Per-Application Server Specialization in Data Centers

Timothy Zhu, Anshul Gandhi, Mor Harchol-Balter, and Michael Kozuch (CMU)

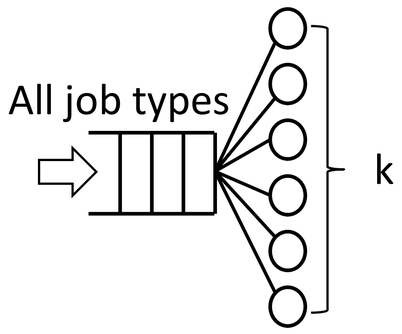
Motivation – Server Specialization

- Data centers often support a mix of applications
- Tailoring servers to specific applications may improve:
  - Individual application performance
  - Aggregate data center performance
  - Aggregate data center energy efficiency

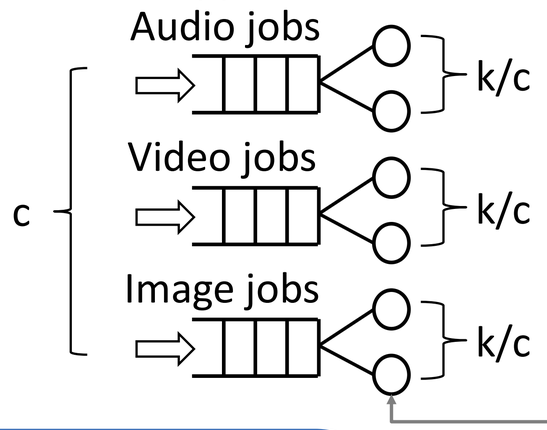
- Application specialization may occur through:
  - Server configuration (more CPU, disk, DRAM, etc)
  - Addition of application-specific accelerators
  - Other (proximity to resources, network connectivity, etc)

A Tale of Two Cluster Types

Homogeneous Cluster
(*k* general-purpose servers)



Heterogeneous Cluster
(Servers specialized to *c* classes of applications)



Specialized servers provide *speedup* for appropriate tasks, but cannot run inappropriate tasks

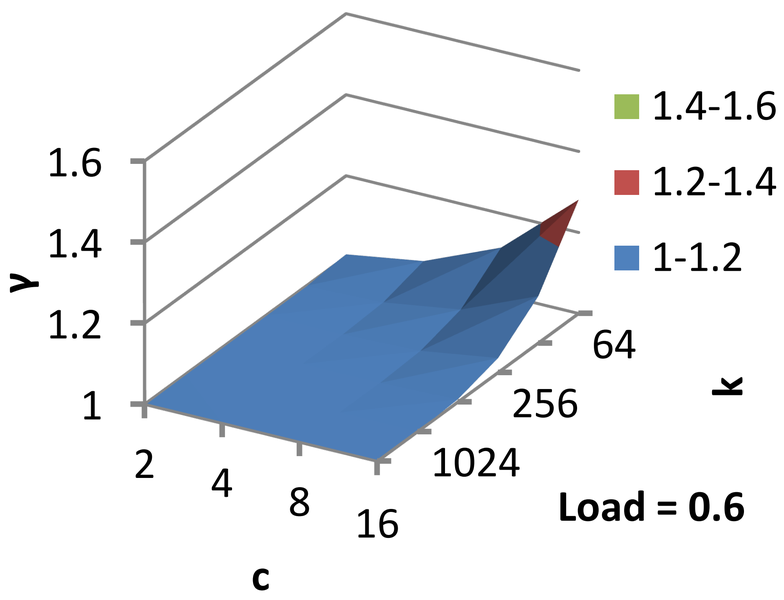
Key Question

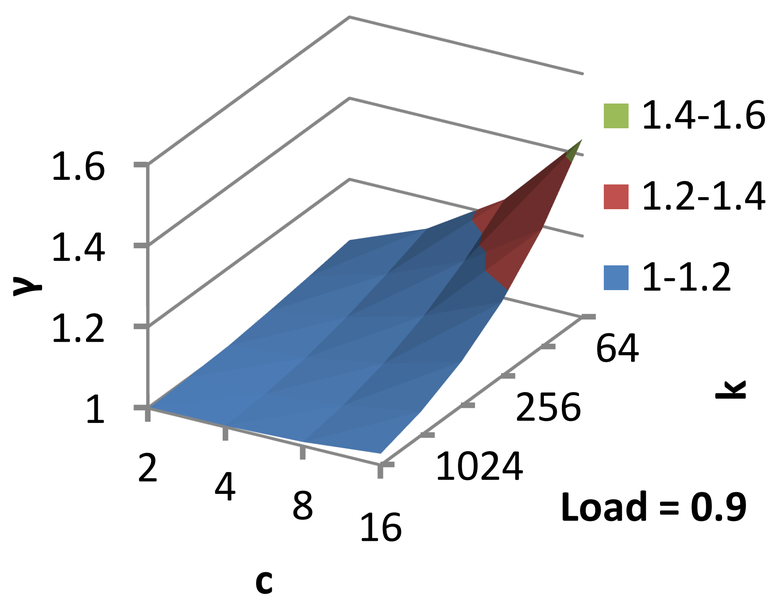
How much faster must specialized servers be to justify a lack of generality?

