Network Performance Evaluation

- Efficiencies gained by sharing resources.

- Bandwidth

- Memory

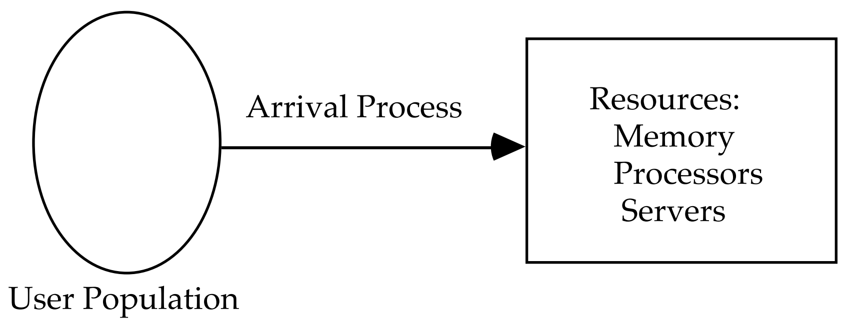
- Processor

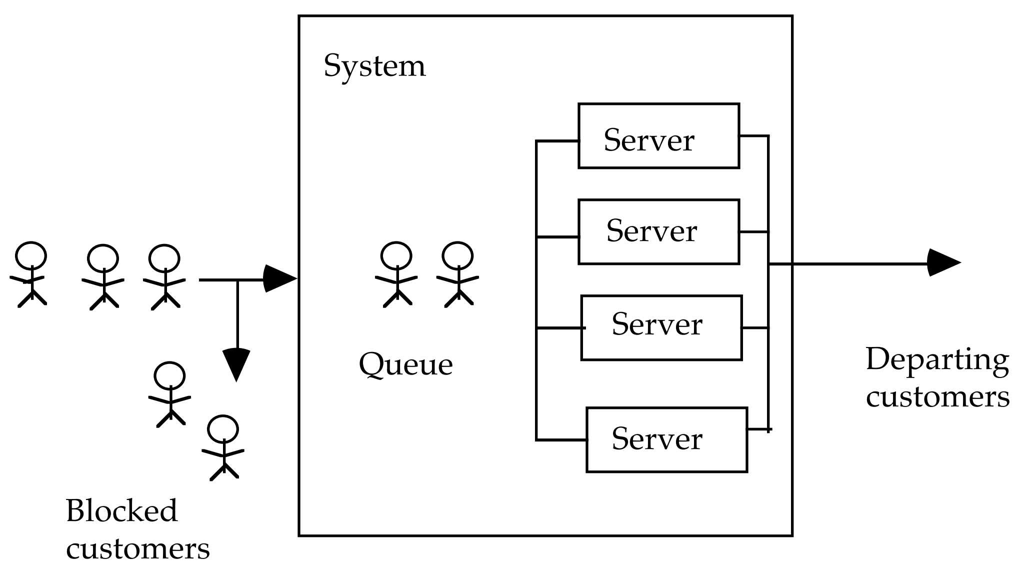
- Demands for these resources are:
  - + Requests for service, arrivals of call or packets
  - + Holding time/request

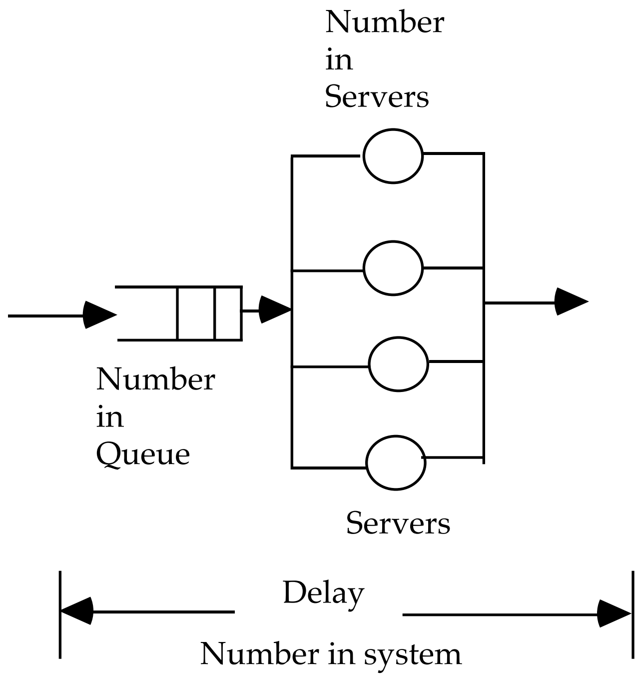
- Model this type of process as a *Queueing System*

- Methods for network analysis:
  + Mathematical analysis
  + Simulation techniques

Elements of a Queueing System







Goals for Network Analysis

Given the nature of the arrival and service processes and the system specification, i.e., buffer size and number of servers, find:

- Delay through the system
- Blocking probability

Specific cases:

- All cases assume Markov arrival and service processes, that means that the interarrival and service times are exponentially distributed.

