HEAVY-TRAFFIC ANALYSIS OF THE GENERALIZED SWITCH UNDER MULTIDIMENSIONAL STATE SPACE COLLAPSE

Daniela Hurtado Lange, Siva Theja Maguluri

Industrial and Systems Engineering Department

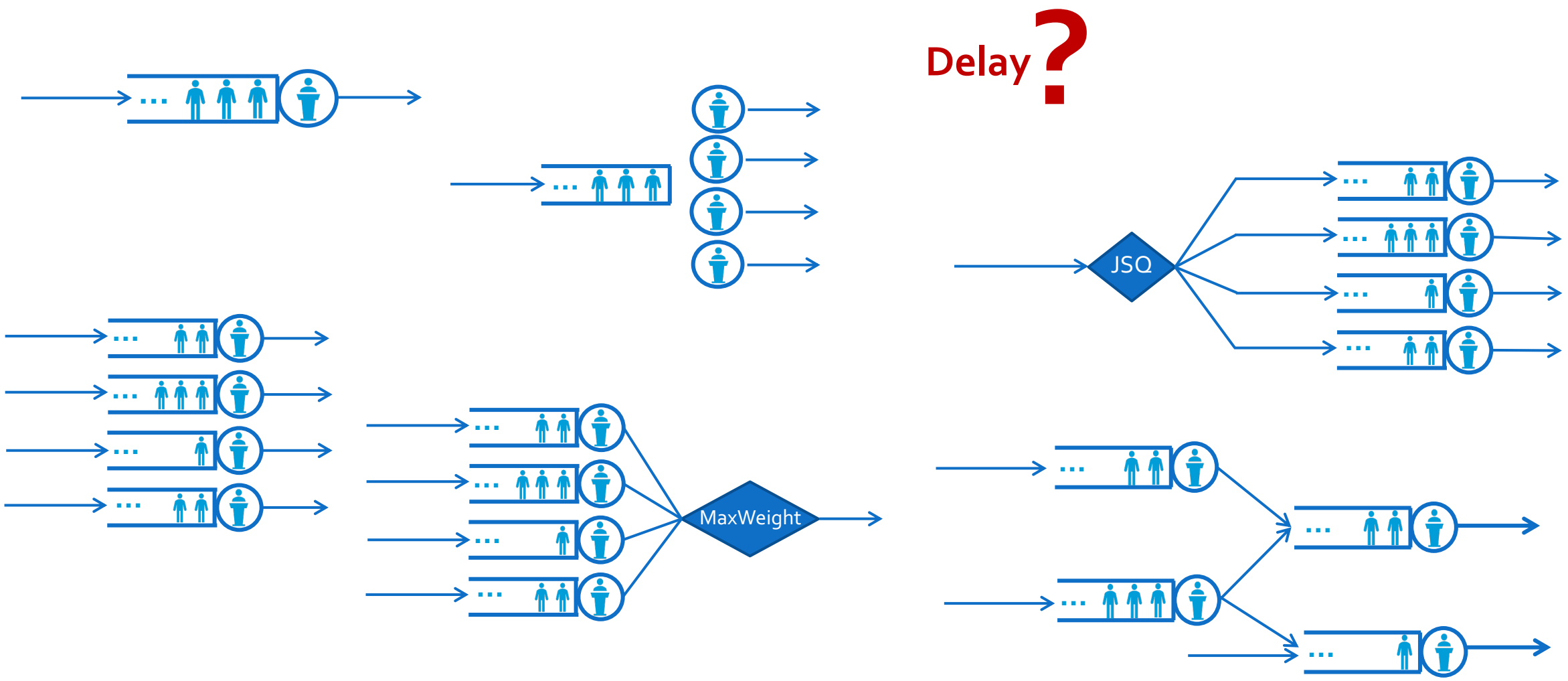
Georgia Institute of Technology

SIGMETRICS 2020

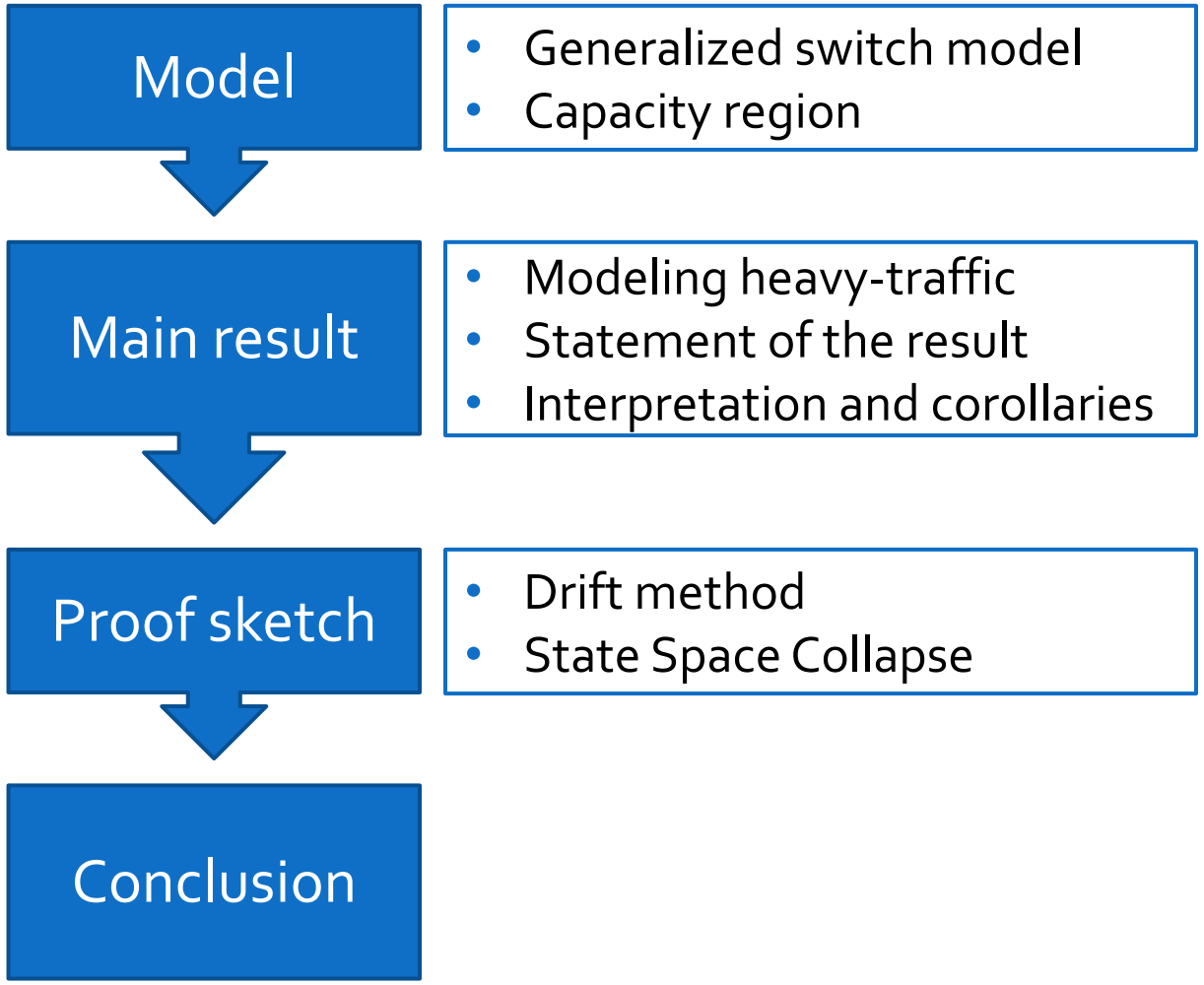
MOTIVATION

Minimize Delay?

STOCHASTIC PROCESSING NETWORKS



OUTLINE

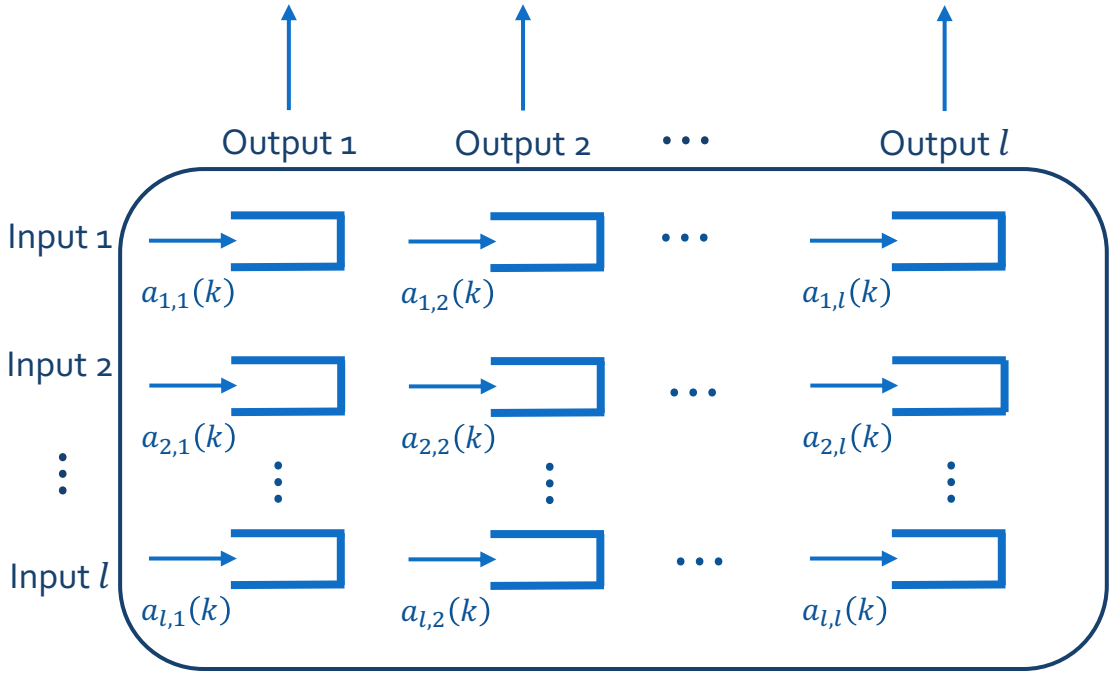

Model
Generalized switch model
Capacity region
Main result
Modeling heavy-traffic
Statement of the result
Interpretation and corollaries
Proof sketch
Drift method
State Space Collapse
Conclusion

INPUT QUEUED SWITCH MODEL

- Discrete time queueing system with $N = l^2$ queues
  - l input ports
  - l output ports
- One queue for each input/output pair

- $a_{i,j}(k)$ = # arrivals to queue (i,j) in time slot k
  - Mean $\lambda_{i,j}$ and variance $\sigma_{i,j}^2$

- $s_{i,j}(k)$ = **offered** service to queue (i,j) in time slot k
  - All packets take one time slot to be served
  - **Constraint**: At most one packet can be processed from each input port (row) and at most one in each output port (column)



Scheduling problem must be solved to compute $\boldsymbol{s}(k)$ for each k

INPUT QUEUED SWITCH MODEL (CONT.)

