Part 2

Queues in the Halfin-Whitt or QED regime

Overview of part 2

1. The Halfin-Whitt regime
   - $M/M/s$ and $M/M/s/s$ queue
   - Heavy-traffic limit (and relaxation time)
2. The $M/D/s$ queue: some history
   - Erlang (roots) and Pollaczek (infinite series)
   - Heavy-traffic limit
3. The Gaussian random walk
   - From infinite series to the Riemann zeta function
4. Corrected diffusion approximations
   - A close look at the Poisson distribution
   - Corrections for the $M/M/s$ and $M/D/s$ queue
5. Refined square-root staffing

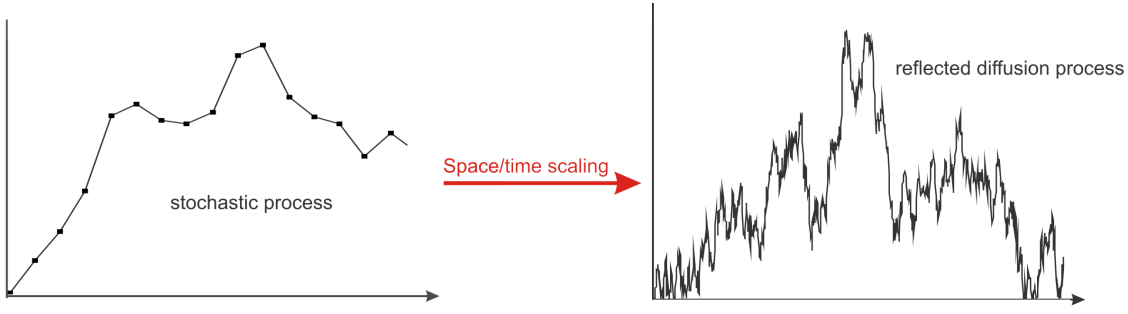
The Halfin-Whitt regime

Heavy traffic

- Consider stochastic processes in their most critical regimes

John Kingman

- *Heavy-traffic theory*, launched by Kingman in the 1960's: system utilization $\approx 100\%$
- Space-time scaling:
  - Speed up time, while shrinking space, so that the CLT kicks in
  - Apply *functional limit theorems* to obtain the scaling limit



Erlang B and C formulas

Consider Poisson arrivals with rate λ and exponential service times with mean $1/\mu$. Erlang (1917) obtained the blocking probability in the $M/M/s/s$ queue:

$$B(s,\lambda) = \frac{\mathbb{P}(\text{Pois}(\lambda) = s)}{\mathbb{P}(\text{Pois}(\lambda) \leq s)}$$

For the delay probability in the $M/M/s$ queue, denoted by $C(s,\lambda)$, it holds that

$$C(s,\lambda)^{-1} = \rho + (1-\rho)B(s,\lambda)^{-1}$$

Large systems

For a constant $\beta \in \mathbb{R}$ such that $s = \lambda + \beta\sqrt{\lambda}$ it holds that

$$B(s, \lambda) = \frac{\mathbb{P}(\mathsf{Pois}(\lambda) = s)}{\mathbb{P}(\mathsf{Pois}(\lambda) \leq s)} \sim \frac{\phi(\beta)}{\Phi(\beta)\sqrt{s}}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2} \mathrm{d}u$

CLT: When λ is a positive integer $\mathsf{Pois}(\lambda)$ can be seen as the sum of λ Poisson random variables with mean one

$$\begin{aligned} \mathbb{P}(\mathsf{Pois}(\lambda) \leq s) &= \mathbb{P}\left(\frac{\sum \mathsf{Pois}(1) - \lambda}{\sqrt{\lambda}} \leq \frac{s - \lambda}{\sqrt{\lambda}} \right) \\ &= \Phi(\beta) + \mathcal{O}(\lambda^{-1/2}) \end{aligned}$$

A balance act

Three regimes for large system (λ and s large):

1. Efficiency driven (ED): Fix $\beta > 0$ and take $s = \lambda + \beta$
2. Quality driven (QD): Fix $\beta > 0$ and take $s = (1 + \beta)\lambda$
3. Quality & Efficiency Driven (QED): Fix $\beta > 0$ and take

$$s = \lambda + \beta\sqrt{\lambda}$$

Halfin-Whitt (1981) derived that

$$\lim_{\lambda \to \infty} C(\lambda + \beta\sqrt{\lambda}, \lambda) = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}$$

This is exactly the right scaling!

Consider the normalized process $X_s(t) = \frac{X(t)-s}{\sqrt{s}}$ with infinitesimal mean $m_s(x)$ that converges as

$$\begin{aligned} m_s(x) &= \frac{-s}{\sqrt{s}} + \frac{\lambda}{\sqrt{s}} \to -\beta, \quad x > 0 \\ &= -\frac{\lfloor \sqrt{s}x + s \rfloor}{\sqrt{s}} + \frac{\lambda}{\sqrt{s}} \to -x - \beta, \quad x < 0 \end{aligned}$$

Theorem
(Halfin-Whitt 1981) Under square-root staffing, $X_s(t) \Rightarrow \hat{X}(t)$ *where* $\hat{X}$ *is a diffusion process with drift*

$$m(x) = \begin{cases} -\beta, & x > 0, \\ -x - \beta, & x < 0, \end{cases} \tag{1}$$

and variance 2.

