Zero Queueing for Multi-Server Jobs

Weina Wang

Carnegie Mellon University

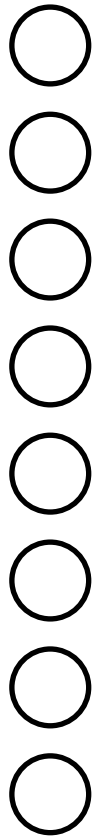
YEQT 2021

Collaborators

Qiaomin Xie
Cornell University

Mor Harchol-Balter
Carnegie Mellon University

What is a multi-server job?



servers

Server needs grow as datacenters grow

Google Borg trace 2019

[Tirmazi et al. 2020], [Wilkes 2019]

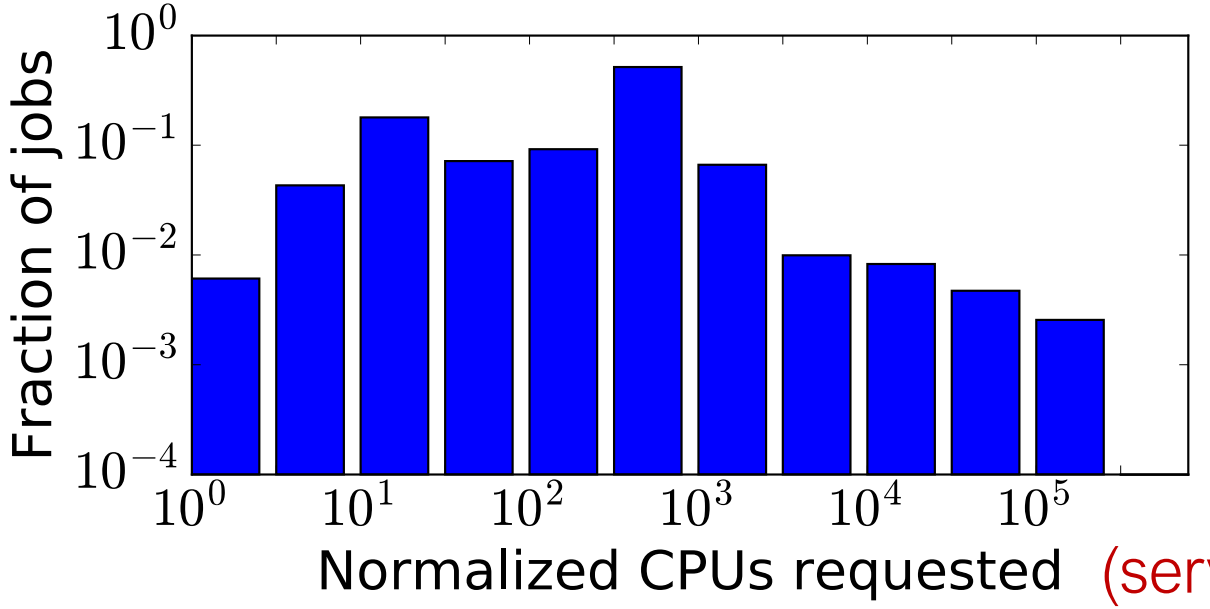


Figure taken from [Grosof, Harchol-Balter, Scheller-Wolf 2020]

Model

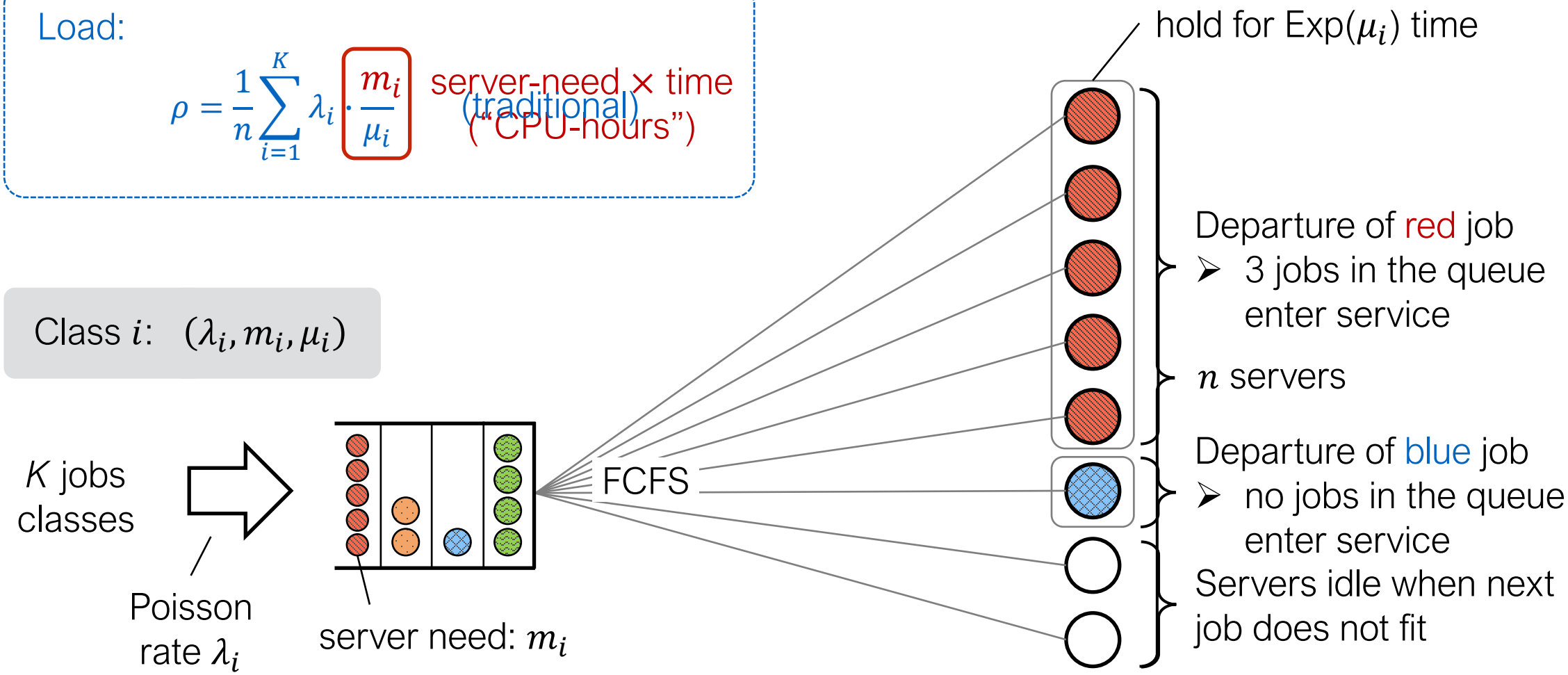
Load:

$$\rho = \frac{1}{n}\sum_{i=1}^{K} \lambda_i \cdot \frac{m_i}{\mu_i}$$

server-need × time ("CPU-hours")

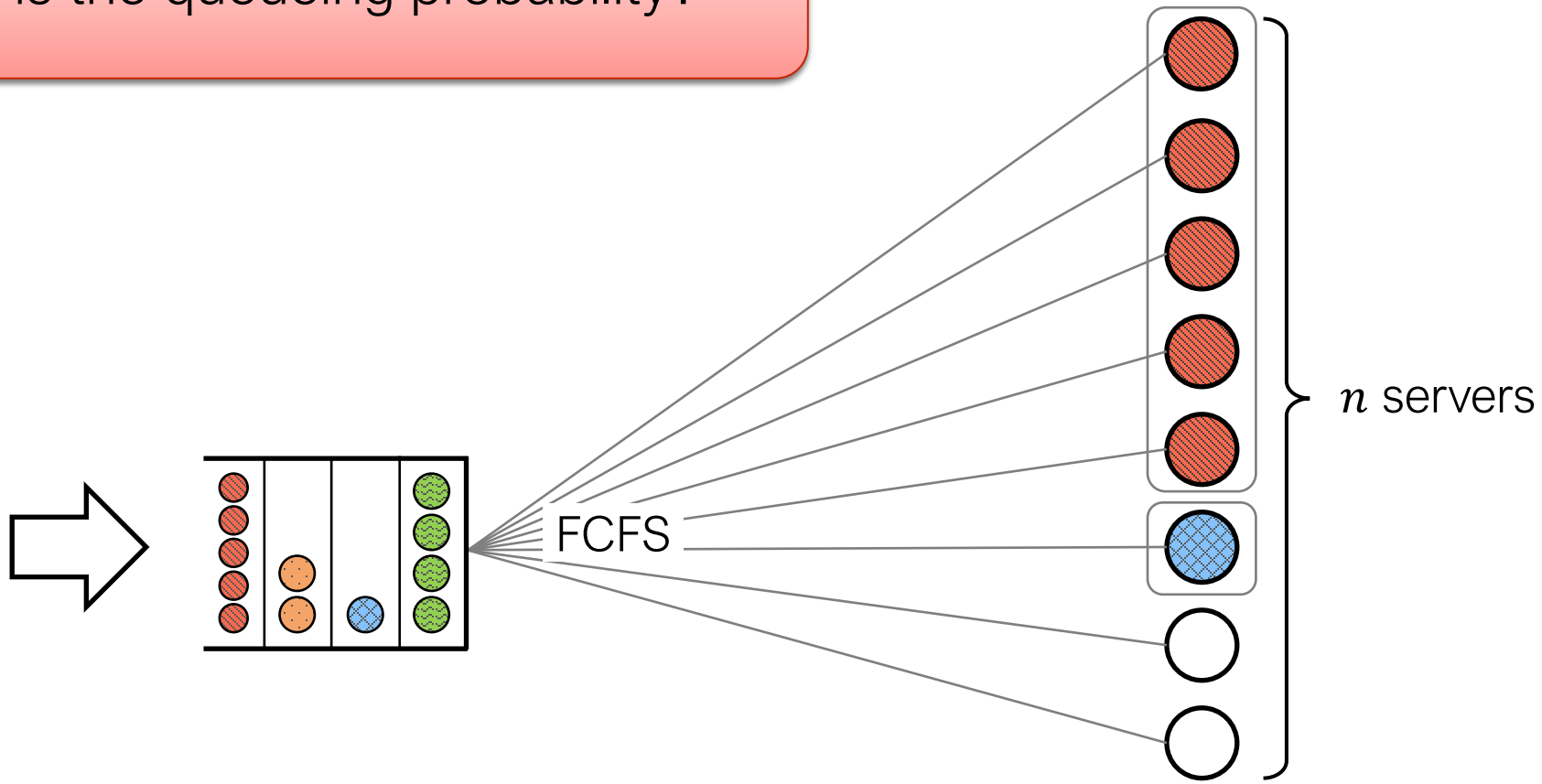
(traditional)

Class i: (λ_i, m_i, μ_i)



Model

What is the queueing probability?



