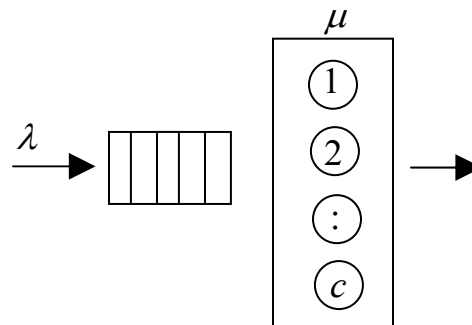
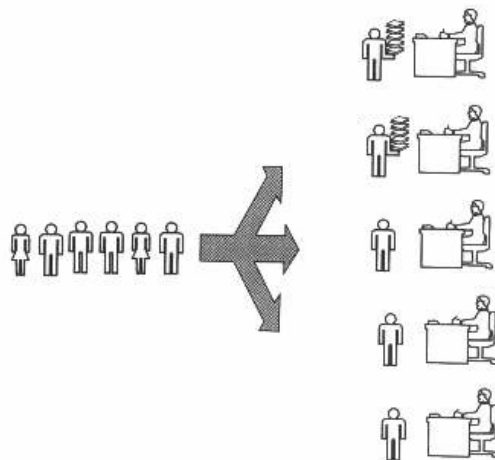


## Queueing Theory (1)

### • What is a queueing system?

- A queueing system consists of “servers” (resources) that provide service to “customers” (entities).
- A Customer requesting service will start service if the required server is not busy. Otherwise, the customer waits in queue until the server is available.
- Queueing (waiting in line) happens because there are not enough resources at certain times.



## • Components of a queueing System

- A queueing system can be composed of one or many service centers or nodes. Customers are “routed” from one node to the other according to certain rules.
- Each node is characterized by three components.
  - (i) The arrival process ;
  - (ii) The service process ;
  - (iii) The queue discipline .
- The arrival process is specified through the random variables  $A_1, A_2, \dots$ , where  $A_i$  is the inter-arrival time between the  $(i - 1)^{\text{st}}$  and the  $i^{\text{th}}$  customer.
- A typical modeling assumption is to assume that  $A_i$ 's are independent and identically distributed (iid). Then, the arrival process is characterized by  $F_A(x) = P\{A < x\}$ , the cdf of  $A$ .
- Important parameters of the arrival process (in addition to  $F_A(\cdot)$ ) are the mean inter-arrival time  $E[A]$ , and the arrival rate  $\lambda = 1 / E[A]$ , the arrival rate.
- The most commonly assumed arrival process is the Poisson process.
- This assumption is realistic (in most cases). In addition, it greatly simplifies the analysis.

- The service process is specified through the random variables  $S_1, S_2, \dots$ , where  $S_i$  is the service time the  $i^{\text{th}}$  customer.
- The  $S_i$  are also typically assumed iid with cdf  $F_S(x)$ .
- Important parameters of the arrival process are the mean service time  $E[S]$  and the service rate  $E[S] = 1/\mu$ .
- Service times are also commonly assumed to be exponential.
- Analytical methods that analyze queues are quite complex without the exponential assumption.
- The queueing or service discipline refers to the rule utilized to select the next customer from the queue when a customer finishes service.
- Typical queueing discipline include first-in, first-out (FIFO), last-in, first out (LIFO), processor sharing (PS), service in random order (SIRO), and priority (PR).
- Under the iid assumptions, a single-node queueing system is generally denoted by  $GI/GI/c$ , where the “ $GI$ ” refers to iid arrival and service processes and  $c$  is the number of servers.
- If the inter-arrival and service times are iid exponential then the queue is denoted by  $M/M/c$ , where the “ $M$ ” refers to the Markovian or *memoryless* property of the exponential distribution.

