Queueing theory primer

Lecturer: Massimo Tornatore

Original material prepared by:
Professor James S. Meditch

Typesetter: Dr. Anpeng Huang

Courtesy of: Prof. Biswanath Mukherjee
A change of focus

• So far we have investigated «static» problems
  – Traffic requests are given and constant in time
    • E.g., Multi commodity flow problem
    • In general, mathematical programming, optimization and graph theory, heuristics…

• Now we move to a class of dynamic problems
  – Random or stochastic flow problems
  – The times at which the demands arrive are uncertain and also the size of the demands are unpredictable
    • Queueing (in our case «traffic») theory
Source

• Notes taken mainly from
  – L. Kleinrock, Queueing Systems (Vol 1: Theory)
    • Chapter 1 and 2
  – L. Kleinrock, Queueing Systems (Vol. 2: Computer Applications)
    • Chapter 1
Delay and Congestion, why?

If $R > C$, we expect congestion (intuitive)
If $R < C$, there might still be congestion (why?)

- The reason for this behaviour is the irregularity (i.e., statistical distribution) of:
  - Arrivals (i.e., interarrival times)
  - Services (i.e., service times)
A. Notation and terminology
$c_n = \text{customer } n \quad \tau_n = \text{arrival time of } c_n$

t_n = \tau_n - \tau_{n-1} = \text{interarrival time} \rightarrow \tilde{t}

w_n = \text{waiting time for } c_n \rightarrow \tilde{W}

x_n = \text{service time for } c_n \rightarrow \tilde{x}

s_n = \text{system time for } c_n \rightarrow \tilde{s}

S_n = w_n + x_n

*Note that these distributions do not depend by $n$ (same distribution for all arrivals/services)*
Arrival process

CDF
\[ A(t) \approx P[\tilde{t} \leq t] \]

pdf
\[ a(t) = \frac{dA(t)}{dt} \]

mean
\[ E[\tilde{t}] = \bar{t} = \frac{1}{\lambda} \]

\[ E[(\tilde{t})^k] = \bar{t}^k \]

Service process

CDF
\[ B(x) \approx P[\tilde{x} \leq x] \]

pdf
\[ b(x) = \frac{dB(x)}{dx} \]

mean
\[ E[\tilde{x}] = \bar{x} = \frac{1}{\mu} \]

\[ E[(\tilde{x})^k] = \bar{x}^k \]
Laplace transform/moment generating function

\[ A^*(s) = \int_0^\infty a(t) e^{-st} \, dt = E[e^{-st}] \]

\[ B^*(s) = E[e^{-s\bar{x}}] \]

\[ A^{*(k)}(0) \approx \left. \frac{d^k A^*(s)}{ds^k} \right|_{s=0} = (-1)^k \bar{t}^k \]
Queueing system performance

Input variables: system defined by $\tilde{t}$ and $\tilde{x}$

Output variables: performance defined by
1. $N(t)$ no. of customers in system at time $t$
2. $w_n$ waiting time
3. $s_n$ system time

HP: $t \to \infty$, $\overline{N}(t) \to \overline{N}$ (statistical equilibrium, stationarity)
\[ \begin{align*}
N & \quad W \quad S \\
F_N(k) &= P[N \leq k] \\
P_k &= P[N = k] \\
E[N] &= \bar{N} \\
E[z^N] &= Q(z) = \sum_{k=0}^{\infty} P_k z^k \\
W(y) &= P[\tilde{w} \leq y] \\
w(y) &= \frac{dW(y)}{dy} \\
E[\tilde{w}] &= W \\
E[e^{-s\tilde{w}}] &= W^*(s) \\
S(y) &= P[\tilde{s} \leq y] \\
s(y) &= \frac{dS(y)}{dy} \\
E[\tilde{s}] &= T \\
E[e^{-s\tilde{s}}] &= S^*(s)
\end{align*} \]
Other performance variables:

\[ I = \text{idle period} \]
\[ D = \text{interdeparture time} \]
\[ G = \text{busy period} \]
\[ N_q = \text{no. of customers in queue} \]
Kendall’s notation for queueing systems

A/B/m

no. of users

no. of servers

A/B/m/K/M

queue size

no. of users

M exponential (Markovian)

E_r r-stage Erlangian

H_r R-stage hyperexponential

D deterministic

G general
B. General results

1. **Utilization factor**

\[ \rho \approx \frac{\text{avg. rate at which work arrives}}{\text{capacity of the system to do work}} \approx \frac{R}{C} \]

\[ \rho = \text{fraction of system capacity in use (on the avg.)} \]

\[ 0 \leq \rho < 1 \]
G/G/1

- Let us start with no assumptions on arrival and service distribution and one single server
- It can be generally proven that:

\[
\rho = \lambda \bar{x} = \frac{\lambda}{\mu} = \frac{\text{avg. no. arrivals / sec}}{\text{avg. rate of service / sec}}
\]

where \( \bar{x} = \frac{1}{\mu} \)
G/G/m

- In case of multiple (m) servers:

\[ \rho = \frac{\lambda \bar{x}}{m} = \frac{\lambda}{m \mu}, \quad \frac{1}{\mu} = \text{avg. service time for each server} \]

\[ \rho = \text{avg. fraction of busy servers} \]

