Introduction to Queuing Theory

Dr. Ali Muqaibel

Coverage

- Basic structure of queuing systems.
- Little's Formula
- M/M/1 system
- Multi-server systems M/M/c, M/M/c/c, and M/M/∞

Introduction- Motivation

- **How to analyze changes in network workloads?**
  - Should I add new terminals? How much?
- **What percentage of calls will be blocked?**
  - Adding more lines would solve the problem?
- **Analysis of system (network) load and performance characteristics**
  - response time
  - throughput
- Performance tradeoffs are often not intuitive
- Queuing theory, although mathematically complex, often makes analysis very straightforward

Queueing Theory

- Operations Research
- The study of waiting
- Back to early twentieth century
  - Danish mathematician A. K. **Erlang** (telephone networks), why?
  - Russian mathematician A. A. Markov
- Applied in a broad variety of applications

Operations research, or **operational research** in British usage, is a discipline that deals with the application of advanced analytical methods to help make better decisions.[1] It is often considered to be a sub-field of mathematics.[2] The terms **management science** and **decision science** are sometimes used as more modern-sounding synonyms http://en.wikipedia.org/wiki/Operations_research

Queuing Jargons

- Kendall's notation
  - Standard notation to describe queuing containing single queue: X/Y/m/n

InterArrival process

Service process

Servers #

Queue capacity

Common distributions

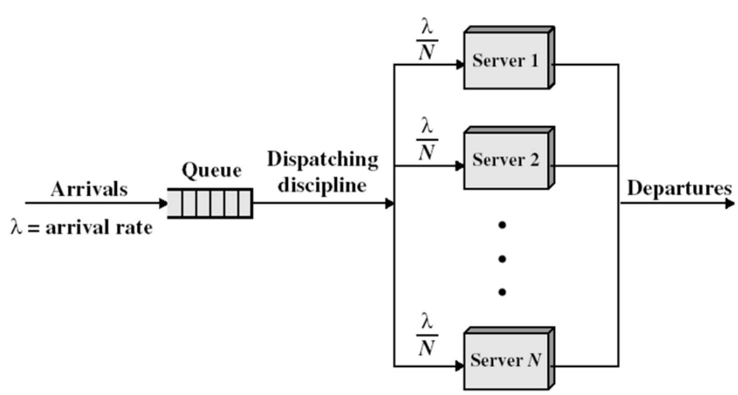
- G = general distribution if interarrival times or service times
- GI = general distribution of interarrival time with the restriction that they are independent
- M = negative exponential distribution (Poisson arrivals)
- D = deterministic arrivals or fixed length service

M/M/1? M/D/1? M/M/1/K? M/M/c/c? M/G/1? M/D/1?

General Characteristics of Queuing Models

- **Item population**
  - generally assumed to be infinite therefore, arrival rate is persistent
- **Queue size**
  - infinite, therefore no loss
  - finite, more practical, but often immaterial
- **Dispatching discipline**
  - FIFO, typical
  - LIFO
  - Relative/Preferential, based on QoS
  - ....

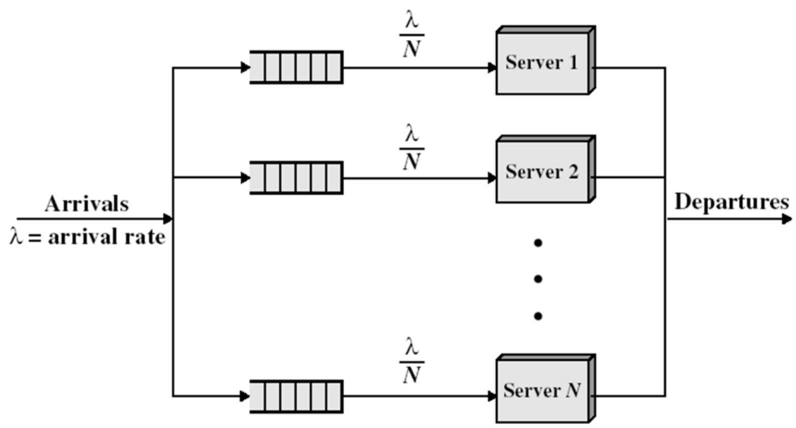
Multiserver Queue



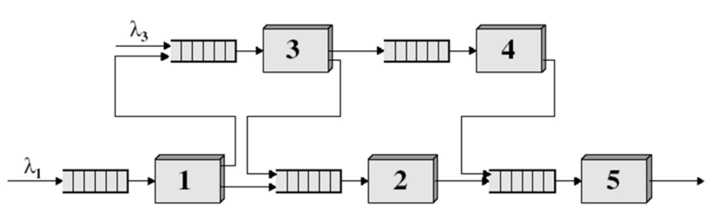
Comments:
1. Assuming N identical servers, and ρ is the utilization of each server.
2. Then, $N\rho$ is the utilization of the entire system, and the maximum utilization is $N \times 100\%$.
3. Therefore, the maximum supportable arrival rate that the system can handle is:

$$\lambda_{max} = N / T_s$$

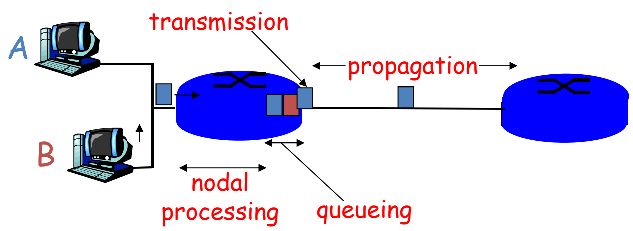
Multiple Single-Server Queues



Network of Queues



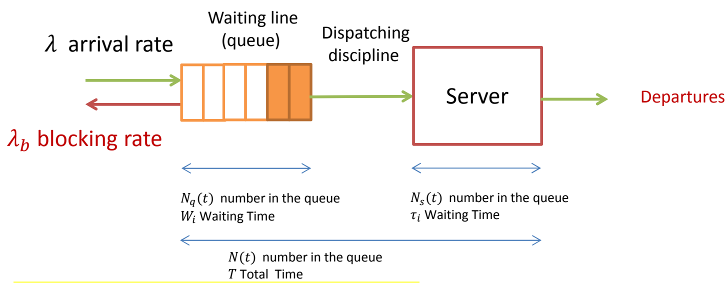
Delay Components



Delay Components (Cont.)

