1.- INTRODUCCIÓN

La teoría de colas es objeto de una amplia bibliografía que aborda desde el estudio de sistemas formado por una cola con un servidor hasta complejas redes de colas de espera.

Los sistemas de colas forman una amplia y útil clase de sistemas de eventos discretos, especialmente aquellos en los que hay recursos compartidos, como son:

- Sistemas de fabricación
- Sistemas de comunicación
- Sistemas informáticos

La teoría de colas ha tenido un énfasis especial en el tratamiento de sistemas estocásticos.

El objetivo de la teoría de colas en la bibliografía es el estudio del comportamiento del sistema bajo ciertas condiciones, dejando en un segundo plano la determinación de políticas óptimas de funcionamiento.

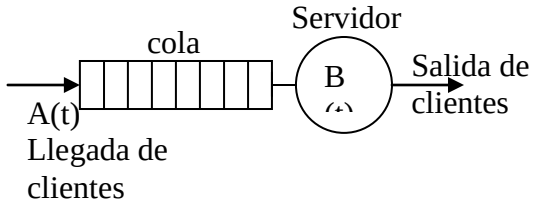
Por tanto, las herramientas desarrolladas son mas bien descriptivas y no de prescripción.

2 ESPECIFICCIONES DE MODELOS DE COLAS

Normalmente son tres las componentes para la especificación del proceso:

- Especificación de modelos estocásticos para los *procesos de llegada* y *de salida.*
- Especificación de los parámetros estructurales del sistema (Capacidad de la cola, número de servidores, etc…)
- Especificación de las políticas de operación, por ejemplo condiciones bajo las cuales un cliente es aceptado, tipo de clientes, …

Una cola con un único servidor se representa gráficamente:



la cola tiene una capacidad infinita.

2.1 MODELOS ESTOCÁSTICOS PARA LOS PROCESOS DE LLEGADA Y DE SERVICIO

Asociada a la llegada del k-ésimo cliente esta la asociada la variable aleatoria Y_K, que representa el tiempo transcurrido entre la legada del cliente k-1 ésimo y el k-ésimo.

Se considera:

- $Y_0 = 0$,
- Y_1=tiempo transcurrido hasta la llegada del primer cliente.
- Y_k = tiempo transcurrido desde la llegada del cliente k-1 hasta el cliente k.

Normalmente, se considera que el intervalo de tiempo entre dos llegadas de clientes sucesivas es una variable aleatoria idénticamente distribuida,es decir $Y_1, Y_2, \ldots$ son independientes e idénticamente distribuidas.

La distribución de probabilidad:

$$A(t) = P[Y \leq t]$$

describe totalmente la sucesión de tiempos entre llegadas. Su esperanza matemática es:

$$E[Y] = \frac{1}{\lambda}$$

de tal manera que λ se interpreta como la tasa promedio de llegada de clientes por unidad de tiempo

De manera similar asociamos a la partida de un cliente del servidor una sucesión estocástica $\{Z_1, Z_2, \ldots\}$, donde Z_k es el tiempo de servicio del k-ésimo cliente.
Se considera que las variables de la sucesión son independientes e idénticamente distribuidas. Se define:

$$B(t) = P[Z \leq t]$$

la esperanza matemática de Z es:

$$E[Z] = \frac{1}{\mu}$$

de tal manera que μ es la tasa promedio de clientes servidos por el servidor (siempre que haya clientes).

2.2 PARÁMETROS ESTRUCTURALES

Los parámetros estructurales de una cola son:

- La capacidad de almacenamiento de la cola; se denota como *K*.
  K=∞ indica capacidad ilimitada de la cola.
- Número de servidores que atienden a la cola; habitualmente se denota como *m*.
  m=1 indica un único servidor

2.3 POLÍTICAS DE OPERACIÓN

Se puede definir diferentes formas de manipular los clientes en la cola así podemos destacar:

- *Número de clases de clientes*: Podemos considerar la existencia de diferentes clientes con diferentes requerimientos del servidor.
- *Políticas de scheduling*: En colas con clases diferentes de servidores, podemos establecer unas prioridades.
- *Disciplinas de la cola*: describen el orden en el que se atienden alos clientes que están en cola.
- *Políticas de admisión*: Incluso en colas con capacidad infinita, se puede rechar a ciertos clientes

2.4 NOTACIÓN

En la teoría de colas se utiliza la notación**:**

A/B/m/K

Donde**:**

- ***A*** es la distribución de probabilidad de tiempo entre dos llegadas consecutivas de clientes.
- ***B*** es la distribución de probabilidad del tiempo de servicio.
- ***M*** es el número de servidores**.**
- ***K*** es la capacidad de almacenamiento de la cola.

Para ***A*** y ***B*** se utiliza la notación:

- G. Indica distribución genérica; no se conoce nada más sobre el proceso
- GI indica una distribución genérica de un proceso de renovación.
- D indica que el proceso es determinista, es decir el tiempo entre dos llegadas consecutivas es fijo en el proceso de llegada o bien el tiempo de servicio es fijo en el proceso de servicio.
- M indica proceso Markoviano. Es decir el tiempo entre dos llegadas consecutivas en el proceso de llegada o bien el tiempo de servicio en el proceso de servicio tienen una distribución exponencial.

3.-COMPORTAMIENTO DE UNA COLA

Definimos:

- Z_k tiempo de servicio del k-ésimo cliente
- Y_k tiempo que transcurre entre la llegada del cliente k-1 y la llegada del cliente k-ésimo
- A_k tiempo de llegada del k-ésimo cliente.
- D_k tiempo de salida del k-ésimo cliente.
- W_k tiempo de espera del k-ésimo cliente. (dede que llega hasta que el servidor comienza a dar servicio)
- S_k tiempo en el sistema del cliente k-ésimo (desde que llega hasta que abandona el sistema)

Se verifica que:

$$S_k = D_k - A_k$$
$$S_k = W_k + Z_k$$
$$D_k = A_k + W_k + Z_k$$

Además definimos las variables aleatorias:

- X(t) longitud de la cola en el instante t (numero de clientes en cola)
- U(t) carga de trabajo en el instante t, es decir, tiempo necesario para vaciar la cola si no llegase ningún cliente más.

En general X(t) es la variable que define el estado de la cola.

El comportamiento estocástico del tiempo de espera W_k proporciona información relevante sobre el comportamiento de la cola. La distribución de probabilidad de W_k depende de k. sin embargo, frecuentemente cuando $k \to \infty$ existe una distribución de probabilidad estacionaria independiente de k, tal que:

$$\lim_{k \to \infty} P[W_k < t] = \lim_{k \to \infty} P[W < t]$$

si este límite existe, la variable aleatoria W describe el tiempo de espera típico de un cliente una vez alcanzado el estacionario. La *esperanza matemática* de esta variable aleatoria, E(W), representa el *tiempo de espera promedio* de los clientes.

De manera análoga, si existe una distribución de probabilidad estacionaria para la *sucesión* $\{S_k\}$, su esperanza matemática E[S] es el *promedio del tiempo en el sistema* de los clientes en estado estacionario.

La misma idea se puede aplicar a los procesos estocásticos {X(t)} y {U(t)}. Si existe una distribución de probabilidad estacionaria para estos procesos cuando $t \to \infty$ entonces las variables aleatorias X y U son utilizadas para describir la longitud de la cola y la carga de trabajo en el sistema en estado estacionario. Utilizaremos la notación Π_n, n= 1,2,…para referirnos a la probabilidad de la longitud de la cola en estacionario, esto es:

$$\Pi_n = P[X = n], \quad n = 0,1,2,...$$

Se verifica que E[X] es la longitud promedio de la cola y E[U] es la carga de trabajo promedio.

Si se asume que se puede alcanzar el estacionario, nos centraremos en el comportamiento de la cola en estado estacionario y tendremos unos índices de medida :

- Promedio del tiempo de espera *E[W]*
- Promedio de tiempo en el sistema *E[S]*
- Promedio de la longitud de la cola, *E[X]*

Que es deseable que sean lo más pequeñas posible.
Y, además,

- *Utilización del sistema*. Fracción del tiempo que el servidor está ocupado.
- *Rendimiento del sistema*, es decir, la tasa a la que los clientes abandonan el sistema tras recibir servicio

Que es deseable que sean tan grandes como sea posible dentro de los rangos posibles:

- La utilización del sistema no puede ser mayor que 1
- El rendimiento del sistema no puede ser mayor que la máxima tasa de servicio del servidor.

Definimos *intensidad de tráfico*:

$$[\text{int}\,ensidad\ \ de\ tráfico] = \frac{[tasa\ \ promedio\ de\ llegada]}{[tasa\ \ promedio\ de\ salida]}$$

En el caso de una cola con un único servidor, la intensidad de tráfico se define:

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

En un sistema con un único servidor Π_0 es la probabilidad de que la cola este vacía en estado estacionario, que se puede interpretar como la fracción de tiempo que el servidor está ocioso. Entonces:

$$[utilización] = [fraccion\ de\ tiempo\ que\ el\ servidor\ está\ ocupado] = 1 - \Pi_0$$
$$[rendimiento] = [tasa\ de\ salida\ de\ clientes\ tras\ recibir\ servicio] = \mu(1 - \Pi_0)$$

en estado estacionario la tasa de clientes que llegan debe ser igual a la tasa de clientes que salen:

$$\lambda = \mu(1 - \Pi_0)$$

de donde,

$$\rho = 1 - \Pi_0$$

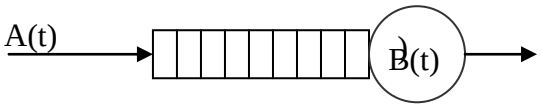
y, por tanto, la *intensidad de trafico* es la *utilización del sistema.*

Si Π_0=1 el sistema está siempre ocioso ya que no llegan clientes. Si Π_0=0 el sistema está siempre ocupado y la longitud de la cola crece indefinidamente (inestabilidad). Se verifica que:

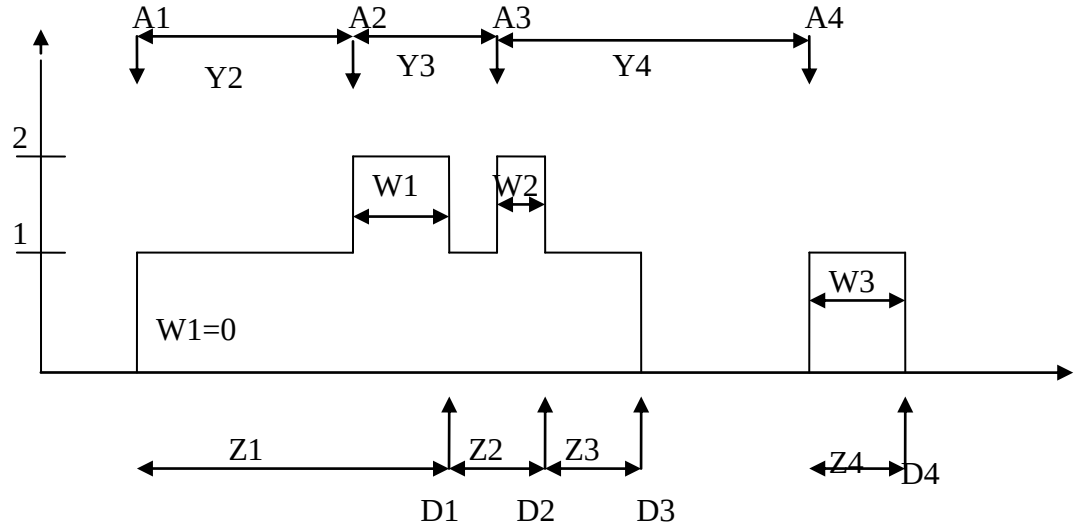
$$0 \le \rho < 1$$

4.- DINÁMICA DE UNA COLA

Vamos a considerar la cola:



Operando bajo una politica FCFS(se atiende a los clientes en el orden en que van llegando). Una posble evolución del número de clientes en la cola puede tener la forma:



Cuando llega el k-ésimo cliente hay dos casos posibles:

1. El sistema está vacio, además Wk=0. El sistema estará vacío cuando:

   $$D_{k-1} \leq A_k$$

   es decir, el cliente anterior sale antes de que el nuevo cliente llegue. Entonces:

   $$D_{k-1} - A_k \leq 0 \Leftrightarrow W_k = 0$$

2. El sistema no está vacío, además $W_k \geq 0.$ En este caso el k-ésimo cliente debe esperar a que el cliente anterior abandone el sistema.Entonces:

   $$D_{k-1} - A_k > 0 \Leftrightarrow W_k = D_{k-1} - A_k$$

Combinando los dos casos podemos obtener:

$$W_k = \begin{Bmatrix} 0 & si\ D_{k-1} - A_k \leq 0 \\ D_{k-1} - A_k & si\ D_{k-1} - A_k > 0 \end{Bmatrix}$$

que podemos expresar como:

$$W_k = \max\{0, D_{k-1} - A_k\}$$

Por otra parte, ya vimos que:

$$D_k = A_k + W_k + Z_k$$

y ,por tanto:

$$D_{k-1} = A_{k-1} + W_{k-1} + Z_{k-1}$$

definiendo

$$Y_k = A_k - A_{k-1}$$

obtenemos que:

$$W_k = \max\{0, W_{k-1} + Z_{k-1} - Y_k\}$$
$$S_K = \max\{0, S_{k-1} - Y_k\} + Z_k$$

Finalmente se obtiene una expresión recursiva para los tiempos de partida:

$$W_k = D_k - A_k - Z_k$$
$$D_k = \max\{A_k, D_{k-1}\} + Z_k$$

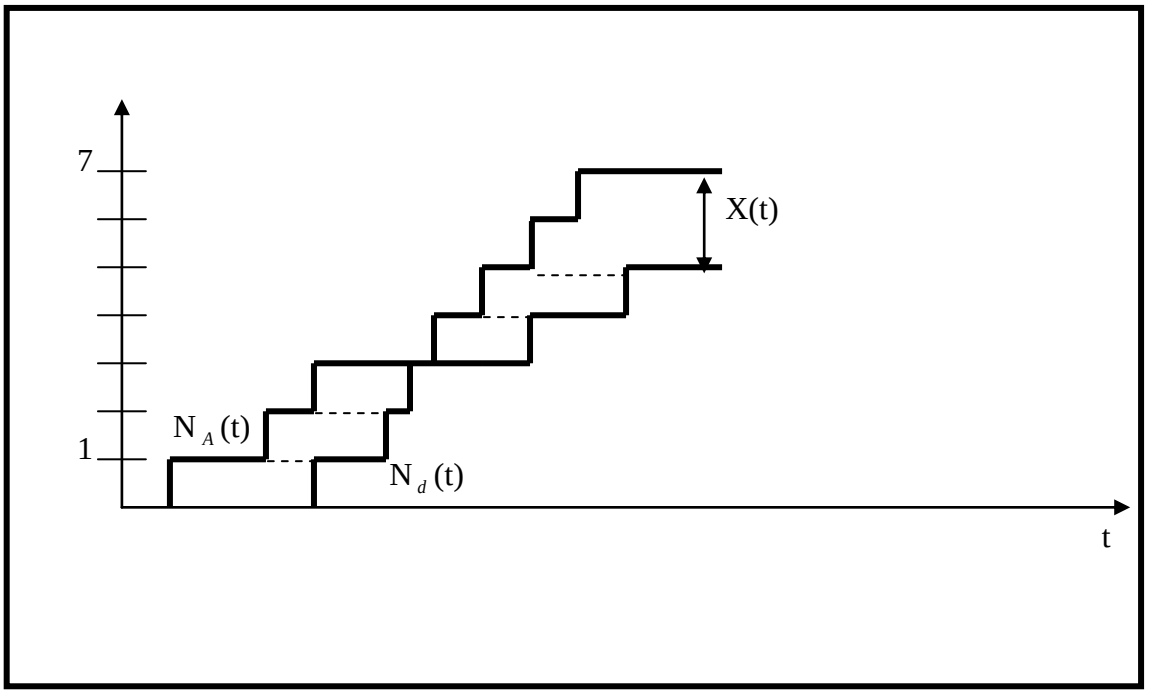
Estas relaciones

- capturan las características de la dinámica de la cola.
- Son de carácter general

5.-LA LEY DE LITTLE

Consideremos un sistema formador una cola con un servidor.

- $N_a(t)$ contabiliza el numero de clientes que llegan
- $N_d(t)$ contabiliza el número de clientes que abandonan el servidor
- la longitud de la cola viene dada por $X(t) = N_a(t) - N_d(t)$



En la figura se aprecian entre las dos líneas rectángulos:

- Su altura es la unidad ya que corresponde a un cliente
- Su anchura indica el tiempo de permanencia del cliente

Dividiendo el área entre ambas líneas por el número de clientes que han llegado en el intervalo (0,t] , $n_a(t)$,obtenemos el promedio de tiempo de permanencia de un cliente en el sistema:

$$\bar{s}(t) = \frac{u(t)}{n_a(t)}$$

Da manera análoga, el número promedio de clientes en el sistema (longitud promedio de la cola)durante el intervalo (0,t] es:

$$\bar{x}(t) = \frac{u(t)}{t}$$

Dividiendo el número total de clientes que han llegado durante el intervalo (0,t] por t, obtenemos la tasa promedio de llegada de clientes:

$$\lambda(t) = \frac{n_a(t)}{t}$$

De las expresiones anteriores, se obtiene:

$$\bar{x}(t) = \lambda(t) \cdot \bar{s}(t)$$

el número promedio de clientes en el sistemas es igual al producto de la tasa de llegada por en tiempo promedio de permanencia en el sistema

Vamos a asumir que:

$$\lim_{t\to\infty} \lambda(t) = \lambda$$

$$\lim_{t\to\infty} \bar{s}(t) = \bar{s}$$

de tal manera que λ y $\bar{s}$ representan la tasa de llegada de clientes y el tiempo en el sistema una vez alcanzado el estacionario. De aquí se obtiene

$$\bar{x} = \lambda \cdot \bar{s}$$

Supongamos que estos limites y relaciones existen para cualquier realización de la dinámica del sistema, siendo fijos λ y $\bar{x}$, es decir, el proceso de llegada, el tiempo en el sistema y la longitud de la cola son *ergódicos.*

En este caso los promedios son las esperanzas matemáticas en estado estacionario y, en concreto,:

$$E[X] = \lambda \cdot E[S]$$

que es la *ley de LITTLE*.
Es importante remarcar que la ley de Little es:

- independiente de los procesos estocásticos
- independiente de las políticas de la cola
- válida para una combinación arbitraria de colas y servidores
- independiente de la configuración de una red de colas
- válida para una cola sin considerar el servidor: $E[X_Q] = \lambda \cdot E[W]$, siendo
  - $E[X_Q]$ la esperanza de clientes en cola (sin servidor)
  - $E[W]$ la esperanza del tiempo de espera.
- Considerando sólo el servidor
  $E[X_s] = \lambda \cdot E[Z]$
  - $E[X_s]$ es el número de clientes en el servidor (0 ó 1)
  - $E[Z]$ es la esperanza del tiempo de servicio

6.- ANÁLISIS DE SISTEMAS DE COLAS MARKOVIANOS

En estos sistemas suponemos:

- Los intervalos de tiempo entre llegadas estan distribuidos exponencialmente con parámetro λ, es decir $G(t) = P[Y_k < t] = 1 - e^{-\lambda t}$
- Los tiempos de servicio están distribuidos exponencialmente con intervalo μ, es decir, $Z(t) = P[Z_k < t] = 1 - e^{-\mu t}$

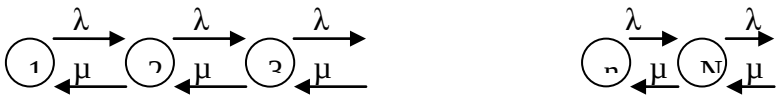
Cuando el proceso de llegada de clientes es un proceso de Poisson, con independencia del proceso de servicio, se verifica que la probabilidad de que al llegar un cliente en el instante t se encuentre con n clientes en la cola es igual a la probabilidad de que en el instante t haya n clientes en la cola

$$\alpha_n(t) = \pi_n(t)$$

6.1-**La cola M/M/1**

- Un único servidor
- Cola de capacidad infinita
- El proceso de llegada de clientes es de Poisson
  - La tasa del proceso de llegada es λ
- Los tiempos de servicio tienen una distribución exponencial
  - La tasa de servicio es μ

Esta cola es realmente un proceso de nacimiento–muerte



se verifica que la probabilidad estacionaria de que la cola esté vacía es:

$$\Pi_0 = \frac{1}{1+\sum_{n=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^n}$$

la serie geométrica del denominador converge si $\frac{\lambda}{\mu} < 1$.

