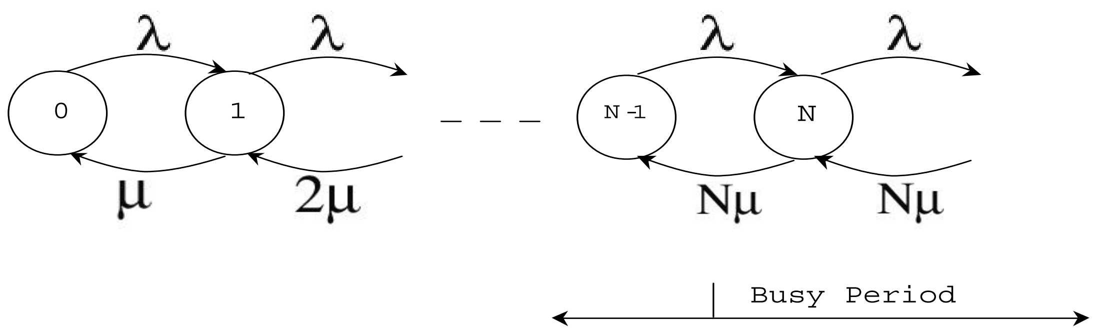
Markovian N-Server Queues

(Birth & Death Models)



Arrivals Poisson (λ) ;

Services exp(μ) $(E(S) = 1/\mu)$

Servers N **statistically identical**, serving FCFS.

Offered load $R = \lambda \times E(S) = \lambda/\mu$ Erlangs

$Q(t)$ $=$ number in system (served + queued) at time t.

$W(k)$ $=$ queueing time of k-th arrival.

Steady State: $Q(\infty)$, $W(\infty)$, when exists.

non-idling $\Rightarrow$ $[Q(t) - N]^{+}$ = queue-length, and

$[Q(t) - N]^{-}$ = number of idle servers

Examples: M/M/N/K, Abandonment, Balking

M/M/∞: $\lambda_k \equiv \lambda, \quad \mu_k = k \cdot \mu, \qquad k \geq 1.$

Steady state $\pi_k = e^{-R} R^k / k!$ Poisson (R)

M/M/N/N: $\lambda_k \equiv \lambda, \quad \mu_k = k \cdot \mu, \quad 0 \leq k \leq N.$

Steady state $$\pi_k = \frac{R^k}{k!} \Bigg/ \sum_{n=0}^{N} \frac{R^n}{n!}$$

Erlang-B $P\{\text{Blocked}\} \doteq \boldsymbol{E_{1,N}} = \pi_N$, by PASTA

M/M/N: $\lambda_k \equiv \lambda, \qquad \mu_k = (k \wedge N) \cdot \mu \quad , \quad k \geq 1.$

Steady state $\Leftrightarrow$ $\rho = \dfrac{\lambda}{N\mu} < 1$, servers' utilization.

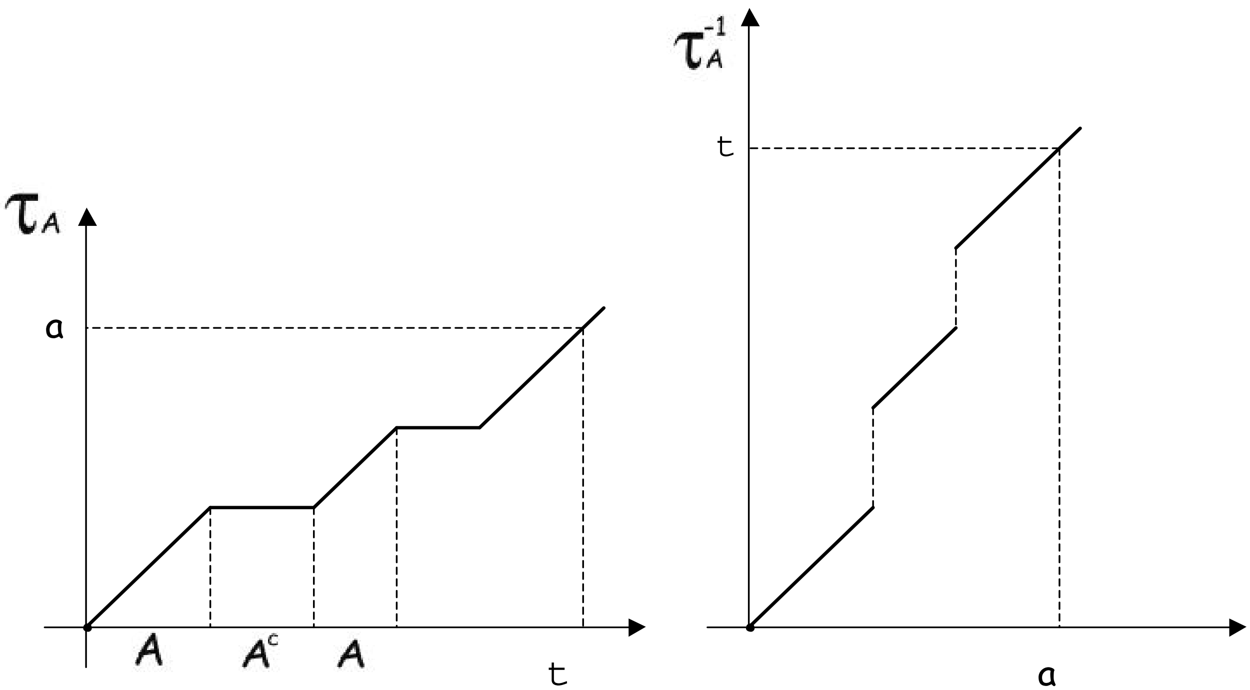
Erlang-C $$P\{\text{Wait} > 0\} \doteq \boldsymbol{E_{2,N}} = \sum_{k \geq N} \pi_k$$

$$W(\infty) \mid W(\infty) > 0 \overset{d}{=} \exp\left(\text{mean} = \frac{1}{N\mu(1-\rho)}\right)$$

