On the Analysis of Queueing Models with Abandonments

Amy Ward
Marshall School of Business, USC

INFORMS 2011 Tutorial

Reference: Ward (2011)
Asymptotic Analysis of Queueing Systems with Reneging: A Survey of Results for FIFO, Single Class Models
Surveys in OR and MS

Queueing models with abandonments arise in many application contexts, such as

Hospital Emergency Rooms

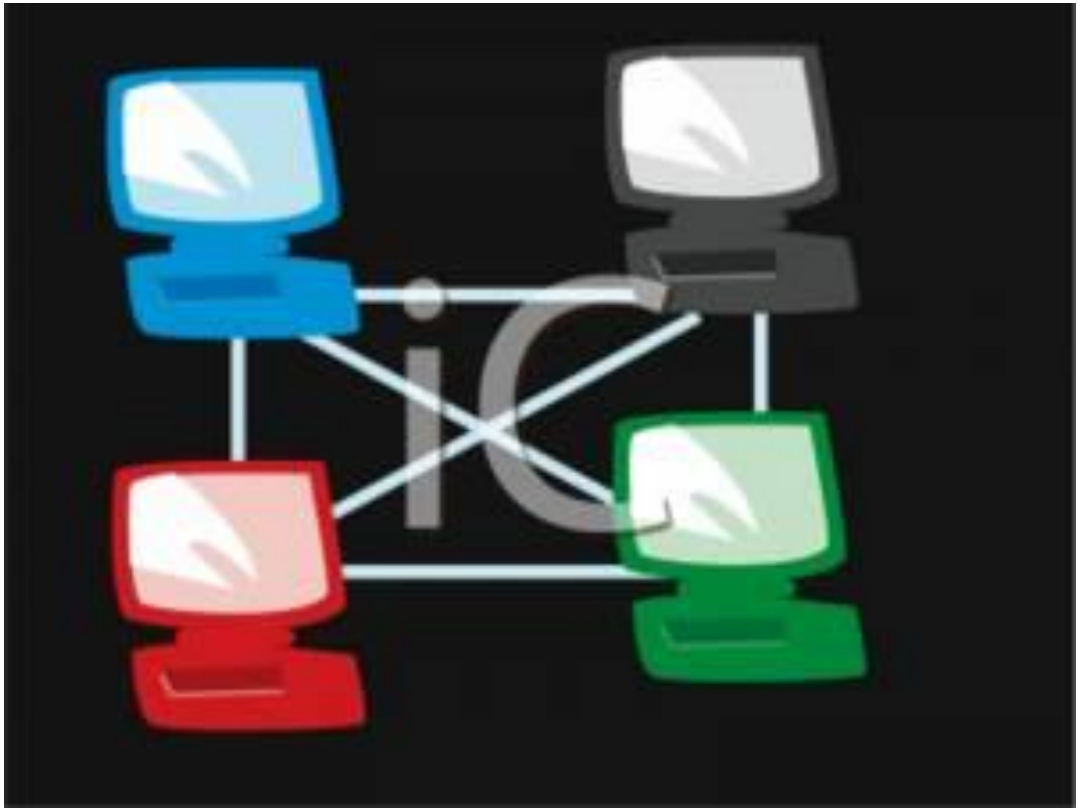
(Green, Soares, Giglio, and Green, 2006)

Queueing models with abandonments arise in many application contexts, such as

Call Centers

(Reference the review papers of Aksin, Armony, and Mehrotra, 2007; Gans, Koole, and Mandelbaum, 2003)

Queueing models with abandonments arise in many application contexts, such as



Communications Networks with Time-Critical Traffic
(Bhattacharya and Ephremides, 1989; Panward, Towsely, and Wolf, 1988)

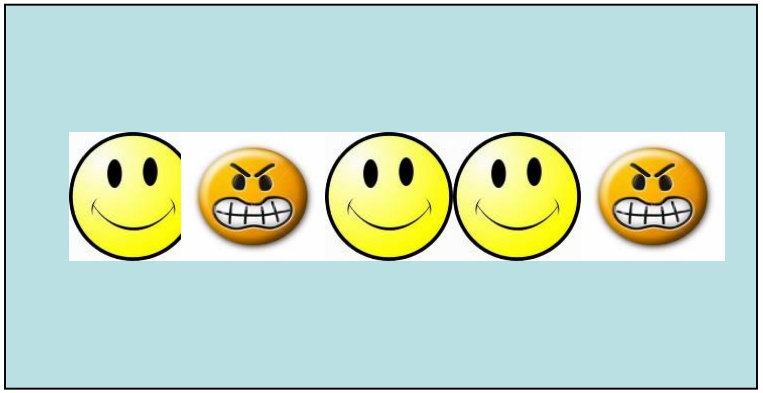
Queueing models with abandonments arise in many application contexts, such as

Inventory System with Perishable Goods

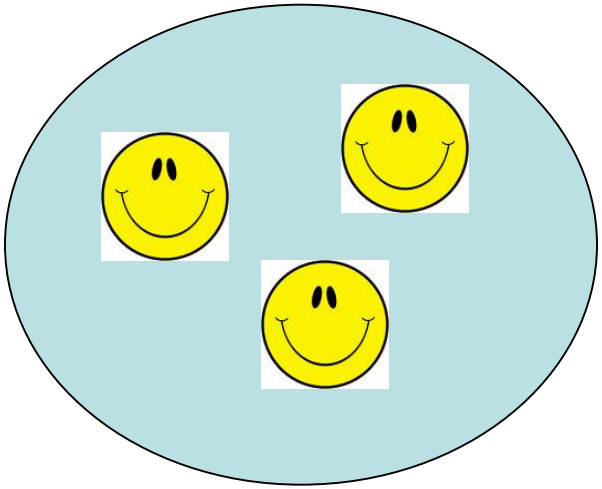
(Nahmias, 1982; Kaspi and Perry, 1983)

