Queues in service systems: customer abandonment and diffusion approximations

Jim Dai

Georgia Institute of Technology

November 13, 2011

Joint work with Shuangchi He (NUS)
Outline

- Single-server queues vs many-server queues
- The QED regime and the square-root staffing rule
- Need of modeling customer abandonment
- Distributional sensitivity
- Diffusion models for many-server queues
Multi-server queues

A $G/GI/n + GI$ queue

- $n$ identical servers working in parallel
  (single-server $n = 1$; many-server $n \gg 1$)
- first-in-first out buffer of infinite size
- a general arrival process ($G$)
- iid service times ($GI$)
- iid patience times ($+GI$)
A many-server queue serves as a building block for modeling large-scale service systems

- Call centers
  - Bank of America, over one thousand agents
  - UPS, several hundred agents

- Hospital beds
  - hundreds of beds

- Web farms/computer clusters
  - up to several thousand servers/CPUs
**Insight**

*The performance of many-server queues is qualitatively different from that of single-server queues or queues with a small number of servers.*

Key performance measures

- delay probability $P_w$
- mean customer waiting time $w$
- fraction of abandonment $P_A$
Performance of a single-server queue

Consider an $M/GI/1$ queue without abandonment

- Poisson arrival process with rate $\lambda$
- Mean service time $m$
- Traffic intensity $\rho := \lambda m$

Assume $\rho < 1$. By PASTA and Pollaczek-Khinchine,

$$P_w = \rho \quad \text{and} \quad w = m \left( \frac{\rho}{1 - \rho} \right) \left( \frac{1 + c_s^2}{2} \right),$$

- SCV of service times $c_s^2$, and waiting time factor

$$f := \frac{w}{m} = \left( \frac{\rho}{1 - \rho} \right) \left( \frac{1 + c_s^2}{2} \right).$$

Since $P_w \to 1$ as $\rho \to 1$, almost all have to wait before being served.
quality: no waiting or very short waiting

efficiency: $\rho \to 1$

However,

$$f \text{ is proportional to } \frac{\rho}{1 - \rho}$$

**Insight**

In a single-server queue, one cannot maintain high server utilization to achieve good quality of service
Quality OR efficiency?

**Figure:** Waiting time factor $f$ vs server utilization $\rho$ in an $M/M/1$ queue
Consider an $M/M/n$ queue. Traffic intensity $\rho := \lambda m/n$. By Erlang-C,

$$P_w = \frac{(n\rho)^n}{n!} \left( (1 - \rho) \sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \right)^{-1}$$

The waiting time factor

$$f = \frac{w}{m} = \frac{P_w}{(1 - \rho)n} \leq \frac{1}{(1 - \rho)n}$$

With $\rho$ fixed, $f \to 0$ as $n \to \infty$
**Many-server queues: quality AND efficiency!**

**Figure:** Delay probability $P_w$ and waiting time factor $f$ vs number of servers $n$, for $M/M/n$ queues with $\rho = 0.95$

If one increases $n$ to 100, then $P_w = 50.7\%$ and $f = 0.101$
**Figure:** Waiting time factor $f$ vs server utilization $\rho$ in an $M/M/1$ queue and an $M/M/18$ queue
The QED regime

The above $M/M/100$ queue with $\rho = 0.95$ achieves both high quality of service and operational efficiency:

- the server utilization close to 1 (efficiency)
- only a fraction of customers need to wait (quality)
- waiting times are relatively short (quality)

The system is operated in the quality- and efficiency-driven (QED) regime, also called the rationalized regime.
The square-root staffing rule in the $M/M/n$ setting

Let $R := \lambda m$ be the offered load. The square-root safety staffing rule recommends

$$n \approx R + \beta \sqrt{R} \quad \text{for some } \beta > 0$$

Erlang-C

$$P_w = \frac{(n\rho)^n}{n!} \left( \frac{1}{n!} \sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \right)^{-1}$$

Halflin and Whitt (1981) proved that

$$P_w \rightarrow \gamma = \frac{1}{\beta \Phi(\beta)/\phi(\beta) + 1} \quad \text{as } R \rightarrow \infty$$

