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 teórico:

Determinar la distribución del número de clientes en el sistema

Objetivo práctico:

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

TEORÍA DE COLAS

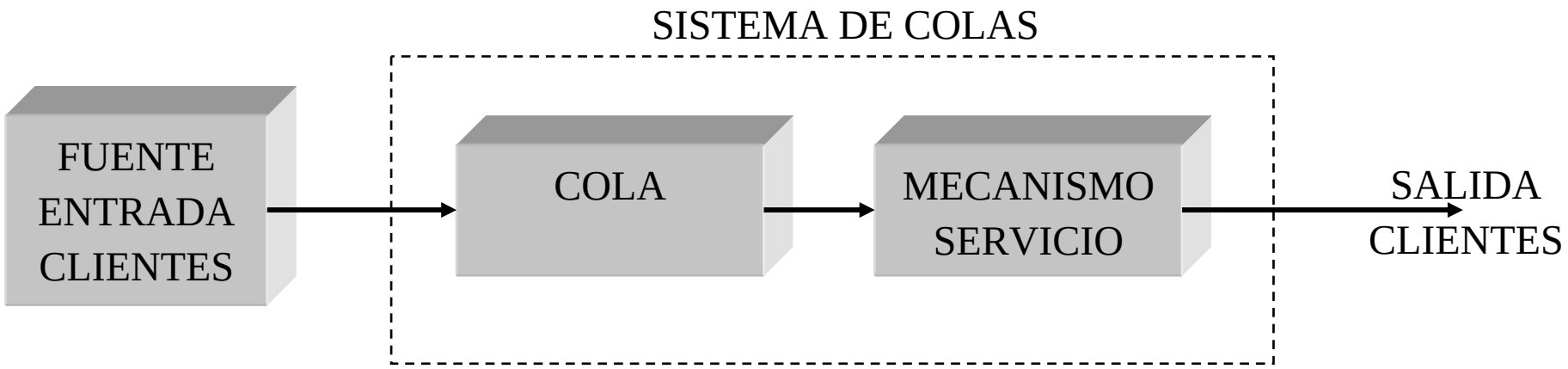
Estudio matemático de las características de los sistemas de colas.

Proceso en una cola

1. Entrada de clientes

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

3. Salida de clientes



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 (hipótesis usual)
- Por bloques

TIEMPO ENTRE LLEGADAS

- Determinista
- Probabilista (hipótesis usual)

Suposición habitual: distribución de probabilidad exponencial y llegadas de clientes independientes e idénticamente distribuidas (IID)

TASA MEDIA DE LLEGADA λ

Número medio de llegadas de clientes por unidad de tiempo

La tasa puede variar en función del número de clientes en la cola

($1/\lambda$ es entonces el tiempo medio entre llegadas)

TASA MEDIA DE ACCESO (O DE LLEGADA EFECTIVA) λ_{EF}

Número medio de entrada de clientes (los que realmente acceden al sistema) por unidad de tiempo

Sólo tiene sentido cuando hay una capacidad de cola

(más adelante se define el cálculo de λ_{EF})

Cola

NÚMERO MÁXIMO DE CLIENTES ADMISIBLE (capacidad de cola)

- Finito (pérdida del cliente o reintento)
- Infinito

Suposición habitual: colas de longitud infinita

NÚMERO DE CANALES (carriles de una calle ante un semáforo) en la cola.
Puede haber interferencia entre canales (movimientos de clientes de un canal a otro)

Disciplina de la cola

ORDEN DE SELECCIÓN de sus miembros para ser atendidos

- FIFO, FIFO con límite (en el tiempo de servicio, de tal forma que si se supera se vuelve a la cola y cuando es de nuevo atendido empieza donde acabó el servicio) (hipótesis usual)
- LIFO
- SIRO (Aleatorio)
- Por prioridad (interruptora o no)

Mecanismo de servicio

SERVIDORES

Proporcionan el servicio al cliente

Número de servidores:

- Uno
- Varios

Puede haber independencia o no entre servidores

TIEMPO DE SERVICIO

- Determinista
- Probabilista (hipótesis usual)

Suposición habitual: distribución de probabilidad exponencial e independencia e idéntica distribución en los tiempos de servicio de un mismo servidor (IID)

TASA MEDIA DE SERVICIO μ

Número medio de clientes que son atendidos en un servidor por unidad de tiempo.
La tasa puede variar en función del número de clientes en la cola
($1/\mu$ es entonces el tiempo medio entre servicios)

TASA MEDIA DE SERVICIO DEL SISTEMA μ_{EF}

Número medio de clientes que son atendidos en el sistema por unidad de tiempo.

