Lecture 6

Prof. Krishna R. Pattipati

Dept. of Electrical and Computer Engineering
University of Connecticut
Contact: krishna@engr.uconn.edu (860) 486-2890

EE 336
Stochastic Models for the Analysis of Computer Systems and Communication Networks

Outline

- Jackson Networks
- Applications of Jackson Networks
  - Computer Systems
  - Delay Analysis in Communication Networks
  - Capacity Assignment Problem
- Single-class Closed Queuing Networks
  - Local and Global Balance Equations
  - Analysis via a two node network (equivalence to M|M|1|N network)
  - Insights from the Model

Product-form Networks

- We considered a general graph $G = (V,E)$ where V is the set of nodes $\{1,2,\ldots..,M\}$ and E is the set of ordered pairs denoting directed arcs.
- Arrival rate λ from the source
  p_{si}= Probability that a customer goes to node i on arrival
  $\Rightarrow$ External arrival rate to node $i = \lambda p_{si}$
- Routing probabilities P_{ji}=Probability {a customer departing node j will go next to node i}; $i = 1,2,\ldots..M$; $j = 1,2,\ldots..M, d$
- Markov chain is irreducible and aperiodic

$$v_i = p_{si} + \sum_{j=1}^{M} P_{ji} v_j \quad \Rightarrow \quad \underline{v} = \underline{p}_s + P^T \underline{v} \quad (OR) \quad \underline{v} = (I - P^T)^{-1} \underline{p}_s$$

v_i = average # of visits to node i by a customer

- Service demand at each node is s_i exponentially distributed
- Service rate functions : $\mu_i(n) = n\mu_i \Rightarrow$ infinite server; $\mu_i(n) = \mu_i \Rightarrow$ single server; $\mu_i(n) = \min(n,m)\, \mu_i \Rightarrow$ multi-server; $\mu_i(n) = \{\, \mu_i(1)\ldots\, \mu_i(m_i)\} \Rightarrow$ state-dependent node

Jackson's Decomposition Theorem

- The steady state distribution of the number of customers at each node $p(n_1, n_2, \ldots\ldots n_M)$ is a **product** of the state probabilities at the individual nodes of the network

$$p(n_1, n_2, \ldots., n_M) = \prod_{i=1}^{M} \frac{(\lambda v_i s_i)^{n_i}}{\prod_{k=1}^{n_i} \mu_i(k)} p_i(o) = \prod_{i=1}^{M} p_i(n_i)$$

- We can apply our earlier results on M/M/1, M/M/m, M/M/∞ and birth-death processes with the following interpretations: $\lambda \to \lambda$; $\mu(n) \to \dfrac{\mu_i(n)}{v_i s_i}$

- Nework Measures:

$$Network\ Queue\ Length: Q = \sum_{i=1}^{M} Q_i$$

$$Network\ Response\ Time\ R = \frac{Q}{\lambda}$$

$$Bottleneck\ node: k = \arg\min_i \left\{ \frac{\mu_i(m_i)}{v_i s_i} \right\}$$

Infinite and Single Server Nodes

- Infinite server nodes:

$$p_i(n_i) = \frac{\rho_i^{n_i} e^{-\rho_i}}{n_i!};\ \rho_i = \frac{\lambda v_i s_i}{\mu_i}$$

$$Q_i = \rho_i$$

$$R_i = \frac{v_i s_i}{\mu_i} \quad \text{(over all visits)}$$

$$U_i = 0$$

Poisson process with rate ρ_i

If need only Q_i, R_i, U_i, don't need $p_i(k)$

- Single server nodes:

$$p_i(n_i) = (1-\rho_i)\rho_i^{n_i}$$

$$Q_i = \frac{\rho_i}{1-\rho_i}$$

$$R_i = \frac{v_i s_i}{\mu_i(1-\rho_i)} \text{ over all visits}$$

$$U_i = \rho_i$$

Modified geometric

If need only Q_i, R_i, U_i, don't need $p_i(k)$

Multi-server Node

- Multi-server node:

$$p_i(n_i) = \begin{cases} \left(\dfrac{\lambda v_i s_i}{\mu_i}\right)^{n_i} \cdot \dfrac{1}{n_i\,!}\, p_i(o); & n_i \le m_i \\ \left(\dfrac{\lambda v_i s_i}{\mu_i}\right)^{m_i} \cdot \dfrac{1}{m_i\,!} \left(\dfrac{\lambda v_i s_i}{\mu_i m_i}\right)^{n_i - m_i} p_i(o); & n_i > m_i \end{cases}$$

$$p_i(o) = \left[1 + \sum_{k=1}^{m_i-1} \left(\frac{\lambda v_i s_i}{\mu_i}\right)^k \cdot \frac{1}{k\,!} + \frac{(m_i \rho_i)^{m_i}}{m_i\,!} \cdot \frac{1}{1-\rho_i}\right]^{-1}; \rho_i = \frac{\lambda v_i s_i}{m_i \mu_i}$$

$$p_i(k) = \begin{cases} \dfrac{\lambda v_i s_i}{k \mu_i}\, p_i(k-1); 1 \le k \le m_i - 1 \\ \rho_i\, p_i(k-1); k \ge m_i \end{cases}$$

$$Q_i = \frac{\rho_i}{1-\rho_i}\left[1 + \sum_{k=1}^{m_i-1} (m_i - n)\, p_i(k-1)\right]$$

Note that we need the distribution for $0 \le k \le m_i$ - 2 only

$$R_i = \frac{Q_i}{\lambda}$$

$$U_i = \sum_{k=1}^{m_i-1} \frac{k}{m_i} p_i(k) + \sum_{k=m_i}^{\infty} p_i(k) = \frac{\lambda v_i s_i}{m_i \mu_i}\left[\sum_{k=1}^{m_i-1} p_i(k-1) + \sum_{k=m_i}^{\infty} p_i(k-1)\right] = \frac{\lambda v_i s_i}{m_i \mu_i} = \rho_i$$

