

Líneas de Espera

Dra. Valeria Soto Mendoza

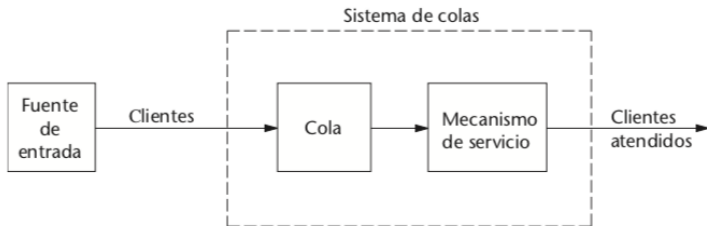
Centro de Investigación en Matemáticas Aplicadas
Universidad Autónoma de Coahuila

4 de marzo de 2020



- 1 Conceptos básicos
- 2 Notación
- 3 Proceso de nacimiento y muerte
- 4 Análisis del proceso de nacimiento y muerte
- 5 Modelos de colas con distribuciones exponenciales
- 6 Modelos de colas con distribuciones no exponenciales
- 7 Modelos sin entradas de Poisson
- 8 Modelos de colas con disciplina de prioridades
- 9 Redes de colas
- 10 Referencias

Los **clientes** que requieren un servicio se generan en el tiempo en una **fuente de entrada**. Luego, entran al **sistema** y se unen a una **cola**. En determinado momento se selecciona un miembro de la cola para proporcionarle el servicio mediante alguna regla conocida como **disciplina de la cola**. Se lleva a cabo el servicio que el cliente requiere mediante un **mecanismo de servicio**, y después el cliente sale del sistema de colas.



- Una característica de la fuente de entrada es su **tamaño**. El tamaño es el **número total de clientes** que pueden requerir servicio en determinado momento (número total de clientes potenciales).
- Esta población a partir de la cual surgen las unidades que llegan se conoce como **población de entrada**.
- Puede suponerse que el tamaño es **infinito** o **finito** (ilimitada o limitada).
- Especificar el patrón estadístico mediante el cual se generan los clientes en el tiempo:
 - el número de clientes que llegan hasta un momento específico tiene una distribución de Poisson (*proceso Poisson*).
 - equivalentemente, la distribución de probabilidad del tiempo que transcurre entre dos llegadas consecutivas es *exponencial*.

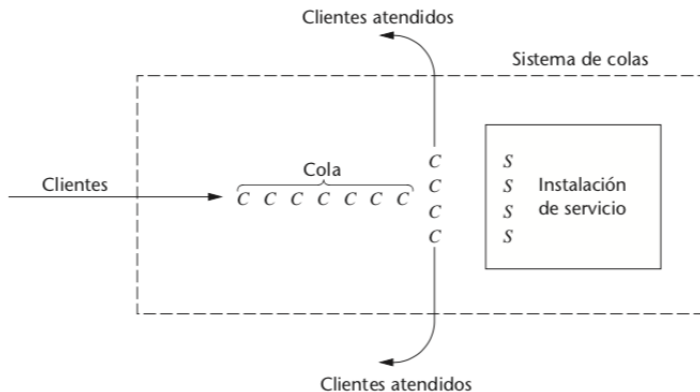
- El tiempo que transcurre entre dos llegadas consecutivas se define como **tiempo entre llegadas**.
- Debe especificarse cualquier otro supuesto no usual sobre el comportamiento de los clientes:
 - cuando se pierde un cliente porque desiste o
 - se rehúsa a entrar al sistema porque la cola es demasiado larga.

- La cola es donde los clientes esperan *antes* de recibir el servicio.
- Una cola se caracteriza por el número máximo permisible de clientes que puede admitir.
- Las colas pueden ser *finitas* o *infinitas*, según si dicho número es finito o infinito.
- El supuesto de una **cola infinita** es el estándar de la mayoría de los modelos.
- En los sistemas de colas en los que la cota superior es tan pequeña que se llega a ella con cierta frecuencia, es necesario suponer una **cola finita**.

- Se refiere al orden en el que sus miembros se seleccionan para recibir el servicio:
 - primero en entrar, primero en salir
 - aleatoria
 - de acuerdo con algún procedimiento de prioridad
 - otro
- En los modelos de colas se supone como normal a la disciplina de **primero en entrar, primero en salir**, a menos que se establezca de otra manera.

- Consiste en una o más estaciones de servicio, cada una de ellas con uno o más canales de servicio paralelos, llamados **servidores**.
- Especificar el arreglo de las estaciones y el número de servidores en cada estación:
 - canales de servicio en serie
 - canales paralelos
- Los modelos más elementales suponen una estación (con un servidor o con un número finito de servidores).
- El tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una estación se llama **tiempo de servicio** (o duración del servicio).