(más adelante se define el cálculo de μ_{EF})

Hipótesis fundamental:

ρ factor de utilización (intensidad de tráfico): proceso no sea explosivo, es decir, que el número de clientes no tenga una tendencia creciente:

$$\rho = \frac{\lambda_{EF}}{\overline{\mu}_{EF}} < 1$$

Siendo $\overline{\mu}_{EF}$ la tasa efectiva cuando los servidores están ocupados (dado que se quiere analizar el comportamiento del sistema cuando existe cola)

El factor de utilización coincide con el porcentaje de tiempo que 1 servidor está ocupado. Para el caso de *s* servidores homogéneos con tasa μ entonces $\overline{\mu}_{EF} = s \cdot \mu$ (más adelante se definirá su cálculo para el caso general), y así:

$\lambda_{EF} = 3$ clientes/hora $\qquad \rho = \frac{3}{2 \cdot 2} < 1$

$\mu = 2$ clientes/hora·servidor

$s = 2$ servidores

Medidas de eficacia de un sistema de colas

N estado del sistema[1], número de clientes en el sistema (cola + servicio)

Hipótesis fundamental: se supone N es un proceso estacionario, es decir, N_t es independiente de t (tasas de llegada y utilización independientes de t, es decir, no hay horas de punta y de valle, son todas unidades de tiempo homogéneas)

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 espera de los clientes en el sistema	
W	tiempo *medio* de espera 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	

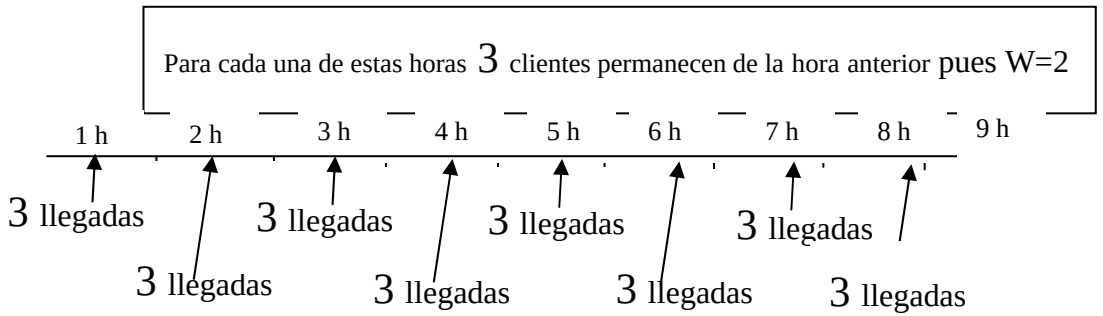
[1]La variable de estado caracteriza indefectiblemente las condiciones en las que el sistema se encuentra.

Fórmulas de Little para relacionar las medidas de eficiencia

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

$$(1)\ L = \lambda_{EF} W \qquad (2)\ L_q = \lambda_{EF} W_q$$

Para entender *(1)*, supóngase W=2 horas, λ_{EF}=3 clientes/hora, entonces el número medio de clientes en el sistema es 3·2=6, tal y como se muestra en la siguiente figura (la hora 1 es despreciable dado que se supone un sistema estacionario y por lo tanto horas homogéneas):



Tiempo medio de los clientes en el sistema = tiempo medio de los clientes en la cola + tiempo medio de servicio de un servidor[2]:

$$(3)\ W = W_q + 1/\mu$$

Nótese que por la definición de esperanza $L = \sum_{n=0}^{\infty} n \cdot p(N = n)$. Si se conoce $p(N = n) = p_n$ se calcula dicha esperanza determinando *W* por *(1)*, seguidamente se saca W_q por *(3)* y finalmente se saca L_q por *(2)*.

