Índice

1 Introducción

La teoría de colas estudia la construcción y el análisis de modelos matemáticos de sistemas que dan servicio a clientes cuyos tiempos de llegada y requisitos de servicio son aleatorios. Las dos cuestiones básicas son el análisis y el diseño de tales sistemas. Las primeras se refieren a evaluar ciertas medidas del comportamiento de sistemas con parámetros y reglas de operación completamente especificadas. Los problemas de diseño se refieren a la determinación de parámetros y reglas de operación que den valores satisfactorios en las medidas de comportamiento.

El modelo básico de colas, a partir del que se pueden construir otros más complicados, consta de tres elementos:

1. *el proceso de entrada* describe las propiedades estadísticas de los instantes entre llegadas de clientes. Típicamente, viene expresado en términos de la distribución entre llegadas.

2. *el mecanismo de servicio* especifica el número de servidores y las propiedades estadísticas de los tiempos de servicio.

3. *la disciplina de cola* describe el comportamiento de los clientes que encuentran todos los servidores ocupados.

Se ha desarrollado una notación específica para describir los sistemas de espera. Se denomina *notación de Kendall*, en honor a David Kendall, al cuál se debe en su mayor parte. Esta notación es de la forma $A/B/c/K/m/Z$ donde A indica la distribución del tiempo entre llegadas, B la distribución del tiempo de servicio, c el número de servidores o canales de servicio, K la capacidad del sistema (máximo número de clientes permitido en el sistema), m el tamaño de la población o fuente de clientes, y Z la disciplina de la cola. En ocasiones, se utiliza la notación abreviada $A/B/c$, suponiéndose $K = \infty$, $m = \infty$ y $Z =$ FIFO (First In, First Out: el primero en llegar es el primero en ser servido).

La discusión habitual en teoría de colas comienza por resultados generales, para discutir después modelos particulares. Sea $N(t)$ el número de clientes en el sistema en el instante t; definamos $P_j(t) = P(N(t) = j)$ como la probabilidad de que el sistema esté en el estado j en el instante t. Es particularmente interesante en colas encontrar la distribución en equilibrio definida por

$$\pi_j = \lim_{t \to \infty} P_j(t)$$

De hecho, una parte importante de la teoría de colas estudia el cálculo de tales probabilidades y su uso en el cálculo de medidas de comportamiento.

Los casos más sencillos de colas corresponden a aquéllos en que las llegadas siguen una distribución de Poisson y los tiempos de servicio siguen una distribución exponencial. En estos casos, en la notación de Kendall $A = B = M$ indicando así, que tanto el tiempo entre llegadas como el tiempo de servicio es exponencial. El estudio de tales sistemas resultará relativamente sencillo si nos apoyamos en los **Procesos de Nacimiento y Muerte**, estudiados en el tema anterior. Si $N(t)$ representa el número de individuos en el sistema en el instante t y suponemos que siempre que hay n individuos en el sistema:

- se producen nuevas llegadas con tasa exponencial λ_n e
- independientemente, se producen salidas del sistema con tasa exponencial μ_n,

un proceso de nacimiento y muerte con tasas de llegada (nacimiento) $\{\lambda_n\}_{n=0}^{\infty}$ y tasas de salida (muerte) $\{\mu_n\}_{n=1}^{\infty}$ es una cadena de Markov en tiempo continuo con espacio de estados $\{0,\ 1, \ldots\}$, tasas de permanencia en cada estado

$$\begin{aligned} v_0 &= \lambda_0 \\ v_i &= \lambda_i + \mu_i, \quad i > 0 \end{aligned}$$

y probabilidades de transición

$$\begin{aligned} P_{01} &= 1 \\ P_{i,i+1} &= \frac{\lambda_i}{\lambda_i + \mu_i}, \quad i > 0 \\ P_{i,i-1} &= \frac{\mu_i}{\lambda_i + \mu_i}, \quad i > 0 \\ P_{i,j} &= 0, \quad i > 0,\ j \neq i+1,\ i-1 \end{aligned}$$

Es posible analizar la situación de equilibrio, que se corresponde al sistema de ecuaciones en diferencias:

$$\lambda_j \pi_j = \mu_{j+1} \pi_{j+1}, \quad j = 0, 1, 2, ..$$

que, con la condición de que las probabilidades sumen 1, tiene la solución

$$\pi_n = \frac{\lambda_0 ... \lambda_{n-1}}{\mu_1 ... \mu_n} \pi_0, \quad n = 1, 2, ...$$

y

$$\pi_0 = \left(1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}\right)^{-1}$$

Una condición necesaria y suficiente para que existan tales probabilidades límite es que

$$\sum_{n=1}^{\infty} \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n} < \infty$$

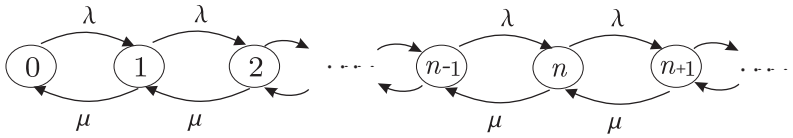
2 Sistema de Colas Exponencial con un Servidor: $M/M/1$

Supongamos que los clientes llegan a un sistema con un solo servidor según un proceso de Poisson de tasa λ. Cada cliente, al llegar, pasa directamente a ser servido si el servidor está libre y, si no, se une a la cola. Cuando el servidor completa el procesamiento de un cliente, éste abandona el sistema y el siguiente cliente en la cola, si hay alguno, pasa a ser servido.

Si $N(t)$ representa el número de individuos en el sistema en el instante t, entonces $\{N(t), t \geq 0\}$ es un proceso de nacimiento y muerte con

$$\begin{aligned} \lambda_n &= \lambda, \quad n \geq 0 \\ \mu_n &= \mu, \quad n \geq 1 \end{aligned}$$

Así, el diagrama de transición de estados para un sistema de colas $M/M/1$ es



que nos conduce al sistema de ecuaciones de equilibrio

$$\begin{aligned} \lambda\pi_0 &= \mu\pi_1 \\ (\lambda + \mu)\pi_n &= \lambda\pi_{n-1} + \mu\pi_{n+1} \\ \sum_{n=0}^{\infty} \pi_n &= 1 \\ \pi_n &\geq 0, \quad \forall n = 0, 1, \ldots \end{aligned}$$

Resolviendo el sistema o utilizando la fórmula vista para un proceso de nacimiento y muerte general, obtenemos

$$\pi_n = \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}\pi_0 = \left(\frac{\lambda}{\mu}\right)^n \pi_0, \quad n \geq 1$$

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{\frac{1}{1-\frac{\lambda}{\mu}}} = 1 - \frac{\lambda}{\mu} \text{ si } \frac{\lambda}{\mu} < 1$$

Observemos que la condición de estabilidad es $\frac{\lambda}{\mu} < 1$. Así,

$$\pi_n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n, \quad n \geq 0$$

Como el uso del servidor es $\rho = \frac{\lambda}{\mu}$ (probabilidad de que el servidor esté ocupado), equivalentemente podemos escribir

$$\pi_n = (1-\rho)\rho^n, \quad n \geq 0$$

que es la función de masa o probabilidad de una variable aleatoria geométrica generalizada de parámetro $1-\rho$ ($p = 1-\rho$, $q = \rho$) cuya media es

$$L = \frac{q}{p} = \frac{\rho}{1-\rho}$$

Directamente,

$$L = \sum_{n=1}^{\infty} n\pi_n = (1-\rho)\sum_n n\rho^n = (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

De las fórmulas de Little obtenemos el tiempo medio de respuesta o tiempo medio en el sistema,

$$W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$$

Como $W = W_q + W_s$, tenemos que el tiempo medio en cola es

$$W_q = W - W_s = \frac{1}{\mu(1-\rho)} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$$

Aplicando de nuevo las fórmulas de Little obtenemos el número medio de clientes en cola y en el servidor:

$$L_q = \lambda W_q = \frac{\lambda\rho}{\mu(1-\rho)} = \frac{\rho^2}{1-\rho}$$

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

Ejemplo 1: Supongamos que llegan trabajos a una unidad central según un proceso de Poisson de tasa 1 cada 12 minutos, y que los tiempos de servicio son exponenciales con tasa 1 sevicio cada 8 minutos. Calcular L y W.