- One server infinite memory

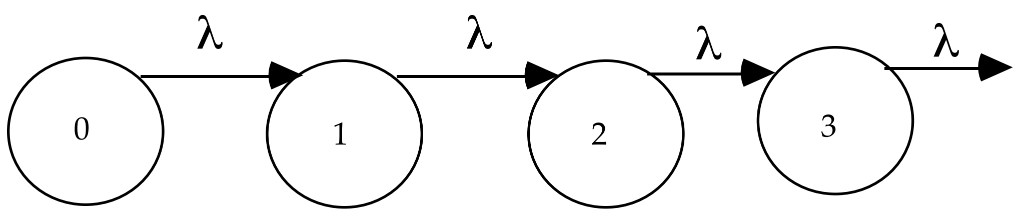
- One server finite memory

- S servers and a system size of S

Approach

- Analysis of a pure birth process to characterize arrival processes
- Extension to general birth/death processes, a common model for queueing systems
- Specialization to the specific cases to find the probability of system occupancy, average buffer size, delay, and blocking probability.

Analysis of a Pure Birth Process



Only Births Allowed

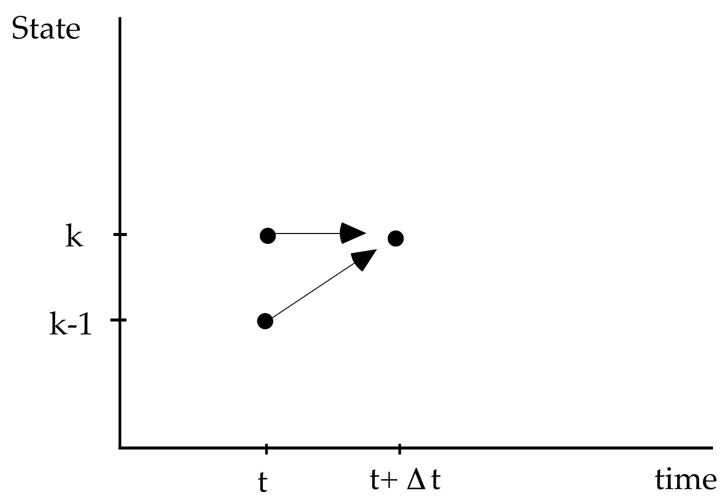
- The number represent the *State* of the system. In networks this is usually the number in the buffer plus the number in service. The system includes the server. The time to clock the message bits onto the transmission facility is the service time. The server is the model for the transmission facility.

- Assumptions:

  + Prob[1 arrivals in Δt sec] = $\lambda\Delta t$

  + Prob[0 arrivals in Δt sec] = $1-\lambda\Delta t$

  + Number of arrivals in non-overlapping intervals of times are statistically independent random variables



Using the assumptions

Probability of k in the system at time $t+\Delta t$ = Prob[k, $t+\Delta t$]

Prob[k, $t+\Delta t$] = Prob[(k in the system at time t and 0 arrivals in Δt) or (k-1 in the system at time t and 1 arrivals in Δt)]

$$\text{Prob}[k, t+\Delta t] = \text{Prob}[k,t](1-\lambda\Delta t) + \text{Prob}[k-1,t]\lambda\Delta t$$

rearranging terms

$$(\text{Prob}[k, t+\Delta t] - \text{Prob}[k,t])/\Delta t + \lambda\text{Prob}[k,t] = \lambda\text{Prob}[k-1,t]$$

Letting Δt --> 0 results in the following differential equation that can be solved for k=0 then k=1 and so on.

$$\frac{d\text{Prob}[k,t]}{dt} + \lambda\text{Prob}[k,t] = \lambda\text{Prob}[k-1,t]$$