Related work

- Exact solutions are unknown
  - Only known for two-server systems [Brill, Green 1984], [Filippopoulos, Karatza 2007]
- Even stability is hard
  - All classes have the same service rate [Rumyantsev, Morozov 2017], [Afanaseva, Bashtova, Grishunina 2019]
  - Two-class system [Grosof, Harchol-Balter, Scheller-Wolf 2020]

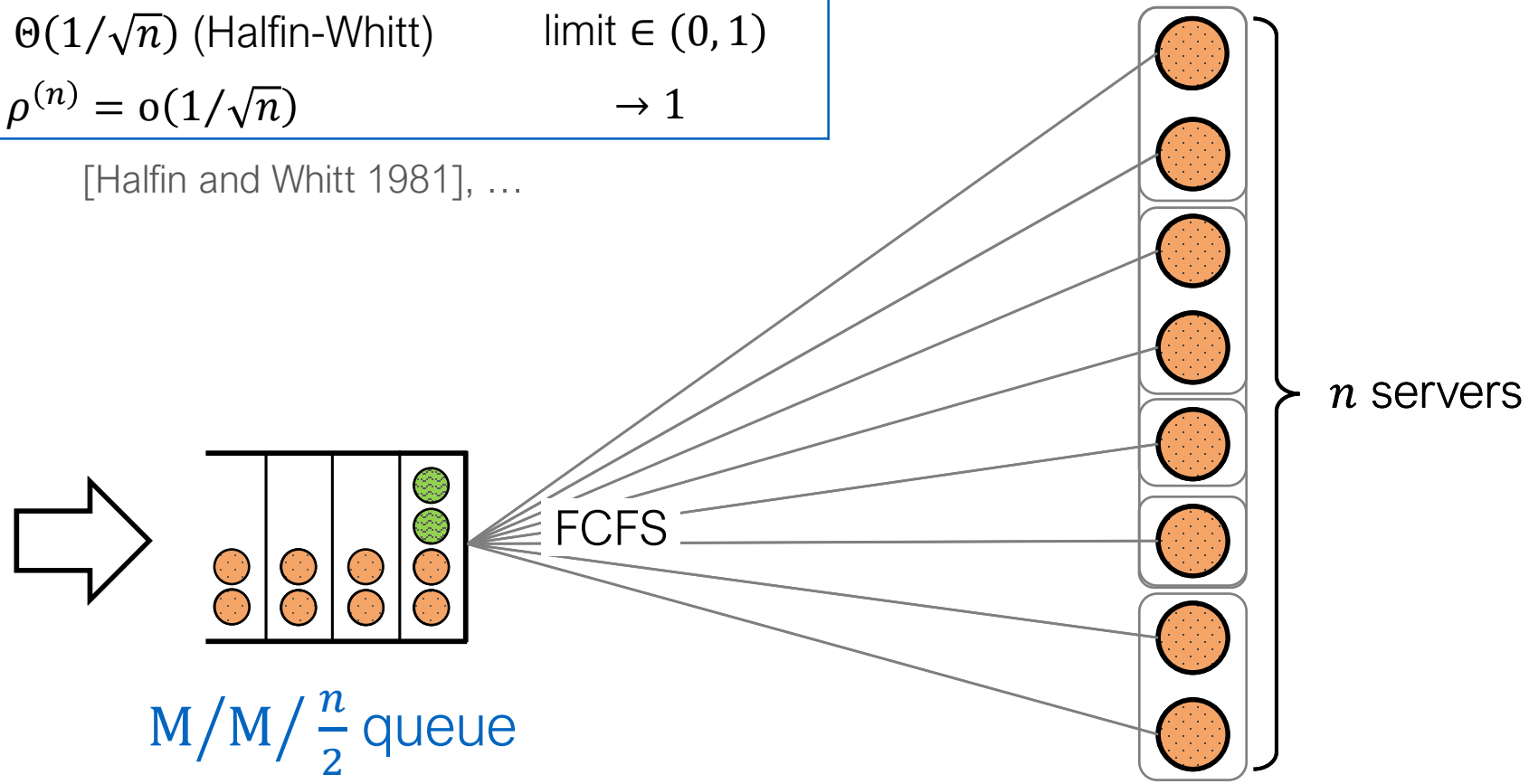
- If we remove the queue …
  - [Arthurs, Kaufman 1979], [Hunt, Kurtz 1993], [Bean, Gibbens, Zachary 1995], [Dasylva, Srikant 1999]…
- Implications in VM (Virtual Machine) scheduling
  - [Maguluri, Srikant, Ying 2012, 2014], [Psychas, Ghaderi 2017, 2019], …

n is large!

scaling regime	queueing prob.
$1-\rho^{(n)}=\omega(1/\sqrt{n})$	$\to 0$
$1-\rho^{(n)}=\Theta(1/\sqrt{n})$ (Halfin-Whitt)	limit $\in(0,1)$
$1-\rho^{(n)}=\mathrm{o}(1/\sqrt{n})$	$\to 1$