Se tiene $\lambda = \dfrac{1}{12}$ trabajos/minuto y $\mu = \dfrac{1}{8}$ trabajos/minuto, por lo que

$$L = \frac{\lambda}{\mu - \lambda} = 2$$

$$W = \frac{L}{\lambda} = 24$$

Así, hay 2 trabajos de media en el sistema y cada trabajo está en el sistema 24 minutos de media.

Supongamos ahora que la tasa de llegada aumenta un 20%, de manera que $\lambda = \dfrac{1}{10}$. Entonces

$$L = 4 \quad \text{y} \quad W = 40$$

es decir, un aumento del 20% en la tasa de llegada o, equivalentemente, en el uso del servidor ρ, dobla el número medio de clientes en el sistema y aumenta en un 67% el tiempo medio de respuesta. □

Si representamos L y W en función de ρ obtenemos las gráficas de la Figura 1. Observemos

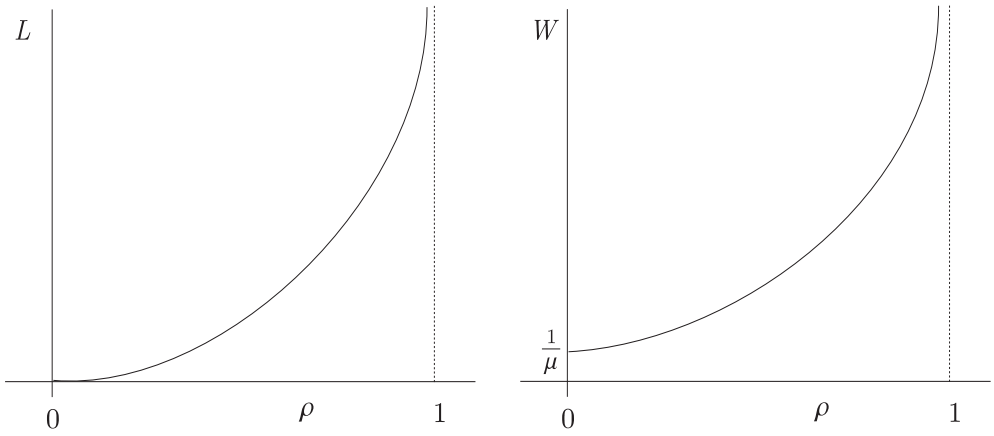


Figura 1: Representación gráfica de L y W

que si $\rho = \dfrac{\lambda}{\mu}$ está cerca de 1, un aumento ligero de ρ conduce a un mayor aumento en L y W. Para $\rho > 0.8$ crece muy rápidamente el valor de L y W. Además, como $W = W_q + W_s$, W_q crece también (si W_s permanece constante).

Con todo esto, tenemos bien caracterizado el comportamiento estacionario de la cola. Una característica interesante del sistema $M/M/1$ es que, además, es relativamente sencillo proporcionar la distribución del tiempo que un cliente pasa en la cola y en el sistema.

Sea w la variable aleatoria (v.a.) tiempo que pasa un cliente en el sistema y $W(t)$ su función de distribución. Para obtener $W(t)$ condicionamos en el número de clientes en el sistema a la llegada del cliente. Sea N la v.a. número de clientes en el sistema.

$$W(t) = P(w \leq t) = \sum_{n=0}^{\infty} P(w \leq t/N = n)P(N = n)$$

Ahora bien,

- Si $n = 0$, el cliente que llega estará en el sistema su tiempo de servicio.
- Si $n \geq 1$, habrá un cliente en servicio y $n - 1$ esperando delante de él. Debido a la pérdida de memoria de la exponencial (para el tiempo de servicio del primer cliente), el cliente deberá esperar un tiempo que es suma de $n + 1$ exponenciales independientes de parámetro μ, que sabemos que sigue una distribución Gamma con parámetros $a = \mu$ y $p = n + 1$ cuya función de densidad es

$$f(x) = \frac{\mu^{n+1} e^{-\mu x} x^n}{n!}, \quad x \geq 0$$

Así,

$$\begin{aligned} W(t) &= \sum_{n=0}^{\infty} \left(\int_0^t \mu e^{-\mu x} \frac{(\mu x)^n}{n!} dx \right) \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = \\ &= \int_0^t (\mu - \lambda) e^{-\mu x} \sum_{n=0}^{\infty} \frac{(\lambda x)^n}{n!} dx = \\ &= \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)x} dx = 1 - e^{-(\mu-\lambda)t} \end{aligned}$$

Por lo tanto, el tiempo que un cliente pasa en el sistema es una v.a. exponencial de parámetro $\mu - \lambda$, cuya media sabemos que es

$$E(w) = \frac{1}{\mu - \lambda}$$

como ya habíamos obtenido.

Sea ahora w_q el tiempo que pasa un cliente en la cola. Para obtener la función de distribución $W_q(t)$ procedemos de manera similar a como lo hacíamos para obtener $W(t)$.

$$W_q(t) = P(0 \leq w_q \leq t)$$

Notemos primero que

$$W_q(0) = P(w_q = 0) = P(0 \text{ clientes en el sistema}) = 1 - \rho$$

ya que ρ es la probabilidad de que el sistema esté ocupado.

Para $t > 0$, tenemos que

$$W_q(t) = P(w_q = 0) + P(0 < w_q \leq t)$$

$$P(0 < w_q \leq t) = \sum_{n=1}^{\infty} P(w_q \leq t/N = n)P(N = n)$$

Si un cliente llega cuando hay n clientes en el sistema ($n-1$ delante de él en la cola y 1 en servicio), tendrá que esperar en cola un tiempo que es suma de n tiempos exponenciales, i.e., de n v.a.i.i.d. según una exponencial de tasa μ, que es una Gamma de parámetros $a = \mu$ y $p = n$, cuya función de densidad es

$$f(x) = \frac{\mu^n e^{-\mu x} x^{n-1}}{(n-1)!}, \quad x \geq 0,\ n \geq 1$$

Así,

$$\begin{aligned} P(0 < w_q \leq t) &= \sum_{n=1}^{\infty} \left(\int_0^t \mu e^{-\mu x} \frac{(\mu x)^{n-1}}{(n-1)!} dx \right) \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = \\ &= \int_0^t (\mu - \lambda) \frac{\lambda}{\mu} e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\lambda x)^{n-1}}{(n-1)!} dx = \\ &= \frac{\lambda}{\mu} \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)x} dx = \rho(1 - e^{-(\mu-\lambda)t}) \end{aligned}$$

con lo que, si $t > 0$

$$\begin{aligned} W_q(t) &= P(w_q = 0) + P(0 < w_q \leq t) = \\ &= 1 - \rho + \rho(1 - e^{-(\mu-\lambda)t}) = 1 - \rho e^{-(\mu-\lambda)t} = \\ &= 1 - \rho e^{-\mu(1-\rho)t} = 1 - \rho e^{-t/W} \end{aligned}$$

Igual que antes podemos comprobar que $E(w_q) = W_q$.

El tiempo medio de espera en cola para aquellos clientes que deben esperar, i.e., hacer cola, es de particular interés para los sistemas en los que los clientes son personas. En ese caso, si se debe esperar durante mucho tiempo en la cola, los clientes pueden abandonar sin recibir servicio e ir a buscar tal servicio en otro sistema.

$$W_q = P(w_q = 0)E(w_q/w_q = 0) + P(w_q > 0)E(w_q/w_q > 0) = (1 - \rho) \times 0 + \rho E(w_q/w_q > 0)$$

por lo que

$$E(w_q/w_q > 0) = \frac{W_q}{\rho} = \frac{W_s}{1 - \rho} = W$$

Como $W = W_q + W_s$, en media, los clientes que deben esperar, estarán en la cola un tiempo de servicio más que el tiempo de espera en cola del cliente medio.