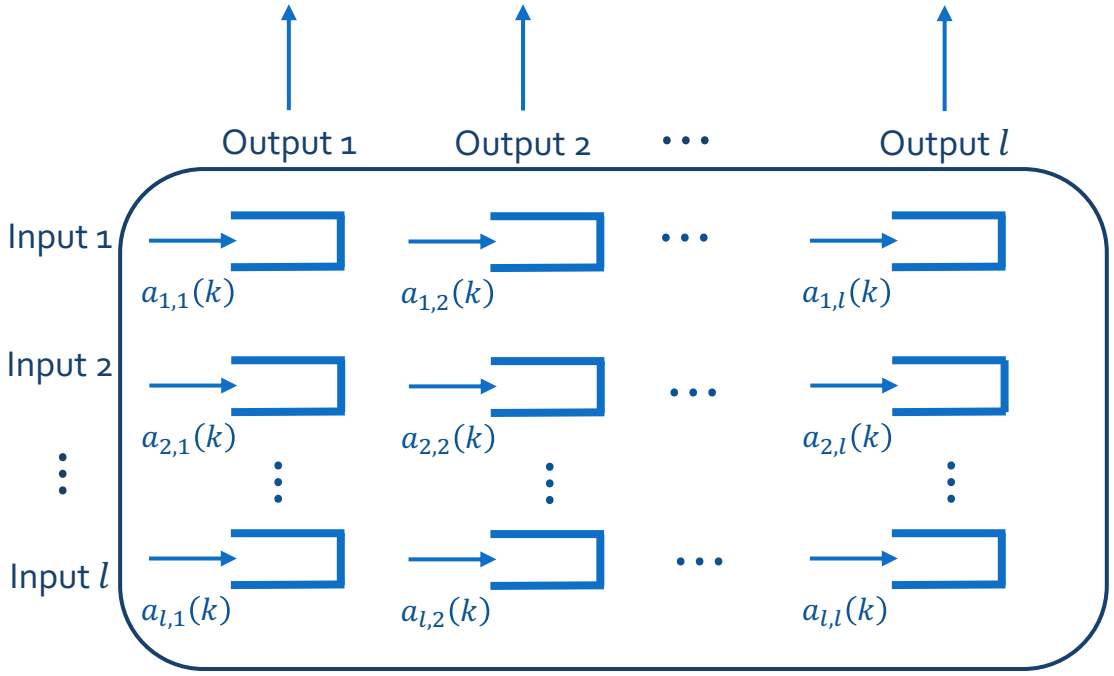
- Capacity region:

$$\mathcal{C} = \left\{ \boldsymbol{\lambda} \in \mathbb{R}_{+}^{lxl} : \sum_{i} \lambda_{i,j} \leq 1 \ \ \forall j \ \ \& \sum_{j} \lambda_{i,j} \leq 1 \ \ \forall i \right\}$$

- Constraint ⇒ $\boldsymbol{s}(k)$ is a permutation matrix

- S = set of lxl permutation matrices

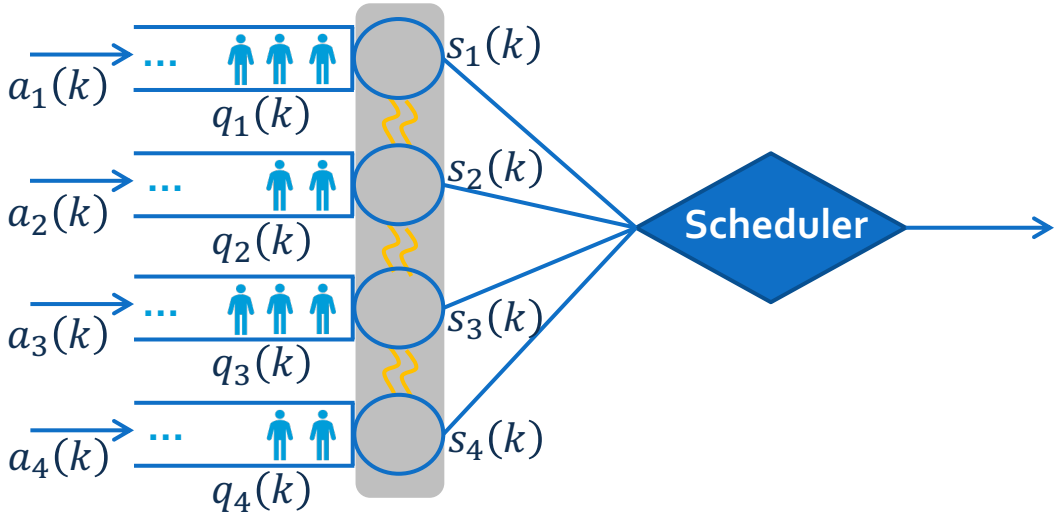
$$\Rightarrow \mathcal{C} = ConvexHull(S)$$



Constraint: At most one packet can be processed from each row and at most one in each column

GENERALIZED SWITCH MODEL

- Discrete time model with N queues
- Each queue has its own arrival process
  - i.i.d. across k
- $s_i(k)$: **offered** service from queue i in time slot k



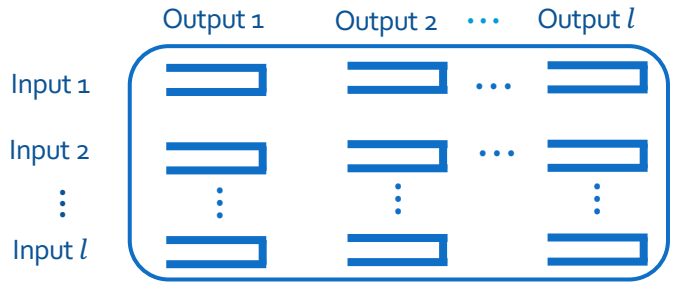
Channel state $M(k) = m$
Feasible service rates: $S^{(m)}$

	Input queued switch	Generalized switch
Constraints on $\boldsymbol{s}(k)$	At most one service in each row and each column	There are **some interference constraints** among servers
Environment	Fixed, does not affect constraint	**Channel state**: Interference constraints **change** with channel state

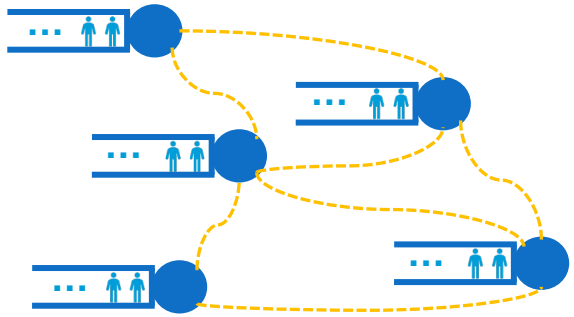
i.i.d. across time

EXAMPLES

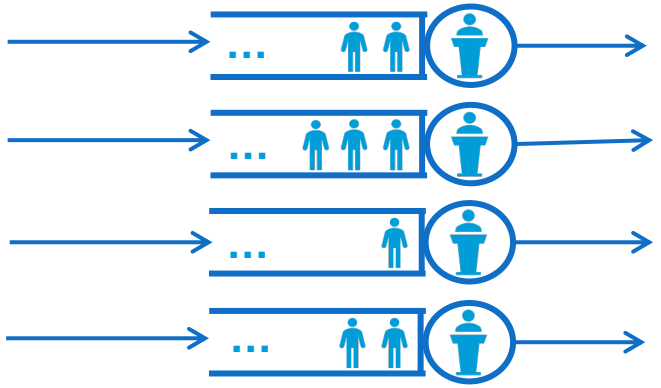
- Fixed channel state
  - Input queued switch



  - Ad hoc wireless network



- Channel state changing with time
  - Ad hoc wireless networks in presence of fading
  - Parallel server system



