History-Based Harvesting of Spare Cycles and Storage in Large-Scale Datacenters

Yunqi Zhang, George Prekas, Giovanni Matteo Fumarola,
Marcus Fontoura, Íñigo Goiri, Ricardo Bianchini

UNIVERSITY OF MICHIGAN

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

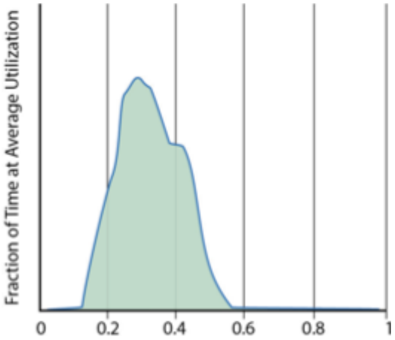
Microsoft

Microsoft Research

Datacenters are underutilized

- Datacenters are massive

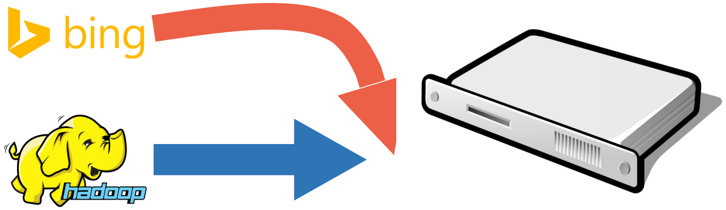
- Overprovision resources
    - Low tail latency requirement
    - Provisioned for peak load
    - Unexpected load spikes and failures
- Underutilization wastes money



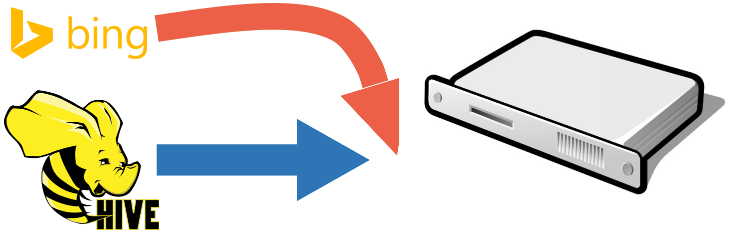
Server Utilization Distribution of a Google Cluster.

Harvesting spare resources

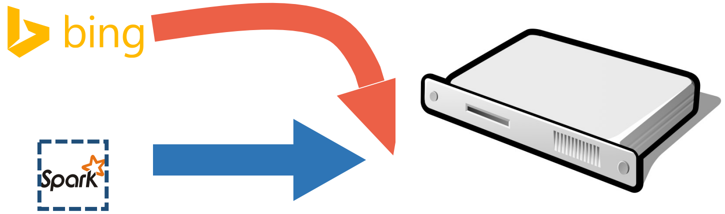
- Interactive services + batch
  - Low priority batch tasks



