Heavy-Traffic Limits for Stationary Network Flows

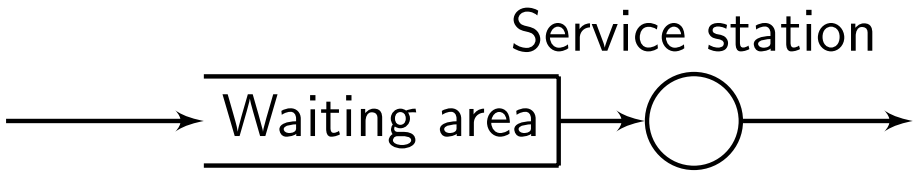
Wei You
(Joint work with Ward Whitt)

Columbia University and HKUST

Chinese Academy of Sciences, Beijing

Dec. 17, 2019

Single-server queue

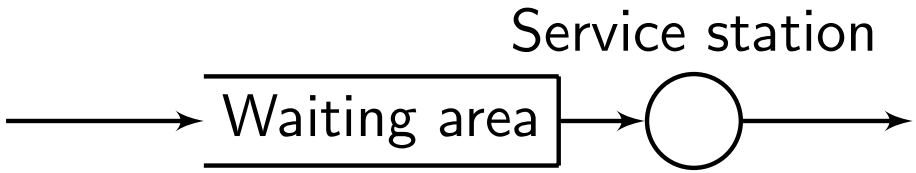


Kendall's notation

GI/GI/1 queue

- "GI" means *independent* and identically distributed with *general* distributions.
- Interarrival times are GI.
- Service time are GI.
- One server in the service station.

Single-server queue

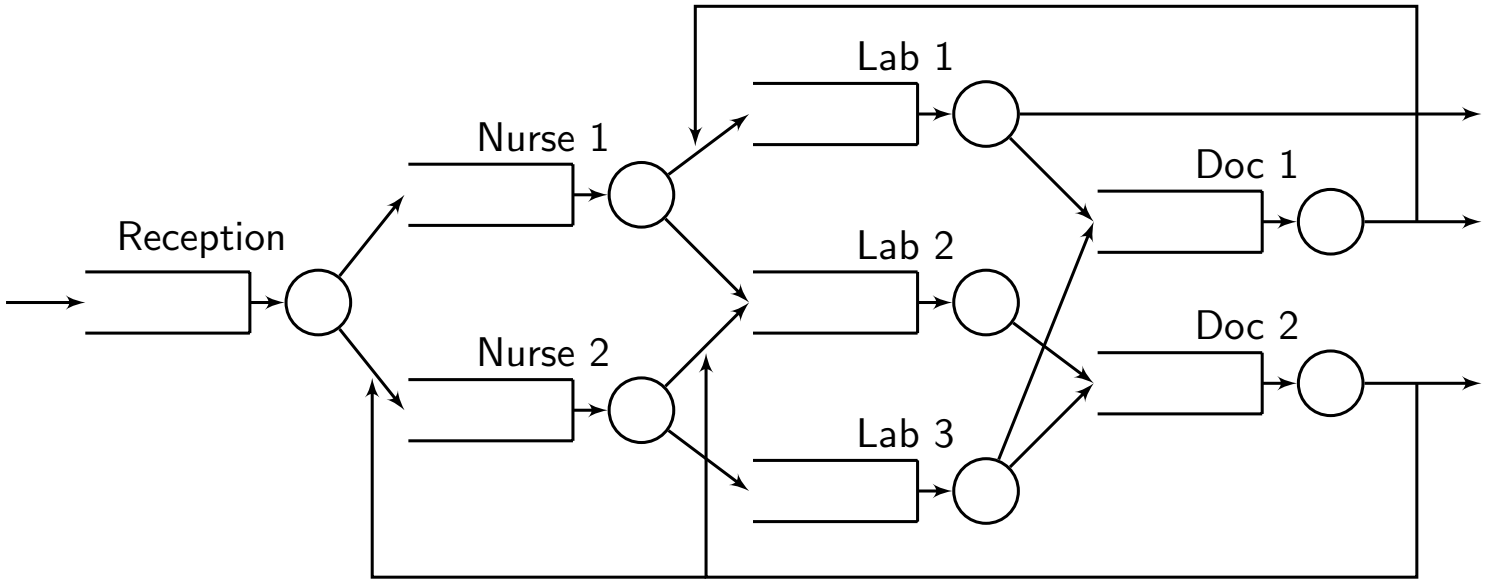


Kendall's notation

GI/GI/1 queue

- "GI" means *independent* and identically distributed with *general* distributions.
- Interarrival times are GI.
- Service time are GI.
- One server in the service station.

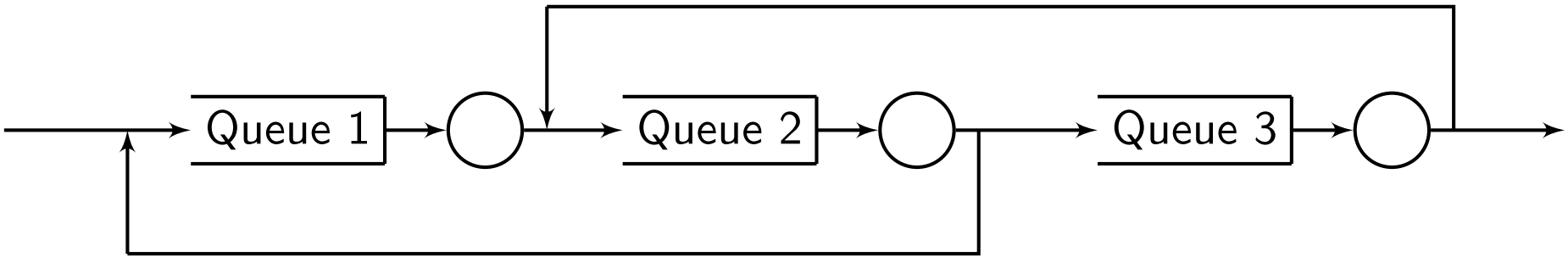
Open Queueing Networks



Service systems

- cloud computing networks
- ride-sharing platforms
- manufacturing lines
- call centers

