedule such as electricity prices, [4] shows that customers respond to average price. Through the encompassing test, the author shows that average price has a significant effect on consumption, while marginal price and expected marginal price have statistically insignificant effects. In fact, customers' perception of their electricity prices are extremely similar to the average price. The author suggests that nonlinear electricity pricing is an unsuccessful strategy in achieving policy goals of energy conservation, until more customers are educated on and provided real-time information about their price schedules and consumption levels. Given the implications of [4], we model our study to generate average prices, whether for individual customers or groups of customers. However, in modeling DERs, we also consider time-varying rate schedules as these make ownership of DERs financially attractive to electric customers.

III. PROBLEM SETUP

Utilities conduct procurement of electricity from wholesale markets in two-stages: the day-ahead and real-time market. In the day-ahead market, the utility forecasts its customers aggregate consumption for the next day and purchases an amount of electricity related to that forecast at the day-ahead price. The following day, the utility purchases from the real-time market the residual amount of demand not purchased in the day-ahead to meet its customer actual demand.

For simplicity, we assume that the utility procures all electricity for its customers at the day-ahead locational marginal price (LMP). This assumption is valid if the electricity market is efficient such that real-time and day-ahead prices are equal. Thus, we can write the cost to the utility for procurement of its customers' aggregate demand, **d**, at the day-ahead price, **p**, on a given day, **k**, as:

$$c_k = p_k^T d_k$$

The rate paid by the utility per unit of electricity ($/kWh) on a given day is then:

$$\lambda_k = \frac{p_k^T d_k}{1^T d_k}$$

The utility can determine the rate it pays to serve an individual customer by examining its historical demand over a period of H days and calculating the rate it pays to procure energy for that customer using historical wholesale market prices. The cost to serve the i[th] customer is then:

$$\widehat{\lambda^{(i)}} = \frac{\sum_{h=1}^{H} p^T d_h^{(i)}}{\sum_{h=1}^{H} 1^T d_h^{(i)}}$$

To recruit and aggregate a group of customers that has a low cost to serve, a selection vector, **u**, identifies customers to minimize the cost to serve metric, λ_u :

$$\lambda_u = \frac{\sum_{h=1}^{H} p_h^T \left(\sum_{i=1}^{N} u_i d_h^{(i)} \right)}{\sum_{h=1}^{H} 1^T \left(\sum_{i=1}^{N} u_i d_h^{(i)} \right)}$$

When BTM solar is introduced, determining the cost to serve metric for an individual or groups of customers is complicated by periods when the customer(s) exports generation back to the grid. In jurisdictions with net metering, the utility pays the customer for this surplus electricity. We, therefore, determined the cost to serve metric using the absolute value of individual customers' or groups' net load:

$$\lambda_u = \frac{\sum_{h=1}^{H} p_h^T \left| \left(\sum_{i=1}^{N} u_i d_h^{(i)} \right) \right|}{\sum_{h=1}^{H} 1^T \left| \left(\sum_{i=1}^{N} u_i d_h^{(i)} \right) \right|}$$

While utilities likely pay their customers a higher retail rate for exported electricity, we simplify this dynamic by assuming the utility will pay the customer the wholesale market price. This assumption results in an underestimate of the cost the utility incurs to serve customers with BTM solar.

In addition to assessing the impact of solar on the cost to serve metric, we also consider the introduction of customer-sited and utility-scale batteries. We first developed optimization models for use cases of each and use the resulting load profiles to calculate the cost to serve. Both types of batteries were modelled by the following equations:

$$\forall t = 1...T : \ 0 \le u_t^+ \le \overline{u}$$

$$\forall t = 1...T : \ 0 \le u_t^- \le \overline{u}$$

$$\forall t = 1...T-1 : x_{t+1} = x_t + \mu u_t^+ - \frac{1}{\mu} u_t^-$$

$$\forall t = 1...T : 0 \le x_t \le B$$

$$x_0 = B/2$$
$$(0.95)B/2 \leq x_T \leq (1.05)B/2$$

Here u_t^+ and u_t^- are the charging and discharging rate of the battery at time t, $\overline{u}$ is the maximum discharge and charge rate, μ is the charge and discharge efficiency of the battery, x_t is the state-of-charge (SOC) of the battery, and B is the capacity of the battery. Due to the computational intensity of modelling battery operations for every customer for each hour of the year, the optimization was conducted for only the first month of the dataset, January 2016, and an end condition constraint was used to ensure the battery ended with a similar SOC. For scenarios in which solar and storage are both modelled, an additional constraint is added such that the battery can only charge using the available solar generation, s_t:

$$\forall t = 1 ... T : \ u_t^+ \leq s_t$$

For a customer on a flat tariff, the use cases for a battery are highly limited and would most likely be used for backup power in the event of an outage. In the near future, the most likely use case of customer-sited batteries is demand charge management and energy arbitrage on a TOU rate [5]. Thus, we optimized the operations of customer-owned batteries to minimize the customer's bill, where l_t is the customer's gross load at time t, q_t is the retail price of energy at time t, and d_c is the demand charge:

$$min \left[\sum_{t=1}^{T} \left(l_t + u_t^+ - u_t^- \right) q_t + max\,(l + u^+ - u^-)\, d_c \right]$$

For a utility scale battery, we first considered the use case of LMP arbitrage. The objective function is to minimize the cost of energy procurement for the aggregate load, L, of the utility's customers:

$$min \left[\sum_{t=1}^{T} \left(L_t + u_t^+ - u_t^- \right) p_t \right]$$

Second, we consider the use of a utility-scale battery for peak shaving. In certain electricity markets, utilities may be incentivized to engage in peak shaving to minimize costs incurred in capacity markets or for transmission services. In this use case, the utility's peak load is minimized with the following objective function.

$$min\,[max\,(L + u^+ - u^-)]$$

The result of this optimization is used to determine the peak load of the utility. The peak load is then used as a constraint while optimizing the battery for LMP arbitrage:

$$\forall t = 1 ... T : (L_t + u_t^+ - u_t^-) \leq peak$$

The resulting load profiles from optimizing these DERs were then applied to the cost to serve algorithm described below. In all objective functions, when solar is incorporated into the analysis, the objective function is adjusted so that net load is calculated net of solar generation.

IV. METHODOLOGY

A. Algorithm

Reference [1] developed a bisection algorithm that identifies a group of households of size *M* that has the lowest cost to serve amongst a larger group of customers. Here, the elements of vector **t** is the cost of procurement for each household, t_i:

$$t_i = p_h^t d_h^{(i)}$$

And, the elements of vector **w** are the demand of each household, w_i:

$$w_i = 1^T d_h^{(i)}$$

The algorithm converges iteratively on a value λ. Each iteration begins with a value for λ that is halfway between the upper bound and lower bound for the value. The algorithm is initialized by setting the upper bound to the cost to serve the most expensive customer in the set and the lower bound to the cost to serve the least expensive customer. For each iteration, the households are sorted in the ranking vector (**t**-λ**w**), and then the selection vector is created by selecting the lowest M households. If the transpose of the ranking vector multiplied by the selection vector is less than 0, then the selection is feasible and the upper bound on the cost to serve is set to λ. If the selection is not feasible, than the lower bound on the cost to serve is raised to λ. The algorithm terminates when the convergence criteria is met. Fig. 1 depicts the bisection algorithm.

Algorithm 1 The LSE can use this alorithm to select the group of M customers who had the lowest cost to serve $\hat{\lambda}_{\mathbf{u}}$ over a period of historical data.

```
Initialize bisection method:
λ̄ ← max{t_1/w_1, ..., t_N/w_N}
λ̲ ← min{t_1/w_1, ..., t_N/w_N}
Set γ                                   ▷ Convergence threshold
Bisection method:
while λ̄ − λ̲ > γ do
    λ = (λ̲ + λ̄)/2                        ▷ Update current λ
    Compute (t − λw)
    Sort (t − λw) in ascending order to obtain {i_1 ... i_N}
    Construct u_{λ,i} = 1 if i ∈ {i_1, ..., i_M}
    if (t − λw)^T u_λ ≤ 0 then            ▷ Feasibility test
        u_λ feasible
    else
        u_λ ← ∅
    end if
    if u_λ = ∅ then
        λ̲ ← λ                 ▷ Infeasible, increase lower bound
    else
        λ̄ ← λ                 ▷ Feasible, decrease upper bound
    end if
end while
Result:
Minimum λ
Associated selection vector u_λ          ▷ 1^T u_λ = M
```

Fig 1. Cost to serve bisection algorithm from [1].