- Find “safe” co-locations
  - Cluster-level

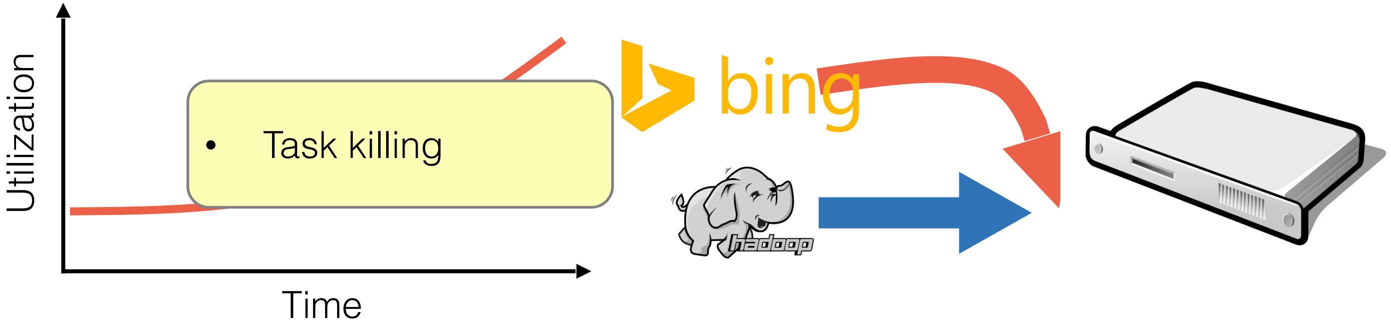


- Performance isolation
  - Server-level



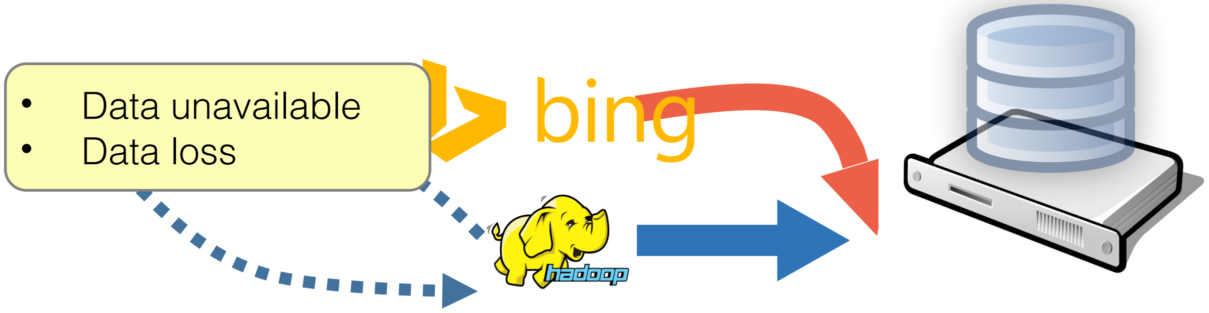
Challenges

- Interactive services “own” the servers
- Resource availability dynamics



- Task killing

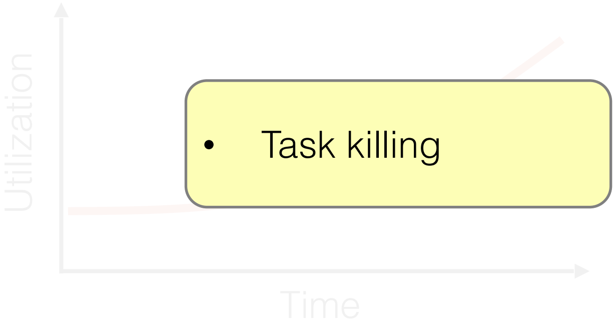
- Data storage co-location



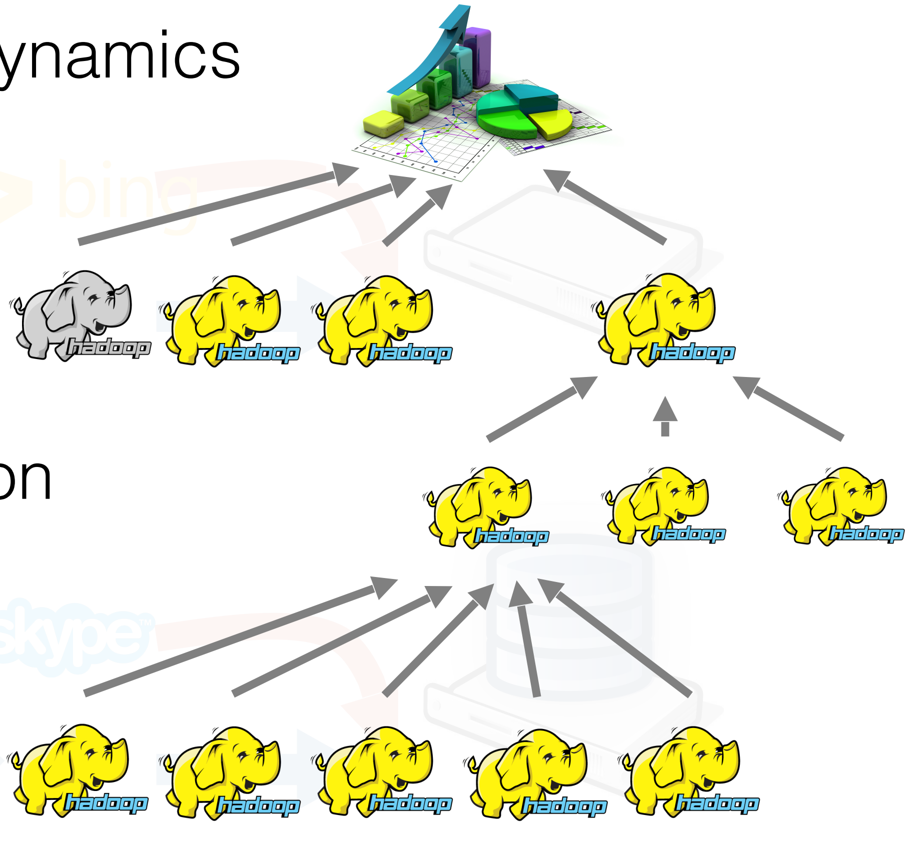
- Data unavailable
- Data loss

Challenges

- Interactive services “own” the servers
- Resource availability dynamics



- Task killing



- Data storage co-location

- Data unavailable
- Data loss

- Distributed data analytics across servers

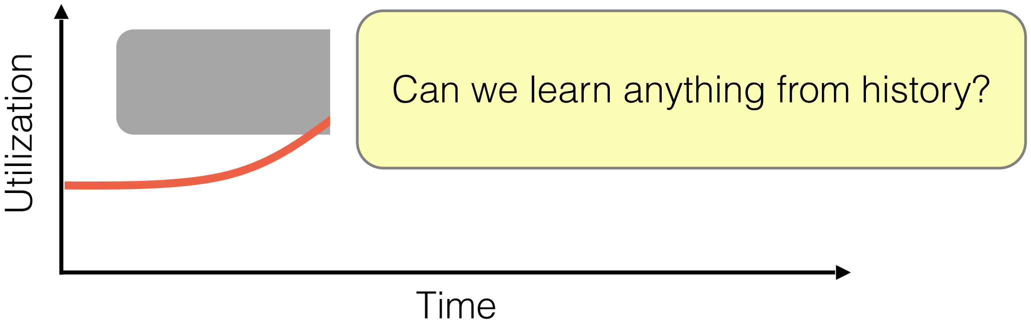
Goals

- Improve efficiency without sacrificing QoS

- Minimize the probability of killing batch tasks

- Maximize data availability and durability

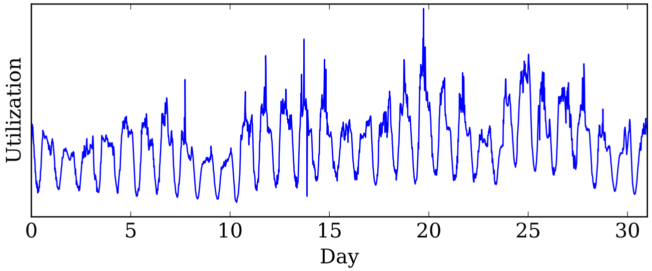
Batch task scheduling

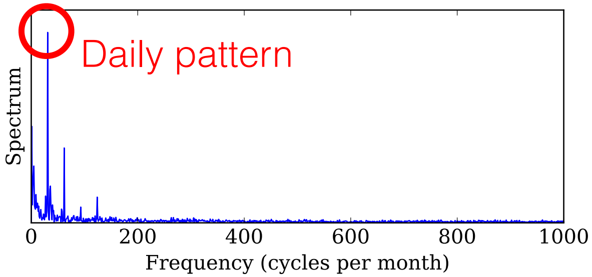


Can we learn anything from history?