CAPACITY REGION

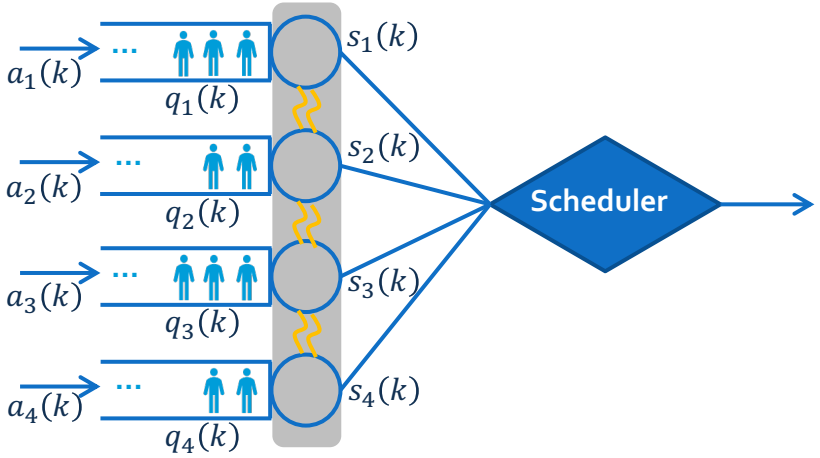
- Channel state pmf:

$$\psi_m = P\{M(k) = m\}$$

- Capacity region:

$$\mathcal{C} = \sum_m \psi_m \, ConvexHull\left(S^{(m)}\right)$$

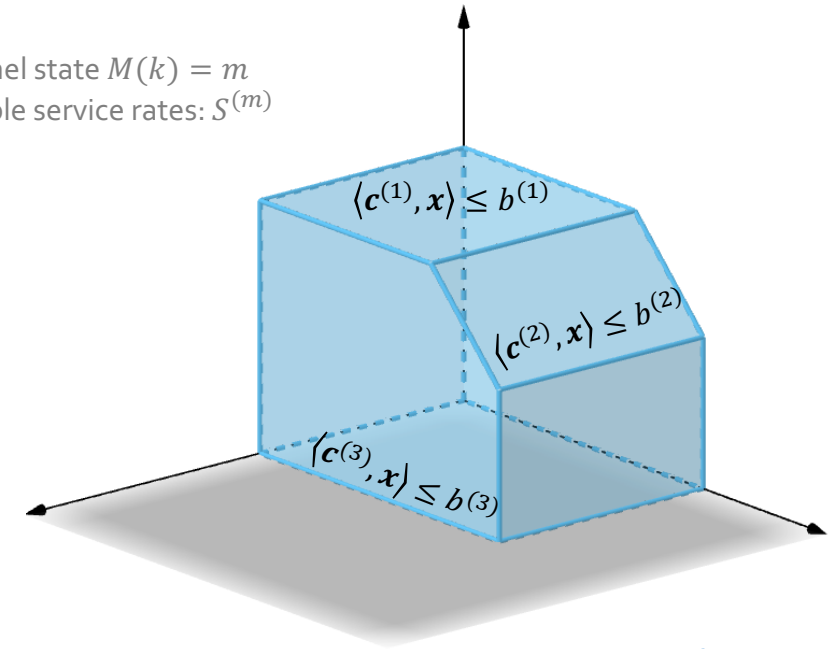
- Assume
  - Channel state has finite state space
  - Each $S^{(m)}$ is finite



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

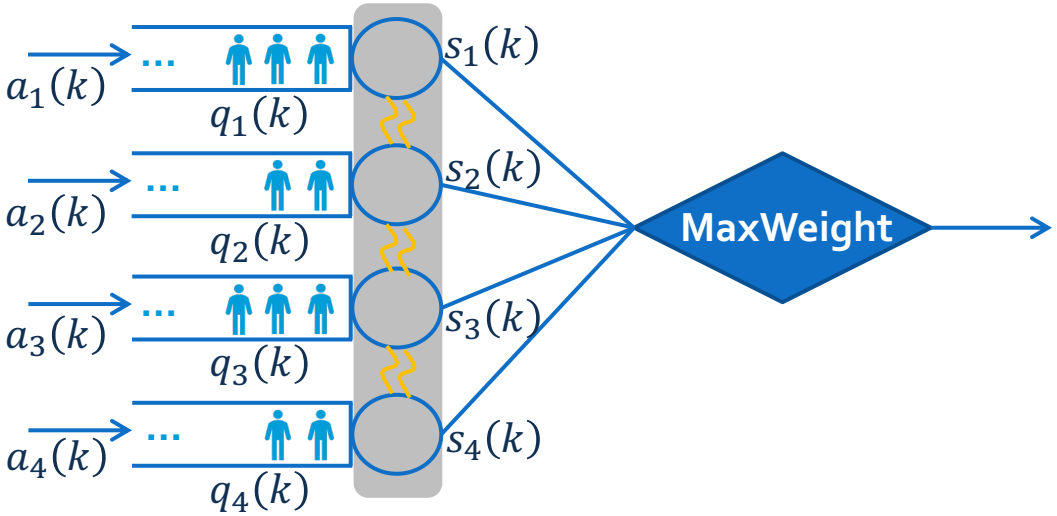
$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}_+^N : \left\langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \right\rangle \leq b^{(\ell)}, \ell = 1, \ldots, L\right\}$$

Polytope



GENERALIZED SWITCH MODEL (cont.)

- Scheduling: MaxWeight
  - Maximize weighted sum of queue lengths
  - Weight = (potential) service rates



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

- Formally, if $M(k) = m$ we choose

$$\boldsymbol{s}(k) \in \arg\max_{\boldsymbol{x} \in S^{(m)}} \langle \boldsymbol{x}, \boldsymbol{q}(k) \rangle$$

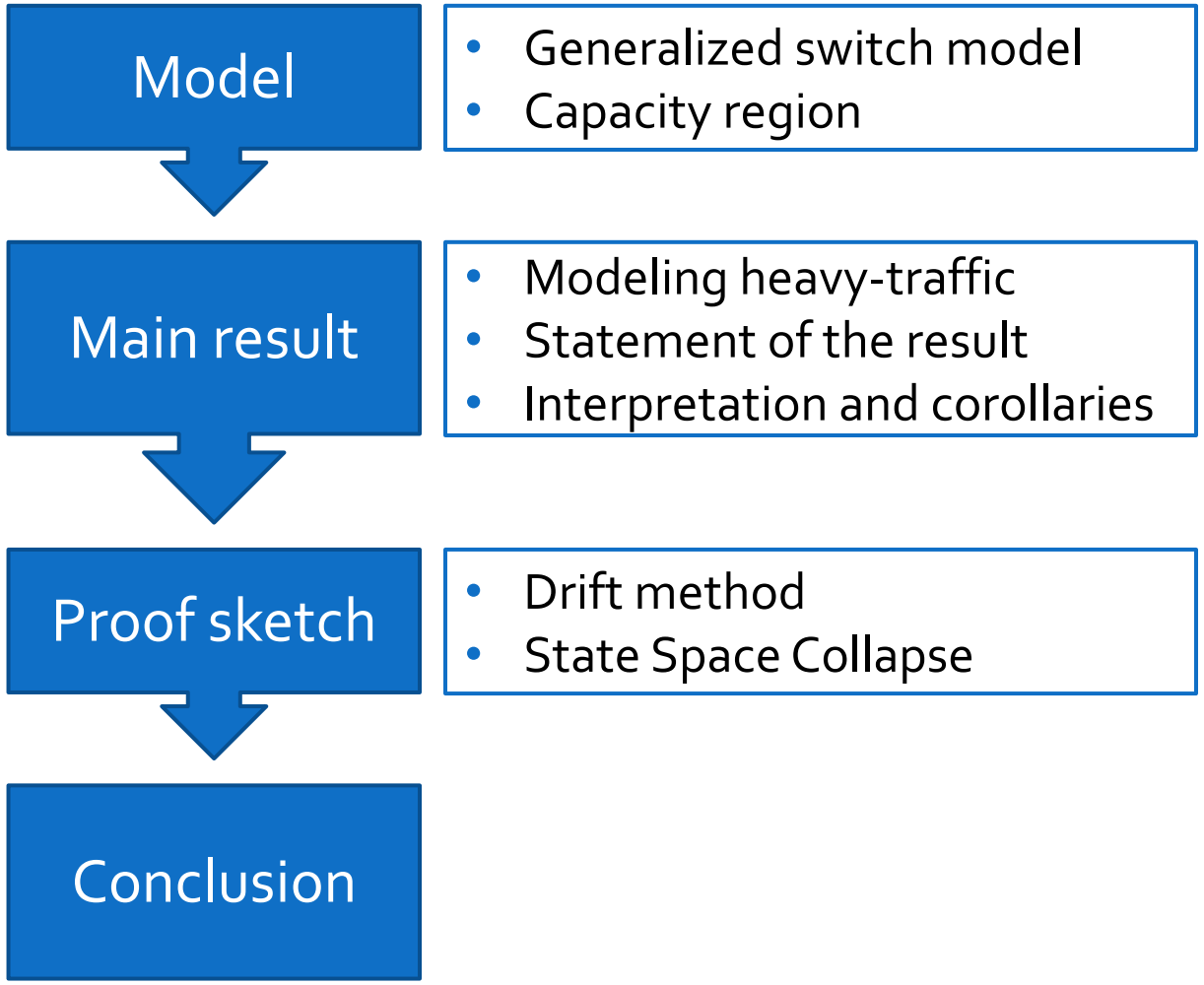
- Dynamics of the queues:

$$\begin{aligned} q_i(k+1) &= [q_i(k) + a_i(k) - s_i(k)]^+ \\ &= q_i(k) + a_i(k) - s_i(k) + u_i(k) \end{aligned}$$

$u_i(k)$: Unused service

$$q_i(k+1)u_i(k) = 0 \text{ but } q_i(k+1)u_j(k) \neq 0 \text{ if } i \neq j$$

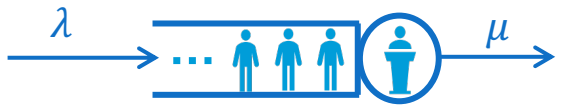
OUTLINE


Model
Generalized switch model
Capacity region
Main result
Modeling heavy-traffic
Statement of the result
Interpretation and corollaries
Proof sketch
Drift method
State Space Collapse
Conclusion

HEAVY-TRAFFIC ANALYSIS

- Load queueing system close to maximum capacity

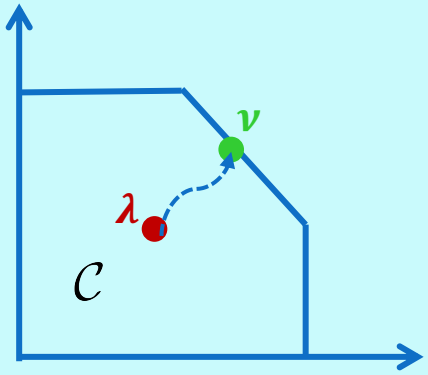
- Capacity region = All arrival rate vectors such that the system can be positive recurrent.