State-dependent Node

- State-dependent node: $\{\mu_i(1), \mu_i(2), \ldots, \mu_i(m_i)\}$

$$Q_i = \frac{\rho_i}{1-\rho_i}\left\{1 + \sum_{k=1}^{m_i-1}\left[\frac{\mu_i(m_i)}{\mu_i(k)} - 1\right] p_i(k-1)\right\}; \quad \rho_i = \frac{\lambda v_i s_i}{\mu_i(m_i)}$$

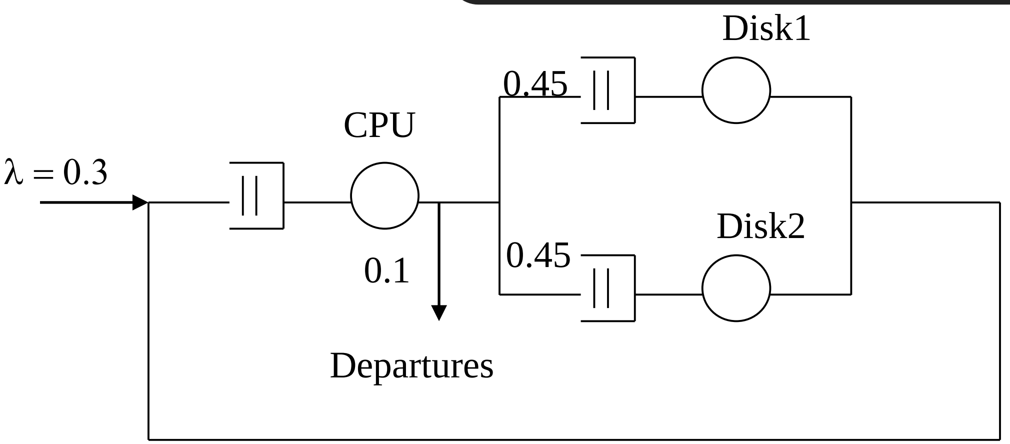
As in multi server case, $p_i(k)$ is obtained from

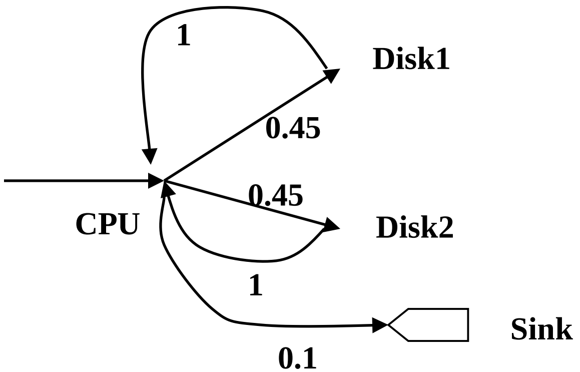
$$p_i(o) = \left[1 + \sum_{k=1}^{m_i-1} \frac{(\lambda v_i s_i)^k}{\prod_{l=1}^{k} \mu_i(l)} + \frac{(\lambda v_i s_i)^{m_i}}{\prod_{l=1}^{m_i} \mu_i(l)} \cdot \frac{1}{1-\rho_i}\right]^{-1}$$

$$\boxed{\begin{aligned} p_i(k) &= \frac{\lambda v_i s_i}{\mu_i(k)} p_i(k-1); \quad 1 \le k \le m_i - 1 \\ p_i(k) &= \frac{\lambda v_i s_i}{\mu_i(m_i)} p_i(k-1); \quad k \ge m_i \end{aligned}}$$

$$\boxed{R_i = \frac{Q_i}{\lambda}; \; U_i = \sum_{k=1}^{m_i-1} \frac{\mu_i(k)}{\mu_i(m_i)} p_i(k) + \sum_{k=m_i}^{\infty} p_i(k) = \frac{\lambda v_i s_i}{\mu_i(m_i)}}$$

Illustrative Example





$$\lambda = 0.3\ Jobs/\sec;\ s_{cpu} = 50000\ Instr;\ s_{D_1} = 50\ blocks;\ s_{D_2} = 50\ blocks$$

$$\mu_{cpu} = 10^6\ Instr/\sec \qquad \mu_{D_1} = 100\ blocks/\sec \qquad \mu_{D_2} = 200\ blocks/\sec$$

$$\left.\begin{aligned} v_{cpu} &= 1 + v_{D_1} + v_{D_2} \\ v_{D_1} &= 0.45 v_{cpu} \\ v_{D_2} &= 0.45 v_{cpu} \end{aligned}\right\} \Rightarrow \begin{aligned} v_{cpu} &= 10 \\ v_{D_1} &= 4.5 \\ v_{D_2} &= 4.5 \end{aligned}$$

$$\Rightarrow \rho_{cpu} = \frac{\lambda v_{cpu} s_{cpu}}{\mu_{cpu}} = \frac{0.3(10)5.10^4}{10^6} = 0.15;\ \rho_{D_1} = \frac{\lambda v_{D_1} s_{D_1}}{\mu_{D_1}} = \frac{0.3(4.5)50}{100} = 0.675$$

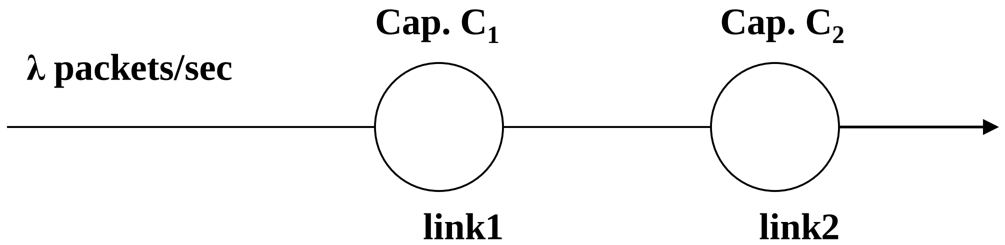
$$\rho_{D_2} = \frac{\lambda v_{D_2} s_{D_2}}{\mu_{D_2}} = \frac{0.3(4.5)50}{200} = 0.3375$$