This diffusion process behaves like a Brownian motion with drift above zero and like an Ornstein-Uhlenbeck process below zero.

- The number of customers is roughly $s + \sqrt{s}\hat{X}(t)$ for s sufficiently large
- The boundary between the Brownian motion and the OU process can be thought of as the number of servers, and $\hat{X}$ will keep fluctuating between these two regions.
- The process mimics a single server queue above zero, and an infinite server queue below zero, for which Brownian motion and the OU process are indeed the respective heavy-traffic limits.
- As β increases, capacity grows and the Halfin-Whitt diffusion will spend more time below zero.

The Halfin-Whitt diffusion $\hat{X}$ is a Markov process on the real line with continuous paths and density $p = p(x, t)$ that satisfies the forward Kolmogorov equation

$$\frac{\partial}{\partial t} p(x,t) = -\frac{\partial}{\partial x}[m(x)p(x,t)] + \frac{\partial^2}{\partial x^2}[p(x,t)]. \qquad (2)$$

with initial condition $p(x, 0) = \delta(x - x_0)$ (the Dirac function) and the boundary conditions $p(\infty, t) = p(-\infty, t) = 0$.

Heavy-traffic scheme

$$
\begin{array}{ccc}
X_s(t) & \Rightarrow & X_s(\infty) \\
\Downarrow & & \Downarrow \\
\hat{X}(t) & \Rightarrow & \hat{X}(\infty)
\end{array}
\qquad\qquad
\begin{array}{c}
M/M/s \text{ queue} \\
\Downarrow \\
\textit{hybrid diffusion process}
\end{array}
$$

Heavy-traffic scheme

$$
\begin{array}{ccc}
X(t) & \Rightarrow & X(\infty) \\
\Downarrow & & \Downarrow \\
\hat{X}(t) & \Rightarrow & \hat{X}(\infty)
\end{array}
$$

queueing process

$\Downarrow$

diffusion or limit process

Literature

Contributions for square-root staffing (Halfin-Whitt regime):

$M/M/s/s$	Erlang (1917)
$M/D/s$	Pollaczek (1930)
$G/M/s$	Halfin-Whitt (1981)
$G/PH/s$	Puhalskii-Reiman ('00)
$G/D/s$	Jelenković-Mandelbaum-Momčilović ('04)
$G/G/s$	Gamarnik-Momčilović ('07)
$G/G/s$	Reed ('07), Kaspi-Ramanan ('08)

Many other applications/extensions including reneging, multi-class customers, control and optimization problems, loss and queueing networks

The $M/D/s$ queue: some history

Godfathers of queueing

A.K. Erlang
1878-1929

F. Pollaczek
1892-1981

A bulk service queue

- Both Erlang (1917) and Pollaczek (1930) studied the queue

$$Q_{n+1} = (Q_n + A_n - s)^+, \quad n = 0, 1, \ldots$$

A bulk service queue

- Both Erlang (1917) and Pollaczek (1930) studied the queue

$$Q_{n+1} = (Q_n + A_n - s)^+, \quad n = 0, 1, \ldots$$

- A_n i.i.d. copies of a Poisson random variable A with mean λ
- Assume that $\lambda < s$
- Let $Q = \lim_{n\to\infty} Q_n$
- Then it comes down to finding the solution to

$$Q \stackrel{d}{=} (Q + A - s)^+$$

Roots

- The equation

$$z^s = e^{\lambda(z-1)}$$

  has s complex roots in the unit disk $|z| \leq 1$ (usually proved using Rouché's theorem and $\lambda < s$)

- Denote the s roots by $z_0 = 1, z_1, \ldots, z_{s-1}$
- Exact expressions:

$$z_k = \sum_{n=1}^{\infty} \mathrm{e}^{-n\rho} \frac{(n\rho)^{n-1}}{n!} (\mathrm{e}^{\frac{2\pi i}{k}})^n, \quad k = 0, 1, ..., s-1$$

Two solutions

The pgf of Q is then given by (Crommelin 1932)

$$\mathbb{E}(z^Q) = \frac{(z-1)(s-\lambda)}{z^s - e^{\lambda(z-1)}} \prod_{k=1}^{s-1} \frac{z - z_k}{1 - z_k}$$

One readily obtains, for instance,

$$\mathbb{P}(Q=0) = \mathrm{e}^{\lambda}(-1)^{s-1}(s-\lambda) \prod_{k=1}^{s-1} \frac{z_k}{1-z_k} \quad \text{(Erlang 1917)}$$

$$= \exp\left\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{m=1}^{\infty} \mathrm{e}^{-l\lambda} \frac{(l\lambda)^{ls+m}}{(ls+m)!} \right\} \quad \text{(Pollaczek 1930)}$$