For k = 0 the solution is:

$$\text{Prob}[0,t]= e^{-\lambda\tau}$$

For k = 1 the solution is:

$$\text{Prob}[1,t]= \lambda e^{-\lambda t}$$

For k = 2 the solution is:

$$\text{Prob}[1,t]= \lambda t e^{-\lambda t}$$

In general the solution is:

$$\text{Prob}[k, t] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

This is a Poisson probability mass function with:

$$E[k] = \lambda t$$

and

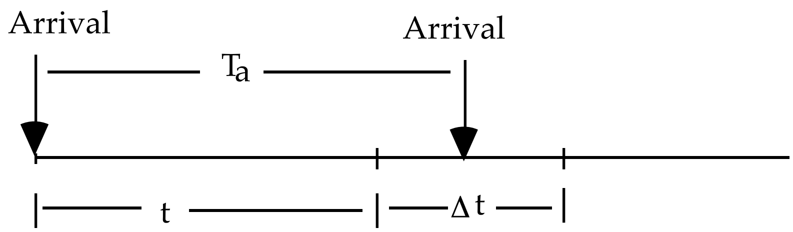
$$\text{Var}[k] = \lambda t$$

MAIN RESULT:

The Poisson Arrival Process

The number of arrivals in any T second interval follows a Poisson probability mass function.

A model for the interarrival time.



$\text{Prob}[t<T_a<t+\Delta t] =$
$\text{Prob}[0 \text{ arrivals in } t \text{ sec and } 1 \text{ arrival in } \Delta t]$

$\text{Prob}[t<T_a<t+\Delta t] =$
$\text{Prob}[k=0,t]\text{Prob}[k=1,\Delta t]$

$\text{Prob}[t<T_a<t+\Delta t] = (e^{-\lambda t})(\lambda \Delta t e^{-\lambda \Delta t})$

Let $\Delta t \to 0$ results in the following

$$\text{Prob}[t < T_a < t + dt] = f_{T_a}(t)dt = \lambda e^{-\lambda t}dt$$

so

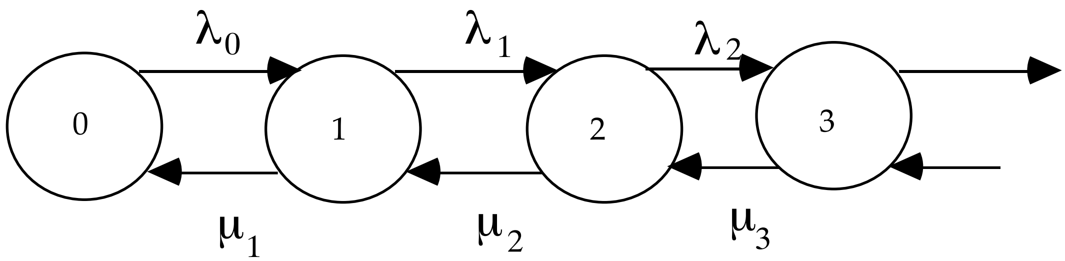
$$f_{T_a}(t) = \lambda e^{-\lambda t} \text{ for } t > 0 \quad f_{T_a}(t) = 0 \text{ for } t < 0$$

MAIN RESULT:
The interarrival time for a Poisson arrival process follows an exponential probability density function.

$E[T_a]=1/\lambda$ and $Var[T_a] = 1/\lambda^2$

λ = Arrival rate

Birth/Death Process, a model of a general queueing system



Birth/Death Process

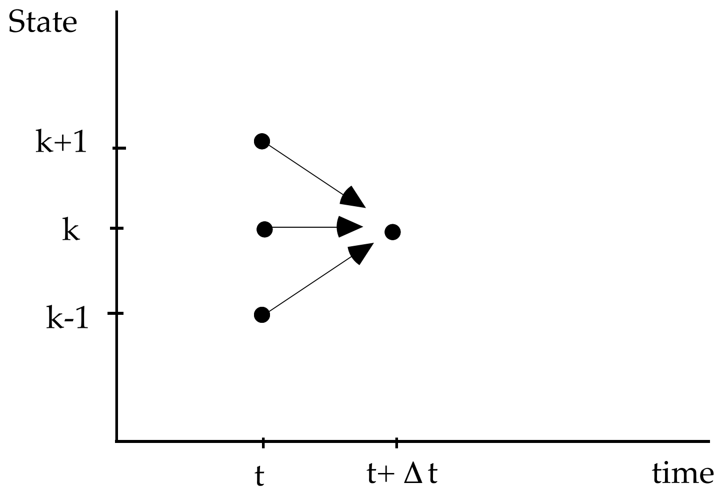
The departure process is Poisson, i.e.,

+ Prob[1 departure in Δt sec when the system is in state k] = $\mu_k \Delta t$
+ Prob[0 departure in Δt sec when the system is in state k] = $1- \mu_k \Delta t$
+ Number of departures in non-overlapping intervals of times are statistically independent random variables

Note that the arrival and service rates are now state dependent.

