Temperature Aware Workload Management in Geo-distributed Datacenters

Hong (Henry) Xu, Chen Feng, Baochun Li

Department of Electrical and Computer Engineering
University of Toronto

USENIX ICAC, San Jose, CA. June 28, 2013

Geo-distributed datacenters

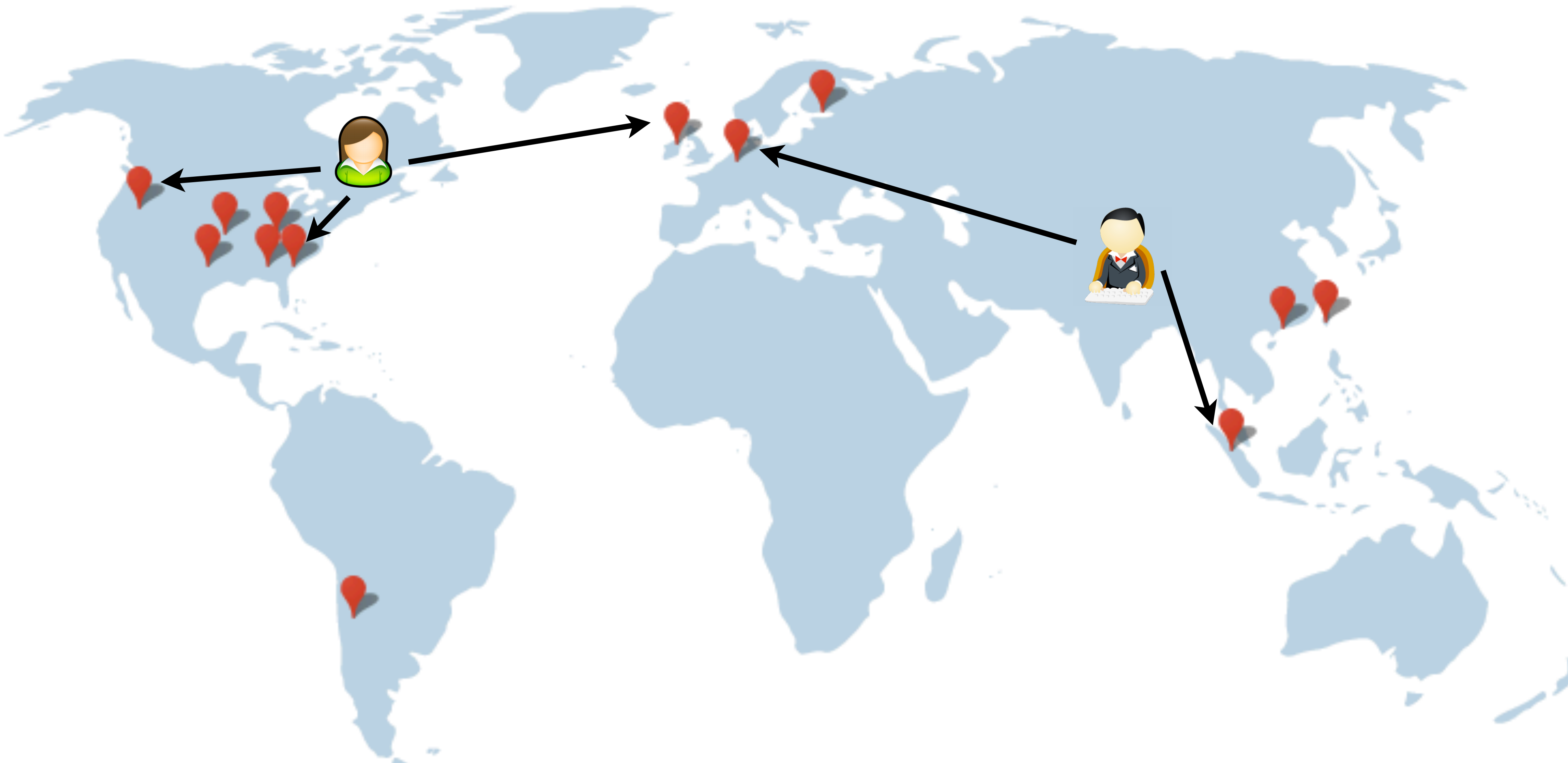
Google

Source: Google

Microsoft

facebook

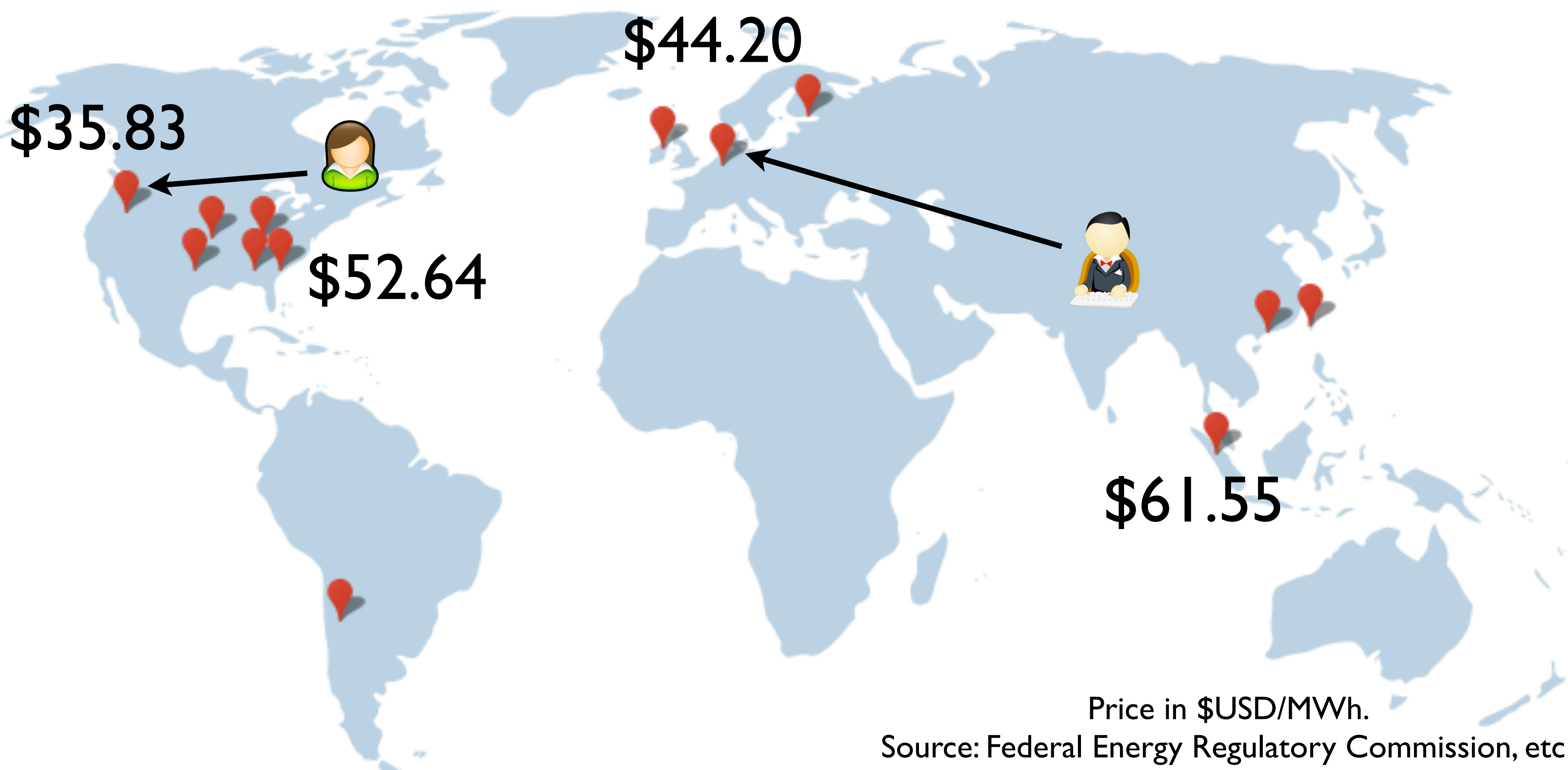
Request routing