Suponiendo que es cierto, se verifica:

$\sum_{n=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^n = \frac{\lambda/\mu}{1-\lambda/\mu}$ y, por tanto, $\Pi_0 = 1-\frac{\lambda}{\mu} = 1-\rho$

La probabilidad estacionaria de que haya n clientes en la cola viene dada por:

$$\Pi_n = \left(\frac{\lambda}{\mu}\right)^n (1-\rho) = (1-\rho)\cdot\rho^n$$

6.1.1 **Utilización y rendimiento**

La utilización se obtiene de forma inmediata:

$$1-\Pi_0=\rho$$

El rendimiento es la tasa de salida del servidor, que es:

$$\mu(1-\Pi_0)=\lambda$$

como cabe esperar ,ya que en estado estacionario las tasas de llegada y de salida están equilibradas

6.1.2 **Longitud media de la cola**

Es la esperanza matemática de la distribución estacionaria del número de clientes en la cola**:**

$$E[X]=\sum_{n=0}^{\infty} n\cdot\Pi_n=(1-\rho)\cdot\sum_{n=0}^{\infty} n\cdot\rho^n$$

teniendo en cuenta que:

$$\sum_{n=0}^{\infty} n\cdot\rho^n=\frac{\rho}{(1-\rho)^2}$$

obtenemos que:

$$E[X]=\frac{\rho}{1-\rho}$$

cabe destacar que cuando $\rho\to 1,\quad E[X]\to\infty$. Esto indica que si intentamos optimizar la utilización del servidor manteniéndolo ocupado tanto como sea posible, el servicio a los clientes empeora porque la cola será más larga.

6.1.3 **Tiempo medio en el sistema**

Aplicando la ley de Little, se obtiene:

$$\frac{\rho}{1-\rho}=\lambda\cdot E[S]$$

que se puede expresar como

$$E[S]=\frac{1/\mu}{1-\rho}$$

se observa:

- Cuando $\rho \to 0, E[S] \to 1/\mu$
  - indica que cuando el servidor está ocupado, el tiempo en el sistema es el tiempo de servicio.
- Cuando $\rho \to 1, E[S] \to \infty$
  - Indica que si el servidor está muy ocupado, el tiempo en el sistema aumenta.

6.1.4 **Tiempo medio de espera**

En estado estacionario se verifica que:

$$E[S]=E[W]+E[Z]=E[W]+\frac{1}{\mu}$$

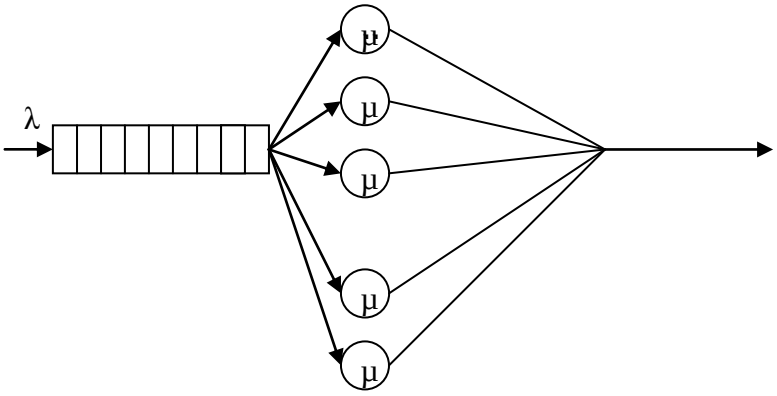
de donde:

$$E[W]=\frac{1/\mu}{1-\rho}-\frac{1}{\mu}=\frac{\rho}{\mu\cdot(1-\rho)}$$

se observa que cuando $\rho \to 1, E[W] \to \infty$

6.2 LA COLA M/M/m

Es una cola con una capacidad infinita y m servidores



Cuando llega un cliente, este es atendido por un servidor que esté libre; si todos están ocupados el cliente espera en la cola hasta que haya un servidor libre.

- Los tiempos entre llegadas de clientes tienen una distribución exponencial de tasa λ
- El tiempo de servicio en cada servidor tiene una distribución exponencial de tasa μ

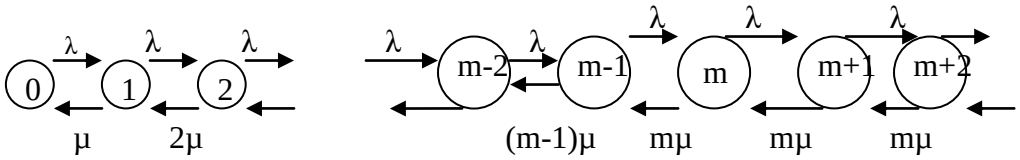
Es de destacar que la tasa efectiva de servicio depende del estado del sistema:

- Si hay n<m clientes en el sistema, hay n servidores ocupados y la tasa de servicio es $n \cdot \mu$
- Si hay n>m clientes en el sistema, entonces la tasa de servicio alcanza su máximo valor, que es $m \cdot \mu$

El sistema se puede modelar como una cadena de nacimiento-muerte de parámetros:

$$\lambda_n = \lambda \quad para\ todo\ n = 0,1,2,\ldots$$

$$\mu_n = \begin{cases} n \cdot \mu & si\ 0 \le n < m \\ m \cdot \mu & si\ n \ge m \end{cases}$$



Podemos obtener la probabilidad estacionaria de que la cola esté vacía:

$$\Pi_0 = \left[1 + \sum_{n=1}^{m-1} \frac{\lambda^n}{(\mu)\cdot(2\mu)\cdots(n\mu)} + \frac{\lambda^{m-1}}{(m-1)!\cdot\mu^{m-1}} \sum_{n=m}^{\infty} \left(\frac{\lambda}{m\mu} \right)^{n-m+1} \right]^{-1}$$

El segundo sumatorio es una progresión geométrica, que converge si $\frac{\lambda}{m\mu} \le 1$, que coincide con la idea intuitiva de que la tasa de llegada no debe superar la máxima tasa de servicio para que la cola sea estable.

Vamos a considerar $\rho = \frac{\lambda}{m\mu}$; la suma infinita es:

$$\sum_{n=m}^{\infty} \left(\frac{\lambda}{m\mu} \right)^{n-m+1} = \sum_{n=m}^{\infty} \rho^{n-m+1} = \frac{\rho}{\rho - 1}$$

de donde:

$$\Pi_0 = \left[1 + \sum \frac{(m\rho)^m}{n!} + \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho} \right]^{-1}$$

Para obtener Π_n podemos distinguir dos casos

- si $n < m$

$$\Pi_n = \left(\frac{\lambda^n}{\mu \cdot (2\mu) \cdots (n\mu)} \right) \cdot \Pi_0 = \frac{\left(\lambda / \mu\right)^n}{n!} \Pi_0$$

- si $n \geq m$

$$\Pi_n = \left(\frac{\lambda^{m-1}}{(\mu) \cdot (2\mu) \cdots (m-1)\mu} \right) \cdot \left(\frac{\lambda^{n-m+1}}{(m\mu)^{n-m+1}} \right) \cdot \Pi_0 = \frac{m^m}{m!} \left(\frac{\lambda}{m\mu} \right)^n \Pi_0$$

que podemos expresar:

$$\Pi_n = \begin{cases} \Pi_0 \dfrac{(m\rho)^n}{n!} & n = 1,2,\ldots,m-1 \\ \Pi_0 \dfrac{m^m}{m!} \rho^n & n = m, m+1, \ldots \end{cases}$$

6.2.1 Utilización y rendimiento

Vamos a considerar la variable aleatoria B que indica el número de servidores ocupados; su esperanza es:

$$E[B] = \sum_{n=0}^{m-1} n \cdot \Pi_n + m \cdot P[X \geq m]$$

$P[X] \geq m$ es la probabilidad de que estén presentes, al menos, m clientes:

$$P[X \geq m] = \sum_{n=m}^{\infty} \Pi_n = \sum_{n=m}^{\infty} \frac{m^m}{m!} \rho^n \Pi_0 = \frac{m^m}{m!} \frac{\rho^m}{1-\rho} \Pi_0$$

de donde:

$$E[B]=\left[\sum_{n=0}^{m-1} n\frac{(m\cdot\rho)^n}{n!}+m\frac{(m\rho)^m}{m!}\frac{1}{1-\rho}\right]\cdot\Pi_0=m\rho\left[1+\sum_{n=2}^{m-1}\frac{(m\cdot\rho)^{n-1}}{(n-1)!}+\frac{(m\cdot\rho)^{m-1}}{m!}\frac{m}{1-\rho}\right]\cdot\Pi_0$$

$$=m\rho\left[1+\sum_{j=1}^{m-1}\left(\frac{(m\rho)^m}{j!}\right)+\frac{(m\rho)}{m!}\frac{1}{1-\rho}\right]\cdot\Pi_0=m\cdot\rho=\frac{\lambda}{\mu}$$

el rendimiento del sistema es λ ya que en estacionario las tasas de llegada y de salida han de ser iguales.

6.2.2 Longitud media de la cola

Es la esperanza matemática de la variable X:

$$E[X]=\sum_{n=0}^{\infty} n\cdot\Pi_n=m\rho+\frac{(m\rho)}{m!}\frac{\rho}{(1-\rho)^2}\Pi_0$$

como cabe esperar:

- cuando $\rho\to 0,\quad E[X]\to 0$
- cuando $\rho\to 1,\quad E[X]\to\infty$

6.2.3 Tiempo medio en el sistema

Aplicando la ley de Little, se obtiene:

$$m\rho + \frac{(m\rho)}{m!}\frac{\rho}{(1-\rho)^2}\Pi_0 = \lambda \cdot E[S]$$

de donde

$$E[S] = \frac{1}{\mu} + \frac{1}{\mu}\frac{(m\cdot\rho)^m}{m!}\frac{\Pi_0}{m(1-\rho)^2}$$

- cuando $\rho \to 0, \quad E[S] \to 1/\mu$, que indica que para valores bajos del tráfico, el cliente es atendido inmediatamente.