Ejemplo 2: Supongamos que llegan mensajes de forma completamente aleatoria con tasa de 240 mensajes por minuto a una central de comunicaciones. La velocidad de transmisión de la línea de la central es 800 caracteres por segundo. La distribución de la longitud de los mensajes es aproximadamente exponencial con longitud media de 176 caracteres. Calcular las principales medidas del sistema, supuesto que el acumulador es suficientemente grande. ¿Cuál es la probabilidad de que haya 10 ó más mensajes esperando a ser transmitidos? ¿Cuál es el tiempo medio de respuesta W si la tasa de tráfico de entrada a la central aumenta en un 10%?

La tasa de llegadas es

$$\lambda = 240 \text{ mensajes/min} = 4 \text{ mensajes/sg}$$

El tiempo medio de servicio es el tiempo medio necesario para transmitir un mensaje o

$$W_s = \frac{1}{\mu} = \frac{\text{longitud media del mensaje}}{\text{velocidad de la línea}} = \frac{176}{800} = 0.22 \text{ sg}$$

La intensidad de tráfico es

$$\lambda W_s = \frac{\lambda}{\mu} = 4 \times 0.22 = 0.88$$

que coincide con el uso del servidor ρ e implica que está transmitiendo el 88% del tiempo.

El paso a través del sistema es en este caso 4 (ya que $\lambda < \mu$).

Las otras medidas de interés son:

$$\begin{aligned} L &= \frac{\rho}{1-\rho} = 7.33 \text{ mensajes} \\ L_q &= \frac{\rho^2}{1-\rho} = 6.45 \text{ mensajes} \\ W &= \frac{W_s}{1-\rho} = 1.83 \text{ segundos} \\ W_q &= \rho W = W - W_s = 1.61 \text{ segundos} \end{aligned}$$

Como hemos obtenido la distribución de las v.a's w y w_q, podemos calcular percentiles. Por ejemplo, el valor del r-ésimo percentil de w, $\pi_w(r)$, está definido por

$$P(w \leq \pi_w(r)) = \frac{r}{100}$$

Calculemos el percentil 90 de w, $\pi_w(90)$: será el valor de t tal que

$$1 - e^{-(\mu-\lambda)t} = 0.9$$

Operando obtenemos

$$t = \frac{ln0.1}{\lambda - \mu} = 4.26 \text{ sg.}$$

Así, $\pi_w(90) = 4.26$ sg. y el 90% de los mensajes pasan menos de 4.26 segundos en el sistema.

Análogamente, podemos calcular el percentil 90 de w_q, $\pi_{w_q}(90)$. Éste será el valor de t tal que

$$1 - \rho e^{-(\mu-\lambda)t} = 0.9$$

Ahora obtenemos $t = 4.01$ sg., con lo que $\pi_{w_q}(90) = 4.01$ sg. y el 90% de los mensajes pasan menos de 4.01 segundos en cola.

Por otra parte, hay 10 ó más mensajes en cola si y sólo si hay 11 ó más mensajes en el sistema, con lo que la probabilidad de que haya 10 ó más mensajes esperando a ser transmitidos es

$$\sum_{n=11}^{\infty} \pi_n = \sum_{n=11}^{\infty} (1-\rho)\rho^n = \rho^{11} = 0.88^{11} \simeq 0.245$$

Si λ se incrementa en un 10%, $\lambda = 4.4$ mensajes/sg., con lo que $\rho = \dfrac{\lambda}{\mu} = 0.968$ (un incremento del 10% en λ equivale a un aumento del 10% en ρ). Así,

$$W = \frac{W_s}{1-\rho} = 6.875 \text{ sg.}$$

con lo que, como ya habíamos comentado con anterioridad, un pequeño aumento en λ, o equivalentemente en ρ conduce a un gran aumento de W (antes era 1.83 sg.). □

Ejemplo 3: Antes de integrar las componentes de un sistema los grupos de componentes deben pasar una revisión. Los grupos llegan al centro de revisión según un modelo de Poisson con tasa λ grupos/hora. La duración de la revisión parece seguir una exponencial con tiempo medio de servicio $\frac{1}{\mu}$ horas/grupo.

Se estima que el tiempo que pasan los grupos en el centro de revisión supone un coste de C_1 euros/grupo por hora.

El coste de adquisición y revisión del equipo es de $C_2\mu$ euros/hora, esté o no funcionando. Hallar el valor de μ que minimiza el coste de la operación de revisión.

Parece razonable suponer que podemos utilizar un sistema $M/M/1$ para el estudio de costes.

En los modelos de coste el objetivo es determinar el nivel de servicio (vía μ o c si el sistema es $M/M/c$) que minimice la suma de costes conflictivos. En este caso nos piden optimizar la tasa de servicio μ. Así, se supone que λ es fija y $\mu = \frac{1}{W_s}$ es controlable. Se tiene que C_1 representa el coste por unidad de tiempo de espera por cliente (por grupo) y C_2 el coste por unidad de incremento de μ por unidad de tiempo. Entonces

C_1L es el coste por unidad de tiempo debido a tener L clientes en el sistema.

$C_2\mu$ es el coste por unidad de tiempo debido a la tasa de servicio μ.

Así, la función de coste total (servicio + espera) por hora es

$$C(\mu) = C_1L + C_2\mu = C_1\frac{\lambda}{\mu - \lambda} + C_2\mu$$

Hay que obtener el valor μ que minimiza $C(\mu)$.

$$\frac{dC}{d\mu} = -\frac{\lambda}{(\mu-\lambda)^2}C_1 + C_2 = 0$$

Despejando μ obtenemos

$$\mu = \lambda \pm \sqrt{\lambda\frac{C_1}{C_2}}$$

Para ver en dónde se alcanza el mínimo, calculamos la derivada segunda.

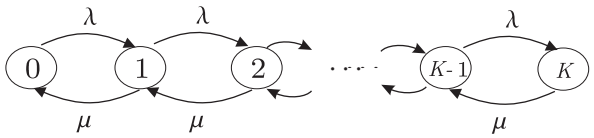
$$\frac{d^2C}{d\mu^2} = \frac{2\lambda}{(\mu-\lambda)^3}C_1$$

que se hace positiva para $\mu = \lambda + \sqrt{\lambda\frac{C_1}{C_2}}$. □

3 Sistema de Colas Exponencial con un Servidor y Capacidad Finita: $M/M/1/K$

En la práctica, suele ocurrir que el sistema tiene capacidad finita K, en el sentido de que si hay K individuos en el sistema no se producen entradas.

El diagrama de tasas de transición es



Las ecuaciones de equilibrio son:

$$\begin{aligned}
\lambda\pi_0 &= \mu\pi_1 \\
(\lambda+\mu)\pi_n &= \lambda\pi_{n-1} + \mu\pi_{n+1}, \quad 1 \le n \le K-1 \\
\mu\pi_K &= \lambda\pi_{K-1} \\
\sum_{n=0}^{K} \pi_n &= 1 \\
\pi_n &\ge 0, \quad \forall n = 0,\, 1, \ldots,\, K
\end{aligned}$$

Sumando cada ecuación con la anterior se pasa al sistema $\lambda\pi_{n-1} = \mu\pi_n$, para $n = 1, \ldots, K$, que unido a la condición de que la suma de las probabilidades tiene que ser 1, obtenemos:

- Si $\lambda = \mu$,

$$\pi_n = \frac{1}{K+1}, \quad \forall n = 0,\, 1, \ldots,\, K$$

- Si $\lambda \neq \mu$,

$$\pi_0 = \frac{1}{\displaystyle\sum_{n=0}^{K} \left(\frac{\lambda}{\mu}\right)^n} = \frac{1 - \dfrac{\lambda}{\mu}}{1 - \left(\dfrac{\lambda}{\mu}\right)^{K+1}}$$

y

$$\pi_n = \frac{1 - \dfrac{\lambda}{\mu}}{1 - \left(\dfrac{\lambda}{\mu}\right)^{K+1}} \left(\frac{\lambda}{\mu}\right)^n, \quad n = 0, \ldots, K$$

Puesto que el tamaño de la cola es acotado, no puede crecer indefinidamente, por lo que el sistema alcanza el equilibrio para todos los valores de λ y μ y no hace falta introducir la condición $\dfrac{\lambda}{\mu} < 1$. Si $\lambda < \mu$, cuando $K \to \infty$, se tiene la convergencia deseada al sistema $M/M/1$.

