Queuing Systems

Queuing theory can be used to predict the effects of some change in load or design.

What does Queuing theory study?

- # number of users
- Arrival Characteristics
- Service Characteristics
- Resources

This leads to an indication of performance:

- Waiting Time
- Blocking

Elements of a Queuing system

- Input
- Queue
- Server
- Output

Models

Customers from some population can arrive at the system at random intervals

- λ is the customer arrival rate
- There are c identical servers
- The j^{th} customer seeks service which requires s_j units of service time from one server.
- If all servers are busy then the arriving customer joins a queue until a server becomes available.

Orders in which customers can leave the queue. 1. FIFO 2. LIFO 3. Priority 4. Fair Queuing 5. etc.

The waiting time, t_{Qj} is the waiting time of a customer in the queue from entrance to the queue to entering service.

Total delay in the queue system $\gamma_j = t_{Qj} + s_j$

n = number of customers in the system

n_q = number of customers in the queue

a/b/m/K notation

a

a represents the type of arrival process

- M for Markov. M denotes Poisson arrivals, so interarrival times are iid, exponential random variables.

b

b represents the service time distribution

- M (Markov) denotes exponentially distributed
- D (Deterministic) denotes constant service time
- G (General) denotes iid service times following some general distribution

m

m denotes the number of servers

K

K denotes the maximum number of customers allowed in the system

Acronyms