Exponential Tail Bounds on Queues

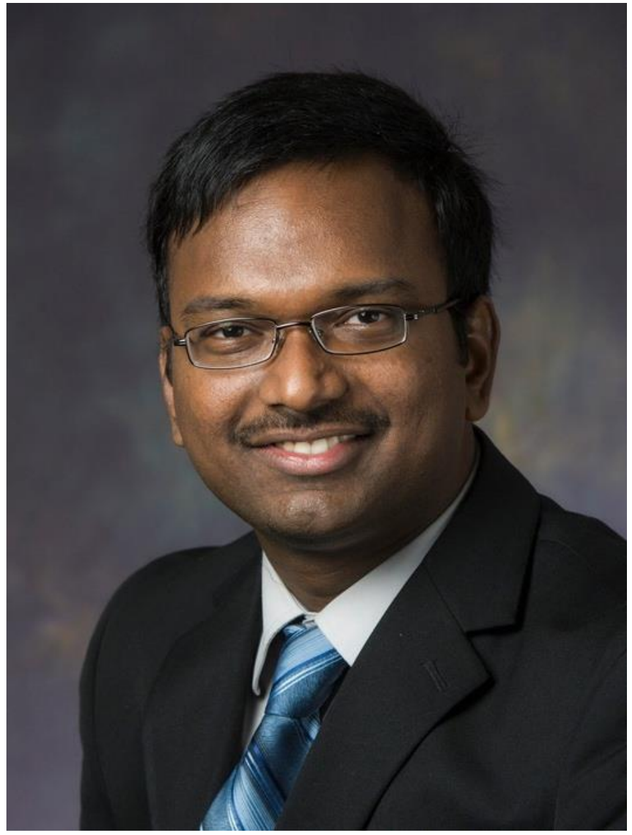
A confluence of non-asymptotic heavy traffic and large deviations

