Lectures 10 & 11

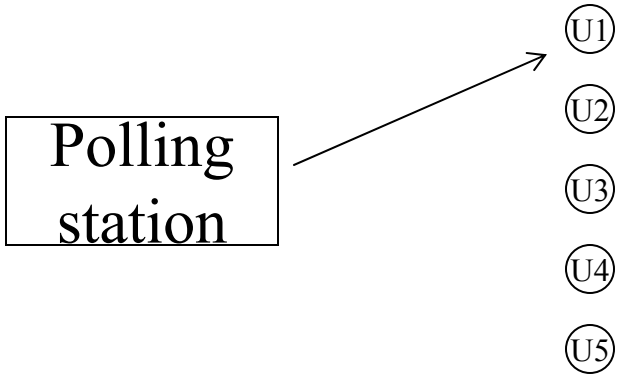
Reservations Systems
M/G/1 queues with Priority

Eytan Modiano
MIT

RESERVATION SYSTEMS

- **Single channel shared by multiple users**
- **Only one user can use the channel at a time**
- **Need to coordinate transmissions between users**

- **Polling systems**
  - **Polling station polls the users in order to see if they have something to send**
  - **A scheduler can be used to receive and schedule transmission requests**

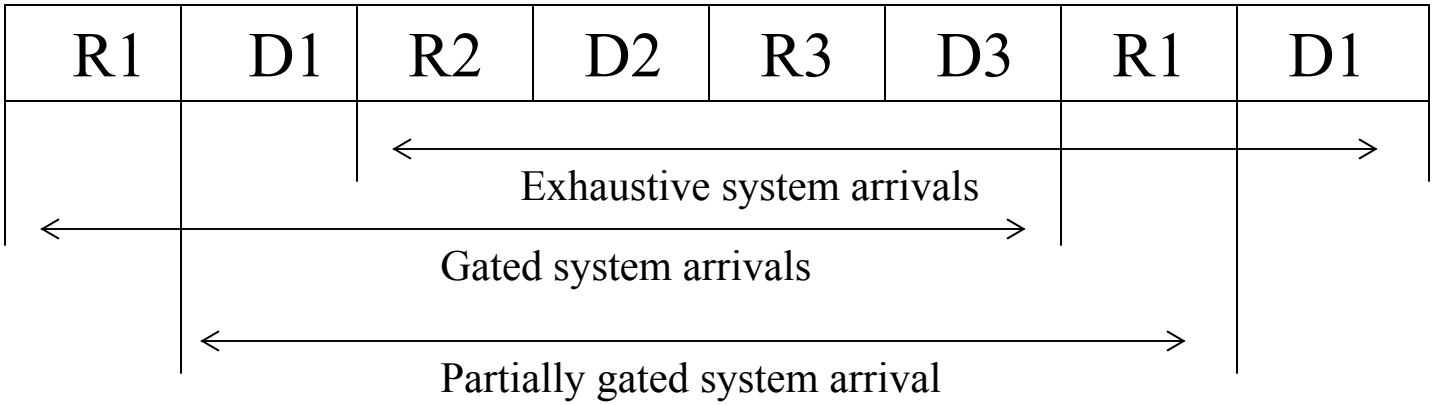


R1	D1	R2	D2	R3	D3	R1	D1

- **Reservation interval (R) used for polling or making reservations**
- **Data interval (D) used for the actual data transmission**

Reservations and polling systems

- **Gated system - users can transmit only those packets that arrived prior to start of reservation interval**
  - **E.g., explicit reservations**
- **Partially gated system - Can transmit all packets that arrived before the start of the data interval**
- **Exhaustive system - Can transmit all packets that arrive prior to the end of the data interval**
  - **E.g., token ring networks**
- **Limited service system - only one (K) packets can be transmitted in a data interval**



Single user exhaustive systems

- Let V_j be the duration of the j^{th} reservation interval
  - Assume reservation intervals are iid

- Consider the i^{th} data packet:

  $E[W_i] = R_i + E[N_i]/\mu$

  R_i = residual time for current packet or reservation interval

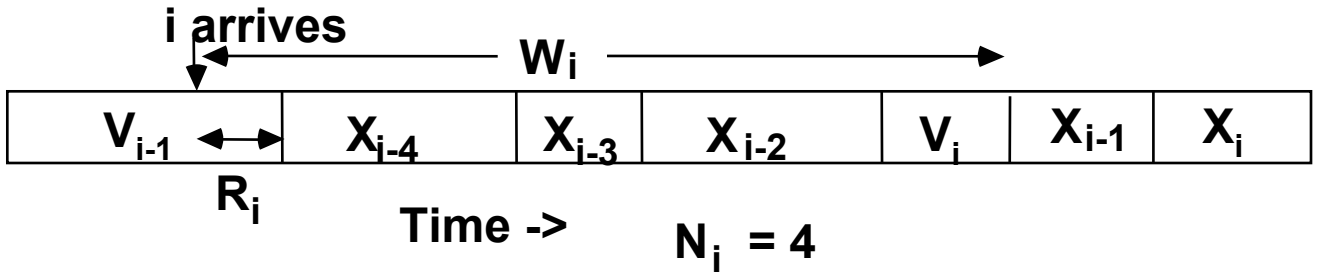
  N_i = Number of packets in queue

- Identical to M/G/1 with vacations

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

$$\text{When V = A (constant)} \Rightarrow W = \frac{\lambda E[X^2]}{2(1-\rho)} + \frac{A}{2}$$

Single user gated system (e.g.,reservations)



$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j + V_i$$

$$E[W_i] = E[R_i] + E[N_i]E[X] + E[V]$$

$$W = R + N_Q E[X] + E[V] \qquad (NQ=\lambda W)$$

$$W = (R + E[V])/(1-\rho)$$

SINGLE USER RESERVATION SYSTEM

- **The residual service time is the same as in the vacation case,**

$$R = \lambda \frac{E[X^2]}{2} + \frac{(1-\rho)E[V^2]}{2E[V]}$$

- **Hence,**

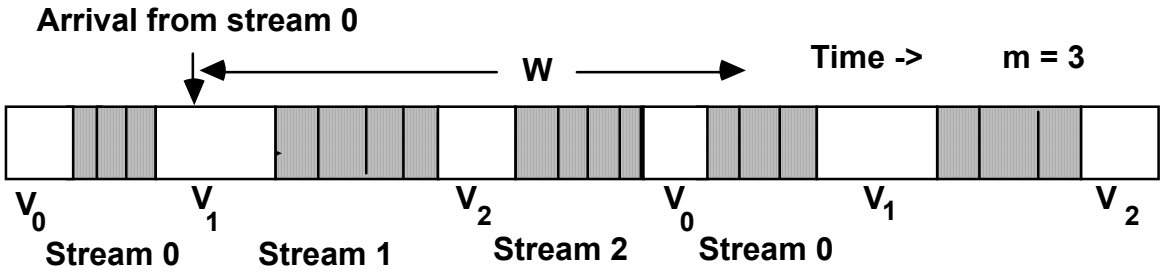
$$W = \lambda \frac{E[X^2]}{2(1-\rho)} + \frac{E[V^2]}{2E[V]} + \frac{E[V]}{1-\rho}$$

- **If all reservation intervals are of constant duration A,**

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

Multi-user exhaustive system

- Consider m incoming streams of packets, each of rate λ/m
- Service times $\{X_n\}$ are IID and independent of arrivals with mean $1/\mu$, second moment $E[X^2]$.
- Server serves all packets from stream 0, then all from stream 1, ..., then all from m-1, then all from 0, etc.
- There is a reservation interval of fixed duration $V_i = V$ (for all i)