6.2.4 Probabilidad de permanencia en cola

Es la probabilidad de que al llegar un cliente todos los servidores estén ocupados y el cliente debe permanecer en la cola; se denota por P_Q. Se verifica que:

$$P_Q = P[X \geq m] = \sum_{n=m}^{\infty}\Pi_n = \frac{(m\cdot\rho)^m}{m!}\frac{\Pi_0}{1-\rho}$$

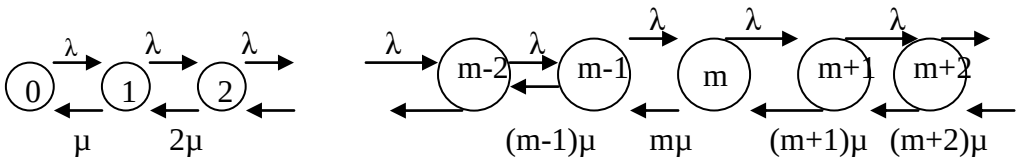
6.3 LA COLA M/M/∞

Esta cola puede concebirse como un caso particular de la cola M/M/m, cuando el número de servidores es ∞. El proceso es de nacimiento-muerte, con parámetros:

$\lambda_n = \lambda \quad$ *para todo* $n = 0,1,\ldots$

$\mu_n = n\mu \quad$ *para todo* $n = 1,2,\ldots$

siendo λ la tasa de llegada de los clientes y μ la tasa de servicio de cada uno de los servidores



La probabilidad estacionaria de que la cola esté vacía, viene dada por:

$$\Pi_0 = \left[1 + \sum_{n=1}^{\infty} \frac{\lambda^n}{(\mu)\cdot(2\mu)\cdots(n\mu)}\right]^{-1} = \left[1 + \sum_{n=1}^{\infty} \frac{\left(\lambda/\mu\right)^n}{n!}\right]^{-1}$$

si $\rho = \lambda/\mu < \infty$ la serie converge y se obtiene:

$$\Pi_0 = e^{-\rho}$$

se debe recalcar que ρ ***no es*** la intensidad de tráfico.

Por otra parte:

$$\Pi_n = e^{-\rho}\frac{\rho^n}{n!} \quad n = 0,1,2\ldots$$

es la distribución estacionaria del número de clientes en el sistema; se observa que es una distribución de Poisson con parámetro ρ.

6.3.1 Utilización y rendimiento

La utilización del sistema se obtiene de forma inmediata:

$$1-\Pi_0 = 1-e^{-\rho}$$

el rendimiento es igual ala tasa de llegada λ

6.3.2 Longitud media de la cola

Es la esperanza de la variable aleatoria X, que como hemos visto tiene una distribución de Poisson:

$$E[X] = \rho = \frac{\lambda}{\mu}$$

en este caso, no existe una cola física y , de hecho, X es el número de servidores ocupados.

6.3.3 Tiempo medio en el sistema

Aplicando la ley de Little, se obtiene:

$$\rho = \lambda \cdot E[S \Leftrightarrow E[S]] = \frac{1}{\mu}$$

Obviamente, como siempre hay un servidor libre para un nuevo cliente, éste solo está en el sistema el tiempo de servicio

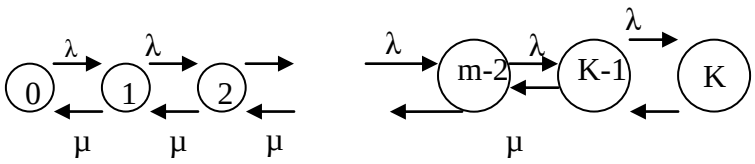
6.4 LA COLA M/M/1/K

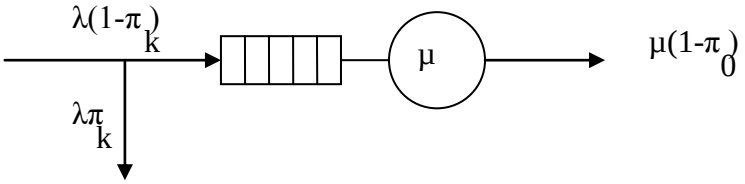
En este caso, la longitud de la cola está limitada a K clientes, de manera que si llega alguno más se rechaza; este cliente rechazado se considera perdido, y la situación se conoce como bloqueo.

Por otra parte, la cola es semejante a la cola M/M/1 y podemos utilizar un modelo de nacimiento-muerte para analisis; la tasa de llegada es $\lambda_n = \lambda$ para todo n=0,1,2 …K-1 y $\lambda_n = 0$ para n $\geq K$

$$\lambda_n = \begin{cases} \lambda & si\ 0 \leq n < K \\ 0 & si\ n \geq K \end{cases}$$

$$\mu_n = \mu \quad para\ todo\ n = 1,2,\ldots,K$$





La probabilidad π_0 es:

$$\pi_0 = \left[1 + \sum_{n=1}^{K} \left(\frac{\lambda}{\mu} \right)^n \right]^{-1}$$

la suma es una serie geométrica finita que podemos evaluar:

$$\sum_{n=1}^{K} \left(\frac{\lambda}{\mu} \right)^n = \frac{\left(\lambda / \mu \right) \cdot \left(1 - \lambda / \mu \right)}{1 - \lambda / \mu}$$

intuitivamente , si $\lambda > \mu$, se pierden clientes, pero la longitud de la cola permanece acotada y no se produce una situación de inestabilidad (crecimiento indefinido de la cola).

Vamos a definir $\rho = \lambda / \mu$, tenemos que:

$$\pi_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

$$\pi_n = \begin{cases} \dfrac{1 - \rho}{1 - \rho^{K+1}} \rho^n & si\ 0 \le n \le K \\ 0 & si\ n > K \end{cases}$$

que nos proporciona la distribución de probabilidad estacionaria.

6.4.1 Utilización y rendimiento

La utilización del servidor viene dada por:

$$1-\pi_0 = \rho\frac{1-\rho^K}{1-\rho^{K+1}}$$

dado que ρ puede ser tan grande como se quiera se observa que cuando $\rho \to \infty$ la utilización tiende a 1.

El rendimiento es dado por la tasa de salida:

$$\mu(1-\pi_0) = \lambda\frac{1-\rho^K}{1-\rho^{K+1}}$$

que es inferior a la tasa de llegada λ

6.4.2 Probabilidad de bloqueo

Quizá la medida mas importante de comportamiento de estas colas sea la probabilidad de que al llegar un clienta, sea rechazado por encontrarse la cola llena. Esta probabilidad se denomina probabilidad de bloqueo:

$$P_B = \pi_k = (1-\rho)\frac{\rho^K}{1-\rho^{K+1}}$$

si mantenemos ρ<1 y hacemos $k \to \infty$ el modelo se aproxima al de la cola estable M/M/1 y $P_B \to 0$.

- Se verifica que $\mu(1-\pi_0) = \lambda(1-P_b)$ que, simplemente refleja el equilibrio de flujo en el sistema.

6.4.3 Longitud media de la cola

Es el valor esperado de X, es decir:

$$E[X] = \sum_{n=0}^{K} n \cdot \pi_n = \frac{1-\rho}{1-\rho^{K+1}} \sum_{n=0}^{K} n \cdot \rho^n = \frac{\rho}{1-\rho^{K+1}} \left[\frac{1-\rho^K}{1-\rho} - K \cdot \rho^K \right]$$

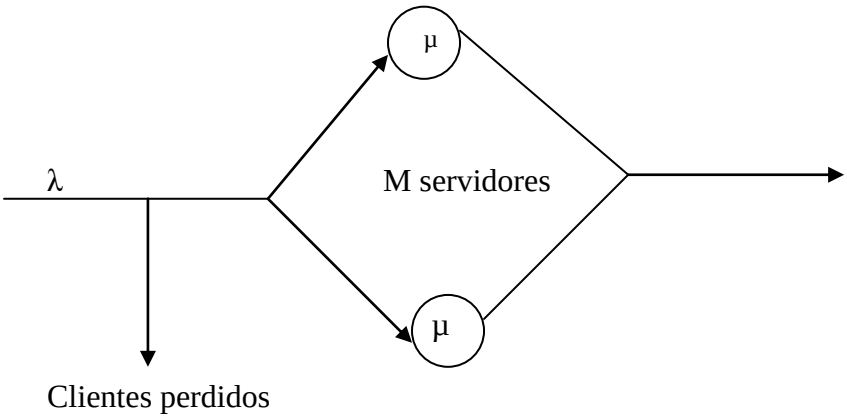
aplicando la ley de Little:

$$E[X] = \lambda(1-\pi_K) \cdot E[S]$$

donde $\lambda(1-\pi_K)$ es la tasa de clientes admitidos que es diferente de λ, ya que esta es la tasa de clientes que llegan antes de que algunos clientes puedan ser rechazados

6.5 LA COLA M/M/m/m

Esta cola está formada por m servidores idénticos y sin ningún espacio de almacenamiento para los clientes, de tal manera que si un cliente llega cuando están ocupados los m servidores el cliente es bloqueado y, por tanto, se pierde.



La tasa de servicio depende del número de clientes en ele sistema, de tal manera que el modelo de cadena de nacimiento-muerte es:

$$\lambda_n = \begin{cases} \lambda & si\ 0 \le n < m \\ 0 & si\ n \ge m \end{cases}$$

$$\mu_n = n\mu \quad para\ todo\ n = 1,2\ldots,m$$

El diagrama de transición de estados es:

La probabilidad estacionaria de que la cola esté vacía viene dada por:

$$\pi_0 = \left[1 + \sum_{n=1}^{m} \frac{\lambda^n}{(\mu)\cdot(2\mu)\cdots(n\mu)}\right]^{-1} = \left[1 + \sum_{n=1}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-.1}$$

Tomamos $\rho = \lambda / \mu$, teniendo en cuenta que no representa la intensidad de tráfico. La suma es finita, pero aparecen problemas de convergencia ya que ρ pudede ser arbitrariamente grande; esto implica que crece el número de clientes son bloqueados a medida que $\lambda > \mu$, pero el número de clientes el el sistema está acotado por m. Se obtiene la distribución de probabilidad:

$$\pi_0 = \left[\sum_{n=0}^{m} \frac{\rho^n}{n!}\right]^{-1}$$

$$\pi_n = \begin{cases} \dfrac{1}{\sum_{j=0}^{m} \dfrac{\rho^j}{j!}} \dfrac{\rho^n}{n!} & si\ 0 \le n \le m \\ 0 & si\ n > m \end{cases}$$

6.5.1 Probabilidad de bloqueo

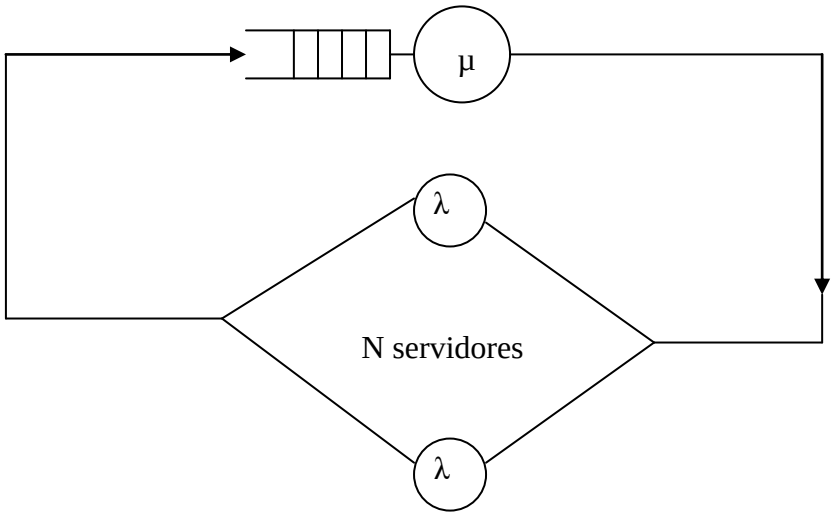
La probabilidad de que un cliente sea bloqueado es la probabilidad de que al llegar un cliente, haya m clientes en el sistema :

$$P_B = \pi_m = \frac{\left(\rho^m / m! \right)}{\sum_{j=0}^{m} \frac{\rho^j}{j!}}$$

Esta ecuación se conoce como fórmula de Erlang y es ampliamente utilizada para el análisis de sistemas de telefonía.