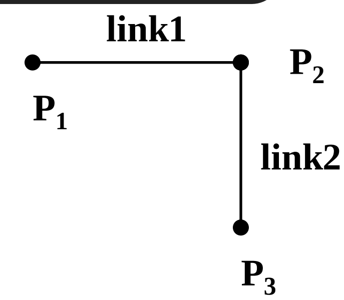
$$Q_{cpu} = \frac{0.15}{0.85} = 0.18; \quad Q_{D_1} = \frac{0.675}{0.325} = 2.07; \quad Q_{D_2} = \frac{0.34}{0.62} = 0.52$$

$$R_{cpu} = 0.60\ \sec; \quad R_{D_1} = 6.9\ \sec; \quad R_{D_2} = 1.71\ \sec$$

$$R = R_{cpu} + R_{D_1} + R_{D_2} = 9.21\ \sec; \quad Q = 2.77\ ;\ Bottleneck\ node: Disk\ 1;\ \lambda_{sat} = \frac{\mu_{D_1}}{v_{D_1} s_{D_1}} = \frac{1}{2.25} = 0.44\ \ Jobs/\sec$$

Delays in Communication Networks -1





- Packet lengths are exponentially distributed with mean *s*
- Inter-arrival times are independent of packet length
- First link is M/M/1 queue. Second link is not M/M/1. why?
  - *The service times at the two links are strongly correlated, since the same message must go through both links.*
  - Indeed, inter arrival times at the second link are strongly correlated with the packet lengths. To see this, consider the busy period of link 1.
    - Inter arrival time at link 2 between two such packets = transmission time of second packet. so, long packet will wait less time at the second link, since their transmission time at the first link takes longer, thereby giving the second link more time to empty out.

Fast cars ○ ○ ○ ○ ○ sees lot of empty space

slow truck

No analytical solutions exist for such dependent queuing processes

Delays in Communication Networks -2

- It is even worse for communication networks ⇒ need to make some assumptions
- Consider several packet streams following different paths

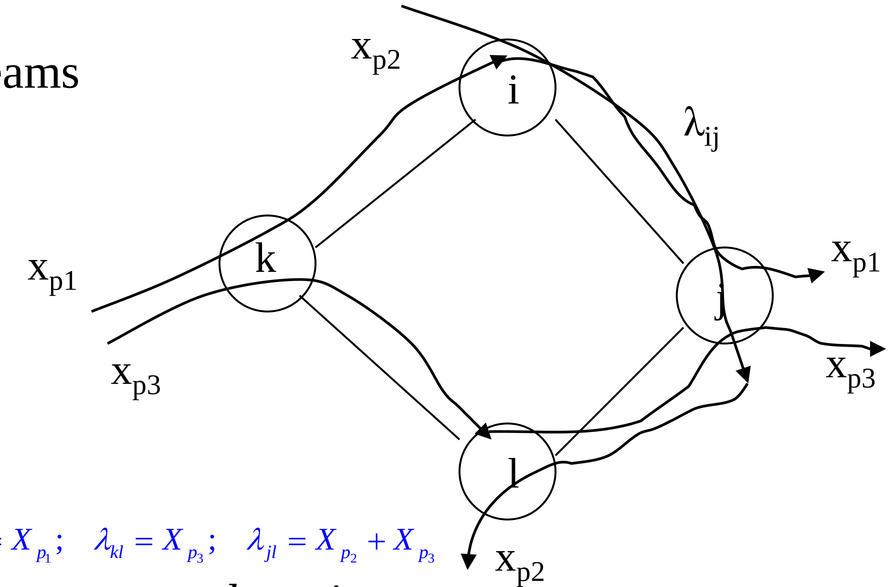
  A path p consists of a sequence of links :

  $p_1 = \{(k, i),(i, j)\}$

  $p_2 = \{(i, j),(j, l)\}$

  $p_3 = \{(k, l),(l, j)\}$

  (links are bi-directional)



- Link flows : $\lambda_{ij} = X_{p_1} + X_{p_2}$; $\lambda_{ki} = X_{p_1}$; $\lambda_{kl} = X_{p_3}$; $\lambda_{jl} = X_{p_2} + X_{p_3}$
- Link flows depend on *input streams and* *routing*
- In general, $\lambda_{ij} = \sum_{\substack{all\ p\ traversing \\ link\ (i,j)}} x_p$

We have just seen that even for two link tandem queue, even if the packet streams are Poisson with independent packet lengths at their point of entry into the network, this property is lost after the first transmission line

Delays in Communication Networks -3

- Kleinrock's independence assumption ⇒ Make it into a Jackson network
  - Based on simulation results, it was found that *merging of several packet streams on a link has an effect similar to restoring the independence of inter arrival times and packet lengths.* Indeed, this assumption is quite accurate for networks with
    1. Poisson arrivals to nodes (external traffic)
    2. Packet lengths are exponentially distributed
    3. Densely-connected networks
    4. Moderate-to-heavy traffic loads.

