Performance Analysis of Data Center Networks: Drift Method and Transform Techniques

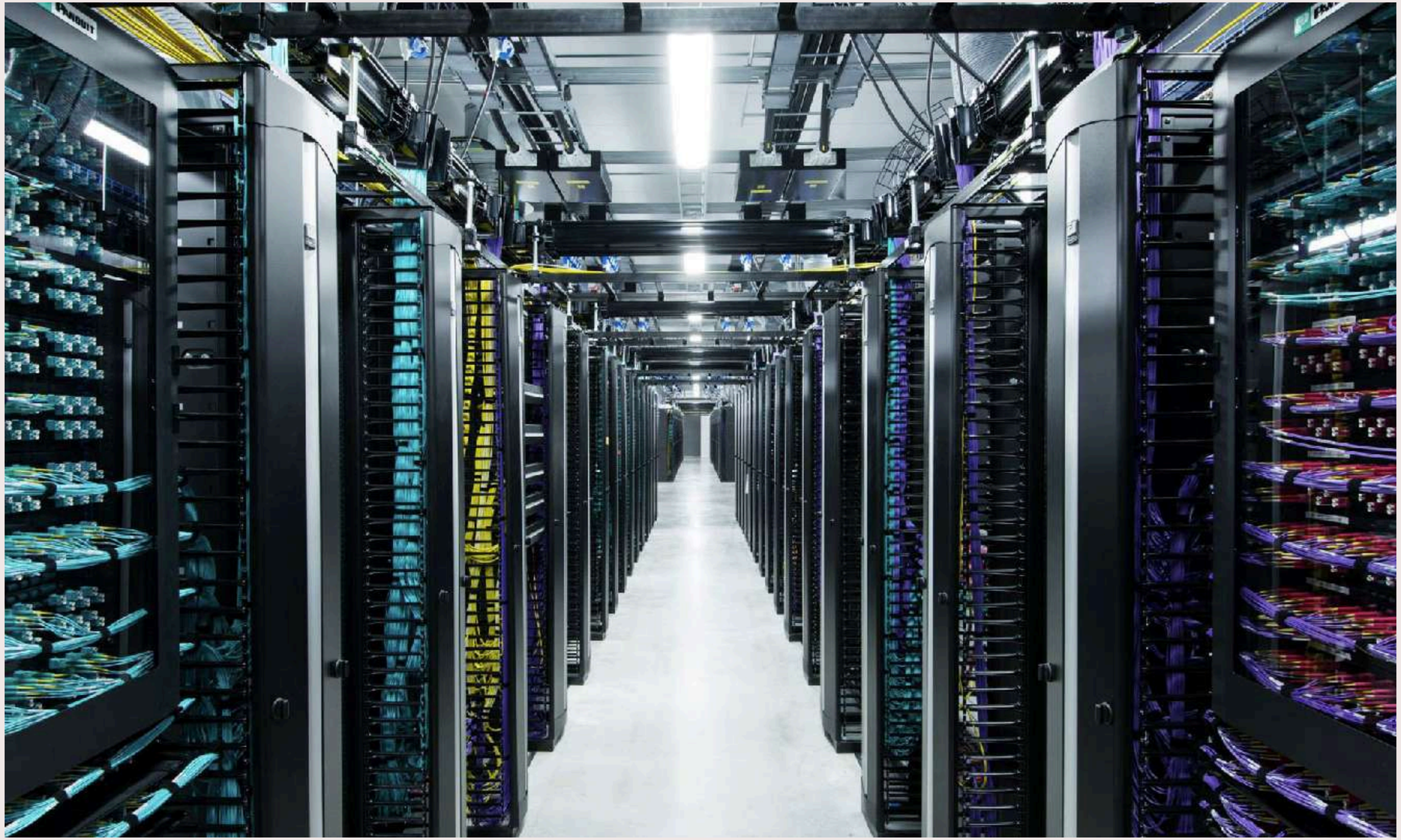
Daniela Hurtado-Lange
William & Mary
Joint work with Siva Theja Maguluri

Kellogg School of Business, October 6th 2022

Contact information: dahurtadolange@wm.edu

Motivation

Data Centers



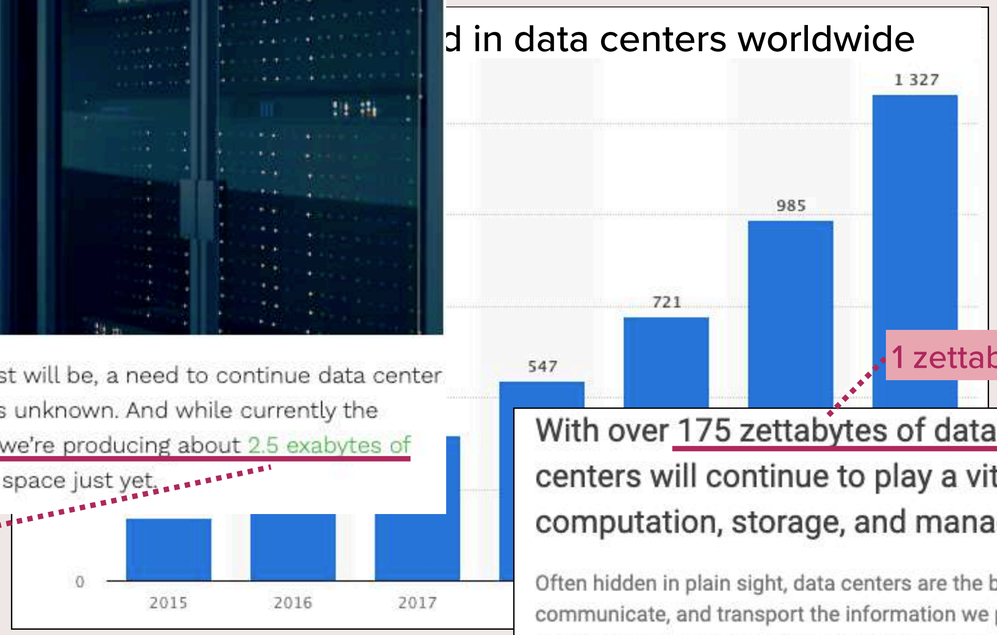
Data Storage and Production

Are There Data Center Storage Capacity Constraints?

There's no argument in the data world that there is, or at least will be, a need to continue data center and data storage expansion, but what that future looks like is unknown. And while currently the world's data centers store about 1,327 exabytes of data, and we're producing about 2.5 exabytes of data daily, we're not in danger of running out of data storage space just yet.

Source: Are data centers running out of storage?, VXCHNGE, available at https://www.vxchnge.com/blog/are-data-centers-running-out-of-storage

1 exabyte = 1M terabytes



Source: https://www.statista.com/statistics/

1 zettabyte = 1B terabytes

With over 175 zettabytes of data expected by 2025, data centers will continue to play a vital role in the ingestion, computation, storage, and management of information.

Often hidden in plain sight, data centers are the backbone of our internet. They store, communicate, and transport the information we produce every single day. The more data we create, the more vital our data centers become.

But many of today's data centers are clunky, inefficient, and outdated. To keep them running, data center operators, from FAMGA to colocation providers, are working on upgrading them to fit our ever-changing world.

Source: The future of data center, CB insights, available at https://www.cbinsights.com/research/future-of-data-centers/

29 terabytes per second

Significance of Delay

Google

400 ms delay implies:

- ↓0.59% # searches per user
- ↓20% traffic

Bing

Additional 1.5 s delay implies:

- ↓1.8% queries per user
- ↓4.3% revenues per user
- ↓3.8% overall satisfaction

amazon

100 ms delay implies:

↓1% revenues ≈ ↓$745 million/year

Goal: Minimize delay

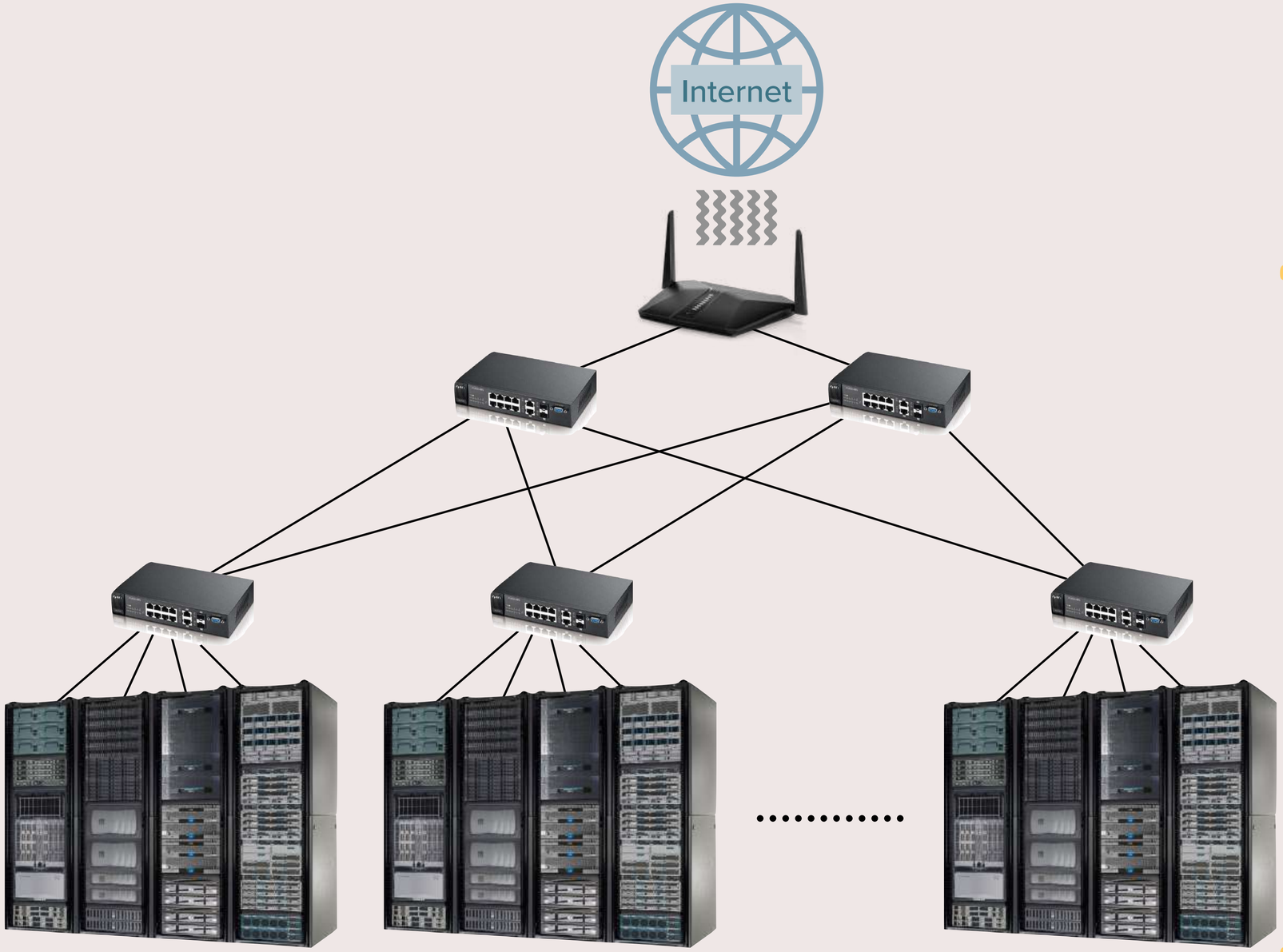
This talk:

Understand probabilistic behavior of delay

Outline

Question 1: **Expected delay and Drift method**	• Expected delay in data centers in heavy-traffic • General result • Proof sketch
Question 2: **Tail bounds and Transform techniques**	• The single server queue • Systems with a single bottleneck • The load balancing system
Overview of other work	• Rate of convergence to heavy traffic • Load balancing with heterogeneous servers • The many-server heavy-traffic regime • Healthcare systems
Conclusion and future work	

Data Centers



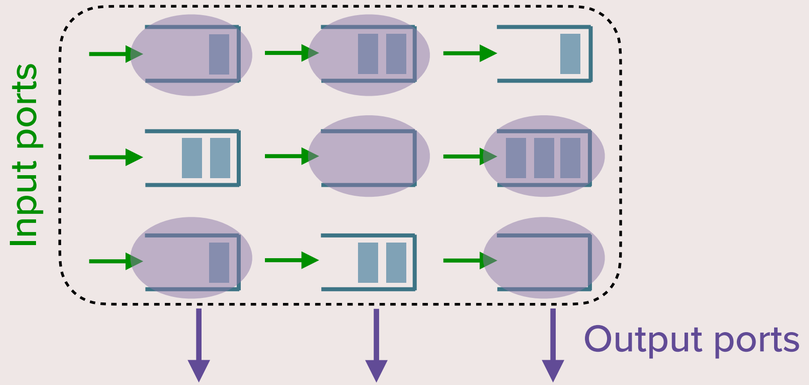
Modeled as a huge switch

Expected delay in a switch?

Diagram based on https://www.researchgate.net/figure/A-data-center-network-with-aggregation-and-TOR-switches_fig1_278021021

The Input-Queued Switch

- Discrete time model
- n input ports and n output ports
  - Jobs arrive at input ports, and go to the desired output port
  - Jobs are processed at output ports and take exactly one time slot
- **Constraint:** At most **one** job can be processed from each input port and at each output port
- **Scheduling:**
  - **Which queues to serve?** Largest total queue length



Shah, Tsitsiklis, Zhong (2011):

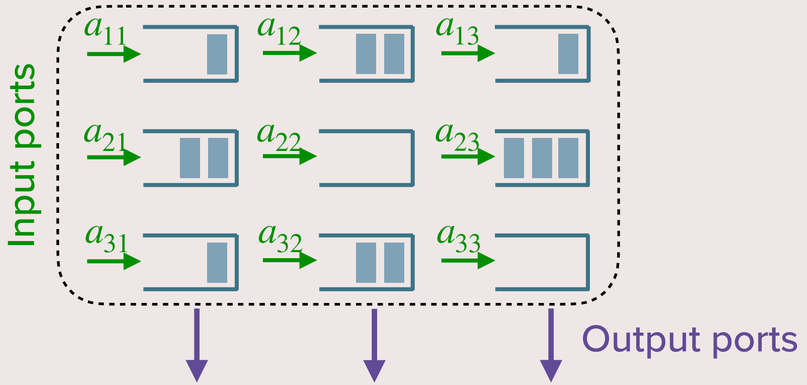
*Switches are, in our opinion, the **simplest non-trivial** example of a stochastic processing network. Over the years, it has served as a **guiding example** for designing as well as analyzing scheduling policies.*

- **First breakthrough:** Maguluri and Srikant (2016): Input-queued switch under **independent** arrival processes
- But the arrival processes are **not independent in data centers**

The Input-Queued Switch

Heavy-traffic analysis:

- Load system close to its maximum capacity
- Define
  $\epsilon := 1 -$ Arrival rate to each row/column
- Take $\epsilon \downarrow 0$



Theorem: [HL, Maguluri '21]

In an $n \times n$ input-queued switch,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i,j} q_{i,j}\right] = \frac{n}{2} \sum_{ij} \left[\underbrace{\sum_{i'} \frac{Cov[a_{ij}, a_{i'j}]}{n}}_{\text{Row average}} + \underbrace{\sum_{j'} \frac{Cov[a_{ij}, a_{ij'}]}{n}}_{\text{Column average}} - \underbrace{\sum_{i'j'} \frac{Cov[a_{ij}, a_{i'j'}]}{n^2}}_{\text{Total average}} \right]$$

Heavy-traffic behavior of queue lengths

What is the expected delay?

Bernoulli arrivals, under independence assumption:

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\text{Delay}\right] = 1 - \frac{3}{2n} + \frac{1}{2n^2}$$

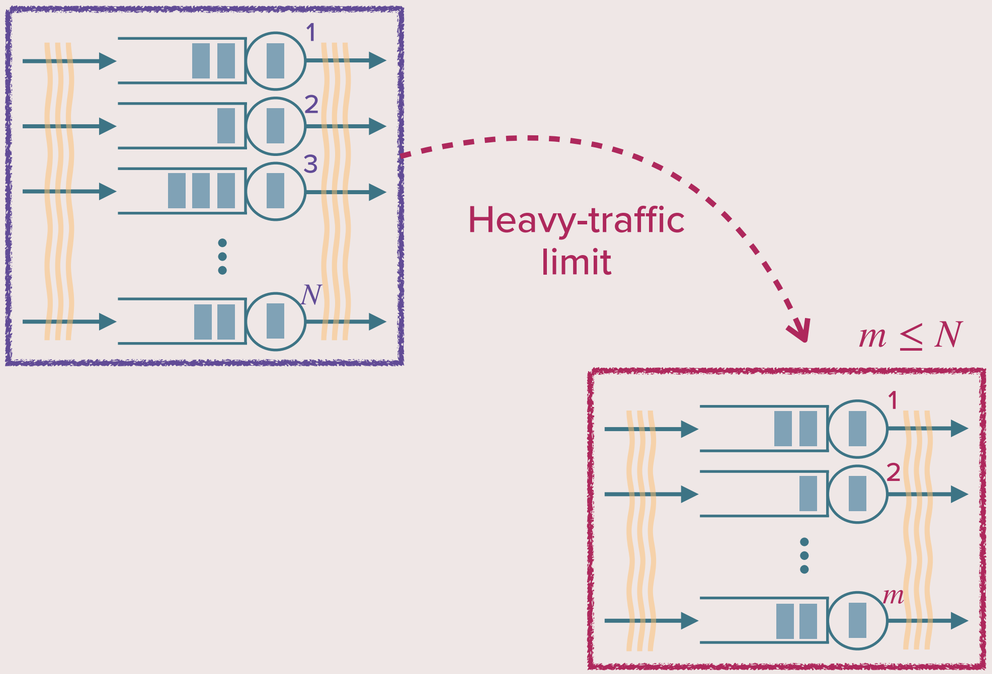
Does not grow with the size of the switch

Heavy-Traffic Analysis

- Load the system close to maximum capacity:

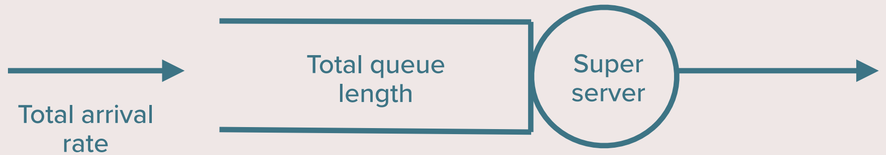
Arrival rate ≈ Service rate

- State Space Collapse (SSC)



Complete Resource Pooling (CRP):

- SSC to a one-dimensional subspace
- System behaves as a single server queue
- All servers “pool” together and behave as a super server



The CRP Condition

Under CRP:

- Multi-dimensional queueing system behaves as single-server queue in heavy traffic
- Vast literature:
  - **Diffusion limits:**
    Kingman (1962); Harrison (1988); Harrison (1998); Williams (1998); Williams (2000); Harrison & López (1999); Stolyar (2004); Gamarnik and Zeevi (2006)…
  - **Drift method:**
    Eryilmaz and Srikant (2013)
  - **Transform methods:**
    HL and Maguluri (2020)
    Second part of this talk

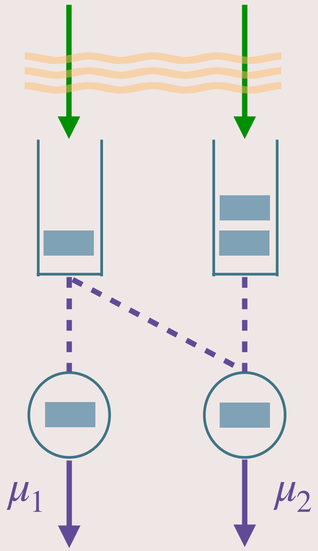
CRP is **not satisfied** in data center networks

CRP not satisfied:

- Multi-dimensional queueing system in heavy traffic
- Open question for a long time
- Results for particular systems:
  - **Input-queued switch:**
    Maguluri and Srikant (2016); Maguluri, Burle, Srikant (2018)
  - **Bandwidth sharing networks:**
    Wang, Maguluri, Srikant and Ying (2018)

The $\mathcal{N}$-system

- Model for production systems
- Each queue has an arrival process
  - They might be correlated
- Two types of servers:
  - Server 1: Dedicated to jobs from queue 1
  - Server 2: Flexible



Ghamami and Ward (2013):

*The $\mathcal{N}$-system is one of the simplest parallel server system models that **retains much of the complexity** inherent to more general models*

Heavy-traffic analysis:

- Define
  $\epsilon :=$ Total service rate $-$ Total arrival rate
- Take $\epsilon \downarrow 0$

Theorem: [HL, Maguluri '21]

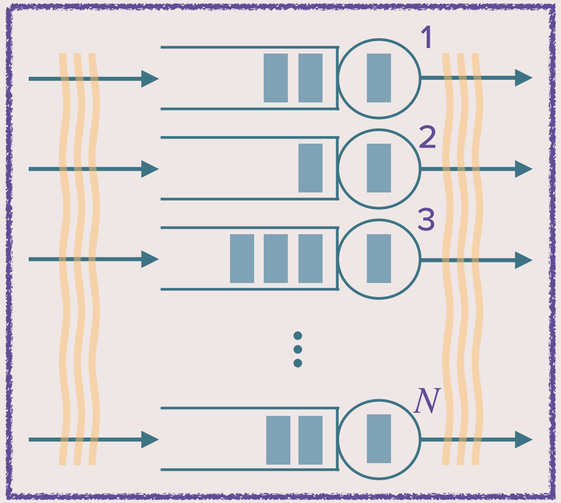
In an $\mathcal{N}$-system, with fixed service rate

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\mu_1 q_1 + \mu_2 q_2\right] = \frac{\sigma_{a1}^2 + \sigma_{a2}^2}{2}$$

Does not depend on correlation between arrival processes!

