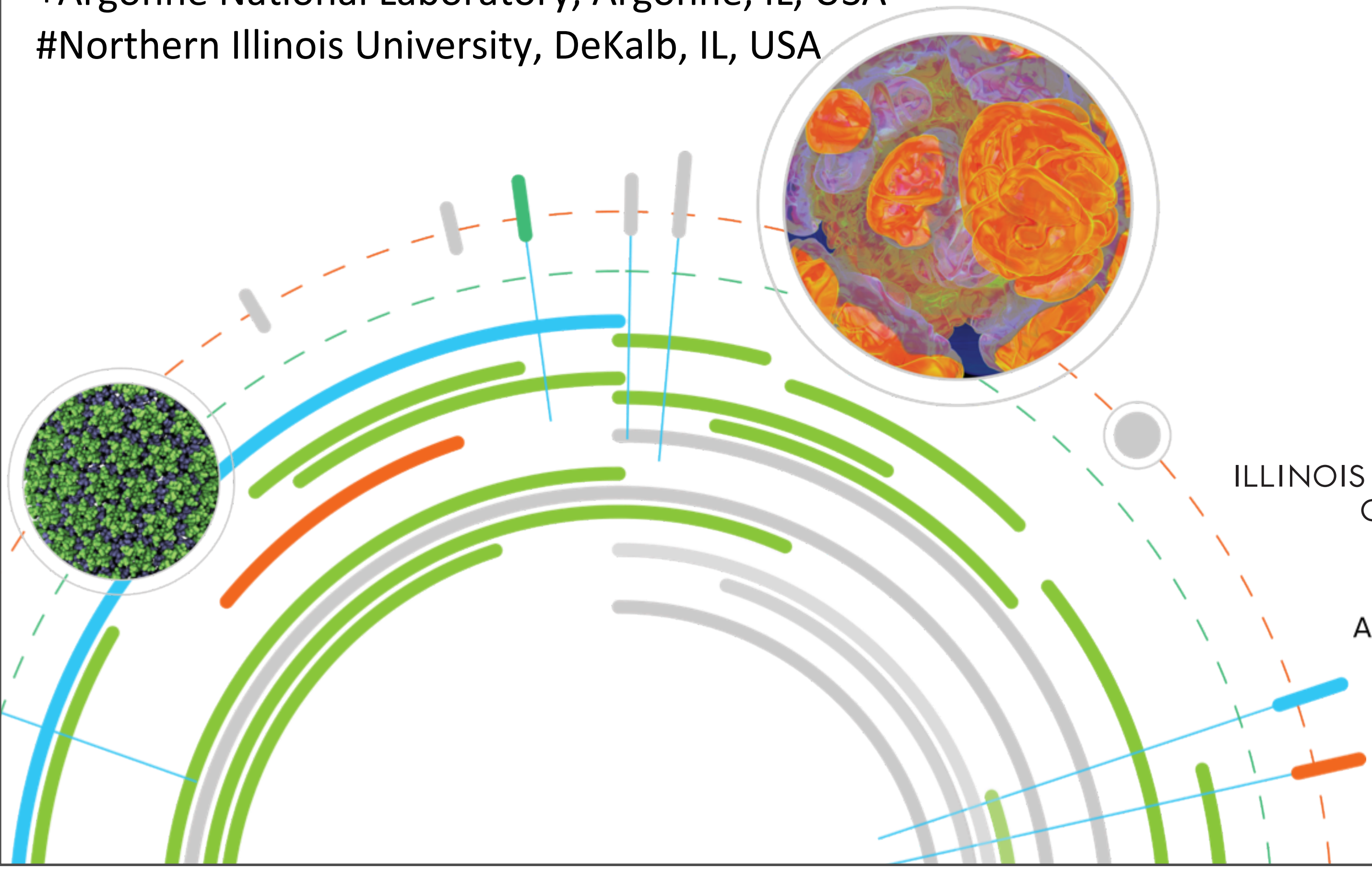
Integrating Dynamic Pricing of Electricity into Energy Aware Scheduling for HPC Systems

Xu Yang[*], Zhou Zhou[*], **Sean Wallace**[*], Zhiling Lan[*],
Wei Tang[+], Susan Coghlan[+], Michael E. Papka[+#]

*Illinois Institute of Technology, Chicago, IL, USA
+Argonne National Laboratory, Argonne, IL, USA
#Northern Illinois University, DeKalb, IL, USA

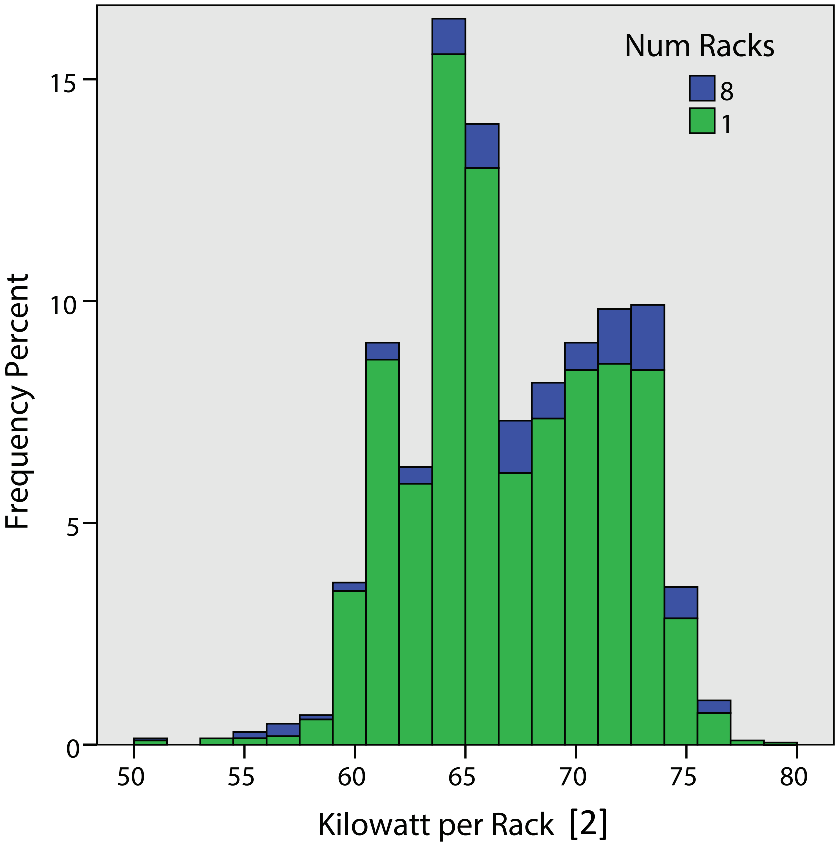


Motivation

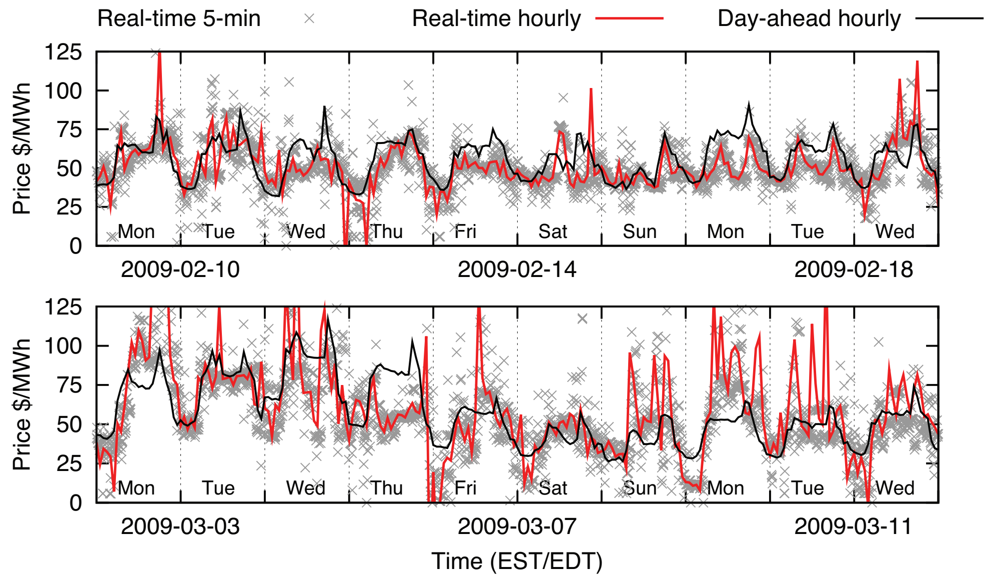
- Energy consumption/cost of HPC systems is increasing.
  - Current petascle systems on average consume 2-7 MW of power per year.
  - Argonne's Leadership Computing Facility (ACLF) budgets approximately $1M annually for electricity costs.
  - Consider if exascale systems were capped at 20MW.
    - Current super computers need to scale by a factor of **60** while increasing power by only a factor of **2**.
- Hardware can not solve this problem alone, software has a key role to play.
  - Dynamic voltage and frequency scaling (DVFS).
  - Power capping.
  - Energy or thermal aware scheduling.
- **Our work is complimentary to the above approaches!**

Key Observations in HPC

1. Distinct job power profiles



2. Dynamic electricity pricing



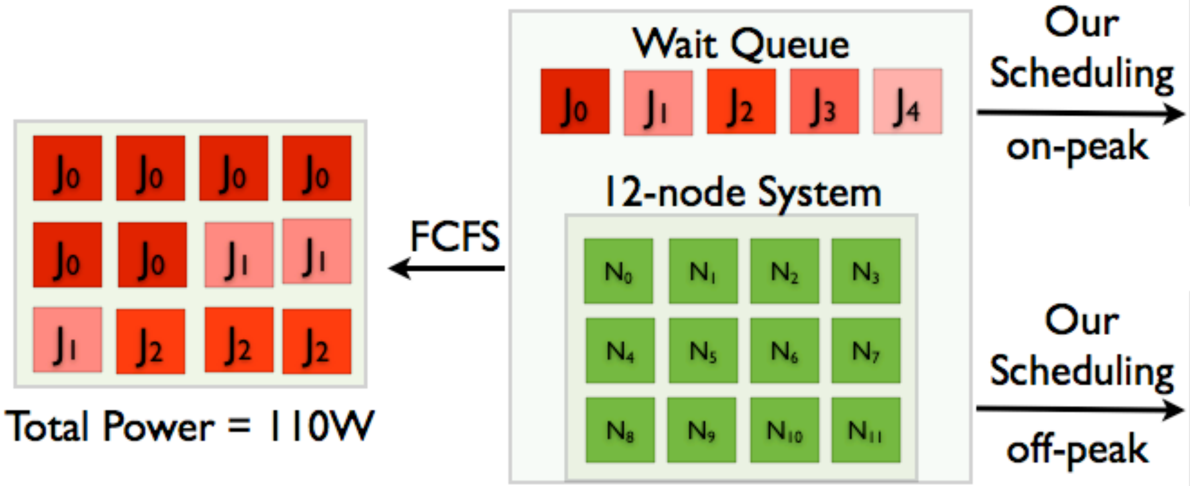
Comparing price variation in different whole-sale markets, for the New York City hub [1].

