

# HEAVY-TRAFFIC ANALYSIS OF THE GENERALIZED SWITCH WITH NO COMPLETE RESOURCE POOLING

---

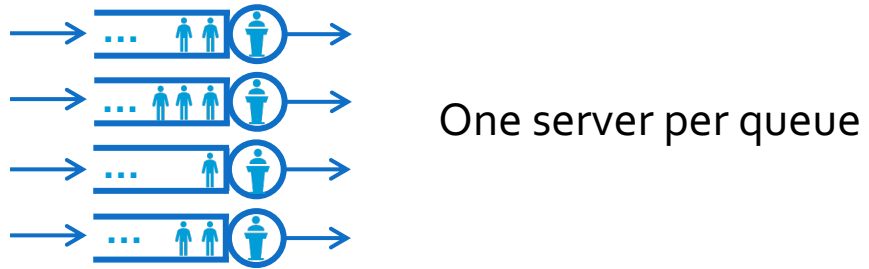
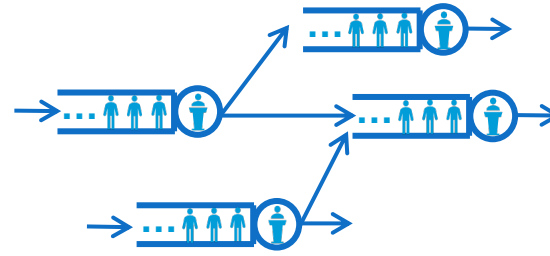
**Daniela Hurtado Lange, Siva Theja Maguluri**

Industrial and Systems Engineering Department

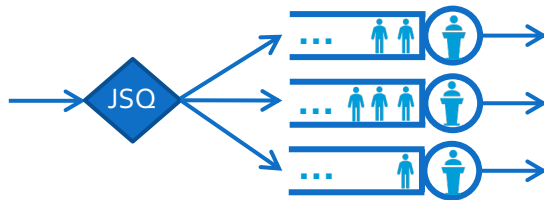
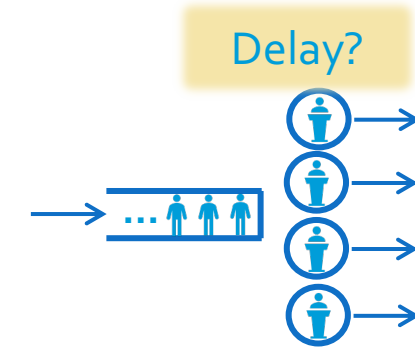
Georgia Institute of Technology

