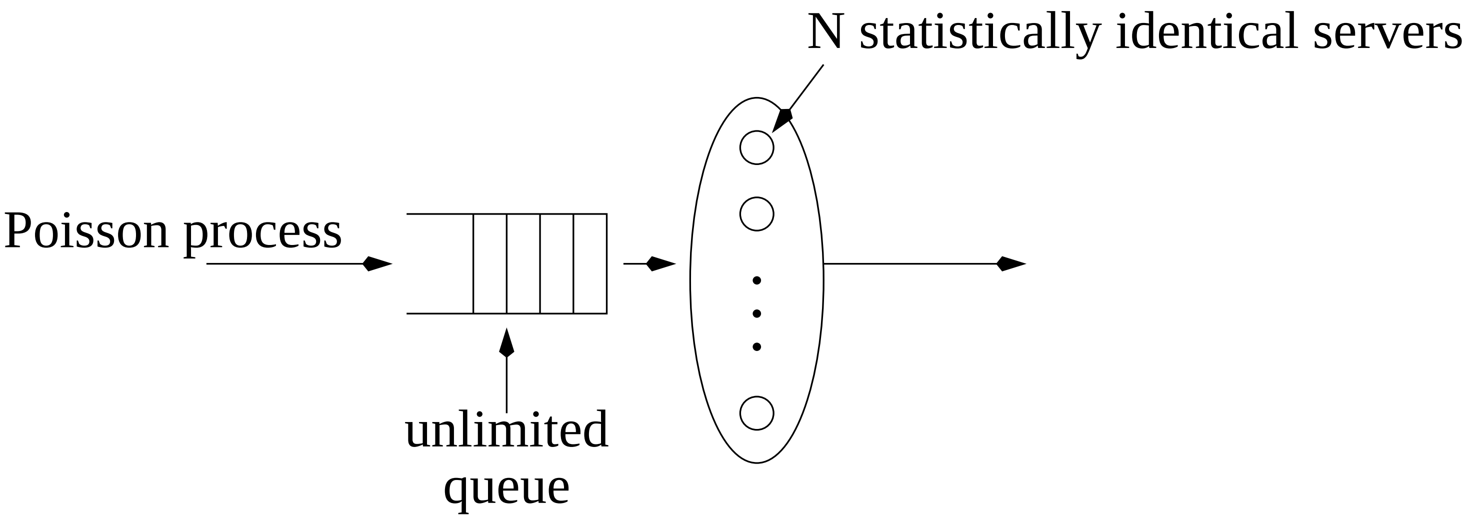
A Comparison of Two Heavy Traffic Regimes, I: Multiserver Queues

Marty Reiman

Bell Labs, Lucent Technologies

Murray Hill, NJ

The M/M/N (Erlang C) Queue



This model assumes:

- A Poisson arrival process with a constant rate, λ
- N statistically identical servers
- Exponentially distributed service times, rate μ
- Unlimited queue size

Traffic intensity: $\rho = \lambda / N\mu$

$Q(t)$ = number of customers in the system at time t.

W_k = waiting time (in queue) of k^{th} customer to enter.

Formulas for M/M/N 'Steady-State' Performance Measures

If $\rho = \frac{\lambda}{N\mu} < 1$, then

$$Q(t) \xrightarrow{d} Q \text{ as } t \to \infty \quad \text{and} \quad W_k \xrightarrow{d} W \text{ as } k \to \infty\,.$$

From the birth-death equations, we can calculate

$$P\{Q = m\} = \begin{cases} G^{-1} \frac{r^m}{m!}, & 0 \le m \le N \\ G^{-1} \frac{r^m}{N!N^{(m-N)}}, & m \ge N \end{cases}$$

where $r = \frac{\lambda}{\mu} = N\rho$, and

$$G = \sum_{m=0}^{N-1} \frac{r^m}{m!} + \frac{r^N}{N!(1-\rho)}\,.$$

Also,

$$P\{W > 0\} = \alpha = \frac{r^N/N!}{r^N/N! + (1-\rho)\sum_{k=0}^{N-1} r^k/k!}$$

and

$$E[Q-N]^+ = \frac{\alpha\rho}{1-\rho}\,.$$

Note that, as $\rho \uparrow 1$, $E[Q-N]^+ \to \infty$, and

$$(1-\rho)E[Q-N]^+ \to \alpha\,.$$

The Central Limit Theorem

Let x_1, $x_2, \ldots$ be i.i.d. real-valued random variables with

$$E[x_i] = \bar{x}$$

and

$$E[(x_i - \bar{x})^2] = \sigma^2, \quad 0 < \sigma^2 < \infty\,.$$

For $n \geq 1$, let

$$S_n = \sum_{i=1}^{n} x_i$$

and

$$Y_n = \frac{S_n - n\bar{x}}{\sqrt{n}}\,.$$

Let $N(\beta, \sigma^2)$ denote a normally distributed random variable with mean β and variance σ^2.

