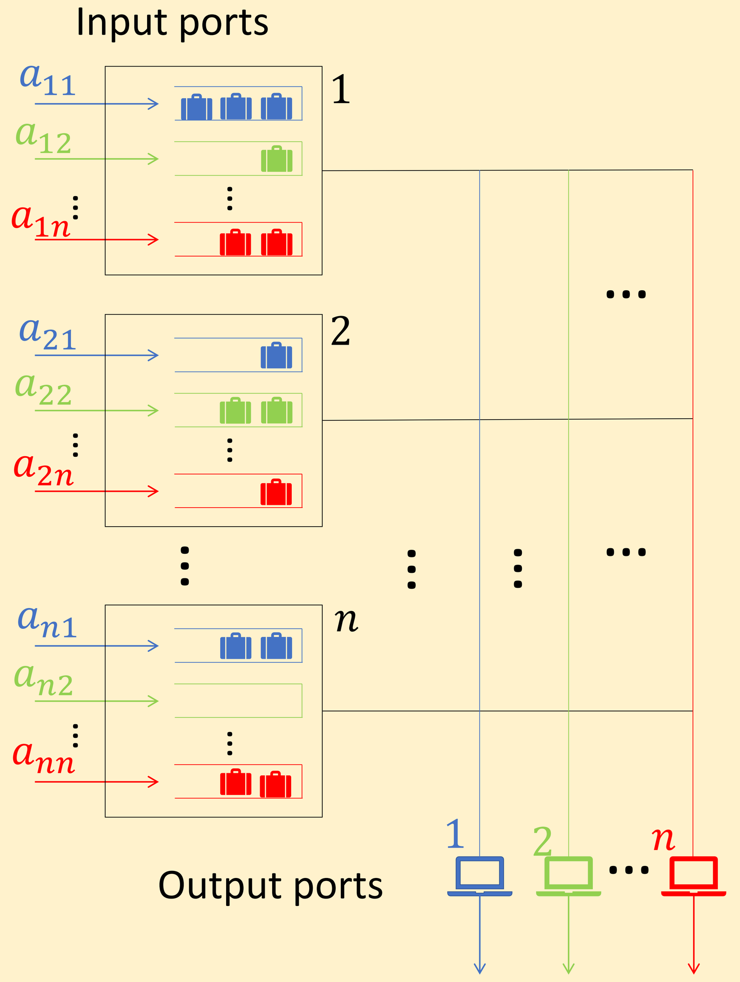
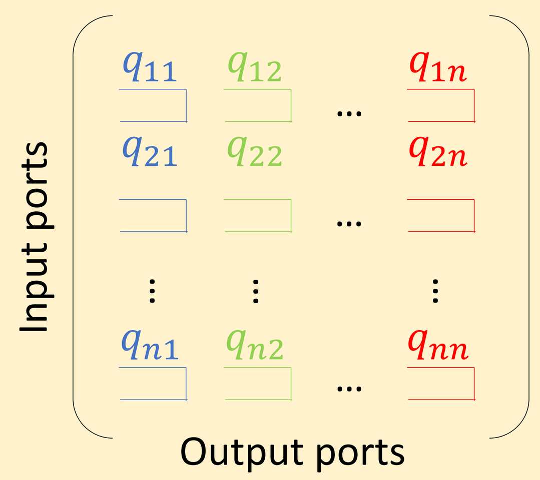
A Novel View of the Drift Method for Heavy-Traffic Limits of Queuing Systems

Daniela Hurtado, Siva Theja Maguluri
School of Industrial and Systems Engineering, Georgia Institute of Technology

Switch





There are n^2 queues but in each time slot we only can serve n that form a permutation
$\Longrightarrow$ Scheduling policy: MaxWeight

For input port i and output port j $(i,j \in \{1,\dots n\})$ let:
$q_{i,j}(k)$: # of packets in system at the beginning of time slot k
$a_{i,j}(k)$: # of arrivals in time slot k
$s_{i,j}(k)$: # of offered services in time slot k
$u_{i,j}(k)$: Unused service in time slot k