Restriction to a Set via Time-Change

X Markov , $X(\infty) \sim \pi$

$\tau_A(t) = \int_0^t 1_{\{X(u)\in A\}} du$ time in A



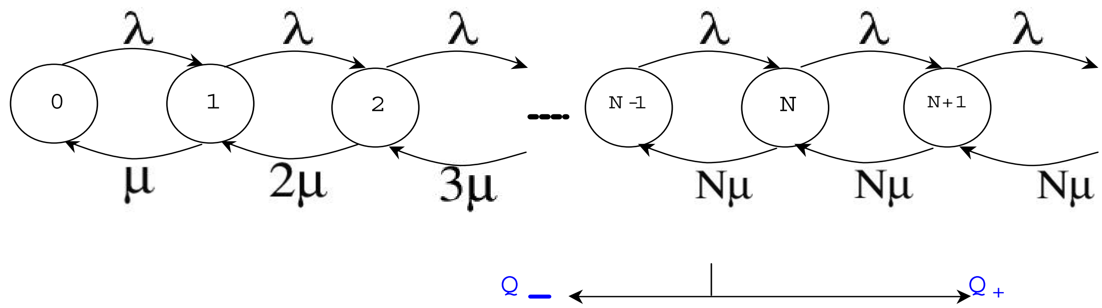
$X_A(t) = X\left(\tau_A^{-1}(t)\right)$ X restricted to A: Markov

$X_A(\infty) \sim X(\infty)|X(\infty) \in A$

Example: M/M/N/N $\stackrel{d}{=}$ M/M/∞ restricted to $\{0, 1, \ldots, N\}$.

$$\begin{aligned} E_{1,N} &= Pr\{X_R = N\}/Pr\{X_R \le N\}, \\ X_R &\sim \text{Poisson } (R) \end{aligned}$$

Example: M/M/N (Erlang-C)



$Q(t) =$ number in system at time $t \geq 0$

$Q_- = Q$ restricted to $\{0, 1, \ldots, N-1\}$: M/M/N-1/N-1 ; λ, μ.

$Q_+ = Q$ restricted to $\{N, N+1, \ldots\}$: M/M/1 ; $\lambda, N\mu$.

Evolution of Q: Alternates between M/M/1 (Q_+) and M/M/N-1/N-1 (Q_-).

$$P(\text{Wait} > 0) = \boldsymbol{E_{2,N}} = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} \quad , \qquad \text{by PASTA,}$$

where (after conditioning on the first step):

$$T_{N,N-1} = \frac{1}{\mu_N \pi_+(0)} = \frac{1}{N\mu(1-\rho)} \quad ; \quad \rho = \frac{\lambda}{N\mu}$$

$$T_{N-1,N} = \frac{1}{\lambda_{N-1}\pi_-(N-1)} = \frac{1}{\lambda\, E_{1,N-1}} \quad .$$

Hence,

$$E_{2,N} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \frac{1-\rho}{\rho\, E_{1,N-1}}\right]^{-1},$$

in which

$$\boldsymbol{E_{1,N}} = \frac{R^N}{N!} \Bigg/ \sum_{k=0}^{N} \frac{R^k}{k!} \qquad ; \qquad R = N\rho = \frac{\lambda}{\mu}$$

M/M/N/N (Erlang-B) with Many Servers: $\mathrm{N} \uparrow \infty$

Assume: μ fixed, while $\lambda_N \uparrow \infty$ as $N \uparrow \infty$.

Recall: $R = R_N = \lambda_N / \mu$, $\rho = \rho_N = R_N / N$.

Erlang-B: $E_{1,N} = Pr\{X_R = N\}/Pr\{X_R \le N\}$,

where $X_R \stackrel{d}{=}$ Poisson (R).

R large $X_R \stackrel{d}{\approx} R + Z\sqrt{R}$, where $Z \stackrel{d}{=} N(0,1)$,

suggesting $N \sim R + \beta\sqrt{R}$, $-\infty < \beta < \infty$.

Then

$$
\begin{aligned}
E_{1,N} &= Pr\{N-1 < X_R \le N\}/Pr\{X_R \le N\} \\
&\sim Pr\left\{\beta - \frac{1}{\sqrt{R}} < Z \le \beta\right\}/Pr\{Z \le \beta\} \\
&\sim \frac{1}{\sqrt{R}}\,\varphi(\beta)/\phi(\beta) \approx \frac{1}{\sqrt{N}}\,\frac{\varphi(\beta)}{\phi(\beta)} = \\
&= \frac{1}{\sqrt{N}}\,\frac{\varphi(-\beta)}{1-\phi(-\beta)} = \frac{1}{\sqrt{N}}\,h(-\beta), \quad h = \text{hazard rate.}
\end{aligned}
$$