The Underlying Queueing Model

Waiting Area



Service Area

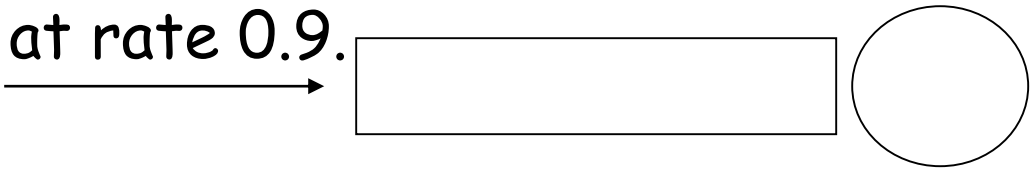


Do the abandonments matter from a modeling perspective? After all, it should be that few customers abandon.

Yes, Abandonments Matter

Exponential Service with mean 1.

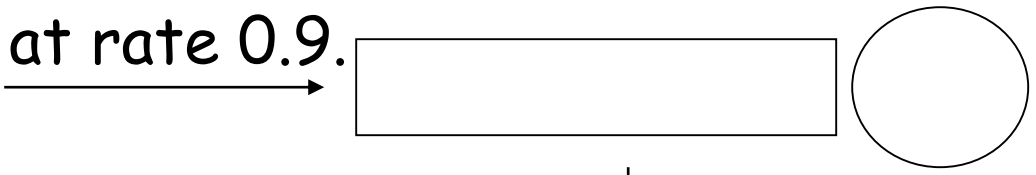
Poisson arrivals at rate 0.9.



Expected steady-state waiting time = 10.

Exponential Service with mean 1.

Poisson arrivals at rate 0.9.



Exponential Abandonment times at rate 0.01.

4.52% of customers abandon.
Expected steady-state waiting time = 5.48.

(Palm (1937) is the earliest to model customer abandonment.)

How did we predict performance?

Let Q be the steady-state number of customers in system.

M/M/1 Steady-state distribution

$$P(Q=n)=\rho^n(1-\rho),\quad n=1,2,3,\ldots$$

M/M/1+M Steady-state distribution

$$P(Q=n)=\frac{\lambda^n}{\prod_{j=0}^{n-1}(\mu+j\gamma)}P(Q=0)$$

Normalization constant

What if the distributions are not exponential?

If we do not assume exponential distributions,

There are some non-asymptotic results available.

- Baccelli, Boyer, Hebuterne (1984)
- Bae, Kim, and Lee (2001)
- Barrer (1957)
- Boots and Tijms (1999)
- Boxma, Perry, and Stadje (2010)
- Boxma, Perry, Stadje, and Zacks (2009)
- Brandt and Brandt (1999)
- Finch (1960)
- Gavish and Schweitzer (1977)
- Gnedenko and Kovalenko (1968)
- Jurkevic (1971)
- Movaghar (1998)
- Perry and Asmussen (1995)
- Rao (1967)
- Stanford (1979)

However, obtaining convenient closed form expressions for either steady-state or transient performance measures is in general not possible.

Are there convenient analytic approximation formulae when we do not assume exponential distributions?

Our Objective and Methdology

Our objective:

To develop simple performance measure approximation formulae for the GI/GI/N+GI queue.

Our methodology:

To find analytically tractable diffusion processes that approximate the GI/GI/N+GI queue-length process