The Erlang D formula

Consider the $M/D/s$ queue: Poisson arrivals with rate λ, deterministic service times, and s servers

We have considered $Q \stackrel{d}{=} (Q + A - s)^+$, and

$$D(s, \lambda) = \mathbb{P}(\text{no waiting time}) \approx \mathbb{P}(Q = 0)$$

Some further reasoning yields

$$D(s, \lambda) = \frac{s - \lambda}{\prod_{k=1}^{s-1}(1 - z_k)} \quad \text{(Erlang 1917, Crommelin 1932)}$$

$$= \exp\Big\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{m=0}^{\infty} \mathrm{e}^{-l\lambda} \frac{(l\lambda)^{ls+m}}{(ls+m)!} \Big\} \quad \text{(Pollaczek 1930)}$$

Computational burden increases with s and $\rho = \lambda/s$

Dealing with large-capacity systems

Pollaczek (1930) obtained the asymptotic result

$$D(s,\lambda) = 1 - \frac{1}{1-\rho}\frac{(\rho e^{1-\rho})^s}{\sqrt{2\pi s}}\left(1 + \mathcal{O}(s^{-1})\right)$$

Pollaczek comments:

> *Cette formule approximative devient inutilisable dans le cas le plus important où, le nombre s des lignes parallèles étant grand, le coefficient de rendement ρ tend vers l'unité, c'est-à-dire où, pour un grand faisceau de lignes, l'on tend à approcher de l'état idéal d'une utilisation parfaite.*

Heavy traffic

Pollaczek proposed to scale the system such that $\rho = 1 - \gamma/\sqrt{s}$, with γ kept constant, and proves

$$D(s,\lambda) = \frac{1}{2\pi \mathrm{i}} \oint_C \log\left(1 - e^{z^2/2+\gamma z}\right) \frac{\mathrm{d}z}{z} + \mathcal{O}(s^{-1})$$

with C a contour to the left of and parallel to the imaginary axis

An equivalent scaling is $s = \lambda + \beta\sqrt{\lambda}$ (square-root staffing), β fixed and $\lambda \to \infty$

$$\beta = \frac{s-\lambda}{\sqrt{\lambda}}, \quad \gamma = \frac{s-\lambda}{\sqrt{s}} = \beta\rho^{\frac{1}{2}}$$

The heavy-traffic limit

Set $s = \lambda + \beta\sqrt{\lambda}$, $\beta > 0$ fixed, and again start from

$$Q_\lambda \stackrel{d}{=} (Q_\lambda + \mathrm{Pois}(\lambda) - s)^+$$

Then

$$\begin{aligned}
\frac{1}{\sqrt{\lambda}} Q_\lambda &\stackrel{d}{=} \frac{1}{\sqrt{\lambda}} (Q_\lambda + \mathrm{Pois}(\lambda) - s)^+ \\
&= \left(\frac{1}{\sqrt{\lambda}} Q_\lambda + \frac{\mathrm{Pois}(\lambda) - s}{\sqrt{\lambda}} \right)^+ \\
&= \left(\frac{1}{\sqrt{\lambda}} Q_\lambda + \frac{\sum \mathrm{Pois}(1) - \lambda}{\sqrt{\lambda}} - \beta \right)^+
\end{aligned}$$

Hence, for $Q^* = \lim_{\lambda\to\infty} Q_\lambda/\sqrt{\lambda}$ we have $Q^* \stackrel{d}{=} (Q^* + N(-\beta, 1))^+$

Gaussian random walk

$$\begin{aligned} Q^* &\stackrel{d}{=} (Q^* + N(-\beta, 1))^+ \\ &\stackrel{d}{=} \max\{0, X_1, X_1 + X_2, \ldots\} =: M_\beta \end{aligned}$$

with $X_1, X_2, \ldots$ independent and normally distributed random variables with mean $-\beta < 0$ and variance 1. This implies

$$\lim_{\lambda \to \infty} D(\lambda + \beta\sqrt{\lambda}, \lambda) = \mathbb{P}(M_\beta = 0)$$

and Pollaczek already proved that

$$\mathbb{P}(M_\beta = 0) = \frac{1}{2\pi \mathrm{i}} \oint_C \log\left(1 - e^{z^2/2 + \beta z}\right) \frac{\mathrm{d}z}{z}$$

$$
\begin{array}{ccc}
Q_{\lambda,n} & \Rightarrow & Q_{\lambda,\infty} \\
\Downarrow & & \Downarrow \\
Q_n^* & \Rightarrow & Q_\infty^*
\end{array}
$$

$$
\begin{array}{c}
M/D/s \text{ queue} \\
\Downarrow \\
\text{Gaussian random walk}
\end{array}
$$

