UNIVERSIDAD PONTIFICIA
ICAI ICADE
COMILLAS
MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
DEPARTAMENTO DE ORGANIZACIÓN INDUSTRIAL

Teoría de Colas

José María Ferrer Caja
Universidad Pontificia Comillas

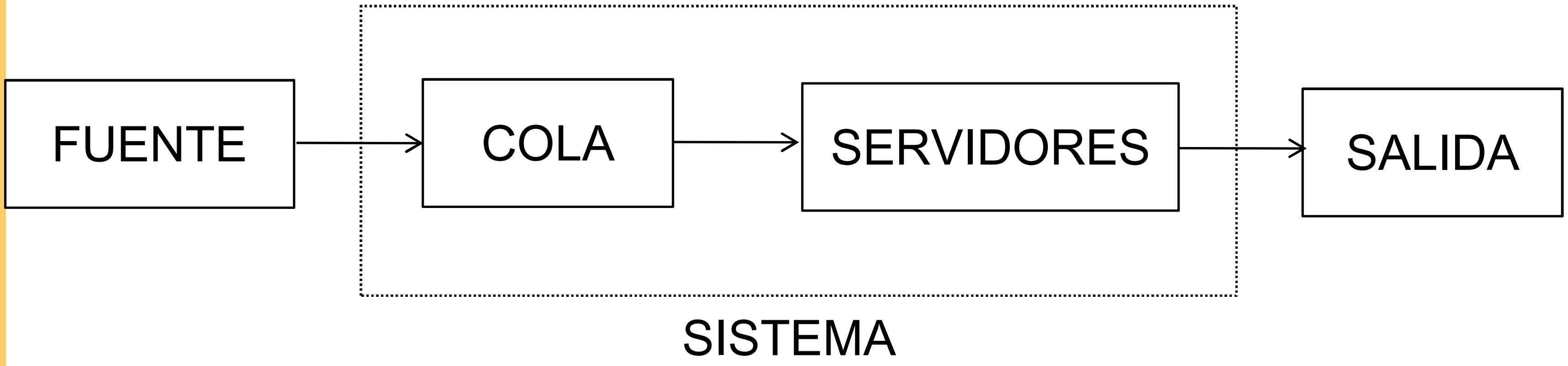
Introducción

- Cola: Conjunto de clientes en espera de recibir un servicio
- Se produce cuando los clientes llegan a un servidor ocupado y permanecen en espera
- Teoría de Colas: Análisis del comportamiento de un sistema de colas a lo largo del tiempo

- Herramientas:
  - ✓ Teoría de Probabilidades
  - ✓ Optimización
  - ✓ Simulación

Elementos

- Fuente: Origen de los clientes
- Cola
- Centro de servicio: Conjunto de servidores
- Sistema: Cola(s)+Servidor(es)
- Salida: Destino de los clientes



Características: Llegadas

- Fuente
  - ✓ Finita → Sistema cerrado
  - ✓ Infinita → Sistema abierto
- Número de fuentes
  - ✓ Una
  - ✓ Varias
- Forma de llegada
  - ✓ Unitaria
  - ✓ En bloques
- Tiempo entre llegadas
  - ✓ Determinista
  - ✓ Aleatorio (llegadas independientes)
  - ✓ Aleatorio (llegadas dependientes)

Características: Cola

- Número de canales
- Interferencia entre canales
  - ✓ Posibilidad de cambiar de canal
  - ✓ Imposibilidad de cambiar de canal
- Capacidad del sistema
  - ✓ Finita
  - ✓ Infinita
- Disciplina
  - ✓ FIFO: El primero que entra es el primero en ser atendido
  - ✓ FIFO con límite: El tiempo de servicio es limitado
  - ✓ LIFO: El último que entra es el primero en ser atendido
  - ✓ SIRO: Los clientes se seleccionan de forma aleatoria
  - ✓ PRI: Se atiende antes a los clientes prioritarios

Características: Mecanismo de servicio

- Número de servidores
- Relación entre servidores
  - ✓ Servidores independientes
  - ✓ Servidores dependientes
- Homogeneidad entre servidores
  - ✓ Servidores homogéneos
  - ✓ Servidores heterogéneos
- Tiempo de servicio
  - ✓ Determinista
  - ✓ Aleatorio (tiempos independientes)
  - ✓ Aleatorio (tiempos dependientes)