Approach

Evaluate *Breakeven Speedup*, γ, the speedup needed such that the mean response time for arriving jobs is the same for both the specialized and homogeneous data centers, when presented with the same load.

Analysis of Idealized Workload – Effect of Load





Observation 1
A little speedup goes a long way

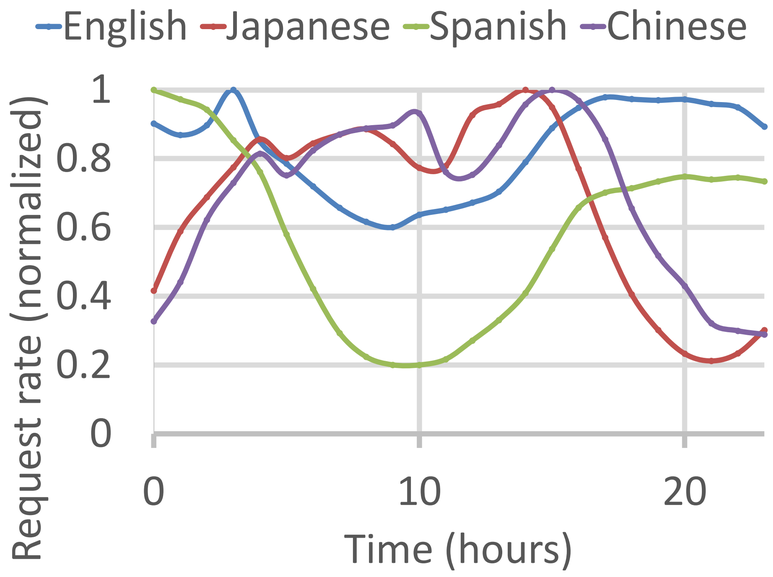
Observation 2
As c increases, γ increases
As k increases, γ decreases

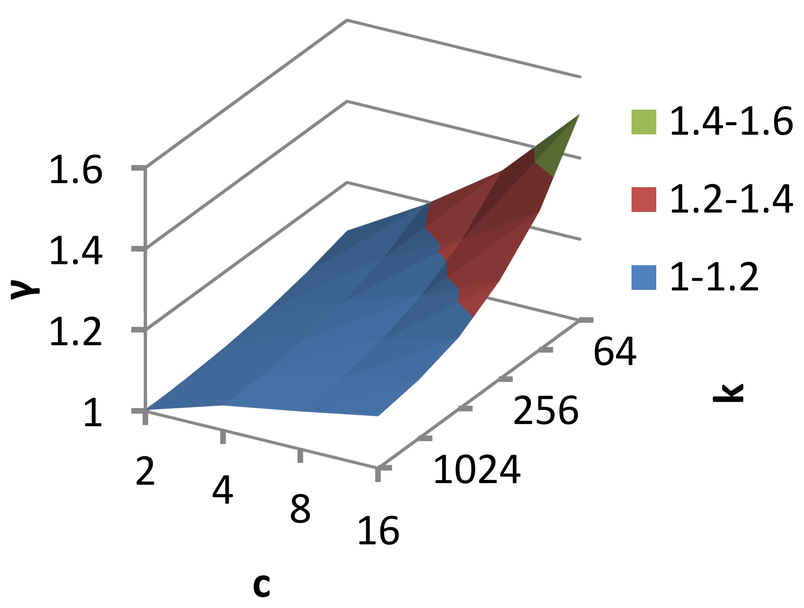
Observation 3
As load increases, γ increases, but not by much

Observation 4
Different, time-varying arrival processes permit the homogeneous data center to "time-shift" resources; γ is greater than for idealized patterns.

Other Observations (not shown)
When *speedup* > γ, the excess computational capacity enables a dramatic reduction in the number of servers required.

Analysis using Real-World Workload: Wikipedia





Intel Science & Technology Center for Cloud Computing
Carnegie Mellon University
Georgia Tech
intel
PRINCETON UNIVERSITY
UC Berkeley
UNIVERSITY of WASHINGTON