July 14<sup>th</sup>, 2020

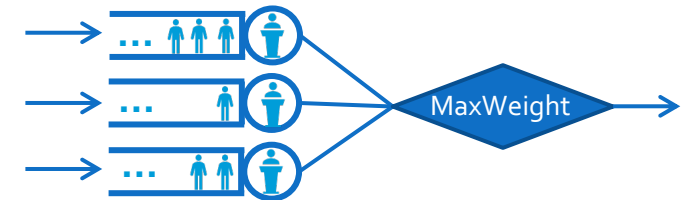
# STOCHASTIC PROCESSING NETWORKS



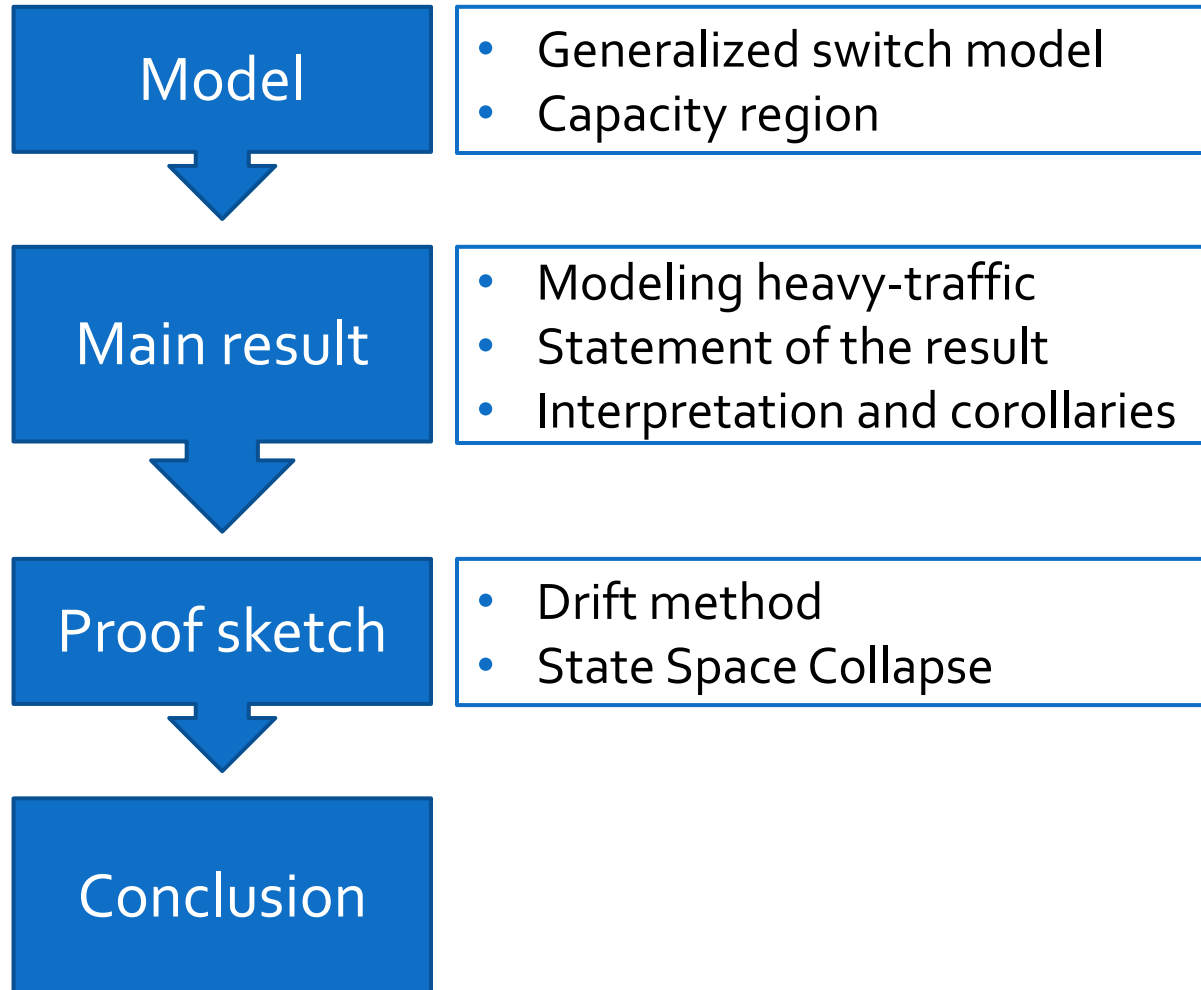
Generalized switch



Control at service

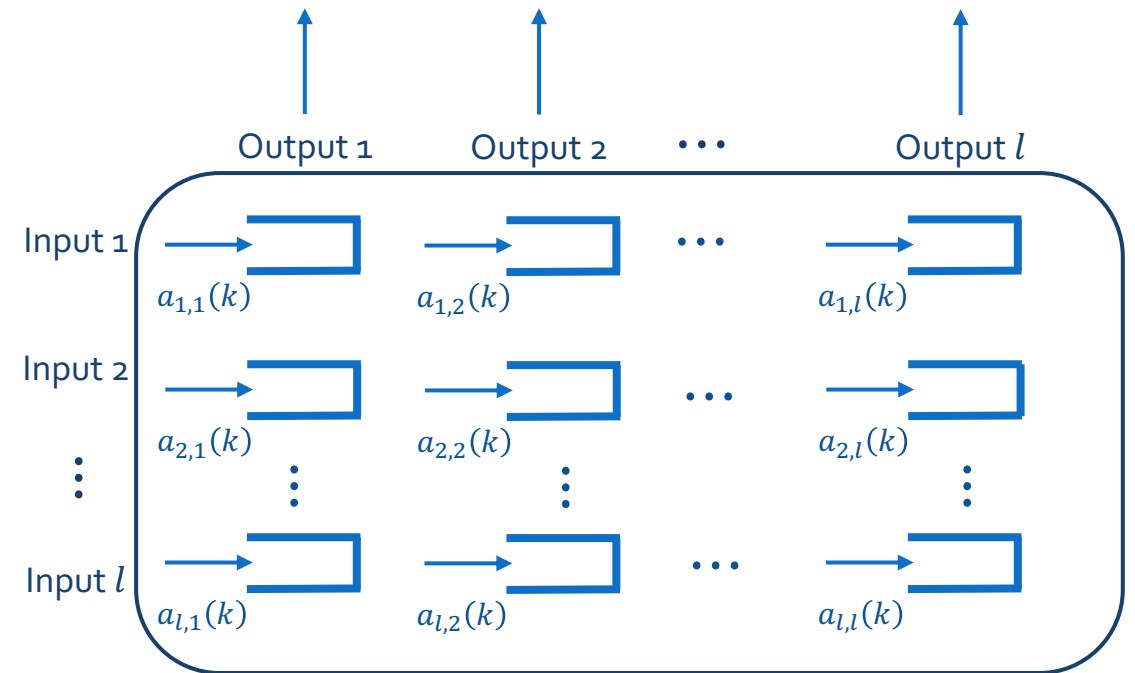


# OUTLINE



# 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  $s(k)$  for each  $k$

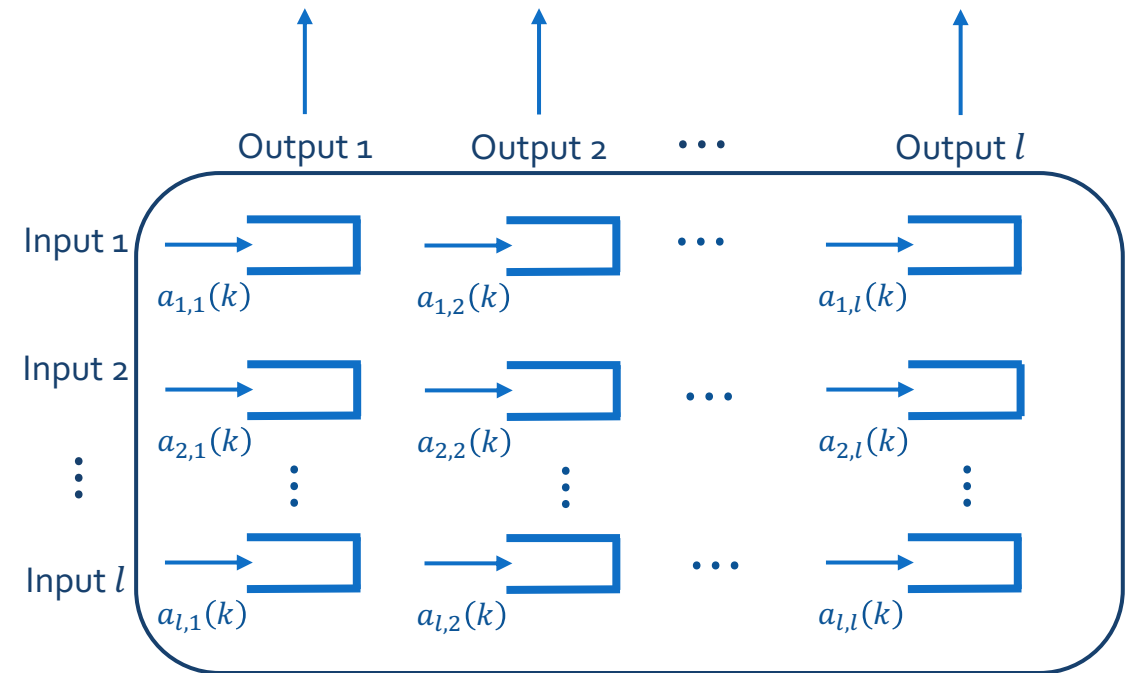
# INPUT QUEUED SWITCH MODEL (CONT.)

- Capacity region:

$$\mathcal{C} = \left\{ \lambda \in \mathbb{R}_+^{l \times l} : \sum_i \lambda_{i,j} \leq 1 \quad \forall j \text{ \& } \sum_j \lambda_{i,j} \leq 1 \quad \forall i \right\}$$

- Constraint  $\Rightarrow \mathbf{s}(k)$  is a permutation matrix

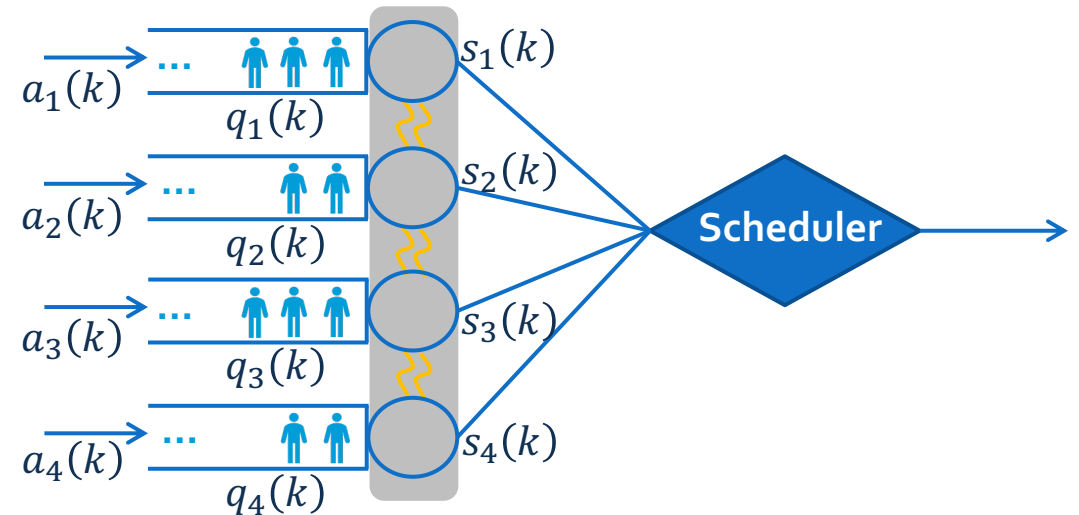
- $S$  = set of  $l \times l$  permutation matrices  
 $\Rightarrow \mathcal{C} = \text{ConvexHull}(S)$



**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  $k$
  - $\Sigma_a$  is the covariance matrix of  $\mathbf{a}(k)$
- $s_i(k)$ : **offered** service from queue  $i$  in time slot  $k$



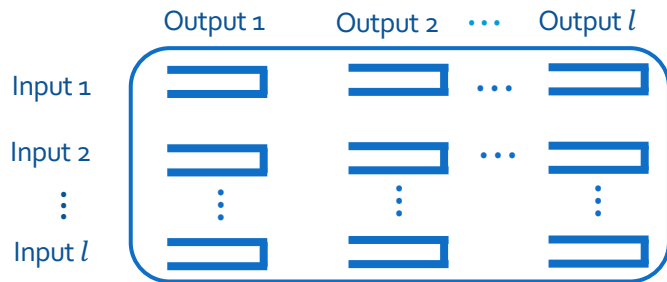
Channel state  $M(k) = m$   
Feasible service rates:  $S^{(m)}$

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

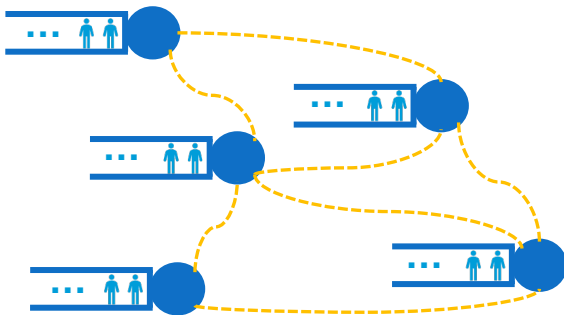
i.i.d. across time

# EXAMPLES

- Fixed channel state
  - Input queued switch

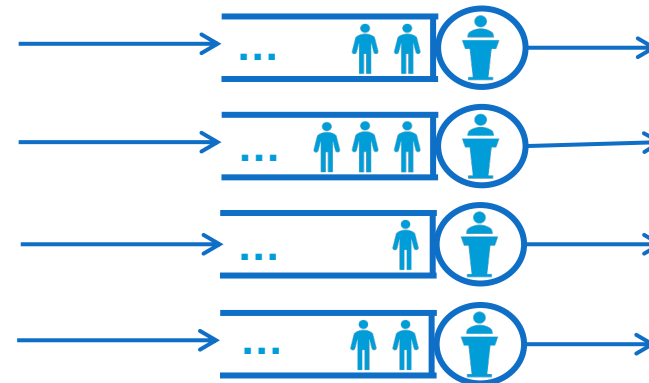


- Ad hoc wireless network



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

- Parallel server system



# CAPACITY REGION

- Channel state pmf:

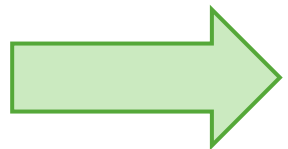
$$\psi_m = P\{M(k) = m\}$$

- Capacity region:

$$\mathcal{C} = \sum_m \psi_m \text{ConvexHull}(S^{(m)})$$

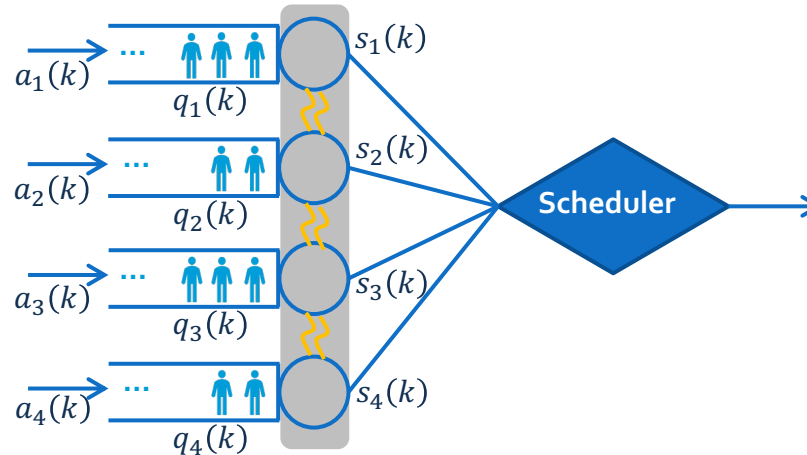
- Assume

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

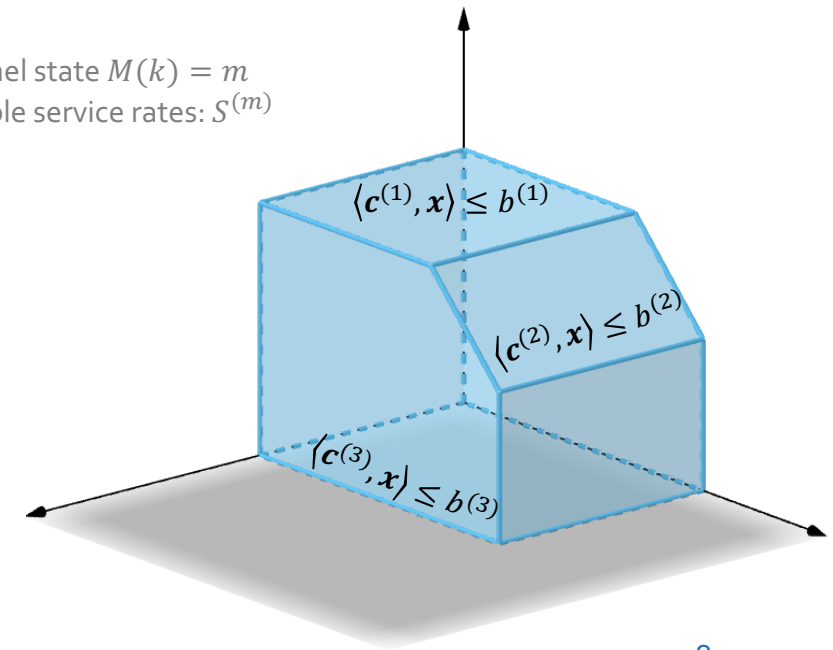


$$\mathcal{C} = \{x \in \mathbb{R}_+^N : \langle c^{(\ell)}, x \rangle \leq b^{(\ell)}, \ell = 1, \dots, L\}$$