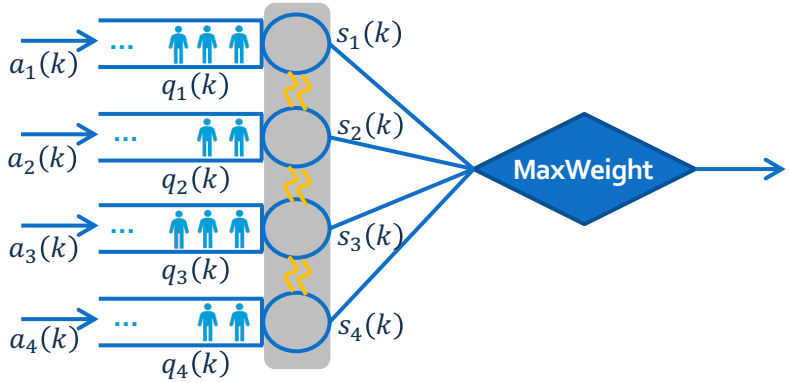
$$\mathcal{C} = \{\lambda : \lambda < \mu\}$$

Heavy-traffic limit:
$\lambda \to \mu$

In general…
$\mathcal{C}$ =Capacity region
$\boldsymbol{\nu} \in \partial\mathcal{C}$
$\boldsymbol{\lambda} = (1-\epsilon)\boldsymbol{\nu}$, with $\epsilon \in (0,1)$
Take limit as $\epsilon \downarrow 0$



MAIN RESULT



Linear combination of the queue lengths

Theorem: [HL, Maguluri '19]

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle] = \frac{1}{2}\left(\mathbf{1}^T (H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T\left((C^T C)^{-1} \circ \Sigma_{cs}\right)\mathbf{1}\right)$$

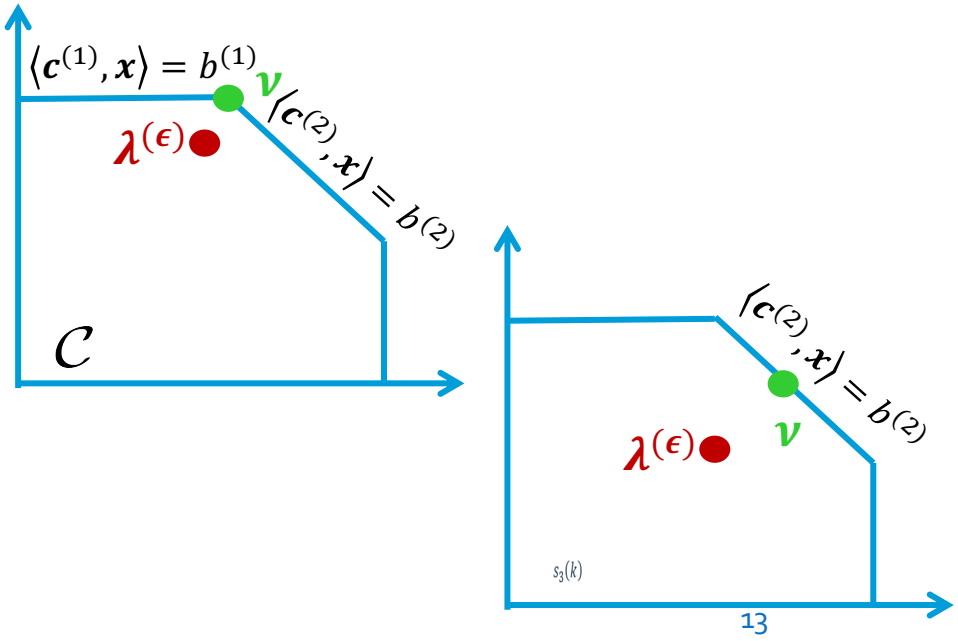
$$\mathcal{C} = \left\{\boldsymbol{x} \in \mathbb{R}^n : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \leq b^{(\ell)}, \ell = 1, \ldots, L\right\}$$

Linear combination of covariance matrix of arrivals

Linear combination of variability of service process (channel state)

Projection matrix on subspace where SSC occurs

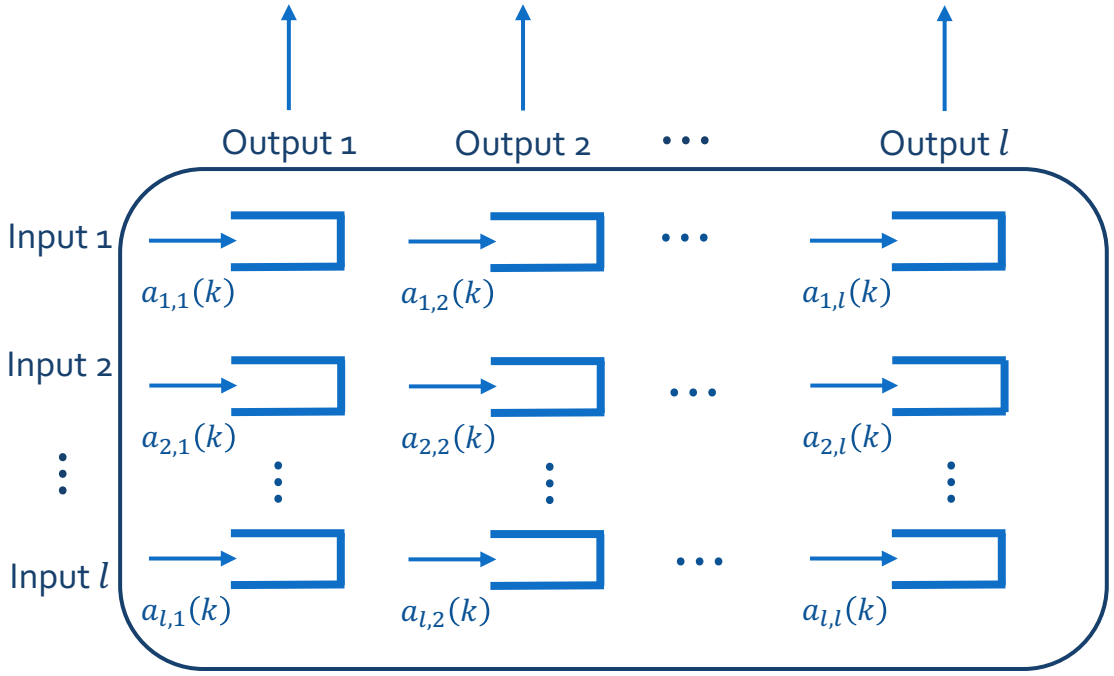
Matrix where columns are defined by facets that intersect at $\boldsymbol{\nu}$



INTERPRETATION AND EXAMPLES

INPUT QUEUED SWITCH

$$\lim_{\epsilon \downarrow 0} E\left[\epsilon \sum_{i,j} q_{ij}\right]$$
$$= \sum_{i,j}\left[\left(1-\frac{1}{2l}\right)\sigma_{ij}^2 + \left(\frac{l-1}{2l}\right)\left(\sum_{i' \neq i} Cov(a_{ij}, a_{i'j}) + \sum_{j' \neq j} Cov(a_{ij}, a_{ij'})\right)\right.$$
$$\left. - \sum_{i' \neq i, j' \neq j} Cov(a_{ij}, a_{i'j'})\right]$$



Maguluri, S. T., & Srikant, R. (2016):

- Heavy-traffic limit of total queue length under **independent arrivals**:

$$\lim_{\epsilon \downarrow 0} E\left[\epsilon \sum_{i,j} q_{ij}\right] = \left(1-\frac{1}{2l}\right)\sum_{ij} \sigma_{ij}^2$$

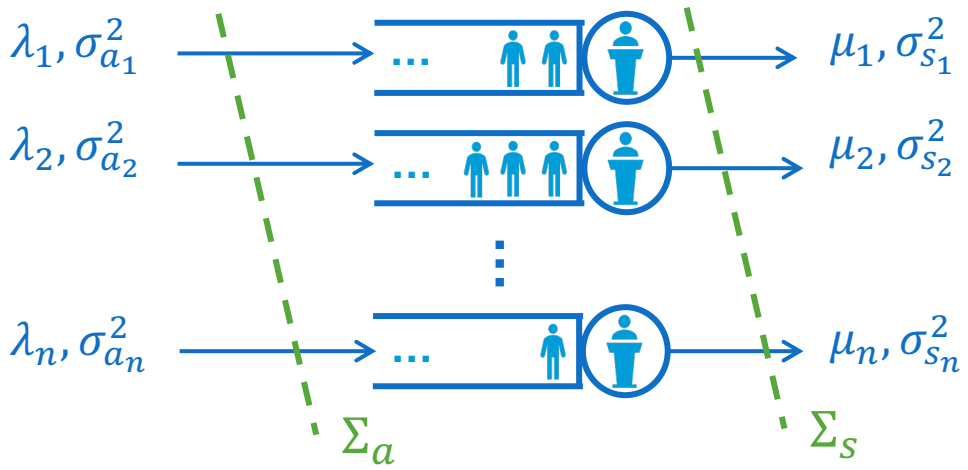
- But this is **unrealistic** for real data centers

INTERPRETATION AND EXAMPLES (cont.)

PARALLEL SERVER SYSTEM

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \boldsymbol{q}, \boldsymbol{\mu} \rangle] = \frac{1}{2}\left(\|\boldsymbol{\sigma_a}\|^2 + \|\boldsymbol{\sigma_s}\|^2\right)$$

- Even if queues are correlated, the answer only depends on the variances
- Recover Kingman's bound



INTERPRETATION AND EXAMPLES (cont.)