[1] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. 2009. Cutting the electric bill for internet-scale systems. SIGCOMM Comput. Commun. Rev. 39, 4 (August 2009), 123-134. DOI=10.1145/1594977.1592584 http://doi.acm.org/10.1145/1594977.1592584

[2] S. Wallace, V. Vishwanath, S. Coghlan, J. Tramm, Z. Lan, and M. Papka, "Application power profiling on IBM Blue Gene/Q". In *IEEE International Conference on Cluster Computing 2013*, Indianapolis, USA, September 2013.

Solution Overview

- Dispatch jobs with greatest power consumption during the off-peak period, jobs with least power consumption during on-peak period.



Darker colors mean more power.

Job	Power Consumption (Watts)	Job Size
J0	50	6
J1	20	3
J2	40	3
J3	30	3
J4	10	6

Methodology

Waiting Queue

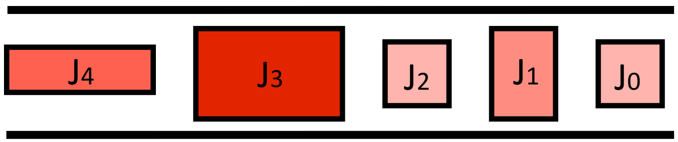
Size of boxes intended to represent runtime horizontally and size vertically.

Color used as in previous slide to represent power requirements.

Naively - Jobs enter Waiting Queue

Methodology

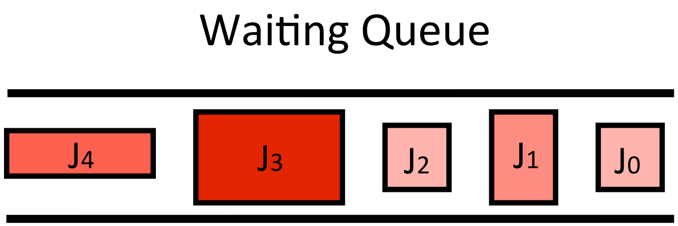
Waiting Queue



Size of boxes intended to represent runtime horizontally and size vertically.
Color used as in previous slide to represent power requirements.

Naively - Jobs enter Waiting Queue

Methodology



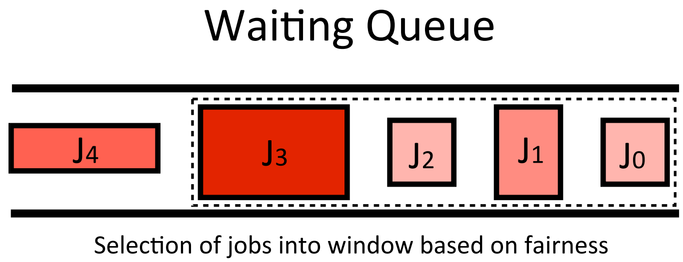
Size of boxes intended to represent runtime horizontally and size vertically.
Color used as in previous slide to represent power requirements.

Step 1 - Scheduling Window maintained at front of queue

Methodology

Size of boxes intended to represent runtime horizontally and size vertically.

Color used as in previous slide to represent power requirements.

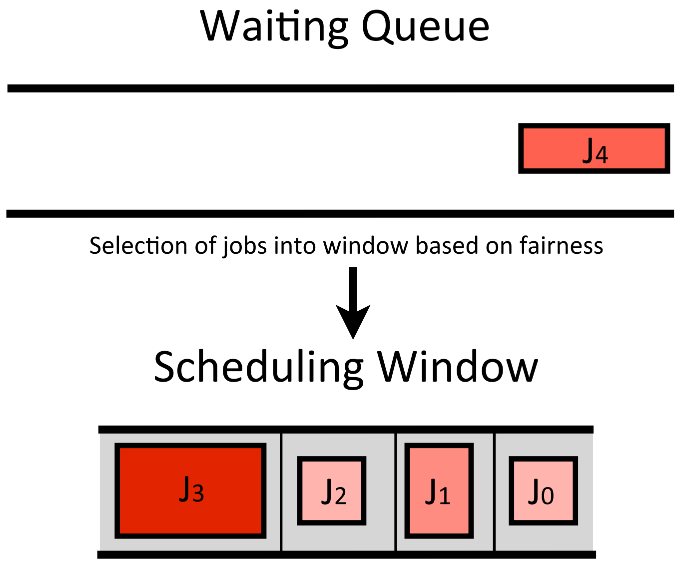


Step 1 - Scheduling Window maintained at front of queue

Methodology

Size of boxes intended to represent runtime horizontally and size vertically.

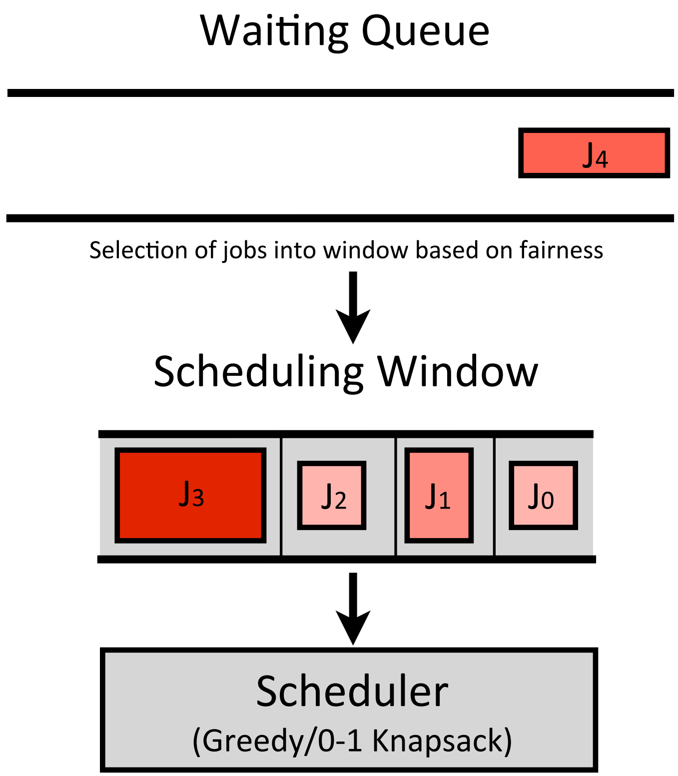
Color used as in previous slide to represent power requirements.



Step 1 - Scheduling Window maintained at front of queue

Methodology

Size of boxes intended to represent runtime horizontally and size vertically.
Color used as in previous slide to represent power requirements.

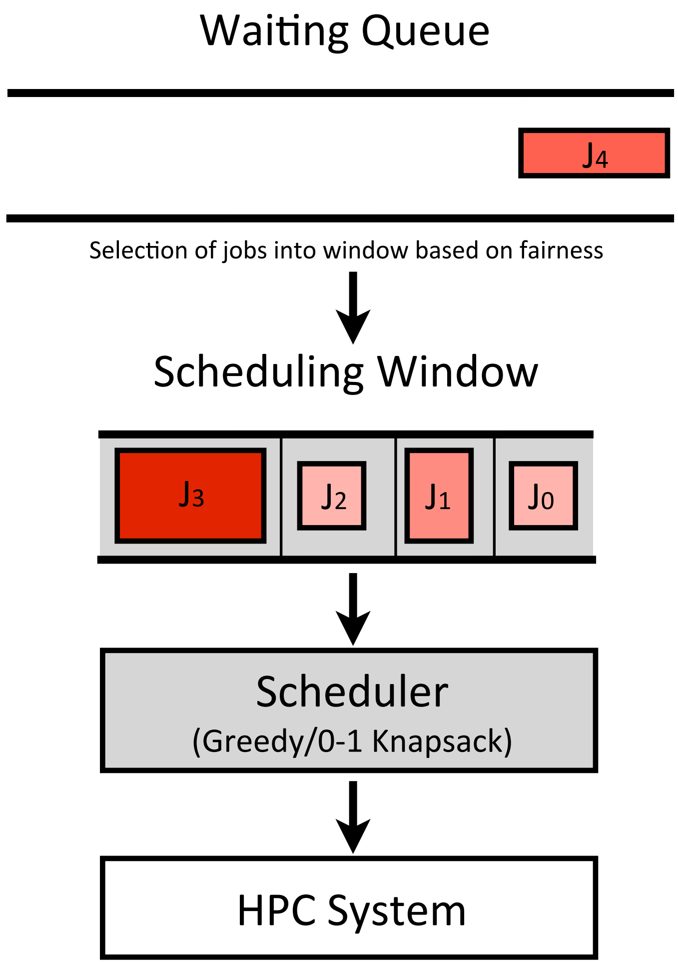


Step 2 - Scheduler picks job from Scheduling Window (Greedy/Knapsack)

Greedy heuristics are investigated as scheduling is an NP-Complete Problem.

Methodology

Size of boxes intended to represent runtime horizontally and size vertically.
Color used as in previous slide to represent power requirements.



Step 3 - Jobs are allocated based on system status, electricity price, etc.

Scheduling Algorithms

- Greedy
  - During the on-peak time, all jobs in the scheduling window are sorted in a decreasing order based on their power profiles.
  - During the off-peak time, jobs are sorted in an increasing order based on power profiles.
  - Complexity O(nlgn)
- Knapsack
  - Number of available nodes in the system (N_t) is used as the knapsack's size.
  - For each job, its power profile (measured in W/node or kW/rack) is used as its value and the number of required nodes is used as the weight.
  - During the on-peak period, the goal is to minimize the value. During off-peak, the goal is to maximize the value.
  - Complexity $O(nN_t)$

Evaluation Metrics

- Electricity Bill
  - The relative different between the electricity bill using our design and FCFS to measure the saving achieved by our design.
  - Measures total dollar savings achieved by our design.
- System Utilization Rate
  - The ratio of the node-hours that are used for useful computation to the elapsed system node-hours.
  - The primary objectives of this work is to save costs on electricity bill <u>without</u> degradation to utilization.
- Job Average Wait Time
  - Job wait time refers to the time elapsed between the moment it is submitted to the moment it is allocated to run.