Daniela Hurtado-Lange
Northwestern University

Joint work with

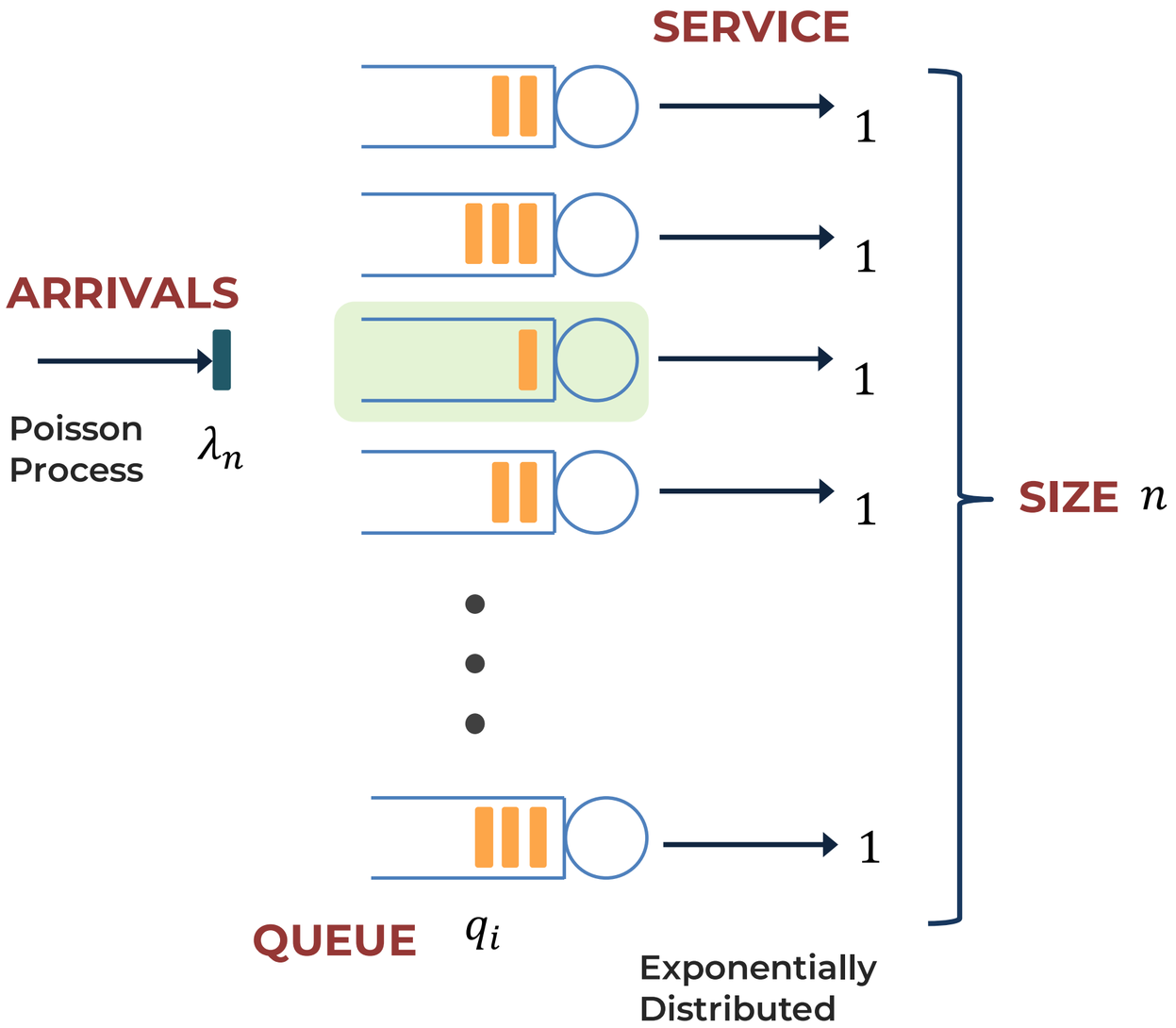
Prakirt Jhunjhunwala

On the academic job market!



Siva Theja Maguluri

Join-the-Shortest Queue (JSQ)



Throughput Optimality:

Stable: $\lambda_n < n$

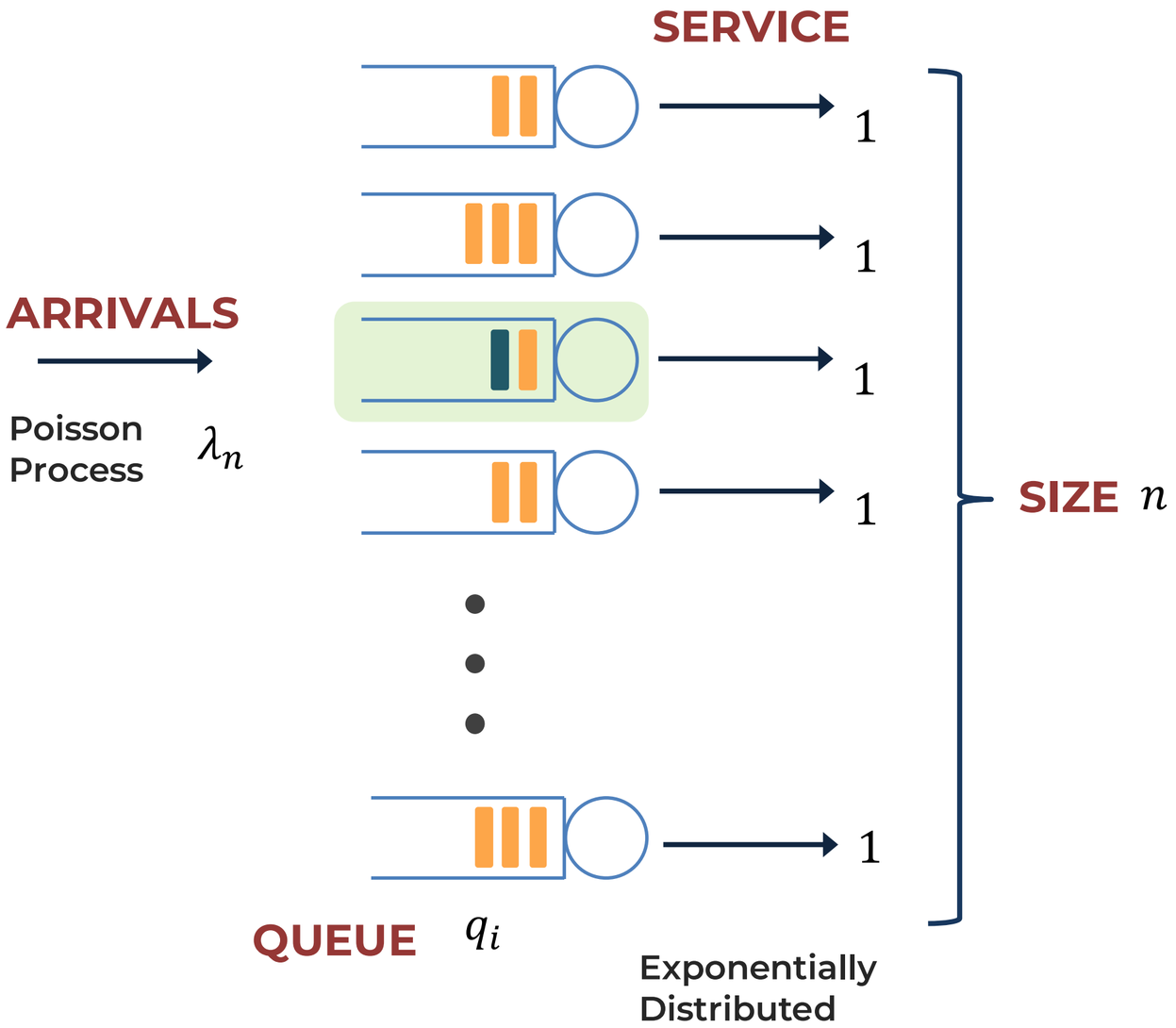
Algorithm Optimality:

Join-the-Shortest Queue (JSQ) **maximizes** the number of customers served

[Winston 1977]

[Weber 1978]

Join-the-Shortest Queue (JSQ)



Analysis Questions

1. How does the system behave w.r.t. SIZE?

Size = n

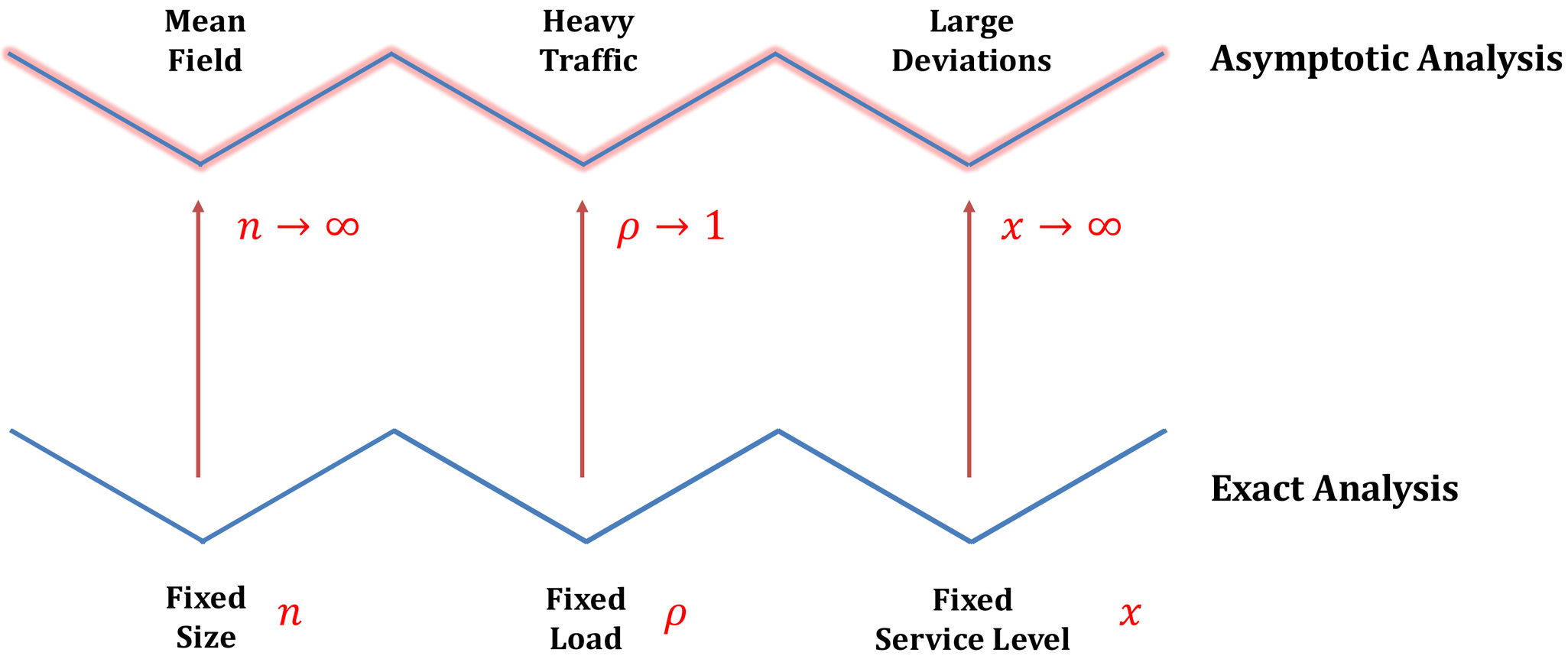
2. How does the system behave w.r.t. LOAD?