Podemos calcular también el número medio de individuos en el sistema.

- Si $\lambda = \mu$,

$$L = \sum_{n=0}^{K} n\pi_n = \frac{1 + 2 + \ldots + K}{K+1} = \frac{K}{2}$$

- Si $\lambda \neq \mu$,

$$\begin{aligned} L = \sum_{n=0}^{K} n\pi_n &= \frac{1 - \dfrac{\lambda}{\mu}}{1 - \left(\dfrac{\lambda}{\mu}\right)^{K+1}} \sum_{n=0}^{K} n \left(\frac{\lambda}{\mu}\right)^n = \\ &= \frac{1-a}{1-a^{K+1}} \sum_{n=0}^{K} n a^n = \\ &= \frac{a(1 + Ka^{K+1} - (K+1)a^K)}{(1-a^{K+1})(1-a)} \end{aligned}$$

ya que

$$\begin{aligned} \sum_{n=0}^{K} n a^n &= a \sum_{n=1}^{K} n a^{n-1} = a \sum_{n=1}^{K} \frac{da^n}{da} = a \frac{d}{da} \sum_{n=0}^{K} a^n = \\ &= a \frac{d}{da} \left(\frac{1 - a^{K+1}}{1-a} \right) = a \frac{-(1-a)(K+1)a^K + (1-a^{K+1})}{(1-a)^2} \end{aligned}$$

Para calcular W hay que tener cierto cuidado con lo que se entiende por clientes en el sistema. Todo el tráfico que llega al sistema no entra ya que si hay K clientes en el sistema no se admite a los clientes que llegan. Así, consideramos que los clientes son aquellos que entran en el sistema y lo hacen con tasa

$$\lambda_a = \lambda(1 - \pi_K)$$

por lo que

$$W = \frac{L}{\lambda_a} = \frac{L}{\lambda(1 - \pi_K)}$$

Además, podemos calcular L_s, para ello sea N_s la v.a. número de clientes en el servidor. Entonces,

$$L_s = E(N_s) = E(N_s/N=0)P(N=0)+E(N_s/N>0)P(N>0) = 0\times\pi_0+1\times(1-\pi_0) = 1-\pi_0$$

Así,

$$L_q = L - L_s = L - (1-\pi_0)$$

y

$$W_q = \frac{L_q}{\lambda(1-\pi_K)}$$

El uso verdadero del servidor es

$$\rho = \lambda_a W_s = \lambda(1-\pi_K)W_s = \frac{\lambda}{\mu}(1-\pi_K)$$

También pueden derivarse las distribuciones de w y w_q, aunque de forma algo más compleja.

Ejemplo 4: Supongamos que en el Ejemplo 2 de la central de comunicaciones, se desea diseñar el sistema de manera que el tamaño del acumulador sea mínimo y la probabilidad de que el sistema esté completo sea menor que 0.005. Para ese tamaño calcular L, L_q, W y W_q.

La probabilidad de que el sistema esté lleno cuando la capacidad del sistema es K está dada por

$$\pi_K = \frac{1-\dfrac{\lambda}{\mu}}{1-\left(\dfrac{\lambda}{\mu}\right)^{K+1}}\left(\frac{\lambda}{\mu}\right)^K$$

con $\dfrac{\lambda}{\mu} = 0.88$ y $\pi_K < 0.005$. Resolviéndolo por ejemplo con el programa MATHEMATICA obtenemos $K > 25.1426$, es decir, $K = 26$ y el tamaño mínimo del acumulador es $K-1 = 25$. Sustituyendo en las fórmulas tenemos

$$\begin{aligned} L &= 6.449 \text{ mensajes} \\ L_q &= 5.573 \text{ mensajes} \\ W &= 1.62 \text{ segundos} \\ W_q &= 1.40 \text{ segundos} \end{aligned}$$

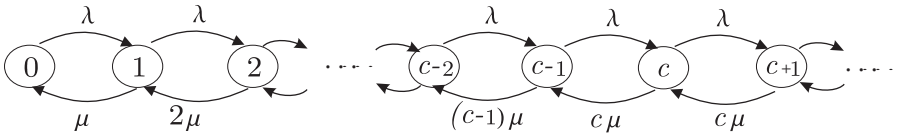
Si comparamos los resultados con los obtenidos para el sistema $M/M/1$, vemos que son mejores los del sistema $M/M/1/26$, pero a costa de no admitir el $100\pi_{26} = 0.4464\%$ de los mensajes. El 0.4464% de los mensajes es rechazado y deben de ser enviados en otro momento. □

4 Sistema de Colas Exponencial con c Servidores: $M/M/c$

Tenemos c servidores idénticos. Este sistema es un proceso de nacimiento y muerte con tasas

$$\begin{aligned} \lambda_n &= \lambda, \quad n \geq 0 \\ \mu_n &= \begin{cases} n\mu, & n = 1, 2, \ldots, c-1 \\ c\mu, & n \geq c \end{cases} \end{aligned}$$

El diagrama de tasas de transición es



y las ecuaciones de equilibrio son

$$\begin{aligned} \lambda\pi_0 &= \mu\pi_1 \\ (\lambda + n\mu)\pi_n &= \lambda\pi_{n-1} + (n+1)\mu\pi_{n+1}, \quad 1 \leq n \leq c-1 \\ (\lambda + c\mu)\pi_n &= \lambda\pi_{n-1} + c\mu\pi_{n+1}, \qquad n \geq c \\ \sum_{n=0}^{\infty} \pi_n &= 1 \\ \pi_n &\geq 0, \quad \forall n \geq 0 \end{aligned}$$

cuya solución es

$$\pi_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}\pi_0 = \begin{cases} \left(\dfrac{\lambda}{\mu}\right)^n \dfrac{1}{n!}\pi_0, & n = 0, 1, \ldots, c-1 \\ \left(\dfrac{\lambda}{c\mu}\right)^n \dfrac{c^c}{c!}\pi_0, & n \geq c \end{cases}$$

$$\pi_0 = \left(\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \sum_{n=c}^{\infty} \left(\frac{\lambda}{c\mu}\right)^n \frac{c^c}{c!}\right)^{-1} = \left(\sum_{n=0}^{c-1} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} + \left(\frac{\lambda}{\mu}\right)^c \frac{1}{c!} \frac{1}{1 - \dfrac{\lambda}{c\mu}}\right)^{-1}$$

y, para la estabilidad, tiene que ser $\rho = \dfrac{\lambda}{c\mu} < 1$.