- **Performance measures and general relations**

- Consider a  $GI/GI/c$  queue (to simplify things).
- In the following we define “steady state” measures, which are statistical measures after the system has been operational for a time which is large enough .
- An important measure is the *traffic intensity*,  $\rho = \lambda/(c\mu)$ .
- If  $\rho \geq 1$ , then it can be shown that the queue length will increase indefinitely as time passes.
- In “stable” systems,  $\rho < 1$ .
- For a single-server system,  $\rho$  is the *mean server utilization*.
- The *stationary system size distribution* is

$$P_n = \lim_{t \rightarrow \infty} P\{L(t) = n\} ,$$

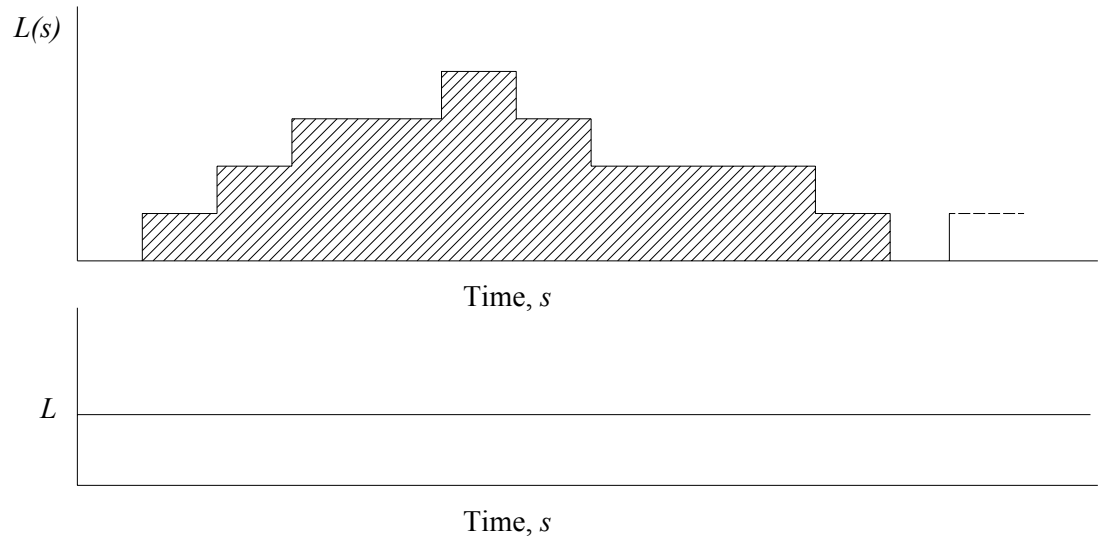
where  $L(t)$  is the number of customers in the system at time  $t$ .

- The *mean number in the system* is

$$L = \sum_{n=0}^{\infty} nP_n .$$

- It can be shown that  $L$  can be estimated differently as

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(s) ds .$$



- The *mean waiting time in the system* is

$$W = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n W_i}{n},$$

where  $W_i$  is the waiting time of customer  $i$  in queue plus the time the customer spend in service.

- Among the most important queueing theory results is *Little's law*

$$L = \lambda W.$$

- Other measures concern waiting in queue.
- The *mean number in the queue* is

$$L_q = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L_q(s) ds,$$

where  $L_q(s)$  is the number of customers in queue at time  $s$ .

- $L_q$  can be also written as

$$L_q = \sum_{n=c}^{\infty} (n - c) P_n.$$

- The *mean waiting time in the queue* (or the *mean delay*) is

$$W_q = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n W_q^i}{n},$$

where  $W_q^i$  is the waiting time of customer  $i$  in queue plus the time the customer spends in service.

- Little's law implies that

$$L_q = \lambda W_q.$$

- Furthermore,  $W$  and  $W_q$  are related by

$$W = W_q + 1/\mu.$$

- Multiplying by  $\lambda$  and applying Little's law we get

$$L = L_q + \lambda/\mu.$$

- Here,  $\lambda/\mu$  can be seen as the mean number of busy servers.

