Elementary queueing system

- Kendall notation
- Little’s Law
- PASTA theorem
- Basics of M/M/1 queue
- M/M/1 with preemptive-resume priority
- M/M/1 with non-preemptive priority
History of queueing theory

• A long history of research
  – Started in 1909, by Agner Erlang (to model the Copenhagen telephone exchange in Denmark)
  – Booming after 1950’s, to model computer/communication/...
  – David G. Kendall introduced an A/B/C queueing notation in 1953

• A branch of operations research
Model and analysis tool

- Modeling and analyze many systems
  - Telephony systems, exchanger of telephone lines
  - Computer systems
  - Communication systems
  - Transportation systems
  - Production systems
  - Hospital/banks/...

- Limitation
  - Restrictive, some assumptions
  - Use other alternative methods, simulation/tools,...
A general queueing system

• A general queueing system

• A generic model for
  – Machine, computer system, communication system, intersection of roads, etc.
The basic elements of queueing system

• Arrival process of customers
  – Interarrival time i.i.d., e.g., Poisson arrival
  – Arrival one by one, or in batches, etc.

• Behavior of customers
  – Patient or impatient; different type of customers

• Service time
  – i.i.d., e.g., exponential; load/state dependent;
The basic elements of queueing system (cont.)

• Service discipline
  – FCFS: first come first serve, LCFS: last come first serve (stack), RS: randomly serve, priority, SRPT: shortest remaining processing time first, PS: processor sharing

• Service capacity (number of servers)
  – Single server or a group of servers

• Waiting room (system capacity)
  – Finite or infinite; buffer size design
Kendall notation

- A family of notation symbols for different categories of queues
- Proposed by David G. Kendall in 1953
  - Professor of Oxford Univ. (1946–1962) and Cambridge Univ. (1962–1985)

David G. Kendall
1918-2007, England
Kendall notation (cont.)

• A family of notation symbols for different categories of queues

• A/B/C/K/N/D
  – A: distribution of interarrival time, M or G or D
  – B: distribution of service time, M or G or D
  – C: number of servers
  – K: system capacity, infinite by default
  – N: number of total customers, infinite by default
  – D: the service discipline
  – For default, $K = \infty$, $N = \infty$ and $D = \text{FCFS}$
Example of Kendall notation

- $M/M/1; M/G/1; G/M/1; G/G/1; M/D/1; M/Er/1; M/PH/1; MAP/M/1$
- $M/M/c; M/M/1/B; M/M/\infty//N$
- $M/M/c/K/N/$
- $M/M/c/K/\!//LCFS$
- $M/M/1///PS$
A joke

Different of type of queues

Quote from Internet
Utilization factor

• For a queue with single server
  – $\lambda$: the arrival rate of customers
  – $\bar{x}$: the mean service time

• **Utilization factor (or traffic intensity):** $\rho := \frac{\lambda}{\bar{x}}$

• Physical meaning:
  – Single server: time fraction that server is busy
  – Multiple server: fraction of busy servers $\rho = \frac{\lambda}{\bar{x}} / m$

• **Offered load** is defined as $r := \frac{\lambda}{\bar{x}}$ for single or multiple servers
Performance measures

• Distribution of performance measures
  – Distribution of waiting time $W$ and sojourn time $T$ of customers; $E\{W\}$, $\text{Var}\{W\}$, $\text{Pr}\{W>t\}$
  – Distribution of number of customers in the queue $N_q$ or in the system $N$; $E\{N\}$, $\text{Var}\{N\}$, $\text{Pr}\{N>n\}$
  – Distribution of busy period of the server $BP$; $E\{BP\}$, $\text{Var}\{BP\}$, $\text{Pr}\{BP>t\}$

• Mean of performance measures
  – Mean waiting time, mean sojourn time/mean response time;
  – Average number of customers, average queue length
  – Average length of busy period
  – Throughput of the system
    • Open system: equals the arrival rate;
    • Closed system: needs calculation and analysis
The Little’s Law

- For any stable queueing system
  - \( L \): average number of customers in the system
  - \( \lambda \): arrival rate, average number of arrivals per unit of time
  - \( T \): mean system time/response time/sojourn time of customers

\[
L = \lambda T
\]

The first PhD in OR in USA, supervised by Philip M. Morse who is considered to be the father of operations research in the U.S.

His son, John N. Little, founder of Mathworks

Ronald Howard, Morse’s student, main contributor of policy iteration in MDP
The Little’s Law (cont.)

