Lecture 10

Prof. Krishna R. Pattipati

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

ECE 336

Outline of Lecture 10

- M|G|1 Queues with vacations
- Application of M|G|1 Results to Reservations & Polling
- Application to Token Ring Networks
- Extension to Non-product form Networks with M|G|1 Nodes
- M|G|1 Queues with Priorities
- Extensions to Queuing Networks with Priorities
- G|G|1 Queues

Queues with Vacations

- Suppose that at the end of a busy period, the server goes on "vacation" for some random interval of time. Thus, a new arrival to an idle system, rather than going into service immediately, waits for the end of a vacation period.

X_1	X_2	X_3	X_4	V_1	V_2	X_5	X_6	V_3	X_7	V_4	V_5	X_8

Busy Period — Vacation — Busy — Vacation — Busy — Vacation — Busy

Usual *M/G/1*: Alternate **busy-idle-busy** cycles

M/G/1 with vacations: $\text{busy} - \underbrace{\text{vacation} - - -}_{\text{can be multiple vacation cycles}} \ \text{busy}$

- Variations:
  - The server may continue taking vacations until, on return from a vacation, it finds at least one customer … multiple vacations model
  - The server takes <u>exactly</u> one vacation. Single vacation model
    - Busy-vacation-idle-busy-vacation-busy- …cycles

Application of Queues with Vacations -1

Applications:

1. Machine breakdown:
   - → Serving Customers (Primary customers)
   - → Maintaining the system when machine fails (secondary customers)

Can also be viewed as a priority model:
- Two priorities
- Breakdowns have preemptive priority over the primary customers.

2. Maintenance in production systems:

- During idle periods, we do preventive maintenance on the system. The system is assumed to never breakdown during production (i.e., busy period)
- This is a single vacation model.

Application of Queues with Vacations -2

3. Maintenance in computer and communication systems:

- Processor in computer and communication systems perform considerable testing and maintenance to improve reliability. There exist several ways in which the maintenance functions are scheduled in these situations:
- Maintenance is divided into short segments. Whenever the primary jobs are absent, the processor does a segment of the maintenance work. If upon completion of the segment, there are no primary jobs, the system continues with the next segment of maintenance. This corresponds to "multiple vacation models" case

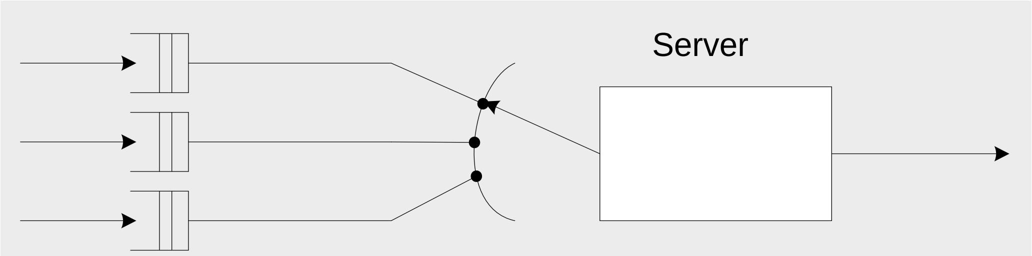
- There exist variations on this model. Some examples include:
  - Always make sure that a certain amount of time δ is spent on maintenance
  - For every *m* primary jobs, do one maintenance job. This is a limited service vacation model, in which the server takes vacation on becoming idle or after serving *m* consecutive primary jobs.
  - For every *T* seconds on primary jobs, spend on one segment on the maintenance job.
  - Periodic maintenance => secondary jobs with preemptive or non preemptive priority over primary jobs

Application of Queues with Vacations - 3

4. Cyclic server queues:

- These arise naturally as models of schedules in computer systems and communication networks, e.g., task processing in computer systems, scheduling virtual circuits or ports in a communication system.
- The basic model here has *m* classes of customers, each with its own queue
- These *m* queues are served by a single server cyclically.

Question: When does the server move from one queue to the next?



- Exhaustive Service: The server leaves a queue when it is empty
  $\Rightarrow$ Multiple vacation model
- Gated Service: The server upon arrival to a queue, closes a gate behind the waiting customers in that queue, and leaves that queue when the customers present before the gate is closed are served.

Application of Queues with Vacations - 4

- Limited Service: There is a limit R_i placed on the number of customers served on each visit to queue i. The server leaves queue i when that queue is empty or when R_i customers have been served during the current visit.

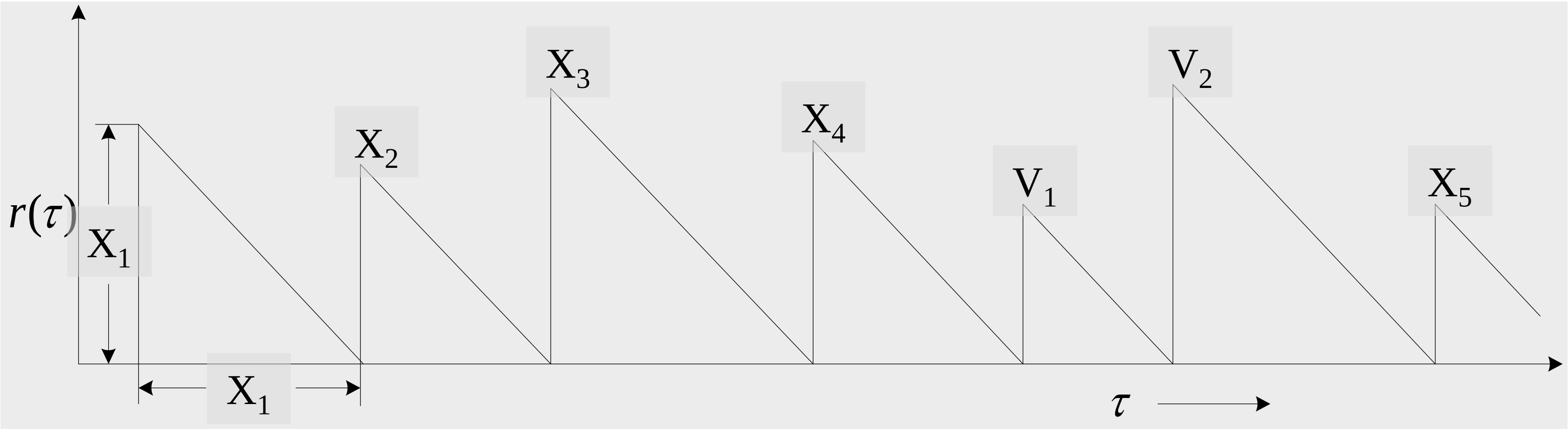
- We will consider *M/G/1* queue with vacations and applications of cyclic server queues to communication networks.
  - Poisson arrival process
  - $V_1, V_2, V_3, \ldots$ are i.i.d random variables.
  - Service times are i.i.d random variables.

- *M/G/1* multiple vacation case:
  - What does a new arrival do?
    - Wait in the queue for the completion of current service and then the service of *all* customers waiting before it.
    - Wait for vacation

$$\Rightarrow \quad W = \frac{X_R}{1-\rho}$$

X_R = mean residual time of completion of service or vacation in process when the i^{th} customer arrives.

M|G|1 Queue with Multiple Vacations - 1



Time averages = Ensemble Averages

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{t}\left[\frac{1}{2}\sum_{i=1}^{M(t)} X_i^2 + \frac{1}{2}\sum_{i=1}^{L(t)} V_i^2\right]$$

M(t) = number of service completions in *(0,t)*
L(t) = number of vacations in *(0,t)*

As $t \to \infty$,

$$\overline{X_R} = \lim_{t\to\infty}\frac{1}{t}\int_0^t r(\tau)d\tau$$

$$= \frac{1}{2}\left[\lim_{t\to\infty}\frac{M(t)}{t}\cdot\frac{1}{M(t)}\sum_{i=1}^{M(t)} X_i^2 + \frac{1}{2}\lim_{t\to\infty}\frac{L(t)}{t}\cdot\frac{1}{L(t)}\sum_{i=1}^{L(t)} V_i^2\right]$$

M|G|1 Queue with Multiple Vacations - 2

As $t \to \infty$, fraction of time occuoied with vacations is $(1-\rho)$

Total vacation time $= (1-\rho)t$

Average vacation time $\overline{V} = \dfrac{(1-\rho)t}{L(t)}$

or $\boxed{\dfrac{L(t)}{t} = \dfrac{1-\rho}{\overline{\overline{V}}}}$

Also, $\lambda = \dfrac{M(t)}{t}$

So, $\overline{X_R} = \dfrac{1}{2}\lambda\overline{X^2} + \dfrac{1}{2}(1-\rho)\dfrac{\overline{V^2}}{\overline{\overline{V}}}$

$$\boxed{\begin{aligned} W_{M/G/1/V_M} &= \frac{1}{2}\frac{\lambda\overline{X^2}}{(1-\rho)} + \frac{1}{2}\frac{\overline{V^2}}{\overline{\overline{V}}} \\ &= W_{M/G/1} + \text{ Residual Vacation Time} \end{aligned}}$$

Indeed, This decomposit ion is valid in a wider generality.