Each link is an M/M/1 queue with arrival rate λ_{ij} packets/sec, capacity of link μ_{ij} bits/sec and packet lengths s bits/packet

$$\Rightarrow \quad \rho_{ij} = \frac{\lambda_{ij} s}{\mu_{ij}}$$

$$Q_{ij} = \frac{\rho_{ij}}{1-\rho_{ij}} = \frac{\lambda_{ij} s}{\mu_{ij} - \lambda_{ij} s}$$

Delays in Communication Networks -4

- Total number of customers in the network : $Q = \sum_{(i,j)} Q_{ij} = XR$

  X = Throughput in packets/sec = $\sum_{p} X_p$ = total external traffic = γ
- Average response time (or delay) per packet

$$\Rightarrow \quad \boxed{R = \frac{1}{\gamma} \sum_{(i,j)} \frac{\lambda_{ij} s}{\mu_{ij} - \lambda_{ij} s}}$$

- If there is a propagation delay and processing delay of d_{ij} sec/bit

$$\boxed{R = \frac{1}{\gamma} \sum_{(i,j)} \left[\frac{\lambda_{ij} s}{\mu_{ij} - \lambda_{ij} s} + \lambda_{ij} s d_{ij} \right]}$$

- Response time over path p is given by

$$\boxed{R_p = \sum_{\substack{all\,(i,j)\,on \\ path\ p}} \left[\frac{s}{\mu_{ij} - \lambda_{ij} s} + d_{ij} s \right] = \sum_{\substack{all\,(i,j)\,on \\ path\ p}} \left[\frac{s}{\mu_{ij}(\mu_{ij} - \lambda_{ij} s)} + \frac{s}{\mu_{ij}} + d_{ij} s \right]}$$

- Research issues :
  - Independence assumption is crucial. Can we relax this?
  - Can we relax exponential packet length assumption? **Only approximately.**

Capacity Assignment Problem - 1

- Problem : Optimize link capacities

Know $\lambda_{ij} \Rightarrow$ Know routing. Want to find the best μ_{ij}

$$\boxed{\begin{aligned}&\min_{\mu_{ij}} \sum_{(i,j)} \mu_{ij} c_{ij}; \quad c_{ij} = \text{cost of link } (i,j)\\ &s.t. \frac{1}{\gamma} \sum_{(i,j)} \left[\frac{\lambda_{ij} s}{\mu_{ij} - \lambda_{ij} s} + \lambda_{ij} s d_{ij} \right] \le \bar{R}_1\end{aligned}}$$

- Equivalent problem :

$$\boxed{\begin{aligned}&\min_{\mu_{ij}} \sum_{(i,j)} \mu_{ij} c_{ij}\\ &s.t. \frac{1}{\gamma} \sum_{(i,j)} \frac{\lambda_{ij} s}{\mu_{ij} - \lambda_{ij} s} \le \bar{R}; \qquad \bar{R} = \bar{R}_1 - \frac{1}{\gamma} \sum_{(i,j)} \lambda_{ij} s d_{ij}\end{aligned}}$$

- Append the constraint with a Lagrange multiplier β>0. At optimum, strict equality.

Capacity Assignment Problem - 2

$$L(\beta,\mu_{ij}) = \sum_{(i,j)}\left[\mu_{ij}c_{ij} + \frac{\beta}{\gamma}\cdot\frac{\lambda_{ij}s}{\mu_{ij}-\lambda_{ij}s} - \beta\overline{R}\right]$$

$$\frac{\partial L}{\partial \mu_{ij}} = 0 \Rightarrow c_{ij} - \frac{\beta}{\gamma}\cdot\frac{\lambda_{ij}s}{(\mu_{ij}-\lambda_{ij}s)^2} = 0$$

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \frac{1}{\gamma}\sum_{(i,j)}\frac{\lambda_{ij}s}{\mu_{ij}-\lambda_{ij}s} = \overline{R}$$

From first equation :

$$\mu_{ij} = \lambda_{ij}s + \sqrt{\frac{\beta\lambda_{ij}s}{\gamma c_{ij}}}$$

From second equation :

$$\overline{R} = \frac{1}{\gamma}\sum_{(i,j)}\frac{\lambda_{ij}s}{\sqrt{\frac{\beta\lambda_{ij}s}{\gamma c_{ij}}}} = \sum_{(i,j)}\sqrt{\frac{c_{ij}\lambda_{ij}s}{\gamma\beta}}$$

or

$$\sqrt{\beta} = \frac{1}{\overline{R}}\sum_{(i,j)}\sqrt{\frac{c_{ij}\lambda_{ij}s}{\gamma}}$$

Capacity Assignment Problem - 3

So, $$\mu_{ij} = \lambda_{ij} s + \frac{1}{\bar{R}} \left(\sqrt{\frac{\lambda_{ij} s}{\gamma c_{ij}}} \right) \cdot \sum_{(m,n)} \sqrt{\frac{c_{mn} \lambda_{mn} s}{\gamma}}$$

$$= \lambda_{ij} s + \frac{1}{\gamma \bar{R}} \left(\sqrt{\frac{\lambda_{ij} s}{c_{ij}}} \right) \cdot \sum_{(m,n)} \sqrt{c_{mn} \lambda_{mn} s}$$

$$= \lambda_{ij} s \left[1 + \frac{1}{\gamma \bar{R}} \frac{\left(\sum_{(m,n)} \sqrt{c_{mn} \lambda_{mn} s} \right)}{\sqrt{\lambda_{ij} s c_{ij}}} \right]$$

Square-root channel capacity assignment

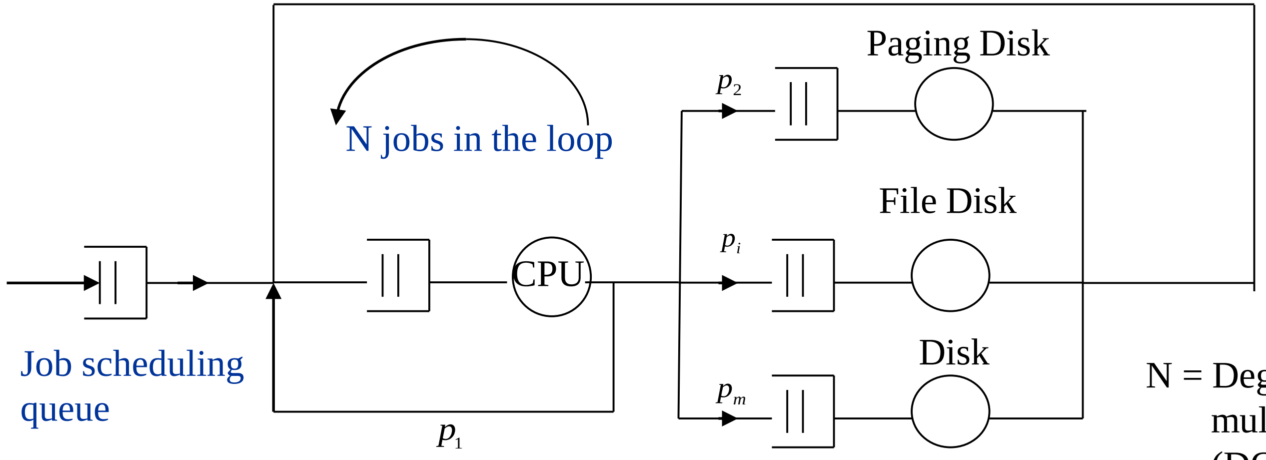
$$Optimal \quad Cost = \sum_{(i,j)} \lambda_{ij} s c_{ij} + \frac{1}{\gamma \bar{R}} \left[\sum_{(m,n)} \sqrt{c_{mn} \lambda_{mn} s} \right]^2$$

