Service Engineering

Class 11

Non-Parametric Models of a Service System; GI/GI/1, GI/GI/n: Exact & Approximate Analysis.

- G/G/1 Queue: Virtual Waiting Time (Unfinished Work).
- GI/GI/1: Lindley's Equations and Stability.
- M/GI/1 (=M/G/1): The Khintchine-Pollaczek Formula.
- G/G/1 and G/G/n: Allen-Cunneen Approximation; Kingman's Exponential Law.
- Call Centers: The M/G/n+G queue.
- Queueing Systems with Priorities (Recitation).

The G/G/1 Queue

Non-Parametric model of a service station.
"Exact" model but Approximate analysis (vs. Markovian queues).

We start with Single-Server models.
(Will be generalized to Multi-Servers.)

Building Blocks:

- **Arrivals**: Counting Process $A = \{A(t), t \geq 0\}$, with **arrivals** (jumps) at $\{A_1, A_2, \ldots, A_i, \ldots\}$.
- **Services**: $\{S_1, S_2, \ldots, S_i, \ldots\}$ denote **service durations**.
- First Come First Served (FCFS = FIFO here).

Work arriving up to time t: $\Sigma_{i=1}^{A(t)} S_i$, $t \geq 0$.

Virtual Waiting Time $V = \{V(t), t \geq 0\}$: V(t) is the amount of time that a (possibly virtual) arrival *at time t* would have to wait.

(Sometimes referred to as *Unfinished Work*, but Virtual Waiting Time is more appropriate for multi-server queues.)

It is possible to create the sample paths of V from those of Work.
We prefer a direct visual (seesaw) construction:
Assume $V(0) = 0$,...

GI/GI/1

Number in system is NOT a Markov process (in contrast to Markovian queues).

For some analysis need some minimal **Assumptions:**

- Arrival times $A_1, A_2, \ldots, A_n, \ldots$ are jumps of a **renewal process**:
  - **Inter-arrival times** $T_i = A_i - A_{i-1},\ i \geq 1$, are iid ($A_0 = 0$).
  - $\mathrm{E}[T_1] = 1/\lambda;\quad C^2(T_1) = C_a^2.$
  - Note: λ = Arrival rate.
- **Service durations** $S_1, S_2, \ldots, S_n, \ldots$ are iid.
  - $\mathrm{E}[S_1] = 1/\mu;\quad C^2(S_1) = C_s^2.$
  - Note: μ = Service rate.
- Independence between arrivals and services.
- Service discipline is First Come First Served .

G/G/1: Lindley's Equations (1952)

Let $W_q(n)$ = Waiting Time of customer number n, n=1,2,...

Recursion:

$$W_q(n) = \begin{cases} 0, & A_n \geq \overbrace{A_{n-1} + W_q(n-1) + S_{n-1}}^{\text{departure time of customer } (n-1)} \\ A_{n-1} + W_q(n-1) + S_{n-1} - A_n, & \text{otherwise.} \end{cases}$$

In short,

$$\begin{aligned} W_q(n) &= [A_{n-1} - A_n + W_q(n-1) + S_{n-1}]^+ \quad (\ x^+ = x \vee 0) \\ &= [W_q(n-1) + X_{n-1}]^+ , \end{aligned}$$

where $X_{n-1} = S_{n-1} - T_n, \quad n \geq 2$ (iid in GI/GI/1).

Note: $X_1, X_2, \ldots$ are data, enabling calculation of successive waiting times via

Lindley's Equations:

$$\boxed{\begin{aligned} &W_q(n) = [W_q(n-1) + X_{n-1}]^+, \ n \geq 2, \\ &W_q(1) = 0. \end{aligned}}$$

Useful: Recursion amenable for spreadsheet calculations.

GI/GI/1 Queue: Stability

Explicit representation of $W_q(n)$:

$$W_q(n) \;=\; \max(0, X_{n-1}, X_{n-1}+X_{n-2}, \ldots, X_{n-1}+X_{n-2}+\ldots+X_1)\,.$$

(Try unfolding the recursion with, say, $n = 3$.)

We have $GI/GI/1$, in which X_i's are iid. Hence,

$$W_q(n) \;\stackrel{d}{=}\; \max(0, X_1, X_1 + X_2, \ldots, X_1 + X_2 + \ldots + X_{n-1})\,.$$

Define the **Random Walk** $Y_k = \sum_{i=1}^{k} X_i, \;\; Y_0 = 0.$

Fact: $\mathrm{P}\{\sup_{k\geq 0} Y_k < \infty\} = 1 \Leftrightarrow E[X_1] < 0.$

Proof: By the Strong Law of Large Numbers,

$$\lim_{k\to\infty} \frac{1}{k}(X_1 + \ldots X_k) = \mathrm{E}[X_1] < 0.$$

Hence, $\lim_{k\to\infty} Y_k = -\infty$, and their sup = max is finite.