See B.T. Doshi, Journal of Applied Probability, Vol. 22, pp.419-428, 1985

M|G|1 Queue with a Single Vacation - 1

- M/G/1 queue with a single vacation: HW problem. See Doshi (1985) and Fuhurmann, Operations Research, 1984, pp.1368-1373

Result :

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\lambda \bar{V}}{B_V(\lambda) + \lambda \bar{V}} \cdot \frac{\overline{V^2}}{2\bar{V}}$$

where $B_V(\lambda) = \int_0^\infty e^{-\lambda V} f_v(V)\,dV$

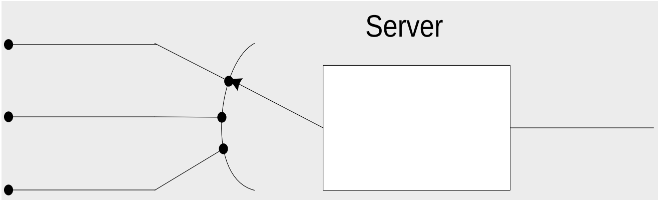
Hint:

$$(1-\rho)t = L(t)(\bar{V} + \bar{I})$$

$$I = \begin{cases} 0 \text{ if } \tau_a < V \\ \tau_a - V \text{ if } \tau_a > V \end{cases}$$

$\tau_a = \text{inter-arrival time}$

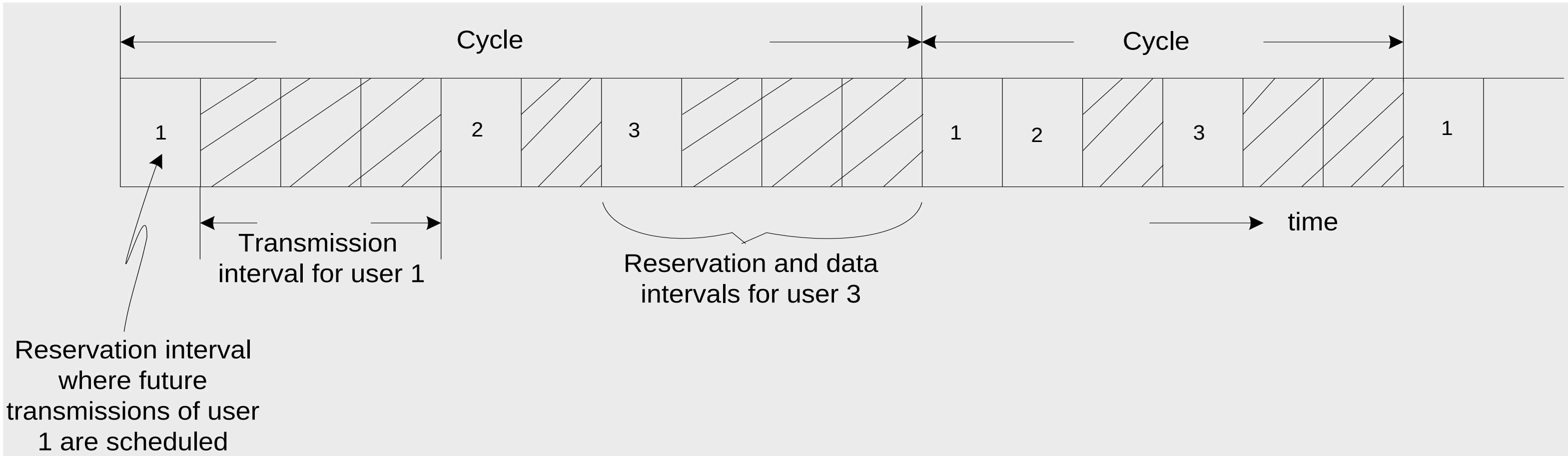
- Application to Communication Networks: Cyclic queues
  - A communication channel is accessed by several spatially separated users.
  - Only one user can transmit successfully on the channel at one time => a multi-access channel
  - Communication resource of the channel can be divided into two portions:



Packet transmissions … data intervals

Reservation (or Polling) messages that schedule future packet Transmissions … reservation intervals

Cyclic Queues - 1



- m users
- Assume that each data interval contains packets of a single user
- Reservations for these packets are made in the immediately preceding reservation interval
- All users are taken up in cyclic order (1,2,3,…,m,1,2,3,…)
- Three versions depending on how packets are transmitted during the data interval of each user

<u>Exhaustive system:</u> A packet of a user that occurs during the user's reservation or data interval is transmitted in the same data interval $\Rightarrow$ channel goes to the next user only after completing the transmission of all the packets of the current user … Token ring

Cyclic Queues - 2

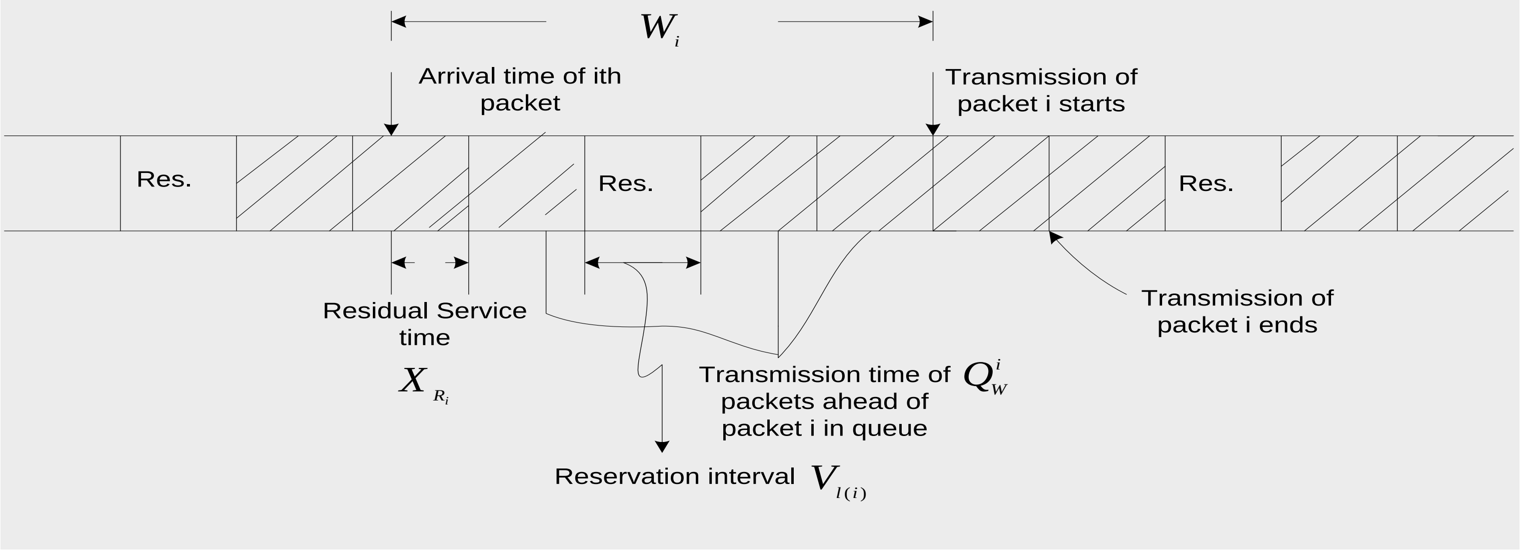
Partially gated system: Only packets that arrived until the end of the reservation interval are transmitted during the current data interval.