Características: Comportamiento del cliente

- Comportamiento al encontrar el servidor ocupado
  - ✓ Entra al sistema y permanece en la cola
  - ✓ Reintenta la entrada tras un periodo de tiempo
  - ✓ Renuncia al servicio
- Selección del canal
  - ✓ Selección aleatoria
  - ✓ Selección bajo algún criterio
  - ✓ Adjudicación automática de canal
- Comportamiento tras un periodo de tiempo en la cola
  - ✓ Se mantiene en el canal
  - ✓ Cambia de canal
  - ✓ Renuncia al servicio

Parámetros

- Tasa de llegadas → λ: Número medio de clientes que llegan al sistema por unidad de tiempo
- Tiempo medio entre llegadas → $1/\lambda$
- Tasa de entradas → λ_{ef}: Número medio de clientes que entran al sistema por unidad de tiempo
- Tasa de servicio → μ: Número medio de clientes que son atendidos por un servidor por unidad de tiempo
- Tiempo medio de servicio → $1/\mu$
- Tasa de servicio del sistema → μ_{ef}: Número medio de clientes que son atendidos por unidad de tiempo
- Número de servidores → s
- Capacidad del sistema → k
- Factor de utilización, intensidad de tráfico → $\rho = \lambda_{ef} / \mu_{ef}$

Estado estacionario: Condición

- $N(t)$: Número de clientes en el sistema en el instante t
- El sistema puede estabilizarse tras un periodo de tiempo → Estado estacionario
- Distribución estacionaria: Distribución de probabilidad de $N(t)$ en el estado estacionario $N(t) \rightarrow N$
- Condición para que exista distribución estacionaria:

$$\rho<1$$

(Tasa de entrada<Tasa de servicio del sistema)

Estado estacionario: Variables

- N: Número de clientes en el sistema
- p_n=P(N=n): Probabilidad de que haya n clientes en el sistema
- N_q: Número de clientes en la cola
- T: Tiempo de un cliente en el sistema
- T_q: Tiempo de un cliente en la cola

Estado estacionario: Medidas de eficiencia

- L: Número medio de clientes en el sistema $L = \mathrm{E}[N]$
- L_q: Número medio de clientes en la cola $L_q = \mathrm{E}[N_q]$
- W: Tiempo medio de clientes en el sistema $W = \mathrm{E}[T]$
- W_q: Tiempo medio de clientes en la cola $W_q = \mathrm{E}[T_q]$
- $\overline{c}$: Número medio de servidores ocupados
- t_c: Tiempo medio de servidores desocupados

Estado estacionario: Relaciones

- Fórmulas de Little

$$L = \lambda_{ef} W$$

$$L_q = \lambda_{ef} W_q$$

- Otras relaciones

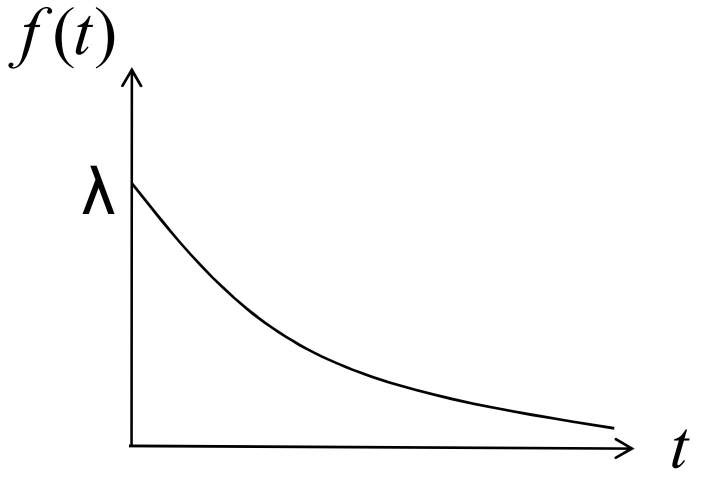
$$W = W_q + \frac{1}{\mu}$$

$$L = L_q + \frac{\lambda_{ef}}{\mu}$$

$$\bar{c} = L - L_q = \frac{\lambda_{ef}}{\mu}$$

Distribución exponencial

$$T \approx \exp(\lambda)$$



- Función de densidad:

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

- Media, varianza, probabilidad:

$$\mathrm{E}[T]=\frac{1}{\lambda} \qquad \mathrm{V}[T]=\frac{1}{\lambda^2} \qquad t>0 \Rightarrow \mathrm{P}(T>t)=e^{-\lambda t}$$