Load = $\rho = \frac{\lambda_n}{n}$

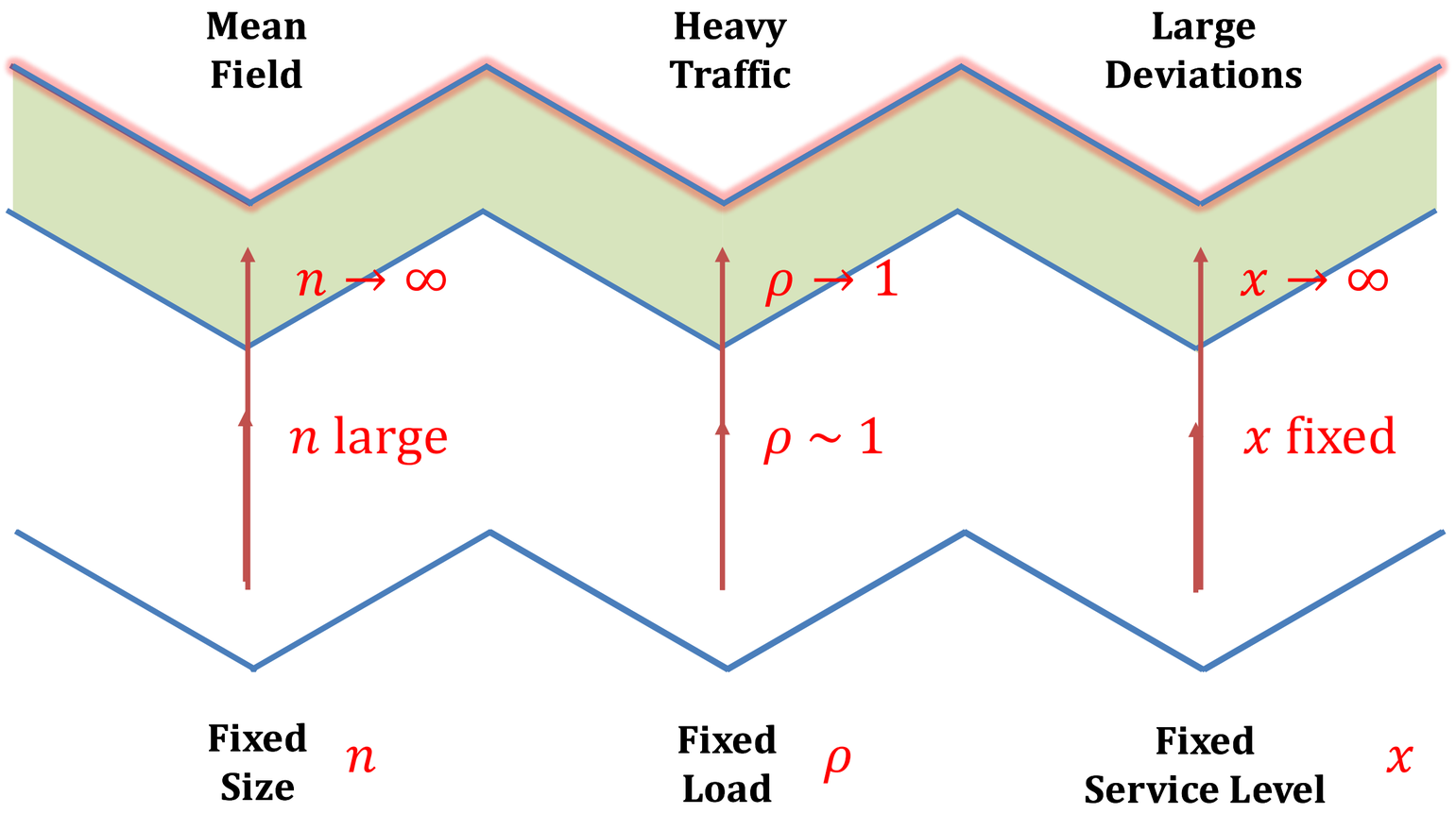
3. How system's SERVICE LEVEL change?

