Queueing Theory - A primer

Harry Perros

- Queueing theory deals with the analysis of queues (or waiting lines) where customers wait to receive a service.
- Queues abound in everyday life!
  - *Supermarket checkout*
  - *Traffic lights*
  - *Waiting for the elevator*
  - *Waiting at a gas station*
  - *Waiting at passport control*
  - *Waiting at a a doctor's office*
  - *Paperwork waiting at somebody's office to be processed*

- There are also queues that we cannot see (unless we use a software/hardware system), such as:
  - *Streaming a video:* Video is delivered to the computer in the form of packets, which go through a number of routers. At each router they have to waiting to be transmitted out
  - *Web services:* A request issued by a user has to be executed by various software components. At each component there is a queue of such requests.
  - *On hold at a call center*

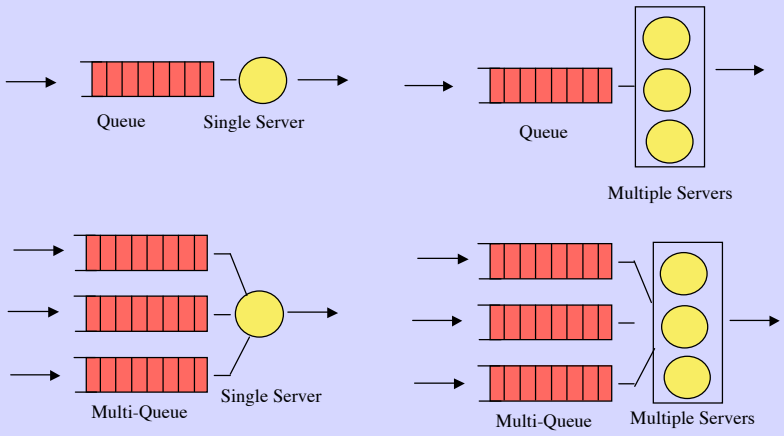
Measures of interest

- *Mean waiting time*
- *Percentile of the waiting time*, i.e. what percent of the waiting customers wait more than *x* amount of time.
- *Utilization of the server*
- *Throughput*, i.e.number of customers served per unit time.
- *Average number of customers waiting*
- *Distribution of the number of waiting customers*, i.e. Probability [n customers wait], n=01,1,2,…

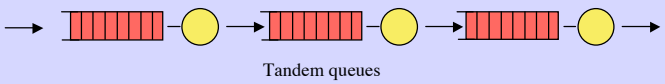
Reality vs perception

- Queueing theory deals with actual waiting times.
- In certain cases, though, it's more important to deal with the perception of waiting.. For this we need a *psychological perspective* ! (Famous example, that "minimized" waiting time for elevators!!)

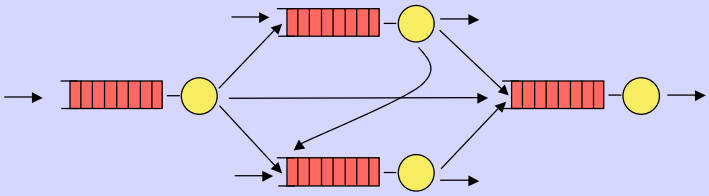
Notation - single queueing systems



Notation - Networks of queues

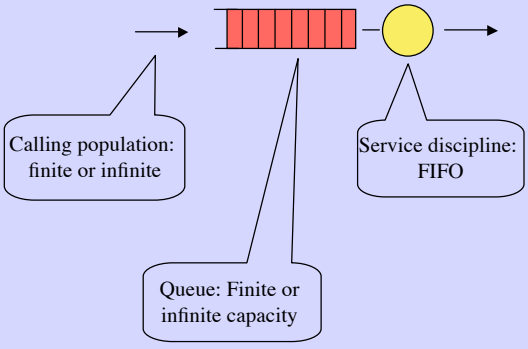


Tandem queues



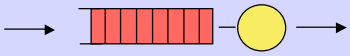
Arbitrary topology of queues

The single server queue



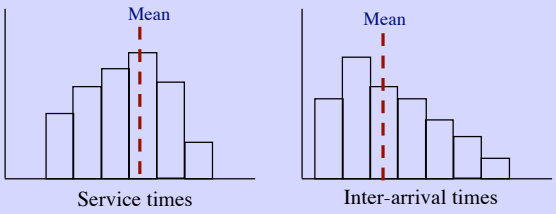
Queue formation

- A queue is formed when customers arrive faster than they can get served.

