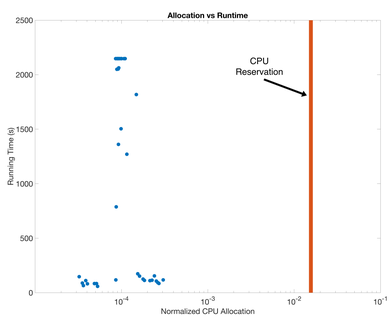
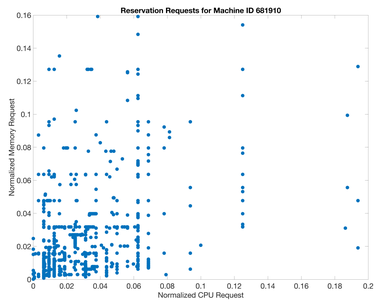
Heterogeneous Resource Allocation in Warehouse Scale Computers

Zach Ankenman, Ben Berg

Motivation

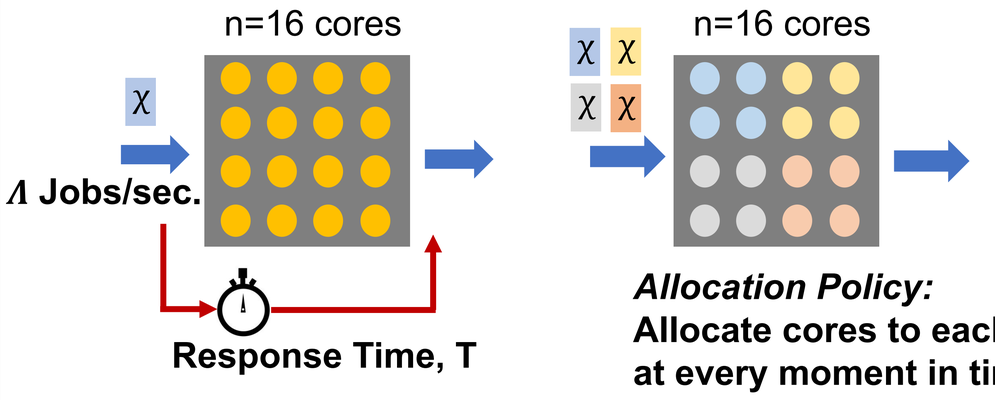
- Warehouse scale computers (WSC's) used reservation-based management
- WSC workload is diverse
- Utilization is low, so *latency* is more important than *throughput*
- Current trend is towards *performance-based* management and *colocation*



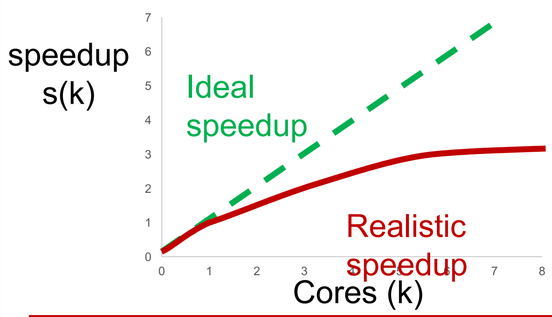


<u>Question:</u> How should we allocate resources to a diverse, latency-critical workload?

Model

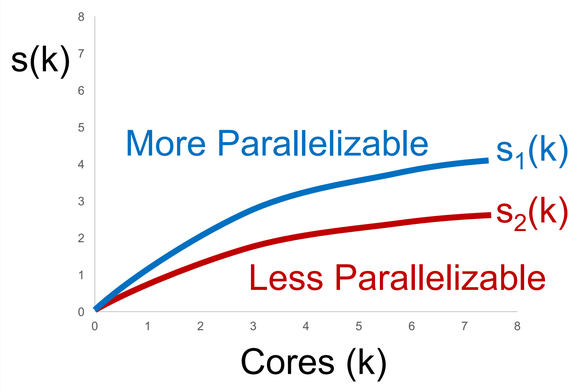


Allocation Policy:
Allocate cores to each job at every moment in time



Jobs receive a sublinear speedup from additional cores

Problem



Consider *heterogeneous* stream of jobs from *two classes:* Class 1 and Class 2

Assume Amdahl's law:

$$s_i(k) = \frac{1}{\frac{p_i}{k} + (1 - p_i)}$$

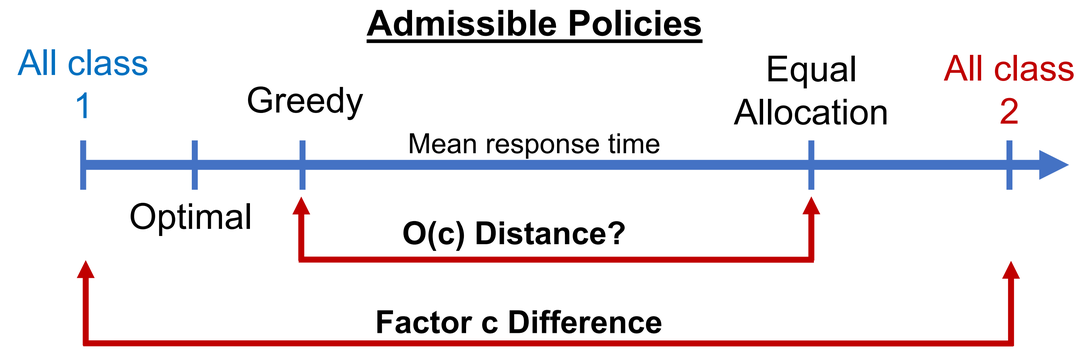
<u>Goal:</u> Describe allocation policy to minimize mean response time

Carnegie Mellon University

Theoretical Results

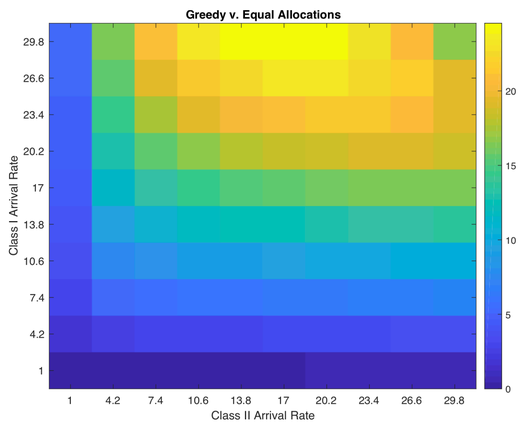
- **In practice, cluster managers use *Greedy policy:* maximize throughput at all times**
  - **Provably suboptimal, but how bad is it?**
- **An *Admissible Policy* is:**
  - **Better than if *all* jobs were class 2**
  - **Worse than if *all* jobs were class 1**
  - **Greedy is admissible**
- **<u>Theorem:</u> If $\frac{s_1(k)}{s_2(k)} \leq c$, any admissible policy becomes a c-approximation of Optimal as n becomes large**

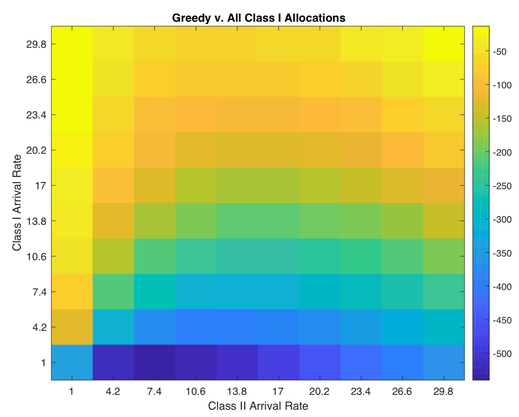
<u>Admissible Policies</u>

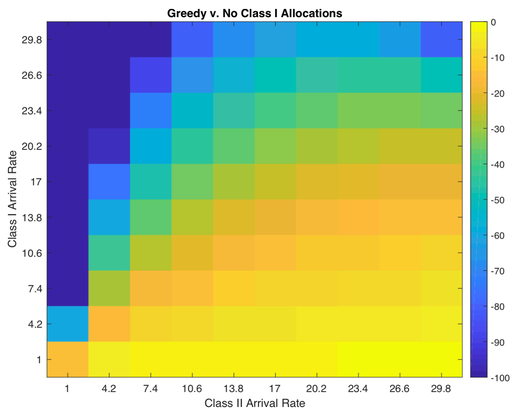


- **Is Greedy much better than the Equal Allocation policy?**

Simulation Results







- **Simulations for n=64 cores, p1=0.9, p2=0.1**
- **Plots show percent difference in latency between policies**
- **Evaluated at various job arrival rates**

Future Work

- **Based on above comparisons, try and and prove bounds**
- **Adapt to handle additional job classes**
- **Use this information to make colocation decisions**
  - **What workloads can be colocated efficiently?**