- Especificar la distribución de probabilidad de los tiempos de servicio de cada servidor (y tal vez de los distintos tipos de clientes)
- Es común suponer la misma distribución para todos los servidores.
- La distribución del tiempo de servicio que más se usa en la práctica es la **distribución exponencial**.
- Otras distribuciones de tiempos de servicio importantes:
 - distribución degenerada (tiempos de servicio constantes)
 - la distribución Erlang (gamma)



$$(a/b/c) : (d/e/f)$$

donde:

a = Distribución de tiempos entre llegadas

b = Distribución de tiempos entre salidas (tiempo de servicio)

c = Número de servidores

d = Disciplina de la cola

e = Número máximo permitido en el sistema (finito o infinito)

f = Tamaño de la fuente solicitante (finita o infinita)

$$(a/b/c) : (d/e/f)$$

La notación estándar para representar las distribuciones de las llegadas y salidas (símbolos a y b) es:

M = Distribución markoviana (o de Poisson) de llegadas y salidas
(equivalentemente, distribución exponencial del tiempo entre llegadas y de servicio)

D = Tiempo constante (determinístico)

E_k = Distribución Erlang o gama del tiempo
(equivalentemente, la suma de distribuciones exponenciales independientes)

GI = Distribución general (genérica) del tiempo de llegadas

G = Distribución general (genérica) del tiempo de servicio

$$(a/b/c) : (d/e/f)$$

La notación estándar para la disciplina en colas (símbolo d) incluye:

FCFS = Primero en llegar, primero en ser servido

(first in, first served)

LCFS = Último en llegar, primero en ser servido

(last in, first served)

SIRO = Servicio en orden aleatorio

(service in random order)

GD = Disciplina general (cualquier tipo de disciplina)

(general discipline)

Estado del sistema = número de clientes en el sistema.

Longitud de la cola = número de clientes que esperan servicio.

= estado del sistema – número de clientes a quienes se les da el servicio.

$N(t)$ = número de clientes en el sistema de colas en el tiempo t ($t \geq 0$).

$P_n(t)$ = probabilidad de que exactamente n clientes estén en el sistema en el tiempo t ,
dado el número en el tiempo 0.

s = número de servidores (canales de servicio en paralelo) en el sistema de colas.


λ_n = tasa media de llegadas (número esperado de llegadas por unidad de tiempo) de nuevos
clientes cuando hay n clientes en el sistema.

μ_n = tasa media de servicio en todo el sistema (número esperado de clientes que completan su
servicio por unidad de tiempo) cuando hay n clientes en el sistema.

λ = cuando λ_n es constante para toda n .

μ = cuando la tasa media de servicio por servidor ocupado es constante para toda $n \geq 1$.

$\rho = \frac{\lambda}{(s\mu)}$ es el factor de utilización de la instalación de servicio

(fracción esperada de tiempo que los servidores individuales están ocupados). 

P_n = probabilidad de que haya exactamente n clientes en el sistema.

$L_s(L)$ = número esperado de clientes en el sistema = $\sum_{n=0}^{\infty} nP_n$

L_q = longitud esperada de la cola (excluye los clientes que están en servicio) = $\sum_{n=s}^{\infty} (n - s)P_n$

$W_s(W)$ = tiempo de espera en el sistema (incluye tiempo de servicio) para cada cliente.

$$W_s = E(W_s)$$

W_q = tiempo de espera en la cola (excluye tiempo de servicio) para cada cliente.

$$W_q = E(W_q)$$

Suponga que λ_n es una constante λ para toda n . Se ha demostrado que en un proceso de colas en estado estable,

$$L = \lambda W$$

Además, la misma demostración prueba que:

$$L_q = \lambda W_q$$

Si las λ_n no son iguales, entonces λ se puede sustituir en estas ecuaciones por $\bar{\lambda}$, la **tasa promedio entre llegadas a largo plazo**.

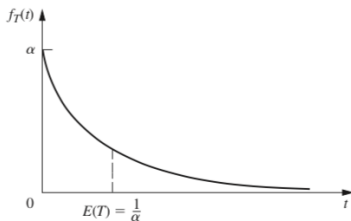
Ahora suponga que el tiempo medio de servicio es una constante $\frac{1}{\mu}$, para toda $n \geq 1$. Se tiene entonces que

$$W = W_q + \frac{1}{\mu}$$

La operación de los sistemas de colas se rige por dos propiedades estadísticas: la distribución de probabilidad de los tiempos entre llegadas y la distribución de probabilidad de los tiempos de servicio.

Suponga que una variable aleatoria T representa ya sea los tiempos entre llegadas o los tiempos de servicio. Se dice que esta variable aleatoria tiene una distribución exponencial con parámetro α si su función de densidad de probabilidad es:

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & \text{para } t \geq 0 \\ 0 & \text{para } t < 0 \end{cases}$$

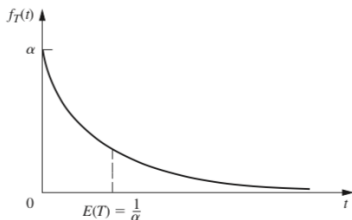


Las probabilidades acumuladas son:

$$\begin{aligned} P\{T \leq t\} &= 1 - e^{-\alpha t} \\ P\{T > t\} &= e^{-\alpha t} \end{aligned} \quad (t \geq 0),$$

el valor esperado y la varianza de T son:

$$\begin{aligned} E(T) &= \frac{1}{\alpha}, \\ \text{var}(T) &= \frac{1}{\alpha^2} \end{aligned}$$



Propiedad 1: $f_T(t)$ es un función de t estrictamente decreciente de $t(t \leq 0)$.