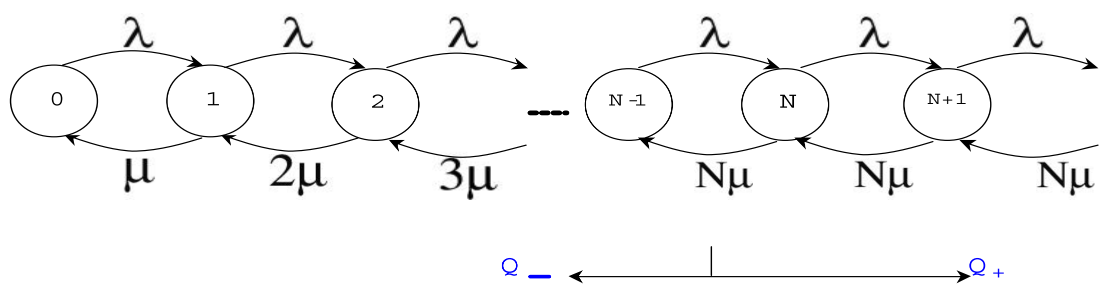
Algebra: μ fixed, $N \uparrow \infty$.

QED:

$$
\begin{aligned}
& N \sim R + \beta\sqrt{R}, \quad \text{for some } \beta, -\infty < \beta < \infty \\
\Leftrightarrow\ & \lambda_N \sim \mu N - \beta\mu\sqrt{N} \\
\Leftrightarrow\ & \rho_N \sim 1 - \frac{\beta}{\sqrt{N}}, \quad \text{namely} \quad \lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) = \beta.
\end{aligned}
$$

Theorem: (Jagerman, 1974) QED $\Leftrightarrow \lim_{N\to\infty} \sqrt{N}\, E_{1,N} = h(-\beta)$.

M/M/N (Erlang-C) with Many Servers: $\mathrm{N} \uparrow \infty$



$Q(0) = N$: all servers busy, no queue.

Recall $$E_{2,N} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} = \left[1 + \frac{1-\rho_N}{\rho_N E_{1,N-1}}\right]^{-1}.$$

Here $$T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1/\mu}{h(-\beta)\sqrt{N}}$$

which applies as $\sqrt{N}\,(1-\rho_N) \to \beta,\ -\infty < \beta < \infty$.

Also $$T_{N,N-1} = \frac{1}{N\mu(1-\rho_N)} \sim \frac{1/\mu}{\beta\sqrt{N}}$$

which applies as above, but for $0 < \beta < \infty$.

Hence, $$E_{2,N} \sim \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1},$$ assuming $\beta > 0$.

QED: $N \sim R + \beta\sqrt{R}$ for some β, $0 < \beta < \infty$

$\Leftrightarrow \quad \lambda_N \sim \mu N - \beta\mu\sqrt{N}$

$\Leftrightarrow \quad \rho_N \sim 1 - \dfrac{\beta}{\sqrt{N}}$, namely $\lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) = \beta$.

Theorem (Halfin-Whitt, 1981) QED $\Leftrightarrow \lim_{N\to\infty} E_{2,N} = \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}$.

QED M/M/N: Steady-State

Theorem (Halfin-Whitt, 1981)

Consider a sequence M/M/N, N = 1,2,. . .. Then the following are equivalent

- $\lim_{N\to\infty} P\{W_N(\infty) > 0\} = \alpha,$ for some $0 < \alpha < 1$;
- $\lim_{N\to\infty} \sqrt{N}\,(1-\rho_N) = \beta,$ for some $0 < \beta < \infty$;
- $\lambda_N = \mu N - \mu\beta\sqrt{N} + o\left(\sqrt{N}\,\right),$ i.e., $N \sim R_N + \beta\sqrt{R_N}$,

in which case

- $\alpha = \alpha(\beta) = \left[1 + \dfrac{\beta}{h(-\beta)}\right]^{-1}$ Halfin-Whitt function
- $\sqrt{N}\, W_N(\infty) \xrightarrow{d} \widehat{W}(\infty)$, $\widehat{W}(\infty) \mid \widehat{W}(\infty) > 0 \overset{d}{=} \exp(\mu\beta)$.

Moreover (Queue must be order $\sqrt{N}$)

$$\left(\frac{1}{\sqrt{N}}\,[Q_N(\infty) - N]^+, \sqrt{N}\,W_N(\infty)\right) \xrightarrow{d} \left(\widehat{Q}^+(\infty),\ \widehat{W}(\infty)\right)$$

where $\widehat{Q}^+(\infty) = \mu\widehat{W}(\infty) \overset{d}{=} \exp(\beta)$.

Proof: Let $\wedge_N(\cdot)$ be Poisson(λ_N). Then divide by $\sqrt{N}$ and take $N \uparrow \infty$ in

$$\wedge_N[W_N(\infty)] \overset{d}{=} [Q_N(\infty) - N]^+ \ \text{(Haji+Newell, 1971)}.$$

QED M/M/N : Process View

Framework : Sequence of M/M/N systems,

indexed by $N = 1, 2, \ldots$

- $Q_N = \{Q_N(t),\ t \geq 0\}$ number in system
- $V_N = \{V_N(t),\ t \geq 0\}$ virtual waiting time, under FCFS

Parameters	$\lambda_N\ ,\ \mu_N \equiv \mu$
Offered load	$R_N = \lambda_N \times \frac{1}{\mu} = \lambda_N / \mu$
Traffic intensity	$\rho_N = R_N / N$

Each M/M/N is a Birth & Death process, which is ergodic iff $\rho_N < 1$, in which case $\rho_N =$ servers' utilization.

