An Overview Of Queueing Network Models

Why worry about modeling?

- Understand the behavior of today's complex computer systems
  - During design and implementation
  - During sizing and acquisition
  - During evolution of the configuration and workload

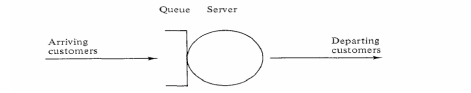
What is a queueing network model?

- Represent a system as a network of queues evaluated analytically
  - Service centers, which represent system resources
  - Customers, which represent users or transactions

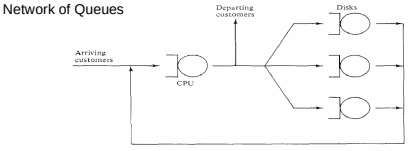
Possible Models (open networks)

Parameters: workload intensity, service demand

A Single Service Center

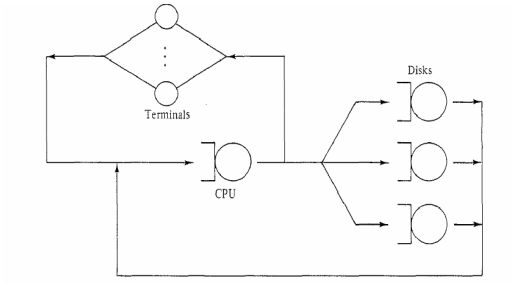


A Network of Queues



Performance Measures: utilization, residence time, queue length, throughput

A Model with a User/Terminal-Driven Workload (closed network)



Basic Quantities

T=length of time we observe the system, A=number of arrivals observed, C=number of completions observed, B=length of time the resource was busy

Arrival rate: lambda = A/T
Throughput: X = C/T
Utilization: U = B/T
Service requirement per request: S = B/C

Fundamental Laws

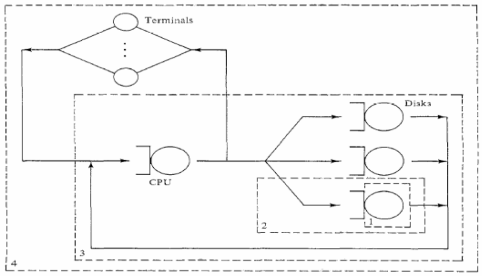
Notation:

T	length of an observation interval
A_k	number of arrivals observed
C_k	number of completions observed
λ_k	arrival rate
X_k	throughput
B_k	busy time
U_k	utilization
S_k	service requirement per visit
N	customer population
R_k	residence time
Z	think time of a terminal user
V_k	number of visits
D_k	service demand

Fundamental Laws:

The Utilization Law:	$U_k = X_k S_k = XD_k$
Little's Law:	$N = XR$
The Response Time Law:	$R = \frac{N}{X} - Z$
The Forced Flow Law:	$X_k = V_k X$

Little's Law Applied at Four Levels



Queuing Notation

- Arrival process: interarrival time (e.g., independent and identically distributed (IID) and exponentially distributed assumption common)
- Service time distribution
- Number of servers
- System capacity
- Population size
- Service discipline: e.g., FCFS, LCFS, LCFS-PR

Markov Process

- Future state of a process independent of the past and dependent only on the current state
- Markov chain: discrete state Markov process
- Birth-death process: transitions are restricted to neighboring state only
- Poisson processes: IID and exponentially distributed interarrival times-> number of arrivals over a given interval has a poisson distribution

Markov model (M/M/1 queues)

- Traffic intensity, t – service time/inter-arrival time (also U, utilization)
- Probability that the system is idle, $p_0 = 1-t$
- Probability of n jobs in the system, $p_n = t^n p_0$
- Probability that the queue is non-empty – $1-p_1-p_0 = 1 – (1-t) – t(1-t) = t^2$
- Expectation of number of customers in the service center, N – sum over all states multiplied by probabilities – $t/(1-t)$
- Expectation of number of customers in the queue, N-1 – sum over all states-1 multiplied by probabilities – $t^2/(1-t)$

Analysis of Open Queueing Networks

- Inputs:
  - X = external arrival rate, system throughput
  - Si = service time per visit to the ith device
  - Vi = number of visits to the ith device
  - M = number of devices (not including terminals)
- Outputs:
  - Qi = mean number of jobs at the ith device
  - Ri = response time of the ith device
  - R = system response time
  - Ui = utilization of the ith device
  - N = mean number of jobs in the system

Open Queueing Networks

- Applications – transaction processing systems such as banking or airline reservations
- Arrival rate independent of the load on the computer system
- Fixed capacity service center (single server with exponentially distributed service time and arrival time, Q_i is the mean number of jobs at the ith device)
  - $R_i = S_i(1+Q_i)$
  - $X=\lambda$
  - $X_i = XV_i$
  - $U_i = X_iS_i = XV_iS_i = \lambda D_i$
  - $Q_i = X_iR_i = X_iS_i (1+Q_i) = U_i (1+Q_i)$
  - $Q_i = U_i/(1-U_i)$
  - $R_i = S_i/(1-U_i)$
- Delay centers (infinite servers with exponentially distributed service time) also possible Ri = Si, Qi = Ui

Closed Networks – Mean Value Analysis (MVA)

- Inputs:
  - N = number of users
  - Z = think time
  - M = number of devices (not including terminals/users)
  - Si = service time per visit to the ith device
  - Vi = number of visits to the ith device
- Outputs:
  - X = system throughput
  - Qi = average number of jobs at the ith device
  - Ri – response time of the ith device
  - R = system response time
  - Ui = utilization of the ith device

MVA Algorithm

Initilization:
For i = 1 to M $Q_i = 0$
Iterations:
for n = 1 to N
  for i = 1 to M $R_i = S_i(1+Q_i)$
  R = 0; for i = 1 to M $R \mathrel{+}= R_iV_i$
  X = N/(Z+R)
  for i = 1 to M $Q_i = XV_iR_i$

Device throughputs: $X_i = XV_i$
Device utilizations: $U_i = XS_iV_i$