UMBC

AN HONORS UNIVERSITY IN MARYLAND

IS 709/809: Computational Methods in IS Research

Queueing Theory Introduction

Nirmalya Roy

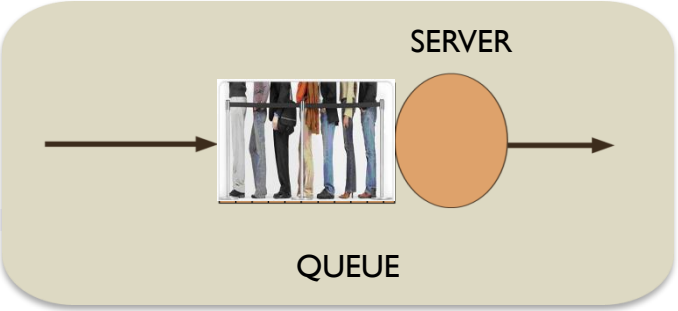
Department of Information Systems

University of Maryland Baltimore County

Introduction: Statistics of things Waiting in Lines

- Wait in line in our cars in traffic jam or at toll booths
- Wait on hold for an operator to pick up telephone calls
- Wait in line at supermarkets to check out
- Wait in line at fast food restaurants
- Wait in line at bank and post offices
- ……..
- More demand for service than there is facility for service available
  - Shortage of available servers; economically infeasible; space limit

Introduction



- Limitations can be removed partially
  - Need to know how much service should then be made available
  - How long must a customer wait? And how many people will form in the line?
- Queueing theory attempt to answer those questions
  - Detailed mathematical analysis
- Probability models to characterize empirical processes
  - mathematical analysis to calculate performance measures
  - probability distributions to model router's interarrival and service times
    - help to determine average queueing delay/queue length

Characteristics of Queueing Process

- Arrival patterns of customers
- Service patterns of servers
- Queue disciplines
- System capacity
- Number of service channel
- Number of service stages

Arrival patterns of customers

- Process of arrivals is stochastic
- Necessary to know the probability distribution describing the times between successive customer arrivals
  - Interarrival times
  - Batch or Bulk arrivals
- Reaction of a customer upon entering the system
  - Irrelevant to the waiting time
  - Decides not to enter the queue upon arrival --- *balked*
  - Enter the queue, but after some time loose patience and decide to leave --- *reneged*

Arrival Patterns

- Arrival pattern changes or does not change with time
  - An arrival pattern that does not change with time
    - Probability distribution describing the input process is time-independent
      - A stationary arrival pattern
  - Not time-independent
    - A non-stationary arrival pattern

Service Patterns

- A probability distribution is needed to describe the sequence of customer service times
- Service may be single or batch
  - One customer being served at a time by a given server
  - Customer may be served simultaneously by the same server
    - A computer with parallel processing
    - Sightseers on a guided tour
    - People boarding a train

Queue Discipline

- The manner in which customers are selected for service
  - First come first served (FCFS)
  - Last come first served (LCFS)
    - Stored items in a warehouse, easier to reach the nearest items
- A variety of priority schemes
  - Higher priorities will be selected for service ahead of those with lower priorities
  - *Preemptive*
    - Higher priority is allowed to enter service immediately
  - *Non-Preemptive*
    - Highest priority customer goes to the head of the queue but cannot get into service until the customer presently in service is completed

System Capacity

- Physical limitation to the amount of waiting room
  - Line reaches a certain length, no further customers are allowed to enter until space becomes available
  - Finite queueing situations
  - Limited waiting room can be viewed as one with force balking

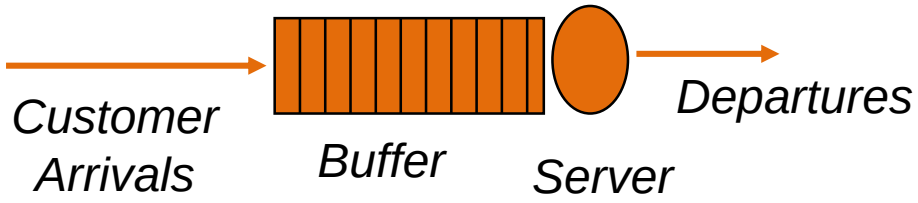
Number of Service Channels

- Design multi-server queueing systems to be fed by a single line
- Specify the number of service channels
  - Typically refer to the number of parallel service stations

Single server queue

Single server

(arrivals when server is full is queued)



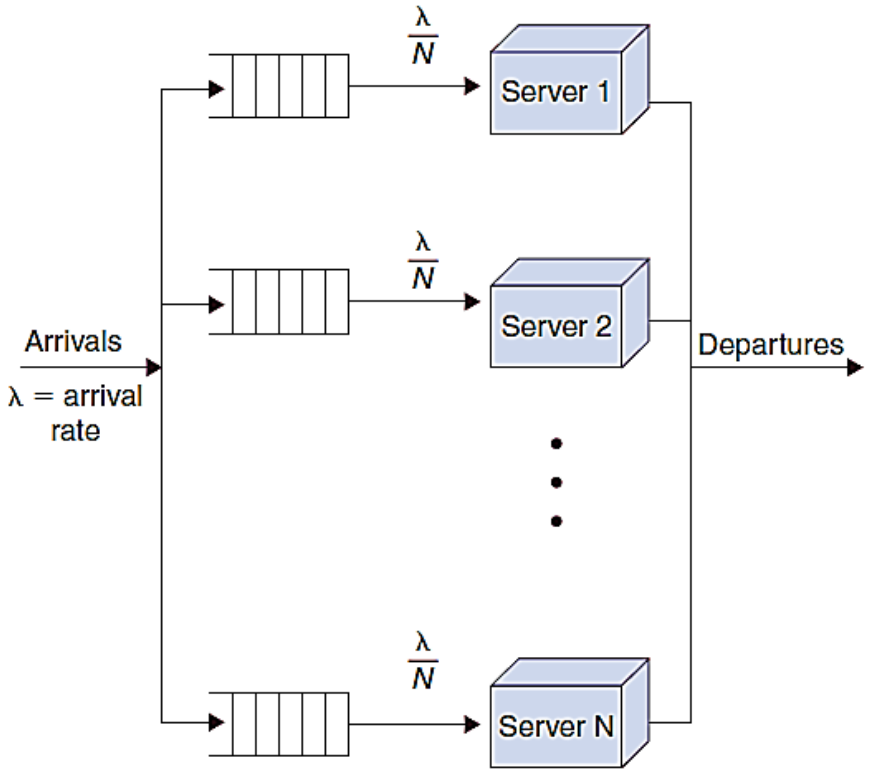
Example:

Small Grocery Store

Multiple single-server queue

Multiple single-server queues

(A queue for each channel)