- Packet delay the sum of delays on each link on the path traversed by the packet.
- Each link delay in turns consists of
  - Processing delay: between the time the packet is correctly received at the head node of the link and the time the packet is assigned to an outgoing link queue; is independent of traffic carried.
  - Queueing delay: between the time the packet is assigned to a queue for transmission and the time it starts being transmitted.
  - Transmission delay: between the times that the first and last bits of the packet are transmitted.
  - Propagation delay: between the time the last bit is transmitted at the head node of the link and the time the last bit is received at the tail node; depends on the physical characteristics of the link.

The Elements of A Queuing System



Given:

- Arrival times $S_1, S_2, .. S_i, ...$
- Arrival rate, λ
- Blocking rate, λ_b
- Service rate $\lambda_d = \lambda - \lambda_b$
- Service time τ_i or T_s.
- Number of servers, *m*

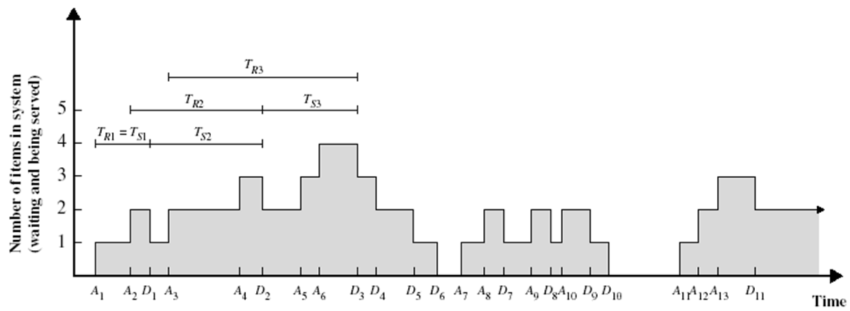
Determine:

- Items waiting, *w*
- Waiting time, W_i or T_w
- Items queued, N_q
- Total number, *N*
- Total Delay, $T_i = W_i + \tau_i$

Two Points of View

- **Customer's point of view**, the performance of the system is given by:
  - Statistics of the waiting time *W*
  - The total delay *T*,
  - The proportion of customers that are blocked, $\frac{\lambda_b}{\lambda}$
- **From the point of view of resource allocation**, the performance of the system is measured by:
  - The proportion of time that each server is utilized.
  - The rate at which customers are serviced by the system, $\lambda_d = \lambda - \lambda_b$.
- **These quantities area function of:**
  - $N(t)$, the number of customers in the system at time *t*.
  - $N_q(t)$, the number of customers in queue at time *t*.

Queuing Process - Example



For item i:
A_i = Arrival time
D_i = Departure time
T_{Ri} = Residence time
T_{Si} = Service time

Transient versus steady-state behaviour

- **Transient behaviour (from $t = 0$)**

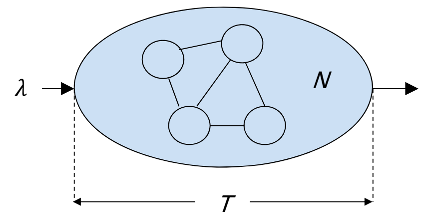
  performance indicators such as average waiting time, average number of customers in queue, etc. are dependent of the time, e.g. $W(t)$, $N_q\ (t)$

- **Steady-State (stationary) behaviour (t $\rightarrow \infty$)**

  performance indicators such as average waiting time are not dependent of the time anymore; the probability that the system is in a certain state is completely independent of time, e.g. W, N_q.

Little's Formula

For systems that reach steady state, the average number of customers in a system is equal to the product of the average arrival rate and the average time spent in the system.

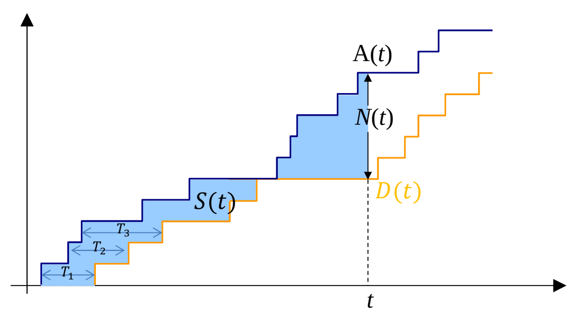


- λ: customer arrival rate
- $E[N]$: average number of customers in system
- $E[T]$: average delay per customer in system

➡ Little's Formula: System in steady-state

➡ $E[N] = \lambda E[T]$

Counting Processes of a Queue



- The system begins empty at time $t = 0$.
- $N(t)$: number of customers in system at time t.
- $A(t)$: number of customer arrivals till time t.
- $D(t)$: number of customer departures till time t
- T_i : time spent in system by the i^{th} customer
- $S(t)$: the cumulative area between $A(t)$ and $D(t)$