대기행렬 이론 개론

IV. Introduction to Queueing Systems

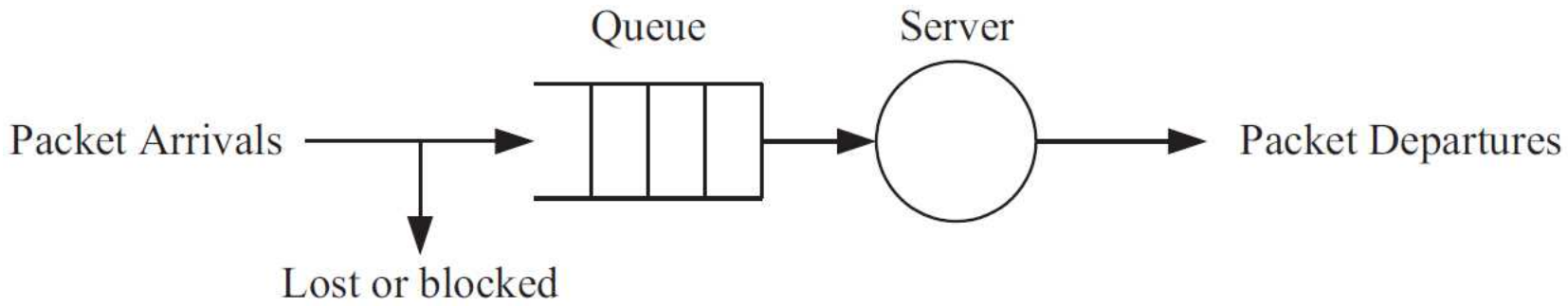
학습에 앞서

- 학습 목표
  - 대기행렬시스템의 특성 요소들 및 성능 측도를 학습한다.
  - Little의 법칙을 학습한다.

- 목차
  - 1. Introduction
  - 2. Queueing System Classification
  - 3. Little's Formula

