Demand Response and Workload Management for Data Centers with Increased Renewable Penetration

Junshan Zhang, ASU; R. Srikant, UIUC; Steven Low, Caltech, Lei Ying, Univ. of Michigan

Strategic Demand Response and Workload Management towards Sustainable Data Center

Smart grid will evolve into world's largest and most complex IoT
Challenges: Supply/demand vary frequently and randomly
Solutions: I) Data center demand response (DR); II) real-time feedback-based optimal power flow and distributed DER-based frequency control; III) load balancing algorithms

I) Data center demand response: Data centers have more market power and can be strategic players in energy market
- Traditional approach: passive price taking;
- Our approach: a bargaining approach for data center DR

II) Developed algorithms for time-varying nonconvex optimization with provable guarantee on tracking error and applied them to real-time feedback-based AC OPF problems; distributed algorithms for DER-based frequency control of multi-area power grid with provable guarantee on satisfaction of operational constraints and line limits

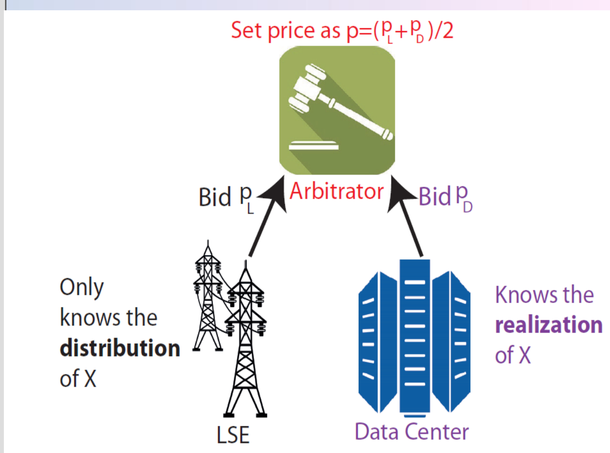
III) Load balancing: Established a university scaling of queue-length distributions of large-scale data centers under a general class of load balancing algorithms; provided sufficient conditions for achieving zero waiting for incoming jobs.

Project Outcome and Impacts

•Honors:
- R. Srikant received the IEEE Technical Field Award "2019 IEEE Koji Kobayashi Computers and Communications Award" (https://en.wikipedia.org/wiki/IEEE_Koji_Kobayashi_Computers_&_Communications_Award)

•Awards from DOE ARPA-e:
- Stochastic Optimal Power Flow for Real-time Management of Distributed Renewable Generation and Demand Response

•Technology transfer via startup activities:
- **Smartiply Inc (**https://www.smartiply.com/**): founded by Junshan Zhang**
- **Powerflex Inc** (https://www.powerflex.com/**): founded by Steven Low**



Our algorithms lead to data center demand response products that are made based on the workload management algorithms that balance quality of service and energy efficiency and determine the supply functions. The workload management algorithms optimize quality of service under the electric load constraints imposed by demand response. Three unique contributions are outlined below:

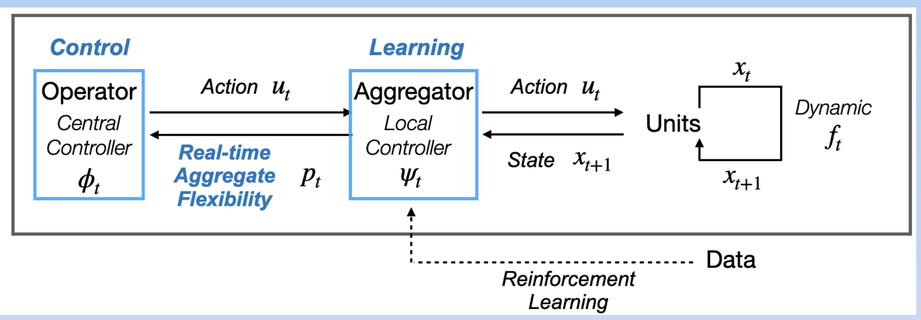
1) new market programs with strategic participation of data centers in DR

2) fundamental understanding of the impacts of power network constraints on data center DR

3) high-performance dynamic server provisioning and load balancing algorithms for large scale data centers under time-varying and stochastic electric load constraints and on-site renewable generation.