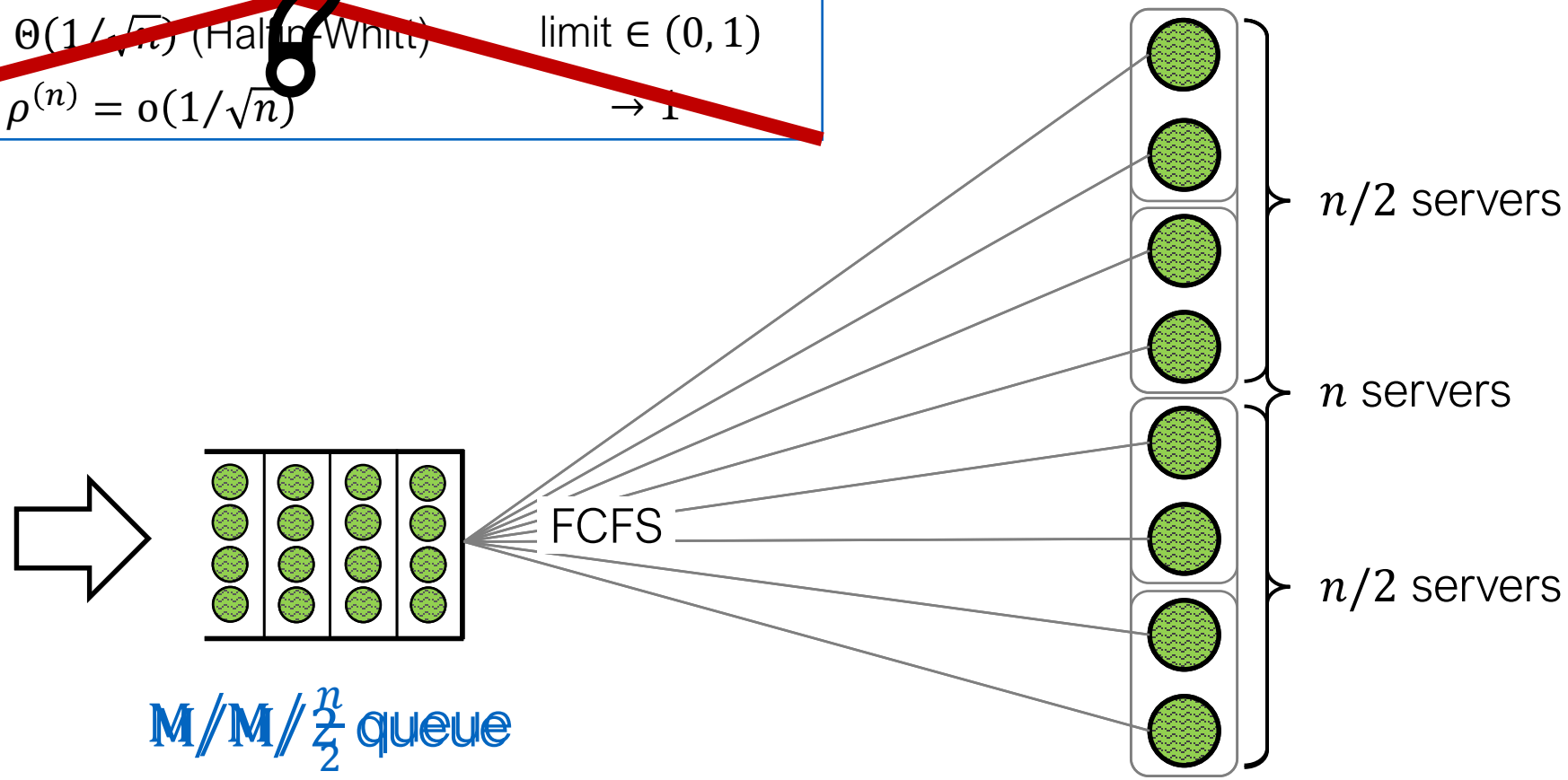
ZERO queueing

[Halfin and Whitt 1981], …



n is large! Server-needs are also large!

