Queuing Theory

Queuing Theory

- Queuing theory is the mathematics of waiting lines.
- It is extremely useful in predicting and evaluating system performance.
- Queuing theory has been used for operations research. Traditional queuing theory problems refer to customers visiting a store, analogous to requests arriving at a device.

Long Term Averages

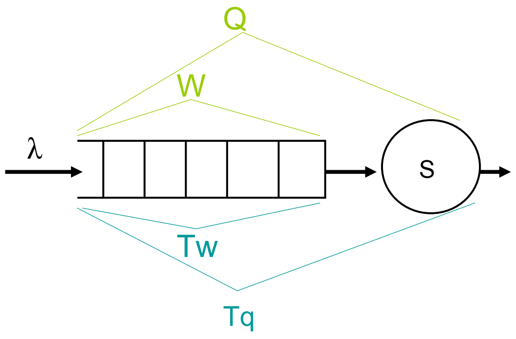
- Queuing theory provides long term average values.
- It does not predict when the next event will occur.
- Input data should be measured over an extended period of time.
- We assume arrival times and service times are random.

Assumptions

- Independent arrivals
- Exponential distributions
- Customers do not leave or change queues.
- Large queues do not discourage customers.

Many assumptions are not always true, but queuing theory gives good results anyway

Queuing Model



Interesting Values

- Arrival rate (λ) — the average **rate** at which customers arrive.
- Service time (s) — the average **time** required to service one customer.
- Number waiting (W) — the average **number** of customers waiting.
- Number in the system (Q) — the average total **number** of customers in the system.

More Interesting Values

- Time in the system (Tq) the average **time** each customer is in the system, both waiting and being serviced.
- Time waiting (Tw) the average **time** each customer waits in the queue.

$$Tq = Tw + s$$

Arrival Rate

- The arrival rate, λ, is the average rate new customers arrive measured in arrivals per time period. Common units are access/second
- The inter-arrival time, **a**, is the average time between customer arrivals. It is measured in time per customer. A common unit would be seconds/access.

$$a = 1 / \lambda$$

Random Values

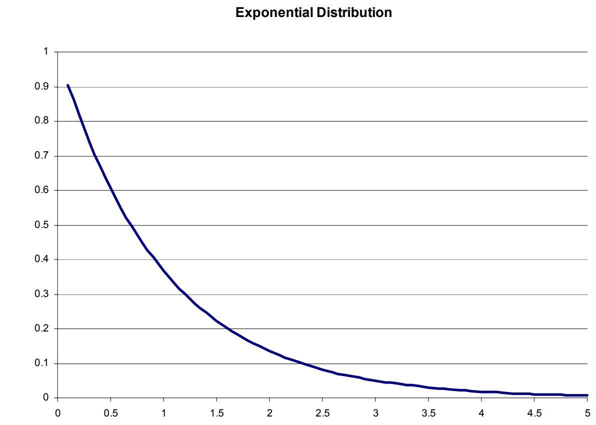
- We assume that most of the events we are interested in occur randomly.
  - Time of a request to a device
  - Time to service a request
  - Time user makes a request
- Although events are random, we may know the average value of the times and their distribution.
- If you flip a coin, you will get heads 50% of the time.

Exponential Distribution

- Many of the random values are exponentially distributed.

Frequency of Occurrence = e^{-t}

- There are many small values and a few large values.
- The inter-arrival time of customers is naturally exponentially distributed.



Poisson Arrival Rate

If customers are arriving at the exponentially distributed rate λ, then the probability that there will be *k* customers after time *t* is:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Math Notes

$$0! = 1! = 1$$

$$X^0 = 1$$

$$X^1 = X$$

Poisson Example

- A networked printer usually gets 15 print jobs every hour. The printer has to be turned off for 10 minutes for maintenance. What is the probability that nobody will want to use the printer during that time?

Poisson Solution

- A networked printer usually gets 15 print jobs every hour. The printer has to be turned off for 10 minutes for maintenance. What is the probability that nobody will want to use the printer during that time?
- The arrival rate is 15/60 = 0.25 jobs/min.

$$P_0(10) = \frac{(0.25*10)^0}{0!} e^{-0.25*10} = 0.082$$

Expected Number of Arrivals

If customers are arriving at the exponentially distributed rate λ, how many customers should you expect to arrive in time t?

$$\text{Expected} = \lambda * t$$

For the printer problem with an arrival rate λ = 0.25, in 10 minutes we should expect 2.5 jobs to arrive

Queuing Models

Queuing systems are usually described by three values separated by slashes

Arrival distribution / service distribution / # of servers

where:

- M = **M**arkovian or exponentially distributed
- D = **D**eterministic or constant.
- G = **G**eneral or binomial distribution

Common Models

- The simplest queuing model is **M/M/1** where both the arrival time and service time are exponentially distributed.
- The **M/D/1** model has exponentially distributed arrival times but fixed service time.
- The **M/M/n** model has multiple servers.