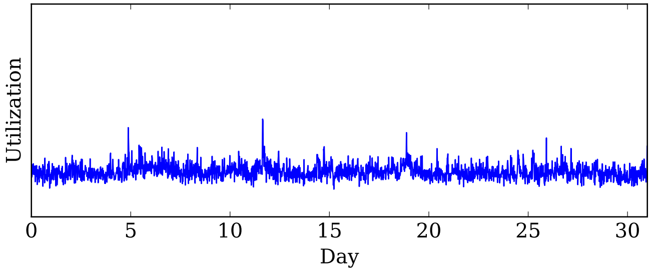
Batch task scheduling

Periodic

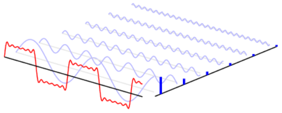


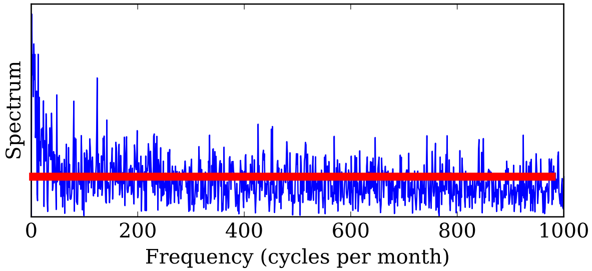

Daily pattern

Constant

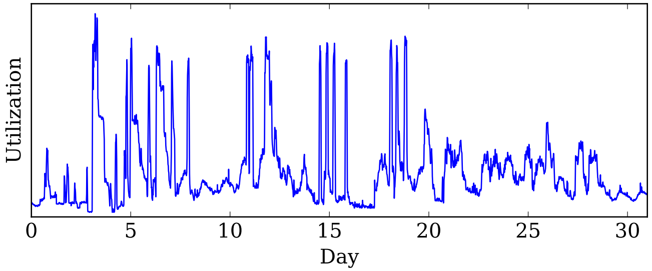


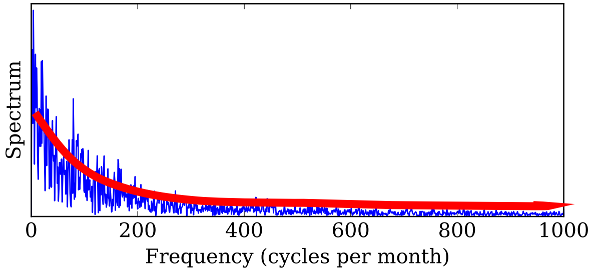
Fourier Transform





Unpredictable

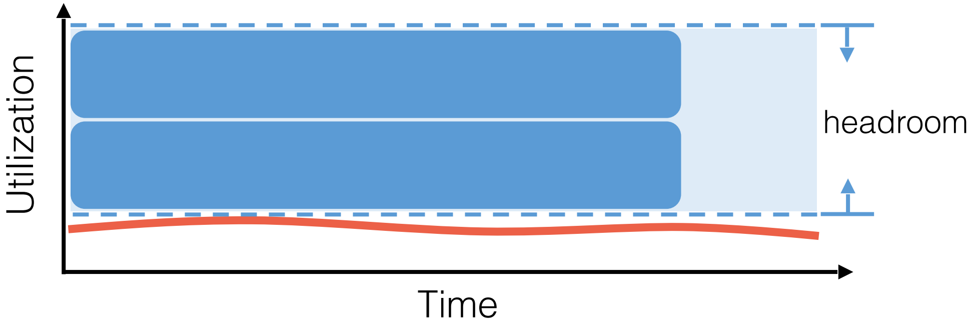




History-based task scheduling

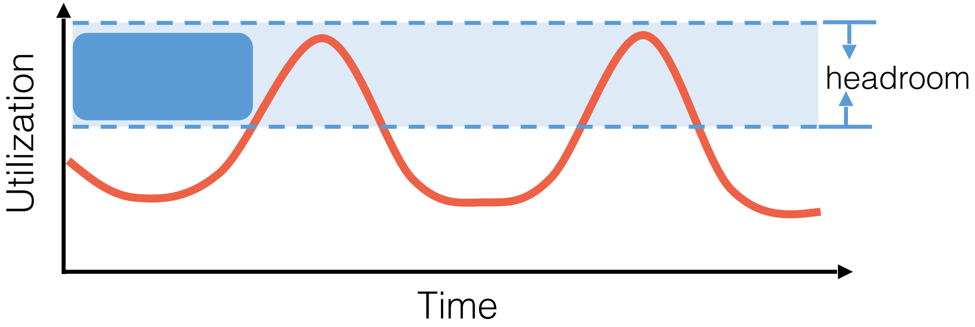
Long Jobs

- Constant
- 1 – MAX(Peak, Current)



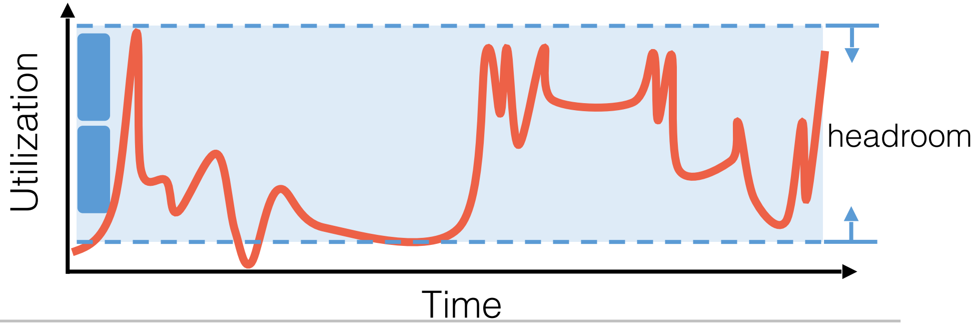
Medium Jobs

- Periodic
- 1 – MAX(Average, Current)



