Queuing Systems - Introduction

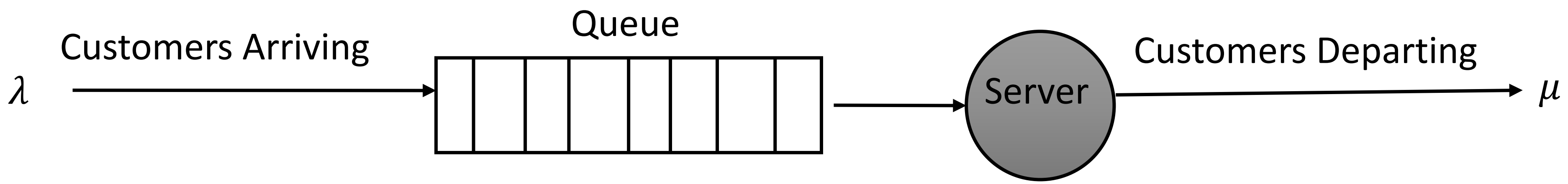
- How much time is spent in one’s daily activities waiting in some form of a queue?
- For example:
  - stopped at a traffic light;
  - delayed at a supermarket checkout stand;
  - standing in line for a ticket at a bus stand;
  - holding the telephone as it rings, and so on.
- One thing common to all the systems is the flow of customers requiring service and there being some restriction on the service that can be provided.
- For example, patients arriving at an out-patient’s clinic to see a doctor, the restriction on service is that only one patient can be served at a time. This is a case of single server queue. An example of multi server queue is a queue for having goods checked at a supermarket.
- In the above examples , the restriction on service is that not more than a limited number of customers can be served at a time, and congestion arises because the unserved customers must queue up and await their turn for service.
- Queueing Theory provides a mathematical basis for understanding and predicting the behavior of a system in general and communication network in particular.

Queuing Systems - Introduction

- Queuing systems are used for analyzing systems performance(qualitative measure) and estimating average packet delay(quantitative measure).
- Queuing arises naturally in both packet-switched and circuit-switched networks.
- Much of the theory of queuing was developed from the study of telephone traffic at Copenhagen telephone exchange in 1910 by A. K. Erlang.
- In networking, a single buffer forms a queue of packets.
- A single queue of packets is an accumulation of packets at certain router or at entire network.
- A queuing buffer is a physical system that stores incoming packets and a server which can be viewed as a mechanism to process and route packets to desired destination.
- A queuing system consists of a queuing buffer of various sizes and one or more servers.

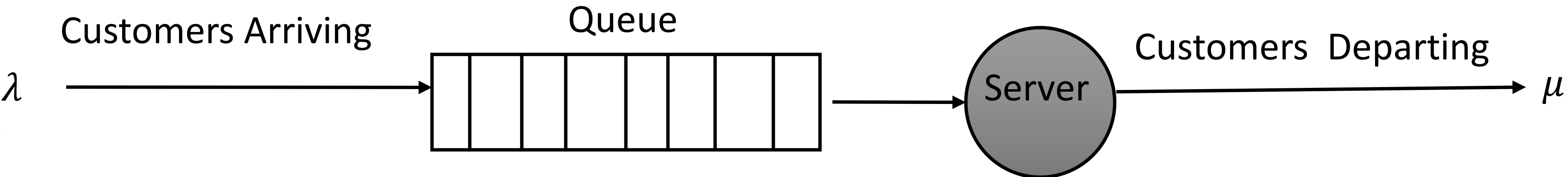
What is a Queue?

- The simplest queue in which customers arrives randomly at an average rate of λ customers per second.
- The customers are held in a queue while a server deals with them at a rate of μ per second and then they leave the system.
- This type of system is known as a single-server queue, although there may be more than one server in a system.
- It is important that the arrival rate λ is not allowed to exceed the service rate μ, or the queue will build up.