Motivation

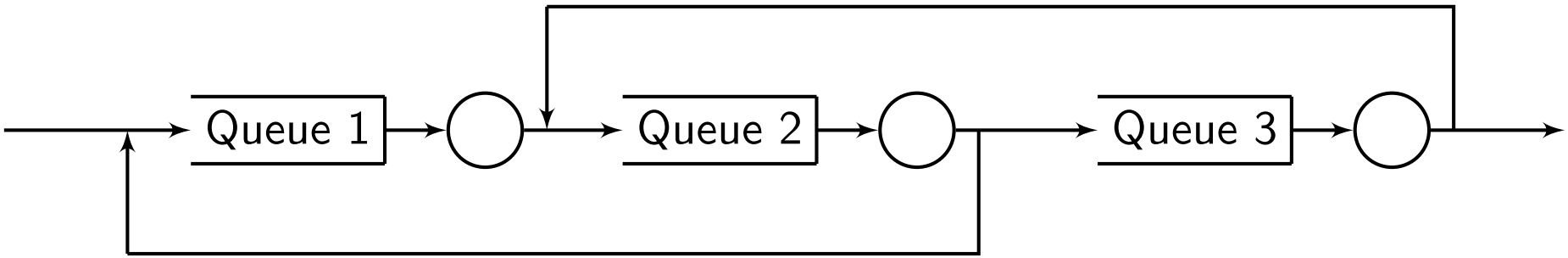


In service operations management, a decision maker cares about the performance measures

- system throughput $\Rightarrow$ profit
- queue length $\Rightarrow$ holding cost of the jobs
- waiting time, delay probability $\Rightarrow$ customer satisfaction
- system workload $\Rightarrow$ maintenance cost incurred by surges of workload

An accurate characterization of the performance measure is essential in any queueing-related applications.

Motivation



In service operations management, a decision maker cares about the performance measures

- system throughput $\Rightarrow$ profit
- queue length $\Rightarrow$ holding cost of the jobs
- waiting time, delay probability $\Rightarrow$ customer satisfaction
- system workload $\Rightarrow$ maintenance cost incurred by surges of workload

An accurate characterization of the performance measure is essential in any queueing-related applications.

Open Queueing Network

Jackson Networks

- Network of M/M/1 queues with Markovian routing

"M" = "Memoryless" = exponential distributions.

- Admits close-form formula for the steady-state performance measures.
  - The steady-state queue length vector have product-form distribution with exponential marginal distributions.

However, realistic models deviate significantly from the tractable structure of the Jackson networks.

- Studying the customer arrival flows turned out to be quite useful in determining/approximating the performance measures.

Performance Measure and the Customer Arrival Flow

- Analyzing the customer arrival flow can be very useful in determining/approximating the performance measures.

As an illustration, the steady-state mean workload in the classical $GI/M/1$ model is

$$E[Z] = \frac{\rho}{\mu(1-\sigma)},$$

where σ is the unique root in $(0,1)$ of the equation

$$\hat{f}(\mu(1-\sigma)) = \sigma,$$

where $\hat{f}$ is the Laplace transform $\hat{f}(s) = \int_0^\infty e^{-st} f(t)dt$ of the interarrival-time pdf f.

Performance Measure and the Customer Arrival Flow

The mean steady-state workload can be re-written as

$$E[Z] = \rho/s^*,$$

where s^* is the unique root in $(0, \rho^{-1})$ of $\hat{V}(s) = \frac{2(1-\rho)}{\rho s^3} - \frac{1}{s^2}$, and $\hat{V}$ is the Laplace transform of the variance function $V(t) = Var(A(t))$ of the arrival flow.

Theorem (Ordering of the mean steady-state workload)

Consider two $GI/M/1$ queues with rate-1 arrival processes A_1 and A_2 and mean service time ρ. If

$$V_1(t) \geq V_2(t), \quad \textit{for all} \quad t \geq 0,$$

then the steady-state workload satisfies

$$E[Z_{1,\rho}] \geq E[Z_{2,\rho}], \quad \textit{for all} \quad \rho \in (0, 1).$$

Performance Measure and the Customer Flows

Going beyond $GI/M/1$ or $M/GI/1$, closed-form solution of the system performance are rarely available.

The customer arrival flow, especially the variance function of it, plays an decisive role.

This inspires us to explore general **customer flows**

- the departure flows,
- the internal arrival flows from one queue to another,
- the total arrival flows.

Performance Measure and the Customer Flows

Going beyond $GI/M/1$ or $M/GI/1$, closed-form solution of the system performance are rarely available.

The customer arrival flow, especially the variance function of it, plays an decisive role.

This inspires us to explore general **customer flows**

- the departure flows,
- the internal arrival flows from one queue to another,
- the total arrival flows.

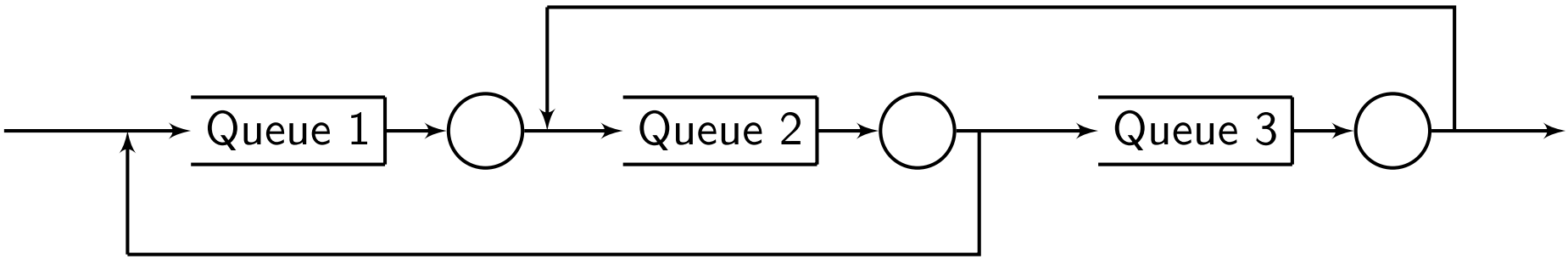
Performance Measure and the Customer Flows

Going beyond $GI/M/1$ or $M/GI/1$, closed-form solution of the system performance are rarely available.

The customer arrival flow, especially the variance function of it, plays an decisive role.

This inspires us to explore general **customer flows**

- the departure flows,
- the internal arrival flows from one queue to another,
- the total arrival flows.



Performance Measure and the Customer Flows

Flows can be useful.

- Whitt and You (2018) proposed new **Robust Queueing** algorithm that rely on the **variance function of the flows** to approximate the steady-state performance in OQN.

However, flows can be complicated.

- To make things worse, the RQ algorithm relies on the **stationary** version of the customer flows.

How do we approximate the stationary flows in open queueing networks?

Performance Measure and the Customer Flows

Flows can be useful.

- Whitt and You (2018) proposed new **Robust Queueing** algorithm that rely on the **variance function of the flows** to approximate the steady-state performance in OQN.