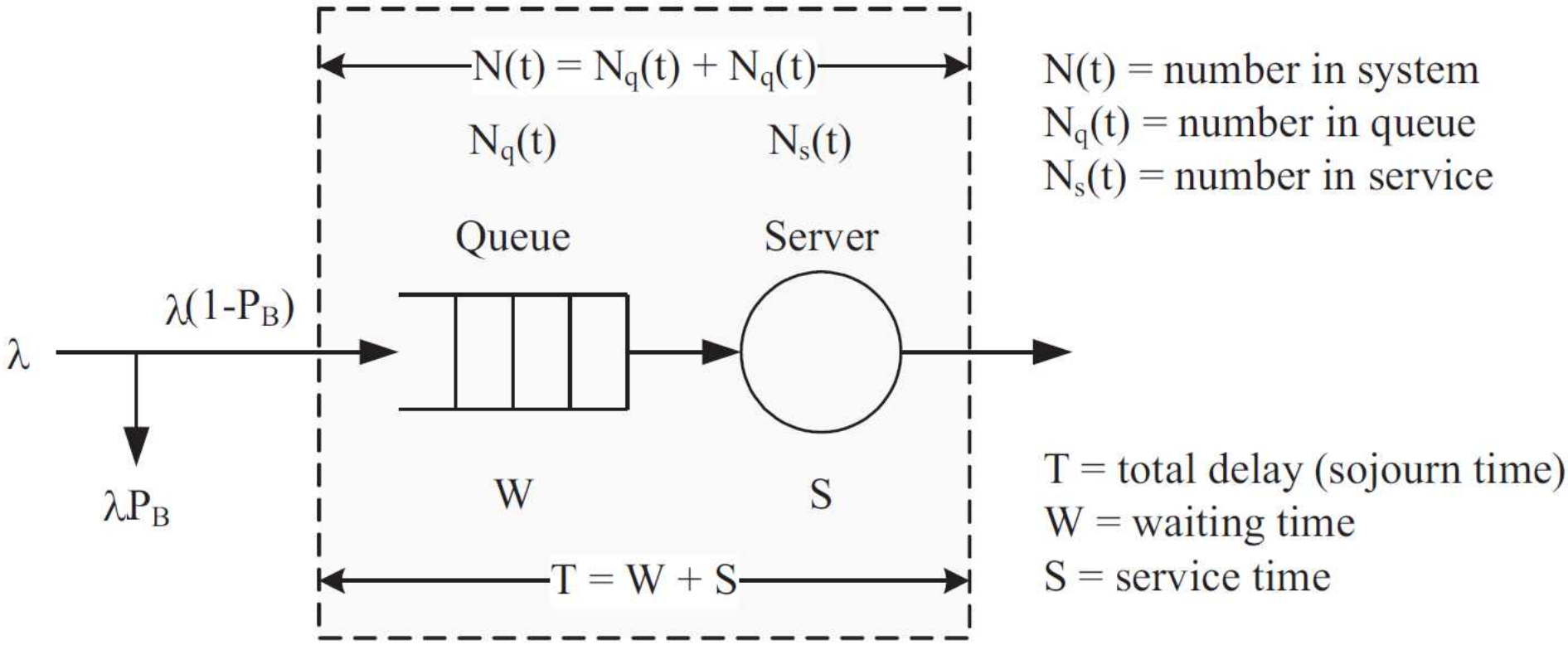
1. Introduction

- The following figure shows a basic queueing model for a delay/loss system. Customers arrive to the system according to some arrival pattern, e.g., Poisson process. The customer spends some time T in the system, e.g., the service time T is exponentially distributed. After this time, the customer departs the system. If the server has not enough resources, customers that arrive at the system when it is in this state are blocked or lost.



A simplified queueing model

Queueing System Variables (1/2)



Queueing System Variables (2/2)

Measure	Descriptions	Equation
Offered load (a)	Average offered load	$a = \lambda E(S)$
Carried load (a')	Average processed load	$a' = a(1 - P_B)$
Traffic intensity (ρ)	Average offered load per server	$\rho = a/c$
Server utilization (ρ')	Average process load per server	$\rho' = a'/c$
Queue size ($N_q(t)$)	Number of customers in the queue excluding serving customers	
Number in service ($N_s(t)$)	Number of serving customers	
Busy period	The time period that a server is busy	

where λ is the arrival rate, $E(S)$ is the average service time, P_B is the blocking probability, and c is the number of servers. If $c = 1$, $\rho = \lambda E(S)$.

Performance Measure

Measure	Descriptions	Equation
Throughput	Average number of customers/second that pass through the system	
Blocking probability (P_B)	Fraction of arriving customers/second that are lost or blocked	
System size ($N(t)$)	Number of customers in the system including serving customers	$N(t) = N_q(t) + N_s(t)$
System sojourn time (T)	Total time that a customer stays in the system until the customer departs the system	$T = W + S$, where S is service time
Waiting time (T_q)	Time that a customer waits until a server begins to service the customer	

Queueing System Characteristics (1/2)

Some common service disciplines are:

- First in, First out (FIFO): a customer that finds the service center busy goes to the end of the queue.
- Last in, First out (LIFO): a customer that finds the service center busy proceeds immediately to the head of the queue. She will be served next, given that no further customers arrive.
- Random Service: the customers in the queue are served in random order
- Round Robin: every customer gets a time slice. If her service is not completed, she 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. This scheme can use pre-emption or not.

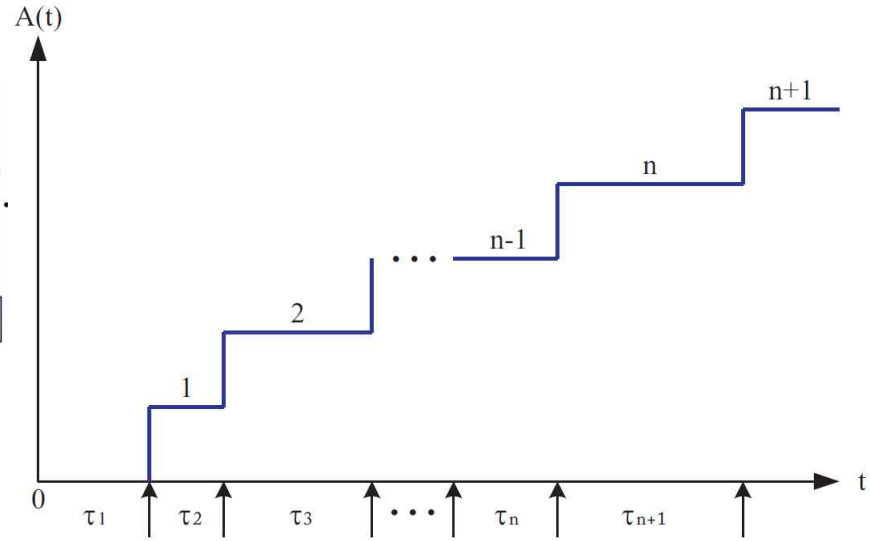
Queueing System Characteristics (2/2)