- $\phi$ is the standard normal probability density
- $\Phi$ is the standard normal cumulative distribution function
The square-root staffing rule in the $M/M/n$ setting

Fix $\beta > 0$ and set $n \approx R + \beta \sqrt{R}$. As $R$ increases,

- $P_w$ stabilizes at $\gamma \in (0, 1)$
- $\rho = R/n \to 1$
- $f = P_w/(\sqrt{n}\beta)$ is on the order of $1/\sqrt{n}$

The system is operated in the QED regime!
Performance analysis using formula (1)

Given staffing level $n$ and utilization level $\rho < 1$, set

$$\beta = \sqrt{n(1 - \rho)}$$

Then,

$$P_w \approx \frac{1}{\beta \Phi(\beta)/\phi(\beta) + 1} \quad \text{and} \quad f = \frac{P_w}{\sqrt{n} \beta}$$

<table>
<thead>
<tr>
<th></th>
<th>Exact</th>
<th>Approx. by (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_w$</td>
<td>50.7%</td>
<td>50.5%</td>
</tr>
<tr>
<td>$f$</td>
<td>0.101</td>
<td>0.101</td>
</tr>
</tbody>
</table>

**Table:** Performance measures in an $M/M/100$ queue with $\rho = 0.95$
Suppose $P_w$ is required to be less than some $\gamma \in (0, 1)$. First solve for $\beta$ by

$$\gamma = \frac{1}{\beta \Phi(\beta)/\phi(\beta) + 1}$$

Then,

$$n \approx R + \beta \sqrt{R}$$
Consider a $GI/GI/n$ queue. Let

$$n = R + \beta \sqrt{R} \quad \text{for some } \beta > 0$$

Reed (2009) proved that as $R$ increases, this staffing level drives the queue to the QED regime.
Historical remarks

- The origin of that can be traced back to **Erlang** (1923)
- Erlang reported that it had been in use at the Copenhagen Telephone Company since **1913**
- **Whitt** (1992) formally proposed and analyzed this rule
Customer abandonment

- Human’s patience is always limited!
- Customer abandonment is present in most service systems

**Insight**

*For a service system with significant customer abandonment, any queueing model that ignores the abandonment phenomenon is likely irrelevant to operational decisions*
One must model abandonment explicitly!

<table>
<thead>
<tr>
<th></th>
<th>$M/M/50 + M$</th>
<th>$M/M/50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean service time</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean patience time</td>
<td>2</td>
<td>N/A</td>
</tr>
<tr>
<td>Arrival rate</td>
<td>55</td>
<td>$55 \times (1 - 10.2%) = 49.39$</td>
</tr>
<tr>
<td>Abandonment fraction</td>
<td>10.2%</td>
<td>N/A</td>
</tr>
<tr>
<td>Server utilization</td>
<td>98.8%</td>
<td>98.8%</td>
</tr>
<tr>
<td>Mean waiting time (in sec.)</td>
<td>12.5</td>
<td>87.7</td>
</tr>
<tr>
<td>Mean queue length</td>
<td>11.2</td>
<td>72.2</td>
</tr>
</tbody>
</table>

**Table:** Queues with and without customer abandonment
Why customer abandonment matters?

- Customers who experience long waiting tend to abandon the system.
- With abandonment, the system can reach a steady state even if the arrival rate is larger than the service capacity.
- Some performance measures in a queue with abandonment is better than in a queue without abandonment.
- To meet certain service levels without considering abandonment, one tends to overestimate the staffing level.
The square-root staffing rule is still applicable

**Insight**

*In the presence of customer abandonment, the square-root safety staffing rule can still lead the system to the QED regime and yield high server utilization, short waiting times, and a very small abandonment fraction.*
The square-root staffing rule in the $M/M/n + M$ setting

As argued by Garnett et al. (2002), with

$$n \approx R + \beta \sqrt{R} \quad \text{for} \quad \beta \in \mathbb{R} \quad \text{and} \quad R \text{ large},$$

one has

$$P_w \approx \left(1 + \frac{h(\beta \sqrt{\mu/\alpha})}{\sqrt{\mu/\alpha} h(-\beta)}\right)^{-1}$$

