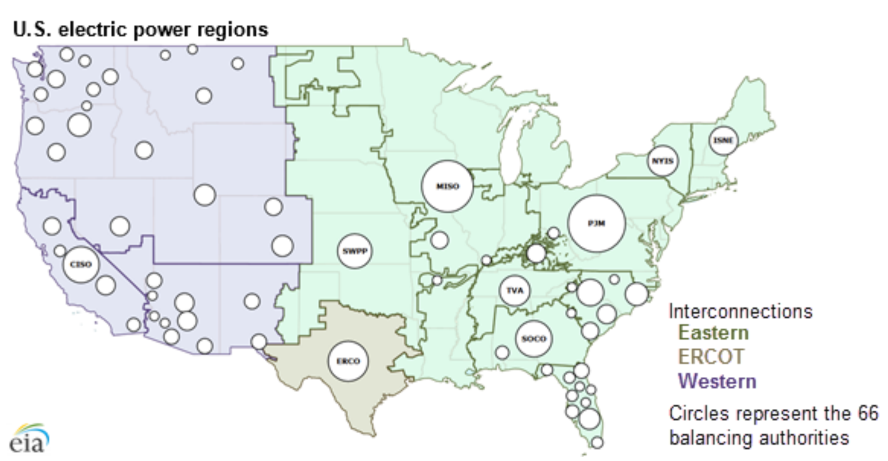
Geodiversification for Small Data Centers

Jared Polonitza and David Chiu (Advisor)
Computer Science / Mathematics
University of Puget Sound
Tacoma, WA
Made Possible by The Washington Space Grant and NASA

Est. 1888 UNIVERSITY of PUGET SOUND

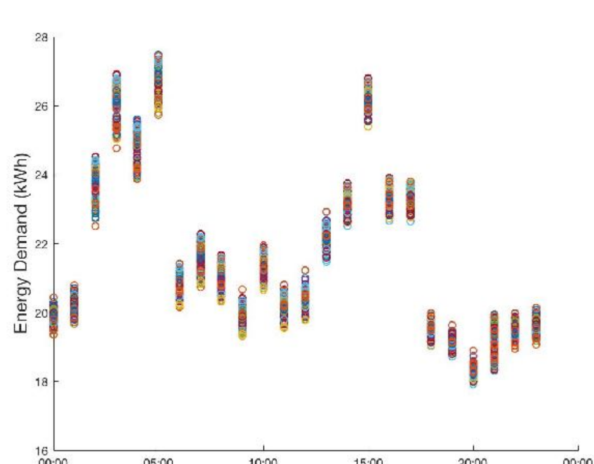
A Turbulent Power Grid

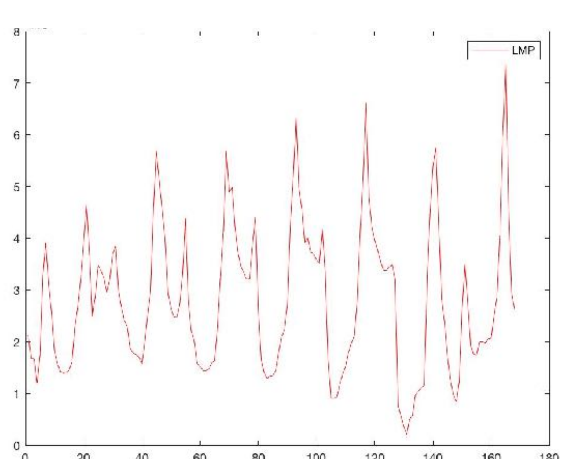


» The US energy grid is made of 3 major Interconnections: Eastern Interconnection, Western Interconnection, and ERCOT

» Over the years there have been many functional issues and changes made to the power grid, many of these changes have been due to the rapid growth of the US population, and the recent efforts to incorporate green energy solutions into US energy distribution