Stability requires \( 0 \leq \rho < 1 \)

(except for D/D/m where \( 0 \leq \rho \leq 1 \))
2. System time

\[ \tilde{s} = \tilde{x} + \tilde{w} \]

\[ E[\tilde{s}] = E[\tilde{x}] + E[\tilde{w}] \]

\[ T = \bar{x} + W \]
3. Little’s result

\[ \lambda \rightarrow \bar{N}, T \rightarrow \lambda \]

\[ \bar{N} = \lambda T \]

\[ \bar{N}_q = \lambda W \]

The average number of customers in a queueing system is equal to the average arrival time of customers to that system, times the average time spent in that system.
Figure 2.3 Arrivals and departures.

\[ \alpha(t) = \text{no. arrivals in } (0, t) \]
\[ \delta(t) = \text{no. departures in } (0, t) \]
\[ N(t) = \alpha(t) - \delta(t) \geq 0 \]

\[ \gamma(t) \equiv \int_0^t N(s)ds \quad \text{(customer - sec) total time all customers} \]
\text{have spent in the system up to time } t \text{ (i.e, the grey area!)}

\[ T_t \equiv \text{system time per customer averaged over all customers} \]
\text{who arrived during } (0, t)

\[ \overline{N}_t \equiv \text{avg. no. of customers in the system during the interval } (0, t) \]
\[ \lambda(t) \approx \frac{\alpha(t)}{t} \left[ \frac{\text{customers}}{\text{sec}} \right] \quad \text{Avg. arrival rate in} \ (0, t) \]

\[ T_t = \frac{\gamma(t)}{\alpha(t)} \left[ \frac{\text{sec}}{\text{customers}} \right] \quad \text{Avg. systemtime per customer in} \ (0, t) \]

\[ \bar{N}_t \equiv \frac{\gamma(t)}{t} \quad \text{Avg. no. customers in system in} \ (0, t) \]

But we can \( \bar{N}_t = \frac{\gamma(t)}{t} \times \frac{\alpha(t)}{\alpha(t)} \)

\[ \Rightarrow \bar{N}_t = \lambda_t T_t \]
If $\lambda = \lim_{t \to \infty} \lambda_t$ and $T = \lim_{t \to \infty} T_t$ exist, then $\overline{N} = \lambda T$ (q.e.d) Similarly, it can be proven that:

\[
\begin{align*}
\overline{N}_q &= \lambda W \\
\overline{N}_s &= \lambda \bar{x}(= \rho) \\
\overline{N} &= \overline{N}_s + \overline{N}_q
\end{align*}
\]

and we also get $T = \bar{x} + W$

$G / G / m$: $\overline{N}_q = \overline{N} - m \rho$, \quad \rho = \frac{\lambda}{m \mu}$
C. Poisson process

Interarrival times: \( t_i \) \{ independent
exponentially distributed,\}

\[
P[\tilde{t} \leq t] = 1 - e^{-\lambda t}, \quad t \geq 0, \quad \tilde{t} = \frac{1}{\lambda}
\]

\[
P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, \ldots
\]

\[
= \Pr \left[ k \text{ arrivals in } (0,t) \right]
\]
1. Derivation of the Poisson process

\[
\text{Pr}[1 \text{ arrival in } \Delta t] = \lambda \Delta t + o(\Delta t)
\]

\[
\text{Pr}[0 \text{ arrival in } \Delta t] = 1 - \text{Pr}[1 \text{ arrival in } \Delta t]
\]

\[
= 1 - \lambda \Delta t + o(\Delta t)
\]

\[
\lambda \approx \text{process intensity (arrival rate)}
\]

\[
P_k(t + \Delta t) = P_k(t)[1 - \lambda \Delta t] + P_{k-1}(t)\lambda \Delta t + o(\Delta t)
\]

\[
P_k(t + \Delta t) - P_k(t) = -\lambda P_k(t) \times \Delta t + \lambda P_{k-1}(t) \Delta t + o(\Delta t)
\]
If I divide by «$dt$», I obtain:

$$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t) \quad k = 1, 2, \ldots$$

Similarly for $P_0$:

$$P_0(t + \Delta t) = P_0(t)[1 - \lambda \Delta t] + o(\Delta t)$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \quad \text{note that } P_0(0) = 1$$

Solving the first-order differential equation (2):

$$P_0(t) = e^{-\lambda t}$$

then inserting $P_0(t)$ in (1) for $k=1$:

$$P_1(t) = \lambda te^{-\lambda t}$$

then continuing by induction:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k \geq 0, \quad t \geq 0$$

*Final note (relation with exp. neg. distribution!!) $P_0(t) = e^{-\lambda t} = 1 - P[\bar{t} \leq t]$
2. Properties of the Poisson Process

i. \( \bar{N}(t) = \sum_{k=0}^{\infty} kP_k(t) = \lambda t \)
\( \lambda = \text{avg. arrival rate} \)

ii. \( \sigma_{N(t)}^2 = \lambda t \)

iii. \( Q(z,t) = \sum_{k=0}^{\infty} P_k(t)z^k = E[ z^{N(t)} ] = e^{\lambda t(z-1)} \)
\[ \frac{dQ(z,t)}{dz} \bigg|_{z=1} = \lambda t \times e^{\lambda t(z-1)} \bigg|_{z=1} = \lambda t \]

iv. \( A(t) = P[\bar{t} \leq t] = 1 - e^{-\lambda t} \)
\( a(t) = \lambda e^{-\lambda t} \)

