Load Management for Price-based Demand Response Scheduling — a Block Scheduling Model

Ding Li, Sudharman K. Jayaweera, Olga Lavrova and Ramiro Jordan
Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87113, USA
Email: {liding, jayaweera, olavrova, rjordan}@ece.unm.edu

***Abstract*—Demand Response (DR) plays an important role in electricity market design, in both reducing utility's investment on peak generation and improving electricity bill savings and incentive payments earned by customers. Improved resource-efficiency of electricity production is achieved by closer alignment of electricity pricing information with energy consumption behaviors. In this paper, a block scheduling model of load management for price-based *Demand Response* is presented under two different real-time pricing schemes: linear pricing scheme and threshold pricing scheme. For linear pricing, the problem is formulated as a convex optimization problem and the optimal demand response profile is given as a two-dimensional *water-filling solution* either with flat water levels or different water levels for different customers. From the perspectives of the customers as a whole or as selfish individuals, the demand-response computations lead to centralized or distributed optimizations, respectively. A trade-off strategy which attempts to balance these competing objectives is also provided. This trade-off strategy divides customers into local groups within which group-wise distributed optimization is performed to improve the overall performance so that the Price of Anarchy (PoA) is reduced. For threshold pricing, which might be more applicable in certain scenarios, detailed characterization of different optimal load profiles are given assuming a discrete load unit model. A search algorithm is also proposed to find the optimal load profiles for both constant and dynamic pricing threshold scenarios. The effect of dynamic pricing threshold on customers' electricity consumption behaviors is highlighted.**

***Index Terms*—Demand response (DR), real-time pricing, load management, two-dimensional water-filling, block scheduling.**

I. Introduction

In most current electricity markets, fixed pricing schemes with constant rates are being widely used. Customers face retail electricity prices that are flat over months or even years [1]. A problem with fixed pricing schemes is the disconnection between short-term marginal electricity production costs and retail rates paid by customers, which leads to inefficient overall resource usage. Due to lack of information on generation costs, electricity consumption behavior of customers may not adjust to supply-side conditions. Thus fixed constant pricing results in suboptimal customer behavior as well as higher electricity costs than they would otherwise be in an optimally efficient system [2].

There is a growing consensus that Demand Response (DR) can play an important role in market design [3]. Lack of DR has been shown to be a major contributing factor for energy-market meltdowns [4]. In [1], for example, DR is defined as *"Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized."* DR not only reduces the capacity investments in peak generation units to serve occasional heightened demand, but also provides short-term reliability benefits as it can offer load relief to resolve system and local capacity constraints. There are two basic demand response options: Price-based demand response and incentive-based demand response. Price-based demand response includes real-time pricing (RTP), critical-peak pricing (CPP), and time-of-use (TOU) rates. Customers can respond to the price structure with changes in energy use, reducing their electricity bills if they adjust the timing of their electricity usage to take advantage of lower-priced periods and avoid consuming when prices are higher [1]. Incentive-based demand response schemes pay participants to reduce their loads at times requested by the program sponsor, triggered either by a grid reliability problem or high electricity prices. DR programs typically specify a method for establishing customers baseline energy consumption level below which demand reductions are not allowed. In power systems, the energy requests that customers send to utility consist of two parts: nonflexible load request and flexible load request [5]. The nonflexible part is the minimum amount of energy that utility needs to provide at a specific time. The flexible part can be reallocated over time according to a certain load management strategy. For any load management strategy there are two common primary goals: peak load shaving and load profile flattening. Under real-time pricing, the electricity price is determined by real time load information.

This paper presents a block scheduling model of load management for price-based demand response scheduling. In this model, the size of the time block is set to be small enough so that all load shifting within the time block can be considered as cost free and acceptable to customers. The solution to this block processing problem can then be the basis for implementations of arbitrarily long scheduling periods. Two types of real-time pricing schemes, linear pricing and threshold pricing, are discussed in this paper. We consider optimal demand-response when customers cooperate as a group as well as when each customer is only interested in minimizing its own cost. Naturally these two scenarios, as shown to lead to centralized and distributed optimizations.

The rest of this paper is organized as follows: In Section II, the system model and the problem formulation for block

scheduling are presented. The block scheduling for linear pricing is presented and solved in Section III. Water-filling solutions for both centralized and distributed scenarios are analyzed and compared. Section IV presents block scheduling model and its solutions for threshold pricing scheme assuming discrete load units. Simulation results of constant and dynamic threshold scenarios are contained in Section V. The conclusions from this study are given in Section VI.