LMP as a Response

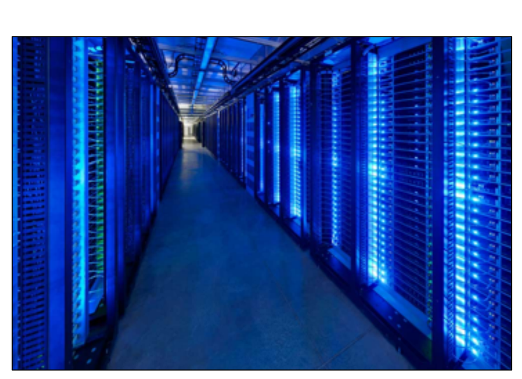


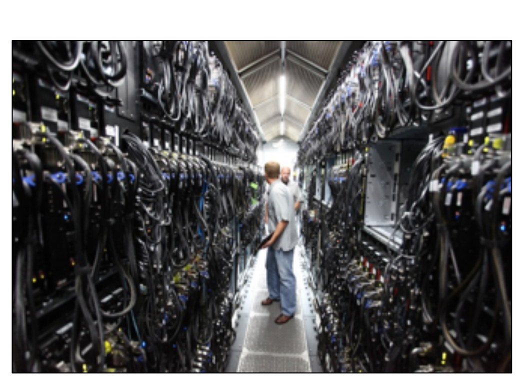


- No way to predict when energy is generated
- No good way to store excess energy
- Too expensive to store excess energy
- Puts the onus of energy management and cost reduction on the participants in the power grid

How Does This Affect Data Centers?

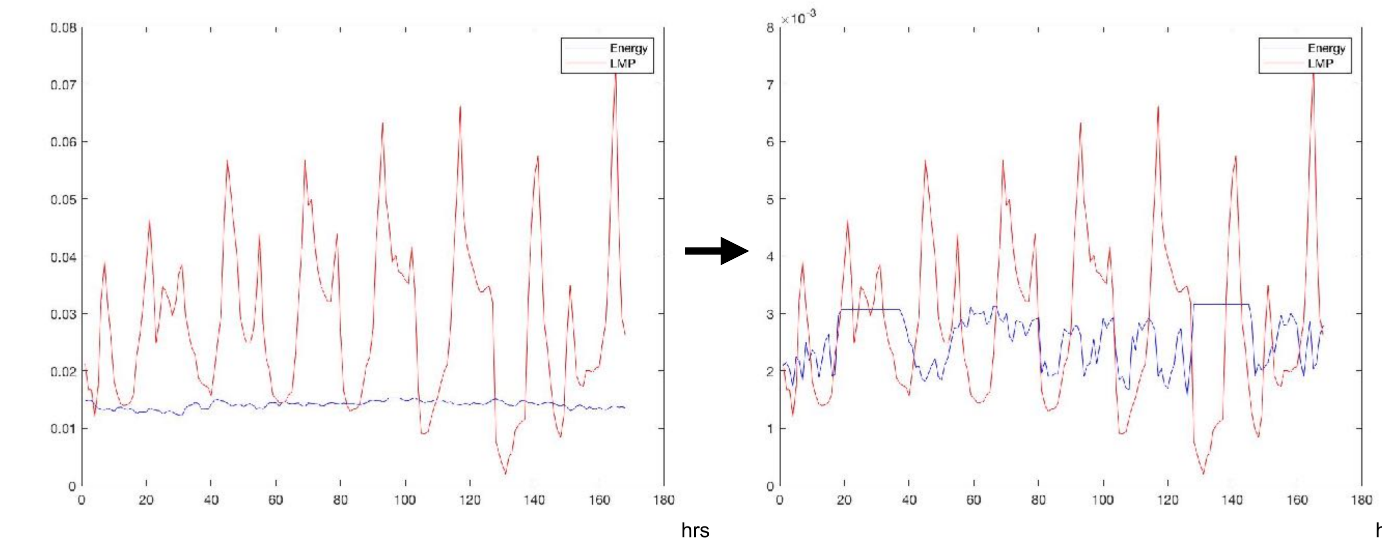
- Data Centers consume a massive amount of power
  - Account for about 2% of total energy used per year
  - Most costs incurred are hard to subvert; nearly 50% of costs are incurred due to cooling
  - This makes it difficult for small data centers to cut costs
- Large companies like Amazon and Microsoft take advantage of LMP via geo-redistribution of work.
- Use dynamic energy pricing as mechanism for demand-response





What if small data centers could copy Amazon?

- Most data centers in the US are small and cannot geodiversify the way Amazon and Microsoft can. The goal of our markets is to allow just that
  - Given a set of collaborating data centers, find out which jobs should be migrated to a certain location to minimize overall costs
- **Quick Thought Experiments:**
  - What if data centers could determine times that they, or their surrounding area are under peak load and off load jobs to reduce cost?
  - Would having a vast grid of Data Centers working together to process jobs rival something of the likes of AWS?