QED Scaling : $\lambda_N = N\mu - \beta\mu\sqrt{N}$, namely

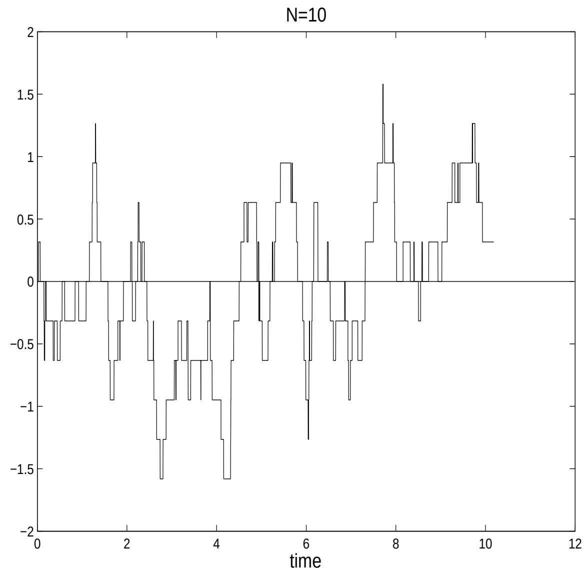
$$\rho_N = 1 - \frac{\beta}{\sqrt{N}}$$

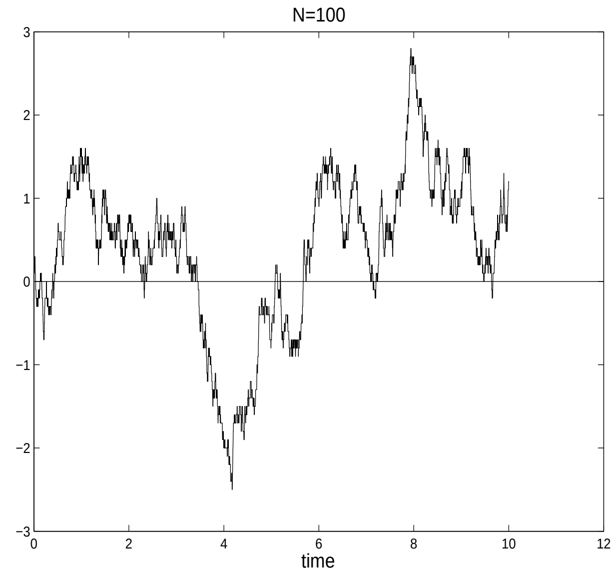
Approximations of Process and Stationary Distribution, as $N \uparrow \infty$:

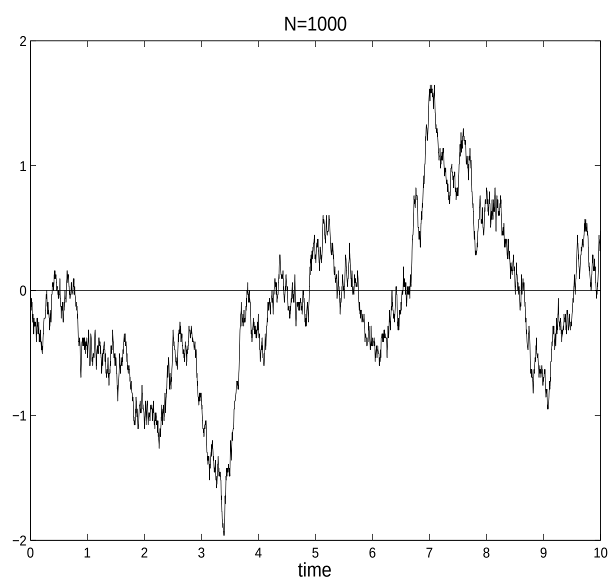
- $\hat{Q}_N(t) = \frac{1}{\sqrt{N}}[Q_N(t) - N]\ ,\quad 0 \leq t \leq \infty$
- $\hat{V}_N(t) = \sqrt{N} V_N(t).$

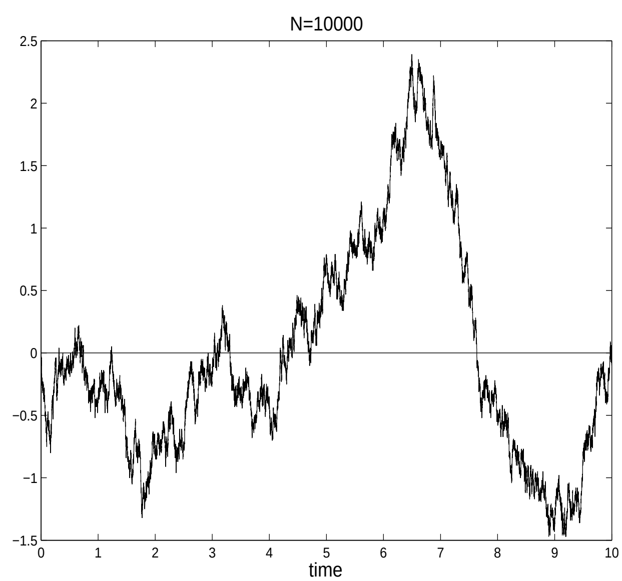
Number in System, Centered and Rescaled: $\hat{Q}_N(t)$

$(\mu = 1, \beta = 0.5)$









Approximating Queueing and Waiting

- $Q_N = \{Q_N(t),\ t \geq 0\}$: $Q_N(t) =$ number in system at $t \geq 0$.

