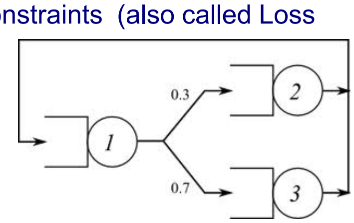
Queueing Networks II

Network Performance

David Tipper
Associate Professor
Graduate Telecommunications and Networking Program
University of Pittsburgh
Slides 6

Networks of Queues

- Many communication systems must be modeled as a set of interconnected queues – which is termed a queueing network.
- Systems modeled by queueing networks can roughly be grouped into four categories
  - Open networks
  - Closed networks
  - Networks with population constraints (also called Loss Networks)
  - Mixed network



Networks with Population Constraints (Loss Networks)

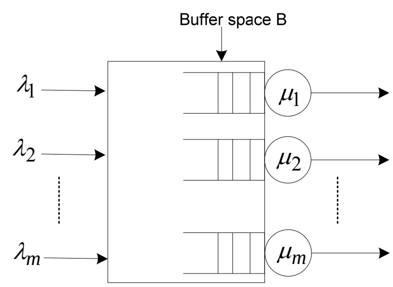
Consider *M* queue system

Customers arrive from outside the network according to a Poisson process with rate λ_i to queue *i*.

Exponential service distribution with rate μ_i at queue *i*

Total system size (waiting space) is *B*

Simple example: *M* output queues at an output buffer of a packet switch.



Networks with Population Constraints

This process is a finite state space *M* dimensional Markov process with state space

$$S = \left\{ (n_1, n_2, \dots, n_m):\ 0 \le n_i \le B \ \forall i;\ \sum_{i=1}^{m} n_i \le B \right\}$$

The steady state probability

$$P(n) = \lim_{t\to\infty} P\left\{ \tilde{n}_1(t) = n_1, \tilde{n}_2(t) = n_2, \dots, \tilde{n}_m(t) = n_m \right\}$$

P(*n*) has a product form

$$P(n) = \frac{1}{G} \prod_{i=1}^{m} \rho_i^{n_i}$$

where *G* is the normalization constant found by $\sum_{n\in S} P(n) = 1$

$$G = \sum_{n\in S} \prod_{i=1}^{m} \rho_i^{n_i}$$

$\Rightarrow$

Networks with Population Constraints

From $P(n)$ one can determine various mean performance measures.

L_i – Average number of customers in queue i.

$$L_i = \sum_{j=0}^{B} j \left(\sum_{n_i = j;\, n \in S} P(n) \right)$$

$$LN = \sum_{i=1}^{M} L_i$$

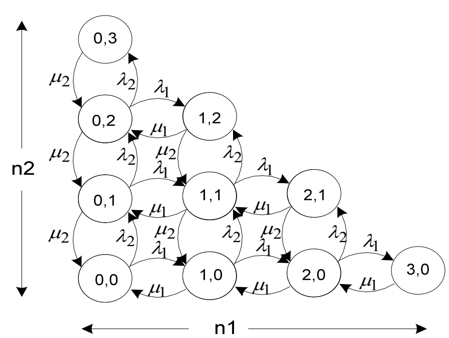
W_i – Average delay at queue i found by Little's Law

5

Networks with Population Constraints

Example: $B = 3$, $M = 2$

State diagram (n_1, n_2) : S = { (n_1, n_2) ; $0 \le n_1 \le 3$, $0 \le n_2 \le 3$, $n_1 + n_2 \le 3$ }



$$P(n_1, n_2) = \frac{1}{G} \prod_{i=1}^{m} \rho_i^{n_i} = \frac{1}{G} \rho_1^{n_1} \rho_2^{n_2}$$

Let $\lambda_1 = 0.5$, $\lambda_2 = 1$,

$\mu_1 = 1$, $\mu_2 = 1$

$\rho_1 = 0.5$, $\rho_2 = 1$

G = 6.125

$$G = 1 + \rho_1 + \rho_1^2 + \rho_1^3 + \rho_2 + \rho_1 \rho_2 + \rho_1^2 \rho_2 + \rho_2^2 + \rho_1 \rho_2^2 + \rho_2^3$$

6

Networks with Population Constraints

$$P(n_1,n_2) = \frac{1}{6.125}(0.5)^{n_1}(1)^{n_2}$$

P(2, 1) = 0.0408

$$L_1 = \sum_{n_1=0}^{3} n_1 \left(\sum_{j=n_1;\, n \in S} P(j, n_2) \right) = 1\big(P(1,0) + P(1,1) + P(1,2)\big) + 1\big(P(2,0) + P(2,1)\big) + 3P(3,0)$$