Multi-user exhaustive system

- **Consider arbitrary packet i**
- **Let Y_i = the duration of whole reservation intervals during which packet i must wait ($E[Y_i] = Y$)**

$$W = R + \rho W + Y$$

- **Packet i may arrive during the reservation or data interval of any of the m streams with equal probability (1/m)**
  - **If it arrives during its own interval $Y_i = 0$, etc…, hence,**

$$Y_i = \{iV \quad w.p. \quad 1/m \quad 0 \le i < m$$

$$Y = E[Y_i] = \frac{V}{m}\sum_{i=0}^{m-1} i = \frac{V(m-1)}{2}$$

$$W = \frac{R+Y}{(1-\rho)}, \qquad R = \frac{(1-\rho)V^2}{2V} + \frac{\lambda E[X^2]}{2}$$

Multi-user exhaustive system

$$W = \frac{(1-\rho)V + \lambda E[X^2] + V(m-1)}{2(1-\rho)},$$

$$= \frac{V}{2} + \frac{V(m-1)}{2(1-\rho)} + \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{\lambda E[X^2]}{2(1-\rho)} + \frac{V(m-\rho)}{2(1-\rho)}$$

- **In text, V = A/m and hence,**

$$W = \frac{A}{2m} + \frac{A(m-1)}{2m(1-\rho)} + \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{\lambda E[X^2]}{2(1-\rho)} + \frac{A(1-\rho/m)}{2(1-\rho)}$$

Gated System

- **When a packet arrives during its own reservation interval, it must wait m full reservation intervals**

$$Y_i = \{iV \quad w.p. \quad 1/m \quad 1 \le i \le m$$

$$Y = E[Y_i] = \frac{V}{m}\sum_{i=1}^{m} i = \frac{V(m+1)}{2}$$

$$W = \frac{V}{2} + \frac{V(m+1)}{2(1-\rho)} + \frac{\lambda E[X^2]}{2(1-\rho)}$$

With $V = A/m$,

$$\frac{\lambda E[X^2]}{2(1-\rho)} + \frac{A}{2m} + \frac{A(1+1/m)}{2(1-\rho)} = \frac{\lambda E[X^2]}{2(1-\rho)} + \frac{A}{2}\left(\frac{1+(2-\rho)/m}{(1-\rho)}\right)$$

M/G/1 Priority Queueing

- **Priority classes 1, …, n (class 1 highest and n lowest)**

$$\lambda_k = arrival rate\ for\ class\ k$$

$$\mu_k = service\ rate\ for\ class\ k$$

$$E[X_k^2] = \text{sec}\, ond\ moment\ of\ service time (class\ k)$$

- **Non-preemptive system: Customer receiving service is allowed to complete service without interruption**

$$W_k = \frac{\sum_{i=1}^{i=n} \lambda_i E[X_i^2]}{2(1-\rho_1 - \ldots - \rho_{k-1})(1-\rho_1 - \ldots - \rho_k)}, \quad \rho_i = \frac{\lambda_i}{\mu_i}$$

- **Notice that the waiting time of high priority traffic is affected by lower priority traffic**

Preemptive-resume systems

- When a higher priority customer arrives, lower priority customer is interrupted
  - Service is resumed when no higher priority customers remain
  - Notice that the delay of high priority customers is no longer affected by that of lower priority customers
  - Preemption is not always practical and usually involves some overhead

- Consider a class k arrival and let,

  - W_k = waiting time for customers of class k or higher priority classes (1..K-1) already in the system
    - R_k = residual time for class k or higher customers
    - Notice that lower priority customers in service don't affect W_k because they are preempted
  - W_l = Waiting time for higher priority customers that arrive while priority k customer is already in the system
  - T_K = Average system time for priority K customer

$$T_k = W_k + W_l + 1/\mu_{\square}$$

Preemptive-resume, continued…

$$W_k = \frac{R_k}{1-\rho_1-\ldots-\rho_k}, \quad R_k = \frac{\sum_{i=1}^{k} \lambda_i E[X_i^2]}{2}$$

$$W_I = \sum_{i=1}^{k-1} (\lambda_i / \mu_i) T_k = \sum_{i=1}^{k-1} (\rho_i) T_k$$

$$T_k = \frac{1}{\mu_k} + \frac{R_k}{1-\rho_1-\ldots-\rho_k} + T_k \sum_{i=1}^{k-1} \rho_i$$

$$T_k = \left(\frac{1}{\mu_k}\right) \frac{(1-\rho_1-\ldots-\rho_k)+R_k}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)}$$

