Teoría de colas

Andrés Ramos
Universidad Pontificia Comillas
http://www.iit.comillas.edu/aramos/
Andres.Ramos@comillas.edu

Sistemas de colas

- Una cola se produce cuando la demanda de un *servicio* por parte de los *clientes* excede la capacidad del servicio.
- Se necesita conocer (predecir) el *ritmo de entrada* de los clientes y el *tiempo de servicio* con cada cliente.

Objetivo:

Equilibrar los costes de capacidad del servicio y el "coste" de una espera larga.

TEORÍA DE COLAS

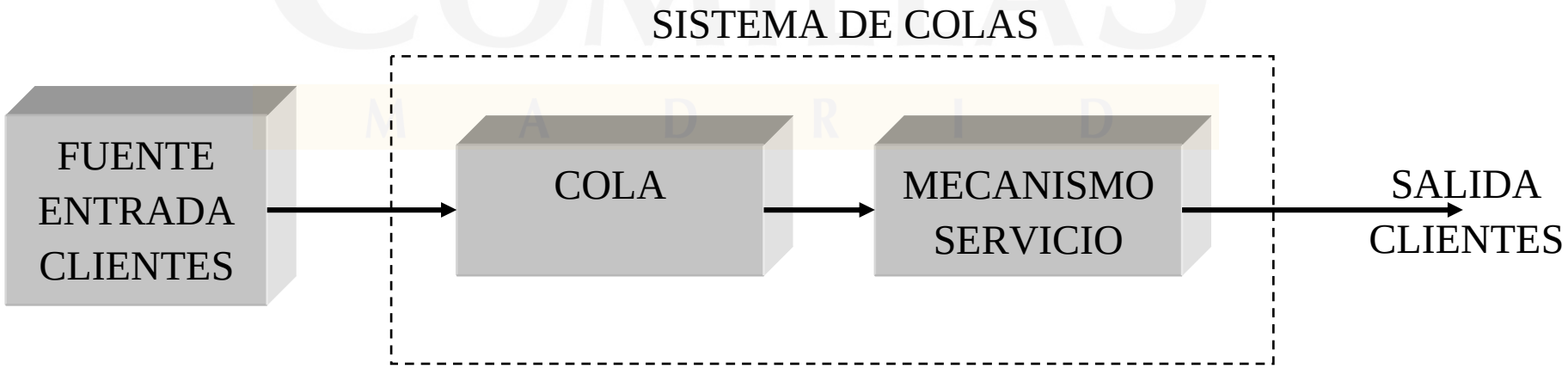
Estudio matemático de las características de los sistemas de colas.

Proceso en una cola

1. Entrada de clientes

2. Sistema de colas { cola o línea de espera
   mecanismo de servicio

3. Salida de clientes



Ejemplos

Clientes	Servicio	Servidores
Clientes tienda	Venta artículo	Dependiente
Clientes banco	Operación financiera	Ventanilla
Clientes supermercado	Cobro compra	Caja
Automóvil	Llenar depósito	Surtidor
Automóvil	Reparación avería	Operarios taller
Avión	Aterrizaje / despegue	Pista
Llamadas telefónicas	Conversación	Centralitas
Enfermos	Atención médica	Médico
Cajas	Transporte	Robot de almacenamiento
Juicios pendientes	Juicio	Jueces

Entrada de clientes

TAMAÑO

Número total de clientes potenciales (población de entrada):

- Finito (fuente limitada) (sistema cerrado)
- Infinito (fuente ilimitada) (sistema abierto)

Suposición habitual: tamaño infinito (es decir, el número de clientes en la cola NO afecta el número potencial de clientes fuera de ella)

ENTRADA O FUENTE

- Unitaria
- Por bloques

TIEMPO ENTRE LLEGADAS

- Determinista
- Probabilista (distribución de probabilidad exponencial)

TASA MEDIA DE LLEGADA λ

Número medio de entrada de clientes por unidad de tiempo

Llegadas de clientes son independientes e idénticamente distribuidas (IID)

Cola

Número máximo de clientes admisible

- Finito
- Infinito

Suposición habitual: colas de longitud infinita (pérdida del cliente o reintento)