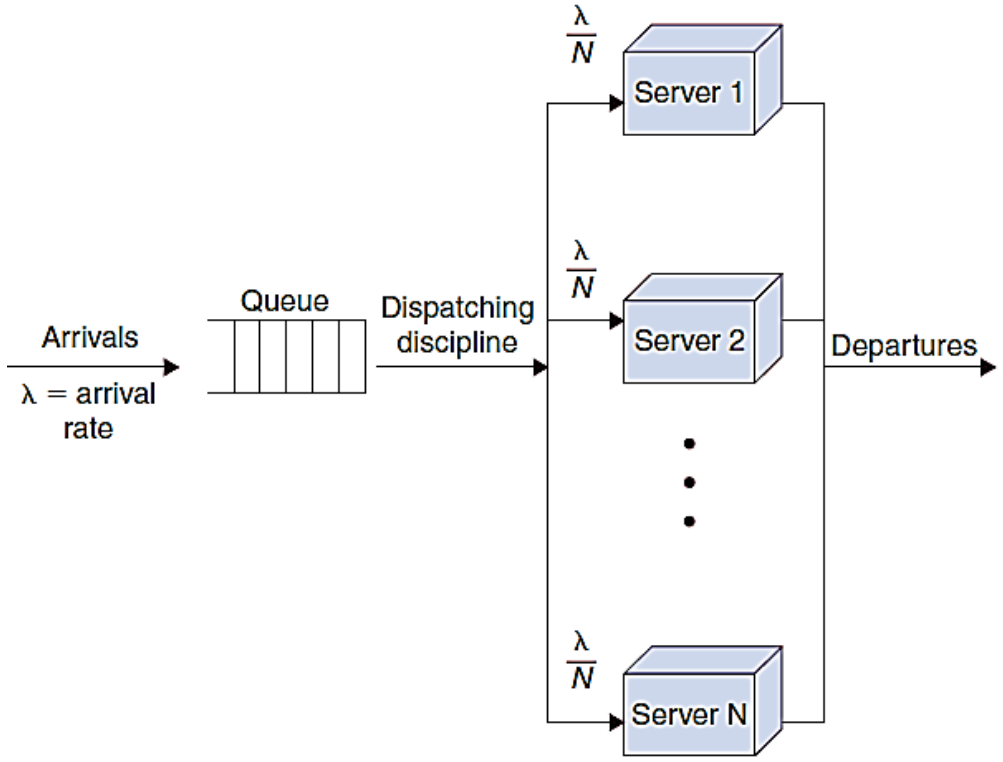
Example:

Large supermarket with multiple cashiers

Multiple single-server queue

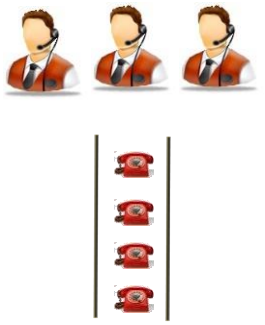
Multiple server

(single queue feeds into multiple servers)



Example:
Bank Teller

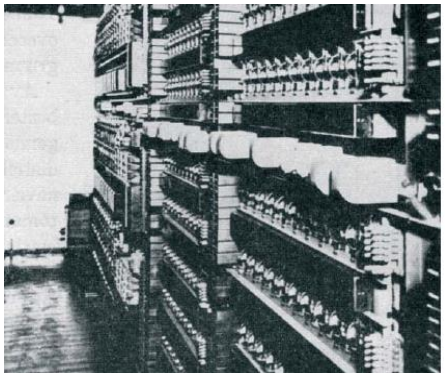
Example:
Call Center



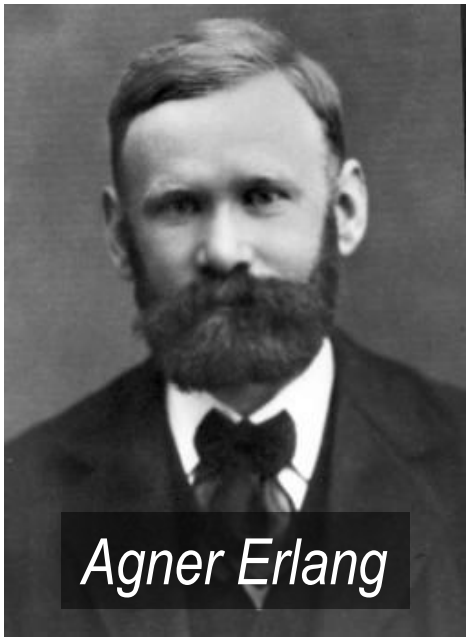
Stages of Service

- Single stage of service
  - Hair styling salon
- Multistage of service
  - Physical examination procedure
  - Recycling or feedback may occur
    - Manufacturing process, quality control
  - Telecommunication network may process messages through a randomly selected sequence of nodes
    - Some messages may require rerouting on occasion through the same stage

History of Queueing Theory



How many trunk lines are required to provide service to a town?



Father of the field of queueing theory and teletraffic engineering

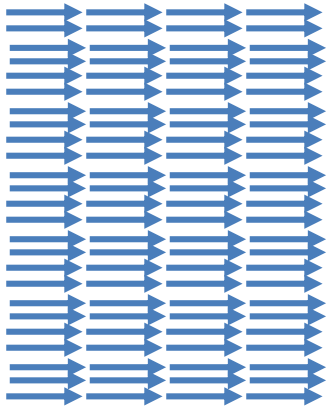
Telecommunication Network Design (TND)

Imagine a small village having a population of 100 telephone users;
how many 'trunk lines' are needed to connect this village's telephone exchange to a long-distance telephone exchange?

Local exchange

Single trunk line

long waits and blocked calls



A trunk line for every user

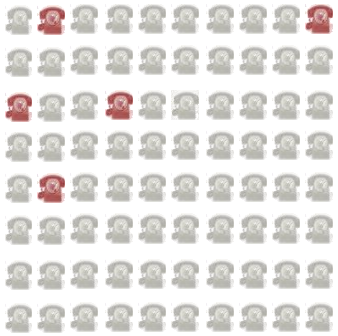
economically inefficient

Trunk line: *a circuit connecting telephone switchboards or other switching equipment.*

Solution

It will be highly unlikely that all 100 users will want to use their service all the time

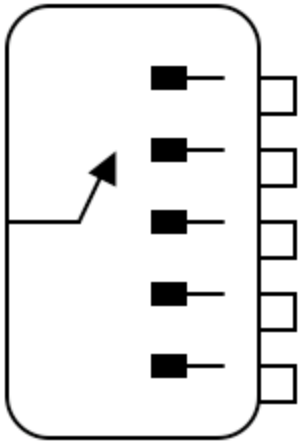
What is the average demand like?



*Assume that **on average** there are **5 calls/ hr** in the busiest hour of the day and the average call lasts for an hour each.*

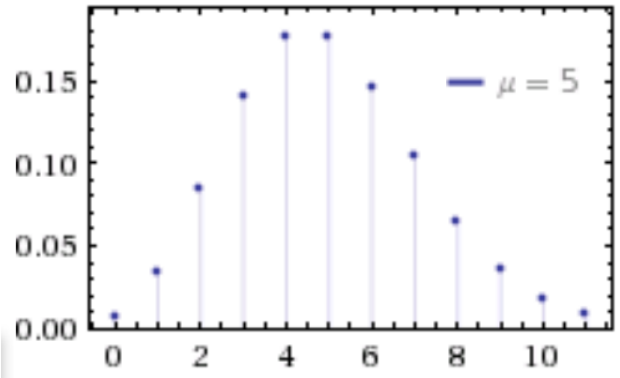
Solution

Now, that we know there are 5 calls/hr on average in the busy-hour-time, we should just put 5 trunk lines. ***RIGHT?***



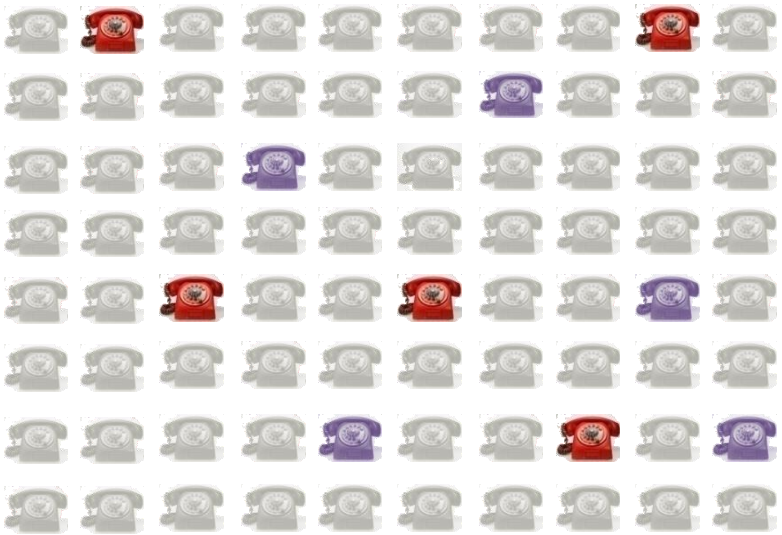
WRONG!
Calls bunch up!