- **Notice independence of lower priority traffic**

Stability of Queueing Systems

- **Possible Definitions**
  - **Average Delay is bounded**

    E(delay) < infinity)
  - **Delay is finite with probability 1**

    P(delay < infinity) = 1
  - **Existence of a stationary occupancy distribution**

    Occupancy does not drift to infinity

E(delay) < Infinity

- **Example: M/M/1 queue**

$$T = \frac{1}{\mu - \lambda} < \infty \ \forall \ \lambda < \mu \Rightarrow \ \rho < 1$$

- **Example: M/G/1 queue**

$$T = \frac{1}{\mu} + \frac{\lambda E[X^2]}{2(1-\rho)} < \infty \ \ if \ (\rho < 1) \ and \ (E[X^2] < \infty)$$

P(Delay< Infinity) = 1

- **Slightly weaker definition than E[delay] < infinity**
- **P(delay < infinity) = 1 even if E(delay) = infinity**
- **Example:**

$$f_D(d) = \frac{2}{\pi(1+d^2)},\ d > 0$$

$$E[Delay] = \int_0^\infty \frac{2d}{\pi(1+d^2)} = \frac{Log[1+d^2]}{\pi}\Big|_0^\infty \Rightarrow \infty$$

$$P[Delay < x] = \int_0^x \frac{2}{\pi(1+d^2)} = \frac{2\arctan(x)}{\pi} \underset{x\to\infty}{\to} 1$$

- **In general it can be shown that for any G/G/1 queue**
  - **Arrival and service time distributions may even be correlated!**

If $\lambda < \mu$, P(delay < Infinity) = 1 even if E(delay) not finite

Existence of a stationary occupancy distribution

- **Irreducible and Aperiodic Markov chain**
  - **Pj > 0 for all states j => all states are visited infinitely often**
- **Drift:** $$D_i = E\left[X_{n+1} - X_n \mid X_n = i\right] = \sum_{k=i}^{\infty} kP_{(i,i+k)}$$
- **When in state I,**

  D_i > 0 => state tends to increase
  D_i < 0 => state tends to decrease
- **Intuitively, we don't want the state to drift to infinity, hence for large enough states the drift better get negative!**
- **Lemma: If D_i < infinity for all i and for some δ > 0 and i' > 0,**

  D_i < - δ for all i > i' , then the Markov chain has a stationary distribution

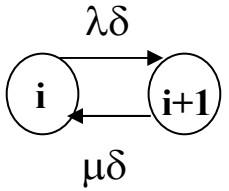
Irriducible: all states communicate (I.e., positive probability of getting from every state to every other state)
Periodic state : self transitions are possible only after a number of transitions (n)
that is a multiple of some constant d (I.e., n = 3, 6, 9, …). Aperiodic => no state is periodic

Examples

- **M/M/1**

$$D_i = E\left[X_{n+1} - X_n \mid X_n = i\right] = 1(\lambda\delta) - 1(\mu\delta) = (\lambda - \mu)\delta$$

$$D_i < 0 \Rightarrow \lambda < \mu$$



- **M/M/m**

$$D_i = E\left[X_{n+1} - X_n \mid X_n = i\right] = 1(\lambda\delta) - 1(m\mu\delta) \quad \forall i \geq m$$

$$D_i < 0 \Rightarrow \lambda < m\mu \quad \forall i \geq m$$

- **M/M/Inf**

$$D_i = E\left[X_{n+1} - X_n \mid X_n = i\right] = 1(\lambda\delta) - 1(i\mu\delta)$$

$$D_i < 0 \Rightarrow \lambda < i\mu$$

$$For\ any\ \lambda < \infty\ and\ 1/\mu < \infty\ \exists i'\ \ s.t., D_i < 0\ \forall i > i'$$