Components of Queuing Systems

- Generally a queueing system can be characterized by the following components:
- Customers – Entities that receive service from server e.g. a process, a transaction , a packet or a message etc.
- Server – Hardware or software providing service e.g. a CPU , an I/O device , software routine or a router etc.
- Major parameters are:
  - Interarrival Time Distribution – arrival pattern of packets
  - Service Time Distribution - length of time that a packet spends in the service facility.
  - No. of Servers - m
  - Queuing Disciplines - order in which packets are taken from the queue
  - No. of buffers – amount of buffer space present in the queue.

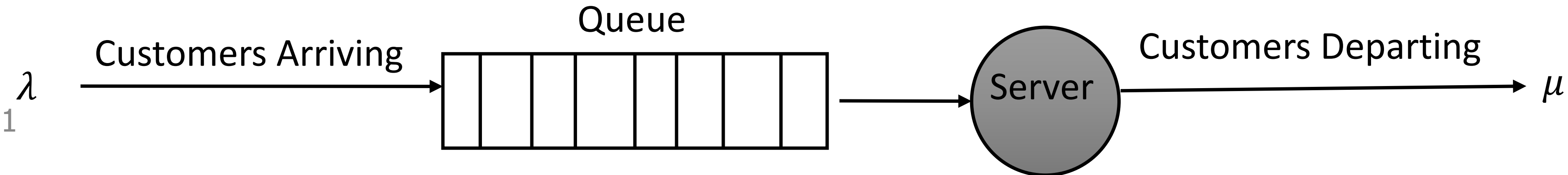


Queuing Disciplines

- Queueing systems may not only differ in their distributions of the interarrival and service times, but also in the number of servers, the size of the waiting line (infinite or finite), the service discipline.
- Some common service disciplines are:
- FIFO: (First in, First out): Customers in the queue are served in the order they arrive.
- LIFO: (Last in, First out): Customers in the queue are served in reverse order of their arrival.
- Random Service: Customers in the queue are served in random order.
- Round Robin: Every customer gets a time slice and if the service is not completed, the customer will re-enter the queue.
- Priority Disciplines: Every customer has a (static or dynamic) priority, the server selects always the customers with the highest priority.
- Preemption: The customer currently being served can be interrupted and preempted if the new customer in the queue has a higher priority.

Kendall’s Notation

- Queuing systems are described by Kendall’s notation as A/B/m/K where :
  - A – distribution of interarrival time of customers
  - B – distribution of service time
  - m – No. of servers
  - K – total capacity of system
- If a system reaches its full capacity, the arriving customer K+1 is blocked.
- A and B are represented with following symbols:
  - M – Exponential Distribution (M = Markov)
  - D – Deterministic Distribution (constant)
  - G – General or arbitrary Distribution



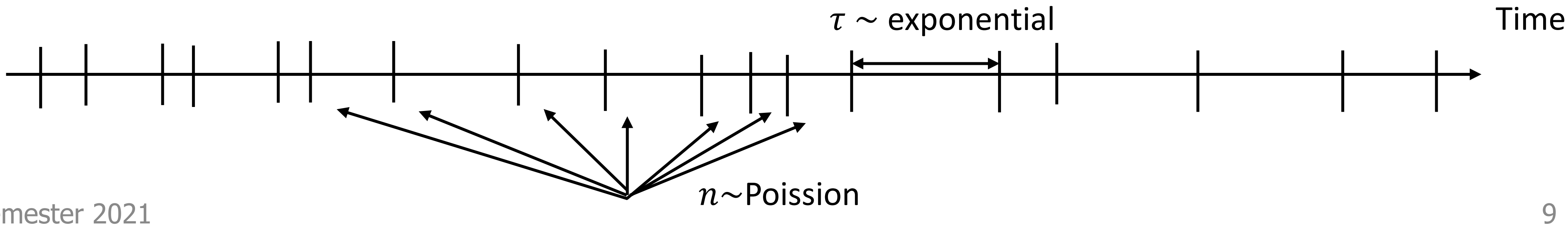
The Poisson Arrival Pattern

- Random arrivals or Poisson arrivals - The simplest arrival pattern mathematically, and the most commonly used in all applications of queueing theory is the random or Poisson arrival process.
- If the interarrival times are exponentially distributed, number of arrivals in any given interval are Poisson distributed.