- $\widehat{Q}_N = \{\widehat{Q}_N(t), t \geq 0\}$: stochastic process obtained by centering and rescaling:

$$\widehat{Q}_N = \frac{Q_N - N}{\sqrt{N}}$$

- $\widehat{Q}_N(\infty)$: stationary distribution of $\widehat{Q}_N$

- $\widehat{Q} = \{\widehat{Q}(t), t \geq 0\}$: process defined by: $\widehat{Q}_N(t) \xrightarrow{d} \widehat{Q}(t)$.

$$\begin{tikzcd}
\widehat{Q}_N(t) \arrow[r, "t \to \infty"] \arrow[d, "N \to \infty"'] & \widehat{Q}_N(\infty) \arrow[d, "N \to \infty"] \\
\widehat{Q}(t) \arrow[r, "t \to \infty"'] & \widehat{Q}(\infty)
\end{tikzcd}$$

Approximating (Virtual) Waiting Time

$$\widehat{V}_N = \sqrt{N}\ V_N \Rightarrow \widehat{V} = \left[\frac{1}{\mu}\ \widehat{Q}\right]^+ \qquad \text{(Puhalskii, 1994)}$$

Diffusion Processes in $\mathbb{R}^1$

$X = \{X_t,\ t \geq 0\}$ Markov process with continuous sample paths

Kolmogorov/Feller/Dynkin: characterized (Strong Markov, Continuous) by

- Drift function $\mu_t(x)$ (infinitesimal mean)

$$E\left[X_{t+\epsilon} - X_t \mid X_t = x\right] = \mu_t(x)\epsilon + o(\epsilon), \quad \epsilon \downarrow 0;$$

- Diffusion function $\sigma_t(x)$ (infinitesimal variance)

$$\mathrm{Var}\left[X_{t+\epsilon} - X_t \mid X_t = x\right] = \sigma_t(x)\epsilon + o(\epsilon), \quad \epsilon \downarrow 0;$$

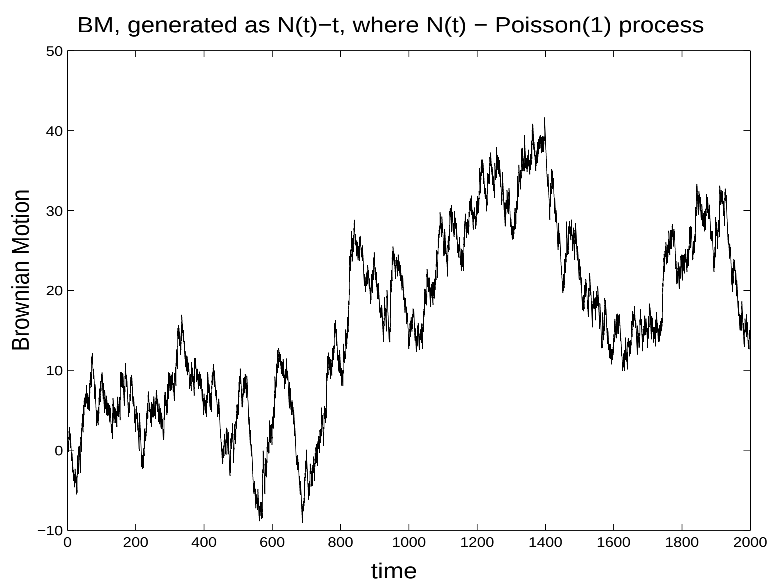
- Boundary behavior: **inaccessible**; absorbing, sticky, **reflecting**

Time-homogeneous: $\mu_t(x) \equiv \mu(x), \qquad \sigma_t(x) \equiv \sigma(x).$

Examples:

1. $\mu(x) \equiv \mu$; $\sigma(x) \equiv \sigma > 0$ Brownian Motion: BM (μ, σ^2)

   $\mu = 0 \qquad \sigma = 1$ Standard BM (SBM)



Diffusions: Examples (Continued)

2. $\mu(x) = a - bx,\ \sigma(x) \equiv \sigma$ Orenstein-Uhlenbeck (OU)

   Stationary distribution $\Leftrightarrow\ b > 0 : N(\frac{a}{b}, \frac{\sigma^2}{2b})$

3. Reflected BM (μ, σ^2), reflected at 0. (Harrison's book)

$$\begin{cases} dX_t = \mu dt + \sigma dB_t + dY_t, \quad X_0 \geq 0 \\ X \geq 0,\ dY \geq 0,\ XdY = 0 \end{cases}$$

   Stationary distribution $\Leftrightarrow \mu < 0 :\ \exp(2|\mu|/\sigma^2)$

4. Reflected OU on $[0, \infty)$, (Glynn and Ward), or generally (Dupuis)

$$\begin{cases} dX_t = \mu_t(X_t)dt + \sigma_t(X_t)dB_t + dY_t; \quad X_0 \geq 0 \text{ given} \\ X \geq 0,\ dY \geq 0,\ XdY = 0 \end{cases}$$

Ito: Characterized by solutions to stochastic differential equations

$$dX_t = \mu_t(X_t)dt + \sigma_t(X_t)dB_t, \quad t \geq 0; \quad X_0 \text{ given.}$$

(Reference: Karlin and Taylor, "2nd Course"; Karatzas and Shreve.)

