Approximate Performance with Non-Exponential Holding Times

Ward Whitt

Columbia University

May 8, 2003
The Bread-and-Butter Model of Call Centers
The Erlang C Model

The $M/M/s/$\infty Queue
The $M/M/s/\infty$ Queue

- $s$ servers (with large $s$; e.g., $s = 100$)
- unlimited waiting room
- Poisson arrival process
- IID exponential holding (service) times
Important Extensions

- multiple types - skill-based routing
- abandonment, blocking and retrials
- time-varying arrival rates
- multiple sites - networking
- non-Poisson arrival process
- non-exponential holding times
How to evaluate performance in customer call centers?
The Holy Grail of Call Centers

Service Level
Service Level

80/20 rule

80% of calls answered in 20 seconds

or the $X/Y$ rule
Why is Service Level so widely used?

Service Level
Why is Service Level so widely used?

• relates to the real performance goal

• introduces reasonable slack for agents

• \( P(\text{Wait} > 0) \) hard to measure

• ?

• ?

• robust for holding-time distribution
Service Level

Why?
Why are the $M/M/s/\infty$ model and the 80/20 service level so widely applied?
Proposed Partial Answer:

The $M/M/s/\infty$ model and the service level are robust to model detail e.g., to the holding-time distribution.
Possible Explanations

- evaluate through sizing

- **insensitivity** in $M/G/s/\infty$ and $M/G/s/0$

- $s$ large: deterministic **fluid approximation**

- $s$ large: **diffusion limit** for $G/H_2^*/s$ **

- the **infinite-server view** **
The Infinite-Server View

Look at the $M/G/\infty$ Model
\( M/G/\infty \)

an approximation for

\( M/G/s/\infty + A \)

(with customer abandonment)

when

time to abandon \( \approx \) holding time
\[ M/G/s/\infty + A \]

\[ S = \text{service time} \]

\[ G(t) = P(S \leq t), \quad t \geq 0 \]

\[ A = \text{time to abandon} \]

\[ H(t) = P(A \leq t), \quad t \geq 0 \]

Assume that \( H = G \).
\[ M/G/\infty \]

an approximation for

\[ M/G/s/\infty + A \]

when

\[ H = G \]

(exact for exponential)
Does this make sense?

We are interested in

\[ P(\text{Wait} > 20 \text{ seconds}) = 0.2 \]

There is no waiting in \( M/G/\infty \).
Approximate conditional waiting time given that customer does not abandon in $M/G/s/\infty + A$, where $H = G$, by the first passage time down to $s - 1$ starting in steady state in model $M/G/\infty$ with cdf $G$. 
**Notation**

\[ W = \text{waiting time in } M/G/s/\infty + A \]

\[ T_{s-1} = \text{first passage time down to } s - 1 \text{ in } M/G/\infty \]

\[ Q = \text{number in system in } M/G/\infty \]

\[ R_i = \text{remaining service time of customer } i \text{ in service in } M/G/\infty \]
More Notation

$\lambda = \text{arrival rate}$

$S = \text{time to serve or abandon}$

$G(t) = P(S \leq t)$

$G_c(t) = 1 - G(t)$

$G_e = \text{stationary-excess distribution}$

$G_e(t) = \frac{1}{ES} \int_0^t G_c(x) \, dx$

Let $ES = 1$. 
Now exploit the wonders of the infinite-server model.
Key Fact Number 1

$Q$ has a Poisson distribution with mean $\lambda$

\[ P(Q = k) = \frac{e^{-\lambda} \lambda^k}{k!} \]
Key Fact Number 2

Remaining holding times $R_i$ of customers in service are IID with distribution $G_e$

$$P(R_i \leq t) = G_e(t) = \int_0^t G^c(x) \, dx$$

(Recall that $ES = 1$.)
Key Fact Number 3

\[ P(T_{s-1} > t) = P(Z_t > s - 1), \]

where \( Z_t \) is Poisson with mean \( \lambda G^c_e(t) \)

(independent thinning with probability \( G^c_e(t) \))

\[ P(Z_t = k) = \frac{e^{-\lambda G^c_e(t)}(\lambda G^c_e(t))^k}{k!} \]
Normal Approximation

\[ P(W > t) \approx P(T_{s-1} > t) = P(Z_t > s - 1) \approx P\left( N(0, 1) > \frac{s - 0.5 - \lambda G_e^c(t)}{\sqrt{\lambda G_e^c(t)}} \right) \]
Capacity Planning: Use the service level

Choose the capacity $s$ so that

$$P(W > t) = 0.20 \quad \text{for} \quad t = 20 \text{ seconds}$$

by the 80/20 rule
Since $P(N(0, 1) > 0.84) = 0.2,$

$$s - 0.5 - \lambda G^c_e(t) \sqrt{\lambda G^c_e(t)} = 0.84$$

or

$$s \approx \lambda G^c_e(t) + 0.84\sqrt{\lambda G^c_e(t)}$$
What about the time $t$?

Since $ES = 1$ and 20 seconds is typically much smaller than the mean service time,

$$G_e(t) = \int_0^t G^c(x) \, dx \approx G^c(0)t = t$$

**independent of the holding-time cdf** $G$. 
Staffing Formula

Given that $G_e(t) \approx t$, $G_e^c(t) \approx 1 - t$.

If $t \approx 0.05$, then $G_e^c(t) \approx 0.95$ and

$$s \approx \lambda G_e^c(t) + 0.84 \sqrt{\lambda G_e^c(t)}$$

$$\approx 0.95\lambda + 0.84\sqrt{0.95\lambda},$$

independent of the holding-time distribution $G$. 
More on Staffing

We can see impact of a change from 80/20 to X/Y, e.g., 90/40:

\[ s \approx \lambda G_e^c(t) + z(X)\sqrt{\lambda G_e^c(t)} \]
\[ \approx 0.95\lambda + 0.84\sqrt{0.95\lambda} \quad (80/20) \]
\[ \approx 0.90\lambda + 1.28\sqrt{0.90\lambda} \quad (90/40) \]

independent of the holding-time distribution \( G \).
Summary

From the $M/G/\infty$ perspective, we see that service level is a robust performance measure.
Heavy-Traffic Limit for $M/H^*_2/s/\infty$

The approximate delay probability $P(W > 0)$
depends on the service distribution only through its mean,
even though the diffusion limit is not the same as for $M/M/s/\infty$. 