λ = arrival rate ; Δt = time interval

- The probability that one arrival occurs between t and $t + \Delta t$ is independent of arrivals in earlier intervals.
- The probability of exactly n customers arriving during an interval of length t is given by:

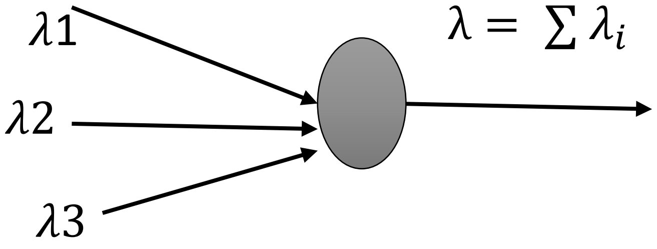
$$P(n\ arrivals\ in\ t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$



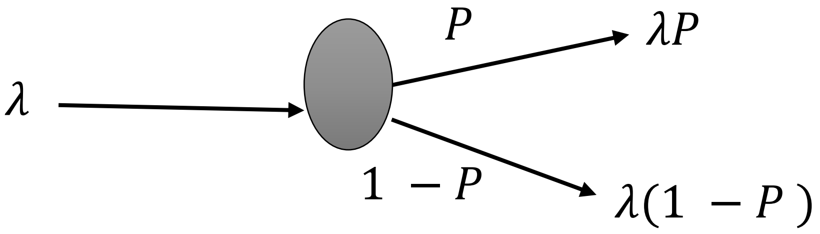
Properties of Poisson

- M (Markov property) - memoryless arrival or Poisson arrival i.e. Previous history does not help in predicting the future.
- Distribution of the time until the next arrival is independent of when the last arrival occurred.
- Merging - Let $A1, A2, \ldots\ Ak$ be independent Poisson Processes of rate $\lambda 1, \lambda 2, \ldots\ \lambda k$ then

$A = \sum A_i$ is also Poisson of rate = $\lambda = \sum \lambda_i$



- Splitting - Suppose that every arrival is randomly routed with probability P to stream 1 and $(1 - P)$ to stream 2 then streams 1 and 2 are Poisson of rates $P\,\lambda$ and $(1 - P)\,\lambda$ respectively.



Key Variables in Queuing Systems

λ = Arrival rate (customers/sec) ; C_n = Nth customer enters the system

μ = Service rate (customers/sec)

N = Average no. of customers in the system

N_q = Average no. of customers waiting in queue

T = Average customer time in system(includes queueing delay plus service time)

W_q= Average customer waiting time in the queue (does not include service time)

W_s = Service time ; reciprocal of service rate = $\frac{1}{\mu}$

τ = Interarrival time; reciprocal of arrival rate = $\frac{1}{\lambda}$

Key Variables in Queuing Systems

ρ = Traffic Intensity or utilization of server ; fraction of time the server is busy

$= \frac{mean\ service\ time}{mean\ interarrival\ time} = \frac{\frac{1}{\mu}}{\frac{1}{\lambda}} = \frac{\lambda}{\mu}$ (for single server)

$= \frac{\lambda}{m\mu}$ (for ‘m’ servers)

For equilibrium or steady state;

$$\rho < 1$$

i.e. number of customers arriving in a finite time is equal to the number of customers leaving the system. Otherwise, system will be unstable.

Little's Theorem

- Provides basis for queuing
- Also known as Little's Formula
- For a network to reach steady state, the average number of packets in a system (N) is equal to the product of the average arrival rate λ, and the average time (T) spent in queueing system.

$$N = \lambda T$$

- The usefulness of this theorem is that it applies to almost every queuing system.
- For example, slow moving traffic (large T) produces crowded streets (large N);
- The theorem can also be used to find the average number of packets in a queue rather than the overall system.
- If we define the following:

W_q= the average time spent waiting in the queue

Nq = the average number of packets found waiting in the queue

- Then Little's theorem leads to:

$$N_q = \lambda W_q$$

Little’s Theorem

Example: A fast-food restaurant is operating with a single person serving customers who arrive at an average rate of two per minute and wait to receive their order for an average of 3 minutes. On average, half of the customers eat in the restaurant and the other half eat take-away. A meal takes an average of 20 minutes to eat. Determine the average number of customers queuing and the average number in the restaurant.

