Sharp Zero-Queueing Bounds for Multi-Server Jobs

Yige Hong†

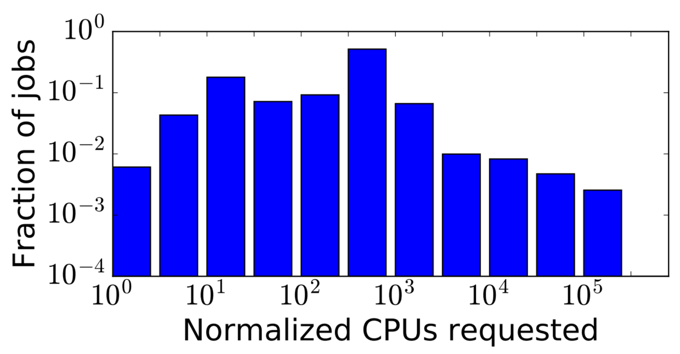
† Carnegie Mellon University

Motivation

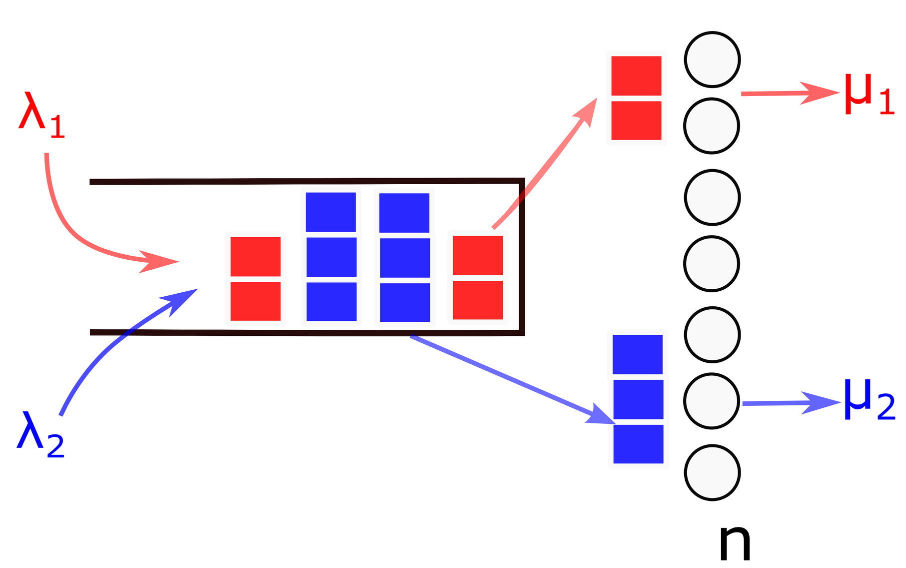
Multi-server job: occupy multiple servers

server need varies across orders of magnitudes

Google Borg:



Model



arrival rate λ_i, service rate μ_i, server need ℓ_i.

Objective

Understand the minimal achievable mean waiting time in steady-state

$$\min \mathbb{E}\left[W(\infty)\right]$$

Proportional to total queue length

Related Work

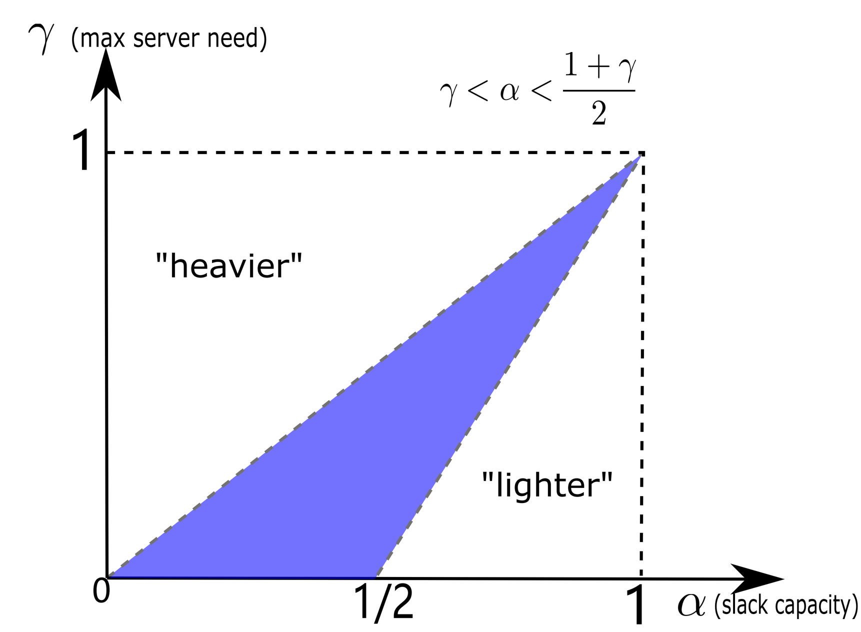
- Exact solutions are only known for two-server systems [Brill, Green 1984], [Filippopoulos, Karatza 2007]; even stability is hard [Rumyantsev, Morozov 2017], [Afanaseva, Bashtova, Grishunina 2019], [Grosof, Harchol-Balter, Scheller-Wolf 2020]
- Under scaling regimes, queueing probability is characterized [Wang, Xie, Harchol-Balter, 2021]