Gated system: Only packets that arrived prior to the reservation interval can be transmitted.

- Analysis:
  - Arrival processes of all users are Poisson with rate λ/m
  - 1st and 2nd moments of packet transmission times $\overline{X} = \frac{1}{\mu}$ and $\overline{X^2}$ (i.i.d. random variables)
  - Inter-arrival times and packet transmission times are independent
  - Reservation intervals of different users can have different distributions, but we assume it to be the same for simplicity.

Cyclic Queues - 3

- Single-user system: *m=1* (Gated)



$V_l = l^{th}$ *reservation interval. Successive reservation intervals are i.i.d with* 1^{st} *and* 2^{nd} *moments* $\overline{V}$ *and* $\overline{V^2}$

$\{V_l, X_i, \tau_a\}$ *are independent.* τ_a = *inter-arrival time*

$$\boxed{E[W_i] = E[X_{R_i}] + E[Q_{\omega_i}]\overline{X} + E[V_{l(i)}]}$$

Cyclic Queues - 4

Consider a gated system

Similar to *M/G/1* with vacations
$\Rightarrow$ a vacation starts when all previous arrivals are served.

$$X_R = \frac{\lambda \overline{X^2}}{2} + \frac{(1-\rho)\overline{V^2}}{2\overline{V}}$$

$$W = \frac{\lambda \overline{X^2}}{2} + \frac{(1-\rho)\overline{V^2}}{2\overline{V}} + \rho W + \overline{V}$$

$$\text{So,} \quad W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} + \underbrace{\frac{\overline{V}}{1-\rho}}_{\text{Cyclelength}} \ \textit{(single user, gated)}$$

Suppose $\overline{V} = A$ *(deterministic), then*

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{A}{2} + \frac{A}{1-\rho}$$

$$= W_{M/G/1} + \frac{A}{2}\left[\frac{3-\rho}{1-\rho}\right]$$

Exhaustive system: $\Rightarrow M/G/1$ with vacations

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} \qquad \overline{V} = A \Rightarrow W = W_{M/G/1} + \frac{A}{2}$$

Cyclic Queues - 5

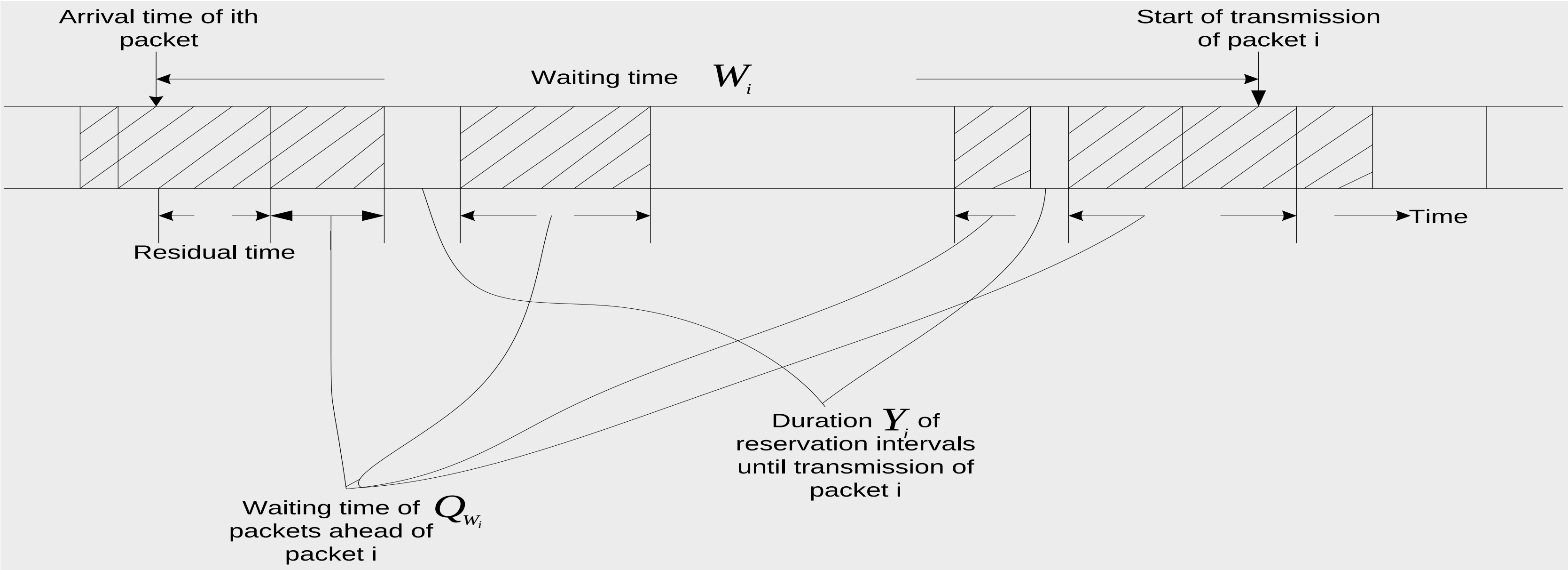
- Partially gated:

$$W = X_R + \rho W + \rho \overline{V}$$

$$(or)\, W = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{\overline{V^2}}{2\overline{V}} + \frac{\rho \overline{V}}{1-\rho}$$

$$\overline{V} = A \; deterministic \;\Rightarrow\; W = W_{M/G/1} + \frac{A}{2}\left(\frac{1+\rho}{1-\rho}\right)$$

- Multi-user system:



Cyclic Queues - 6

- *user data rates* λ/m *for users 0,1,2,...,m-1*
- l^{th} *reservation interval is used to make reservations for user* $l \bmod(m) = l - \left\lfloor \frac{l}{m} \right\rfloor .m$

  and the subsequent l^{th} *data interval is used to send packets corresponding to those reservations*
- *Consider packet* i

$$E\{W_i\} = E\{X_{Ri}\} + E\{Q_{\omega i}\}\overline{X} + E\{Y_i\}$$

as $i \to \infty$

$$W = \overline{X_R} + \rho W + Y$$

$$\Rightarrow W = \frac{\overline{X_R} + Y}{1-\rho}$$

Know $\overline{X_R} = \frac{\lambda}{2}\overline{X^2} + \frac{(1-\rho)\sum_{k=0}^{m-1}\overline{V_k^2}}{2\sum_{k=0}^{m-1}\overline{V_k}} = \frac{\lambda}{2}\overline{X^2} + \frac{(1-\rho)\overline{V^2}}{2\overline{V}}$

<u>Need to compute Y</u>: $m = 1 \Rightarrow Y = \begin{cases} 0 \text{ exhaustive} \\ \rho\overline{V} \text{ partially gated} \\ \overline{V} \text{ gated} \end{cases}$

Cyclic Queues - 7

- What happens when *m>1* ? Consider Exhaustive case

Let

$$\alpha_{ij} = E\left\{ \begin{matrix} Y_i \mid packet\ \ i\ \ arrives\ \ in\ \ user\ \ l\text{'s}\ \ reservation\ \ or\ \ data\ \ interval \\ and\ \ belongs\ \ to\ \ user\,(l+j) \bmod m \end{matrix} \right\}$$

$$\Rightarrow \alpha_{ij} = \begin{cases} 0; \ j=0 \\ \bar{V}_{(l+1)\bmod m} + \ldots + \bar{V}_{(l+j)\bmod m} & ; j = 1,2,..,m-1 \end{cases}$$

since packet i belongs to any user with probability $\frac{1}{m}$ *, we have*

$$E\{Y_i \mid packet\ \ i\ \ arrives\ \ in\ \ user\ \ l\text{'s}\ \ reservation\ \ or\ \ data\ \ interval\}$$

$$= \sum_{j=1}^{m-1} \frac{m-j}{m} \bar{V}_{(l+j)\bmod m}$$

Finally, a packet will arrive during l's data interval with probability $\frac{\rho}{m}$

a packet will arrive during l's reservation interval with probability $\frac{(1-\rho)\bar{V}_l}{\sum_{k=0}^{m-1} \bar{V}_k}$