Solution :

Arrival rate = $\lambda = 2/\text{min}$

Customers who eat in the restaurant stay on average for 23 minutes

Customers who take-away for 3 minutes.

Average customer time in restaurant = $T = 0.5 \times 23 + 0.5 \times 3 = 13\, minutes$

Average time in queue = $W_q = 3\, minutes$

From Little’s Theorem;

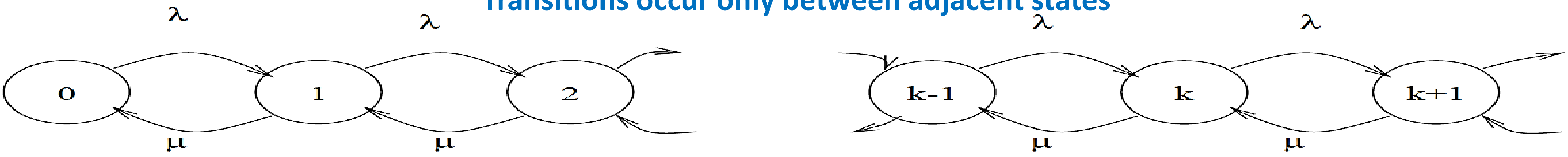
Average number of customers queuing = $Nq = \lambda W_q = 2 \times 3 = 6$ ✓

Average number in restaurant = $N = \lambda T = 2 \times 13 = 26$ ✓

Markovian Queuing Systems

- The common characteristic of all markovian systems is that the distribution of the interarrival times and the distribution of the service times are exponential distributions and thus exhibit the Markov (memoryless) property. Examples are : M/M/1, M/M/1/b and M/M/∞.
- The M/M/1 Queue has interarrival times, which are exponentially distributed with parameterλ and also service times with exponential distribution with parameter μ. The system has only a single server and uses the FIFO service discipline. The waiting buffer is of infinite size.
- The M/M/1 system is a pure birth/death process, where at any point in time at most one event occurs, with an event either being the arrival of a new customer(birth) or the completion of a customer's service(death).
- In a birth/death process, at any given state k can connect to only state $k-1$ with rate μ_k or to a state $k+1$ with rate λ_k

Transitions occur only between adjacent states



M/M/1 Queue Formulas

In this system, $\lambda_k = \lambda\,; k = 0,1,2,\ldots$ and $\mu_k = \mu\,; k = 0,1,2,\ldots$

We say that the state $A_k$is occupied if there are k customers in the queue including the one is being served.

Utilization Factor = $\rho = \frac{\lambda}{\mu}$

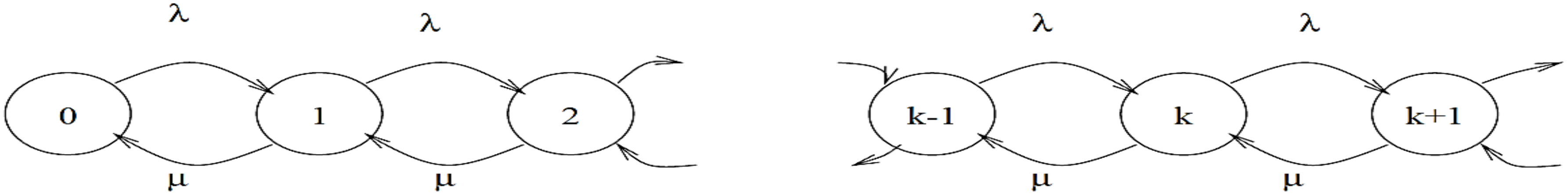
Probability of 'k' customers in the system = $\mathrm{P}(k) = \rho^k(1-\rho)$