- Research Problems :

1. Channel capacities come in discrete quantities ⇨ Integer programming problem
2. Want to min. w.r.t λ_{ij} (i.e. routing) and μ_{ij}
3. May want to include reliability constraints w.r.t. connectivity

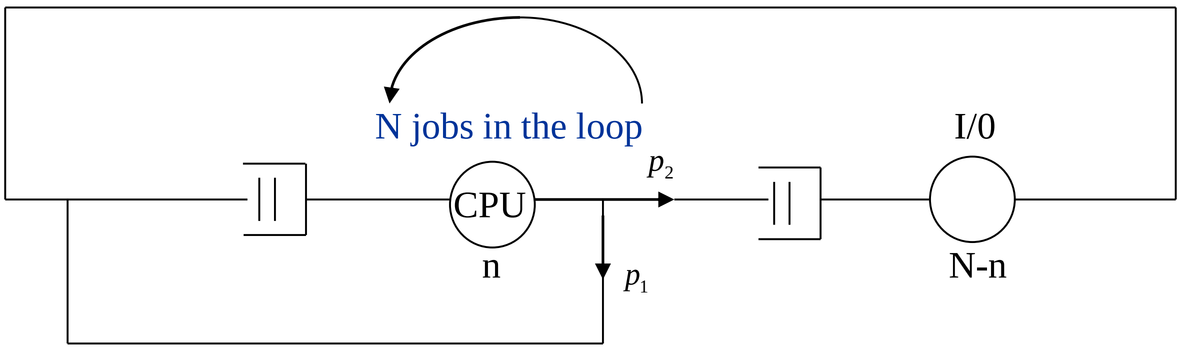
Closed Queuing Networks

- Central server model



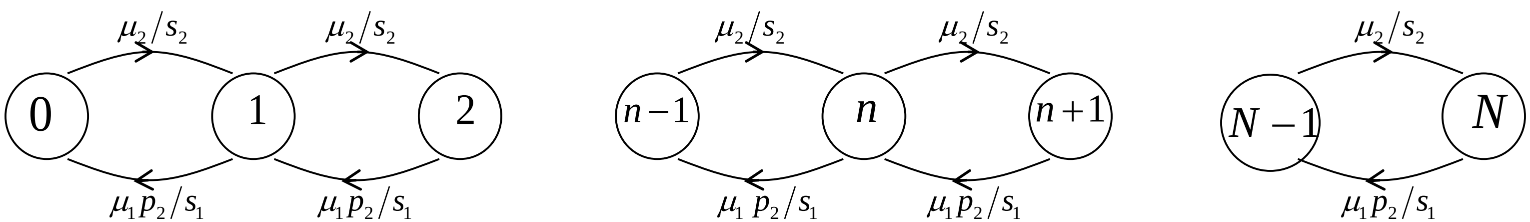
N = Degree of multi-programming (DOM)

- **Let us consider a s**imple two-node closed queuing network first



Assumptions of the Model

- Lengths of successive CPU bursts are exponentially distributed random variables with mean s_1 instructions. Instruction execution rate of the CPU is μ_1 instructions/sec $\Rightarrow$ service time per visit = s_1/μ_1
- Successive I/O bursts are also exponentially distributed with mean data transfer of s_2 words. Transfer rate is μ_2 words/sec $\Rightarrow$ service time per visit = s_2/μ_2
- Routing : At the end of CPU bursts, a program completes execution with probability p_1 or requires an I/O operation with probability $p_2 = (1 - p_1)$. As soon as a program completes execution, another statistically equivalent program enters the system so that the number in the system, termed the degree of multi-programs is constant



Similar to M/M/1/N queue

Detailed Balance Equations -1

$$\frac{\mu_1 p_2}{s_1} p(n/N) = \frac{\mu_2}{s_2} p(n-1/N) \Rightarrow p(n/N) = \frac{\mu_2 s_1}{\mu_1 p_2 s_2} p(n-1/N) = \rho p(n-1/N)$$

$$\rho = \frac{s_1}{\mu_1 p_1} \cdot \frac{p_1 \mu_2}{p_2 s_2} = \frac{\text{CPU service time}}{\text{I/O service time}}$$

$$\sum_{n=0}^{N} p(n/N) = 1 \Rightarrow p(0/N) = \frac{1}{1+\rho+\rho^2+....+\rho^N} = \frac{1-\rho}{1-\rho^{N+1}} = \frac{1}{G(N)}$$

$$p(0/N) = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}} & \rho < 1 \\ \dfrac{1}{N+1} & \rho = 1 \\ \dfrac{\rho - 1}{\rho^{N+1} - 1} & \rho > 1 \end{cases}$$