Consequence:
$W_q(n) \stackrel{d}{\to} \sup_{k\geq 0} Y_k < \infty$ as $n \to \infty$
if and only if $\lambda < \mu$ $(\rho = \lambda/\mu < 1)$.

Conclude: **GI/GI/1 stable if and only if $\rho < 1$.**

From now on, all models are in steady-state.

M/GI/1 (=M/G/1) in Steady-State
The Khintchine-Pollaczek Formula

M/G/1 Queue: Poisson arrivals,
generally distributed (iid) service durations.

Theorem. (Khintchine-Pollaczek)

$$E(W_q) = E(S) \cdot \frac{\rho}{1-\rho} \cdot \frac{1+C^2(S)}{2}.$$

Remarks:

- A remarkable second-moment formula quantifying congestion.
- **"Congestion Index"** $= E(W_q)/E(S)$ (unitless).
- Decomposes "Congestion" into two multiplicative components (the two congestion-drivers, in our simple M/G/1 context):
  - **Server-Utilization**: ρ ;
  - **Stochastic-Variability**, arising from Services: $C(S)$; ("Where are the Arrivals"? - to be discussed momentarily).
- Quantifies the effect of the service-time distribution (via its CV); for example, changing from a human-service to a robot.
- The Number-in-System is not Markov; however at instants of service completions it is an (embedded) Markov-chain.

Illuminating derivation, with the ingredients:
Little, PASTA, Biased sampling; Wald.

Congestion Index : $\frac{E W_q}{E S} = \frac{\bar{L}_q}{N p}$ ← observable   מש/ק

מעריב 25/II/92

כך דופקים את האזרח

משרד הרישוי חולון

מחכים שעתיים בשביל 19 שניות

מאת אבי פלד

[illegible] למשרד הרישוי בחולון [illegible] דרך [illegible] מקונה ישראלי – [illegible] נמשכת 45 דקות. בשעה 08.09 נמצאים באולם ההמתנה כ־60 איש. באשנב המודיעין יושב פקיד. איפה התור לבדיקת ההיסטוריה של מכונית שאני רוצה לקנות?

"תעודת זהות, רשיון רכב או יפוי כוח, באשנב שלוש" – ענה הפקיד המיומן.

אני לא מכיר את בעל הרכב ואני רוצה לבדוק אם הרכב מעוקל.

"תעודת זהות, רשיון רכב או יפוי כוח, באשנב שלוש", קיבלתי את התשובה באותה אדישות.

אבל...

"תעודת זהות, רשיון רכב או יפוי כוח, באשנב שלוש", חזר הפקיד תקליט. צריך להבין, באשנב שלוש אין סיכוי לקבל מידע על רכב, אם אין לך את רשיון הרכב.

בשעה 08.17 נפל המחשב במשרד הרישוי. אחרי שבע דקות נשמעה במערכת הכריזה ההודעה הבאה: "עקב תקלה במחשב יופסק מתן השירות. לא ידוע מתי יתחדש. ניתן לקבל הארכת תוקף לרשיון. [illegible] איתכם הסליחה". הפקידות המשועממות קפצו על ההזדמנות וניצלו את הזמן הפנוי בשיחה עם [illegible], עם אמא או עם חברה.

בתור לאשנב שלוש, שבו צעיר, שתפס מקום כבר ב־07.30 בבוקר. [illegible] הוא הדליק סיגריה. [illegible] עצמו מקרב הממתינים, הודיע שהשירות לא יתחדש עד הצהריים. אחר "מבין במחשבים", הסביר ש"צ'יפ אחד התבלבל".

אבל ב־09.18 החל המחשב לגלות סימני חיים. הבשורה המשמחת התפשטה במהירות והמונים הצטרפו לתור שבסופו מנפיקים כפל רשיון נהיגה.

לאחר 42 דקות הגיע תורי. חשבון פשוט גילה שטיפול באדם נמשך בממוצע כארבע דקות. הצגתי תעודת עיתונאי וביקשתי כפל רשיון נהיגה. בתוך 19 שניות, על השעון, עזבתי את המקום עם המסמך המבוקש. שיפור ניכר ביחס לממוצע.

ובעניין בירור ההיסטוריה של המכונית: מלשכת דובר משרד התחבורה נמסר, שאין מוסרים פרטים על רכב אלא ברשות בעליו, מחשש לפגיעה בצינעת הפרט.

