
Temas 10 y 11

Sistemas de reservas y colas M/G/1 con prioridad

**Eytan Modiano
MIT**

SISTEMAS DE RESERVAS

- Un solo canal compartido por múltiples usuarios
- Sólo un usuario puede utilizar el canal en cada momento
- Es necesario coordinar las transmisiones entre los usuarios
- **Sistemas de sondeos:**

- El centro de sondeos realiza un sondeo entre los usuarios para ver si tienen algo que enviar
- Se puede utilizar un gestor para recibir y gestionar las solicitudes de transmisión

Centro de sondeos

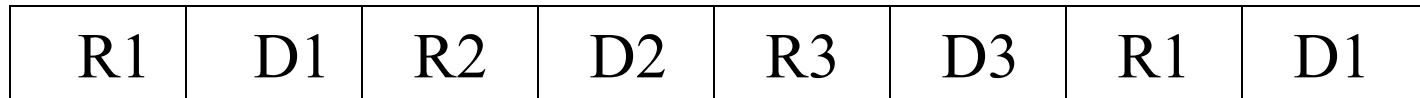
U1

U2

U3

U4

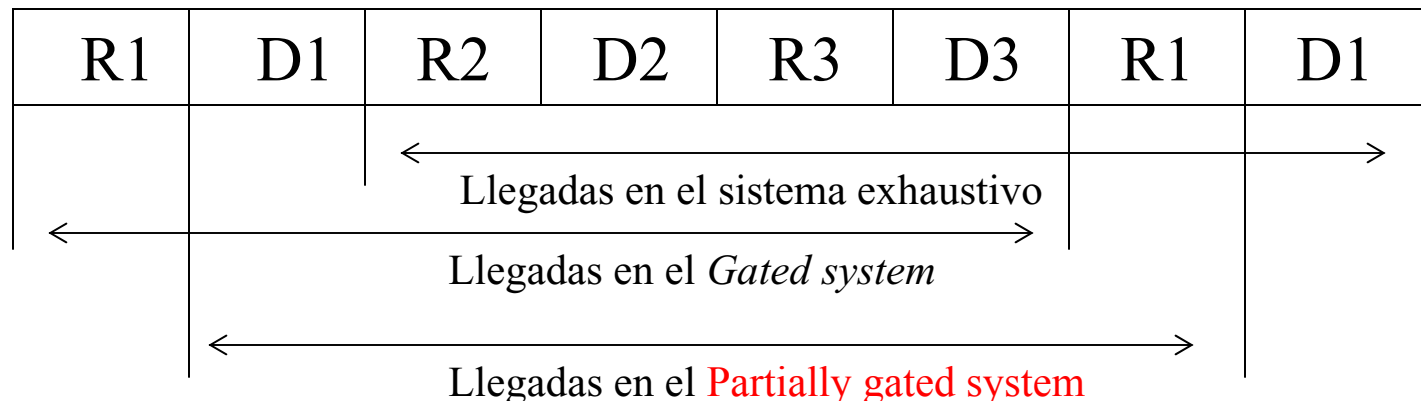
U5



- Intervalo de reserva (R) utilizado para realizar sondeos o reservas
- Intervalo de datos (D) utilizado para la transmisión de los datos en sí

Sistemas de reservas y sondeos

- **Gated system:** los usuarios sólo pueden transmitir los paquetes que hayan llegado antes del intervalo de reserva
 - Ej.: reservas explícitas
- **Partially gated system:** se pueden transmitir todos los paquetes que hayan llegado antes del intervalo de datos
- **Sistema exhaustivo:** se pueden transmitir todos los paquetes que hayan llegado antes de que finalice el intervalo de datos
 - Ej.: redes *token ring*
- **Sistema de servicio limitado:** sólo se puede transmitir un número de paquetes (K) en cada intervalo de datos



Sistemas exhaustivos de un solo usuario

- Sea V_j la duración del intervalo de reserva número j :
 - Suponer que los intervalos de reserva son IID
- Examinamos el paquete de datos número i :

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

R_i = tiempo residual del paquete en servicio o el intervalo de reserva

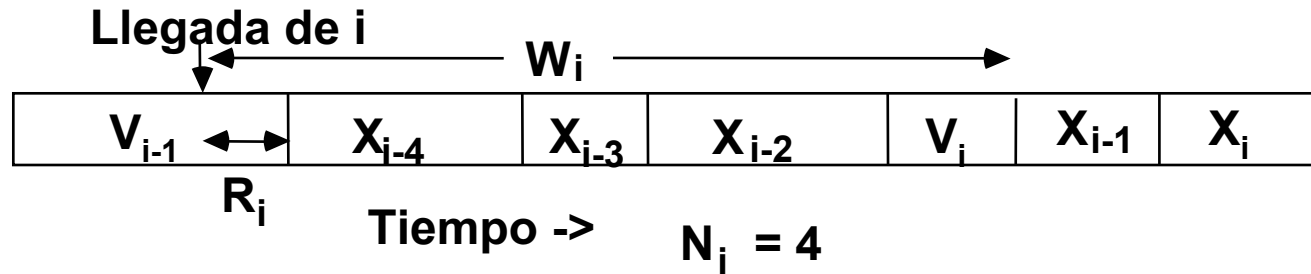
N_i = Número de paquetes presentes en la cola