Número medio de clientes en el sistema = número medio de clientes en la cola + número medio de servidores ocupados:

$$L = L_q + \lambda_{EF}/\mu$$

[2]Nótese que se divide por la tasa de servicio de cada servidor, en lugar de por la tasa efectiva del sistema. Esto se debe a que las medidas indicadas se refieren al tiempo que pasa un cliente en el sistema, que lógicamente es atendido en un único servidor.

Nótese que para comprender la anterior ecuación, dado que ρ es el porcentaje de tiempo que 1 servidor esté ocupado, el valor $c \cdot \rho = \lambda_{EF}/\mu$ es el porcentaje de tiempo que están ocupados los servidores, o lo que es lo mismo el número de clientes siendo atendidos.

Número medio de servidores ocupados en el sistema = Número medio de clientes en el sistema - número medio de clientes en la cola:

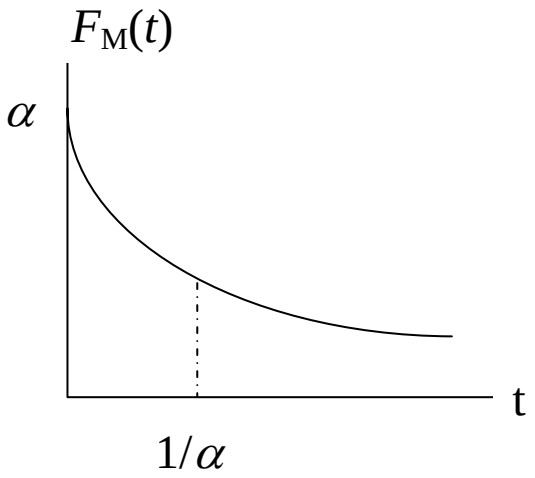
$$\overline{c} = L - L_q = \lambda_{EF}/\mu$$

Distribución exponencial[3]

M variable aleatoria tiempo entre llegadas o tiempo de servicio

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

estrictamente decreciente en t



$$E(M) = 1/\alpha$$

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 $P\{M > t + \Delta t \mid M > \Delta t\} = e^{-\alpha t}$

[3]Transparencias 11 y 12 justifican el uso de la distribución exponencial para modelar el tiempo entre eventos siempre que éstos se produzcan IID.

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.

$S(t)$ 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)

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

Por la falta de memoria de la exponencial, la probabilidad de ocurrencia de un suceso en un intervalo (pequeño) de tiempo de longitud Δt sabiendo que no se ha producido hasta el momento t es directamente proporcional a $\alpha \Delta t$, de hecho $P\{S(t+\Delta t) - S(t) = 1\} = \alpha \cdot \Delta t \cdot e^{-\alpha \cdot \Delta t}$, independiente de t. Esta hipótesis es creíble en sucesos (llegadas/salidas) IID por lo que la Poisson también es creíble para modelar el número de ocurrencias.

Modelo general: tasas dependientes de *N*

Hipótesis:

- Distribución IID de llegadas y salidas *exponencial*.
- Considérese $\lambda_n \;\; n = 0,1,\ldots$ la tasa efectiva de llegada de clientes al sistema dado que hay n clientes $N(t) = n$.
- Considérese $\mu_n \;\; n = 0,1,\ldots$ la tasa efectiva de salida de clientes del sistema dado que hay n clientes $N(t) = n$
- Disciplina *FIFO*

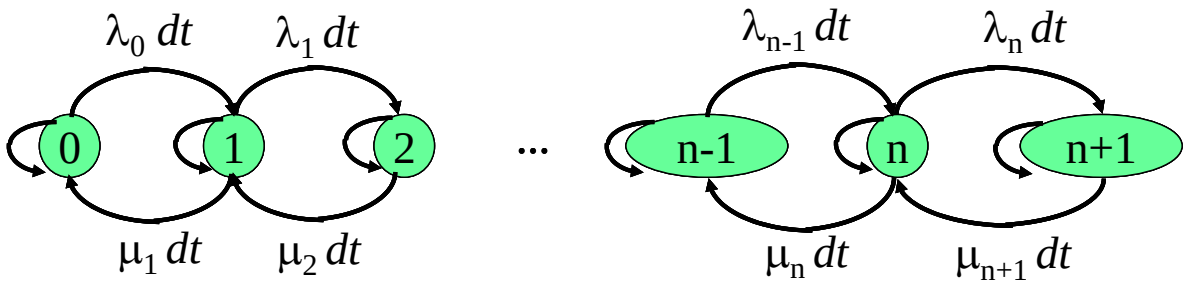
Objetivo (necesario para aplicar fórmulas de Little):

Obtener la probabilidad p_n.