גישה

תור התחלה 60 איש

יחס

מחשב נפל

הכתבה הנ"ל אינה מספרת את "איכות המתנה" ו"[illegible]", מעוררת ואחר כך [illegible]

- חוסר [illegible] / מתן [illegible] לגבי הזמן

- [illegible] בפירים והתנהגות [illegible] הזמן

Derivation of Khintchine-Pollaczek

For customer $n = 1, 2, \ldots,$ denote

$$\begin{aligned} W_q(n) &= \text{waiting-time of } n\text{-th customer.} \\ R(n) &= \text{residual service time, at time of the } n\text{-th arrival;} \\ (&= 0, \text{for arrivals without waiting).} \\ L_q(n) &= \#\text{ of customers in queue, at time of } n\text{-th arrival.} \\ \{S_n\} &= \text{sequence of service-times.} \end{aligned}$$

$$\begin{aligned} W_q(n) &= R(n) + \sum_{k=n-L_q(n)}^{n-1} S_k \,, && n \geq 1\,. \\ EW_q(n) &= ER(n) + E(S_1) \cdot EL_q(n) \,, && \text{by } \textbf{Wald}, \\ E(W_q) &= E(R) + E(S_1)E(L_q) \,, && n \uparrow \infty\,, \text{ assuming} \\ & && \exists \text{ limit} + \textbf{ PASTA}, \\ &= E(R) + \lambda E(S_1)E(W_q) \,, && \text{by } \textbf{Little}, \\ E(W_q) &= E(R) + \rho E(W_q) \,, && \rho < 1 \Leftrightarrow \exists \text{ steady-state}, \\ E(W_q) &= E(R)/(1-\rho)\,. \end{aligned}$$

Left to calculate E$[R]$?
Via **Biased Sampling** (see next page):

- $\rho =$ Prob. of arriving to a busy server. (**PASTA+Little**)

- $E(R) = (1-\rho) \cdot 0 + \rho \cdot E(S) \cdot \dfrac{1 + C^2(S)}{2}.$ q.e.d.

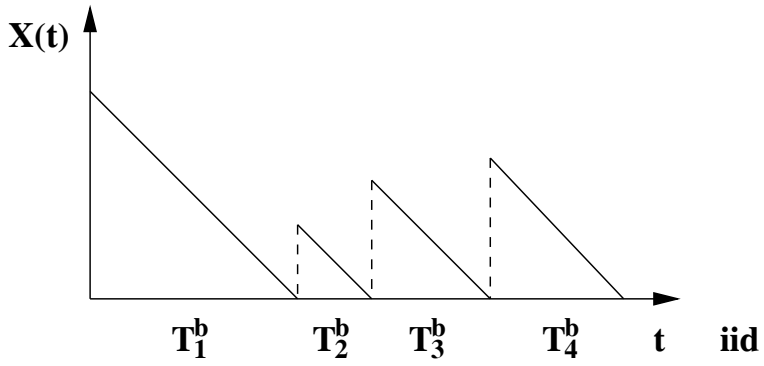
Biased Sampling (via PASTA)

A *renewal process* is a counting process with iid interarrivals.

Descriptions:	$R = \{R(t),\ t \geq 0\}$ or	$\{T_1, T_2, \ldots\}$ iid, or	$\{S_1, S_2, \ldots\}$
Example:	Poisson	exponential	Erlang

Story: Buses arrive to a bus stop according to a renewal process $R_b = \{R_b(t), t \geq 0\}$.
T_i^b — times between arrivals of the buses.
Passengers arrive to the bus stop in a completely random fashion (Poisson).
S_i^p — arrival times of the passengers.

Question: How long, on average, do they wait? Plan service-level.



$A = \{A(t),\ t \geq 0\}$ = Poisson arrivals of passengers.
$X = \{X(t),\ t \geq 0\}$ = state = *Virtual waiting time*.

PASTA: $\lim_{T\uparrow\infty} \frac{1}{T} \int_0^T X(t)dt = \lim_{N\uparrow\infty} \frac{1}{N} \sum_{n=1}^{N} X(S_n^p-) = \bar{\tau}$

$$\begin{aligned}
\Rightarrow \bar{\tau} &= \frac{1}{T} \cdot \text{ (area under } X\text{, over } [0,T]) \\
&\approx \frac{1}{T} \cdot \left(\frac{1}{2}(T_1^b)^2 + \frac{1}{2}(T_2^b)^2 + \cdots + \frac{1}{2}(T_{R_b(T)}^b)^2 \right) \\
&= \frac{R_b(T)}{T} \cdot \frac{1}{2} \cdot \frac{T_1^2 + \cdots + T_{R_b(T)}^2}{R_b(T)} \xrightarrow[T\uparrow\infty]{} \frac{1}{E(T_1^b)} \cdot \frac{1}{2} \cdot E(T_1^b)^2 \text{, by SLLN} \\
&= \underbrace{\frac{1}{2}E(T_1^b)}_{\text{“Deterministic” answer}} \quad \underbrace{[1 + c^2(T_1^b)]}_{\text{Bias, due to variability}}, \qquad c = \frac{\sigma}{E} \text{ coefficient of variation.}
\end{aligned}$$

Check Poisson bus arrivals to derive Paradox:

1(“stochastic” answer) = $\frac{1}{2}$ (“deterministic” answer).