The Gaussian random walk

An open problem

> *Despite the apparent simplicity of the problem, there does not seem to be an explicit expression even for* $\mathbb{E}M_\beta$*..., but it is possible to give quite sharp inequalities and asymptotic results for small* β *(John Kingman, 1965).*

Kingman showed that for $\beta \downarrow 0$

$$\mathbb{E}M_\beta = \frac{1}{2\beta} - c + \mathcal{O}(\beta)$$

where

$$c \;=\; \frac{1}{\sqrt{2\pi}} \sum_{n=1}^{\infty} \frac{1}{\sqrt{\sqrt{n}(\sqrt{n}+\sqrt{n-1})^2}} \;=\; 0.5826...$$

Riemann zeta function

The Riemann zeta function ζ is defined as

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}, \quad \operatorname{Re} s > 1$$

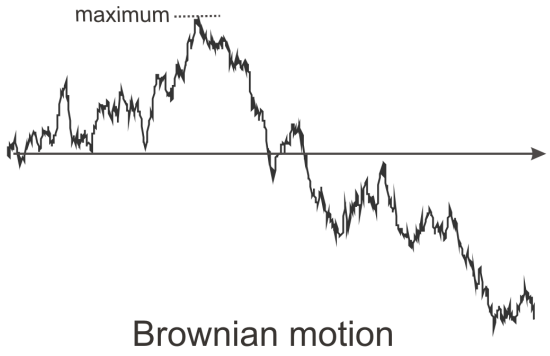
This definition is extended by analytic continuation to the entire complex plane except $s = 1$, where ζ has a simple pole

In Janssen-JvL ('07) we derived for $0 < \beta < 2\sqrt{\pi}$

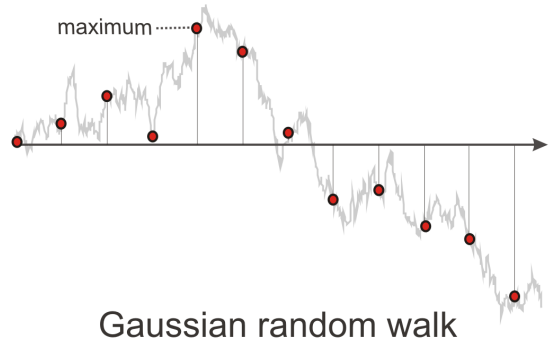
$$\mathbb{E}M_\beta = \frac{1}{2\beta} + \frac{\zeta(\frac{1}{2})}{\sqrt{2\pi}} + \frac{1}{4}\beta + \frac{\beta^2}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-\frac{1}{2} - r)}{r!(2r+1)(2r+2)} \left(\frac{-\beta^2}{2}\right)^r$$

- Chernoff (1965) and Kingman (1965) identified $\zeta(1/2)/\sqrt{2\pi}$
- Siegmund (1985) identified $\beta/4$
- Chang & Peres (1997) identified $\beta^2\zeta(-1/2)/2(2\pi)^{1/2}$

Sampled version of Brownian motion



Brownian motion



Gaussian random walk

Brownian motion with negative drift $-\beta$:

$$\mathbb{E}M_{\mathrm{BM}} = \frac{1}{2\beta}$$

$$\mathbb{E}M_{\beta} = \frac{1}{2\beta} + \frac{\zeta(\frac{1}{2})}{\sqrt{2\pi}} + \frac{1}{4}\beta + \frac{\beta^2}{\sqrt{2\pi}} \sum_{r=0}^{\infty} \frac{\zeta(-\frac{1}{2}-r)}{r!(2r+1)(2r+2)} \left(\frac{-\beta^2}{2}\right)^r$$

The Gaussian random walk gives heavy-traffic approximations for the bulk service queue. We know that

$$Q^* \approx \sqrt{\lambda} M_\beta$$

and

$$M_\beta \stackrel{d}{=} \mathrm{Exp}(2\beta)$$

Corrected diffusion approximations
(with Janssen and Zwart (OR, 2011))

A close look at the Poisson distribution

A well known relation:

$$\mathbb{P}(\text{Pois}(\lambda) \leq s) = \sum_{j=0}^{s} e^{-\lambda} \frac{\lambda^j}{j!} = \frac{\Gamma(s+1,\lambda)}{\Gamma(s+1)} = \frac{1}{s!} \int_{\lambda}^{\infty} e^{-t} t^s \mathrm{d}t$$

We want to bring this into Gaussian form...

$$\begin{aligned}
\mathbb{P}(\text{Pois}(\lambda) \leq s) &= \frac{1}{s!} \int_{\lambda/s}^{\infty} e^{-su} (su)^s s \mathrm{d}u \\
&= \frac{s^{s+1} e^{-s}}{s!} \int_{\lambda/s}^{\infty} e^{s(1-u)} u^s \mathrm{d}u \\
&= \frac{p(s)\sqrt{s}}{\sqrt{2\pi}} \int_{\rho}^{\infty} e^{s(1-u+\ln u)} \mathrm{d}u
\end{aligned}$$

with $\rho = \lambda/s$ and $p(s) = s^s e^{-s} \sqrt{2\pi s}/s!$

$$\mathbb{P}(\mathrm{Pois}(\lambda) \leq s) = \frac{p(s)\sqrt{s}}{\sqrt{2\pi}} \int_{\rho}^{\infty} e^{s(1-u+\ln u)} \mathrm{d}u$$