Interarrival times are exponentially distributed (memory less!)
\( \bar{t} = \frac{1}{\lambda} \quad \sigma^2 = \frac{1}{\lambda^2} \)
Random process

• Poisson Process is a stochastic (or random) process
• Now we will use some more advanced stochastic processes (Markov chains)
• But… what is a stochastic process?
  – It is «family» of random variables $X(t)$ indexed by time $t$
  – A possible (intuitive) case is when the random process represents the sum of simple random variables at instant $t$
Example of a random process

- $S_1(t), S_2(t), S_3(t), \ldots$ are random variables
- $X(t)$ is a random process
- $X(t, \omega) =$ Number of servers busy at time $t$ of realization $\omega$ of the process = one realization of the process

Assumptions
- Stationarity
  - $E_{[t_0, t_0 + \tau]}[X(t, \omega)] = A_\tau(t_0, \omega) = A(\omega)$
- Ergodicity
  - $A(\omega) = A$
D. Markov chains

• A random process is called a “chain” if the state space is discrete (i.e., if $X(t)$ can assume only discrete values)
  – Note: $S = \{0, 1, \ldots, K\}$ finite state
  $S = \{0, 1, \ldots\}$ infinite state (countable)

• A chain can be
  – Discrete-time
    • If $t$ can assume only discrete values, i.e., if $X(t)$ can change values on at discrete instants in time
  – Continuous-time
    • If $t$ can assume any continuous value

• A chain is a Markov chain if … see next slide
D.1. Discrete-time Markov chains

1) It’s a chain since: \( X_n \in S, S = \{0, 1, ..., N\}, N \leq \infty \)

2) It’s a discrete-time chain since the time («x-axis») is slotted

3) It’s a Markov Chain since:

\[
P[X_n = j|X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0] = P[X_n = j|X_{n-1} = i_{n-1}]
\]
State probabilities

Let us define $\pi_i^{(n)} = P[X_n = i]$ \quad $\pi^{(n)} = [\pi_0^{(n)} \pi_1^{(n)} \cdots ]$

$$\sum_i \pi_i^{(n)} = 1$$

(One-step) transition probabilities

Let us define $P_{ij}^{(n-1)} = P[X_n = j|X_{n-1} = i]$ \quad $P^{(n-1)} = [P_{ij}^{(n-1)}]$ \quad $\sum_j P_{ij}^{(n)} = 1 \ \forall \ i, n$
State equations

\[ \pi^{(n)} = \pi^{(n-1)} P^{(n-1)} \]

\[ n = 1, 2, \ldots \quad \pi^{(0)} \text{ given} \]

Homogeneous chain

\[ P_{ij}^{(n)} = P_{ij}, \text{i.e., independent of time } n \]

\[ P_{ij} = P[X_n = j | X_{n-1} = i] \quad \forall n \]

\[ P = [P_{ij}] \quad \sum_j P_{ij} = 1 \text{ for each } i \]
\[ \pi^{(n)} = \pi^{(n-1)} P \quad n = 1, 2, \ldots \]
\[ \pi^{(n)} = \pi^{(0)} P^n \]
\[ \pi^{(0)} \text{ given} \quad \sum_i \pi_i^{(n)} = 1 \]

If \( \lim_{n \to \infty} \pi^{(n)} = \pi \) exists, then

\[ \pi = \pi P \rightarrow \text{equilibrium probabilities ("steady state")} \]
\[ \sum_i \pi_i = 1 \]
Proof of $\pi^{(n)} = \pi^{(n-1)} P$
Proof- Contd.

\[
\pi_i^{(n)} = \pi_0^{(n-1)} P_{0i} + \pi_1^{(n-1)} P_{1i} + \pi_2^{(n-1)} P_{2i} + \\
+ \ldots + \pi_s^{(n-1)} P_{si} + \ldots = \\
\]

\[
= [\pi_0^{(n-1)} \pi_1^{(n-1)} \pi_2^{(n-1)} \ldots \pi_s^{(n-1)} \ldots ] \begin{bmatrix}
P_{0i} \\
P_{1i} \\
P_{2i} \\
\vdots \\
P_{si} \\
\vdots 
\end{bmatrix}
\]

\[i-th\ column\ of\ P \rightarrow\]
Example

Discrete-time birth-death (arrival-departure) process with no waiting room.

\[ P(1 \text{ arrival at any time } n) = a \]
\[ P(1 \text{ departure at any time } n) = d \]
\[ s = \{0,1\} \quad \pi = [\pi_0 \pi_1] \]
Example – Contd.

\[ P = \begin{bmatrix} 1-a & a \\ d & 1-d \end{bmatrix} \]

\[ \pi^{(n)} = \pi^{(n-1)} P \]

\[ n \to \infty \quad \Rightarrow \quad \pi = \pi \ P \]

\[ \pi_0 = (1-a) \pi_0 + d\pi_1 \]

\[ \pi_1 = a \pi_0 + (1-d)\pi_1 \]
Example – Contd.

