II. QUEUING THEORY

(a) General Concepts

- queuing theory useful for considering performance analysis of packet switching and circuit switching

General model of a queue:

- in practice, queue size is finite (i.e., number of packets that can be queued is limited $\Rightarrow$ extra packets discarded $\Rightarrow$ "blocking")

- if $\lambda > \mu \Rightarrow$ # queued packets will grow until queue saturated (remains full) or if queue size allowed to be $\infty$ (in theory), # queued packets will grow without bound

- $\rho = \lambda/\mu = \text{utilization or traffic intensity}$
- as $\rho \to 1$, queue becomes unstable

- factors of interest: time delay, blocking performance, packet throughput (packets/time to get through)

- queue modelled by considering
  (1) packet arrival statistics
  (2) service time distribution (i.e., packet length distribution)
  (3) service discipline - FIFO, priority discipline
  (4) buffer size
  (5) input population (finite or $\infty$)

(b) Poisson Process

- arrival process (eg. packets generated at input to packet switch network or call initiated in circuit switch network) are often assumed to be Poisson

- continuous time:

- discrete time:
- divide time $t$ into $n$ intervals of length $\Delta t$ (very small)

- let probability of arrival to queue in interval $\Delta t = p_+$ and assume all arrival events are independent (i.e., memoryless)

- assume $\Delta t$ is small enough so that probability of $\geq 2$ arrivals in $\Delta t$ is negligible, i.e., $p_+ \ll 1$, then $p_+ \approx \lambda \Delta t$ (recall $\lambda =$ arrival rate)

- rationale:

- average # of arrivals in an interval $t = \lambda t = np_+$

  $$\Rightarrow p_+ = \lambda t/n$$

- probability of exactly $k$ arrivals in $n = t/\Delta t$ intervals

  $$P_k(n) =$$

  (binomial distribution)
- hence,

- for fixed $t$, let $\Delta t \to 0 \Rightarrow n \to \infty$ since $t = n \cdot \Delta t$
  (i.e., making discrete case continuous)
\[ \lim_{n \to \infty} P_k(n) = P_k(t) \]

(discrete)    (continuous)

- probability of \( k \) arrivals in a time \( t \)

\[ P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \]

Poisson Distribution

Notes:
(c) M/M/1 Queue

Interarrival Time

What is distribution?

- consider arbitrary point in time $t_0$ and define $t_0 = 0$

$$P(\text{arrival at time } t) = P(\text{no arrival in interval (0, } t)) \times P(\text{arrival in interval (} t, t + \Delta t))$$

- using independence

= 
- consider graph of \( f(t) = \lambda e^{-\lambda t} \)

- since \( f(\tau) \Delta t = \text{probability} = \text{area under} f(t) \) then \( f(t) = \lambda e^{-\lambda t} \)
is probability density function

- now \( t_0 = 0 \) can represent any arbitrary point in time, so since it can represent an arrival event point, the variable \( t \) represents an interarrival time

\[ \therefore \text{interarrival time has exponential distribution with} \]

\[ \text{pdf } f(t) = \lambda e^{-\lambda t} \]

Note:
Departures

- assume packets in queue and let $p_\cdot = $ probability of departure in interval $\Delta t$

- define $p_\cdot = \mu \Delta t$ (recall $\mu = $ service or departure rate)

$\therefore P(\text{departure after } n \text{ intervals})$

$= $ (geometric distribution)

$\therefore$ service/departure time pdf

$f(t) = \mu e^{-\mu t}$ Exponential (same as arrivals)
**M/M/1 Queue:**

\[ M / M / 1 \]
→ Markov Arrivals / Markov Departures / One Server

- Markov process → memoryless process

- \( M/M/1 \) ⇒ Poisson arrivals, exponential service times, one server

**(d) Discrete Model of M/M/1 Queue**

- let \( k = \# \) packets in queue including packet being served

- hence, \( k \) is a random variable and can be considered to be queue state

- now divide time into small intervals of \( \Delta t \)
State diagram:

- state transition for every interval

\[ P_k = \text{probability system in state } k \text{ in an interval} \]

⇒

Lemma 1

- by definition of pdf

Lemma 2

Theorem
- an interpretation

Proof of theorem by induction:

Base Case:
Induction Case:

- show if it is true for $k - 1$, it is true for $k$

Note:
- if you know one state probability and transition probabilities, you can determine probability of being in any state
- expect $P_k \rightarrow 0$ as $k \rightarrow \infty$ or queue will blow up
What is mean # of customers in queue?

\[ \bar{k} = \frac{\rho}{1 - \rho} \]
What is variance of $k$? (variance is a measure of spread)

$$\sigma_k^2 = \frac{\rho}{(1 - \rho)^2}$$

- conservation of customers for M/M/1 ($\infty$ size)

What is time spent in queue?