II. System Model and Problem Formation

We assume that customers send their electric energy requests to the utility at the beginning of each processing time block which consists of T time intervals, for $t = 1, 2, \ldots, T$. For each time t, the load requests consist of two parts: a *nonflexible part* which has to be satisfied at the specific time, and a *flexible part* for which certain amount of reallocation within the current block is acceptable. Different customers may have different weights on the two parts of the load request. For example, hospitals might have a high demand of nonflexible loads while a load request from a household could have a significant flexible portion.

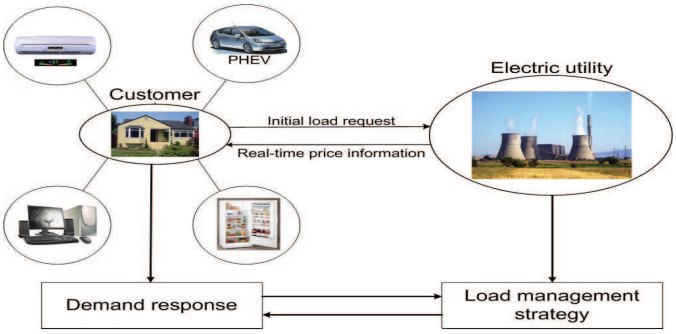

Fig. 1. System model: Communications between the utility and customers.

We denote by $l^F_{t,k}$ and $l^N_{t,k}$ respectively the flexible and nonflexible loads requested by customer k, for $k = 1, 2, \ldots, U$, for time interval t, where U is the total number of customers in the market. Denote by l_F the total requested flexible loads, so that $l_F = \sum_{k=1}^{U} \sum_{t=1}^{T} l^F_{t,k}$. The amount of flexible load of customer k that the utility will schedule to satisfy at time t after load reallocation is denoted by $x_{t,k}$, where $x_{t,k} \geq 0$ and $\sum_{k=1}^{U} \sum_{t=1}^{T} x_{t,k} = l_F$.

Customers send their initial load requests $l^F_{t,k}$'s and $l^N_{t,k}$'s for the current processing time block to the utility. The utility then optimally reallocate flexible loads from all customers at each time instant $l^F_t = \sum_{k=1}^{U} l^F_{t,k}$, while supporting all nonflexible load requests at each time interval $l^N_t = \sum_{k=1}^{U} l^N_{t,k}$ to minimize the overall generation cost. Based on the minimum-cost generation schedule, the utility may determine a pricing scheme for all customers. Note that such price determinations may not be performed at each processing block but in a longer-term basis, which is beyond the scope of this paper. The interaction between customer and utility is shown in Fig. 1.

Based on the pricing information, the customers will attempt to optimize their load scheduling. Since electricity consumers usually show certain clustering effects, for example, a community with a large number of customers might consider coordinating among themselves the consumption behaviors so that instead of minimizing individual consumption cost of each customer, minimizing the total consumption cost of the whole community becomes the goal.

III. Block Scheduling for Linear Pricing Scheme

Under linear pricing, the unit price for customer k at time interval t is given by $P_{t,k} = K_p(l^N_{t,k} + x_{t,k})$ \$/kWh, where K_p is a positive price scaling factor. The cost of customer k at time interval t is then $C_{t,k} = P_{t,k}(l^N_{t,k} + x_{t,k}) = K(l^N_{t,k} + x_{t,k})^2$, and the cost of customer k over the processing time block is $C_k = \sum_{t=1}^{T} C_{t,k}$. Since electricity consumers usually show certain clustering characteristics, there are two reasonable approaches to minimizing the electricity consumption costs of customers. One is the centralized optimization, in which the objective is to minimize the total cost of all customers on the market as a whole, which can be interpreted as the *social optimal*. However, from the perspective of an individual customer, the k-th customer may be interested in minimizing its own cost $C_k = \sum_{t=1}^{T} K(l^N_{t,k} + x_{t,k})^2$ during the processing time block. This clearly leads to a distributed optimization problem.