\[ d \pi_1 = a \pi_0 \quad \rightarrow \quad \pi_1 = \frac{a}{d} \pi_0 \]

\[ \pi_0 + \pi_1 = 1 \quad \rightarrow \quad (1 + \frac{a}{d})\pi_0 = 1 \]

\[ \pi_0 = \frac{d}{a + d} \quad \pi_1 = \frac{a}{a + d} \]
D.2. Continuous-time Markov chains

1) It’s a chain since: \( X(t) \in S, S = \{0, 1, ..., N\}, N \leq \infty \)

2) It’s a continuous-time chain since the times \( t \) can assume any real value

3) It’s a Markov Chain since:

**Def.** For all \( n \) and any \( 0 \leq t_1 < t_2 < ... < t_n < t_{n+1} \)

\[
P[X(t_{n+1}) = j \mid X(t_n) = i_n, ..., X(t_1) = i_1] = P[X(t_{n+1}) = j \mid X(t_n) = i_n]
\]
Time spent in a state

Property of continuous-time Markov Chain

\[ \tau = \text{time in state } E_i \in S \text{ is a r.v. with} \]

\[ P[\tau_i \leq t] = 1 - e^{-\mu_i t} \text{ (negative exp.)} \]

Note: in a discrete-time Markov Chain time spent in a state is a geometric r.v.
Neg. Exp. Is «Memoryless»

\[
\Pr\{T \leq t + t_0 \mid T > t_0\} = \frac{\Pr\{t_0 < T \leq t + t_0\}}{\Pr\{T > t_0\}}
\]

\[
= \frac{\Pr\{T \leq t + t_0\} - \Pr\{T \leq t_0\}}{\Pr\{T > t_0\}}
\]

\[
= \frac{1 - e^{-\lambda(t + t_0)} - \left(1 - e^{-\lambda t_0}\right)}{1 - \left(1 - e^{-\lambda t_0}\right)} = 1 - e^{-\lambda t}
\]

\[
= \Pr\{T \leq t\}
\]

\[
\Pr\{T > t + t_0 \mid T > t_0\} = \Pr\{T > t\}
\]
D.2.a. General Case

D.2.a.1. Transition probabilities

\[ p_{ij}(s,t) = P[X(t) = j \mid X(s) = i] \quad t \geq s \]

\[ P(t) = [p_{ij}(t, t + \Delta t)] \]

\[ Q(t) = \lim_{\Delta t \to 0} \frac{P(t) - I}{\Delta t} = \text{transition rate matrix} \]

\[ q_{ii}(t) = \lim_{\Delta t \to 0} \frac{p_{ii}(t, t + \Delta t) - 1}{\Delta t} \]

\[ q_{ij}(t) = \lim_{\Delta t \to 0} \frac{p_{ij}(t, t + \Delta t)}{\Delta t} \quad i \neq j \]

\[ \sum_j q_{ij}(t) = 0 \quad \text{for each } i \]

If these limits do not exist, we do not have a continuous-time Markov chain.
D.2.a.2 State Probabilities

\[ \pi_j(t) = P[X(t) = j] \quad j \in S \]

\[ \pi(t) = [\pi_0(t) \quad \pi_1(t) \quad \ldots \quad \pi_N(t)], \quad N \leq \infty \]

\[
\frac{d\pi(t)}{dt} = \pi(t) Q(t) \\
\pi(0) \text{ given } \sum_j \pi_j(t) = 1
\]

\[ \pi(t) = \pi(0) \exp\left[\int_0^t Q(s)ds\right] \]
D.2.b. Homogenous case

\[ p_{ij}(t) = p_{ij}(s, s + t) \quad \textit{indep. of } s \]
\[ q_{ij}(t) = q_{ij} = \text{const.} \]
\[ Q = [q_{ij}] \]

\[ \frac{d\pi(t)}{dt} = \pi(t) Q \rightarrow \pi(t) = \pi(0) e^{Qt} \]
Homogenous case – Contd.

**Steady – state (if it exists)**

\[
\lim_{{t \to \infty}} \pi_j(t) = \pi_j
\]

independent of \( \pi(0) \):

\[
\pi Q = 0 \\
\sum \pi_j = 1
\]
Models used in this course
Birth-death processes

• We want to apply the previously seen random processes to model queues in telecom equipment

• The state changes only
  – When a packet arrives (birth)
  – When a packet leaves (death)

• Basically you can only move between adjacent states
  – State $k \Rightarrow$ State $k+1$ (birth)
  – State $k \Rightarrow$ State $k-1$ (death)

• Continuous time
  – Packet arrive and depart at any time!