Why is there Queuing?

- The arrivals come at random times.
- Sometimes arrivals are far apart. Sometimes many customers arrive at almost the same time. When more customers arrive in a short period of time than can be serviced, queues form.
- If the arrival rate was not random, queues would not be created.

Utilization

- Utilization (represented by the Greek letter rho, ρ) is the fraction of time the server is busy.
- Utilization is always between zero and one

$$0 \leq \rho \leq 1$$

- If a bank teller spends 6 hours out of an 8 hour day counting money, her utilization is 6/8 = 0.75

Calculating Utilization

- Utilization can be calculated from the arrival rate and the service time.

$$\rho = \lambda * s$$

It is important that the units of both the arrival rate and the service time be identical. It may be necessary to convert these values to common units.

Little's Formula

- The number in the system is equal to the arrival rate times the average time a customer spends in the system.

$$Q = \lambda * Tq$$

- This is also true for just the queue.

$$W = \lambda * Tw$$

M/M/1 Formulas

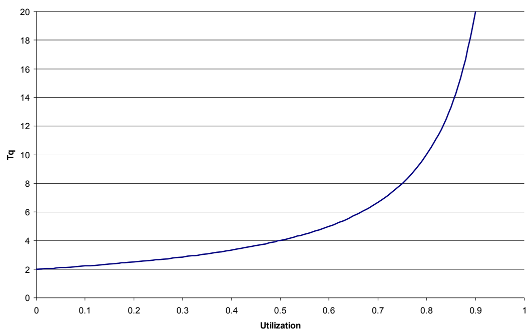
$$Tq = \frac{s}{1-\rho} \qquad Q = \frac{\rho}{1-\rho}$$

$$Tw = \frac{s\rho}{1-\rho} \qquad W = \frac{\rho^2}{1-\rho}$$

Application of Little's Formula

- Multiplying the formulas on the left by λ gives the formula on the right.

$$\lambda Tq = \frac{\lambda s}{1-\rho} = \frac{\rho}{1-\rho} = Q$$



Solution Process

1. Determine what quantities you need to know.
2. Identify the server
3. Identify the queued items
4. Identify the queuing model
5. Determine the service time
6. Determine the arrival rate
7. Calculate ρ
8. Calculate the desired values

Example

- Consider a disk drive that can complete an average request in 10 ms. The time to complete a request is exponentially distributed. Over a period of 30 minutes, 117,000 requests were made to the disk. How long did it take to complete the average request?
  What is the average number of queued requests?

Solution

- Determine what quantities you need to know.
  - The average request time is Tq
  - The number of queued jobs is W
- Identify the server
  - The disk drive is the server
- Identify the queued items
  - Disk requests
- Identify the queuing model
  - M/M/1

Solution (cont.)

- Determine the service time
  - ❖S = 10 ms = 0.01 sec / request
- Determine the arrival rate
  - ❖λ = 117,000 request / (30 min * 60 sec/min) = 65 requests / sec
- Calculate ρ
  - ❖$\rho = \lambda * s$ = 0.01 sec/request * 65 req/sec = 0.65

Solution (cont.)

- Time to complete the average request

$$T_Q = \frac{s}{1-\rho} = \frac{0.01}{1-0.65} = 28.6ms$$

The average length of the queue

$$W = \frac{\rho^2}{1-\rho} = \frac{0.65^2}{1-0.65} = 1.21$$

Number in the System

- The value Q represents the average number of jobs in the system, both waiting and being served.
- There are not always Q jobs in the system. Sometimes there are more, sometimes less. Q is the average.