- $1/\alpha$ is the mean patience time
- $h(x) = \phi(x)/(1 - \Phi(x))$ is the standard normal hazard rate

The fraction of abandonment

$$P_A \approx \frac{1}{\sqrt{R}} \left(\sqrt{\alpha/\mu} h(\beta \sqrt{\mu/\alpha}) - \beta\right) \left(1 + \frac{h(\beta \sqrt{\mu/\alpha})}{\sqrt{\mu/\alpha} h(-\beta)}\right)^{-1}$$
Fix $\beta = \sqrt{n}(1 - \rho)$. As $n$ increases,

- the mean waiting time decreases at rate $1/\sqrt{n}$ in $M/M/n$ queues
- Garnett et al (2002) confirmed the same decreasing rate in $M/M/n + M$ queues

When $n$ is large,

- waiting times are relatively short
- the patience time distribution $F$, outside a small neighborhood of the origin, barely has any influence on the system dynamics
Sensitivity on $F$ with fixed $\alpha = F'(0+)$

Consider an $M/M/100 + GI$ queue

- with different $F$
- but with the same $\alpha = F'(0+)$
- $\lambda = 105$ and $m = 1$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Abandonment fraction</th>
<th>Mean queue length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp</td>
<td>Uniform</td>
</tr>
<tr>
<td>$\alpha = 0.1$</td>
<td>0.0497</td>
<td>0.0498</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td>0.0603</td>
<td>0.0607</td>
</tr>
<tr>
<td>$\alpha = 1$</td>
<td>0.0670</td>
<td>0.0676</td>
</tr>
<tr>
<td>$\alpha = 2$</td>
<td>0.0739</td>
<td>0.0748</td>
</tr>
<tr>
<td>$\alpha = 10$</td>
<td>0.0886</td>
<td>0.0902</td>
</tr>
</tbody>
</table>

**Table:** Performance insensitivity to patience time distributions $F$
Sensitivity on $F$ with mean patience time $m_p$ fixed

<table>
<thead>
<tr>
<th></th>
<th>Abandonment fraction</th>
<th></th>
<th>Mean queue length</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp</td>
<td>Uniform</td>
<td>$H_2$</td>
<td>Exp</td>
</tr>
<tr>
<td>$m_p = 0.1$</td>
<td>0.0886</td>
<td>0.0840</td>
<td>0.0926</td>
<td>0.9301</td>
</tr>
<tr>
<td>$m_p = 0.5$</td>
<td>0.0739</td>
<td>0.0676</td>
<td>0.0794</td>
<td>3.882</td>
</tr>
<tr>
<td>$m_p = 1$</td>
<td>0.0670</td>
<td>0.0608</td>
<td>0.0730</td>
<td>7.031</td>
</tr>
<tr>
<td>$m_p = 2$</td>
<td>0.0603</td>
<td>0.0550</td>
<td>0.0682</td>
<td>12.67</td>
</tr>
<tr>
<td>$m_p = 10$</td>
<td>0.0497</td>
<td>0.0481</td>
<td>0.0543</td>
<td>52.18</td>
</tr>
</tbody>
</table>

**Table:** Mean patience time is a **wrong** statistic!
**Insight**

In the QED regime, the system performance is generally invariant with the patience time distribution as long as its density at the origin is fixed and positive.

- For a $G/GI/n + GI$ queue in the QED regime, it is generally accurate to replace $F$ with an exponential distribution with rate $\alpha = F'(0^+)$.
- The matrix-analytic method can be used to evaluate $GI/Ph/n + M$ queues.
Dai and He (2010) proved that

\[ A(t) \approx \alpha \int_0^t Q(s) \, ds \]

- \( A(t) \) is the number of abandonments by time \( t \)
- \( Q(t) \) is the number of waiting customers at time \( t \)

This justifies the replacement of \(+ GI\) with \(+ M\).
Non-exponential distributions

The exact analysis of a many-server queue has largely been limited to the $M/M/n + M$ model. However...