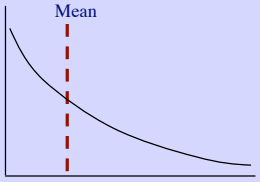


- Examples:
  - Service time = 10 minutes, a customer arrives every 15 minutes ---> No queue will ever be formed!
  - Service time = 15 minutes, a customer arrives every 10 minutes ---> Queue will grow for ever (bad for business!)

- Service times and inter-arrival times are rarely constant.
- From real data we can construct a histogram of the service time and the inter-arrival time.



- If real data is not available, then we assume a theoretical distribution.
- A commonly used theoretical distribution in queueing theory is the exponential distribution.

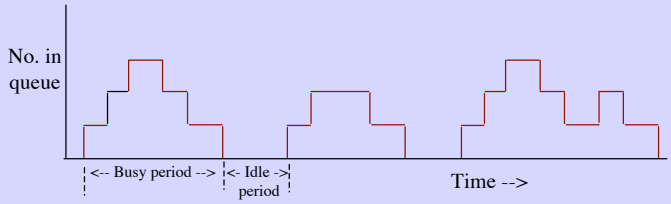


Stability condition

- A queue is stable, when it does not grow to become infinite over time.
- The single-server queue is stable if on the average, the service time is less than the inter-arrival time, i.e.

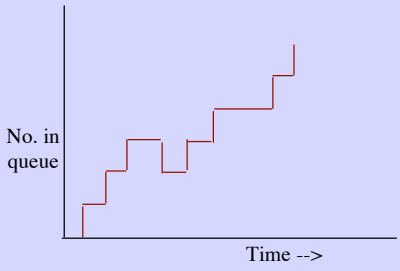
mean service time < mean inter-arrival time

Behavior of a stable queue
Mean service time < mean inter-arrival time



When the queue is stable, we will observe busy and idle periods continuously alternating

Behavior of an unstable queue
Mean service time > mean inter-arrival time



Queue continuously increases..
This is the case when a car accident occurs on the highway

Arrival and service rates: definitions

- *Arrival rate is the mean number of arrivals per unit time = 1/ (mean inter-arrival time)*
  - If the mean inter-arrival = 5 minutes, then the arrival rate is 1/5 per minute, i.e. 0.2 per minute, or 12 per hour.
- *Service rate is the mean number of customers served per unit time = 1/ (mean service time)*
  - If the mean service time = 10 minutes, then the service rate is 1/10 per minute, i.e. 0.1 per minute, or 6 per hour.

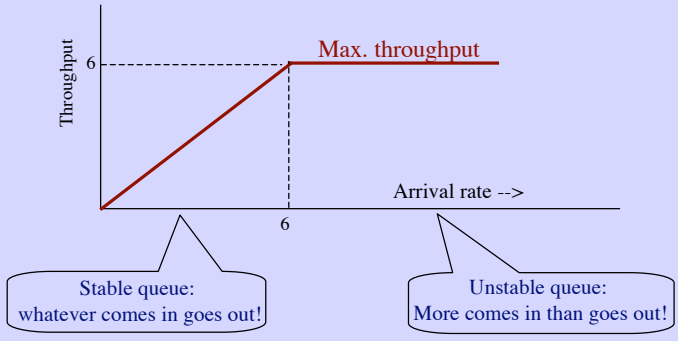
Throughput

- This is average number of completed jobs per unit.
- Example:
  - The throughput of a production system is the average number of finished products per unit time.
- Often, we use the *maximum throughput* as a measure of performance of a system.

Throughput of a single server queue

- This is the average number of jobs that depart from the queue per unit time (after they have been serviced)
- Example: The mean service time =10 mins.
  - What is the maximum throughput (per hour)?
  - What is the throughput (per hour) if the mean inter-arrival time is:
    - 5 minutes ?
    - 20 minutes ?

Throughput vs the mean inter-arrival time. Service rate = 6



Server Utilization= Percent of time server is busy = (arrival rate) x (mean service time)

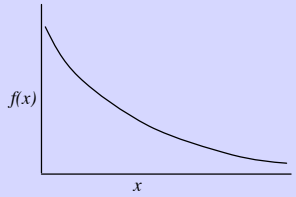
- Example:
  - Mean inter-arrival = 5 mins, or arrival rate is 1/5 = 0.2 per min. Mean service time is 2 minutes
  - Server Utilization = Percent of time the server is busy:

    0.2x2=0.4 or 40% of the time.
  - Percent of time server is idle?
  - Percent of time no one is in the system (either waiting or being served)?