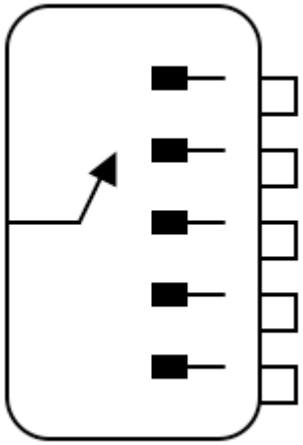
The number of calls in a hour is a random variable and follows a Poisson distribution with expected value of 5 calls/hour



Solution

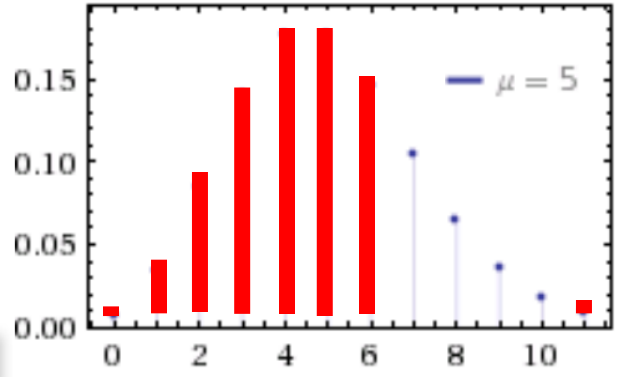
Now, that we know there are 5 calls/ hr on average in the busy-hour-time, we should just put 5 trunk lines. ***RIGHT?***





WRONG!
Calls bunch up!

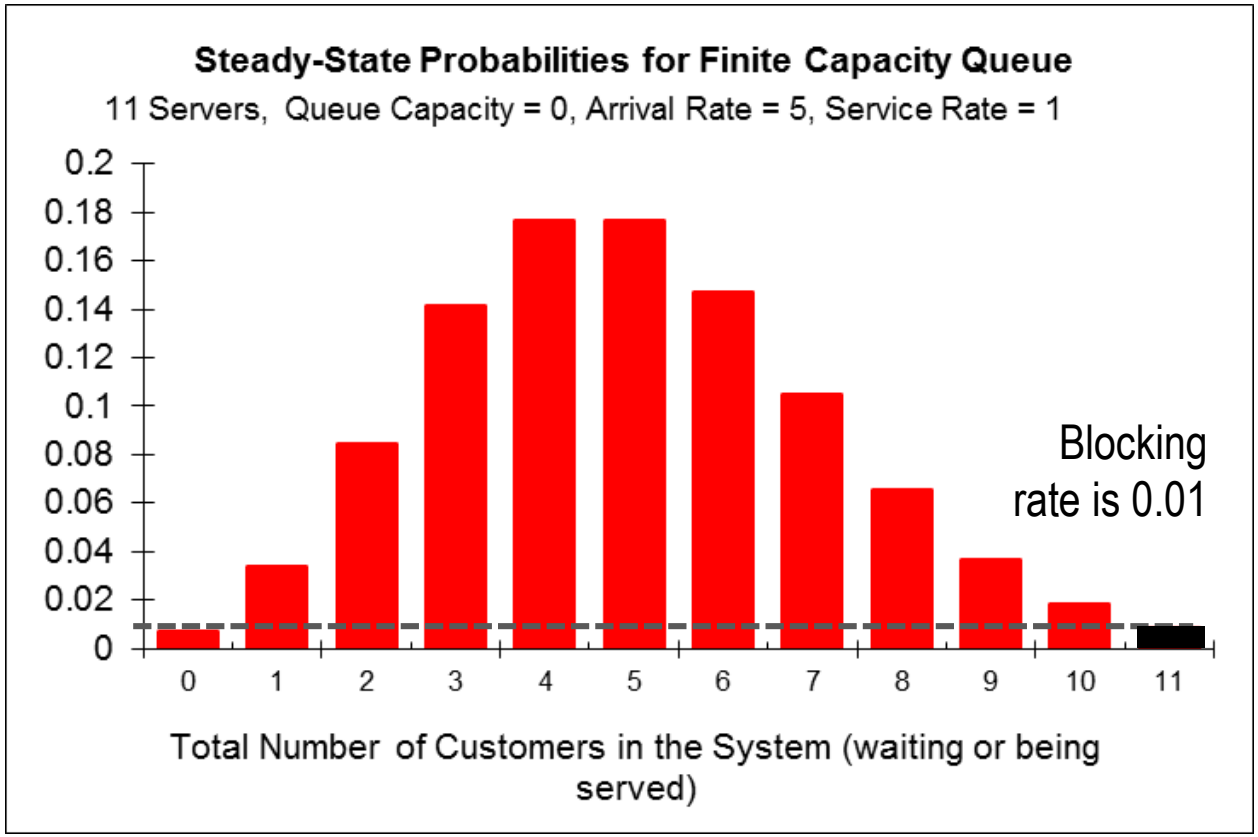
The number of calls in a hour is a random variable and follows a Poisson distribution with expected value of 5 calls/ hour



Erlang B Table

Ch	1%						
7	2.50						
8	3.12						
9	3.78						
10	4.46						
11	5.16						
12	5.87						
13	6.60						
14	7.35						
15	8.10						
16	8.87						
17	9.65						
18	10.43						
19	11.23						
20	12.03						
21	12.83						
22	13.65						
23	14.47						
24	15.29						
25	16.12						
26	16.95	18.38	20.94	48	36.11	38.39	42.54
27	17.97	19.26	21.90	49	37.00	39.32	43.53
28	18.64	20.15	22.86	50	37.90	40.25	44.53

11

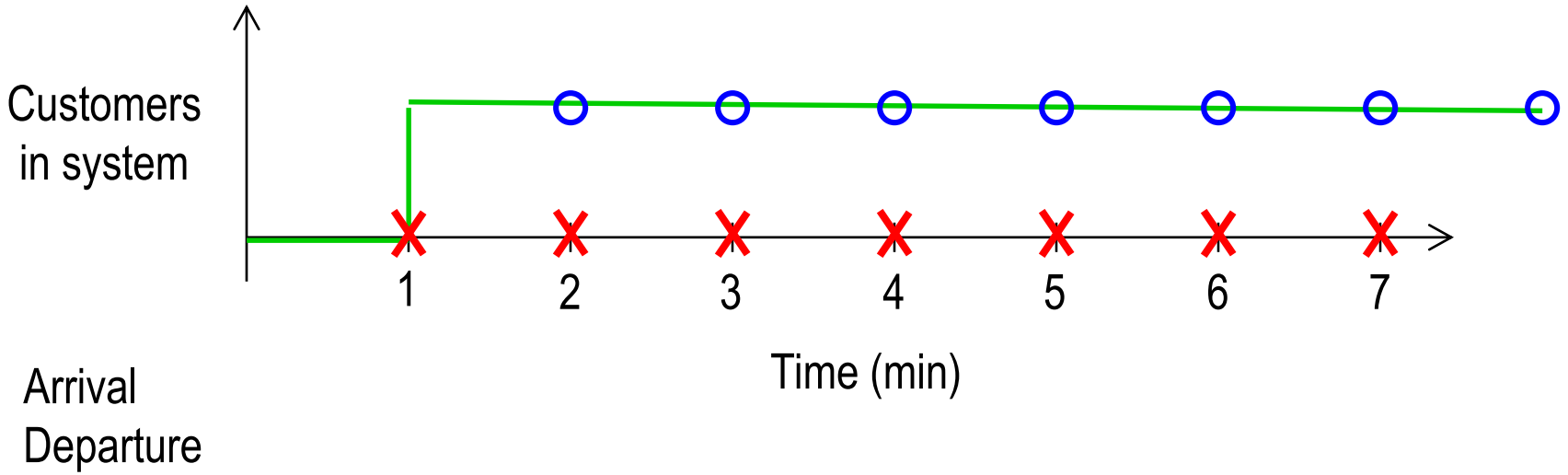


To carry 5 erlangs of traffic (5 calls/ hr with average call duration of 1 hour) with blocking probability of 0.01 only, 11 trunk lines are needed