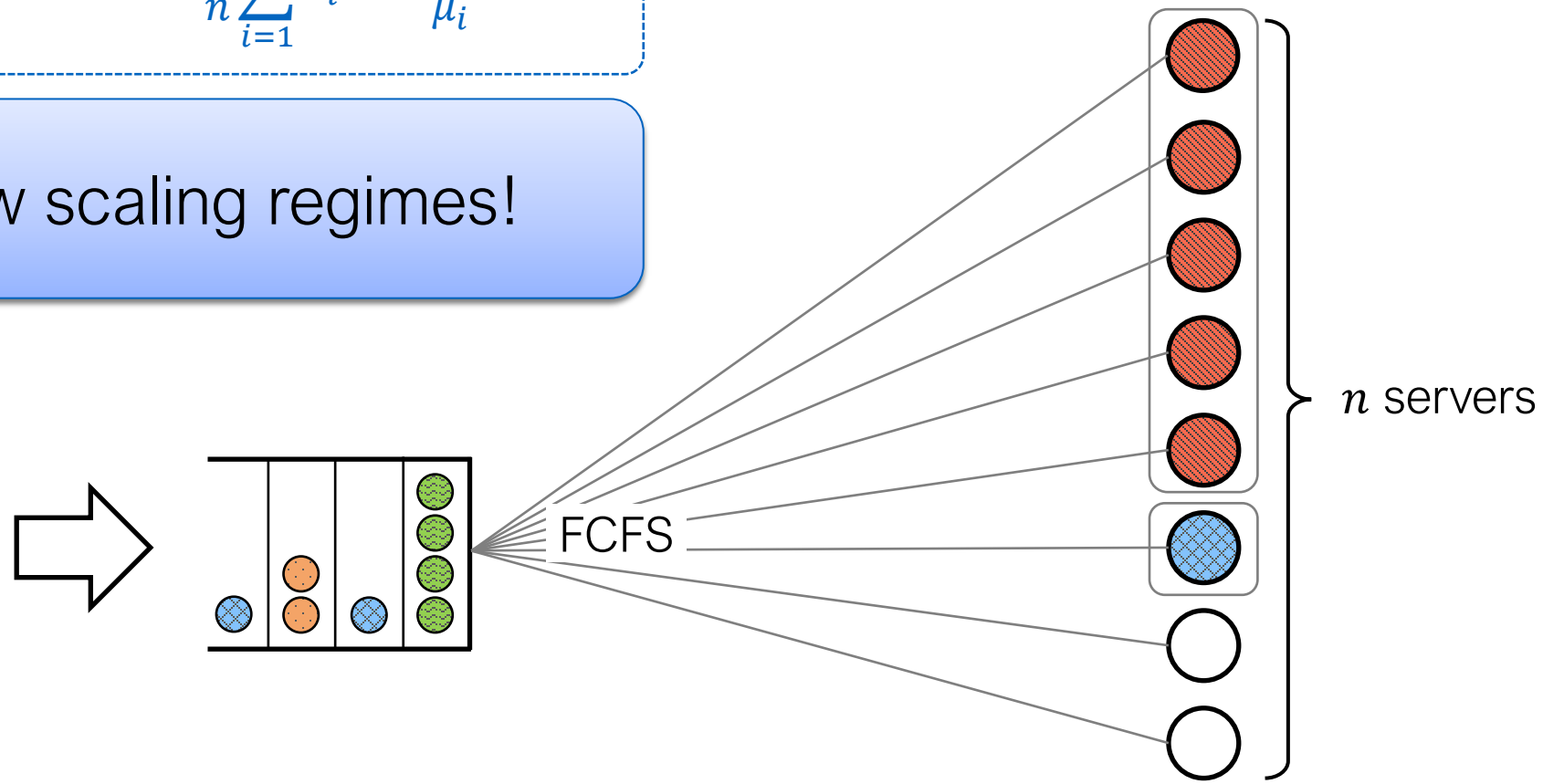
scaling regime	queueing prob.
$1-\rho^{(n)}=\omega(1/\sqrt{n})$	$\to 0$
$1-\rho^{(n)}=\Theta(1/\sqrt{n})$ (Halfin-Whitt)	limit $\in(0,1)$
$1-\rho^{(n)}=o(1/\sqrt{n})$	$\to 1$



n is large! Server-needs are also large!

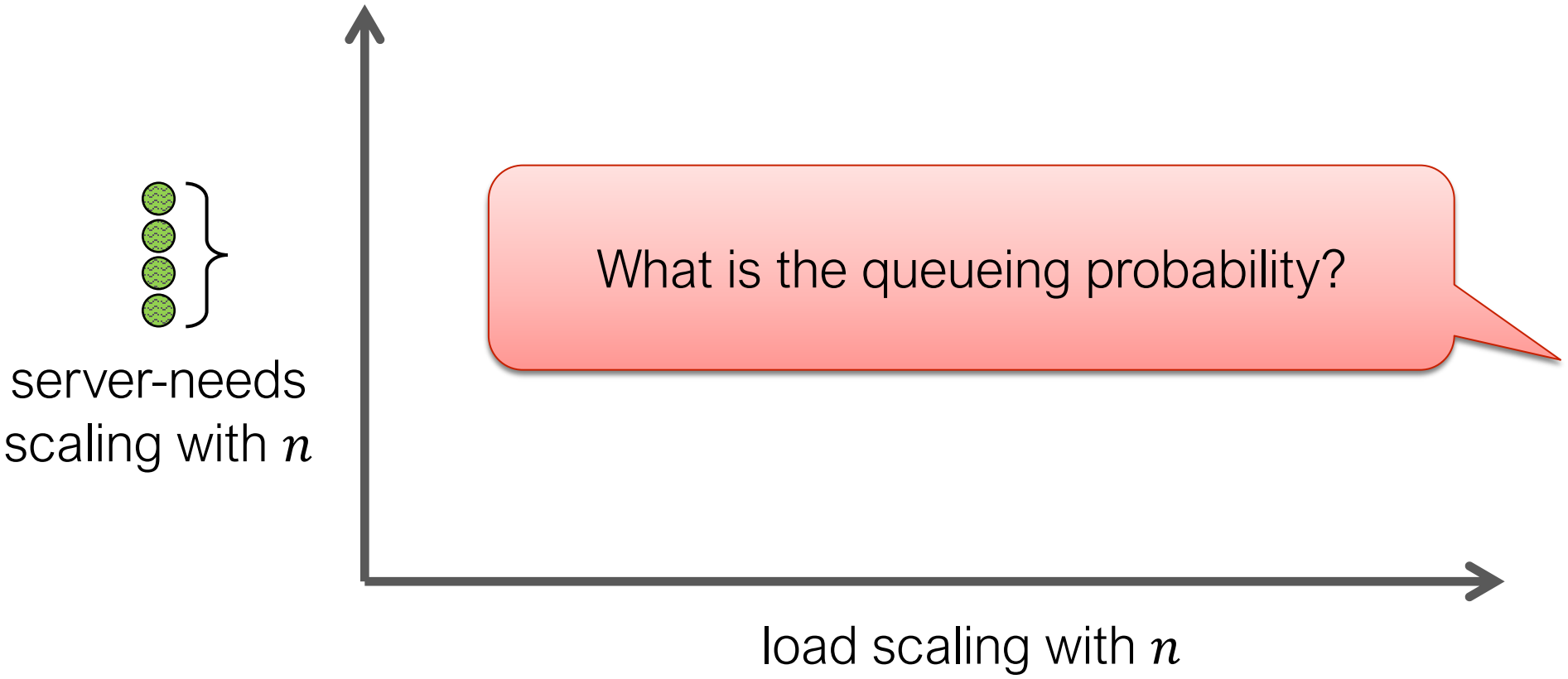
Load: $\rho^{(n)} = \frac{1}{n}\sum_{i=1}^{K} \lambda_i^{(n)} \cdot \frac{m_i^{(n)}}{\mu_i}$