Podemos entonces deducir algunos de los parámetros clave del sistema $M/M/c$. Primeramente, calculemos la probabilidad de que un cliente tenga que hacer cola al llegar. Esta probabilidad coincide con la probabilidad de que el cliente encuentre c o más clientes en el sistema. La fórmula que da esta probabilidad se denota por $C(c,a)$ con $a = \dfrac{\lambda}{\mu}$ y se denomina *fórmula C de Erlang o de retraso de Erlang*, ya que es importante y se utiliza para calcular otras medidas importantes del sistema.

$$\begin{aligned} C(c,a) &= \sum_{n=c}^{\infty} \pi_n = 1 - \sum_{n=0}^{c-1} \pi_n = 1 - \pi_0 \sum_{n=0}^{c-1} \frac{a^n}{n!} = \\ &= 1 - \frac{\displaystyle\sum_{n=0}^{c-1} \frac{a^n}{n!}}{\displaystyle\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)}} = \\ &= \frac{\dfrac{a^c}{c!}}{(1-\rho)\displaystyle\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!}} = \frac{\dfrac{a^c}{c!}}{1-\rho}\pi_0 \end{aligned}$$

Hay algoritmos para calcular $C(c,a)$, que de paso ayudan a calcular otras medidas de interés. Por ejemplo, la probabilidad de que el sistema esté ocioso

$$\pi_0 = \frac{C(c,a)c!(1-\rho)}{a^c},$$

la probabilidad de que el cliente no tenga que hacer cola

$$W_q(0) = P(w_q = 0) = 1 - C(c,a),$$

el número medio de clientes en cola

$$\begin{aligned} L_q &= \sum_{n=c}^{\infty}(n-c)\pi_n = \sum_{n=0}^{\infty} n\pi_{n+c} = \sum_{n=0}^{\infty} n\rho^{n+c}\frac{c^c}{c!}\pi_0 = \\ &= \pi_0 \frac{a^c}{c!}\sum_{n=0}^{\infty} n\rho^n = \pi_0 \frac{a^c}{c!}\frac{\rho}{(1-\rho)^2} = \\ &= \frac{\rho C(c,a)}{1-\rho} \end{aligned}$$

Además, por la fórmula de Little obtenemos

$$W_q = \frac{L_q}{\lambda} = \frac{C(c,a)}{c\mu(1-\rho)}$$

con lo que

$$W = W_s + W_q = \frac{1}{\mu}\left(1 + \frac{C(c,a)}{c(1-\rho)}\right)$$

y $L = \lambda W$.

Se pueden calcular también $W_q(t)$ y $W(t)$, con ideas parecidas a las ya empleadas en el sistema $M/M/1$. Por ejemplo, para la función de distribución de w_q obtenemos lo siguiente.

Si $t > 0$ un cliente tiene que esperar en cola si hay c o más clientes en el sistema. Puesto que, en ese caso, todos los servidores están ocupados, el tiempo de complección de servicios es exponencial de parámetro $c\mu$ (mínimo de c exponenciales independientes de parámetro μ). Si al llegar el cliente encuentra n clientes en el sistema, hay c clientes recibiendo servicio y $n-c$ esperando en cola, con lo que el cliente debe esperar la complección de $n-c+1$ servicios. Así, el tiempo de espera en la cola es la suma de $n-c+1$ v.a's exponenciales independientes de parámetro $c\mu$, que sabemos se distribuye como una Gamma de parámetros $a = c\mu$ y $p = n-c+1$.

$$\begin{aligned}
W_q(t) &= W_q(0) + P(0 < w_q \leq t) = W_q(0) + \sum_{n=c}^{\infty} P(w_q \leq t/N = n)\pi_n = \\
&= W_q(0) + \sum_{n=c}^{\infty}\left(\int_0^t \frac{(c\mu)^{n-c+1}}{(n-c)!}e^{-c\mu x}x^{n-c}dx\right)\frac{a^n}{c!c^{n-c}}\pi_0 = \\
&= W_q(0) + \sum_{n=c}^{\infty}\frac{\pi_0 a^n}{c!c^{n-c}}\int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!}e^{-c\mu x}dx = \\
&= 1 - C(c,a)e^{-\mu t(c-a)}
\end{aligned}$$

Análogamente, se pueden encontrar fórmulas para $W(t)$.

Ejemplo 5: Una compañia de aviación introduce un nuevo sistema de reservas. Cada agente tiene un terminal y puede atender al cliente típico en 5 minutos, distribuyéndose los tiempos exponencialmente. Las llamadas son completamente aleatorias y el sistema tiene un gran acumulador para las llamadas no atendidas. En el pico de actividad, se esperan 36 llamadas por hora. Se introducen tres criterios de diseño:

1. La probabilidad de que una llamada encuentre todos los agentes acupados no debe exceder 0.1.

2. El tiempo medio de espera, para aquéllos que deben esperar, no debe superar 1 minuto.

3. Menos del 5 % de los clientes deben esperar más de un minuto para conseguir un agente.

¿Cuántos agentes deben incluirse?

La tasa de llegadas (en el pico de actividad) es

$$\lambda = 36 \text{ llamadas/hora} = 0.6 \text{ llamadas/minuto}$$

La tasa de servicio es

$$\mu = \frac{1}{5} \text{ llamadas/minuto} = 0.2$$

La intensidad de tráfico es $a = \dfrac{\lambda}{\mu} = 3$ erlangs.

Para que el sistema sea estable se debe verificar

$$\rho = \frac{\lambda}{c\mu} < 1$$

con lo que c debe ser mayor que 3 y se deben incluir al menos 4 agentes.

Las condiciones de diseño son:

1. $\displaystyle\sum_{n=c}^{\infty} \pi_n = C(c,3) \leq 0.1$
2. $E(w_q/w_q > 0) \leq 1$
3. $P(w_q > 1) \leq 0.05$

y queremos minimizar c.

Utilizando la gráfica de $C(c,a)$ (ver Figura 4 en el apéndice) obtenemos que para satisfacer la primera condición debe ser $c \geq 6$.

Para la segunda condición tenemos que

$$E(w_q) = E(w_q/w_q = 0)P(w_q = 0) + E(w_q/w_q > 0)P(w_q > 0) = E(w_q/w_q > 0)P(w_q > 0)$$

con lo que

$$E(w_q/w_q > 0) = \frac{E(w_q)}{P(w_q > 0)} = \frac{W_q}{C(c,a)} = \frac{1}{c\mu(1-\rho)}$$

que debe ser menor o igual que 1, de donde obtenenos $c \geq 8$.

Por otra parte, sabemos que

$$P(w_q > t) = C(c,a)e^{-\mu t(c-a)}$$

y si calculamos la $P(w_q > 1)$ con $c = 8$ observamos que se verifica la tercera condición. Efectivamente,

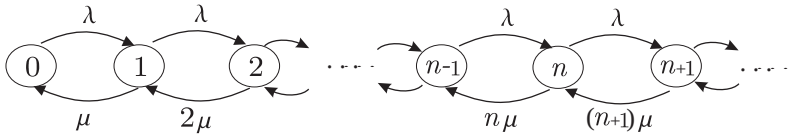
$$P(w_q > 1) = C(8,3)e^{-0.2(8-3)} = 0.00476 < 0.05$$

Así, $c = 8$, con lo que vemos que $\rho = 3/8$ y los agentes están ocupados 3/8 del tiempo. □

5 Sistema de Colas Exponencial con Infinitos Servidores: $M/M/\infty$

Consideramos un sistema de colas con llegadas y servicios exponenciales en el que se proporciona servicio al cliente en cuanto llega. Puede utilizarse, por ejemplo, para estimar el número de líneas en uso en una gran red de comunicaciones o como estimación en sistemas $M/M/c$ (y $M/M/c/c$ que veremos a continuación) para grandes valores de c. Además, su distribución estacionaria coincide con la del sistema $M/G/\infty$, por lo que puede utilizarse en varios contextos.

El diagrama de tasas de transición es



En este caso las ecuaciones de equilibrio son

$$\begin{aligned} \lambda\pi_0 &= \mu\pi_1 \\ (\lambda + n\mu)\pi_n &= \lambda\pi_{n-1} + (n+1)\mu\pi_{n+1}, \quad n \geq 1 \\ \sum_{n=0}^{\infty} \pi_n &= 1 \\ \pi_n &\geq 0, \quad \forall n \geq 0 \end{aligned}$$

cuya solución es

$$\pi_n = \frac{\lambda_0 ... \lambda_{n-1}}{\mu_1 ... \mu_n}\pi_0 = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\pi_0, \quad n \geq 0$$

$$\pi_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}} = e^{-\frac{\lambda}{\mu}}$$

Por tanto,

$$\pi_n = e^{-\frac{\lambda}{\mu}} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}, \quad n \geq 0$$

y N sigue una distribución de Poisson de parámetro $\frac{\lambda}{\mu}$.

