COMP 190-088: Systems Performance Analysis

System Modeling - III

Jasleen Kaur

Department of Computer Science
The University of North Carolina at Chapel Hill

Spring 2005

Markov Processes

- Future states of the process depend only on the present state
  - Independent of past trajectory and states
- Markov Chain:
  - A Discrete-state Markov process
- For continuous-time processes, current state knowledge suffices
  - Do not need to know how long system has been in current state
  - Possible only if state-time has an Exponential distribution
    - ☹: Limited applicability
- Can be used to model M/M/m queues
  - Time spent by a job in the queue: Markov Process
  - Number of jobs in queue: Markov Chain

Exponential Distributions Are Memoryless

$$f(x) = \lambda e^{-\lambda x} \qquad F(x) = P(\tau < x) = 1 - e^{-\lambda x} \qquad mean = 1/\lambda ; \sigma = 1/\lambda$$

- If there is an arrival at t = 0, the mean time to next arrival is: 1/λ
  - Suppose we don't see an arrival till t = x
  - Distribution of remaining time till next arrival

$$P(\tau - x < t \mid \tau > x) = \frac{P(x < \tau < x + t)}{P(\tau > x)}$$
$$= \frac{\left(1 - e^{-\lambda (x+t)}\right) - \left(1 - e^{-\lambda x}\right)}{e^{-\lambda x}}$$
$$= \frac{e^{-\lambda x}\left(1 - e^{-\lambda t}\right)}{e^{-\lambda x}} = 1 - e^{-\lambda t}$$

- Same as distribution of time till next arrival at t = 0
  - Mean time to next arrival is still : 1/λ, regardless of x

Time can be factored out of the analysis!

Birth-death Processes

- Discrete-space Markov processes in which the transitions are restricted to neighboring states only
- Possible to represent states by integers,
  - A process in state n can transition only to states (n+1) or (n-1)
- eg, number of jobs in a single-server queue with non-bulk arrivals
  - Arrivals result in a transition to state (n+1)
  - Departures result in a transition to state (n-1)

Poisson Processes

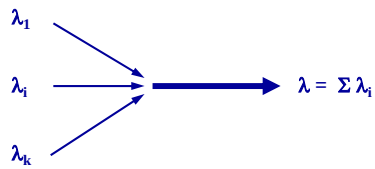
$$P(X = x) = \lambda_p{}^x \frac{e^{-\lambda_p}}{x!}$$

- If inter-arrival times are IID **exponentially** distributed with rate λ, number of arrivals over a given interval (t, t+i) have a Poisson distribution with mean: $\lambda_p = \lambda * i$
- Popular in queuing theory since arrivals are memoryless
- P.A.S.T.A.
  - **P**oisson **A**rrivals **S**ee **T**ime **A**verages
  - S1: Sample of system state collected at **all** time instances
  - S2: Sample of system state collected just before a job arrival
  - If arrivals are Poisson, averages of the two samples will be the same
    - Not true for other arrival processes !

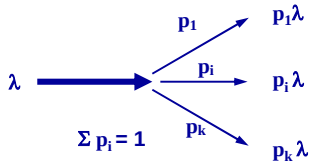
Discrete-time Markov Chains can be used to estimate continuous-time averages

Additional Useful Properties of Poisson Streams

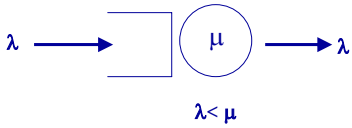
Merging of Poisson streams results in a Poisson stream



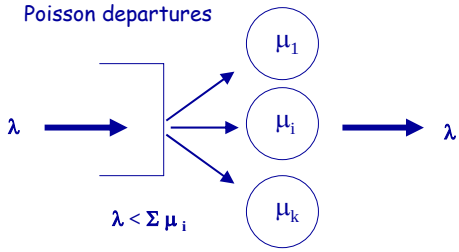
Probabilistic splitting of Poisson streams results in Poisson sub-streams



Poisson arrivals to a stable single exponential server result in Poisson departures



Poisson arrivals to a stable set of m exponential servers result in Poisson departures

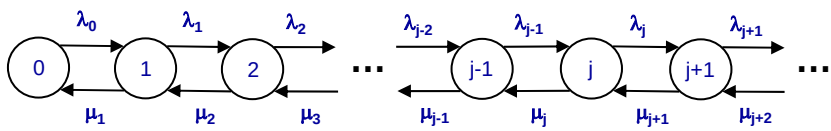


System Modeling Overview

- Queuing Basics
- **Single-server Analysis**
- Multiple-server Analysis
- Operational Laws
- Case Studies
  - Processor Scheduling
  - Disk Scheduling
  - Memory Management
- Network of Queues

Birth-death Processes

- Jobs arrive one at a time (no batch arrivals)
- State of the system can be represented by the number of jobs, n(t)
  - Arrival of a job changes state to (n+1)
  - Departure of a job changes state to (n-1)
- State transition diagram:



- Inter-arrival times and service times are exponentially distributed

What is the probability of system being in state n?