However, flows can be complicated.

- To make things worse, the RQ algorithm relies on the **stationary** version of the customer flows.

How do we approximate the stationary flows in open queueing networks?

Heavy-traffic Approximations

$$Q_\rho(t) \xrightarrow{\lim_{\rho\to 1}(1-\rho)Q_\rho((1-\rho)^{-2}t)} Z(t)$$

Heavy-traffic limits of open queueing network (OQN)

- A major source of approximation.
- Feed-forward networks:
  - Iglehart and Whitt (1970a,b); Harrison (1973, 1978).
- Open queueing networks:
  - Reiman (1984); Chen and Mandelbaum (1991a,b).
- Reflected Brownian motion (RBM):
  - Harrison (1978); Harrison and Reiman (1981); Dai and Harrison (1992).

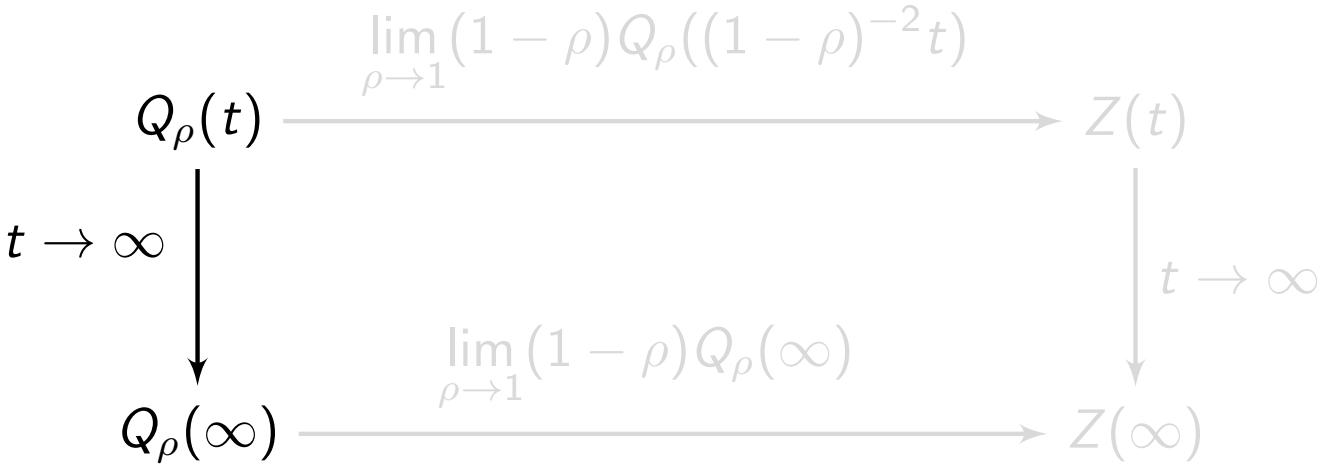
Heavy-traffic Approximations

$$Q_\rho(t) \xrightarrow{\lim_{\rho\to 1}(1-\rho)Q_\rho((1-\rho)^{-2}t)} Z(t)$$

Heavy-traffic limits of open queueing network (OQN)

- A major source of approximation.
- Feed-forward networks:
  - Iglehart and Whitt (1970a,b); Harrison (1973, 1978).
- Open queueing networks:
  - Reiman (1984); Chen and Mandelbaum (1991a,b).
- Reflected Brownian motion (RBM):
  - Harrison (1978); Harrison and Reiman (1981); Dai and Harrison (1992).

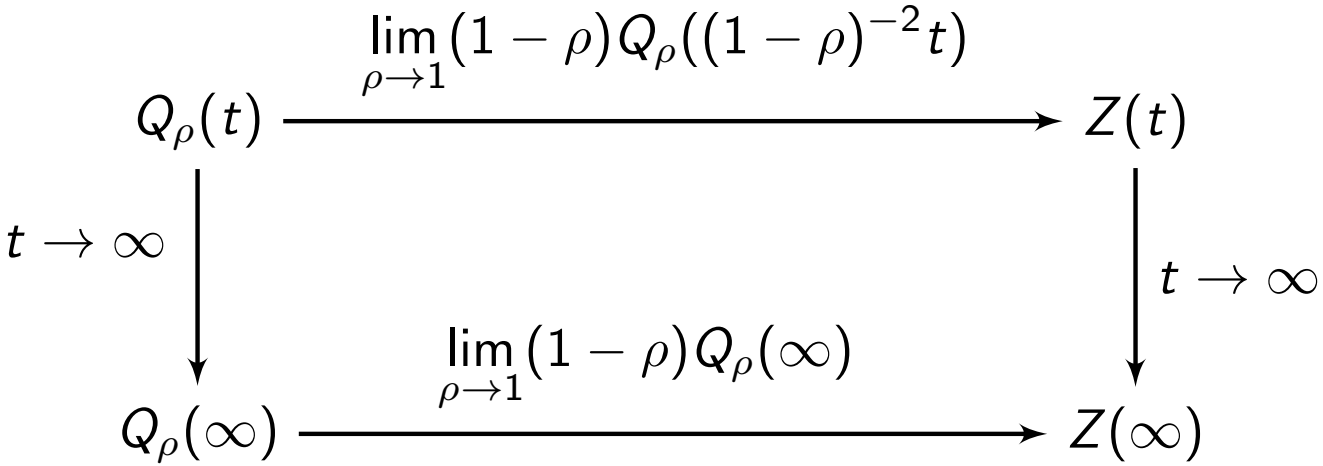
Motivation



Heavy-traffic approximation of the **steady-state queue length** in OQN

- Interchange of limits:
  - Gamarnik and Zeevi (2006); Budhiraja and Lee (2009); Braverman et al. (2017).

Motivation



Heavy-traffic approximation of the **steady-state queue length** in OQN

- Interchange of limits:

  - Gamarnik and Zeevi (2006); Budhiraja and Lee (2009); Braverman et al. (2017).

Motivation

So far, the heavy-traffic literature has focused on the queue length, busy time, waiting time and workload processes.

Little is known regarding the HT limit of the customer **flows**.

In This Talk

- We establish the existence of unique stationary flows in generalized Jackson networks and the convergence to it as time increases.
- We establish heavy-traffc limits for the stationary flows, allowing an arbitrary subset of the queues to be critically loaded.
- We demonstrate the approximation of the flows with numerical examples.

The OQN Model

Consider a queueing network with K single-server stations with unlimited waiting space and the first-come first-served (FCFS) discipline.