B. ARIMA model

While recruiting and and conducting procurement on behalf of a single customer would in theory eliminate the cost sharing burden to that customer, this not tenable in practice and would also result in large forecasting errors which could force utilities to purchase a significant amount of electricity at the more expensive real-time price. With larger groups of customers, we expect the forecasting error to decrease, but as seen in the bisection algorithm, a larger group will result in a higher cost-to-serve. Thus, there is a tradeoff in applying this algorithm in selecting a reasonable group size so that forecast errors are minimized as well as the cost to serve the selected group.

To determine the optimal group size M, we calculated the CV of the load forecast for groups of various sizes to analyze the relationship between group size and forecasting error. First, we established an autoregressive integrated moving average (ARIMA) model with daily temperature as an exogenous variable to predict the daily total consumption for the last three months of the year using consumption data from the first nine months. Then, we established a Vector ARIMA model, with hourly temperature as an exogenous variable, to predict the daily load profile vector s for the last three months, using the normalized load profile for the first nine months. Each vector contains the fraction of hourly consumption for each day. Finally, we multiplied the predicted daily total consumptions by predicted normalized load profile, $y s$, to obtain hourly consumption for the last three months of each group of different sizes.

To tune the ARIMA model and the Vector ARIMA model, an augmented Dickey-Fuller (ADF) test was applied to determine the degree of differencing. The result shows hourly consumption data are stationary when the degree of differencing is 1. The number of observations and the order of moving average included in the ARIMA model are 10 and 5, respectively, as determined by the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) tests. Similarly, the order of the Vector ARIMA model for the number of AR and MA parameters have been chosen as (1,1).

The forecasting error (CV) can be calculated as follows, where $d(h)$ is the actual hourly consumption at time h for each group, and $\hat{d}(h)$ is the predicted hourly consumption:

$$CV = 100\frac{\sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(d(h)-\hat{d}(h)\right)^2}}{\frac{1}{H}\sum_{h=1}^{H}d(h)}$$

Fig. 2 shows the relationship between the group size and the forecasting error. We applied the ARIMA model and Vector ARIMA model to both randomly constructed groups of size M and groups determined by selecting the M customers with the lowest cost to serve. For both group types, the CV curve is non-monotonic and exhibits complicated behaviors, suggesting that CV is a function of both the selection vector and the group size. For non-random groups, before the group size reaches 20, the CV curve decreases monotonically. As group size exceeds 20, the CV curve increases in general. Thus, a group size of 20 has been chosen for the analysis. For simplicity, all groups will have the size of 20 in this analysis.

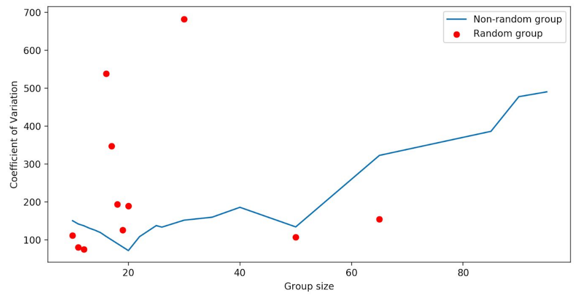

Fig. 2 The CV for sorted group (blue line) and the CV for randomly constructed groups (red dots)

C. Application of the algorithm to DER use cases

The cost to serve algorithm was applied through multiple use cases of solar and/or storage. First, the algorithm was applied for groups of various sizes for the use case in which the customers share a 400 kW community solar resource. The solar resource was distributed to all customers through three different allocation approaches: the customers receive equal shares of the resource, the customers receive a share of the resource proportional to their gross load, and the customers receive a share of the resource proportional to the cost to serve their gross load. The net load was calculated by subtracting the share of the hourly solar profile allocated to that customer from its gross load profile in each of the three allocation approaches. The algorithm was applied to the resulting net demand profiles for each customer.