Polytope



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



# GENERALIZED SWITCH MODEL (cont.)

- Scheduling: MaxWeight
  - Maximize weighted sum of queue lengths
  - Weight = feasible service rates

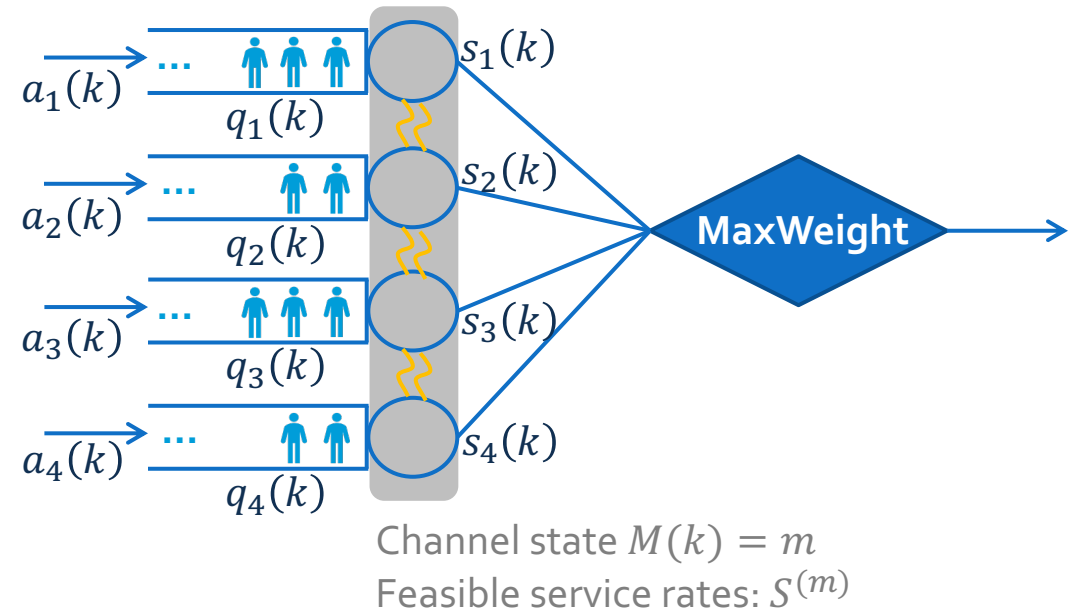
- Formally, if  $M(k) = m$  we choose

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

- Dynamics of the queues:

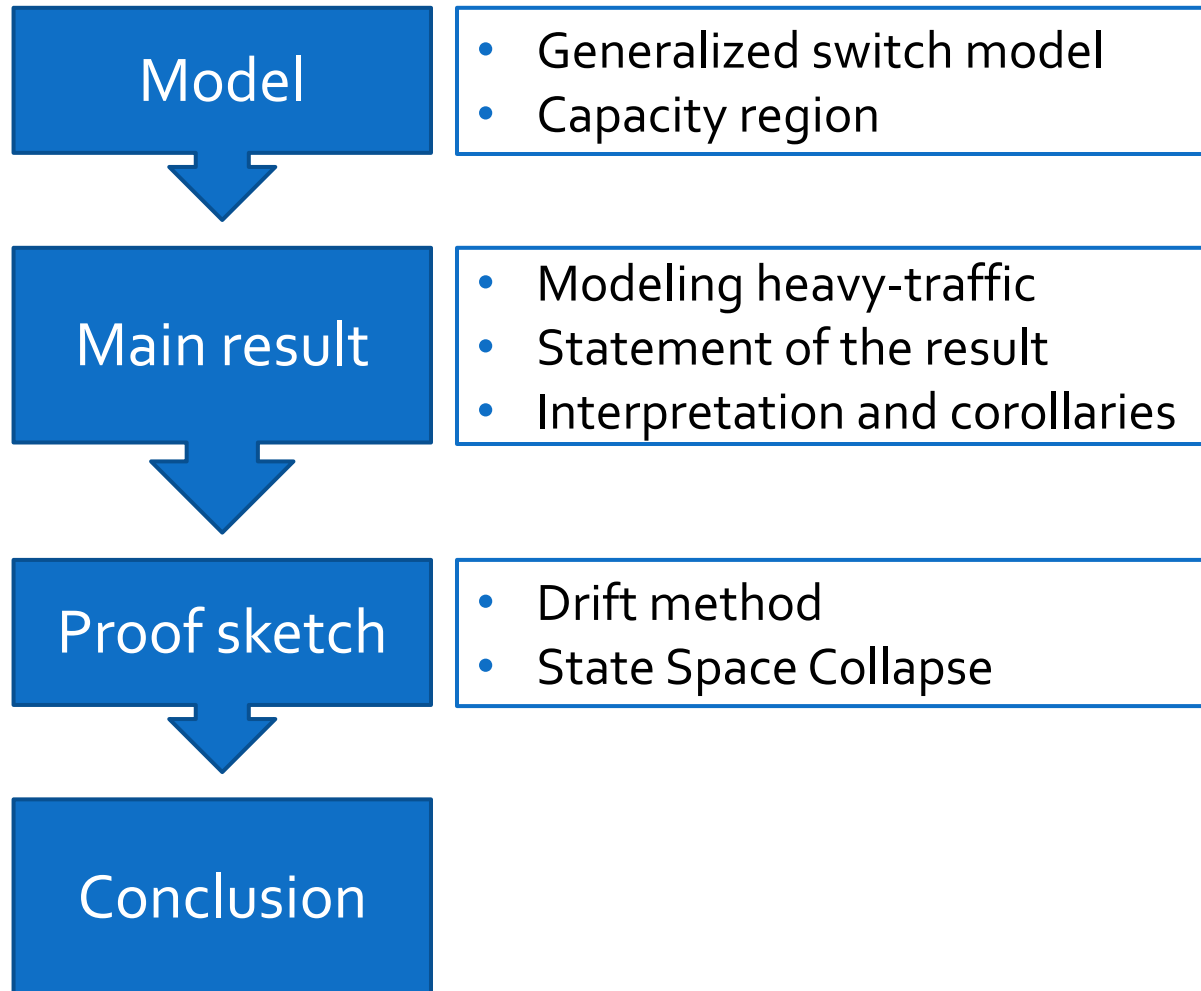
$$\begin{aligned} q_i(k+1) &= [q_i(k) + a_i(k) - s_i(k)]^+ \\ &= q_i(k) + a_i(k) - s_i(k) + u_i(k) \end{aligned}$$

→ Unused service



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

# OUTLINE



# 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:  
 $\lambda \rightarrow \mu$

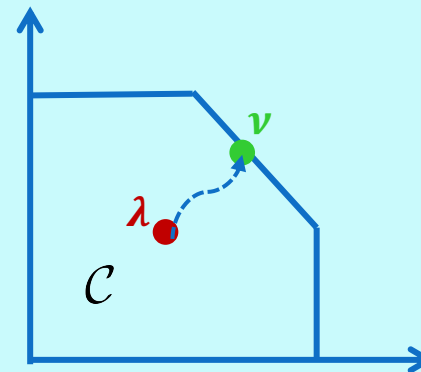
**In general...**

$\mathcal{C}$  = Capacity region

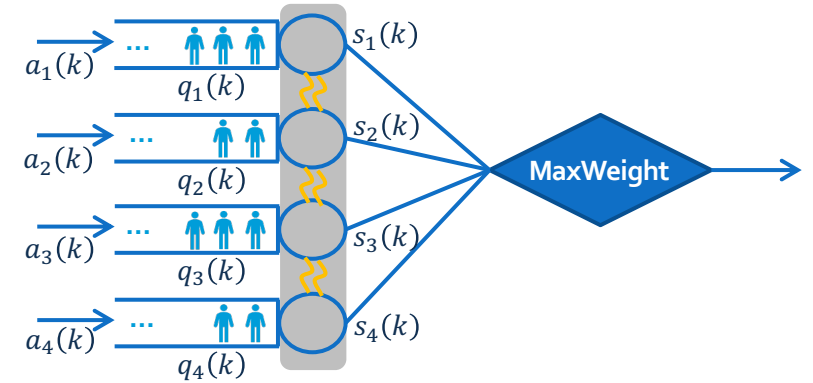
$\mathbf{v} \in \partial\mathcal{C}$

$\lambda = (1 - \epsilon)\mathbf{v}$ , with  $\epsilon \in (0,1)$

Take limit as  $\epsilon \downarrow 0$



# MAIN RESULT



Linear combination of the queue lengths

**Theorem:** [HL, Maguluri '19]

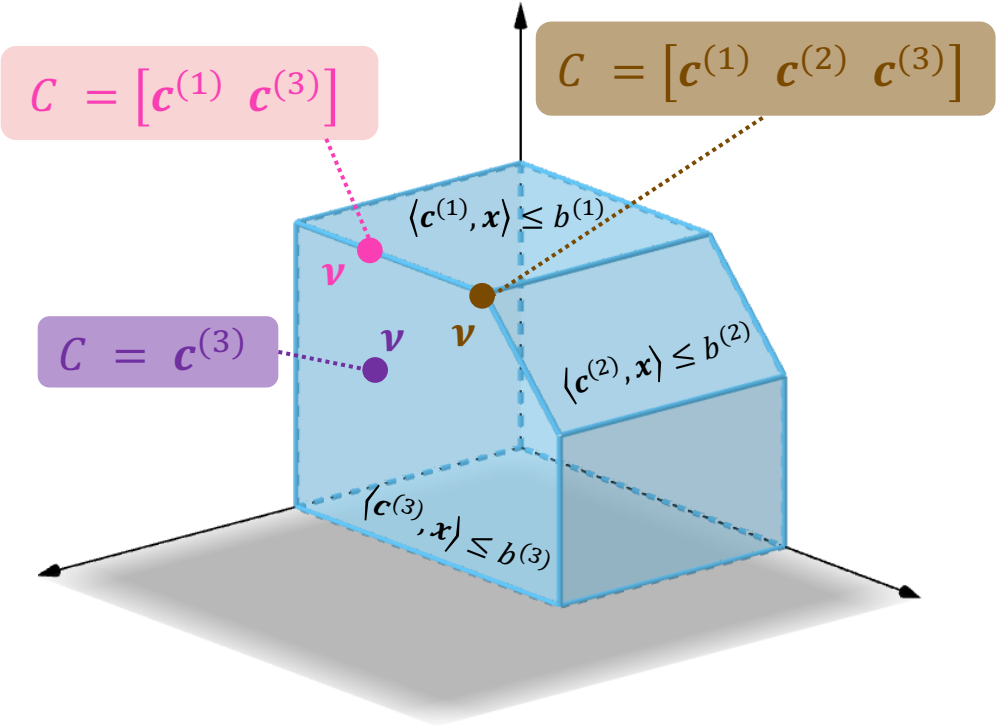
$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \mathbf{q}, \mathbf{v} \rangle] = \frac{1}{2} \left( \mathbf{1}^T (H \circ \Sigma_a) \mathbf{1} + \mathbf{1}^T ((C^T C)^{-1} \circ \Sigma_{CS}) \mathbf{1} \right)$$