- $A_{0,i}(t)$: external arrival point process at station i.
- $S_i(t)$: (uninterrupted) service point (counting) process

$$S_i(t) = \max_{n \geq 0} \left\{ \sum_{l=1}^{n} V_i^l \leq t \right\}, \quad t \geq 0,$$

  where $\{V_i^l : l \geq 1\}$ is the sequence of service times at station i.
- $D_i(t)$: departure flow from station i
  - $D_i(t) = S_i(B_i(t))$, where $B_i(t)$ is the cumulative busy time of server i up to time t.

The OQN Model

Consider a queueing network with K single-server stations with unlimited waiting space and the first-come first-served (FCFS) discipline.

- $A_{0,i}(t)$: external arrival point process at station i.
- $S_i(t)$: (uninterrupted) service point (counting) process

$$S_i(t) = \max_{n \geq 0} \left\{ \sum_{l=1}^{n} V_i^l \leq t \right\}, \quad t \geq 0,$$

  where $\{V_i^l : l \geq 1\}$ is the sequence of service times at station i.
- $D_i(t)$: departure flow from station i
  - $D_i(t) = S_i(B_i(t))$, where $B_i(t)$ is the cumulative busy time of server i up to time t.

The OQN Model

Consider a queueing network with K single-server stations with unlimited waiting space and the first-come first-served (FCFS) discipline.

- $A_{0,i}(t)$: external arrival point process at station i.
- $S_i(t)$: (uninterrupted) service point (counting) process

$$S_i(t) = \max_{n \geq 0} \left\{ \sum_{l=1}^{n} V_i^l \leq t \right\}, \quad t \geq 0,$$

  where $\{V_i^l : l \geq 1\}$ is the sequence of service times at station i.
- $D_i(t)$: departure flow from station i
  - $D_i(t) = S_i(B_i(t))$, where $B_i(t)$ is the cumulative busy time of server i up to time t.

The OQN Model

- $\Theta_{i,j}(n)$: number of customers routed to j from i among the first n-th departure from queue i.
- $A_{i,j}$: internal arrival flows

$$A_{i,j}(t) = \Theta_{i,j}(D_i(t)).$$

- $A_i(t)$: total arrival process at station i

$$A_i(t) = A_{0,i}(t) + \sum_{j=1}^{K} A_{j,i}(t)$$

- $Q_i(t)$: queue length process

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t))$$

Assumption 1

Assumption (Generalized Jackson Network)

1. *The external arrival process at station i is a renewal processes with finite rates λ_i. The interarrival times have finite squared coefficient of variation (scv) $c^2_{a_{0,i}}$.*
2. *The service times are i.i.d. random variables with means $1/\mu_i$ and finite scv $c^2_{s_i}$.*
3. *The routing is Markovian with routing matrix $P = (p_{i,j})_{1 \le i,j \le K}$ such that $p_{i,j} \ge 0$, $p_{i,0} \equiv 1 - \sum_{j=1}^{K} p_{i,j} \ge 0$ and $I - P'$ is invertible;.*
4. *The arrival, service and routing processes are mutually independent.*

Markov Representation

- Let $U(t)$ denote the vector of residual external arrival times at time t.
- Let $V(t)$ be the vector of residual service times at time t, set to 0 when the server is idle.
- Let the **system-state process** be

$$\mathcal{S}(t) \equiv (Q(t), U(t), V(t)), \quad t \geq 0.$$

One can show with Davis (1984) that

Theorem

For GJN, the system state process $\mathcal{S}$ is a strong Markov process.

- Because $\mathcal{S}$ is a piece-wise deterministic Markov process.

Assumption 2

- λ: effective arrival rate

$$\lambda = (I - P')^{-1}\lambda_0,$$

  where λ_0 is the external arrival rate.
- $\rho_i = \lambda_i/\mu_i$: traffic intensity

Assumption

The traffic intensities satisfy $\max_i \rho_i < 1$.

Assumption 3

We make the key assumption to obtain the Harris recurrence as in Sigman (1988, 1990), Dai (1995) and Ch. VII of Asmussen (2003).

Assumption

Each external interarrival-time distribution is **unbounded above** *and* **spread out**.

- **Spread out**: for a distribution F, there exist a integer $i > 0$ such that the i-fold convolution F^{*i} has an absolutely continuous component (has a density) with respect to the Lebesgue measure.
  - In applied context, spread out is practically the same as non-lattice.

Existence, Uniqueness and Convergence

Theorem (System-state process)

Under Assumptions 1-3, the system state stochastic process $\mathcal{S}$ is a positive Harris recurrent Markov process. There exists a unique stationary distribution π and for every initial condition and the distribution of $\mathcal{S}(t)$ converges to π as $t \to \infty$.

- Theorem 2 of Gamarnik and Zeevi (2006) or Theorem 5.1 of Dai (1995) or Theorem 6.2 of Dai and Meyn (1995), which extend earlier work on stability for OQNs in Borovkov (1986), Sigman (1990) and Foss (1991).

Existence, Uniqueness and Convergence

For the system state processes,

- let $Q_s(t) = Q(s+t), U_s(t) = U(s+t)$ and $V_s(t) = V(s+t)$;
- so that $\mathcal{S}_s \equiv (Q_s, U_s, V_s)$ is the system state process start at time t.

Corollary (Weak convergence of the system state process)

Under Assumptions 1-3, there exist stationary processes (Q_e, U_e, V_e) *such that*

$$\mathcal{S}_s \Rightarrow \mathcal{S}_e \equiv (Q_e, U_e, V_e), \quad as \quad s \to \infty,$$

where $\Rightarrow$ *denote weak convergence.*

- Proved using a generalized version of Theorem 12.6 in Billingsley (1999): the system-state process live in a subset of $\mathcal{D}^3$ with nice sample path, where convergence on a countable dense subset of time implies convergence in SJ_1 topology.
- Convergence of the finite-dimensional distribution via strong Markov property.

Existence, Uniqueness and Convergence

Now, we turn to the stationary flows.

- Define the customer flows $\mathcal{F}$ by

$$\mathcal{F}(t) \equiv (A_0(t), S(t), A_{\text{int}}(t), A(t), D(t)) .$$

  - We use $\mathcal{F}_s(t)$ to denote the flows that starts at time s.

Theorem (Existence of and convergence to the stationary flows)

Under Assumptions 1-3, there exists unique stationary and ergodic cumulative process $\mathcal{F}_e$ such that

$$(\mathcal{S}_s, \mathcal{F}_s) \Rightarrow (\mathcal{S}_e, \mathcal{F}_e) \equiv (Q_e, U_e, V_e, A_{0,e}, S_e, A_{\text{int},e}, A_e, D_e), \quad \text{as} \quad s \to \infty,$$

where $\Rightarrow$ denote weak convergence.