$P(\text{Queue} > x)$

Asymptotic Analysis



Non-Asymptotic Analysis



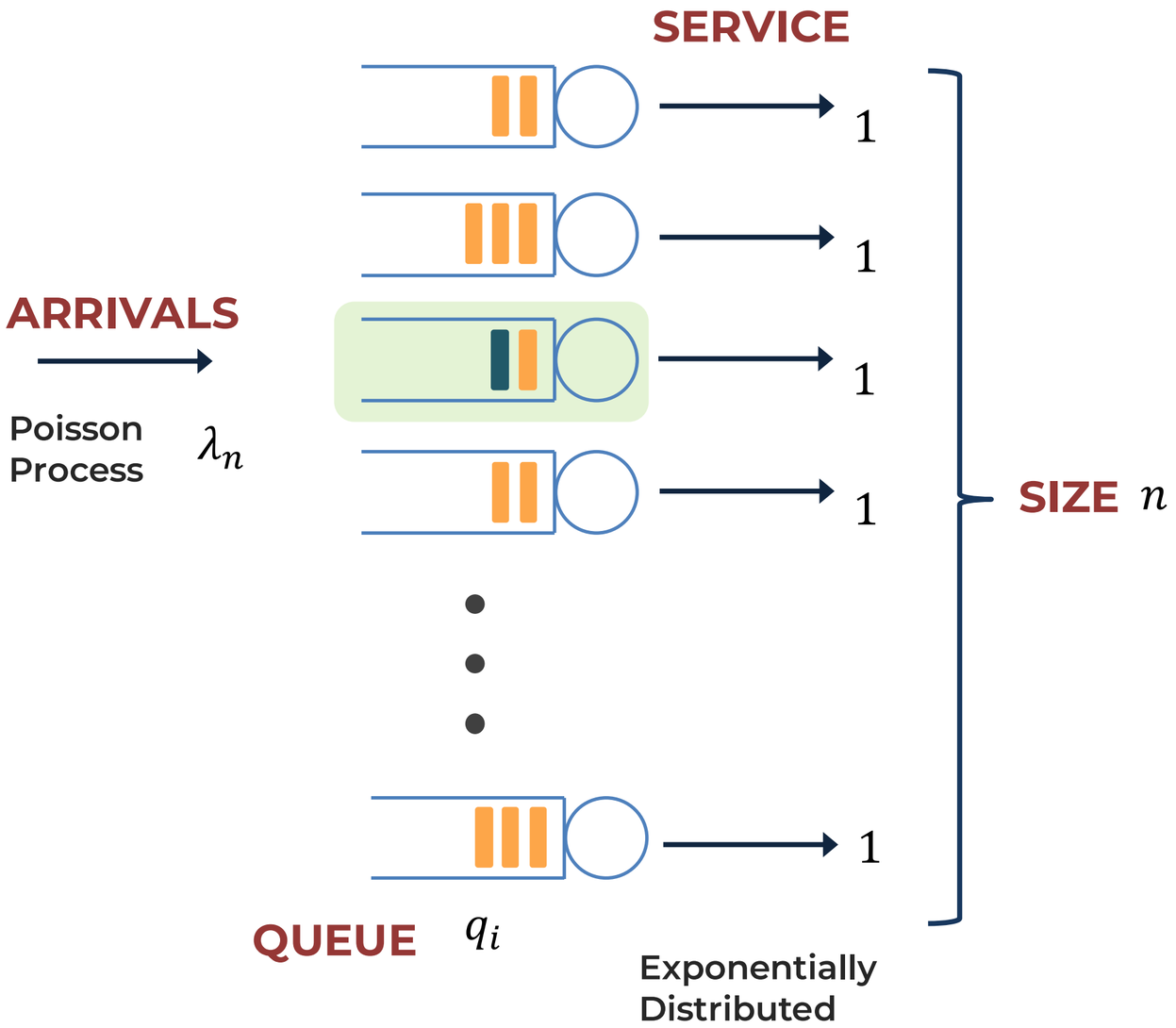
Asymptotic Analysis

Non-Asymptotic Analysis

[Rate of Convergence]

Exact Analysis

Join-the-Shortest Queue (JSQ) Parametrization



Heavy Traffic (HT) parameter:

Stability: $\lambda_n < n$

Many Server-HT: $\lambda_n = n - n^{1-\alpha}, \alpha > 0$

Super-Slowdown: $\alpha > 1$

Classic-HT: Fix n, $\lambda = n(1-\epsilon)$

$\lambda_n = n - n^{1-\alpha}, \alpha \to \infty$

Related Work

Asymptotic Analysis

[Mukherjee et al. 2018]
[Mitzenmacher 2001]
[Liu, Ying 2020]
[Varma, Castro, Maguluri 2022]
[Eschenfeldt, Gamarnik 2018]
[Banerjee, Mukherjee 2019]
[Braverman 2020]
[Liu, Ying 2022]
[Zhao, Banerjee, Mukherjee 2021]

Almost Zero queueing delay

Finite delay

[Gupta, Walton 2019]

[Eryilmaz, Srikant 2012]
[**HL**, Maguluri 2020]
[**HL**, Maguluri 2022]

Infinite queueing delay

$n^{-\alpha} q \xrightarrow{d} \text{Exp}(1)$

Mean Field $\alpha = 0$ $\alpha = 1$ $\alpha = 2$ $\alpha > 2$ $\alpha \to \infty$ **Heavy Traffic**

[Jhunjhunwala, HL, Maguluri 2023]