The cost to serve algorithm was also applied to the net load profiles generated through customer-sited resources in three cases: solar, storage, and solar and storage. The net load profiles for the customer solar case were determined by subtracting the solar generation of the customers' resources from their gross loads. Not all customers in the dataset have BTM solar. For these customers, the hourly generation profile is always 0 and their load net of solar is their gross load. The absolute value of the net load profiles was used for the bisection algorithm as described in the problem setup. The net loads for the storage and solar and storage cases were constructed from the results of optimizing the customers' batteries to a TOU as previously described.

We also compared the impact to the total cost of procurement for the utility of customer-sited DERs and utility-owned DERs. The profiles of all customers following the TOU optimization were aggregated together and multiplied by the hourly LMPs to determine the total cost to the utility if every customer optimizes its resources according to the signal provided by the TOU rate. The utility was then given a solar resource equivalent to the total generation of its customers and a battery with the aggregate capacity of the customers' batteries. These resources were then optimized to conduct peak shaving and/or LMP arbitrage. The total costs from the resulting optimizations were compared to the total cost incurred when customers optimize for the TOU rate.

Finally, we examined how the cost to serve a group formed based upon its gross load changes when it adopts DERs. The bisection algorithm was applied using the gross load of the customers to sort all customers into groups of 20 according to their cost to serve. We then optimized the aggregated load and DERs of the group to a TOU rate. The cost to serve the customers with DERs was also calculated where the individual customers managed their resources independently and sought to minimize their utility bill.

D. Data

Hourly smart meter data including net load use, gross load use and solar generation for 339 customers in Houston, Texas, in 2016 are obtained from dataport.cloud, a database maintained by Pecan Street Inc. For the ARIMA model, the dataset of 8784 hourly load records was separated into a training set (the first nine months, 75%) and a testing set (the last three months, 25%), to determine the optimal group size. When aggregating customers into groups and conducting the battery optimizations, only load data from the first month was used due to the computational intensity of optimizing DERs for all customers for every hour of the year. Hourly temperature data over 2016 in Houston, Texas, was obtained from NOAA National Centers for Environmental Information and used as an exogenous variable in the establishment of ARIMA models to reduce seasonal bias [6].

Locational Marginal Price (LMP) data from 2016 was obtained from ERCOT day-ahead market archive, in order to calculate the cost to serve each individual customer as well as aggregated groups [7]. For community solar modelling, an hourly generation profile was obtained from the National Renewable Energy Laboratory's PVWatts for Houston, Texas [8].

V. RESULTS

The bisection algorithm was applied to the three community solar allocation cases to find the average cost to serve groups of varying sizes. In all cases, the average cost to serve the group increases with group size. The bisection algorithm identifies the cheapest households to serve, so increasing the group size incorporates increasingly expensive customers. The average cost to serve curves shown in Fig. 3 are remarkably similar, especially in the cases where solar is proportioned according to the cost to serve and according to consumption. While the cost to serve decreases with the introduction of solar, a large decrease is not observed. The size of the community solar array was approximately 32% of the peak load for the community and likely too small to create a large change in the cost to serve.

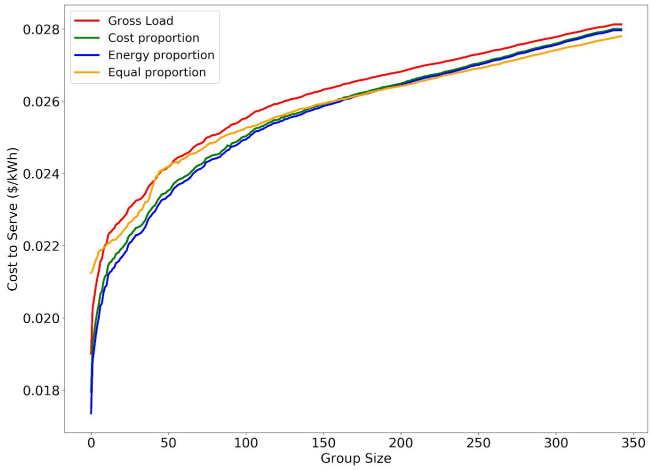

Fig. 3 The average cost to serve the cheapest group of customers based on group size. Three community solar allocation scenarios are modeled and compared to the gross load case.