A. Centralized Block Scheduling

In the following we first consider the centralized optimization problem to minimize the total cost of all U users in the market:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & C(\mathbf{x}) = K \sum_{k=1}^{U} \sum_{t=1}^{T} (x_{t,k} + l^N_{t,k})^2 \qquad (1) \\ \text{subject to} \quad & -x_{t,k} \leq 0, \quad t = 1, 2, \ldots, T, k = 1, 2, \ldots, U, \\ & \sum_{k=1}^{U} \sum_{t=1}^{T} x_{t,k} - l_F = 0. \end{aligned}$$

The objective function (1) is quadratic and constraints are linear. Thus the above optimization problem is convex. The Lagrangian associated with the primal problem is $L(\mathbf{x}, \lambda, v) = K \sum_{k=1}^{U} \sum_{t=1}^{T} (x^2_{t,k} + 2l^N_{t,k} x_{t,k} + (l^N_{t,k})^2) - \sum_{k=1}^{U} \sum_{t=1}^{T} \lambda_{t,k} x_{t,k} + v(\sum_{k=1}^{U} \sum_{t=1}^{T} x_{t,k} - l_F)$, where $\lambda_{t,k}$'s are the Lagrange multipliers associated with inequality constraints and v is the Lagrange multiplier associated with the equality constraint. With the primal problem being convex, the optimal primal and dual solutions are achieved if and only if the following Karush-Kuhn-Tucker (KKT) conditions are held [6]:

$$\begin{aligned} \sum_{k=1}^{U} \sum_{t=1}^{T} x^*_{t,k} - l_F = 0 \qquad (2) \\ -x^*_{t,k} \leq 0, \forall t, k \quad , \\ \lambda^*_{t,k} \geq 0, \forall t, k \quad , \\ -\lambda^*_{t,k} x^*_{t,k} = 0, \forall t, k \quad , \\ \frac{\partial L(\mathbf{x}^*, \lambda^*, v^*)}{\partial x^*_{t,k}} = 2Kx^*_{t,k} + 2Kl^N_{t,k} - \lambda^*_{t,k} + v^* = 0, \forall t, k \quad . \end{aligned}$$

By solving the above set of equations, it can be shown that the optimal load profile is given by

$$x^*_{t,k} = \begin{cases} 0 & \text{if } w^* < l^N_{t,k} \\ w^* - l^N_{t,k} & \text{if } w^* \geq l^N_{t,k} \end{cases} , \\ = [w^* - l^N_{t,k}]^+ \quad . \tag{3}$$

where $[a]^+ = \max(0, a)$ and w^* is the unique solution to

$$\sum_{k=1}^{U} \sum_{t=1}^{T} \max(0, w^* - l^N_{t,k}) = l_F \quad . \tag{4}$$

Note that, the left hand side of (4) is a piecewise-linear increasing function of w^*, with breakpoints at $l^N_{t,k}$, ensuring the uniqueness of its solution. In general, there may not be a closed form solution for w^*, requiring numerical computation. However, this solution structure (3) is well known in information theory and is referred to as the *water-filling solution* [7]: We can think of $l^N_{t,k}$ as the height of the bottom level at location (t, k) on a two-dimensional plane. Starting from zero, we allocate flexible loads to the location with the lowest nonflexible load. As flexible loads increase, some of the them are put into locations with higher nonflexible loads. We continue to allocate flexible loads in this way until we have allocated all of l_F. At this time, the height of the flat flexible load level would be the solution w^* of (3). This process is similar to the way in which water distributes itself in a vessel. The depth of water at location (t, k) is then the optimal value $x^*_{t,k}$. Figure 2 shows an example of the two dimensional water-filling solution of the above optimization problem. In our simulation setup, for each time interval t, the flexible and nonflexible loads are generated according to uniform distributions $U(0, u_t)$ and u_t's are parameters that we can change.

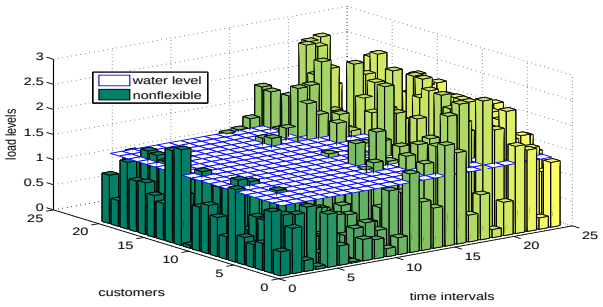

Fig. 2. Two-dimensional water-filling solution that indicates how loads from different customers are scheduled over the processing time block.