Non- Asymptotic Analysis

[Braverman 2023],
[**HL**, Maguluri 2022],
[**HL**, Varma, Maguluri, 2021],
[Mukherjee 2022]

$$\mathrm{E}[f(\epsilon q)] = \text{Limiting Value} + O(\epsilon)$$
$$\mathrm{P}(\epsilon q > x) = \text{Limiting Value} + O(\epsilon)$$

Main Contribution: Informal

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime, with large n

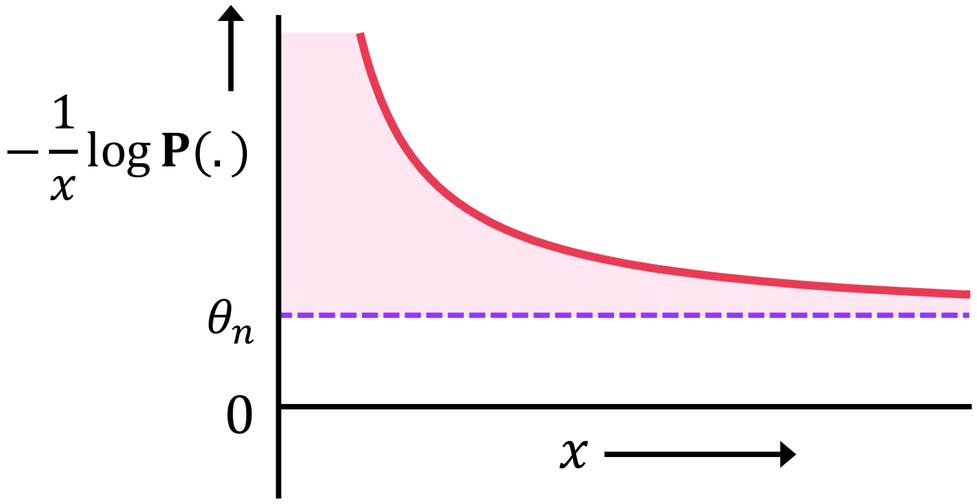
$$\mathrm{Exp}(-\theta_n x) \quad \leq \quad \mathbf{P}(\text{Scaled Queue} > x) \leq \left[1 + \tilde{O}(n^{1-\alpha})\right] \quad 2ex \quad \mathrm{Exp}(-\theta_n x)$$

Main Contribution: Informal

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime, with large n

$$\text{Exp}(-\theta_n x) \quad \leq \quad \mathbf{P}(\text{Scaled Queue} > x) \leq \left[1 + \tilde{O}(n^{1-\alpha})\right] \quad 2ex \quad \text{Exp}(-\theta_n x)$$



Large Deviation:

$$\lim_{x\to\infty} -\frac{1}{x}\log \mathbf{P}(.) = \theta_n$$

Many-Servers

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{Exp}(-\theta_n x) \quad \leq \quad \mathbf{P}(\text{Scaled Queue} > x) \leq \left[1 + \tilde{O}(n^{1-\alpha})\right] \ 2ex \ \mathrm{Exp}(-\theta_n x)$$

Lower Bound

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\text{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(\text{Scaled Queue} > x) \leq \left[1 + \tilde{O}(n^{1-\alpha})\right] \ 2ex \ \text{Exp}(-\theta_n x)$$

Pool all servers together

Lower bound:

For all n, $\quad \mathbf{P}(n^{-\alpha} q > x) \geq e^{-\theta_n x}$

Tail decay rate for SSQ: $\theta_n = n^{\alpha} \log \dfrac{1}{1 - n^{-\alpha}}$

State Space Collapse

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime [$\lambda_n = n - n^{1-\alpha},\ \alpha > 1$], with large n [$n^{1-\alpha} \log n$ is small]

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ \left[1 + \tilde{O}(n^{1-\alpha})\right] \ 2ex \ \mathrm{Exp}(-\theta_n x)$$

State Space Collapse Violation

State Space Collapse

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime [$\lambda_n = n - n^{1-\alpha},\ \alpha > 1$], with large n [$n^{1-\alpha} \log n$ is small]

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ [1 + K n^{1-\alpha} \log n] \ \ 2ex \ \ \mathrm{Exp}(-\theta_n x)$$

JSQ with $n = 2$

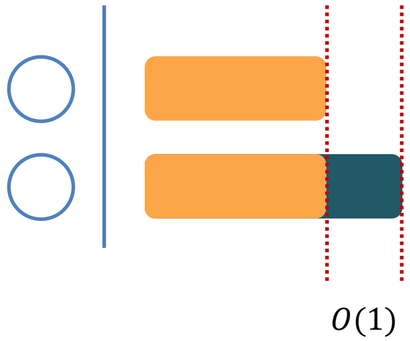
State Space Collapse

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ [1 + K n^{1-\alpha} \log n] \ \ 2ex \ \ \mathrm{Exp}(-\theta_n x)$$