Scaling Regime

A sequence of systems with $n, \lambda_i, \ell_i \to \infty$.

- Slack capacity: $\delta = n - \sum_{i=1}^{I} \frac{\lambda_i \ell_i}{\mu_i}$
- Maximal server need: $\ell_{\max} = \max_{i\in[I]} \ell_i$

Special case: $\delta = n^{\alpha}$, $\ell_{\max} = n^{\gamma}$.



Results

Q1: What's waiting time under First Come First Serve (FCFS)?

First Come First Serve (FCFS): Fit the jobs one by one from the head of line until no enough servers available

Theorem 1: FCFS (Informal)

Under FCFS, the mean waiting time is
$$\mathbb{E}\left[W(\infty)\right] = \Theta\left(n^{\gamma-\alpha}\right).$$

Q2: Is FCFS a good idea?

Theorem 2: Lower Bound (Informal)

Under any policy, the mean waiting time is
$$\mathbb{E}\left[W(\infty)\right] = \Omega\left(n^{-\alpha}\right).$$

Q3: Is the low bound achievable?

Theorem 3: Achievability (Informal)

There is a policy with mean waiting time
$$\mathbb{E}\left[W(\infty)\right] = O\left(n^{-\alpha}\right).$$

Preliminary

Approach:

Analyze steady-state expectation $\mathbb{E}\left[\sum_{i=1}^{I} X_i\right]$

X_i: number of type-i jobs in the system

Work-conserving policies:

- either fit all jobs into servers,
- or keep idle servers $\leqslant \ell_{\max}$

Examples: FCFS, priority policy

Workload as an easier surrogate

workload $\sum_{i=1}^{I} \frac{\ell_i}{\mu_i} X_i$.

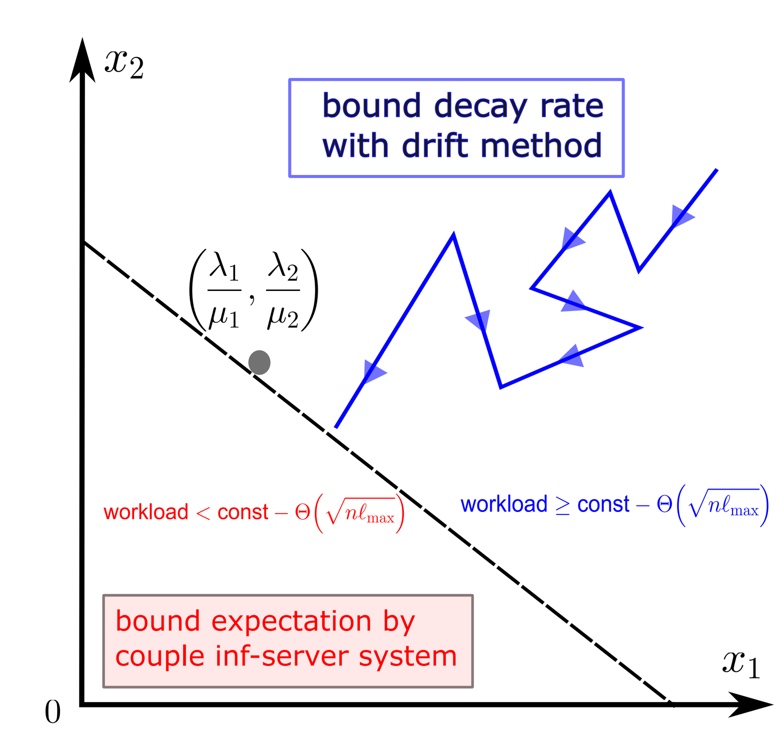
Under any work-conserving policy, the decrease rate of workload is