GI/GI/1
The Allen-Cunneen Approximation

Assume General Arrivals (renewal) and General Services (iid):

$$E(W_q) \approx E(S) \cdot \frac{\rho}{1-\rho} \cdot \frac{C^2(A)+C^2(S)}{2}.$$

Mean Service Time | **Utilization Availability** | **Stochastic Variability**

Facts:

- Exact for M/G/1.
- Upper bound in general.
- Asymptotically exact as $\rho \uparrow 1$ - in **Heavy Traffic**. (But then can actually say much more - momentarily).

Internalize: Assume $C^2(A) = C^2(S) = 1$, as in M/M/1:

$$\frac{E(W_q)}{E(S)} = \frac{\rho}{1-\rho}.$$

Now substitute $\rho = 0.5\ (1), 0.8\ (4), 0.9\ (10), 0.95\ (19)$.
Finally think in terms of "5 minute telephone service-time" (or "1 week job-shop processing-time").

Other Measures of (Average) Performance:

$$E(W) = E(S) + E(W_q), \quad E(L_q) = \lambda E(W_q),$$
$$E(L) = \lambda E(W) = E(L_q) + \rho.$$

GI/GI/1
Kingman's Exponential Law

Fact (Kingman, 1961):
In heavy-traffic, **"Waiting-Time is Exponential"**.
Get its mean from the Allen-Cunneen approximation.

Formally: **Kingman's Exponential Law of Congestion:**

$$\frac{W_q}{E(S)} \approx \begin{cases} \exp\left(\text{mean } = \dfrac{1}{1-\rho} \cdot \dfrac{C^2(A)+C^2(S)}{2}\right) & , \text{ wp } \rho, \\ 0 & , \text{ wp } 1-\rho, \end{cases}$$

Remarks:

- **"Congestion Index"** $= E(W_q)/E(S)$ (unitless):
  The Allen-Cunneen Approximation.
- Decomposes "Congestion" into two multiplicative components (the two congestion-drivers, in our simple G/G/1 context):
  - **Server-Utilization**: ρ ;
  - **Stochastic-Variability**, which arises from **Arrivals** - $C(A)$ and **Services** - $C(S)$.
- Both ρ and $C(S)$ effect congestion non-linearly – draw congestion curves.
- M/M/1 – Special case in which $C^2(A) = C^2(S) = 1$: Exact.
  M/G/1 – Only $E(W_q)$ is Exact.

Justifying the Law of Congestion: Why $W_q \approx \exp\left(\text{mean } = \frac{1}{\mu}\ \frac{\rho}{1-\rho}\ \frac{C_a^2+C_s^2}{2}\right)$?

via *Strong Approximations*. (*Heavy Traffic* Theory)

$$
\begin{aligned}
S(t) &= \sum_{n=1}^{\lfloor t \rfloor} S_n \approx \frac{1}{\mu}\, t + \sigma_s B_s(t) && \text{(Donsker for partial sums)} \\
A(t) &\approx \lambda t + \lambda^{3/2}\sigma_a B_a(t) = \lambda t + \lambda^{1/2} C_a B_a(t) && \text{(for renewals)} \\
L(t) &= S[A(t)] \approx \frac{1}{\mu}\left[\lambda t + \lambda^{3/2}\sigma_a B_a(t)\right] + \sigma_s B_s(\lambda t) && (B - \text{fluctuations}) \\
&= \frac{\lambda}{\mu}\, t + \frac{\lambda^{1/2}}{\mu}\, C_a B_a(t) + \frac{\lambda^{1/2}}{\mu}\, C_s \lambda^{-1/2} B_s(\lambda t) \\
X(t) &= L(t) - t \approx -(1-\rho)t + \frac{\lambda^{1/2}}{\mu}\left[C_a B_a(t) + C_s\, \frac{1}{\sqrt{\lambda}}\, B_s(\lambda t)\right] \\
&\overset{d}{=} -(1-\rho)t + \frac{\lambda^{1/2}}{\mu}(C_a^2 + C_s^2)^{1/2} B(t) \\
&\quad \uparrow \\
&\quad \text{sum of independent } BM \overset{d}{=} BM \text{ ; selfsimilarity (both by characterization)} \\
&= -(1-\rho)t + \sigma B(t), \qquad \text{where} \quad \sigma^2 = \frac{1}{\mu}\,\rho(C_a^2 + C_s^2)
\end{aligned}
$$

Recall: V obtained from X through *reflection*, and reflection is Lipshitz *continuous*.

$V \approx RBM(-(1-\rho), \sigma)$ with stationary distribution $\exp\left(\text{mean } = \frac{\sigma^2}{2(1-\rho)}\right)$.

Hence, $V(\infty) \overset{d}{\approx} \exp\left(\text{mean } = \frac{1}{\mu}\ \frac{\rho}{1-\rho}\ \frac{C_a^2+C_s^2}{2}\right)$

$\uparrow$ v. significant

$\uparrow$ Generalized $P-K$, for EW_q

- Approximation improves as $\rho \uparrow 1$ (heavy traffic)
- $EW \approx \frac{1}{\mu}\left[1 + \underbrace{\frac{\rho}{1-\rho}}_{\text{utilization}} \cdot \underbrace{\frac{C_a^2 + C_s^2}{2}}_{\text{stoch. variability}}\right]$

  cost of *congestion* ≥ 0

  strictly convex, increasing in ρ, C_a, C_s.