MAIN RESULT:
The message length is assumed to have an exponential probability density function.

Following the procedure used for the pure birth process using the transitions shown below



results in the following:

$$\frac{dP[0,t]}{dt} = -\lambda_0 P[0,t] + \mu_1 P[1,t] \; for \; k = 0$$

and

$$\frac{dP[k,t]}{dt} = -(\lambda_k + \mu_k) P[k,t]$$
$$+ \lambda_{k-1} P[k-1,t] + \mu_{k+1} P[k+1,t]$$

$$for \quad k > 0$$

We also know that

$$\sum_{k=0}^{\infty} P[k,t] = 1$$

The above relationships can be used to find P[k,t] for all k and t for some cases. However, the of most interest is the *steady state* result, i.e., when t --> infinity P[k,t] --> P[k]. The state occupancy is P[k] which can be used to determine the performance of the system.

If P[k,t] = P[k] then

$$\frac{dP[k,t]}{dt} = 0$$

With $\frac{dP[k,t]}{dt} = 0$

then

flow into state k = $\lambda_{k-1}P[k-1,t] + \mu_{k+1}P[k+1,t]$

=

flow out of state k $= (\lambda_k + \mu_k)P[k,t]$

As t--> infinity then an equilibrium or balance condition is reached where

flow into state k = flow out of state k.

The balance condition also requires that,

flow from state k to k-1 =
flow from state k-1 to k

or

$$\mu_k P[k] = \lambda_{k-1}P[k-1]$$

Now state occupancy probabilities P[k] , in steady state, can be found.

$$for\ k = 1 \quad P[1] = \frac{\lambda_0}{\mu_1} P[0]$$

$$for\ k = 2 \quad P[2] = \frac{\lambda_1}{\mu_2} P[1] = \frac{\lambda_1}{\mu_2}\frac{\lambda_0}{\mu_1} P[0]$$

$$for\ k = 3 \quad P[3] = \frac{\lambda_2}{\mu_3} P[2] = \frac{\lambda_2}{\mu_3}\frac{\lambda_1}{\mu_2}\frac{\lambda_0}{\mu_1} P[0]$$

$$for\ k \quad P[k] = \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j} P[0]$$

Next P[0] is found knowing that

$$\sum_{k=0}^{\infty} P[k] = 1$$

$$\sum_{k=0}^{\infty} P[k] = 1 = P[0] + \sum_{k=1}^{\infty} P[k] = P[0] + \sum_{k=1}^{\infty} P[0] \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j} =$$

$$P[0](1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j})$$

so

$$P[0] = \frac{1}{(1 + \sum_{k=1}^{\infty} \prod_{j=1}^{k} \frac{\lambda_{j-1}}{\mu_j})}$$

Specific queueing systems are modeled by setting λk and μk and solved using the above result.

Queueing System Notation
(Kendall's notation)

a/b/m/K/L

a = type of arrival process
b = type of service process
m = number of servers
K = maximum number of elements allowed in the system
L = population size

Special cases:

a = M => the arrival process is Poisson and the interarrival times are independent, identically distributed exponential random variables.

b = M => the departure process is Poisson and the interdeparture times are independent, identically distributed exponential random variables.

a or b = G=> times are independent, identically distributed general random variables.

a or b = D => times are deterministic, i.e., fixed times

Case 1: The M/M/1
(or M/M/1/ ∞ / ∞)queue

No limitation on buffer size means that the arrival rate is independent of state or

$$\lambda_k = \lambda$$

Only one server means that the service rate is independent of state or

$$\mu_k = \mu$$

Solving for the state occupancy probabilities

$$P[k] = \rho^k (1 - \rho)$$

where

$$\rho = \frac{\lambda}{\mu} = \lambda \mu_X$$

where

$$\mu_X = \text{ Average message length in sec } = \frac{L}{C}$$

where

L = Average message length in bits

C = Link capacity in b / s

Note ρ is the traffic intensity and $\rho < 1$.