• Mathematical treatment follows in the next slides
E. Continuous-time birth-death processes
aka arrival-departure processes

(1) Continuous-time Markov chain dealing with a population of size $N$ at time $t$
\[ P_k(t) = P[N(t) = k] \quad k \in S \]
\[ S = \{0, 1, ..., L\} \quad L \leq \infty \]

(2) System state changes by at most one (up or down, or no change) in $\Delta t$
\[ \lambda_k dt = p_{k-1,k} \]
\[ \mu_k dt = p_{k+1,k} \]
\[ p_{k,j} = 0 \quad \text{if} \quad |k - j| > 1 \]
(3) Births and deaths independent
   – Follows from Markovianity

(4) Transitions

\[ P[\text{exactly } 1 \text{ b in } (t, t + \Delta t) \mid N(t) = k] = \lambda_k \Delta t + o(\Delta t) \]
\[ P[\text{exactly } 0 \text{ b in } (t, t + \Delta t) \mid N(t) = k] = 1 - \lambda_k \Delta t + o(\Delta t) \]
\[ \lambda_k = \text{birth rate} \]

\[ P[\text{exactly } 1 \text{ d in } (t, t + \Delta t) \mid N(t) = k] = \mu_k \Delta t + o(\Delta t) \]
\[ P[\text{exactly } 0 \text{ d in } (t, t + \Delta t) \mid N(t) = k] = 1 - \mu_k \Delta t + o(\Delta t) \]
\[ \mu_k = \text{death rate} \]
Transitions - State Eqns.

- 3 (set of) equations regulate the birth-death process

1) \[ P_k(t + \Delta t) = P_k(t)p_{k,k}(\Delta t) + P_{k-1}(t)p_{k-1,k}(\Delta t) \]
   \[ + P_{k+1}(t)p_{k+1,k}(\Delta t) \quad k \geq 1 \]

2) \[ P_0(t + \Delta t) = P_0(t)p_{00}(\Delta t) + P_1(t)p_{10}(\Delta t) \quad k = 0 \]

3) \[ \sum_{k=0}^{L} P_k(t) = 1 \quad (L \leq \infty) \]
State Eqns. – Contd.

- Solving of previous equations

\[
P_k(t + \Delta t) = P_k(t)[1 - \lambda_k \Delta t + o(\Delta t)][1 - \mu_k \Delta t + o(\Delta t)] \\
+ P_{k-1}(t)[\lambda_{k-1} \Delta t + o(\Delta t)] + P_{k+1}(t)[\mu_{k+1} \Delta t + o(\Delta t)]
\]

\[
P_0(t + \Delta t) = P_0(t)[1 - \lambda_0 \Delta t + o(\Delta t)] + P_1(t)[\mu_1 \Delta t + o(\Delta t)]
\]

*Similar derivation as seen in the Poisson process*
Solving these equations is quite complex, so, in practice, the approach used is the one shown in the next slide (inspection over a state diagram)
State-Transition Rate Diagrams

Note: a simple «inspection» technique to find the same equations

Notion of probability flow

Flow into $E_k = \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t)$

Flow out of $E_k = (\lambda_k + \mu_k)P_k(t)$

$$\frac{dP_k(t)}{dt} = \text{Flow into } E_k - \text{Flow out of } E_k$$

$$= \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t) - (\lambda_k + \mu_k)P_k(t) \quad k = 0, 1, \ldots$$
if we want flows across boundary balanced (equilibrium)
\[
\frac{dP_k(t)}{dt} = 0 \quad \text{Hence, } P_k(t) = P_k
\]

Note:
\[
\mu_k P_k = \lambda_{k-1} P_{k-1}
\]
\[
P_k = \frac{\lambda_{k-1}}{\mu_k} P_{k-1} \quad k = 1, 2, \ldots
\]
(Generic M/M/1) Queue

System

Representation

\( \lambda_k = \text{arrival rate} \quad \text{jobs/sec} \)

\( \mu_k = \text{service rate} \quad \text{jobs/sec} \)

Equilibrium behavior \( t \to \infty \),

\( \frac{dP_k(t)}{dt} = 0 \quad k = 0, 1, \ldots \)
1. State dependent $\lambda_k$ and $\mu_k$; $P_k(t) \rightarrow p_k$

\[ 0 = -(\lambda_k + \mu_k) p_k + \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1}, \quad k \geq 1 \]

\[ 0 = -\lambda_0 p_0 + \mu_1 p_1 \quad \quad k = 0 \]

\[ \sum_{k=0}^{\infty} p_k = 1 \]

\[ p_1 = \frac{\lambda_0}{\mu_1} p_0, \quad p_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0, \ldots \]

\[ p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \]

\[ p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \]
Classical M/M/1

\[ \lambda_k = \lambda, \, \mu_k = \mu, \, \text{all } k \]

\[ p_k = \left( \frac{\lambda}{\mu} \right)^k \, p_0 \quad k \geq 1 \]

\[ p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k} \]
If \( \frac{\lambda}{\mu} < 1 \),
\[
1 + \sum_{k=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^k = \sum_{k=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^k = \frac{1}{1 - \frac{\lambda}{\mu}}
\]

\[
\rho = \frac{\lambda}{\mu} \quad 0 \leq \rho < 1
\]

\[
p_0 = 1 - \rho \quad \quad p_k = (1 - \rho) \rho^k
\]

\[
p_k = (1 - \rho) \rho^k \quad k \geq 0
\]
\[ \overline{N} = \sum_{k=0}^{\infty} k \ p_k = (1 - \rho) \sum_{k=0}^{\infty} k \ \rho^k \]