Central Limit Theorem: $Y_n \overset{d}{\to} Y = N(0, \sigma^2)$ as $n \to \infty$, where $\overset{d}{\to}$ denotes convergence in distribution.

Brownian Motion

Let $\{Z(t), t \geq 0\}$ be a continuous path stochastic process with $Z(0) = 0$.

Then $\{Z(t), t \geq 0\}$ is Brownian motion with drift c and variance σ^2 if, for
$0 = t_0 < t_1 < t_2 < \cdots < t_k < \infty$,

$$Z(t_1) - Z(t_0), \quad Z(t_2) - Z(t_1), \ldots, Z(t_k) - Z(t_{k-1})$$

are independent, and

$$Z(t_i) - Z(t_{i-1}) = N(c[t_i - t_{i-1}], \quad \sigma^2[t_i - t_{i-1}]) \,.$$

We denote such a process by $BM(c, \sigma^2)$.

The Functional Central Limit Theorem

For $n \geq 1$ and $t \geq 0$, let

$$Z_n(t) = Y_{\lfloor nt \rfloor} = \frac{S_{\lfloor nt \rfloor} - \lfloor nt \rfloor \bar{x}}{\sqrt{n}} .$$

By the Central Limit Theorem, with t fixed,

$$Z_n(t) \xrightarrow{d} N(0, \sigma^2 t) ,$$

and with $0 < s < t$,

$$Z_n(t) - Z_n(s) \xrightarrow{d} N(0, \sigma^2 [t - s]) .$$

Functional Central Limit Theorem (Donsker's Theorem):

$$Z_n \xrightarrow{d} Z = BM(0, \sigma^2) \quad \text{as} \quad n \to \infty .$$

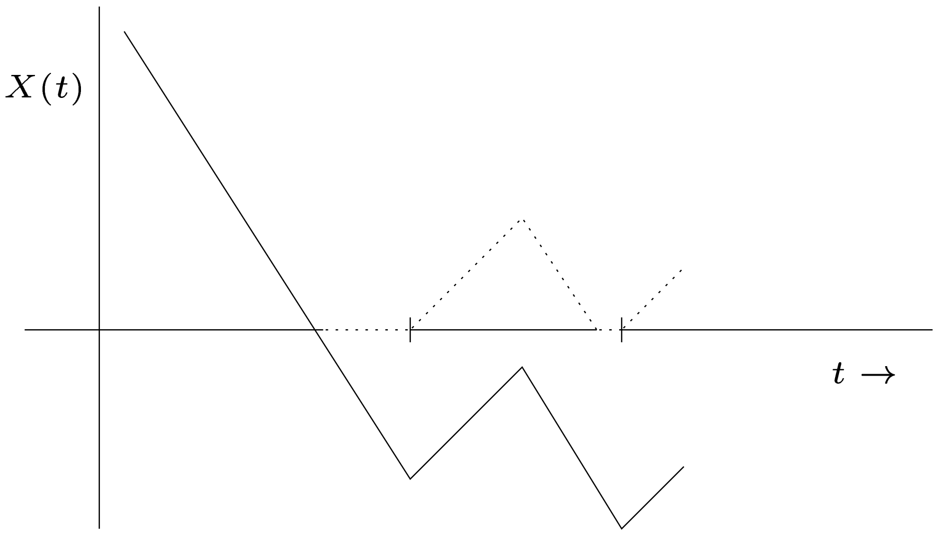
Reflected Brownian Motion

Let $Z = BM(c, \sigma^2)$, and let

$$X(t) = Z(t) - \inf_{0 \le s \le t} \{Z(s)\} \wedge 0\,.$$

Then $\{X(t), t \ge 0\}$ is reflected Brownian motion with drift c and variance σ^2:

$$X = RBM(c, \sigma^2)$$



If $c < 0$, then X is positive recurrent, $X(t) \to \tilde{X}$ as $t \to \infty$, with

$$P\{\tilde{X} \le x\} = 1 - e^{2cx/\sigma^2}\,.$$

'Classical' Heavy Traffic for the M/M/N Queue

Consider a sequence of systems, indexed by $n \geq 1$.