We want to see energy usage become something more dynamic, taking flat energy usage, and transforming it into something that adapts to the modern market

Market 1: Geodiversification via Sale

- In order to simulate the diversification system used by large companies with diverse placements of data centers, we propose allowing small centers to sell work to each other
- When jobs are sold, the center that was the original owner is paid for the work already done, then the job is packaged and sent to the next center
  - Tests have shown reductions to both incurred costs, and energy usage
  - This market allows centers to avoid working during peaks
  - Reduces the amount of wasted time centers spend working on jobs that will eventually be killed by system or end user

Market 2: Uber for the Computational Market

- Create a market equivalent to uber, ie individual users can sign their clusters up to participate in the market
- Use online market, similar to AWS in function, to accept jobs from users:
  - As jobs come in, system will monitor which centers are under the least stress / which are experiencing cheap energy prices
  - Jobs sent to the cheapest centers available at all times
  - Rather than reducing time spent on jobs that will be killed, this focuses solely on taking advantage of price fluctuations
  - Estimating realistic buy/sell sizes for transactive energy markets

Methodology:

- In order to test a real world market environment, we created a simulation model of the us power grid. Within this power grid we modeled the real time processing a data center does to complete tasks, and based off that calculated cost of operation on a minute by minute basis, tracking things like cost revenue, jobs completed, jobs failed, etc
- To best facilitate jobs being both completed, and sold in a reasonable fashion, we based our market engine on optimizing the amount of slack that each transferred job has. This means the engine will only suggest jobs that have the highest chance of being completed upon transferral to the end user

$$\underset{M}{\textbf{Maximize}} \quad slack = \sum_{j=1}^{m} \left(dl_j - (t + remaining\_time_j) \right)$$

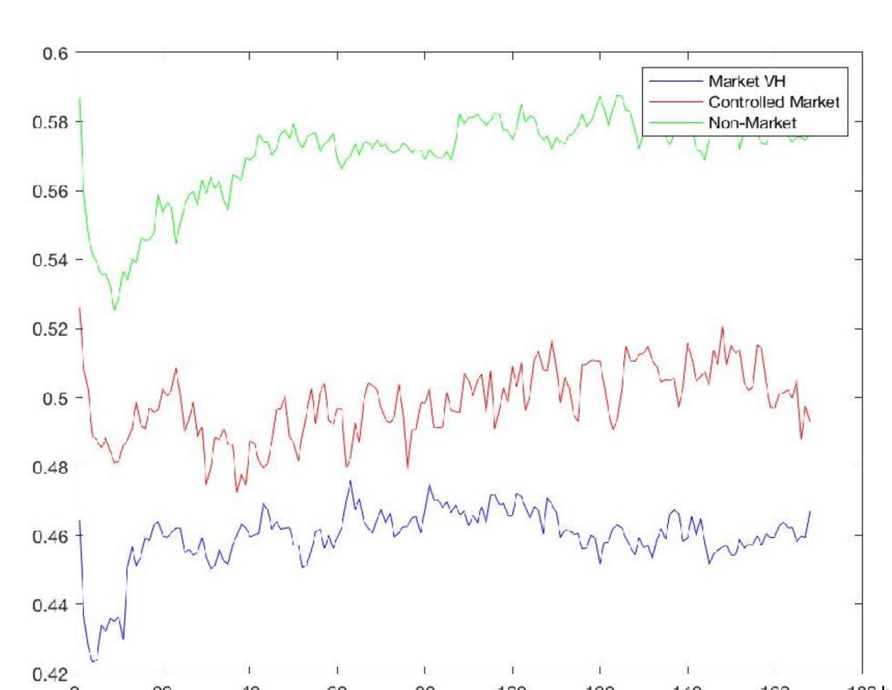
$$\textbf{Subject to} \quad \sum_{j=1}^{m} M_j \le m$$

$$\sum_{j=1}^{m} \left(M_j \cdot \sum_{s=1}^{S_j} p_s(t) \right) \ge \frac{budget}{\tau} \cdot \frac{1}{r(t) - r(t-\tau)}$$