Existence, Uniqueness and Convergence

Proof sketch

- $A_{0,s}$ and S_s are vectors of delayed renewal process with first interval distributed as $U_s(0)$ and $V_s(0)$, respectively.
- $(A_{\text{int}}(t), A(t), D(t))$ are piece-wise constant with unit jumps. The jumps corresponds to the jumps in $\mathcal{S}_s$.
- Consider the sequence of jump times and jump types in $\mathcal{S}_s$, it is a continuous function of $\mathcal{S}_s$.
- Recover the flows from the jumps of $\mathcal{S}_s$ using inverse map, which is continuous.

HT Limits of the Stationary Flows

Having established the existence and convergence of stationary flows, we turn to the heavy-traffic limit of the stationary flows.

- We now assume that the system is in stationarity.
  - Suppress the subscript e to simplify the notation.
- $\mathcal{H} \subset \{1, 2, \ldots, K\}$: an arbitrary pre-selected subset of bottleneck queues.
  - So that $\rho_i \uparrow 1$ for $i \in \mathcal{H}$ and $\rho_i < 1 - \epsilon$ for $i \notin \mathcal{H}$.

Notation Via an Equivalent Network

Consider an alternative network, where non-bottleneck queues $i \in \mathcal{H}^c$ act as **instantaneous switches**.

- Let $P_{\mathcal{I},\mathcal{J}}$ collect the routing probabilities from stations in $\mathcal{I}$ to the ones in $\mathcal{J}$.
  - Define $P_{\mathcal{H},\mathcal{H}^c}, P_{\mathcal{H}^c,\mathcal{H}}$ and $P_{\mathcal{H}} \equiv P_{\mathcal{H},\mathcal{H}}, P_{\mathcal{H}^c} \equiv P_{\mathcal{H}^c,\mathcal{H}^c}$
- The new routing matrix for the bottleneck stations, denoted by $\hat{P}_{\mathcal{H}}$, is
$$\hat{P}_{\mathcal{H}} = P_{\mathcal{H}} + P_{\mathcal{H},\mathcal{H}^c} \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c}.$$
- Let $\hat{P}_{\mathcal{H}^c,\mathcal{H}}$ denote the probabilities that the first visit to a bottleneck queue of an external arrival at a non-bottleneck queue $i \in \mathcal{H}^c$ is at $j \in \mathcal{H}$, then we have
$$\hat{P}_{\mathcal{H}^c,\mathcal{H}} = \left(I_{\mathcal{H}^c} - P_{\mathcal{H}^c}\right)^{-1} P_{\mathcal{H}^c,\mathcal{H}}.$$
- The new external arrival rate $\hat{\lambda}_{0,\mathcal{H}}$ is
$$\hat{\lambda}_{0,\mathcal{H}} = \lambda_{0,\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} \lambda_{0,\mathcal{H}^c},$$
- The new total arrival rates remain unchanged
$$\hat{\lambda}_{\mathcal{H}} = (I - \hat{P}'_{\mathcal{H}})^{-1} \hat{\lambda}_{0,\mathcal{H}} = \lambda_{\mathcal{H}}.$$

We consider the usual HT scaling:

- Scale time by $(1-\rho)^{-2}$, scale space by $(1-\rho)$.

$$
\begin{aligned}
A^*_{0,i,\rho}(t) &\equiv (1-\rho)[A_{0,i}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_{0,i}t],\\
S^*_{i,\rho}(t) &\equiv (1-\rho)[S_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\mu_{i,\rho}t],\\
\Theta^*_{i,j,\rho}(t) &\equiv (1-\rho)\left[\Theta_{i,j,\rho}\left(\lfloor(1-\rho)^{-2}t\rfloor\right) - p_{i,j}(1-\rho)^{-2}t\right],\\
A^*_{i,j,\rho}(t) &\equiv (1-\rho)[A_{i,j,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i p_{i,j}t],\\
A^*_{i,\rho}(t) &\equiv (1-\rho)[A_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t],\\
D^*_{i,\rho}(t) &\equiv (1-\rho)[D_{i,\rho}((1-\rho)^{-2}t) - (1-\rho)^{-2}\lambda_i t],\\
Q^*_{i,\rho}(t) &\equiv (1-\rho)Q_{i,\rho}((1-\rho)^{-2}t),\\
Z^*_{i,\rho}(t) &\equiv (1-\rho)Z_{i,\rho}((1-\rho)^{2}t).
\end{aligned}
$$

Furthermore, let $\mathcal{F}^*_\rho$ collects all the flows, defined as

$$\mathcal{F}^*_\rho(t) \equiv (A^*_{0,\rho}(t), S^*_\rho(t), A^*_{\text{int},\rho}(t), A^*_\rho(t), D^*_\rho(t)).$$

Theorem (Heavy-Traffic FCLT)

Under Assumption 1-3,

$$(Q^*_\rho, Z^*_\rho, \Theta^*_\rho, \mathcal{F}^*_\rho) \Rightarrow (Q^*, Z^*, \Theta^*, \mathcal{F}^*), \quad \rho \uparrow 1.$$

1. $A^*_{0,i}, S^*_i$ *and* $\Theta^* = (\Theta^*_i : 1 \le i \le K)$ *are mutually independent BMs.*
2. $Q^*_{\mathcal{H}^c} \equiv 0$ *and* $Q^*_{\mathcal{H}}$ *is a stationary* $|\mathcal{H}|$*-dimensional RBM*

$$Q^*_{\mathcal{H}} \equiv \psi_{I-\hat{P}_{\mathcal{H}}}\left(\hat{X}^*_{\mathcal{H}}\right),$$
$$\hat{X}^*_{\mathcal{H}} = Q^*_{\mathcal{H}}(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}\right)\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) - (I - \hat{P}_{\mathcal{H}})S^*_{\mathcal{H}} - \hat{\lambda}_{0,\mathcal{H}} e.$$

3. *The total arrival process* A^**, the internal arrival process* $A^*_{i,j}$ *and the departure process* D^*

$$A^* = (I - P')^{-1}\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) + P'(I - P')^{-1}\left(Q^*(0) - Q^*\right),$$
$$D^* = (I - P')^{-1}\left(Q^*(0) - Q^* + A^*_0 + (\Theta^*)'\,\mathbf{1}\right),$$
$$A^*_{i,j} = p_{i,j} D^*_i + \Theta^*_{i,j} \circ \lambda_i e, \quad \textit{for} \quad 1 \le i, j \le K.$$

4. $Z^*_i = \lambda_i^{-1} Q^*_i$.