- Throughput

$$\text{At CPU}: X_1(N) = \frac{\mu_1}{s_1} \sum_{n=1}^{N} p(n/N) = \frac{\mu_1}{s_1}(1 - p(0 \mid N)) = \frac{\mu_1}{s_1} \cdot \frac{\rho(1-\rho^N)}{1-\rho^{N+1}}$$

$$\text{At Disk}: X_2(N) = \frac{\mu_2}{s_2} \sum_{n=1}^{N} p(N-n/N) = \frac{\mu_2}{s_2}(1 - p(N \mid N)) = \frac{\mu_2}{s_2} \cdot \frac{(1-\rho^N)}{1-\rho^{N+1}}$$

$$\text{Note}: \text{Job completion rate}: X(N) = X_1(N) p_1 \Rightarrow X_1(N) = \frac{X(N)}{p_1}; \quad X_2(N) = X(N) \cdot \frac{p_2}{p_1}$$

Detailed Balance Equations- 2

- Utilization :

CPU : $U_1 = \dfrac{X_1(N)s_1}{\mu_1} = 1 - p(0|N) = \begin{cases} \dfrac{\rho(1-\rho^N)}{1-\rho^{N+1}} & \rho \neq 1 \\ \dfrac{N}{N+1} & \rho = 1 \end{cases}$

I/O : $U_2 = \dfrac{X_2(N)s_2}{\mu_2} = 1 - p(N/N) = \begin{cases} \dfrac{1-\rho^N}{1-\rho^{N+1}} & \rho \neq 1 \\ \dfrac{N}{N+1} & \rho = 1 \end{cases}$

- Queue Length :

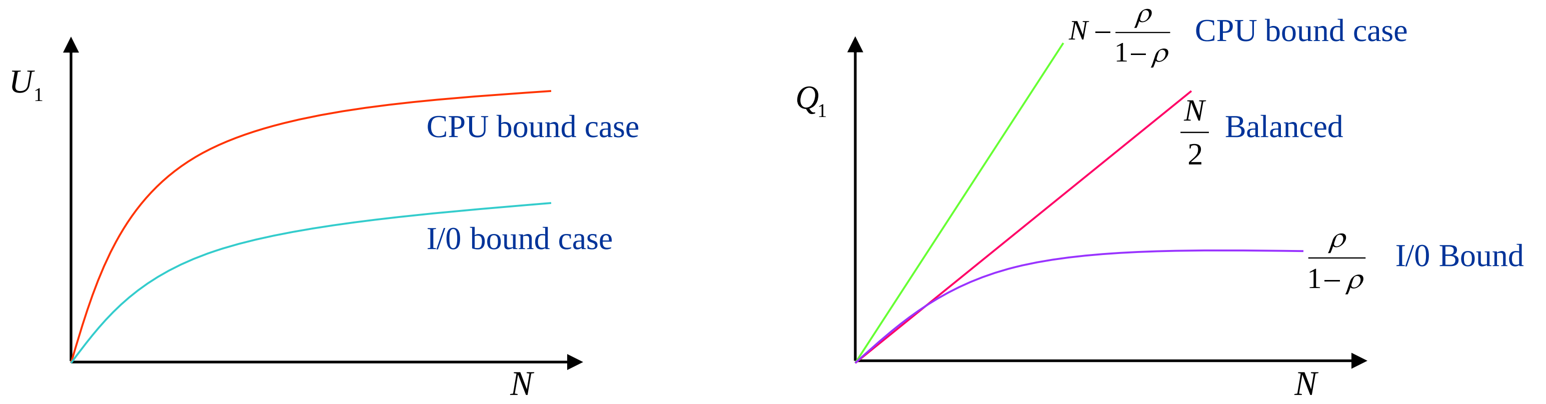
$$Q_1 = \sum_{n=1}^{N} np(n/N) = \sum_{n=0}^{N} \frac{n(1-\rho)}{1-\rho^{N+1}}\rho^n$$

$$= \frac{(1-\rho)\rho}{1-\rho^{N+1}} \cdot \sum_{n=0}^{N} \frac{d}{d\rho}(\rho^n) = \frac{(1-\rho)\rho}{1-\rho^{N+1}} \cdot \frac{d}{d\rho} \frac{\rho(1-\rho^N)}{(1-\rho)}$$

$$= \frac{(1-\rho)\rho}{1-\rho^{N+1}} \cdot \left[\frac{1-\rho^N}{1-\rho} + \frac{-N\rho^N}{1-\rho} + \frac{\rho(1-\rho^N)}{(1-\rho)^2}\right] = \left(\frac{\rho}{1-\rho}\right) \cdot [1 - \frac{(N+1)(1-\rho)\rho^N}{1-\rho^{N+1}}]$$

$$= \frac{\rho}{1-\rho}[1-(N+1)p(N|N)] \Rightarrow \textit{identical to } M|M|1|N \textit{ result (see Lecture 4)}$$

Insights from the Model



2

$\rho < 1$ As $N \to \infty, p_0 = 1-\rho, U_1 = \rho, Q_1 = \frac{\rho}{1-\rho}$ since M/M/1/N →M/M/1 queue $U_2 = 1, Q_2 = N - \frac{\rho}{1-\rho}$

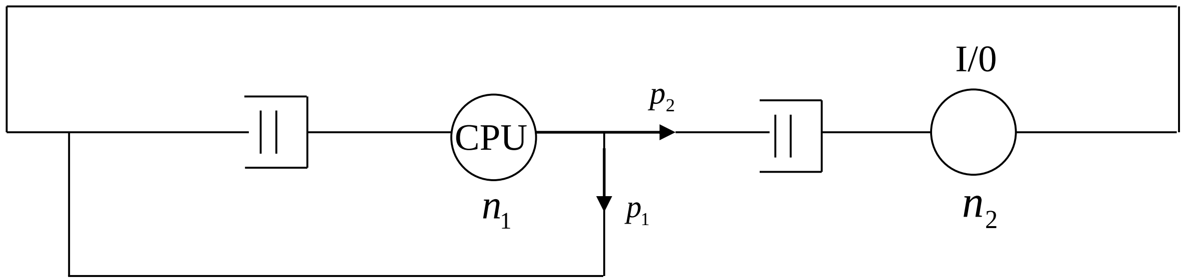
ρ < 1 ⇒ CPU service rate > I/O service rate (OR) system is I/O bound
⇒ queue length at the I/O gets arbitrarily large.
⇒ utilization of I/O → 1 look at $\frac{1-\rho^N}{1-\rho^{N+1}} \to 1$ as N→∞
⇒ I/O device becomes a Poisson source with rate $\frac{\mu_2}{s_2}$
⇒ $Q = \frac{\rho}{1-\rho}; p_0 = 1-\rho; M/M/1$ queue with arrival rate $\frac{\mu_2}{s_2}$ and service rate $\frac{\mu_1 p_1}{s_1}$

