Class 11

A Single-Server Service-Station in Steady State;
Multi-Server Service-Stations in Steady State;
Laws of Congestion.

A Non-Parametric Model of A Single-Server Service-Station

- Analytical models (vs. Simulation/4CallCenters):

  "Approximate" analysis of Exact models – Today;

  vs. "Exact" analysis of Approximate models – Birth & Death Queues, most notably Erlang-A/C/B (as well as Fluid Models).

- A Non-Parametric Model: the GI/GI/1 Queue.

  Lindley's Equations; Stability.

  Tentative: MOP's; Brummelle's Formula.

  Khinchine-Pollaczek Formula (with an illuminating proof: Hall, pages 168-169).

  Allen-Cunneen Approximation (for averages: (5.69) on page 153 in Hall).

  Kingman's Exponential Law of Congestion.

  Approximations (Framework for).

  Tentative: Priorities: Non-Preemptive, Preemptive.

  Tentative: On Optimal Scheduling: The $c\mu$-rule.]

Models of a Multi-Server Service-Station:
Non-Parametric (GI/GI/m) and Markovian (M/M/m)

- Congestion Curves
- From M/M/m to G/G/m; (Laws of congestion: Kingman, Allen-Cunneen)
- Strategic Queueing Theory
  - Economies of Scale (EOS) Simply Cases, more Subtle Cases, City Bank
  - Efficiency-Driven Service Operations
  - Pooling in a Queueing Network - Part I
    - Pooling Servers(Capacity): One Fast vs. Several Slow
    - Pooling Queues (Geography): Virtual Call Centers
    - Polling Tasks (Services): Job Design (Perhaps Later)
  - Kleinrock's Cycle: Scale-Up (Pooling Queues), then Technological Improvement (Pooling Servers)
- Tentative: Introduction to QED Services Operations

Laws of Congestion

Recitation 12: MJP Models of Service.

Service Engineering

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).

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 ($A_0 = 0$).
  - $E[T_1] = 1/\lambda;\ \ C^2(T_1) = C_a^2$.
  - Note: λ = Arrival rate.
- **Service durations** $S_1, S_2, \ldots, S_n, \ldots$ are iid.
  - $E[S_1] = 1/\mu;\ \ C^2(S_1) = C_s^2$.
  - Note: μ = Service rate.
- Independence between arrivals and services.
- Service discipline is First Come First Served .

1

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 : $\frac{E W_q}{E S} = \frac{L_q}{N \rho}$ ← observable

25/II/92

כך דופקים את האזרח

משרד הרישוי חולון

מחכים שעתיים בשביל 19 שניות

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):

- ρ = 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.

GI/GI/1
The Allen-Cunneen Approximation

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

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

↑ **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:

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

3

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.

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:

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

Allen-Cunneen Approximation:

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

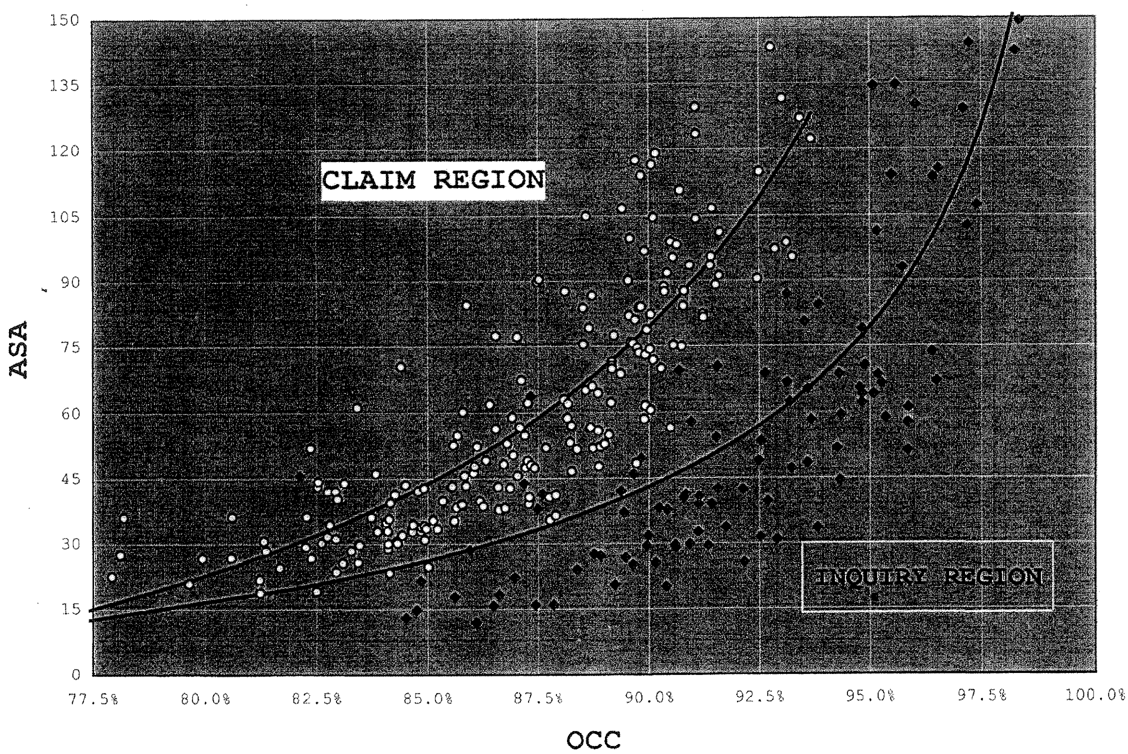
or equivalently,

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

- 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.

4

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



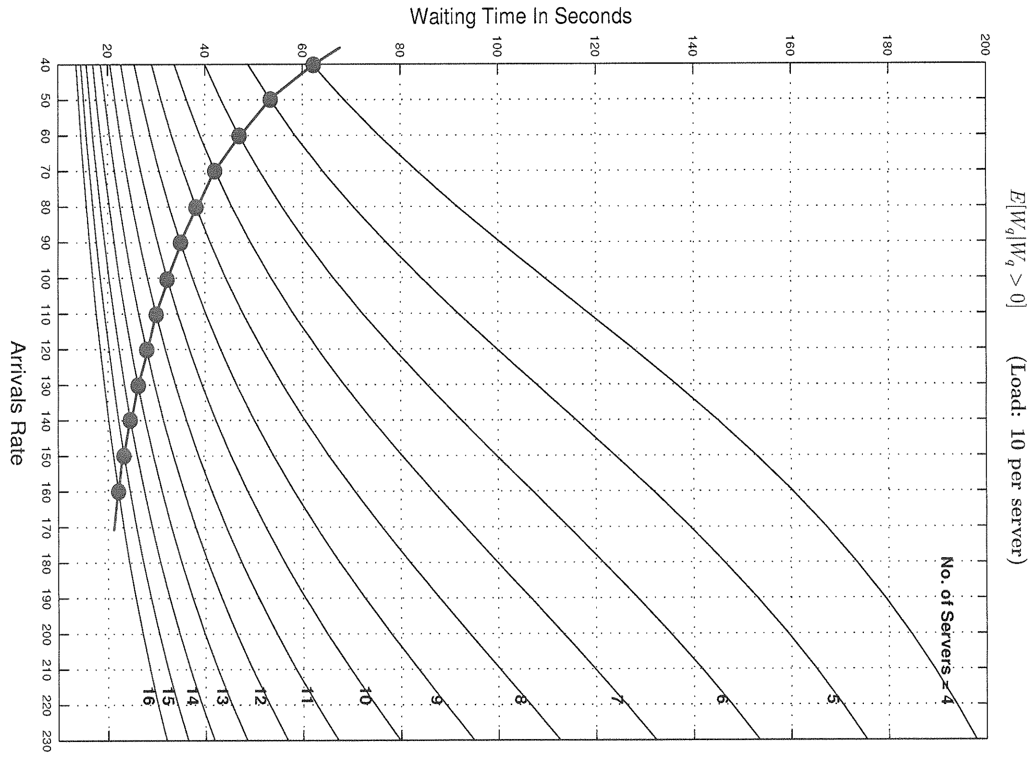
$$\frac{\overline{W_q}}{\overline{S}} \approx \frac{1}{N} \cdot \frac{\rho}{1-\rho} \cdot \bullet$$

index — efficiency — ?

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)



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).

6