Cyclic Queues - 8

Let $i \to \infty$

$$\Rightarrow Y = \sum_{l=0}^{m-1}\left(\frac{\rho}{m} + \frac{(1-\rho)\overline{V}_l}{\sum_{k=0}^{m-1}\overline{V}_k}\right)\sum_{j=1}^{m-1}\frac{m-j}{m}\overline{V}_{(l+j)\bmod m} = \frac{\rho(m-1)\overline{V}}{2} + \frac{(1-\rho)m\overline{V}}{2} - \frac{(1-\rho)\sum_{k=0}^{m-1}\overline{V}_k^{\,2}}{2m\overline{V}}$$

$$\Rightarrow W_{exh} = \frac{\lambda\overline{X^2}}{2(1-\rho)} + \frac{\sigma_V^2}{2\overline{V}} + \frac{m-\rho}{2}\frac{\overline{V}}{(1-\rho)};\sigma_V^2 = \frac{\sum_{k=0}^{m-1}(\overline{V_k^2} - \overline{V}_k^{\,2})}{m};\overline{V} = \frac{\sum_{k=0}^{m-1}\overline{V}_k}{m}$$

See Bertsekas & Gallagher, pp. 200 for details

Partially gated system: *Same as exhaustive, except that if a packet arrives during user's own data interval, it is delayed by an additional* $m\overline{V}$.

This occurs with probability $\dfrac{\rho}{m}$

$$\Rightarrow Y_{\rho G} = Y_{exh} + \rho\overline{V}$$

$$W_{\rho G} = W_{exh} + \frac{\rho\overline{V}}{1-\rho}$$

Cyclic Queues - 9

Gated System : If a packet arrives during a user's own reservation or data interval, it is delayed by an additional $m\bar{V}$ time units. This occurs with probability $\frac{1}{m}$

$$Y_G = Y_{exh} + \bar{V} \Rightarrow W_G = W_{exh} + \frac{\bar{V}}{1-\rho}$$

suppose $\bar{V} = \frac{A}{m}$, then

$$W_{exh} = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{A}{2}\frac{(1-\rho/m)}{(1-\rho)} = W_{M/G/1} + \frac{A}{2}\frac{(1-\rho/m)}{(1-\rho)} = W_{M/G/1/V_m} + \frac{A}{2}.\frac{m-1}{m}.\frac{\rho}{(1-\rho)}$$

$$W_{\rho G} = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{A}{2}\frac{(1+\rho/m)}{(1-\rho)}$$

$$W_G = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \frac{A}{2}\frac{(1+(2-\rho)/m)}{(1-\rho)}$$

As $m \to \infty$

$$W_{exh} = W_{\rho G} = W_G = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \underbrace{\frac{A}{2(1-\rho)}}_{1/2 \text{ cycle length}}$$

Cyclic Queues - 10

Limited service Systems: $k_i = 1$ case

- *In each user's data interval, only the first packet of the user waiting in queue(if any) is transmitted rather than all waiting packets.*
- *Consider only gated and partially gated systems (exhaustive case doesn't make sense here)*

As before

$$W = \bar{X}_R + \rho W + Y_L$$

What is Y_L ?

Consider partially gated system

A packet arriving during user l's data or reservation interval will belong to any one of the users with probability $\frac{1}{m}$. In steady state, the average number of packets waiting in the individual queue of the user that owns the arriving packet $= \frac{\lambda W}{m}$

Cyclic Queues - 11

Each of these packets cause an extra cycle of resrvations of length $m\bar{V}$*. So,*

$$Y_{L,\rho G} = Y_{\rho G} + \frac{\lambda W}{m}.m\bar{V}$$

$$\Rightarrow W_{L,\rho G} = \frac{X_R + Y_{\rho G}}{(1-\rho-\lambda\bar{V})}$$

$$= W_{\rho G}.\frac{(1-\rho)}{(1-\rho-\lambda\bar{V})}$$

Gated System :

$$Y_{L,G_i} = Y_{L,\rho G_i} + m\bar{V}\ \text{Prob}\left\{\begin{array}{l}\text{packet i arrives during the reservation interval of its owner}\\ \text{and the subsequent data interval is empty}\end{array}\right\}$$

$$\text{Prob}\{\text{packet i arrives during the resrvation interval}\} = 1-\rho$$

$$\text{Let Prob}\{\text{reservation interval followed by an empty data interval}\} \stackrel{\Delta}{=} p$$

$$\text{Prob}\{\text{reservation interval followed by a nonempty data interval}\} = 1-p$$

Cyclic Queues - 12

$$\therefore (1-p)\frac{\overline{X}}{\overline{V}} = \frac{\rho}{1-\rho} \Rightarrow \rho\overline{V} = (1-p)\overline{X} - (1-p)\rho\overline{X}$$

$$\Rightarrow 1-p = \frac{\lambda\overline{V}}{1-\rho} \Rightarrow p = \frac{(1-\rho-\lambda\overline{V})}{(1-\rho)} = 1-\frac{\lambda\overline{V}}{1-\rho}$$

So,

$$Y_{L,G_i} = Y_{L\rho G_i} + \frac{m\overline{V}(1-\rho-\lambda\overline{V})}{(1-\rho)}\frac{(1-\rho)}{m}$$

$$W_{L,G} = \left[\overline{X_R} + Y_{L,\rho G} + \overline{V}(1-\rho-\lambda\overline{V})\right] \Big/ (1-\rho-\lambda\overline{V})$$

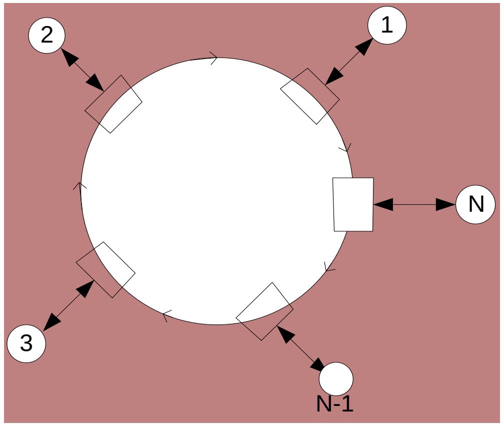
$$= W_{L,\rho G}\left(\frac{1-\rho}{1-\rho-\lambda\overline{V}}\right) + \overline{V}$$

- *Note that we need* $\lambda(\overline{X}+\overline{V})<1$ *for stability*

Application to Token Ring Networks

Application to Token ring networks:

- m *poisson streams with rate* $\dfrac{\lambda}{m}$
- $\overline{V}$ *propagation delay + relaying delay per step* $\Rightarrow \overline{V} = \dfrac{A}{m}$
- $\overline{X} = 1 \Rightarrow \rho = \lambda$



Exhaustive System:

$$W = \frac{\lambda \overline{X^2}}{2(1-\lambda)} + \frac{A}{2}\frac{m-\rho}{m(1-\rho)} = W_{M/G/1} + \frac{\overline{V}(m-\lambda)}{2(1-\lambda)}$$

$$\overline{V} = \frac{\dot{A}}{m}$$

Partially gated, limited service system:

$$W = \frac{\lambda \overline{X^2} + (m+\lambda)\overline{V}}{2(1-\lambda-\lambda\overline{V})}$$

$$\Rightarrow \text{stable if } \quad \lambda < \frac{1}{1+\overline{V}}$$

TDM vs. FDM vs. SFDM

- *Slotted time-division and frequency-division multiplexing:*

FDM: • *m poisson streams with rate* $\frac{\lambda}{m}$

• *transmission time of each packet = m time units.*

Each channel is an M/D/1 queue $\Rightarrow W_{FDM} = \frac{\lambda m}{2(1-\lambda)}$