$\boldsymbol{q}(k)$: Vector with elements $q_{i,j}(k)$ for $i,j \in \{1,\dots n\}$

$$q_{i,j}(k+1) = \left[q_{i,j}(k) + a_{i,j}(k) - s_{i,j}(k)\right]^+ = q_{i,j}(k) + a_{i,j}(k) - s_{i,j}(k) + u_{i,j}(k)$$

$$\Longrightarrow q_{i,j}(k+1)u_{i,j}(k) = 0$$

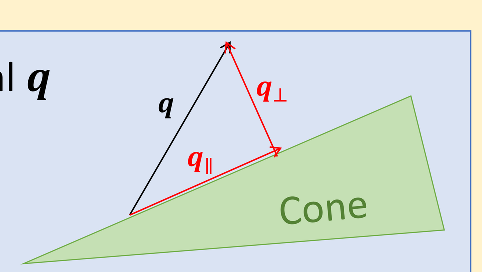
Capacity region:
Let $\lambda_{i,j} = E[a_{i,j}(1)]$

$$\mathcal{C} = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{n^2} : \sum_{i=1}^{n} \lambda_{i,j} \le 1\ \forall j, \sum_{j=1}^{n} \lambda_{i,j} \le 1\ \forall i\}$$

Heavy traffic (HT): Let $\boldsymbol{\lambda}$ be such that $\sum_j \lambda_{i,j} = 1-\epsilon$ and $\sum_i \lambda_{i,j} = 1-\epsilon$. The limit of the system as $\epsilon \to 0$ is called HT.

HT limit: Queue lengths go to ∞ as ϵ
To understand behavior, study: $\lim_{\epsilon\to 0} \epsilon\ E\left[\sum q_{i,j}\right]$

State space collapse (SSC): In the HT limit, the n^2-dimensional $\boldsymbol{q}$ collapses into a $(2n-1)$-dimensional cone. Let $\boldsymbol{q}_{\parallel}$ be the projection of $\boldsymbol{q}$ on the cone and $\boldsymbol{q}_{\perp}$ is the error $\boldsymbol{q}_{\perp} = \boldsymbol{q} - \boldsymbol{q}_{\parallel}$. Then, for each $r = 1,2,\dots$ $E[\|\overline{\boldsymbol{q}}_{\perp}\|^r] \le M_r$



- Diffusion limits
- RBM on a cone

Theorem: Let $\boldsymbol{\sigma}^2$ be the vector of variances of the arrival process. Then,

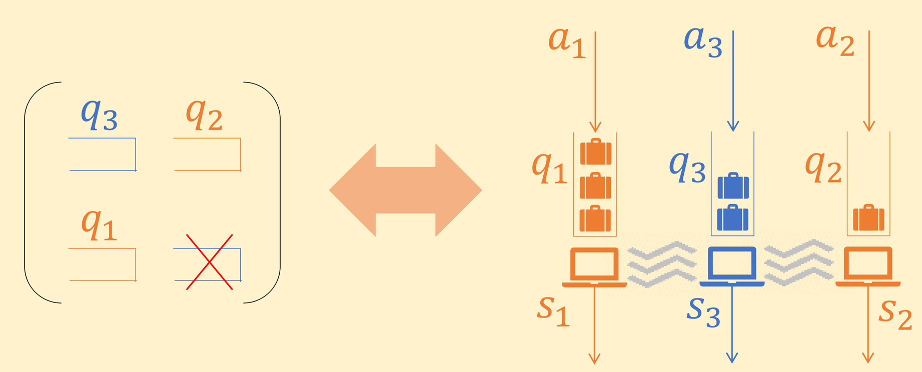
$$\lim_{\epsilon\to 0} \epsilon\ E\left[\sum_{i,j} \bar{q}_{i,j}\right] = \lim_{\epsilon\to 0} \epsilon\ E\left[\sum_{i,j} \bar{q}_{\parallel\, i,j}\right] = \left(1 - \frac{1}{2n}\right)\sum_{i,j} \sigma_{i,j}^2$$

CHALLENGE!

Joint distribution of $\epsilon \boldsymbol{q}$ in steady state ?

Stationary distribution of RBM in a cone ?