Simulator and Job Traces

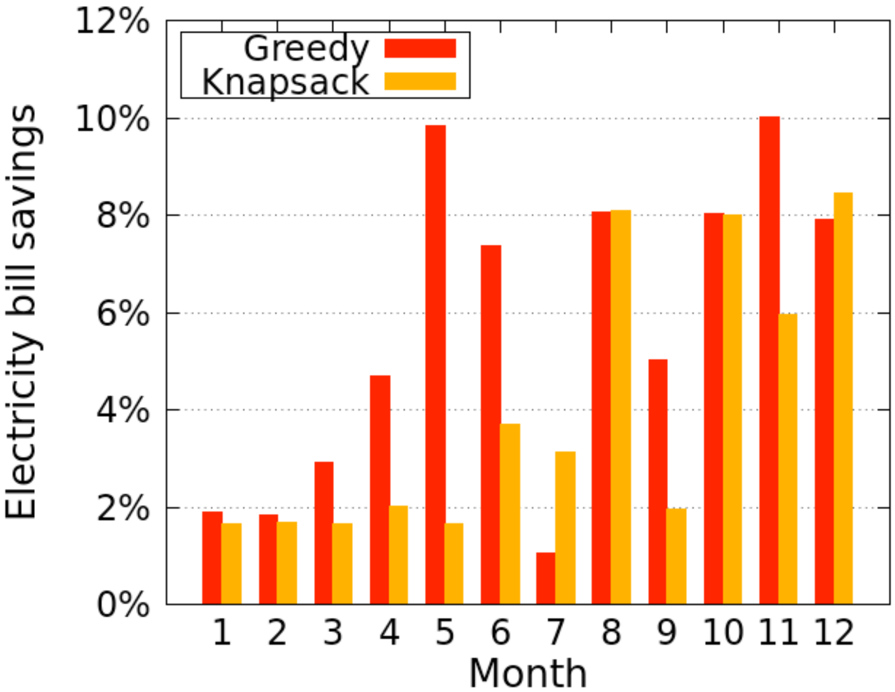
- Simulator
  - CQSim: a trace-based, event-driven (job submission, job start, and job end) scheduling simulator.
- Job Traces
  - ANL-BGP: contains 26,012 jobs, represents **capability computing** where the computing power is used to solve larger problems.
    - Trace at 40 rack scale, however for simplification we look at 2048 nodes at a time.
  - SDSC-Blue: contains 144,830 jobs, represents **capacity computing** where the computing power is utilized to solve a large number of small problems.
    - Trace at 1,152-process scale.
- Job Traces and Power Consumption from Mira (48 Racks of IBM Blue Gene/Q) - A Case Study

Trace Based Experiments

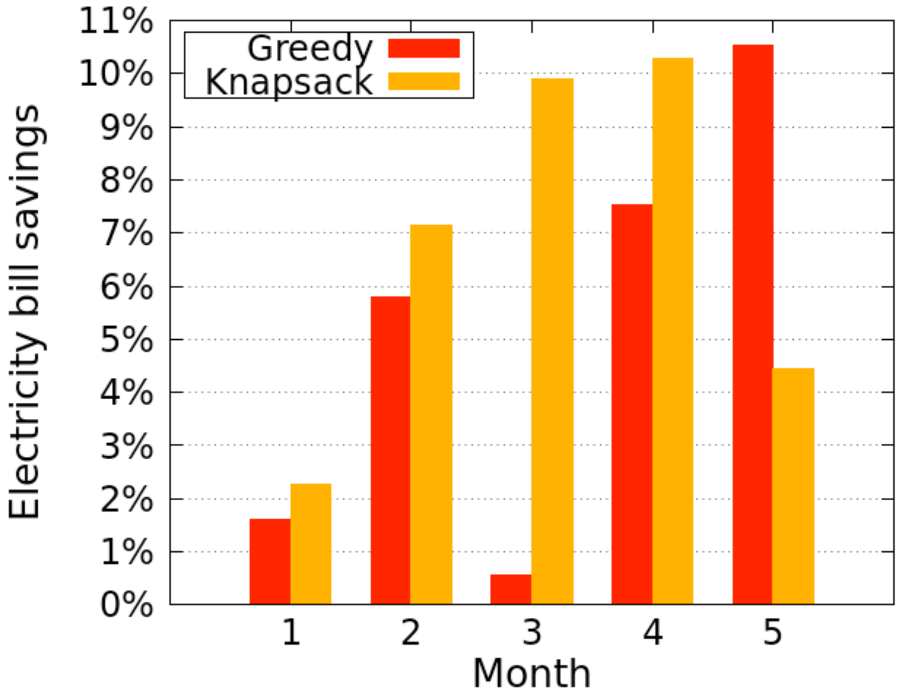
- Three sets of experiments with both traces for both Greedy and Knapsack scheduling:
  - The impact of on-peak/off-peak electricity pricing ratio.
  - The impact of job power profile ratio.
  - The impact of scheduling frequency.

- Base configuration for these experiments:
  - Job power profile ratio - 1:3
  - Off-peak/on-peak pricing ratio - 1:3
  - Scheduling frequency - 10 seconds.

Experimental Results - Electrical Bill Savings



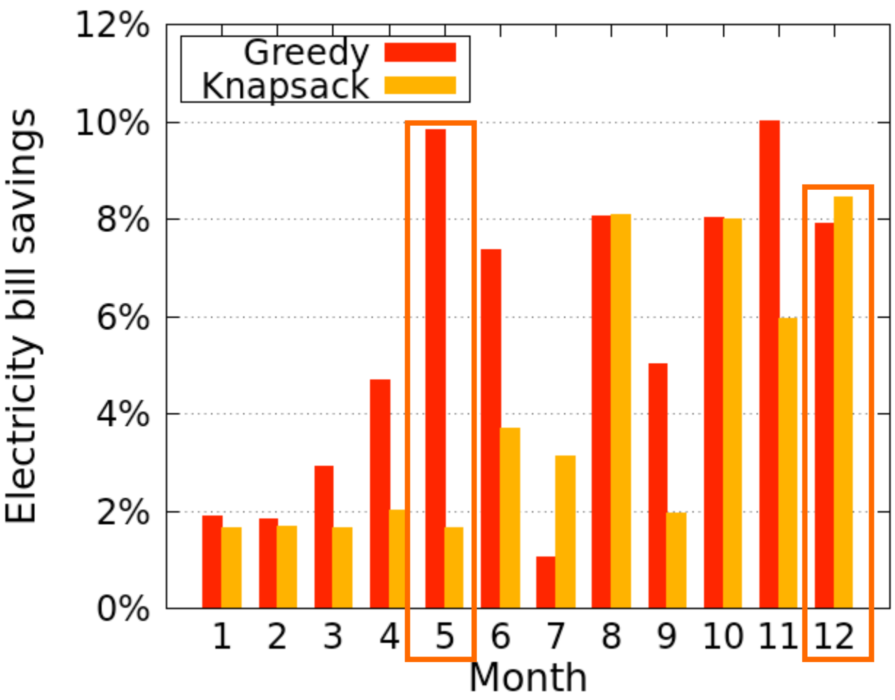
SDSC-Blue



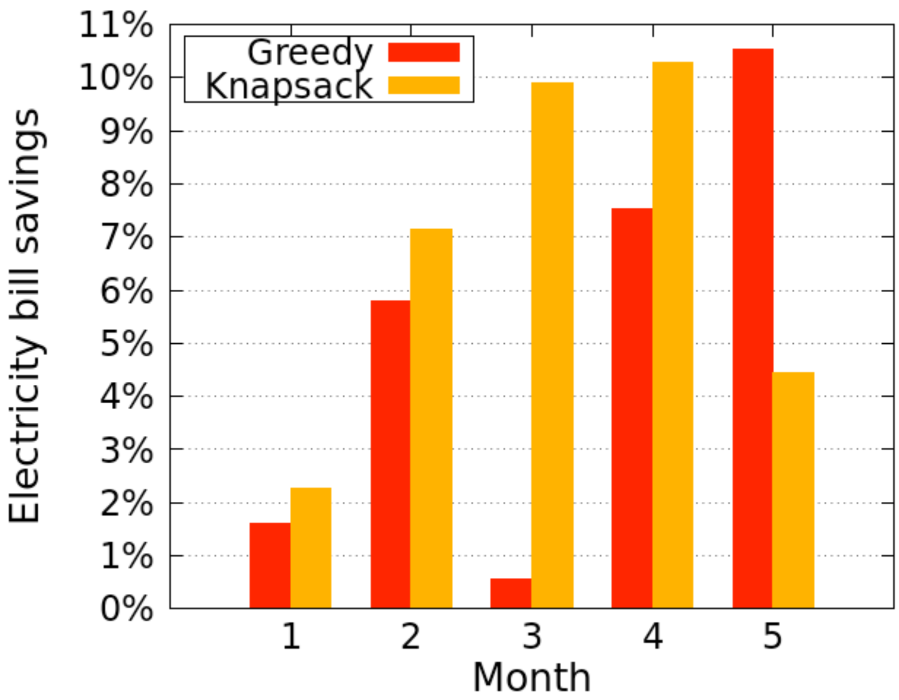
ANL-BGP

Savings with Greedy	0.5% - 10%
Savings with Knapsack	2% - 10%
Average Savings	3.16% - 5.53%