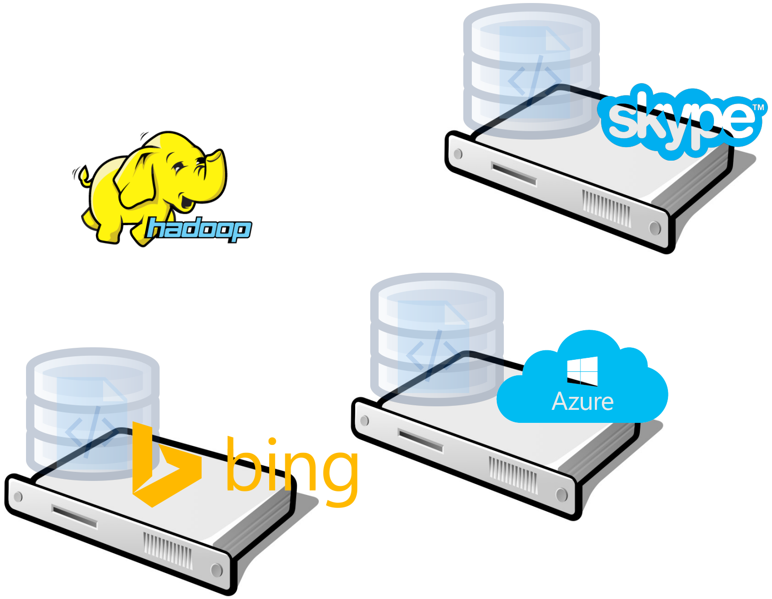
Short Jobs

- Unpredictable
- 1 – Current



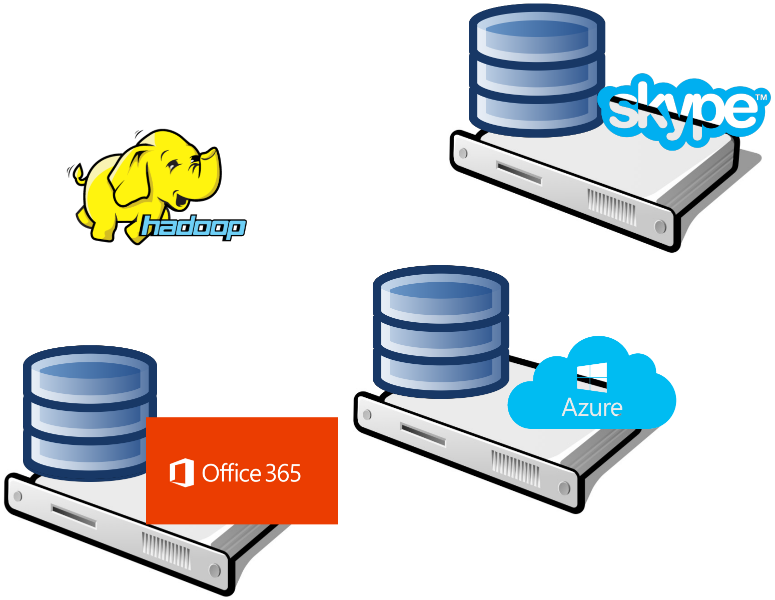
Data storage co-location

Data availability



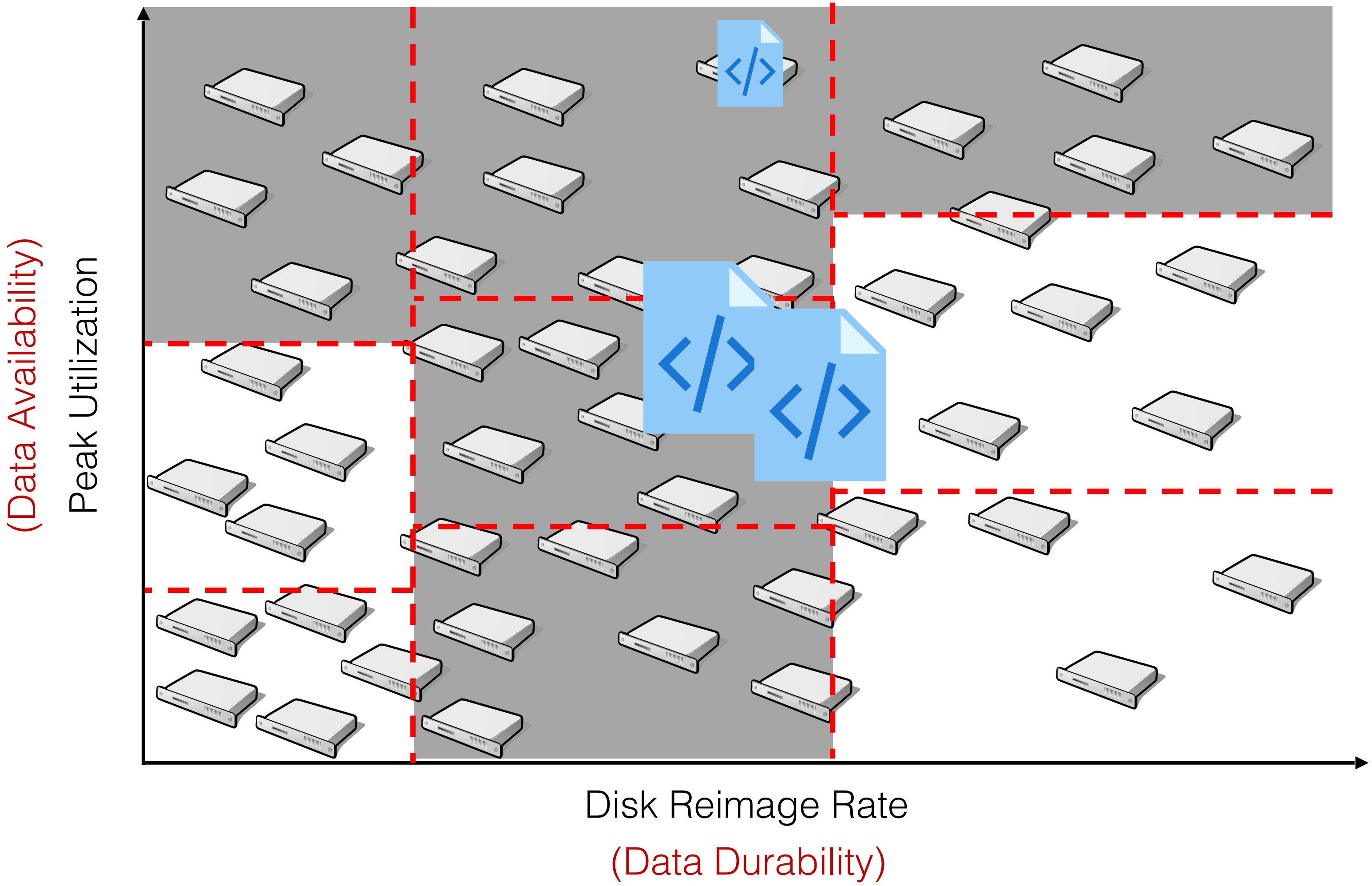
Diverse in utilization pattern.