A Simple Deterministic Queue

Interarrival time is exactly 1 minute and

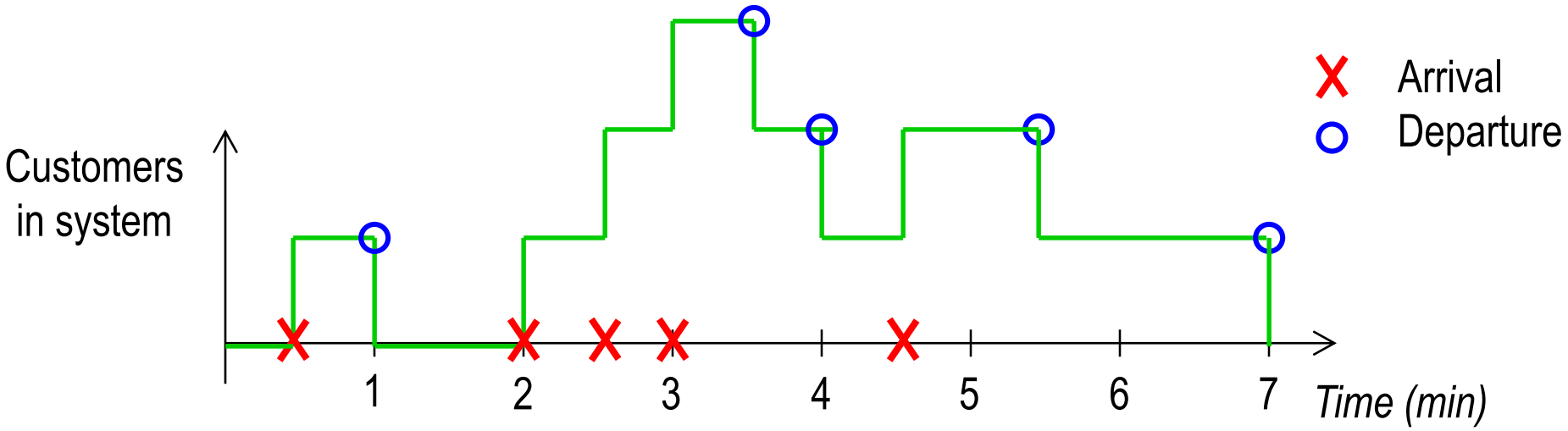
Service times is also exactly 1 minute.



A Stochastic Queue

Interarrival time is .5 minute or 1.5 minute with equal probability
expected interarrival time = 1 minute

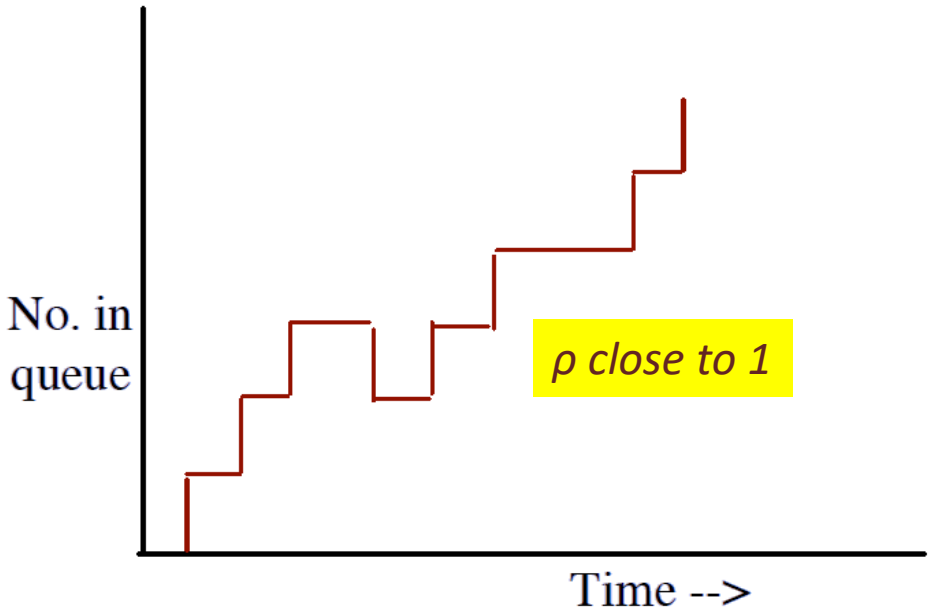
Service time is .5 minute or 1.5 minute with equal probability
expected service time = 1 minute



This queue is unstable since the interarrival time and service time are equal
(Due to stochastic nature of arrivals, arrivals would bunch up)

Stability of Queue

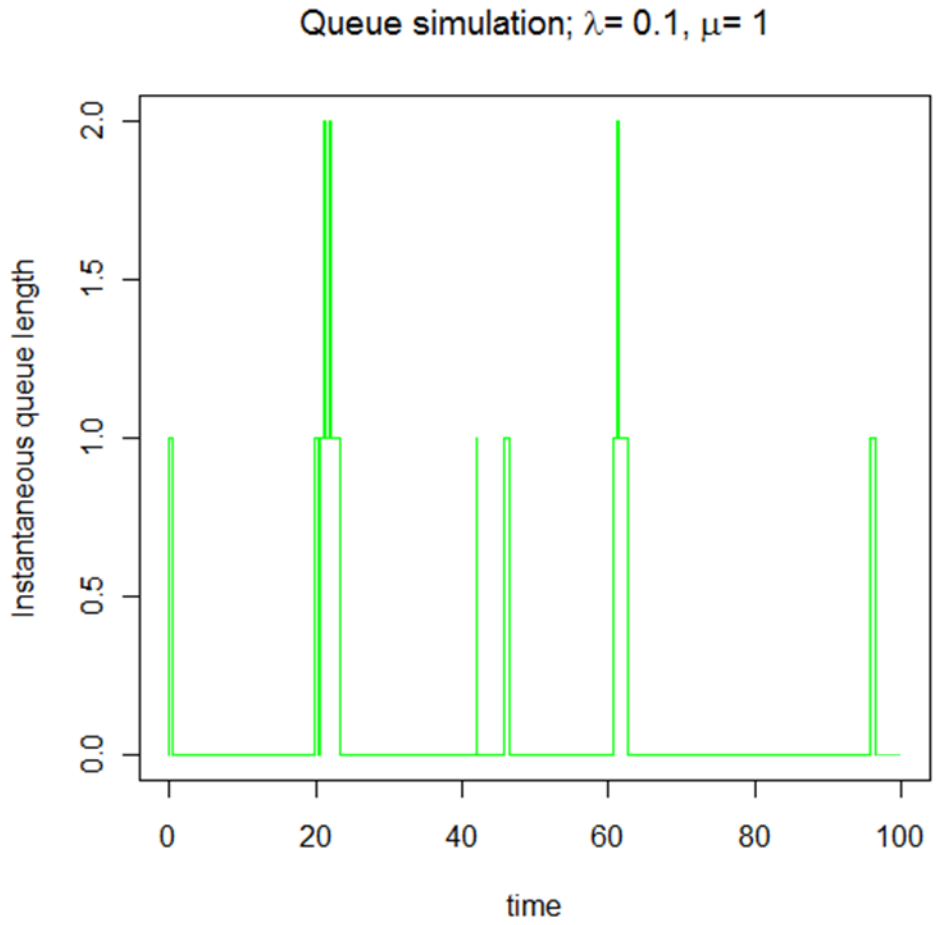
For a stable queue, the average service rate (μ) must be more than the arrival rate (λ); the traffic intensity ρ (= λ/ μ) < 1