Define y as the solution of $y + \ln(1-y) = -\frac{1}{2}x^2, \ x \in \mathbb{C}$

Let $\alpha = \sqrt{-2s(1-\rho+\ln\rho)}$ with $\mathrm{sign}(\alpha) = \mathrm{sign}(1-\rho)$

We get

$$\mathbb{P}(\mathrm{Pois}(\lambda) \leq s) = \frac{p(s)}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{1}{2}x^2} y'(x/\sqrt{s}) \mathrm{d}x$$

We can show that

$$y(x) = \sum_{n=1}^{\infty} a_n x^n, \quad |x| < 2\sqrt{\pi}$$

with

$$a_1 = 1, \; a_2 = -\frac{1}{3}, \; a_3 = \frac{1}{36}, \; a_4 = \frac{1}{270}, \; a_5 = \frac{1}{4320}$$

Combining

$$\mathbb{P}(\text{Pois}(\lambda) \leq s) = \frac{p(s)}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{1}{2}x^2} y'(x/\sqrt{s}) \mathrm{d}x$$

with $y(x) = \sum_{n=1}^{\infty} a_n x^n$ and $p(s) = \frac{s^s e^{-s} \sqrt{2\pi s}}{s!} \sim 1 - \frac{1}{12s} + \ldots$
yields

$$\mathbb{P}(\text{Pois}(\lambda) \leq s) \sim \Phi(\alpha) + \frac{2}{3\sqrt{s}} \phi(\alpha) - \frac{1}{12s} \alpha\phi(\alpha) + \ldots$$

From

$$\mathbb{P}(\text{Pois}(\lambda) \leq s) = \frac{p(s)}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{1}{2}x^2} y'(x/\sqrt{s}) \mathrm{d}x$$

and $\mathbb{P}(\text{Pois}(\lambda) = s) = p(s)\phi(\alpha)$ we get

$$B(s,\lambda)^{-1} = \frac{\sqrt{s}}{\phi(\alpha)} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{1}{2}x^2} y'(x/\sqrt{s}) \mathrm{d}x$$

- If we use $y'(x) \approx 1$ and $\alpha \approx \beta$, we obtain

$$B(s,\lambda)^{-1} \approx \frac{\sqrt{s}}{\phi(\alpha)} \Phi(\alpha) \approx \frac{\sqrt{s}}{\phi(\beta)} \Phi(\beta)$$

- Both approximations can be justified

A real benefit of our approach is that it is possible to obtain bounds for y'

Appropriate bounds for y' can be guessed from the series expansion

$$y'(x) = 1 - \frac{2}{3}x + \frac{1}{12}x^2 + \frac{2}{135}x^3 + \frac{1}{864}x^4 + \ldots, \quad |x| < 2\sqrt{\pi}$$

We can for instance prove that, for $x \leq 0$,

$$\begin{aligned} y'(x) &\leq 1 - \frac{2}{3}x + \frac{1}{12}x^2 \\ y'(x) &\geq 1 - \frac{2}{3}x + \frac{1}{12}x^2 + \frac{2}{135}x^3 \end{aligned}$$

Implications for Erlang B

1. Full asymptotic series expansions for the regime $s = \lambda + \beta\sqrt{\lambda}, s \to \infty$. For example:

$$B(s,\lambda)^{-1} = \frac{\sqrt{s}\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3} + \frac{1}{12\sqrt{s}}\left(\frac{\Phi(\alpha)}{\phi(\alpha)} - \alpha\right) + \mathcal{O}(1/s)$$

2. Bounds that hold for all values of s and λ. For example:

$$B(s,\lambda)^{-1} \geq \frac{\sqrt{s}\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3},$$

$$B(s,\lambda)^{-1} \leq \frac{\sqrt{s}\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3} + \frac{\sqrt{s}}{\phi(\alpha)(12s-1)}$$

Numerical illustration

Values for $s = \lambda + \beta\sqrt{\lambda}$ with $\beta = 1$

s	λ	$B(s,\lambda)$	Erlang	LB	UB
1	0.38	0.2764	0.4653	0.2627	0.2870
2	1.00	0.2000	0.2876	0.1953	0.2044
3	1.69	0.1645	0.2208	0.1620	0.1671
5	3.20	0.1282	0.1606	0.1270	0.1294
10	72.9	0.0910	0.1065	0.0906	0.0914
20	16.0	0.0644	0.0719	0.0643	0.0646
30	25.0	0.0526	0.0575	0.0525	0.0527
50	43.4	0.0407	0.0437	0.0407	0.0408
100	90.5	0.0288	0.0302	0.0288	0.0288
200	186	0.0204	0.0211	0.0204	0.0204
300	283	0.0166	0.0171	0.0166	0.0166
500	478	0.0129	0.0132	0.0129	0.0129