Data durability

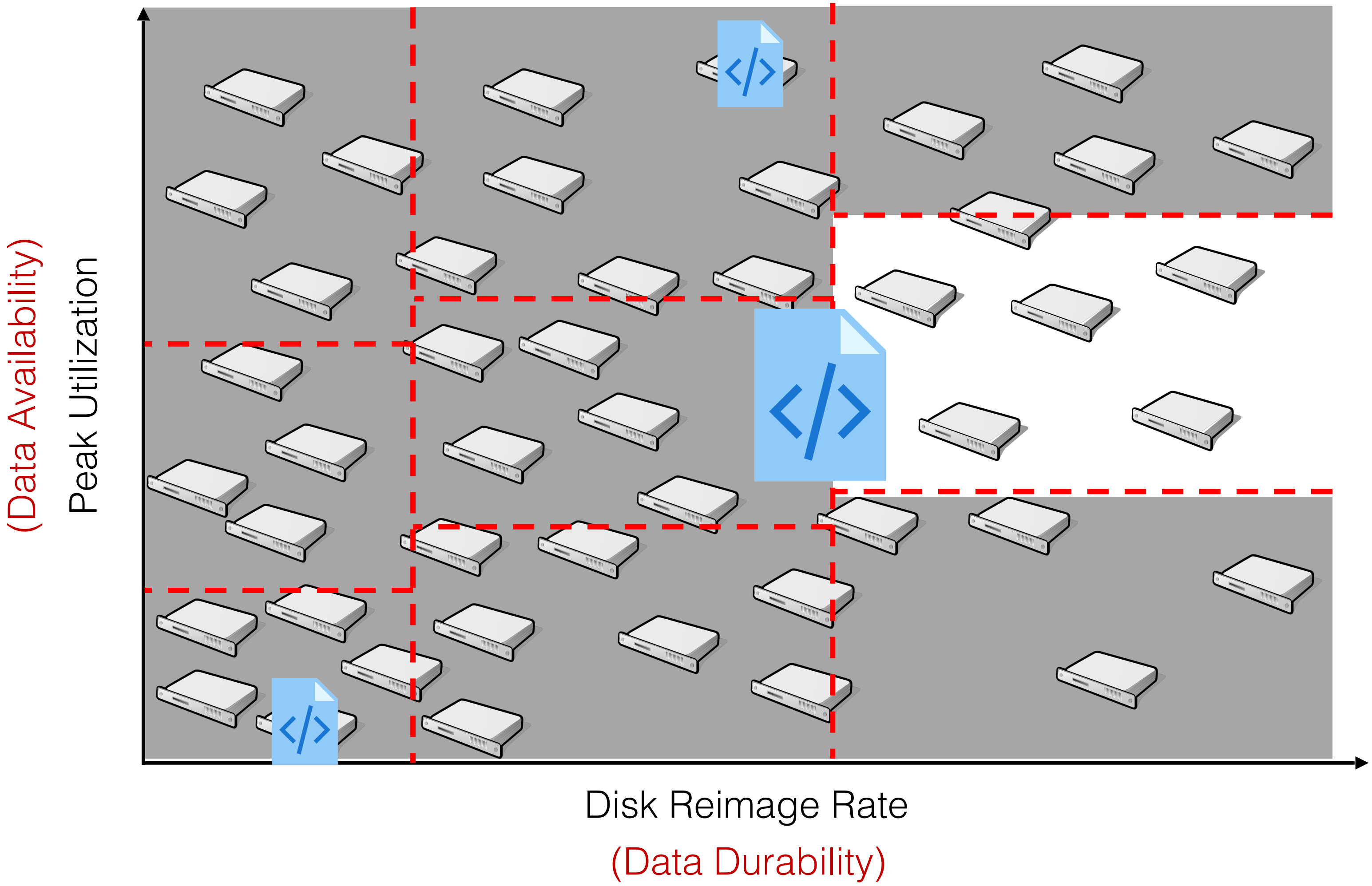


Diverse in reimaging pattern.