Lightly loaded queue

Poisson arrival process (rate λ = 0.1 arrivals/ minute);

Exponential service (rate μ = 1 departure/ minute)

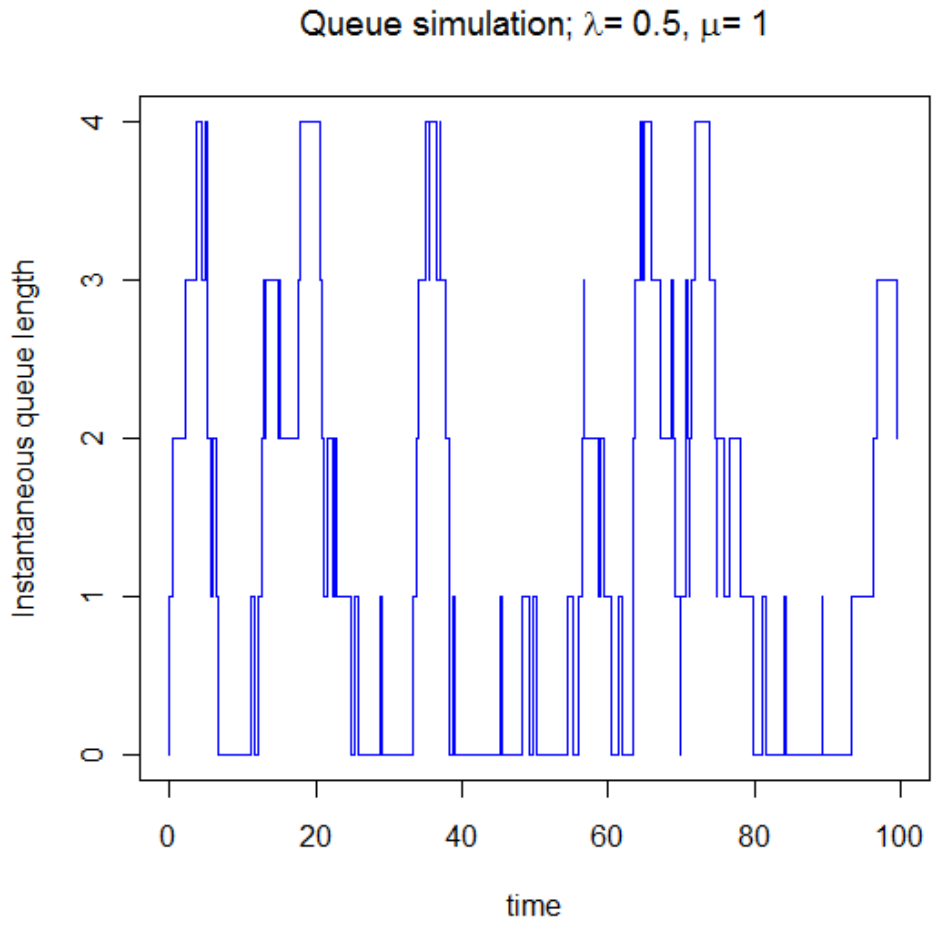


$\rho = \lambda/\mu = 0.1$

Moderately loaded queue

Poisson arrival process (rate =λ = 0.5 arrivals/ minute);

Exponential service (rate= μ = 1 departure/ minute)

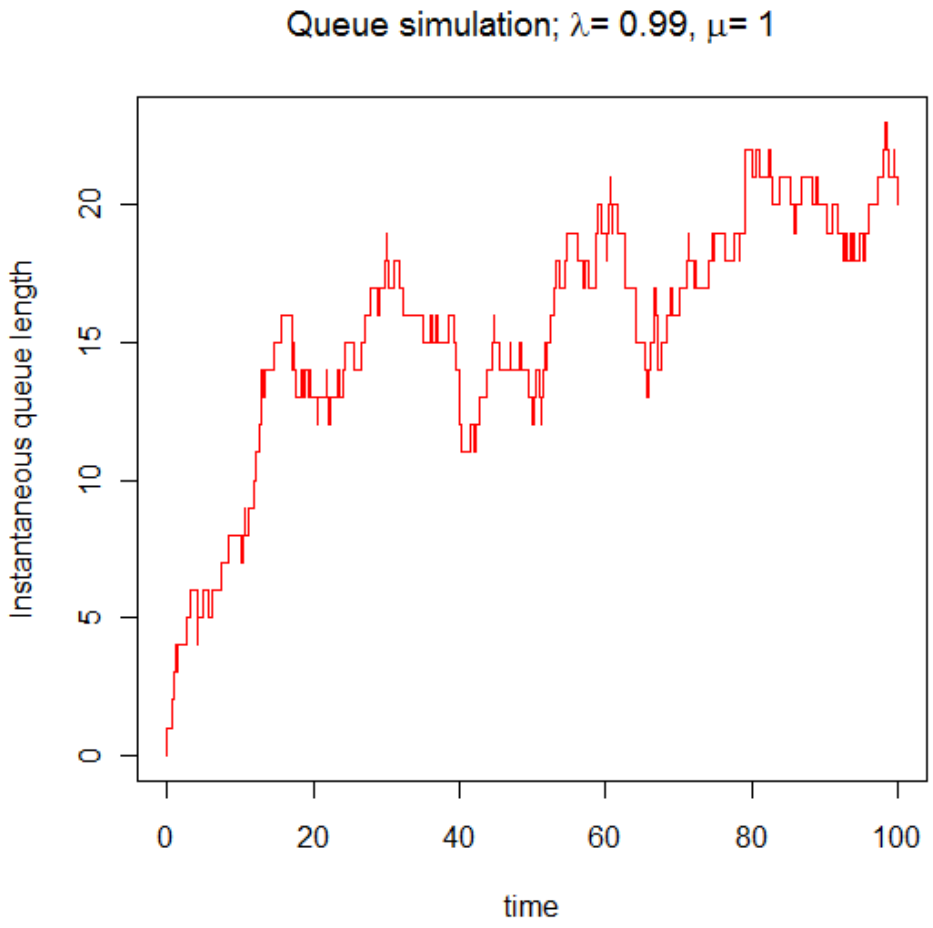


$\rho = \lambda/\mu = 0.5$

Heavily loaded queue

Poisson arrival process (rate =λ = 0.99 arrivals/ minute);

Exponential service (rate= μ = 1 departure/ minute)

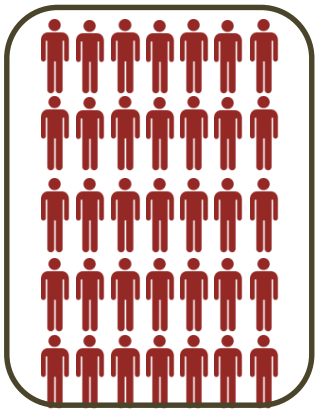


$\rho = \lambda/\mu = 0.99$

Notation of a Queueing System

A/S/m/B/K/SD (Kendall's notation)

