18. Smart Grid Load Balancing

金力 Li Jin
li.jin@sjtu.edu.cn
上海交通大学密西根学院
Shanghai Jiao Tong University UM Joint Institute

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Outline

- Background
- System model
- Stage 1: Power load balancing
- Stage 2: Data workload distribution

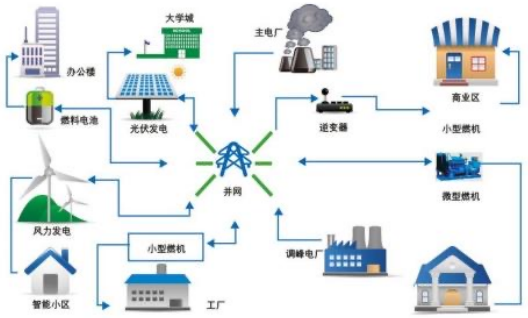
Ref: Hao W , Huang J , Lin X , et al. Exploring smart grid and data center interactions for electric power load balancing[J]. ACM Sigmetrics Performance Evaluation Review, 2014, 41(3):89-94.

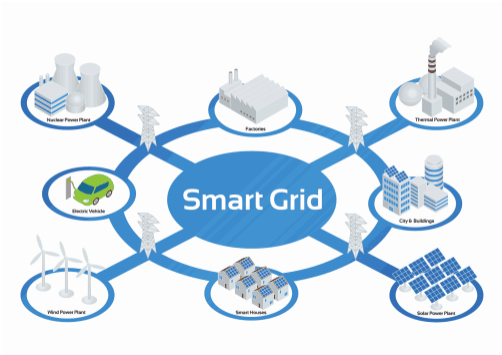
Introduction

https://haokan.baidu.com/v?pd=wisenatural&vid=2981825073698562159

Smart grid

- Smart grid (SG): an enhancement of the 20th century power grid.
- The traditional power grids are generally used to carry power from a few central generators to a large number of users or customers.
- In contrast, the SG uses two-way flows of electricity and information to create an automated and distributed advanced energy delivery network.





Background

- Scenario for load balancing:
  - System operator wants to balance the load on the smart grid over time to reduce operational cost and efficiency.
  - This would require incentivizing consumers to collaborate.
- In this lecture, we consider a special pair of consumer and supplier:
  - Consumer: data centers that can collaboratively use power
  - Supplier: smart grid operator that charges the consumer

Background

- With the fast development of cloud computing services, it is common for a cloud service provider (e.g., Google, Microsoft, or Amazon) to build multiple geographically dispersed large data centers across the country.
- Each data center may include hundreds of thousands of servers, storage equipment, cooling facilities, and power transformers.
- The energy consumption and cost of a single data center hence can be very significant.



Background

- Due to the enormous energy consumption, data centers are expected to have a great influence on the operation of power grid.
- However, conventional power control schemes focused on the study of data centers' energy minimization and cost minimization, without detailed analysis of the two-way impact on the power grid.
- For example, when large data centers suddenly increase their energy consumption in low price regions, they may overload the grid, which can cause various problems, such as a regional or major blackout.

Background

- Smart grid is equipped with advanced communication technologies, and is able to integrate energy suppliers and users more effectively through two-way communications.
- For example, an energy supplier can send real-time price information to the smart meters of users, and the users can change energy consumption in response to the price changes.
- This can effectively coordinate demand with supply, and hence avoid the danger of power overload.