Average number of customers in the system = $N = \frac{\rho}{1-\rho} = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}$

The average amount of time that a customer spends in the system can be obtained from Little's formula = $T = \frac{N}{\lambda} = \frac{\frac{\lambda}{\mu-\lambda}}{\lambda} = \frac{1}{\mu-\lambda}$

Average waiting time in the queue = $W_q = T - W_s = T - \frac{1}{\mu}$

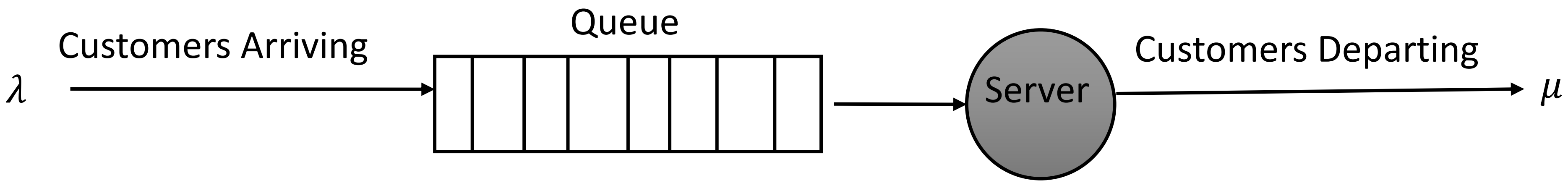
The average number of customers in the queue can be obtained from little's formula = $N_q = \lambda W_q = \frac{\rho^2}{1-\rho}$



M/M/1 Queue

Example:

- Consider a single server queue where the interarrival time is exponentially distributed with an average of 10 minutes and the service time is also exponentially distributed with an average of 8 minutes, find the following:
- Average wait time in the queue,
- Average number of customers in the queue,
- Average wait time in the system,
- Average number of customers in the system and
- Proportion of time the server is idle.



M/M/1 Queue

Solution:

Arrival rate = $\lambda = 1/10$

Service rate $= \mu = \ 1/8$

Utilization of server = $\rho = \frac{\lambda}{\mu} = \frac{8}{10}$

Average number of customers in the queue = $N_q = \frac{\rho^2}{1-\rho} = \frac{(0.8^2)}{1-0.8} = 3.2$

Average wait time in the queue = $W_q = \frac{N_q}{\lambda} = \frac{3.2}{\frac{1}{10}} = 32\ mins$

Average wait time in the system = $T = \frac{1}{\mu-\lambda} = \frac{1}{\left(\frac{1}{8} - \frac{1}{10}\right)} = 40\ mins$

Average number of customers in the system = $N = \lambda T = \left(\frac{1}{10}\right) * 40 = 4$