Approximating G/G/n

Stability condition: $\rho = \frac{\lambda}{n\mu} < 1\,.$

Kingman's Exponential Law:

$$\frac{W_q}{E(S)} \approx \begin{cases} \exp\left(\text{mean} \;\; = \frac{1}{n} \cdot \frac{1}{1-\rho} \;\cdot\; \frac{C^2(A)+C^2(S)}{2}\right) & , \text{ wp } E_{2,n}, \\ 0 & , \text{ otherwise.} \end{cases}$$

In particular, a popular measure for service-level, used to determine the number-of-servers n, is:

$$P\{W_q > x \cdot E(S)\} \approx E_{2,n} \cdot \exp\left(-x \cdot \frac{2n(1-\rho)}{C^2(A)+C^2(S)}\right), \;\; x > 0.$$

Allen-Cunneen Approximation:

$$E(W_q) \approx E(S) \;\cdot \frac{1}{n} \cdot \frac{E_{2,n}}{1-\rho} \;\cdot\; \frac{C^2(A)+C^2(S)}{2}.$$

or equivalently,

$$E(W_q) \approx E(W_{q,M/M/n}) \;\cdot\; \frac{C^2(A)+C^2(S)}{2}.$$

- Above accurate in **Efficiency-Driven (ED)** systems.

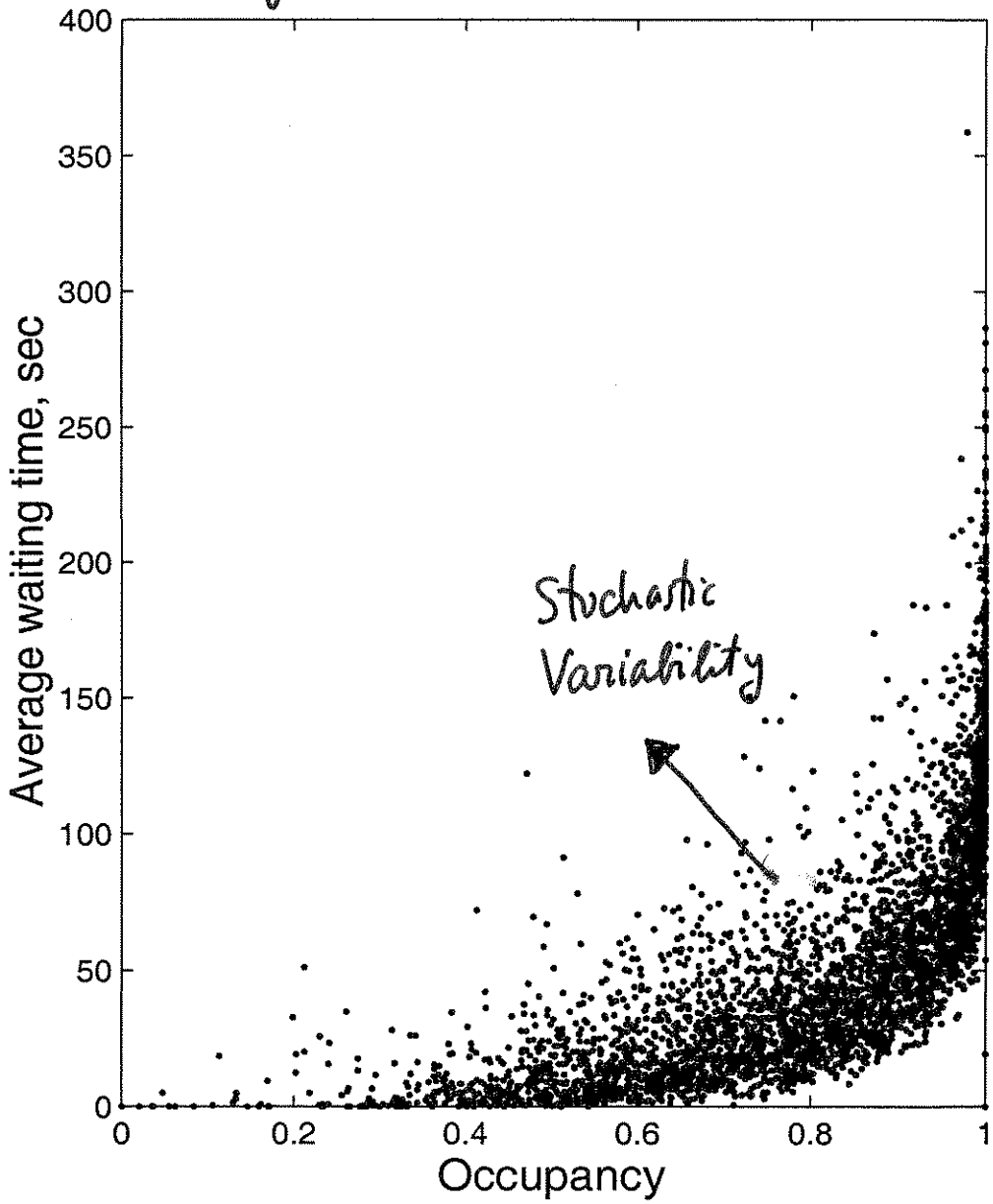
Rules-of-thumb ED-Characterization: In small systems (few servers), over 75% of the customers are delayed in queue prior to service; in large systems (many 10's or several 100's of servers), essentially all customers delayed - more on that in future classes.