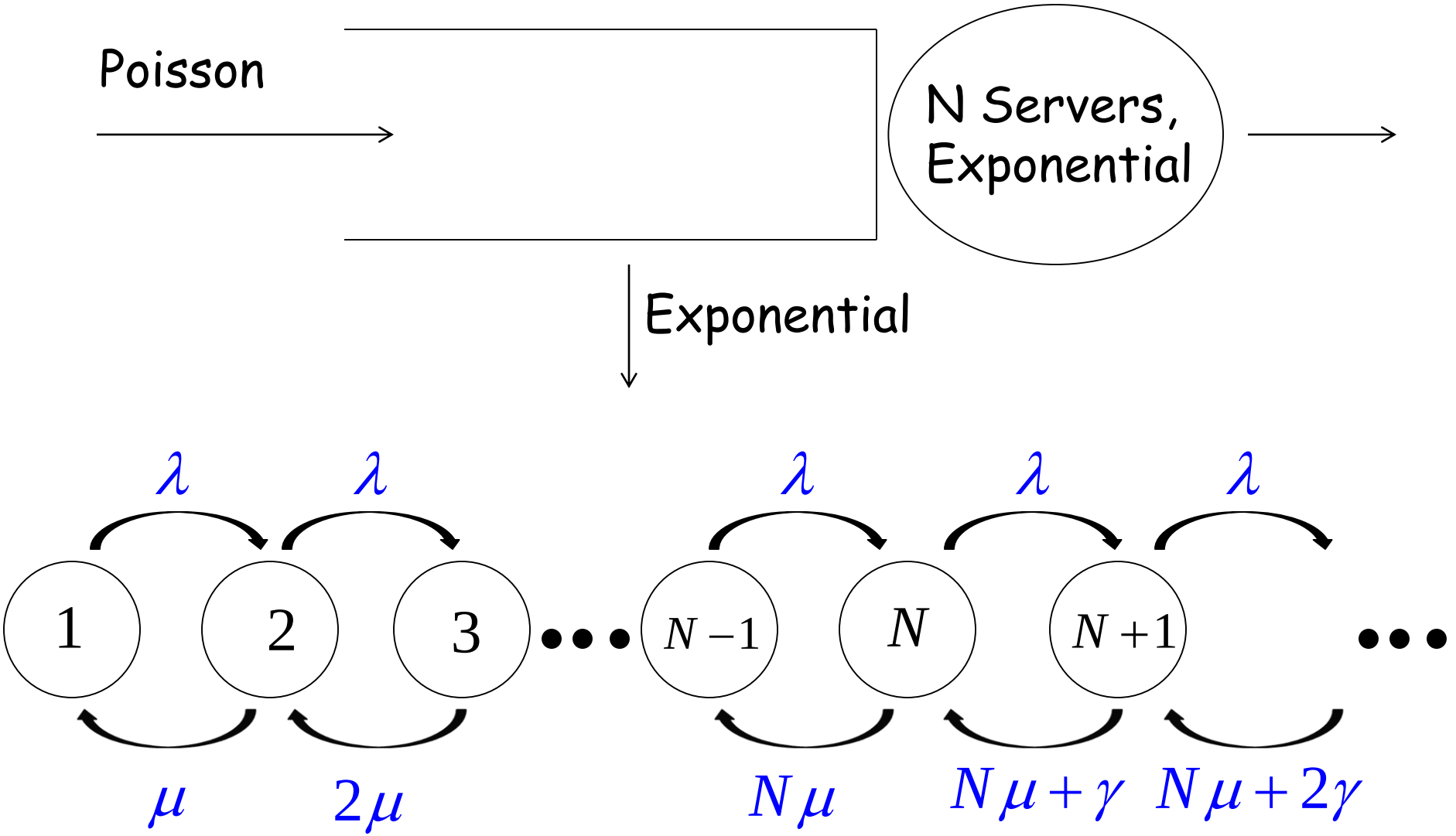
Talk Outline

- The M/M/N+M Queueing Model
- Diffusion approximations for the M/M/N+M Queueing Model
  - Conventional Heavy-Traffic
  - The Halfin-Whitt Many-Server Regime
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- Intermediate Regimes
- Unifying diffusion approximations

Talk Outline

- **The M/M/N+M Queueing Model**
- Diffusion approximations for the M/M/N+M Queueing Model
  - Conventional Heavy-Traffic
  - The Halfin-Whitt Many-Server Regime
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- Intermediate Regimes
- Unifying diffusion approximations

The M/M/N+M Queueing Model



The M/M/N+M Queueing Model

Steady-State Probabilities:

$$P(Q=n)=\begin{cases}\dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^{n}P(Q=0) & n\in\{0,1,\ldots,N\}\\[2ex] \dfrac{1}{N!}\left(\dfrac{\lambda}{\mu}\right)^{N}\dfrac{N\mu}{\lambda}\left(\dfrac{\lambda}{\gamma}\right)^{n-N+1}\dfrac{\Gamma\left(\dfrac{N\mu}{\gamma}\right)}{\Gamma\left(\dfrac{N\mu}{\gamma}+n-N+1\right)}P(Q=0) & n\in\{N+1,N+2,\ldots\}\end{cases}$$

↑ Normalization constant

Customer Abandonment Probability:

The steady-state rate at which customers abandon.

=

The steady-state rate at which customers who abandon enter.

$$\gamma E\left[\left[Q(\infty)-N\right]^{+}\right]=\lambda P_a$$

For other steady-state performance measures of interest, such as the expected wait time conditional on being served, see Whitt (1999).

The M/M/N+M Queueing Model

Transient Analysis:

Suppose we would like to compute $E[Q(t+s) \mid Q(t) = q]$.

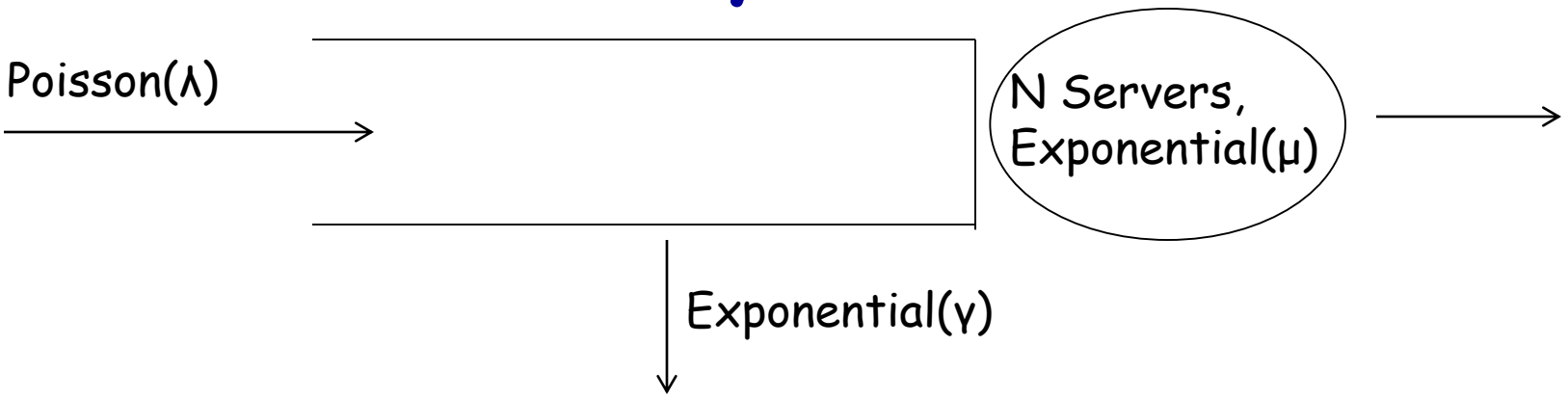
We could solve the Kolmogorov differential equations numerically. But there are no nice formulae.