Full-Dimensional State Space Collapse

- N queues with its own arrival and service processes
  - Arrival process to different queues might be correlated
  - Service process from different queues might be correlated
- SSC occurs into an N-dimensional subspace



Theorem: [HL, Maguluri '21]

Consider a queueing system that satisfies full-dimensional SSC, with fixed service rates. Then,

$$\lim_{\epsilon \downarrow 0} \epsilon \mathbb{E}\left[\sum_{i=1}^{N} q_i\right] = \frac{1}{2}\sum_{i=1}^{N} \sigma_{ai}^2$$

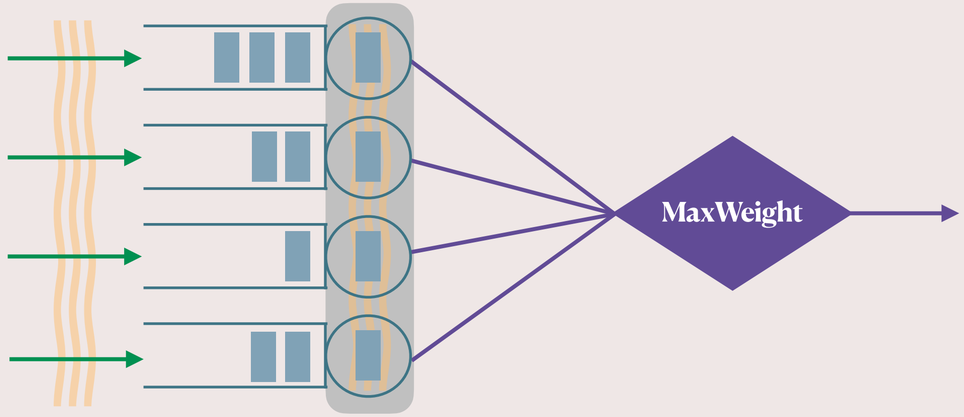
Does not depend on correlation among arrival processes!

General Result: The Model

The generalized switch model

- Subsumes several systems that arise in data centers
- Discrete time model with N queues
- Each queue has its own arrival process
  - i.i.d. across time
  - Correlated among queues: Σ_a is the covariance matrix
- Interference constraints among servers
  - Scheduling problem solved in each time slot
- Channel state: Conditions of the environment
  - Determines interference constraints
  - i.i.d. across time
- **MaxWeight:** Given the channel state and the vector of queue lengths $\boldsymbol{q}$,

$$\text{Service vector} \in \arg\max_{\boldsymbol{x} \in \mathcal{S}^{(m)}} \langle \boldsymbol{x}, \boldsymbol{q} \rangle$$



Channel state: $M(k) = m$
Feasible service rates: $\mathcal{S}^{(m)}$

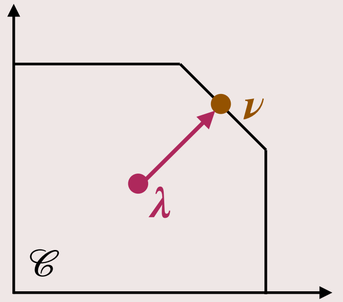
Heavy-traffic analysis:

$\mathscr{C}$ = Arrival rates such that the system can be stable

$\boldsymbol{\nu} \in \partial\mathscr{C}$

$\boldsymbol{\lambda} = (1-\epsilon)\boldsymbol{\nu}$, with $\epsilon \in (0,1)$

Take limit as $\epsilon \downarrow 0$



General Result

Expected weighted queue length

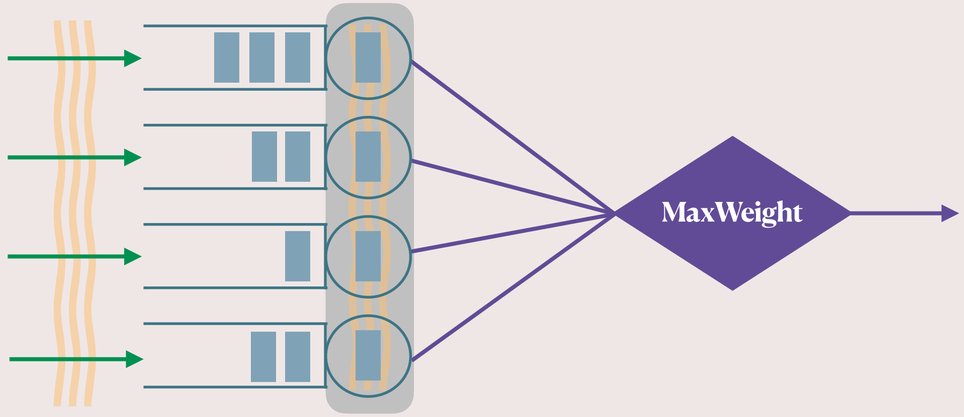
Variability of arrival process

Variability of service process

Theorem: [HL, Maguluri '20]

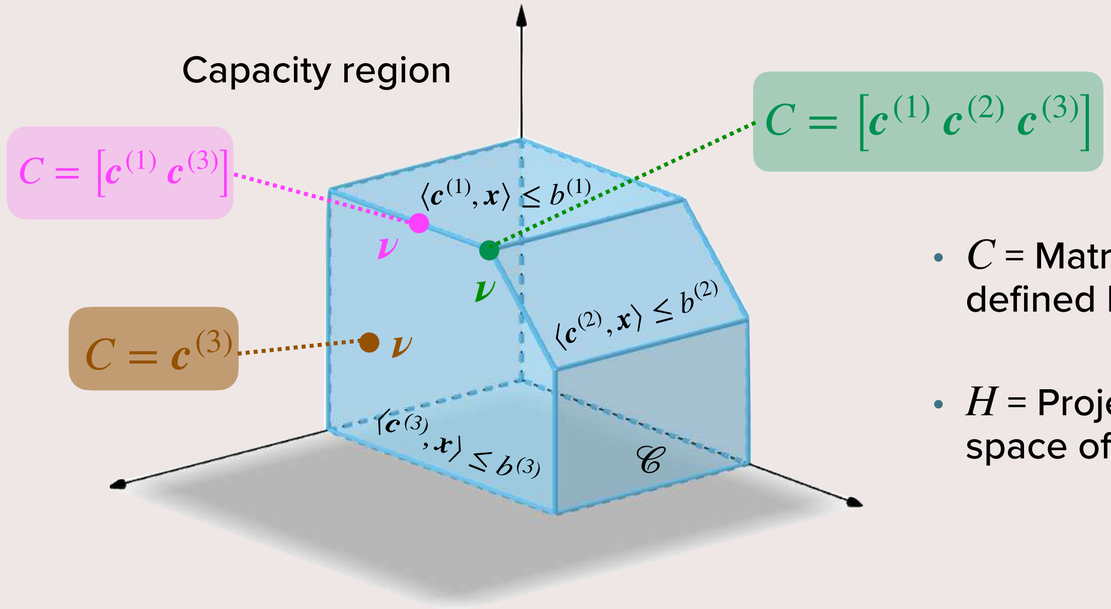
For the generalized switch, we have

$$\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle\right] = \frac{1}{2\epsilon}\left(\mathbf{1}^T(H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T((C^TC)^{-1} \circ \Sigma_{cs})\mathbf{1}\right) + o\left(\frac{1}{\epsilon}\right)$$



Channel state: $M(k) = m$

Feasible service rates: $\mathcal{S}^{(m)}$



- C = Matrix where columns are defined by facets intersecting at $\boldsymbol{\nu}$
- H = Projection matrix on column space of C

Matrices C and H characterize the subspace where SSC occurs

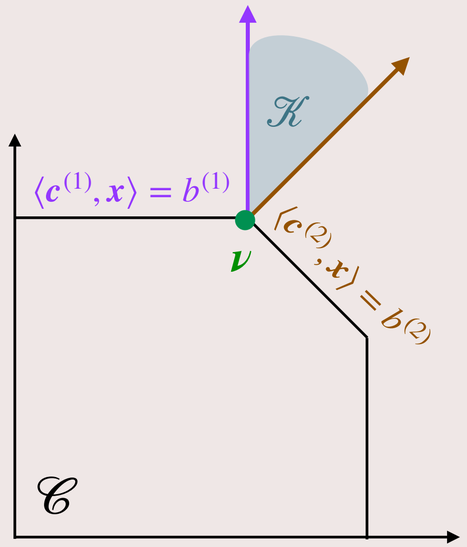
Proof Sketch: Drift Method

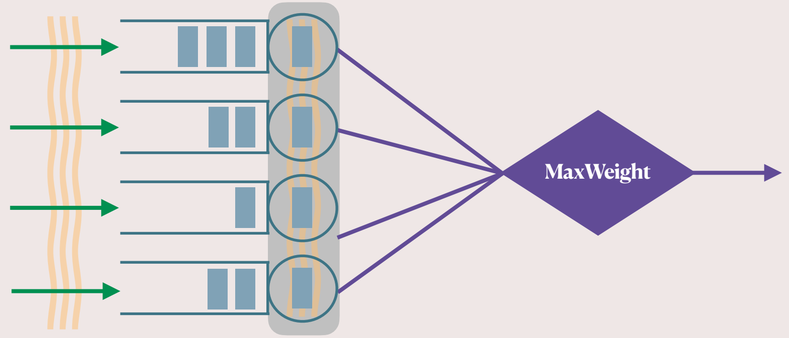
I) State Space Collapse (SSC)