The Efficiency - Quality Tradeoff

Congestion Curves
(Empirical Proof of Khinchine-Pollatcheck Formula)

Service Level vs. Availability

-(Service Level) / Quality



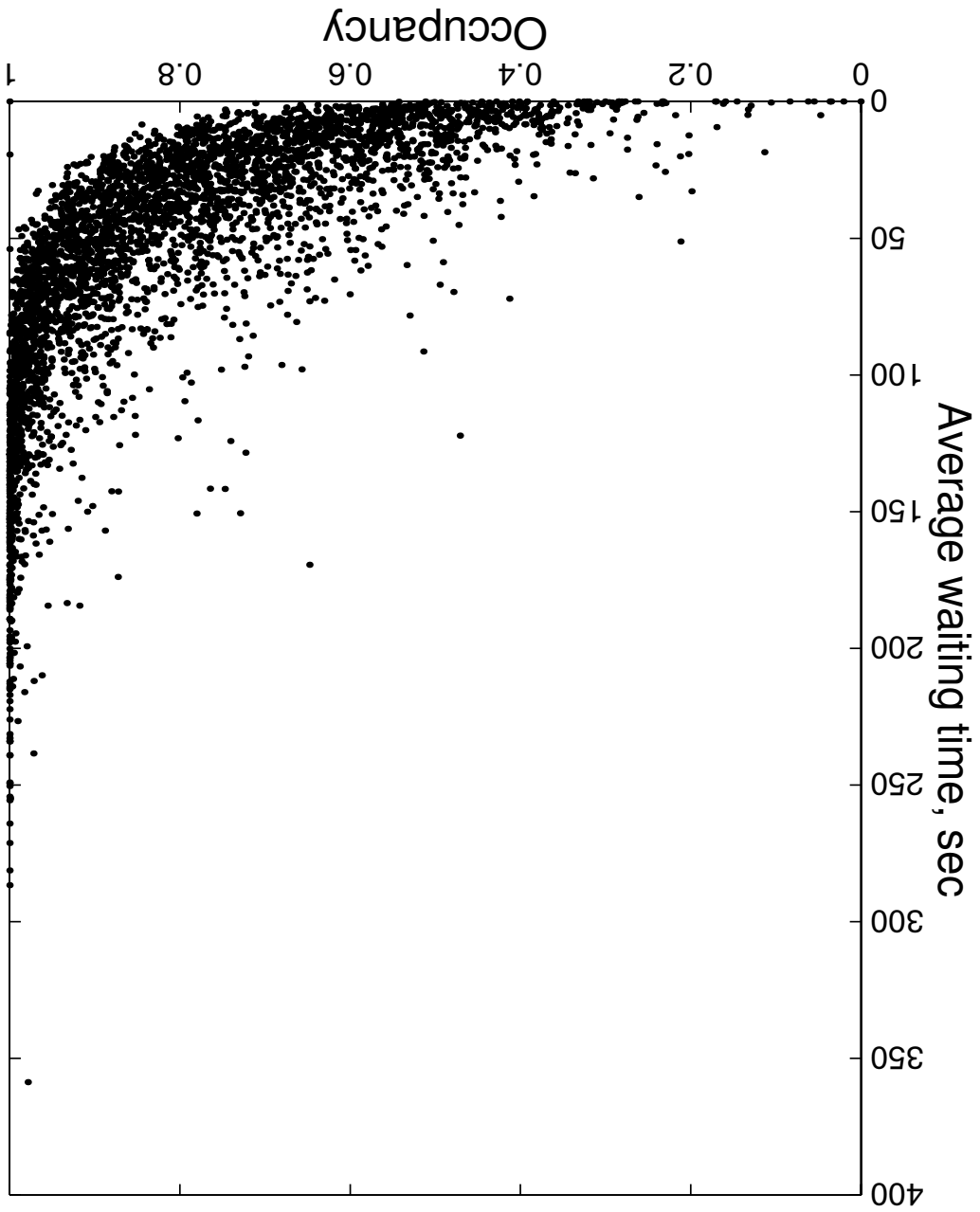
Efficiency

The 2nd Law:

Congestion Index: $\frac{E(W_q)}{E(S)} \approx \frac{1}{N} \frac{\rho}{1-\rho} \frac{C_a^2 + C_s^2}{2}$ (N = number of servers)