Patience time hazard rate of a call center, by Mandelbaum and Zeltyn (2004)
A $GI/GI/n + GI$ queue is difficult to be analyzed because of
- general interarrival/service/patience time distributions
- a large number of servers

As a consequence,
- no analytical solution and no numerical algorithms for performance measures
- usually evaluated by simulation

We use diffusion processes to approximate many-server queues
Let $\{E(t) : t \geq 0\}$ be a Poisson process with rate $\lambda$.

**Figure:** A Poisson sample path with rate $\lambda = 1$
The centered sample path with $\lambda = 1$

Then, $\{E(t) - \lambda t : t \geq 0\}$ is the centered process

**Figure:** The sample path of the centered process with $\lambda = 1$
A Poisson sample path with $\lambda = 100$

**Figure:** A Poisson sample path with rate $\lambda = 100$
The centered sample path with $\lambda = 100$

**Figure:** The sample path of the centered process with $\lambda = 100$
A Poisson sample path with $\lambda = 10,000$

**Figure:** A Poisson sample path with rate $\lambda = 10,000$
The centered sample path with $\lambda = 10,000$

**Figure:** The sample path of the centered process with $\lambda = 10,000$
Brownian motion and Donsker's theorem

Let

$$\tilde{E}_\lambda(t) = \frac{E(t) - \lambda t}{\sqrt{\lambda}}$$

Donsker's theorem implies that the process $\tilde{E}_\lambda$ is close to a standard Brownian motion when $\lambda$ is large.

**Definition**

A process $B = \{B(t) : t \geq 0\}$ is said to be a $(\mu_B, \sigma_B^2)$-Brownian motion if

- $B(0) = 0$ and almost every sample path is continuous
- $\{B(t) : t \geq 0\}$ has stationary, independent increments
- $B(t)$ is normally distributed with mean $\mu_B t$ and variance $\sigma_B^2 t$ for every $t > 0$

$B$ is a standard Brownian motion if $\mu_B = 0$ and $\sigma_B^2 = 1$
System equation for an $M/M/n + GI$ queue

$$X(t) = X(0) + E(t) - S\left(\mu \int_0^t Z(s) \, ds\right) - A(t)$$

- $X(t)$ is the number of customers in system at time $t$
- $E(t)$ is the number of arrivals by time $t$
- $\{S(t) : t \geq 0\}$ is a Poisson process with rate one
- $\mu = 1/m$ is the service rate
- $Z(t)$ is the number of busy servers at time $t$
- $A(t)$ is the number of abandonments by time $t$
Let
\[ \tilde{E}(t) = \frac{E(t) - \lambda t}{\sqrt{n}} \quad \text{and} \quad \tilde{S}(t) = \frac{S(nt) - nt}{\sqrt{n}} \]

By Donsker’s theorem
\[ \tilde{E} \approx B_E \quad \text{and} \quad \tilde{S} \approx B_S \]

- \( B_E \) is a \((0, \rho \mu)\)-Brownian motion
- \( B_S \) is a \((0, 1)\)-Brownian motion
- \( B_E \) and \( B_S \) are independent
Recall that

\[ \alpha = F'(0+) \]

The abandonment process is approximated by

\[ A(t) \approx \alpha \int_0^t Q(s) \, ds = \alpha \int_0^t (X(s) - n)^+ \, ds \]
Scaled system equations

$$\tilde{X}(t) = \frac{X(t) - n}{\sqrt{n}}, \quad \beta = \sqrt{n}(1 - \rho), \quad \tilde{A}(t) = \frac{A(t)}{\sqrt{n}}$$

The scaled system equation

$$\tilde{X}(t) = \tilde{X}(0) + \tilde{E}(t) - \tilde{S} \left( \mu \int_0^t \frac{Z(s)}{n} ds \right)$$

$$- \beta \mu t + \mu \int_0^t \tilde{X}(s)^- ds - \tilde{A}(t)$$

where

$$\tilde{E} \approx B_E$$

$$\tilde{S} \approx B_S$$

$$\tilde{A} \approx \alpha \int_0^t \tilde{X}(s)^+ ds$$

$$\frac{Z(t)}{n} \approx \rho \land 1$$
A diffusion model for an $M/M/n + GI$ queue