Número de canales (carriles de una calle ante un semáforo) en la cola e interferencia entre ellos

Disciplina de la cola

Orden de selección de sus miembros para ser atendidos

- FIFO, FIFO con límite
- LIFO
- SIRO (Aleatorio)
- Por prioridad (interruptora o no)

Mecanismo de servicio

SERVIDORES

Proporcionan el servicio al cliente

Número de servidores:

- Uno
- Varios

Independencia o no entre servidores

TIEMPO DE SERVICIO

- Determinista
- Probabilista (distribución de probabilidad exponencial)

TASA MEDIA DE SERVICIO μ

Número medio de clientes que son atendidos en un servidor por unidad de tiempo.

Servicios a clientes son independientes e idénticamente distribuidas (IID)

Especificación de un sistema de colas

Distribución del tiempo entre llegadas / Distribución del tiempo de servicio / Número de servidores / Número máximo de clientes en el sistema / Disciplina de la cola

M exponencial
D degenerada (tiempos constantes)
E Erlang (Gamma)
G general

Ejemplos:

M/M/s tiempo entre llegadas exponencial / tiempo de servicio exponencial / s servidores

M/M/s/K/FIFO

M/M/s/s

M/G/1

Medidas de eficacia de un sistema de colas

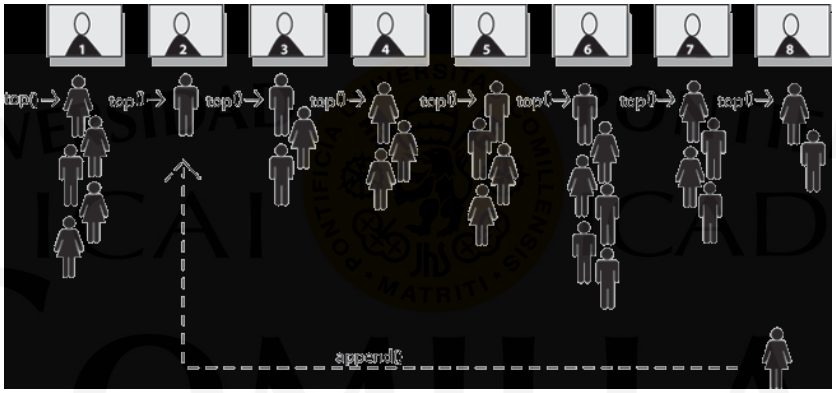
λ	tasa de llegada	$1/\lambda$	tiempo medio entre llegadas consecutivas
μ	tasa de servicio	$1/\mu$	tiempo medio de servicio

ρ factor de utilización (intensidad de tráfico): fracción esperada de tiempo que están ocupados los *s* servidores $\rho = \frac{\lambda}{s\mu}$ habitualmente $\rho < 1$

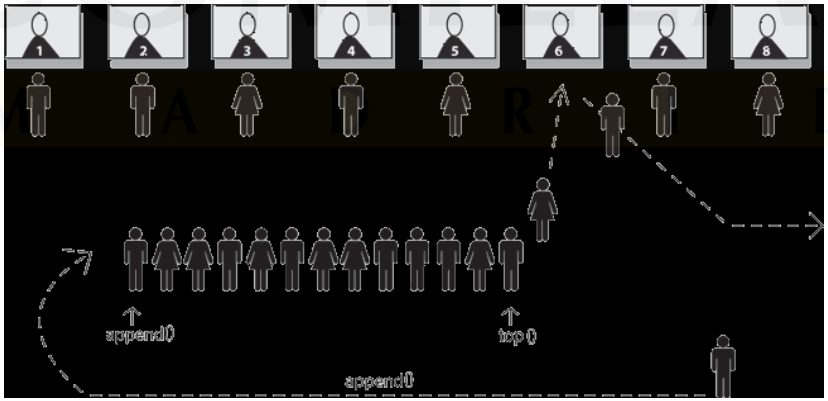
N	estado del sistema, número de clientes en el sistema (cola + servicio)	
L	número *medio* de clientes en el sistema	$L = \mathrm{E}[N]$
N_q	longitud de la cola, número de clientes en la cola	
L_q	número *medio* de clientes en la cola	$L_q = \mathrm{E}[N_q]$
T	tiempo de estancia de los clientes en el sistema	
W	tiempo *medio* de estancia de los clientes en el sistema	$W = \mathrm{E}[T]$
T_q	tiempo de espera de los clientes en la cola	
W_q	tiempo *medio* de espera de los clientes en la cola	$W_q = \mathrm{E}[T_q]$
$\overline{c}$	número *medio* de servidores ocupados	

¿Qué sistema de colas es más efectivo?