New scaling regimes!

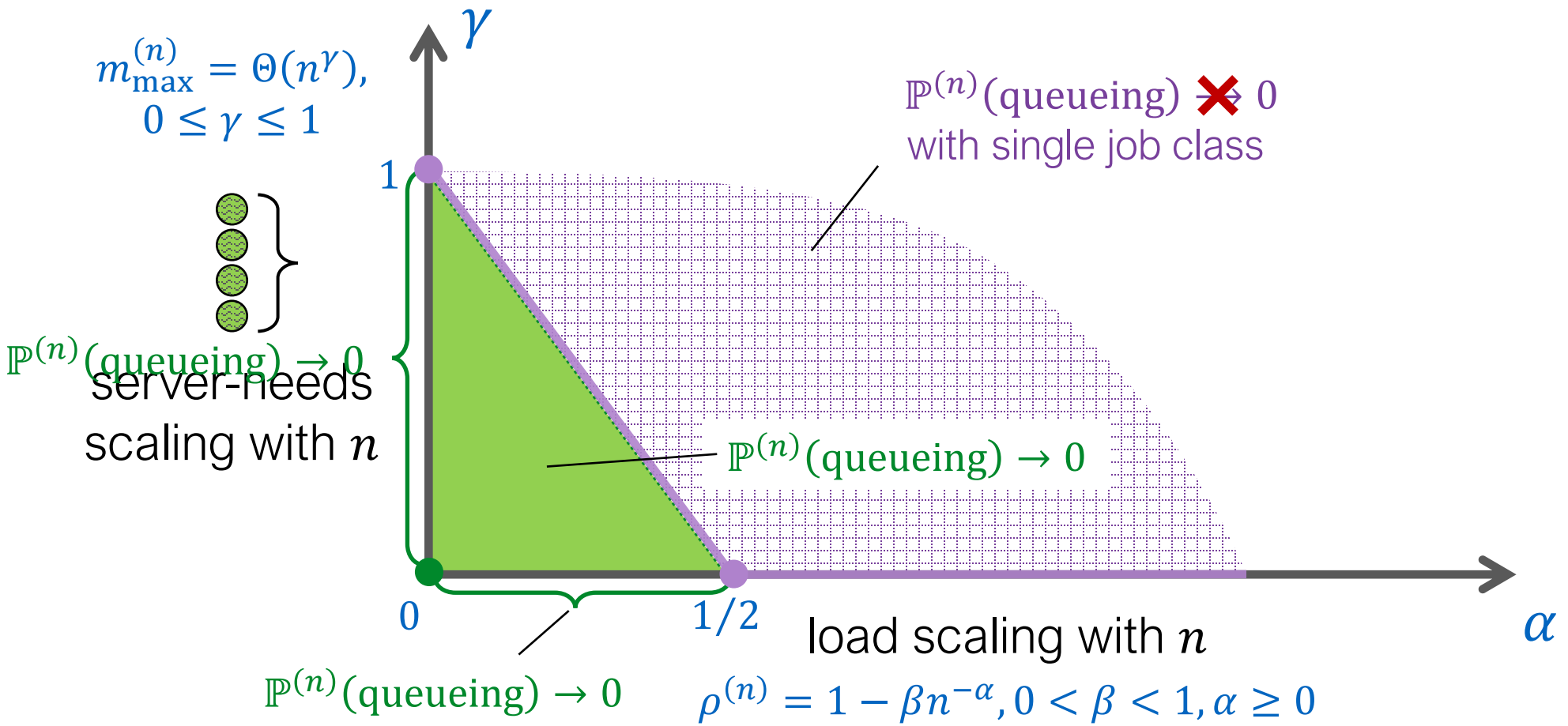


Can we analyze the queueing probability under NEW scaling regimes?