- Prove that the vector of queue lengths collapses to a lower-dimensional subspace $\mathcal{K}$
- $\boldsymbol{q}_{\|}$ = Projection of $\boldsymbol{q}$ on $\mathcal{K}$

II) Set drift of test function to zero

- Test function: $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$
- Set its drift to zero: $\mathbb{E}\left[\Delta V(\boldsymbol{q})\right] = 0$





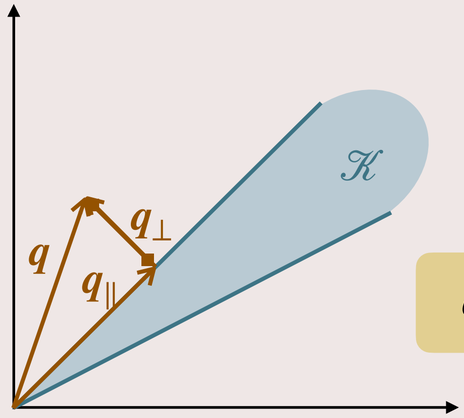
Channel state: $M(k) = m$
Feasible service rates: $\mathcal{S}^{(m)}$

Proposition (SSC): [HL, Maguluri '20]

$\boldsymbol{q}_{\|}(k)$ = Projection of $\boldsymbol{q}$ on $\mathcal{K}$

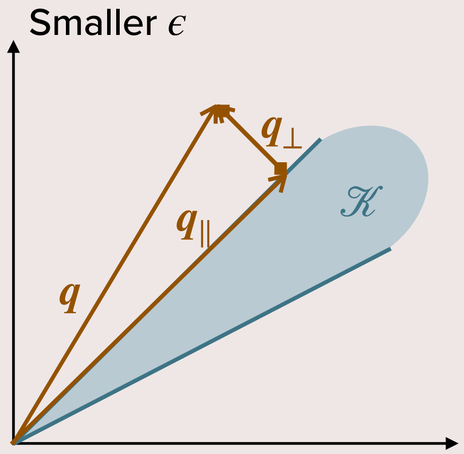
$\boldsymbol{q}_{\perp}(k) = \boldsymbol{q}(k) - \boldsymbol{q}_{\|}(k)$ error of approximating $\boldsymbol{q} \approx \boldsymbol{q}_{\|}$

Then, $\mathbb{E}\left[\|\boldsymbol{q}_{\perp}\|^t\right] \leq T_t$ for all $t = 1,2,3\ldots$



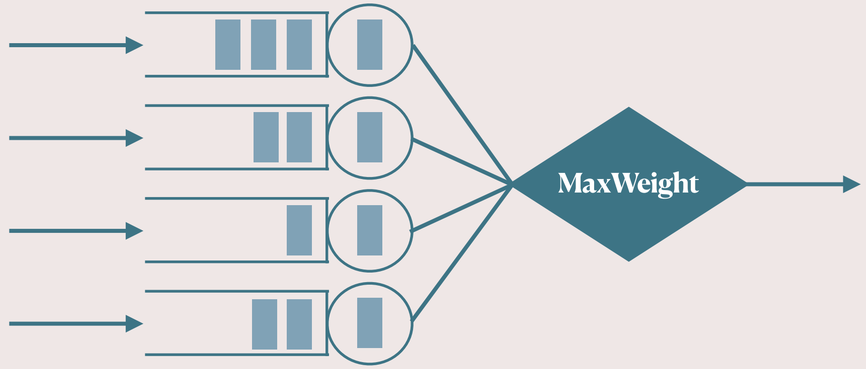
$\boldsymbol{q} \approx \boldsymbol{q}_{\|}$

Smaller ϵ

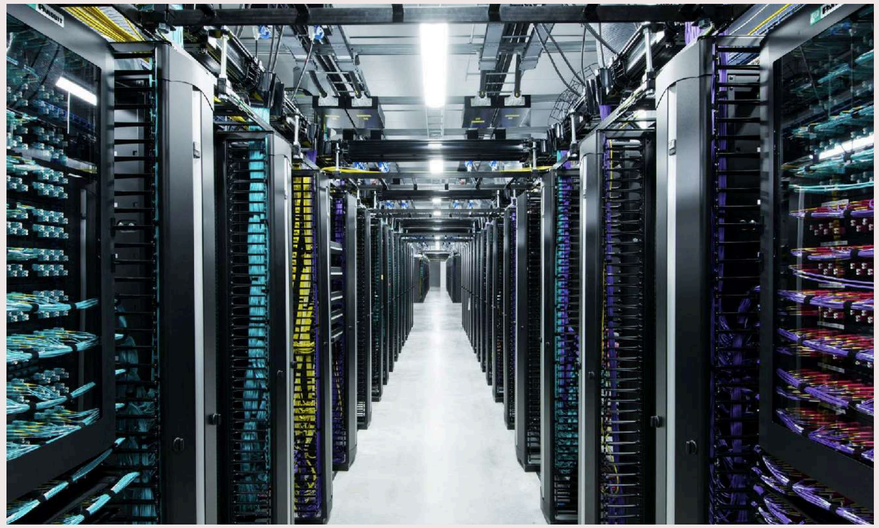


Key Takeaways

- **Generalized switch:** Model that subsumes several queueing systems
  - Input-queued switch with correlated arrivals
  - $\mathcal{N}$-system
  - Parallel-server systems



- When **CRP condition is not satisfied**
  - Open question for long time
  - Our result gives the **mean of linear combinations** of queue lengths
  - Our result is **immediately applicable** in several systems



Service Level Agreement (SLA):

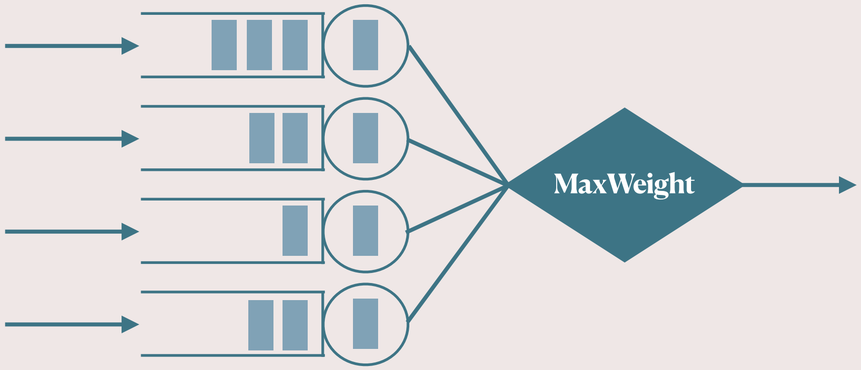
E.g. Delay greater than 1 s less than 5% of the time

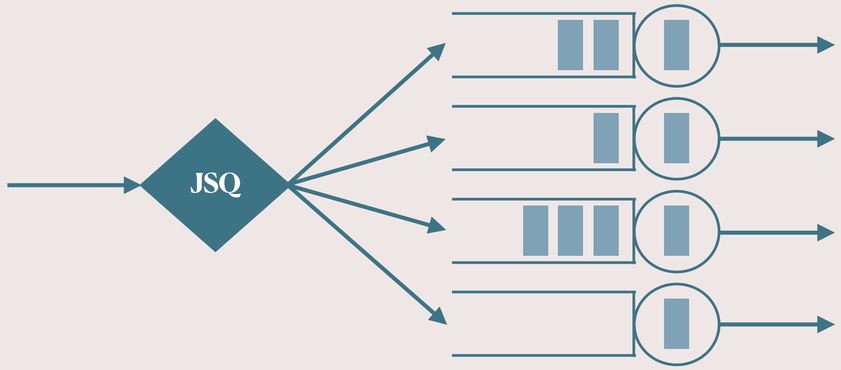
Tail behavior of queue lengths

Outline

Question 1: Expected Delay and Drift method	• Expected delay in data centers in heavy-traffic • General result • Proof sketch
Question 2: Tail bounds and Transform techniques	• The single server queue • Systems with a single bottleneck • The load balancing system
Overview of other work	• Rate of convergence to heavy traffic • Load balancing with heterogeneous servers • The many-server heavy-traffic regime • Healthcare systems
Conclusion and future work	

Drift Method and Transform Techniques





Mean behavior:

Theorem: [HL, Maguluri '21]

$$\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle\right]$$
$$= \frac{1}{2\epsilon}\left(\mathbf{1}^T(H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T((C^TC)^{-1} \circ \Sigma_{cs})\mathbf{1}\right) + o\left(\frac{1}{\epsilon}\right)$$

Theorem: [Eryilmaz and Srikant '13]

$$\mathbb{E}\left[\sum_{i=1}^{N} q_i\right] = \frac{1}{2\epsilon}\left(\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}\right) + o\left(\frac{1}{\epsilon}\right)$$

Tail behavior:

Theorem: [HL, Maguluri '20]

If the generalized switch satisfies SSC along the cone generated by $\boldsymbol{c}$, we have

$$\epsilon \boldsymbol{q} \Rightarrow \boldsymbol{c}Expo\left(\frac{2}{\boldsymbol{c}^T\Sigma_a\boldsymbol{c} + \sigma_{cs}^2}\right)$$

Theorem: [HL, Maguluri '20]

For the load balancing system operating under JSQ, we have

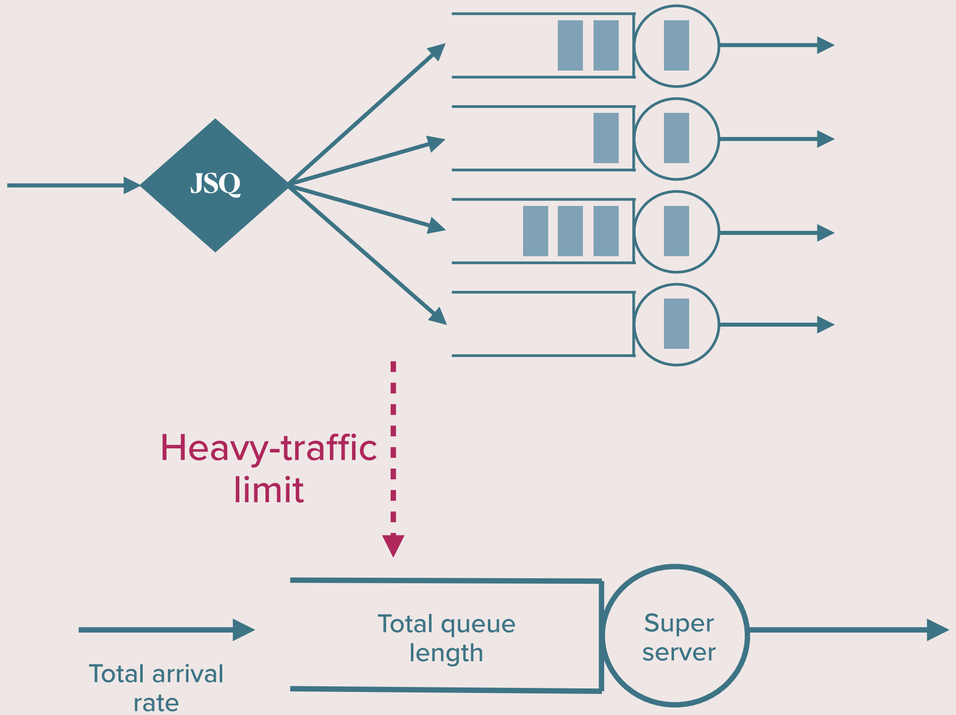
$$\epsilon \boldsymbol{q} \Rightarrow \mathbf{1}Expo\left(\frac{2N}{\sigma_a^2 + \mathbf{1}^T\Sigma_s\mathbf{1}}\right)$$