6.6 LA COLA M/M/1//N

En esta cola tenemos un único servidor, con un espacio de cola infinito con una población de clientes finita y acotada por N. El servidor tiene una distribución exponencial con una tasa de servicio de μ. Cuando un cliente recibe servicio desaloja el servidor y vuelve ala cola para recibir servicio tras un tiempo aleatorio modelado con una distribución exponencial de parámetro
λ.



La cola M/M/1//N puede sermodelada como una cadena de nacimiento-muerte en la cual el estado es el número de clientes en la cola; si este número es n entonces hay N-n clientes que volverán a la cola y que, por tanto, ocupan N-n servidores de parámetro λ. La cadena de nacimiento-muerte es:

Tenemos la superposición de N-n procesos de Poisson y , por tanto, el modelo es:

$$\lambda_n = \begin{cases} (N-n)\lambda & si\ 0 \le n < N \\ 0 & si\ n \ge N \end{cases}$$

Se verifica que:

$$\pi_0 = \left[1 + \sum_{n=1}^{N} \frac{[N\lambda][(N-1)\lambda \cdots [(N-n+1)\lambda]]}{\mu^n} \right]$$

Dado que la suma es finita, no hay problemas de convergencia. Tomamos $\rho = \lambda / \mu$ y la expresión anterior es:

$$\pi_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \rho^n \right]^{-1}$$

Por otra parte

$$\pi_n = \frac{[N\lambda][(N-1)\lambda] \cdots [(N-n+1)\lambda]}{\mu^n} \pi_0 = \frac{N!}{(N-n)!} \rho^n \pi_0$$

de donde la distribución de probabilidad estacionaria es:

$$\pi_n = \begin{cases} \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!}\rho^n\right]^{-1} & si\ n = 0 \\ \pi_0 \frac{N!}{(N-n)!}\rho^n & si\ 1 \le n \le N \\ 0 & si\ n > N \end{cases}$$

6.6.1 Utilización y rendimiento

La utilización del servidor es $1-\pi_0$.
El rendimiento viene dado por π_0

6.6.2 Tiempo medio de respuesta

El *tiempo de respuesta* es una variable aleatoria R definida pòr el tiempo que transcurre desde que un cliente entra en la cola hasta que se ha completado su servicio.

Aplicando la ley de Little al servidor :

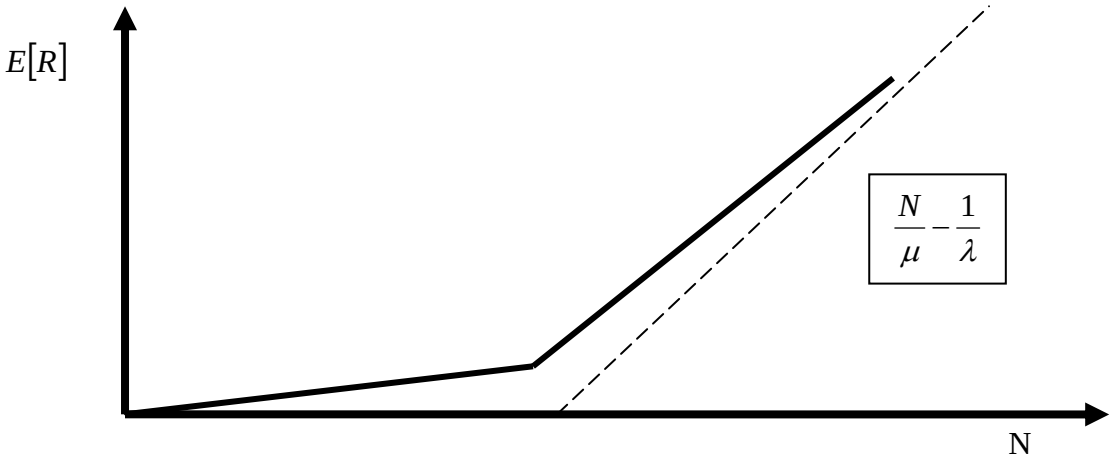
$$E[X] = \mu(1-\pi_0)E[R]$$

Por otra parte, aplicando la ley de Little a las servidores de tasa λ se obtiene:

$$E[N-X] = \mu(1-\pi_0)\frac{1}{\lambda}$$

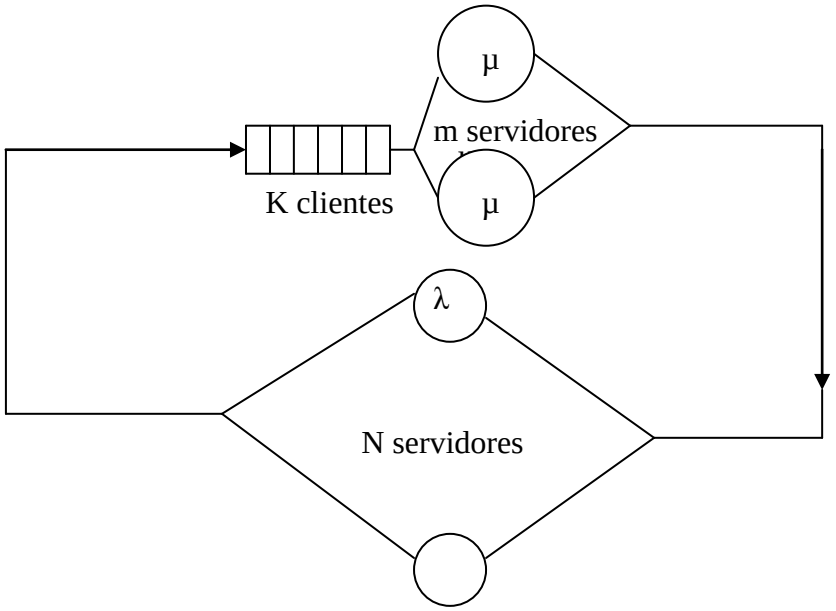
A partir de ambas ecuaciones:

$$E[R] = \frac{N}{\mu(1-\pi_0)} - \frac{1}{\lambda}$$



6.7 LA COLA M/M/m/K/N

Este tipo de cola engloba a todos los tipos anteriores, si permitimos que los parámetros m,k y N tomen valores finitos o infinitos.



Como se observa es similar a la cola que hemos visto an el apartado anterior, pero hay m servidores en vez de uno y el espacio dentro de la cola está limitado a K clientes, de manera que si llega un cliente más, esa rechazado.

El análisis de esta cola es bastante tedioso, así que simplemente expondremos los resultados más importantes.

La cadena de nacimiento-muerte que modela esta cola es:

$$\lambda_n = \begin{cases} (N-n)\lambda & si\ 0 \le n < K \\ 0 & si\ n \ge K \end{cases}$$

$$\mu_n = \begin{cases} n\mu & si\ 0 \le n < m \\ m\mu & si\ n \ge m \end{cases}$$

Es inmediato deducir que $\pi_n = 0$ para todo n>K.
Vamos a considerar $\rho = \lambda / \mu$; la distribución de probabilidad estacionaria es:

$$\pi_n = \begin{cases} \pi_0 \binom{N}{n} \rho^n & n = 1,2,\ldots,m-1 \\ \pi_0 \binom{N}{n} \frac{n!}{m!} m^{m-n} \rho^n & n = m, m+1,\ldots,K \end{cases}$$

y,

$$\pi_0 = \left[1 + \sum_{n=1}^{m-1} \binom{N}{n} \rho^n + \binom{N}{m-1} \rho^{m-1} \sum_{n=m}^{K} \frac{(N-m+1)!}{(N-n)!} \binom{\rho}{m}^{n-m+1} \right]^{-1}$$

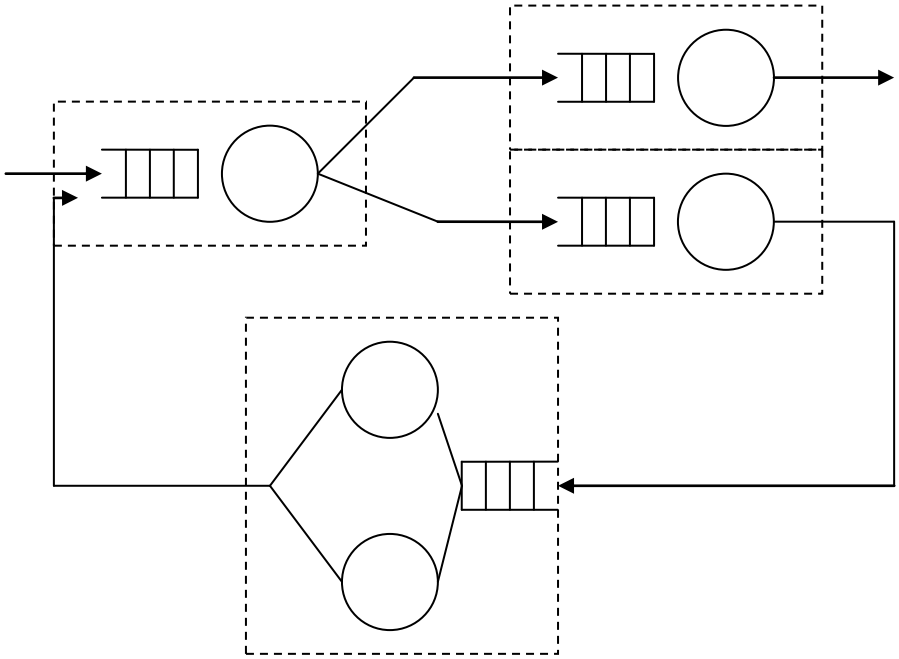
7. REDES DE COLAS MARKOVIANAS

Las colas de de espera que hemos considerado hasta ahora están formadas por una única cola con uno o varios servidores; en algún caso se permite que el cliente tras recibir servicio retorne a la cola.