The scaled system equation

$$\tilde{X}(t) = \tilde{X}(0) + \tilde{E}(t) - \tilde{S} \left( \mu \int_0^t \frac{Z(s)}{n} \, ds \right)$$

$$- \beta \mu t + \mu \int_0^t \tilde{X}(s)^{-} \, ds - \tilde{A}(t)$$

The diffusion model

$$Y(t) = \tilde{X}(0) + B_E(t) - B_S((\rho \wedge 1) \mu t)$$

$$- \beta \mu t + \mu \int_0^t Y(s)^{-} \, ds - \alpha \int_0^t Y(s)^{+} \, ds$$
A piecewise OU process

The diffusion model

\[ Y(t) = \tilde{X}(0) + B_E(t) - B_S((\rho \wedge 1)\mu t) \]
\[ - \beta \mu t + \mu \int_0^t Y(s) - ds - \alpha \int_0^t Y(s) + ds \]

- a piecewise linear drift

\[ b(x) = \begin{cases} 
-\beta \mu - \alpha |x| & \text{when } x \geq 0 \\
-\beta \mu + \mu |x| & \text{when } x \leq 0
\end{cases} \]

- the mean-reverting property

\( Y \) is a piecewise Ornstein-Uhlenbeck (OU) process. It becomes an OU process when \( \alpha = \mu \)
A process $Y = \{ Y(t) : t \geq 0 \}$ is called an **OU process** if it satisfies

$$Y(t) = Y(0) + \sigma B(t) - \beta \mu t - \mu \int_0^t Y(s) \, ds$$

- a **linear drift**
  $$b(x) = -\beta \mu - \mu x$$
- a **normal** stationary distribution
  $$g(z) = \sqrt{\frac{\mu}{\pi \sigma^2}} \exp \left( - \frac{\mu (z + \beta)^2}{\sigma^2} \right) \text{ for } z \in \mathbb{R}$$
Stationary distribution of a piecewise OU process

The diffusion model has a **piecewise linear** drift

\[ b(x) = \begin{cases} 
-\beta \mu - \alpha |x| & \text{when } x \geq 0 \\
-\beta \mu + \mu |x| & \text{when } x \leq 0
\end{cases} \]

It admits a **piecewise normal** stationary distribution

\[ g(z) = \begin{cases} 
a_1 \exp \left( - \frac{\alpha (z + \alpha^{-1} \mu \beta)^2}{\sigma_B^2} \right) & \text{when } z \geq 0, \\
a_2 \exp \left( - \frac{\mu (z + \beta)^2}{\sigma_B^2} \right) & \text{when } z < 0,
\end{cases} \]

- \( \sigma_B^2 = \mu (\rho + \rho \wedge 1) \)
- \( a_1 \) and \( a_2 \) are constants such that

\[ \int_{-\infty}^{\infty} g(z) \, dz = 1 \quad \text{and} \quad g(0-) = g(0+) \]
Performance approximations for \(M/M/n + GI\) queues

- the long-run average queue length

\[ \bar{Q} \approx \sqrt{n} \cdot \mathbb{E}[Y(\infty)^+] = \sqrt{n} \int_{0}^{\infty} xg(x) \, dx \]

- the long-run average number of idle servers

\[ \bar{I} \approx \sqrt{n} \cdot \mathbb{E}[Y(\infty)^-] = -\sqrt{n} \int_{-\infty}^{0} xg(x) \, dx \]

- the abandonment fraction

\[ P_A \approx 1 - \frac{\mu(n - \bar{I})}{\lambda} \]
## Diffusion approximation for the $M/M/100 + M$ queue

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Abandonment fraction</th>
<th>Mean queue length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exp</td>
<td>Diffusion</td>
</tr>
<tr>
<td>$\alpha = 0.1$</td>
<td>0.0497</td>
<td>0.0497</td>
</tr>
<tr>
<td>$\alpha = 0.5$</td>
<td>0.0603</td>
<td>0.0603</td>
</tr>
<tr>
<td>$\alpha = 1$</td>
<td>0.0670</td>
<td>0.0669</td>
</tr>
<tr>
<td>$\alpha = 2$</td>
<td>0.0739</td>
<td>0.0738</td>
</tr>
<tr>
<td>$\alpha = 10$</td>
<td>0.0886</td>
<td>0.0886</td>
</tr>
</tbody>
</table>