Sistema de 8 servidores con 8 colas.



Sistema de 1 cola que abastece a 8 servidores.



Fórmulas de Little para condición estacionaria en sistema M/M/1

La condición estacionaria se produce cuando la *distribución* del número de clientes en el sistema se conserva a través del tiempo.

Número medio de clientes en el sistema/cola = tasa de llegada x tiempo medio de los clientes en el sistema/cola

$$L = \lambda W \qquad L_q = \lambda W_q$$

Tiempo medio de los clientes en el sistema = tiempo medio de los clientes en la cola + tiempo medio de servicio

$$W = W_q + 1/\mu$$

Número medio de clientes en el sistema = número medio de clientes en la cola + factor de utilización (número medio de clientes siendo atendidos)

$$L = L_q + \lambda/\mu$$

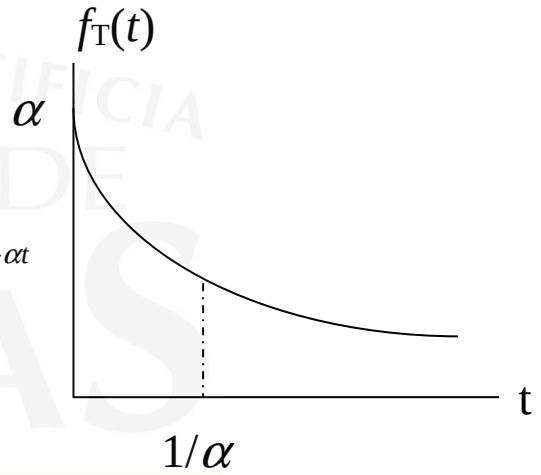
NO PUEDEN UTILIZARSE SI HAY TASAS DE SERVICIO DIFERENTES.

Distribución exponencial

T variable aleatoria tiempo entre llegadas o tiempo de servicio

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

estrictamente decreciente en t



Probabilidad de una llegada después del instante t $P\{T > t\} = e^{-\alpha t}$

$$\text{var}(T) = 1/\alpha^2$$

FALTA DE MEMORIA:

La distribución de la probabilidad del tiempo que falta para que ocurra el evento es siempre la misma independientemente del tiempo que haya pasado

$$P\{T > t + \Delta t \mid T > \Delta t\} = \frac{P\{T > \Delta t \mid T > t + \Delta t\} P\{T > t + \Delta t\}}{P\{T > \Delta t\}} = \frac{e^{-\alpha(t+\Delta t)}}{e^{-\alpha \Delta t}} = e^{-\alpha t} = P\{T > t\}$$

El **mínimo de variables aleatorias exponenciales** tiene distribución **exponencial**.

$$P(B / A) = \frac{P(A / B) \cdot P(B)}{P(A)}$$

Procesos de Poisson

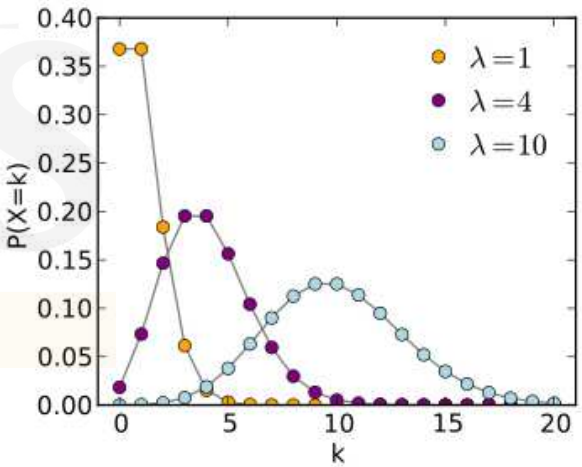
Si los tiempos entre llegadas/servicios se distribuyen según una exponencial el número de llegadas/servicios **hasta un cierto tiempo** es un proceso de Poisson.