Data is replicated across the wide area

How to route requests to datacenters?

Prior work



Price in $USD/MWh.
Source: Federal Energy Regulatory Commission, etc.

Price aware request routing:

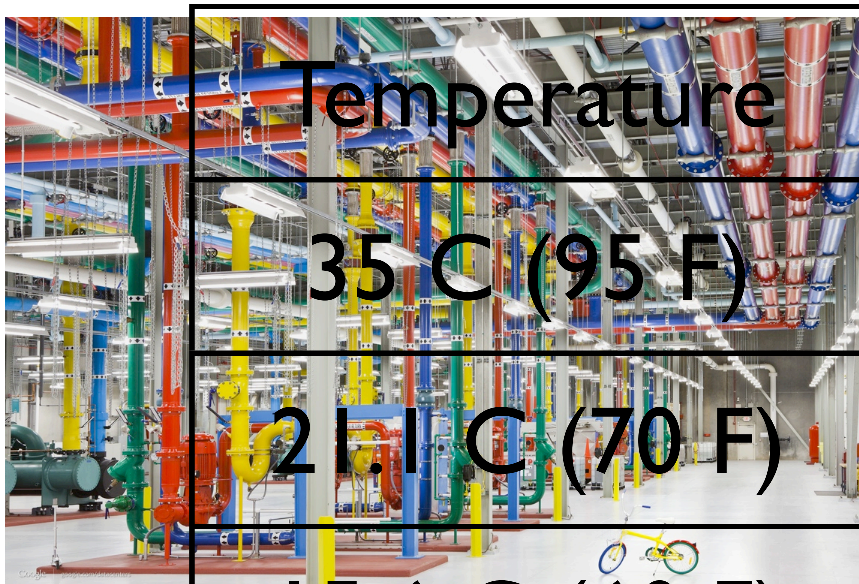
A. Qureshi et al., *Cutting the Electricity Bill for Internet-scale Systems*, SIGCOMM 2009

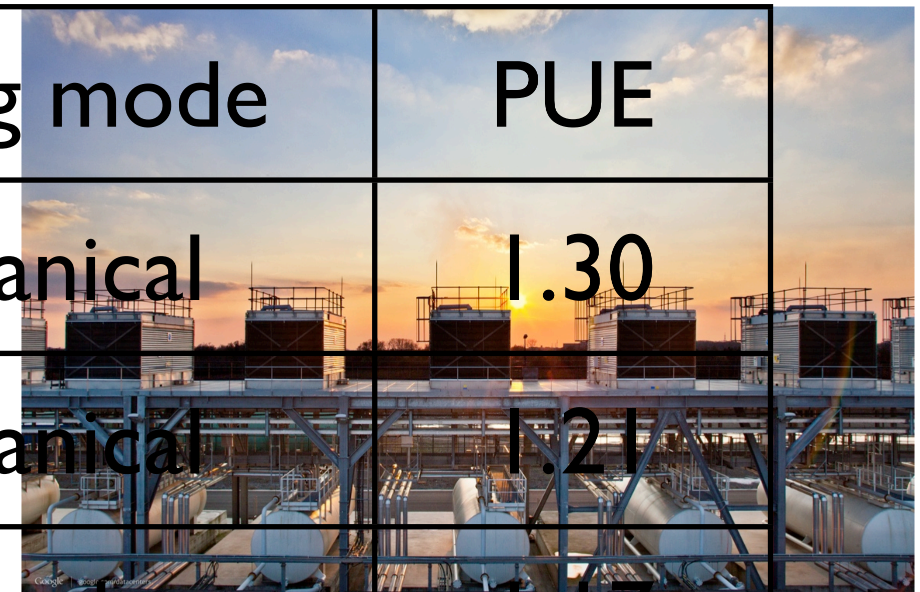
Z. Liu et al., *Greening Geographical Load Balancing*, SIGMETRICS 2011