Keep N fixed, and assume that $\mu^n/n \to \mu$, $\lambda^n/n \to N\mu$ $(0 < \mu < \infty)$, and

$$\sqrt{n}(1-\rho^n) \to \beta, \quad -\infty < \beta < \infty, \quad \text{as} \quad n \to \infty\,.$$

(Recall: $\rho^n = \lambda^n/N\mu^n$)

Let

$$X^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \quad 0 \leq t < \infty\,,$$

and let $X = RBM(-\beta N\mu,\ 2N\mu)$.

Theorem (Iglehart and Whitt, 1970). If $X^n(0) \xrightarrow{d} X(0)$ as $n \to \infty$ then $X^n \xrightarrow{d} X$ on $D[0,\infty)$ as $n \to \infty$.

One-Dimensional Diffusions

A one-dimensional diffusion process $\{Z(t), t \geq 0\}$ with drift $m(\cdot)$ and variance $\sigma^2(\cdot)$ is a continuous path (strong) Markov process with

$$\lim_{\Delta \downarrow 0} \Delta^{-1} E\left(Z(t+\Delta) - Z(t) \middle| Z(t) = x\right) = m(x)$$

and

$$\lim_{\Delta \downarrow 0} \Delta^{-1} E\left([Z(t+\Delta) - Z(t)]^2 \middle| Z(t) = x\right) = \sigma^2(x) \, .$$

If $Z = BM(c, \sigma^2)$, then $m(x) = c$ and $\sigma^2(x) = \sigma^2$.

Heavy Traffic with $N \to \infty$

Consider a sequence of systems, indexed by $N \geq 1$. Assume that $\mu^N \to \mu$ $(0 < \mu < \infty)$ and $\lambda^N/N \to \mu$ with

$$\sqrt{N}(1-\rho^N) \to \beta, \quad -\infty < \beta < \infty, \quad \text{as} \quad N \to \infty\,.$$

Let

$$X^N(t) = \frac{Q^N(t) - N}{\sqrt{N}}, \quad 0 \leq t < \infty\,.$$

Theorem (Halfin and Whitt, 1981). If $X^N(0) \xrightarrow{d} X(0)$ as $N \to \infty$ then $X^N \xrightarrow{d} X$ on $D[0,\infty)$ as $N \to \infty$, where X is a one-dimensional diffusion with infinitesimal drift

$$m(x) = \begin{cases} -\mu\beta, & x \geq 0 \\ -\mu(x+\beta), & x \leq 0 \end{cases}$$

and infinitesimal variance 2μ.

Proof of the Halfin-Whitt Limit Theorem

Infinitesimal drift:

$$m^N(x) = \lim_{\Delta \downarrow 0} \Delta^{-1} E\left[X^N(t+\Delta) - X^N(t) \mid X^N(t) = x\right]$$

$$= \frac{1}{\sqrt{N}} \left[\lambda^N - N\mu^N - \sqrt{N}\mu^N (x \wedge 0)\right]$$

$$\approx -\mu^N \beta - \mu^N (x \wedge 0) \to -\mu\beta - \mu(x \wedge 0) \;\text{ as } N \to \infty$$

Infinitesimal variance:

$$\sigma^2(N,x) = \lim_{\Delta \downarrow 0} \Delta^{-1} E\left[\left(X^N(t+\Delta) - X^N(t)\right)^2 \mid X^N(t) = x\right]$$

$$= \frac{1}{N} \left[\lambda^N + N\mu^N + \sqrt{N}\mu^N (x \wedge 0)\right]$$

$$\to 2\mu \;\text{ as } N \to \infty$$

Apply Stone's Theorem (1963) to obtain the result.

Comparison

N **fixed**	$N \to \infty$
queue length not centered	queue length is centered
limit process is reflected	limit process is unconstrained
limit drift is constant	limit drift is state dependent
time in service: $O(n^{-1}) \to 0$	time in service is $O(1)$

Desired Extensions

1. More general arrival process (e.g. renewal process)

2. More general service time distribution

3. Network setting

4. Time dependent parameters

5. Multiple classes

Classical Heavy Traffic for the GI/GI/N Queue

Renewal arrival process $\{A^n(t),\ t \geq 0\}$:

rate λ^n, variance $\sigma_A^2(n)$

Renewal 'service' process $\{S_i^n(t),\ t \geq 0\}$:

rate μ^n, variance $\sigma_S^2(n)$, $1 \leq i \leq N$

N fixed,

$$\mu^n/n \to \mu,\ \lambda^n/n \to N\mu,$$
$$n^2\sigma_A^2(n) \to \sigma_A^2,\ n^2\sigma_S^2(n) \to \sigma_S^2,$$

with $0 < \mu < \infty,\ 0 \leq \sigma_A^2, \sigma_S^2 < \infty,\ \sigma_A^2 + \sigma_S^2 > 0$.

$$\sqrt{n}(1-\rho^n) \to \beta, \quad -\infty < \beta < \infty, \quad \text{as} \quad n \to \infty\,.$$

(Also assume a Lindeberg condition on interarrival and service times.)