Load Balancing System in the Literature

- The load balancing system under JSQ satisfies the **CRP condition**
  - Diffusion limits: Foschini and Salz (1978)
  - Drift method: Eryilmaz and Srikant (2013)

- **Optimality** of JSQ:
  Winston (1977); Weber (1978); Ephremides, Varaiya and Walrand (1980)

- Improving **performance** and decreasing **complexity** of routing algorithm:
  Chen and Ye (2012); Li, Kong and Wang (2018); Braverman (2020); Zhou, Tan and Schroff (2018); Ying (2016); Ying (2017); Eschenfeldt and Gamarnik (2018); Lu, Xie, Kliot, Geller, Larus and Greenberg (2011); Stolyar (2017); Ying, Srikant and Kang (2017)...



Diffusion limits approach

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

$$\begin{array}{ccc} \epsilon q^{(\epsilon)}(t) & \xrightarrow{t \to \infty} & \epsilon q^{(\epsilon)}(\infty) \\ \epsilon \downarrow 0 \big\downarrow & & \big\downarrow \epsilon \downarrow 0 \\ \widetilde{q}(t) & \xrightarrow{t \to \infty} & q^* \end{array}$$

RBM

Interchange of limits

We propose a proof in 4 lines

The Single-Server Queue

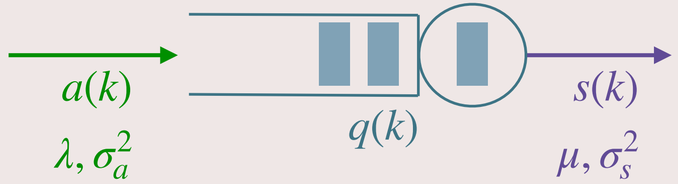
- Discrete-time model: k indexes time
- $q(k)$: # jobs in the system at the beginning of time slot k
- $a(k)$: # arrivals in time slot k
  - $\mathbb{E}[a(k)] = \lambda, \text{Var}[a(k)] = \sigma_a^2$
- $s(k)$: potential service in time slot k
  - $\mathbb{E}[s(k)] = \mu, \text{Var}[s(k)] = \sigma_s^2$
- Dynamics of the queues:

$$q(k+1) = \max\{q(k) + a(k) - s(k), 0\}$$
$$= q(k) + a(k) - s(k) + u(k)$$

Unused service

$$q(k+1)u(k) = 0$$

Key property



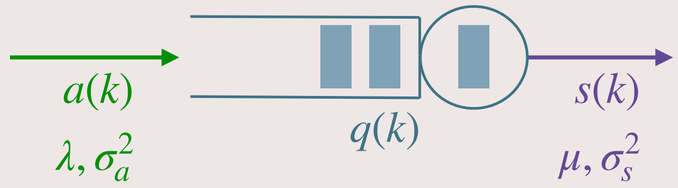
We wish to show:

$$\epsilon q \Rightarrow Expo\left(\frac{2}{\sigma_a^2 + \sigma_s^2}\right)$$

The MGF Method [HL, Maguluri '20]

- Key Lemma:

Step 1:
$$\left(e^{\theta\epsilon q(k+1)} - 1\right)\left(e^{-\theta\epsilon u(k)} - 1\right) = 0$$



Dynamics of the queues:

$$q(k+1) = q(k) + a(k) - s(k) + u(k)$$
$$\Longrightarrow q(k+1)u(k) = 0$$

- From the lemma:

$$\mathbb{E}\left[e^{\theta\epsilon q(k+1)}\right] = \mathbb{E}\left[e^{\theta\epsilon(q(k)+a(k)-s(k))} - e^{-\theta\epsilon u(k)} + 1\right]$$

$$\Longrightarrow \mathbb{E}\left[e^{\theta\epsilon q}\right] = \frac{1 - \mathbb{E}\left[e^{-\theta\epsilon u}\right]}{1 - \mathbb{E}\left[e^{\theta\epsilon(a-s)}\right]}$$

Step 2: Bound unused service and take the heavy-traffic limit

$$\Longrightarrow \lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\theta\epsilon q}\right] = \frac{1}{1 - \theta\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)}$$

MGF of expo. random variable with mean $\frac{\sigma_a^2 + \sigma_s^2}{2}$

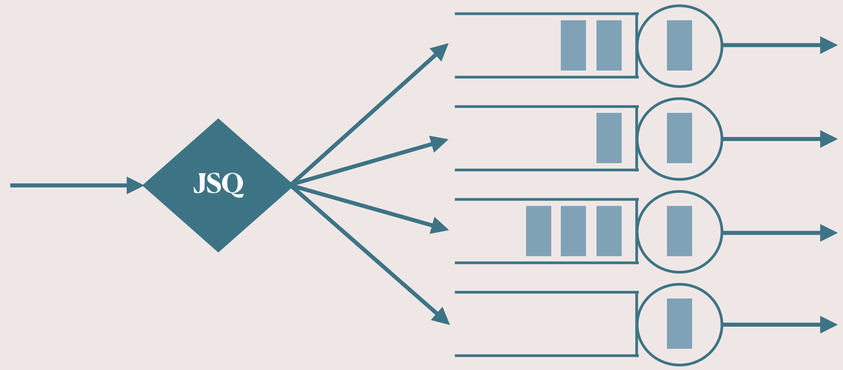
$$\epsilon q \Rightarrow Expo\left(\frac{2}{\sigma_a^2 + \sigma_s^2}\right)$$

Tail Behavior

Theorem: [HL, Maguluri '20]

For the load balancing system operating under JSQ, we have

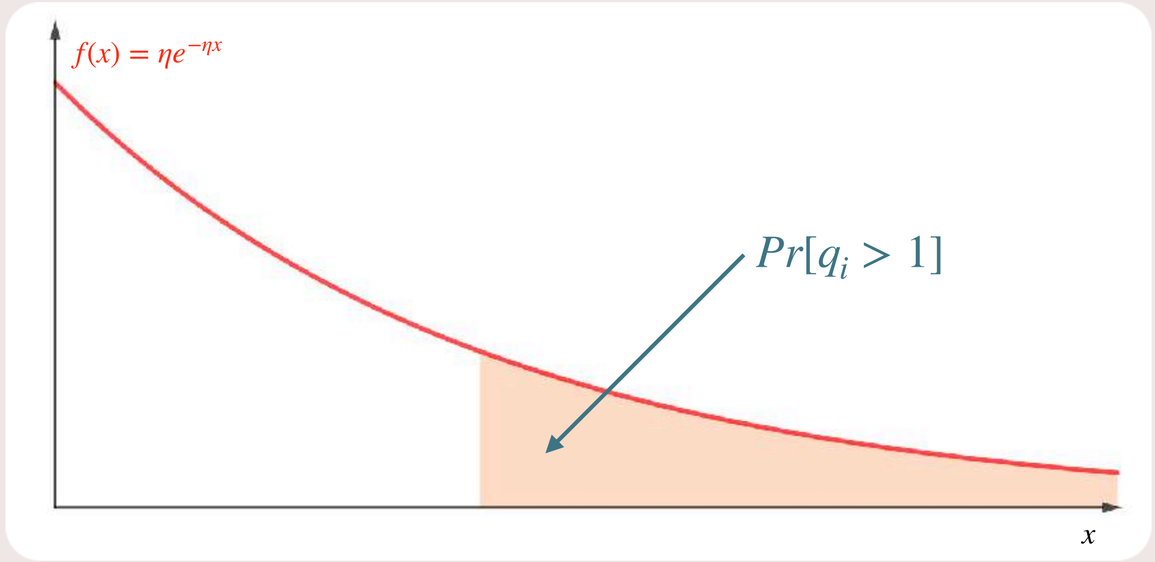
$$\epsilon \boldsymbol{q} \Rightarrow \mathbf{1} Expo\left(\underbrace{\frac{2N}{\sigma_a^2 + \mathbf{1}^T \Sigma_s \mathbf{1}}}_{\eta}\right)$$



What is the tail behavior?