Two missing aspects...

Cooling system

Cooling energy efficiency (PUE) is NOT a constant

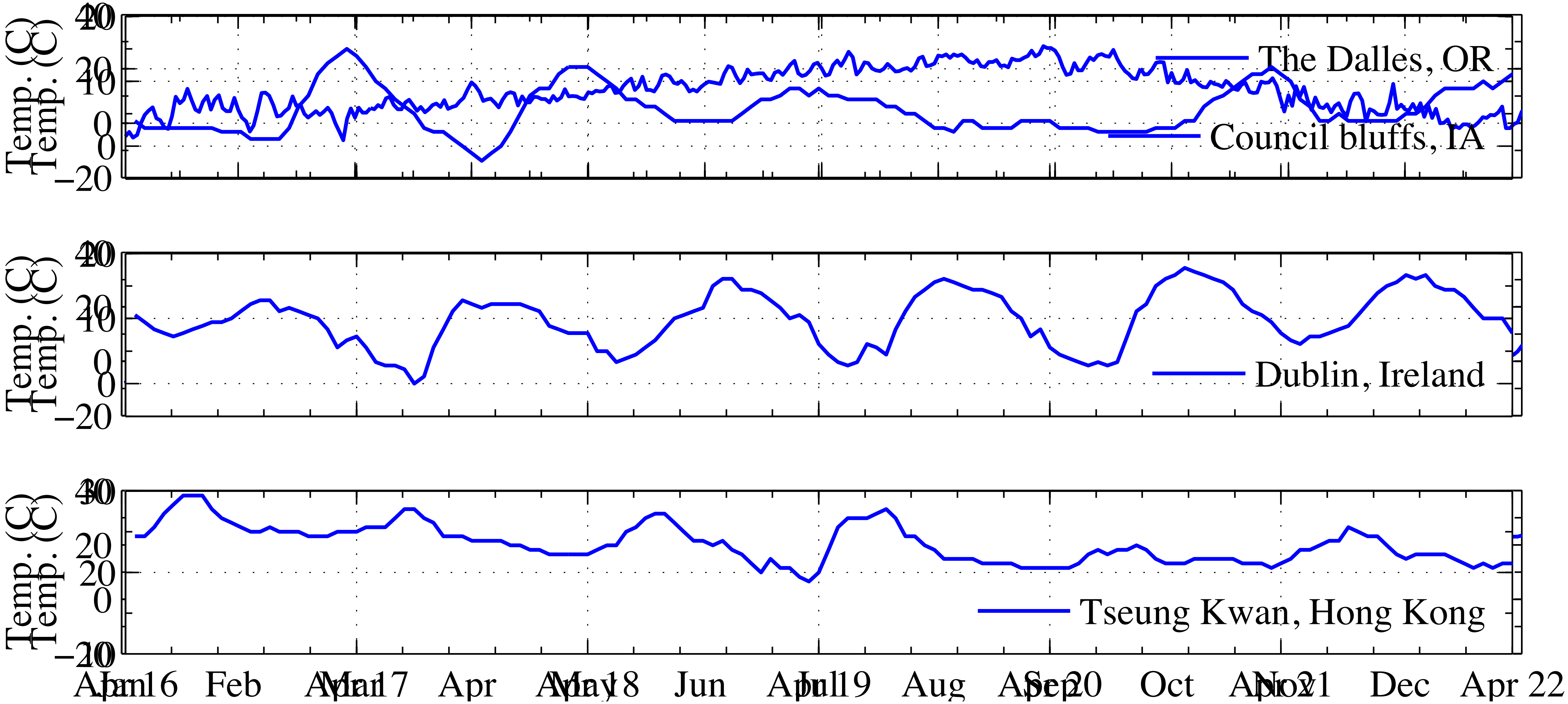




Temperature	Cooling mode	PUE
35 C (95 F)	Mechanical	1.30
21.1 C (70 F)	Mechanical	1.21
15.6 C (60 F)	Mixed	1.17
10 C (50 F)	Outside air	1.10
-3.9 C (25 F)	Outside air	1.05

