Transform Method for Markov-Modulated Queues

Daniela Hurtado-Lange (Kellogg) and Izzy Grosof (IEMS)
Northwestern University

SNAPP Seminar — December 1st, 2025

Waiting in Line

Ultimate goal: Minimize Delay

Line Sitter and Wai… Services in …

★★★★☆

Queuing up f… energy. Hire s…

✓ Yes, line wa… world!

✓ Your professio… as they wait in li… …uted.

Book Now

Source: TaskRabbit website

…ERAL PUBLIC ON SALE

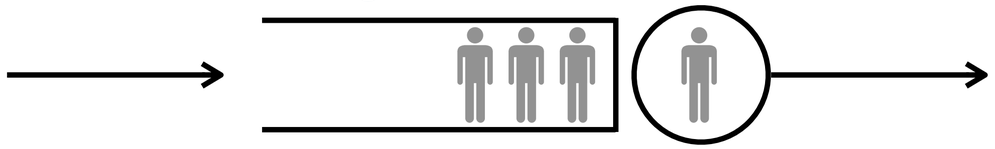
…OURNE CONCERTS STARTS FRI 30 JUN, 2PM (AEST)

Your turn to purchase tickets is coming soon

Next update in 4 seconds

Understanding Delay

Single-Server Queue



✓ Well understood
Distribution of q is known

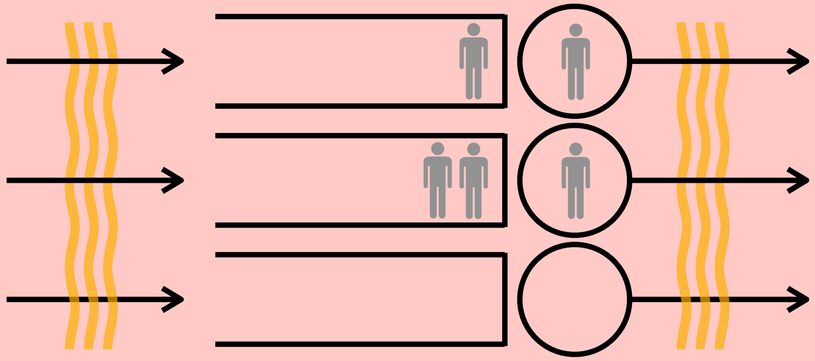
Distribution of $\boldsymbol{q}$?

Mean: $E[\boldsymbol{q}]$? Variance: $Var[\boldsymbol{q}]$?

How about $\sum q_j$?

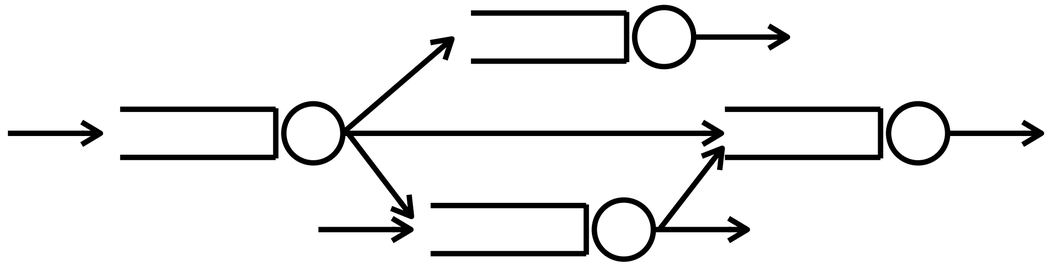
Tail prob: $P(\sum q_j > x)$?

Multiple queues with correlated arrivals and services



This talk

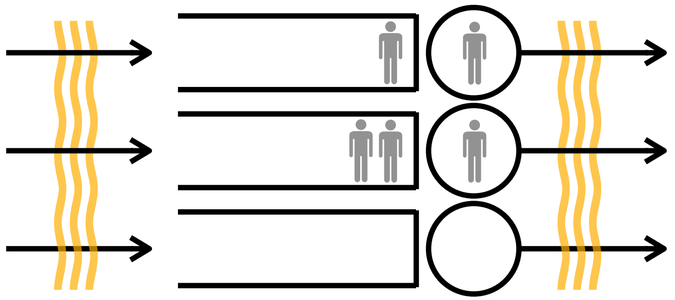
Networks of queues



✓ Jackson networks:
Distribution is product form

In the literature…

Asymptotic Regimes



Large Deviations:
(a.k.a. rare events)
$\lim_{x\to\infty} P(wait > x)$

Ganesh (2004). Big Queues.
Glynn and Whitt (1994)
Puhalskii (1995)
Duffield and O'Connell (1995)
Dupuis and Ellis (1995)
…

Mean Field:
(a.k.a. many servers)
servers $\to \infty$

Vvedenskaya et.al. (1996)
Mitzenmacher (1996, 2001)
Stolyar (2017)
Ying (2017)
Mukkerjee et.al. (2018)
…

Many-Server Heavy-Traffic:
servers $\to \infty$
& Arrival rate $\approx$ Service rate

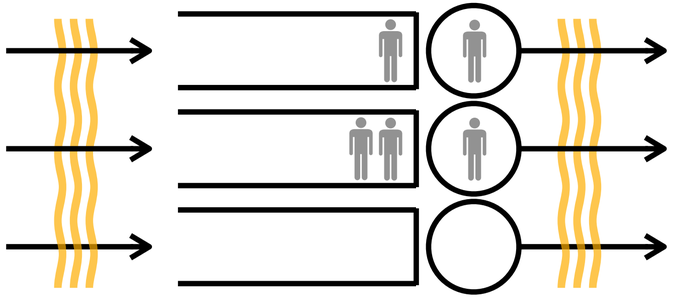
Banerjee et.al. (2019, 2020)
Liu and Ying (2019)
Braverman (2020)
Varma et.al. (2022)
HL, Maguluri (2022)
Jhunjhunwala, **HL**, Maguluri (2023)
…

Heavy Traffic:
Load system to max capacity
$\iff$ Arrival rate $\approx$ Service rate