Experimental Results - Electrical Bill Savings



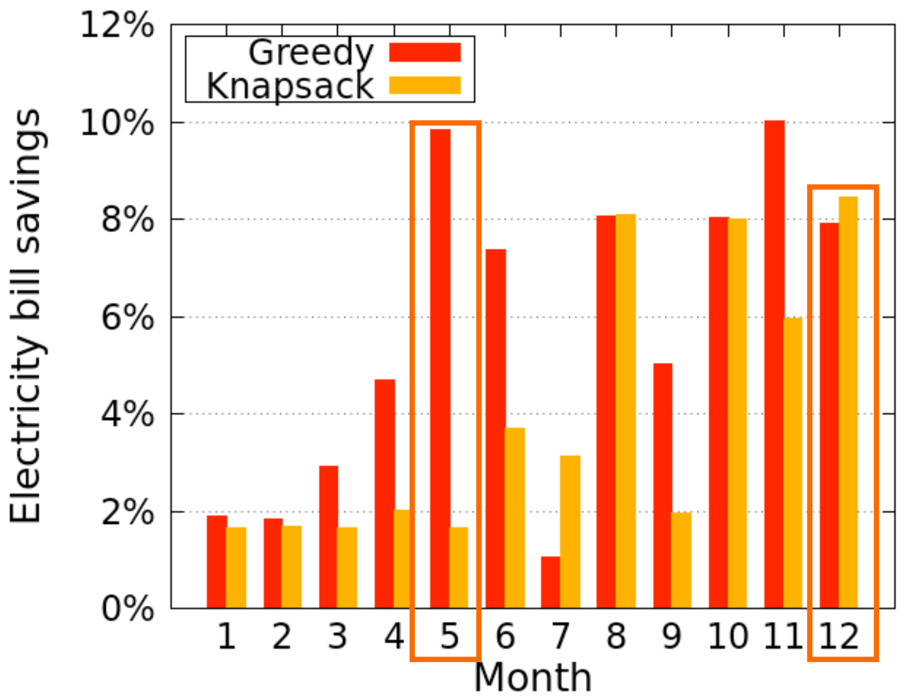
SDSC-Blue



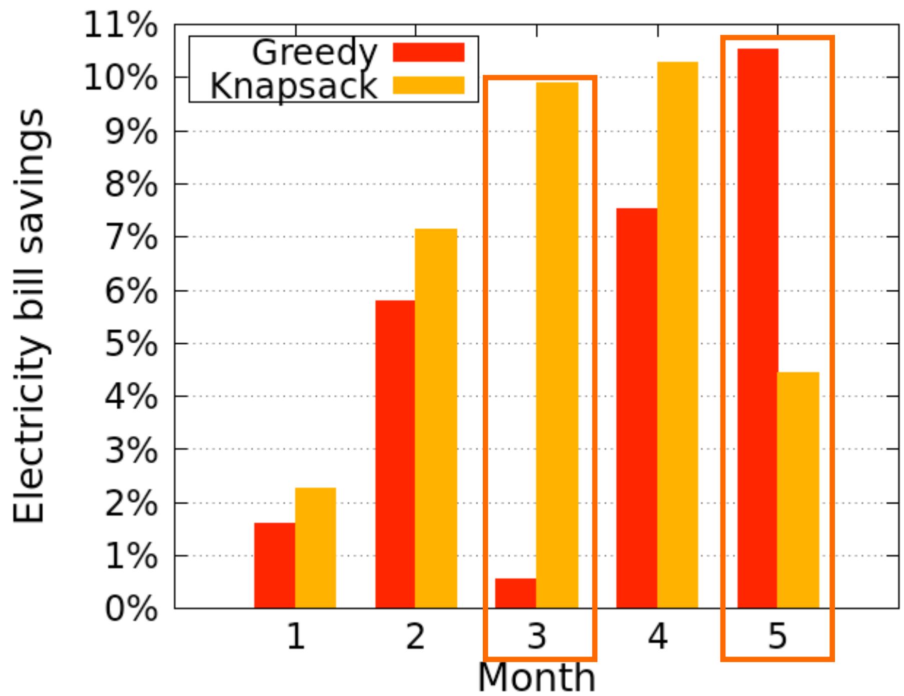
ANL-BGP

Savings with Greedy	0.5% - 10%
Savings with Knapsack	2% - 10%
Average Savings	3.16% - 5.53%

Experimental Results - Electrical Bill Savings



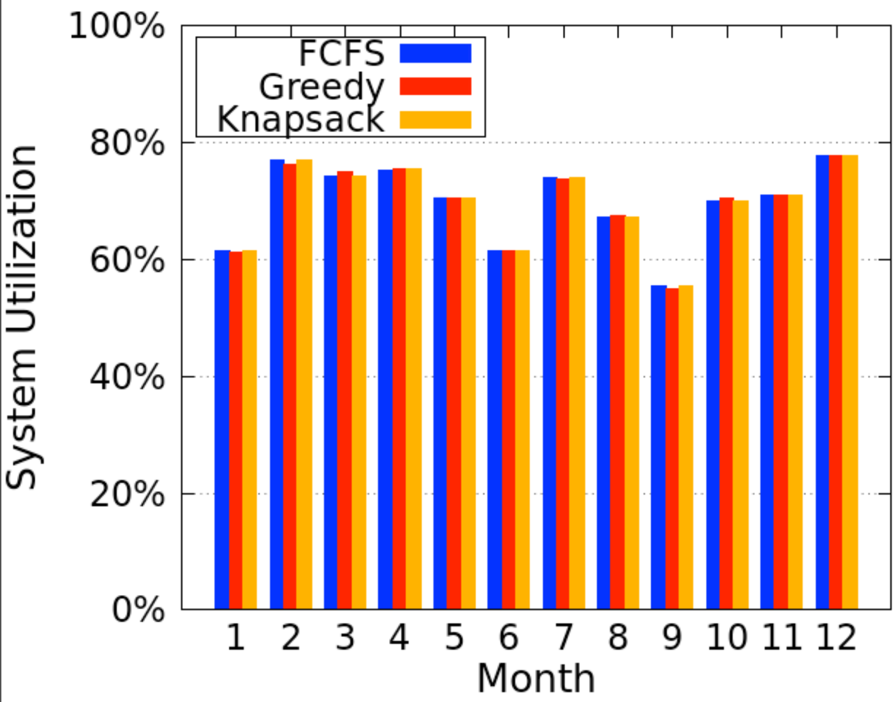
SDSC-Blue



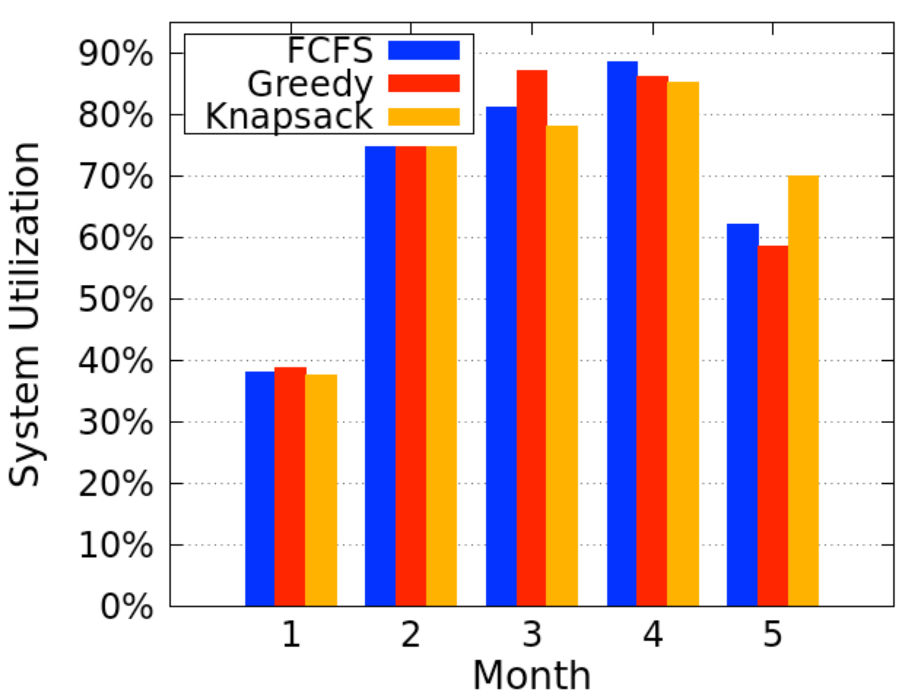
ANL-BGP

Savings with Greedy	0.5% - 10%
Savings with Knapsack	2% - 10%
Average Savings	3.16% - 5.53%