$N(t)$ número de ocurrencias (llegadas o servicios) en el tiempo t $(t \geq 0)$. Se distribuye según una Poisson con parámetro αt (α número medio de ocurrencias por unidad de tiempo)

$$P\{N(t)=n\} = \frac{(\alpha t)^n e^{-\alpha t}}{n!} \qquad n = 0,1,\ldots$$

$$P\{N(t)=0\} = e^{-\alpha t} = P\{T > t\}$$

$$E[N(t)] = \alpha t$$



La probabilidad de ocurrencia de un suceso en el siguiente intervalo (pequeño) de tiempo Δt sabiendo que no se ha producido hasta ese momento t es $\alpha\Delta t$ $\quad P\{T \leq t + \Delta t \mid T > t\} \cong \alpha\Delta t$

Procesos de Poisson

PROPIEDAD REPRODUCTIVA:

La suma de procesos de entrada de Poisson es también un proceso de Poisson siendo la tasa la suma de las tasas respectivas.

DIVISIBILIDAD:

Si las llegadas a un sistema son de tipo Poisson con tasa α y cada llegada es encaminada a un subsistema s con una probabilidad p_i el proceso de llegada a cada subsistema es también de Poisson con tasa αp_i

Modelo general. Proceso estacionario de nacimiento y muerte

Nacimiento = llegada de clientes al sistema

Muerte = salida de clientes una vez servidos

$N(t)$ estado del sistema en tiempo t = número de cliente en el sistema

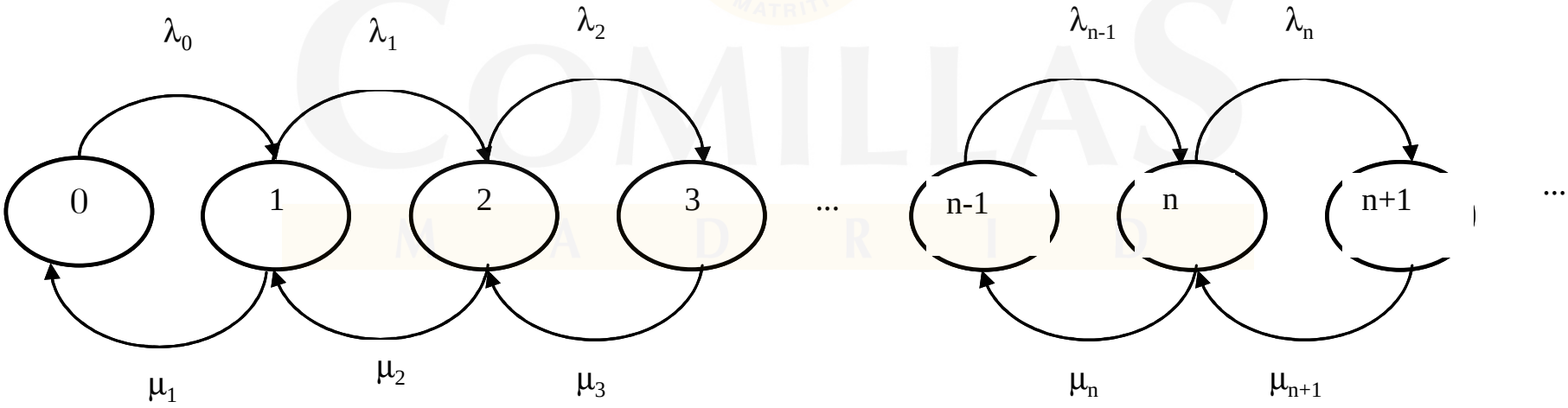
Hipótesis:

- Distribución del tiempo que falta para la llegada es *exponencial* con parámetro λ_n $n = 0,1,\ldots$ siendo λ_n la tasa de llegada de clientes al sistema dado que hay n clientes $N(t) = n$
- Distribución del tiempo que falta para la salida es *exponencial* con parámetro μ_n $n = 0,1,\ldots$ siendo μ_n la tasa de salida de clientes del sistema dado que hay n clientes $N(t) = n$
- *Independencia* entre el tiempo hasta próxima llegada y tiempo hasta próxima salida

Diagrama de transiciones

Por ser proceso de Poisson, la probabilidad de ocurrencia de un suceso en un Δt es proporcional a Δt siendo $\Delta t \to 0$

Tanto la llegada como la salida son procesos de Poisson e independientes, luego de un estado dado sólo se puede pasar a dos posibles estados.