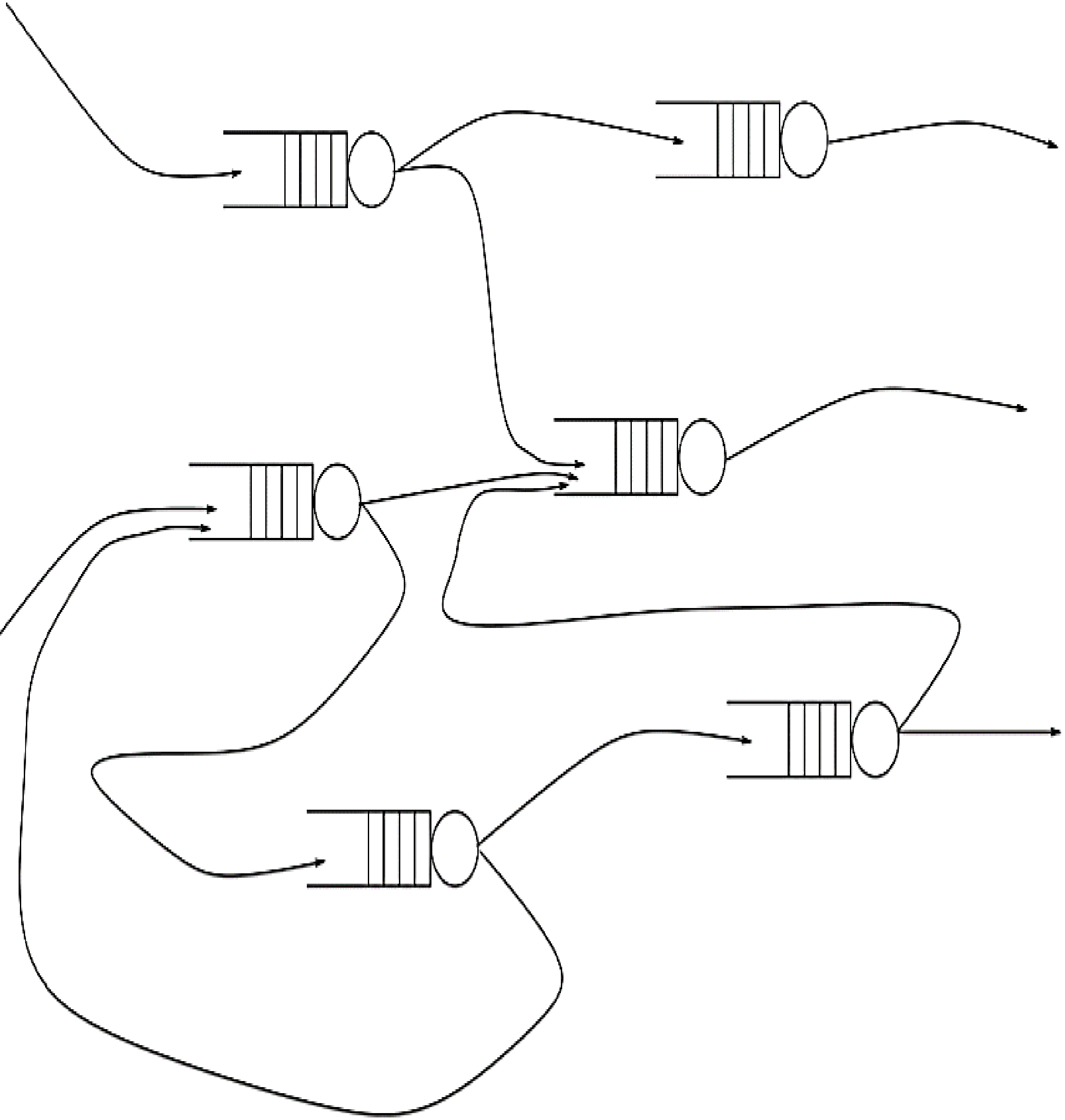
Proportion of time the server is idle = $1 - \rho = 1 - 0.8 = 0.2 = 20\%\ of\ the\ time.$