**Table**: Performance estimates for the $M/M/100 + M$ queue
Beyond exponential service distributions

A two-phase hyperexponential distribution \((H_2)\)

\[ V = \begin{cases} 
\text{Exp}(\nu_1) & \text{with probability } p_1 \\
\text{Exp}(\nu_2) & \text{with probability } p_2 = 1 - p_1
\end{cases} \]

- fraction of phase \(j\) workload

\[ \theta_j = \frac{p_j/\nu_j}{p_1/\nu_1 + p_2/\nu_2}, \quad \theta_1 + \theta_2 = 1 \]

- a special case of phase-type distributions
A diffusion model for an \( M/H_2/n + GI \) queue

Let \( X_j(t) \) be the number of type \( j \) customers in system at time \( t \)

\[
\tilde{X}_j(t) = \frac{X_j(t) - n\theta_j}{\sqrt{n}}
\]

A two-dimensional process \((Y_1, Y_2)\) is used to approximate \((\tilde{X}_1, \tilde{X}_2)\)
A diffusion model for an $M/H_2/n + GI$ queue

$$Y_j(t) = \tilde{X}_j(0) - \beta \mu p_j t + p_j B_E(t) + (-1)^{j-1} B_M(\rho \mu t) - B_j((\rho \wedge 1) \theta_j \nu_j t)$$

$$+ \nu_j \int_0^t (p_j (Y_1(t) + Y_2(t))^+ - Y_j(t)) \, ds$$

$$- p_j \alpha \int_0^t (Y_1(s) + Y_2(s))^+ \, ds$$

- $B_E$ is a $(0, \rho \mu)$-Brownian motion
- $B_1$ and $B_2$ are $(0, 1)$-Brownian motions
- $B_M$ is a $(0, p_1 p_2)$-Brownian motion
- they are all independent

See He and Dai (2011) for diffusion models for a $GI/Ph/n + GI$ queue
Let $Y$ be a $d$-dimensional diffusion process. Assume that $Y$ has a unique stationary density $g$ on $\mathbb{R}^d$. The basic adjoint relationship (BAR) says

$$\int_{\mathbb{R}^d} Gf(x)g(x) \, dx = 0 \quad \text{for all } f \in C^2_b(\mathbb{R}^d)$$

- $G$ is the generator of $Y$
- He and Dai (2011) designed an algorithm to solve the BAR
Example: an $M/H_2/500 + M$ queue

**Figure:** $\rho = 1.045$, $p = (0.9351, 0.0649)$, $1/\nu = (0.1069, 13.89)$, mean patience time = 2
Example: an $M/H_2/20 + M$ queue

Figure: $\rho = 1.112$, $\rho = (0.9351, 0.0649)$, $1/\nu = (0.1069, 13.89)$, mean patience time = 2.
The approximation $A(t) \approx \alpha \int_0^t Q(s) \, ds$ is not always good

- The abandonment process still depends on $F$ in a neighborhood of the origin, not just the origin
- When the patience time changes rapidly near the origin, this abandonment approximation can be inaccurate
- When $\alpha = 0$ and $\rho > 1$, the queue can still reach a steady state thanks to abandonment, but the diffusion model does not have a stationary distribution
How to improve the abandonment approximation?

Consider a neighborhood of the origin rather than the origin itself!

- Exploiting the idea of scaling the patience time hazard rate, proposed by Reed and Ward (2008)
- Assume $F$ has a bounded hazard function

$$h(t) = \frac{f(t)}{1 - F(t)} \quad \text{for } t \geq 0.$$ 

The scaled abandonment process is approximated by

$$\tilde{A}(t) \approx \int_0^t \int_0^{Q(s)/\lambda} h\left(\frac{\sqrt{nu}}{\lambda}\right) \, du \, ds.$$
Intuition on the abandonment rate

- By time $s$, the $i$th customer from the back of the queue has been waiting around $i/\lambda$ minutes.
- This customer will abandon the queue during the next $\delta$ minutes with probability $h(i/\lambda)\delta$.
- The abandonment rate at time $s$ is around $\sum_{i=1}^{Q(s)} h(i/\lambda)$.
- The scaled abandonment rate