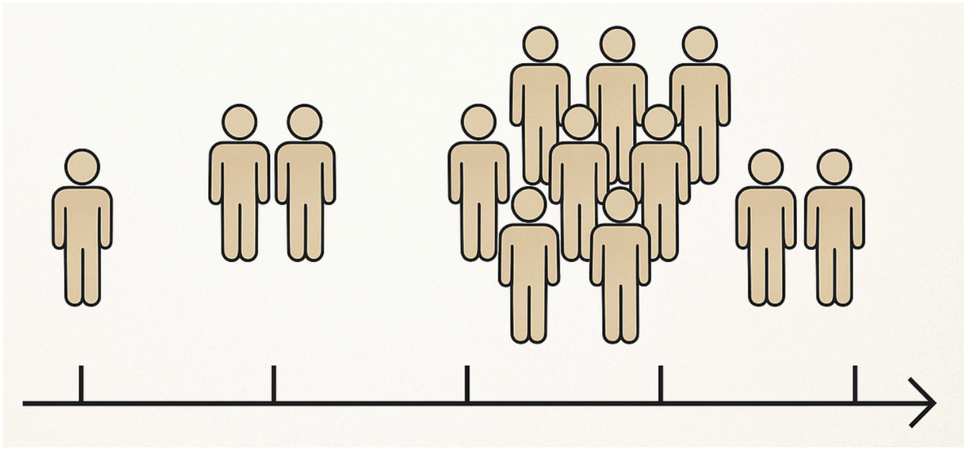
Kingman (1962); Harrison (1998)
Williams (1998, 2000)
Harrison and López (1999)
Stolyar (2004)
Gamarnik and Zeevi (2006)
Eryilmaz and Srikant (2012)
HL, Maguluri (2020, 2021)
HL et.al. (2022)
…

But all these assume constant arrival and service rates

Beyond fixed parameters

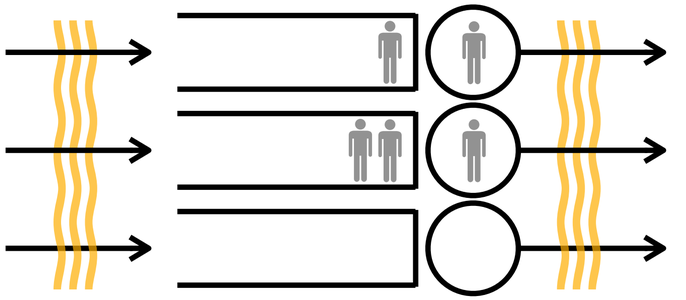


Arrival rate:



Service rate:

Literature (continued)



Asymptotic Regimes

Large Deviations:
(a.k.a. rare events)
$\lim_{x\to\infty} P(wait > x)$

Glynn and Whitt (1994); Puhalskii (1995); Duffield and O'Connell (1995); Dupuis and Ellis (1995); Ganesh (2004)…

Mean Field:
(a.k.a. many servers)
servers $\to \infty$

Many-Server Heavy-Traf
servers $\to \infty$
& Arrival rate $\approx$ Service rate

(2022); **HL**, Maguluri (2022); Jhunjhunwala, **HL**, Maguluri (2023)…

Heavy Traffic:
Load system to max capacity
$\iff$ Arrival rate $\approx$ Service rate

Kingman (1962); Harrison (1998); Williams (1998, 2000); Harrison and López (1999); Stolyar (2004); Gamarnik and Zeevi (2006); Eryilmaz and Srikant (2012); **HL**, Maguluri (2020, 2021); **HL** et.al. (2022)…

Constant arrival and service rates

Our contribution:
✓ Multiple queues and servers
✓ Markov-modulated arrival and service rates, countable state space
✓ Heavy-traffic analysis

Markov Modulated Queues

- Burman, Smith (1986): Markov/M/1 queue in light and heavy traffic. Approximation of mean.
- Prabhu, Zhu (1989): Busy period and mean workload in a Markov/Markov/1 queue
- …91): Markov/G/1 queue analysis
- …kov/G/1 queue-length distribution in
- …m (2005): Single-server queue with …vice times
- …

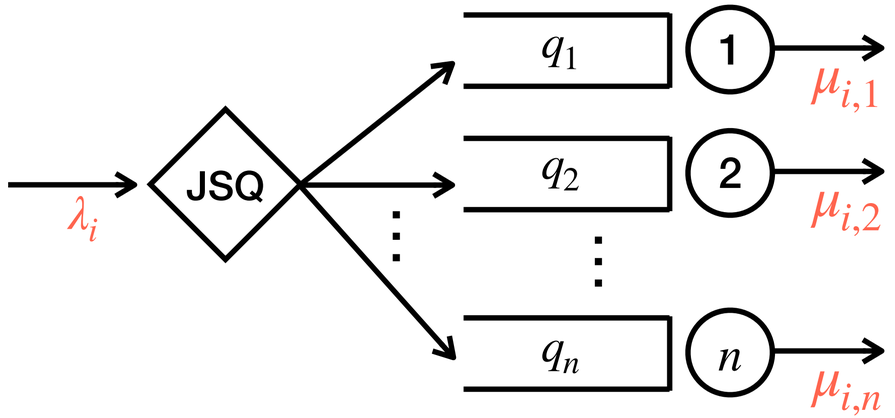
↑ Single queue with finite-state Markov modulation

- Maguluri, Mou (2024): Input-queued switch with Markov-modulated arrivals

Multiple queues, but constant service

JSQ Model

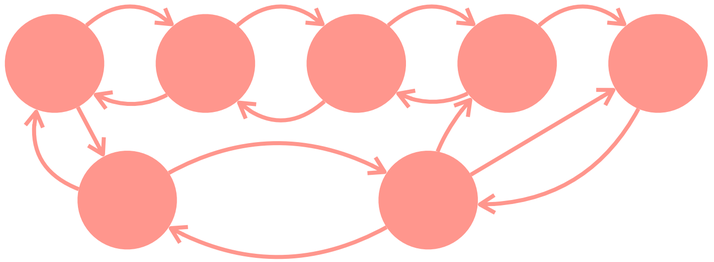
- Load balancing under Join-the-Shortest-Queue (JSQ)
- Exponential inter-arrival and service times
- Heterogeneous servers
- Arrival and service rate are Markov-Modulated



Assumptions on $\{Z(t)\}_t$

- Countable state space
- Stationary distribution exists
- $\lambda_{\max} < \infty$ and $\mu_{\max} < \infty$