Let

$$X^n(t) = \frac{Q^n(t)}{\sqrt{n}}, \quad 0 \leq t < \infty\,,$$

and

$$X = RBM(-\beta N\mu, N^3\mu^3\sigma_A^2 + N\mu^3\sigma_S^2)\,.$$

Theorem (Iglehart and Whitt, 1970). If $X^n(0) \xrightarrow{d} X(0)$ as $n \to \infty$ then $X^n \xrightarrow{d} X$ on $D[0,\infty)$ as $n \to \infty$.

Outline of the Proof of the Iglehart-Whitt Limit Theorem

Let

$$\hat{A}^n(t) = \frac{A^n(t) - \lambda^n t}{\sqrt{n}},$$

$$\hat{S}_i^n(t) = \frac{S_i^n(t) - \mu^n t}{\sqrt{n}},$$

and

$$B^n(t) = \frac{A^n(t) - \sum_{i=1}^{N} S_i^n(t)}{\sqrt{n}} = \hat{A}^n(t) - \sum_{i=1}^{N} \hat{S}_i^n(t) + \frac{(\lambda^n - N\mu^n)}{\sqrt{n}}\, t\,.$$

Under the assumptions, using Prohorov's Theorem (1963), $\hat{A}^n \xrightarrow{d} \hat{A} = BM(0, N^3\mu^3\sigma_A^2)$ and $\hat{S}_i^n \xrightarrow{d} \hat{S}_i = BM(0, \mu^3\sigma_S^2)$, so that

$$B^n \xrightarrow{d} B = BM(-\beta N\mu,\ N^3\mu^3\sigma_A^2 + N\mu^3\sigma_S^2)\,.$$

Use

$$X^n(t) = B^n(t) - \left[\inf_{0 \le s \le t} \{B^n(s)\} \wedge 0\right] + \text{ small error}\,,$$

along with the Continuous Mapping Theorem (Billingsley, 1968) to complete the proof.

Phase-Type Distributions

Continuous time transient Markov process:

$K = \#$ of states (phases)

$(p_1, \ldots, p_K)$: initial distribution, $\sum_{i=1}^{K} p_i = 1$

$(p_{ij},\ 1 \le i, j \le K)$: transition probabilities ($p_{ii} = 0$)

$p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$: absorption probabilities

μ_i^{-1}: mean sojourn time in state i

Auxiliary discrete time $K+1$ state Markov chain:

state space $\{0, 1, \ldots, K\}$

transition probabilities: $\begin{cases} p_{ij}, & 1 \le i \le K,\ 0 \le j \le K \\ p_j, & i = 0, \qquad\quad 1 \le j \le K \end{cases}$

stationary distribution $(\eta_0, \eta_1, \ldots, \eta_K)$

mean service time $\mu^{-1} = \eta_0^{-1} \sum_{i=1}^{K} \frac{\eta_i}{\mu_i}$

Service process:

continuous time Markov process with state 0 instantaneous

stationary distribution $(q_1, \ldots, q_K)$

$$q_i = \mu \frac{\eta_i}{\eta_0 \mu_i}$$

The GI/PH/N Queue in the 'Halfin-Whitt' Regime

Renewal arrival process, rate λ_N, variance $\sigma_A^2(N)$,

$$\lambda_N/N \to \mu \in (0,\infty),\ N^2\sigma_A^2(N) \to \sigma_A^2 \in [0,\infty)$$

Service times are i.i.d., with a phase-type distribution having K phases.

$Q_i^N(t)$ = total number of 'phase i' customers in the system,

and

$$X_i^N(t) = \frac{Q_i^N(t) - Nq_i}{\sqrt{N}},\ 1 \le i \le K,\ t \ge 0\,.$$

Assume that, for some β, with $-\infty < \beta < \infty$,

$$\lambda_N = \mu N - \mu\beta\sqrt{N},\quad N \ge 1$$

so that $\sqrt{N}(1-\rho_N) = \beta$.

(Also: some additional conditions on the initial configuration of customers in the queue, if any.)