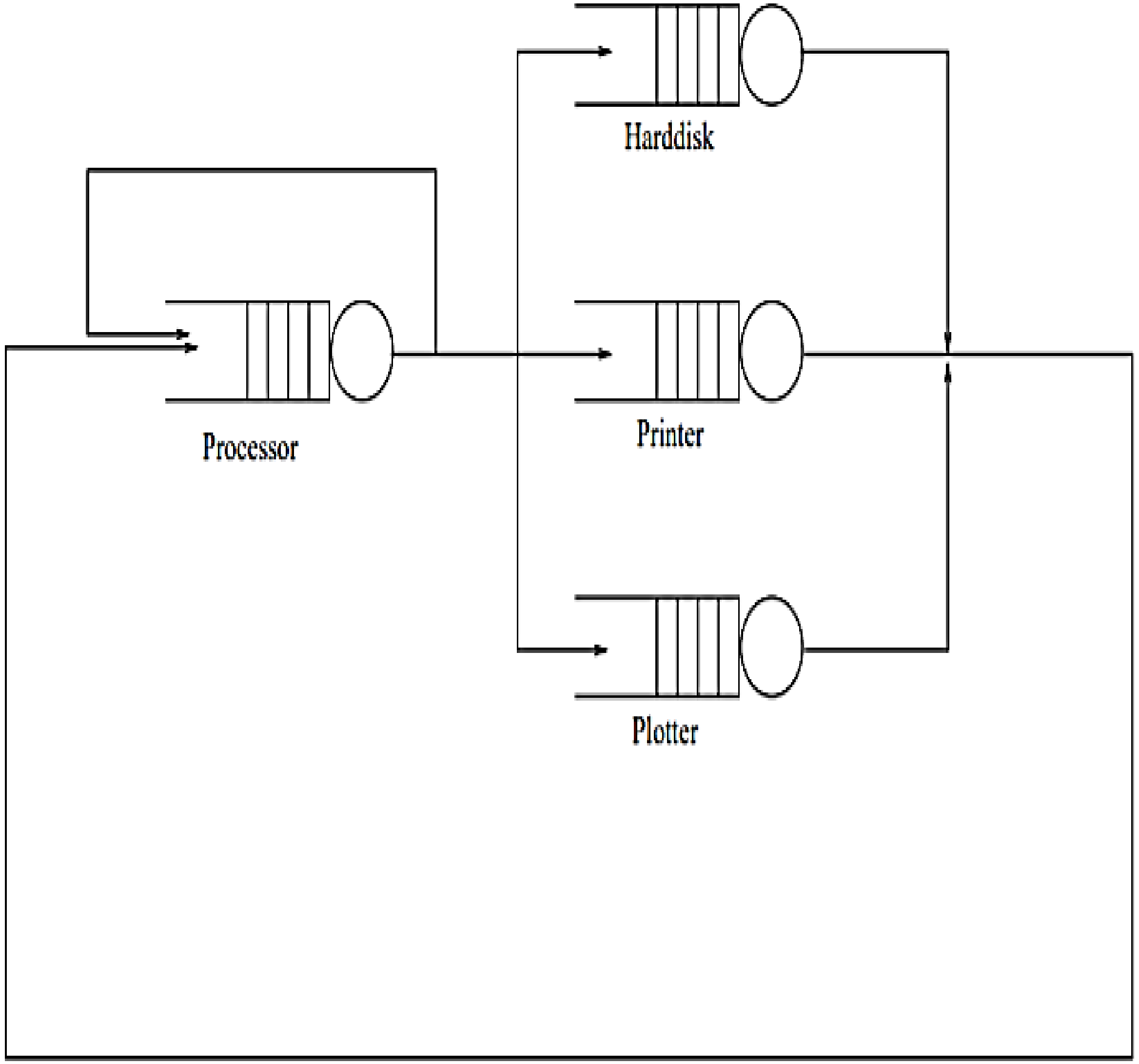
Effect of Errors on Delay

- If errors occur in a system and Automatic Repeat Request (ARQ) is used for error correction to retransmit erroneous packets, then the average transmission time will increase.
- This will result in increased delays and queue lengths.
- In determining an average transmission time, both the error rate and the type of ARQ strategy need to be taken into account.
- The greater the error rate, the more packets will need to be retransmitted, and the greater will be transmission and queuing times and queue lengths.

Queuing Networks

- So far we have only looked at a single standalone queueing system.
- However, most real systems are better represented as a network of queues.
- An example is the Internet, where we can model each outgoing link of each router as a single queueing system, and where an end-to-end path traverses a multitude of intermediate routers.
- One basic classification of queueing networks is the distinction between open and closed queueing networks.
- In an open network new customers may arrive from outside the system (coming from a conceptually infinite population) and later on leave the system. A simple example for an open queueing network may be the internet, where new packets arrive from "outside the system" (in fact, from the users).
- In a closed queueing network the number of customers is fixed and no customer enters or leaves the system. An example for a closed queueing network is the simple central server computer model . There is a fixed set of tasks and each task alternates between states where it performs some computations, thus using the processor and where it performs some I/O, e.g. access a hard disk and so forth.




Harddisk
Processor
Printer
Plotter