El número medio de servidores ocupados es

$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\lambda}{\mu}$$

Como tenemos tantos servidores como clientes en el sistema, $L_q = 0$ y, por supuesto, $W_q = 0$. Así, el tiempo medio en el sistema es el tiempo medio de servicio, es decir, $W = \frac{1}{\mu}$, y la distribución del tiempo en el sistema, $W(t)$, es igual a la distribución del tiempo de servicio, es decir, exponencial con media $\frac{1}{\mu}$.

Ejemplo 6: Llegan llamadas aleatorias al intercambiador de un sistema telefónico con tasa 140 llamadas/hora. Si hay un número muy grande de líneas para atender las llamadas, que duran en media 3 minutos, ¿cuál es el número medio de líneas en uso? Estimar los percentiles 90 y 95 del número de líneas en uso.

Utilizamos un modelo $M/M/\infty$, donde

$$\lambda = 140 \text{ llamadas/hora} = 73 \text{ llamadas/minuto}$$

$$\frac{1}{\mu} = 3 \text{ llamadas/minuto}$$

Así, el número medio de líneas en uso es

$$L = \frac{\lambda}{\mu} = 7$$

El número de líneas en uso N sigue una distribución de Poisson de parámetro 7.

Para calcular los percentiles, o bien usamos tablas o la aproximación normal. Como el parámetro de la distribución es mayor que 5, podemos aproximar $N \sim \mathcal{P}(7)$ con $Y \sim \mathcal{N}(7, \sqrt{7})$.

- Percentil 90.

  Tenemos que calcular $a = \pi_N(90)$ tal que $P(N \leq a) = 0.9$. Así, si Z es una v.a. distribuida según una $\mathcal{N}(0, 1)$,

  $$P(N \leq a) \simeq P(Y \leq a + 0.5) = P(Z \leq \frac{a + 0.5 - 7}{\sqrt{7}}) = 0.9$$

  De donde $a = 6.5 + 1.28\sqrt{7} = 9.88$.

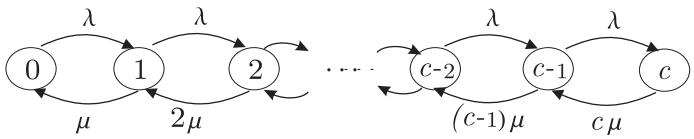
- Percentil 95.

  Análogamente, se obtiene $\pi_N(95) = 6.5 + 1.65\sqrt{7} = 10.86$. □

6 Sistema de Colas Exponencial con varios Servidores y Pérdidas: $M/M/c/c$

El sistema $M/M/c/c$ se denomina de pérdidas porque los clientes que llegan cuando todos los servidores están ocupados han de abandonar el sistema (son pérdidas para el sistema).

El diagrama de transición de estados es



Las ecuaciones de equilibrio son

$$\begin{aligned}
\lambda\pi_0 &= \mu\pi_1 \\
(\lambda + n\mu)\pi_n &= \lambda\pi_{n-1} + (n+1)\mu\pi_{n+1}, \quad 1 \leq n \leq c-1 \\
c\mu\pi_c &= \lambda\pi_{c-1} \\
\sum_{n=0}^{c} \pi_n &= 1 \\
\pi_n &\geq 0, \quad \forall n = 0, \ldots, c
\end{aligned}$$

cuya solución es

$$\pi_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}\pi_0 = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\pi_0, \quad n = 0,\, 1, \ldots,\, c$$

$$\pi_0 = \frac{1}{\displaystyle\sum_{n=0}^{c} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}}$$

Esta distribución se denomina *distribución de Poisson truncada*.

Se perderán clientes si todos los servidores están ocupados, con lo que la proporción de clientes que se pierden es

$$\pi_c = B(c,a) = \frac{\frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c}{1+\frac{\lambda}{\mu}+\cdots+\frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c} = \frac{\frac{a^c}{c!}}{1+a+\frac{a^2}{2!}+\cdots+\frac{a^c}{c!}}$$

A esta fórmula designada por $B(c,a)$ se le llama fórmula B de Erlang.

La tasa de entrada en el sistema es

$$\lambda_a = \lambda(1-B(c,a))$$

Como no se permite esperar a ningún cliente, se tiene $W_q = L_q = 0$. Además,

$$L = \sum_{n=0}^{c} n\pi_n = \pi_0 \sum_{n=1}^{c} n\frac{a^n}{n!} = \pi_0 \sum_{n=1}^{c} \frac{a^n}{(n-1)!} = a\pi_0 \sum_{n=0}^{c-1} \frac{a^n}{n!} = a(1-B(c,a))$$

Por la fórmula de Little

$$W = \frac{L}{\lambda_a} = \frac{a(1-B(c,a))}{\lambda(1-B(c,a))} = \frac{\frac{\lambda}{\mu}}{\lambda} = \frac{1}{\mu} = W_s$$

Además, w tiene la distribución del tiempo de servicio y

$$W(t) = 1 - e^{-\mu t}$$

Todas las fórmulas salvo la de $W(t)$ (que procede de $W(t) = W_s(t) = P(s \leq t)$) son válidas para los sistemas $M/G/c/c$. Es decir, sólo el valor medio del tiempo de servicio es importante. Tales sistemas de colas se denominan "sistemas robustos".

Ejemplo 7: Una compañía decide instalar un sistema de comunicación interno entre sus oficinas de Barcelona y Madrid. Una llamada recibe una señal de ocupado si es realizada cuando todas las líneas están ocupadas. Las llamadas ocurren aleatoriamente a una tasa de 105 llamadas/hora y tardan en promedio 4 minutos en ser servidas.

1. Se deben instalar suficientes líneas para asegurar que la probabilidad de obtener una señal de ocupado no exceda de 0.005. ¿Cuántas líneas son necesarias?

2. ¿Cuántas líneas se requieren si la probabilidad de obtener señal de ocupado es de 0.01?

3. Estudiar el comportamiento de este sistema con 10 líneas.

$$\lambda = 105 \text{ llamadas/hora} = 1.75 \text{ llamadas/minuto}$$

$$\frac{1}{\mu} = 4 \text{ minutos/llamada}$$

La intensidad de tráfico es $a = \dfrac{\lambda}{\mu} = 1.75 \times 4 = 7$.

1. Hay que determinar c de forma que

   $$\pi_c = B(c, 7) \leq 0.005$$

   Utilizando la tabla de la fórmula B de Erlang (ver Figura 5 en el apéndice) se encuentra que $c \geq 15$.

2. Para que $\pi_c = B(c, 7) \leq 0.01$ debe ser $c \geq 14$.

3. Si se introducen 10 líneas, la probabilidad de encontrar todas las líneas ocupadas es $B(10, 7) = 0.07874$. □

De una forma similar se pueden deducir las fórmulas para el sistema de colas $M/M/c/K$.

7 Sistema de Colas Exponencial con un Procesador y Población Finita: $M/M/1/K/K$

Este sistema también se conoce como sistema de reparación de máquinas con un reparador y como modelo de colas cíclico.

Consideramos un sistema que soporta K terminales activos. Suponemos que el tiempo que necesita cada usuario para solicitar trabajo al procesador, *tiempo para pensar*, se distribuye exponencialmente con media $\dfrac{1}{\lambda}$ y el procesador procesa los trabajos según una distribución exponencial con media $\dfrac{1}{\mu}$. El procesador sirve a un usuario a lo sumo. Un terminal que espera resultados se dice que está en espera. En términos de máquinas-reparador, $\dfrac{1}{\lambda}$ es el tiempo medio de operatividad entre roturas o el tiempo medio hasta fallo y $\dfrac{1}{\mu}$ el tiempo medio de reparación.

Este sistema está representado en la Figura 2.

