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

<table>
<thead>
<tr>
<th>R1</th>
<th>D1</th>
<th>R2</th>
<th>D2</th>
<th>R3</th>
<th>D3</th>
<th>R1</th>
<th>D1</th>
</tr>
</thead>
</table>
- 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 the 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

<table>
<thead>
<tr>
<th>R1</th>
<th>D1</th>
<th>R2</th>
<th>D2</th>
<th>R3</th>
<th>D3</th>
<th>R1</th>
<th>D1</th>
</tr>
</thead>
</table>

- Exhaustive system arrivals
- Gated system arrivals
- Partially gated system arrival
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

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

When $V = A$ (constant) $W = \frac{\sqrt{E[X^2]}}{2(1 - r)} + \frac{A}{2}$
Single user gated system (e.g., reservations)

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


\[ W = R + N_Q E[X] + E[V] \quad (N_Q = \bar{W}) \]

\[ W = \frac{R + E[V]}{1 - \bar{Q}} \]
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 $\lambda/m$

• Service times $\{X_n\}$ are IID and independent of arrivals with mean $1/\lambda$, 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 $Y_i = \text{the duration of whole reservation intervals during which packet } i \text{ must wait } (E[Y_i] = Y)$

$$W = R + W + Y$$

- 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,

$$Y_i = \left\{ \begin{array}{ll} iV & \text{w.p. } 1/m \\ 0 & \text{if } i < m \end{array} \right.$$  

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

$$W = \frac{R + Y}{(1\otimes1)}, \quad R = \frac{(1\otimes1)V^2}{2V} + \frac{E[X^2]}{2}$$
Multi-user exhaustive system

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

\[ = \frac{V}{2} + \frac{V(m - 1)}{2(1 - r)} + \frac{\mathbb{E}[X^2]}{2(1 - r)} + \frac{V(m - 1)}{2(1 - r)} \]

* In text, \( V = A/m \) and hence,

\[ W = \frac{A}{2m} + \frac{A(m - 1)}{2m(1 - r)} + \frac{\mathbb{E}[X^2]}{2m(1 - r)} + \frac{A(1 - r)}{2m} \]
Gated System

- When a packet arrives during its own reservation interval, it must wait \( m \) full reservation intervals.

\[
Y_i = \begin{cases} 
  iV & \text{w.p. } 1/m \\
  V & \text{w.p. } 1/m \\
  1 & \text{w.p. } 1/m
\end{cases}
\]

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

\[
W = \frac{V}{2} + \frac{V(m+1)}{2(1+1/m)} + \frac{E[X^2]}{2(1+1/m)}
\]

With \( V = A/m \),

\[
\frac{E[X^2]}{2(1+1/m)} + \frac{A}{2m} + \frac{A(1+1/m)}{2(1+1/m)} = \frac{E[X^2]}{2(1+1/m)} + \frac{A}{2} \left( \frac{1+(2+1)/m}{(1+1/m)} \right)
\]
M/G/1 Priority Queueing

- Priority classes 1, ..., n (class 1 highest and n lowest)
  \[ \lambda_k = \text{arrival rate for class } k \]
  \[ \mu_k = \text{service rate for class } k \]
  \[ E[X_k^2] = \text{second moment of service time (class } k) \]

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

\[
W_k = \frac{\prod_{i=1}^{n} \lambda_i E[X_i^2]}{2(1 \prod_{i=1}^{1} \lambda_i \prod_{i=k+1}^{n} \lambda_i)(1 \prod_{i=1}^{1} \mu_i \prod_{i=k+1}^{n} \mu_i)}
\]

- 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
  - \( W_I \) = 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

\[
T_k = W_k + W_I + \frac{1}{\mu_k}
\]
Preemptive-resume, continued...