$$P\{0 \leq T \leq \Delta t\} > P\{t \leq T \leq t + \Delta t\}$$

Propiedad 2: Falta de memoria.

$$P\{T > t + \Delta t | T > \Delta t\} = P\{T > t\}$$

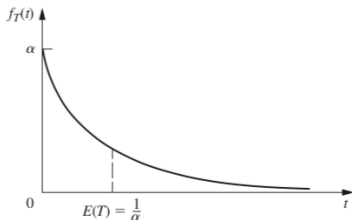
Propiedad 3: El mínimo de diversas variables aleatorias exponenciales independientes tiene una distribución exponencial.

$$U = \min\{T_1, T_2, \dots, T_n\}$$

$$P\{U > t\} = \exp\left(-\sum_{i=1}^n \alpha_i t\right)$$

$$\alpha = \sum_{i=1}^n \alpha_i$$

$$P\{T_j = U\} = \frac{\alpha_j}{\sum_{i=1}^n \alpha_i}, \text{ para } j = 1, 2, \dots, n$$



Propiedad 4: Relación con la distribución de Poisson.

Sea $X(t)$ el número de ocurrencias en el tiempo $t (t \geq 0)$

$$P\{X(t) = n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!},$$

para $n = 0, 1, 2, \dots$

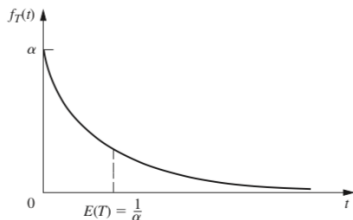
es decir, $X(t)$ tiene una distribución de Poisson con parámetro αt . Para $n = 0$,

$$P\{X(t) = 0\} = e^{-\alpha t}$$

que es exactamente la probabilidad que se obtuvo a partir de la distribución exponencial para que ocurra el primer evento después de un tiempo t . La media de la distribución de Poisson es:

$$E\{X(t)\} = \alpha t$$

de manera que el número esperado de eventos por unidad de tiempo es α . Por lo tanto, se dice que α es la *tasa media* a la que ocurren los eventos. Cuando se cuentan los eventos de manera continua, se dice que el proceso de conteo $\{X(t); t \geq 0\}$ es un **proceso de Poisson** con parámetro α .



Propiedad 5: Para todos los valores positivos de t ,

$$P\{T \leq t + \Delta t | T > t\} \approx \alpha \Delta t,$$

para un Δt pequeño.

Propiedad 6: No afecta agregar o desagregar.

La mayor parte de los modelos elementales de colas suponen que las entradas (llegada de clientes) y las salidas (clientes que se van) del sistema ocurren de acuerdo con un proceso de *nacimiento y muerte*.

Nacimiento = llegada de un nuevo cliente al sistema de colas.

Muerte = salida del cliente servido.

$N(t)$ = estado del sistema en el tiempo $t (t \geq 0)$, es decir, número de clientes que hay en el sistema de colas en el tiempo t .

Los supuestos del proceso de nacimiento y muerte son los siguientes:

Supuesto 1: Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para el próximo nacimiento (llegada) es exponencial con parámetro $\lambda_n (n = 0, 1, 2, \dots)$.

Supuesto 2: Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para la próxima muerte (terminación del servicio) es exponencial con parámetro $\mu_n (n = 0, 1, 2, \dots)$.

Los supuestos del proceso de nacimiento y muerte son los siguientes:

Supuesto 3: Las variables aleatorias del supuesto 1 y supuesto 2 son mutuamente independientes. La siguiente transición del estado del proceso es:

$$n \rightarrow n + 1 \text{ (un solo nacimiento)}$$

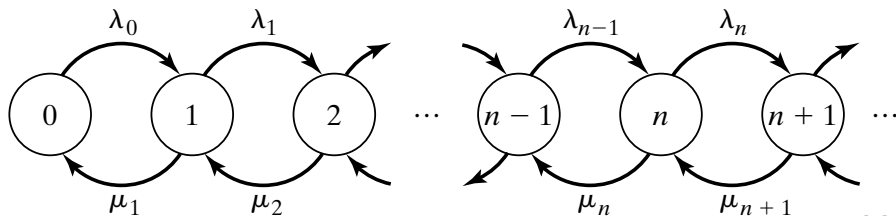
o

$$n \rightarrow n - 1 \text{ (una sola muerte)}$$

lo que depende de cuál de las dos variables es más pequeña.

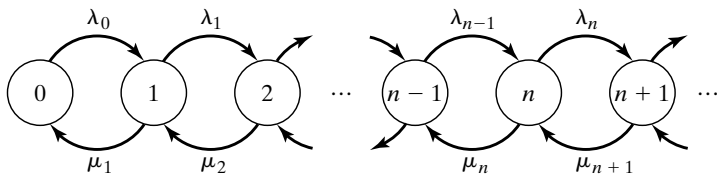
λ_n y μ_n representan, respectivamente, la *tasa media de llegada* y la *tasa media de terminación de servicio*, cuando hay n clientes en el sistema.