Theorem (Puhalskii and Reiman, 2000). If $\boldsymbol{X}^N(0) \stackrel{d}{\to} \boldsymbol{X}(0)$ then $\boldsymbol{X}^N \stackrel{d}{\to} \boldsymbol{X}$ in $D[0,\infty)$, where $\boldsymbol{X}$ is a K dimensional diffusion with infinitesimal drift vector $\boldsymbol{m(x)}$ and an infinitesimal covariance matrix Γ, where

$$m_i(x) = \begin{cases} -\mu\beta p_i + \displaystyle\sum_{j=1,j\neq i}^{K} x_j\mu_{ji} - \mu_i x_i, & \displaystyle\sum_{\ell=1}^{K} x_\ell \leq 0 \\ -\mu\beta p_i + \displaystyle\sum_{j=1,j\neq i}^{K} \left[x_j - p_j \sum_{\ell=1}^{K} x_\ell\right] \mu_{ji} & \\ -\mu_i \left[x_i - p_i \displaystyle\sum_{\ell=1}^{K} x_\ell\right], & \displaystyle\sum_{\ell=1}^{K} x_\ell \geq 0, \end{cases}$$

$$\Gamma_{ii} = 2q_i\mu_i, \quad 1 \leq i \leq K$$

$$\Gamma_{ij} = -q_i\mu_{ij} - q_j\mu_{ji}, \quad 1 \leq i \neq j \leq K$$

and $\mu_{ij} = \mu_i p_{ij}$

Outline of Proof

Let

$$\begin{aligned}
\hat{Q}_i^N(t) &= \#\text{ of customers in service in phase } i,\\
\tilde{Q}_i^N(t) &= \#\text{ of phase } i \text{ customers in queue,}
\end{aligned}$$

and

$$Q_0^N(t) = \text{total } \#\text{ of customers in queue}$$

(Recall:

$$\begin{aligned}
Q_i^N(t) &= \text{total } \#\text{ of phase } i \text{ customers in system}\\
&= \hat{Q}_i^N(t) + \tilde{Q}_i^N(t))
\end{aligned}$$

Consider 2 filtrations:

$\hat{\mathbb{F}}^N$ 'knows' the phase of a customer when it enters the system

$\mathbb{F}^N$ does not know the phase of a customer until it enters service

a) Q_0^N and $\hat{Q}_i^N$ are adapted to both $\hat{\mathbb{F}}^N$ and $\mathbb{F}^N$.

b) $\tilde{Q}_i^N$ is adapted to $\hat{\mathbb{F}}^N$ but not to $\mathbb{F}^N$.

c) The compensator of $\hat{Q}_i^N$ w.r.t. $\hat{\mathbb{F}}^N$ has a term of the form

$$1\{\text{phase } i \text{ customer is at the head of the queue}\}.$$

In its place, the compensator w.r.t. $\mathbb{F}^N$ has p_i.

Main Steps:

1. Obtain semimartingale decompositions for Q_0^N and $\hat{Q}_i^N$ using filtration $\mathbb{F}^N$.

2. Introduce auxiliary processes

$$\bar{Q}_i^N(t) = p_i Q_0^N(t) + \hat{Q}_i^N(t)$$

and show that Q_i^N and $\bar{Q}_i^N$ are close by showing that

$$\sup_{0 \le t \le T} \left| \frac{\tilde{Q}_i^N(t)}{\sqrt{N}} - \frac{p_i Q_0^N(t)}{\sqrt{N}} \right| \xrightarrow{p} 0 \, .$$

(This is an example of 'state-space collapse'.)

3. Obtain semimartingale decomposition for Q_i^N using filtration $\hat{\mathbb{F}}^N$, which yields a stochastic integral for X^N containing an 'error term' that, by step 2, is small.

4. Finish up with martingale FCLT.

A Comparison of Two Heavy Traffic Regimes, II: Multiserver Queueing Networks

Marty Reiman

Bell Labs, Lucent Technologies

Murray Hill, NJ, U.S.A.

The Jackson Network

K stations
N_k servers at station k, $1 \le k \le K$
Poisson arrivals, rates $\lambda_1, \ldots, \lambda_K$
Exponential service times, rates $\mu_1, \ldots, \mu_K$
i.i.d. routing, routing matrix $P = \{p_{ij}\}$

Traffic Intensity:

Let $\boldsymbol{\gamma} \in \mathbb{R}_+^K$ be the unique solution to

$$\boldsymbol{\gamma} = \boldsymbol{\lambda} + \boldsymbol{\gamma} P \ ,$$

so

$$\boldsymbol{\gamma} = \boldsymbol{\lambda}(I - P)^{-1} = \boldsymbol{\lambda} \sum_{l=0}^{\infty} P^l \, .$$

Then

$$\rho_i = \frac{\gamma_i}{N_i \mu_i}$$

$Q_i(t) = \#$ of customers at station i at time t.