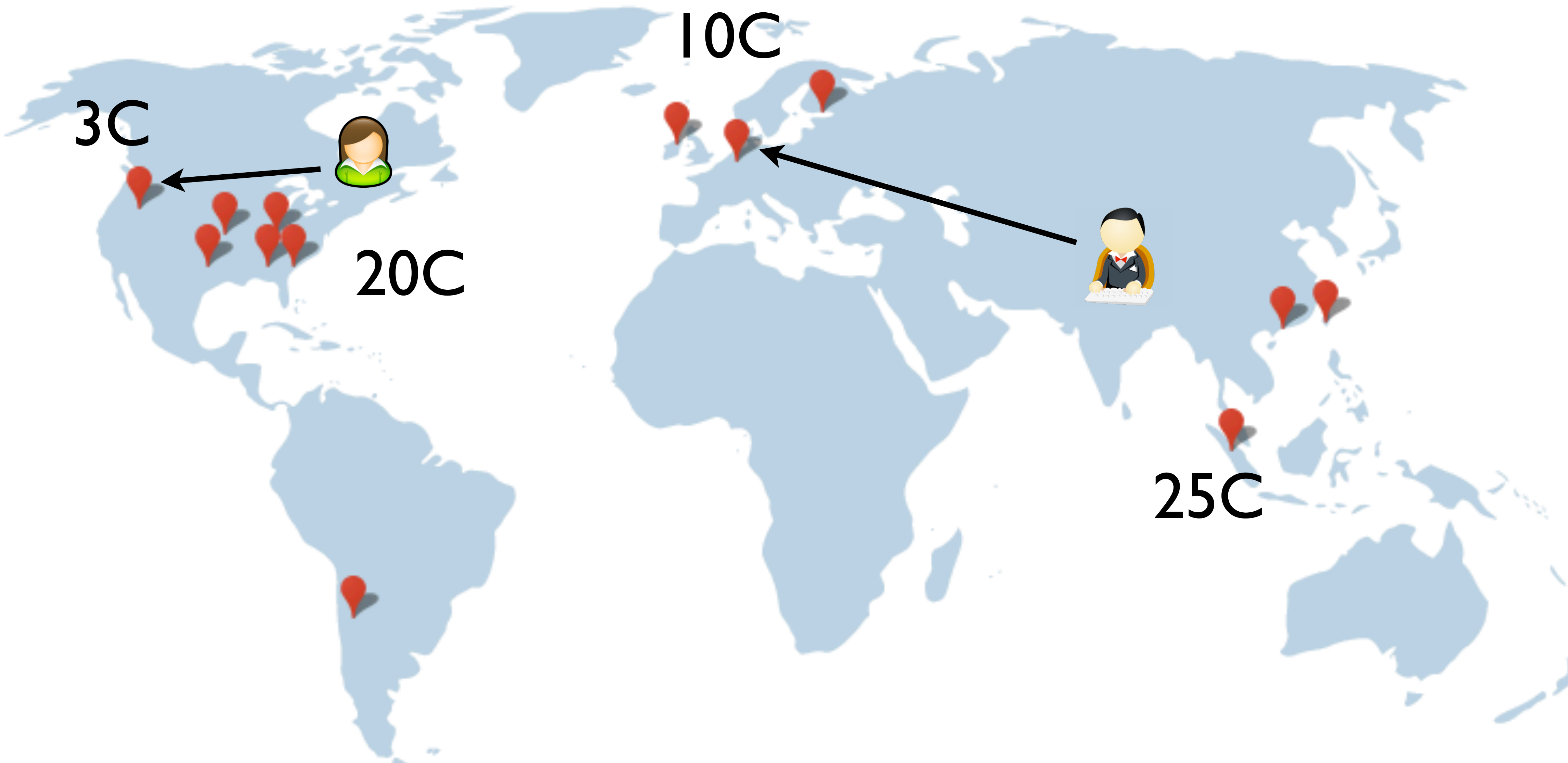
Source: Emerson® Liebert DSE™ cooling system with an EconoPhase air-side economizer

Temperature diversity



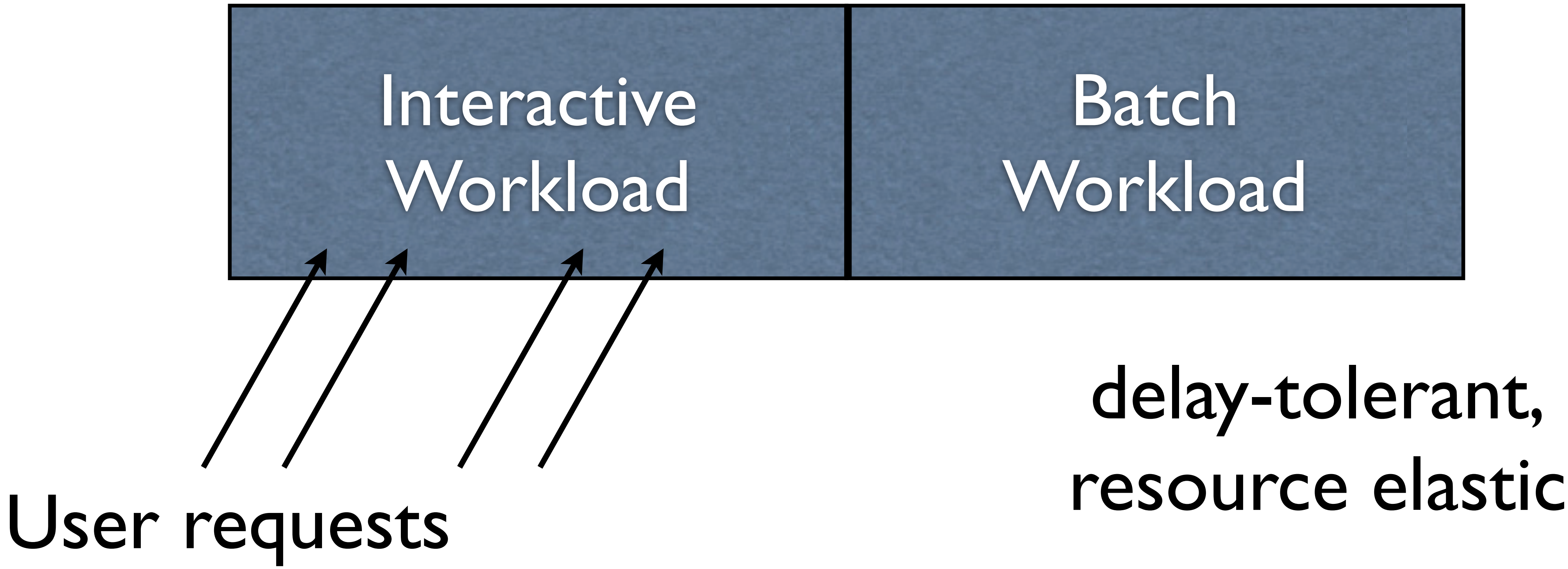
Selected Google DC locations. Source: National Climate Data Center