Characteristics	Descriptions
Customer classification	single class, multiple class, priority
Population types	finite population, infinite population
Arrival process	interarrival time, arrival rate
Service process	single service, batch service, bulk service
Maximum number	Maximum number of customers are allowed in the system
Number of servers	single server, multiple server, infinite number of servers
Service discipline	first come first service (FCFS), last come first service (LCFS), random selection for service (RSS), priority serive, short job first (SJF), shortest remaining processing time (SRPT)
Customer behavior	retrial, blocking
Queue structure	single-queue system, multi-queue system, queueing network

Notations and Definitions

Define

- $A(t)$: the number of arrivals at the system in $(0,t]$.
- $B(t)$: the number of blocked customers in $(0,t]$.
- $D(t)$: the number of customers departures in $(0,t]$



Time of nth arrival = $\tau_1 + \tau_2 + \ldots + \tau_n$
Arrival rate = 1/mean interarrival time

Description	Equations
The number of customers in the system at time t	$N(t) = A(t) - D(t) - B(t)$
Arrival rate	$\lambda = \lim_{t\to\infty} \frac{A(t)}{t}$ customers/second
Throughput	throughput = $\lim_{t\to\infty} \frac{D(t)}{t}$ customers/second
Average number of customers in the system	$E(N) = \lambda E(T)$
Blocking probability	$P_B = \lim_{t\to\infty} \frac{B(t)}{A(t)}$

2. Queueing System Classification

A special notation, called Kendall's notation, is used to describe a queueing system

$$A/B/c/K$$

where

- A: describes the interarrival time distribution
- B: the service time distribution
- c: the number of servers
- K: the size of the system capacity (including the servers).

The symbols traditionally used for A and B are

- M: exponential distribution (M stands for Markov)
- D: deterministic distribution
- G (or GI): general distribution.

Kendall's Notation

Arrival process/ Service time/ Servers/ Max occupancy

Interarrival times	Service times	Number of servers	K customers
M = exponential	M = exponential	1, c, infinite	Unspecified if unlimited
D = deterministic	D = deterministic		
G = general	G = general		
Arrival rate:	Service rate:		
$\lambda = 1/E(\tau)$	$\mu = 1/E(S)$		

Multiplexer models: M/M/1, M/M/1/K, M/D/1, M/G/1
Trunking models: M/M/c/c, M/G/c/c
User activity: M/M/∞ , M/G/∞

Arrival Process

- Define the type of arrival process
- Often it is thought that the interarrival times are independent (renewal process), whence the process is determined by the type of interarrival distribution.

Commonly used symbols are

- M exponential interarrival distribution (M = Markovian, memoryless); Poisson process
- D deterministic, constant interarrival times
- G general (unspecified)
- E_k Erlang-k distribution
- PH phase distribution
- Cox Cox distribution

Service Process

- Defines the distribution of the customer's service time
- The service time is affected by two factors
  - the required work requested by the customer (e.g. the size of a data packet to be sent, kB)
  - the service rate of the server (e.g. kB/s)
  - the service time is the ratio of these
- In Kendall's notation, the type of the service time distribution is indicated by substituting an appropriate symbol; commonly the same symbols (M, D, G, etc.) are being used as for defining the type of the interarrival time distribution

Example (1/2)

Example The queue $M/M/1$

- Poisson arrival process
- exponential service time distribution
- single server
- unlimited number of waiting places

Example The queue $M/M/c/c$

- Poisson arrival process
- exponential service time distribution
- c servers and c system places $\Rightarrow$ no waiting room, so called loss system

Example (2/2)

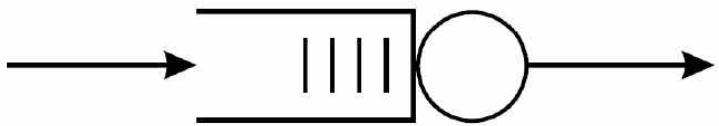
- Example

  We consider an uplink of a wireless communication system with a base station and K users. Each user generates data packets at a Poisson rate λ and the service times of the packets are exponentially distributed with a mean of $1/\mu$. The packets from users will be queued in the base station and the base station services one packet at once in accordance with a first-in-first-out (FIFO) policy. The queue size is assumed to be infinity. Let the state be the number of packets in the system.

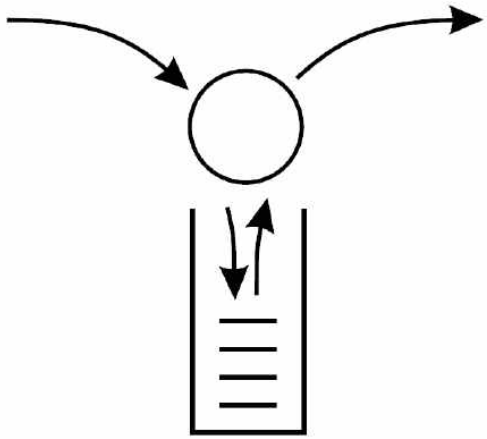
  → This can be modeled by using an $M/M/1$ with the arrival rate of $\Lambda = K\lambda$ and the service rate of μ.