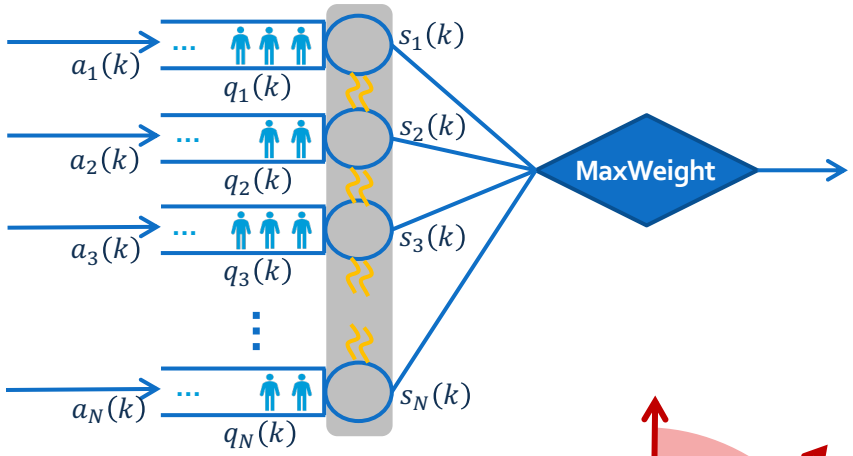
FULL DIMENSIONAL SSC

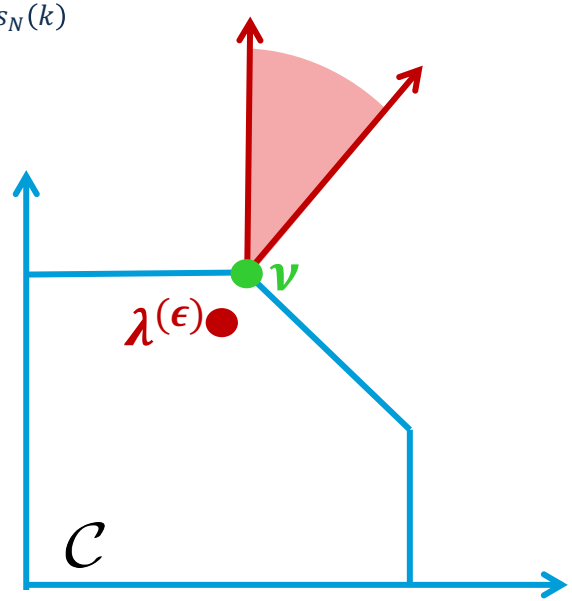
- SSC occurs into an N-dimensional subspace

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle] = \frac{1}{2}\left(\|\boldsymbol{\sigma_a}\|^2 + \mathbf{1}^T\left((C^T C)^{-1} \circ \Sigma_{cs}\right)\mathbf{1}\right)$$

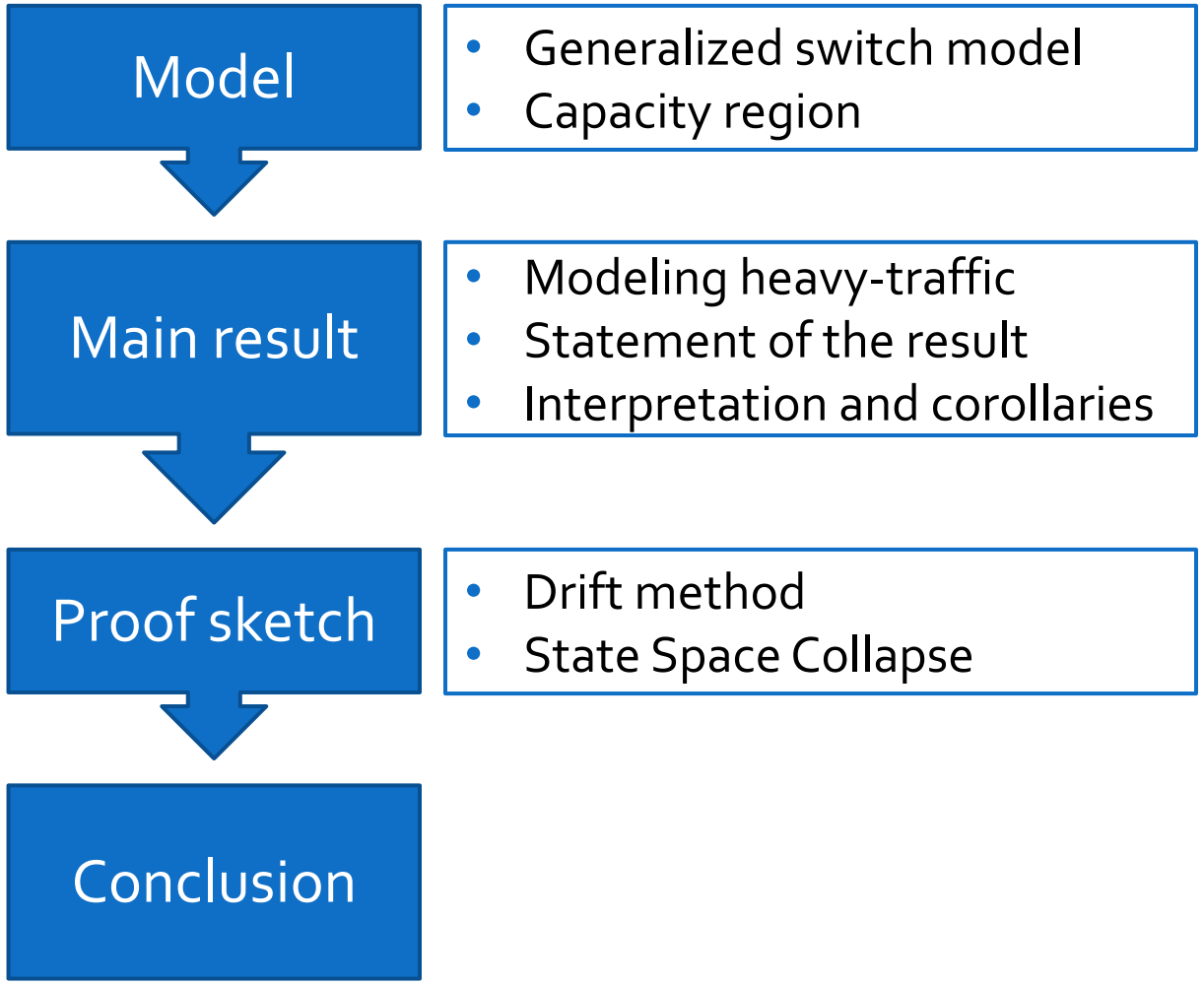
- Even if queues are correlated, the answer only depends on the variances

- Example: Parallel server system





OUTLINE


Model
Generalized switch model
Capacity region
Main result
Modeling heavy-traffic
Statement of the result
Interpretation and corollaries
Proof sketch
Drift method
State Space Collapse
Conclusion

PROOF: DRIFT METHOD

State Space Collapse

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

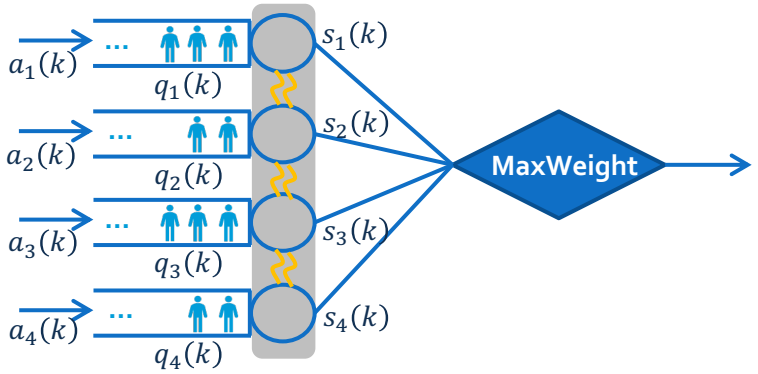
Set drift of test function to zero

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

Heavy-traffic limit of expected linear combination of queue lengths

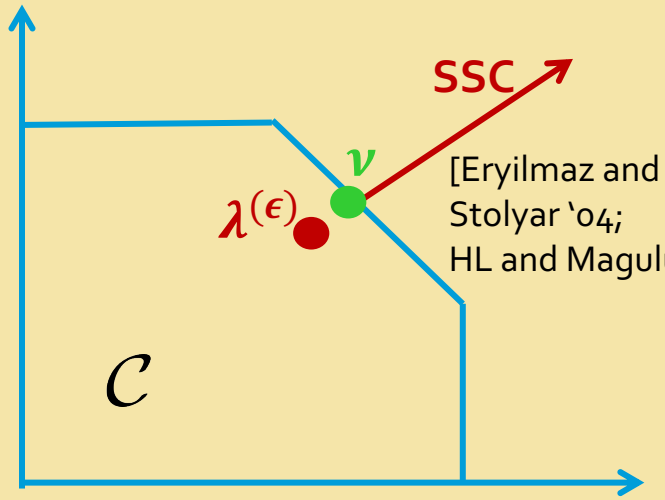
STATE SPACE COLLAPSE

- $\lambda_i = E[a_i(1)]$
- Consider $\boldsymbol{\nu} \in \partial\mathcal{C}$ and define $\boldsymbol{\lambda}^{(\epsilon)} := (1-\epsilon)\boldsymbol{\nu}$, where $\epsilon \in (0,1)$



COMPLETE RESOURCE POOLING (CRP)

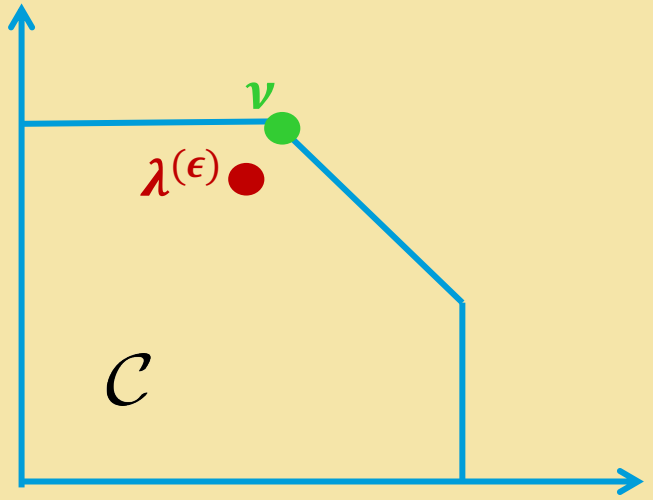
- Approach the interior of a facet



[Eryilmaz and Srikant '13 ; Stolyar '04; HL and Maguluri '19]