L_1 = 0.4694

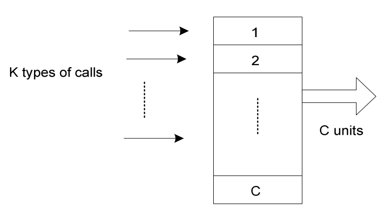
Similarly L_2 = 1.2653 $\Rightarrow LN = 1.7347$

Networks with Population Constraints

Multirate loss system : Multi-dimensional loss systems

Consider a single link in a multi-rate circuit switched network

Various services are offered and each service has different characteristics (call arrival rate, holding time, bandwidth.)



Assume *K* types of connections
each type *i* arrives according to a Poisson process rate λ_i
and have holding time exponentially holding time with a rate μ_i

(results hold for general holding time.)

Each type *i* connection requires m_i basic units of bandwidth.
The total bandwidth available is *C* units.
Chapter 7 or ITU Teletraffic Handbook

Networks with Population Constraints

Let $\tilde{n}_i(t)$ = number of type *i* connection in system at time *t*.

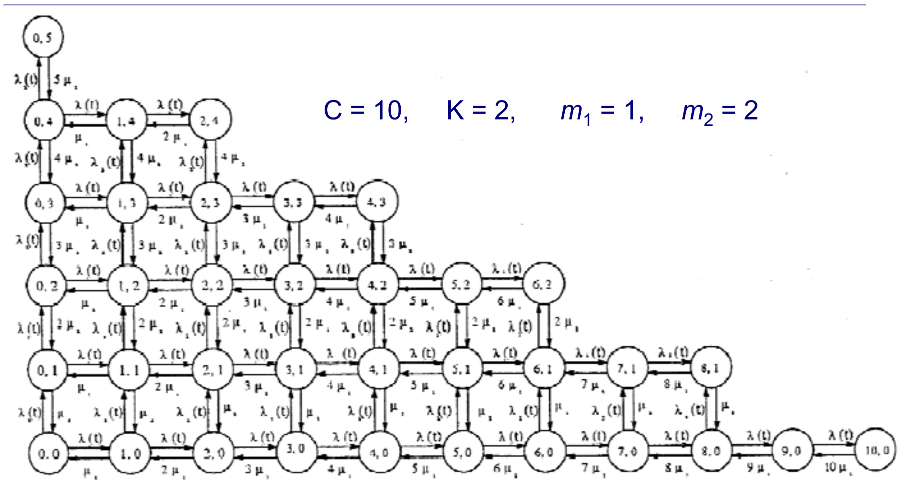
K dimensional Markov process with finite state space *S* $\left(\tilde{n}_1(t), \tilde{n}_2(t), \ldots, \tilde{n}_m(t)\right)$

$0 \le n_i \le \lfloor C/m_i \rfloor$ and $\sum_{i=1}^{K} n_i m_i \le C$

Loss Networks

C = 10, K = 2, m_1 = 1, m_2 = 2

n_2



n_1

Networks with Population Constraints

The steady state probabilities

$$P(n_1, n_2, \ldots, n_K) = \lim_{t\to\infty} P\{\tilde{n}_1(t) = n_1, \tilde{n}_2(t) = n_2, \ldots, \tilde{n}_K(t) = n_K\}$$

The product form exists where $\rho_i = \dfrac{\lambda_i m_i}{\mu_i}$

$$P(n_1, n_2, \ldots, n_K) = \frac{1}{G(k)} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!}$$

⇨ **when *K*=1, get Erlang B model M/G/C/C**
Sometimes called Generalized Erlang eq

$$G(k) = \sum_{n\in S} \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!}$$

Connection blocking rates $PB_i = \sum_{n\in S} P(n_1, n_2, \ldots, n_K)$

n where type *i* blocked ⇐ sum over states where $C - m_i < \sum_{j=1}^{K} n_j m_j$

11

Networks with Population Constraints

$$PB_2 = P(0,5) + P(1,4) + P(2,4) + P(3,3) + P(4,3) + P(5,2) + P(6,2) + P(7,1) + P(8,1) + P(9,0) + P(10,0)$$

$$PB_1 = P(0,5) + P(2,4) + P(4,3) + P(6,2) + P(8,1) + P(10,0)$$

Numerical example:

C = 48, *K*=2, *k*=1 voice 64 Kbps ⇒ $m_1 = 1$

k=2 H_{232} video 384 Kbps ⇒ $m_2 = 6$

$\lambda_1 = 15$, $\lambda_2 = 0.125$, $\mu_1 = 1$, $\mu_2 = 0.5$

Offered load $\sum_{i=1}^{K} \dfrac{\lambda_i m_i}{\mu_i} = 30$

$PB_1 = 0.0248$, $PB_2 = 0.086$

12

Circuit Switched Networks

- Model each link by Erlang B or multi-class loss system
- How to determine end-to-end blocking?
- Consider case of single traffic class (e.g., voice) of N nodes and L links .