The bisection algorithm was then used to find the average cost to serve the cheapest group of any size when customers have BTM resources. The result of the algorithm using only the gross load was compared to three customer-sited resource cases: solar, storage, and both solar and storage. Fig. 4 compares the average costs to serve in these four cases. The cost to serve is lowest in the case of both BTM solar and storage. When only one of the two is used, the average cost to serve the cheapest group is lower with only storage than with only solar.

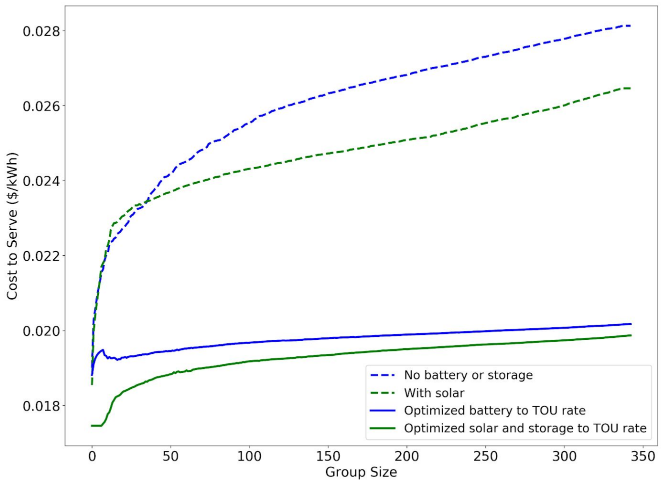

Fig. 4 The average cost to serve the cheapest group of customers based on group size. Four scenarios are modeled to show that the cost to the utility decreases with the implementation of BTM solar and/or storage.

Interestingly, the cheapest group of up to approximately 30 customers is more expensive to serve with BTM solar than without solar or storage. This is likely due to the net metering structure of our study; customers who consume less energy than they produce with their solar resource sell that extra electricity back to the utility. For small group sizes, these customers cause an increase in the average cost to serve the group. However, for groups larger than approximately 30, this phenomenon is no longer observed as cost savings overshadow the initial slight increase in cost. The aggregate size of the BTM resource in our dataset is much larger (approximately 73% of peak load) than the modeled community solar resource, resulting in a greater decrease in the cost to serve.

To investigate how the costs to serve groups changes with the introduction of DERs, the households were aggregated into groups of size 20, which was determined previously to be the optimal group size that minimizes CV, using their gross load (scenario 1) or their net load after solar generation (scenario 2). The customers were then given batteries. In each scenario, the cost to serve the group was compared in three cases: no customers have storage; individual customers optimize their batteries on a TOU rate; and the group optimizes an aggregated shared battery to minimize the group's bill on a TOU rate. Fig. 5 shows that incorporating storage had a significant impact on reducing the cost to serve. While introducing solar and/or storage significantly reduced the cost to serve each group, no consistent benefit to minimizing the cost to serve was found by optimizing DERs on a individual basis or group basis.