\[ W_k = \frac{R_k}{1 \Box \Box_1 \Box \ldots \Box \Box_k}, \quad R_k = \frac{\sum_{i=1}^{k} i \cdot E[X_i^2]}{2} \]

\[ W_I = \sum_{i=1}^{k} \left( \frac{\Box_i}{\Box_i} \right) T_k = \sum_{i=1}^{k} \left( \Box_i \right) T_k \]

\[ T_k = \frac{1}{\Box_k} + \frac{R_k}{1 \Box_1 \Box \ldots \Box_k} + \sum_{i=1}^{k} \Box_i \]

\[ T_k = \left( \frac{1}{\Box_k} \right) \frac{(1 \Box_1 \Box \ldots \Box \Box_k)}{(1 \Box_1 \Box \ldots \Box \Box_k \Box)} \left( 1 \Box_1 \Box \ldots \Box_k \Box \Box_k \right) + R_k \]

• Notice independence of lower priority traffic
Stability of Queueing Systems

• Possible Definitions
  
  – Average Delay is bounded

    \[ E(\text{delay}) < \text{infinity} \]

  – Delay is finite with probability 1

    \[ P(\text{delay < infinity}) = 1 \]

  – Existence of a stationary occupancy distribution

    Occupancy does not drift to infinity
E(delay) < Infinity

- Example: M/M/1 queue

\[ T = \frac{1}{\lambda} < \quad \lambda < \bar{\lambda} \quad \mu < 1 \]

- Example: M/G/1 queue

\[ T = \frac{1}{\lambda} + \frac{\mathbb{E}[X^2]}{2(1 - \rho)} < \quad \text{if } (\rho < 1) \text{ and } (\mathbb{E}[X^2] < \ldots) \]
P(Delay < Infinity) = 1

• Slightly weaker definition than E[delay] < infinity

• P(delay < infinity) = 1 even if E(delay) = infinity

• Example:
  \[
  f_d(d) = \frac{2}{\sqrt{(1 + d^2)}}, \quad d > 0
  \]

  \[
  E[\text{Delay}] = \int_0^\infty \frac{2d}{\sqrt{(1 + d^2)}} = \left[ \log[1 + d^2] \right]_0^\infty
  \]

  \[
  P[\text{Delay} < x] = \int_0^x \frac{2}{\sqrt{(1 + d^2)}} = \frac{2 \arctan(x)}{x} \quad 1
  \]

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

  If \( d < \frac{1}{\sqrt{2}} \), P(delay < Infinity) = 1 even if E(delay) not finite
Existence of a stationary occupancy distribution

- Irreducible and Aperiodic Markov chain
  - \( P_j > 0 \) for all states \( j \) ⇒ all states are visited infinitely often

- Drift:
  \[
  D_i = E \left[ X_{n+1} \mid X_n = i \right] = \sum_{k=i} k P(i, i+k)
  \]

- 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 \( \square > 0 \) and \( i' > 0 \),
  \[
  D_i < - \square \text{ for all } i > i', \text{ then the Markov chain has a stationary distribution}
  \]

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., \( n = 3, 6, 9, \ldots \)). Aperiodic ⇒ no state is periodic
Examples

- **M/M/1**
  \[ D_i = E[X_{n+1} - X_n \mid X_n = i] = 1(l_i) - 1(l_{i+1}) = (r_i) \]
  \[ D_i < 0 \quad \text{if} \quad i < \infty \]

- **M/M/m**
  \[ D_i = E[X_{n+1} - X_n \mid X_n = i] = 1(l_i) - 1(m_i) \quad \text{if} \quad i \geq m \]
  \[ D_i < 0 \quad \text{if} \quad i < m \quad \text{and} \quad i \geq m \]

- **M/M/Inf**
  \[ D_i = E[X_{n+1} - X_n \mid X_n = i] = 1(l_i) - 1(i_i) \]
  \[ D_i < 0 \quad \text{if} \quad i < i' \]
  
  For any \( i \) such that \( 1/i < i' \), \( s.t. \), \( D_i < 0 \) \( i > i' \)