Slotted FDM :Packet transmissions can start only at times 0,m,2m,.... M/D/1 queue with vacations

where $\overline{V} = m, \overline{V^2} = m^2$

$$W_{SFDM} = W_{FDM} + \frac{m}{2} = \frac{m}{2(1-\lambda)}$$

TDM: $W_{TDM} = W_{SFDM} = W_{FDM} + \frac{m}{2} = \frac{m}{2(1-\lambda)}$

Response Times :

$$R_{FDM} = m + \frac{\lambda m}{2(1-\lambda)}$$

$$R_{SFDM} = R_{FDM} + \frac{m}{2}$$

$$R_{TDM} = 1 + \frac{m}{2(1-\lambda)} = R_{FDM} - \left(\frac{m}{2} - 1\right)$$

$$R_{TDM} < R_{FDM} < R_{SFDM} \quad \text{for} \quad m > 1$$

Networks with General Service Times

No product form $\Rightarrow$ *Approximation*

MVA is ideally suited for these approximations

For product-form networks, have

$$R_{ij}(\underline{n}) = \frac{S_{ij}}{\mu_i}\left[1 + Q_i(\underline{n} - \underline{e}_j)\right]$$

At FCFS nodes, need $S_{ij} = S_i \quad \forall \; j$

1) Suppose S_{ij} *is different for different classes j & exponential*

$$\Rightarrow \quad R_{ij}(\underline{n}) \cong \frac{S_{ij}}{\mu_i} + \left[\sum_{k=1}^{J} Q_{ik}(\underline{n} - \underline{e}_j) S_{ik}\right] \cdot \frac{1}{\mu_i}$$

2) General service demand requirement:

$$\Rightarrow \quad R_{ij}(\underline{n}) \cong \frac{S_{ij}}{\mu_i} + \sum_{k=1}^{J}\left[Q_{ik}(\underline{n} - \underline{e}_j) - u_{ik}(\underline{n} - \underline{e}_j)\right]\frac{S_{ik}}{\mu_i} + \sum_{k=1}^{J} u_{ik}(\underline{n} - \underline{e}_j)\frac{\tilde{S}_{ik}}{\mu_i}$$

$$\tilde{S}_{ik} = \text{ residual service demand } = \frac{S_{ik}^2 + \sigma_{ik}^2}{2S_{ik}}$$

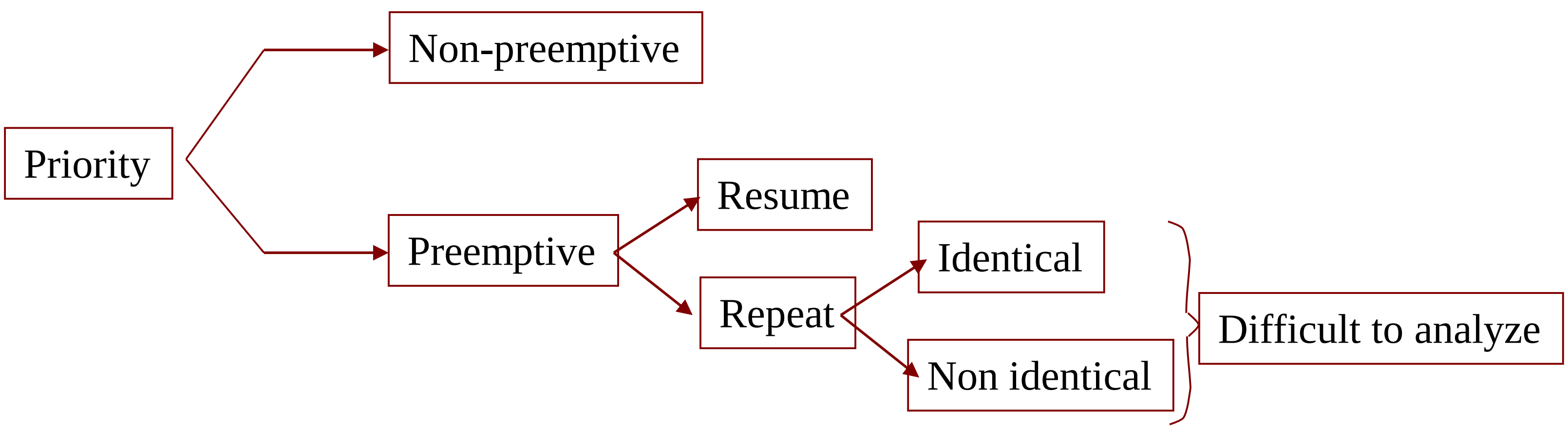
M|G|1 Queues with Priorities -1

- **References:**

  1. Kleinrock, Vol.II, Chapter 3
  2. N.K.Jaiswal, "Priority Queues", Academic Press, 1968

• Customers are divided into *J* different priority classes
  Class 1: higher priority   Class J: least priority
  Static priorities   predetermined (not dependent on waiting time, # in the system, etc.)
• Arrival processes are independent, Poisson & independent of the service times

$$\lambda_k, \quad \overline{X}_k = \frac{1}{\mu_k}, \overline{X_k^2}$$



M|G|1 Queues with Priorities - 2

- **Nonpreemptive priority:** a customer is allowed to complete service without interruption even if a customer of higher priority arrives in the mean time

  - A separate queue for each priority class
  - When the server becomes free, the first customer of the highest nonempty priority queue enters service => Head-Of- the-Line (HOL) priority.
  - Need to compute waiting time of each priority class. We appeal to conceptual reasoning rather than analytic derivation

Q_{Wk}: Average number waiting in queue k
W_k: Average waiting time for priority k customers

$\rho_k = \dfrac{\lambda_k}{\mu_k}$ System Utilization for priority k customers

$\overline{X_R}$: Mean residual time

Assume

$$\rho_1 + \rho_2 + \ldots + \rho_J < 1$$

$$\text{If not } \exists \ a\ k^* \ni \quad W_{k'+i} = \infty \quad i = 0,1,2,\ldots,J-k^*$$

$$W_1 = \overline{X_R} + Q_{W1}\overline{X_1} \quad \Rightarrow \quad W_1 = \frac{\overline{X_R}}{1-\rho_1}$$

M|G|1 Queues with Priorities - 3

$$W_2 = \overline{X}_R + Q_{W1}\overline{X}_1 + Q_{W2}\overline{X}_2 + \overbrace{\lambda_1 W_2 \overline{X}_1}^{\text{Future arrivals}}$$

$$= \overline{X}_R + \rho_1 W_1 + \rho_2 W_2 + \rho_1 W_2 \quad \Rightarrow \quad W_2 = \frac{\overline{X}_R + \rho_1 W_1}{1-\rho_2-\rho_1} = \frac{\overline{X}_R}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

In general,

$$W_k = \overline{X}_R + \rho_1 W_1 + \rho_2 W_2 + \ldots + \rho_k W_k + \left(\sum_{i=1}^{k-1}\rho_i\right) W_k$$

$$\Rightarrow \quad W_k = \frac{\overline{X}_R}{\left[1-\sum_{i=1}^{k-1}\rho_i\right]\left[1-\sum_{i=1}^{k}\rho_i\right]}$$

$$R_k = W_k + \overline{X}_k$$

$\underline{\textit{Need } \overline{X}_R}$:

$$\overline{X}_R = \frac{1}{2}\left(\sum_{i=1}^{J}\lambda_i\right)\overline{X^2}$$

$$\overline{X^2} = \frac{\lambda_1}{\sum_{i=1}^{J}\lambda_i}\overline{X_1^2} + \frac{\lambda_2}{\sum_{i=1}^{J}\lambda_i}\overline{X_2^2} + \ldots + \frac{\lambda_J}{\sum_{i=1}^{J}\lambda_i}\overline{X_J^2}$$

$$\Rightarrow \quad \overline{X}_R = \frac{1}{2}\sum_{i=1}^{J}\lambda_i \overline{X_i^2}$$

Waiting time of a high priority class (*e.g.*, W_1) depends on the arrival rates of lower priority classes

M|G|1 Queues with Priorities - 4

Note:

1) $W_k = \overline{X}_R + \sum_{l=1}^{k} \rho_l W_l + \left[\sum_{l=1}^{k-1} \rho_l\right] W_k$