Even for the birth and death M/M/N+M model, it will be nice to have a simple approximation for the queue-length process. (Remember our original motivation is the GI/GI/N+GI model.)

Talk Outline

- The M/M/N+M Queueing Model
- **Diffusion approximations for the M/M/N+M Queueing Model**
  - **Conventional Heavy-Traffic**
  - The Halfin-Whitt Many-Server Regime
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- Intermediate Regimes
- Unifying diffusion approximations

Weak Convergence Results for the M/M/N+M Queue: Preliminary Calculations



Infinitesimal mean:

$$m_Q(n) := \lim_{h \downarrow 0} \frac{1}{h} E\left[Q^\lambda(h) - Q^\lambda(0) \mid Q^\lambda(0) = n \right]$$

$$= \lambda - \min(N, n)\mu - \gamma[n - N]^+, n \in \{0, 1, 2, \ldots\}$$

Infinitesimal variance:

$$v_Q(n) := \lim_{h \downarrow 0} \frac{1}{h} E\left[\left(Q^\lambda(h) - Q^\lambda(0) \right)^2 \mid Q^\lambda(0) = n \right]$$

$$= \lambda + \min(N, n)\mu - \gamma[n - N]^+, n \in \{0, 1, 2, \ldots\}$$

The RO-U Approximation

Definition: The Conventional Heavy Traffic (HT) Limit Regime

For a given $\beta \in \Re$, define

$$u^{\lambda} := \frac{\lambda + \beta\sqrt{\lambda}}{N},$$

and assume that N is fixed and independent of λ.

The infinitesimal mean of $\frac{Q^{\lambda}}{\sqrt{\lambda}}$ is:

$$\lim_{h \downarrow 0} \frac{1}{h} E\left[\frac{Q^{\lambda}(h)}{\sqrt{\lambda}} - \frac{Q^{\lambda}(0)}{\sqrt{\lambda}} \mid \frac{Q^{\lambda}(0)}{\sqrt{\lambda}} = x \right]$$

$$= \frac{1}{\sqrt{\lambda}} m_Q\left(\sqrt{\lambda} x\right) \rightarrow m_{ROU}(x), \text{ as } \lambda \rightarrow \infty, \text{where } m_{ROU}(x) = -\beta - \gamma x, \text{ for } x \geq 0$$

The infinitesimal variance of $\frac{Q^{\lambda}}{\sqrt{\lambda}}$ is:

$$\lim_{h \downarrow 0} \frac{1}{h} E\left[\left(\frac{Q^{\lambda}(h)}{\sqrt{\lambda}} - \frac{Q^{\lambda}(0)}{\sqrt{\lambda}} \right)^2 \mid \frac{Q^{\lambda}(0)}{\sqrt{\lambda}} = x \right] = \frac{1}{\lambda} v_Q\left(\sqrt{\lambda} x\right) \rightarrow 2, \text{ as } \lambda \rightarrow \infty.$$

The RO-U approximation is a diffusion with mean $-(\beta + \gamma x)$ and variance 2, having state space $[0, \infty)$.

The RO-U Approximation Theorem

Theorem 1: The Conventional Heavy Traffic (HT) Limit

(Ward and Glynn, 2003)

In conventional heavy traffic,

$$\frac{Q^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty \text{ in D},$$

where, for $\sigma^2 = 2$,

Brownian Motion

$$X_{ROU}(t) = X_{ROU}(0) + \int_0^t m_{ROU}\left(X_{ROU}(s)\right)ds + \sigma B(t) + L_{ROU}(t) \geq 0$$

$$L_{ROU}(0) = 0, L_{ROU} \text{ is non-decreasing, and } \int_0^{\infty} X_{ROU}(t)dL_{ROU}(t) = 0.$$

The implication is that: $Q^{\lambda}(\bullet) \approx \sqrt{\lambda} X_{ROU}(\bullet).$

Example

M / M / 1 + M queue with arrival rate $\lambda = 0.9$, service rate $\mu = 1$, and abandonment rate

1/γ	P[abandon]	E[Q] exact	$\lambda^{1/2}E[X_{ROU}]$	% Error
10,000	0.09%	8.84	9.33	5.5%
1,000	0.77%	7.84	8.22	4.8%
100	4.52%	4.93	5.07	2.8%
10	15.78%	2.18	2.13	2.3%
1	34.06%	0.90	0.74	17.5%

(Note that the RBM approximation has 9.5 as the approximated E[Q].)

Is RO-U a tractable process?

- Its steady-state density is that of a normal random variable conditioned to be positive.

  (Browne and Whitt, 1995)

RBM has an exponential tail.

- Its transient distribution is available via its Sturm-Liouville spectral expansion.

  (Linetsky, 2005)

- There is an exact expression for its transient distribution in terms of the hitting time density of an unreflected O-U process.

  (Cox and Rosler, 1983)

- Its transient moments can be expressed in terms of the transient moments of RBM when γ is small.