Linear combination of covariance matrix of arrivals

Linear combination of variability of service process (channel state)

Projection matrix on column space of  $C$

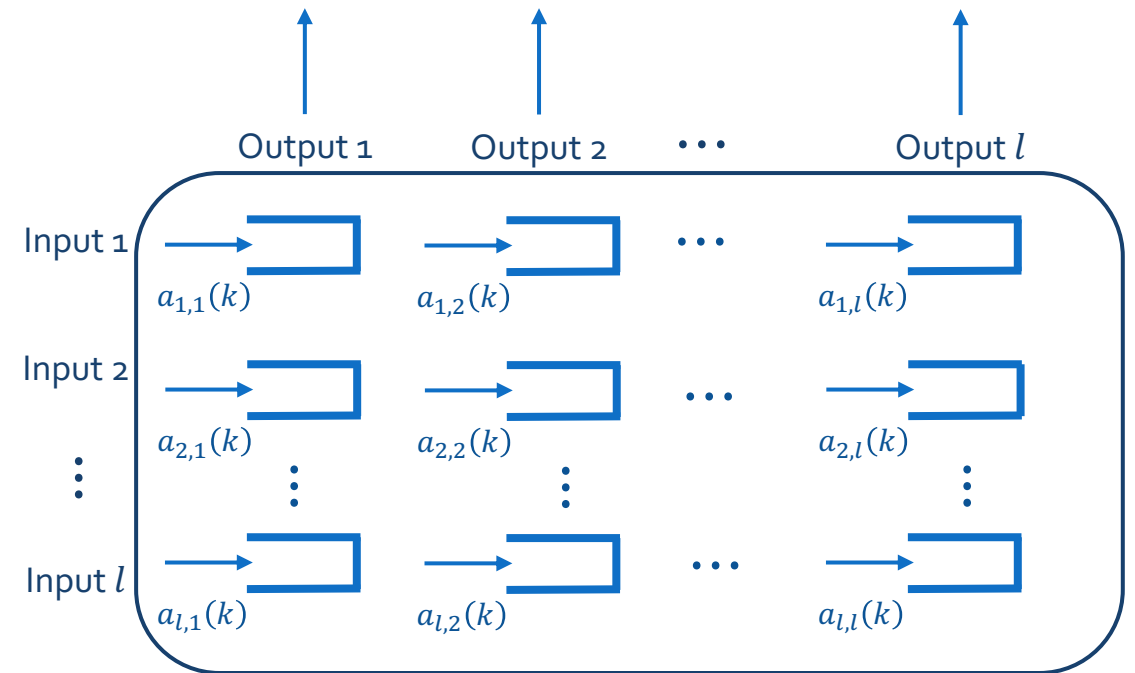
Matrix where columns are defined by facets that intersect at  $\mathbf{v}$



# INTERPRETATION AND EXAMPLES

## INPUT QUEUED SWITCH WITH CORRELATED ARRIVALS

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



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

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

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

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

# INTERPRETATION AND EXAMPLES

## AD HOC WIRELESS NETWORK

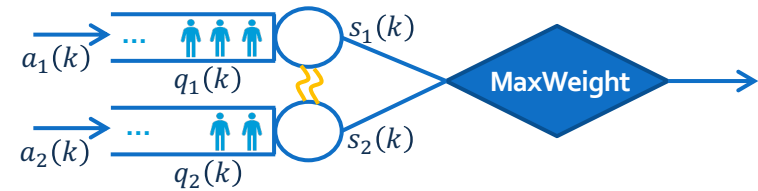
- Feasible service rate vectors:

$$S = \left\{ (1,0); (0,1); \left(\frac{2}{3}, \frac{2}{3}\right) \right\}$$

- Covariance matrix of arrivals:

$$\Sigma_a = \begin{pmatrix} \sigma_1^2 & \varphi \\ \varphi & \sigma_2^2 \end{pmatrix}$$

$$\lim_{\epsilon \downarrow 0} \epsilon E[q_1 + q_2] = \frac{3}{4} (\sigma_1^2 + \sigma_2^2)$$



Covariance between arrivals does not appear

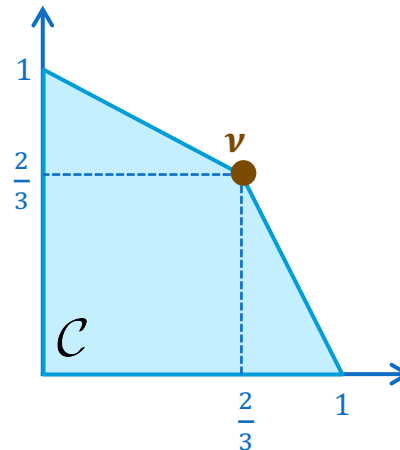
### Proof:

From the theorem, we know:

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

= 0 (channel state is fixed)

$$\text{Also, } C = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \Rightarrow H = I$$



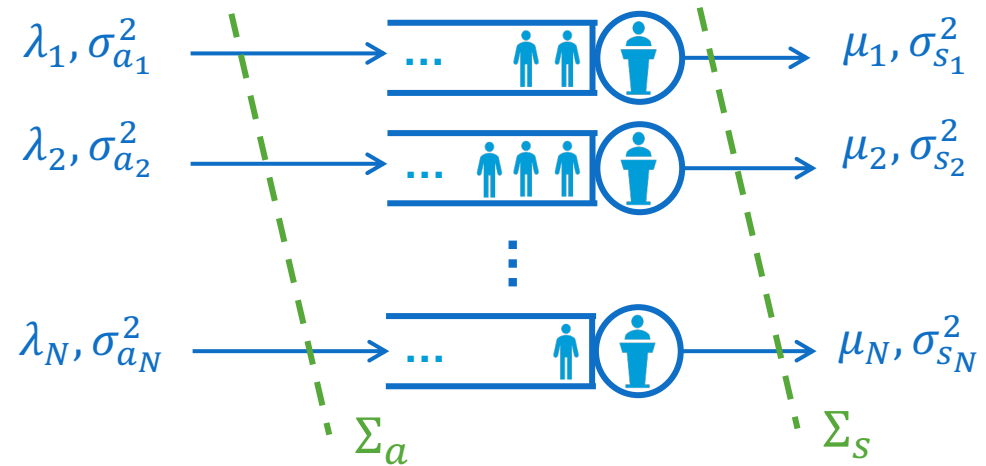
# INTERPRETATION AND EXAMPLES (cont.)

## PARALLEL SERVER SYSTEM

$$\lim_{\epsilon \downarrow 0} \epsilon E[\langle \mathbf{q}, \boldsymbol{\mu} \rangle] = \frac{1}{2} (\|\boldsymbol{\sigma}_a\|^2 + \|\boldsymbol{\sigma}_s\|^2)$$

Covariance does not appear

- Recover Kingman's bound



### Proof:

From the theorem, we know:

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

Capacity region:  $\mathcal{C} = \{\boldsymbol{\lambda} \in \mathbb{R}_+^N : \lambda_i \leq \mu_i\}$