Remarks

The queue length process

$$\hat{X}^*_{\mathcal{H}} = Q^*_{\mathcal{H}}(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}\right)\left(A^*_0 + (\Theta^*)'\,\mathbf{1}\right) - (I - \hat{P}_{\mathcal{H}})S^*_{\mathcal{H}} - \hat{\lambda}_{0,\mathcal{H}} e.$$

- A^*_0 corresponds to the external arrival process, whereas $(\Theta^*)'\,\mathbf{1}$ is there because of the Markovian routing.
- $e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c,\mathcal{H}} e'_{\mathcal{H}^c}$ collects (1) the arrivals *directly* to the bottleneck queues and (2) the arrival to non-bottleneck queues and the directed to bottleneck queues.
- The service at bottleneck queues are adjusted by $I - \hat{P}_{\mathcal{H}}$ to account for immediate feedback to themselves.

Remarks

Proof sketch

- A key step is to show

$$(Q^*_{\mathcal{H},\rho}(0), Q^*_{\mathcal{H}^c,\rho}(0)) \Rightarrow (Q^*_{\mathcal{H}}(0), Q^*_{\mathcal{H}^c}(0)) \quad \text{as} \quad \rho \uparrow 1,$$

  which follows from Budhiraja and Lee (2009) with a slight generalization to cover networks with non-bottleneck queueus.
- The convergence of $Q^*_{\mathcal{H},\rho}, A^*, D^*$ then follows from the system equation for the scaled processes.
- The convergence of internal arrival processes A^*_{int} follows from the functional cental limit theorem for the splitting operation.

Approximation of the Variability in the Flows

We now illustrate how the HT limits for the stationary flows can be applied in queueing approximations.

For convenience, we work with the Index of Dispersion for Counts (IDC)

$$I_a(t) \equiv Var(A(t))/E[A(t)], \quad t \geq 0,$$

- Simply a scaled version of the variance-time curve.
- Robust Queueing algorithm produces approximation of the performance measures using IDC.
- Thus, we focus only on the approximation of the IDC of the stationary customer flows here.

Heavy-Traffic Limit for the Variance Functions

Define the HT-scaled variance function of the stationary departure process

$$V^*_{d,\rho}(t) \equiv Var(D^*_\rho(t)).$$

Theorem (HT limit for the departure variance)

Under uniform integrability conditions, $V^*_{d,\rho}(t)$ *converges to*

$$V^*_d(t) \equiv w^*\left(\lambda t/c_x^2\right) c_a^2 \lambda t + \left(1 - w^*\left(\lambda t/c_x^2\right)\right) c_s^2 \lambda t, \text{ as } \rho \uparrow 1$$

where $c_x^2 = c_a^2 + c_s^2$,

$$w^*(t) = \frac{1}{2t}\left(\left(t^2 + 2t - 1\right)\left(2\Phi(\sqrt{t}) - 1\right) + 2\sqrt{t}\phi(\sqrt{t})\left(1 + t\right) - t^2\right)$$

and ϕ, Φ *are the standard normal pdf and cdf, respectively.*

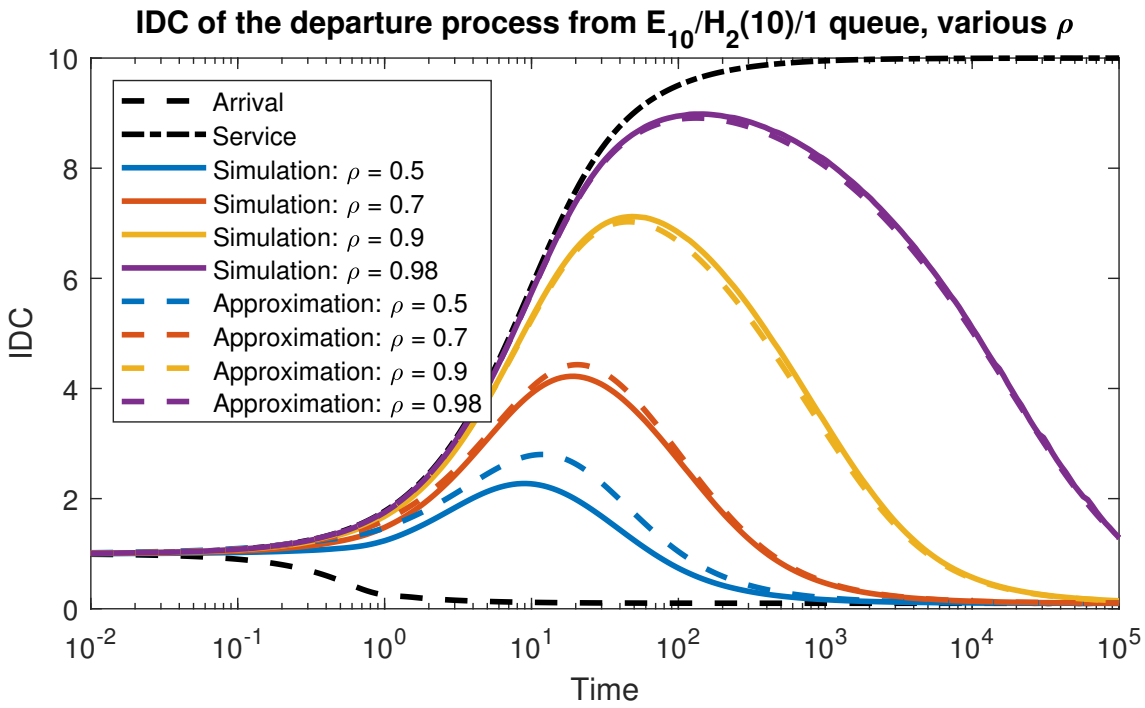
Approximation for Departure IDC

The HT theorem for variance supports the following approximation

$$I_d(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t)) I_s(\rho t), \qquad \text{(Dep)}$$

where

$$w_\rho(t) = w^*((1-\rho)^2 \lambda t/(\rho c_x^2)),$$



Example: Dependent Superposition

Let us look at another simple example.