Outline

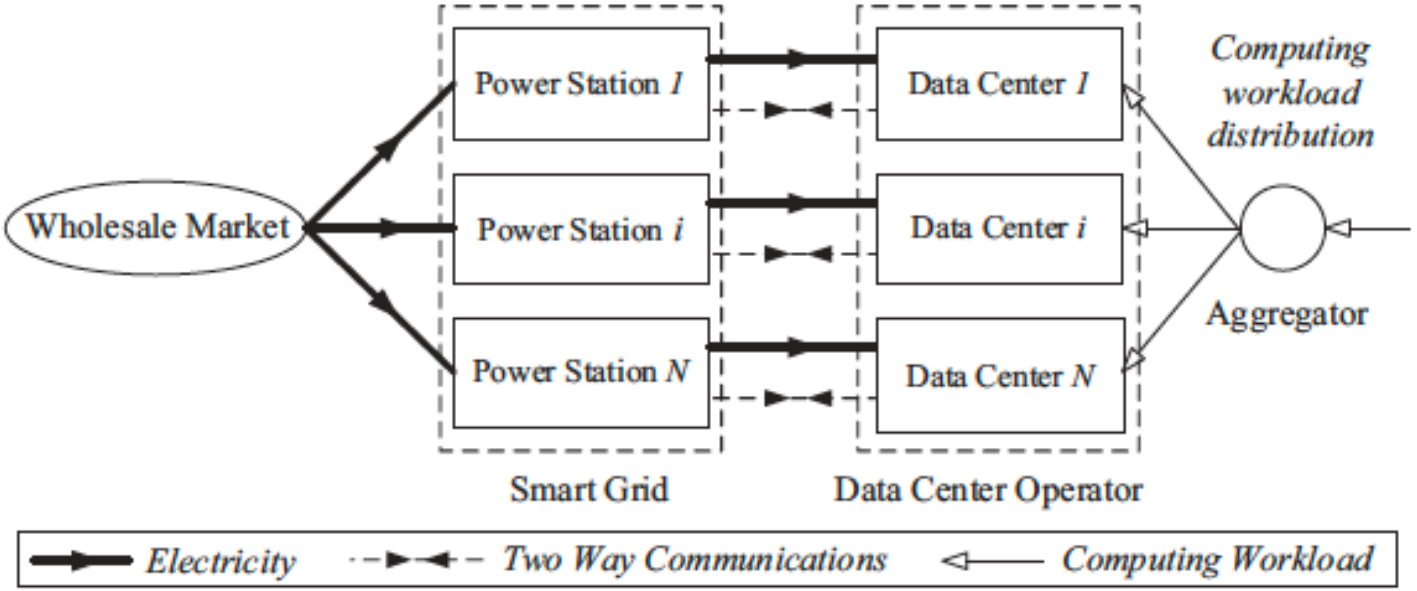
- Background
- System model
- Stage 1: Power load balancing
- Stage 2: Data workload distribution

Discrete-time model

- Time $t = 1,2, \dots, T$
- Set of geographically dispersed data centers $\mathcal{N} = \{1,2, \dots, N\}$
- Data center i has M_i homogeneous servers
- Not all the servers will be turned on during each time slot
- Each data center is powered by a separate power substation
- A traffic aggregator distributes total incoming computing workload L^t at time t to various data centers

Cost and price

- No cost for distributing workload among data centers
- Quality of service must be satisfied