First idea



Route more requests to cooler locations to reduce energy consumption and cost.

Second idea



At cooling efficient locations, allocate more capacity to interactive workload.

Capacity allocation is fixed

This work

Temperature aware workload management

1. System model and formulation
2. A distributed optimization algorithm (ADMM)
3. Trace-driven simulations

System model [1/2]

	In our model	In reality	
User	A unique IP prefix	Common practice, e.g. Akamai [34]	✓
Request traffic	Arbitrarily splittable among datacenters	Common practice, e.g. DNS, HTTP proxies [17,29,34]	✓
Time scale	Hourly optimization	Common practice, traffic predictable, electricity price known [28, 32, 34]	✓

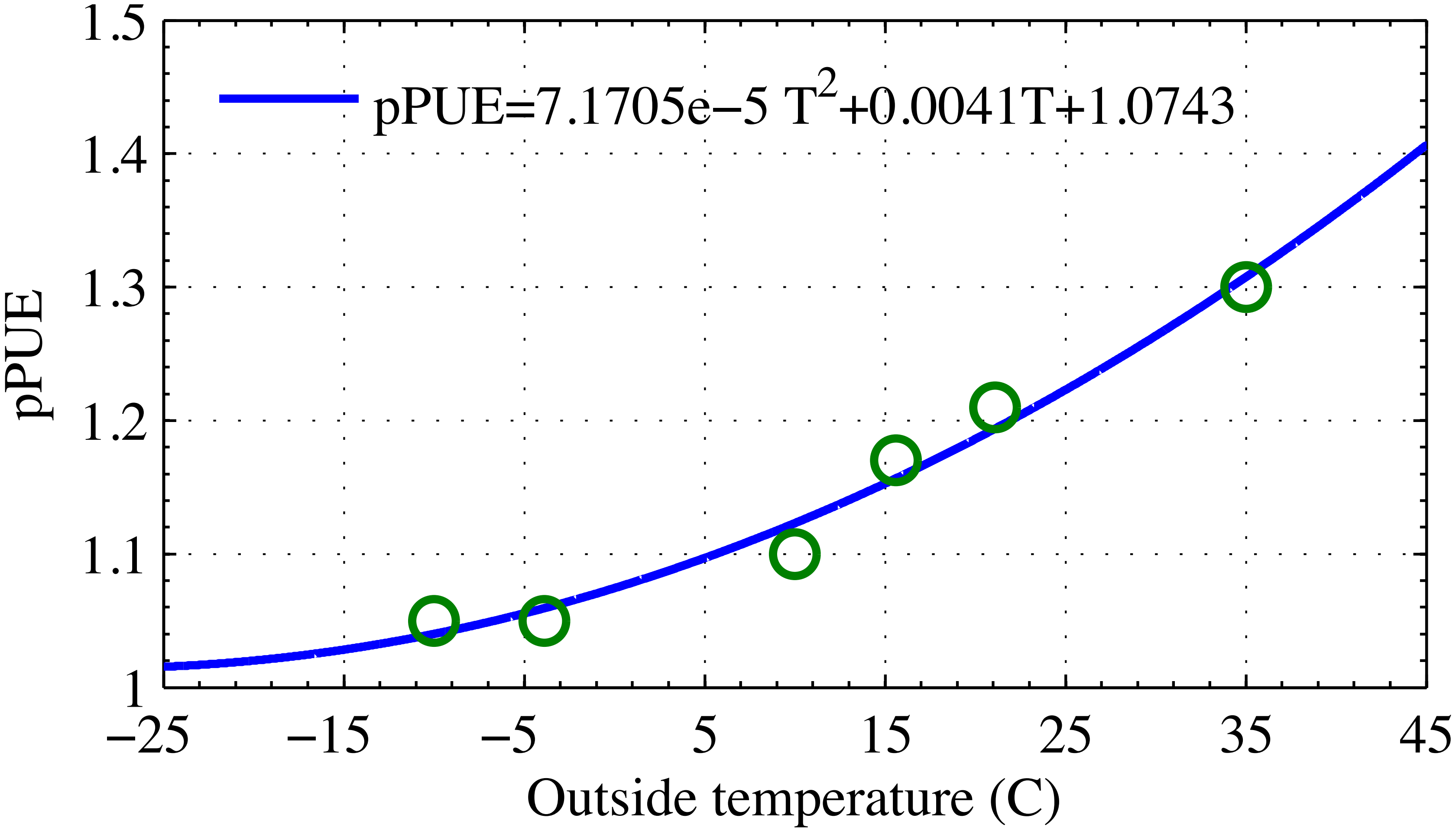
System model [2/2]