$Z(t) = i \sim$ Markov chain



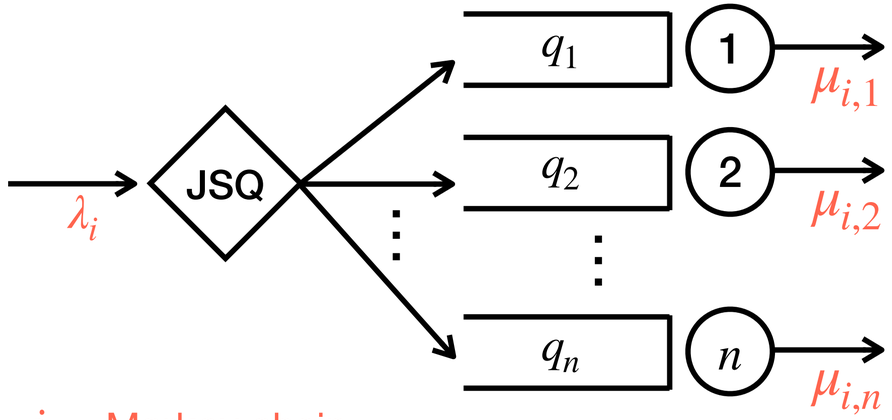
$$\rho_\pi = \frac{\mathbb{E}[\lambda_i \mid i \sim \pi]}{\sum_j \mathbb{E}[\mu_{ij} \mid i \sim \pi]}, \quad \epsilon = 1 - \rho_\pi$$

Asymptotic Distribution of Queue Lengths

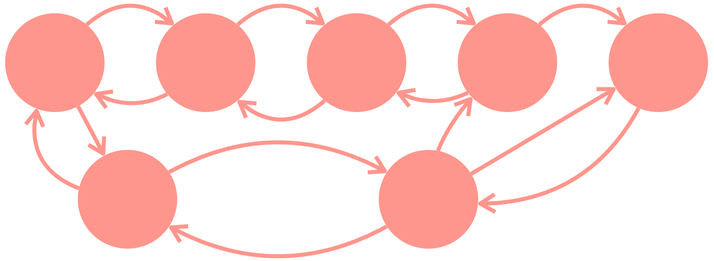
Theorem [HL, Grosof '25]:
If $\lambda_i > 0$ is large enough for each i, then as $\epsilon \downarrow 0$

$$\epsilon \boldsymbol{q} \Longrightarrow Exp\left(\text{Mean} = 1 + \frac{k^*}{\mu_\Sigma}\right)$$

$\approx$ Variance of arrival and service times in steady state



$i \sim$ Markov chain



$$\rho_\pi = \frac{\mathbb{E}[\lambda_i \,|\, i \sim \pi]}{\sum_j \mathbb{E}[\mu_{ij} \,|\, i \sim \pi]}, \quad \epsilon = 1 - \rho_\pi$$

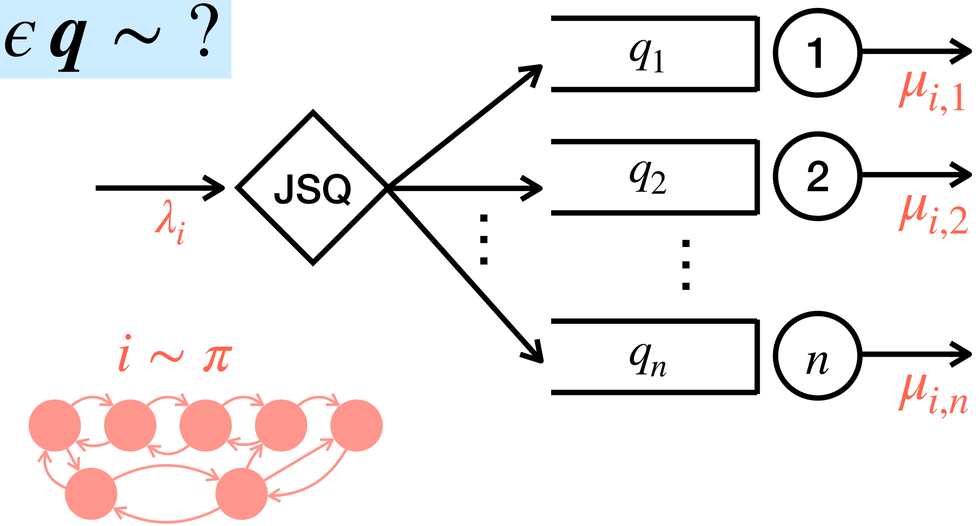
Proof Sketch

Theorem [HL, Grosof '25]:
If $\lambda_i > 0$ is large enough for each i, then as $\epsilon \downarrow 0$

$$\epsilon\, \boldsymbol{q} \Longrightarrow Exp\left(\text{Mean} = 1 + \frac{k^*}{\mu_\Sigma}\right)$$

≈ Variance of arrival and service times in steady state

$\epsilon\, \boldsymbol{q} \sim ?$



Step 1: State Space Collapse (SSC)

$\boldsymbol{q} \approx \left(\frac{q_\Sigma}{n}\right)\mathbf{1}$, so study the following single-server queue:

Step 2: Asymptotic distribution

$$\epsilon\, q_\Sigma \sim ?$$

We use:

- Transform Method
- Poisson Equation

State Space Collapse (SSC)

Proposition [HL, Grosof '25]:

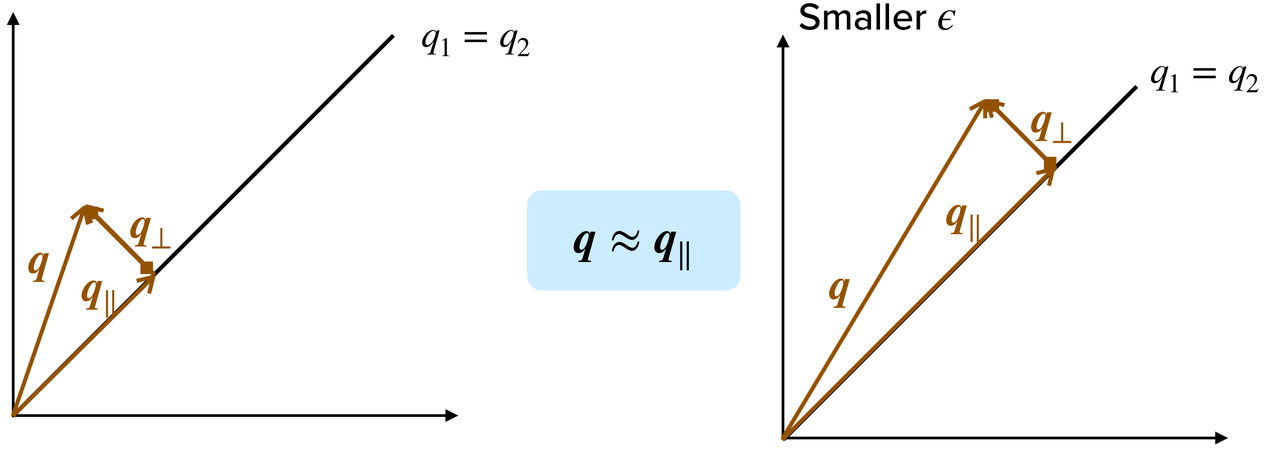
$\boldsymbol{q}_{\parallel}$:= projection of $\boldsymbol{q}$ on $\mathbf{1}$