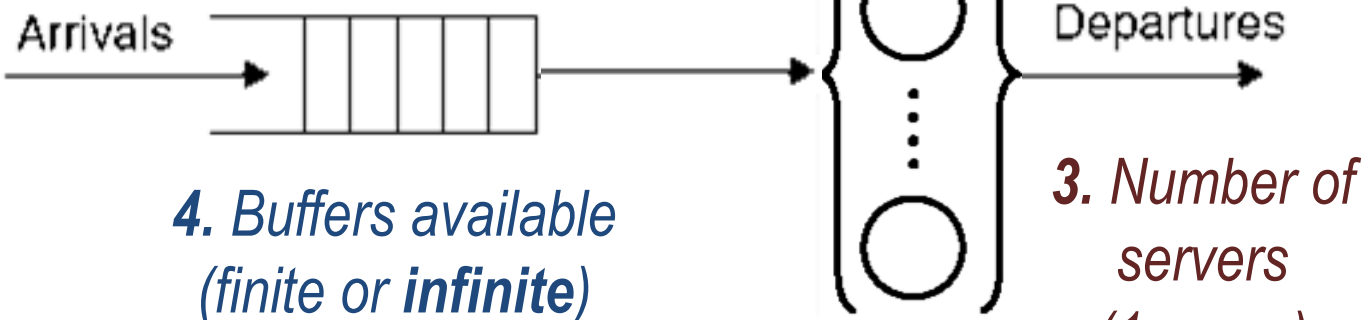
5. *Customer Population (finite or **infinite**)*



6. *Service discipline (e.g., **FIFO**)*

2. *Service time distribution (M,D,G)*

1. *Arrival process distribution (M,D,G)*



4. *Buffers available (finite or **infinite**)*

3. *Number of servers (1, m, ∞)*

Interarrival time and service time is assumed to be IID, therefore, only the family of distributions needs to be specified

Kendall's notation

- *M/G/1: Poisson arrivals; General service distributions; 1 server* *(Infinite buffer, population; FCFS)*
- *G/G/1: General arrival and service distributions; 1 server* *(Infinite buffer, population; FCFS)*
- *M/D/2/∞/FCFS: Poisson arrivals; deterministic service time; two parallel servers; no restrictions on the maximum # allowed in the system; and FCFS queue disciplines*

Measuring System Performance

- Effectiveness of a queueing system
- Generally 3 types of system responses of interest
  - Waiting time
  - Customer accumulation manner
  - A measure of the idle time of the server
- Queueing systems have stochastic elements
  - Measures are often random variables and their probability distributions

Measuring System Performance

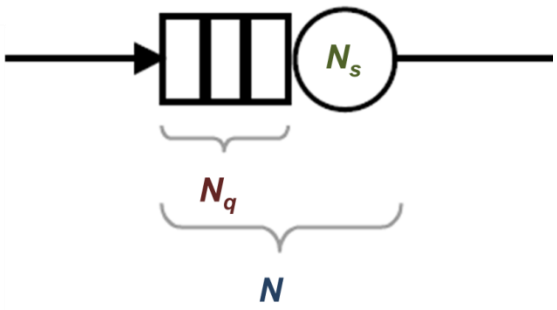
- Two types of customer waiting times
  - Time a customer spends in the queue
    - amusement park
  - Total time a customer spends in the system (queue + service)
    - Machines that need repairs
- Customer accumulation measures
  - Number of customers in the queue
  - Total number of customers in the system
- Idle-service measures of a server
  - Time the entire system is devoid of customers

Little's Theorem

- One of the most powerful relationship in queueing theory
  - Developed by John D. C. Little in early 1960s
  - Related the steady-state mean system sizes to the steady state average customer waiting times
- T_q = time a customer spends waiting in the queue
- T = total time a customer spends in the system= response time
- $T = T_q + S$; where S is the service time and T, T_q, and S all are random variables
- Two often used measures of system performance
  - Mean waiting time in queue; $W_q = E[T_q]$
  - Mean waiting time in the system; $W = E[T]$

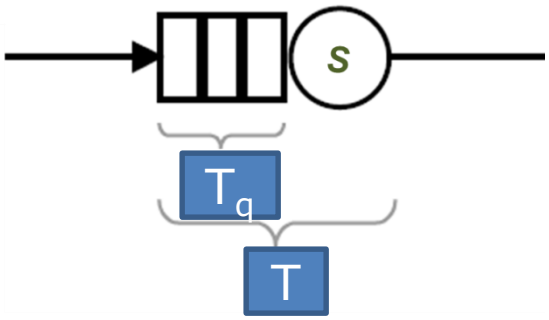
Little's Theorem

Mean number of customers (backlog)



Number of customers in queue: N_q
Number of customers obtaining service: N_s
Number of customers in system $N = (N_q + N_s)$

Average waiting time of customers (delay)



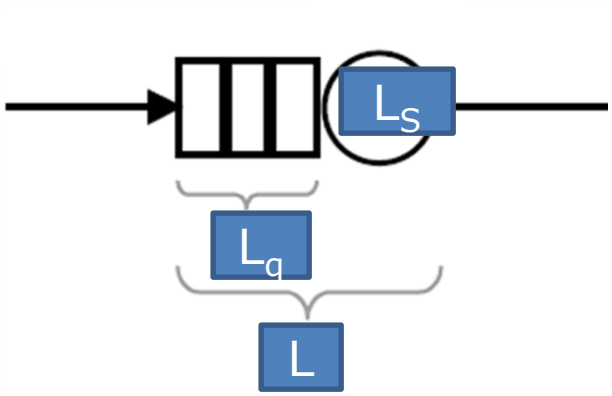
Waiting time for customers in queue: T_q
Service time for customers: S
Total time (response time) of customers in system T
$= (T_q + S)$

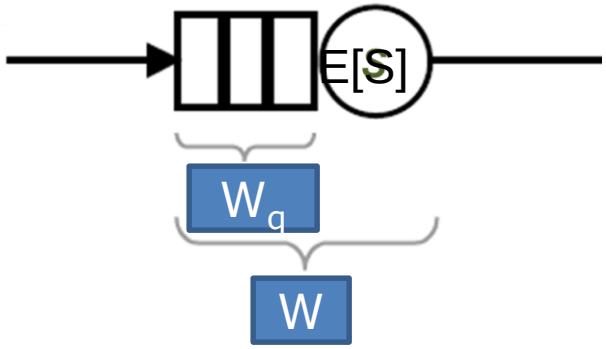
All the above metrics are RVs, and therefore we will calculate their expectations

Little's Law

Little's Law related the two primary performance measures of any queue

Mean number of customers $\propto$ *Average waiting time of customers*

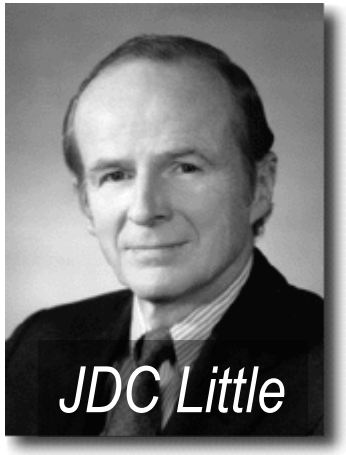


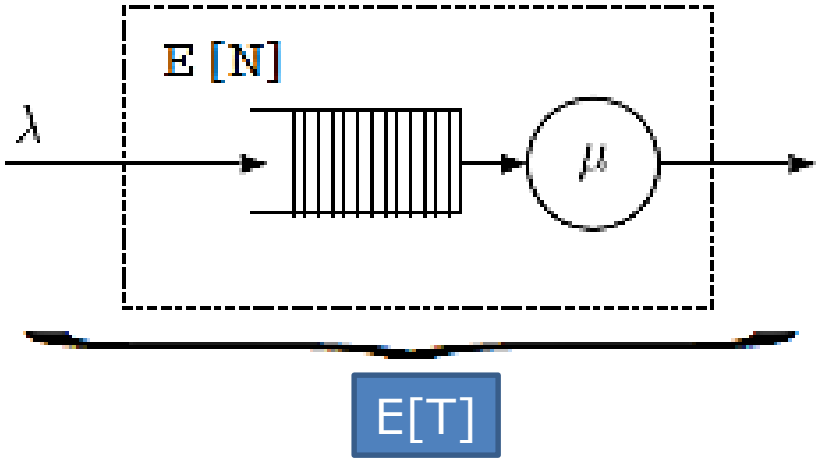


Little's Theorem

***L**ength = **A**rrival-rate x **W**ait-time*

Little's law states that time-average of queue length is equal to the product of the arrival rate and the customer-average waiting time (response time)





$$E[N] = \lambda.E[T]$$

$$\Rightarrow L = \lambda.W$$

Little's Theorem

- Little's formulas are:
  - $L = \lambda W$
  - $L_q = \lambda W_q$
- L_q = mean number of customer in the queue
- L = mean number of customer in the system
- λ = arrival rate
- $E[T] = E[T_q] + E[S] \Rightarrow W = W_q + 1/\mu$