Formalizes the "infinitesimal description":

$$X_{t+\epsilon} - X_t \mid X_t = x \overset{d}{\approx} \mu_t(x) \cdot \epsilon + \sigma_t(x) \cdot (B_{t+\epsilon} - B_t)$$

where $B_{t+\epsilon} - B_t \sim N(0, \epsilon)$, independent of $\{B_u,\ u \leq t\}$.

(Equivalently, $B = \{B_t,\ t \geq 0\}$ is SBM.)

QED M/M/N : Diffusion Approximation

$Q_N =$	$\{Q_N(t),\ t \geq 0\}$	Birth and Death with
$\lambda_N =$	$\mu N - \mu\beta\sqrt{N}$	birth rate, constant
$\mu_N(x) =$	$\mu \cdot (x \wedge N)$	death rate at state x
$\hat{Q}_N(t) =$	$\frac{1}{\sqrt{N}}\,[Q_N(t) - N]$	Scaled queue length of customers $(+)$ or servers $(-)$

Diffusion: Var $\left[\hat{Q}_N(t+\epsilon) - \hat{Q}_N(t) \mid \hat{Q}_N(t) = x\right] = 2\mu \cdot \epsilon + o(\epsilon)$

Drift: $E\left[\hat{Q}_N(t+\epsilon) + \hat{Q}_N(t) \mid \hat{Q}_N(t) = x\right] =$

$$\frac{1}{\sqrt{N}}\ E\left[Q_N(t+\epsilon) - Q_N(t) \mid Q_N(t) = N + x\sqrt{N}\ \right] =$$

$$\frac{1}{\sqrt{N}}\ \left[\lambda_N - \mu_N\left(N + x\sqrt{N}\ \right)\right] \cdot \epsilon + o(\epsilon) =$$

$$\frac{1}{\sqrt{N}}\ \left\{\mu N - \mu\beta\sqrt{N}\ - \mu\left[\left(N + x\sqrt{N}\ \right) \wedge N\right]\right\} \cdot \epsilon + o(\epsilon) =$$

$$\left.\begin{array}{llll} = & -\mu\beta \cdot \epsilon + o(\epsilon) & x \geq 0 & \text{(BM)} \\ = & -\mu(\beta + x) \cdot \epsilon + o(\epsilon) & x \leq 0 & \text{(OU)} \end{array}\right\} = -\mu(\beta - x^-) \cdot \epsilon + o(\epsilon)$$

Expect $\hat{Q}_N \xrightarrow{d} \hat{Q}$ diffusion: $\mu(x) = -\mu(\beta - x^-),\ \sigma(x) = 2\mu$

Proof: Apply Stone (1963), as in Halfin-Whitt (1981).

QED M/M/N : Diffusion Approximation

Theorem (Halfin-Whitt).

Consider a sequence M/M/N, N = 1,2,. . .. Assume QED(β).

Define $\hat{Q}_N(t) = \frac{1}{\sqrt{N}} [Q_N(t) - N]$, $0 \le t < \infty$,

$\hat{V}_N(t) = \sqrt{N}\, V_N(t)$, $0 \le t < \infty$.

If $\left(\hat{Q}_N(0),\, \hat{V}_N(0)\right) \xrightarrow{d} \left(\hat{Q}(0),\, \hat{V}(0)\right)$, then

$$\left(\hat{Q}_N,\, \hat{V}_N\right) \xrightarrow{d} \left(\hat{Q},\, \frac{1}{\mu}\,\hat{Q}^+\right) \quad \text{(Functional CLT)},$$

where $\hat{Q}$ is a diffusion process starting at $\hat{Q}(0)$, with infinitesimal parameters

$$\mu(x) = \begin{cases} -\mu\beta & x \ge 0 \\ -\mu(x+\beta) & x < 0 \end{cases} , \quad \sigma^2(x) = 2\mu \,;$$

and steady-state distribution $\hat{Q}(\infty)$ given by

$$P\{\hat{Q}(\infty) \ge 0\} = \alpha(\beta)\,,$$

$$P\{\hat{Q}(\infty) > x \mid \hat{Q}(\infty) \ge 0\} = e^{-x\beta} \qquad (\exp \leftrightarrow \text{RBM})$$

$$P\{\hat{Q}(\infty) \le x \mid \hat{Q}(\infty) \le 0\} = \phi(x+\beta)/\phi(\beta) \qquad (\text{normal} \leftrightarrow \text{OU})$$

E-Driven M/M/N : Approximations

Recall: 1:1 staffing gave rise to $\lambda_N \sim \mu N - \mu\gamma$

$$\lim_{N\to\infty} N(1-\rho_N) = \gamma, \quad 0 < \gamma < \infty.$$

Applying the usual QED scaling: $\hat{Q}_N(t) = \frac{1}{\sqrt{N}}(Q_N(t) - N)$,

results in a Halfin-Whitt diffusion limit with $\beta = 0$: No stationary distribution for

$$\hat{Q}^+(t) = \lim_{N\to\infty} \frac{1}{\sqrt{N}}(Q_N(t) - N)^+.$$

To identify a more informative scaling, recall:

$$W_N(\infty) \mid W_N(\infty) > 0 \overset{d}{=} \exp\left(\text{mean} = \frac{1}{N\mu(1-\rho_N)}\right).$$