\[ \sum_{k=0}^{\infty} k \ \rho^k = \rho \sum_{k=0}^{\infty} k \ \rho^{k-1} = \rho \left( \sum_{k=0}^{\infty} \frac{d}{d\rho} (\rho^k) \right) = \]

\[ = \rho \frac{d}{d\rho} \left( \sum_{k=0}^{\infty} \rho^k \right) = \rho \frac{d}{d\rho} \left( \frac{1}{1 - \rho} \right) = \frac{\rho}{(1 - \rho)^2} \]

\[ \overline{N} = \frac{\rho}{1 - \rho} \]

\[ \sigma_N^2 = \sum_{k=0}^{\infty} (k - \overline{N})^2 \ p_k \]

\[ \sigma_N^2 = \frac{\rho}{(1 - \rho)^2} \]
\[ \bar{N} = \lambda T \]

\[ T = \frac{1}{\mu} \frac{\mu}{1 - \rho} \]

\[ T = \bar{x} + W = \frac{1}{\mu} + W = \frac{1}{\mu} \frac{\rho}{1 - \rho} \]

\[ W = \frac{\rho / \mu}{1 - \rho} = \frac{1}{\mu} \frac{N}{\bar{N}} \]
Contd.

\[ N_q = \lambda W \]

\[ \overline{N}_q = \frac{\rho^2}{1 - \rho} \]

\[ P[N \geq k] = \sum_{i=k}^{\infty} p_i = \sum_{i=k}^{\infty} (1 - \rho) \rho^i \]

\[ P[N \geq k] = \rho^k \]

Above results are valid iff \( 0 \leq \rho < 1 \)
M/M/m Queue Model

\[ \rho = \frac{\lambda}{m\mu} < 1 \]

Equivalently,

\[ \mu_k = \begin{cases} 
  k\mu & 0 \leq k \leq m \\
  m\mu & k \geq m 
\end{cases} \]
Use state-dependent birth-death model to determine $p_k, k \geq 0$

\[ p_1 = \rho_1 p_0 \quad \lambda p_0 = \mu \rho_1 \]
\[ p_2 = \frac{\rho_1^2}{2} p_0 \quad \lambda p_1 = 2 \mu \rho_2 \]
\[ p_3 = \frac{\rho_1^3}{3!} p_0 \quad \lambda p_2 = 3 \mu \rho_3 \]
\[ p_m = \frac{\rho_1^m}{m!} p_0 \]
\[ p_{m+1} = \frac{\rho_1^{m+1}}{m! m} p_0 \]
\[ p_{m+2} = \frac{\rho_1^{m+2}}{m! m^2} p_0 \]

\[ p_0 = \frac{1}{\sum_{k=0}^{m-1} \frac{\rho_1^k}{k!} + \frac{\rho_1^m}{m!} \frac{1}{1 - \rho_1/m}} \]

\[ \rho = \frac{\rho_1}{m} \]

\[ p_k = \begin{cases} 
\frac{(m\rho)^k}{k!} p_0 & 0 \leq k \leq m \\
\frac{m^m \rho^k}{m!} p_0 & k \geq m 
\end{cases} \]
M/M/m Queue system characteristics

\[ p_k = \begin{cases} 
\frac{(mp)^k}{k!} p_0 & 0 \leq k \leq m \\
\frac{m^m \rho^k}{m!} p_0 & k \geq m
\end{cases} \]

\[ p[m] = \frac{(m\rho)^m}{m!(1-\rho)} p_0 \]

Where

\[ p_0 = \frac{1}{\sum_{k=0}^{m-1} \frac{(mp)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{1-\rho}} \]

\[ p[\text{queueing}] = \sum_{k=m}^{\infty} p_k \triangleq p_m \]

\[ \bar{N} = m\rho + \frac{\rho}{1-\rho} p_m \]

\[ \bar{N}_q = \frac{\rho}{1-\rho} p_m \quad \bar{N}_s = m\rho = \frac{\lambda}{\mu} \]

\[ T = \frac{\bar{N}}{\lambda} \quad W = \frac{\bar{N}_q}{\lambda} \quad x = \frac{\bar{N}_s}{\lambda} = \frac{1}{\mu} \]
Backup slides
M/G/1 Queue

Poisson arrival process: \( \lambda \) customers/sec

General service process: \( B(x), \ x, \ x^k \)

Queueing discipline is FCFS

1. Pollaczek-Khinchin (P-K) relations

\[
\bar{x} = E[x] \quad \bar{x}^2 = E[x^2] \quad \rho = \frac{\lambda \bar{x}}{1} < 1
\]

\[
W = \frac{\lambda \bar{x}^2}{2(1-\rho)} \quad T = \bar{x} + W
\]

\[
\bar{N} = \lambda \bar{x} + \lambda W \quad \bar{N} = \rho + \frac{\lambda^2 \bar{x}^2}{2(1-\rho)}
\]
Ex.

\[ \lambda = 0.4 \text{cust/sec} \]

\[ x = 2 \text{sec} \]

\[ x^2 = \frac{1}{2} \int_1^3 x^2 \, dx = \frac{13}{3} \text{sec}^2 \]

\[ \rho = \lambda x = (0.4)(2) = 0.8 \]

\[ W = \frac{\lambda x^2}{2(1 - \rho)} = \frac{(0.4)(13/3)}{2(0.2)} = \frac{13}{3} \]

\[ W = 4.33 \text{sec} \]

\[ \bar{N} = \rho + \lambda W = 0.8 + 0.4(13/3) \]

\[ \bar{N} = 2.53 \]

\[ T = \bar{x} + W = 2 + 4.33 \quad T = 6.33 \text{sec} \]
2. **P-K transform relations**

a. 