Energy cost at data center *j*:

$$E_j(W_j) = \underbrace{(C_j P_{\text{idle}} + (P_{\text{peak}} - P_{\text{idle}})\, W_j)}_{\text{Google cluster measurements [15]}} \cdot \underbrace{\text{pPUE}(T_j)}_{\text{Our empirical data}} P_j$$

Google cluster measurements [15]

Our empirical data



Formulation

$$\boxed{\text{User } i} \xrightarrow{\alpha_{ij}} \boxed{\text{Datacenter } j} \qquad \beta_j\text{: Batch workload}$$

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}\succeq\mathbf{0}} \quad \sum_j E_j\Big(\sum_i \alpha_{ij}\Big) + \sum_i U_i(L(\boldsymbol{\alpha_i}))$$

Latency

Energy cost + Utility loss

$$+ \sum_j E_j(\beta_j) + \sum_j V_j(\beta_j)$$

Revenue loss due to performance

$$\text{s.t.:} \quad \forall i : \sum_j \alpha_{ij} = D_i,$$

Workload conservation

$$\forall j : \sum_i \alpha_{ij} + \beta_j \leq C_j.$$

Capacity constraint

Challenges

Convex optimization

Large-scale problems

- $O(10^5)$ IP prefixes, $O(10^7)$ variables, $O(10^5)$ constraints

Distributed optimization algorithms

- Dual decomposition with subgradient methods
- Two drawbacks:
  - Delicate step size adjustment
  - Very slow convergence

ADMM

Alternating Direction Method of Multipliers

[S. Boyd et al., 2011]

Fast convergence for large-scale distributed convex optimization in data mining and machine learning

Limitation: It only works for problems with 2 sets of variables linked by an equality constraint

Does NOT work for our problem

Generalized ADMM

Minimize utility loss for interactive

penalty(α^k, a^{k-1}) α^k $\downarrow$ per-user sub-problems

Minimize energy cost for interactive

penalty(a^k, α^k) a^k $\downarrow$ per-DC sub-problems

Minimize total cost for batch

penalty(β^k, α^k) β^k $\downarrow$ per-DC sub-problems

Dual update

Convergence

Theorem: Generalized ADMM converges to the optimal solution.

It works for problems with any sets of variables.

Applicable to problems in other domains.

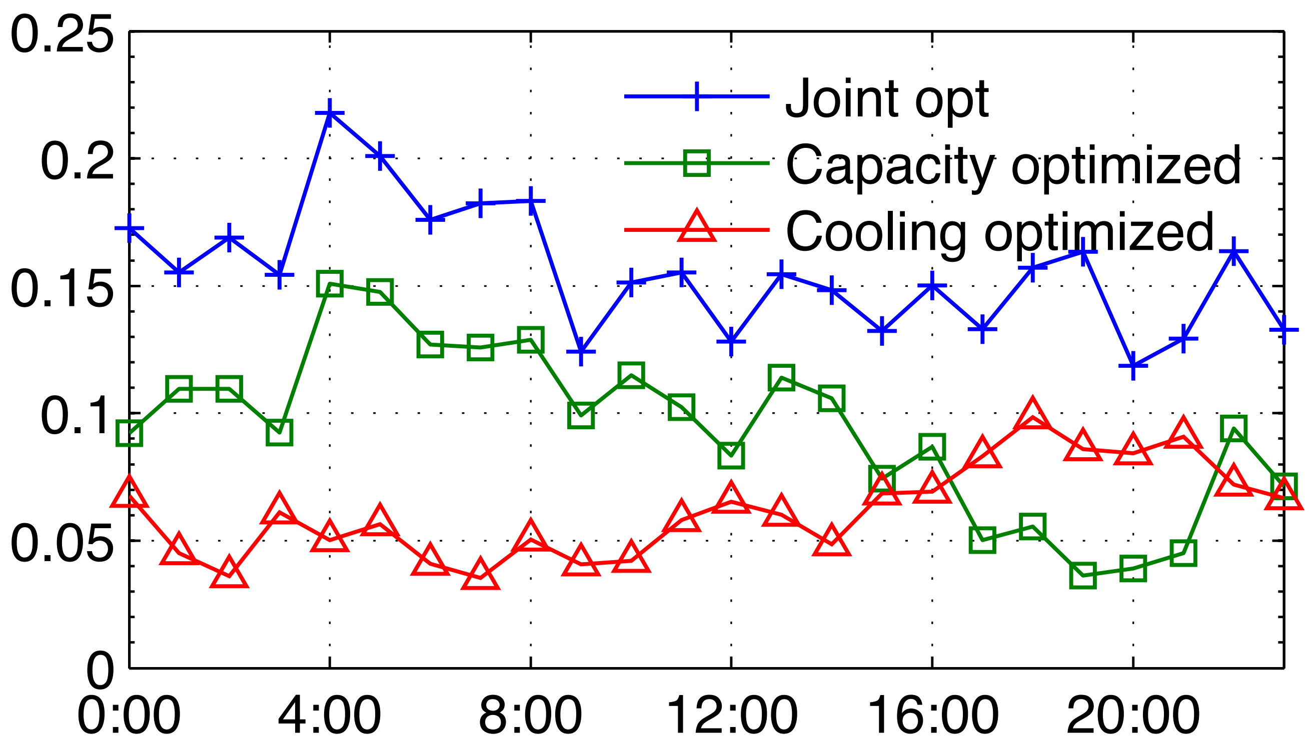
Evaluation: Setup

Google DC locations, Wikipedia request traces, empirical temperature, latency, and electricity price data