Tasa media de llegada al estado n $\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1}$

Tasa media de salida del estado n $\lambda_n P_n + \mu_n P_n$

P_n probabilidad de que haya n clientes en el sistema de manera estacionaria

Por ser el sistema estacionario (tasa medio de llegada = tasa media de salida) para cualquier estado n $\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} = \lambda_n P_n + \mu_n P_n$

$n=0$ $\mu_1 P_1 = \lambda_0 P_0$ $P_1 = \frac{\lambda_0}{\mu_1} P_0$

$n=1$ $\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1$ $P_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$

$n=2$ $\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2) P_2$ $P_3 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$

$$P_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n \mu_{n-1}\cdots\mu_1} P_0 \qquad \sum_{n=0}^{\infty} P_n = 1$$

$$C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n \mu_{n-1}\cdots\mu_1} \qquad n = 1, 2, \ldots$$

$$C_0 = 1 \qquad n = 0$$

$$\sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} C_n P_0 = 1 \qquad P_0 = \frac{1}{\sum_{n=0}^{\infty} C_n}$$

Número medio de clientes en el sistema $$L = \sum_{n=0}^{\infty} nP_n$$

Número medio de clientes en cola con *s* servidores $$L_q = \sum_{n=s}^{\infty} (n-s)P_n$$

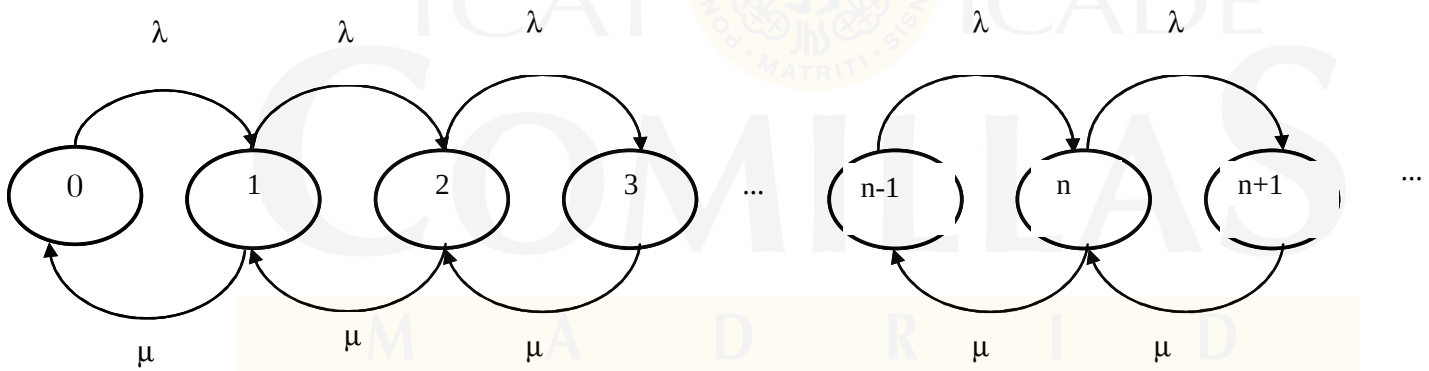
Tasa media de llegadas $$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$

Cola M/M/1

Tasa media de llegada λ constante e independiente del estado del sistema $\lambda_n = \lambda$

Tasa media de servicio μ constante e independiente del estado del sistema $\mu_n = \mu$

Factor de utilización $\rho = \dfrac{\lambda}{\mu}$ Para alcanzar estado estable $\rho < 1$



$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \qquad P_n = \rho^n P_0 \qquad P_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho \qquad P_n = (1-\rho)\rho^n \qquad n = 0,1,2,\ldots$$