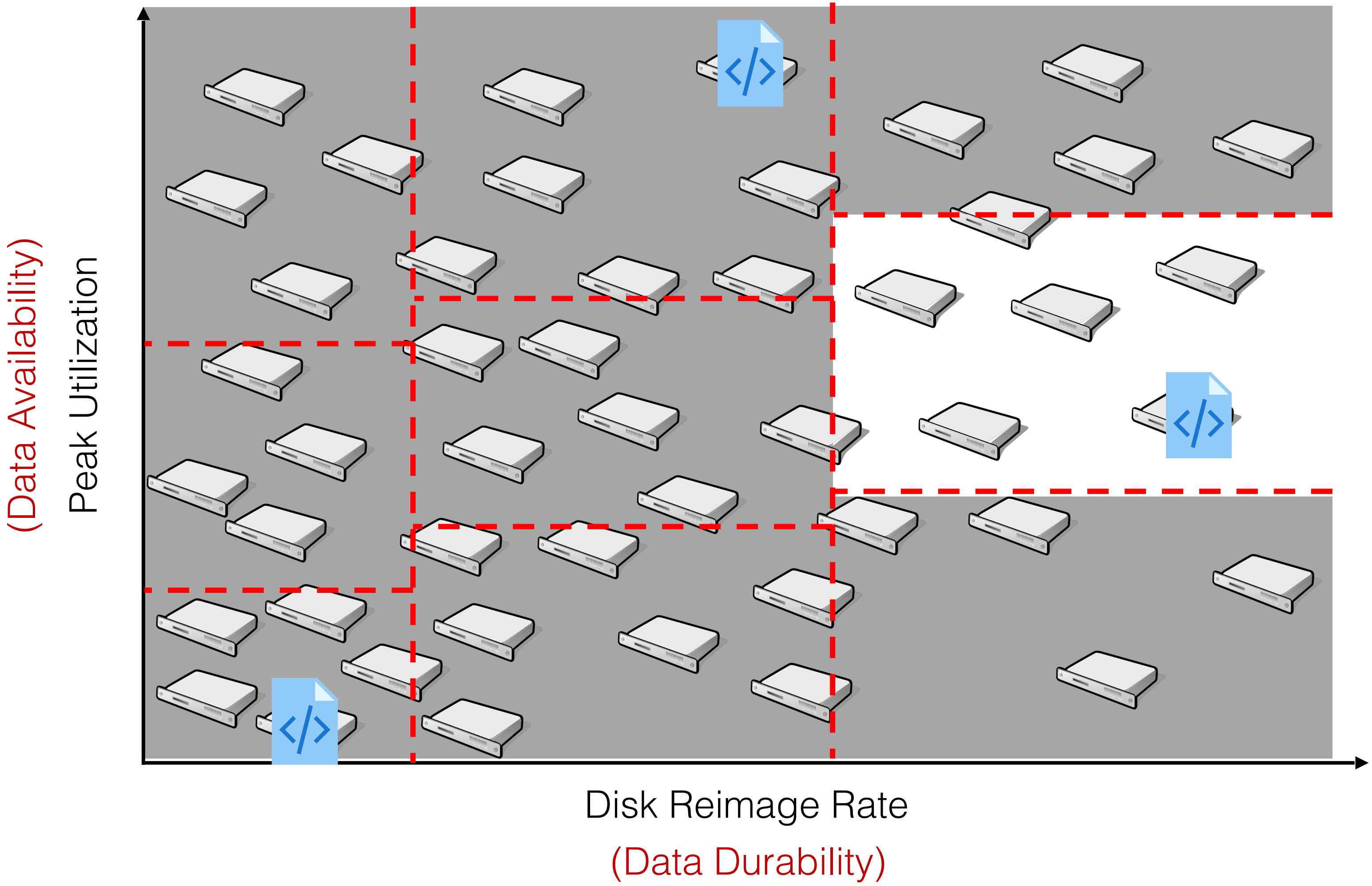
History-based replica placement



History-based replica placement



History-based replica placement



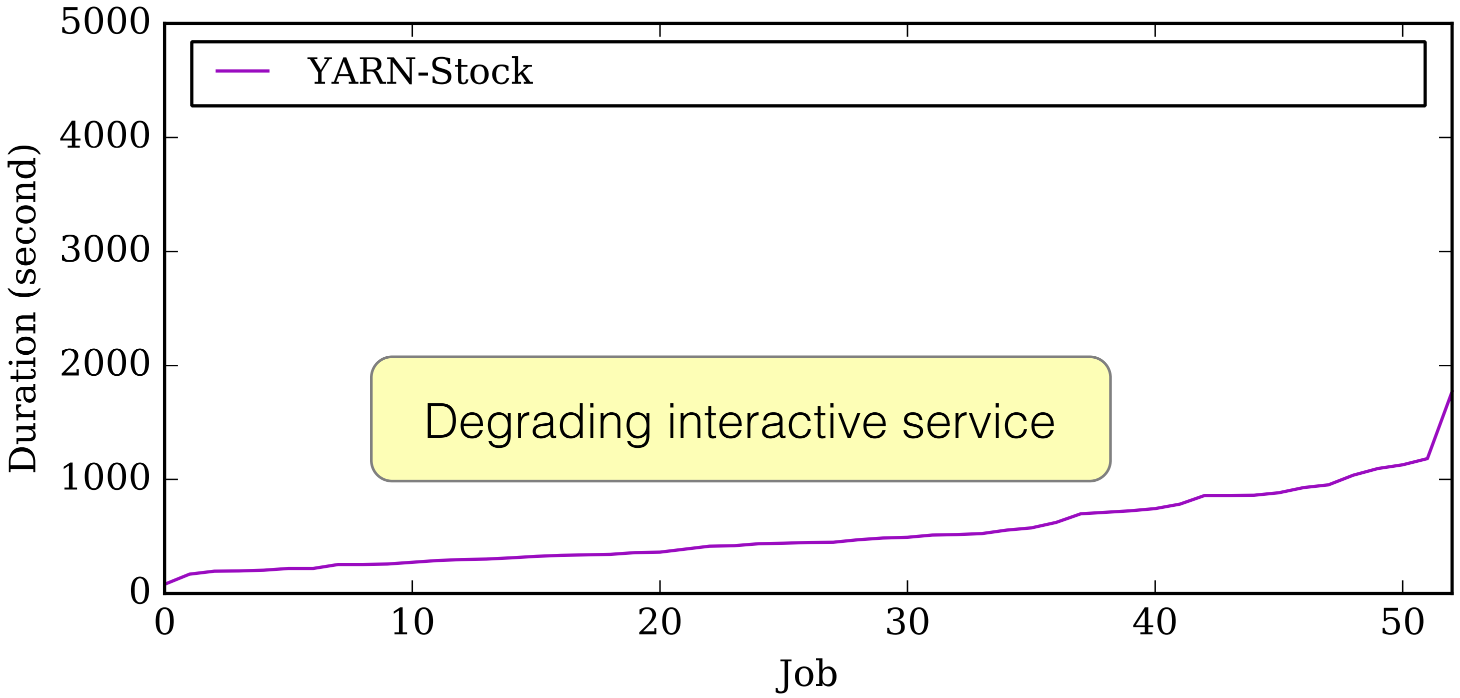
System implementation

- Clustering service
  - Extract utilization and reimaging patterns
- YARN-H
  - Protect interactive services by killing batch tasks
- Tez-H
  - History-based batch task scheduling
- HDFS-H
  - History-based replica placement
  - Protect interactive services by denying accesses