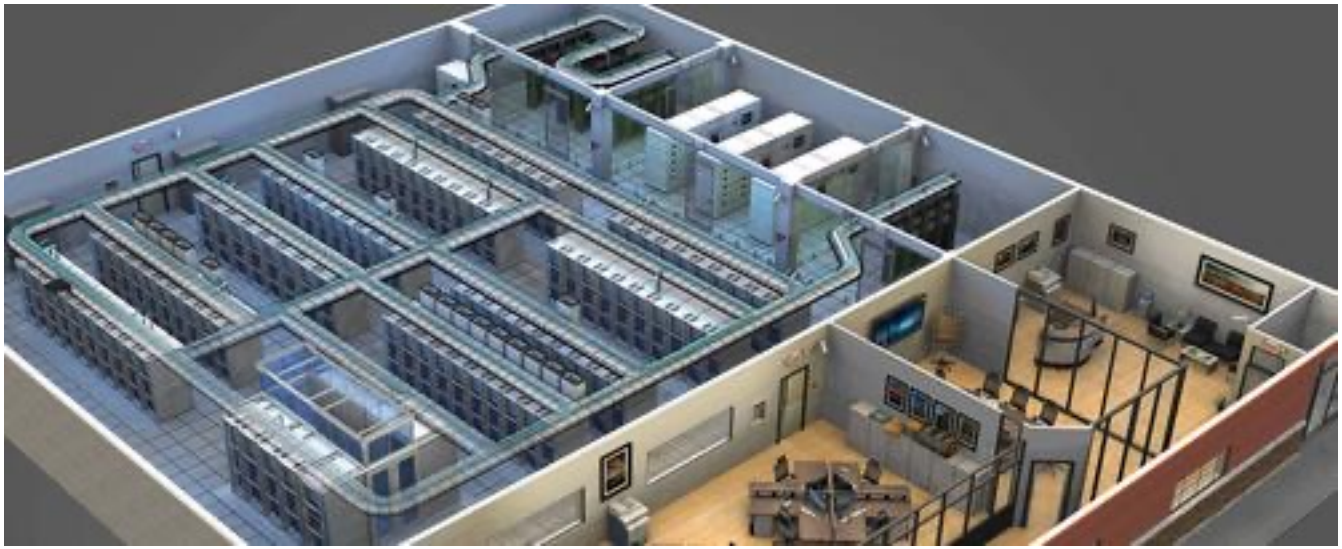
- Data centers (consumers/customers) cannot influence the electricity market price.

Two-stage decision making

- Two decision makers: smart grid (supplier) & data centers (consumer)
- Stage 1: smart grid decision
  - sets the charging threshold s_i^t
  - i.e., the "electricity tariff"
- Stage 2: data center decision
  - data centers belong to the same cloud computing provider
  - cooperate with each other to minimize the total energy cost
  - done by determining the computing workload allocation λ_i^t and the number of active servers x_i^t in each data center i.

Objective

- Get familiar with state-of-the-art smart grid technologies
- Experience formulation and simulation of practically relevant decision-making problems



Outline

- Background
- System model
- Stage 1: Power load balancing
- Stage 2: Data workload distribution

Four-stage formulation

- Data:
  - Substation capacity C_i
  - Background demand B_i^t
  - Energy consumption E_i^t
- Decision variable: charging threshold s_i^t
- Constraints:
  - All demand must be served
  - Price must not exceed marginal price
- Objective
  - Minimize overall load ratio (load balancing)
  - Minimize revenue loss (due to discount)

Load balancing

- Overload has been a major problem in traditional power grid, as it is often difficult to prevent users in a specific region to generate an excessive power demand.
- With the emergence of smart grid, it is possible for the grid operator to communicate with the users and try to incentivize the users to shift load from heavy load regions to light load regions.
- Smart grid optimizes dynamic prices by setting the charging thresholds

$$s = \left[s_i^t\right]_{i,t}$$

- to achieve the power load balancing.

Electric load ratio

- To measure the power load levels in different locations, we define the electric load ratio in location i at time t as

$$r_i^t(s) = \frac{E_i^t(s) + B_i^t}{C_i}$$

- $E_i^t(s)$ = energy consumption of data center i; determined by data centers
- B_i^t = background power load
- C_i = capacity of power substation i
- If the load ratio is high, then the grid operator can better utilize the installed capacity.
- However, high load ratio may also increase the risk of overloading the power system.

Background power

- $q_i^t = C_i - B_i^t$ (maximal available power supply to data center i at time t)
- Usually, conventional electricity users have inflexible usage patterns and their demands can be inelastic.
- However, data centers have a very good flexibility in energy consumption because of the ability of routing workload among different locations.
- Therefore, we denote the total energy usage of conventional users other than data centers as the background energy load B_t^i.
- Assume the background energy load can be predicted accurately as a constant.

Electric load index

- Electric load index (ELI)

$$ELI = \sum_{t=1}^{T} \sum_{i=1}^{N} \left(r_i^t(s) \right)^2 C_i$$

- Motivated by the index measurement techniques used for feeder load balancing in distribution system.
- ELI not only measures the overall load ratio but also gives different weights based on the capacities in different locations.
- The smart grid aims at balancing the load ratio $r_t^i(s)$ at all locations and all time slots by minimizing the ELI.
- Minimizing ELI results in balancing the electric load across all the locations.