Also,

$$\lim_{N\to\infty} P\{W_N(\infty) > 0\} = 1.$$

Hence,

$$W_N(\infty) \xrightarrow{d} \hat{W}(\infty) \overset{d}{=} \exp(\mu\gamma).$$

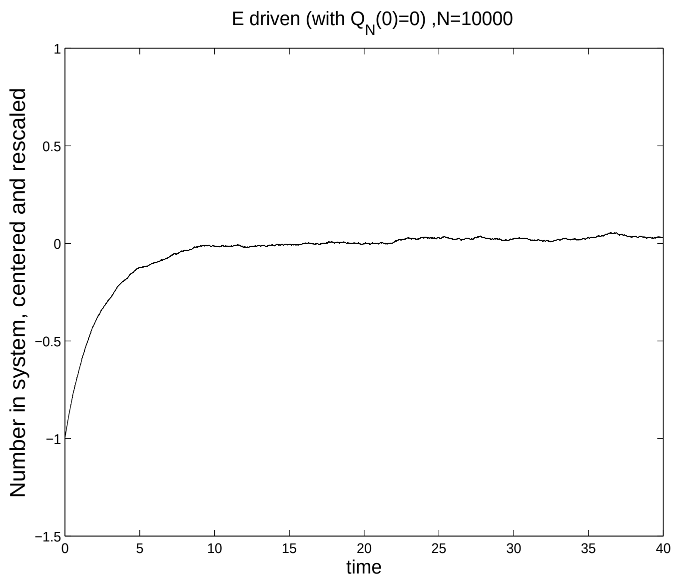
And (by Haji & Newell)

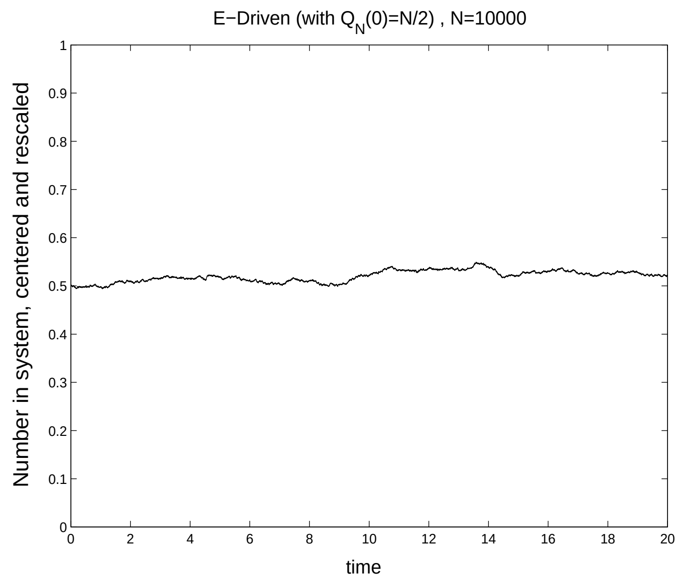
$$\frac{1}{N}\left[Q_N(\infty) - N\right]^+ \xrightarrow{d} \hat{Q}(\infty)^+ \overset{d}{=} \mu\hat{W}(\infty) \overset{d}{=} \exp(\gamma)\,.$$

Expect a non-degenerate behavior of

$$\bar{Q}_N(t) = \frac{1}{N}[Q_N(t) - N] \xrightarrow{d} \bar{Q}(t) \ , \ 0 \le t \le \infty$$

Process View: Simulations of $\bar{Q}_N(t)$





Proposition: Let $\lim_{N\to\infty} N(1-\rho_N) = \gamma, \quad 0 < \gamma < \infty.$

If $\frac{1}{N}[Q_N(0) - N] \xrightarrow{d} \bar{Q}(0)$, then

$$\bar{Q}_N(t) \xrightarrow{d} \bar{Q}(t) = \begin{cases} \bar{Q}(0) \quad , & \text{if } \bar{Q}(0) \geq 0, \\ \bar{Q}(0)e^{-\mu t} , & \text{if } \bar{Q}(0) \leq 0. \end{cases}$$

Proof:

Diffusion: $\quad \text{Var}\left[\bar{Q}_N(t+\epsilon) - \bar{Q}_N(t) \mid \bar{Q}_N(t) = x\right] = 0 + o(\epsilon)$

Drift: $\quad E\left[\bar{Q}_N(t+\epsilon) + \bar{Q}_N(t) \mid \bar{Q}_N(t) = x\right] =$

$$\left.\begin{array}{lll} = & 0 \cdot \epsilon + o(\epsilon) & x \geq 0 \\ = & -\mu x \cdot \epsilon + o(\epsilon) & x \leq 0 \end{array}\right\} = \mu x^- \cdot \epsilon + o(\epsilon).$$

Conclude $\bar{Q}_N \xrightarrow{d} \bar{Q}$: $\mu(x) = \mu x^-, \ \sigma(x) \equiv 0$.

Discovered that $\bar{Q}_N(t), 0 \le t < \infty$ degenerates, as $N \uparrow \infty$. But

$$\bar{Q}_N(\infty) = \frac{1}{N}[Q_N(\infty) - N]^+ \xrightarrow{d} \exp(\gamma).$$

What's going on?

Two explanations (assuming $\bar{Q}(0) \ge 0$, non-random, for simplicity):

1. For $k \ge N$ (with $\lambda = N\mu - \gamma\mu$),

$$T_{k+cN,k} = \theta(N) * T_{k+1,k} = \theta(N) * \theta(1) \to \infty, \quad \text{as } N \uparrow \infty.$$

Hence, $\bar{Q}_N(t)$ takes close to infinite time to move a one-unit distance, suggesting that, in the limit, $\bar{Q}(t)$ freezes.