Steady State for the Jackson Network

If $\rho_i < 1$, $1 \leq i \leq K$, then

$$(Q_1(t), \ldots, Q_K(t)) \stackrel{d}{\to} (Q_1, \ldots, Q_K) \quad \text{as} \quad t \to \infty\,,$$

with

$$P\{Q_1 = m_1, \ldots, Q_K = m_K\} = \prod_{i=1}^{K} P\{Q_i = m_i\}\,,$$

$$P\{Q_i = m_i\} = \begin{cases} G_i^{-1} \frac{r_i^{m_i}}{m_i!}, & 0 \leq m_i \leq N_i \\ G_i^{-1} \dfrac{r_i^{m_i}}{N_i! N_i^{(m_i - N_i)}}, & m_i \geq N_i\,, \end{cases}$$

$r_i = N_i \rho_i$ and

$$G_i = \sum_{m=0}^{N_i - 1} \frac{r_i^m}{m!} + \frac{r_i^{N_i}}{N_i!(1-\rho_i)}\,.$$

Classical Heavy Traffic for the Generalized Jackson Network: Set-Up

Generalized Jackson Network: same as Jackson network, except

Renewal arrivals: rate λ_i^n, variance $\sigma^2_{A,i}(n)$

i.i.d. service times: mean $(\mu_i^n)^{-1}$, variance $\sigma^2_{S,i}(n)$.

Keep $N_1, \ldots, N_K$ fixed, and assume that

$$\mu_i^n/n \to \mu_i, \qquad \lambda_i^n/n \to \lambda_i \ ,$$

$$n^2\sigma^2_{A,i}(n) \to \sigma^2_{A,i}, \qquad n^2\sigma^2_{s,i}(n) \to \sigma^2_{s,i}, \ldots,$$

and

$$\sqrt{n}(\rho_i^n - 1) \to \beta_i, \quad 1 \le i \le K\,.$$

Let $\boldsymbol{c} = \boldsymbol{\beta}(I-P)$ and

$$X_k^n(t) = \frac{Q_k^n(t)}{\sqrt{n}}, \qquad 1 \le k \le K\,.$$

Reflected Brownian Motion on the Nonnegative Orthant

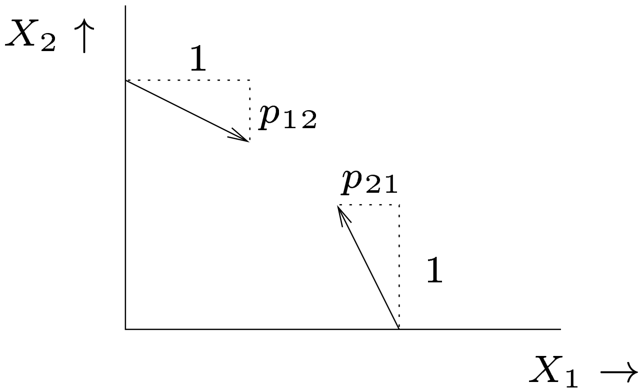
Reflection map:

Given $z : \mathbb{R}_+ \to \mathbb{R}^K$ continuous, let $x : \mathbb{R}_+ \to \mathbb{R}^K$ be such that

1) $x(t) = z(t) + (I - P)^T y(t) \geq 0$

2) $y(t)$ continuous and nondecreasing

3) $\int_0^t x_i(s) dy_i(s) = 0$, $1 \leq i \leq K$



Write $x = \phi(z)$. (x is unique)

Reflected Brownian motion on $\mathbb{R}_+^K$:

$$X = \phi(Z)$$

where $Z = \{Z(t), t \geq 0\}$ is Brownian motion on $\mathbb{R}^K$ with infinitesimal drift vector $\boldsymbol{c}$ and covariance matrix Γ.