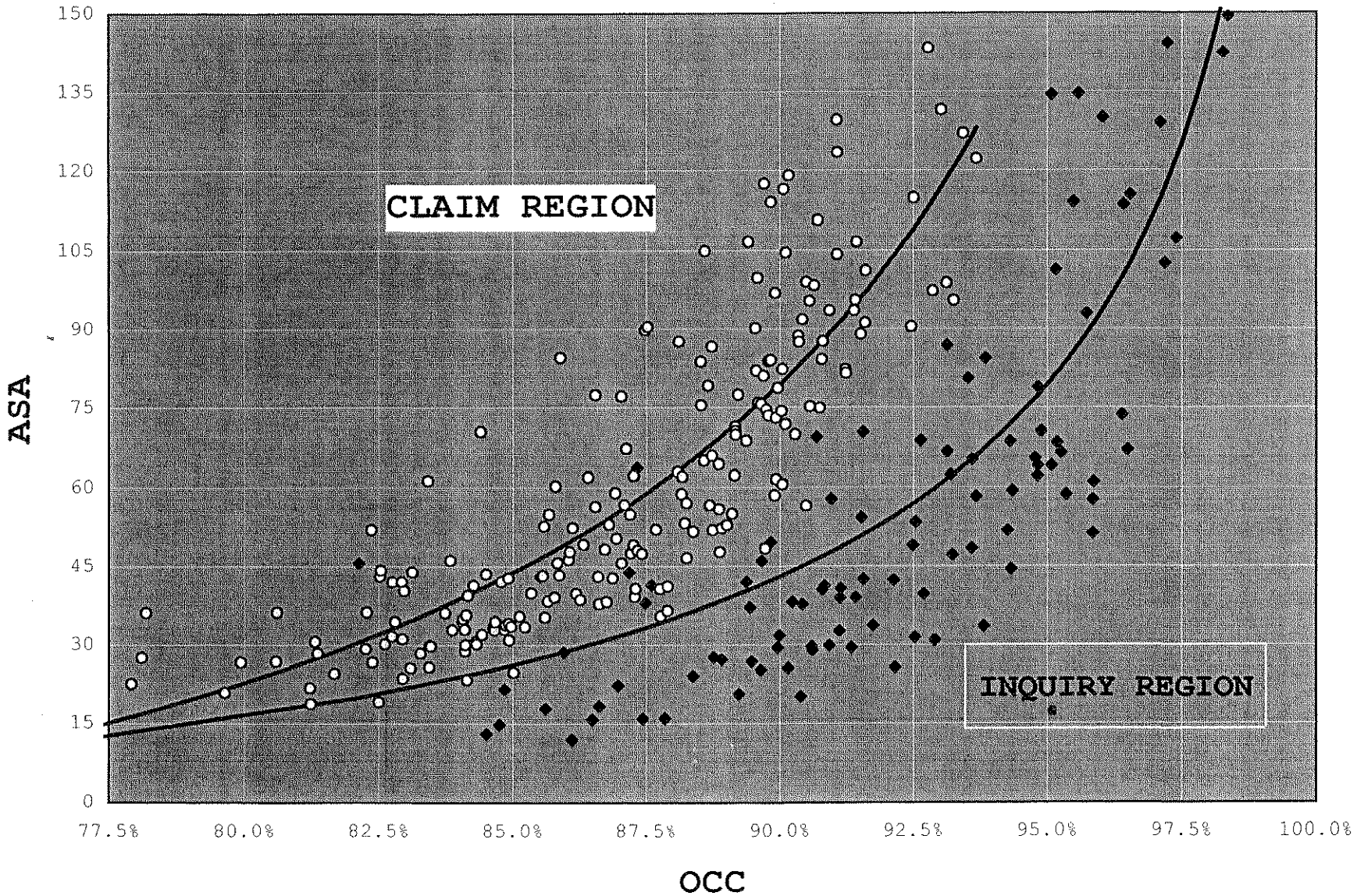
$$= \frac{1}{N} \cdot \frac{\rho}{1-\rho} C^2$$

Performance vs. **Availability** \ **Accessibility**



Queueing Science: Measurements
Model
Validation

K-P / A-C Laws (2 moments; performance averages)



$$\frac{\sigma_q}{s_j} \approx \frac{1}{N} \cdot \frac{p}{1-p} \cdot \bullet \longrightarrow ?$$