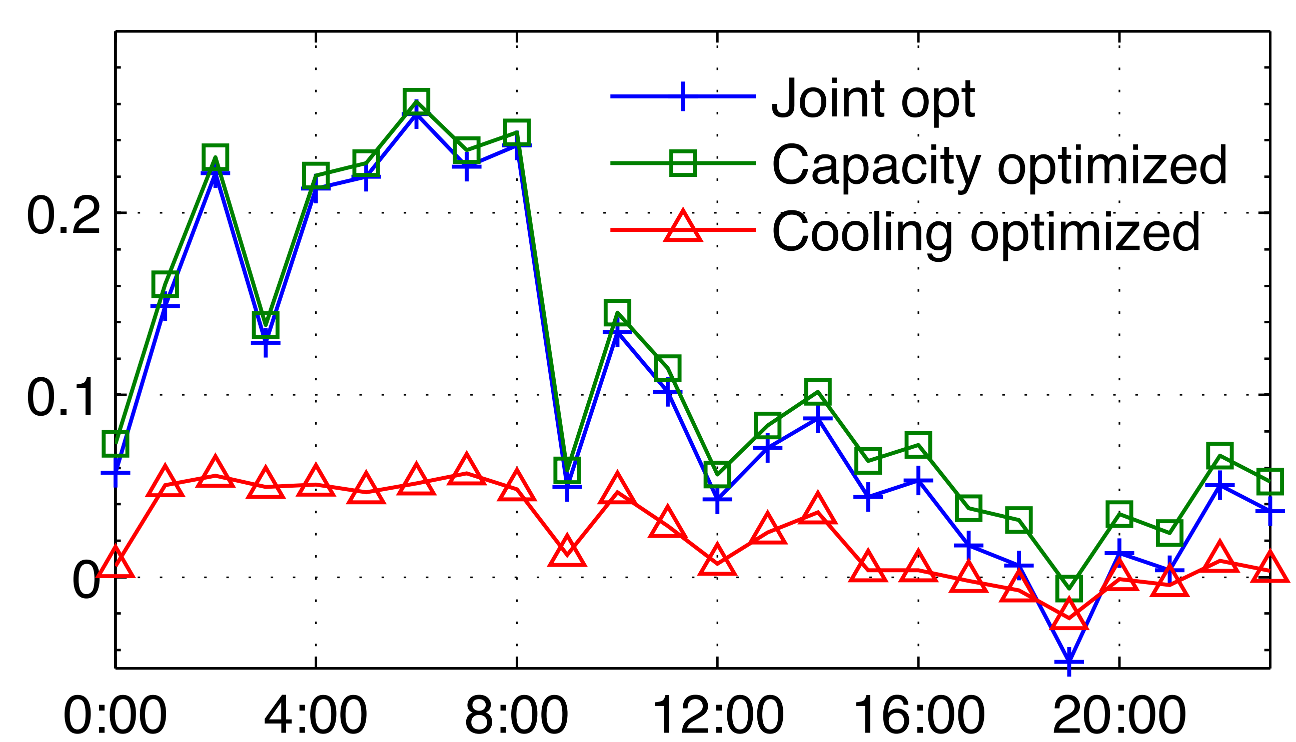
Benchmarks:

- Joint opt: Our work
- Baseline: State-of-the-art, no temperature aware request routing, no capacity allocation
- Cooling optimized
- Capacity optimized