JSQ with $n = 2$



State Space Collapse

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime [$\lambda_n = n - n^{1-\alpha},\ \alpha > 1$], with large n [$n^{1-\alpha} \log n$ is small]

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ [1 + K n^{1-\alpha} \log n] \ \ 2ex \ \ \mathrm{Exp}(-\theta_n x)$$

JSQ with $n = 2$

[Eryilmaz and Srikant 2012]

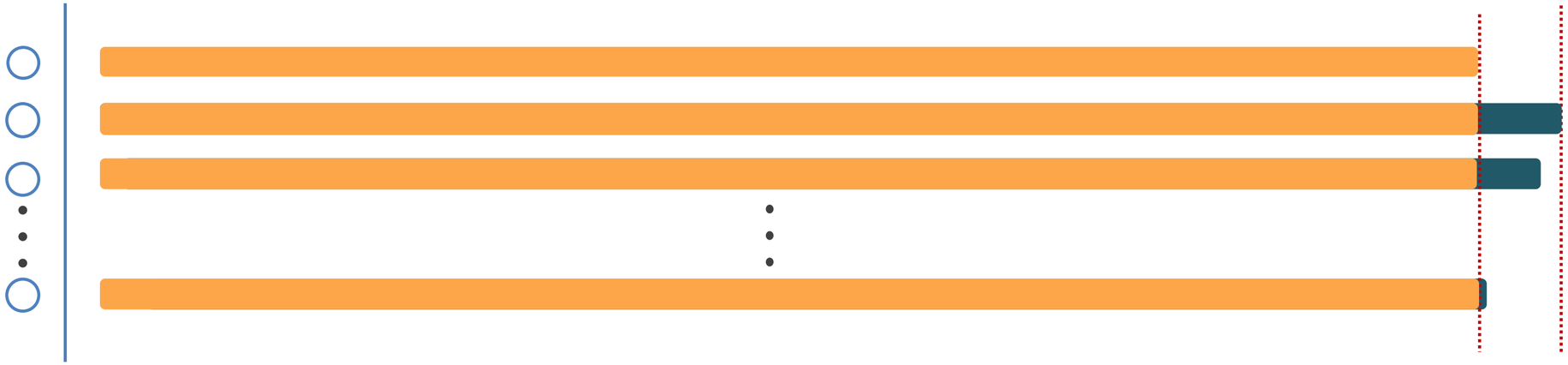
State Space Collapse

JSQ with $n = 2$

[Eryilmaz and Srikant 2012]

JSQ with n large

[Jhunjhunwala, HL and Maguluri 2023] $\boldsymbol{O(1)}$



Improvement: Previously $O(n)$ [Hurtado-Lange and Maguluri 2020]

State Space Collapse

LEMMA

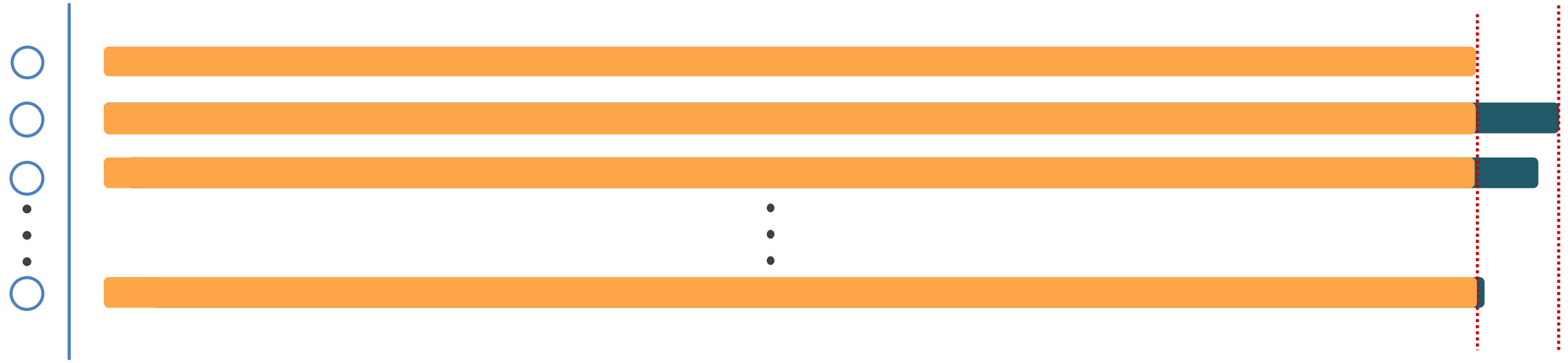
JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{MGF}\left(\mathrm{Q}_{\mathrm{SSQ}}\right) \leq \mathrm{MGF}\left(\mathrm{Q}_{\mathrm{JSQ}}\right) \leq [1 + K n^{1-\alpha} \log n]\ \mathrm{MGF}(\mathrm{Q}_{\mathrm{SSQ}})$$