(This can in fact be also proven by Little's law.)

- Note that knowing one of the four performance measures,  $L$ ,  $W$ ,  $L_q$ , and  $W_q$ , allows determining the other three.

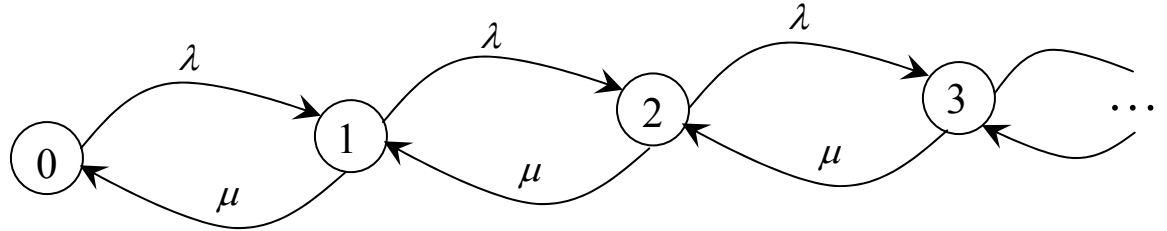
- **The  $M/M/1$  queue**

- Consider a single-server queue with iid exponential inter-arrival and service times (hence called  $M/M/1$ ).
- Let  $\lambda$  and  $\mu$  denote the arrival and service rates and  $\rho = \lambda/\mu$ .

Assume  $\rho < 1$ .



- The number of customers in the  $M/M/1$  system  $L(t)$  is a birth death process with  $\lambda_i = \lambda$ , and  $\mu_i = \mu$ , with the following transition probability diagram:



- Recall that  $\rho = \lambda / \mu$  is the traffic intensity.
- $\rho$  can be also seen as the average server utilization, or the fraction of time the server is busy.
- Applying the general flow balance equation for a birth-death process, the limiting probabilities are given by

$$\begin{aligned}
 P_0 &= \left( 1 + \sum_{n=1}^{\infty} \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_1 \lambda_0}{\mu_n \mu_{n-1} \dots \mu_2 \mu_1} \right)^{-1} = \left( 1 + \sum_{n=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^n \right)^{-1} = \left( 1 + \sum_{n=1}^{\infty} \rho^n \right)^{-1} \\
 &= \left( 1 + \rho \sum_{n=1}^{\infty} \rho^{n-1} \right)^{-1} = \left( 1 + \frac{\rho}{1 - \rho} \right)^{-1} = 1 - \rho. \\
 P_n &= \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_1 \lambda_0}{\mu_n \mu_{n-1} \dots \mu_2 \mu_1} P_0 = \rho^n (1 - \rho), \quad n = 1, 2, \dots
 \end{aligned}$$

- Note that  $\sum_{n=1}^{\infty} \rho^n < \infty$  only if  $\rho < 1$  or equivalently  $\lambda < \mu$ .
- This is the stability condition that should be always satisfies in order for the  $M/M/1$  queue to have a finite congestion level (measured in  $L$ ,  $W$ ,  $W_q$ , etc).

- $P_0$  could be found in another way by noting that

$$P_0 = P\{\text{server is idle}\} = 1 - P\{\text{server is busy}\} = 1 - \rho.$$

- The mean number in the system,  $L$ , is

$$L = \sum_{n=1}^{\infty} nP_n = \sum_{n=1}^{\infty} n\rho^n(1-\rho) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

- Other performance measures are determined as follows:

$$L_q = L - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)},$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda}, \text{ (Little's law)}$$

$$W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)}, \text{ (Little's law)}$$

- Another measure of performance is the probability that the number of customers in the system is  $n$  or more

$$P_{n+} = \lim_{t \rightarrow \infty} P\{L(t) \geq n\} = \sum_{k=n}^{\infty} P_k = \sum_{k=n}^{\infty} \rho^k(1-\rho) = \rho^n.$$

### • Example 1

- Customers arrive at a bank according to a Poisson process with rate 9 customers per hour and request service of a single teller. The teller has an exponential service time with rate 10 customers per hour.
- What is the fraction of time the teller is busy?
- This can be modeled as a  $M/M/1$  with  $\lambda = 9$  customers/hour and  $\mu = 10$  customers/hour.