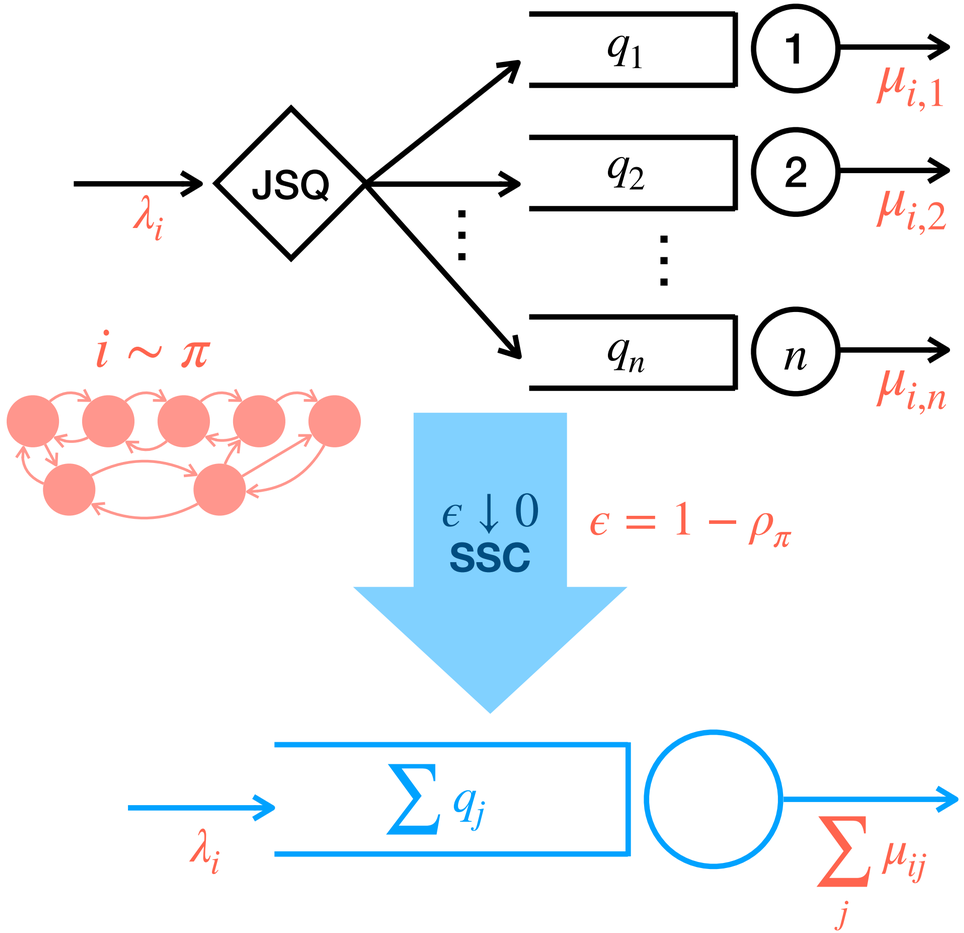
$\boldsymbol{q}_{\perp} := \boldsymbol{q} - \boldsymbol{q}_{\parallel}$

If $\lambda_i > 0$ is large enough for each i, for each $m \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^m\right] \leq M_m$$

Proof: Drift analysis





State Space Collapse (SSC)

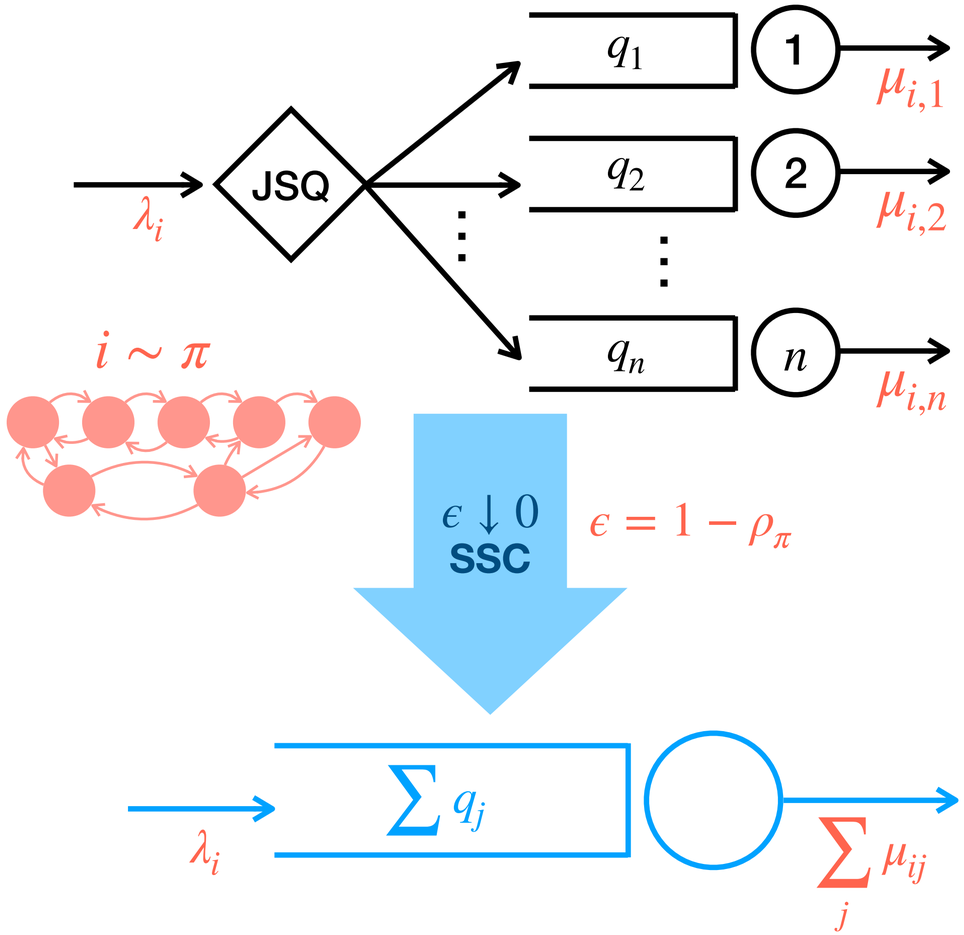
Proposition [HL, Grosof '25]:

$\boldsymbol{q}_{\parallel} :=$ projection of $\boldsymbol{q}$ on $\mathbf{1}$

$\boldsymbol{q}_{\perp} := \boldsymbol{q} - \boldsymbol{q}_{\parallel}$

If $\lambda_i > 0$ is large enough for each i, for each $m \in \mathbb{N}$,

$$\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^m\right] \leq M_m$$



What is λ_i large enough?

$$\lambda_i > \mu_{i\Sigma} - n \min_j \mu_{ij}$$

When does JSQ balance the queue lengths?

- ✓ For homogeneous servers, $\lambda_i > 0$ is enough
- ✓ "Enough" depends on heterogeneity of servers
- ✓ Arrivals need to cover range of service rates
- ✓ JSQ only balances $\boldsymbol{q}$ with *enough arrivals*

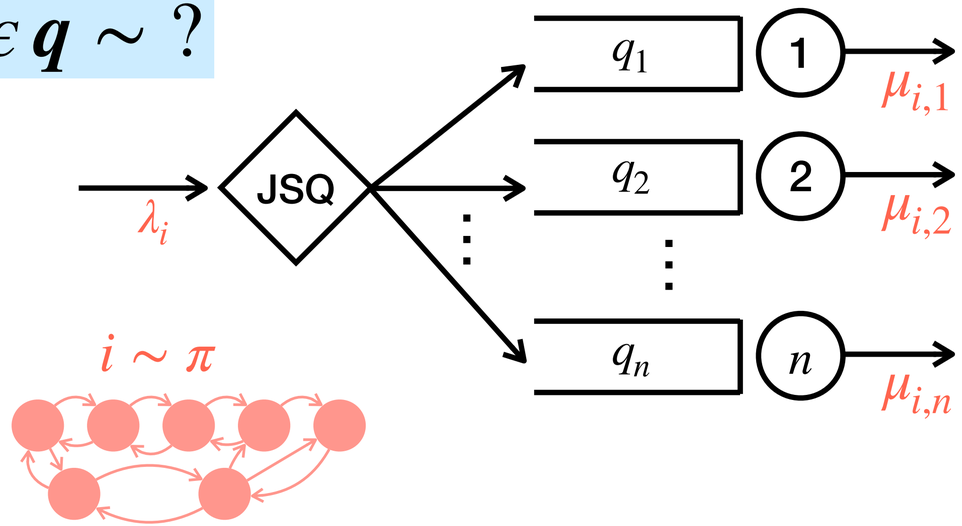
Proof Sketch

Theorem [HL, Grosof '25]:
If $\lambda_i > 0$ is large enough for each i, then as $\epsilon \downarrow 0$

$$\epsilon\, \boldsymbol{q} \Longrightarrow Exp\left(\text{Mean} = 1 + \frac{k^*}{\mu_\Sigma}\right)$$

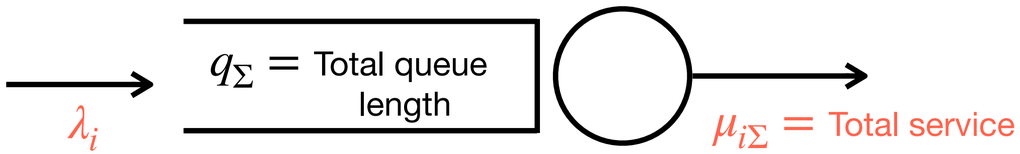
≈ Variance of arrival and service times in steady state

$\epsilon\, \boldsymbol{q} \sim ?$



Step 1: State Space Collapse (SSC)

$\boldsymbol{q} \approx \left(\frac{q_\Sigma}{n}\right)\mathbf{1}$, so study the following single-server queue:



Step 2: Asymptotic distribution

$$\epsilon\, q_\Sigma \sim ?$$

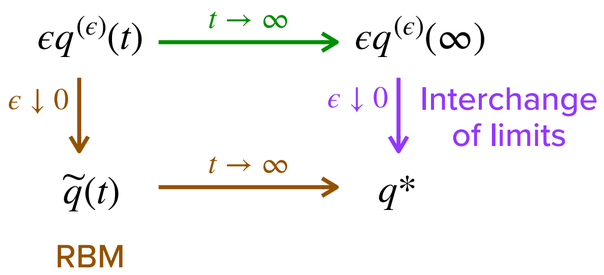
We use:

- Transform Method
- Poisson Equation

Heavy-Traffic Analysis in the Literature

Diffusion Limits Approach

- Most popular
- Introduced by Kingman (1962)
- Show convergence in distribution of queue length scaled heavy-traffic parameter $1-\rho$



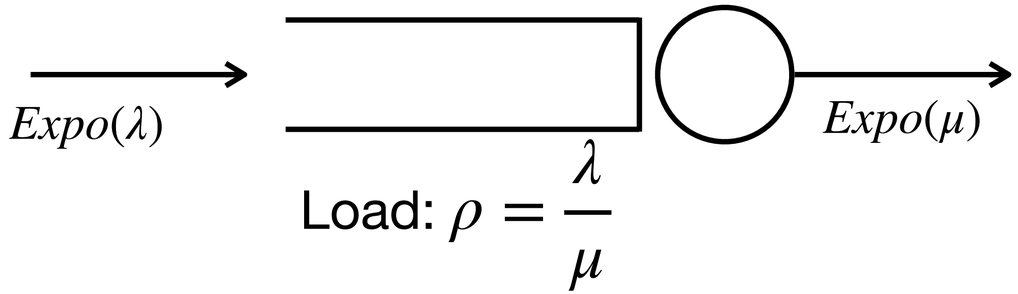
Direct Methods

- **Stein's method**
  Bound Wasserstein distance between $\epsilon q^{(\epsilon)}(\infty)$ and q^*
- **Drift Method**
  Inductively compute $\mathbb{E}\left[(\epsilon q)^k\right]$ for $k \in \mathbb{N}$
- **BAR approach**
  Analyze $e^{\epsilon q(\infty)}$ and bound jumps appropriately
- **Transform Methods**
  Analyze drift of $e^{\epsilon q^{(\epsilon)}(\infty)}$ and directly compute $\mathbb{E}\left[e^{\epsilon q^{(\epsilon)}(\infty)}\right]$
  - ✓ Tractable analysis
  - ✓ Compute distribution
  - ✓ Obtain tail bounds