En varios ámbitos de la vida real, se tiene sistemas formados por varios servidores cada uno con su cola, de tal manera que los clientes circulan de una cola a otra requiriendo diferentes servicios; un ejemplo son los sistemas de fabricación, en los cuales se puede modelar cada máquina por un servidor y las piezas que circulan de una máquina a otra son los clientes. Se justifica así que extendamos nuestro estudio a redes de colas, formadas por varias colas y en las cuales los clientes circulan de una cola a otra.
Las redes de colas pueden ser:

- *Redes de colas abiertas*: Los clientes proceden del exterior y tras recibir los servicios requeridos, abandonan el sistema.
- *Redes de colas cerradas:* El número de clientes en el sistema es fijo y circulan de una cola a otra requiriendo diferentes servicios.

En una rede de colas denominaremos nodo al conjunto de servidores con la cola asociada de tal manera que una red de colas está formada por la interconexión de varios nodos.



En los sistemas simples que hemos considerado hasta ahora, nuestro objetivo era obtener la distribución de probabilidad estacionaria del estado X, donde X es la longitud de la cola.
En una red de colas el estado X es un vector, cuya i-ésima componente corresponde al número de clientes en la cola de i-ésimo nodo:

$$X = [X_1, X_2, \ldots X_M]$$

el principal objetivo del análisis de una red de colas es la obtención de la distribución de probabilidad estacionaria de X:

$$\pi(n_1, n_2, \ldots n_M) = P[X_1 = n_1, X_2 = n_2 \ldots X_M = n_M]$$

Para todos los posibles valores de $n_1, \ldots n_M, \quad n_l = 0,1,2\ldots$

Nuestro objetivo en las secciones siguientes es presentar los principales resultados en el análisis de redes de colas Markovianas; esto implica que la llegada de clientes del exterior y los tiempos de servicio de los servidores son caracterizados por distribuciones exponenciales.
La primera cuestión que vamos a abordar es cómo son los procesos de salida de clientes de una cola Markoviana, ya que la composición de varios junto con los clientes que llegan del exterior conforman el proceso de llegada a un nodo.

7.1 El PROCESO DE SALIDA DE LA COLA M/M/1

En primer lugar nos ocuparemos del proceso de salida de una cola M/M/1; como es evidente si la cola siempre estuviera llena, el proceso sería de Poisson con tasa μ; pero dado que esta situación no es la real, el proceso de salida de clientes no sabemos como es.

Teorema: *El proceso de salida de una cola M/M/1 estable y estacionaria con tasa de llegada λ es un proceso de Poisson de tasa λ.*

El teorema anterior se denomina teorema de Burke y es de una gran importancia, ya que en el caso de redes de colas esta propiedad nos permite desacoplar las colas y estudiarlas por separado.

7.2 REDES DE COLAS ABIERTAS

Vamos a estudiar una red de colas abierta que está formada por varios nodos siendo cada uno de ellos una cola M/M/1. Asumiremos:

- Hay un único tipo de clientes.
- Tenemos M nodos.
- Cada nodo es una cola M/M/1
- Cada nodo está formado por un único servidor y una cola de capacidad infinita.
- Los nodos operan con una política de servir primero al cliente que llega antes.
- Los clientes circulan entre los nodos hasta que finalmente abandonan el sistema.
- Los clientes llegan del exterior al nodo i-ésimo siguiendo un proceso de Poisson de tasa r_i.
- La probabilidad de que un cliente tras recibir servicio en el nodo i-ésimo vaya al nodo j-ésimo viene dada por $p_{i,j}$

La tasa total de llegada al nodo i-ésimo es λ_i, siendo:

$$\lambda_i = r_i + \sum_{j=1}^{M} \lambda_j \; p_{i,j} \qquad i = 1,2,\ldots,M$$

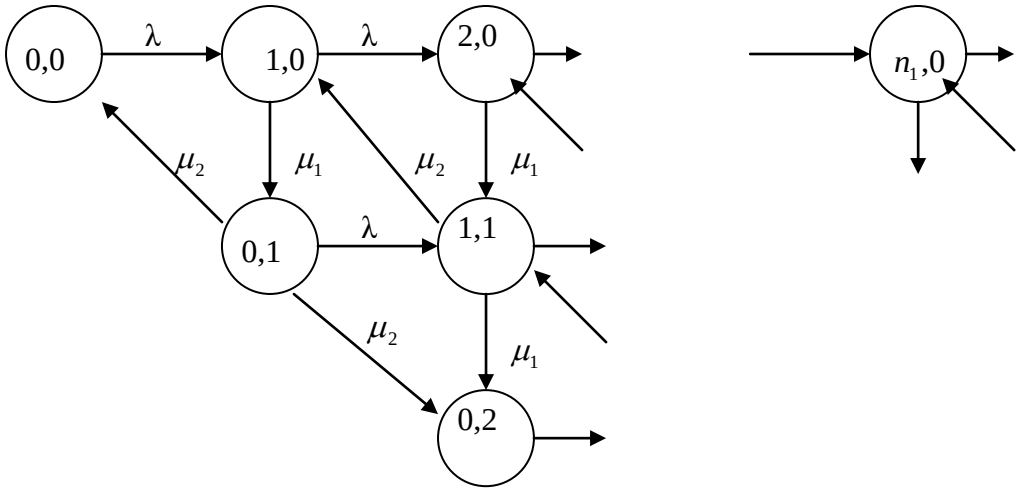
El primer sumando es el flujo de clientes desde el exterior y el siguiente sumando es el flujo de clientes desde otros nodos.

7.2.1 Colas en tandem

En primer lugar vamos a considerar el caso más simple, formado por dos colas M/M/1 en serie. En este caso el vector de estado está formado por dos componentes, conteniendo cada una el número de clientes en la cola $X = [X_1, X_2]$, por otra parte el conjunto de eventos que alteran el estado del sistema está formado por dos elementos:

- Llegada de un cliente a una cola.
- Salida de un cliente del servidor.

Dado que todos los procesos son procesos de Poisson, podemos modelar el sistema como una cadena de Harkov, cuyo diagrama de transición de estados es:



De donde, podemos :

$$\lambda \cdot \pi(n_1 - 1, n_2) + \mu_1 \cdot \pi(n_1 + 1, n_2 - 1) + \mu_2 \cdot (n_1, n_2 + 1) - (\lambda + \mu_1 + \mu_2) \cdot \pi(n_1, n_2) = 0$$

De forma similar, para los estados $(n_1,0)$ *siendo* $n_1 > 0$:

$$\lambda \cdot \pi(n_1 - 1,0) + \mu_1 \cdot \pi(n_1,1) + \mu_2 \cdot (n_1, n_2 + 1) - (\lambda + \mu_1) \cdot \pi(n_1,0) = 0$$

Y para los estados $(0,n_2)$ *siendo* $n_2 > 0$

$$\mu_1 \cdot \pi(1,n_2 -1) + \mu_2 \cdot \pi(0,n_2 +1) - (\lambda + \mu_2) \cdot \pi(0\ ,n_2) = 0$$

Finalmente, para el estado (0,0):

$$\mu\ \cdot \pi(0,1) - \lambda \cdot \pi(0,0) = 0$$

Además la suma de las probabilidades verifican:

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi(i,j) = 1$$

Este conjunto de ecuaciones puede ser resuelto y da lugar a :

$$\pi(n_1,n_2) = (1-\rho_1)\cdot\rho_1^{n_1}\cdot(1-\rho_2)\cdot\rho_2^{n_2}$$

siendo:

$\rho_1 = \frac{\lambda}{\mu_1}, \quad \rho_2 = \frac{\lambda}{\mu_2}$ las intensidades de tráfico de los nodos .

Cabe destacar que se ve a cada uno de los nodos por separado con unas distribuciones de probabilidad

$$\pi_1(n_1) = (1-\rho_1)\cdot\rho_1^{n_1}$$
$$\pi_2(n_2) = (1-\rho_2)\cdot\rho_2^{n_2}$$

Tenemos una solución en forma de producto:

$$\pi(n_1,n_2) = \pi_1(n_1)\cdot\pi_2(n_2)$$

Observar que cuando λ aumenta el nodo con menor tasa de servicio es el primero que causa la inestabilidad; por esta razon se lo denomina “nodo cuello de botella” de la red.

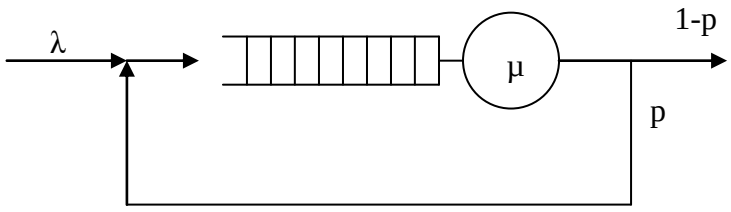
Este resultado es, de hecho, resultado de la aplicación del teorema de Burke. Éste permite el desacoplo de los dos nodos y analizarlos de manera separada como colas M/M/! separadas y, posteriormente, combinar los resultados.

7.2.2 Realimentación de clientes

Es de gran interés poder extrapolar el apartado anterior cuando algunos clientes que abandonan la cola se vuelven a incorporar a ella inmediatamente.

En primer lugar vamos a tratar por qué la realimentación de clientes puede crear problemas:

o



Consideremos la cola representada en la figura, el proceso de llegada de clientes externos es un proceso de Poisson de tasa λ. Los clientes, tras haber recibido servicio abandonan el sistema con una probabilidad 1-p o bien se incorporan de nuevo a la cola con una probabilidad p.
La dificultad aparece en que el proceso de llegada de clientes a la cola no es de Poisson, sin embargo el proceso de salida de clientes es de Poisson.
Jackson estableció que aun en presencia de realimentación de clientes, existe solución en forma de producto.