Medidas de funcionamiento de cola M/M/1

Número medio de clientes en el sistema $L = \sum_{n=0}^{\infty} nP_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$

Número medio de clientes en cola con 1 servidor $L_q = \sum_{n=1}^{\infty} (n-1)P_n = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$

Tiempo medio de los clientes en el sistema $W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$

Tiempo medio de los clientes en cola $W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$

Factor de utilización del servidor $\rho = L - L_q = 1 - P_0$

Probabilidad de tiempo de espera en cola nulo $P_0 = 1 - \rho = P\{W_q = 0\}$

Probabilidad de tiempo de espera en cola > t $P\{W_q > t\} = \rho e^{-\mu(1-\rho)t}$ $t \geq 0$

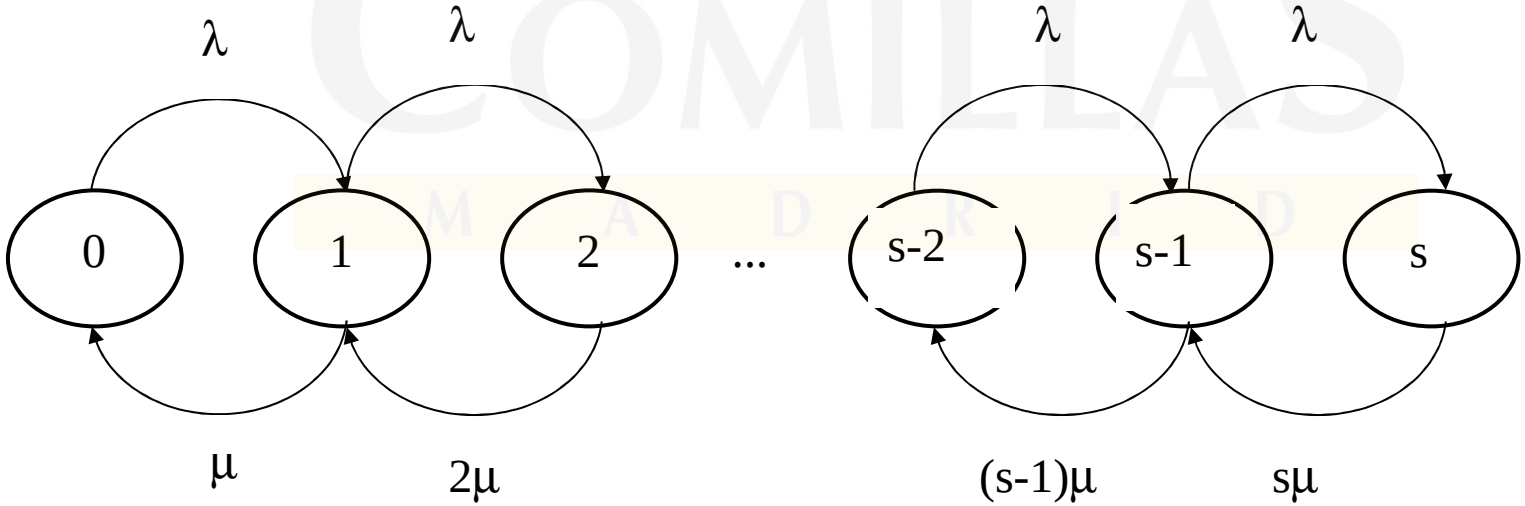
Probabilidad de tiempo de estancia en el sistema > t $P\{W > t\} = e^{-\mu(1-\rho)t}$ $t \geq 0$

Cola M/M/*s*

Tasa media de llegada λ constante e independiente del estado del sistema $\lambda_n = \lambda$

Tasa media de servicio μ $\mu_n = \begin{cases} n\mu & n \le s \\ s\mu & n > s \end{cases}$

Factor de utilización $\rho = \frac{\lambda}{s\mu}$ Para alcanzar estado estable $\rho < 1$



$$C_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n & n \le s \\ \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{n-s} & n > s \end{cases}$$

$$P_0 = \frac{1}{\sum_{n=0}^{\infty} C_n} = \frac{1}{1 + \sum_{n=1}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{\infty} \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{n-s}} = \frac{1}{1 + \sum_{n=1}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{1}{1 - \frac{\lambda}{s\mu}}}$$