Three-queue System



HT:
$\boldsymbol{\lambda} = (\lambda, \lambda, 1-\epsilon-\lambda)$, where $0 < \lambda < 1$

SSC:
$\boldsymbol{q}$ collapses into the 2-dimensional space defined by:
$q_{\parallel 3} = q_{\parallel 1} + q_{\parallel 2}$

Theorem: Let $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ be the vector of variances of the arrival process. Then,

$$\lim_{\epsilon\to 0} \epsilon\ E[q_1+q_2+q_3] = \lim_{\epsilon\to 0} \epsilon\ E[q_{\parallel 1}+q_{\parallel 2}+q_{\parallel 3}] = \frac{2}{3}(\sigma_1^2+\sigma_2^2+\sigma_3^2)$$

Proof Idea:

Set $E[\Delta V(\overline{\boldsymbol{q}})] = 0$, where

$$V(\boldsymbol{q}) = \frac{1}{2}\|\boldsymbol{q}_{\parallel}\|^2 = q_{\parallel 1}^2 + q_{\parallel 2}^2 + q_{\parallel 1}q_{\parallel 2}$$

Goal: All moments of all linear combinations of q_1, q_2, q_3

SSC → $E[\|\overline{\boldsymbol{q}}\|^m] \approx E\left[\|\overline{\boldsymbol{q}}_{\parallel}\|^m\right]$ & $q_{\parallel 3} = q_{\parallel 1} + q_{\parallel 2}$ → Only need linear combinations of moments of: $q_{\parallel 1}, q_{\parallel 2}$

Here we focus on the first moments: $E[\theta_1 q_{\parallel 1} + \theta_2 q_{\parallel 2}] =$ **?**

General quadratic test function: $V(\boldsymbol{q}) = \alpha_1 q_{\parallel 1}^2 + \alpha_2 q_{\parallel 2}^2 + \alpha_3 q_{\parallel 1} q_{\parallel 2}$ ⟷ All quadratic monomials: $V_1(\boldsymbol{q}) = q_{\parallel 1}^2$, $V_2(\boldsymbol{q}) = q_{\parallel 2}^2$, $V_3(\boldsymbol{q}) = q_{\parallel 1}q_{\parallel 2}$

Set $E[\Delta V(\overline{\boldsymbol{q}})] = 0$ or $E[\Delta V_i(\overline{\boldsymbol{q}})] = 0\ \forall i \in \{1,2,3\}$

$$2\lim_{\epsilon\to 0} \epsilon\ E[q_1] = \frac{1}{3}(4\sigma_1^2+\sigma_2^2+\sigma_3^2) - 2\lim_{\epsilon\to 0} E[q_1^+ u_2]$$

$$2\lim_{\epsilon\to 0} \epsilon\ E[q_2] = \frac{1}{3}(\sigma_1^2+4\sigma_2^2+\sigma_3^2) - 2\lim_{\epsilon\to 0} E[q_2^+ u_1]$$

$$\lim_{\epsilon\to 0} \epsilon\ E[q_1+q_2] = \frac{1}{3}(-2\sigma_1^2-2\sigma_2^2+\sigma_3^2) + 2\lim_{\epsilon\to 0} E[q_1^+ u_2 + q_2^+ u_1]$$

Solve for: $\lim_{\epsilon\to 0} \epsilon\ E[q_1]$, $\lim_{\epsilon\to 0} \epsilon\ E[q_2]$, $\lim_{\epsilon\to 0} E[q_1^+ u_2]$, $\lim_{\epsilon\to 0} E[q_2^+ u_1]$ → 3 equations 4 unknowns **!** → Need more equations **!**

References:
[1] W.N. Kang and R.J. Williams. 2012. Diffusion approximation for an input-queued packet switch operating under maximum weight algorithm. *Stochastic systems* (2012).
[2] S. Kumar and P.R. Kumar. 1994. Performance bounds for queuing networks and scheduling policies. *IEE Trans. Automat. Control* 39,8 (Aug 1994), 1600-1611. https://doi.org/10.1109/9.310033
[3] S.T. Maguluri, S.K. Burle, and R. Srikant, 2018. Optimal heavy-traffic queue length scaling in an incompletely saturated switch. *Queueing systems* 88, 3-4 (2018), 279-309.