  (Glynn and Ward, 2003)

Talk Outline

- The M/M/N+M Queueing Model
- **Diffusion approximations for the M/M/N+M Queueing Model**
  - Conventional Heavy-Traffic
  - **The Halfin-Whitt Many-Server Regime**
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- Intermediate Regimes
- Unifying diffusion approximations

Queueing models with abandonments arise in many application contexts, such as

Call Centers

The Many-Server Approximation

Definition: The Halfin-Whitt (HW) Limit Regime

For a given $\beta \in \Re$, define

$$N^\lambda := \frac{\lambda + \beta\sqrt{\lambda}}{u},$$

and assume that μ is fixed and independent of λ.

The infinitesimal mean of $\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}}$ is:

$$\lim_{h \downarrow 0} \frac{1}{h} E\left[\frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \mid \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} = x \right]$$

$$= \frac{1}{\sqrt{\lambda}} m_Q\left(\sqrt{\lambda}x + N^\lambda\right) \to m_{HW}(x), \text{ as } \lambda \to \infty, \text{where } m_{HW}(x) = \begin{cases} -\beta - \gamma x, & \text{if } x \geq 0 \\ -\beta - \mu x, & \text{if } x < 0 \end{cases}.$$

In the upper half of the state space (when there are customers waiting), the diffusion behaves exactly like the one that arises in conventional HT.

In the lower half of the state space, the diffusion drift comes from servers being idle.

The infinitesimal variance of $\frac{Q^\lambda}{\sqrt{\lambda}}$ is:

$$\lim_{h \downarrow 0} \frac{1}{h} E\left[\left(\frac{Q^\lambda(h) - N^\lambda}{\sqrt{\lambda}} - \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} \right)^2 \mid \frac{Q^\lambda(0) - N^\lambda}{\sqrt{\lambda}} = x \right] = \frac{1}{\lambda} v_Q\left(\sqrt{\lambda}x + N^\lambda\right) \to 2, \text{ as } \lambda \to \infty.$$

The Many Server Approximation Theorem

Theorem 2: The Halfin-Whitt (HW) Many Server Limit

(Garnett, Mandelbaum, and Reiman, 2002)

In the HW limit regime,

$$\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}, \text{ as } \lambda \to \infty \text{ in D},$$

where, for $\sigma^2 = 2$,

$$X_{HW}(t) = X_{HW}(0) + \int_0^t m_{HW}\left(X_{HW}(s)\right)ds + \sigma B(t).$$

The implication is that: $Q^\lambda(\bullet) \approx \sqrt{\lambda} X_{HW}(\bullet) + N^\lambda.$

(How large does N need to be? Not very – 10 is generally fine.)

Is X_{HW} a tractable process?

- Its steady-state density can be expressed in terms of the standard normal cdf and pdf.

  (Browne and Whitt, 1995)

- The Laplace transform of its transient distribution is available.

  Knessl and van Leeuwaarden (2008), $\gamma = 0$

  Knessl and van Leeuwaarden (2010), $\gamma > 0$

A Comparison of the Conventional HT and HW Limit Regimes

Conventional HT : $Q^\lambda(\bullet) \approx \sqrt{\lambda} X_{ROU}(\bullet)$	**HW :** $Q^\lambda(\bullet) \approx \sqrt{\lambda} X_{HW}(\bullet) + N^\lambda$
$\rho^\lambda := \dfrac{\lambda}{N\mu^\lambda} \to 1$ as $\lambda \to \infty$ λ is large compared to γ. (Small % of customers abandon.) Waiting times are of order $1/\sqrt{\lambda}$.	$\rho^\lambda := \dfrac{\lambda}{N^\lambda\mu} \to 1$ as $\lambda \to \infty$ λ is large compared to γ. (Small % of customers abandon.) Waiting times are of order $1/\sqrt{\lambda}$.
An arriving customer is almost certain to be delayed. Assumes λ and μ are large, and N is small. (Service times are of order $1/\lambda$.) Delay times >> service times.	The probability an arriving customer is delayed remains strictly between 0 and 1. Assumes λ and N are large, and μ is small (in comparison). (Service times are of order 1.) Delay times << service times.

Talk Outline

- The M/M/N+M Queueing Model
- Diffusion approximations for the M/M/N+M Queueing Model
  - Conventional Heavy-Traffic
  - The Halfin-Whitt Many-Server Regime
- **Diffusion approximations for the GI/GI/N+GI Queueing Models**
- Intermediate Regimes
- Unifying diffusion approximations

The GI/GI/N+GI Queueing Model

In both queues where there are a small number of servers, and queues where there are many servers, it is not clear that exponential distribution assumptions are appropriate. (There is an empirical study in Brown et al (2005).)