Insights

- A major difference between our approximation and Erlang's is the replacement of β by $\alpha = \sqrt{-2s(1-\rho+\ln\rho)}$
- Note that

$$\tfrac{1}{2}\alpha^2 = s\sum_{n=2}^{\infty}\frac{(1-\rho)^n}{n} \quad \Rightarrow \quad \alpha \approx \sqrt{s}(1-\rho) = \gamma \approx \beta$$

- We can show that $\gamma < \alpha < \beta$, and $\alpha \to \beta$ in the HW regime
- α seems most robust choice among the three

Erlang D formula

The no wait probability in the $M/D/s$ queue is known to be

$$D(s,\lambda) = \exp\Big\{ -\sum_{l=1}^{\infty} \frac{1}{l} \sum_{m=0}^{\infty} \mathrm{e}^{-l\lambda} \frac{(l\lambda)^{ls+m}}{(ls+m)!} \Big\} \quad \text{(Pollaczek 1930)}$$

The CLT says for the HW regime that

$$\begin{aligned} D(s,\lambda) &= \exp\Big\{ -\sum_{l=1}^{\infty} \frac{1}{l} \mathbb{P}(\mathrm{Pois}(\lambda l) \geq ls) \Big\} \\ &\approx \exp\Big\{ -\sum_{l=1}^{\infty} \frac{1}{l} \Phi(-\beta\sqrt{l}) \Big\} \;=:\; \mathbb{P}(M_\beta = 0) \end{aligned}$$

Our expansion $\mathbb{P}(\mathrm{Pois}(\lambda) \leq s) \sim \Phi(\alpha) + \frac{2}{3\sqrt{s}}\phi(\alpha) - \frac{1}{12s}\alpha\phi(\alpha) + \ldots$
yields $D(s,\lambda) \approx \mathbb{P}(M_\alpha = 0)$ and arbitrarily many refinements

Erlang C formula

Erlang (1917) obtained the delay probability in the $M/M/s$ queue $C(s,\lambda)$ for which

$$C(s,\lambda)^{-1} = \rho + (1-\rho)B(s,\lambda)^{-1}$$

Halfin-Whitt (1981) derive

$$\lim_{\lambda\to\infty} C(\lambda+\beta\sqrt{\lambda},\lambda) = \left(1+\frac{\beta\Phi(\beta)}{\phi(\beta)}\right)^{-1} =: C_*(\beta)$$

For $\lambda < s$ and $s \in \mathbb{N}$ we have (Janssen-JvL-Zwart '08)

$$C(s,\lambda)^{-1} \le \rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s}} + \frac{1}{\phi(\alpha)}\frac{1}{12s-1}\right)$$

$$C(s,\lambda)^{-1} \ge \rho + \gamma\left(\frac{\Phi(\alpha)}{\phi(\alpha)} + \frac{2}{3}\frac{1}{\sqrt{s}}\right)$$

S. Ramanujan (1887-1920)

A famous problem

In 1911 Ramanujan set the following problem:

$$\xi(n) = \frac{n!}{n^n}\left(\frac{1}{2}e^n - \sum_{k=0}^{n-1}\frac{n^k}{k!}\right), \quad n = 1, 2, \ldots$$

lies between $\frac{1}{2}$ and $\frac{1}{3}$
$\Rightarrow$ Proofs published by Szegö (1928) and Watson (1929)

In his first letter to Hardy in 1913 Ramanujan asserted:

$$\frac{1}{3} + \frac{4}{135(n + 8/45)} \leq \xi(n) \leq \frac{1}{3} + \frac{4}{135(n + 2/21)}$$

$\Rightarrow$ Proved by Flajolet et al. in 1995 using singularity analysis
$\Rightarrow$ We can derive an alternative proof and sharper results

A famous problem

$$\begin{aligned}\xi(n) &= \frac{n!}{n^n}\left(\frac{1}{2}e^n - \sum_{k=0}^{n-1}\frac{n^k}{k!}\right)\\ &= \frac{1}{\mathbb{P}(\mathrm{Pois}(n)=n)}\left(\frac{1}{2} - \mathbb{P}(\mathrm{Pois}(n)\le n)\right) + 1\\ &= \frac{1}{2\mathbb{P}(\mathrm{Pois}(n)=n)} - B(n,n)^{-1} + 1\end{aligned}$$