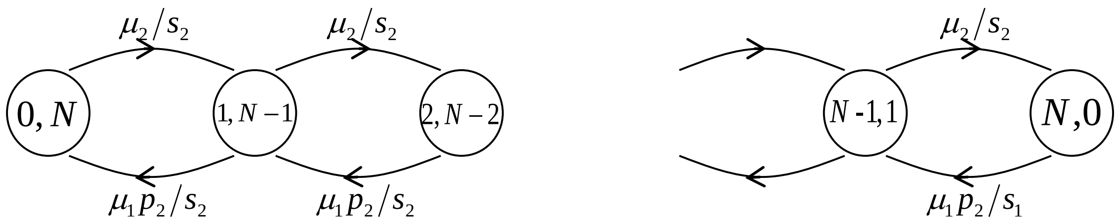
ρ > 1 ⇒ CPU service rate < I/O service rate (or) system is CPU bound as $N \to \infty, p_0 \to 0 \Rightarrow U_1 = 1\,(or)$ CPU is always busy
⇒ CPU becomes a Poisson source with rate $\frac{\mu_1 p_2}{s_1}$
⇒ Each additional increase in N will increase the queue length by 1 $\Rightarrow \frac{dQ_1}{dN} = 1$

$\rho = 1$ ⇒ Balanced ⇒ gradual increase in utilization ⇒ N/2 split in customers ⇒ Maximum Throughput

Global Balance Equations

- Let us look at the queuing system in a slightly different way





Global balance equations :

$$(1)\quad \left[\frac{\mu_2}{s_2}+\frac{\mu_1 p_2}{s_1}\right]p(n_1,n_2)=\frac{\mu_1 p_2}{s_1}p(n_1+1,n_2-1)+\frac{\mu_2}{s_2}p(n_1-1,n_2+1);n_1,n_2>0$$

$$(2)\quad \frac{\mu_2}{s_2}p(0,N)=\frac{\mu_1 p_2}{s_1}p(1,N-1)$$

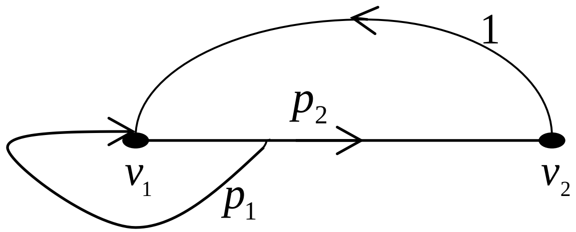
$$(3)\quad \frac{\mu_1 p_2}{s_1}p(N,0)=\frac{\mu_2}{s_2}p(N-1,1)$$

Product Form

Local balance : $\frac{\mu_2}{s_2} p(n_1 - 1, n_2 + 1) = \frac{\mu_1 p_2}{s_1} p(n_1, n_2)$

Note that local balance equation is valid when we multiply LHS and RHS by a constant. It turns out that $\exists$ infinite # of ways of specifying the Local balance equations.

Define variables v_1, v_2 ...... Known as "*visit ratios*" (or) "relative throughput"



$$\Rightarrow v_1 = p_1 v_1 + v_2$$
$$v_2 = p_2 v_1$$
$$= (1 - p_1) v_1$$
$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} p_1 & 1 \\ 1 - p_1 & 0 \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \Rightarrow \underline{v} = P^T \underline{v}$$

Facts

- $\exists$ infinite # of solutions to the equation $\underline{v} = P^T \underline{v}$
- can pick v_1 or v_2 arbitrarily

Can prove that $p(n_1, n_2) = \frac{1}{G(N)} \cdot \left(\frac{v_1 s_1}{\mu_1} \right)^{n_1} \cdot \left(\frac{v_2 s_2}{\mu_2} \right)^{n_2}$ Product form

substitute in local balance equation :

$$\frac{\mu_2}{s_2} \cdot \frac{1}{G(N)} \cdot \left(\frac{v_1 s_1}{\mu_1} \right)^{n_1 - 1} \cdot \left(\frac{v_2 s_2}{\mu_2} \right)^{n_2 + 1} \overset{?}{=} \frac{\mu_1 p_2}{s_1} \cdot \frac{1}{G(N)} \cdot \left(\frac{v_1 s_1}{\mu_1} \right)^{n_1} \cdot \left(\frac{v_2 s_2}{\mu_2} \right)^{n_2}$$

$$\frac{\mu_2}{s_2} \cdot \frac{v_1 s_1}{\mu_1} \cdot \frac{v_2 s_2}{\mu_2} \overset{?}{=} \frac{\mu_1 p_2}{s_1} \Rightarrow \frac{v_2}{v_1} \overset{?}{=} p_2$$

Yes !!

Although v_i can be specified in an infinite # of ways, $\exists$ four popular choices.