The M/M/1 queue

- M implies the exponential distribution (Markovian)
- The M/M/1 notation implies:
  - *a single server queue*
  - *exponentially distributed inter-arrival times*
  - *exponentially distributed service times.*
  - *Infinite population of potential customers*
  - *FIFO service discipline*

The exponential distribution

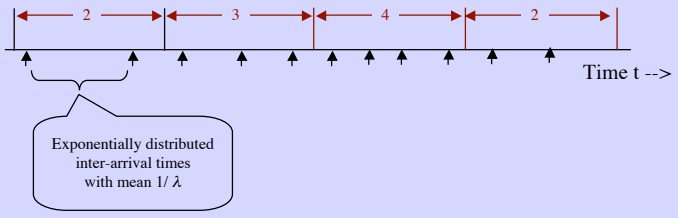


- $f(x) = \lambda e^{-\lambda x}$, where λ is the arrival / service rate.
- Mean = $1/\lambda$
- Memoryless property - Not realistic, but makes the math easier!

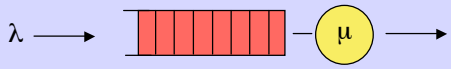
The Poisson distribution

- Describes the number of arrivals per unit time, if the inter-arival time is exponential
- Probability that there *n* arrivals during a unit time:

$$\text{Prob}(n) = (\lambda t)^n e^{-\lambda t}$$



Queue length distribution of an M/M/1 queue



- Probability that there are n customers in the system (i.e., queueing and also is service):

$$\text{Prob}[n] = \rho^n (1-\rho), \quad n=0,1,2,..$$

where ρ is known as the *traffic intensity* and

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

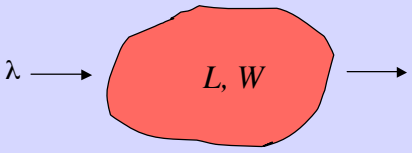
Performance measures

- Prob. system (queue and server) is empty:

$$\text{Prob}[n=0] = 1-\rho$$

- Percent of time server is idle = $1-\rho$
- Server utilization = percent of time server is busy = ρ
- Throughput = λ

Little's Law

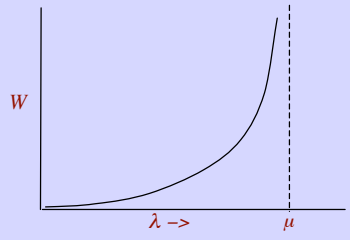


Denote the mean number of customers in the system as L and the mean waiting time in the system as W. Then:

$$\lambda W = L$$

Mean waiting time and mean number of customers

- Mean number of customers in the system: $L = \lambda / (\mu - \lambda)$
- Mean waiting time in the system (queueing and receiving service) obtained using Little's Law: $W = 1 / (\mu - \lambda)$



- *Problem 1:* Customers arrive at a theater ticket counter in a Poisson fashion at the rate of 6 per hour. The time to serve a customer is distributed exponentially with mean 10 minutes.
  - Prob. a customer arrives to find the ticket counter empty (i.e., it goes into service without queueing)
  - Prob. a customer has to wait before s/he gets served
  - Mean number of customers in the system (waiting and being served)
  - Mean time in the system
  - Mean time in the queue
  - Server utilization

1. *Problem 1 (Continued)*

- How many parking spaces we need to construct so that 90% of the customers can find parking?

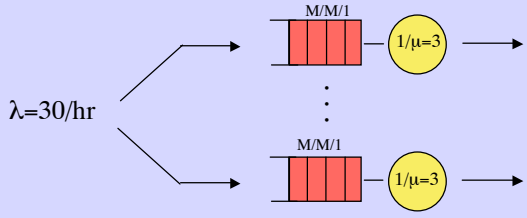
n	$P(n)$	$P(k \leq n)$
0		
1		
2		
3		
4		

- *Problem 2: Design of a drive-in bank facility*

  People who are waiting in line may not realize how long they have been waiting until at least 5 minutes have passed. How many drive-in lanes we need to keep the average waiting time less than 5 minutes?