Take  $\mathbf{v} = \boldsymbol{\mu} \Rightarrow H = I$

# INTERPRETATION AND EXAMPLES (cont.)

$\mathbf{v}$  IS AT A VERTEX OF THE CAPACITY REGION

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

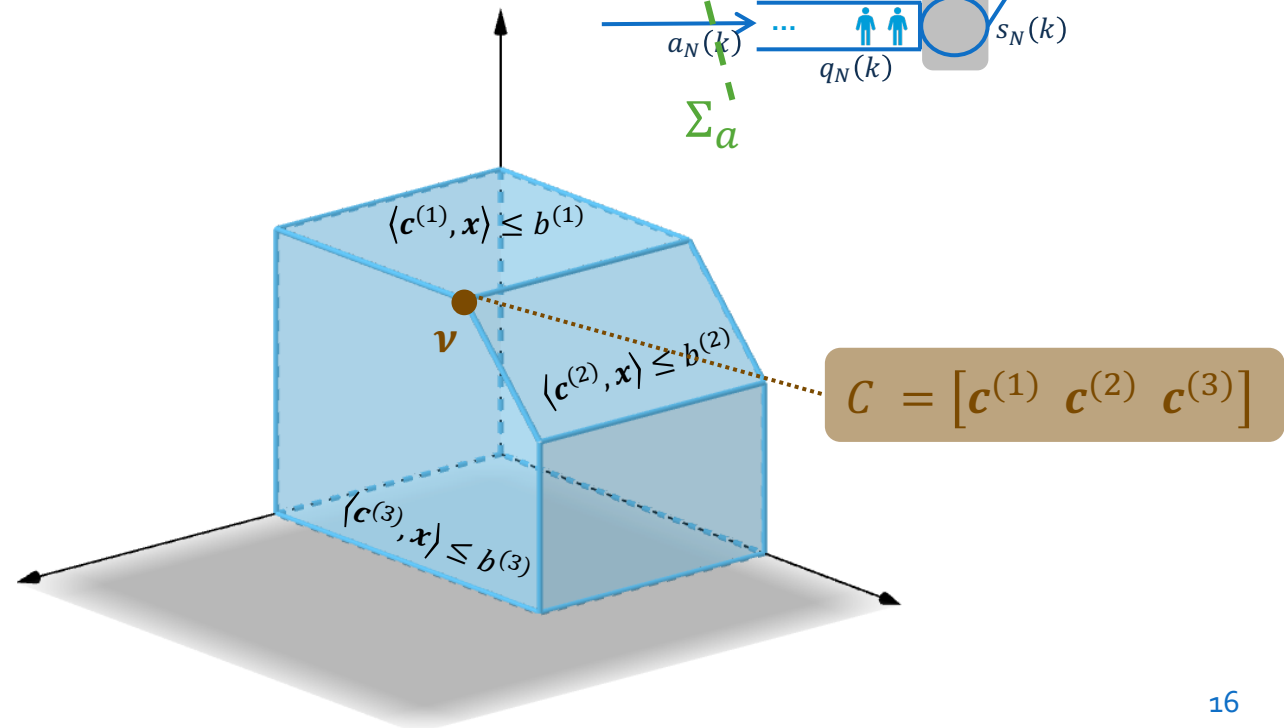
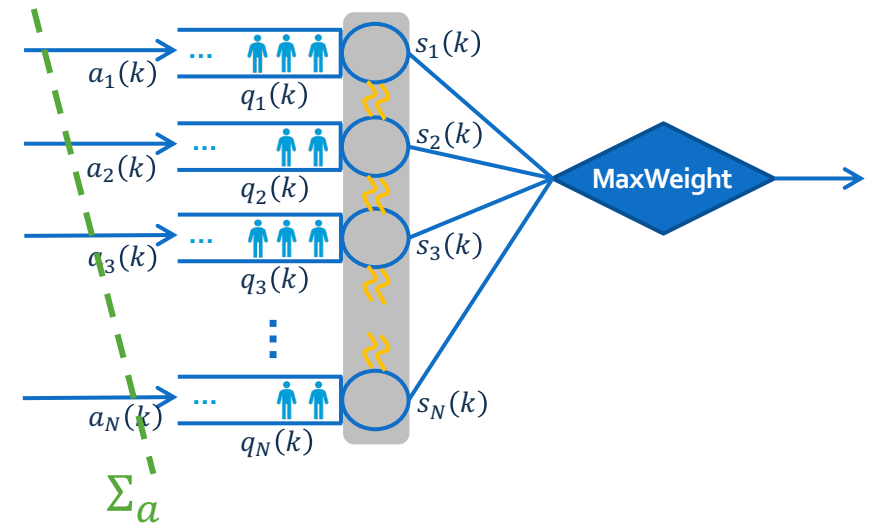
Covariance between arrivals does not appear

**Proof:**

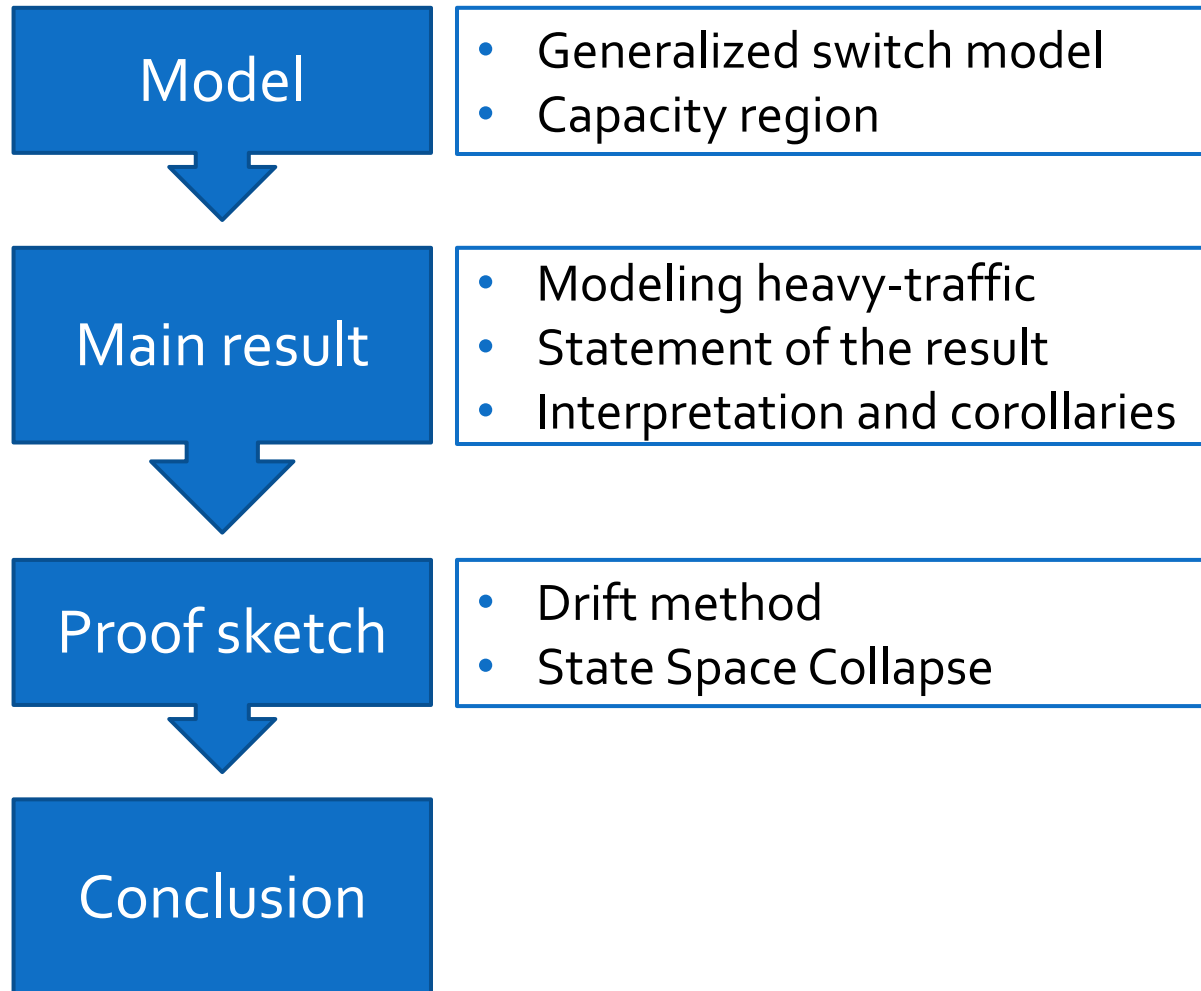
From the theorem, we know:

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

$\mathbf{v}$  is a vertex  $\Rightarrow C$  is square and invertible  
 $\Rightarrow H = I$



# OUTLINE



# PROOF: DRIFT METHOD

State Space Collapse



Set drift of test function to zero

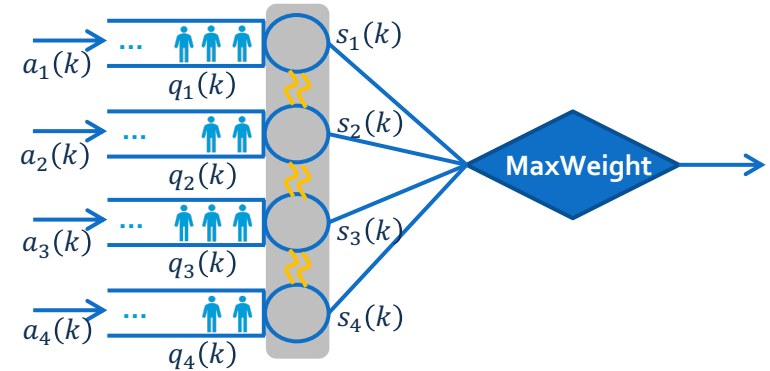


Heavy-traffic limit of expected  
linear combination of queue lengths

- Prove that the vector of queue lengths collapses to a lower dimensional subspace  $\mathcal{K}$
- $\mathbf{q}_{\parallel}$ : projection of vector of queue lengths on  $\mathcal{K}$
- Test function:  $V(\mathbf{q}) = \|\mathbf{q}_{\parallel}\|^2$
- Set its drift to zero:  $E[\Delta V(\mathbf{q})] = 0$

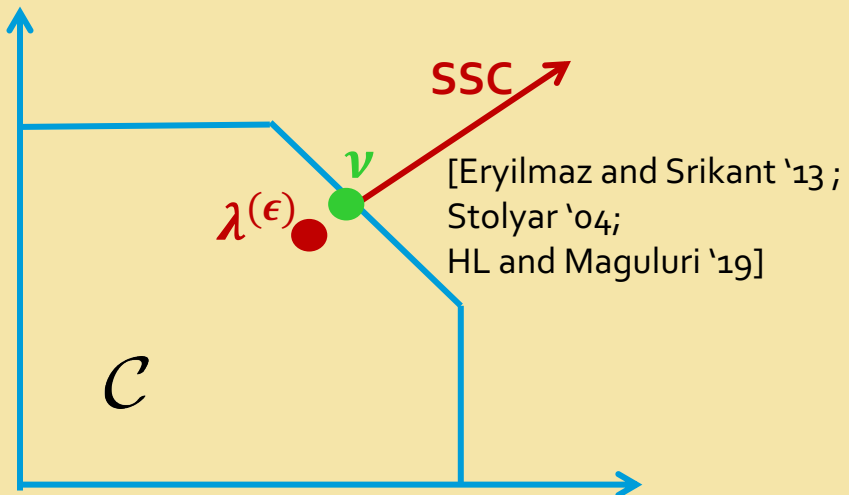
# STATE SPACE COLLAPSE

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



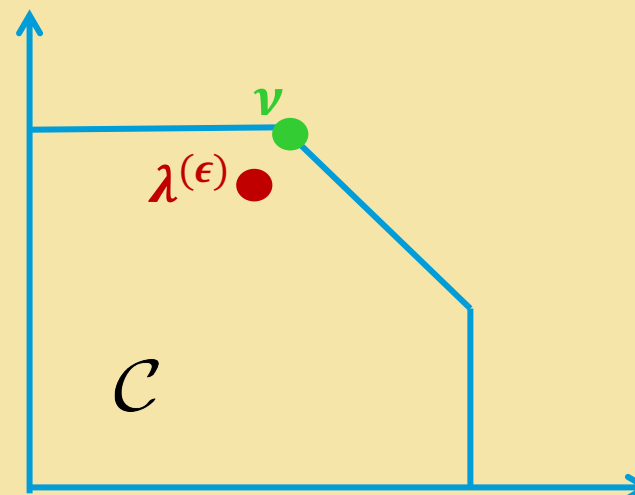
## COMPLETE RESOURCE POOLING (CRP)