Choice of Visit Ratios - 1

- Choices for v_i :

(1) $v_1 = \frac{1}{p_1} =$ # of visits to CPU/Job

$$= 1 + p_2 + p_2^2 + \ldots\ldots = \frac{1}{1-p_2} = \frac{1}{p_1}$$

$$v_2 = p_2 v_1 = \frac{p_2}{p_1}$$

$\Rightarrow$

$\frac{v_1 s_1}{\mu_1} =$ relative service time per job at the CPU

$\frac{v_2 s_2}{\mu_2} =$ relative service time per job at the I/0

$$p(n_1, n_2) = \frac{1}{G_1(N)} \cdot \left(\frac{s_1}{p_1\mu_1}\right)^{n_1} \cdot \left(\frac{s_2 p_2}{p_1\mu_2}\right)^{n_2} = \frac{1}{G_1(N)} \cdot \left(\frac{s_1}{p_1\mu_1}\right)^{n_1} \cdot \left(\frac{s_2 p_2}{p_1\mu_2}\right)^{N-n_1}$$

$$= \frac{1}{G_1(N)} \cdot \left(\frac{s_2 p_2}{p_1\mu_2}\right)^{N} \cdot \left(\frac{\mu_2 s_1}{\mu_1 s_2 p_2}\right)^{n_1}$$

$$\sum_{n_1=0}^{N} p(n_1) = 1 \Rightarrow 1 = \frac{1 - \left(\frac{\mu_2 s_1}{\mu_1 s_2 p_2}\right)^{N+1}}{G_1(N) \cdot \left(1 - \frac{\mu_2 s_1}{\mu_1 s_2 p_2}\right)} \left(\frac{s_2 p_2}{p_1\mu_2}\right)^{N}$$

$$\therefore G_1(N) = \frac{1-\rho^{N+1}}{1-\rho} \cdot \left(\frac{s_2 p_2}{p_1\mu_2}\right)^{N};\ \rho = \frac{\dfrac{\mu_2}{s_2}}{\dfrac{\mu_1 p_2}{s_1}}$$

$$\therefore p(n_1) = \frac{1-\rho}{1-\rho^{N+1}} \cdot \rho^{n_1}$$

independent of how we select v_i

Choice of Visit Ratios - 2

Indeed, the performance measures are independent of how we choose v_i's ; choice of v_i affects only the normalization constant $G(N)$.

$$U_1 = U_{cpu} = 1 - p(0) = \frac{\rho\left(1-\rho^N\right)}{1-\rho^{N+1}}$$

$$\text{Similarly}, U_2 = \frac{\left(1-\rho^N\right)}{1-\rho^{N+1}}$$

- Some observations :

$$(i) \quad U_1 = \frac{\rho\left(1-\rho^N\right)}{1-\rho^{N+1}} = \frac{\mu_2 s_1}{\mu_1 s_2 p_2}\cdot\frac{\left(1-\rho^N\right)}{1-\rho^{N+1}} = \frac{s_1}{p_1\mu_1}\cdot\frac{\mu_2 p_1}{s_2 p_2}\cdot\frac{\left(1-\rho^N\right)}{1-\rho^{N+1}} = \left(\frac{v_1 s_1}{\mu_1}\right)\frac{G_1(N-1)}{G_1(N)}$$

$$X(N) = \frac{G_1(N-1)}{G_1(N)} \qquad \text{Throughput} = \frac{\text{normalization constant with } (N-1) \text{ customers}}{\text{normalization constant with } N \text{ customers}}$$

(ii)

$$p(n_1) = \frac{1-\rho}{1-\rho^{N+1}}\rho^{n_1} \;\square\; p(n_1/N)$$

$$p(n_1-1/N-1) = \frac{1-\rho}{1-\rho^N}\rho^{n_1-1}$$

$$\frac{p(n_1/N)}{p(n_1-1/N-1)} = \frac{1-\rho^N}{1-\rho^{N+1}}\rho = U_1$$

$$= \frac{v_1 s_1}{\mu_1} X(N)$$

$$\therefore \boxed{p(n_1/N) = \frac{v_1 s_1}{\mu_1} X(N)\, p(n_1-1/N-1)} \Leftarrow \text{Basis of MVA}$$

$$Q_1(N) = \sum_{n_1=1}^{N} n_1 p(n_1/N); R_1(N) = \frac{Q_1(N)}{X(N)}, \quad \text{we have}$$

$$\boxed{R_1(N) = \frac{v_1 s_1}{\mu_1}\left[1 + Q_1(N-1)\right]} \qquad \text{MVA equation}$$

Choice of Visit Ratios - 3

(iii) $$G_1(N)=\sum_{n_1=0}^{N}\left(\frac{v_1 s_1}{\mu_1}\right)^{n_1}\left(\frac{v_2 s_2}{\mu_2}\right)^{N-n_1}$$

$$G_1^{(2)}(N-n_1)=G_1^{(2)}(n_2)=\left(\frac{V_2 s_2}{\mu_2}\right)^{N-n_1}=\left(\frac{V_2 s_2}{\mu_2}\right)^{n_2}$$

$\Rightarrow$ *Normalization constant with node 1 removed and $(N-n_1)$ customers*

$$\boxed{G_1(N)=G_1^{(1)}(N)*G_1^{(2)}(N)}$$

- Homework :

Choice 2 : $v_1=\frac{\mu_1}{s_1}\Rightarrow v_2=\frac{\mu_1}{s_1}p_2\Rightarrow\frac{v_1 s_1}{\mu_1}=1\Rightarrow$ all utilization will be scaled by CPU utilization. $\frac{v_2 s_2}{\mu_2}=\frac{1}{\rho}$

Choice 3 : $v_1 = 1,\ v_2 = p_2$ → CPU is the reference node with 1 visit. Lavenberg's book uses this.

Choice 4 : $v_1 + v_2 = 1,\ v_2 = p_2 v_1$ …… Probability interpretation less common

Prove that all the choices lead to same utilization, throughput, etc.

Summary

- Jackson Networks
- Applications of Jackson Networks
  - Computer Systems
  - Delay Analysis in Communication Networks
  - Capacity Assignment Problem
- Single-class Closed Queuing Networks
  - Local and Global Balance Equations
  - Analysis via a two node network (equivalence to M|M|1|N network)
  - Insights from the Model