  - Service time: exponentially distributed with mean 3 mins

  - Arrival rate: Poisson with a rate of 30 per hour



- *Answer:*

  Mean waiting time in the system (queueing and receiving service) : $W = 1 /(\mu-\lambda).$

  Mean waiting time in the queue (excluding service time) : $W_q = 1 /(\mu-\lambda) - 1/\mu = \lambda /\mu(\mu-\lambda)$

  Results

  1 Teller -> M/M/1 queues -> W_q=

  2 Tellers -> 2 M/M/1 queues -> W_q=

  3 Tellers -> 3 M/M/1 queues -> W_q=

  4 Tellers -> 4 M/M/1 queues -> W_q=

- *Problem 3:*

  If both customers and servers are employees of the same company, then the cost of employees waiting (lost productivity) and the cost of the service is equally important to the company.

  Employees arrive at a service station at the rate of 8/hour. The cost of providing the service is a function of the service rate, i.e. $C_s = f$(service rate)$=10\mu$ per hour. The cost to the company for an employee waiting in the system (queue and also getting served) is C_w = \$50/hour.

  Assuming an M/M/1 queue what is the value of the mean service time that minimizes the total cost ?

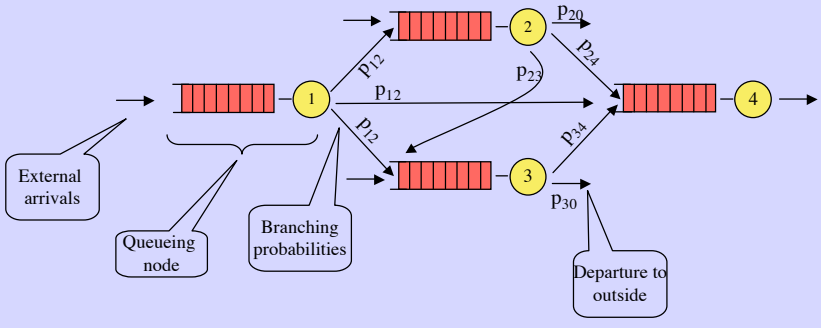
- Service cost $C_s = 10\mu$
- Waiting cost $C_w = 50\ \lambda\ W = 400\ [1/(m-8)]$
- Total cost: $C_s + C_w$

μ	10μ	400/(μ-8)	Sum
10	100	200	300
11	110	133.333	243.333
12	120	100	220
13	130	80	210
14	140	66.6666	206.666
15	150	57.1428	207.142
16	160	50	210
17	170	44.4444	214.444
18	180	40	220
19	190	36.3636	226.363
20	200	33.3333	233.333
21	210	30.7692	240.769
22	220	28.5714	248.571

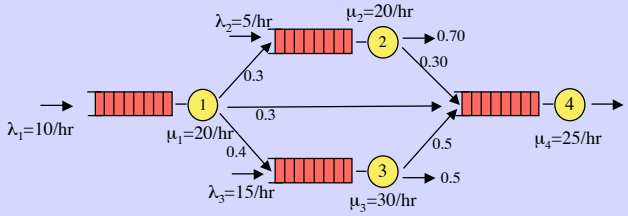
Networks of queues

- Typically, the flow of customers through a system may involve a number of different service stations.
- Examples:
  - Flow of IP packets through a computer network
  - Flow of orders through a manufacturing system
  - Flow of requests/messages through a web service system
  - Flow of paperwork through an administration office
- Such systems can be depicted by a network of queues

Queues can be linked together to form a network of queues which reflect the flow of customers through a number of different service stations

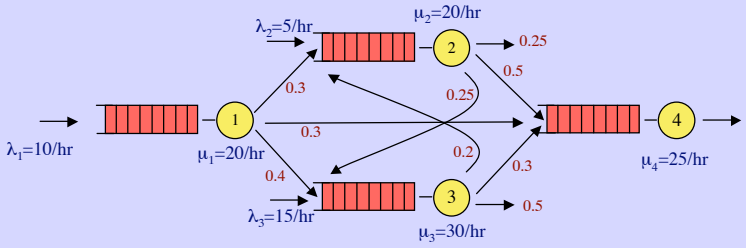


Traffic flows



- What is the total (effective) arrival rate to each node?
- Is each queue stable?
- What is the total departure rate from each node to the outside ?
- What is the total arrival to the network?
- What is the total departure from the network?

Traffic flows - with feedbacks



- What is the total (effective) arrival rate to each node?
- Is each queue stable?
- What is the total departure rate from each node to the outside ?
- What is the total arrival to the network?
- What is the total departure from the network?