Little's Theorem Proof

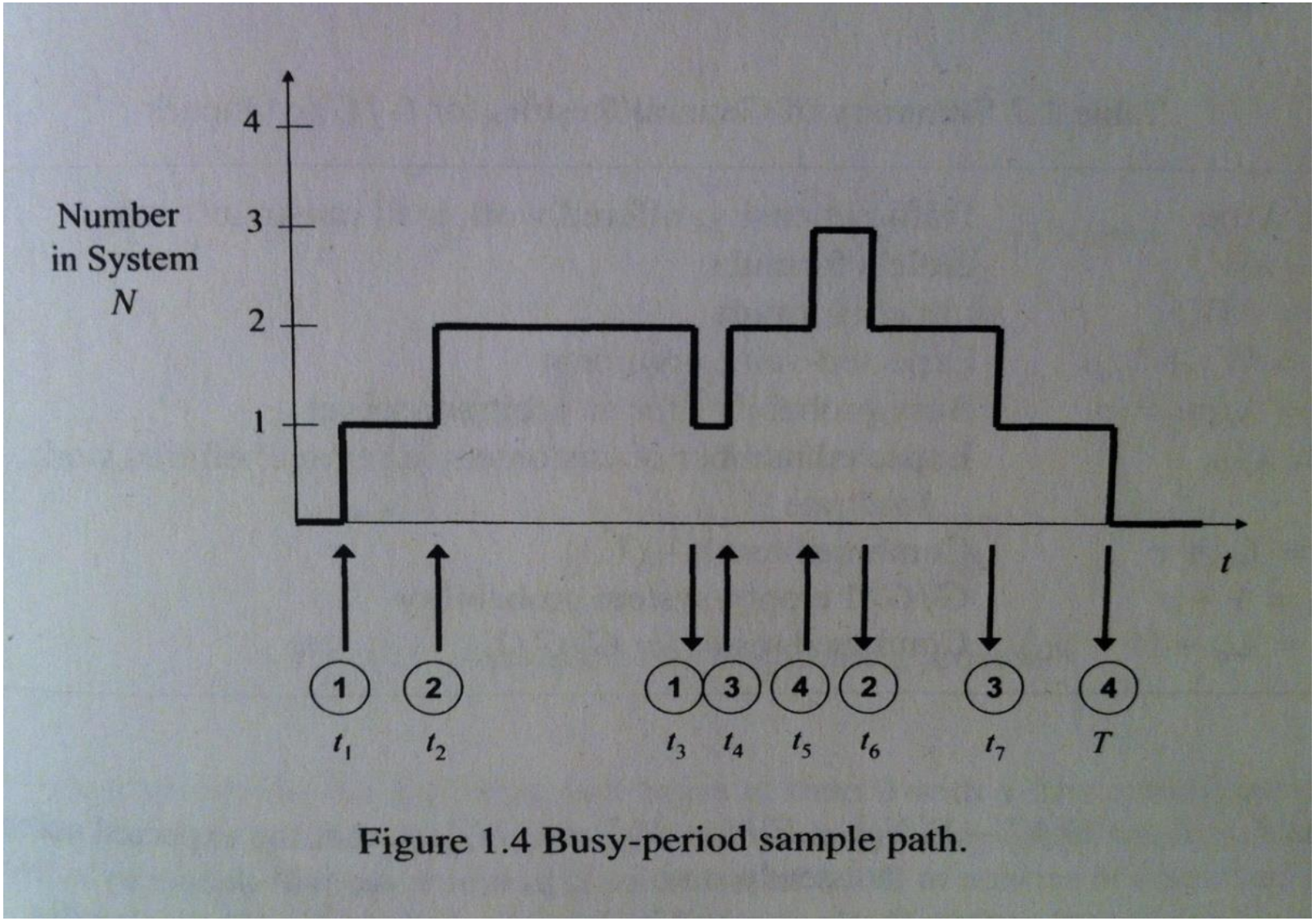


Figure 1.4 Busy-period sample path.

Little's Theorem Proof

- Number of customers (say N) arrive over the time period (0,T) is 4
- Find out the value of L and W from the graph

$$L = [1(t_2 - t_1) + 2(t_3 - t_2) + 1(t_4 - t_3) + 2(t_5 - t_4) + 3(t_6 - t_5) + 2(t_7 - t_6) + 1(T - t_7)] / T$$
$$= (\text{area under curve}) / T$$
$$= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1) / T$$

$$W = [(t_3 - t_1) + (t_6 - t_2) + (t_7 - t_4) + (T - t_5)] / 4$$
$$= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1) / 4$$
$$= (\text{area under curve}) / N$$

Little's Theorem Proof

$$LT = WN \Rightarrow L = \frac{N}{T}W \Rightarrow L = \lambda W$$

where fraction (N/T) is the number of customers arriving over the time T and which is for this period, the arrival rate λ

$$L - L_q = \lambda(W - W_q) = \lambda(1/\mu) = \frac{\lambda}{\mu}$$

$$L - L_q = E[N] - E[N_q] = E[N - N_q] = E[N_s] = \frac{\lambda}{\mu}$$

So expected no. of customers in service in the steady state is λ/μ, also denoted by r. For a single server system r=ρ

Summary of General Results for G/G/c queues

$\rho = \frac{\lambda}{c\mu}$ Traffic intensity; offered work load rate to a server

$L = \lambda W$ Little's formula

$L_q = \lambda W_q$ Little's formula

$W = W_q + 1/\mu$ Expected - value argument

$p_b = \lambda/c\mu = \rho$ Busy probability for an arbitrary server

$r = \frac{\lambda}{\mu}$ offered work load rate

$L = L_q + r$

$p_0 = 1 - \rho$ G/G/1 empty system probability

$L = L_q + (1 - p_0)$

Example: Little's Law

Assume that you receive 50 emails every day and that you archive your messages after responding to your email. The number of un-responded emails in your inbox varies between ~100 to ~200 and its average value is 150.

How long do you take to answer your emails?

$$L = 150 \text{ emails}$$

$$\lambda = 50 \text{ emails/ day}$$

$$W = L/\lambda = 3 \text{ days}$$

Example: Little's Law

Suppose that 10,800 HTTP requests arrive to a web server over the course of the busiest hour of the day. If we want to limit the mean waiting time for service to be under 6 seconds, what should be the largest permissible queue length?