- Igual a las M/G/1 con vacaciones:

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

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

Gated system de un solo usuario (ej.: reservas)



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

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

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

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

SISTEMA DE RESERVAS DE UN SOLO USUARIO

- El tiempo de servicio residual es el mismo que en el caso de las vacaciones:

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

- Así:

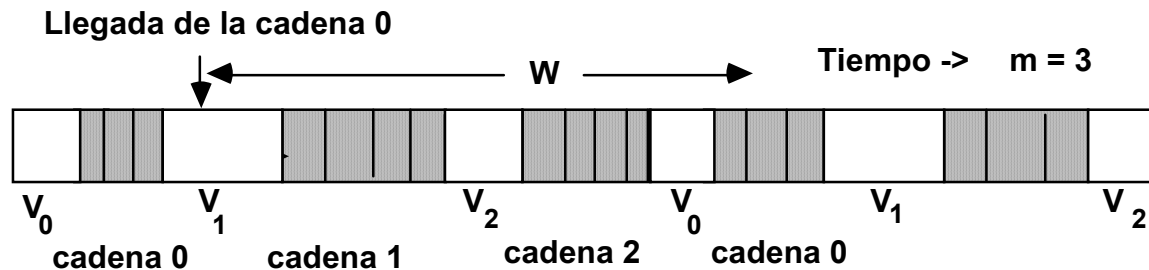
$$W = \lambda \frac{E[X^2]}{2(1-\rho)} + \frac{E[V^2]}{2E[V]} + \frac{E[V]}{1-\rho}$$

- Si todos los intervalos de reserva tienen una duración constante A:

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

Sistema exhaustivo multiusuario

- Supongamos m cadenas de paquetes entrantes, todos ellos con una tasa de λ/m
- Los tiempos de servicio $\{X_n\}$ son IID e independientes de las llegadas con media $1/\mu$, segundo momento $E[X^2]$.
- El servidor sirve todos los paquetes de la cadena 0; luego, todos los de la cadena 1, etc. A continuación, sirve todos los de la $m-1$, luego todos los de la 0, etc.
- Hay un intervalo de reserva de duración establecida $V_i = V$ (para todo i)



Sistema exhaustivo multiusuario

- Supongamos un paquete arbitrario i
- Sea Y_i = la duración de la totalidad de los intervalos de reserva durante los cuales el paquete i debe esperar ($E[Y_i] = Y$)

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

- El paquete i puede llegar durante el intervalo de reserva o el de datos de cualquiera de las m cadenas con la misma probabilidad ($1/m$):
 - Si llega durante su propio intervalo $Y_i = 0$, etc. Así:

$$Y_i = \{iV \quad \text{con prob. } 1/m \quad 0 \leq i < m$$

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

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

Sistema exhaustivo multiusuario

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

- En el texto, $V = A/m$ y, por tanto:

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

Gated System

- Cuando un paquete llega durante su propio intervalo de reserva, debe esperar m intervalos de reserva completos:

$$Y_i = \{iV \quad \text{con prob. } 1/m \quad 1 \leq i \leq m$$

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

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

Con $V = A/m$:

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

Prioridad en las colas M/G/1

- **Clases de prioridad 1, ... n (la clase 1 es la superior y la n la inferior)**

$\lambda_k =$ tasa de llegada de la clase k

$\mu_k =$ tasa de servicio de la clase k

$E[X_k^2] =$ segundo momento del tiempo de servicio (clase k)

- **Sistema sin preferencias: el cliente que se está sirviendo puede completar su servicio sin interrupción:**

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

- **Obsérvese que el tiempo de espera del tráfico de prioridad alta se ve afectado por el tráfico de menor prioridad**