B. Distributed Block Scheduling

While above centralized optimization minimizes the total cost for all customers in a global manner, this strategy might not be optimal for all individual customers. Let us denote by $C^* = \min_{t,k} \sum_{k=1}^{U} \sum_{t=1}^{T} K(l^N_{t,k} + x_{t,k})^2$ and $\tilde{C}^* = \sum_{k=1}^{U} \tilde{C}^*_k$ the minimum costs from centralized and distributed optimizations. Due to space limitations, in the following we assume that the total load requests from all customers at each time instant is below the maximum capacity of the utility, so that the individual customer optimization problems can be decoupled. With this assumption the distributed problem for customer k becomes

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \tilde{C}^*_k = \min_{x_{t,k}} \sum_{t=1}^{T} K(l^N_{t,k} + x_{t,k})^2 \quad , \\ \text{subject to} \quad & -x_{t,k} \leq 0, \quad t = 1, 2, \ldots, T \quad , \\ & \sum_{t=1}^{T} x_{t,k} - l_F = 0 \quad . \end{aligned} \tag{5}$$

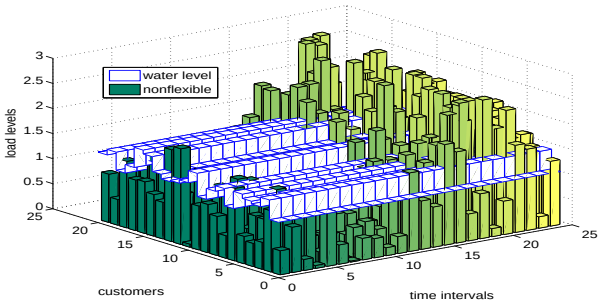

Fig. 3. Two-dimensional water-filling with different water levels for different customers.

By following a similar method to that above, we can show that the solution of the distributed optimization problem is also given by a two dimensional water-filling result but with different water levels for different customers, as shown in Fig. 3 (for the same set of initial load requests that generates Fig. 2). It should be noted that the optimal solutions for different customers in the distributed optimization problem (5) will be coupled if the total loads of all customers in any time interval were to violate the maximum capacity of the utility . In this case the distributed optimization problem will lead to a non-cooperative game among customers. As mentioned above, for simplicity in this paper we do not consider this situation. In general, the total cost is increased by going from the centralized to distributed optimization, so that $C^* \leq \tilde{C}^*$. In game theory, this degradation, which is caused by the selfish behavior of customers, is referred to as the Price of Anarchy (PoA) compared to the global optimal [8]. We may characterize this inefficiency of the distributed solution by $\frac{\tilde{C}^* - C^*}{C^*}$, which is the normalized extra cost of opting for distributed optimization over centralized optimization. If we sum up optimal load requests over all users in Figs. 2 and 3, we get Figs. 4 and 5 showing the overall load requests for each time interval. Fig. 6 shows the consumption cost comparison of the two optimization schemes for two sets of initial load requests. Note that, although for this particular example, the price of anarchy seem to be small, it could change with different initial load requests.

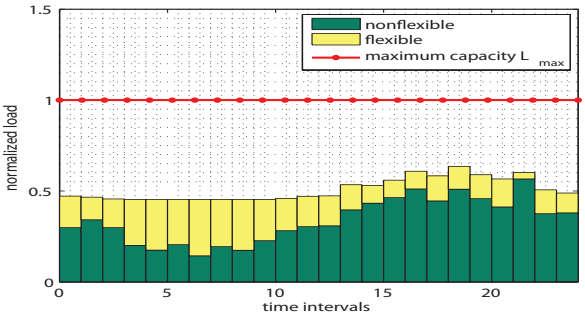

Fig. 4. The centralized optimal load profile over time intervals.

The mismatch in the centralized vs. distributed optimization goals makes it necessary to consider a tradeoff between the two optimization objectives. A tradeoff optimization scheme that balances both the total cost and the individual cost would be welcome from both whole-market and individual perspectives. One way of doing this is to divide customers into distributed groups. For each customer group (instead of each individual customer) a local water-filling solution can be obtained within that group. By doing this the cost of each group is minimized and different groups could have different water levels. Figure 7 shows the normalized extra cost as the customer group size increases from 1 (corresponding to customer-wise distributed optimization) to 20 (corresponding to centralized optimization, assuming we have a total of 20 customers). It can be observed from Fig. 7 that the price of anarchy decreases monotonically as the group size increases.

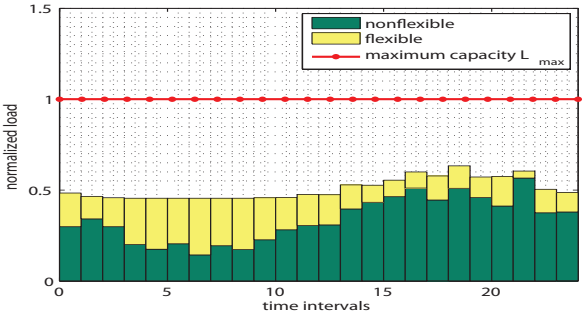

Fig. 5. The distributed optimal load profile over time intervals.