Transform Method for $M/M/1$ queue

Step 1: Drift of exponential test function

$$\varphi_\theta(q) = e^{\theta q},\ \theta \in \mathbb{R}$$



Drift:

$$\Delta\varphi_\theta(q) = \mathbb{E}\left[e^{\theta q(t+)} - e^{\theta q} \,\middle|\, q(t) = q\right]$$

Arrival Departure

$$\Delta\varphi_\theta(q) = \lambda(e^{\theta(q+1)} - e^{\theta q}) + \mu 1_{\{q>0\}}(e^{\theta(q-1)} - e^{\theta q})$$

$$= e^{\theta q}\left(e^{-\theta} - 1\right)\left(\mu - \lambda e^{\theta}\right) - \mu\left(e^{-\theta} - 1\right) e^{\theta q} 1_{\{q=0\}}$$

$$= 1_{\{q=0\}}$$

Step 2: Set drift to zero

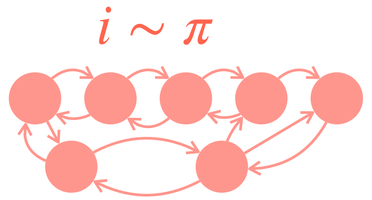
$$\mathbb{E}\left[e^{\theta q}\right] = \frac{\mathbb{P}[q=0]}{1-\rho e^{\theta}}$$

$$\mathbb{E}\left[e^{\theta q}\right] = \frac{1-\rho}{1-\rho e^{\theta}} \iff q \sim \text{Geometric}(1-\rho)$$

Step 3: Set $\theta = 0$

$$\mathbb{P}[q=0] = 1-\rho$$

Transform Method Markov-Mod JSQ



Step 1: Drift of exponential test function

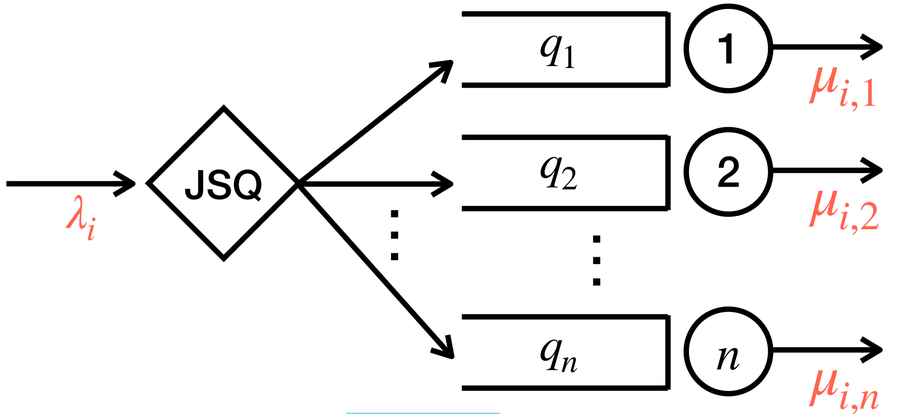
$$\varphi_s(i,\boldsymbol{q}) = e^{-s\epsilon q_\Sigma},\ s>0$$

$$\Delta\varphi_s(i,\boldsymbol{q}) = \left(e^{-s\epsilon}-1\right)e^{-s\epsilon q_\Sigma}\left(\lambda_i - \mu_{i\Sigma}e^{s\epsilon}\right)$$

$$-\left(e^{-s\epsilon}-1\right)e^{-s\epsilon q_\Sigma}\left(\sum_j \mu_{ij}1_{\{q_j=0\}}\right)$$

$$\approx \sum_j \mu_{ij}\,e^{-s\epsilon n q_j}1_{\{q_j=0\}} \quad \text{(SSC)}$$

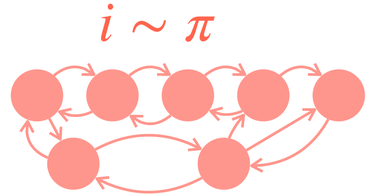
$$= \sum_j \mu_{ij}1_{\{q_j=0\}}$$



$\epsilon \downarrow 0$
SSC

$\epsilon := 1-\rho_\pi$

Transform Method Markov-Mod JSQ



Step 1: Drift of exponential test function

$$\varphi_s(i,\boldsymbol{q}) = e^{-s\epsilon q_\Sigma},\ s>0$$

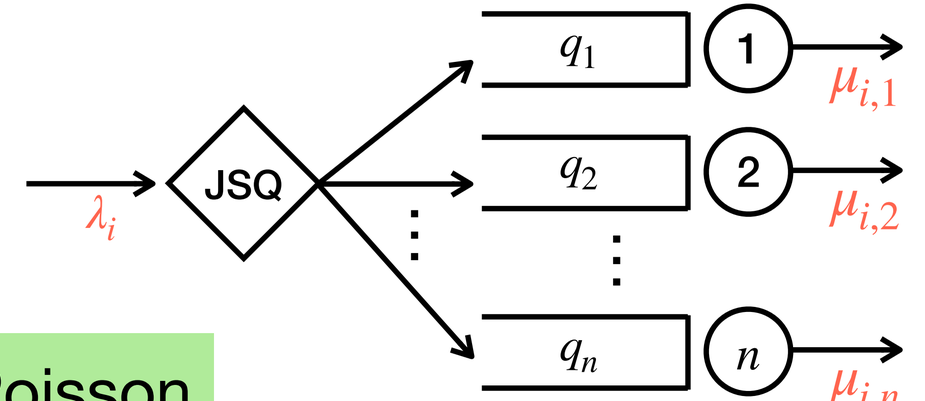
Not independent!