If ϵ is small,

$$Pr\left[q_i > 1\right] = Pr\left[\epsilon q_i > \epsilon\right] \approx \exp\left(-\eta\epsilon\right)$$



Transform Techniques

Step 1: Prove an exponential version of $q(k+1)u(k) = 0$

Step 2: Bound unused service and take heavy-traffic limit

Moment Generating Function

$$\mathbb{E}\left[e^{\theta\epsilon q}\right]$$

- $\theta \in \mathbb{R}$
- Two-sided Laplace transform of stationary distribution
- Must exist for θ in an interval around the origin

  ➡Must prove

Characteristic Function

$$\mathbb{E}\left[e^{i\theta\epsilon q}\right]$$

- $\theta \in \mathbb{R}, i = \sqrt{-1}$
- Fourier transform of stationary distribution
- Must use complex numbers

One-Sided Laplace Transform

$$\mathbb{E}\left[e^{\theta\epsilon q}\right]$$

- $\theta < 0$
- Always exists because $q \geq 0$

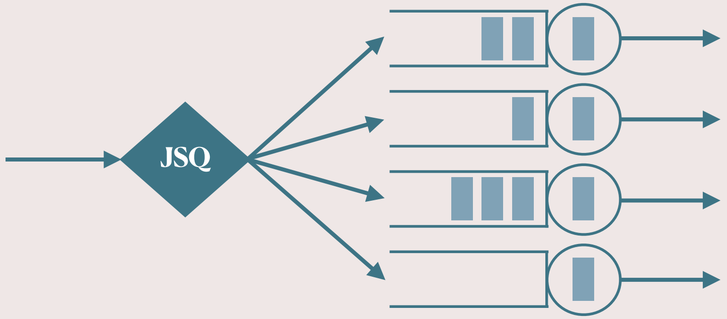
Key Takeaways

- Tail bounds on queue length behavior
- Transform techniques:
  - 2-step proof
  - Flexible
- More than just heavy-traffic results
  - Non-asymptotic results
  - Rate of convergence
- It's all about handling **unused service**!
- Useful in many contexts
  - Ride-sharing systems

Theorem: [HL, Maguluri '20]

For the single server queue,

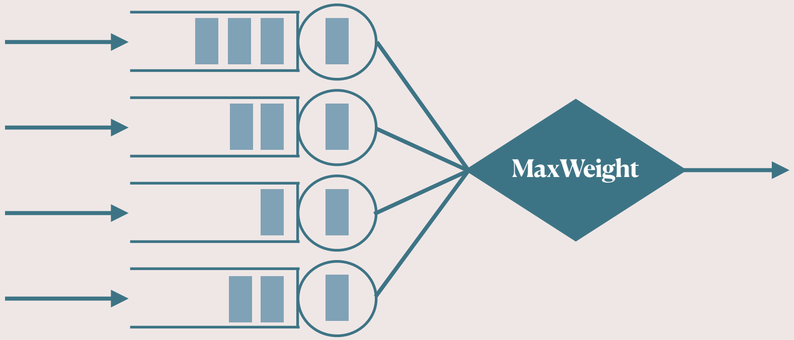
$$\epsilon q \Rightarrow Expo\left(\frac{2}{\sigma_a^2+\sigma_s^2}\right)$$



Theorem: [HL, Maguluri '20]

For the load balancing system operating under JSQ, we have

$$\epsilon \boldsymbol{q} \Rightarrow \mathbf{1}Expo\left(\frac{2N}{\sigma_a^2+\mathbf{1}^T\Sigma_s\mathbf{1}}\right)$$



Theorem: [HL, Maguluri '20]

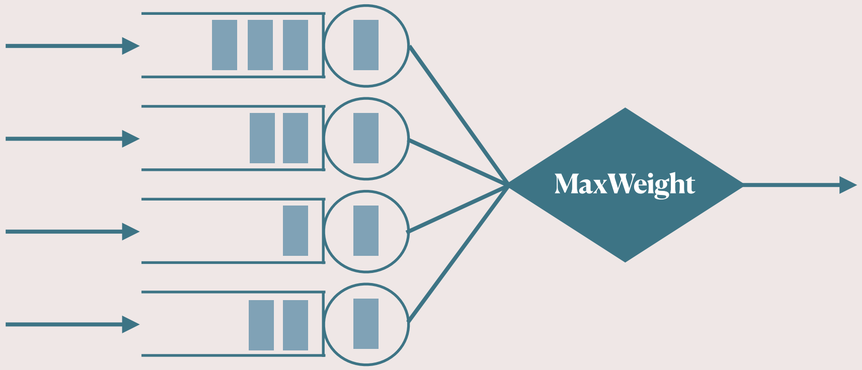
If the generalized switch satisfies SSC along the cone generated by $\boldsymbol{c}$, we have

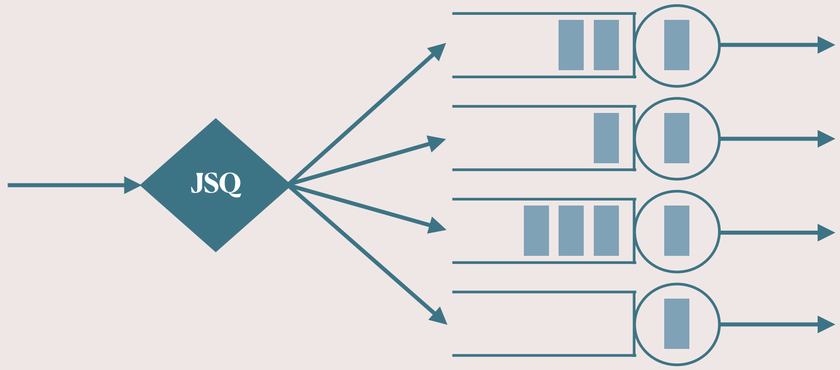
$$\epsilon \boldsymbol{q} \Rightarrow \boldsymbol{c}Expo\left(\frac{2}{\boldsymbol{c}^T\Sigma_a\boldsymbol{c}+\sigma_{cs}^2}\right)$$

Outline

Question 1: Expected Delay and Drift method	• Expected delay in data centers in heavy-traffic • General result • Proof sketch
Question 2: Tail bounds and Transform techniques	• The single server queue • Systems with a single bottleneck • The load balancing system
Overview of other work	• Rate of convergence to heavy traffic • Load balancing with heterogeneous servers • The many-server heavy-traffic regime • Healthcare systems
Conclusion and future work	

Rate of Convergence





Theorem: [HL, Maguluri '20]

$$\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle\right] = \frac{1}{2\epsilon}\left(\mathbf{1}^T(H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T((C^TC)^{-1} \circ \Sigma_{cs})\mathbf{1}\right) + o\left(\frac{1}{\epsilon}\right)$$

Theorem: [Eryilmaz and Srikant '13]

$$\mathbb{E}\left[\sum_{i=1}^{N} q_i\right] = \frac{1}{2\epsilon}\left(\sigma_a^2 + \sum_{i=1}^{N} \sigma_{si}^2\right) + o\left(\frac{1}{\epsilon}\right)$$

What is this term?

Theorem: [HL, Varma, Maguluri '21]

$$\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle\right] = \frac{1}{2\epsilon}\left(\mathbf{1}^T(H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T((C^TC)^{-1} \circ \Sigma_{cs})\mathbf{1}\right) + \log\left(\frac{1}{\epsilon}\right)$$

Theorem: [HL, Varma, Maguluri '21]

$$\mathbb{E}\left[\sum_{i=1}^{N} q_i\right] = \frac{1}{2\epsilon}\left(\sigma_a^2 + \sum_{i=1}^{N} \sigma_{si}^2\right) + \log\left(\frac{1}{\epsilon}\right)$$

Load Balancing Under Heterogeneous Servers

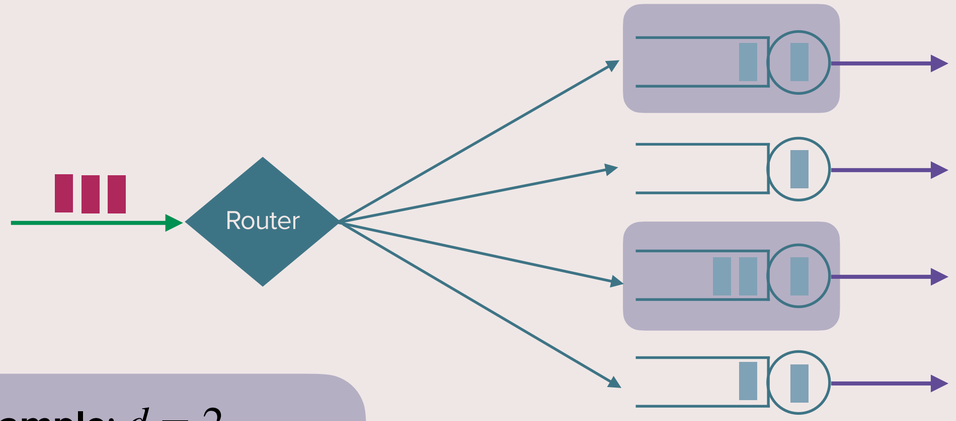
Is JSQ good in large systems?

No! Need to **find** the shortest queue

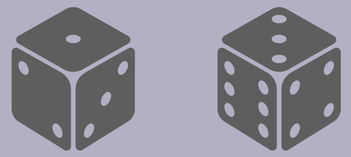
Power-of-d choices

- Select d queues **uniformly** at random
- Route to the shortest among them
- Well studied **if servers' rates are equal:** Vvedenskaya, Dobrushin and Karpelevich (1996); Mitzenmacher M (1996); Mitzenmacher M (2001); Akgun, Righter and Wolff (2011); Chen and Ye (2012); Maguluri, Srikant and Ying (2014) ...

But servers are not identical in data centers



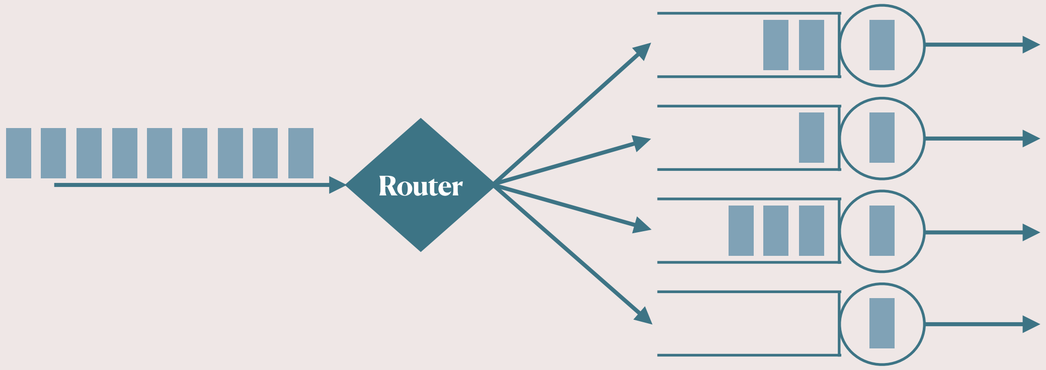
Example: $d = 2$



Result: [HL, Maguluri '21]

- Characterize **imbalance** among service rates
- Throughput and heavy-traffic optimality
- Use the MGF method