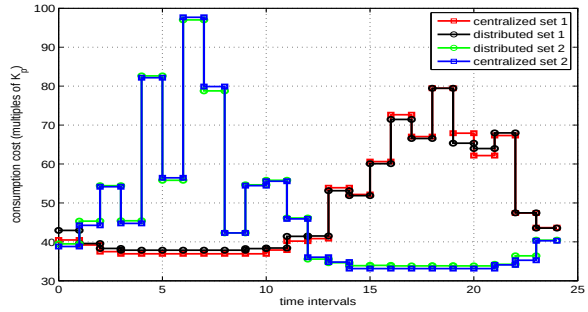

Fig. 6. Comparison of customers' consumption costs under centralized and distributed optimization schemes. Set 1 is based on the same initial load request information set that generates the two-dimensional water-filling result. Set 2 is based on another set of initial load request information for comparison. The two sets initial loads were generated according to different distributions.

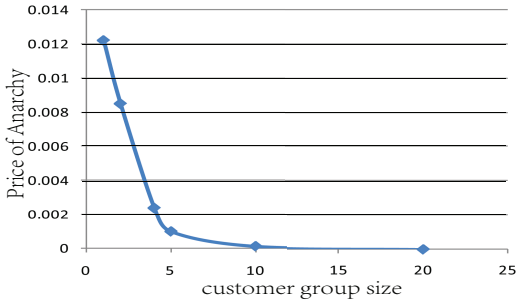

Fig. 7. Price of anarchy vs customer group size.

IV. BLOCK SCHEDULING FOR THRESHOLD PRICING SCHEME

A. Problem Formation

Under certain assumptions, threshold pricing schemes might be more realistic in practice compared to the linear pricing scheme [9]. Given initial load requests from customers, for a small enough load tuple Δl, all loads $\{l^F_{t,k}\}$'s and $\{l^N_{t,k}\}$'s can be represented as multiples of Δl. We label the m-th load tuple of customer k at time interval t by $e^m_{t,k}$, for $t = 1, 2, \ldots, T$, $k = 1, 2, \ldots, U$ and $m = 1, 2, \ldots, M_{t,k}$, where $M_{t,k} = \frac{l^F_{t,k}+l^N_{t,k}}{\Delta l}$. For customer k at time interval t, there is a threshold $L_{t,k}$ which can also be represented as multiples of Δl, say $L_{t,k} = \tilde{M}_{t,k}\Delta l$. We denote the price level of $e^m_{t,k}$ by $n^m_{t,k}$. The price level $n^m_{t,k}$ for tuple $e^m_{t,k}$ is given by

$$n^m_{t,k} = \begin{cases} 0 & \text{if } m \le \tilde{M}_{t,k} \\ m - \tilde{M}_{t,k} & \text{if } m > \tilde{M}_{t,k} \end{cases}, \\ = [m - \tilde{M}_{t,k}]^+ \quad .$$

Denote by $L_{\max}$ the maximum load capacity of utility and $L_{\max} = M_{\max}\Delta l$. Then we have that $n^m_{t,k} \le M_{\max}$ for $\forall t, m$. Fig. 8 is a example of a slice for a certain customer.

The threshold pricing scheme can be described as follows: For each customer k at time interval t, a constant basic unit price P_0 (\$/kWh) applies for all $e^m_{t,k}$'s below threshold $L_{t,k}$. The unit price for the m-th load tuple $e^m_{t,k}$ of customer k at time interval t above threshold $L_{t,k}$ is given by $P^m_{t,k} = P_0 + n^m_{t,k}\Delta P$, where ΔP (\$/kWh) is the increment in unit price. Assume that $x_{t,k}$ is the flexible load request that utility will schedule to satisfy for customer k at time interval t. The consumption cost of customer k at time interval t is given by $C_{t,k} = P_0(l^N_{t,k} + x_{t,k})$, if $l^N_{t,k} + x_{t,k} \le L_{t,k}$; and $C_{t,k} = P_0 L_{t,k} + \sum_{n^m_{t,k}=1}^{M_{t,k}-\tilde{M}_{t,k}} (P_0 + n^m_{t,k}\Delta P)\Delta l$, if $l^N_{t,k} + x_{t,k} > L_{t,k}$. To minimize the total consumption cost of all customers (centralized optimization), we need to minimize $C_{total} = \sum_{t=1}^{T}\sum_{k=1}^{U} C_{t,k}$.