Nótese que con el modelo general:

$$\lambda_{EF} = \sum_{n=0}^{\infty} \lambda_n \cdot p_n ; \quad \mu_{EF} = \sum_{n=0}^{\infty} \mu_n \cdot p_n ; \quad \overline{\mu}_{EF} = \sum_{n=0}^{\infty} \mu_n \cdot p\left(Z = n\right)$$ siendo Z la variable aleatoria condicional $\left\{ N_q / N_q > 0 \right\}$ (dado que para el factor de utilización se supone cola)

Sea *dt* tan pequeño que de un estado dado sólo se puede pasar a dos posibles estados[4] (**diagrama de tasas de transición**).



Por ser proceso de Poisson, la probabilidad de ocurrencia de un suceso en un Δt es proporcional a *dt* (llegada proporcional a $\lambda_n \cdot dt$, salida proporcional a $\mu_n \cdot dt$)

[4]Se asume que la simultaneidad de eventos no puede darse al considerar que su probabilidad es nula (proporcional a dt^2 .

Por ser el sistema estacionario es obvio que se tiene $\frac{d}{dt} p_n(t) = 0$ (siendo $p_n(t) = P(N(t) = n)$).

Como para un sistema estacionario o no se cumple:

$$\frac{d}{dt} P_n(t) = \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t) - (\lambda_n + \mu_n) P_n(t)$$ (ver apuntes)

entonces en concreto para los sistemas estacionarios se tiene:

$$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = \lambda_n P_n + \mu_n P_n$$

(ecuaciones de balance de probabilidades de entrada y salida)

Tasa media de llegada al estado n $\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}$ (ver diagrama de tasas)

Tasa media de salida del estado n $\lambda_n P_n + \mu_n P_n$ (ver diagrama de tasas)

tasa medio de llegada = tasa media de salida

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

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

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

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

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

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

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

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

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

Aplicar fórmulas de Little para calcular el resto de medidas de eficiencia

Notación kendall de un sistema de colas

NOTACION: A/B/s/m/d

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

Para **A y B**: M exponencial, D degenerada (tiempos constantes), E Erlang (Gamma), G general

Si **m** no aparece por defecto infinito.

Si **d** no aparece por defecto FIFO

Ejemplos:

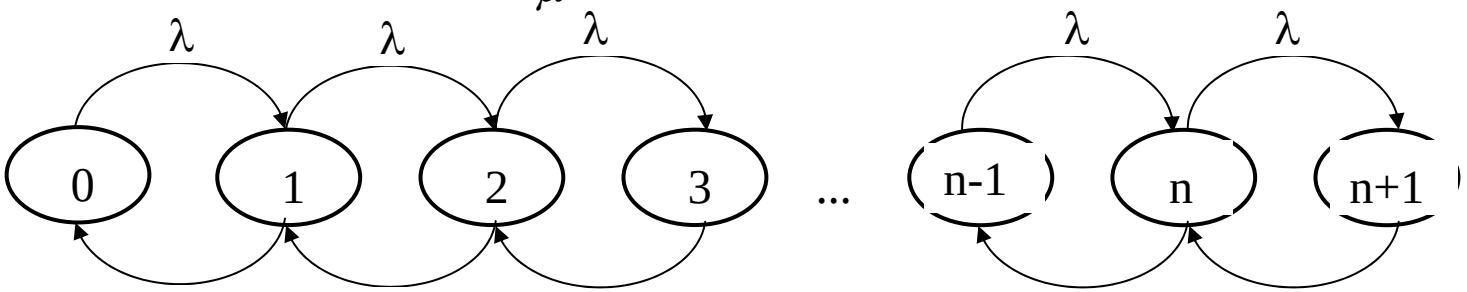
M/M/s	tiempo entre llegadas exponencial / tiempo de servicio exponencial / s servidores
M/M/s/K/FIFO	
M/M/s/s	
M/G/1	

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 $\rho = \frac{\lambda}{\mu}$ Para alcanzar estado estable $\rho < 1$



$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \qquad P_n = \rho^n P_0 \qquad P_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho \qquad P_n = (1-\rho)\rho^n \qquad n = 0,1,2,\ldots$$ [5]

[5] Nótese que $\sum_{k=0}^{m} x^k = \frac{1-x^{m+1}}{1-x}$ y por tanto $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ pero solo cuando $0 \le x < 1$.