Unified Framework for Many-Server Heavy-Traffic



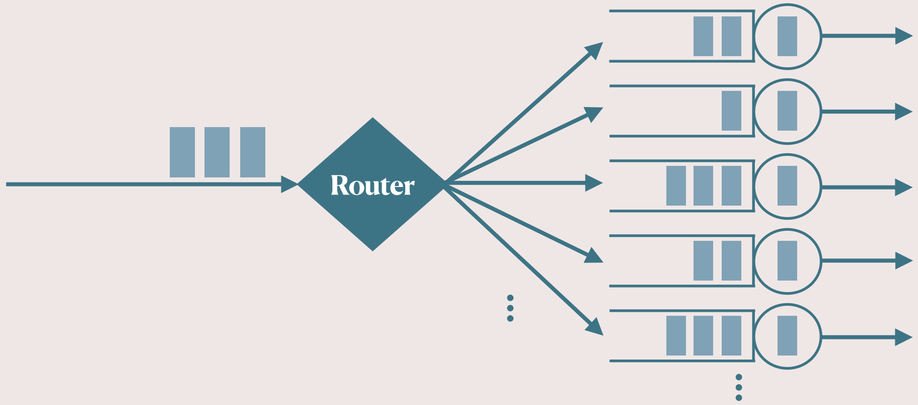
Heavy traffic

↑ load to max capacity
Fixed number of servers

Increase load and #servers together:
$\epsilon := N^{-\alpha}$, $\alpha > 0$

Many-server heavy-traffic regime

- In the literature: $\alpha \leq 1$
  Gurvich (2012, 2014); Eschenfeldt and Gamarnik (2018); Braverman (2018); Banerjee and Mukherjee (2019)...
- **Our result:** [HL and Maguluri '21]
  - $\alpha > 1$
  - **How fast** can the number of servers grow with respect to the load, to **observe heavy-traffic** behavior?
  - Use the MGF method

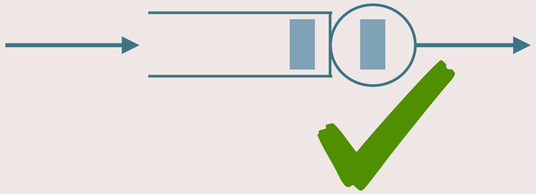


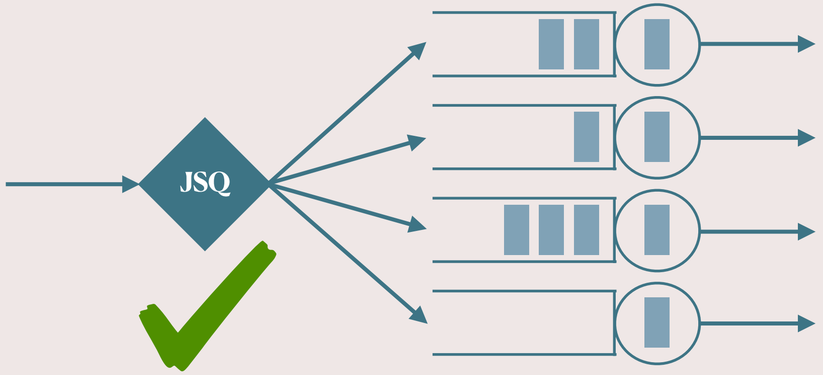
Mean field

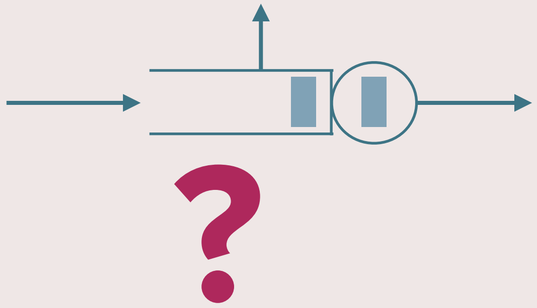
Fixed load
↑ number of servers

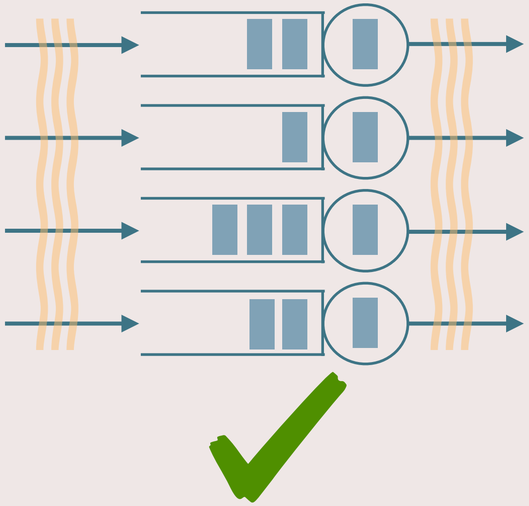
Future work:
Unified method for all $\alpha > 0$

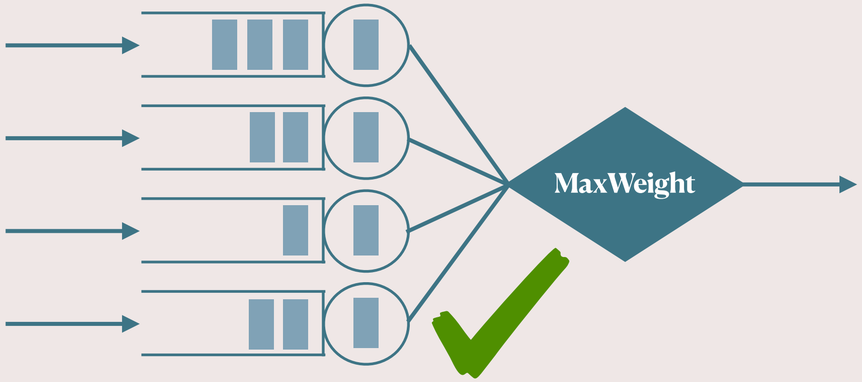
Summary

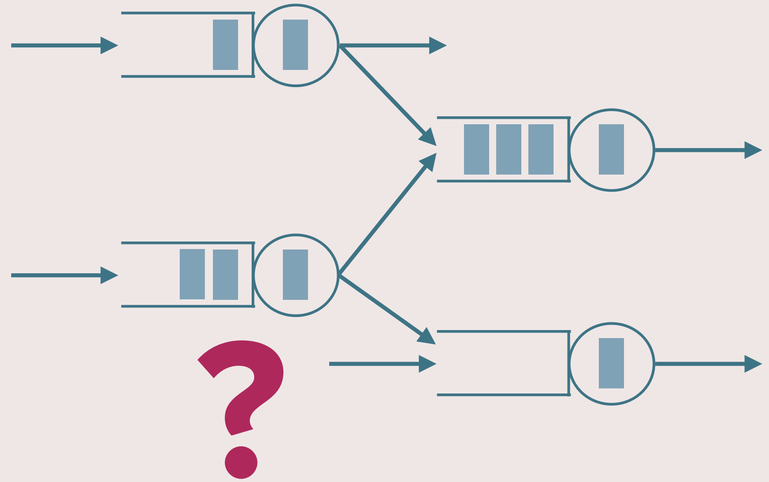




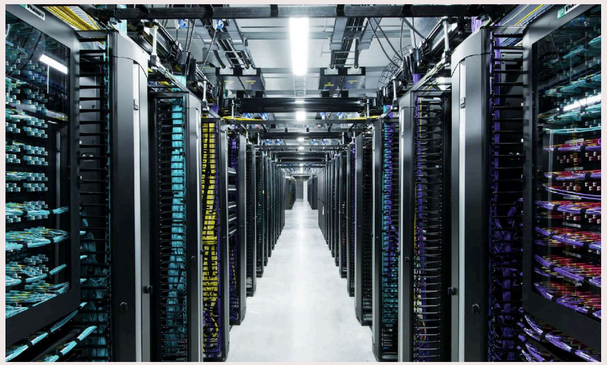








Future Work: Transform Techniques

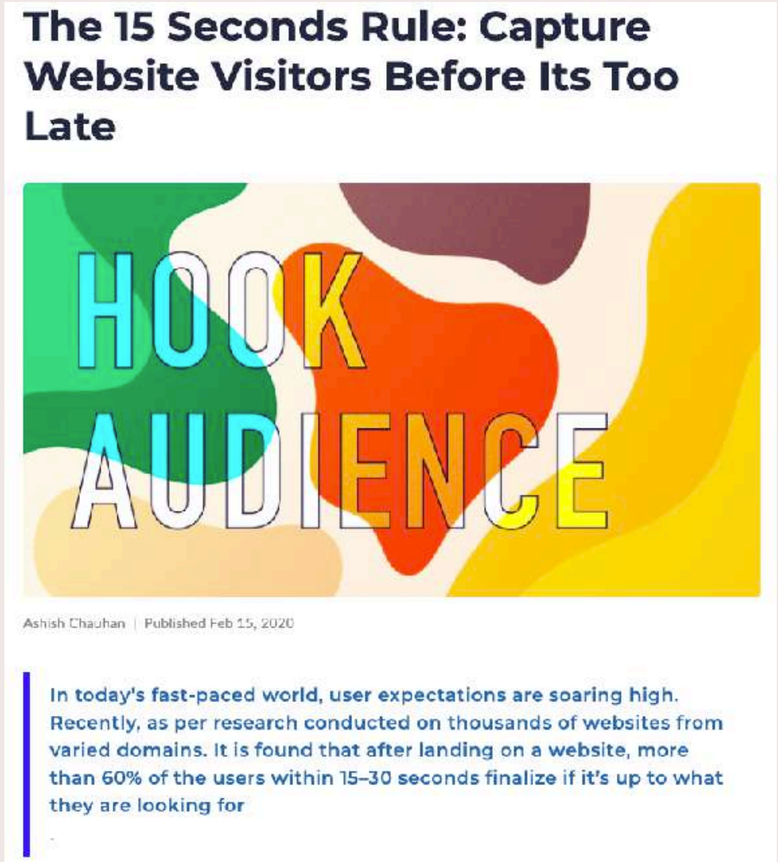


Systems that do not satisfy CRP

- Beyond the mean behavior

Other asymptotic regimes

- Unified method for many-server heavy-traffic



Source: https://www.makerobos.com/topic/ask-an-expert/15-seconds-rule-to-engage-website-visitors

Source: https://www.mediapost.com/publications/article/284398/many-visitors-abandon-mobile-sites-if-load-time-to.html

More practical constraints

- Abandonment
- Redundancy
- Multi-hop systems

Survey paper on transform techniques

Ongoing Work: Healthcare Systems

In collaboration with Pengyi Shi

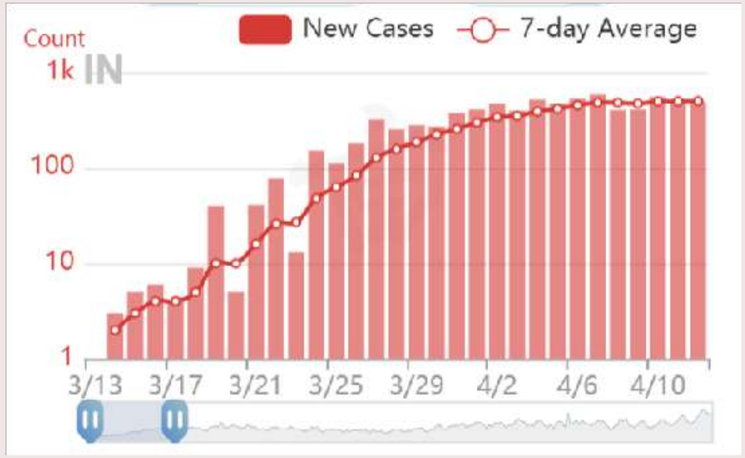
Goal: Estimate arrival rate to ICU using number in system

We know:

- Arrivals to ICU during pandemic follow a Cox process
- Cox process:
  - Doubly stochastic Poisson Process
  - Intensity rate is an SDE

$Cox/G/m$ queue

- Compute pmf of number-in-system
  - Maximum Likelihood Estimator
- We know: $Cox/M/\infty$
  - Good approximation in (very) light traffic



https://coronavirus.1point3acres.com/

Thanks!