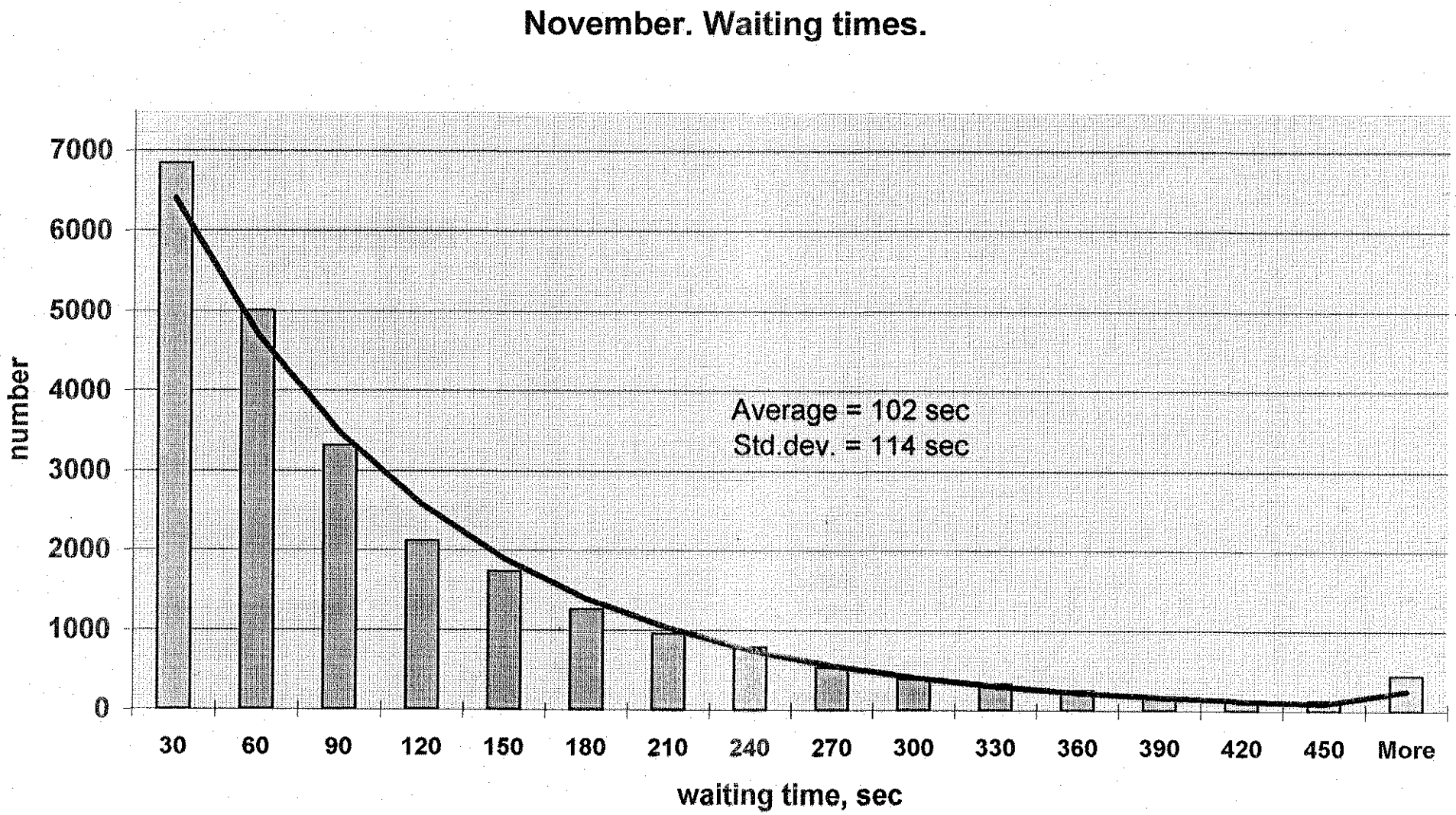
index

efficiency

Kingman's Exponential Invariance Law for the Distribution of Congestion : waitcha

The 3rd Law :

$$P(W_q > T \mid W_q > 0) \approx e^{-T/\overline{W_q}}$$



- $W_q \mid W_q > 0 \sim$ exponential (heavy traffic) ← Kingman, Iglehart-Whitt, ...
- $\exists \eta, \alpha \ni e^{\eta x} P(W_q > x) \to \alpha$, as $x \to \infty$. (Exponential decay) ← Whitt 93, §4.2

M/G/n+G: The Basic Call Center Model

Why fundamental? since, in call centers, and elsewhere,

- **Arrivals** reasonably-approximated by **Poisson**,
- **Services** typically **not Exponential**,
- **(Im)Patience** typically **not Exponential**.

From M/G/n+G to M/M/n+M (Erlang-A):

1. M/**M**/n+G: "Assume" Exponential service times with the same mean (Whitt, 2005, via simulations);
2. M/M/n+**M**: "Assume" Exponential (im)patience times;
3. Estimate the patience-parameter θ via P{Ab}/E[W_q] (with Zeltyn, 2005).

Possible inaccuracies in the exponential approximation for service times, when

- Very large or very small $C(S)$;
- Very patient customers (very small θ).

Theoretical Congestion Curves: Staffing Tools (4CallCenters)

Economies of Scale
Average Waiting Time - But Only of Those Who Wait

$E[W_q|W_q > 0]$ **(Load: 10 per server)**