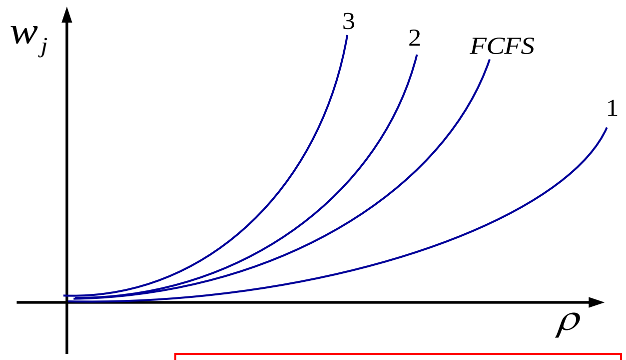
$$\sum_{k=1}^{J} \rho_k W_k = \rho \overline{X}_R + \sum_{k=1}^{J}\sum_{l=1}^{k} \rho_l \rho_k W_l + \sum_{k=1}^{J}\sum_{l=1}^{k-1} \rho_l \rho_k W_k$$

$$\sum_{k=1}^{J} \rho_k W_k = \rho \overline{X}_R + \sum_{l=1}^{J}\left(\sum_{k=l}^{J} \rho_k\right) \rho_l W_l + \sum_{k=1}^{J}\sum_{l=1}^{k-1} \rho_l \rho_k W_k$$

$$= \rho \overline{X}_R + \sum_{k=1}^{J}\left(\sum_{l=1}^{J} \rho_l\right) \rho_k W_k$$

$$\Rightarrow \sum_{k=1}^{J} \rho_k W_k = \frac{\rho \overline{X}_R}{1-\rho} \qquad \textit{independent of priorities}$$

M/G/1 Conservation law



Some customer classes do better than FCFS, Others do worse.

For multi-server queues, see Buzen, Operations Research, 1983

2) *Suppose it costs* $\$C_k$ *per unit of wait in the queue by a class k customer*

Want to minimize

$$\sum_{k=1}^{J} C_k Q_{Wk} = \sum_{k=1}^{J} \frac{C_k}{\overline{X}_k}(\rho_k W_k)$$

$$\Rightarrow \quad \min \sum_{k=1}^{J} \left(\frac{C_k}{\overline{X}_k}\right) \rho_k W_k$$

The μC Rule

Suppose classes are ordered according to [1] [2] ... [k][k+1] [k+2]...[J]... Optimal $\Rightarrow$ *Cost*$=\sum_{k=1}^{J}\frac{C_{[k]}}{\bar{X}_{[k]}}\rho_{[k]}W_{[k]}$

Suppose we interchange [k] and [k+1]

Does not affect waiting times of [1] [2] ... [k-1] & [k+2] [k+3]...[J]

$$\Rightarrow \frac{C_{[k]}}{\bar{X}_{[k]}}\rho_{[k]}W^*_{[k]}+\frac{C_{[k+1]}}{\bar{X}_{[k+1]}}W^*_{[k+1]}\rho_{[k+1]}\le\frac{C_{[k+1]}}{X_{[k+1]}}W_{[k+1]}\rho_{[k+1]}+\frac{C_{[k]}}{X_{[k]}}W_{[k]}\rho_{[k]}$$

$$\frac{C_{[k+1]}}{X_{[k+1]}}\rho_{[k+1]}\Delta W_{[k+1]}\le-\frac{C_{[k]}}{X_{[k]}}\rho_{[k]}\Delta W_{[k]};\Delta W_{[k]}=W^*_{[k]}-W_{[k]}<0;\Delta W_{[k+1]}=W^*_{[k+1]}-W_{[k+1]}>0$$

$$\Rightarrow\left[\frac{C_{[k+1]}}{X_{[k+1]}}-\frac{C_{[k]}}{X_{[k]}}\right]\rho_{[k+1]}\Delta W_{[k+1]}\le 0\quad \textit{since}\quad \sum_{k=1}^{J}\rho_{[k]}\Delta W_{[k]}=0$$

$$\Rightarrow\frac{C_{[k+1]}}{X_{[k+1]}}\le\frac{C_{[k]}}{X_{[k]}}\quad \textit{since}\quad \Delta W_{[k+1]}>0\Rightarrow C_{[k+1]}\mu_{[k+1]}\le C_{[k]}\mu_{[k]}$$

or $\quad C_{[k+1]}X_{[k]}\le C_{[k]}X_{[k+1]}$

or $\quad \frac{X_{[k]}}{C_{[k]}}\le\frac{X_{[k+1]}}{C_{[k+1]}}$

arrange priorities according to μC *rule or weighted shortest-processing time rule (WSPT)*

$$\mu_1 C_1\ge\mu_2 C_2\ge\ldots\ge\mu_J C_J$$

-μC *rule minimizes expected waiting cost*

M/G/1 with Preempt–resume Priority - 1

KEY: The waiting time of a high priority customer class is *independent of* the arrival rates of lower priority classes (unlike non-preemptive priority)

* service of a lower priority customer is interrupted when a high priority customer arrives, and is resumed from the point of interruption once all customers of higher priority have been served

Here, we find it convenient to calculate the response time rather than the waiting time. Consider class j. R_j consists of :

$$R_j = Term_a + Term_b + Term_c$$

Term a: Average service time $\bar{x}_j = 1/\mu_j$ (since preempt-resume)

Term b: Average time required, upon arrival of a priority j customer, to service customers of priorities 1 to j, that are already in the system $\Rightarrow$ average unfinished work corresponding to priorities 1 through j.

M/G/1 with Preempt–resume Priority - 2

What is Term b?

The average waiting time of an *M/G/1* queue with arrivals due to classes *1, 2, ..., j* (priorities *j+1, j+2, ..., J* are neglected)

$$Term_b = \frac{\sum_{i=1}^{j} \lambda_i \overline{x_i^2}}{2(1-\sum_{i=1}^{j}\rho_i)} = \frac{\bar{x}_{Rj}}{(1-\sum_{i=1}^{j}\rho_i)}$$

Term c: Average waiting time for customers of priorities 1 through *(j-1)* who arrive while the customer of class *j* is in the system

What is Term c?

$$Term_c = \sum_{i=1}^{j-1} \frac{1}{\mu_i}\lambda_i R_j = (\sum_{i=1}^{j-1}\rho_i) R_j$$

$$\therefore R_j = \frac{1}{\mu_j} + \frac{\bar{x}_{Rj}}{(1-\sum_{i=1}^{j}\rho_i)} + (\sum_{i=1}^{j-1}\rho_i)R_j \;\Rightarrow R_j = \frac{\frac{1}{\mu_j}(1-\sum_{i=1}^{j}\rho_i) + \bar{x}_{Rj}}{(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i)}$$

M/G/1 with Preempt–resume Priority - 3

- Can recursively evaluate R_j :

```
RHOSUM=0
```
$\overline{x_R} = 0$

Do $j = 1, \ldots, J$

TEMP=RHOSUM

RHOSUM = RHOSUM + ρ_j

$$\overline{x_R} = \overline{x_R} + \frac{\lambda_j \overline{x_j^2}}{2}$$

$$R_j = \frac{[\frac{1}{\mu_j}(1 - RHOSUM) + \overline{x_R}]}{(1 - RHOSUM)(1 - TEMP)}$$

End do

- Extension to multiple servers: Buzen, ***Operations Research***, 1983
  Agrawal ***Metamodeling,*** MIT Press 1985

 Lecture 10

Extension to Multi-class Queuing Networks - 1

- Two approximations:
  - Shadow approximation, due to *Sevcik*, valid for only preempt-resume discipline
  - *Bryant*, *Lakshmi*, *Chandy* and *Krzenski* approximation (also termed MVA approximation) …. ***The Best***

Assume only single server nodes. Infinite server nodes are easy; multi-server & state-dependent nodes are research issues.