NO CRP

- Approach intersection of facets

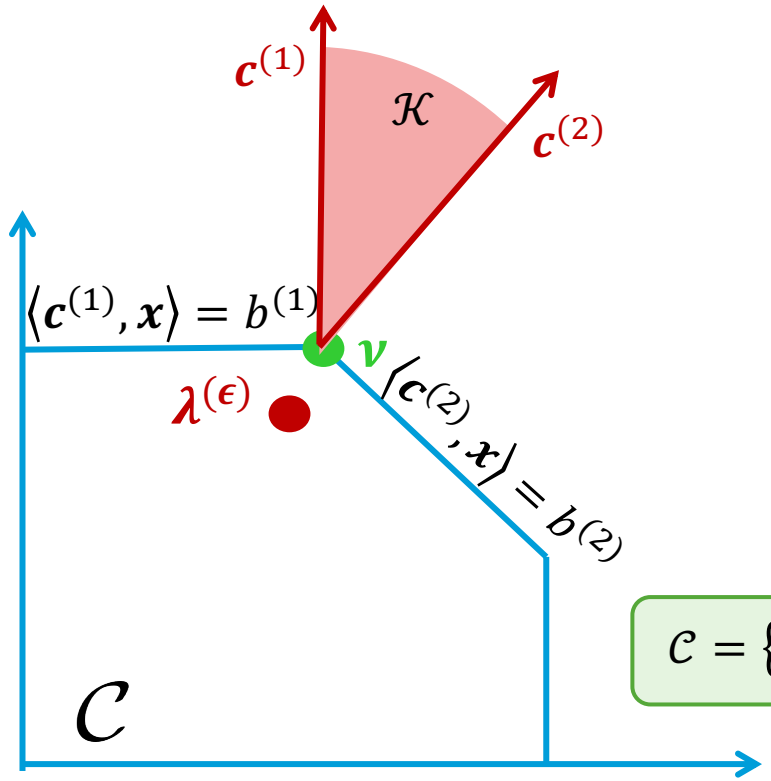


STATE SPACE COLLAPSE (cont.)

- P= set of indexes of the facets that intersect at $\boldsymbol{\nu}$

$$P = \{\ell \in [L]: \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{\nu} \rangle = b^{(\ell)}\}$$

- Cone $\mathcal{K}$ generated by $\boldsymbol{c}^{(\ell)}$ for $\ell \in P$



Proposition: (SSC)
$\boldsymbol{q}_{\parallel}(k)$: projection of $\boldsymbol{q}(k)$ on $\mathcal{K}$
$\boldsymbol{q}_{\perp}(k) \coloneqq \boldsymbol{q}(k) - \boldsymbol{q}_{\parallel}(k)$: error of approximating $\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$
Then, $E[\|\boldsymbol{q}_{\perp}\|^t] \leq T_t$ for all $t = 1,2, \ldots$

$\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$ as $\epsilon \downarrow 0$

$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^N: \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \leq b^{(\ell)}, \ell = 1, \ldots, L\}$$

IDEA OF THE PROOF

- Set to zero the drift of $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$
- We obtain

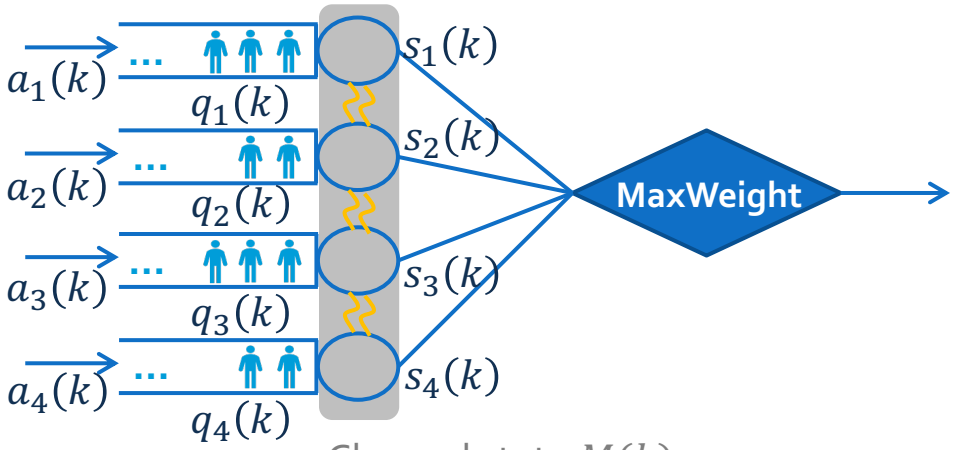
$$\underbrace{2E\left[\left\langle \boldsymbol{q}_{\|}(k), \boldsymbol{s}_{\|}(k) - \boldsymbol{a}_{\|}(k)\right\rangle\right]}_{\mathcal{T}_1 \xrightarrow{\epsilon \to 0} 2\epsilon E[\langle \boldsymbol{q}, \boldsymbol{\nu}\rangle]}$$

$$= \underbrace{E\left[\left\|\boldsymbol{a}_{\|}(k) - \boldsymbol{s}_{\|}(k)\right\|^2\right]}_{\mathcal{T}_2} - \underbrace{E\left[\left\|\boldsymbol{u}_{\|}(k)\right\|^2\right]}_{\mathcal{T}_3} + \underbrace{2E\left[\left\langle \boldsymbol{q}_{\|}(k+1), \boldsymbol{u}_{\|}(k)\right\rangle\right]}_{\mathcal{T}_4}$$

(solve least squares problem) $\mathcal{T}_2 \xrightarrow{\epsilon \to 0}$ RHS

$\mathcal{T}_3 \xrightarrow{\epsilon \to 0} 0$

$\mathcal{T}_4 \xrightarrow{\epsilon \to 0} 0$ (using SSC)



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

$$q_i(k+1) = \max\{q_i(k) + a_i(k) - s_i(k), 0\}$$
$$= q_i(k) + a_i(k) - s_i(k) + u_i(k)$$

$q_i(k+1)u_i(k) = 0$ but
$q_i(k+1)u_j(k) \neq 0$ if $i \neq j$

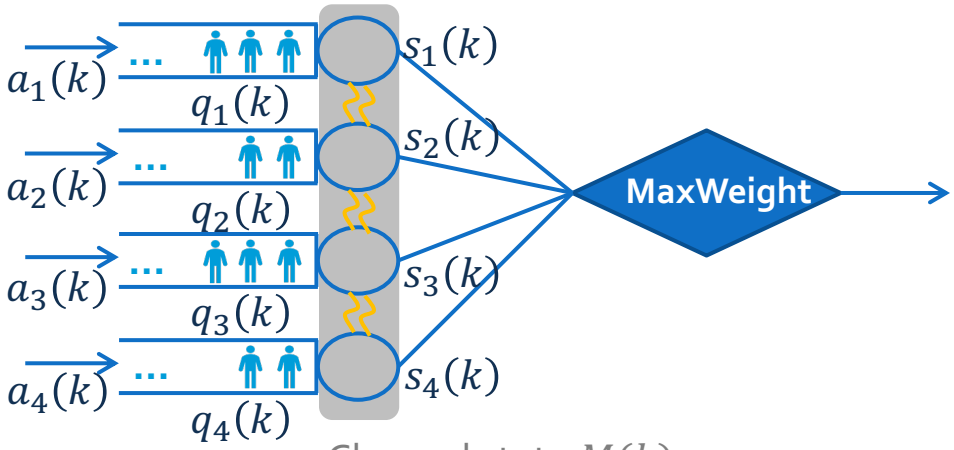
DISCUSSION

- So

$$\lim_{\epsilon \downarrow 0} \epsilon E[\boldsymbol{q}] = ??$$

- Can we obtain other linear combinations of the queue lengths?

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \boldsymbol{q}, \boldsymbol{w} \rangle] = ??$$



Channel state $M(k) = m$

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

CASE 1:

$\boldsymbol{w} = \boldsymbol{\nu} + \boldsymbol{y}$ with $\boldsymbol{y} \parallel \boldsymbol{q}_{\perp}$

(by SSC)

CASE 2:

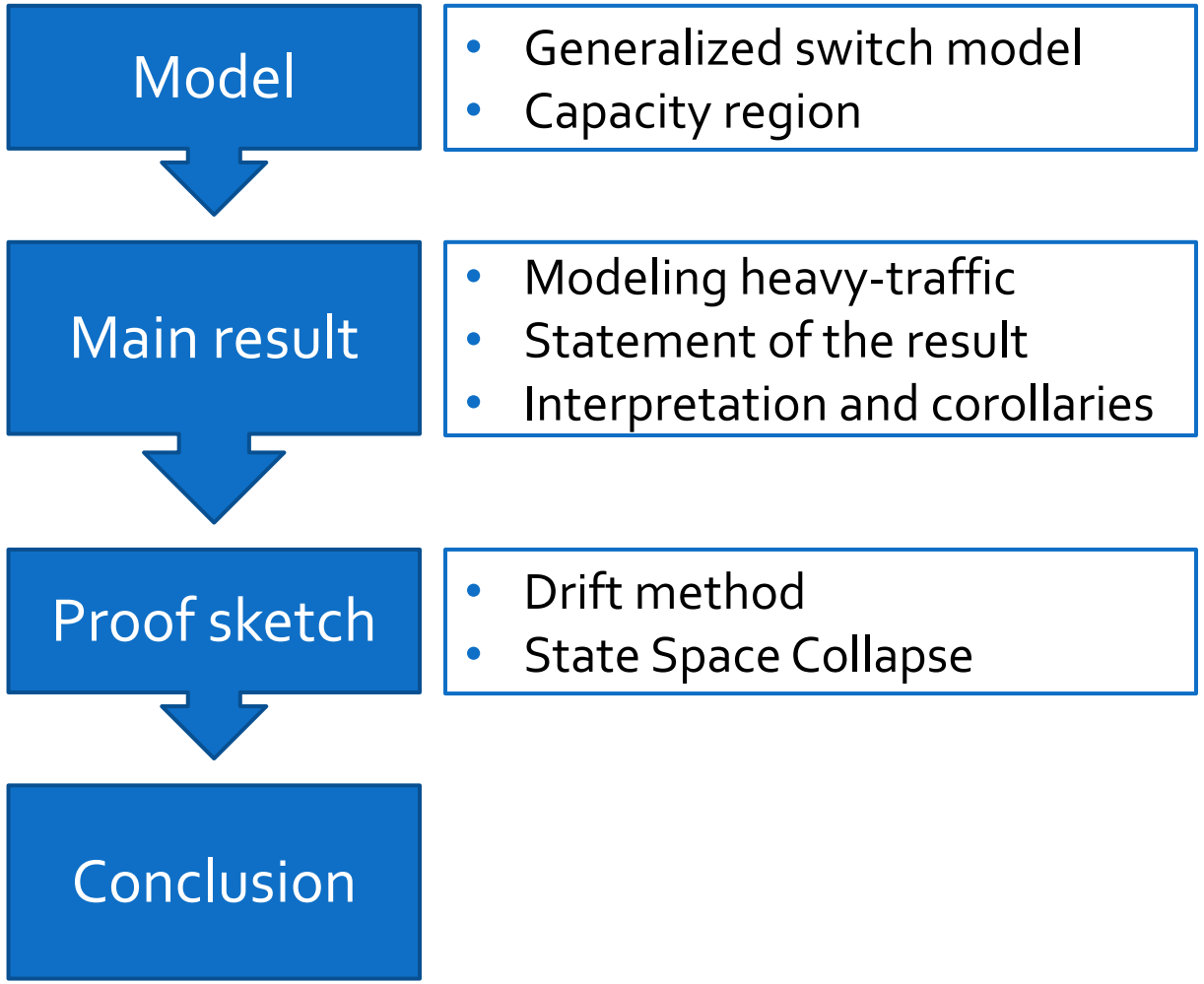
Other $\boldsymbol{w}$

? Not with the Drift method !

Theorem:

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \boldsymbol{q}, \boldsymbol{\nu} \rangle]$$
$$= \frac{1}{2}\left(\mathbf{1}^T (H \circ \Sigma_a)\mathbf{1} + \mathbf{1}^T \left((C^T C)^{-1} \circ \Sigma_{cs}\right)\mathbf{1}\right)$$

OUTLINE


Model
Generalized switch model
Capacity region
Main result
Modeling heavy-traffic
Statement of the result
Interpretation and corollaries
Proof sketch
Drift method
State Space Collapse
Conclusion

CONCLUSION

- Generalized switch
  - One of the most general single-hop queueing systems
- Heavy-traffic analysis using the drift method
  - State Space Collapse
  - Handle unused service $q_i(k+1)u_i(k) = 0$
- But we only get specific linear combinations of queue lengths

HEAVY-TRAFFIC ANALYSIS OF THE GENERALIZED SWITCH UNDER MULTIDIMENSIONAL STATE SPACE COLLAPSE

Daniela Hurtado Lange, Siva Theja Maguluri

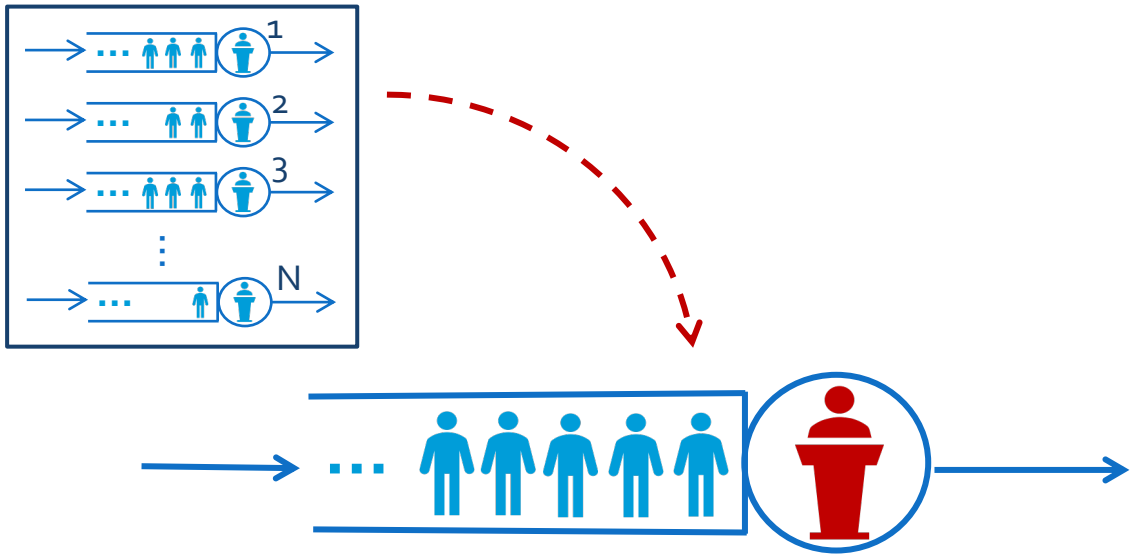
Industrial and Systems Engineering Department

Georgia Institute of Technology

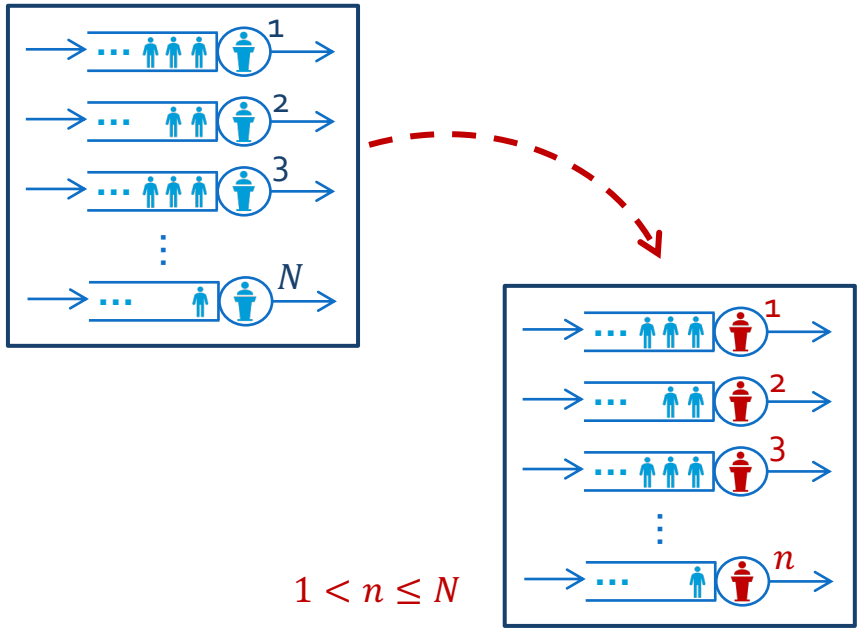
STATE SPACE COLLAPSE (SSC)

- Queueing system behaves as a lower-dimensional system

COMPLETE RESOURCE POOLING (CRP)



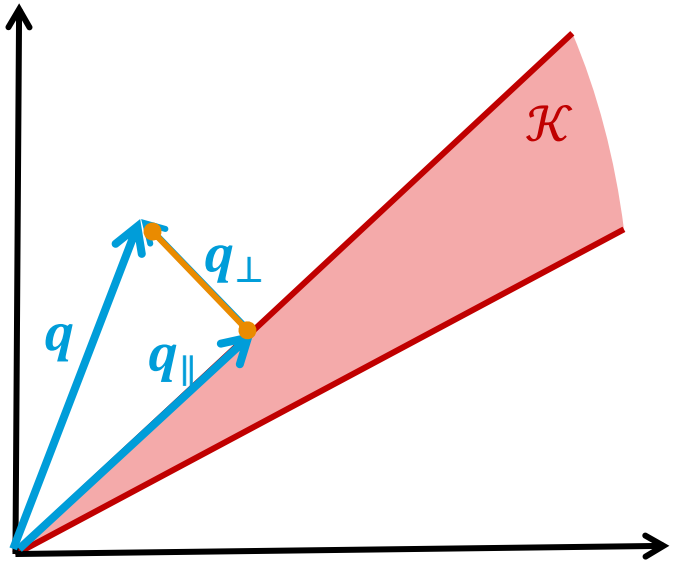
NO CRP



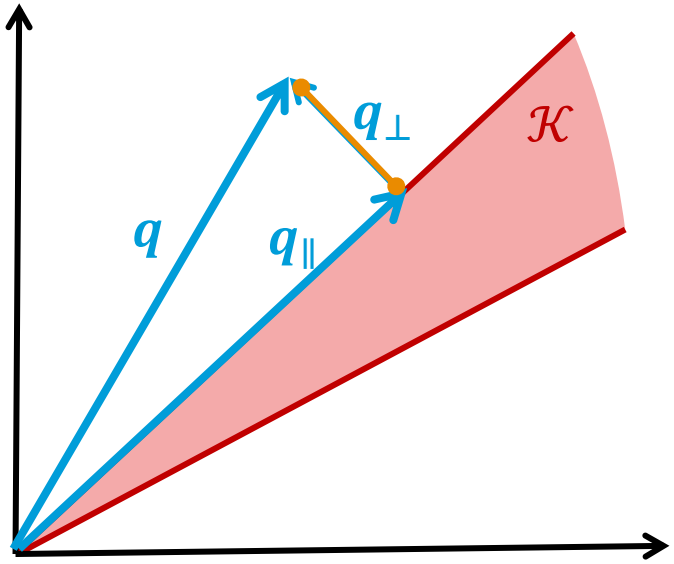
$1 < n \leq N$

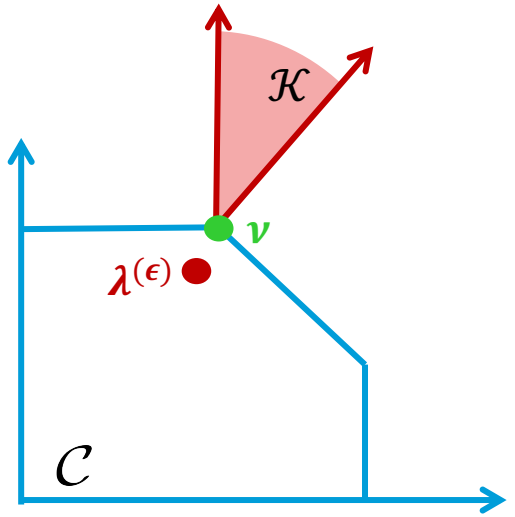
SSC INTERPRETATION

Proposition:
$\boldsymbol{q}_{\parallel}(k)$: projection of $\boldsymbol{q}(k)$ on $\mathcal{K}$
$\boldsymbol{q}_{\perp}(k) \coloneqq \boldsymbol{q}(k) - \boldsymbol{q}_{\parallel}(k)$: error of approximating $\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$
Then, $E[\|\boldsymbol{q}_{\perp}\|^t] \leq T_t$ for all $t = 1,2, \ldots$