If ρ is not less than 1 then the systems is not stable and the buffer occupancy grows without bound.

The expected number in the systems is

$$E[K] = \frac{\rho}{1-\rho}$$

and the variance is

$$Var[K] = \frac{\rho}{(1-\rho)^2}$$

Case 2: A finite buffer: the M/M/1/N queue

Only one server means that the service rate is independent of state or

$$\mu_k = \mu$$

The limitation on buffer size means that the arrival rate is dependent of state or

$\lambda_k = \lambda$ for k < N

$\lambda_k = 0$ for k $\geq$ N

Arrivals to a full system are blocked so there can be no arrivals to a full system.

Solving for the state occupancy probabilities

$$P[k] = \frac{(1-\rho)\rho^k}{1-\rho^{N+1}} \ \ for\ k \leq N$$

$$P[k] = 0 \ \ for\ k > N$$

The Quality of Service (QoS) metric in this case is the probability of blocking.

For a M/M/1/N queue the blocking probability is given by

Probability of Blocking = P_B =

$$P_{k=N} = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}$$

Case 3 M/M/S/N

The limitation on buffer size means that the arrival rate is dependent of state or

$\lambda_k = \lambda$ for k < N

$\lambda_k = 0$ for k $\geq$ N

Arrivals to a full system are blocked so there can be no arrivals to a full system.

Multiple servers means that

$$\mu_k = k\mu \ \ for\ k \leq S$$
$$\mu_k = S\mu \ \ for\ k > S$$

This model is difficult to solve in general. The case of special interest is S=N. This case models a circuit switch system with N transmission facilities. A call arriving to the system with all transmission facilities busy is blocked.

Solving for the state occupancy probabilities

$$P[k] = \frac{\dfrac{\rho^k}{k!}}{\displaystyle\sum_{n=0}^{N} \frac{\rho^n}{n!}}$$

and

$$P_B = P[k = N] = \frac{\dfrac{\rho^N}{N!}}{\displaystyle\sum_{n=0}^{N} \frac{\rho^n}{n!}}$$

This blocking equation is called the Erlang B formula.

Example:

1. A department has 140 phone, each phone is busy 10% of the time during the busy hour.

How many line do you need to buy from the phone company to keep the blocking probability less than 2%.

Solution:

Traffic intensity = 14 Erlangs
From Erlang Table 14 Erlangs & 2% blocking ==> 21 lines

Example:

Design of a building phone system. The design goal is to minimize the number of lines needed between the building and the phone company. The blocking QoS is specified to be 5%.

A building has four floors, on each floor is a separate department. Each department has 22 phones, each busy 10% of the time during the busy hour.

Consider the following two design alternatives :

a) Acquire one telephone switch for each floor.

2.2 Erlangs/floor & B=5% gives
5 lines/floor or 20 lines for the building.

b) Acquire one telephone switch for the building.

88 phones @ .1 Erlang/phone = 8.8 Erlangs

8.8 Erlangs & B=5% gives
13 lines for the building

Select alternative b)

Delay Analysis for the M/M/1 Queue

Average delay =
average transmission time + queueing time

queueing time = (average number of messages in queue ahead)(average time to serve each message)

$$\text{Average number ahead} = \frac{\rho}{1-\rho}$$

$$\text{Aveage time / message} = \frac{1}{\mu} = \frac{L}{C} = \mu_X$$

$$\text{E[T]} = \text{Average delay} = \frac{1}{\mu} + \frac{1}{\mu}\left(\frac{\rho}{1-\rho}\right)$$

$$\text{E[T]} = \frac{1}{\mu(1-\rho)} = \frac{\mu_X}{(1-\rho)}$$

Little's Result

$$E[N]=\lambda E[T]$$

Example:

Find the average delay for the following case:

L = 1000 bits
C = 1200 b/s
λ = 1/2 packets/sec

$$\rho = \frac{(0.5 \text{ packets / sec})(1000\text{bits / packet})}{1200\text{bits / sec}} = \frac{5}{12}$$

$$E[N] = \frac{\rho}{1-\rho} = \frac{5}{7}$$

$$E[T] = \frac{E[N]}{\lambda} = \frac{\frac{5}{7}\text{ packets}}{0.5 \text{ packets / sec}} = 1.43\,\text{sec}$$