  Let C_i be the capacity of link i and a_i be the load in Erlangs at link i ,

  $B_i(C_i, a_i)$ is the call blocking rate on link i

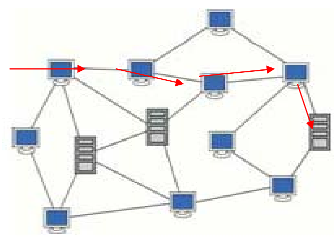
  End to End Call Blocking along a path P_{ij}

$$End \ \ to \ \ end \ \ Blocking \ \ \le \ \ 1 - \prod_{i \in P_{ij}} (1 - B_i(C_i, a_i))$$

Assumes load independent on each link

if load of single flow is small fraction on each link

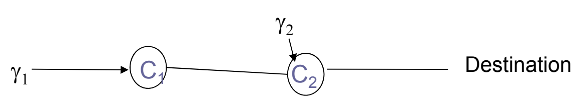
O.K. approximation



13

End to End Blocking

- Two T1 line example, offered load $\gamma_1 = 20$, $\gamma_2 = 6$ Erlangs, $C_1 = C_2 = 24$



$B_1 = B(20, 24) = .066$, $B_2 = B(26, 24) = .189$

Estimate end to end blocking $\le 1 – (1- B_1)(1- B_2) = .2425$

Note assumes traffic is independent on each link

Can improve approximation by reducing the load on link 2 to account for blocking at link 1

Thus load on link 2 $= 20*(1- B_1) + 6 = 24.68$ Erlangs ➔ $B_2 = B(24.68, 24) = 0.161$

This is called a *Modified Load Approximation or Reduced Load Approximation*

Yields → End to end blocking $\le 1 – (1- B_1)(1- B_2) = .2164$

Assumes load from source 1 *thinned* on first line before being carried on second line

14

Erlang Fixed Point Approximation

- Let A^i be the offered load in Erlang sfrom source *i* to a path from i to j
- In reality the number of calls active from source *i* to destination must be the same on each link along the path as signaling will reserve end to end resources before call is connected
- Use reduced load approximation to get estimate of load at each link

$$a_s^l \leq \quad A^i \prod_{i \in P_{ij}} (1 - B_i(C_i, a_i)) /(1 - B_s(C_s, a_s))$$

- Get a set of coupled non-linear equations – that are solved iteratively for a solution until B_i converge at each link - initialize by computing every link independently
- Can be extended to multi-class of traffic, routing etc.
- See Chapter 5. ``K. Ross , Multiservice Loss Models for Broadband Communication Networks," Springer-Verlag, 1995.

15

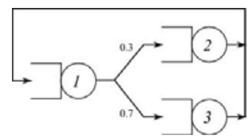
Loss Networks

- Many generalizations of Loss Systems
- General Service Times, PS queueing discipline etc.
- Several algorithms for efficient computation of G(K)

16

Networks of Queues

- Systems modeled by queueing networks can roughly be grouped into categories
  - Open networks
  - Networks with population constraints
  - Closed networks



- Looked at cases where state probabilities P(n) have a product form solution. Where C determined from normalization condition

$$P(n) = C\prod_{i=1}^{m} \rho_i^{n_i}$$

- What about networks without product form?
  - Limited results – mainly special cases or approximations

17

Remember G/G/1 KLB Approximation

- KLB approximation based on two moments of the arrival and service time distributions.

- This approximation is often used to determine the effects of increased utilizations on systems where measurement data is available to determine C_a^2 and C_s^2

$$L \approx \rho + \frac{\rho^2 (C_a^2 + C_s^2) J}{2(1-\rho)}$$

where

$J = \text{scaling factor}$

$$J = \begin{cases} e^{\frac{-2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2 + C_s^2)}} & ; C_a^2 \le 1 \\ e^{\frac{-(1-\rho)(C_a^2 - 1)}{(C_a^2 + 4C_s^2)}} & ; C_a^2 > 1 \end{cases}$$

18

QNA

Assume arbitrary network of *M* queues, define

λ_i – Total mean customer arrival rate to queue *i*.