Maximum Entropy Feedback for DR

Motivation
- DERs provide flexibility to a power system operator (SO)
- … to help the SO achieve her objectives
- … while respecting their own private constraints.
- The SO and aggregators thus form a closed-loop system to optimize grid objectives subject to DERs' private constraints.
- What to feed back from aggregator to operator that describes DERs' aggregate flexibility yet protecting privacy?



Answer: Max entropy feedback ψ_t that uniquely solves

$$\max_{p_1,\dots,p_T} \sum_{t=1}^{T} \mathbb{H}(U_t|U_{<t}) \text{ subject to } U \in \mathsf{S}$$

Theorem
Let conditional probability $p_t^* := p_t^*(\cdot\,|u_{<t})$ be the max entropy feedback at time t.
1. *Feasibility*: For any control trajectory $\mathrm{u} := (u_1, \dots, u_T)$, if $p_t^*(\cdot\,|u_{<t}) > 0$, then u is feasible.
2. *Flexibility*: if $p_t^*(u_t|u_{<t}) \geq p_t^*(u_t'|u_{<t})$, then u_t provides more flexibility going forward.

Online algorithm

Data: Sequentially arrived cost functions and MEF
Result: Actions $u = (u_1, \dots, u_T)$
for $t = 1, \dots, T$ **do**
  Choose an action u_t by minimizing:

$$u_t = \phi_t(p_t) := \arg\inf_{u_t \in \mathrm{U}} (c_t(u_t) - \beta_t \log p_t(u_t|u_{<t}))$$

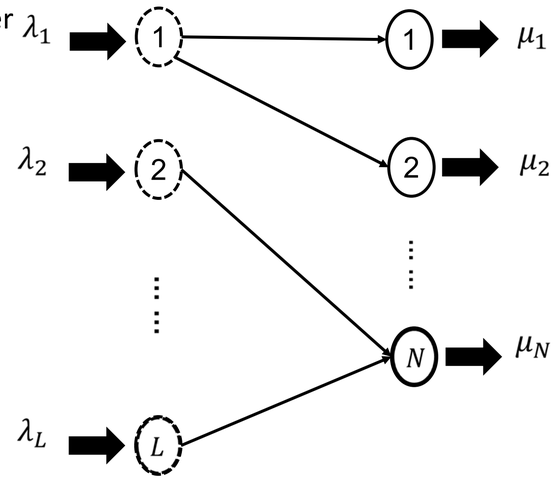
end
Return u;

Theorem
There exist β_t s.t. the algorithm generates controls $(u_1^*, \dots, u_T^*)$ that are optimal

Resource Allocation in Data Centers

- goes to infinity? How does the topology of the data center network impact the delay performance? Massive numbers of servers, 1000s to millions
- Traditional problem: Many choices, how to select a small number for load-balancing purposes?
- In practice, the number of choices is limited
- Key Questions: What resource allocation algorithms are asymptotically delay optimal in the limit as the size of the data centers

Problem:
- Jobs generated at various servers, destined for other servers
- Route jobs to minimize total response time
- The network is assumed to be sufficiently well connected
- We are interested in a mean-field regime in which the number of servers goes to infinity



Algorithm
- Join the fastest of the shortest queues (JFSQ)
- Route a job to the shortest queue, but if there is a tie, choose the fastest among them
- Important fact: No need to know the exact service rates of various servers, relative service rates are sufficient, i.e., only need to know which among a set of servers is fastest, the actual processing rate of each server is not used

Optimality: For a **well-connected** system,

JSQ has asymptotically zero delays for homogeneous servers

JFSQ is asymptotically optimal for heterogeneous servers

Derived a delay bound for finite systems

References:
- X Liu, L Ying. Steady-state analysis of load-balancing algorithms in the sub-Halfin–Whitt regime. Journal of Applied Probability 57 (2), 578-596, 2020.
- X. Liu and L. Ying. On Universal Scaling of Distributed Queues under Load Balancing. arXiv preprint arXiv:1912.11904
- W. Weng, X. Zhou and R. Srikant. Optimal Load Balancing with Locality Constraints. Proceedings of ACM SIGMETRICS 2021