This is the case $\lambda = s = n$ and so $\alpha = 0$. We have

$$B(s,\lambda)^{-1} \le \frac{\Phi(\alpha)\sqrt{s}}{\phi(\alpha)} + \frac{2}{3} + \frac{\Phi(\alpha) - \alpha\phi(\alpha)}{12\phi(\alpha)\sqrt{s}}$$

$$B(s,\lambda)^{-1} \ge \frac{\Phi(\alpha)\sqrt{s}}{\phi(\alpha)} + \frac{2}{3} + \frac{\Phi(\alpha) - \alpha\phi(\alpha)}{12\phi(\alpha)\sqrt{s}} - \frac{4+2\alpha^2}{135s}$$

A famous problem

$$\begin{aligned}\xi(n) &= \frac{n!}{n^n}\left(\frac{1}{2}e^n - \sum_{k=0}^{n-1}\frac{n^k}{k!}\right)\\ &= \frac{1}{\mathbb{P}(\mathrm{Pois}(n)=n)}\left(\frac{1}{2} - \mathbb{P}(\mathrm{Pois}(n)\le n)\right) + 1\\ &= \frac{1}{2\mathbb{P}(\mathrm{Pois}(n)=n)} - B(n,n)^{-1} + 1\end{aligned}$$

This is the case $\lambda = s = n$ and so $\alpha = 0$. We have

$$B(n,n)^{-1} \le \frac{\sqrt{2\pi n}}{2} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{n}}$$

$$B(n,n)^{-1} \ge \frac{\sqrt{2\pi n}}{2} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{n}} - \frac{4}{135n}$$

A famous problem

$$\xi(n) = \frac{n!}{n^n}\left(\frac{1}{2}e^n - \sum_{k=0}^{n-1}\frac{n^k}{k!}\right)$$
$$= \frac{1}{\mathbb{P}(\mathrm{Pois}(n)=n)}\left(\frac{1}{2} - \mathbb{P}(\mathrm{Pois}(n)\leq n)\right) + 1$$
$$= \frac{1}{2\mathbb{P}(\mathrm{Pois}(n)=n)} - B(n,n)^{-1} + 1$$

This is the case $\lambda = s = n$ and so $\alpha = 0$. We have

$$B(n,n)^{-1} \leq \frac{\sqrt{2\pi n}}{2} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{n}}$$
$$B(n,n)^{-1} \geq \frac{\sqrt{2\pi n}}{2} + \frac{2}{3} + \frac{\sqrt{2\pi}}{24\sqrt{n}} - \frac{4}{135n}$$

The bounds are not sharp enough. The ones doing the trick follow from...

$$y'(x) \leq 1 - \frac{2}{3}x + \frac{1}{12}x^2 + \frac{2}{135}x^3 + \frac{1}{864}x^4 - \frac{1}{2835}x^5, \quad x \leq 0$$

$$y'(x) \geq 1 - \frac{2}{3}x + \frac{1}{12}x^2 + \frac{2}{135}x^3 + \frac{1}{864}x^4 - \frac{1}{2835}x^5 - \frac{139}{777600}x^6 - \frac{1}{25515}x^7 - \frac{571}{261273600}x^8 + \frac{281}{151559100}x^9, \quad x \leq 0$$

Refined square root staffing
(with Janssen and Zwart (OR, 2011) and Bo Zhang (OR, 2012))

Square root staffing principle

$C_*(\beta)$ is easier to work with than $C(s, \lambda)$, especially for large λ

Example: How many agents are necessary so that at least 50 percent of calls are answered immediately?

- The Halfin-Whitt result can be applied as follows: Determine β so that $C^*(\beta) = 0.5$
- Choose the number of agents $s = \lceil \lambda + \beta\sqrt{\lambda} \rceil$
- The same β can be used for different values of the traffic volume λ

Possible issue

Square root staffing is based on asymptotic theory. How large should a system be before the asymptotics kick in?

Refined square root staffing

- The Halfin-Whitt approximation suggests to use β_* which satisfies $C_*(\beta_*) = \epsilon$ and staffing level $s_* = \lambda + \beta_* \sqrt{\lambda}$
- We don't take β_*, but

$$\beta = \beta_* + \frac{\beta_\bullet}{\sqrt{\lambda}}$$

- This leads to a refinement of the form

$$s_\bullet \;=\; \lambda + \left(\beta_* + \frac{\beta_\bullet}{\sqrt{\lambda}}\right)\sqrt{\lambda} \;=\; \lambda + \beta_*\sqrt{\lambda} + \beta_\bullet$$

- The correction term $\beta_\bullet$ follows from the refinements of the Halfin-Whitt approximation for $C(s, \lambda)$:

$$\beta_\bullet(\epsilon) = \beta_*(\epsilon)\frac{(1-\epsilon)\left(\frac{1}{2}\beta_*(\epsilon) + \frac{1}{6}\beta_*(\epsilon)^3\right) + \epsilon\left(\frac{1}{3}\beta_*(\epsilon) + \frac{1}{6}\beta_*(\epsilon)^3\right)}{1 - \epsilon + \beta_*(\epsilon)^2}$$

λ	Exact	s_*	Difference	$s_\bullet$	Difference
1	3	2.4	0	2.9	0
2	5	4.0	0	4.5	0
5	9	8.1	0	8.7	0
10	16	14.4	-1	15.0	0
20	27	26.3	0	26.9	0
50	61	60.0	0	60.6	0
100	115	114.2	0	114.7	0
200	221	220.0	0	220.6	0
500	533	531.7	-1	532.3	0
1000	1046	1044.9	-1	1045.4	0

Results for $\epsilon = 10^{-1}$; $\beta_* = 1.4202$ and $\beta_\bullet = 0.5666$.