Pricing

- To balance the electric power load, the grid operator provides rewards (price discounts) to incentivize users to shift their electricity usage.
- As power grid is usually a regulated industry, the regulator sets the upper bound and lower bound to the unit price π_i^t [e.g. RMB per unit kW].
- First, the unit price π_i^t should be non-negative, and thus $\pi_i^t \geq 0$.
- Second, smart grid only provides price discount (instead of charging extra) to data centers, the unit price should be lower than the local marginal price (LMP) α_i^t:

$$\pi_i^t \leq \alpha_i^t$$

Dynamic pricing

$$\pi_i^t = \alpha_i^t + \beta_i\left(E_i^t - s_i^t\right)$$

- E_i^t is data center's the electricity consumption, s_i^t is the charging threshold, $\beta_i > 0$ is the sensitivity parameter, and $\alpha_i^t > 0$ denotes the locational marginal price (LMP), all in the location i at time t.
- The dynamic pricing scheme is motivated by the tiered pricing which has been widely implemented in the power markets such as US, Japan, and China.
- The key idea is to set several tiers of energy consumption, and the unit price per unit of energy increases with the tiers progressively.

Constraints for s

$$E_i^t \le s_i^t \le \frac{\alpha_i^t}{\beta_i} + E_i^t \qquad 1 \le i \le N, 1 \le t \le T$$

- β_i = sensitivity parameter at data center i
  - Unit price $\pi_i^t = \alpha_i^t + \beta_i\left(E_i^t - s_i^t\right)$
  - Penalizes energy exceeding the threshold s_i^t
  - If excess increases by 1 unit, unit price increases by β_i units.
  - Total price = $E_i^t\left(\alpha_i^t + \beta_i\left(E_i^t - s_i^t\right)\right)$ (nonlinear in consumption!)
- Upper & lower bounds dependent of s; not constants!
- Hence, not box constraints (high-dimensional intervals), but are general inequalities.

ELI as objective function

- The ELI is not only a technical measure, but also an economic indicator.
- The higher ELI is, the more costly to maintain the stability of the power system.
- Furthermore, if the load ratio r_i^t is close to 100%, then the grid operator must invest more in electricity facilities to prevent the demand from exceeding the capacity, and such capacity investment is extremely costly.
- Meanwhile, we also note that the smart grid needs to give discounts to the users (through a proper choice of s) to achieve load balancing, and such discounts lead to the grid operator's revenue loss.

Stage 1 objective

- Therefore, the smart grid will aim at minimizing the weighted sum of the cost represented by ELI and revenue loss caused by offering discounts to the data centers.
- Stage 1: Electric power load balancing
- $\min_{s} \sum_{t=1}^{T} \sum_{i=1}^{N} \theta C_i \left(r_i^t(s)\right)^2 + (1-\theta)\beta_i \left(s_i^t - E_i^t(s)\right) E_i^t(s)$
- s.t. $E_i^t \le s_i^t \le \frac{\alpha_i^t}{\beta_i} + E_i^t,\ 1 \le i \le N, 1 \le t \le T$
- θ = tradeoff coefficient between ELI & discounts.

Four-step formulation

- Data:
  - Substation capacity C_i
  - Background demand B_i^t
  - Energy consumption E_i^t
- Decision variable: charging threshold s
- Constraints:
  - All demand must be served
  - Price must not exceed marginal price
- Objective
  - Minimize overall load ratio (load balancing)
  - Minimize revenue loss (due to discount)

Example

Problem 1: Suppose that we have 3 data centers $\{1,2,3\}$ and consider $T = 3$. Consider the following parameters:

Data center index i	Substation capacity C_i	Background demand B_i^t	Local marginal price α_i^t	Sensitivity β_i
1	1	0.5	0.1	0.05
2	2	0.5	0.2	0.05
3	2	1+0.05t	0.1+0.01t	0.05

a) Use a random number generator to generate E_i^t such that $0 \le E_i^t \le q_i^t$ for all i and for all t.

b) Arbitrarily select s such that

$$E_i^t \le s_i^t \le \frac{\alpha_i^t}{\beta_i} + E_i^t \qquad 1 \le i \le N, 1 \le t \le T$$

c) Compute the objective function with your selection of s in part b). Assume $\theta = 0.5$. (No optimization needed.)