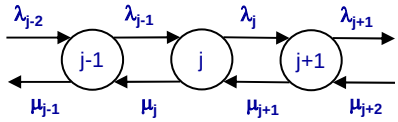
Formulating State Transition Probabilities

- If system is in state j, in the next (small) time interval Δt,
  - System can move to state (j+1) or state (j-1) with probabilities:

$$P\{n(t+\Delta t)=j+1 \mid n(t)=j\}=\lambda_j \Delta t \qquad P\{n(t+\Delta t)=j-1 \mid n(t)=j\}=\mu_j \Delta t$$

  - System will stay in state j (no arrivals/departures) with probability:

$$P\{n(t+\Delta t)=j \mid n(t)=j\}=1-\lambda_j \Delta t-\mu_j \Delta t$$



- If $p_j(t) = P[n(t)=j]$,

$$p_0(t+\Delta t)=(1-\lambda_0 \Delta t)p_0(t)+\mu_1 \Delta t.p_1(t)$$
$$p_1(t+\Delta t)=\lambda_0 \Delta t.p_0(t)+(1-\mu_1 \Delta t-\lambda_1 \Delta t)p_1(t)+\mu_2 \Delta t.p_2(t)$$
$$p_2(t+\Delta t)=\lambda_1 \Delta t.p_1(t)+(1-\mu_2 \Delta t-\lambda_2 \Delta t)p_2(t)+\mu_3 \Delta t.p_3(t)$$
$$\ldots$$
$$p_j(t+\Delta t)=\lambda_{j-1} \Delta t.p_{j-1}(t)+(1-\mu_j \Delta t-\lambda_j \Delta t)p_j(t)+\mu_{j+1} \Delta t.p_{j+1}(t)$$

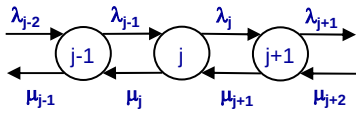
Steady-state Transition Probabilities

- j-th equation can be rewritten as:

$$\lim_{\Delta t \to 0} \frac{p_j(t+\Delta t)-p_j(t)}{\Delta t}=\frac{dp_j(t)}{dt}=\lambda_{j-1}p_{j-1}(t)+\mu_{j+1}p_{j+1}(t)-(\mu_j+\lambda_j)p_j(t)$$

- In steady-state:
  - Probability of being in a state j does not change with time
  - Amount of flow into a state is equal to the amount of flow out

$$\lim_{t \to \infty} p_j(t)=p_j \qquad \lim_{t \to \infty} \frac{dp_j(t)}{dt}=0$$



- Substitution leads to:

$$p_{j+1}=\left(\frac{\mu_j+\lambda_j}{\mu_{j+1}}\right)p_j-\frac{\lambda_{j-1}}{\mu_{j+1}}p_{j-1}, \quad j=1,2,3,\ldots \qquad p_1=\frac{\lambda_0}{\mu_1}p_0$$

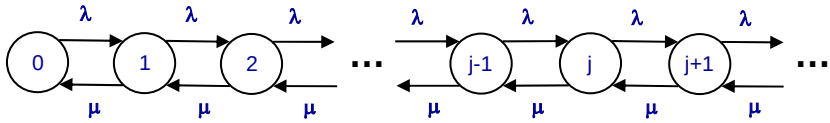
- Solution:

$$p_n=\frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}p_0, \quad n=1,2,\ldots,\infty$$

$$p_0=\frac{1}{1+\sum_{n=1}^{\infty}\frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}}$$

M/M/1 Queues

- M/M/1 ⇒ M/M/1/∞/∞/FCFS
  - Special case of a birth-death process
  - Analysis needs only arrival rate (λ) and service **rate** (μ)



- Steady-state probabilities (if ρ= λ/μ):
  - Obtained by substituting in the formulation for birth-death processes

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 = \rho^n p_0, \quad n = 1,2,\ldots,\infty \qquad p_0 = \frac{1}{1+\rho+\rho^2+\cdots+\rho^\infty} = 1-\rho$$

- ρ is called the traffic intensity
- n is geometrically distributed

$$\boxed{p_n = \rho^n(1-\rho), \quad n = 0,1,2,\ldots,\infty}$$

Deriving Properties of M/M/1 Queues

- Server utilization:
  - Probability of having at least one job in the system: = 1 – p_0 = ρ
- Mean number of jobs in the system: $E[n] = \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$
- Variance in n: $Var[n] = E[n^2] - (E[n])^2 = \frac{\rho}{(1-\rho)^2}$
- Mean number of jobs in queue: $E[n_q] = \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$
- Mean response time (E[r]):
  - Use Little's Law: E[r] = E[n] / λ
- Probability of n or more jobs in the system:

$$P(\geq n.\,jobs) = \sum_{j=n}^{\infty} p_j = \sum_{j=n}^{\infty} (1-\rho)\rho^j = \rho^n$$

Response and Wait Times in an M/M/1 Queue

- CDF of response time: $F(r) = 1 - e^{-r\mu(1-\rho)}$
- CDF of waiting time: $F(w) = 1 - \rho e^{-w\mu(1-\rho)}$
- Response time grows exponentially with traffic intensity
  - Sharp rise in response time as utilization approaches 100%
  - For an M/M/1 queue to be stable, ρ must be less than 1