B. Solution Searching Algorithm

We define the *vacancy* (shown in Fig. 8) value for customer k at time interval t as $v_{t,k} = M_{t,k} + 1$. Based on the observation that the only way of decreasing the generation cost is to shift some $e^m_{t,k}$'s from higher price levels to vacancies with lower price levels, we have the following proposition:

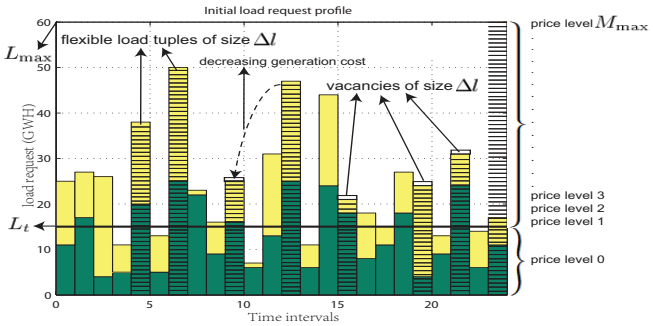

Fig. 8. An illustration of threshold pricing scheme with discrete load tuples and vacancies. Please note 1) This Figure is a one-slice example for a certain customer. 2) In this Figure the threshold $L_{t,k}$ is set to be constant over all time intervals which could be dynamic in general. 3) The discrete load tuples and vacancies are shown only for several time intervals.

Proposition 1: In the block scheduling for threshold pricing scheme, the total consumption cost of all customers is minimized if and only if $\max_{t,k} M_{t,k} \leq \min_{t,k} v_{t,k}$.

Proof: See Appendix 1. ■

For the threshold pricing scheme, the optimal load profiles are of two categories according to whether the increment in unit price applies or not.

1) No increment in unit price applies: In the initial load profile, if all $e^m_{t,k}$'s above the threshold can be allocated into vacancies below the threshold, then all the flexible loads in the optimal load profile will be in price level 0. All optimal load profiles that satisfy this property are considered as being optimal. (i.e. The optimal solution is not unique.)

2) Increment in unit price applies: In the initial load profile, if the $e^m_{t,k}$'s above the threshold are more than the vacancies below the threshold, then some flexible load tuples will cause price increments at some time intervals in the optimal load profile.

A slight variation of proposition *1* tells more about the optimal load profile in this case: Noticing that $\max_{t,k} M_{t,k} \leq \min_{t,k} v_{t,k} \Leftrightarrow \max_{t,k} v_{t,k} \leq \min_{t,k} v_{t,k} + 1$, the optimal load profile is flat in a Δl-flat sense. By "Δl-flat" we mean that $\max_{(t_1,k_1),(t_2,k_2)} |(l^N_{t_1,k_1} + x_{t_1,k_1}) - (l^N_{t_2,k_2} + x_{t_2,k_2})| \leq \Delta l$. As $\Delta l \to 0$, we have $\max_{t,k} v_{t,k} = \min_{t,k} v_{t,k}$. Thus, the optimal load profile again converges to a two-dimensional water-filling result. Hence, the optimization problem to minimize the utility generation cost can be stated as follows: *Given initial load request information: price levels $n^m_{t,k}$'s ($M_{t,k}$'s) and vacancy levels $v_{t,k}$'s for $t = 1, 2, \ldots, T$ with threshold level $L_{t,k}$, by doing a load reallocation which is also an updating process of $n^m_{t,k}$'s and $v_{t,k}$'s, we can minimize the total consumption cost of all time intervals if and only if the achieved load profile (possibly not unique) with $M^*_{t,k}$'s and $v^*_{t,k}$'s satisfy the optimization condition:* $\max_t M^*_{t,k} \leq \min_{t,k} v^*_{t,k}$.

To find the optimal load profile, we may start from $\min_{t,k} v_{t,k}$ and search upward to $\max_{t,k} v_{t,k}$ until the testing level $v^*_{t,k}$ satisfies the following conditions:

1. In the initial load request profile, the number of $e^m_{t,k}$'s above the testing level $v^*_{t,k}$ is strictly less than the number of all vacancies on and below testing level $v^*_{t,k}$.