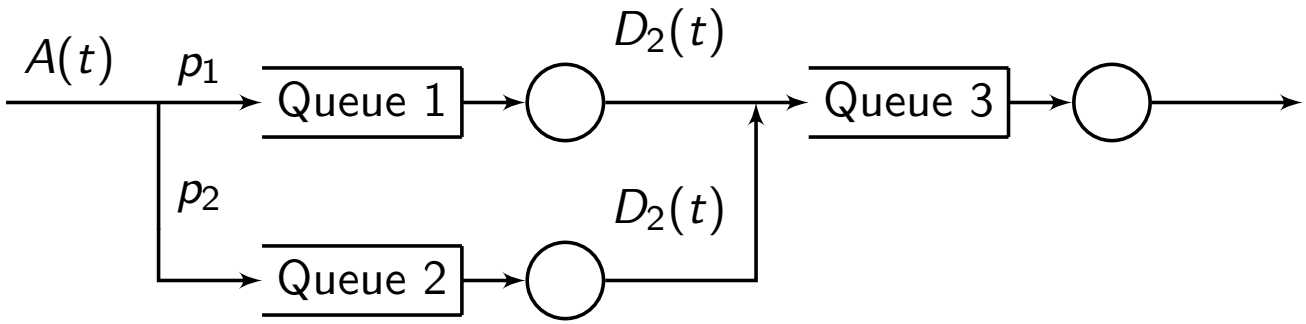


Figure: A re-combining after splitting example.

To approximate the IDC of the total arrival process at queue 3, we write

$$
\begin{aligned}
I_{a,3,\rho}(t) &\equiv \frac{\mathrm{Var}(A_{3,\rho}(t))}{E[A_{3,\rho}(t)]} = \frac{\mathrm{Var}\left(D_{1,\rho}(t) + D_{2,\rho}(t)\right)}{E[A_{3,\rho}(t)]} \\
&= \frac{\mathrm{Var}\left(D_{1,\rho}(t)\right)}{E[A_{3,\rho}(t)]} + \frac{\mathrm{Var}\left(D_{2,\rho}(t)\right)}{E[A_{3,\rho}(t)]} + \mathrm{cov}\left(D_{1,\rho}(t), D_{2,\rho}(t)\right) / E[A_{3,\rho}(t)] \\
&= p_1 I_{d,1,\rho}(t) + p_2 I_{d,2,\rho}(t) + \beta_\rho(t).
\end{aligned}
$$

Example: Dependent Superposition

In general, exact characterization of β_ρ is not readily available. We propose the following approximation

$$\begin{aligned}\beta_\rho(t) &\approx 2\mathrm{cov}\left(D_1^*((1-\rho)^2 t), D_2^*((1-\rho)^2 t)\right)/(\lambda(1-\rho)^2 t) \\ &= 2p_1(1-p_1)(c_{a_0}^2 - 1)w^*((1-\rho)^2 p_1 \lambda t / c_{x_1}^2))\end{aligned}$$

Let $\beta_\rho^*(t) = \beta_\rho\left((1-\rho)^{-2} t\right)$ be the HT-scaled correction term.

Corollary

Under mild conditions, we have

$$\beta_\rho^*(t) \to 2p_1(1-p_1)(c_{a_0}^2 - 1)w^*\left(p_1 \lambda t / c_{x_1}^2\right)$$

uniformly on bounded intervals.

Example: Dependent Superposition

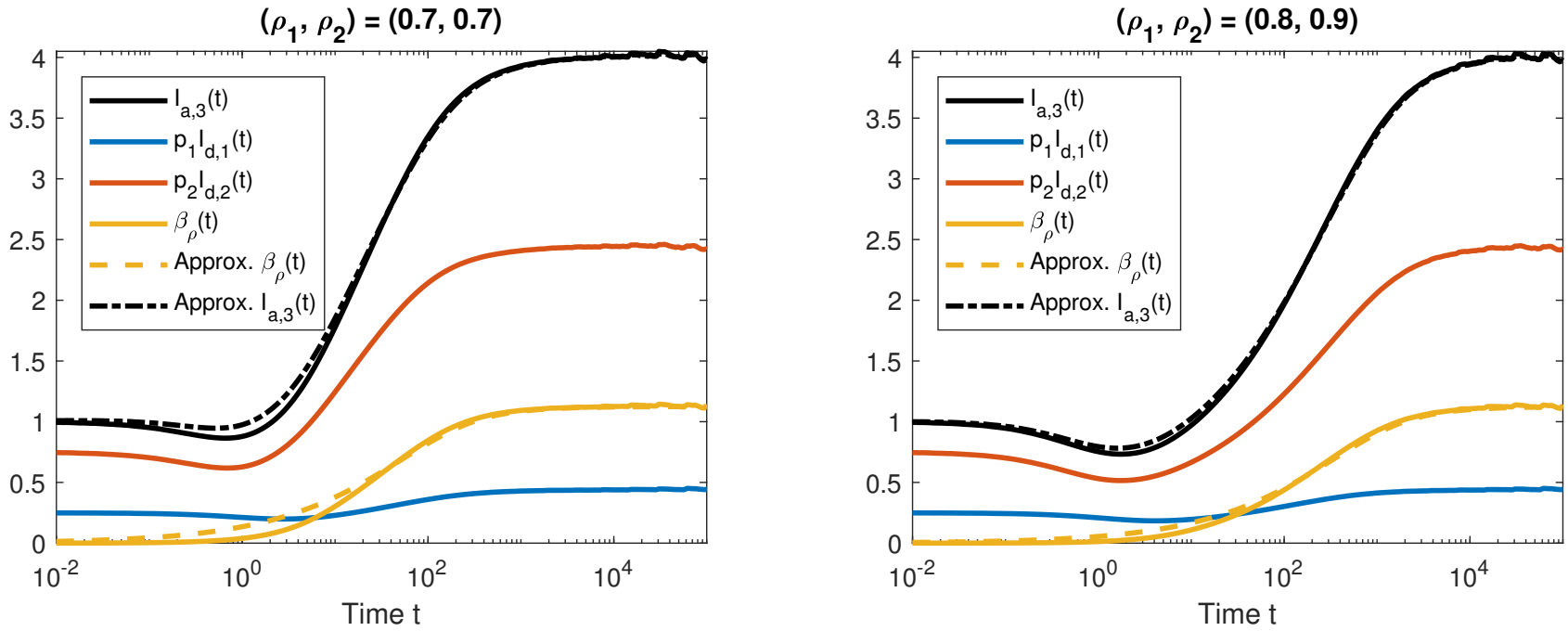


Figure: Approximation of the IDC of the total arrival process at station 3. The external arrival process is hyperexponential and the service distribution is Erlang.