Experimental Results - System Utilization



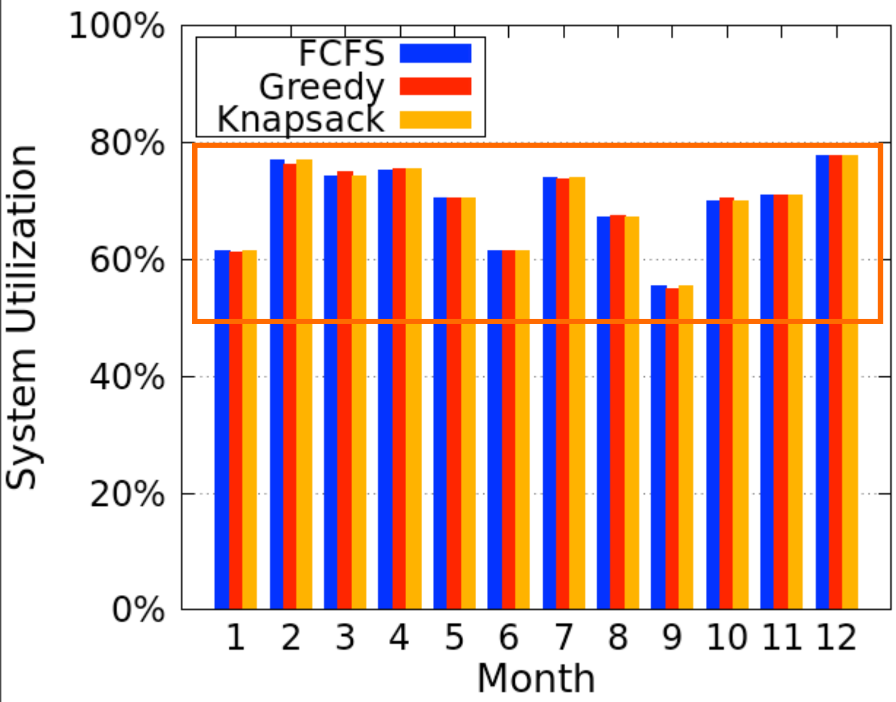
SDSC-Blue



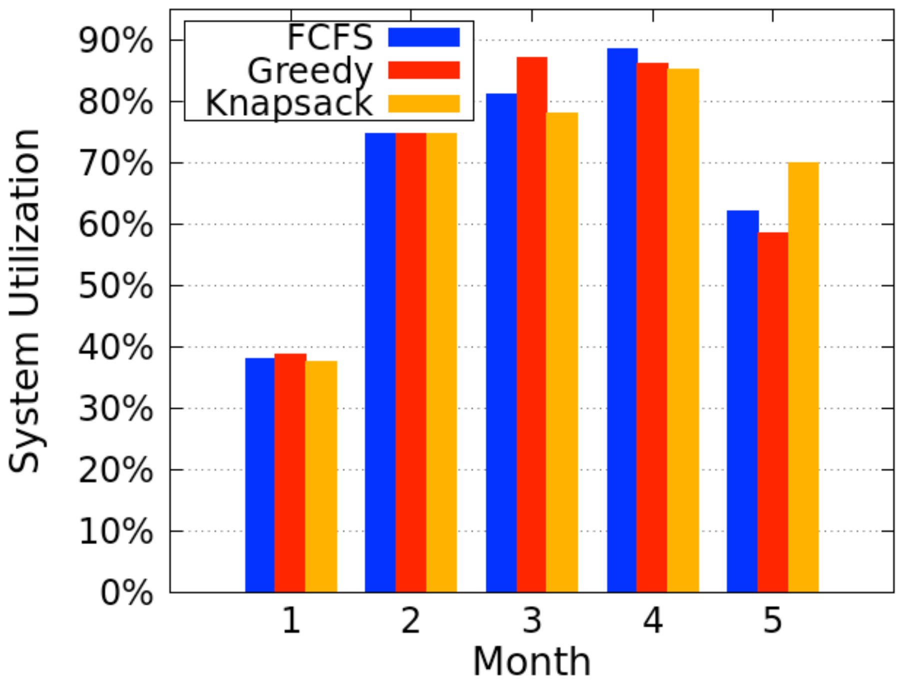
ANL-BGP

Average monthly utilization degradation introduced by our design is always less than **5%**.

Experimental Results - System Utilization



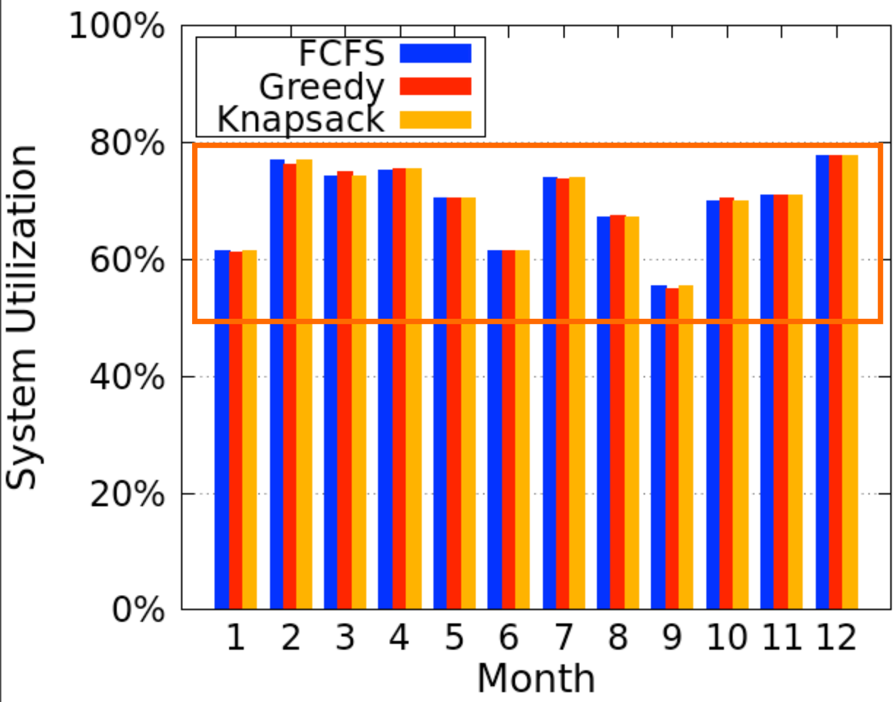
SDSC-Blue



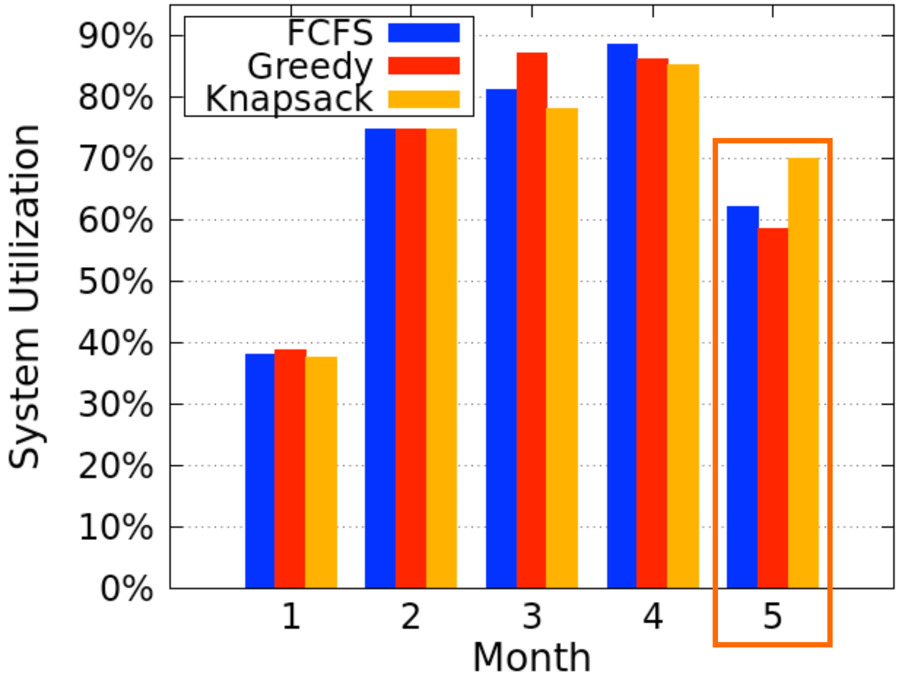
ANL-BGP

Average monthly utilization degradation introduced by our design is always less than **5%**.

Experimental Results - System Utilization



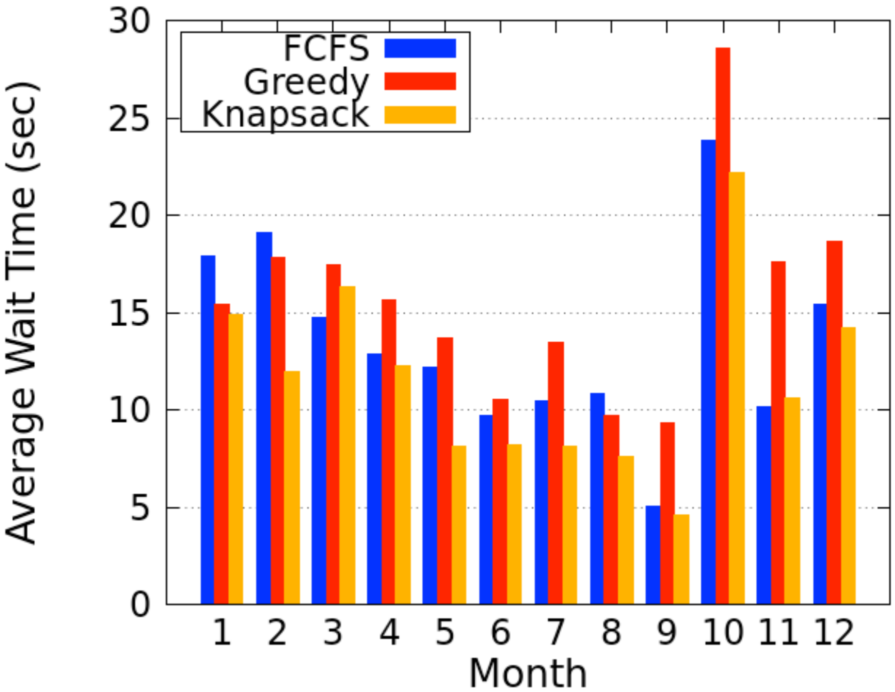
SDSC-Blue



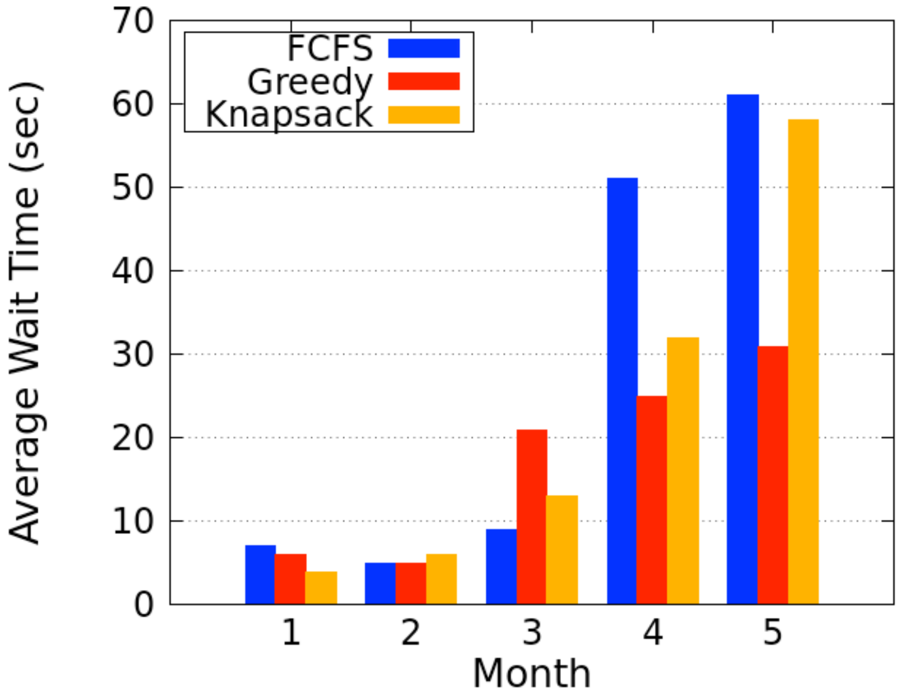
ANL-BGP

Average monthly utilization degradation introduced by our design is always less than **5%**.