The Classical Heavy Traffic Limit for the Generalized Jackson Network

Let $\{X(t), t \geq 0\}$ be a reflected Brownian motion on $\mathbb{R}_+^K$ with infinitesimal drift $(N_1 c_1 \mu_1, \ldots, N_K c_K \mu_K)$ and covariance matrix Γ, where

$$\Gamma_{ii} = \lambda_i^3 \sigma_{A,i}^2 + N_i \mu_i^3 \sigma_{S,i}^2 (1 - 2p_{ii})$$

$$+ \sum_{j=1}^{K} N_j \mu_j p_{ji} (1 - p_{ji} + p_{ji} \mu_j^2 \sigma_{S,j}^2), \; 1 \leq j \leq K$$

and

$$\Gamma_{ij} = -\Big[N_i \mu_i^2 \sigma_{S,i}^2 p_{ij} + N_j \mu_j^2 \sigma_{S,j}^2 p_{ji}$$

$$+ \sum_{k=1}^{K} N_k \mu_k p_{ki} p_{kj} (1 - \mu_k^2 \sigma_{S,k}^2) \Big], \; 1 \leq i < j < K \, .$$

Theorem (Reiman, 1984; Chen and Shanthikumar, 1994). In a generalized Jackson network, if $\boldsymbol{X}^n(0) \stackrel{d}{\to} \boldsymbol{X}(0)$, then $\boldsymbol{X}^n \stackrel{d}{\to} \boldsymbol{X}$ in $D[0, \infty)$.

Markovian Service Networks

Consider the process $\{Q_1(t), \dots, Q_K(t), t \geq 0\}$ defined by

$$\boldsymbol{Q}(t) = \boldsymbol{Q}(0) + \sum_{i=1}^{I} A_i \left(\int_0^t \alpha_i(\boldsymbol{Q}(s)) ds \right) \boldsymbol{v}_i$$

with

A_i, $1 \leq i \leq I$: mutually independent rate 1 Poisson processes

α_i: transition rates

$\boldsymbol{v}_i$: jump vectors

Example: For the $M/M/N$ queue, $K = 1$, $I = 2$,

$$\alpha_1(x) = \lambda, \quad v_1 = 1$$

$$\alpha_2(x) = \mu(x \wedge N), \quad v_2 = -1$$

$$Q(t) = Q(0) + A_1(\lambda t) - A_2 \left(\int_0^t \mu(Q(s) \wedge N) ds \right).$$

The Jackson Network in the Halfin-Whitt Regime

The Jackson network is a Markovian service network:

$$\boldsymbol{Q}(t) = \boldsymbol{Q}(0) + \sum_{i=1}^{K} \Bigg[A_i^a(\lambda_i t)\boldsymbol{e}_i$$

$$+ \sum_{j=0}^{K} A_{ij}^c \left(\int_0^t (Q_i(s) \wedge N)\mu_i p_{ij} ds \right) (\boldsymbol{e}_j - \boldsymbol{e}_i) \Bigg]$$

where $\boldsymbol{e}_0 = \boldsymbol{0}$ and $p_{i0} = 1 - \sum_{j=1}^{K} p_{ij}$. Assume that

$$N_i^N/N \to \bar{N}_i, \; 0 < \bar{N}_i < \infty,$$

$$\mu_i^N \to \mu_i, \qquad \lambda_i^N/N \to \lambda_i,$$

and that

$$\sqrt{N}(\rho_i^N - 1) \to \beta_i, -\infty < \beta_i < \infty, 1 \le i \le K\,.$$

Let

$$X_k^N(t) = \frac{Q_k^N(t) - N_k^N}{\sqrt{N}}, \quad 1 \le k \le K\,.$$

Theorem (Mandelbaum, Massey, Reiman, 1998). In a Jackson network, if $\boldsymbol{X}^N(0) \stackrel{d}{\to} \boldsymbol{X}(0)$, then $\boldsymbol{X}^N \stackrel{d}{\to} \boldsymbol{X}$ in $D[0,\infty)$, where $\boldsymbol{X}$ is a K dimensional diffusion with infinitesimal drift vector $\boldsymbol{m(x)}$ and a covariance matrix Γ, where

$$m_i(x) = \mu_i \bar{N}_i c_i + \mu_i x_i^- - \sum_{j=1}^{K} p_{ji} \mu_j x_j^-, \ 1 \le i \le K,$$

$$\Gamma_{ii} = \lambda_i + \bar{N}_i \mu_i (1 - 2p_{ii}) + \sum_{j=1}^{K} \bar{N}_j \mu_j p_{ji},$$

and

$$\Gamma_{ij} = -\left[\bar{N}_i \mu_i p_{ij} + \bar{N}_j \mu_j p_{ji}\right],$$

with

$$x_i^- = \begin{cases} -x_i, & x_i < 0 \\ 0, & x_i \ge 0 \end{cases}$$

The $M/H_2/N$ Queue is not an MSN

H_2: $K = 2$, $p_{12} = p_{21} = 0$

Recall, $Q_i(t)$ = total # of 'phase i' customers in system.