- M nodes $\{1, 2, ..., M\}$
- J Classes $\{1, 2, ..., J\}$, class 1 has highest priority,…, class J has lowest
- visits, v_{ij} ;
- Mean service time per visit : $s_{ij} = 1/\mu_{ij}$
- R_{ij} = Response time over all visits;
- Q_{ij} = Queue length at node i for class j;
- X_j = Throughput of class j customers

Extension to Multi-class Queuing Networks - 2

- Shadow approximation:

> *K. Sevcik*, "**Priority scheduling disciplines in QN models of computing systems**", *Proc. IFIP congress*, North Holland, 1977, pp. 565-574

- Preempt resume service discipline. Assume a single PR center
- Key idea:

1. Replace each priority center by J shadow centers, where J is the number of priority classes
2. *Each shadow service center is visited by one class only*
3. Service time per visit of class j customers at the shadow service center is

$$\boxed{s_{sj} = \frac{s_{pj}}{1 - \sum_{k=1}^{j-1} \rho_{pk}}; \quad \rho_{pk} = x_k s_{pk}}$$

Solve (M+J-1) node product-form network

Extension to Multi-class Queuing Networks - 3

- Algorithm:

Initialize ρ_{pk}

While ρ_{pk} not converged Do

$$s_{sj} = \frac{s_{pj}}{1 - \sum_{k=1}^{j-1} \rho_{pk}}$$

Solve (*M*+*J*-*1*) product form network $\Rightarrow x_j$

Compute ρ_{pk}

End

- Can easily extend to multiple *preempt-resume* service centers

Extension to Multi-class Queuing Networks - 4

- Algorithm:

Initialize ρ_{pk} at all $p \in P_R$

While ρ_{pk} , $p \in P_R$ not converged Do

$$s_{sj} = \frac{s_{pj}}{1 - \sum_{k=1}^{j-1} \rho_{pk}};$$

Solve $(|P_R|J + M - |P_R|)$ node

product form network

Compute ρ_{pk}

End

- Errors can be as high as 40%!!

Extension to Multi-class Queuing Networks - 5

- MVA approximation:

  Bryant, M.S.Lakshmi, K.M.Chandy and A.E.Krzenski "**MVA Priority Approximation**", *ACM Trans. on Comp. Systems*, Feb. 1983

- Recall product-form MVA equations

  Repeat $\forall \ \underline{n} \ \in \ \underline{0} \le \underline{n} \le \underline{N}$

$$R_{ij}(\underline{n}) = v_{ij} s_{ij} [1 + Q_i(\underline{n} - \underline{e}_j)]$$

$$x_j(\underline{n}) = n_j / \sum_{i=1}^{M} R_{ij}(\underline{n})$$

$$Q_{ij}(\underline{n}) = x_j(\underline{n}) R_{ij}(\underline{n})$$

  End Loop

- Restrictions:

QD	$\sim$	PS or LCFS PR
QD	$\sim$	FCFS $\Rightarrow s_{ij} = s_i$ indpendent of j

How do we extend these results to queuing networks with priority nodes?

Extension to Multi-class Queuing Networks - 6

- Suppose we have an isolated open node with arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_J$ with visits $v_j = 1.$ Then,
  - Preempt-resume

$$R_j = s_j + \sum_{k=1}^{j} (Q_k - \rho_k) s_k + \sum_{k=1}^{j-1} R_j \lambda_k s_k + \overline{s}_{Rj};$$

$$\overline{s}_{Rj} = \frac{\sum_{k=1}^{j} \lambda_k \overline{s_k^2}}{2} = \sum_{k=1}^{j} \frac{\rho_k s_k [1 + C_{s_k}^2]}{2}; \; C_{s_k} = \frac{\sigma_{s_k}}{s_k}$$

$$\Rightarrow \quad R_j = \frac{\left[s_j + \sum_{k=1}^{j} \left\{ (Q_k - \rho_k) s_k + \frac{\lambda_k \overline{s_k^2}}{2} \right\} \right]}{1 - \sum_{k=1}^{j-1} \rho_k} \qquad *$$

$$= \frac{\left[s_j + \sum_{k=1}^{j} \left\{ Q_k s_k + \frac{\rho_k s_k (C_{s_k}^2 - 1)}{2} \right\} \right]}{1 - \sum_{k=1}^{j-1} \rho_k}$$

$$\textit{For exponential case:} \quad R_j = \frac{\left[s_j + \sum_{k=1}^{j} Q_k s_k \right]}{1 - \sum_{k=1}^{j-1} \rho_k}$$

Extension to Multi-class Queuing Networks - 7

– Non-preemptive:

$$W_j = \underbrace{\sum_{k=1}^{j}(Q_k - \rho_k)s_k}_{\substack{\text{waiting time due}\\ \text{to customers ahead of}\\ \text{our tagged customer}}} + \underbrace{\overline{s_R}}_{\substack{\text{remaining}\\ \text{service}\\ \text{time}}} + \underbrace{\sum_{k=1}^{j-1} W_j \lambda_k s_k}_{\substack{\text{waiting time due to}\\ \text{arrivals after our tagged}\\ \text{customer came in}}}$$

$$= \sum_{k=1}^{j}(Q_k - \rho_k)s_k + \sum_{k=1}^{J}\frac{\rho_k s_k[1+C_{sk}^2]}{2} + W_j \sum_{k=1}^{j-1}\rho_k$$

$$\Rightarrow R_j = W_j + s_j = s_j + \frac{\sum_{k=1}^{j}(Q_k - \rho_k)s_k + \sum_{k=1}^{J}\frac{\rho_k s_k[1+C_{sk}^2]}{2}}{1-\sum_{k=1}^{j-1}\rho_k} \qquad **$$

$$For\ Exponential\ case: R_j = s_j + \frac{\sum_{k=1}^{j}Q_k s_k + \sum_{k=j+1}^{J}\rho_k s_k}{1-\sum_{k=1}^{j-1}\rho_k}$$

Equation (*) and (**) form the basis of MVA equations for priority networks

Extension to Multi-class Queuing Networks - 8

- Consider preempt-resume priority case first
  - Assumption 1: Poisson arrival s at the nodes ------ not true in networks

$$R_{ij}(\underline{n}) = \frac{v_{ij}[s_{ij} + \sum_{k=1}^{j} Q_{ik}^{(j)}(\underline{N}) s_{ik} + \sum_{k=1}^{j} \frac{\rho_{ik}^{(j)} s_{ik}[c_{sk}^2 - 1]}{2}}{[1 - \sum_{k=1}^{j-1} \rho_{ik}^{(j)}(\underline{N})]}$$

  - $Q_{ik}^{(j)}(\underline{N}) =$ average number of class k customers at node i as seen by an arrival of a customer of class j … $Q_{ik}(\underline{N} - \underline{e}_j)$ for product-form networks
  - Assumption 2: we assume that the arrival theorem is valid

$$\Rightarrow \quad Q_{ik}^{(j)} = Q_{ik}(\underline{N} - \underline{e}_j)$$

$$R_{ij}(\underline{n}) = \frac{v_{ij}[s_{ij} + \sum_{k=1}^{j} Q_{ik}(\underline{N} - \underline{e}_j) s_{ik} + \sum_{k=1}^{j} \frac{\rho_{ik}^{(j)} s_{ik}[c_{sk}^2 - 1]}{2}}{[1 - \sum_{k=1}^{j-1} \rho_{ik}^{(j)}(\underline{N})]}$$

  - Need means of computing ρ_{ik}

Extension to Multi-class Queuing Networks - 9

– Assumption 3: Know

$$\rho_{ik}^{(j)}(\underline{N}) = x_k^{(j)}(\underline{N}) s_{ik}(\underline{N}) v_{ik}$$

$$\rho_{ik}^{(j)}(\underline{N}) = \rho_{ik}(\underline{N} - \underline{e}_j) \qquad \textit{Not Good !!}$$

$$\rho_{ik}^{(j)}(\underline{N}) = \rho_{ik}(\underline{N}) \qquad \textit{Not good when utilization of server} > 0.7$$

(Bryant−Krzenski approximation)

$$\rho_{ik}^{(j)}(\underline{N}) = \rho_{ik}(\underline{N} - Q_{ik}\underline{e}_k) \qquad \textit{Best approximation !}$$

(Chandy-Lakshmi approximation)

When there are Q_{ik} customers at node *i*, the arrival rate of class *k* at node *i* is determined by the $(N_k - Q_{ik})$ customers in the network

$$\boxed{\Rightarrow \rho_{ik}(\underline{N} - Q_{ik}\underline{e}_k) = x_{ik}(\underline{N} - Q_{ik}\underline{e}_k) s_{ik} v_{ik}}$$

– Errors generally less than 10%. Extension to non-preemptive is easy (**)

- Open problems:
  Method validated for exponential service times. General service times open.
  Extension to multi-server & state-dependent server modes
  B-S and C-N approximations for priority MVA