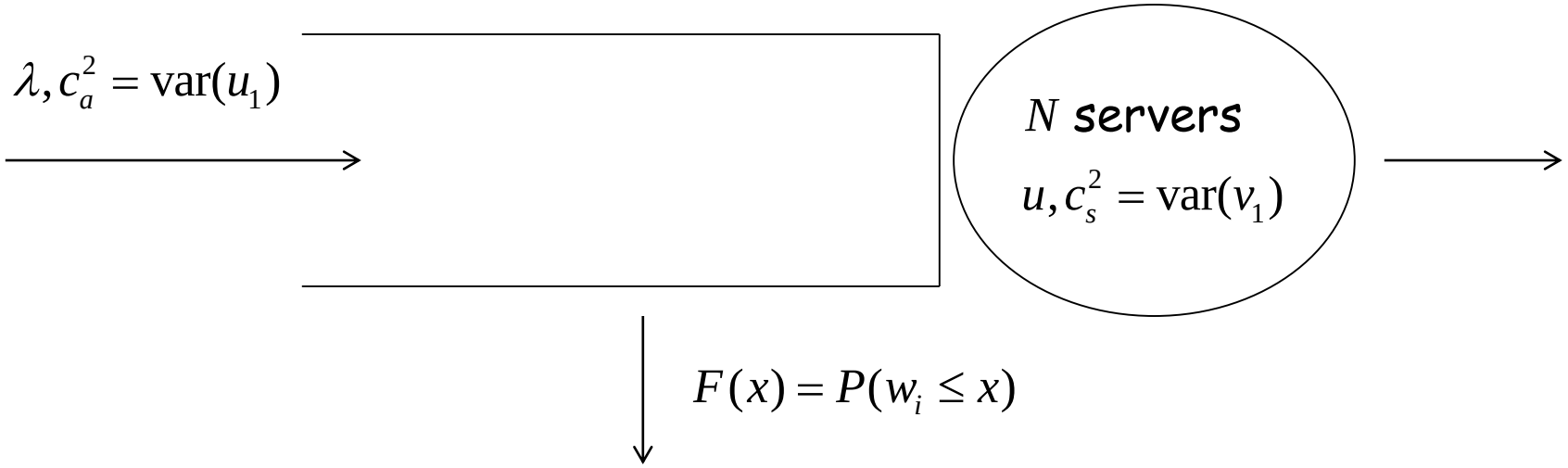
The GI/GI/N+GI Queueing Model

There are 3 independent i.i.d. sequences $\{u_i, i \geq 1\}, \{v_i, i \geq 1\}, \{w_i, i \geq 1\}$, all having mean 1 and finite variance.

The ith customer arrives at time $\sum_{j=1}^{i} \frac{u_j}{\lambda}$, has service time $\frac{v_i}{\mu}$,

and will abandon if his service does not begin within w_i time units.

Service is FIFO.



We would like simple approximations for the queue-length and abandonment probability. Are X_{ROU} and X_{HW} relevant?

The RO-U Approximation Theorem

Theorem 3: The Conventional Heavy Traffic (HT) Limit
(Ward and Glynn, 2005)

In conventional heavy traffic,

$$\frac{Q^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{ROU}, \text{ as } \lambda \to \infty \text{ in D},$$

where, for $\sigma^2 = c_a^2 + c_s^2$,

$$X_{ROU}(t) = X_{ROU}(0) - \int_0^t m_{ROU}\left(X_{ROU}(s)\right)ds + \sigma B(t) + L_{ROU}(t) \geq 0$$

$$L_{ROU}(0) = 0, L_{ROU} \text{ is non-decreasing, and } \int_0^{\infty} X_{ROU}(t)dL_{ROU}(t) = 0,$$

and

$$m_{ROU}(x) = \beta + F'(0)x, x \geq 0.$$

The same process that approximates the M/M/N+M queue in conventional HT also approximates the GI/GI/N+GI queue.

The Probability a Customer Abandons

$L^{\lambda}(t)$ is the cumulative number of customer abandonments up to time t > 0.

In conventional HT, $L^{\lambda}(t) \approx \sqrt{\lambda}\int_0^t F'(0) X_{ROU}(s)ds.$

$F'(0)$ governs the instantaneous customer abandonment rate.

It is also true in the many-server limit regime that

$F'(0)$ governs the instantaneous customer abandonment rate.

Theorem 4: The Cumulative Number of Abandonments (Dai and He, 2010 and Mandelbaum and Momcilovic, 2010)

For any $T > 0$,

$$\frac{1}{\sqrt{\lambda}} \sup_{0 \le t \le T} \left| L^{\lambda}(t) - F'(0)\int_0^t \left[Q^{\lambda}(s) - N^{\lambda} \right]^+ ds \right| \to 0, \text{ i.p., as } \lambda \to \infty.$$

Note that $\frac{L^{\lambda}(t)}{A^{\lambda}(t)}$ approximates the customer abandonment probability.

The Many Server Approximation Theorem

Theorem 5: The Halfin-Whitt (HW) Many Server Limit

(Mandelbaum and Momcilovic, 2010)

In the HW limit regime, for the GI/M/N+GI queue,

$$\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}, \text{ as } \lambda \to \infty \text{ in D},$$

where, for $\sigma^2 = c_a^2 + 1$,

$$X_{HW}(t) = X_{HW}(0) + \int_0^t m_{HW}\left(X_{HW}(s)\right)ds + \sigma B(t),$$

and

$$m_{HW}(x) = \begin{cases} -\beta - F'(0)x & \text{if } x \geq 0 \\ -\beta - \mu x & \text{if } x < 0 \end{cases}.$$

M&M allows service times to be GI. But the limit process is much more complicated, and so we do not state it here.

The same process that approximates the M/M/N+M queue in the HW limit regime also approximates the GI/M/N+GI queue.

Recap

We have proposed two approximations for the queue length process.

Conventional HT,GI/GI/N+GI:

$$Q^{\lambda}(\bullet) \approx \sqrt{\lambda} X_{ROU}(\bullet)$$

HW many server regime, GI/M/N+GI:

$$Q^{\lambda}(\bullet) \approx \sqrt{\lambda} X_{HW}(\bullet) + N^{\lambda}$$

X_{ROU} has infinitesimal drift $m_{ROU}(x) = -\beta - F'(0)x$ for $x > 0$.

X_{HW} has infinitesimal drift $m_{ROU}(x) = \begin{cases} -\beta - F'(0)x \text{ for } x > 0 \\ -\beta - \mu x \text{ for } x \le 0 \end{cases}$.