$\theta \in (0, \theta_n)$

JSQ with n large

[Jhunjhunwala, HL and Maguluri 2023] $\boldsymbol{O(1)}$



Transform Method

$$n^{-\alpha}\ \mathrm{q}_{\mathrm{JSQ}} \approx n^{-\alpha}\ (\mathrm{q}_{\mathrm{SSQ}} + O(n))$$

Transform Method

LEMMA

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{MGF}\left(\mathrm{Q}_{\mathrm{SSQ}}\right) \leq \mathrm{MGF}\left(\mathrm{Q}_{\mathrm{JSQ}}\right) \leq [1 + K n^{1-\alpha} \log n]\ \mathrm{MGF}(\mathrm{Q}_{\mathrm{SSQ}})$$

$\theta \in (0, \theta_n)$

Transform Method

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ [1 + K n^{1-\alpha} \log n] \ \ 2ex \ \ \mathrm{Exp}(-\theta_n x)$$

LEMMA

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha} \log n$ is small$]$

$$\mathrm{MGF}\left(\mathrm{Q}_{\mathrm{SSQ}}\right) \leq \mathrm{MGF}\left(\mathrm{Q}_{\mathrm{JSQ}}\right) \leq [1 + K n^{1-\alpha} \log n]\ \mathrm{MGF}(\mathrm{Q}_{\mathrm{SSQ}})$$

$\theta \in (0, \theta_n)$

Transform Method

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime [$\lambda_n = n - n^{1-\alpha},\ \alpha > 1$], with large n [$n^{1-\alpha}\log n$ is small]

$$\mathrm{Exp}(-\theta_n x) \ \le\ \mathbf{P}(n^{-\alpha} q > x) \le\ [1 + K n^{1-\alpha}\log n]\ \ 2ex\ \ \mathrm{Exp}(-\theta_n x)$$

Pre-exponent term

Pre-limit tail

[Markov's Inequality]

$$\mathbf{P}(n^{-\alpha} q > x) \le [1 + K n^{1-\alpha}\log n] \quad \inf_{\theta \in (0,\theta_n)} e^{-\theta x}\ \mathrm{MGF}(\mathrm{Q}_{\mathrm{SSQ}})$$

Main Contribution: Formal

THEOREM: Non-Asymptotic Tail Bound [JHLM23]

JSQ in super slowdown regime $[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$, with large n $[n^{1-\alpha}\log n$ is small$]$

$$\mathrm{Exp}(-\theta_n x) \ \leq \ \mathbf{P}(n^{-\alpha} q > x) \leq \ [1 + K n^{1-\alpha}\log n] \ \ 2ex \ \ \mathrm{Exp}(-\theta_n x)$$

Non-Asymptotic Many Server-HT

$[\lambda_n = n - n^{1-\alpha},\ \alpha > 1]$

$n^{1-\alpha}\log n$ is small

Non-Asymptotic Large Deviation

Pre-Exponent Terms

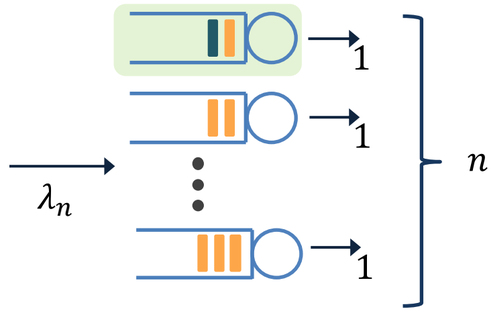
$[1 + K n^{1-\alpha}\log n]\ 2ex$

Rate of Convergence of Tail Decay

$$\theta_n = n^{\alpha}\log\frac{1}{1-n^{-\alpha}}$$

SUMMARY

Thanks!

	JSQ: Many-Server 	
Regime	**Many-Server-HT:** $\lambda_n = n - n^{1-\alpha},\ \alpha > 1$	
	Asymptotic: $n \to \infty$	**Non-Asymptotic:** n is large
SSC	**1-dim**	**SSC Violation:** $[1 + O(n^{1-\alpha} \log n)]$
Transform Eq.	**Explicit Linear Eq.**	**MGF Bound**
Limiting Distribution	 **Exponential**	**Non-Asymptotic Tail Bounds and Large Deviation**