Queue Size Probabilities

The probability that there are exactly N jobs in the system is given by

$$\text{Prob}[Q = N] = (1-\rho)\rho^N$$

Summing the probabilities for individual cases gives the probability of N or less customers in the system

$$prob[Q \le N] = \sum_{i=0}^{N} (1-\rho)\rho^i$$

Large Queue Probabilities

- The probability that there are more than N customers in the system is the sum of the probabilities from N-1 to ∞.
- Remembering that the sum of all probabilities is one, the probability that there are more than N customers in the system is:

$$prob[Q > N] = 1 - \sum_{i=0}^{N} (1-\rho)\rho^i$$

Example Continued

- In the previous example, what is the probability that a request does not get queued?
- A job can get serviced immediately if there are only zero or one jobs in the system.

$$P[Q = 0 or 1] = (1-\rho)*\rho^0 + (1-\rho)*\rho^1 = 0.35 + .2275 = 0.58$$

Accuracy and Significant Digits

- Just because my calculator displays a 10 digit number does not mean the answer is accurate to 10 digits.
- Your answer can only be as accurate as your input data. If your data has three significant digits, your answer cannot have more than three digits.
- Always use as much accuracy as possible in these calculations and round off only at the end.

Constant Service Time

- In some systems the service time is always a constant. The M/D/1 model is used for constant service time.
- There is less randomness in the system.
- The wait time will be less.

M/D/1 Formulas

$$Tq = \frac{s(2-\rho)}{2(1-\rho)} \qquad Q = \frac{\rho^2}{2(1-\rho)} + \rho$$

$$Tw = \frac{s\rho}{2(1-\rho)} \qquad W = \frac{\rho^2}{2(1-\rho)}$$

M/D/1 Example

- An ATM network sends 53 byte packets over a 155 Mb/sec line. It always takes 2.74 ms to send a packet. Each second 145,000 packets are sent. How long does a packet wait to be sent?

M/D/1 Solution

- Determine what quantities you need to know.

The average time spent in the queue, Tw.

- Identify the server

The transmission line.

- Identify the queued items

Packets (not bits or bytes)

- Identify the queuing model

M/D/1

M/D/1 Solution

- Determine the service time

2.74×10^{-6} seconds

- Determine the arrival rate

145,000 packets/second

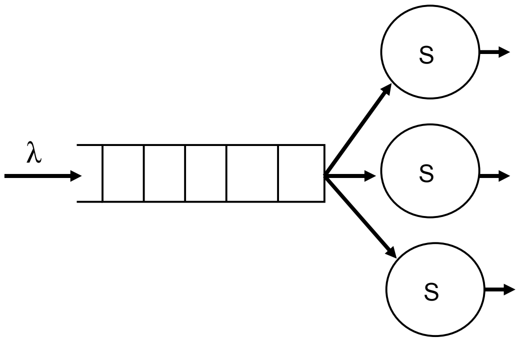
- Calculate ρ

$\rho = 145{,}000 * 2.74\times10^{-6} = 0.3973$

- Calculate the desired values

$$Tw = \frac{s\rho}{2(1-\rho)} = \frac{2.75 * 10^{-6} * 0.3973}{2(1-0.3973)} = 9.03 * 10^{-7} \text{ sec}$$

Multiple Servers



Multiple Servers

- Customers arrive and join a single queue.
- Whenever any of the servers is idle, it serves the first customer on the single queue.
- All of the servers must be identical. Any customer can be served by any server.
- When there are N servers, the model is

M/M/N

Multiple Server Utilization

- The server utilization for an N server system is:

$$\rho = \lambda s / N$$

This is the average utilization for all N servers.

Intermediate Value K

- To make calculations easier, we first compute the value K.

$$K = \frac{\sum_{i=0}^{N-1} \frac{(\lambda s)^i}{i!}}{\sum_{i=0}^{N} \frac{(\lambda s)^i}{i!}}$$

K Calculation

- The first term (i = 0) is always 1
- Note that the value in the denominator is equal to the numerator plus the last term.
- Since the denominator is always larger than the numerator, the value K must always be less than 1.
- The value K is an intermediate that simplifies calculations. It has no intrinsic meaning.

Multiple Servers Busy

The probability that all servers are busy is

$$C = \frac{1-K}{1-\frac{\lambda s K}{N}}$$

This is the probability that a new customer will have to wait in the queue.

M/M/N formulas

$$Tq = \frac{Cs}{N(1-\rho)} + s \qquad Q = C\frac{\rho}{1-\rho} + \lambda s$$

$$Tw = \frac{Cs}{N(1-\rho)} \qquad W = C\frac{\rho}{1-\rho}$$

note: $\rho = \lambda s / N$

Example

- Assume that you have a printer that can print an average file in two minutes. Every two and a half minutes a user sends another file to the printer. How long does it take before a user can get their output?

Slow Printer Solution

- Determine what quantities you need to know.