2. In the initial load request profile, the number of $e^m_{t,k}$'s on and above the testing level $v^*_{t,k}$ is equal to or greater than the number of all vacancies below testing level $v^*_{t,k}$.

Thus the above centralized optimization can be solved again by two-dimensional water-filling result but with the water level flat in "Δl"-flat sense. The solution for distributed optimization can be characterized similarly to the two-dimensional water-level solution with different water levels for different customers, and the details are omitted due space limit.

V. Simulation Results

In the following, we simulated the proposed load management strategy for an electric utility with the threshold generation cost model during a period of $T = 24$ hours with each time interval 1 hour, and for a electricity market of 20 customes. Load tuple size is set to be $\Delta l = 1$ GWh. The utility has a maximum capacity $L_{\max} = 60$ GWh. The threshold was set to be $L_t = 15$ GWh, $\forall t$. For each time interval, the flexible and nonflexible loads were generated according to uniform distributions $U(0, u_t)$ and u_t's were adjustable. As $L_{\max}$ is normalized to 1, all loads can be expressed as certain percentages of maximum utility capacity.

Scenario *1*: Figure 9 shows that in the initial load request profile, all flexible loads can be reallocated to vacancies under the threshold and all nonflexible are below the threshold. Thus a constant unit generation cost applies for all load tuples.

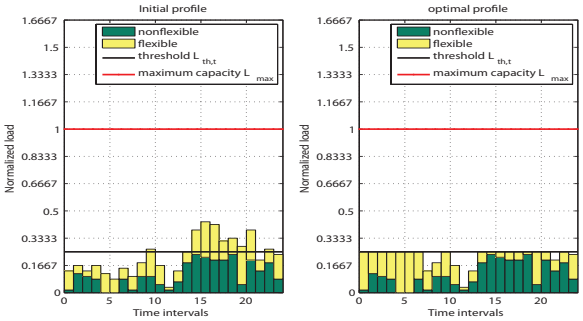

Fig. 9. Constant unit generation cost applies for all load tuples.

Scenario *2*: Figure 10 shows that in the initial load request profile, the number of flexible loads above the threshold ($L_t = 15$ GWh) is greater than the number of vacancies below the threshold. Thus, in the optimal load profile, extra increment in unit generation cost will apply. However, since the nonflexible loads are not high, the optimal load profile keep "Δl"-flat.

Scenario *3*: Figure. 11 presents the situation in which the nonflexible loads are very high at some time intervals. In this case, the optimal load profile is not "Δl"-flat.

Recall that the two principal goals of DR are peak-load shaving and load-profile flattening. Thus based on the principle that heavy load hours corresponds to higher unit price and vise versa, we may set lower threshold $L_{t,k}$'s for heavy load time intervals (for example, during day and evening hours) and higher threshold $L_{t,k}$'s for lighter load time intervals (for example, during midnight hours) [10]. Such dynamic

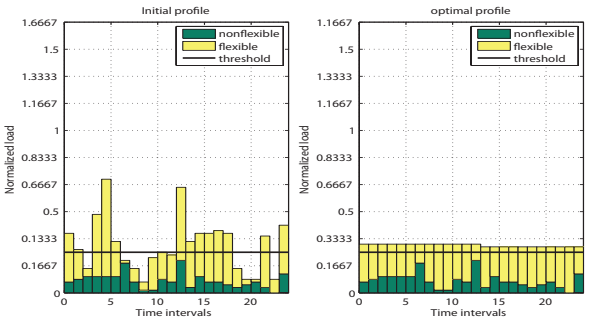

Fig. 10. "Δl"-flat optimal profile with increment in unit generation cost.

price-thresholding can naturally incentivize the customers to schedule their demand-responses in a way that will lead to peak-load shaving and load profile flattenning. An example of the dynamic threshold pricing is shown in Fig. 12. From the perspective of the customers, dynamic threshold model can have a great influence on customers' electricity consumption behaviors: Customers are encouraged to make use of off-peak time periods. Indeed, Fig. 12 (b) shows how customers' optimal response shifts most of their flexible loads to off-peak hours.

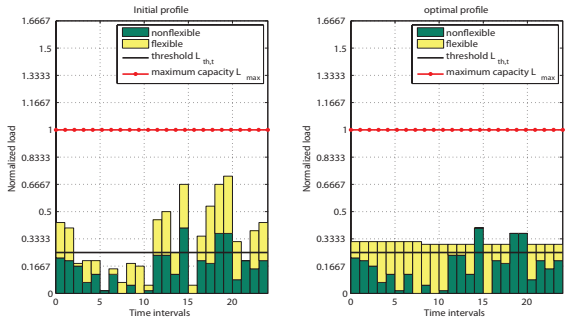

Fig. 11. Optimal profile is not "Δl"-flat and with high nonflexible loads.