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} + \frac{(s\rho)^s}{s!(1-\rho)}}$$

$$P_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & n \le s \\ \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^n \frac{1}{s^{n-s}} P_0 & n > s \end{cases}$$

Medidas de funcionamiento de cola M/M/*s*

Número medio de clientes en cola con *s* servidores $L_q = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0$

Número medio de clientes en el sistema $L = L_q + \frac{\lambda}{\mu}$

Tiempo medio de los clientes en cola $W_q = \frac{L_q}{\lambda}$

Tiempo medio de los clientes en el sistema $W = \frac{L}{\lambda} = W_q + \frac{1}{\mu}$

Probabilidad de tiempo de estancia en el sistema > *t*

$$P\{W > t\} = e^{-\mu t}\left[1 + \frac{P_0(\lambda/\mu)^s}{s!(1-\rho)} \frac{1 - e^{-\mu t(s-1-\lambda/\mu)}}{s-1-\lambda/\mu}\right] \quad t \geq 0$$

Probabilidad de tiempo de espera en cola > *t* $P\{W_q > t\} = \left[1 - P\{W_q = 0\}\right] e^{-s\mu(1-\rho)t} \quad t \geq 0$

Probabilidad de tiempo de espera en cola nulo $P\{W_q = 0\} = \sum_{n=0}^{s-1} P_n$

Cola M/M/s/K

K número máximo de clientes en el sistema (por ejemplo, lugares disponibles para los clientes –camillas-)

No se permite la entrada cuando el sistema está lleno.

Tasa media de llegada $\lambda_n = \begin{cases} \lambda & n = 0,1,2,\ldots,K-1 \\ 0 & n \geq K \end{cases}$

Número de servidores inferior al número máximo de clientes $s \leq K$

$$C_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n & n = 0,1,2,\ldots,s \\ \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{n-s} & n = s, s+1,\ldots,K \\ 0 & n > K \end{cases} \qquad P_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & n = 0,1,2,\ldots,s \\ \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \left(\frac{\lambda}{s\mu}\right)^{n-s} P_0 & n = s, s+1,\ldots,K \\ 0 & n > K \end{cases}$$

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

Número medio de clientes en cola $$L_q = \frac{(\lambda/\mu)^s \rho}{s!(1-\rho)^2} P_0 \left[1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho)\right]$$

Número medio de clientes en el sistema $$L = \sum_{n=0}^{s-1} nP_n + L_q + s\left(1 - \sum_{n=0}^{s-1} P_n\right)$$

Tasa media de llegada (entrada efectiva) $$\lambda_{EF} = \lambda(1 - P_K)$$

Tiempo medio de los clientes en cola $$W_q = \frac{L_q}{\lambda_{EF}}$$

Tiempo medio de los clientes en el sistema $$W = \frac{L}{\lambda_{EF}}$$

Cola M/G/1

Tiempos entre llegadas independientes y distribución exponencial con tasa de llegada λ

Tiempos de servicio independientes y distribución **general** $F(\bullet)$ con media $\frac{1}{\mu}$ y varianza σ^2

No se puede aplicar el proceso generalizado de nacimiento y muerte.

Fórmula de Pollaczek-Khintchine: $L = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)}$ siendo $\rho = \frac{\lambda}{\mu}$.

Sistema cerrado con cola M/M/1

Fuente finita de tamaño m. Clientes una vez servidos vuelven a la fuente.