7.2.3 Solución en forma de producto

En el anterior subapartado se ha establecido que no es necesario en un nodo que el proceso de llegada sea de Poisson, pero el nodo se comporta como si fuesen procesos de Poisson y, por tanto, se tratan como si fuesen nodos M/M/1.
En general tenemos soluciones en forma de producto:

$$\pi(n_1, n_2, \ldots, n_m) = \pi_1(n_1) \cdot \pi_2(n_2) \cdots \pi_n(n_m)$$

Siendo $\pi_i(n_i)$ la solución de la i-ésima cola, que será una cola $M/M/m_i$ con tasa de servicio μ_i y tasa de llegada λ_i.
Para garantizar la existencia de solución de distribución de probabilidad estacionaria imponemos en cada nodo la condición:

$$\lambda_i = r_i + \sum_{j=1}^{M} \lambda \cdot p_{i,j} < m_i \mu_i$$

7.3 REDES DE COLAS CERRADAS

Una red de colas cerrada es aquella en la que el número de clientes en el sistema permanece fijo, no entran clientes desde el exterior ni salen clientes hacia el exterior.
Desde el punto de vista de modelado, una red de colas cerradas es una red de colas abierta tomando:

- $r_i = 0 \quad para\ todo\ i = 1,2,\ldots,M$
- $\sum_{j=1}^{M} p_{i,j} = 1 \quad para\ todo\ i = 1,2,\ldots,M$

Bajo estas condiciones, la suma del número de clientes en cada cola, esto es la suma de las variables de estado permanece constante:

$$\sum_{i=1}^{M} X_i = N$$

El espacio de estado es finito y está formado por las posibles asignaciones de N clientes entre M nodos; el número de asignaciones viene dado por:

$$\binom{M+N-1}{M-1} = \frac{(M+N-1)!}{(M-1)!\cdot N!}$$

Por otra parte, la expresión de balance de flujo implica:

$$\lambda_i = \sum \lambda_j \cdot p_{i,j} \quad i = 1,2\ldots,M$$

Existe una diferencia importante entre esta ecuación y la análoga para redes abiertas de colas. En las redes de colas abiertas tenemos M ecuaciones linealmente independientes. Por otra parte, en ausencia de clientes que procedan del exterior tenemos M-1 ecuaciones linealmente

independientes. Esto implica que tenemos sólo M-1 ecuaciones linealmente independientes por tanto, la solución $\lambda_1, \lambda_2, \ldots, \lambda_M$ contiene una constante libre cuya interpretación es la siguiente, supongamos que elegimos λ_1 como constante, entonces λ_i, $i \neq 1$ se interpreta como el rendimiento relativo del nodo i con respecto al del primer nodo.

En este tipo de redes tambien existe una solución en forme de producto para la distribución de probabilidad estacionaria $\pi(n_1, n_2, \ldots, n_M)$ con la restricción $\sum_{i=1}^{M} n_i = N$.

El punto de partida para la obtención de la distribución de probabilidad estacionaria es las ecuaciones de flujo de la cadena de Markov que modela la red cerrada de colas, teniendo en cuenta:

- Cualquier transición del estado $(n_1, n_2, \ldots n_M)$ se debe a la salida de un cliente de un nodo. Entonces el flujo de probabilidad que sale de este estado viene dado por $\sum_{i:n_i>0} \mu_i \cdot \pi(n_1, n_2, \ldots, n_M)$
- Cualquier transición hacia este estado se debe a la salida de un cliente de un nodo j con $n_j > 0$ el cual es enrutado hacia el nodo i. La tasa de transición de estado es $p_{i,j} \cdot \mu_j$.
- El estado resultante de la transición anterior es $(n_1, n_2, \ldots n_M)$, entonces el estado cuando el evento de tasa $p_{i,j} \cdot \mu_j$ tiene lugar será $(n_1, n_2, \ldots, n_j + 1, \ldots, n_i - 1, \ldots n_M)$; es decir, el cliente abandona el nodo j y va al nodo i. El flujo de probabilidad hacia $(n_1, n_2, \ldots n_M)$ es:

$$\sum_{j:n_j>0}\sum_{i} p_{i,j}\cdot\mu_j\cdot\pi\left(n_1,n_2,\ldots,n_j+1,\ldots,n_i-1,\ldots n_M\right)$$

Por tanto, la ecuación de balance de flujo es:

$$\sum_{i:n_i>0}\mu_i\cdot\pi(n_1,n_2,\ldots,n_M)=\sum_{j:n_j>0}\sum_{i} p_{i,j}\cdot\mu_j\cdot\pi\left(n_1,n_2,\ldots,n_j+1,\ldots,n_i-1,\ldots n_M\right)$$

para todo $n_1,n_2,\ldots,n_M$ que satisfaga $\sum_{i=1}^{M} n_i = N$.

La solución a estas ecuaciones es:

$$\pi(n_1,n_2,\ldots,n_M)=\frac{1}{C(N)}\cdot\rho_1^{n_1}\cdot\rho_2^{n_2}\cdots\rho_M^{n_M}$$

Siendo $\rho_i = \lambda_i / \mu_i$ con λ_i obtenido de $\lambda_i=\sum\lambda_j\cdot p_{i,j}\quad i=1,2\ldots,M$ con una constante libre elegida arbitrariamente.

C(N) es una constante que depende del número de clientes N y que se obtiene de la condición

$$\frac{1}{C(N)}\cdot\sum_{n_1,\ldots n_M}\rho_1^{n_1}\cdot\rho_2^{n_2}\cdots\rho_M^{n_M}=1$$

Entonces para obtener la distribución de probabilidad estacionaria se debe seguir los pasos:

- Resolver las ecuaciones lineales

  $$\lambda_i=\sum\lambda_j\cdot p_{i,j}\quad i=1,2\ldots,M$$

- Fijar una constante arbitraria
- Obtener C(N) de la expresión

  $$\frac{1}{C(N)}\cdot\sum_{n_1,\ldots n_M}\rho_1^{n_1}\cdot\rho_2^{n_2}\cdots\rho_M^{n_M}=1$$

La obtención de C(N) no es una tarea trivial; a continuación se describen algunos procedimientos.

7.3.1 Obtención de la constante C(N)

Para la computación de C(N) existen varios procedimientos computacionales; además estos algoritmos permiten evaluar varios aspectos de comportamiento de la red de colas sin que sea necesario hallar $\pi(n_1, n_2, \ldots, n_M)$

Uno de los más sencillos se debe a Buzen y está basado en la relación recursiva:

$$C_i(k) = C_{i-1}(k) + \rho_i \cdot C_i(k-1), \qquad i = 1,2,\ldots,M \ , \quad k = 2,3,\ldots,N$$

con condiciones iniciales

$$C_1(k) = \rho_1^k \qquad k = 1,2,\ldots,N$$

$$C_i(1) = 1 \qquad i = 1,2,\ldots,M$$

de donde:

$$C(N) = C_M(N)$$

Puede demostrarse que la utilización del nodo i cuando el número de clientes es N, viene dado por:

$$\mu_i[1 - \pi_i(0)] = \rho_i \frac{C(N-1)}{C(N)}$$

7.3.2 Análisis del valor medio

Supongamos hora que estamos interesados sólo en la obtención de medidas de comportamiento, como son el rendimiento y los valores medios de longitud de la cola. En este caso Reiser y Lavenberg desarrollaron un procedimiento que permite obviar la computación de C(N).

Consideremos un cliente que llega al nodo i y sea $\bar{S}_i$ el tiempo medio del cliente en el sistema; por otra parte, sea $\bar{X}_i$ la longitud media de la cola cuando llega el cliente, se verifica que:

$$\bar{S}_i = \frac{1}{\mu_i} + \bar{X}_i \frac{1}{\mu_i}$$

Siendo $1/\mu_i$ el tiempo medio de servicio en el nodo i.

Se puede probar que en una cola cerrada con N clientes $\bar{X}_i$ es igual que la longitud media de la cola en el i-ésimo nodo en una red con N-1 clientes. Si :

- $\bar{X}_i(N)$ es la longitud media de la cola en el nodo i
- $\bar{S}_i(N)$ el tiempo medio en el sistema en el nodo i

cuando tenemos N clientes podemos escribir la expresión recursiva:

$$\bar{S}_i(N) = \frac{1}{\mu_i}\left[1 + \bar{X}_i(N-1)\right], \qquad i = 1,2,\ldots,M$$

con condiciones iniciales

$$\bar{X}_i(0) = 0, \qquad i = 1,2,\ldots,M$$

Por otra parte utilizando la ley de Little tenemos:

$$N = \Lambda_N \sum_{i=1}^{M} \bar{X}_i(N)$$

siendo Λ_N el rendimiento y N el número de clientes en la red.

Se debe tener en cuenta que el rendimiento en estado estacionario debe ser el mismo en todos los nodos aplicando la ley de Little a un nodo y tomando

$$\bar{X}_i(N) = \Lambda_N \cdot \bar{S}_i(N) \qquad i = 1,2,\ldots,M$$

Obtenemos un conjunto de ecuaciones que definen un algoritmo mediante $\bar{X}_i(N), \bar{S}_i(N)$ y Λ_N puede ser evaluado para varios valores de N=1,2,…

7.4 REDES CON SOLUCIÓN EN FORMA DE PRODUCTO

Las redes que hemos visto en los apartados anteriores reciben el nombre de redes con solución en forma de producto debido a que la distribución de probabilidad estacionaria puede ser expresada como un producto de términos asociados cada uno de ellos a un nodo.

Se puede pensar que es la naturaleza Markoviana del sistema la razón fundamental para que la solución en forma de producto sea una descomposición de soluciones asociadas a los diferentes nodos. Además hemos considerado un único tipo de clientes y una política de atender antes al cliente que llega antes.

Existen redes de colas mucho más complejas que también admiten solución en forma de producto; este hecho sugiere que no es la naturaleza Markoviana del proceso sino más bien la estructura de la red de colas es lo que permite una descomposición.

La red con solución en forma de producto más destacada es la denominada BMCP debida a Baskett,Candí,Munt y Palacios (1975).