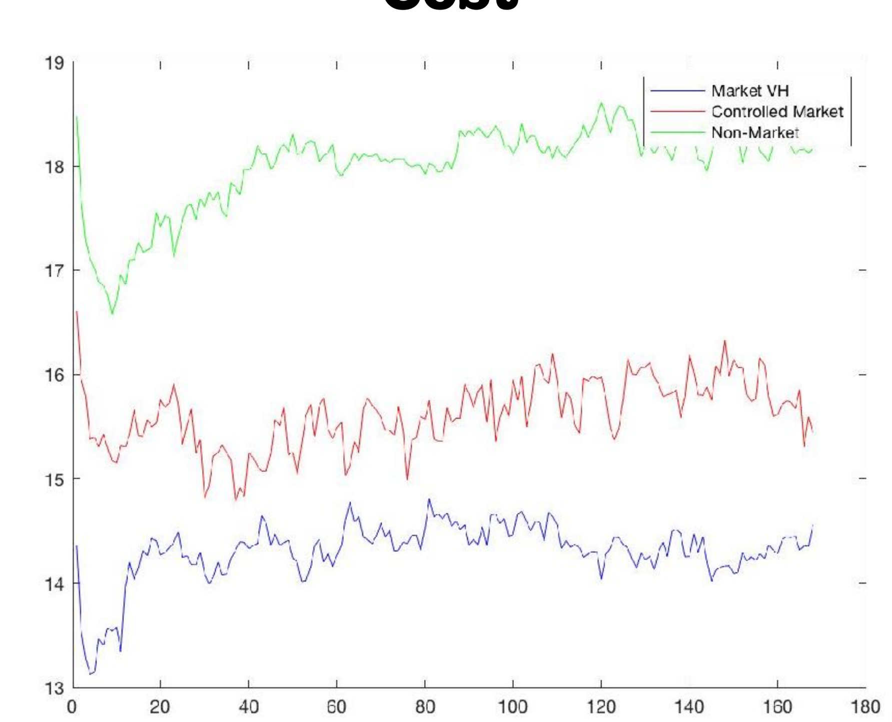
$$remaining\_time_j = \hat{t}_j(S_j) + M_j \cdot mig\_time_j, \ \forall\, 1 \le j \le m$$

Results: Energy / Cost / Revenue

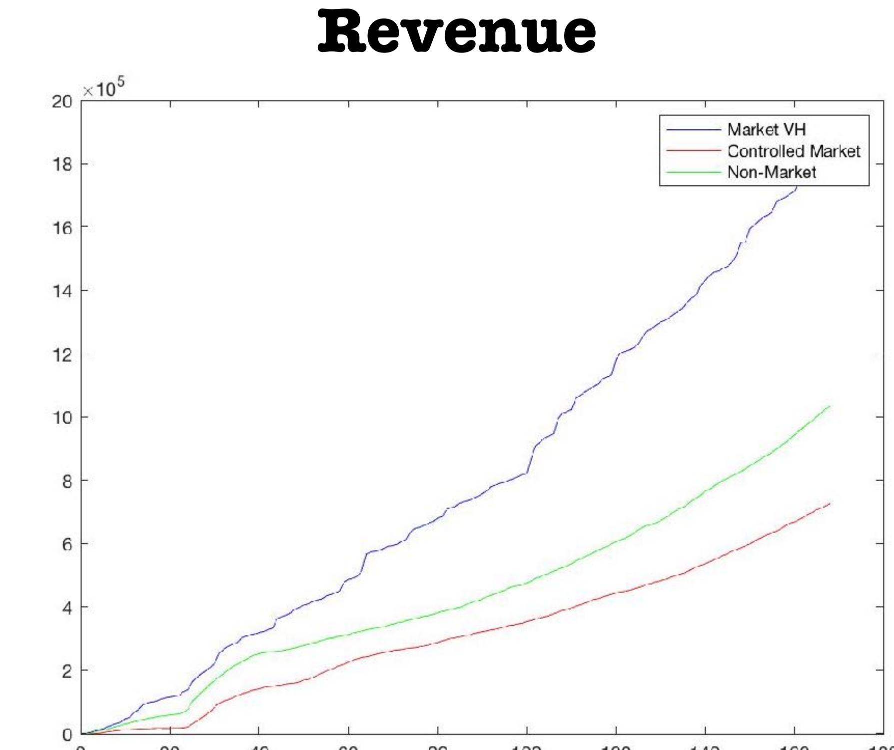
Energy



Cost



Revenue



These are the averages of running the simulation five times. Each of these lines represents the aggregate values for 25 data centers of varying size (1-5) clusters, participating from all across the US. The green line represents the average results for clusters not participating in either market. The red line is the uber analogue, and the blue line is our sales market, where each center participated at random intervals between every one to three hours.

Both of the markets have merit in terms of global reductions to cost and energy usage for data centers. These savings in turn show promising increases in revenue for participating centers.