- The traffic intensity is  $\rho = 9/10 = 0.9$ , which is also the fraction of time the server is busy.

- What is the mean number of customers in the bank?

$$L = \rho / (1 - \rho) = 0.9/0.1 = 9 \text{ customers.}$$

- What is the mean number of customers waiting in line?

$$L_q = L - \rho = 9 - 0.9 = 8.1 \text{ customers.}$$

- What is the mean time a customer spends in the bank?

$$W = L / \lambda = 9 / 9 = 1 \text{ hour.}$$

- What is the mean delay in queue?

$$W_q = L_q / \lambda = 8.1 / 9 = 0.9 \text{ hour} = 54 \text{ mins.}$$

- For what fraction of time the number of customers in the system exceeds 3?

$$P_{4+} = \rho^4 = 0.9^4 = 0.656 .$$

- What do you think of this system performance?

- Not good. Mean delay time is too long.

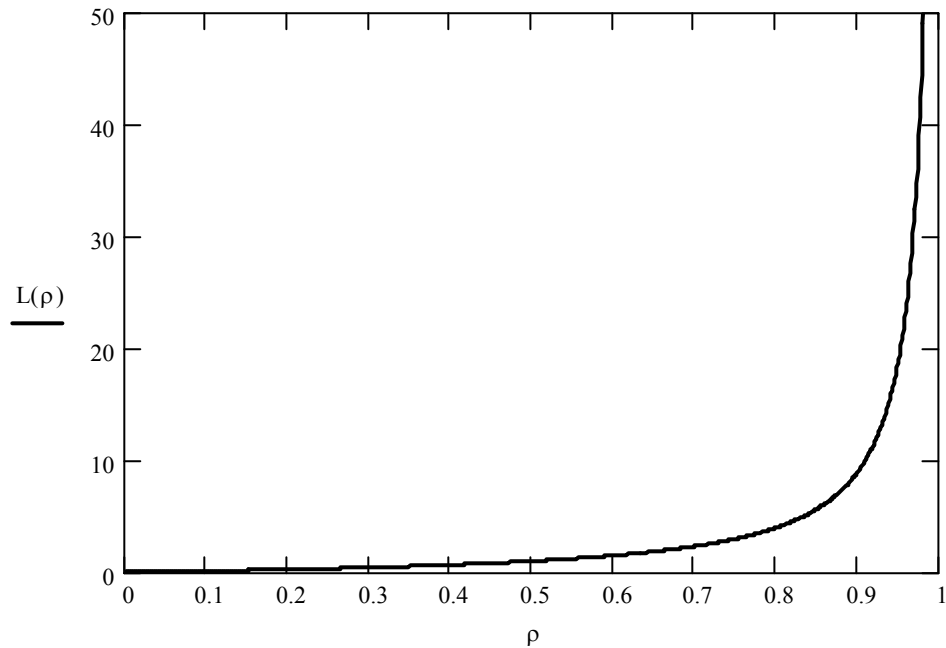
- How would you improve the performance?

- Add one or more servers, or train the teller (if possible) so she handles customers faster (i.e. increase  $\mu$ ).

- What other measures of performance would you estimate?

- $P\{\text{waiting time} > \bar{t}\} < \alpha$ , where  $\alpha$  is small.

- Beware of the nonlinear behavior of queues!



- For example, if  $\rho$  is increased from 0.9 to 0.945 (by 5%) in the bank example,  $L$  increases from 9 to 17.18 (by about 100%).