- Approach the interior of a facet



## NO CRP

- Approach intersection of facets

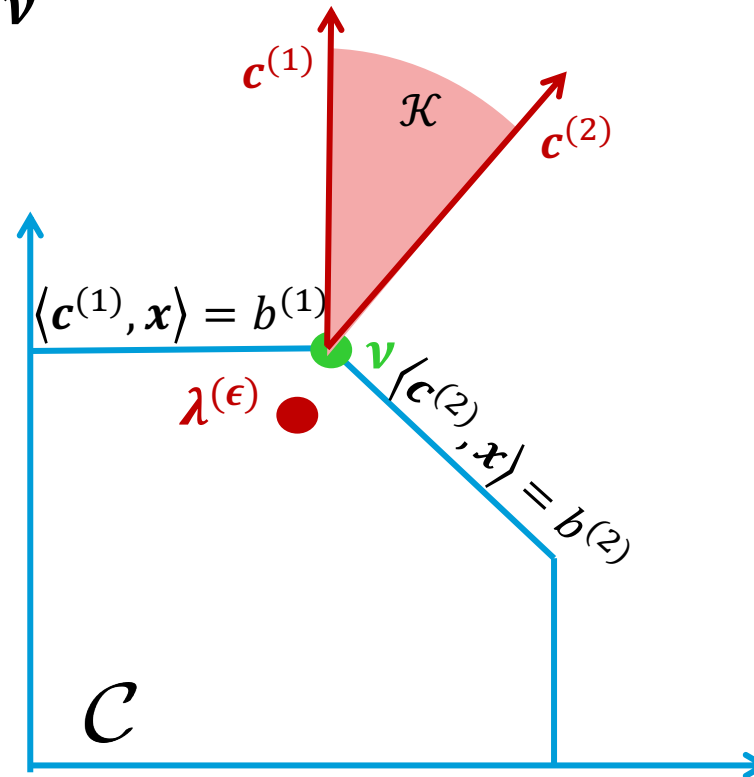


# STATE SPACE COLLAPSE (cont.)

- $P$  = set of indexes of the facets that intersect at  $\mathbf{v}$   
$$P = \{\ell \in [L]: \langle \mathbf{c}^{(\ell)}, \mathbf{v} \rangle = b^{(\ell)}\}$$
- Cone  $\mathcal{K}$  generated by  $\mathbf{c}^{(\ell)}$  for  $\ell \in P$

## Proposition: (SSC)

The vector of queue lengths  $\mathbf{q}$  collapses to the cone  $\mathcal{K}$  in heavy traffic.



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

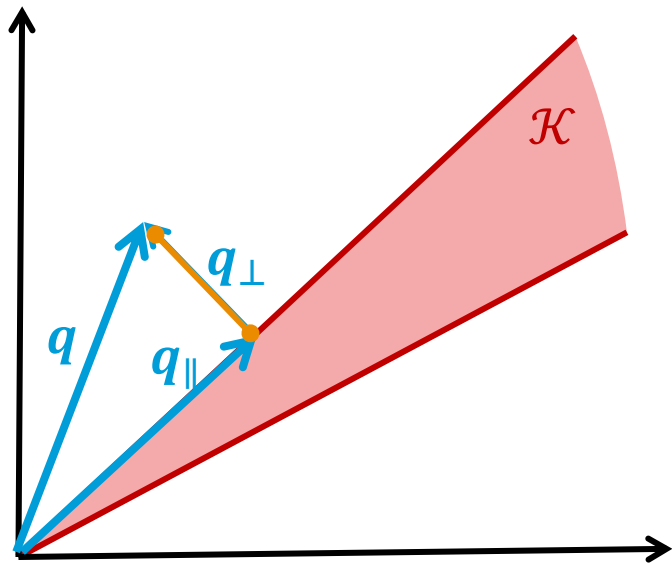
# SSC INTERPRETATION

## Proposition:

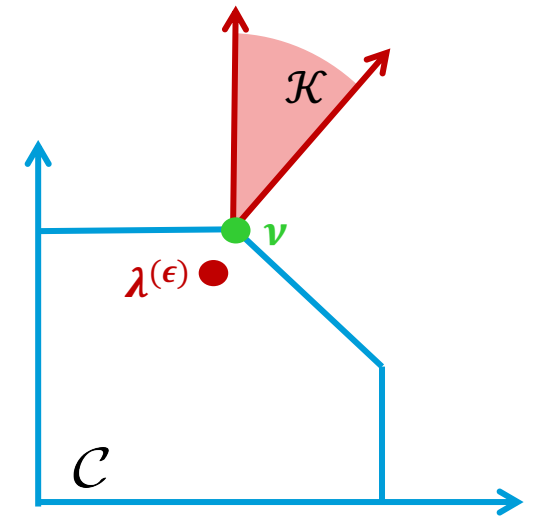
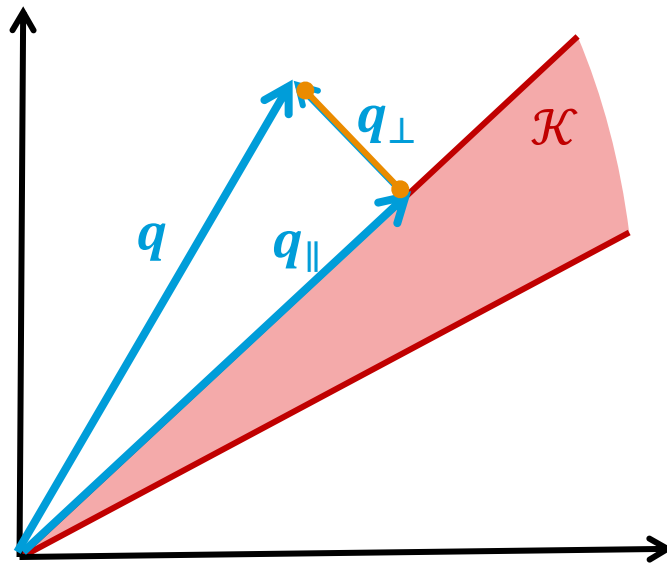
$\mathbf{q}_{\parallel}(k)$ : projection of  $\mathbf{q}(k)$  on  $\mathcal{K}$

$\mathbf{q}_{\perp}(k) := \mathbf{q}(k) - \mathbf{q}_{\parallel}(k)$ : error of approximating  $\mathbf{q} \approx \mathbf{q}_{\parallel}$

Then,  $E[\|\mathbf{q}_{\perp}\|^t] \leq T_t$  for all  $t = 1, 2, \dots$



Smaller  $\epsilon$ :



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

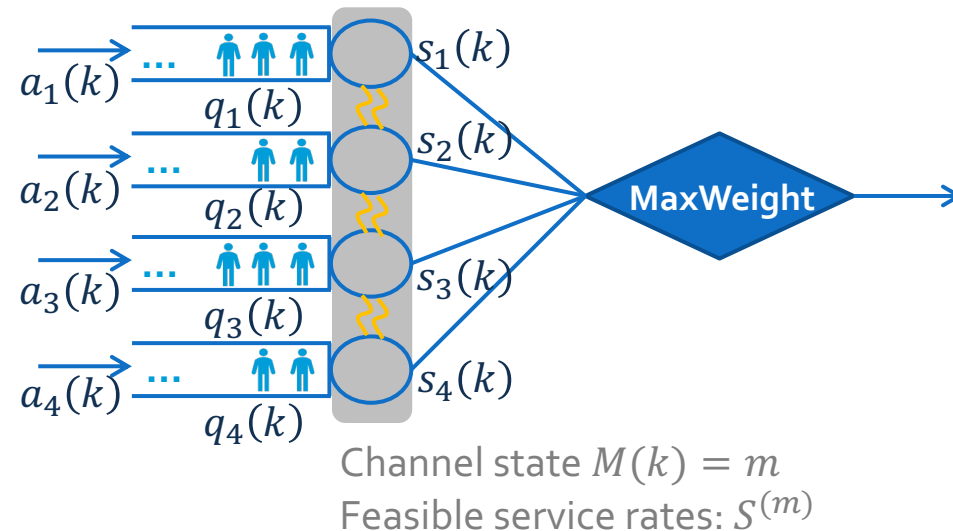
$q \approx q_{\parallel}$  as  $\epsilon \downarrow 0$

# IDEA OF THE PROOF

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

- We obtain

$$\begin{aligned}
 & \mathcal{J}_1 \xrightarrow{\epsilon \rightarrow 0} 2\epsilon E[\langle \mathbf{q}, \mathbf{v} \rangle] \\
 & \underbrace{2E[\langle \mathbf{q}_{\parallel}(k), \mathbf{s}_{\parallel}(k) - \mathbf{a}_{\parallel}(k) \rangle]}_{\mathcal{J}_2} \\
 = & E[\|\mathbf{a}_{\parallel}(k) - \mathbf{s}_{\parallel}(k)\|^2] - E[\|\mathbf{u}_{\parallel}(k)\|^2] + 2E[\langle \mathbf{q}_{\parallel}(k+1), \mathbf{u}_{\parallel}(k) \rangle] \\
 & \underbrace{\hspace{10em}}_{\mathcal{J}_3} \xrightarrow{\epsilon \rightarrow 0} 0 \\
 & \underbrace{\hspace{10em}}_{\mathcal{J}_4} \xrightarrow{\epsilon \rightarrow 0} 0 \quad (\text{using SSC})
 \end{aligned}$$