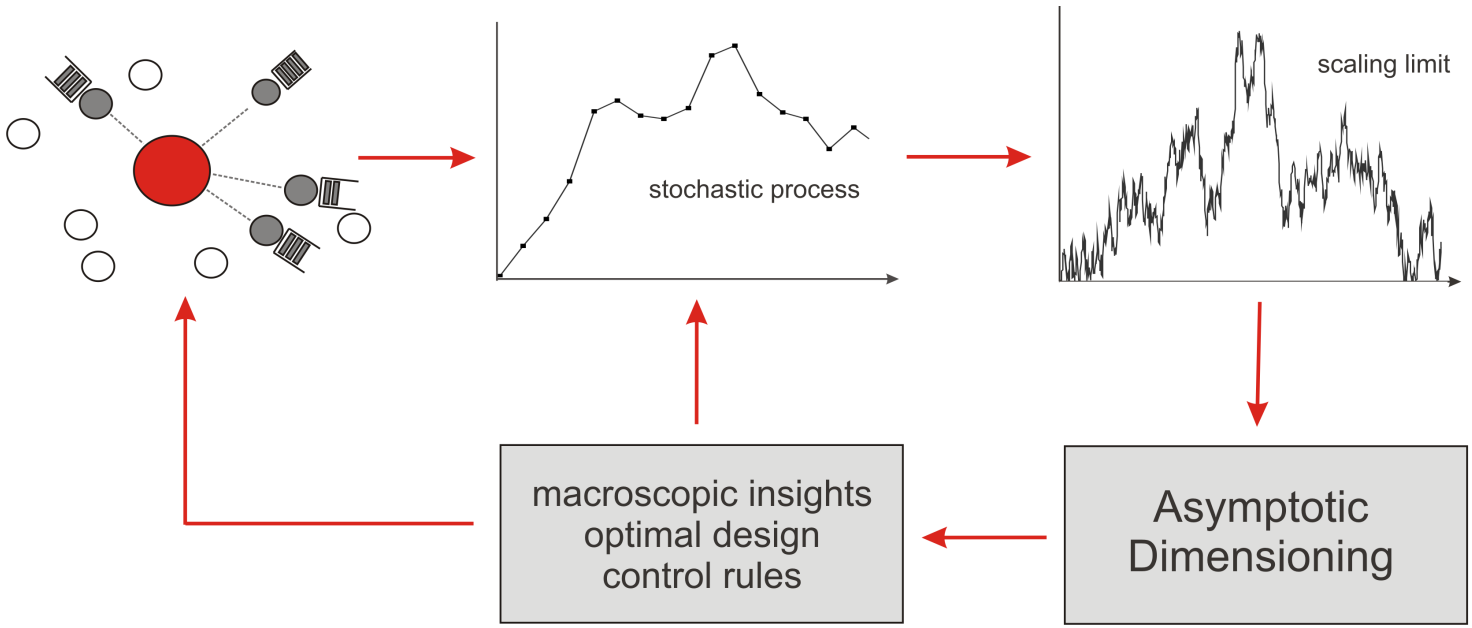
λ	Exact	s_*	Difference	$s_\bullet$	Difference
1	6	4.1	-1	6.0	1
2	9	6.4	-2	8.3	0
5	14	11.9	-2	13.8	0
10	22	19.8	-2	21.7	0
20	36	33.9	-2	35.8	0
50	74	72.0	-1	73.9	0
100	134	131.1	-2	133.0	0
200	246	244.0	-1	245.9	0
500	572	569.6	-2	571.5	0
1000	1101	1098.5	-2	1100.4	0

Results for $\epsilon = 10^{-3}$; $\beta_* = 3.1153$ and $\beta_\bullet = 1.9197$.

λ	Exact	s_*	Difference	$s_\bullet$	Difference
1	10	5.7	-4	9.8	0
2	13	8.7	-4	12.8	0
5	20	15.6	-4	19.7	0
10	29	25.0	-3	29.1	1
20	46	41.2	-4	45.3	0
50	88	83.6	-4	87.7	0
100	152	147.6	-4	151.7	0
200	272	267.3	-4	271.4	0
500	611	606.4	-4	610.5	0
1000	1155	1150.5	-4	1154.6	0

Results for $\epsilon = 10^{-6}$; $\beta_* = 4.7615$ and $\beta_\bullet = 4.0979$.

Some sort of summary


stochastic process
scaling limit
macroscopic insights
optimal design
control rules
Asymptotic
Dimensioning