λ=10,800 requests/ hour

λ= 3 requests/ second

W=6 seconds

Since $L_q = \lambda W_q$

=> L_q= 18

Poisson Process & Exponential Distribution

- Stochastic queueing model
  - Assume interarrival times and service times obey the Exponential distribution
  - Equivalently arrival rate and service rate follow a Poisson distribution
- First derive the Poisson distribution
  - See whiteboard
- Show that assuming no. of occurrences in some time interval to be a Poisson random variable is equivalent to assuming time between successive occurrences to be an exponentially distributed random variable

Derivation of Poisson Distribution

- Consider an arrival counting process {N(t), t ≥ 0}
  - N(t) denotes the total no. of arrivals up to time t
- Assumption:
  - Probability that an arrival occurs between time t and time (t + Δt) = λΔt + o(Δt)
    - Pr{an arrival occurs between t and t + Δt} = λΔt + o(Δt)
      - where λ is a constant independent of N(t)
      - Δt is an incremental element
      - o(Δt) denotes a negligible quantity when as Δt → 0; $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$
  - Pr{more than one arrival between t and t + Δt} = o(Δt)
  - Number of arrivals in non-overlapping intervals are statistically independent

Derivation of Poisson Distribution

- Calculate $p_n(t)$ = probability of n arrivals in a time interval of length t, where $n \geq 0$
- Develop *differential – difference* equations for the arrival process
- $p_n(t + \Delta t)$ = Pr{*n arrivals in t and none in Δt*} +
  Pr{*n-1 arrivals in t and 1 in Δt*} +
  Pr{*n-2 arrivals in t and 2 in Δt*} + ……..
  + Pr{*no arrivals in t and n in Δt*} .

$$p_n(t + \Delta t) = p_n(t)[1 - \lambda\Delta t - o(\Delta t)] + p_{n-1}(t)[\lambda\Delta t + o(\Delta t)] + o(\Delta t)$$

where the last term $o(\Delta t)$ represents the terms

Pr{n -j arrivals in t and j in Δt; $2 \leq j \leq n$ }

Derivation of Poisson Distribution

- $p_n(t + \Delta t) = p_n(t)[1 - \lambda\Delta t - o(\Delta t)] + p_{n-1}(t)[\lambda\Delta t + o(\Delta t)] + o(\Delta t)$
  - For n = 0; $p_0(t + \Delta t) = p_0(t)[1 - \lambda\Delta t - o(\Delta t)]$
  - Rewriting;

$p_0(t + \Delta t) - p_0(t) = -\lambda\Delta t p_0(t) + o(\Delta t)$ and

$p_n(t + \Delta t) - p_n(t) = -\lambda\Delta t p_n(t) + \lambda\Delta t p_{n-1}(t) + o(\Delta t)$ for $(n \geq 1)$

  - Divide by Δt and take the limit as $\Delta t \rightarrow 0$ to obtain the *differential-difference* equations

$$\lim_{\Delta t \to 0}\left[\frac{p_0(t+\Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t}\right]$$

$$\lim_{\Delta t \to 0}\left[\frac{p_n(t+\Delta t) - p_n(t)}{\Delta t} = -\lambda p_n(t) + \lambda p_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}\right] \quad (n \geq 1)$$

Derivation of Poisson Distribution

- Reduces to

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t)$$

$$\frac{dp_n(t)}{dt} = -\lambda p_n(t) + \lambda p_{n-1}(t) \quad (\mathrm{n} \geq 1)$$

- We have an infinite set of linear, first-order ordinary differential equations to solve

Derivation of Poisson Distribution

$p_0(t) = Ce^{-\lambda t}$ where C = 1 since $p_0(0) = 1$

For n =1; $\frac{dp_1(t)}{dt} + \lambda p_1(t) = \lambda p_0(t) = \lambda e^{-\lambda t}$

The solution to this equation is: $p_1(t) = Ce^{-\lambda t} + \lambda t e^{-\lambda t}$

Using the boundary condition $p_n(0) = 0$ for all n > 0 yields C=0

$$p_1(t) = \lambda t e^{-\lambda t}$$

Therefore; $p_2(t) = \frac{(\lambda t)^2}{2} e^{-\lambda t}$, $p_3(t) = \frac{(\lambda t)^3}{3!} e^{-\lambda t}$

Derivation of Poisson Distribution

- General formula is

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

which is a well known formula for a Poisson probability distribution with mean λt

- The random variable defined as the number of arrivals to a queueing system by time t
  - this random variable has a Poisson distribution with a mean of λt arrivals or with a mean arrival rate of λ.

Markovian Property of Exponential Distribution

- Markov process is characterized by its unique property of memoryless
  - the future states of the process are independent of its past history and depends solely on its present state
- Poisson processes constitute a special class of Markov processes
  - event occurring patterns follow the Poisson distribution
  - the inter-arrival times and service times follow the exponential distribution
- Exponential distribution is the continuous distribution that possesses the unique property of memoryless-ness

Markovian Property of Exponential Distribution

- Recall the conditional probability law that

$$P(A \mid B) = P(A \cap B) \Big/ P(B)$$

given the event B, the probability of the event A is equal to the joint probability of A and B divided by the probability of the event B

- Let T be the variable representing the random inter-arrival time between two successive arrivals at two time points
  - we have the following probabilities for the two mutually-exclusive events

Memorylessness

$$p_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

- No arrival yet for a period of t seconds: $P(T \geq t) = e^{-\lambda t}$
- Having an arrival within a period of t seconds: $P(T \leq t) = 1 - e^{-\lambda t}$
- Prove that $P(T \leq t + \Delta t \mid T \geq t) = P(0 \leq T \leq \Delta t)$
  - left-hand-side represents the probability of having an arrival by waiting Δt seconds longer under the condition that no arrival has occurred during the past waiting period of t seconds with t ≥ 0
  - right-hand-side represents the probability of having an arrival if waiting for another Δt seconds
- The equation states that the probability of having an arrival during the next Δt seconds is independent of when the last arrival occurred

Memorylessness

- Prove that $P(T \le t + \Delta t \mid T \ge t) = \dfrac{P[(T \le t + \Delta t) \cap (T \ge t)]}{P(T \ge t)}$

$$= \frac{e^{-\lambda t} - e^{-\lambda (t+\Delta t)}}{e^{-\lambda t}} = 1 - e^{-\lambda \Delta t} = P(0 \le T \le \Delta t)$$

- Consider there has been no arrival during the last 10 seconds. Then the probability of having an arrival within the next 2 seconds is independent of how long there has been no arrival so far, namely,

$$P(T \le 10 + 2 \mid T \ge 10) = P(T \le 2)$$

- Do not mistakenly think that

$$P(T \le 10 + 2 \mid T \ge 10) = P(T \le 12)$$

Stochastic Process

- Stochastic process is the mathematical abstraction of an empirical process
  - governed by the probabilistic laws (such as the Poisson process)
- A Family of Random Variables (RV) $X = \{X(t) \mid t \in T\}$ defined over a given probability space, indexed by parameter t that varies over index set T, is called Stochastic Process
  - the set T is the time range and X(t) denotes the state of the process at time t
  - the process is classified as a discrete-parameter or continuous parameter process

Stochastic Process

- If T is countable sequence like T = {0, ±1, ±2, ….} or T = {0, 1, 2, ….}
  - Stochastic process {X(t), t ∈ T} is said to be a discrete-parameter process
- If T is an interval for example T = {t: -∞ < t < +∞} or T = {t: 0 < t < +∞}
  - Stochastic process {X(t), t ∈ T} is called a continuous-parameter process

Markov Process

- A discrete-parameter stochastic process

  {X(t), t = 0, 1, 2, …..} or continuous-parameter stochastic process {X(t), t > 0} is a Markov process when

  - the conditional distribution of $X(t_n)$ given the values of $X(t_1)$, $X(t_2)$, $X(t_3)$, ……., $X(t_{n-1})$ depends only on the preceding value $X(t_{n-1})$; mathematically;
    - $Pr\{X(t_n) \leq x_n \mid X(t_1) = x_1, \ldots\ldots, X(t_{n-1}) = x_{n-1}\} = Pr\{X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}\}$

- More intuitively, given the “present” condition of the process, the future is independent of the “past”
  - The process is thus “memoryless”

Markov Process

- Markov processes are classified according to:
  - The nature of the index set of the process (discrete or continuous parameter)
  - The nature of the state space of the process (discrete or continuous parameter)

State Space	***Index set T (Type of Parameters)***	
	Discrete	Continuous
Discrete	Discrete parameter Stochastic/Markov chain	Continuous parameter Stochastic/Markov chain
Continuous	Discrete parameter Continuous Stochastic/ Markov process	Continuous parameter Continuous Stochastic/ Markov process

Questions

?