2. Consider Q_N^+: the restriction of the Q_N to $\{N, N+1, ..\}$.

$Q_N^+ \stackrel{d}{=} M/M/1$ with $\lambda^+ = N\mu - \gamma\mu$, $\mu^+ = N\mu$.

$\stackrel{d}{=} M/M/1$ with $\bar{\lambda} = \frac{\lambda^+}{N} \to \mu$, $\bar{\mu} = \frac{\mu^+}{N} \to \mu$, accelerated by N.

Formally, $\bar{Q}_N^+(t) = \frac{1}{N}Q_N^+(t) \xrightarrow{d} \bar{Q}(0)^+$ is a **fluid limit** , since

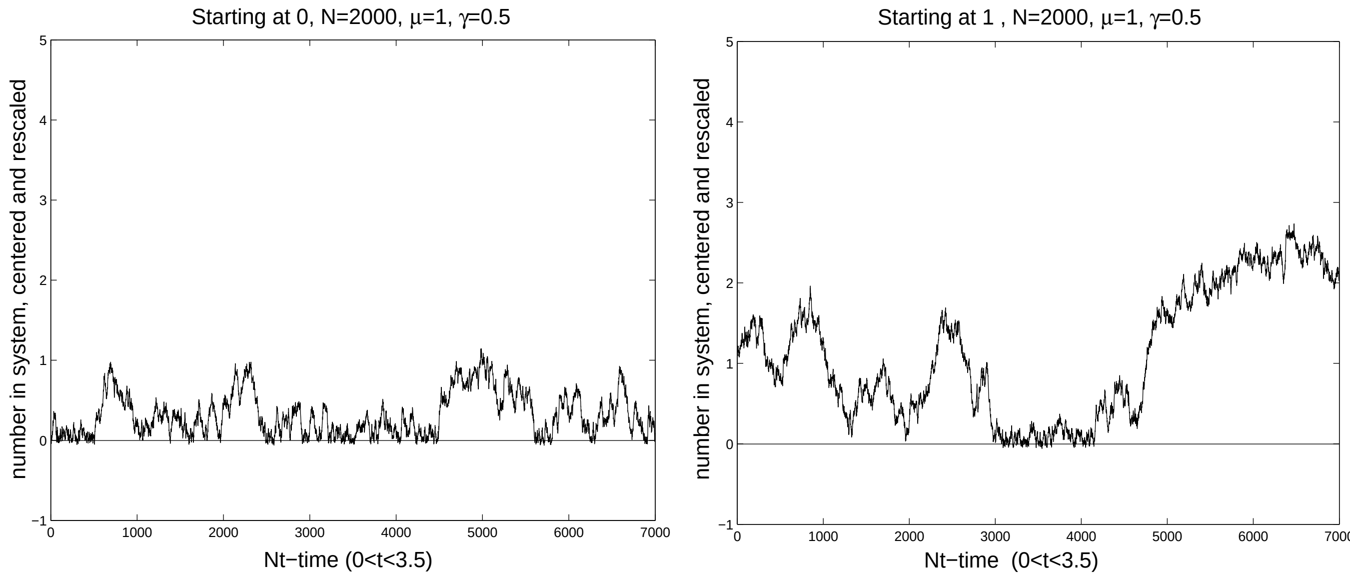
$$\bar{Q}_N^+(t) = \frac{1}{N}Q_N^+(t)(\text{ **with** } \lambda^+, \mu^+) \quad \stackrel{d}{=} \quad \frac{1}{N}Q_N^+(Nt)(\text{ **with** } \bar{\lambda}, \bar{\mu})$$

The limit is deterministic, hence a degenerate stationary distribution.

To get a **diffusion limit**, **accelerate**:

$$\hat{Q}_N^+(t) = \frac{1}{N}Q_N^+(Nt)(\text{ **with** } \lambda^+, \mu^+) \stackrel{d}{=} \frac{1}{N}Q_N^+(N^2 t)(\text{ **with** } \bar{\lambda}, \bar{\mu})$$

Simulations of $\hat{Q}_N(t) = \frac{1}{N}[Q_N(Nt) - N]$



Theorem: Let $\lim_{N\to\infty} N(1-\rho_N) = \gamma, \quad 0 < \gamma < \infty$.

Define $\hat{Q}_N(t) = \frac{1}{N}[Q_N(Nt) - N]$. Then

$$\hat{Q}_N(t) \xrightarrow{d} \hat{Q}(t) \stackrel{d}{=} \mathrm{RBM}(-\gamma\mu, 2\mu).$$

Proof:

Diffusion: $\sigma_N(x) \to \sigma(x) = (2 - x^-)\mu$

Drift: $\mu_N(x) \to \mu(x) = \begin{cases} -\mu\gamma & x \geq 0 \\ +\infty & x \leq 0 \end{cases}$

Via random time-change: $\hat{Q}(t) \stackrel{d}{=}$ RBM $(-\gamma\mu, 2\mu)$.

(RBM $\geq$ 0: Since the infinitesimal drift "is $+\infty$" at $x \leq 0$.)

Note: $\hat{Q}(\infty) \stackrel{d}{=} exp(\gamma)$, establishing convergence over $0 \leq t \leq \infty$.

Note: $V_N(t) \xrightarrow{d} \hat{V}(t) \stackrel{d}{=} \frac{1}{\mu}\hat{Q}(t),\ 0 \leq t \leq \infty$: no scaling needed.

(Check at $t = \infty$, using Haji & Newell).