Evaluation

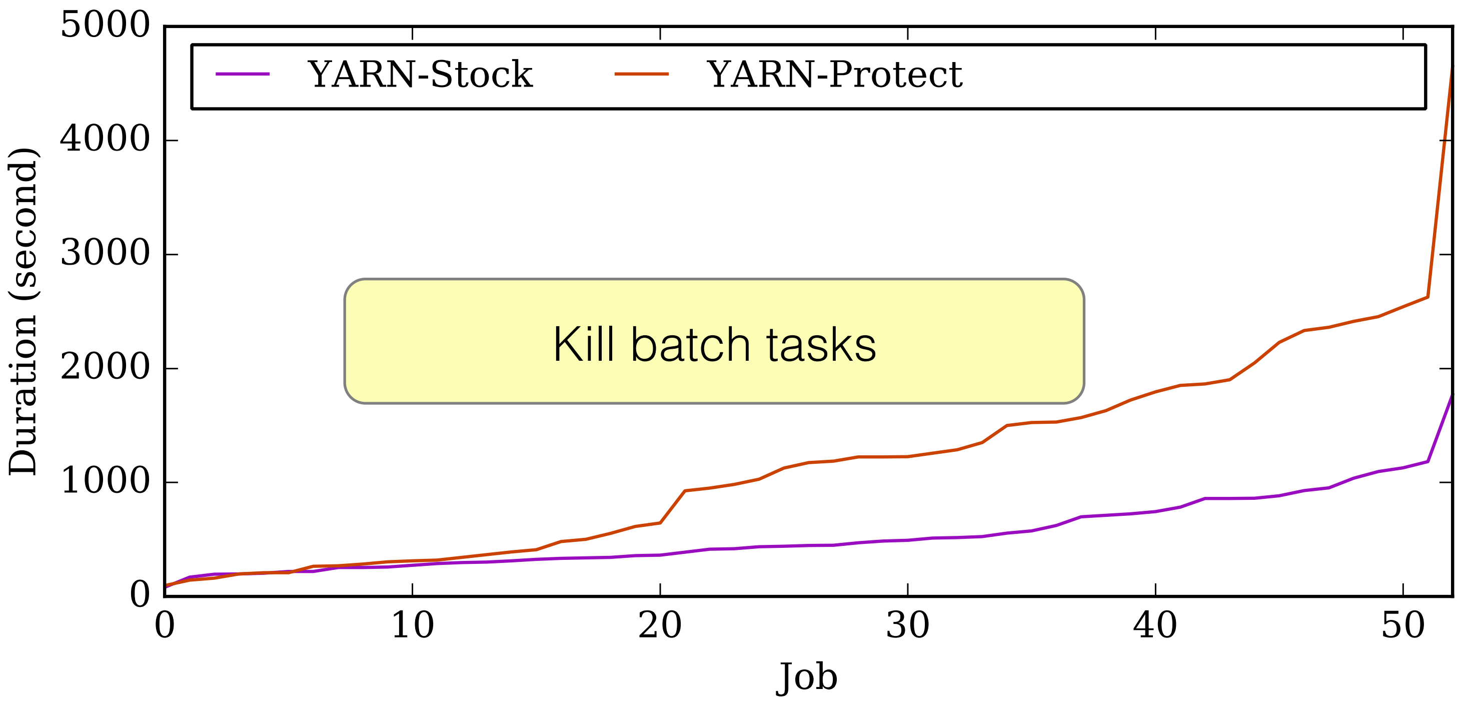
- Real-system deployment
  - 102-server cluster
  - Interactive service: Lucene with utilization trace
  - Batch task: TPC-DS queries on Hive
- Large-scale simulation
  - Trace from 10 production datacenters at Microsoft
  - Full datacenters for one month
- Production environment deployment
  - Data replica placement