Both processes have infinitesimal variance $\sigma^2 = c_a^2 + c_s^2$.

Both processes are analytically tractable.

How do we approximate waiting times?
A transient Little's law holds in the limit (W&G 2005 and M&M 2010).

Possible Objection

The mean and variance of both the inter-arrival and service times are required for the proposed approximation.

The only knowledge we need of the abandonment distribution is the value of its density at 0. That is not a very robust statistic. (And what do we do if the abandonment density either equals 0 or explodes at 0?)

Is there an approximation that requires more knowledge of the abandonment distribution?

An Approximation Refinement: Hazard Rate Scaling

An approximate Markovian abandonment rate for the jth customer from the end of the queue is: $h(j/\lambda)$.

Then, the instantaneous abandonment rate from the queue is: $\sum_{j=1}^{Q^\lambda(t)} h(j/\lambda)$.

If $h^\lambda(x) := h\left(\sqrt{\lambda}x\right)$ for all $x \geq 0$, then, assuming $\frac{Q^\lambda}{\sqrt{\lambda}} \Rightarrow \hat{Q}$ in D,

$$\frac{1}{\sqrt{\lambda}} \sum_{j=1}^{Q^\lambda(t)} h^\lambda(j/\lambda) = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{\sqrt{\lambda}\left(\frac{Q^\lambda(t)}{\sqrt{\lambda}}\right)} h(j/\sqrt{\lambda}) \rightarrow \int_0^{\hat{Q}(t)} h(x)dx$$

This observation suggests replacing X_{ROU} and X_{HW} by X_{ROU-HS} and X_{HW-HS}, that have infinitesimal drifts

$$m_{ROU-HS}(x) = -\beta - \int_0^x h(y)dy \text{ for } x \geq 0, m_{HW-HS}(x) = \begin{cases} -\beta - \int_0^x h(y)dy & \text{for } x \geq 0 \\ -\beta - \mu x & \text{for } x < 0 \end{cases}.$$

An Approximation Refinement: Hazard Rate Scaling

Theorem 6: Hazard Rate Scaling

(Reed and Ward, 2008, and Reed and Tezcan, 2010)

In conventional HT, for the GI / GI / N+GI queue,

$$\frac{Q^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{ROU-HS}, \text{ as } \lambda \to \infty \text{ in D.}$$

In the HW limit regime, for the GI / M / N+GI queue,

$$\frac{Q^{\lambda} - N^{\lambda}}{\sqrt{\lambda}} \Rightarrow X_{HW-HS}, \text{ as } \lambda \to \infty \text{ in D.}$$

Why is this a refinement?

From a Taylor series expansion, noting that $h(0) = F'(0)$,

$$-\beta - \int_0^x h(y)dy = -\beta - \left[h(0)x + \frac{1}{2}h'(0)x^2 + \sum_{j=3}^{\infty} h^{(j)}(0)\frac{x^{j+1}}{(j+1)!} \right].$$

As in the drift terms for X_{ROU} and X_{HW}.

Are $X_{ROU\text{-}HS}$ and $X_{HW\text{-}HS}$ analytically tractable processes?

There are analytic expressions for their steady-state distributions; see R&W and R&T. However, we are not aware of any results on their transient distributions.

Note also that the numeric results in R&W and R&T show that these approximations are quite accurate.

Talk Outline

- The M/M/N+M Queueing Model
- Diffusion approximations for the M/M/N+M Queueing Model
  - Conventional Heavy-Traffic
  - The Halfin-Whitt Many-Server Regime
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- **Intermediate Regimes**
- Unifying diffusion approximations

A Comparison of the Conventional HT and HW Limit Regimes

Conventional HT : $Q^\lambda(\bullet) \approx \sqrt{\lambda} X_{ROU}(\bullet)$	**HW :** $Q^\lambda(\bullet) \approx \sqrt{\lambda} X_{HW}(\bullet) + N^\lambda$
$\rho^\lambda := \dfrac{\lambda}{N\mu^\lambda} \to 1$ as $\lambda \to \infty$ λ is large compared to γ. (Small % of customers abandon.) Waiting times are of order $1/\sqrt{\lambda}$.	$\rho^\lambda := \dfrac{\lambda}{N^\lambda\mu} \to 1$ as $\lambda \to \infty$ λ is large compared to γ. (Small % of customers abandon.) Waiting times are of order $1/\sqrt{\lambda}$.
An arriving customer is almost certain to be delayed. Assumes λ and μ are large, and N is small. (Service times are of order $1/\lambda$.) Delay times >> service times.	The probability an arriving customer is delayed remains strictly between 0 and 1. Assumes λ and N are large, and μ is small (in comparison). (Service times are of order 1.) Delay times << service times.

Is there a regime in which service and delay times are of the same order?

Definition: An Intermediate Limit Regime

For a given $\beta \in \Re$, define

$u^\lambda := \mu\left(\lambda^{1-\alpha} + \lambda^{1/2-\alpha}\beta\right)$ and $N^\lambda = \lambda^\alpha / \mu$ for $\alpha \in [0,1]$.

$\alpha = 0$ is conventional HT and $\alpha = 1$ is the HW limit regime, with $\beta = 0$.

It is still true that $\rho^\lambda = \dfrac{\lambda}{N^\lambda \mu^\lambda} \to 1$, as $\lambda \to \infty$.

When $\alpha = 1/2$, delays and service times are of the same order.