Como consecuencia de los supuestos 1 y 2, el proceso de nacimiento y muerte es un tipo especial de *cadena de Markov de tiempo continuo*. Como la propiedad 4 de la distribución exponencial implica que las λ_n y μ_n son tasas medias, estos supuestos se pueden resumir en el diagrama de tasas que se muestra. Las flechas de este diagrama muestran las únicas transiciones posibles en el estado del sistema (supuesto 3) y el elemento junto a cada flecha es la tasa media de esa transición (supuestos 1 y 2) cuando el sistema se encuentra en el estado que hay en la base de la flecha.



Para cualquier estado $n(n = 0, 1, 2, \dots)$ del sistema, la tasa media de entrada = tasa media de salida.

La ecuación que expresa este principio se llama **ecuación de balance** del estado n . Después de construir las ecuaciones de balance de todos los estados en términos de las probabilidades P_n desconocidas, se puede resolver este sistema de ecuaciones para encontrarlas.



Considere el estado 0:

La tasa media de entrada es:

$$\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1$$

La tasa media de salida es: $\lambda_0 P_0$, por lo que la ecuación de balance del estado 0 es:

$$\mu_1 P_1 = \lambda_0 P_0$$

Estado	Tasa de entrada = Tasa de salida
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$
\vdots	\vdots
$n - 1$	$\lambda_{n-2} P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n$
\vdots	\vdots

Para simplificar lo anterior, sea

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \dots \lambda_0}{\mu_n \mu_{n-1} \dots \mu_1}, \text{ para } n = 1, 2, \dots$$

y después se define $C_n = 1$ para $n = 0$.

Las probabilidades de estado estable son:

$$P_n = C_n P_0, \text{ para } n = 0, 1, 2, \dots$$

con la condición:

$$\sum_{n=0}^{\infty} P_n = 1 \longrightarrow \left(\sum_{n=0}^{\infty} C_n \right) P_0 = 1, \text{ así}$$

$$P_0 = \left(\sum_{n=0}^{\infty} C_n \right)^{-1}$$

Cuando un modelo de líneas de espera se basa en el proceso de nacimiento y muerte, de manera que el estado del sistema n representa el número de clientes en el sistema de colas, las medidas clave de desempeño del sistema (L , L_q , W y W_q) se pueden obtener de inmediato después de calcular las P_n mediante las fórmulas anteriores.

$$L = \sum_{n=0}^{\infty} nP_n, L_q = \sum_{n=s}^{\infty} (n - s)P_n$$

$$W = \frac{L}{\bar{\lambda}}, W_q = \frac{L_q}{\bar{\lambda}},$$

donde $\bar{\lambda}$ es la tasa de llegadas *promedio* a largo plazo.

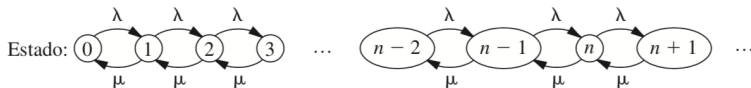
$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Estos resultados de estado estable obedecen al supuesto de que los parámetros λ_n y μ_n tienen valores tales que el proceso, en realidad, puede *alcanzar* la condición de estado estable. Este supuesto siempre se cumple si $\lambda_n = 0$ para algún valor de n mayor que el estado inicial, de forma que sólo son posibles un número finito de estados. También se cumple siempre cuando λ y μ están definidas y $\rho = \frac{\lambda}{s\mu} < 1$. No se cumple si $\sum_{n=0}^{\infty} C_n = \infty$.

Supone que todos los *tiempos entre llegadas* son independientes e idénticamente distribuidos de acuerdo con una distribución exponencial, que todos los *tiempos de servicio* son independientes e idénticamente distribuidos de acuerdo con otra distribución exponencial y que el número de servidores es s .

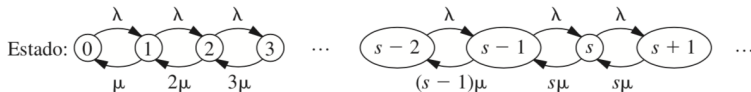
a) Caso de un solo servidor ($s = 1$)

$$\begin{aligned} \lambda_n &= \lambda, & \text{para } n = 0, 1, 2, \dots \\ \mu_n &= \mu, & \text{para } n = 1, 2, \dots \end{aligned}$$



b) Caso de varios servidores ($s > 1$)

$$\begin{aligned} \lambda_n &= \lambda, & \text{para } n = 0, 1, 2, \dots \\ \mu_n &= \begin{cases} n\mu, & \text{para } n = 1, 2, \dots, s \\ s\mu, & \text{para } n = s, s+1, \dots \end{cases} \end{aligned}$$