How to compute $\mathbb{E}[\Delta\varphi_s(i,\boldsymbol{q})]$?

$$\Delta\varphi_s(i,\boldsymbol{q}) \approx \left(e^{-s\epsilon}-1\right) e^{-s\epsilon q_\Sigma}\left(\lambda_i - \mu_{i\Sigma} e^{s\epsilon}\right)$$

$$-\left(1-e^{s\epsilon}\right)\Big(\underbrace{\sum_j \mu_{ij} 1_{\{q_j=0\}}}_{\mathbb{E}\left[\sum_j \mu_{ij} 1_{\{q_j=0\}}\right] = \mu_\Sigma \epsilon}\Big)$$

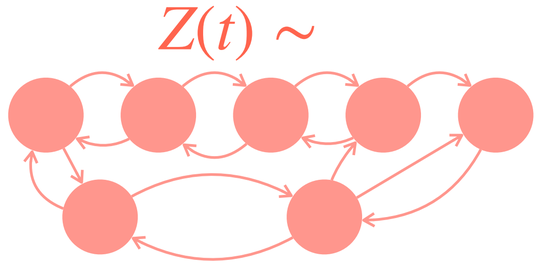
We use the Poisson equation!



$\epsilon \downarrow 0$
SSC

$\epsilon := 1-\rho_\pi$

Poisson Equation



Poisson equation:

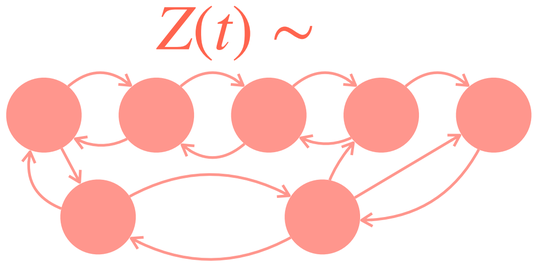
Let $\{Z(t)\}_t$ be a CTMC with countable state space $\mathcal{Z}$ and transition rates $\alpha_{i,i'}$

Consider a function $f : \mathcal{Z} \to \mathbb{R}$ and let $\bar{f} = \mathbb{E}[f(Z)]$. Then, there exists a function $V_f : \mathcal{Z} \to \mathbb{R}$ such that

$$V_f(i) = \frac{f(i) - \bar{f}}{\alpha_{i.}} + \sum_{i' \neq i} \frac{\alpha_{ii'}}{\alpha_{i.}} V_f(i')$$

$$\iff \underbrace{\sum_{i' \neq i} \alpha_{ii'} \left[V_f(i') - V_f(i) \right]}_{\text{Drift of solution } V_f(i)} = \underbrace{\bar{f} - f(i)}_{\text{Difference between average and } f(i)}$$

Poisson Equation for Transform Method



Poisson equation:

Let $\{Z(t)\}_t$ be a CTMC with countable state space $\mathcal{Z}$ and transition rates $\alpha_{i,i'}$

Consider a function $f : \mathcal{Z} \to \mathbb{R}$ and let $\bar{f} = \mathbb{E}[f(Z)]$. Then, there exists a function $V_f : \mathcal{Z} \to \mathbb{R}$ such that

$$V_f(i) = \frac{f(i) - \bar{f}}{\alpha_{i.}} + \sum_{i' \neq i} \frac{\alpha_{ii'}}{\alpha_{i.}} V_f(i')$$

Theorem [HL, Grosof '25]**:**

For any function $f : \mathcal{Z} \to \mathbb{R}$ such that $V_f(i)$ exists and $\mathbb{E}\left[|V_f(i)|^{1+\eta} \right] \leq C$ for some $\eta > 0$,

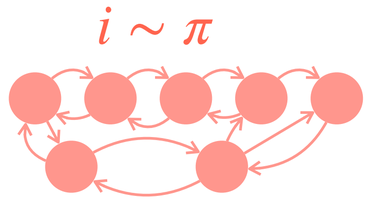
$$Cov\left(e^{-s\epsilon q_\Sigma}, f(i)\right) = \mathbb{E}\left[e^{-s\epsilon q_\Sigma} f(i)\right] - \mathbb{E}\left[e^{-s\epsilon q_\Sigma}\right] \bar{f} = (e^{-s\epsilon} - 1)\mathbb{E}\left[e^{-s\epsilon q_\Sigma} V_f(i) (\lambda_i - \mu_{i\Sigma})\right] + O\left(\epsilon^{2 - \frac{1}{1+\eta}}\right)$$

✓Product of expectations

Solution to Poisson equation

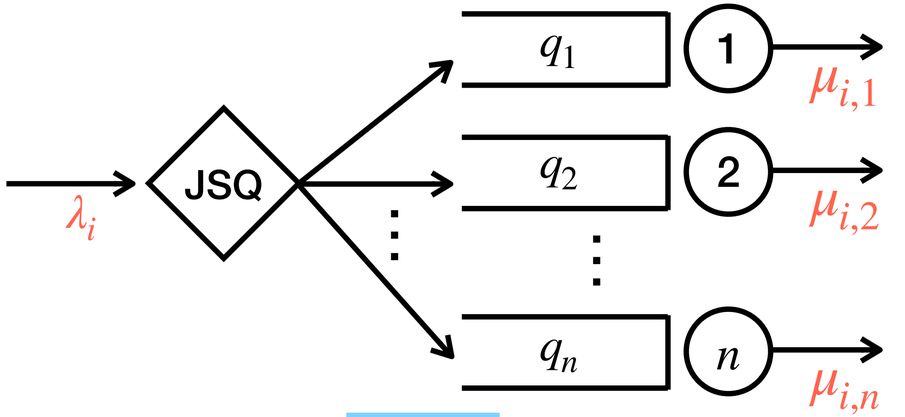
Error term

Transform Method Markov-Mod JSQ



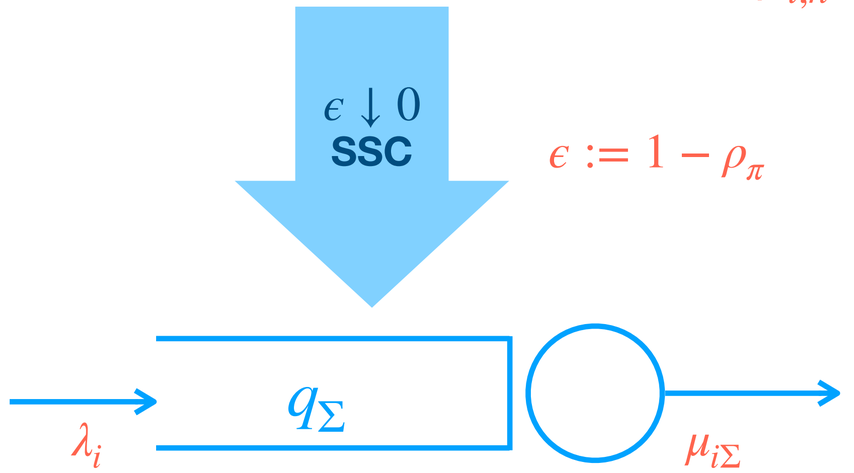
Step 1: Drift of exponential test function

$$\Delta\varphi_s(i,\boldsymbol{q}) \approx \left(e^{-s\epsilon}-1\right) e^{-s\epsilon q_\Sigma}\left(\lambda_i - \mu_{i\Sigma}\, e^{s\epsilon}\right)$$
$$-\left(1-e^{s\epsilon}\right)\Big(\sum_j \mu_{ij} 1_{\{q_j=0\}}\Big)$$