Queueing Discipline/Scheduling (1/3)

- Ordinary queue, service in the order of arrivals
  - FCFS First Come First Served
  - FIFO First In First Out



- Stack, the latest arrival is being served first
  - LIFS Last Come First Served
  - LIFO Last In First Out

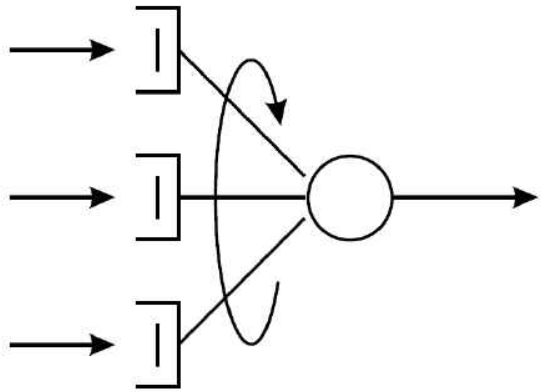


Queueing Discipline/Scheduling (2/3)

- There are three sub-cases of a stack
  - pre-emptive resume
    the arriving customer pre-empts the ongoing service, which is then resumed when the interrupted customer is again taken into the server, continuing from the same point on as at the time of interruption
  - pre-emptive restart
    the arriving customer pre-empts the ongoing service; the service is started from the beginning when the interrupted customer is again taken into the server
  - non-pre-emptive
    the arriving customer waits until the ongoing service is finished before being taken into the server

Queueing Discipline/Scheduling (3/3)

- Service in rotating order
  RR Round robin
  - each customer receives, in turn, a small time slice of service
  - polling



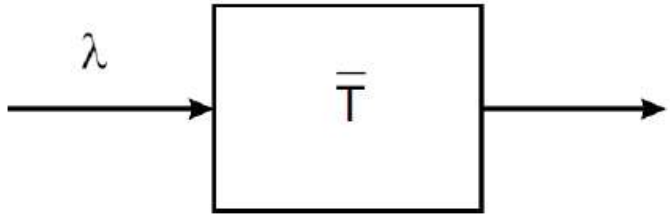
- Other service disciplines are e.g.
  - SIRO (Service In Random Order)
  - SSF (Shortest Jobs First): the service time has to be known in advance; this minimizes the mean waiting time

3. Little's Formula

Little's result or Littles theorem is a very simple (but fundamental) relation between the arrival rate of customers, average number of customers in the system and the mean sojourn time of customers in the system.

Result (Little's formula) Consider any system as a "black box". Let $L = E(N)$ be the average number of customers in the system and let $W = E(T)$ be the average sojourn time of a customer. Then,

$$E(N) = \lambda E(T)$$
$$L = \lambda W.$$



Now consider a system in which customers can be blocked. Then, since the actual arrival rate into a system with blocking is $\lambda(1 - P_B)$, Little's formula for a system with blocking is

$$E(N) = \lambda(1 - P_B)E(T).$$

Traffic Intensity

- Traffic intensity (load)

Consider some elements in a network, e.g. trunks, ports or any logical units such that each customer in the system reserves one such element. Let

$$\begin{cases} \lambda = \text{the arrival rate of customers into the system} \\ \overline{T} = \text{average holding time of an element.} \end{cases}$$

The quantity $a = \lambda \overline{T}$ is called the (offered) *traffic intensity* or *load* of the traffic.

- Traffic intensity is a pure number, but in order to emphasize the context one often denotes as its "unit" erlang or erl (as a tribute to the Danish pioneer of traffic theory, A.K. Erlang).

By Little's result, the traffic intensity is the same as the average number of simultaneously reserved elements.

Example

Example In the trunk group from a PBX (private branch exchange) to the central office there are on the average 150 calls per hour. The mean holding time of a call is on 3 min. The traffic intensity is

$$a = 150h^{-1} \times 3\text{min} = 7.5\text{erl}$$

In other words, in the trunk group there are on the average 7.5 calls in progress simultaneously. □

Summary

- Queueing System Performance Measure
  - Throughput, Blocking probability, System size, System sojourn time, Waiting time

- Queueing System Classification
  - Kendall's notation: A/B/c/K
    - Arrival process/Service time/Servers/Max occupancy

- Little's Formula
  - Relation between the arrival rate of customers, average number of customers in the system and the mean sojourn time of customers in the system

$$E(N) = \lambda E(T)$$

$$L = \lambda W.$$