$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n,$$

para $n = 0, 1, 2, \dots$

$$P_n = \rho^n P_0 = (1 - \rho) \rho^n,$$

para $n = 0, 1, 2, \dots$

$$P_0 = 1 - \rho$$

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

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

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

$$P\{W_q = 0\} = P_0 = 1 - \rho$$

$$P\{W_q > t\} = \rho e^{-\mu(1-\rho)t},$$

para $t \geq 0$

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

Modelo (M/M/s) (s > 1)

$$C_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & \text{para } n = 1, 2, \dots, s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} & \text{para } n = s, s+1, \dots \end{cases}$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$$

$$P_0 = 1 \left/ \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1 - \lambda/(s\mu)} \right] \right.$$

$$\text{si } \lambda < s\mu (\rho = \lambda/(s\mu) < 1)$$

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

$$W = W_q + \frac{1}{\mu}$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0, & \text{si } 0 \leq n \leq s \\ \frac{(\lambda/\mu)^n}{s!s^{(n-s)}} P_0, & \text{si } n > s \end{cases}$$

$$L = L_q + \frac{\lambda}{\mu}$$

$$P\{W > t\} = e^{-\mu t} \left[\frac{1 + P_0(\lambda/\mu)^s}{s!(1-\rho)} \left(\frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s-1-\lambda/\mu} \right) \right]$$

$$P\{W_q > t\} = (1 - P\{W_q = 0\})e^{-s\mu(1-\rho)t}$$

$$P\{W_q = 0\} = \sum_{n=0}^{s-1} P_n$$

A veces los sistemas de colas tienen una *cola finita*; esto es, no se permite que el número de clientes en el sistema exceda un número especificado (K), por lo que la capacidad de la cola es $K - s$.

$$\lambda_n = \begin{cases} \lambda & \text{para } n = 0, 1, 2, \dots, K - 1 \\ 0 & \text{para } n \geq K \end{cases}$$

$$C_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} = \rho^n & \text{para } n = 0, 1, 2, \dots, K \\ 0 & \text{para } n > K \end{cases}$$

$$P_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

para $\rho \neq 1$

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

para $n = 0, 1, 2, \dots, K$

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

$$L_q = L - (1 - P_0), \text{ cuando } s = 1$$

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

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$\bar{\lambda} = \lambda(1 - P_K)$$

$$C_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} & \text{para } n = 0, 1, 2, \dots, s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} & \text{para } n = s, s+1, \dots, K \\ 0 & \text{para } n > K \end{cases}$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & \text{para } n = 1, 2, \dots, s \\ \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0 & \text{para } n = s, s+1, \dots, K \\ 0 & \text{para } n > K \end{cases}$$

$$P_0 = 1 / \left[\sum_{n=0}^s \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s+1}^K \left(\frac{\lambda}{s\mu} \right)^{n-s} \right]$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2} [1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)], \text{ donde } \rho = \lambda/(s\mu)$$

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s \left(1 - \sum_{n=0}^{s-1} P_n \right)$$

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

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$\bar{\lambda} = \lambda(1 - P_K)$$

Modelo (M/M/s) con variación de fuente de entrada finita

Ahora, la *fuentes de entrada está limitada*; es decir, el tamaño de la *población potencial es finito*. En este caso, sea N el tamaño de esa población. Cuando el número de clientes en el sistema de colas es n ($n = 0, 1, 2, \dots, N$), existen sólo $N - n$ clientes potenciales restantes en la fuente de entrada.

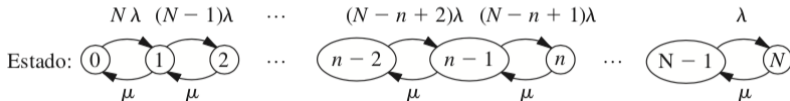
La aplicación más importante de este modelo es el problema de reparación de máquinas, en el que se asigna a uno o más técnicos la responsabilidad de mantener en operación cierto grupo de N máquinas dando servicio a cada una de las que se descomponen. Se considera que un técnico de mantenimiento es un servidor individual en el sistema de colas si trabaja en forma independiente en máquinas diferentes, mientras que los miembros de una cuadrilla completa se toman como un servidor si trabajan unidos en cada máquina. Las máquinas constituyen la población potencial. Cada una se considera un cliente en el sistema de colas cuando está descompuesta en espera de ser reparada, mientras que cuando está en operación está fuera del sistema.

Modelo (M/M/s) con variación de fuente de entrada finita

a) Caso de un solo servidor ($s = 1$)

$$\lambda_n = \begin{cases} (N - n)\lambda, & \text{para } n = 0, 1, 2, \dots, N \\ \infty, & \text{para } n \geq N \end{cases}$$

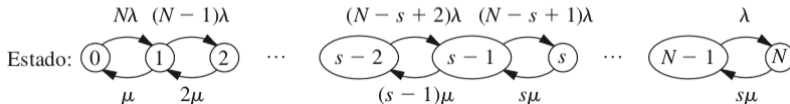
$$\mu_n = \mu, \quad \text{para } n = 1, 2, \dots$$



b) Caso de varios servidores ($s > 1$)

$$\lambda_n = \begin{cases} (N - n)\lambda, & \text{para } n = 0, 1, 2, \dots, N \\ \infty, & \text{para } n \geq N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & \text{para } n = 1, 2, \dots, s \\ s\mu, & \text{para } n = s, s + 1, \dots \end{cases}$$



Modelo (M/M/1) con variación de fuente de entrada finita

$$C_n = \begin{cases} \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n & \text{para } n \leq N \\ 0 & \text{para } n > N \end{cases}$$

$$P_0 = 1 / \sum_{n=0}^N \left[\frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n \right]$$

$$P_n = \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n P_0, \text{ si } n = 1, 2, \dots, N$$