\[
Q(z) \triangleq \sum_{k=0}^{\infty} p_k z^k, \quad p_k = Z^{-1}[Q(z)]
\]

\[
B^*(s) = L[b(x)] = \int_0^{\infty} b(x)e^{-sx} dx
\]

\[
Q(z) = B^*(\lambda - \lambda z) \frac{(1 - \rho)(1 - z)}{B^*(\lambda - \lambda z) - z}
\]

where

\[
B^*(\lambda - \lambda z) = B^*(s) \bigg|_{s = \lambda - \lambda z}
\]

b. 

\[
w(y) = \text{prob. density fn. of waiting time}
\]

\[
W^*(s) = L[w(y)]
\]

\[
W^*(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda B^*(s)}
\]

\[
w(y) = L^{-1}[W^*(s)], \quad y \geq 0
\]

\[
s(y) = \text{prob. density fn. of system time}
\]

\[
S(s) = L[s(y)]
\]

\[
S^*(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda B^*(s)} B^*(s)
\]

\[
s(y) = L^{-1}[S^*(s)]
\]
\[ S^*(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda \mu} \frac{\mu}{(s + \mu)} \]

\[ = \frac{s(1 - \rho) \mu}{s^2 + s(\mu - \lambda) - \lambda \mu + \lambda \mu} \]

\[ = \frac{(1 - \rho) \mu}{s + \mu - \lambda} \frac{\mu - \lambda}{s + \mu - \lambda} \]

\[ s(y) = (\mu - \lambda) e^{-(\mu - \lambda) y} \]
3. Modeling
   a. Imbedded Markov Chain

   \[ q_n = \text{no. customers left behind by } C_n \]

   \[ v_n = \text{no. customers which enter during } x_n \]

   \[ q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & q_n > 0 \\ v_{n+1} & q_n = 0 \end{cases} \]

   \( \{q_n, n = 0, 1, \ldots\} \) = Cont. –time Markov chain

   b. Tagged job

   \[ W = R + (\lambda W)x \]

   \[ R = \text{Residual work (residual life time)} \]

   \[ W = R + (\lambda x)\bar{W} = R + \rho W \]

   \[ (1 - \rho)W = R \]

   \[ W = R / (1 - \rho) \]
Residual Life - R

Service process:
\[ b(x), \bar{x}, x^2 \]

\[ F_Z(x) = P[Z \leq x | \text{system busy at time of arrival}] \]
\[ r = E[Z | \text{system busy}] \]

Intuition: \[ r = \frac{x}{2} \quad ? \text{Wrong!} \]
Residual Life - Cont’d

Chances are higher that x arrives during longer periods

\[ f_Z(x) = \text{density fn. for length of service period during which arrival occurs, given system busy} \]

\[ f_Z(x)\Delta(x) = Kx b(x)\Delta(x) \]

\[ \int_0^\infty f_Z(x) \, d(x) = K \int_0^\infty x \, b(x) \, dx = k\bar{x} = 1, \quad K = \frac{1}{\bar{x}} \]

\[ f_Z(x) = \frac{x \, b(x)}{\bar{x}} \]

\[ r = \int_0^\infty \left( \frac{x}{2} \right) \left( \frac{x \, b(x)}{\bar{x}} \right) \, d(x) = \frac{x^2}{2\bar{x}} \]
Residual Life - Cont’d

\[ R = E[Z \mid \text{system busy}] E[\text{system busy}] \]

\[ = \frac{x^2}{2\bar{x}} \cdot \lambda\bar{x} = \frac{1}{2} \lambda \frac{x^2}{\bar{x}} \]

\[ \therefore W = \frac{1}{2} \lambda\frac{x^2}{\bar{x}} = \frac{\lambda x^2}{2(1 - \rho)} \]

Markovian queueing Networks

- N-node interconnection of queueing systems
- Queueing and service at each node
- Exponential service times at each node
- Model multi-service processes
1. Open Networks

N nodes

Each node – single queue, \( m_i \) servers

Exponential service time \( x_i = \frac{1}{\mu_i} \) sec
- External arrivals – Poisson process (indep.) \( \gamma_i \) jobs/second
- \( r_{ij} = P[\text{that a job which completes service at node } i \text{ will proceed next to node } j] \)
- \( 1 - \sum_{j=1}^{N} r_{ij} = P[\text{that a job which completes service at node } i \text{ will leave the network}] \)
- \( \lambda_i = \text{Avg. arrival rate at node } i \text{ from both external sources and other nodes (including itself)} \)

\[ \lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j r_{ji} \]

Flow Equations:
\[ \lambda = [\lambda_1 \lambda_2 \ldots \lambda_N] \]
\[ \gamma = [\gamma_1 \gamma_2 \ldots \gamma_N] \]
\[ R = [r_{ji}] \]
\[ \lambda = \gamma + \lambda R \]
Jackson’s theorem (1957)