Step 2: Expectation under steady state

$$\underbrace{\mathbb{E}\left[e^{-s\epsilon q_\Sigma}\left(e^{s\epsilon}\mu_{i\Sigma} - \lambda_i\right)\right]}_{\text{Step 3: Our theorem!}} = e^{s\epsilon}\,\underbrace{\mathbb{E}\Big[\sum_j \mu_{ij} 1_{\{q_j=0\}}\Big]}_{=\mu_\Sigma\,\epsilon} + O(\epsilon^2)$$



Applying the Poisson Equation

Set drift to zero: $\underbrace{\mathbb{E}[e^{-s\epsilon q_\Sigma}(e^{s\epsilon}\mu_{i\Sigma} - \lambda_i)]} = e^{s\epsilon}\,\mu_\Sigma\epsilon + O(\epsilon^2)$

$$= \mathbb{E}[e^{-s\epsilon q_\Sigma}(\mu_{i\Sigma} - \lambda_i)] + s\epsilon\mathbb{E}[e^{-s\epsilon q_\Sigma}\mu_{i\Sigma}] + O(\epsilon^2)$$

Poisson equation theorem on lhs with $f(i) = h(i) := \mu_{i\Sigma} - \lambda_i$ and $f(i) = \ell(i) := \mu_{i\Sigma}$:

$$LHS = \mathbb{E}[e^{-s\epsilon q_\Sigma}](1+s)\epsilon\mu_\Sigma + (1 - e^{-s\epsilon})\,\mathbb{E}[e^{-s\epsilon q_\Sigma}V_h(i)(\mu_{i\Sigma} - \lambda_i)] + O\left(\epsilon^{2-\frac{1}{1+\eta}}\right)$$

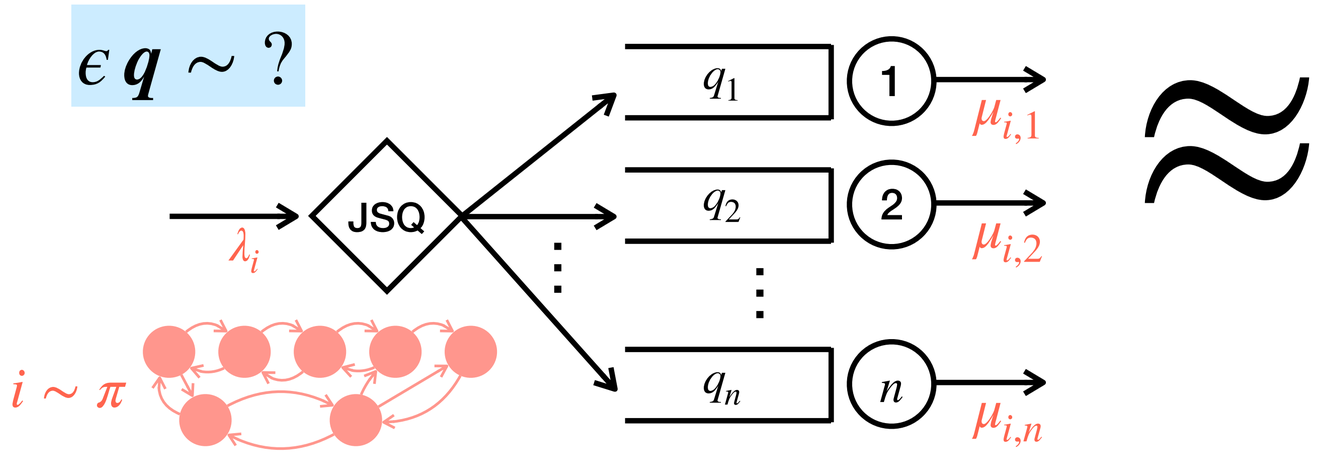
Poisson equation theorem on rhs with $f(i) = k(i) := V_h(i)(\mu_{i\Sigma} - \lambda_i)$:

$$\mathbb{E}[e^{-s\epsilon\, q_\Sigma}V_h(i)(\mu_{i\Sigma} - \lambda_i)] = \mathbb{E}[e^{-s\epsilon q_\Sigma}]\,\mathbb{E}[k(i)] + O(\epsilon)$$

Then, put everything together. ■

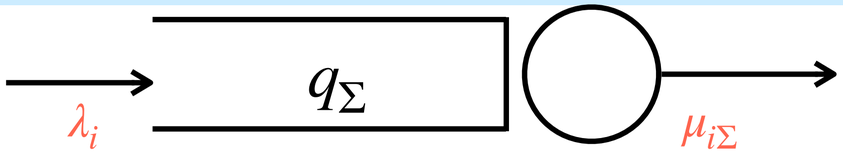
Key Takeaways

Thanks! Questions?



$\approx$

Step 1: State Space Collapse



- ✓ Only need λ_i large enough so JSQ balances queue lengths for all i
- ✓ Queue can be unstable for some i

Theorem [**HL**, Grosof '25]:

$$\mathbb{E}\left[e^{-s\epsilon q_\Sigma}\right] = \frac{1}{1 + s\left(1 + \frac{\mathbb{E}[k(i)]}{\mu_\Sigma}\right)} + O\left(\epsilon^{2-\frac{1}{1+\eta}}\right)$$

with $k(i) = V_h(i)(\mu_{i\Sigma} - \lambda_i)$ and $h(i) = \mu_{i\Sigma} - \lambda_i$

Step 2: Asymptotic distribution

Transform Method:

$$\varphi_s(i, \boldsymbol{q}) = e^{-s\epsilon q_\Sigma},\ \mathbb{E}[\Delta\varphi_s(i, \boldsymbol{q})] = 0$$

+ Poisson Equation