Tiempos entre llegadas independientes y distribución exponencial con tasa de llegada dependiente del número de clientes en el sistema $\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$

Probabilidad de cada estado

$$P_n = \frac{m!}{(m-n)!}\rho^n P_0 = (m-n+1)\rho P_{n-1} \quad 0 < n \leq m \qquad \text{y} \quad P_0 = \left[1 + \sum_{n=1}^{m} \frac{m!\rho^n}{(m-n)!}\right]^{-1}$$

$$P_n = 0 \qquad n > m$$

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

Tasa media de llegada al sistema $\lambda_{EF} = (m - L)\lambda$

Número medio de clientes en cola $L_q = m - \frac{1+\rho}{\rho}(1 - p_0)$

Número medio de clientes en el sistema $L = m - \frac{1 - p_0}{\rho}$

Tiempo medio de los clientes en cola $W_q = \frac{L_q}{(m-L)\lambda} = \frac{1}{\mu}\left[\frac{m}{1-p_0} - \frac{1+\rho}{\rho}\right]$

Tiempo medio de los clientes en el sistema $W = \frac{L}{(m-L)\lambda}$

Sistema cerrado con cola M/M/s

Fuente finita de tamaño m. Clientes una vez servidos vuelven a la fuente.

Tiempos entre llegadas independientes y distribución exponencial con tasa de llegada dependiente del número de clientes en el sistema $\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$

Tasa media de servicio μ $\qquad \mu_n = \begin{cases} n\mu & 0 \leq n \leq s \\ s\mu & s \leq n \leq m \end{cases}$

Probabilidad de cada estado

$$P_n = \begin{cases} \binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n P_0 & 0 \leq n \leq s \\ \binom{m}{n}\frac{n!(\lambda/\mu)^n}{s!\,s^{n-s}} P_0 & s \leq n \leq m \end{cases} \quad \text{siendo } \rho = \frac{\lambda}{s\mu}$$

Tasa media de llegada al sistema $\qquad \lambda_{EF} = (m - L)\lambda$

Cola M/M/s/s

Capacidad del sistema es igual número de servidores (centrales telefónicas).

Probabilidad de que el sistema esté saturado (número de clientes igual a número de servidores) $P_s = \frac{(s\rho)^s / s!}{\sum_{i=0}^{s} (s\rho)^i / i!}$

Cola M/M/∞

El sistema tiene un número muy grande de servidores (sistemas de autoservicio, visitas a una ciudad).

Tasa de llegadas $\lambda_n = \lambda$

Tasa de servicios $\mu_n = n\mu$

Probabilidad de cada estado $p_n = e^{-\lambda/\mu}\dfrac{(\lambda/\mu)^n}{n!} \quad n = 0,1,...$

Medidas de funcionamiento de la cola $L = \dfrac{\lambda}{\mu}; \quad L_q = 0; \quad W = \dfrac{1}{\mu}; \quad W_q = 0$

Diseño óptimo de los sistemas de colas

Objetivo:

Determinar el nivel de servicio que minimiza la suma de costes incurridos por proporcionar el servicio + costes de los clientes por estar en el sistema (Número medio de clientes en el sistema L por coste de estancia de cada cliente C_c)

Coste de los clientes:

- Pérdidas de ganancia por pérdida de clientes
- Coste social del servicio
- Pérdida de productividad

Decisiones:

- Número de servidores por instalación s
- Eficiencia de los servidores μ
- Número de sistemas en servicio (instalaciones) λ

Optimizar el número de servidores

μ, λ conocidos y fijos

C_s coste por servidor por unidad de tiempo

$$\min E\left[CT(s)\right] = sC_s + C_c L(s) \qquad s \in N$$

$$CT(s-1) \geq CT(s) \leq CT(s+1)$$

$$\Rightarrow \quad L(s) - L(s+1) \leq \frac{C_s}{C_c} \leq L(s-1) - L(s)$$

Optimizar la tasa de servicio

λ conocida y fija
C_μ coste por unidad de tasa de servicio por unidad de tiempo

$$\min E\left[CT(\mu)\right] = \mu C_\mu + C_c L(\mu)$$

Para cola M/M/1

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

$$\frac{\partial E\left[CT(\mu)\right]}{\partial \mu} = 0 \quad \Rightarrow \quad \mu = \lambda + \sqrt{\frac{C_c \lambda}{C_\mu}}$$

Optimizar la tasa de servicio y la capacidad del sistema

λ	conocida y fija
C_K	coste por unidad de capacidad por unidad de tiempo
C_p	coste por clientes perdidos por unidad de tiempo

$$E\left[CT(\mu,K)\right] = \mu C_\mu + C_c L(\mu,K) + KC_K + \lambda P_K C_p \qquad K \in N$$