Experimental Results - Average Wait Time



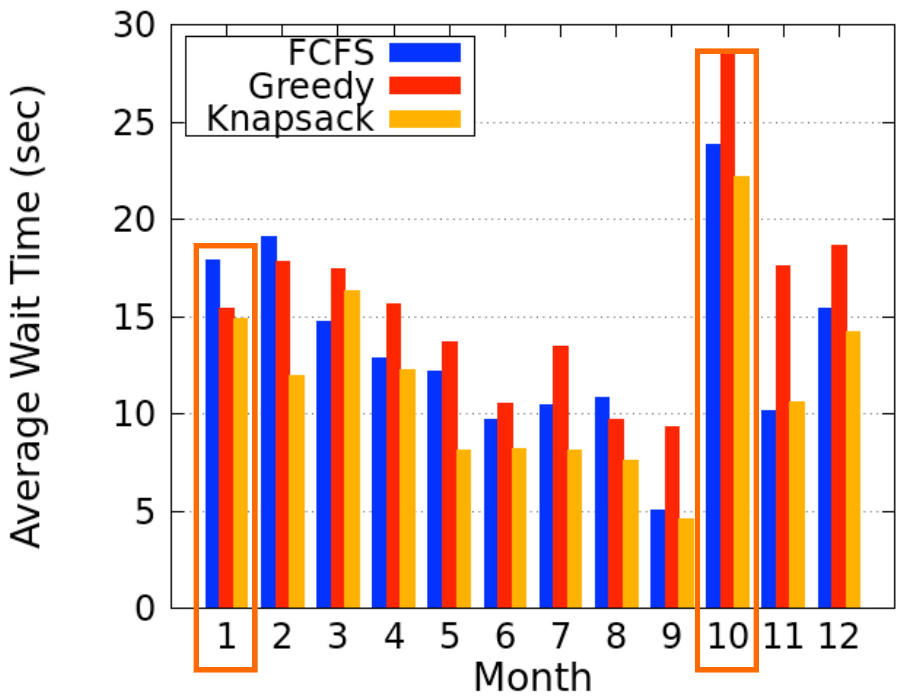
SDSC-Blue



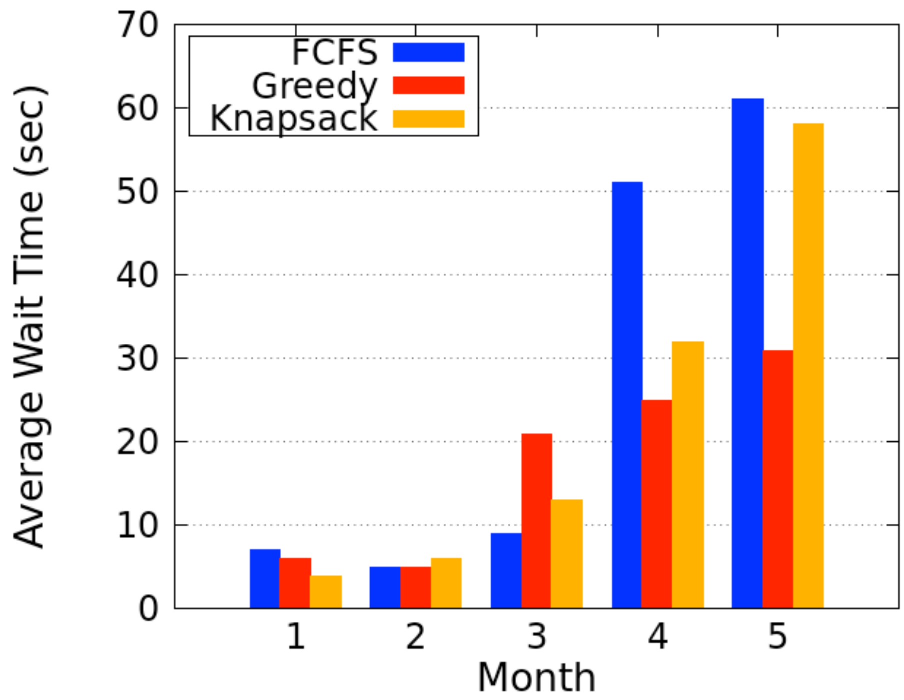
ANL-BGP

In both cases, maximum average monthly penalty incurred by our design over FCFS is **10 seconds**.

Experimental Results - Average Wait Time



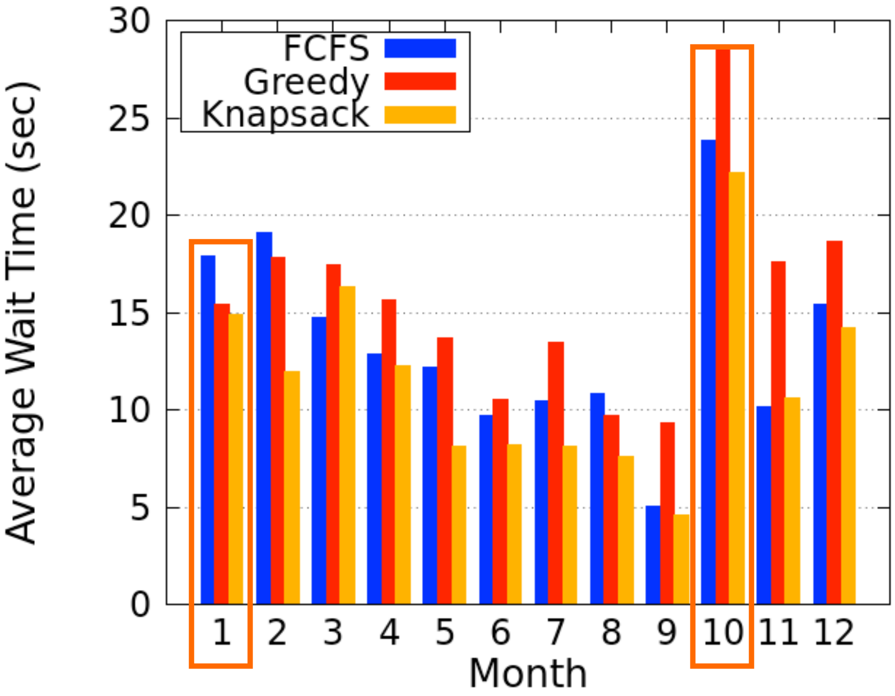
SDSC-Blue



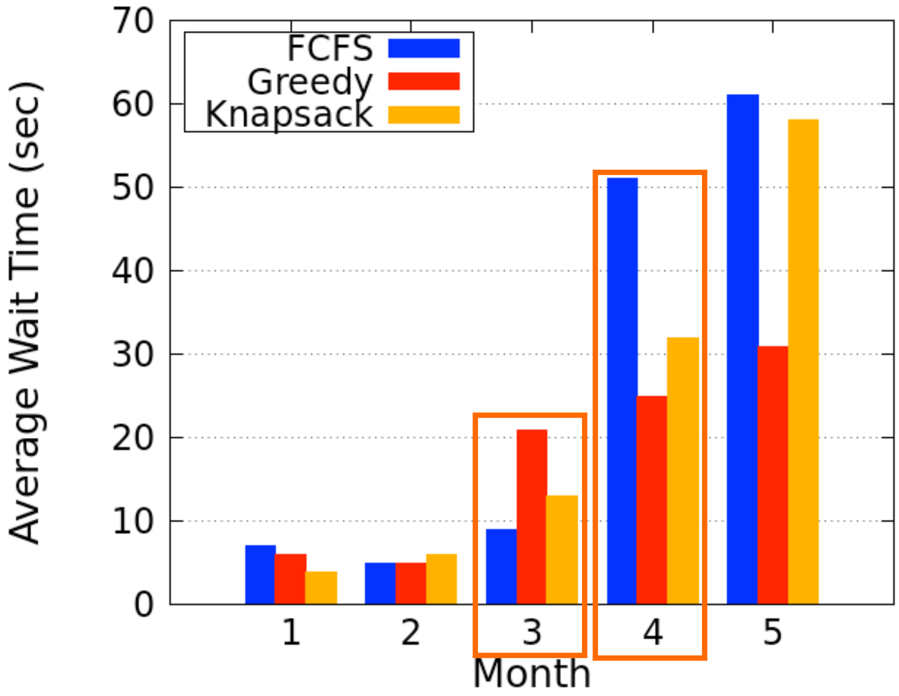
ANL-BGP

In both cases, maximum average monthly penalty incurred by our design over FCFS is **10 seconds**.

Experimental Results - Average Wait Time



SDSC-Blue



ANL-BGP

In both cases, maximum average monthly penalty incurred by our design over FCFS is **10 seconds**.

Impacts of Electricity Prices on Job Power Profiles

Power Ratio		Pricing Ratio: ANL-BGP 1:3	ANL-BGP 1:4	ANL-BGP 1:5	SDSC-Blue 1:3	SDSC-Blue 1:4	SDSC-Blue 1:5
1:2	Greedy	3.54%	4.33%	4.79%	3.84%	4.84%	6.19%
	Knapsack	4.18%	5.07%	5.64%	2.39%	3.01%	3.85%
1:3	Greedy	5.06%	6.13%	6.85%	4.33%	5.46%	6.98%
	Knapsack	5.35%	6.48%	7.25%	3.16%	3.98%	5.10%
1:4	Greedy	6.27%	7.58%	8.48%	5.55%	6.98%	8.95%
	Knapsack	7.21%	8.52%	9.86%	3.05%	3.84%	4.92%

Power ratio is the ratio between power per node at highest profile and lowest profile.
Price ratio is the ratio between off-peak and on-peak.