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{Q(s)} h\left(\frac{i}{\lambda}\right) \approx \int_0^{\sqrt{n}Q(s)} h\left(\frac{\sqrt{n}u}{\lambda}\right) du$$
The refined diffusion model

$$Y_j(t) = \tilde{X}_j(0) - \beta \mu p_j t + p_j B_E(t) + (-1)^{j-1} B_M(\rho \mu t) - B_j((\rho \wedge 1) \theta_j \nu_j t)$$

$$+ \nu_j \int_0^t (p_j(Y_1(t) + Y_2(t))^+ - Y_j(t)) \, ds$$

$$- p_j \int_0^t \int_0^s (Y_1(u) + Y_2(u))^+ h\left(\frac{\sqrt{nu}}{\lambda}\right) \, du \, ds$$
Example: an $M/H_2/500 + H_2$ queue

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th>Diffusion</th>
<th>Refined diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean queue length</td>
<td>6.413</td>
<td>1.475</td>
<td>6.359</td>
</tr>
<tr>
<td>Abandonment fraction</td>
<td>0.05512</td>
<td>0.05863</td>
<td>0.05517</td>
</tr>
<tr>
<td>$\mathbb{P}[X(\infty) &gt; 480]$</td>
<td>0.8881</td>
<td>0.8663</td>
<td>0.8929</td>
</tr>
<tr>
<td>$\mathbb{P}[X(\infty) &gt; 500]$</td>
<td>0.4720</td>
<td>0.3192</td>
<td>0.4822</td>
</tr>
<tr>
<td>$\mathbb{P}[X(\infty) &gt; 520]$</td>
<td>0.1050</td>
<td>$9.274 \times 10^{-5}$</td>
<td>0.1074</td>
</tr>
</tbody>
</table>

**Table:** Performance measures of the $M/H_2/500 + H_2$ queue.

- traffic intensity: $\rho = 1.045$
- service time distribution: $p = (0.5915, 0.4085)$ and $\nu = (5.917, 0.454)$
- patience time distribution: $p = (0.9, 0.1)$ and $\nu = (1, 200)$
Example: an $M/H_2/500 + E_3$ queue

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th>Refined diffusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean queue length</td>
<td>119.1</td>
<td>119.5</td>
</tr>
<tr>
<td>Abandonment fraction</td>
<td>0.04337</td>
<td>0.04340</td>
</tr>
<tr>
<td>$P[X(\infty) &gt; 480]$</td>
<td>0.9940</td>
<td>0.9946</td>
</tr>
<tr>
<td>$P[X(\infty) &gt; 500]$</td>
<td>0.9756</td>
<td>0.9770</td>
</tr>
<tr>
<td>$P[X(\infty) &gt; 600]$</td>
<td>0.6645</td>
<td>0.6733</td>
</tr>
</tbody>
</table>

Table: Performance measures of the $M/H_2/n + E_3$ queue.

- $\rho = 1.045$ and $\alpha = 0$, the first diffusion model fails!
- Service time distribution: $\rho = (0.5915, 0.4085)$ and $\nu = (5.917, 0.454)$
- Mean patience time 1 minute
Summary

- Single-server queues and many-server queues are qualitatively different.
- Follow the square-root staffing rule to operate your system in the QED regime.
- Model customer abandonment explicitly.
- In the QED regime, the patience density at the origin has the most impact on system performance.
- Diffusion models is a useful tool to evaluate a many-server queue’s performance.
Survey of call centers


Selected references

The QED regime and the square-root staffing rule

- W. K. Grassmann. Finding the right number of servers in real-world queuing systems. *Interfaces*, 1988
Selected references

Customer abandonment

- J. G. Dai and S. He. Customer abandonment in many-server queues. *MOR*, 2010
- J. E. Reed and A. R. Ward. Approximating the $GI/GI/1 + GI$ queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *MOR*, 2008
- S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems*, 2005
**Selected references**

**Diffusion approximations for many-server queues**

- J. Reed and T. Tezcan. Hazard rate scaling for the $GI/M/n + GI$ queue. *Preprint*, 2009