γ_i – Mean arrival rate from outside of network to queue *i*, ← external arrivals

r_{ij} – Routing probability customer leaving queue *i* goes to queue *j*.

$r_{i(m+1)}$ – Probability customer leaving queue *i* exits the network.

μ_i – Mean service rate at queue *i*.

Co_i^2 – Squared coefficient of variation of outside arrivals to *i*.

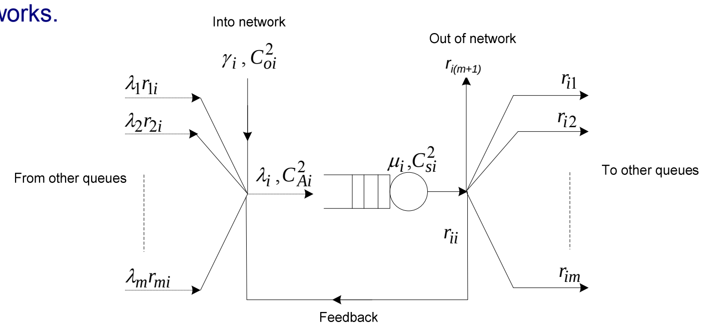
Cs_i^2 – Squared coefficient of variation of service process at *i*.

C_{Ai}^2 – Squared coefficient of variation of arrival process at *i*.

Open Network of G/G/1 Queue

Whitt's method for open's network of G/G/1 queues - Queueing Network Analysis (QNA)

- The basic idea is to use the KLB G/G/1 two moment approximation at each queue *i* in the network.
- The model of queue *i* is similar to the arbitrary queue studied in Jackson networks.



QNA

As in the Open Jackson network case, find mean arrival rate at each queue *i* by the flow conservation equation

$$\lambda_i = \gamma_i + \sum_{j=1}^{m} r_{ji}\lambda_j \qquad \Rightarrow \qquad \lambda = \gamma(I-R)^{-1}$$

To apply KLB equation need $C_{A_i}^2$ at each queue.

This requires the application of three approximations (similar to Jackson network approach) for

1. Departure process approximation
2. Spitting process approximation
3. Merging process approximation

QNA - Departure Process Approximation

Mean departure rate = λ

C_d^2 = Squared coefficient of variation of departure process.

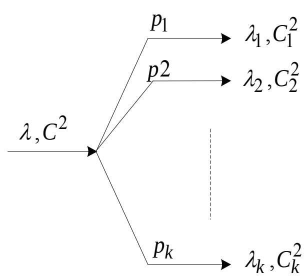
Based on renewal process approximation

$$C_d^2 \approx \rho^2 C_s^2 + (1-\rho^2)C_A^2$$

QNA – Splitting Process Approximation

If a process with mean λ and C^2 is probabilistically split into *K* stream with probabilities

$$p_i \quad \Rightarrow \quad \sum_{i=1}^{K} p_i = 1$$



We can approximate C_i^2 as below

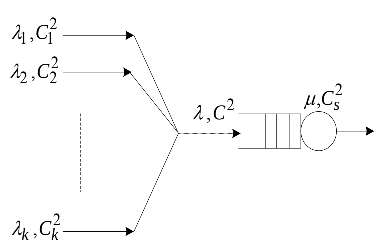
$$\lambda_i = p_i \lambda$$

$$C_i^2 \approx p_i C^2 + (1 - p_i)$$

23

QNA – Merging Process Approximation

The C^2 of a merger of *K* streams is approximated by $\lambda = \sum_{i=1}^{K} \lambda_i$



$$C^2 = 1 - W + W\left[\sum_{j=1}^{M} \frac{\lambda_j C_i^2}{\lambda}\right]$$

$$W = \left[1 + 4(1-\rho)^2 \left(\sum_{j=1}^{M} \frac{\lambda_i^{\,2}}{\lambda^2} - 1\right)\right]^{-1}$$