Modelo (M/M/1) con variación de fuente de entrada finita

$$L_q = N - \frac{\lambda + \mu}{\lambda}(1 - P_0)$$

$$L = N - \frac{\mu}{\lambda}(1 - P_0)$$

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

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$\bar{\lambda} = \lambda(N - L)$$

Modelo (M/M/s) ($s > 1$) con variación de fuente de entrada finita

$$C_n = \begin{cases} \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n & \text{para } n = 0, 1, 2, \dots, s \\ \frac{N!}{(N-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n & \text{para } n = s, s+1, \dots, N \\ 0 & \text{para } n > N \end{cases}$$

$$P_n = \begin{cases} \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n P_0 & \text{si } 0 \leq n \leq s \\ \frac{N!}{(N-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n P_0 & \text{si } s \leq n \leq N \\ 0 & \text{si } n > N \end{cases}$$

$$P_0 = 1 / \left[\sum_{n=0}^{s-1} \frac{N!}{(N-n)!n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^N \frac{N!}{(N-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n \right]$$

Modelo (M/M/s) ($s > 1$) con variación de fuente de entrada finita



$$L_q = \sum_{n=s}^N (n-s)P_n$$

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s \left(1 - \sum_{n=0}^{s-1} P_n \right)$$

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

$$W_q = \frac{L_q}{\bar{\lambda}}$$

$$\bar{\lambda} = \lambda(N-L)$$

Supone que el sistema de colas tiene un *servidor* y un *proceso de entradas de Poisson* con una tasa media llegadas *fija* λ . Como siempre, se supone que los clientes tienen tiempos de servicio *independientes* con la misma distribución de probabilidad, pero no se imponen restricciones sobre cuál debe ser esta distribución de tiempos de servicio. En realidad, sólo es necesario conocer (o estimar) la media $1/\mu$ y la varianza σ^2 de esta distribución.

Cualquier sistema de líneas de espera de este tipo podrá alcanzar, en algún momento, una condición de estado estable si $\rho = \lambda/\mu$.

$$P_0 = 1 - \rho$$

$$L_q = \frac{\lambda^2 \sigma^2 + \rho^2}{2(1 - \rho)} \text{ (fórmula de Pollaczek-Khintchine)}$$

$$L = \rho + L_q$$

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

$$W = W_q + \frac{1}{\mu}$$

Cuando el servicio consiste básicamente en la misma tarea rutinaria que el servidor realiza para todos los clientes, tiende a haber poca variación en el tiempo de servicio que se requiere. Muchas veces, el modelo $M/D/s$ proporciona una representación razonable de este tipo de situaciones porque supone que todos los tiempos de servicio son iguales a una constante fija (la distribución de tiempos de servicio degenerada) y que tiene un proceso de entradas de Poisson con tasa media de llegadas fija λ .

Cuando sólo se tiene un servidor, el modelo $M/D/1$ es un caso especial del modelo $M/G/1$, donde $\sigma^2 = 0$, con lo que la fórmula de Pollaczek-Khintchine se reduce a

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

donde a partir de este valor de L_q se pueden obtener L , W_q y W como ya se demostró anteriormente. En el caso de la versión de más de un servidor de este modelo ($M/D/s$) se dispone de un método complicado para obtener la distribución de probabilidad de estado estable del número de clientes en el sistema y su media; por lo que existen tabulaciones de estos resultados para muchos casos.

Otro tipo de distribución teórica de tiempos de servicio con cuerda con la **distribución de Erlang**, la cual es:

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t}, \text{ para } t \geq 0$$

donde μ y k son parámetros estrictamente positivos de la distribución y k está restringido a valores enteros. Su media y desviación estándar son:

$$\text{Media} = \frac{1}{\mu}$$

y

$$\text{Desviación estándar} = \frac{1}{\sqrt{k}} \frac{1}{\mu}$$

k es el parámetro que especifica el grado de variabilidad de los tiempos de servicio con relación a la media. Por lo general, se hace referencia a k como el *parámetro de forma*.

$$L_q = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu - \lambda)}$$

$$W_q = \frac{1+k}{2k} \frac{\lambda}{\mu(\mu - \lambda)}$$

$$W = W_q + \frac{1}{\mu}$$

$$L = \lambda W$$

Los modelos de colas presentados previamente suponen un proceso de entradas de Poisson (tiempos entre llegadas exponenciales). No obstante, este supuesto se viola si las llegadas se programan o regulan de alguna forma que evite que ocurran de manera aleatoria, en cuyo caso se necesita otro modelo.

Se dispone de tres modelos de este tipo siempre que los tiempos de servicio tengan distribución exponencial con un parámetro fijo. Estos modelos se obtienen al invertir las distribuciones supuestas de tiempos entre llegadas y tiempos de servicio de los tres modelos anteriores.

- ($GI/M/s$) no impone restricciones sobre el tipo de distribución de los tiempos entre llegadas. En este caso se dispone de algunos resultados de estado estable de las dos versiones del modelo de uno y varios servidores, pero no son ni cercanamente tan convenientes como las expresiones sencillas del modelo $M/G/1$.
- ($D/M/s$) supone que todos los tiempos entre llegadas son iguales a una constante fija, que representaría un sistema de colas en el que se programan las llegadas a intervalos regulares.
- ($E_k/M/s$) supone una distribución de Erlang de los tiempos entre llegadas que maneja el espacio intermedio entre llegadas regulares programadas (constante) y completamente aleatorias (exponencial).