Let \( p(k_1, k_2, \ldots, k_N) \) = \( p(k_1 \text{ jobs at node 1, } k_2 \text{ jobs at node 2, } \ldots, k_N \text{ jobs at node N}) \) and

\[ p_i(k_i) = p[k_i \text{ jobs at node i}] \]

Then

\[ p(k_1, k_2, \ldots, k_N) = p_1(k_1)p_2(k_2)\ldots\ldots p_N(k_N) \]

Each node in the network can be treated as an \( M/M/m_i \) queue, \( m_i \geq 1 \)
\[
\overline{N} = \sum_{i=1}^{N} \overline{N}_i \\
\gamma = \sum_{i=1}^{N} \gamma_i \\
T = \frac{\overline{N}}{\gamma}
\]
Example:

\begin{align*}
\gamma & = 1 \text{ job/sec} \\
\gamma_1 & = 0.2 \quad \Rightarrow \quad \gamma_{12} = 0.8 \\
\gamma_2 & = 0.4 \quad \Rightarrow \quad \gamma_{23} = 0.6 \\
\frac{1}{\mu_1} & = 0.1 \text{ sec} \\
\frac{1}{\mu_2} & = 0.05 \text{ sec} \\
\frac{1}{\mu_3} & = 0.9 \text{ sec}
\end{align*}
Node 1

\[ \lambda_1 = \gamma + r_{11} \lambda_1 = 1 + 0.2 \lambda_1 \rightarrow \lambda_1 = 1.25 \]

\[ \rho_1 = \frac{\lambda_1}{\mu_1} = 0.125 \quad \overline{N}_1 = \frac{\rho_1}{1 - \rho_1} = 0.1429 \]

Node 2

\[ \lambda_2 = r_{12} \lambda_1 + r_{22} \lambda_2 = 0.8(1.25) + 0.4 \lambda_2 \rightarrow \lambda_2 = 1.6667 \]

\[ \rho_2 = \frac{\lambda_2}{\mu_2} = 0.0833 \quad \overline{N}_2 = \frac{\rho_2}{1 - \rho_2} = 0.0909 \]

Node 3

\[ \lambda_3 = r_{22} \lambda_2 = 0.6(1.6667) = 1 = \gamma \]

\[ \rho_3 = \frac{\lambda_3}{\mu_3} = 0.9 \quad \overline{N}_3 = \frac{\rho_3}{1 - \rho_3} = 0.9 \]

\[ \overline{N} = \overline{N}_1 + \overline{N}_2 + \overline{N}_3 = 9.2338 \]

\[ T = \frac{\overline{N}}{\gamma} = 9.23 \text{ sec} \]
2. Closed Networks

N nodes; K jobs circulating through the network
No external arrivals or departures
Each node – single queue, \( m_i \) servers
Exponential service time, \( x_i = \frac{1}{\mu_i} \) sec

\[
\begin{align*}
\gamma_i &= 0 \quad \sum_{j=1}^{N} r_{ij} = 1 \quad \text{for each } i \\
\sum_{i=1}^{N} K_i &= K \\
\lambda &= [\lambda_1, \lambda_2, \ldots, \lambda_N] \quad \text{(flow vector)} \\
\lambda &= \lambda R
\end{align*}
\]
Gorjon and Newell (1967)

\[ p(k_1, k_2, \ldots, k_N) = \frac{1}{G(K)} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)} \]

where

(1) \{x_i\} satisfy

\[ \mu_i x_i = \sum_{j=1}^{N} \mu_j x_j r_{ji} \quad i = 1, 2, \ldots, N \]

(2) \[ G(K) = \sum_{k \in A} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)} \]

with \( k = (k_1, k_2, \ldots, k_N) \) and

\[ A = \text{set of all } k \text{ vectors for which } k_1 + k_2 + \ldots + k_N = k \]

(3) \[ \beta_i(k_i) = \begin{cases} k_i! & \text{if } k_i \leq m_i \\ m_i! m_i^{k_i - m_i} & \text{if } k_i \geq m_i \end{cases} \]

Utilization at node i: \( \rho_i = \frac{x_i}{m_i} \quad i = 1, 2, \ldots, N \)
Example: (Application-interactive computing)

\[ \gamma = \frac{N_1}{\tau} = \text{avg. "think" time at each terminal (exp. distribution think times)} \]

\[ T = \frac{K}{\gamma} - \tau \text{ seconds} \]

\[ T = \text{avg. response time} \]
2. M/M/1 \ (\lambda_k = \lambda, \mu_k = \mu)

Equilibrium

\[ \mu P_k = \lambda P_{k-1} \]

\[ k = 1, 2, \ldots \]

\[ P_k = \frac{\lambda}{\mu} P_{k-1} \]

This leads to

\[ P_1 = \frac{\lambda}{\mu} P_0 \]

\[ P_2 = \frac{\lambda}{\mu} P_1 = \left(\frac{\lambda}{\mu}\right)^2 P_0 \]

\[ \vdots \]

\[ P_k = \left(\frac{\lambda}{\mu}\right)^k P_0 \]

subject to

\[ \sum_{k=0}^{\infty} P_k = 1 \]