Little's Theorem: $$\bar{k} = \lambda \bar{T}$$

where $\bar{T} =$ average time spent in system (including service time)
intuition:

- if serviced in $T$ and still $\bar{k}$ customers in queue, then for equilibrium $\bar{k} / \bar{T} = \lambda$

- makes intuitive sense but will not formally prove

- holds for M/M/1 and many other queues as well

(e) Queues with Finite Buffers

$M/M/1$ queue of size $N \rightarrow M/M/1/N$
- buffer overflow occurs when $k = N$ and packet arrives

- can use same state analysis as previous, except only $N+1$ states, instead of infinite number of states

- so

- mean of $k$
Conservation perspective:

\[ \gamma = \mu(1-P_0) \]

rate of servicing \hspace{1cm} fraction of time customer being served

\[ \therefore \mu(1-P_0) = \lambda(1-P_B) \]

\[ \rho = \frac{\lambda}{\mu} = \frac{1-P_0}{1-P_B} \]

\[ \therefore P_B = \frac{P_0 - (1-\rho)}{\rho} \]

- recall

\[ \therefore P_B = P_N \]
(f) Extensions of M/M/1 Queue

(i) Multiple Sources

- combining two or more Poisson processes
  ⇒ Poisson process

(ii) Multiple Servers

$M/M/m$ Queue:
- let $k$ represent packets served and packets waiting and consider 2 cases:

(1) $k \leq m$  (i.e., all customers being served)

(2) $k \geq m$  (i.e., $m$ customers being served, $k - m$ waiting)
- for \( k \leq m \)

- for \( k \geq m \)
What is probability customer arrives at system and must wait to be served?

\[ P_w = \]

Erlang C Formula for M/M/m queue

(iii) Feedback

- simple communication system model
in this case arrivals to Q2 are Poisson but loss may occur

- to minimize loss use feedback → feedback channel to shut off transmitter when receiver full

- arrivals to Q2 are now not Poisson, although $\mu_2 \geq \lambda_1$ or queue Q1 will blow up

(g) M/G/1 System

- often exponential service time is not an accurate model

   eg. in ATM, fixed size cell ⇒ deterministic service time of cell size / link rate
- "G" represents general distribution for service time $\tau$ with known mean and variance

- let $T =$ time in system, $W =$ time waiting in queue

so

$$T = W + \tau$$

$$E\{W\} = \frac{\lambda E\{\tau^2\}}{2(1 - \rho)}$$

and average customers in queue given by $\bar{k} = \lambda \bar{T}$ (Little's Theorem)
Special Cases:

Example Queuing Problem:
(h) Queuing Network Examples

- Communication networks are, in fact, complex network of queues

Example 1:

Example 2:
Example 3:
Closed Queuing Networks

Aside: Norton equivalent of queuing network

$N$ packets circulating around closed queuing network

- service rate dependent on number in queue

- derived by short circuitry $A \rightarrow B$ and allowing $n$ customers to circulate
Example 4:

Sliding Window Flow Control with window size $N$
- assume all queues have same average service rate (i.e., same average packet sizes and link rates)

- assume ACKs are sent on high priority zero delay channel and are sent for every packet

- queue $M+1$ is an artificial queue used to represent generation of packets to send

- equivalent network:

What is $u(n)$?
\[ P_n = \]

and then from (2)

\[ P_0 = \]

- now throughput

\[ \gamma = \]
- using Little's formula, average total delay

→ could determine average delay from parameters \( N, M, \mu, \lambda \) for sliding window flow control

- consider scenario where \( \lambda \to \infty \) (i.e., packets served in zero time for queue \( M+1 \) implying data packets sent immediately following ACK)