The IDC Equations

In fact, we can derive a set of IDC equations

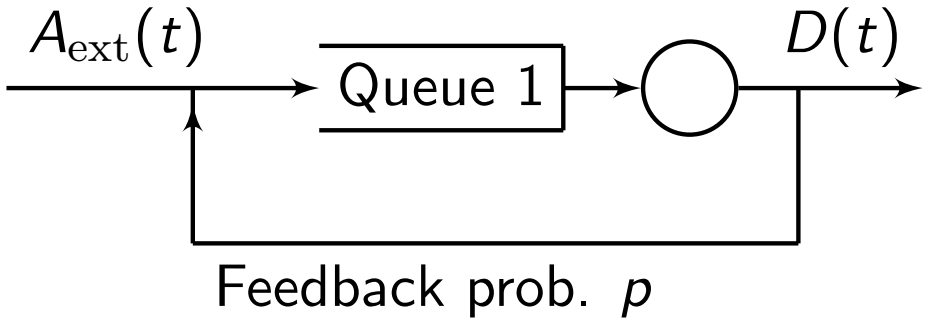
$$l_{d,i}(t) = w_i(t) l_{a,i}(t) + (1 - w_i(t)) l_{s,i}(\rho t), \tag{Dep}$$

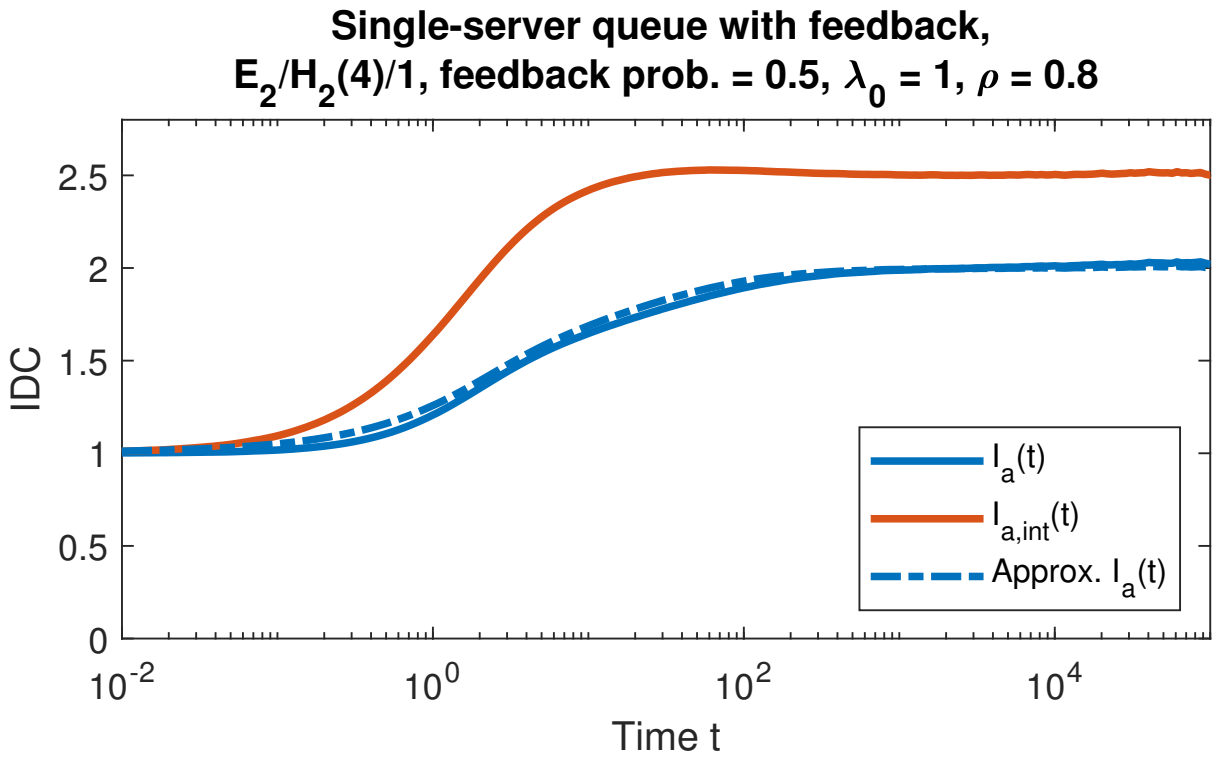
$$l_{a,i,j}(t) = p_{i,j} l_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t), \tag{Spl}$$

$$l_{a,i}(t) = \sum_{j=0}^{K} (\lambda_{j,i}/\lambda_i) l_{a,j,i}(t) + \beta_i(t). \tag{Sup}$$

- A system of linear equations for each fixed t;
- The IDC equations have a unique solution if every customer eventually leave the system.

Example: Dependent Splitting





Thank You!

Asmussen, S. (2003). *Applied Probability and Queues*. Springer, New York, second edition.

Borovkov, A. A. (1986). Limit theorems for queueing networks, I. *Theory of Probability & Its Applications*, 31(3):413–427.

Braverman, A., Dai, J. G., and Miyazawa, M. (2017). Heavy traffic approximation for the stationary distribution of a generalized Jackson network: the BAR approach. *Stochastic Systems*, 7(1):143–196.

Budhiraja, A. and Lee, C. (2009). Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, 34(1):45–56.

Chen, H. and Mandelbaum, A. (1991a). Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.*, 16(2):408–446.

Chen, H. and Mandelbaum, A. (1991b). Stochastic discrete flow networks: diffusion approximations and bottlenecks. *The Annals of Probability*, 19(4):1463–1519.

Dai, J. (1995). On the positive Harris recurrence for multiclass queueing networks. *Ann Appl Probab*, 5:49–77.

Dai, J. and Meyn, S. P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control*, 40(11):1889–1904.

Dai, J. G. and Harrison, J. M. (1992). Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, pages 65–86.

Davis, M. H. A. (1984). Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic processes. *J. Roy. Stat.Soc. B*, 46(3):353–388.

Foss, S. (1991). Ergodicity of queueing networks. *Siberian Math. J.*, 32:183–202.

Gamarnik, D. and Zeevi, A. (2006). Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Advances in Applied Probability*, 16(1):56–90.

Harrison, J. M. (1973). The heavy traffic approximation for single server queues in series. *Journal of Applied Probability*, 10(3):613–629.

Harrison, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10(4):886–905.

Harrison, J. M. and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. *The Annals of Probability*, pages 302–308.

Iglehart, D. L. and Whitt, W. (1970a). Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*, 2(1):150–177.

Iglehart, D. L. and Whitt, W. (1970b). Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability*, 2(2):355–369.

Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458.

Sigman, K. (1988). Queues as Harris recurrent Markov chains. *Queueing Systems*, 3(2):179–198.

Sigman, K. (1990). The stability of open queueing networks. *Stochastic Processes and their Applications*, 35(1):11–25.

Whitt, W. and You, W. (2018). Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199.