Outline

- Background
- System model
- Stage 1: Power load balancing
- Stage 2: Data workload distribution

Four-stage formulation

- Data:
  - Substation capacity C_i
  - Background demand B_i^t
  - Threshold
- Decision variable: Energy consumption E_i^t
- Constraints:
  - Workload constraint
  - Quality of service constraint
- Objective
  - Minimize overall load ratio (load balancing)
  - Minimize revenue loss (due to discount)

Data center's action

- A cloud computing provider (such as Google) wants to minimize the total energy cost of multiple data centers.
- At time t, the smart grid charges the data center i with the following regional electricity price π_i^t (per unit of energy)

$$\pi_i^t = \alpha_i^t + \beta_i\left(E_i^t - s_i^t\right)$$

- Here E_i^t is the decision variable by data center i
- s_i^t is the decision variable that is determined by the smart grid in Stage 1.
- The value of s_i^t is assumed to be fixed and known in Stage 2.
- The unit price π_i^t will be lower than the LMP benchmark if the threshold s_i^t is set to be larger than the energy consumption E_i^t .

Workload constraint:

- In each time slot, the N data centers should work together to complete the total workload of L^t with the allocation to data center i as λ_i^t
- Summation: $\sum_{i=1}^{N} \lambda_i^t = L^t$
- Non-negativity: $\lambda_i^t \geq 0$ for $1 \leq i \leq N, 1 \leq t \leq T$

Quality of service (delay) constraint

- It is important for data centers to provide QoS guarantees to the users, and one important QoS metric is delay.
- Consider both the transmission delay (incurred before the request arriving at the data centers) and the queuing delay (caused by the processing in the data centers).
- To model the transmission delay, we let d_i^t denote the transmission delay experienced by a computing request from the aggregator to the data center i during time slot t.
- Notice that d_i^t is usually much less than the length of a time slot.

Quality of service (delay) constraint

- To model the queuing delay, we use queuing theory to analyze the average processing time in data center i when there are x_i^t active servers processing workload λ_i^t with a service rate μ per server, and the average waiting time (delay) is

$$\frac{1}{\mu x_i^t - \lambda_i^t}$$

- To meet the QoS requirement, the total time delay experienced by a computing request should satisfy some delay bound D, which is the maximum waiting time that a request can tolerate.
- For simplicity, in this paper we will assume homogeneous requests that have the same delay bound D. Therefore, we have the following QoS constraint

$$d_i^t + \frac{1}{\mu x_i^t - \lambda_i^t} \leq D, \mu x_i^t > \lambda^t, 1 \leq i \leq N, 1 \leq t \leq T$$

Server constraint

- At each data center i, there are tens of thousands of servers providing cloud computing services to meet users' requests.
- Let M_i denote the maximum number of available servers.
- Since the number of servers is usually large, we can relax the integer constraint of number of active servers without significantly affecting the optimal result.
- Therefore, we have the following server constraint

$$0 \leq x_i^t \leq M_i, 1 \leq i \leq N, 1 \leq t \leq T$$

Energy consumption constraint

- The energy consumption of a data center mainly depends on its computing workload and the number of active servers.
- More precisely, the energy consumption of data center i at time slot t is

$$E_i^t = x_i^t\left(P_{idle} + \left(E_{usage} - 1\right)P_{peak}\right) + x_i^t\left(P_{peak} - P_{idle}\right)\gamma_i^t + \xi$$

- P_{idle} and P_{peak} represent the average idle power and average peak power of a single server, respectively.
- Power usage effectiveness (PUE), denoted by E_{usage}, measures the energy efficiency of the data center, and is defined as the ratio of the data center's total energy consumption to the energy consumption of servers.