Traffic equations

- M nodes; λ_i is the external arrival rate into node i, and p_{ij} are the branching probabilities. Then the effective arrival rates can be obtained as follows:

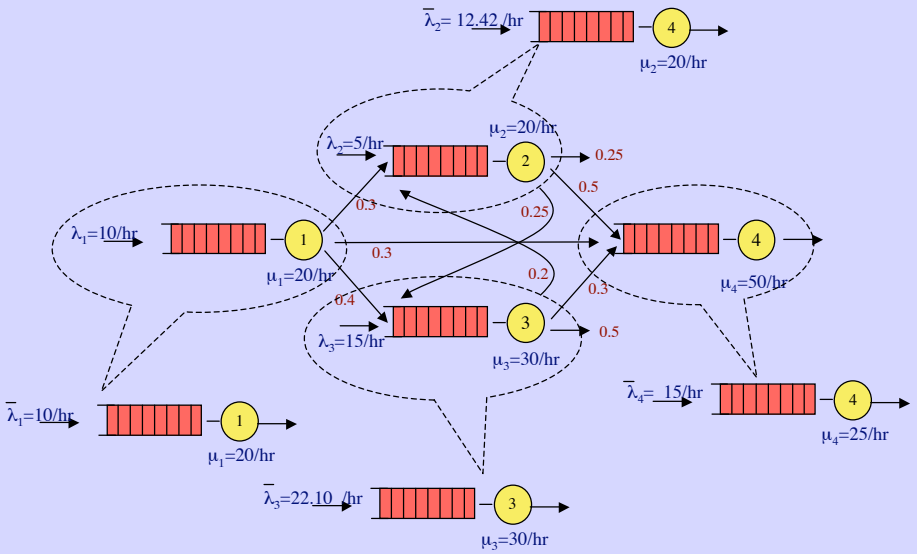
$$\bar{\lambda}_1 = \lambda_1 + \sum_{i=1}^{M} \bar{\lambda}_i p_{i,1}$$

...

$$\bar{\lambda}_M = \lambda_M + \sum_{i=1}^{M} \bar{\lambda}_i p_{i,M}$$

Solution

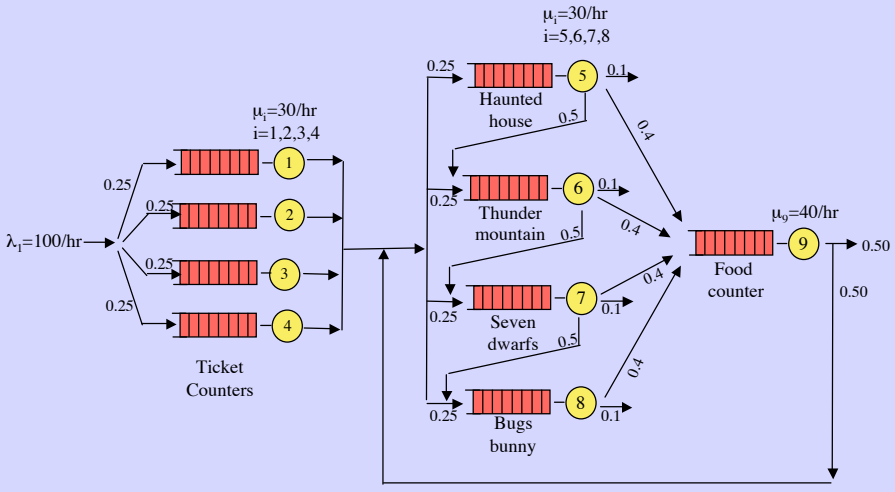
- Assumptions:
  - *Each external arrival stream is Poisson distributed.*
  - *The service time at each node is exponentially distributed.*
- Each node can be analyzed separately as an M/M/1 queue with an arrival rate equal to the effective (total) arrival rate to the node and the same original service rate



- For the previous example calculate
  - Prob. each node is empty
  - Mean number of customers in each node
  - Mean waiting time in each node
  - Assume a customer arrives at node 1, then it visits node 2, 3, and 4, and then it departs. What is the total mean waiting time of the customer in the network?
  - What is the probability that at each node it will not have to wait?

Problem: Having fun at Disneyland !

- A visit at Disney Land during the Christmas Holidays involves queueing up to buy a ticket, and then once in the park, you have to queue up for every theme.
- Does one spends more time queueing up or enjoying the themes? That depends on the arrival rates of customers and service times (i.e. the time to see a theme)



- Questions
  - Are all the queues stable?
  - What is the utilization of each ticket counter?
  - What is the probability that a customer will get served immediately upon arrival at the food court?
  - What is the mean waiting time at each theme?
  - How long would it take on the average to visit all themes once, and of that time how much one spends standing in a queue?
  - How long on the average a customer spends waiting for