Benefits breakdown

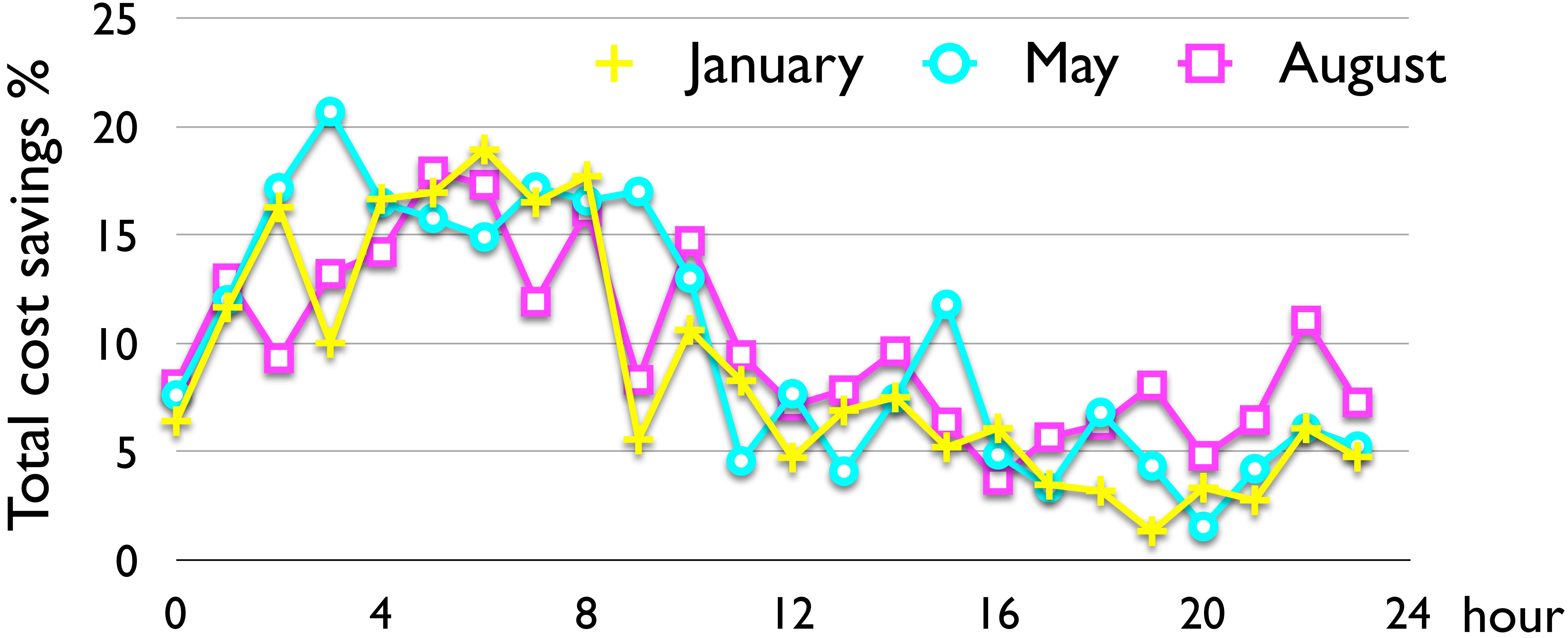


Cooling energy savings



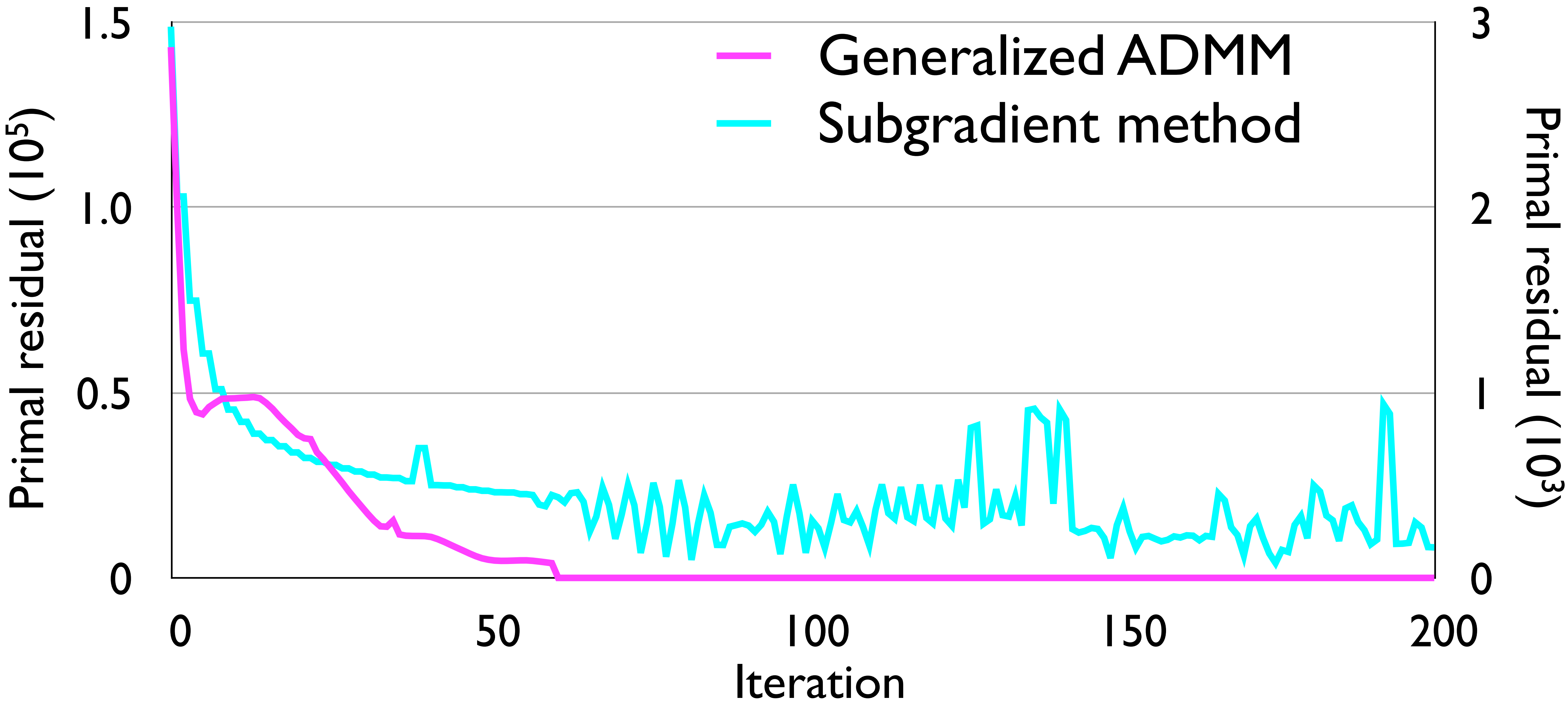
Utility loss reductions

Overall improvement



Result: 5%-20% total cost savings, consistent across seasons

Convergence



Result: Generalized ADMM converges much faster than existing algorithms.

Related work

- Workload management in geo-distributed DCs
    - A. Qureshi et al., *Cutting the Electricity Bill for Internet-scale Systems*. SIGCOMM, 2009
    - Z. Liu et al., *Greening Geographical Load Balancing*. SIGMETRICS, 2011
    - Gao et al., *It's not easy being green*. SIGCOMM, 2012
- ADMM
    - Han et al., *A note on the alternating direction method of multipliers*. J. Optim. Theory Appl. 155:227-238, 2012
    - Hong et al., *On the linear convergence of the alternating direction method of multipliers*. arXiv, August 2012

Thank you!

Google “Henry Xu”