Smaller ϵ:



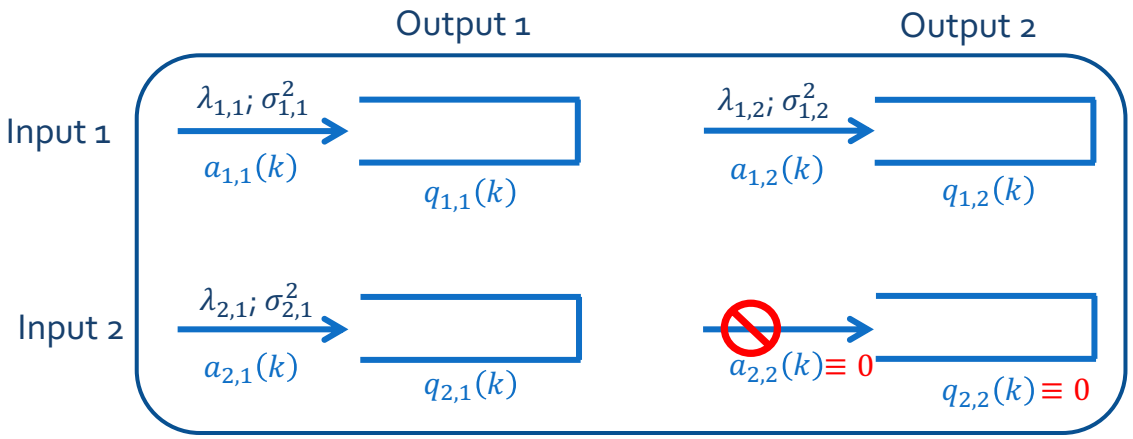


$$\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^N : \langle \boldsymbol{c}^{(\ell)}, \boldsymbol{x} \rangle \leq b^{(\ell)}, \ell = 1, \ldots, L\}$$

$\boldsymbol{q} \approx \boldsymbol{q}_{\parallel}$ as $\epsilon \downarrow 0$

NEW VIEW OF THE DRIFT METHOD [Hurtado Lange, Maguluri '19]

- Simplest no-CRP queueing system
  - 2x2 switch operating under MaxWeight
  - No arrivals to queue 2,2
  - Arrivals to other queues with mean $\lambda_{i,j}$ and variance $\sigma^2_{i,j}$



- Heavy-traffic: $\epsilon > 0$
  - $\lambda_{1,1} = 1 - \lambda - \epsilon$
  - $\lambda_{1,2} = \lambda_{2,1} = \lambda$

- State Space Collapse [Maguluri et.al. '18]

2-dimensional
⇒No-CRP

$$\mathcal{K} = \left\{ \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & 0 \end{bmatrix} : x_{1,1} = x_{1,2} + x_{2,1} \right\}$$

Theorem [Maguluri et.al. '18]:

$$\lim_{\epsilon \downarrow 0} \epsilon E\left[q_{1,1} + q_{1,2} + q_{2,1}\right] = \frac{2}{3}\left(\sigma^2_{1,1} + \sigma^2_{1,2} + \sigma^2_{2,1}\right)$$

Proof: Set to zero the drift of $V(\boldsymbol{q}) = \|\boldsymbol{q}_{\|}\|^2$ and use SSC.

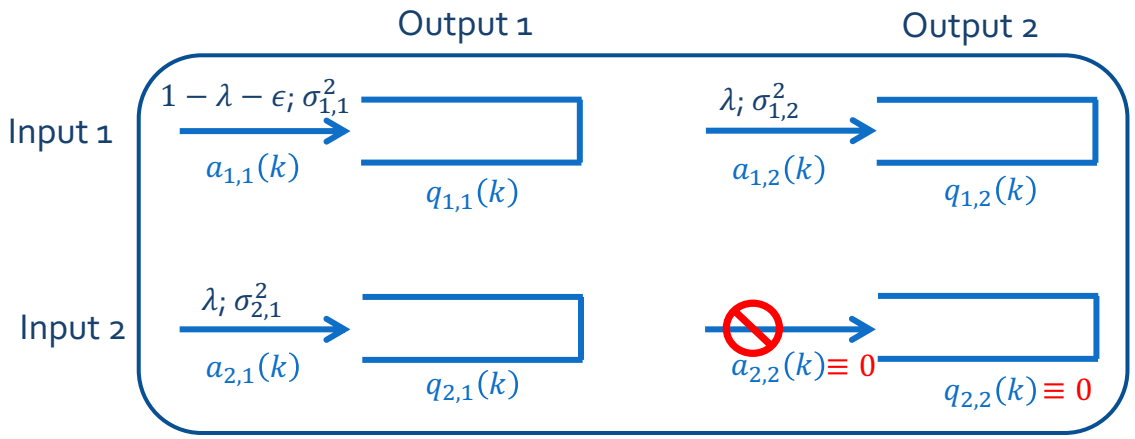
OTHER LINEAR COMBINATIONS?

- SSC:
  - $E[\|\boldsymbol{q}\|^2] \approx E\left[\|\boldsymbol{q}_{\|}\|^2\right]$
  - $q_{\|1,1} = q_{\|1,2} + q_{\|2,1}$

- Most general quadratic test function

$$V(\boldsymbol{q}) = \alpha_1 q_{\|1,2}^2 + \alpha_2 q_{\|2,1}^2 + \alpha_3 q_{\|1,2} q_{\|2,1}$$

$$V_1(\boldsymbol{q}) = q_{\|1,2}^2$$
$$V_2(\boldsymbol{q}) = q_{\|2,1}^2$$
$$V_3(\boldsymbol{q}) = q_{\|1,2} q_{\|2,1}$$



$$\mathcal{K} = \left\{ \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & 0 \end{bmatrix} : x_{1,1} = x_{1,2} + x_{2,1} \right\}$$

OTHER LINEAR COMBINATIONS? (cont.)

- Set to zero the drift of these 3 test functions:

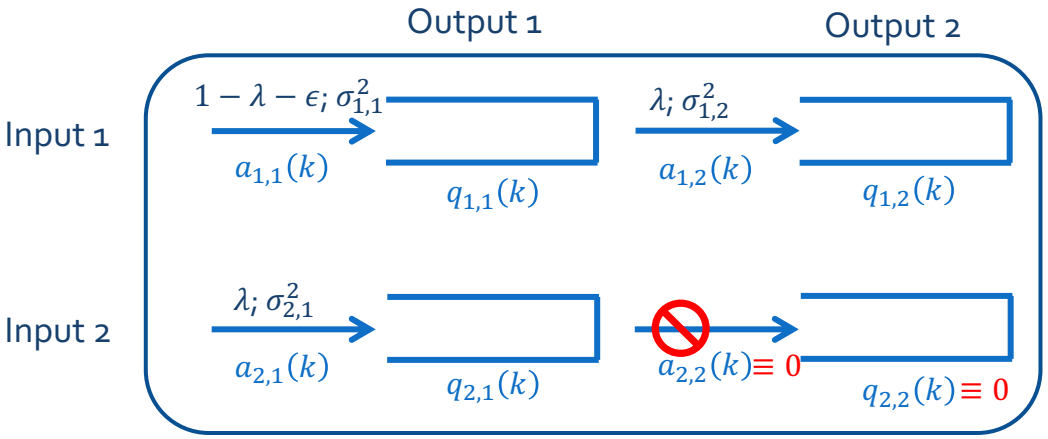
$$2\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{1,2}\right]=\frac{\sigma_{1,1}^2+4\sigma_{1,2}^2+\sigma_{2,1}^2}{3}-2\lim_{\epsilon\downarrow 0}E\left[q_{1,2}^{+}u_{2,1}\right]$$

$$2\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{2,1}\right]=\frac{\sigma_{1,1}^2+\sigma_{1,2}^2+4\sigma_{2,1}^2}{3}-2\lim_{\epsilon\downarrow 0}E\left[q_{2,1}^{+}u_{1,2}\right]$$

$$\lim_{\epsilon\downarrow 0}\epsilon E\left[q_{1,2}+q_{2,1}\right]=\frac{\sigma_{1,1}^2-2\sigma_{1,2}^2-2\sigma_{2,1}^2}{3}+2\lim_{\epsilon\downarrow 0}E\left[q_{1,2}^{+}u_{2,1}+q_{2,1}^{+}u_{1,2}\right]$$

4 unknowns
3 equations !

Need more equations !



$$\mathcal{K}=\left\{\begin{bmatrix}x_{1,1} & x_{1,2}\\ x_{2,1} & 0\end{bmatrix}: x_{1,1}=x_{1,2}+x_{2,1}\right\}$$

$$V_1(\boldsymbol{q})=q_{\|1,2}^2$$
$$V_2(\boldsymbol{q})=q_{\|2,1}^2$$
$$V_3(\boldsymbol{q})=q_{\|1,2}q_{\|2,1}$$