Si ni los tiempos entre llegadas ni el tiempo de servicio de un sistema de colas tienen distribución exponencial, entonces existen tres modelos de colas adicionales para los que también se tienen resultados.

- $(E_k/E_k/s)$ supone una distribución de Erlang de ambos tiempos.
- $(E_k/D/s)$ y $(D/E_k/s)$ suponen que uno de estos tiempos tiene una distribución de Erlang y el otro es igual a una constante fija.

En los modelos con disciplina de prioridades, la disciplina de la cola se basa en un *sistema prioritario*. El orden en el que se seleccionan los clientes para darles el servicio se basa en sus prioridades asignadas.

Estos modelos suponen:

- existen N clases de prioridad (la clase 1 tiene la prioridad más alta y la clase N la más baja) y que siempre que un servidor queda libre para comenzar el servicio de un nuevo cliente, el cliente que se selecciona es el miembro de la clase prioritaria más alta representada en la cola y que haya esperado más.
- Cada clase prioritaria está sometida a un proceso de entradas de Poisson y tiempos de servicio exponenciales.
- Tiempo medio de servicio es el mismo para todas las clases prioritarias, pero permite que la tasa media de llegadas difiera entre ellas.

Existen dos tipos de modelos de colas que manejan prioridades:

Modelo de **prioridades sin interrupción** no se puede regresar a la cola a un cliente que se encuentra en servicio (interrumpirlo) si entra un cliente de prioridad más alta al sistema de colas. Por lo tanto, una vez que el servidor comienza a atender a un cliente, el servicio debe terminar sin interrupción.

Modelo de **prioridades con interrupción**, el cliente de prioridad más baja que se encuentre en servicio es interrumpido (va de regreso a la cola) cada vez que entra un cliente con prioridad más alta al sistema de colas.

Sea W_k el tiempo esperado de espera en el sistema en estado estable (lo cual incluye el tiempo de servicio) de un miembro de la clase de prioridad k . Entonces,

$$W_k = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu}, \text{ para } k = 1, 2, \dots, N$$

donde

$$A = s! \frac{s\mu - \lambda}{s\mu} \sum_{j=0}^{s-1} \frac{r^j}{j!} + s\mu,$$

$$B_0 = 1,$$

$$B_k = 1 - \frac{\sum_{i=1}^k \lambda_i}{s\mu},$$

s = número de servidores,

μ = tasa media de servicio por servidor ocupado,

λ_i = tasa media de llegadas de la clase de prioridad i ,

$$\lambda = \sum_{i=1}^N \lambda_i,$$

$$r = \frac{\lambda}{\mu}.$$

$$L_k = \lambda_k W_k, \text{ para } k = 1, 2, \dots, N$$

$$W_{k_q} = W_k - \frac{1}{\mu}$$

$$L_{k_q} = W_k \lambda_k.$$

Variación de un servidor en el modelo de prioridades sin interrupción

μ_k = tiempo medio de servicio de la clase de prioridad k , para $k = 1, 2, \dots, N$

$$W_k = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k}, \text{ para } k = 1, 2, \dots, N$$

donde

$$a_k = \sum_{i=1}^k \frac{i}{\mu_i^2},$$

$$b_0 = 1,$$

$$b_k = 1 - \sum_{i=1}^k \frac{\lambda_i}{\mu_i},$$

válido mientras $\sum_{i=1}^k \frac{\lambda_i}{\mu_i} < 1$.

En este tipo de modelos se necesita retomar el supuesto de que el tiempo esperado de servicio es el mismo para todas las clases de prioridad. Con la misma notación que en el modelo anterior, el hecho de poder interrumpir cambia el tiempo esperado total en el sistema (incluido el tiempo total de servicio) a

$$W_k = \frac{1/\mu}{B_{k-1}B_k}, \text{ para } k = 1, 2, \dots, N$$

Cuando $s = 1$ (caso de 1 servidor):

$$L_k = \lambda_k W_k, \text{ para } k = 1, 2, \dots, N$$

Hasta ahora se han estudiado nada más los sistemas de colas que tienen una estación de servicio con uno o más servidores, pero, en realidad, los sistemas de colas que se encuentran en el mundo real a veces son redes de colas, es decir, redes de instalaciones de servicio en las que los clientes solicitan el servicio de algunas o todas ellas.

Propiedad de equivalencia: suponga que una instalación de servicio tiene s servidores, un proceso de entradas Poisson con parámetro λ y la misma distribución de los tiempos de servicio de cada servidor con parámetro μ (el modelo $M/M/s$), en donde $s\mu > \lambda$. Entonces, la salida en estado estable de esta instalación de servicio también es un proceso de Poisson con parámetro λ .