Energy consumption constraint

- Average server utilization of data center i at time t, denoted by γ_i^t , is represented as

$$\gamma_i^t = \frac{\lambda_i^t}{\mu x_i^t}$$

- The parameter ξ is an empirical constant.
- The term $x_i^t\left(P_{idle} + \left(E_{usage} - 1\right)P_{peak}\right)$ represents the base energy consumption, which only depends on the number of active servers.
- The term $x_i^t\left(P_{peak} - P_{idle}\right)\gamma_i^t$ represents the incremental consumption, which depends on the workload.

Energy consumption constraint

- We can rewrite E_i^t in the equivalent form as

$$E_i^t = a\lambda_i^t + bx_i^t + c, \qquad 1 \le i \le N, 1 \le t \le T$$

- An affine function in terms of the number of active servers x_i^t and the computing workload λ_i^t
- The coefficients are

$$a = \frac{P_{peak} - P_{idle}}{\mu},$$
$$b = P_{idle} + \left(E_{usage} - 1\right)P_{peak},$$
$$c = \xi.$$

Energy consumption constraint

- Note that a data center i needs to share the power supply with the background consumption, i.e., electricity usage by other industrial and residential users, in the same location.
- Since the total power supply capacity is limited at this station and the background load is time varying, then we limit the maximum power that can be consumed by data center i at time t as

$$E_i^t \leq q_i^t$$

- q_i^t denotes the available power supply to data center i at time t.

Data center's optimization

Stage 2: Total energy cost minimization

$$\min_{\lambda,x} ECost = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_i^t + \beta_i\left(E_i^t - s_i^t\right)\right)E_i^t$$

s.t. $\sum_{i=1}^{N}\lambda_i^t = L^t,\ \lambda_i^t \geq 0$ for $1 \leq i \leq N, 1 \leq t \leq T$

$$d_i^t + \frac{1}{\mu x_i^t - \lambda_i^t} \leq D, \mu x_i^t > \lambda^t, 1 \leq i \leq N, 1 \leq t \leq T$$

$$0 \leq x_i^t \leq M_i, 1 \leq i \leq N, 1 \leq t \leq T$$

$$E_i^t = a\lambda_i^t + bx_i^t + c,\ 1 \leq i \leq N, 1 \leq t \leq T$$

$$E_i^t \leq q_i^t,\ 1 \leq i \leq N, 1 \leq t \leq T$$

$$E_i^t \leq s_i^t \leq \frac{\alpha_i^t}{\beta_i} + E_i^t \qquad 1 \leq i \leq N, 1 \leq t \leq T$$

Example

Problem 2 Consider again the model in problem 1. Suppose that

$$L_t = 1, 1 \le t \le T$$
$$d_i^t = 1, 1 \le t \le T, 1 \le i \le N$$
$$D = 20$$
$$M_1 = M_2 = 3, M_3 = 5$$
$$\mu = 5$$
$$a = 0.1, b = 0.02, c = 0.1$$
$$s_i^t = 1, 1 \le t \le T, 1 \le i \le N$$

a) Construct a feasible solution λ_i^t, x_i^t for all i, t. Compute the corresponding objective value.

b) Find another feasible solution that improves the objective value with respect to that in part a).

Example

Problem 3 Suppose now we want to solve the two-stage problem. We use a **heuristic** algorithm.

a) Consider your results in problem 2 part b). Keep improving your solution to the stage-2 problem until either (i) you have completed 5 iterations or (ii) you can no longer improve your solution. Let's use this solution as the "optimal" solution associated with $s_i^t = 1, 1 \leq t \leq T, 1 \leq i \leq N$. Note that you can compute an objective value for the stage-1 problem now.

b) Find another s such that, with the updated value s', if you repeat the procedures in part a), you obtain a better objective value for the stage-1 problem than part a).

c) Conduct another 3 iterations of s