M/G/1 Busy periods

$$P_0 = Prob\ \ idle = \lim_{n\to\infty} \frac{I_1 + I_2 + \cdots I_n}{(I_1 + I_2 + \cdots I_n) + (B_1 + B_2 + \cdots B_n)} = \frac{E(I)}{E(I) + E(B)}$$

- For M/G/1

$$E(I) = \frac{1}{\lambda}$$

Also, know $1 - P_0 = \lambda E(x) = \lambda \bar{x} = \rho$

$$\Rightarrow\ 1 - \lambda \bar{x} = \frac{1/\lambda}{1/\lambda + \bar{B}} \qquad \Rightarrow \qquad \bar{B} = \frac{\bar{x}}{1 - \lambda \bar{x}}$$

$$\text{Average \# of customs served per busy period} = \frac{1}{1-\rho}$$

M/G/1 with Batch Arrivals

- Batch arrivals:

$$\alpha_j = Prob\{batch\ size = j\}$$

$$Expected\ \ Batch\ \ Size = E(N) = \sum_{j=0}^{\infty} j\alpha_j$$

$$W = \bar{X}_R + \lambda E(N)\bar{X}W \Rightarrow W = \frac{\bar{X}_R}{1 - \lambda E(N)\bar{X}}$$

$\bar{X}_R$ = remaining service time of customer in service + waiting time due to those in batch $= \dfrac{\lambda E(X^2)E(N)}{2} + E(W_B)$

$$E(W_B) = \sum_j E(W_B \mid batch\ size = j)Prob\{batch\ size = j\}$$

$$\Pr ob\{batch\ size = j\} = \frac{j\alpha_j}{\sum_j j\alpha_j} = \frac{j\alpha_j}{E(N)}$$

$$E(W_B \mid batch\ size = j) = \sum_{i=1}^{j} (i-1)\bar{X}\frac{1}{j} = \frac{j-1}{2}\bar{X} \Rightarrow E(W_B) = \sum_j (j-1)\alpha_j j\frac{\bar{X}}{2E(N)} = \frac{\bar{X}[E(N^2) - E(N)]}{2E(N)}$$

$$so, W = \frac{\left\{\dfrac{\lambda E(X^2)E(N)}{2} + \dfrac{\bar{X}[E(N^2) - E(N)]}{2E(N)}\right\}}{1 - \lambda E(N)\bar{X}}$$

G/G/1 Waiting Time Bound - 1

- Can get only bounds on waiting time

$$W \le \frac{\lambda(\sigma_a^2 + \sigma_x^2)}{2(1-\rho)} \qquad \textit{equality as } \rho \to 1$$

σ_a^2 : *variance of inter-arrival times*

σ_x^2 : *variance of service times*

λ: *arrival rate*

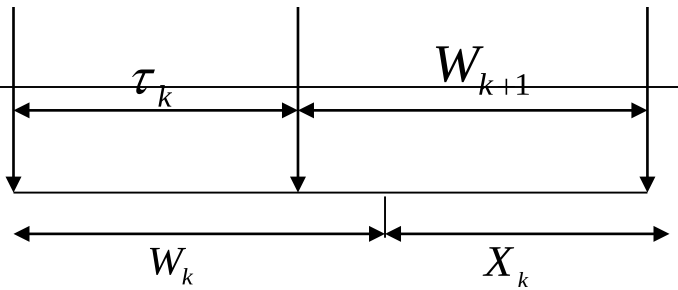
$$\rho = \frac{\lambda}{\mu}$$

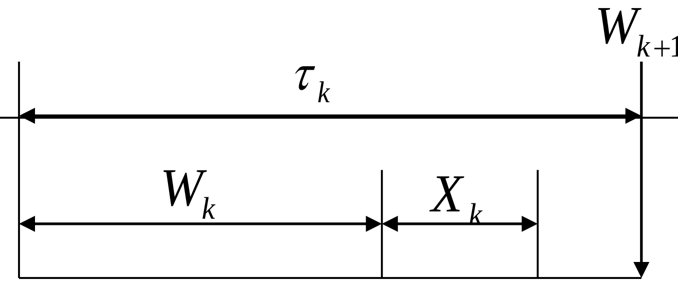
Let W_k : waiting time of k^{th} customer

X_k : service time of k^{th} customer

τ_k : inter-arrival time between k^{th} and $(k+1)^{th}$ customer

G/G/1 Waiting Time Bound - 2

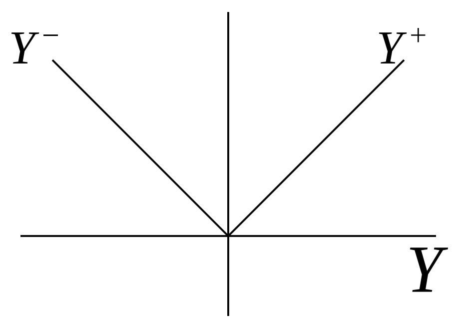




$$W_{k+1} = \max(0, \;\; W_k + X_k - \tau_k)$$
$$= \max(0, \;\; W_k + V_k)$$

- Some identities:

$$Y^+ = \max(0, Y) \qquad\qquad Y^- = -\min(0, Y)$$
$$\Rightarrow Y = Y^+ - Y^- \;\; \& \;\; Y^+ Y^- = 0$$
$$\bar{Y} = \bar{Y}^+ - \bar{Y}^- \qquad\qquad \sigma_Y^2 = \sigma_{Y^+}^2 + \sigma_{Y^-}^2 + 2\bar{Y}^+\bar{Y}^-$$



G/G/1 Waiting Time Bound - 3

$$W_{k+1} = (W_k + V_k)^+ \;\; ; V_k = X_k - \tau_k \;\; ; I_k = (W_k + V_k)^-$$

$$\sigma^2_{(W_k+V_k)} = \sigma^2_{(W_k+V_k)^+} + \sigma^2_{(W_k+V_k)^-} + 2\overline{(W_k+V_k)^+}\;\overline{(W_k+V_k)^-}$$

$$= \sigma^2_{W_k} + \sigma^2_{V_k} = \sigma^2_{W_k} + \sigma^2_a + \sigma^2_X \;\; \textit{(since } W_k \textit{ and } V_k \textit{ are independant)}$$

$$\sigma^2_{W_k} + \sigma^2_a + \sigma^2_X = \sigma^2_{W_{k+1}} + \sigma^2_{I_k} + 2\overline{W}_{k+1}\overline{I}_{k+1}$$

$$\textit{As } k \to \infty,\; \overline{W}_k \to \overline{W};\; \overline{I}_k \to \overline{I};\; \sigma^2_{W_k} \to \sigma^2_W$$

$$\Rightarrow \qquad \overline{W} = \frac{\sigma^2_a + \sigma^2_X}{2\overline{I}} - \frac{\sigma^2_I}{2\overline{I}}$$

$$\textit{Average idle time } \overline{I} = \frac{(1-\rho)}{\lambda} \Rightarrow \quad \overline{W} \le \frac{(\sigma^2_a + \sigma^2_X)\lambda}{2(1-\rho)}$$

$$\textit{as } \rho \to 1,\; \sigma^2_I \to 0 \textit{ since } I \to 0 \textit{ with probability } 1$$

- Special case: M/G/1

$$W = \frac{(\sigma^2_a + \sigma^2_X - \sigma^2_I)}{2\overline{I}} = \frac{\lambda(\sigma^2_X + \mu^{-2})}{2(1-\rho)}$$

$$\Rightarrow \sigma^2_I = \frac{1}{\lambda^2} - \frac{1}{\mu^2}$$

$$\textit{Neglected item in the bound: } \frac{\lambda\sigma^2_I}{2(1-\rho)} = \frac{\lambda\mu(\mu^2-\lambda^2)}{2\lambda^2\mu^2(\mu-\lambda)} = \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{\mu}\right) < \frac{1}{\lambda} \textit{ for } \rho < 1$$

Summary

- M|G|1 Queues with vacations
- Application of M|G|1 Results to Reservations & Polling
- Application to Token Ring Networks
- Extension to Non-product form Networks with M|G|1 Nodes
- M|G|1 Queues with Priorities
- Extensions to Queuing Networks with Priorities
- G|G|1 Queues