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

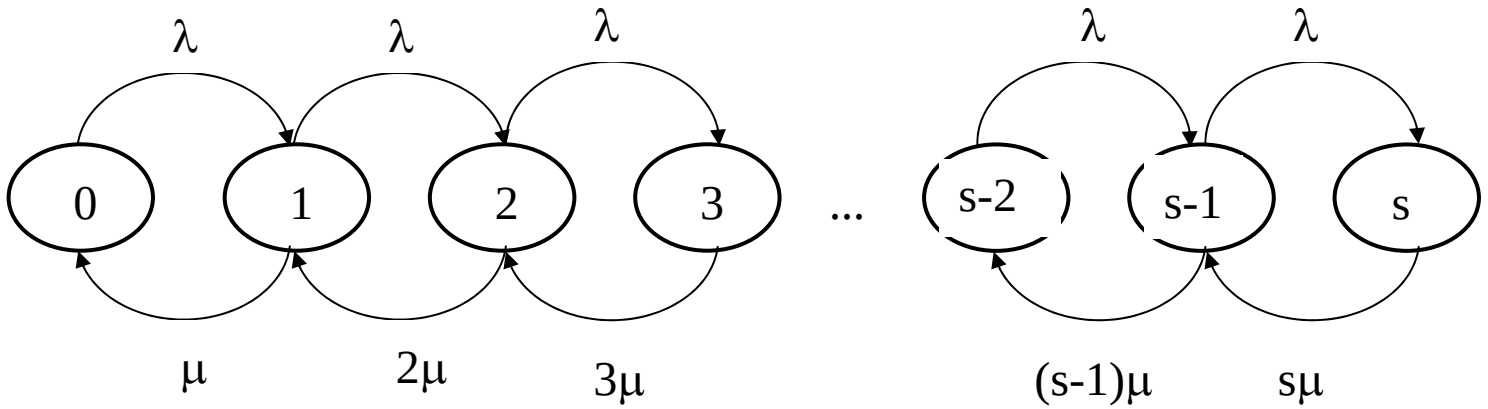
Factor de utilización del servidor $$\overline{c} = L - L_q = 1 - P_0$$

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 $\rho = \dfrac{\lambda}{s\mu}$ Para alcanzar estado estable $\rho < 1$



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

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

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

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

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

Factor de utilización del servidor $\overline{c} = L - L_q = \frac{\lambda}{\mu}$

Caso particular de M/M/s: "Cola" M/M/∞ (s a infinito)

El sistema tiene un número muy grande de servidores (sistemas de autoservicio, visitas a una ciudad: cada visitante se da su servicio y no hay cola).

Tasa de llegadas $\lambda_n = \lambda$

Tasa de servicios $\mu_n = n\mu$ (dado que en este caso μ es el número de veces que el mismo individuo es servido en 1 día)

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

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

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 y salida $\lambda_n = \begin{cases} \lambda & n = 0,1,2,\ldots,K-1 \\ 0 & n \geq K \end{cases}$ $\mu_n = \begin{cases} \mathrm{n}\mu & n \leq s \\ \mathrm{s}\mu & \mathrm{s} \leq n \leq K \end{cases}$