Esta red es una red de colas cerrada con K tipos de clientes; cada clase de cliente está caracterizada por su propia probabilidad de enrutado y su propia tasa de servicio, es decir:

- $p_{i,j}^{k}$ es la probabilidad de que un cliente de clase k sea enrutado desde el nodo i al nodo j
- μ_i^k es la tasa de servicio de clientes de clase k en el nodo i

Se permite la existencia de cuatro tipos de nodos:

1. Nodos con un único servidor con tiempos de servicio distribuidos exponencialmente y $\mu_i^k = \mu_i$ para todas las clases de clientes. A politica de servicio es dar servicio antes al cliente que llega primero.
2. Nodos con un único servidor y cualquier distribución del tiempo de servicio, con posibilidad de que sea diferente para cada clase, siempre y cuando la distribución sea diferenciable. La política de servicio puede ser de tipo compartir procesador , esto es , cada cliente recibe un periodo de tiempo fijo de servicio y cuando ha terminado el tiempo para el cliente, si no se ha finalizado el servicio, el cliente se incorpora de nuevo a la cola.
3. El mismo tipo de nodos que en el caso 2, pero la política de servicio es atender primero al cliente que ha llegado el último con reemplazo, es decir, un cliente es desalojado del procesador si llega un nuevo cliente.

4. Nodos con un infinito número de servidores y cualquier distribución del tiempo de servicio, posiblemente diferente para cada clase de cliente siempre y cuando la distribución sea diferenciable.

En este tipo de redes el estado en cada nodo es de la forma: $X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,k}]$ siendo $X_{i,k}$ el número de clientes de clase k en el nodo i. El vector de estado de la red es: $X = [X_1, X_2, \ldots, X_M]$. Asumiendo que el número de clientes de clase k es N_k se debe verificar para todo k que $\sum_{i=1}^{M} X_{i,k} = N_k$.

Aunque la notación puede ser complicada lo mas impòrtante es indicar que existen soluciones en forma de producto para redes con nodos con distribución no exponencial y con diferentes políticas de servicio.

8 COLAS NO MARKOVIANAS

En el estudio de las colas que hemos desarrollado hasta ahora ha sido de vital importancia que podamos modelar el comportamiento como procesos Markovianos y más concretamente como cadenas de nacimiento-muerte.
A continuación vamos a considerar colas de espera con procesos no- Markovianos , más concretamente procesos semimarkovianos generalizados. La mayor complicación surge del hecho de que el estado del sistema no puede ser descrito solamente por el número de clientes en la cola sino que ha de tenerse en cuenta el tiempo, ya que los procesos ahora no carecen de memoria.

Este hecho es más claro si tenemos en cuenta si hacemos referencia al mecanismo para determinación del evento E' que provoca la transición en un proceso semi-Markoviano generalizado generalizado . Sea x el estado actual con un conjunto factible de eventos I'(x). Cada evento $j \in \Gamma(x)$ tiene un valor de reloj Y_j; Entonces, la probabilidad de que el evento que provoca la transición sea un $i \in \Gamma(x)$ es dada por la probabilidad de que el evento i tenga el menor valor del tiempo de reloj de entre todos los eventos en $\Gamma(x)$:

$$P[E' = i] = P\left[Y_j = \min_{j \in \Gamma(x)} \{Y_j\} \right]$$

Para la determinación de esta probabilidad necesitamos información de las variables aleatorias $Y_j,\ j \in \Gamma(x).$ Es en el caso de cadenas de Markov en el que la propiedad de no tener memoria nos permite obtener:

$$P[E' = i] = \frac{\lambda_i}{\Lambda(x)}$$

siendo λ_i la tasa de Poisson del evento i y $\Lambda(x) = \sum_{j \in \Gamma(x)} \lambda_i$. En este caso no es necesario tener información de los valores de reloj; por ejemplo en el caso de una cola M/M/1 con eventos de llegada a y eventos de salida d la probabilidad para todos los estados es:

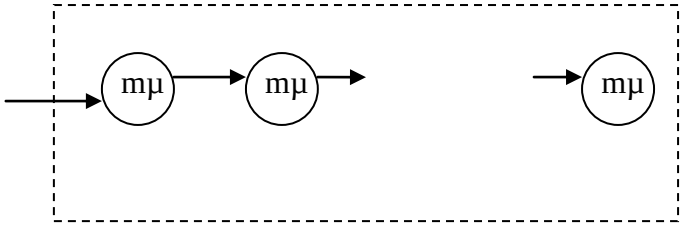
$$P[E' = a] = \frac{\lambda}{\lambda + \mu} \qquad P[E' = d] = \frac{\mu}{\lambda + \mu}$$

Para tratar procesos de eventos no Markovianos, hay dos posibles técnicas:

- Construir el proceso no Markoviano a partir de la combinación de procesos Markovianos, de tal manera que los procesos Markovianos constituyen bloques que sirven para construir modelos más complicados.
- Aprovechar las propiedades estructurales que no son influenciadas por la naturaleza de los procesos de los eventos; este procedimiento es más complicado.

8.1 MÉTODO DE LOS PASOS

Vamos a considerar un conjunto de eventos $e_1, e_2, \ldots, e_m$ cada uno de ellos generado por un proceso de Poisson de tasa λ. La manera más simple de combinar estos eventos para obtener uno más complejo es que ocurran en serie, es decir el evento e ocurre cuando los eventos $e_1, e_2, \ldots, e_m$ han tenido lugar de forma consecutiva. Sea z_i el tiempo de vida del evento e_i, entonces el tiempo de vida del evento e viene dado por $Z = Z_1 + Z_2 + \cdots + Z_m$. Para obtener la distribución del tiempo de vida de este nuevo evento vamos a considerar m servidores en serie:



Servidor de m pasos

Se considera que cuando un cliente entra en el primero de los pasos ocupa todos los demás hasta que ha recibido servicio en todos los pasos. Si z_i es el tiempo de servicio en el i-ésimo paso se verifica que:

$$Z = Z_1 + Z_2 + \cdots + Z_m$$

Dado que z_i tiene una distribución exponencial con tasa $m\mu$, se verifica que:

$$E[Z] = m\left(\frac{1}{m \cdot \mu}\right) = \frac{1}{\mu}$$

Además el tiempo promedio de servicio en este sistema es $\frac{1}{\mu}$.

A continuación vamos a determinar la distribución del tiempo de servicio. Asumiendo que cada uno de los pasos es independiente del resto, podemos obtener la fución de densidad de probabilidad Z como la convolución de m funciones de densidad de probabilidad exponenciales cada una con tasa $m \cdot \mu$. Omitiendo los detalles:

$$f_z(t) = \frac{1}{(m-1)!} \cdot m \cdot \mu \cdot (m \cdot \mu \cdot t)^{m-1} e^{-m\mu t} \qquad t > 0$$

$$F_Z(t) = 1 - e^{m\mu t} \sum_{i=0}^{m-1} \frac{(m \cdot \mu \cdot t)^i}{i!} \qquad t > 0$$

Esta distribución recibe el nombre de distribución de Erlang y se denota como E_m; ésta distribución tiene dos parámetros que son m y μ; sin embargo su media sólo de pende de μ. La importancia de esto reside en que podemos construir una variedad de tiempos de servicio no exponenciales con la misma esperanza matemática,pero diferentes momentos de orden superior, por ejemplo la varianza de Z es:

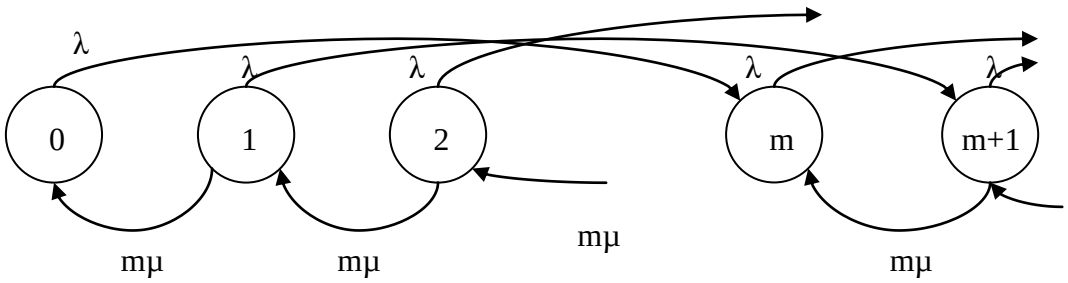
$$Var[Z] = \frac{1}{m \cdot \mu^2}$$

Mediante la variación de m podemos generar distribuciones de servicio que van desde la puramente exponencial (m=1) hasta la determinista cuando $m, \to \infty$. La varianza de la distribución que podemos construir no puede ser mayor que la de una distribución exponencial, $1/\mu^2$.

El servidor de Erlang permite modelar una clase de eventos no Markovianos preservando una estructura Markoviana. La cola construida se denota como $M/E_m/1$ en la que la llegada de clientes es un proceso de Poisson de tasa λ que son servidos por un servidor de Erlang de m pasos; en este caso la información para especificar el estado del sistema tiene dos partes:

- La longitud de la cola $X, X \in \{1,2,\ldots\}$
- El paso en el que se encuentra el cliente que está siendo servido.

Sin embargo esta descripción del estado puede ser condensada en una única variable de estado $\tilde{x}$ definida como el número total de pasos en el sistema, es decir: $\tilde{X} = 0 \; si \; X = 0 \; y \; \tilde{X} = (X-1)m + (m-K+1) = MX - K + 1 \; si \; X > 0$ el diagrama de estados que resulta es



El modelo es una cadena de Markov, pero no es una cadena de nacimiento-muerte debido a que una llegada de cliente causa una transición desde el estado n al estado n+m.
Una de las limitaciones del servidor de Erlang de varios pasos es que las distribuciones generadas no pueden tener una varianza mayor que $1/\mu^2$

8.2 ANÁLISIS DEL VALOR MEDIO DE UNA COLA M/G/1

La cola más simple no Markoviana que puede ser analizada en detalle es la cola M/G/1 en la cual los clientes llegan según un proceso de Poisson, pero los tiempos de servicio forman una sucesión de variables aleatorias idénticamente distribuidas con una distribución de probabilidad arbitraria. Es posible obtener de forma explícita el estado (número de clientes en la cola); sin embargo, nos ocuparemos sólo de la obtención de la longitud promedio de la cola E[X] que se denomina análisis de valor medio.

El resultado obtenido es conocido como la formula de *Pollaczek-Kinchin* que tiene una gran cantidad de aplicaciones:

$$E[X] = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}\left(1-\mu^2\sigma^2\right)$$

Siendo:

- $1/\mu$ el tiempo medio de servicio.
- σ^2 la varianza de la distribución del tiempo de servicio.
- $\rho = \lambda/\mu$ es la intensidad de tráfico.
- λ es la tasa del proceso de llegada de Poisson