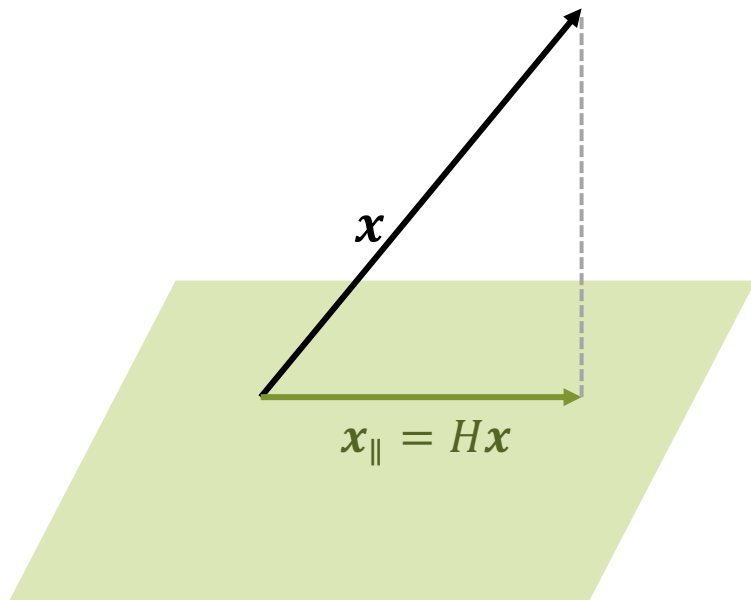


$$\begin{aligned}
 q_i(k+1) &= \max\{q_i(k) + a_i(k) - s_i(k), 0\} \\
 &= q_i(k) + a_i(k) - s_i(k) + u_i(k)
 \end{aligned}$$

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

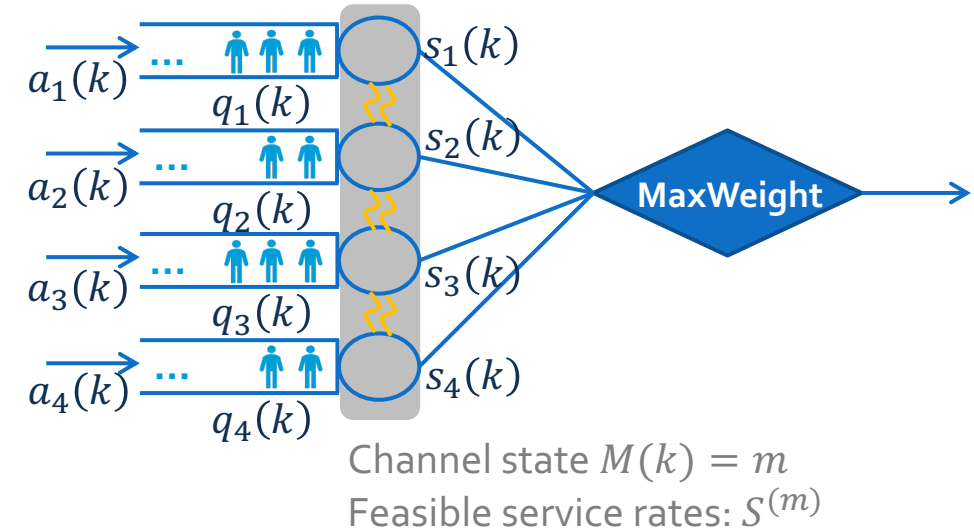
# IDEA OF THE PROOF (CONT.)

- $\mathcal{J}_2 = E \left[ \|\mathbf{a}_{\parallel}(k) - \mathbf{s}_{\parallel}(k)\|^2 \right]$
- Least squares problem:
  - $H$  = Projection matrix on subspace generated by  $\mathcal{K}$



Therefore,  
 $\mathcal{J}_2 \rightarrow RHS$

Subspace generated by  $\mathcal{K}$



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

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

$$q_i(k+1)u_i(k) = 0 \text{ but}$$

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

# DISCUSSION

- So

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

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

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

## CASE 1:

$\mathbf{w} = \mathbf{v} + \mathbf{y}$  with  $\mathbf{y} \parallel \mathbf{q}_\perp$



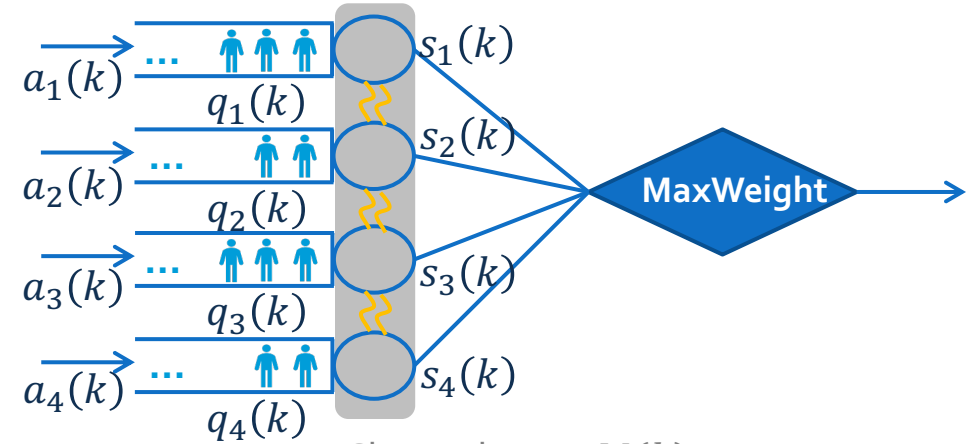
(by SSC)

## CASE 2:

Other  $\mathbf{w}$



Not with the Drift method!

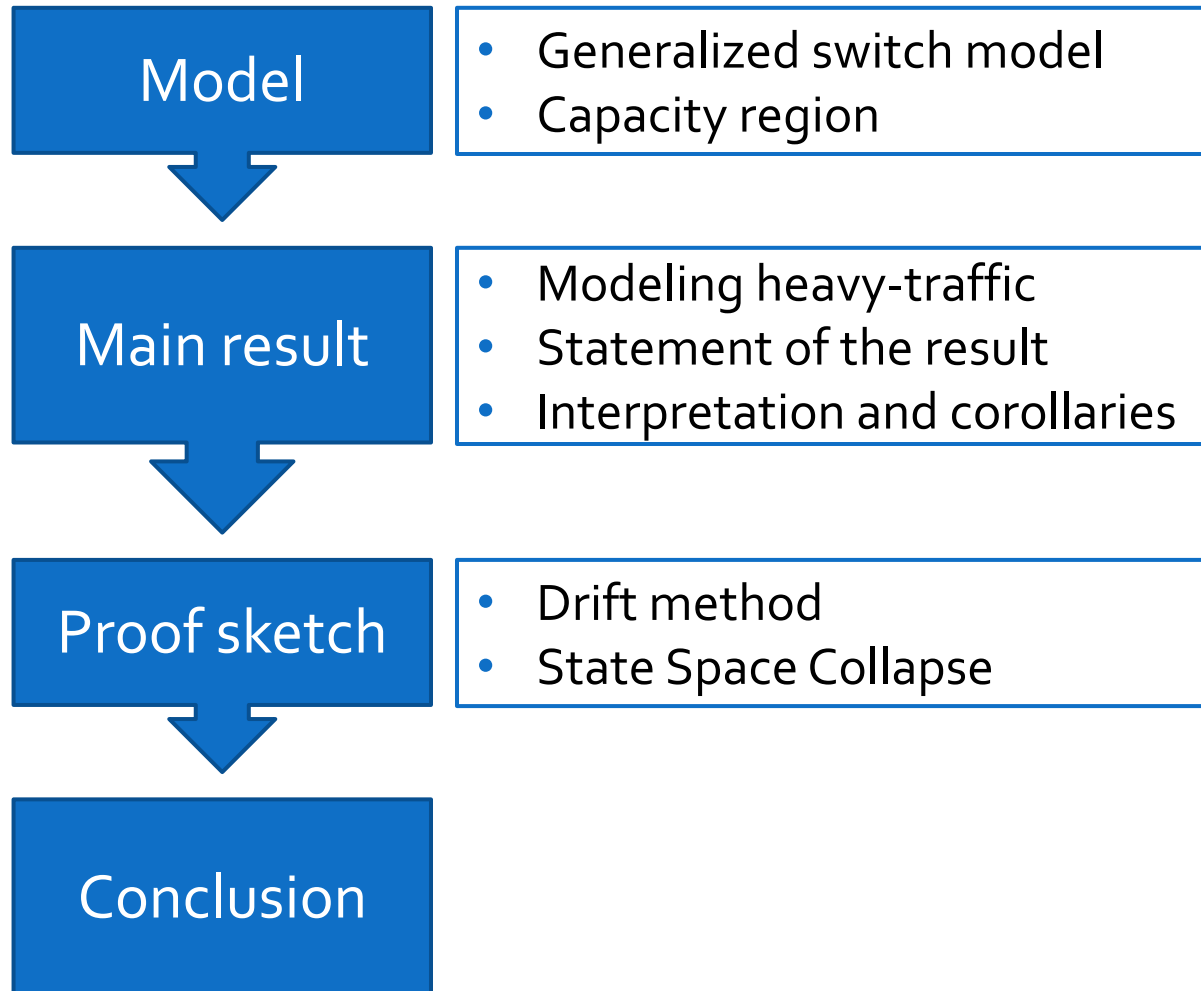


Channel state  $M(k) = m$   
Feasible service rates:  $S^{(m)}$

## Theorem:

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

# OUTLINE



# CONCLUSION

- Generalized switch
  - One of the most general single-hop queueing systems with control on the service process
- Heavy-traffic analysis using the Drift method
  - Input-queued switch with correlated arrivals
  - Full-dimensional SSC: Limit does not depend on correlation among arrivals
- Proof:
  - 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 WITH NO COMPLETE RESOURCE POOLING

---

**Daniela Hurtado Lange, Siva Theja Maguluri**

Industrial and Systems Engineering Department

Georgia Institute of Technology

July 14<sup>th</sup>, 2020

# NEW VIEW OF THE DRIFT METHOD [HL, Maguluri '19]

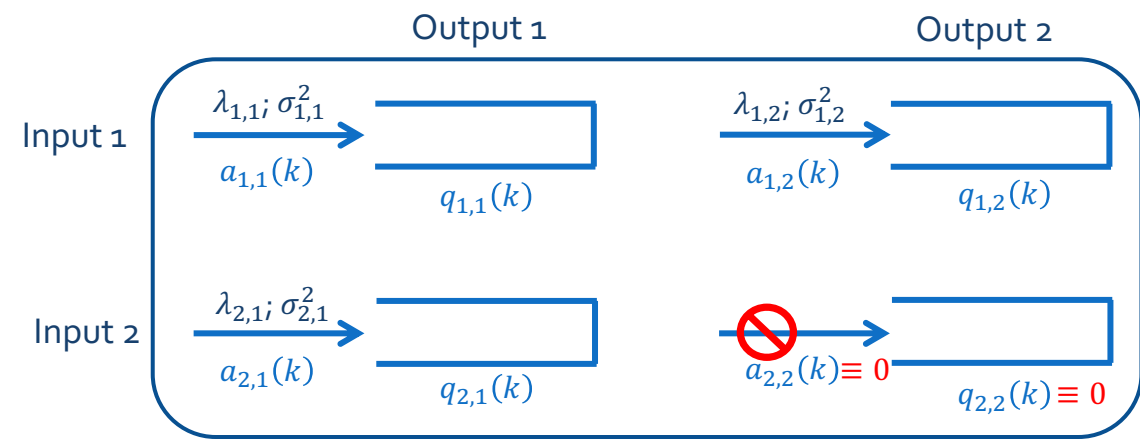
- Simplest no-CRP queueing system
  - 2x2 switch operating under MaxWeight
  - No arrivals to queue 2,2
  - Arrivals to other queues with mean  $\lambda_{i,j}$  and variance  $\sigma_{i,j}^2$

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

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

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

2-dimensional  
⇒ No-CRP



**Theorem** [Maguluri et.al. '18]:

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

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

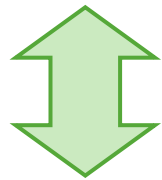
# OTHER LINEAR COMBINATIONS?