$$\sum_{i=1}^{I} \frac{\ell_i}{\mu_i}(\lambda_i - \mu_i Z_i) = \sum_{i=1}^{I} \frac{\ell_i}{\mu_i}\lambda_i - \sum_{i=1}^{I} \ell_i Z_i \approx \sum_{i=1}^{I} \frac{\ell_i}{\mu_i}\lambda_i - n = -\delta$$

where Z_i is number of type-i jobs in service

Lemma: Workload (Informal)

Under any work-conserving policy, the workload is
$$\mathbb{E}\left[\sum_{i=1}^{I} \frac{\ell_i}{\mu_i}\left(X_i - \frac{\lambda_i}{\mu_i}\right)\right] = \Theta\left(n^{1+\gamma-\alpha}\right),$$
(Note that $\mathbb{E}\left[X_i - \frac{\lambda_i}{\mu_i}\right] = \mathbb{E}[Q_i]$)



Assumptions on traffic:

$\lambda_I \ell_I = \Theta(n)$, $\lambda = \Theta(n)$

Proof

Queue Length Lower Bound

Consider the following LP-relaxation

$$\min_{\{q_i:\, i\in[I]\}} \quad \mathbb{E}[\text{ total queue length }] \qquad (1)$$
subject to $\quad$ workload constraint

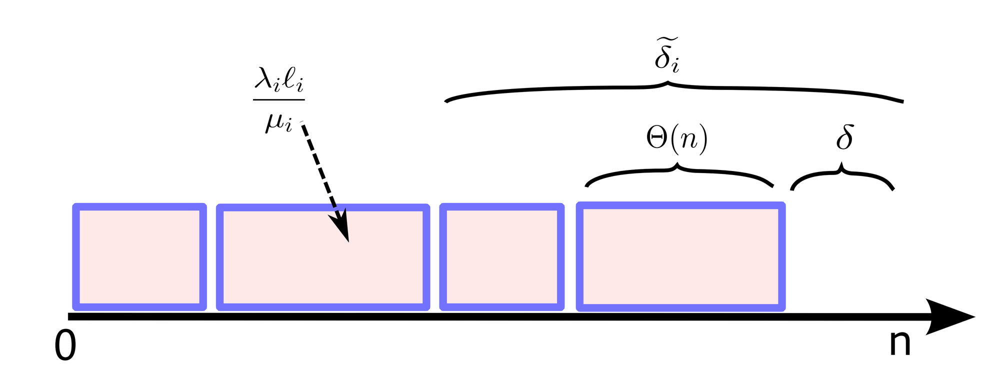
FCFS

- Modified-FCFS: serve a new customer only when at least $\ell_{\max}$ servers available
- Under Modified-FCFS, $\mathbb{E}[Q_i^{(U)}] \geqslant \mathbb{E}[Q_i]$
- Independence $\mathbb{E}[Q_i^{(U)}] = \frac{\lambda_i}{\lambda}\mathbb{E}[Q^{(U)}]$.
- Derive total queue length from workload in Modified FCFS.
- $(n+\ell_{\max})$—server under Modified-FCFS can serve as lower bounding system

Priority

Suppose $\ell_1 \leqslant \ell_2 \leqslant \dots \ell_I$, prioritize lower indices

1. For $i \in [I]$, the first i types of jobs form a system with slack capacity $\widetilde{\delta}_i$ and maximal server need $\widetilde{\ell_{\max}}_i = \ell_i$
2. $\mathbb{E}[Q_i] \leqslant \frac{\mu_{\max}}{\ell_i}\mathbb{E}$ [workload for first i types]
3. Upper bound queue length using upper bound on workload
4. Queue length $\mathbb{E}[Q_i]$ small for any $i < I$, last job type $\mathbb{E}[Q_I] \leqslant O\left(\frac{n}{\delta}\right)$



Future work

Find a practical non-preemptive policy that achieves the lower bound