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

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

$$P_0 = \begin{cases} \dfrac{1}{\sum_{n=0}^{s} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \sum_{n=s+1}^{K} \left(\frac{\lambda}{c\mu}\right)^{n-s}} & \rho = 1 \\ \dfrac{1}{\sum_{n=0}^{s} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s (k-s+1)} & \rho \neq 1 \end{cases}$$

$$L_q = \begin{cases} \dfrac{\left(\dfrac{\lambda}{\mu}\right)^s (k-s)(k-s+1)}{2s!} p_0 & \rho = 1 \\ p_0 \dfrac{\left(\dfrac{\lambda}{\mu}\right)^s \dfrac{\lambda}{s\mu}}{s!\left(1-\dfrac{\lambda}{s\mu}\right)^2}\left[1-\left(\dfrac{\lambda}{s\mu}\right)^{k-s+1} - (k-s+1)\left(1-\dfrac{\lambda}{s\mu}\right)\left(\dfrac{\lambda}{s\mu}\right)^{k-s}\right] & \rho \neq 1 \end{cases}$$

Tasa media de llegada (entrada efectiva) $\lambda_{EF} = \sum_{n=0}^{\infty} \lambda_n \cdot p_n = \lambda \cdot \sum_{n=0}^{K-1} p_n = \lambda(1 - P_K)$

Número medio de clientes en el sistema $L = L_q + \dfrac{\lambda_{EF}}{\mu}$

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

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

Factor de utilización del servidor $\overline{c} = L - L_q = \dfrac{\lambda_{EF}}{\mu}$

Caso particular de M/M/s/K: “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) $P_s = \dfrac{\left(\dfrac{\lambda}{\mu}\right)^s / s!}{\sum_{i=0}^{s}\left(\dfrac{\lambda}{\mu}\right)^i / i!}$

Sistema cerrado con cola M/M/1

Fuente finita de tamaño m. Clientes una vez servidos vuelven a la fuente.
En este caso (sistema cerrado) λ es la tasa de retorno de UN cliente, NO la tasa de llegadas de los clientes al sistema. La tasa de retorno es entonces el número de servicios solicitados por unidad de tiempo y por UN cliente

La tasa de llegada al sistema es entonces $\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$

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 $\rho = \frac{\lambda}{\mu}$ que es simple notación, no el factor de utilización.

Tasa media de llegada al sistema $\lambda_{EF} = \sum_{n=0}^{\infty} \lambda_n \cdot p_n = \sum_{n=1}^{m} (m-n)\lambda \cdot p_n = (m-L)\lambda$

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 $W = \frac{L}{(m-L)\lambda}$

Factor de utilización del servidor $\overline{c} = L - L_q = 1 - p_0$

Sistema cerrado con cola M/M/s

Fuente finita de tamaño m. Clientes una vez servidos vuelven a la fuente.
λ es la tasa de retorno de un cliente

La tasa de llegada al sistema es entonces $\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$

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

Probabilidad de cada estado:

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

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

Número medio de clientes en cola $L_q = \sum_{n=0}^{m} (m - n)\, p_n$ (no existe expresión analítica)

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

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

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

Factor de utilización del servidor $\overline{c} = L - L_q = \frac{\lambda_{EF}}{\mu}$

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 $\rho = \frac{\lambda}{\mu}$ que tendrá que ser <1 para que el sistema sea estacionario

Diseño óptimo de los sistemas de colas (objetivo práctico)

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.

Coste de los clientes:

- Pérdidas de ganancia/productividad por pérdida de clientes

Decisiones:

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

Optimizar la tasa de servicio

λ conocida y fija
C_μ coste por unidad de tasa de servicio por unidad de tiempo
C_c coste por cliente en el sistema por unidad de tiempo

$$\min CT(\mu) = C_\mu \cdot \mu + C_c \cdot L(\mu)$$

Para cola M/M/1

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

$$\frac{\partial CT(\mu)}{\partial \mu} = 0 \quad \Rightarrow \quad \mu = \lambda + \sqrt{\frac{C_c \lambda}{C_\mu}}$$

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

$$\min CT(\mu, K) = C_\mu \cdot \mu + C_c \cdot L(\mu, K) + C_K \cdot K + C_p \cdot \lambda \cdot P_K \qquad K \in N$$

Optimizar el número de servidores

μ, λ conocidos y fijos

C_s coste por servidor por unidad de tiempo

$$\min CT(s) = C_s \cdot s + C_c \cdot L(s) \qquad s \in N$$

En el óptimo se tiene que cumplir que $CT(s-1) \geq CT(s) \leq CT(s+1)$

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