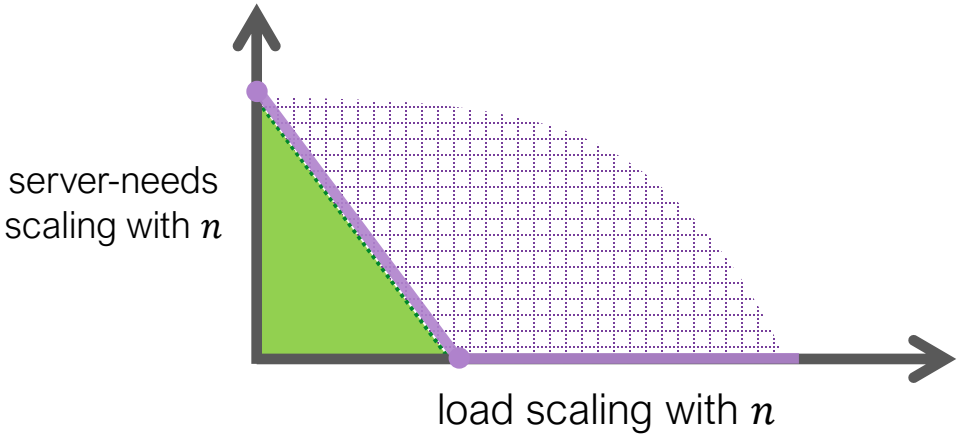
NEW joint scaling



Our results


$m_{\max}^{(n)} = \Theta(n^{\gamma})$,
$0 \le \gamma \le 1$
γ
1
$\mathbb{P}^{(n)}(\text{queueing}) \not\to 0$
with single job class
$\mathbb{P}^{(n)}(\text{queueing}) \to 0$
server-needs
scaling with n
$\mathbb{P}^{(n)}(\text{queueing}) \to 0$
0
1/2
α
load scaling with n
$\mathbb{P}^{(n)}(\text{queueing}) \to 0$
$\rho^{(n)} = 1 - \beta n^{-\alpha}, 0 < \beta < 1, \alpha \ge 0$

Our results



THEOREM

$$\mathbb{P}^{(n)}(\text{queueing}) \leq \frac{1}{1-\rho^{(n)}}\left(3K\sqrt{\frac{m_{\max}^{(n)}}{n}}+\frac{m_{\max}^{(n)}}{n}\right)$$

Assumptions: $m_{\max}^{(n)} = o(n)$, and $\rho^{(n)} < 1 - \frac{m_{\max}^{(n)}}{n}$ (stability)

Our results

THEOREM

dominant term

$$\mathbb{P}^{(n)}(\text{queueing}) \le \frac{1}{1-\rho^{(n)}}\left(3K\sqrt{\frac{m_{\max}^{(n)}}{n}} + \frac{m_{\max}^{(n)}}{n}\right)$$

Assumptions:

$$m_{\max}^{(n)} = o(n),$$

$$\rho^{(n)} < 1 - \frac{m_{\max}^{(n)}}{n}$$

$$\mathbb{P}^{(n)}(\text{queueing}) \to 0 \quad \Leftarrow \quad 1-\rho^{(n)} \gg \sqrt{\frac{m_{\max}^{(n)}}{n}}$$

$$\Leftrightarrow \quad n\left(1-\rho^{(n)}\right) \gg \sqrt{n m_{\max}^{(n)}}$$

$n\left(1-\rho^{(n)}\right)$: # spare servers

$\sqrt{n m_{\max}^{(n)}}$: ~std(# busy servers) in infinite-server system

Existing drift-based frameworks

- Tail bounds implied by drift

  [Hajek 1982], [Bertsimas, Gamarnik, and Tsitsiklis 2001], [Maguluri and Srikant 2016], …