Batch task scheduling -- real system



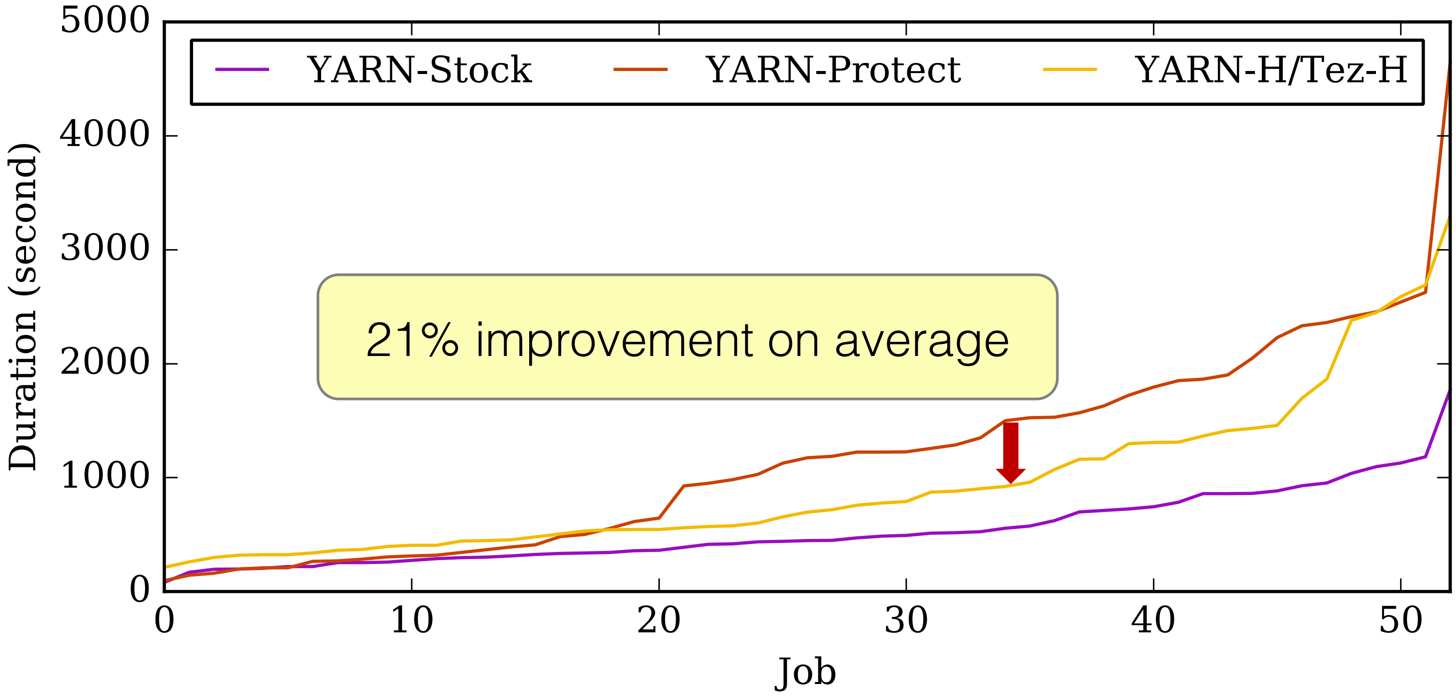
Degrading interactive service

Batch task scheduling -- real system



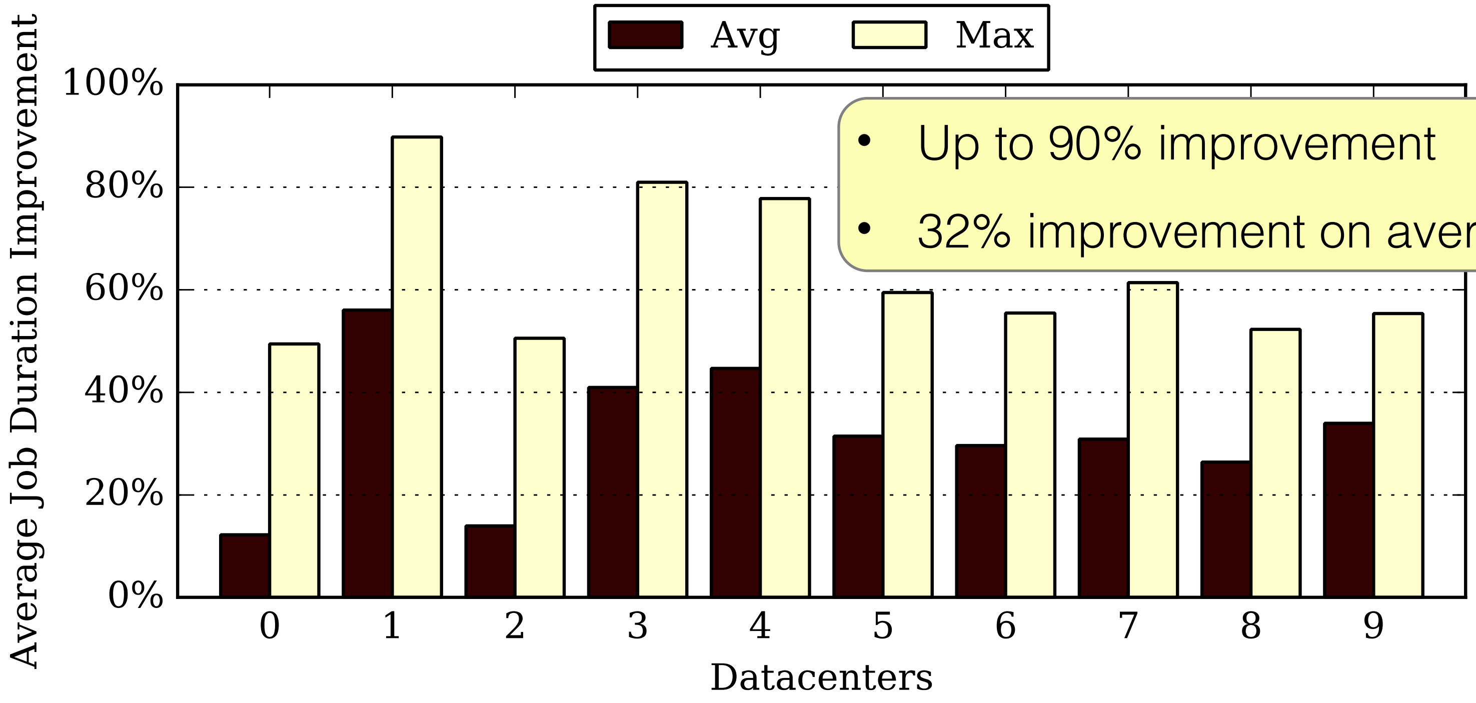
Kill batch tasks

Batch task scheduling -- real system



21% improvement on average

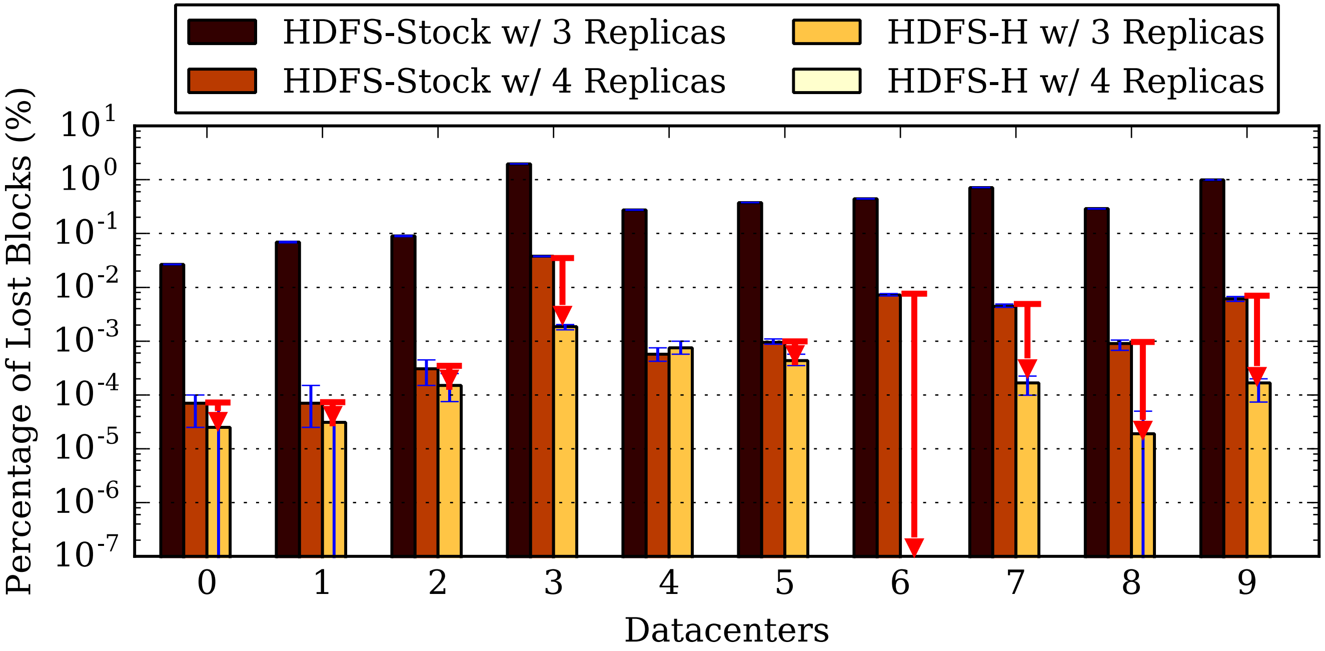
Batch task scheduling -- simulation



- Up to 90% improvement
- 32% improvement on average

Replica placement -- durability

- >2 orders of magnitude improvement
- Higher durability with fewer replicas



- Deployed to thousands of production servers for almost a year
- Eliminated data losses except minor bugs and not enough diversity

Lessons learned from deployment

- Placement diversity and disk space utilization
- Synchronous operations and unavailability
- Simplicity is critical in production systems
- More lessons in the paper

Conclusion

- History-based resource harvesting
    - Resource utilization dynamics
    - Data storage co-location
    - Complex data analytics distributed across servers
- Significantly improve datacenter efficiency
    - Deployed in production datacenters
    - Contributed to open-source community

History-Based Harvesting of Spare Cycles and Storage in Large-Scale Datacenters

Yunqi Zhang, George Prekas, Giovanni Matteo Fumarola,
Marcus Fontoura, Íñigo Goiri, Ricardo Bianchini

UNIVERSITY OF MICHIGAN

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Microsoft

Microsoft Research