Sistemas con preferencias

- Cuando llega un cliente con más prioridad, se interrumpe al cliente de menos prioridad:
 - El servicio de éste se retoma cuando ya no quedan clientes de más prioridad
 - Obsérvese que la espera de los clientes de prioridad alta ya no se ve afectada por la de los clientes de menos prioridad
 - El uso de preferencias no siempre es práctico y, a menudo, implica algún costo
- Supongamos una llegada de la clase k y establezcamos:
 - W_k = tiempo de espera de los clientes de la clase k o de clases de prioridad superior ($1..K-1$) ya en el sistema:
 - R_k = tiempo residual de los clientes de la clase k o superior
 - Obsérvese que los clientes de menos prioridad que están en servicio no influirán en W_k ya que éstos últimos tienen preferencia
 - W_1 = tiempo de espera de los clientes de más prioridad que lleguen cuando un cliente de prioridad k se encuentra ya en el sistema
 - T_k = promedio de tiempo del sistema para un cliente de prioridad K

$$T_k = W_k + W_1 + 1/\mu$$

Sistemas con preferencias (continuación)

$$W_{k\Box} = \frac{R_{k\Box}}{1 - \rho_1 - \dots - \rho_k}, \quad R_{k\Box} = \frac{\sum_{i=1}^{k\Box} \lambda_i E[X_i^2]}{2}$$

$$W_{I\Box} = \sum_{i\Box=1}^{k\Box+1} (\lambda_i / \mu_i) T_{k\Box} = \sum_{i\Box=1}^{k\Box+1} (\rho_i) T_{k\Box}$$

$$T_{k\Box} = \frac{1}{\mu_{k\Box}} + \frac{R_{k\Box}}{1 - \rho_1 - \dots - \rho_{k\Box}} + T_{k\Box} \sum_{i\Box=1}^{k\Box+1} \rho_{i\Box}$$

$$T_{k\Box} = \left(\frac{1}{\mu_{k\Box}} \right) \frac{(1 - \rho_1 - \dots - \rho_k) + R_{k\Box}}{(1 - \rho_1 - \dots - \rho_{k\Box+1})(1 - \rho_1 - \dots - \rho_k)}$$

- Obsérvese que hay una independencia del tráfico de menos prioridad

Estabilidad de los sistemas de colas

- **Definiciones posibles:**

- **Promedio de espera delimitado:**

$$E(\text{espera}) < \text{infinito}$$

- **La espera es finita, con probabilidad 1**

$$P(\text{espera} < \text{infinito}) = 1$$

- **Existencia de una distribución estacionaria de la ocupación:**

La ocupación no tiende a infinito

E(espera) < infinito

- **Ejemplo: cola M/M/1**

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

- **Ejemplo: cola M/G/1**

$$T = \frac{1}{\mu} + \frac{\lambda E[X^2]}{2(1 - \rho)} < \infty \quad \text{si} \quad (\rho < 1) \quad \text{y} \quad (E[X^2] < \infty)$$

P(espera < infinito) = 1

- Definición ligeramente más débil que: $E[\text{espera}] < \text{infinito}$
- $P(\text{espera} < \text{infinito}) = 1$ incluso si $E(\text{espera}) = \text{infinito}$

• Ejemplo:

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

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

$$P[\text{espera} < x] = \int_0^x \frac{2}{\pi(1+d^2)} = \frac{2 \arctan(x)}{\pi} \xrightarrow{x \rightarrow \infty} 1$$

- En general, se puede demostrar que para cualquier cola G/G/1:
 - ¡Las distribuciones de las llegadas y el tiempo de servicio pueden incluso ser correlativos!

Si $\lambda < \mu$, $P(\text{espera} < \text{Infinito}) = 1$ incluso si $E(\text{espera})$ no es finito

Existencia de una distribución estacionaria de la ocupación

- Cadena de Markov no periódica e irreducible:
 - $P_j > 0$ para todos los estados $j \Rightarrow$ todos los estados se visitan a menudo infinitamente
- Tendencia:
$$D_i = E[X_{n+1} - X_n | X_n = i] = \sum_{k=i}^{\infty} kP_{(i,i+k)}$$
- En el estado i :
 - $D_i > 0 \Rightarrow$ el estado tiende a aumentar
 - $D_i < 0 \Rightarrow$ el estado tiende a disminuir
- Intuitivamente, no queremos que el estado tienda a infinito; así, para los estados lo suficientemente grandes, ¡es mejor que la tendencia sea negativa!
- Lema: Si $D_i < \infty$ para todo i y para alguno $\delta > 0$ e $i' > 0$:

$D_i < -\delta$ para todo $i > i'$, entonces la cadena de Markov tiene una distribución estacionaria

Irreducible: todos los estados se comunican (es decir, la probabilidad de pasar de cualquier estado a otro es positiva)

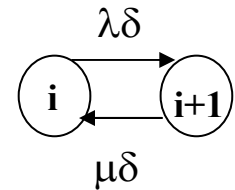
Estado periódico: las transiciones a uno mismo son posibles sólo tras un número de transiciones (n), múltiplo de alguna constante d (es decir, $n = 3, 6, 9 \dots$). Aperiódico \Rightarrow ningún estado es periódico

Ejemplos

- **M/M/1:**

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

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



- **M/M/m:**

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

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

- **M/M/Inf:**

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

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

Para cualquier $\lambda < \infty$ y $1/\mu < \infty \exists \bar{i}$ tal que, $D_i < 0 \forall i > \bar{i}$