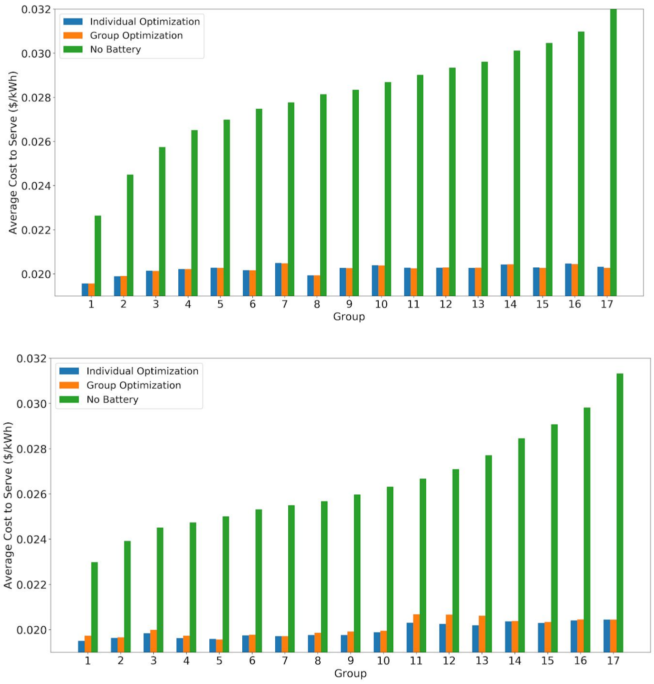

Fig. 5 Comparison of average cost to the utility to serve each customer group using (top) storage only and (bottom) both solar power and storage. For both, the average cost to serve each group of 20 customers is modeled under three battery scenarios: (1) Optimization of individual customers, (2) Group optimization under a TOU rate with demand charge in groups of size 20, and (3) No battery.

We further compare cost to the utility in three different use cases of storage only versus a combination of solar and storage: (1) energy arbitrage on a TOU rate in which each individual customer owns a battery, (2) LMP arbitrage in which the utility has a battery whose size is equal to that of the aggregated individual batteries, and (3) peak shaving followed by LMP arbitrage with the utility-scale battery. Fig. 6 compares the total cost to the utility and peak load of the system in the six scenarios. As expected, utility ownership of solar and/or storage reduces its cost of procurement. The utility's total cost under (2) LMP arbitrage and (3) peak shaving followed by LMP arbitrage decreased when utility-scale solar is added. However, total cost increased in case (1), individual customer battery optimization, when BTM solar is added due to the utility having to purchase unused solar from their customers that was not modelled to be sold elsewhere. However, the average cost per kWh the utility incurs in that scenario is $0.0200/kWh, slightly less than the average cost without solar, $0.0202/kWh. This suggests that if our model incorporated the sale of customer generated power to other customers, the cost to the utility in the solar scenario would actually be less. The system peak load in all three cases decreased when solar was incorporated, indicating that the utility could benefit from the addition of DERs by not having as high of a maximum load to serve. Although the peak shaving case resulting in a significant decrease in system peak load, it also resulted in slightly higher energy procurement costs compared to the LMP arbitrage alone case.

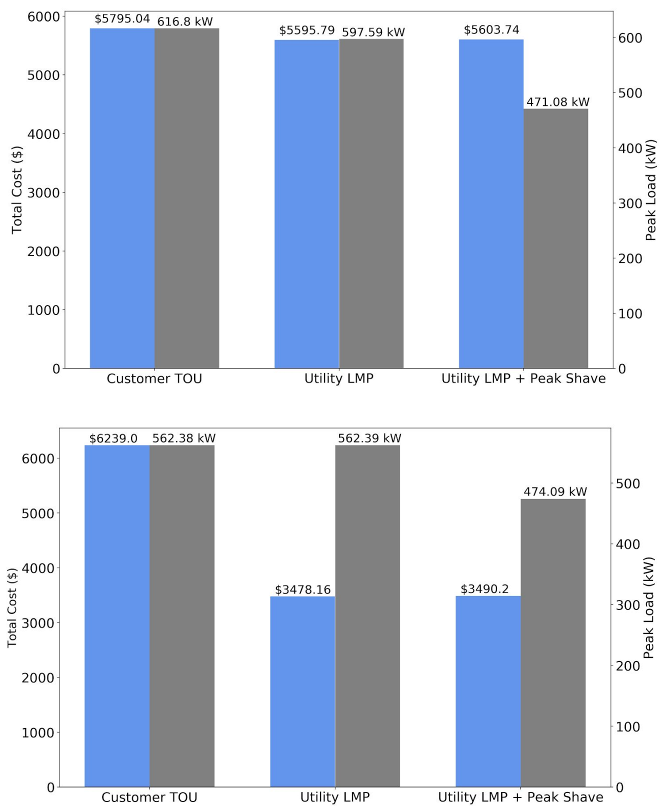

Fig. 6 Total cost to the utility to serve all of its customers and its peak load using (top) storage only and (bottom) both solar power and storage. For both, three scenarios of battery optimization are modeled: (1) Each individual customer has a battery to manage their demand charge on a TOU rate, (2) The utility is given a battery equal to the size of its customers' aggregated batteries that is operated for LMP arbitrage, and (3) The utility operates its battery for both LMP arbitrage and aggregate peak demand shaving.