Savings at least ~**4%** for best algorithms for ANL-BGP and SDSC-Blue
At least **$40,000** savings on $1M

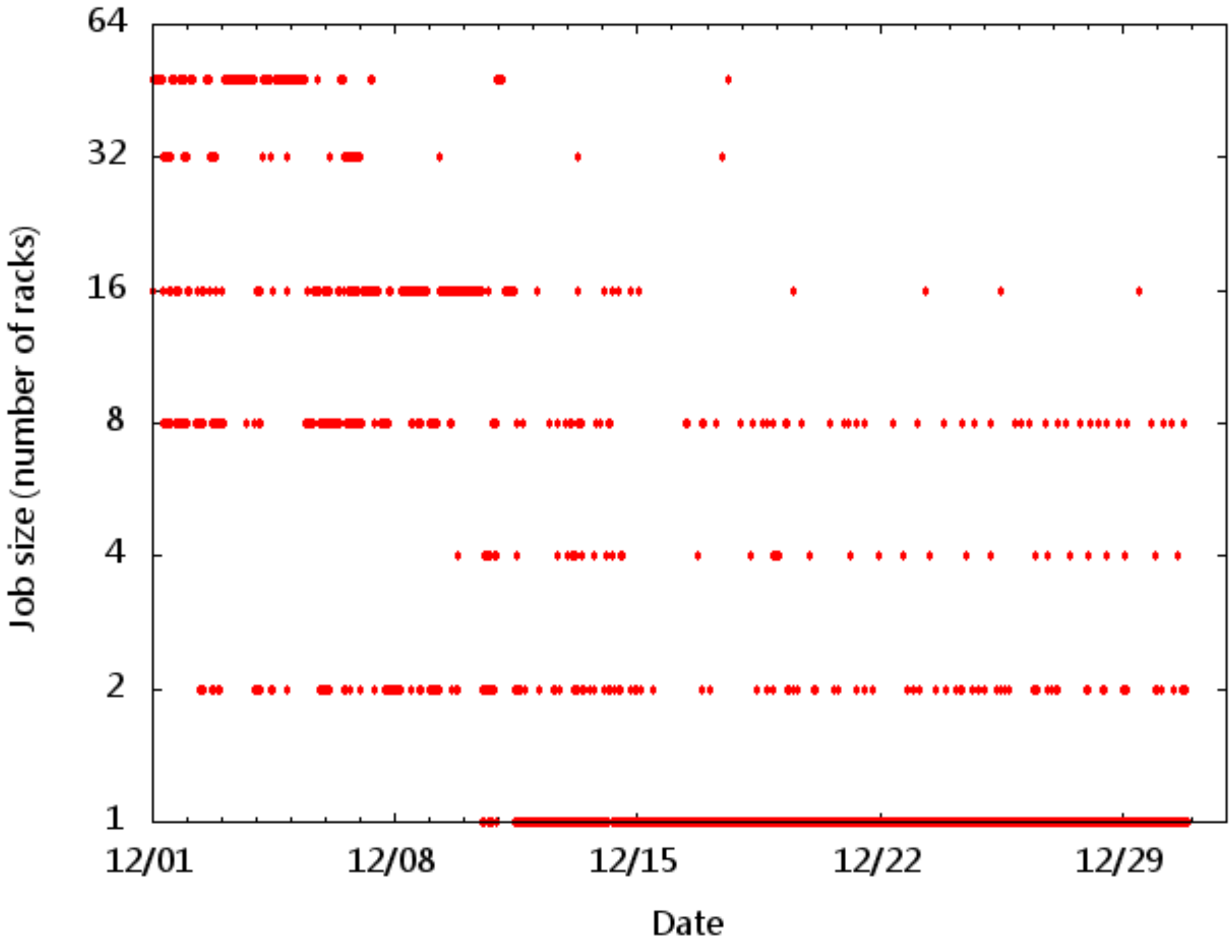
Impact of Scheduling Frequencies

Frequency	Greedy	Knapsack	FCFS	Greedy	Knapsack
	Scheduling Policy				
10 Seconds	7.49%	7.13%	70.0%	69.70%	69.07%
	4.33%	3.16%	69.59%	69.53%	69.50%
20 Seconds	10.07%	8.91%	68.56%	69.03%	65.97%
	9.70%	9.80%	68.56%	69.25%	65.06%
30 Seconds	17.52%	**22.43%**	63.77%	60.42%	60.84%
	19.69%	**23.06%**	67.38%	68.85%	66.21%
	Electrical bill savings obtained by our scheduling policies with different scheduling frequencies. In each cell, the top number is on ANL-BGP and the bottom number is on SDSC-Blue		**System utilization rate** under different scheduling frequencies. In each cell, the top number is on ANL-BGP and the bottom number is on SDSC-Blue		

Savings not less than 7%, even at lowest frequency.

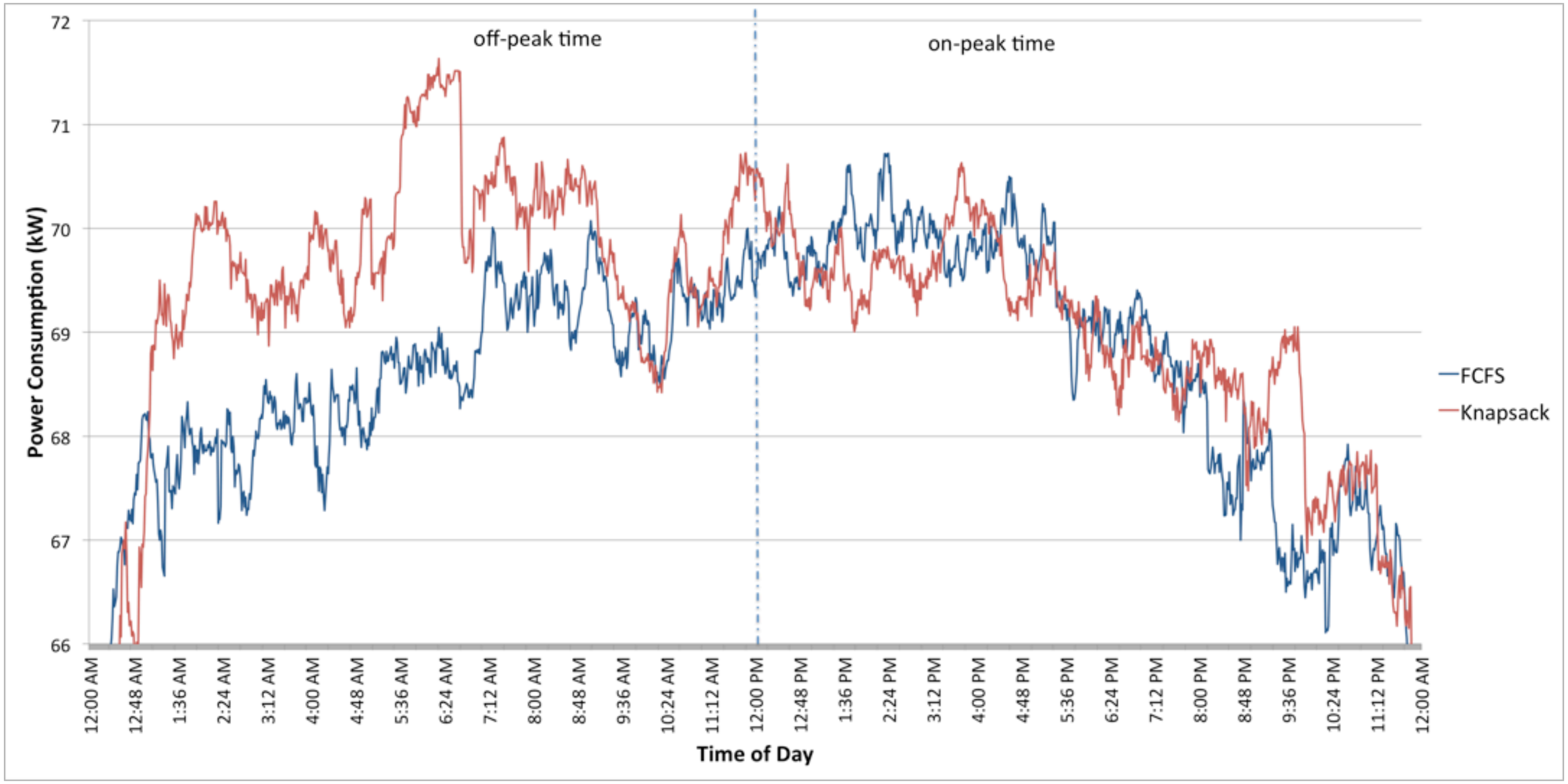
Case Study - Mira

Job traces from Mira were collected during December 2012. Totally, there were 3,333 jobs executed on the machine.



Case Study - Mira

Average daily power consumption. Power consumption at each point in time is calculated as the average over the month.



Our design aims to increase power consumption during off-peak and decrease during on-peak, as shown.

Savings vs. FCFS was **9.98%**.

Conclusions

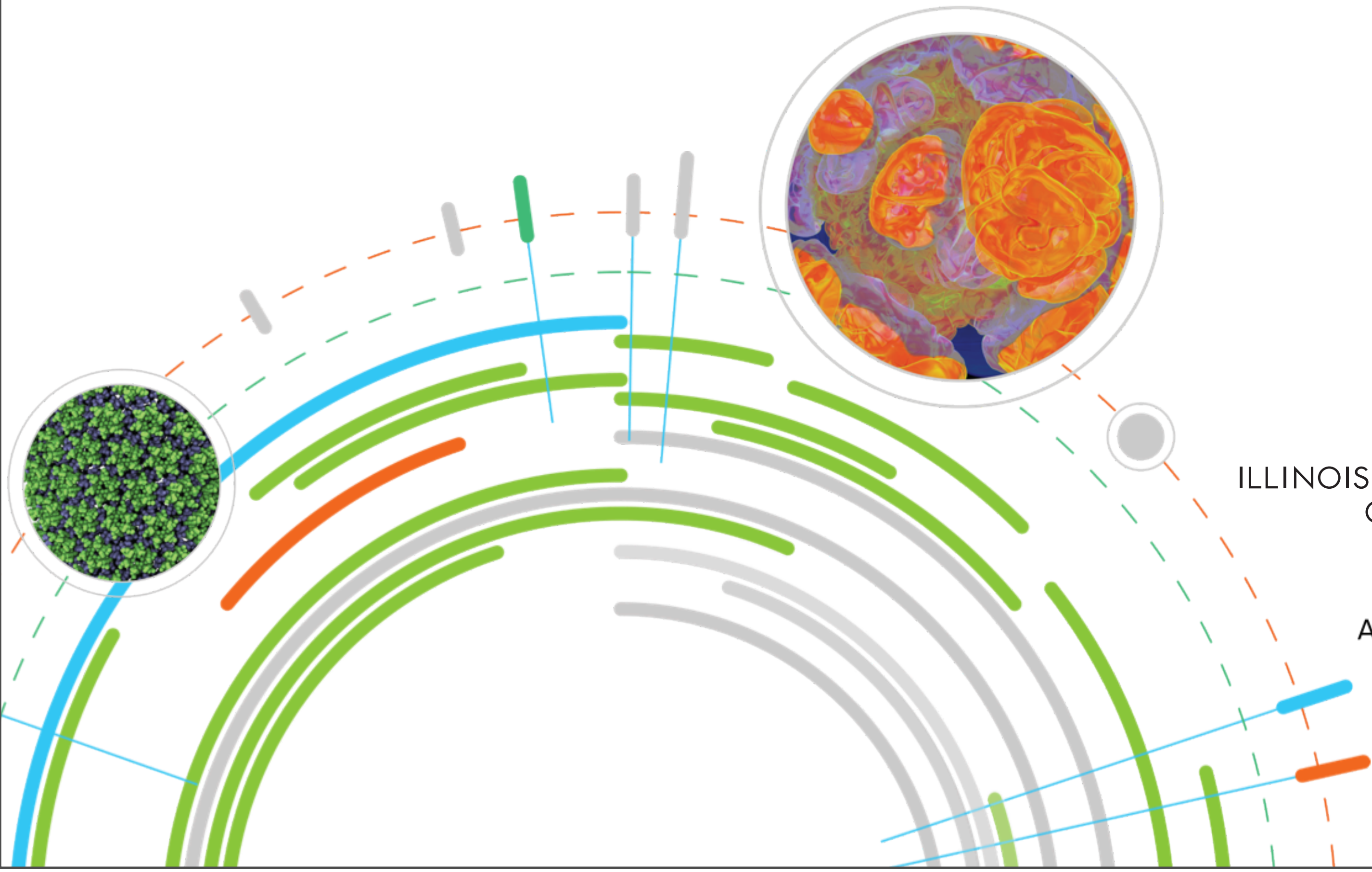
- Novel energy aware scheduling design proposed.
- Design capable of cutting electricity bill by up to **23%** without impact on utilization.
- Most effective with big capability workloads, though improvement possible with other systems.
- Knapsack outperforms Greedy for capability computing.