Given $(Q_1(t), Q_2(t))$ with $Q_1(t) + Q_2(t) > N$, we do not know how many of each phase are in service.

Let $\tilde{Q}_i^N(t)$ = # of phase i customers in queue

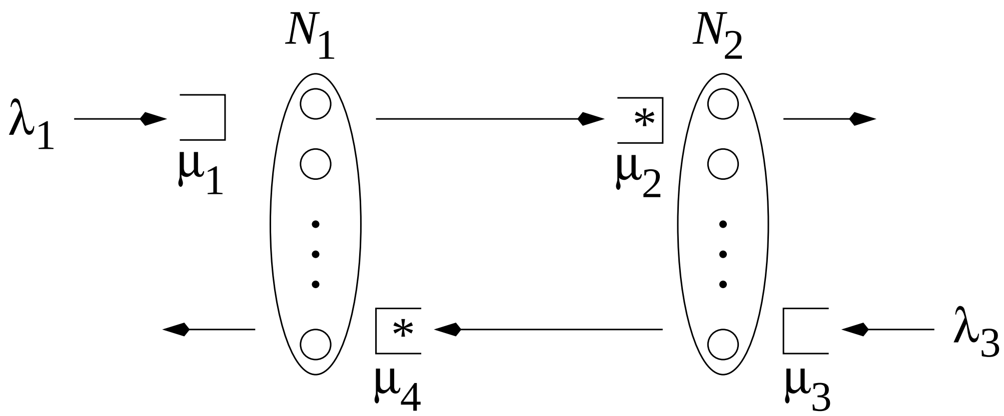
$Q_0^N(t)$ = total # of customers in queue

Lemma. For $0 < T < \infty$, $1 \le i \le 2$,

$$\sup_{0\le t\le T} \left| \frac{\tilde{Q}_i^N(t)}{\sqrt{N}} - p_i \frac{Q_0^N(t)}{\sqrt{N}} \right| \xrightarrow{P} 0\,.$$

This is an example of 'state-space collapse'

Multiclass Networks: An Example (The 'Rybko–Stolyar Network')



$*$: This customer class receives preemptive priority

$$\boldsymbol{Q}(t) \;=\; (Q_1(t), Q_2(t), Q_3(t), Q_4(t))\,, \quad t \geq 0$$

$$\rho_1 \;=\; N_1^{-1}\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_3}{\mu_4}\right), \quad \rho_2 = N_2^{-1}\left(\frac{\lambda_1}{\mu_2} + \frac{\lambda_3}{\mu_3}\right)$$

With $N_1 = N_2 = 1$, $\lambda_1 = \lambda_3 = 1$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_4 < 2$, Rybko and Stolyar (1992) showed that, even if $\rho_1 = \rho_2 < 1$, this system is not stable.

Their argument carries through with $N_1 = N_2 = N > 1$, $\lambda_1 = \lambda_3 = N$, $\mu_1 = \mu_3$, and $\mu_2 = \mu_4 < 2$.

The Rybko-Stolyar Network in the Halfin-Whitt Regime

Assume that $\mu_1^N = \mu_3^N$, $\mu_2^N = \mu_4^N$, $\lambda_1^N = \lambda_3^N$, $N_i^N = N$, $\mu_i^N \to \mu_i$, $i = 1, 2$, $\lambda_1^N / N \to \lambda_1$, and that

$$\sqrt{N}(\rho_1^N - 1) \to \beta, \quad -\infty < \beta < \infty\,.$$

Let

$$\begin{aligned} q_1 &= \frac{\lambda_1/\mu_1}{\lambda_1/\mu_1 + \lambda_3/\mu_4}, \quad q_2 = \frac{\lambda_1/\mu_2}{\lambda_1/\mu_2 + \lambda_3/\mu_3} \\ q_3 &= 1 - q_2, \quad q_4 = 1 - q_1\,, \text{ and} \\ X_i^N(t) &= \frac{Q_i^N(t) - q_i N}{\sqrt{N}}, \quad 1 \le i \le 4, \quad t \ge 0. \end{aligned}$$

Theorem (Mandelbaum, Massey, Reiman, 1998). If $\boldsymbol{X}^N(0) \overset{d}{\to} \boldsymbol{X}(0)$, then $\boldsymbol{X}^N \overset{d}{\to} \boldsymbol{X}$ in $D[0, \infty)$, where $\boldsymbol{X}$ is a 4 dimensional diffusion with drift vector $\boldsymbol{m}(\cdot)$ and covariance matrix Γ, where ...