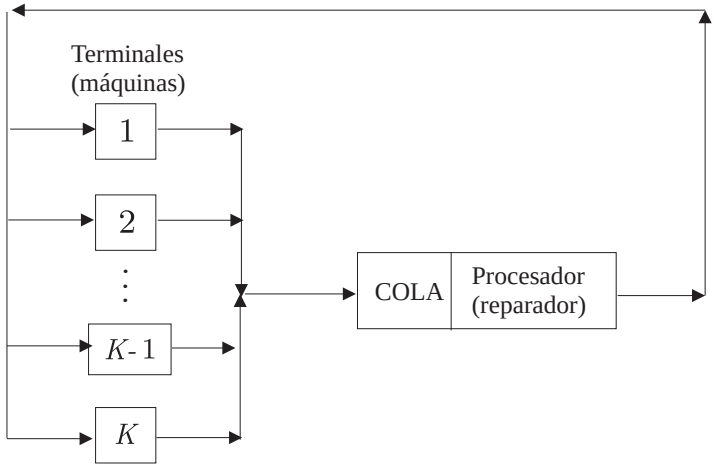
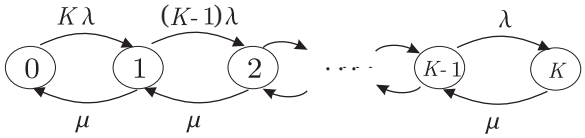


Figura 2: Modelo de Colas Cíclico

Las máquinas operativas están fuera del sistema de colas y sólo entran en el sistema cuando se rompen y necesitan, por tanto, ser reparadas. Cuando hay n máquinas estropeadas, $K-n$ están operativas y el tiempo hasta que la siguiente máquina se estropea es el mínimo de $K-n$ exponenciales independientes de parámetro μ, es decir, una exponencial de parámetro $(K-n)\lambda$. Por tanto, la tasa media de llegadas al sistema es $\lambda_n = (K-n)\lambda$, $\forall n = 0, 1, \ldots, K-1$ y el diagrama de transición de estados es el siguiente:



El sistema de colas siempre alcanza el estado estacionario porque no puede haber más de K clientes en el sistema.

Las ecuaciones de equilibrio son

$$\begin{aligned}
K\lambda\pi_0 &= \mu\pi_1 \\
((K-n)\lambda+\mu)\pi_n &= (K-(n-1))\lambda\pi_{n-1} + \mu\pi_{n+1}, \quad 1 \le n \le K-1 \\
\mu\pi_K &= \lambda\pi_{K-1} \\
\sum_{n=0}^{K} \pi_n &= 1 \\
\pi_n &\ge 0, \quad \forall n = 0, \ldots, K
\end{aligned}$$

$$\pi_n = \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}\pi_0 = \frac{K\lambda(K-1)\lambda \cdots (K-n+1)\lambda}{\mu \cdots \mu}\pi_0 = \left(\frac{\lambda}{\mu}\right)^n \frac{K!}{(K-n)!}\pi_0, \quad n = 0,\, 1, \ldots,\, K$$

con lo que

$$\pi_0 = \frac{1}{\displaystyle\sum_{n=0}^{K}\left(\frac{\lambda}{\mu}\right)^n \frac{K!}{(K-n)!}} = \frac{1}{\displaystyle\sum_{n=0}^{K}\frac{K!}{(K-n)!}a^n}$$

$$\pi_n = \frac{K!}{(K-n)!}a^n\pi_0$$

Dividiendo y multiplicando por $\dfrac{1}{a^K}$ y llamando $r = \dfrac{1}{a}$, obtenemos

$$\pi_n = \frac{\dfrac{r^{K-n}}{(K-n)!}}{\displaystyle\sum_{i=0}^{K}\frac{r^{K-i}}{(K-i)!}}, \quad n = 0,\, 1, \ldots,\, K$$

llamando $P_{K-n} = \pi_n$, $n = 0,\, 1, \ldots,\, K$, tenemos

$$P_n = \frac{\dfrac{r^n}{n!}}{\displaystyle\sum_{i=0}^{K}\frac{r^i}{i!}}, \quad n = 0,\, 1, \ldots,\, K$$

que son las probabilidades estacionarias del sistema $M/M/K/K$, con lo que podemos utilizar $B(K, r)$ para los cálculos. Por ejemplo, la probabilidad de que el procesador (reparador) esté ocupado, i.e. el uso del procesador, es

$$\rho = 1 - \pi_0 = 1 - P_K = 1 - B(K, r)$$

Puesto que $\rho = \lambda_a W_s$ se tiene que la tasa media de llegada al sistema es

$$\lambda_a = \frac{\rho}{W_s}$$

Para calcular las medidas de comportamiento, razonamos como sigue. Para cada una de las K máquinas o terminales un ciclo completo consiste en

- tiempo para pensar (periodo operativo),
- tiempo en cola y
- tiempo de servicio.

Así, la tasa media de entradas al sistema (tasa a la que se rompen las máquinas) es

$$\lambda_a = \frac{K}{\frac{1}{\lambda} + W_q + W_s} = \frac{K}{\frac{1}{\lambda} + W}$$

de manera que

$$W = \frac{K}{\lambda_a} - \frac{1}{\lambda}$$

Además,

$$\begin{aligned} W_q &= W - W_s \\ L_q &= \lambda_a W_q \\ L &= \lambda_a W \end{aligned}$$

Claramente, es esencial calcular $B(K, r)$ para lo que disponemos de tablas y programas.

Ejemplo 8: Un técnico atiende 4 máquinas. Para cada máquina el tiempo entre fallos sigue una distribución exponencial con un valor medio de 10 horas. El tiempo de reparación parece seguir la misma distribución y tiene un valor medio de 2 horas. Cuando una máquina está esperando o siendo reparada el tiempo perdido tiene un valor de 12 euros/hora y el tiempo de servicio del técnico cuesta 30 euros/día. Suponer jornadas de 10 horas.

1. Calcular el número esperado de máquinas funcionando.
2. ¿Sería deseable tener dos mecánicos para que cada uno atienda sólo a dos máquinas?

Tenemos un sistema $M/M/1/4/4$ con

$$\begin{aligned} \frac{1}{\lambda} &= \text{10 horas, es decir, } \lambda = \frac{1}{10} \text{ fallos/hora} \\ \frac{1}{\mu} &= \text{2 horas, es decir, } \mu = \frac{1}{2} \text{ reparaciones/hora} \end{aligned}$$

1. Como π_n es la probabilidad de que haya n clientes en el sistema, i.e., probabilidad de que haya n máquinas estropeadas, el número esperado de máquinas funcionando es

$$4 - \sum_{n=1}^{4} n\pi_n = 4 - L$$

Tenemos $r = \dfrac{\lambda}{\mu} = 5$, con lo que

$$\rho = 1 - B(4, 5) = 0.602$$

$$\lambda_a = \mu\rho = 0.301$$
$$W = \frac{4}{0.301} - 10 = 3.289$$
$$L = \lambda_a W = 0.99$$

y por tanto el número medio de máquinas funcionando es $4 - 0.99 = 3.01$.

Alternativamente, podríamos haber calculado este valor directamente a partir de las probabilidades estacionarias cuyo valor es:

$$\pi_0 = 0.398,\ \ \pi_1 = 0.32,\ \ \pi_2 = 0.191,\ \ \pi_3 = 0.076,\ \ \pi_4 = 0.015$$

2. Para contestar a este apartado, calculamos primero el coste asociado al sistema considerado hasta ahora. El coste total es la suma de un coste fijo (coste del mecánico) y de un coste variable (pérdida por máquina parada). El coste fijo es de 30 euros al día, mientras que el variable es $12 \times 10 \times L = 118.8$ euros al día. Así, el coste total por día es 148.8 euros.

   El modelo propuesto consta de dos sistemas $M/M/1/2/2$ independientes. Para cada uno de ellos tenemos que el número medio de máquinas estropeadas es $L = 0.378$ y el coste por día es 75.36 euros, con lo que el coste total asociado a los dos sistemas es 150.72 euros por día.