- SSC:

- $E[\|\mathbf{q}\|^2] \approx E[\|\mathbf{q}_{\parallel}\|^2]$
- $q_{\parallel 1,1} = q_{\parallel 1,2} + q_{\parallel 2,1}$

- Most general quadratic test function

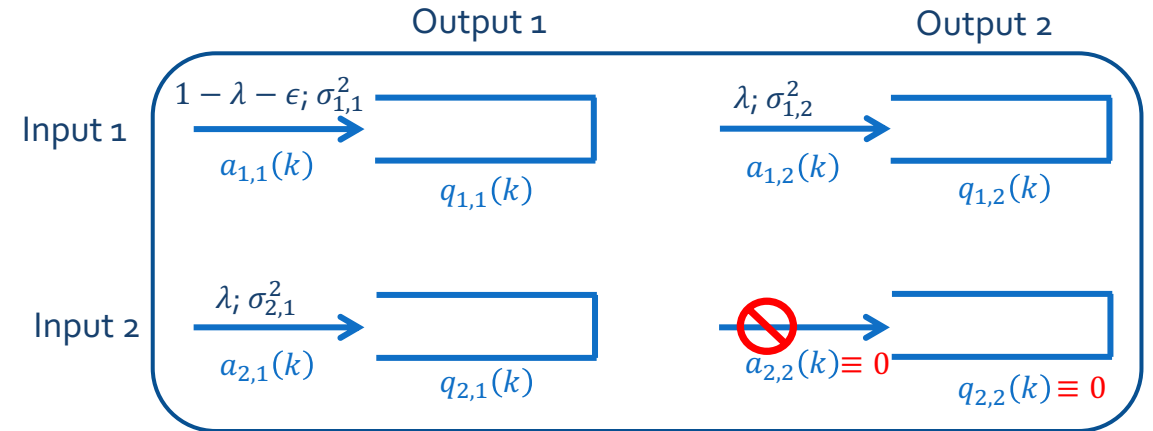
$$V(\mathbf{q}) = \alpha_1 q_{\parallel 1,2}^2 + \alpha_2 q_{\parallel 2,1}^2 + \alpha_3 q_{\parallel 1,2} q_{\parallel 2,1}$$



$$V_1(\mathbf{q}) = q_{\parallel 1,2}^2$$

$$V_2(\mathbf{q}) = q_{\parallel 2,1}^2$$

$$V_3(\mathbf{q}) = q_{\parallel 1,2} q_{\parallel 2,1}$$



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

# OTHER LINEAR COMBINATIONS? (cont.)

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

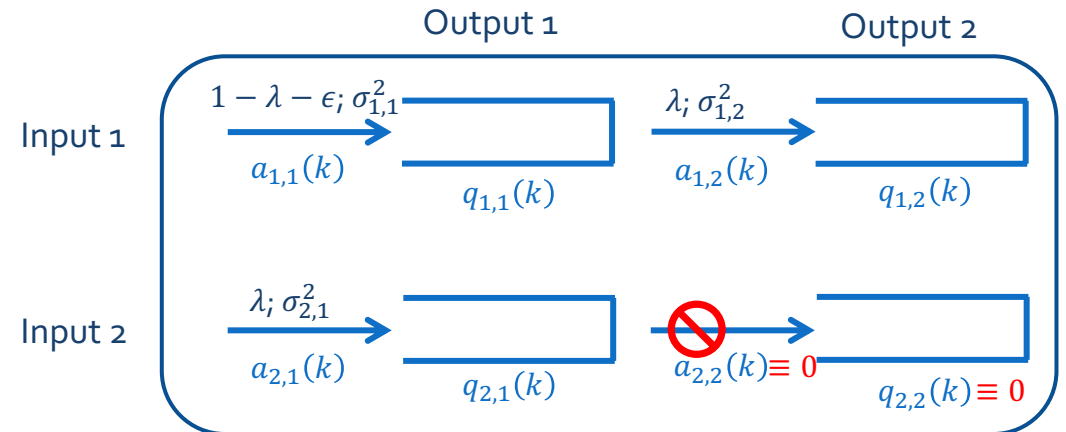
$$2 \lim_{\epsilon \downarrow 0} \epsilon E[q_{1,2}] = \frac{\sigma_{1,1}^2 + 4\sigma_{1,2}^2 + \sigma_{2,1}^2}{3} - 2 \lim_{\epsilon \downarrow 0} E[q_{1,2}^+ u_{2,1}]$$

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

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

4 unknowns  
3 equations !

Need more  
equations !



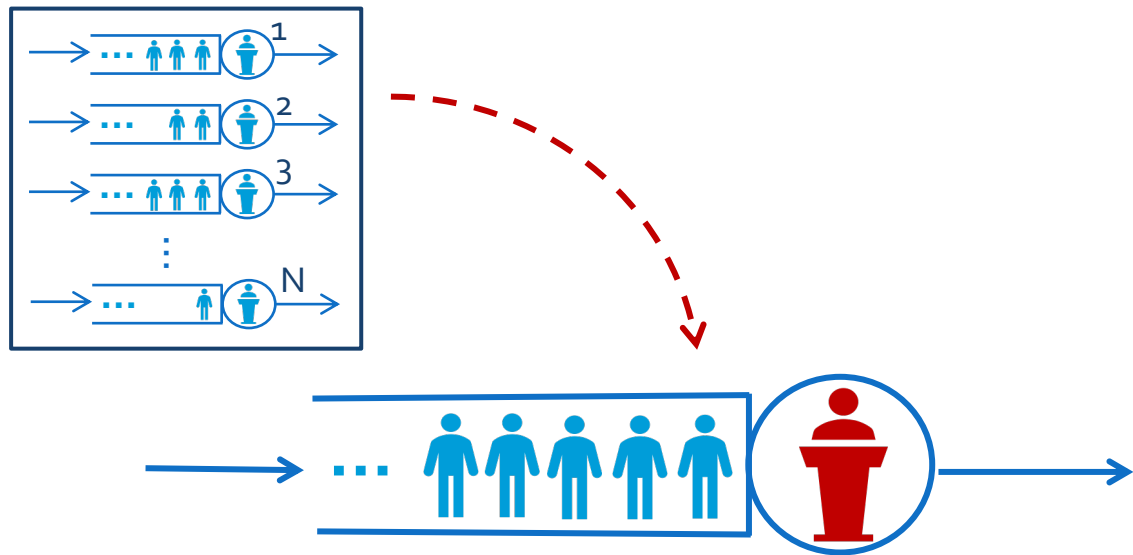
$$\mathcal{K} = \left\{ \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & 0 \end{bmatrix} : x_{1,1} = x_{1,2} + x_{2,1} \right\}$$

$$\begin{aligned} V_1(\mathbf{q}) &= q_{\parallel 1,2}^2 \\ V_2(\mathbf{q}) &= q_{\parallel 2,1}^2 \\ V_3(\mathbf{q}) &= q_{\parallel 1,2} q_{\parallel 2,1} \end{aligned}$$

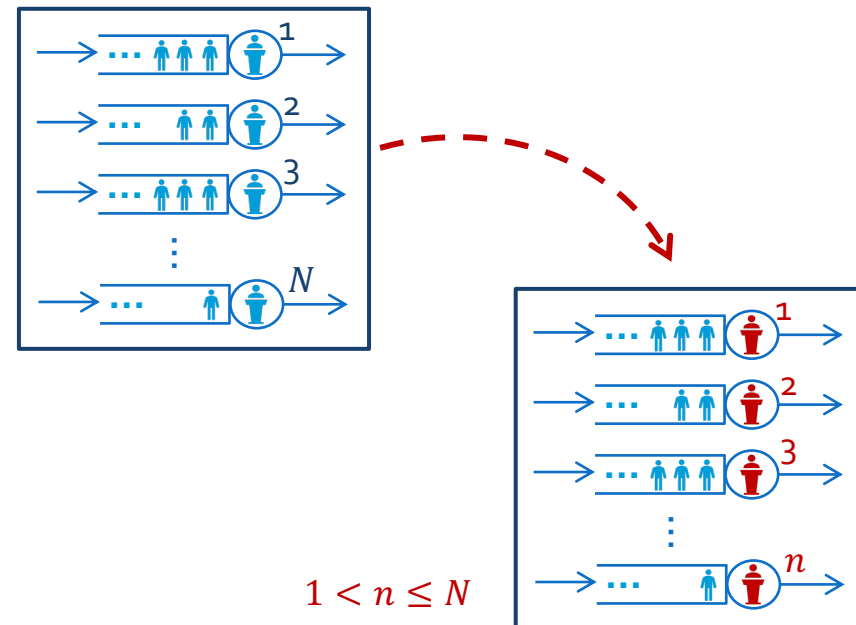
# STATE SPACE COLLAPSE (SSC)

- Queueing system behaves as a lower-dimensional system

## COMPLETE RESOURCE POOLING (CRP)



## NO CRP



$$1 < n \leq N$$