- Future work:
  - Integration of our design with the work on environmental data analysis tools to automatically obtain job power profiles.

- Acknowledgements

  - US NSF grant CNS-0843514 and OCI-0904679 DOE Contract DEAC02-06CH11357

Questions?

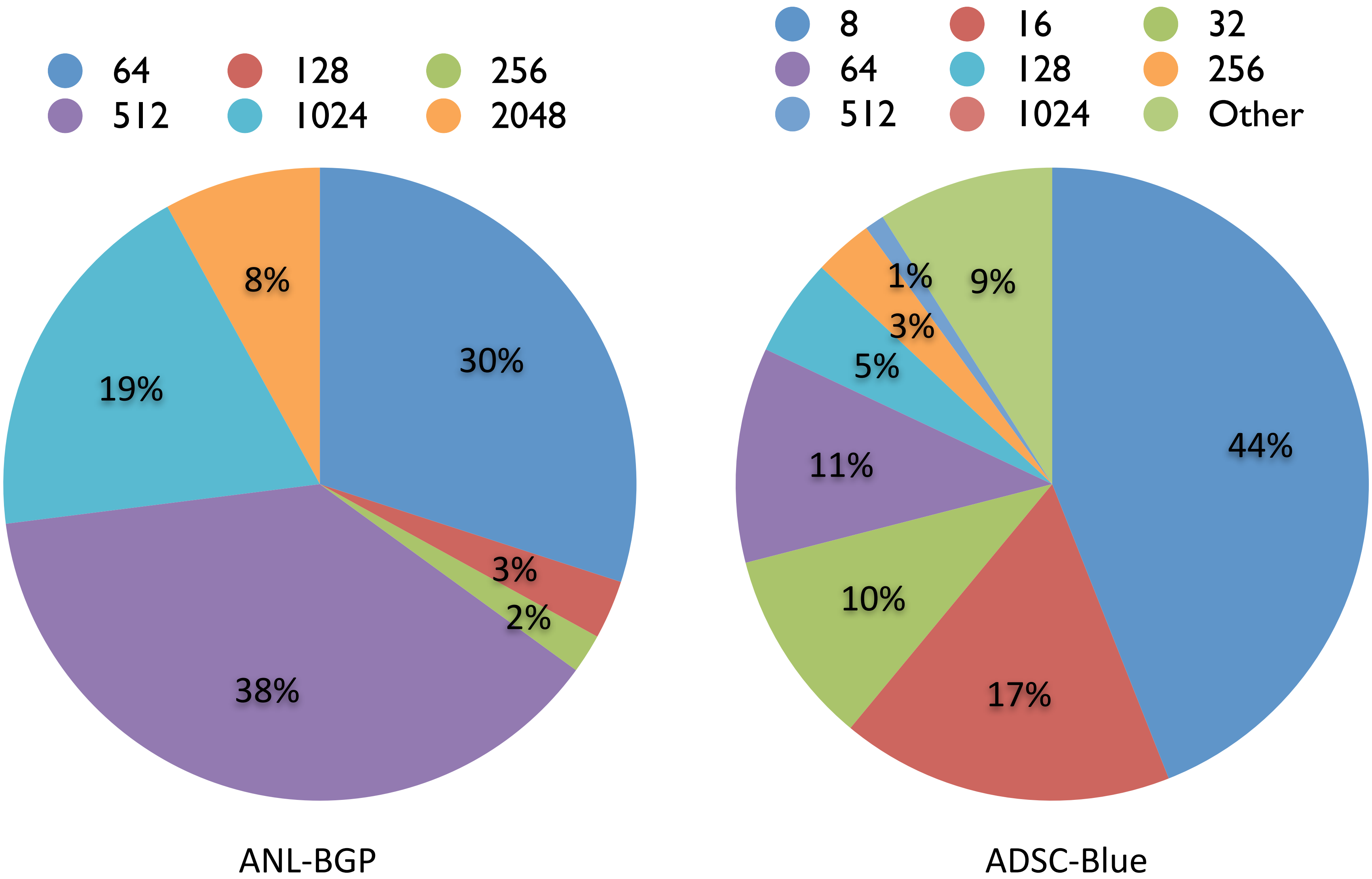


ILLINOIS INSTITUTE OF TECHNOLOGY

Argonne **Leadership Computing** Facility

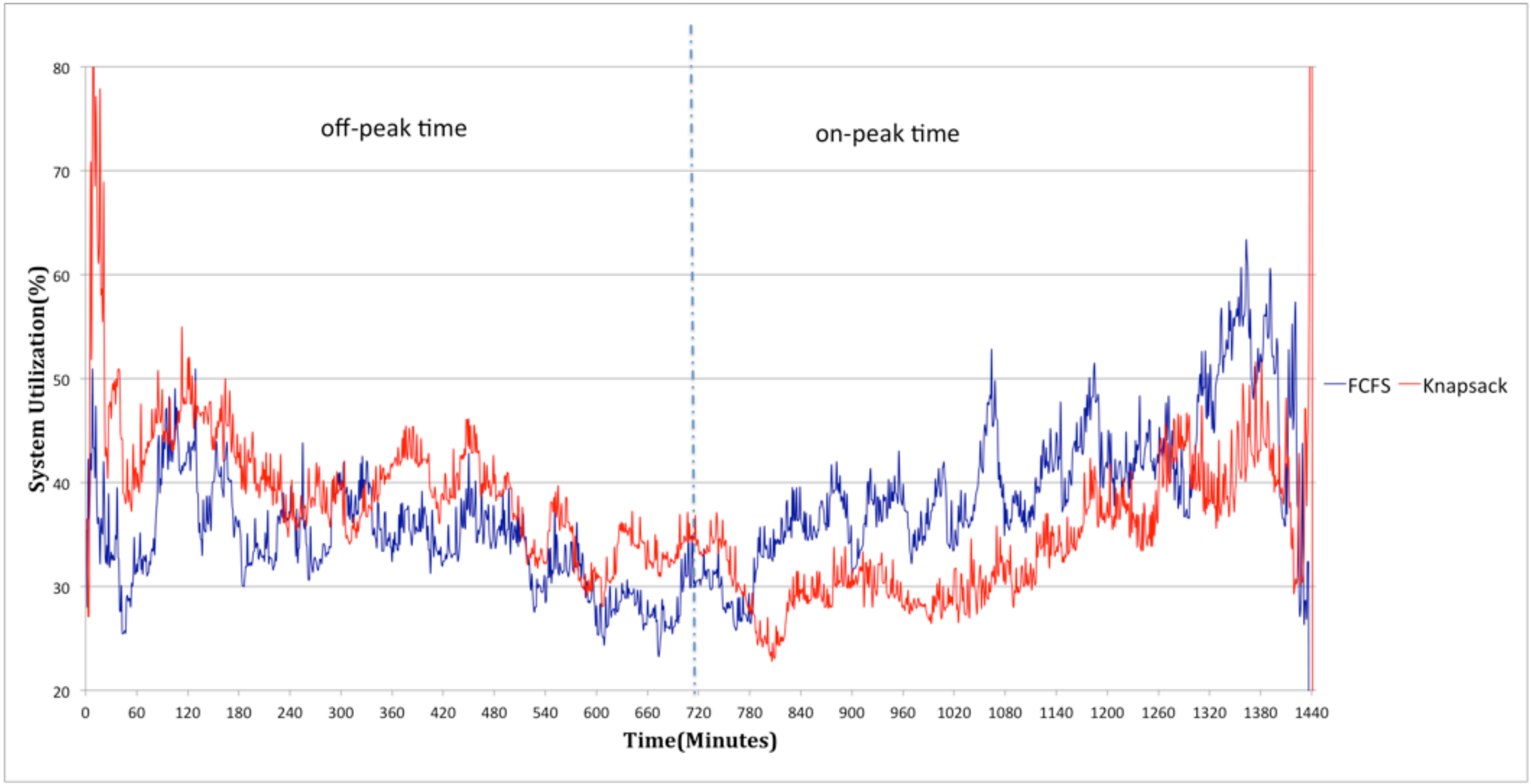
Argonne NATIONAL LABORATORY

Trace Size Distributions



Case Study - Mira

Average utilization within a day. Here system utilization at each time point is calculated as the average over the month.



Utilization is higher during off-peak as our design attempts to allocate as many jobs as possible with high power profiles. During off-peak our design schedules large jobs with low power profiles, leaving some idle nodes that are not sufficient for other jobs.