Theorem 7: The Intermediate Limit

(Atar, 2010)

In the intermediate limit regime, for the M/M/N+M queue,

$$\frac{Q^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU} \text{ in } D.$$

Talk Outline

- The M/M/N+M Queueing Model
- Diffusion approximations for the M/M/N+M Queueing Model
  - Conventional Heavy-Traffic
  - The Halfin-Whitt Many-Server Regime
- Diffusion approximations for the GI/GI/N+GI Queueing Models
- Intermediate Regimes
- **Unifying diffusion approximations**

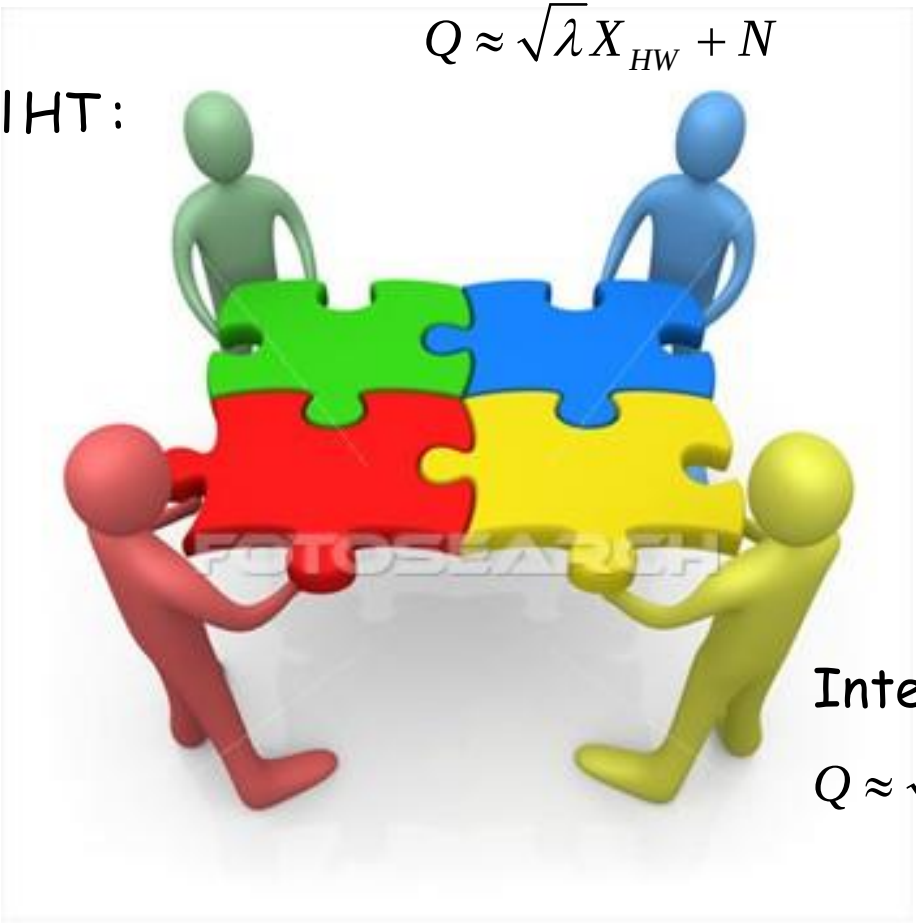
Can we bring the different diffusion approximations together?

Halfin-Whitt Many Server Limit Regime:

$$Q \approx \sqrt{\lambda} X_{HW} + N$$

Conventional HT:

$$Q \approx \sqrt{\lambda} X_{ROU}$$



Intermediate Limit Regime:

$$Q \approx \sqrt{\lambda} X_{ROU} + N$$

Towards a unifying approximation.

Define

$$X(t) = X(0) + \int_0^t m_X(X(s))ds + \sqrt{\lambda\left(c_a^2 + c_s^2\right)}B(t) + L(t) \geq 0$$

$$L(0) = 0, L \text{ is non-decreasing}, \int_0^\infty X(t)dL(t) = 0,$$

where $m_X(x) := m_Q(x)$. ⟵ Depends on λ, N, and μ.

We match the infinitesimal drift of X to the queue-length process of a M/M/N+M queue.

Theorem 8: A Unifying Approximation

(Ward, 2011)

In conventional HT, $\frac{X^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}$ in D, as $\lambda \to \infty$.

In the Halfin-Whitt limit regime, $\frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{HW}$ in D, as $\lambda \to \infty$.

In the intermediate limit regime, $\frac{X^\lambda - N^\lambda}{\sqrt{\lambda}} \Rightarrow X_{ROU}$ in D, as $\lambda \to \infty$.

The Proposed Unifying Approximation

The proposed unifying approximation is:

$$Q(\cdot) \approx X(\cdot).$$

We do not need to scale.

We do not need to determine which regime we are in.

M/M/N+M queue with arrival rate $\lambda = 50$, abandonment rate $\gamma = 1$, and $\lambda/(N\mu) = 1$, and varying values of λ and μ.

N	μ	P[abandon]	E[Q] exact	E[X] approximate
1	50	9.8%	5.80	6.10
10	5	7.8%	13.12	13.13
20	5/2	6.9%	22.02	22.10
30	5/3	6.3%	31.31	31.27
40	5/4	5.9%	40.64	40.60
50	1	5.6%	50.00	50.00

Caveat: This is not consistent with the hazard rate scaling.
Abandonment probability approximations are not so good.

Conclusions

- We have surveyed diffusion approximation results for the GI/GI/N+GI Queue.
- Our purpose was to develop simple approximations for the queue-length, customer abandonment probability, and wait time.
- We have proposed a unifying approximation.