• Apply it to the queue (excluding the server)
  – $L_q$: average number of customers in the queue
  – $\lambda$: arrival rate, average number of arrivals per unit of time
  – $W$: mean waiting time of customers

\[ L_q = \lambda W \]
The Little’s Law (cont.)

• Apply it to the server only
  – \( \rho \): average number of customers in the server,
  – \( \lambda \): average number of arrivals to the server per unit of time
  – \( \bar{x} \): mean service time of customers

\[
\rho = \lambda \bar{x}
\]

• For M/M/1, we have

\[
\rho = \frac{\lambda}{\mu}
\]
The Little’s Law (cont.)

• Applicability
  – Very general, G/G/c
  – Applicable to any queue which is STABLE
  – Applicable to any subsystem of the queue

• Limitation
  – Inapplicable to unstable queue
  – Only for mean metrics, inapplicable to study the transient metrics or distributions
PASTA Theorem

• Poisson Arrivals See Time Averages
• For queues with Poisson arrivals, M/./.:
  – The arriving customers find the same mean measures as that observed by an outside observer at an arbitrary point of time
  – Intuitively explained by the fact that Poisson arrivals occurs completely random in time (purely random sampling)
    • Analog: sampling theorem in signal processing?
PASTA Theorem (cont.)

• Applicable to any queues with Poisson arrival
  • M/M/1, M/G/1, etc.

• Not valid for some queues, e.g., D/D/1
  • Empty at time 0, arrive at 1, 3, 5, ..., service time is 1.
  • Arrivals see empty queue, while average number of customers is 1/2

• Little’s Law and PASTA theorem is very fundamental and important in queueing theory
  – E.g., Mean Value Analysis (MVA) for queueing networks
    • Calculate the mean performance metrics, L, W, T, etc.
M/M/1 queue

• System parameters
  – Poisson arrival rate $\lambda$, service rate $\mu$
  – Infinite capacity, FCFS

• Study the performance metrics
Dynamics of M/M/1 queue

Li Xia, Tsinghua Univ.
State transition rate diagram of M/M/1

- State transition rate diagram
  - state: the number of customers in the system

- A simple birth-death process

Li Xia, Tsinghua Univ.
Equilibrium behavior of M/M/1

• When the system reaches steady
  – Global balance equation
    \[ \pi_0 \lambda = \pi_1 \mu \]
    \[ \pi_n (\lambda + \mu) = \pi_{n+1} \mu + \pi_{n-1} \lambda, \quad n = 1, 2, \ldots \]
  – Local balance equation
    \[ \pi_n \lambda = \pi_{n+1} \mu, \quad n = 0, 1, 2, \ldots \]
  – Normalization equation
    \[ \sum_{n=0}^{\infty} \pi_n = 1 \]
Equilibrium behavior of M/M/1 (cont.)

• Chalk writing
  – Derive the steady state distribution
    \[ \pi_n = \frac{1}{G} \rho^n = (1 - \rho) \rho^n \]

• G is called the normalization constant, \( G = 1/(1-\rho) \)
  – 3 more ways to solve the local balance equation
    • recursion, Z-transform, direct approach of solving difference equations
  – Very simple to solve the local balance equation
Key performance metrics of M/M/1 (time average metrics)

• Average number of customers in the system, $L$
  - We have
  - Variance is $\rho/(1-\rho)^2$
  - $\Pr\{n \geq k\} = \sum_{i=k}^{\infty} \pi_i = \rho^k$

• Average queue length (excluding the customer being served)
  - We have

• Chalk writing
  - The curve of $L$ w.r.t. $\rho$

\[
L = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}
\]
\[
L_q = \sum_{n=1}^{\infty} (n-1)\pi_n = L - \rho = \frac{\rho^2}{1-\rho}
\]

Trick: use dev. to avoid integration by parts!
Key performance metrics of M/M/1 (customer average metrics)

• Mean response time of customers $T$
  
  – By Little’s law, we have $T = L / \lambda = \frac{1}{\mu - \lambda}$

• Mean waiting time of customers (excluding the service time)
  
  – We have $W = T - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu - \lambda} = \rho T$

• Chalk writing
  
  – The curve of $T$, $W$ w.r.t. $\rho$
  
  – Example, double $\lambda$ and $\mu$, how are $L, T, W$ changing?
Another way to calculate mean performance metrics (MVA)

• Mean Value Analysis (MVA)
  – Combine the Little’s law and PASTA theorem
  – No need to know the distribution of steady state