Suponga que todos los clientes deben recibir servicio en **una serie** de m instalaciones, en una secuencia fija. Además, que cada instalación tiene una cola infinita (el número de clientes que acepta no tiene límite), de manera que las instalaciones en serie forman un sistema de **colas infinitas en serie**. Suponga, además, que los clientes llegan a la primera instalación de acuerdo con un proceso de Poisson con parámetro λ y que cada instalación i ($i = 1, 2, \dots, m$) tiene la misma distribución exponencial de tiempos de servicio con parámetro μ_i de sus s_i servidores, donde $s_i \mu_i > \lambda$. Debido a la propiedad de equivalencia se puede decir que (en condiciones de estado estable) cada instalación de servicio tiene entrada Poisson con parámetro λ . Entonces, se puede usar el modelo elemental $M/M/s$ para analizar cada instalación de servicio en forma independiente de las otras.

Al poder usar el modelo $M/M/s$ para obtener las medidas de desempeño de cada instalación independiente, en lugar de analizar la interacción entre las instalaciones, se logra una simplificación enorme.

Por ejemplo, la probabilidad de tener n clientes en una instalación en particular está dada por la fórmula de P_n del modelo $M/M/s$. La probabilidad conjunta de n_1 clientes en la instalación 1, n_2 clientes en la instalación 2, etc., es, entonces, el producto de las probabilidades individuales obtenidas de esta manera sencilla. En particular, esta probabilidad conjunta se puede expresar como:

$$P\{(N_1, N_2, \dots, N_m) = (n_1, n_2, \dots, n_m)\} = P_{n1}P_{n2}P_{nm}$$

(Esta forma sencilla de solución se llama **solución en forma de producto**). De manera similar, el tiempo de espera total esperado y el número esperado de clientes en el sistema completo se pueden obtener con sólo sumar las cantidades correspondientes que se obtuvieron de cada instalación.

Los sistemas de colas infinitas en serie no son las únicas redes de colas en las que se puede usar el modelo $M/M/s$ para analizar cada instalación de servicio de manera independiente. Otro tipo importante de redes con esta propiedad son las redes de Jackson.

Las características de una red de Jackson son las mismas supuestas para el sistema de colas infinitas en serie, excepto que ahora los clientes visitan las instalaciones en diferente orden (y pueden no llegar a todas). Para cada instalación, los clientes que llegan provienen tanto de afuera del sistema (de acuerdo con un proceso de Poisson) como de otras instalaciones.

Una **red de Jackson** es un sistema de m instalaciones de servicio en donde la instalación i ($i = 1, 2, \dots, m$) tiene:

- 1 Una cola infinita
- 2 Clientes que llegan de afuera del sistema según un proceso de entrada Poisson con parámetro a_i
- 3 s_i servidores con distribución exponencial de tiempos de servicio con parámetro μ_i .

Un cliente que deja la instalación i se encamina después a la instalación j ($j = 1, 2, \dots, m$) con probabilidad p_{ij} o sale del sistema con probabilidad

$$q_i = 1 - \sum_{j=1}^m p_{ij}$$

Cualquier red de este tipo tendrá la siguiente propiedad:

En condiciones de estado estable, cada instalación j ($j = 1, 2, \dots, m$) de una red de Jackson se comporta como si fuera un sistema de colas $M/M/s$ *independiente* con tasa de llegadas

$$\lambda_j = a_j + \sum_{i=1}^m \lambda_i p_{ij},$$

donde $s_i \mu_i > \lambda_j$

Cada instalación i , los procesos de entrada desde las diferentes fuentes (externa y de otras instalaciones) son *procesos de Poisson independientes*, de manera que el proceso de entrada *agregado* es de Poisson con parámetro λ_i . La propiedad de equivalencia dice entonces que el proceso de *salida agregado* de la instalación i debe ser Poisson con parámetro λ_i . Al desagregar este proceso de salida, el proceso de los clientes que salen de la instalación i a la instalación j debe ser Poisson con parámetro $\lambda_i p_{ij}$. Este proceso se convierte en uno de los procesos de *entrada* de Poisson de la instalación j , lo que ayuda a mantener la serie de procesos de Poisson en todo el sistema.

La ecuación anterior para obtener λ_j se basa en el hecho de que λ_i es tanto la *tasa de salida* como la *tasa de entrada* de todos los clientes que utilizan la instalación i . Como p_{ij} es la proporción de clientes que salen de la instalación i para ir a la instalación j , la tasa a la que estos clientes de la instalación i llegan a la instalación j es $\lambda_i p_{ij}$. Al sumar estos productos sobre toda i y después agregar esta suma a a_j , se obtiene la *tasa de llegadas total* a la instalación j desde todas las fuentes.

Para calcular λ_j a partir de esta ecuación se requiere conocer las λ_i para $i \neq j$, pero estas λ_i también son incógnitas dadas por las ecuaciones correspondientes. Por lo tanto, el procedimiento es obtener *simultáneamente* $\lambda_1, \lambda_2, \dots, \lambda_m$ mediante la solución simultánea de todo el sistema de ecuaciones lineales, con λ_j para $j = 1, 2, \dots, m$.

- ① Taha, H. A. Investigación de operaciones. Pearson Educación.
- ② Hillier, F. S., Lieberman, G. J., & Osuna, M. A. G. Introducción a la Investigación de Operaciones. McGraw-Hill.