Performance Analysis of Data Center Networks: Drift Method and Transform Techniques

Daniela Hurtado-Lange
William & Mary
Joint work with Siva Theja Maguluri

Kellogg School of Business, October 6th 2022

Contact information: dahurtadolange@wm.edu

Proof Sketch: Drift Method

Set drift of test function to zero

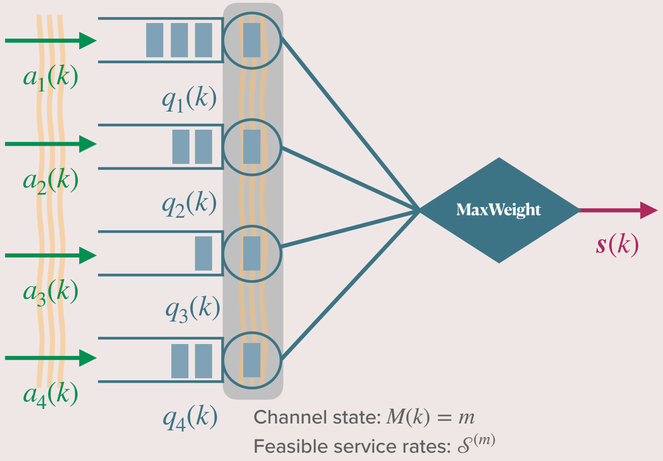
- Test function: $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$
- Set its drift to zero: $\mathbb{E}\left[\Delta V(\boldsymbol{q})\right] = 0$

We obtain: $\epsilon \downarrow 0$ → $2\epsilon\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu}\rangle\right]$

$$2\mathbb{E}\left[\langle \boldsymbol{q}_{\|}(k), \boldsymbol{s}_{\|}(k) - \boldsymbol{a}_{\|}(k)\rangle\right]$$

$$= \mathbb{E}\left[\|\boldsymbol{a}_{\|}(k) - \boldsymbol{s}_{\|}(k)\|^2\right] - \mathbb{E}\left[\|\boldsymbol{u}_{\|}(k)\|^2\right] + 2\mathbb{E}\left[\langle \boldsymbol{q}_{\|}(k+1), \boldsymbol{u}_{\|}(k)\rangle\right]$$

- $\epsilon \downarrow 0$: Least squares problem → RHS
- $\epsilon \downarrow 0$ → 0
- $\epsilon \downarrow 0$: Using SSC → 0



Theorem: [HL, Maguluri '20]

For the generalized switch, we have

$$\mathbb{E}\left[\langle \boldsymbol{q}, \boldsymbol{\nu}\rangle\right] = \frac{1}{2\epsilon}\left(\mathbf{1}^T(H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T((C^TC)^{-1} \circ \Sigma_{cs})\mathbf{1}\right) + o\left(\frac{1}{\epsilon}\right)$$

Dynamics of the queues:

$$q_i(k+1) = q_i(k) + a_i(k) - s_i(k) + u_i(k)$$

$$\implies q_i(k+1)u_i(k) = 0$$

Single-Server Queue and Drift Method

- In steady-state ($k \to \infty$):

$$\mathbb{E}\left[q^2(k+1)\right] = \mathbb{E}\left[q^2(k)\right]$$

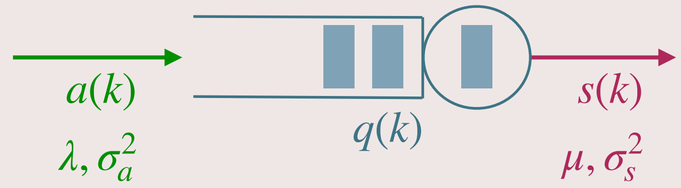
- Yields:

$$\mathbb{E}[q] = \frac{\mathbb{E}\left[(a-s)^2\right]}{2\epsilon} - \frac{\mathbb{E}\left[u^2\right]}{2\epsilon}$$

Small compared to the first term

- Therefore,

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\left[\epsilon q\right] = \frac{\sigma_a^2 + \sigma_s^2}{2}$$



Dynamics of the queues:

$$q(k+1) = q(k) + a(k) - s(k) + u(k)$$
$$\implies q(k+1)u(k) = 0$$

Use $e^{\theta \epsilon q}$ as test function

Drift Method:

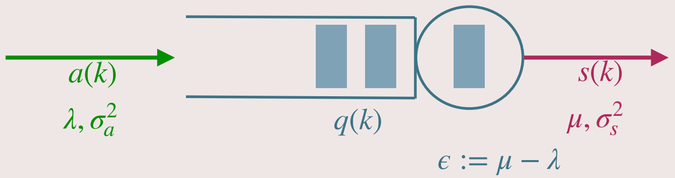
Test function: $1 + \epsilon q + \frac{1}{2}\epsilon^2 q^2 + \frac{1}{3!}\epsilon^3 q^3 + \cdots + \frac{1}{m!}\epsilon^m q^m + \cdots = e^{\epsilon q}$

Obtain: $1 + \mathbb{E}\left[\epsilon q\right] + \frac{1}{2}\mathbb{E}\left[\epsilon^2 q^2\right] + \cdots + \frac{1}{(m-1)!}\mathbb{E}\left[\epsilon^{m-1} q^{m-1}\right] + \cdots = \mathbb{E}\left[e^{\epsilon q}\right]$

$$\lim_{\epsilon \downarrow 0}(\quad) = \frac{\sigma_a^2 + \sigma_s^2}{2} \quad 2\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^2 \quad (m-1)!\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)^{m-1}$$

$$\epsilon q \Rightarrow Expo\left(\frac{\sigma_a^2 + \sigma_s^2}{2}\right)$$

The MGF Method [HL, Maguluri '20]

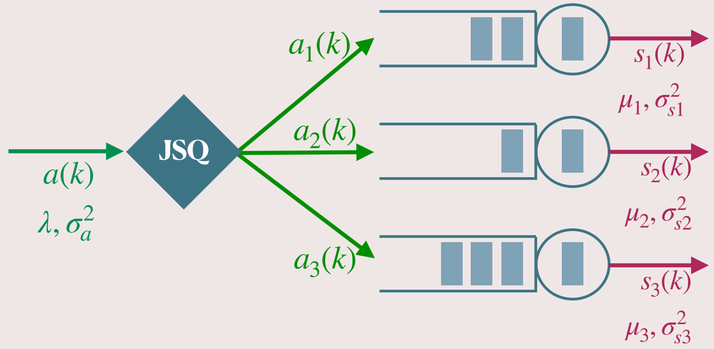


Set the drift of $V(q) = e^{\theta \epsilon q}$ to zero

Key Lemma: $\left(e^{\theta \epsilon q(k+1)} - 1\right)\left(e^{-\theta \epsilon u(k)} - 1\right) = 0$

Yields: $\lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\theta \epsilon q}\right] = \dfrac{1}{1 - \theta\left(\dfrac{\sigma_a^2 + \sigma_s^2}{2}\right)}$

$\epsilon q \Rightarrow Expo\left(\dfrac{\sigma_a^2 + \sigma_s^2}{2}\right)$



$\epsilon := \sum_i \mu_i - \lambda$

$\boldsymbol{q}_{\|} := \mathbf{1}\left(\dfrac{\sum_j q_j}{N}\right)$

& $\boldsymbol{q}_{\perp} := \boldsymbol{q} - \boldsymbol{q}_{\|}$

Set the drift of $V(\boldsymbol{q}) = e^{\theta \epsilon \sum_i q_i}$ to zero

Key Lemma: $\mathbb{E}\left[\left(e^{\theta \epsilon \sum_i q_i(k+1)} - 1\right)\left(e^{-\theta \epsilon \sum_i u_i(k)} - 1\right)\right]$ is $o(\epsilon^2)$

Yields: $\lim_{\epsilon \downarrow 0} \mathbb{E}\left[e^{\theta \epsilon \sum_i q_i}\right] = \dfrac{1}{1 - \theta\left(\dfrac{\sigma_a^2 + \sum_i \sigma_{si}^2}{2}\right)}$

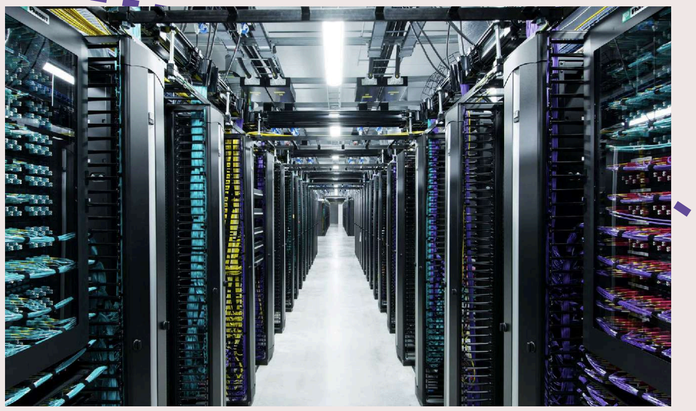
Then, $\epsilon \boldsymbol{q}_{\|} \Rightarrow \dfrac{\mathbf{1}}{N} Expo\left(\dfrac{\sigma_a^2 + \sum_i \sigma_{si}^2}{2}\right)$

SSC implies $\epsilon \boldsymbol{q}_{\perp} \Rightarrow \mathbf{0}$

$\epsilon \boldsymbol{q} \Rightarrow \dfrac{\mathbf{1}}{N} Expo\left(\dfrac{\sigma_a^2 + \sigma_s^2}{2}\right)$

Research Summary

Applied Probability



Performance Analysis

Resource Allocation

Queue Length Distribution

Scheduling Algorithms

Minimize delay

Routing Algorithms

Tail Guarantees

Practical Constraints

Optimization