• Analysis process
  – Seen by an arriving customer, mean response time is (should be equivalent to T by PASTA)
    • $T = L/\mu + 1/\mu$  Called arrival relation
  – Little’s law: $L = \lambda T$
  – Combine to obtain: $T = 1/(\mu - \lambda)$, $L = \rho/(1 - \rho)$, etc.
Distribution of sojourn time

• We focus on an arriving customer
  – \( S \): the sojourn time of the arriving customer
  – \( L^a \): # of customers in the system seen by the arrival
  – \( B_k \): service time of the \( k \)th customer, \( k=1,..., L^a \)
  – We have
    \[
    S = \sum_{k=1}^{L^a+1} B_k
    \]
  • Since \( B_k \) and \( L^a \) are independent, we further have
    \[
    P(S > t) = P \left( \sum_{k=1}^{L^a+1} B_k > t \right) = \sum_{n=0}^{\infty} P \left( \sum_{k=1}^{n+1} B_k > t \right) P(L^a = n)
    \]
Distribution of sojourn time (cont.)

- By PASTA theorem,  \( P(L^a = n) = \pi_n = (1 - \rho) \rho^n \)
- \( P\left(\sum_{k=1}^{n+1} B_k > t\right) \) is a \( n+1 \) stage Erlang distribution
- So,

\[
P(S > t) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{(\mu t)^k}{k!} e^{-\mu t} (1 - \rho) \rho^n
\]
\[
= \sum_{k=0}^{\infty} \sum_{n=k}^{\infty} \frac{(\mu t)^k}{k!} e^{-\mu t} (1 - \rho) \rho^n
\]
\[
= \sum_{k=0}^{\infty} \frac{(\mu \rho t)^k}{k!} e^{-\mu t}
\]
\[
= e^{-\mu (1 - \rho) t}, \quad t \geq 0.
\]

exponential distribution with parameter \( \mu(1-\rho)=(\mu-\lambda) \)

Mean sojourn time \( T = 1/(\mu-\lambda) \)
Distribution of sojourn time (cont.)

- The sojourn time is an exponential distribution with parameter \((\mu - \lambda)\). Interesting!

- Comparison 1
  - Weighted sum of Erlang distribution could be exp. distr.
  - Weighted sum of exponential distr. could be ??

- Comparison 2
  - Sum of K number of exponential distributions
    Erlang-K distribution
  - Sum of random number of exponential distributions could be exponential distribution
Another easy way is using Laplace transform

\[ \tilde{S}(s) = E(e^{-sS}) \]
\[ = \sum_{n=0}^{\infty} P(L^\alpha = n) E(e^{-s(B_1+...+B_{n+1})}) \]
\[ = \sum_{n=0}^{\infty} (1 - \rho)\rho^n E(e^{-sB_1}) \cdots E(e^{-sB_{n+1}}). \]

Since \( E(e^{-sB_k}) = \frac{\mu}{\mu + s} \), we have

\[ \tilde{S}(s) = \sum_{n=0}^{\infty} (1 - \rho)\rho^n \left( \frac{\mu}{\mu + s} \right)^{n+1} = \frac{\mu(1 - \rho)}{\mu(1 - \rho) + s}, \]

- The above is an exponential distribution with parameter \( \mu(1-\rho) = (\mu - \lambda) \).

\[ P(S > t) = e^{-\mu(1-\rho)t} \]
Distribution of waiting time, $W$

• Since $S = W + B$, so 
  $$\tilde{S}(s) = \tilde{W}(s) \cdot \tilde{B}(s) = \tilde{W}(s) \cdot \frac{\mu}{\mu + s}. $$

• We have 
  $$\tilde{W}(s) = \frac{(1 - \rho)(\mu + s)}{\mu(1 - \rho) + s} = (1 - \rho) \cdot 1 + \rho \cdot \frac{\mu(1 - \rho)}{\mu(1 - \rho) + s}. $$

  – $W=0$ with probability $1 - \rho$;
  – $W$ is exponentially distributed with parameter $\mu(1 - \rho)$ with probability $\rho$.

Very Special! one part is pmf, the other part is pdf
• IP(idle period) is exponentially distributed with parameter $\lambda$

• Mean busy period:
  
  \[ E(BP) = \frac{1}{\mu - \lambda} \]

  So, it equals $T$.

Li Xia, Tsinghua Univ.
Distribution of busy period

• It is complicated
  – use Laplace-transform and recursion analysis
  – Omitted for simplicity

• pdf of BP:
  
  \[ f_{BP}(t) = \frac{1}{t\sqrt{\rho}} e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu}), \quad t > 0, \]

  – Where \( I_1(.) \) is the modified Bessel function of the first kind of order 1,

  \[ I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!(k+1)!}. \]