   Por tanto, nos interesa más el primer sistema al ser más económico. □

Cuando se estudian sistemas interactivos mediante el sistema $M/M/1/K/K$, se suele considerar el tiempo de respuesta normalizado.

$$\begin{aligned}\frac{W}{W_s} &= \frac{1}{W_s}\left(\frac{K}{\lambda_a} - \frac{1}{\lambda}\right) = \frac{1}{W_s}\left(\frac{W_s K}{\rho} - \frac{1}{\lambda}\right) = \\ &= \frac{K}{\rho} - r = \frac{K}{1 - B(K,r)} - r\end{aligned}$$

- Cuando $K = 1$, como sólo hay un terminal activo no hay cola y $W = W_s$.
- Cuando $K \to \infty$, $\rho \to 1$ (cuando K crece mucho, se espera que el sistema esté ocupado la mayoría del tiempo, por lo que $\rho \approx 1$), con lo que

$$\frac{W}{W_s} \to K - r$$

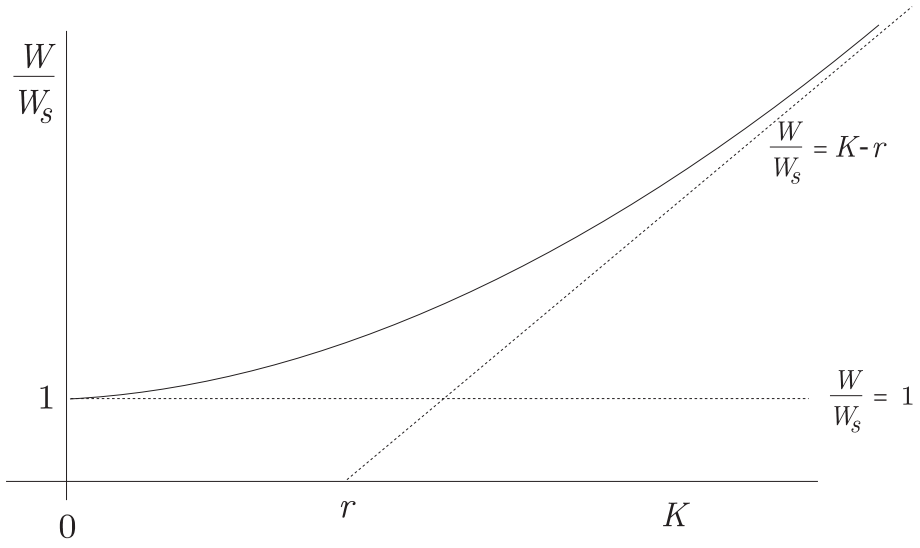


Figura 3: Gráfica de W/W_s en función de K

La gráfica de $\dfrac{W}{W_s}$ como función de K está representada en la Figura 3.

Ambas asíntotas se cortan en el punto $(K^* = 1+r, 1)$. K^* se denomina *número de saturación* y tiene la siguiente interpretación: si cada usuario de los terminales utiliza de tiempo para pensar, exactamente, $\dfrac{1}{\lambda}$ unidades de tiempo y emplea, exactamente, $\dfrac{1}{\mu}$ unidades de tiempo de servicio por interacción, entonces con sincronización perfecta[1], K^* es el número máximo de terminales que el sistema soporta sin que haya interferencias mutuas.

Así, si $K = 1$ no hay interferencia mutua, cuando K es pequeño apenas hay interferencia mutua. Si K supera K^*, estamos seguros de que habrá interferencias. La fórmula

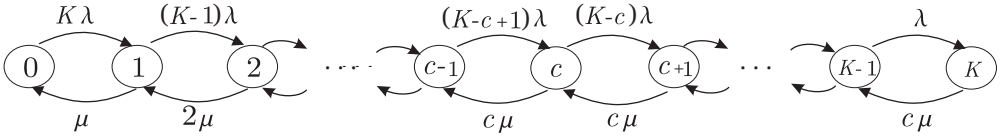
$$W = KW_s - \frac{1}{\lambda}$$

para K grande ($\rho = 1$) muestra interferencia completa, es decir, cuando el número de usuarios es muy grande, cada usuario retrasa a cada uno de los otros usuarios por un tiempo medio de servicio.

8 Sistema Exponencial con c Procesadores y Población Finita: $M/M/c//K$ o $M/M/c/K/K$

Se define como en el modelo de colas anterior, salvo que ahora tenemos c servidores. Tenemos un proceso de nacimiento y muerte con diagrama de tasas de transición:

[1]esto es, entra un terminal y cuando acaba, inmediatamente entra el siguiente. Esto es lo que hace que K^* no coincide con el K' que hace $\rho = 1$.



La solución estacionaria es

$$\pi_n = \frac{\lambda_0 \ldots \lambda_{n-1}}{\mu_1 \ldots \mu_n}\pi_0 = \begin{cases} \binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n \pi_0, & n = 0,\, 1, \ldots,\, c \\ \\ \frac{n!}{c^{n-c}c!}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n \pi_0, & n = c+1, \ldots,\, K \end{cases}$$

$$\pi_0 = \left(\sum_{n=0}^{c}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c+1}^{K}\frac{n!}{c^{n-c}c!}\binom{K}{n}\left(\frac{\lambda}{\mu}\right)^n\right)^{-1}$$

Además,

$$L_q = \sum_{n=c+1}^{K}(n-c)\pi_n$$

$$\lambda_a = \frac{K}{\frac{1}{\lambda} + W}$$

$$W_q = \frac{L_q}{\lambda_a} = \frac{L_q}{K}\left(\frac{1}{\lambda} + W_q + W_s\right) = \frac{L_q}{K - L_q}\left(\frac{1}{\lambda} + W_s\right)$$

y

$$\lambda_a = \frac{L_q}{W_q}$$

$$W = W_q + W_s$$

$$L = \lambda_a W$$

Análogamente, pueden calcularse otras características del comportamiento estacionario de la cola.

Ejemplo 9: Consideremos el Ejemplo 8 analizado en la sección anterior, permitiendo ahora que los dos mecánicos considerados en el apartado 2, reparen cualquier máquina. Determinar el coste de este sistema.

Tenemos por tanto un sistema $M/M/2/4/4$, con

$$\lambda = \frac{1}{10},\ \ \mu = \frac{1}{2},\ \ \frac{\lambda}{\mu} = 15$$

En este caso, las probabilidades estacionarias son

$$\pi_0 = 0.47,\ \ \pi_1 = 0.38,\ \ \pi_2 = 0.12,\ \ \pi_3 = 0.023,\ \ \pi_4 = 0.0023$$

Así, el número medio de máquinas estropeadas es $L = \pi_1 + 2\pi_2 + 3\pi_3 + 4\pi_4 = 0.6982$, con lo que el coste total del sistema es $60 + 12 \times 10 \times 0.6982 = 113.78$ euros por día y, por tanto, si lo comparamos con los sistemas analizados en el Ejemplo 8, este sistema proporciona la solución más barata. □

Bibliografía

La preparación de este tema se ha apoyado en los siguientes textos:

Allen, A.O. (1990) *Probability, Statistics, and Queueing Theory with Computer Science Applications.* Academic Press.

Gross, D., Harris, C.M. (1985) *Fundamentals of Queueing Theory.* Wiley.

Kleinrock, L. (1975) *Queuing Systems, Volume I: Theory.* Wiley.

Kleinrock, L. (1976) *Queuing Systems, Volume II: Computer Applications.* Wiley.

Leung, C.H.C. (1988) *Quantitative Analysis of Computer Systems.* Wiley.

Ross, S. (2001) *Introduction to Probability Models.* Academic Press.

Apéndice

Gráfico de la fórmula C de Erlang

Figura 4: Probabilidad $C(c, u)$ de que todos los c servidores estén ocupados en un sistema de colas $M/M/c$ frente a la intensidad de tráfico $u = \lambda W_s$

Gráfico de la fórmula B de Erlang

Figura 5: Probabilidad $B(c, u)$ de que todos los servidores estén ocupados en un sistema de pérdidas $M/M/c/c$