How long for job to exit the system, Tq

- Identify the server

The printer

- Identify the queued items

Print job

- Identify the queuing model

M/M/1

Slow Printer Solution

- Determine the service time

Print a file in 2 minutes, s = 2 min

- Determine the arrival rate

new file every 2.5 minutes. $\lambda = 1/2.5 = 0.4$

- Calculate ρ

$\rho = \lambda * s = 0.4 * 2 = 0.8$

- Calculate the desired values

Tq = s / (1- ρ) = 2 / (1 - 0.8) = 10 min

Add a Second Printer

- To speed things up you can buy another printer that is exactly the same as the one you have. How long will it take for a user to get their files printed if you had two identical printers?

- All values are the same, except the model is M/M/2 and $\rho = \lambda * s / 2 = 0.4$

Calculate K

$$K = \frac{\sum_{i=0}^{N-1} \frac{(\lambda s)^i}{i!}}{\sum_{i=0}^{N} \frac{(\lambda s)^i}{i!}} = \frac{\frac{(\lambda s)^0}{0!} + \frac{(\lambda s)^1}{1!}}{\frac{(\lambda s)^0}{0!} + \frac{(\lambda s)^1}{1!} + \frac{(\lambda s)^2}{2!}}$$

K = 0.849057

Calculate M/M/N Solution

$$C = \frac{1-K}{1-\frac{\lambda s K}{N}} = 0.22857$$

$$Tq = \frac{Cs}{N(1-\rho)} + s = 2.57 \min$$

Note that with twice as many printers this example runs about 4X as fast.

Faster Printer

- Another solution is to replace the existing printer with one that can print a file in an average of one minute. How long does it take for a user to get their output with the faster printer?

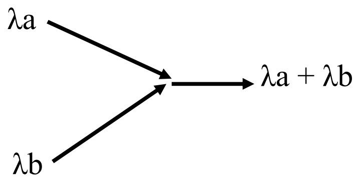
- M/M/1 queue with λ = 0.4 and s = 1.0

  Tq = s / (1- ρ) = 1 / (1 - 0.4) = 1.67 min

A single fast printer is better, particularly at low utilization. 6X better than slow printer.

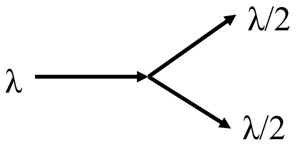
Multiple Arrival Streams

- Exponentially distributed arrival streams can be merged. The total arrival rate is the sum of the individual arrival rates.



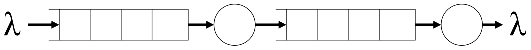
Dividing Customer Streams

- An exponentially distributed arrival stream can be divided. The sum of the separated arrival rates must equal to original arrival rate.



Linking Multiple Queues

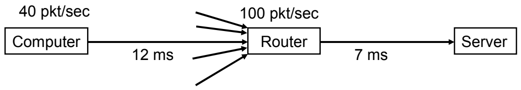
- The exit rate of a system is equal to the arrival rate.
- The output from one queuing system can feed into another.



The time through the system is the sum of the time through each queuing component.

Multiple Queue Example

- Consider a network with a computer connected to a router which is connected to a server. The computer can send a packet in 12 ms while the router can send it to the server in 7 ms. Programs on the computer generate 40 packets/second. The router receives a total of 100 packets/second.
- How long does it take for a packet to get to the server?



Multiple Queue Solution

- Determine what quantities you need to know

Sum of Tq for both networks

- Identify the servers

The computer transmitter and the router

- Identify the queued items

Packets

- Identify the queuing model

both M/M/1 queues

Multiple Queue Solution

- Determine the service time

12ms for the computer, 7 ms for the router

- Determine the arrival rate

40 pkt/sec for computer, 100 pkt/sec for router

- Calculate ρ

$\rho_{computer}$= 40*0.012 = 0.48 ρ_{router}=100*0.007 = 0.7

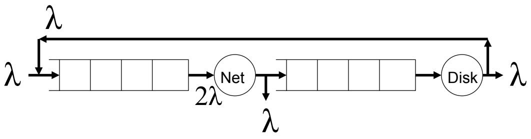
- Calculate the desired values

$Tq_{computer}$ =0.012/(1-0.48) = 0.0231 sec

Tq_{router} =0.007/(1-0.7) = 0.0233 sec Total=46.4ms

Reusing a Server

- Consider a file server. Requests use the network to get to the server, then use the disk, then the network again to get back.
- The load on the network is doubled.


λ
λ
Net
2λ
λ
Disk
λ