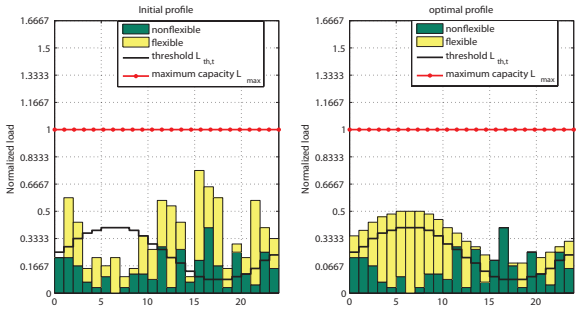

Fig. 12. dynamic threshold scenario.

VI. Conclusion

In this paper, a block scheduling model of load management for price-based Demand Response was presented under two real-time pricing schemes: For linear pricing, the problem was formed as a convex optimization problem and the optimal load profile was obtained as a two dimensional water-filling with a flat water level in the case of centralized optimization. The distributed optimization problem was also solved by a two-dimensional water-filling result but with different water levels for different customers. While the centralized optimization minimized the total *social cost* of all customers globally, the distributed optimization minimized the cost for each individual customers. Further analysis on Price of Anarchy (PoA) and a tradeoff strategy that balances the two optimization goals were also provided. For threshold pricing, which may be more realistic in practice, we presented and solved the discrete load unit model and proposed a search algorithm to obtain the optimal solution for centralized optimization. In the end, a dynamic threshold pricing scheme was proposed in order to encourage the customers adapt an optimal demand-response profile that will naturally lead to peak-load shaving and load profile flattening, and its effectiveness was shown with a representative numerical example.

Appendix

A. Proof of Proposition 1

Proof: Necessity: Assume we have minimized the generation cost but there are some time-customer pairs (t_1, k_1) and (t_2, k_2) such that $M_{t_1,k_1} > v_{t_2,k_2}$, then by shifting the load tuple $e_{t_1,k_1}^{M_{t_1}}$ from price level M_{t_1,k_1} to the vacancy v_{t_2,k_2} we can further decrease the cost, this contradicts the minimum cost assumption.

Sufficiency: If the price cost function is not minimized, then there exists some load shifting strategy that enable us to further decrease the generation cost. Thus there exists time-customer pairs (t_3, k_3) and (t_4, k_4) such that $M_{t_3,k_3} > v_{t_4,k_4}$. Thus if $\max_{t,k} M_{t,k} \leq \min_{t,k} v_{t,k}$ holds, there will be no load shifting strategy that could further decrease the cost function, meaning the current generation cost is the minimum. ■

References

[1] "Benefits of demand response in electricity markets and recommendations for achieving them," Tech. Rep., U.S. Department of Energy, Feb. 2006.

[2] N. Lu, D. P. Chassin, and S. E. Widergren, "Modeling uncertainties in aggregated thermostatically controlled loads using a state queueing model," *IEEE Trans. Power Syst.*, vol. 20, pp. 725–733, May 2005.

[3] S. H. S. Han and K. Sezaki, "Development of an optimal vehicle-to-grid aggregator for frequency regulation," *IEEE Trans. Smart Grid*, vol. 1, pp. 65–72, Apr. 2010.

[4] F. Rahimi and A. Ipakchi, "Demand response as a market resource under the smart grid paradigm," *IEEE Trans. Smart Grid*, vol. 1, pp. 82–88, Jun. 2010.

[5] D. P. Chassin, D. J. Hammerstrom, and J. G. DeSteese, "The Pacific Northwest demand response market demonstration," *IEEE PES GM*, pp. 1 – 6, Jul. 2008.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2008.

[7] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & sons, 2008.

[8] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[9] E. Celebi and J. D. Fuller, "A model for efficient consumer pricing schemes in electricity markets," *IEEE Trans. Power Syst.*, vol. 22, pp. 60 – 67, Feb. 2007.

[10] K. Mets, T. Verschueren, W. Haerick, C. Develder, and F. D. Turck, "Optimizing smart energy control strategies for plug-in hybrid electric vehicle charging," in *IEEE/IFIP NOMS Wksps*, (Osaka, Japan), pp. 293 – 299, Apr. 2010.