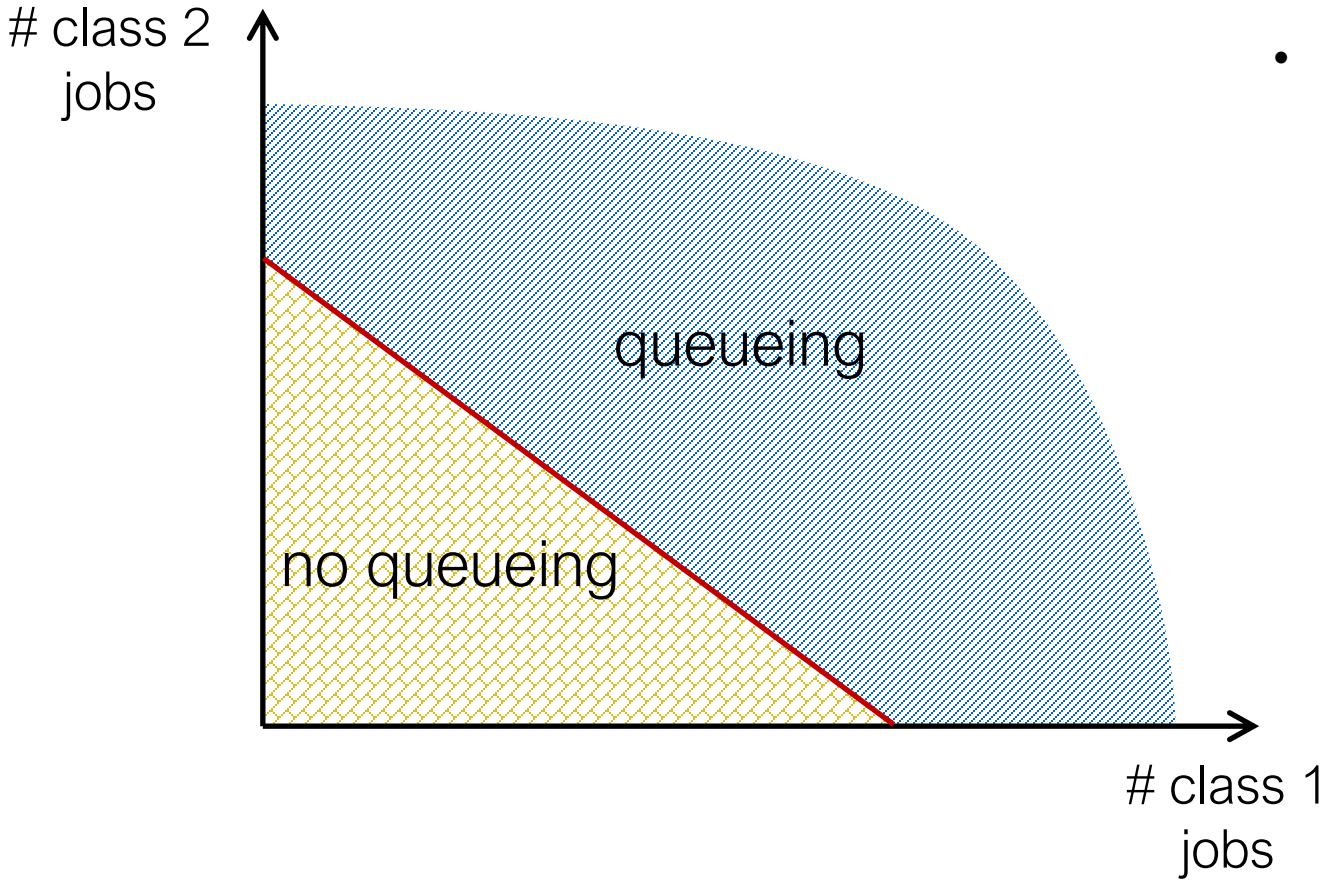
  - Challenge: hard to get proper drift bounds

- Expectation bounds via Stein's equation

  [Braverman, Dai, and Feng 2017], [Liu and Ying 2019], [Liu and Ying 2020], …

  - Challenge: hard to find the solution function

Our drift-based approach

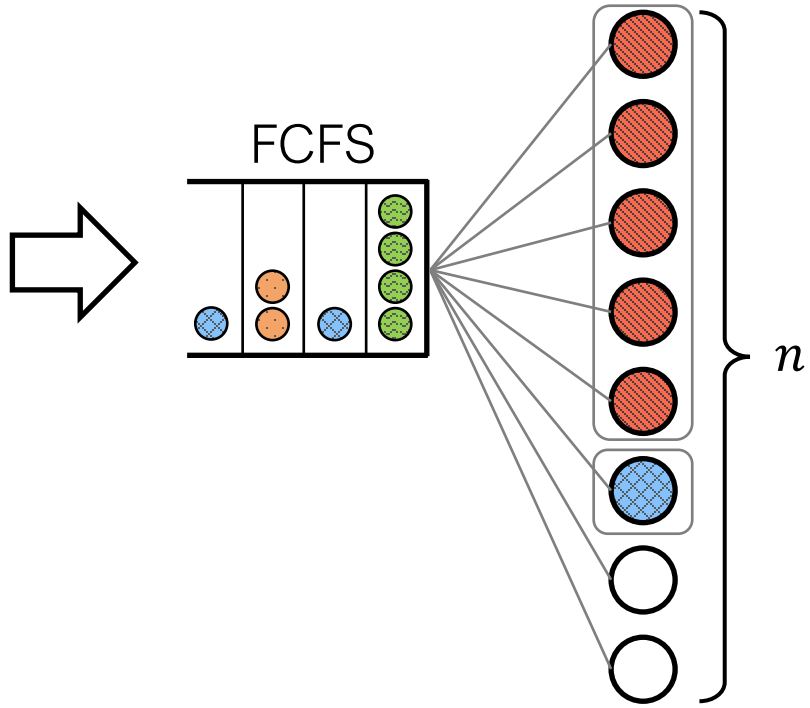


- Use a Lyapunov function
  - Drift in "queueing" region is negative
  - $\mathbb{E}$ (drift in "no queueing" region) can be bounded using existing drift methods
  - $\mathbb{E}(\text{total drift}) = 0$

    $\Rightarrow \mathbb{P}^{(n)}(\text{queueing}) \leq$ upper bound

Conclusions and future directions

New scaling regimes:
joint scaling of server-needs & load



THEOREM

$$\mathbb{P}^{(n)}(\text{queueing}) \leq \frac{1}{1-\rho^{(n)}}\left(3K\sqrt{\frac{m_{\max}^{(n)}}{n}}+\frac{m_{\max}^{(n)}}{n}\right)$$

What is the expected queueing time?

Which scheduling algorithm minimizes queueing asymptotically?

How to deal with Lyapunov functions with large and small jumps?

THANK YOU!

weinaw@cs.cmu.edu