Combining the three approximations to determine C_{Ai}^2 at each queue *i*

24

QNA – Merging Process Approximation

$$C_{Ai}^2 = 1 - W_i + W_i\left[\frac{\gamma_i}{\lambda_i}C_{oi}^2 + \sum_{j=1}^{M}\frac{\lambda_j r_{ji}}{\lambda_i}\left[r_{ji}\left(\rho_j^2 C_{sj}^2 + (1-\rho_j^2)C_{Aj}^2\right) + (1-r_{ji})\right]\right]$$

yields a system of linear equations to solve for C_{Ai}^2

where

$$W_i = \left[1 + 4(1-\rho_i)^2\left(\sum_{j=1}^{M}\frac{(\lambda_i r_{ji})^2}{\lambda_i^2} - 1\right)\right]^{-1}$$

Once C_{Ai}^2 approximation solved can treat each queue independently and determine the mean metrics for each queue from the KLB approximation for G/G/1 queue and the network-wide measures *LN*, *WN*, etc

QNA Summary

- Given γ_i, C_{oi}^2, μ_i, C_{Si}^2, r_{ij}
- Solve for λ_i ➔ ρ_i

$$\lambda = \gamma(I-R)^{-1}$$

- Then solve for C_{Ai}^2

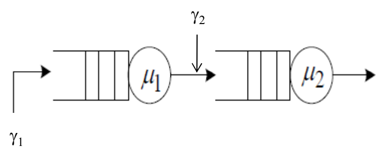
$$C_{Ai}^2 = 1 - W_i + W_i\left[\frac{\gamma_i}{\lambda}C_{oi}^2 + \sum_{j=1}^{M}\frac{\lambda_j r_{ji}}{\lambda}\left[r_{ji}\left(\rho_j C_{sj}^2 + (1-\rho_j^2)C_{Aj}^2\right) + (1-r_{ji})\right]\right]$$

$$W_i = \left[1 + 4(1-\rho_i)^2\left(\sum_{j=1}^{M}\frac{(\lambda_i r_{ji})^2}{\lambda_i^2} - 1\right)\right]^{-1}$$

- Use KLB approximation to find mean behavior for each queue
- QNA approximation tends to do pretty well on network-wide measures *LN*, *WN*, etc…, but not so well for individual queues.
- QNA implemented in several software packages (QNAP) , (RAQS)

Example

- Consider tandem queueing model below.
- Customers arrive to the first queue according to a Poisson process with mean rate γ_1 = 1.0, and $C_{o1}^2 = 1$
- Outside customers arrive to the second queue according to a deterministic process with mean rate γ_2 = 1.0, and $C_{o2}^2 = 0$
- Service process at queue one is Erlang$_2$ distributed with μ_1 = 1.2, $C_{S1}^2 = ½$ Service process at queue two is exponential with rate μ_2 = 2.2, $C_{S2}^2 = 1$



$\lambda = \gamma (I - R)^{-1}$ $\lambda_1 = 1.0$, $\lambda_2 = 2$, ➔ $\rho_1 = 1/1.2 = .833$ $\rho_2 = 2/2.2 = .9091$

27

Example

From the figure $C_{A1}^2 = C_{o1}^2 = 1$

$$W_i = \left[1 + 4(1-\rho_i)^2\left(\sum_{j=1}^{M}\frac{(\lambda_i r_{ji})^2}{\lambda_i^2} - 1\right)\right]^{-1}$$ ➔ $W_2 = 1/.9752 = 1.0254$

$$C_{Ai}^2 = 1 - W_i + W_i\left[\frac{\gamma_i}{\lambda_i}C_{oi}^2 + \sum_{j=1}^{M}\frac{\lambda_j r_{ji}}{\lambda_i}\left[r_{ji}\left(\rho_j^2 C_{sj}^2 + (1-\rho_j^2)C_{Aj}^2\right) + (1 - r_{ji})\right]\right]$$ ➔ $C_{A2} = .3093$

From KLB equation get

$$L \approx \rho + \frac{\rho^2 (C_a^2 + C_s^2) J}{2(1-\rho)}$$ ➔ $J_1 = 1$, $L_1 = 3.9583$, $J_2 = .9023$, $L_2 = 2.1774$

where

J = scaling factor

$$J = \begin{cases} e^{\frac{-2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2 + C_s^2)}} & ; C_a^2 \le 1 \\ e^{\frac{-(1-\rho)(C_a^2 - 1)}{(C_a^2 + 4C_s^2)}} & ; C_a^2 > 1 \end{cases}$$

Get LN = $L_1 + L_2$ = 6.1358, γN = 2 ➔
WN = LN/γN = 3.067

28

Summary

- Networks with Population Constraints
  - Multi-class links
  - Multi-rate links
- Networks of multi –class or rate systems
- QNA Approximation for G/G/ 1 networks

29