- Falta de memoria:

$$\mathrm{P}\left(T>t_0+t \,/\, T>t_0\right)=\mathrm{P}(T>t)$$

Distribución de Poisson

- N≡Número de éxitos/unidad de tiempo → $N \approx \text{Poisson}(\lambda)$
- Función de probabilidad

$$P(N=n)=\frac{e^{-\lambda}\lambda^n}{n!} \quad n\in\{0,1,2,...\}$$

- Media, varianza:

$$E[N]=\lambda \qquad\qquad V[N]=\lambda$$

- Reproductividad: La suma de variables independientes de Poisson es otra variable de Poisson
- Proporcionalidad: El número medio de éxitos es proporcional al tiempo
- Relación con la exponencial: El número de éxitos sigue una distribución de Poisson ⇔ el tiempo entre dos éxitos consecutivos sigue una distribución exponencial

Proceso poissoniano

- Proceso estocástico: Colección de variables aleatorias $\{N(t)\}$
- Proceso de conteo:
  - ✓ $N(t) \in \{0, 1, 2, \ldots\}$
  - ✓ $s \leq t \Rightarrow N(s) \leq N(t)$
- Proceso markoviano:
  - ✓ Incrementos estacionarios:

    $N(t) - N(s)$ sólo depende *de* $t\text{-}s$
  - ✓ Incrementos independientes
- $N(0) = 0$ c.s.
- $N(t) - N(s) \approx \text{Poisson}(\lambda(t\text{-}s)),\ N(t) \approx \text{Poisson}(\lambda t)$
- En un intervalo cuya longitud tiende a 0, la probabilidad de que ocurra más de un éxito tiende a 0

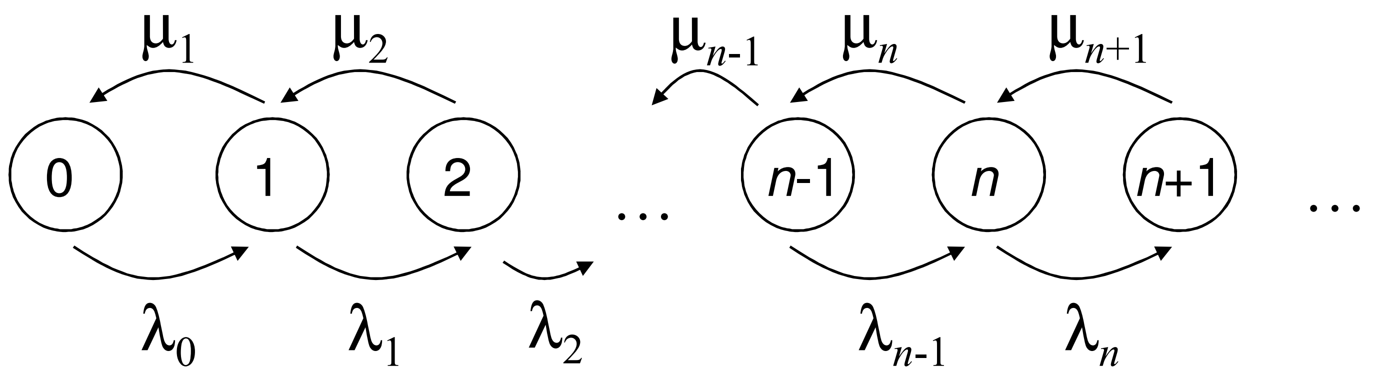
Proceso de nacimiento y muerte: Planteamiento

- Nacimiento: Entrada de un cliente
- Muerte: Salida de un cliente una vez servido
- El tiempo entre llegadas es exponencial
- El tiempo entre salidas es exponencial, e independiente del tiempo entre nacimientos
- $N(t)$: Número de clientes en el sistema en el instante t
- λ_n: Tasa de entradas si hay n clientes en el sistema
- μ_n: Tasa de salidas si hay n clientes en el sistema
- Objetivo → Obtener la distribución estacionaria N

  $p_n=\mathrm{P}(N=n)$: Probabilidad de que haya n clientes en el sistema, en el estado estacionario, $n = 0, 1, 2,\ldots$

Proceso de nacimiento y muerte: Diagrama de transiciones

- Los procesos que rigen el número de llegadas y el número de salidas son poissonianos
- De cada estado n sólo es posible pasar a dos estados:
  - ✓ $n+1$ si se produce una llegada
  - ✓ $n-1$ si se produce una salida



Proceso de nacimiento y muerte: Estado estacionario

Supuesto alcanzado el estado estacionario:

- 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: $\lambda_n\, p_n + \mu_n\, p_n$
- Tasa media de llegada al estado n = Tasa media de salida del estado n:

$$\lambda_{n-1}\, p_{n-1} + \mu_{n+1}\, p_{n+1} = \lambda_n\, p_n + \mu_n\, p_n$$

- Para el estado $n = 0$: $\mu_1\, p_1 = \lambda_0\, p_0$
- A partir de estas expresiones podemos obtener cada p_n en función de p_0:

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \Rightarrow p_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 \Rightarrow p_n = \frac{\lambda_{n-1} \ldots \lambda_1 \lambda_0}{\mu_n \ldots \mu_2 \mu_1} p_0$$

- Para obtener p_0 basta usar que $\sum_{n=0}^{\infty} p_n = 1$

Proceso de nacimiento y muerte: Medidas de eficiencia

- Número medio de clientes en el sistema

$$L = \sum_{n=0}^{\infty} np_n$$

- Número medio de clientes en cola (para un sistema con *s* servidores)

$$L_q = \sum_{n=s}^{\infty} (n - s) p_n$$

- Tasa media de llegadas

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

- Tiempo medio de permanencia en el sistema

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

- Tiempo medio de permanencia en cola

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

Modelos clásicos: Notación de Kendall

- Especificaciones del modelo

  A / B / c / m / d

- A: Distribución tiempo entre llegadas. Puede ser
  - ✓ M → Exponencial
  - ✓ D → Constante
  - ✓ E_k → Erlang de parámetro k
  - ✓ G → Genérica
- B: Distribución tiempo de servicio. Puede ser
  - ✓ M, D, E_k, G
- c: Número de servidores
- m: capacidad del sistema
- d: Disciplina de la cola

Modelos clásicos: Hipótesis generales

- Una única fuente de tamaño infinito
- No hay impaciencia: Todo cliente que llega al sistema, entra (a no ser que se haya alcanzado la capacidad máxima), y una vez dentro no lo abandona hasta haber sido servido.
- Un único canal para todos los servidores
- Servidores independientes y homogéneos
- Independencia entre llegadas y servicios
- Por defecto, se supone capacidad infinita del sistema y disciplina FIFO

Modelos clásicos: M/M/1

- Hipótesis:
  - ✓ Tiempos entre llegadas independientes, distribuidos según una exponencial de parámetro λ
  - ✓ Tiempos de servicio independientes, distribuidos según una exponencial de parámetro μ
  - ✓ Un único servidor: $s=1$
  - ✓ Capacidad ilimitada
  - ✓ Disciplina FIFO
- Factor de utilización: $\rho = \dfrac{\lambda}{\mu} \Rightarrow p_n = \rho^n p_0 \Rightarrow p_0 \sum_{n=0}^{\infty} \rho^n = 1$
- Estado estacionario $\Leftrightarrow \rho < 1 \Leftrightarrow \sum_{n=0}^{\infty} \rho^n = \dfrac{1}{1-\rho} \Leftrightarrow p_0 = 1-\rho$
- Distribución estacionaria: $p_n = \rho^n (1-\rho),\ n = 1,2,3,\ldots$

Modelos clásicos: M/M/1

- Medidas de eficiencia :

$$L = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

$$L_q = \sum_{n=1}^{\infty} (n-1)p_n = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

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

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

- Distribución del tiempo en el sistema y en cola:

$$T \approx \exp(\mu-\lambda) \approx \exp(\mu(1-\rho))$$

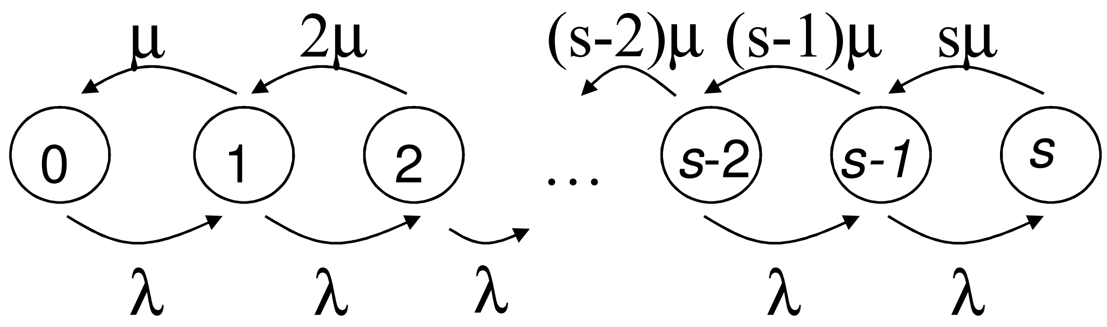
T_q : Distribución mixta

$$P(T_q = 0) = p_0 = 1-\rho$$

$$t > 0 \Rightarrow f(t) = \rho(1-\rho)\mu\, e^{-(1-\rho)\mu t}$$

Modelos clásicos: M/M/s

- Hipótesis:
  - ✓ Tiempos entre llegadas independientes, distribuidos según una exponencial de parámetro λ
  - ✓ s servidores independientes y homogéneos
  - ✓ Tiempos de servicio independientes para cada servidor, distribuidos según una exponencial de parámetro μ
  - ✓ Capacidad ilimitada
  - ✓ Disciplina FIFO
- Tasas: $\lambda_n = \lambda \qquad \mu_n = \begin{cases} n\mu & n \le s \\ s\mu & n > s \end{cases} \qquad \rho = \dfrac{\lambda}{s\mu}$
- Diagrama de transiciones:



Modelos clásicos: M/M/s

- Distribución estacionaria $\Leftrightarrow \rho < 1$

$$p_0 = \frac{1}{\dfrac{(s\rho)^s}{s!(1-\rho)} + \displaystyle\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!}} \qquad p_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & 1 \le n \le s \\ \dfrac{1}{s!\,s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & \mathrm{n} \ge \mathrm{s} \end{cases}$$

- Medidas de eficiencia:

$$L = \frac{(s\rho)^s \rho}{s!(1-\rho)^2} p_0 + s\rho$$

$$L_q = L - s\rho = \frac{(s\rho)^s \rho}{s!(1-\rho)^2} p_0$$

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

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

Modelos clásicos: M/M/s

Probabilidades

- ✓ Probabilidad de no hacer cola

$$P(T_q = 0) = \sum_{n=0}^{s-1} p_n = 1 - \frac{(s\rho)^s}{s!(1-\rho)} p_0$$

- ✓ Probabilidad de permanecer en cola un tiempo mayor que t

$$P(T_q > t) = \frac{(s\rho)^s}{s!(1-\rho)} p_0 \, e^{-s\mu(1-\rho)t}$$

Modelos clásicos: M/M/∞

- Hipótesis:
  - ✓ Caso particular del modelo M/M/s para un número ilimitado de servidores
  - ✓ No hay cola, cada cliente que llega es servido directamente
- Distribución estacionaria (existe siempre):

$$N \approx \text{Poisson}(\lambda/\mu)$$

- Medidas de eficiencia:

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

$$L_q = 0$$

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

$$W_q = 0$$

Modelos clásicos: M/M/s/k

- **Hipótesis**:
  - ✓ Tiempos entre llegadas independientes, distribuidos según una exponencial de parámetro λ
  - ✓ s servidores independientes y homogéneos
  - ✓ Tiempos de servicio independientes para cada servidor, distribuidos según una exponencial de parámetro μ
  - ✓ Capacidad limitada a k clientes, $k \geq s$
  - ✓ Disciplina FIFO
- Distribución estacionaria (existe siempre). Para $\rho = \dfrac{\lambda}{s\mu} \neq 1$

$$p_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & 1 \leq n \leq s \\ \dfrac{1}{s!\,s^{n-s}}\left(\dfrac{\lambda}{\mu}\right)^n p_0 & s \leq n \leq k \ldots \end{cases} \qquad \text{con } p_0 \text{ tal que } \sum_{n=0}^{k} p_n = 1$$

Modelos clásicos: M/M/s/k

- Tasa de entradas:

$$\lambda_{ef} = \lambda(1 - p_k)$$

- Medidas de eficiencia :

$$L_q = p_0 \frac{(s\rho)^s \rho}{s!(1-\rho)^2}\left[1 - \rho^{k-s+1} - (k-s+1)(1-\rho)\rho^{k-s}\right]$$

$$W_q = \frac{L_q}{\lambda_{ef}}$$

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

$$L = W\lambda_{ef}$$

Modelos clásicos: M/M/s/s

- Hipótesis:
  - ✓ Caso particular del modelo M/M/s/k cuando la capacidad del sistema coincide con el número de servidores
  - ✓ No hay cola, cuando un cliente llega o es servido directamente o no puede entrar en el sistema
- Probabilidad de que el sistema esté saturado

$$p_s = \frac{(s\rho)^s / s!}{\sum_{n=0}^{s} (s\rho)^n / n!}$$

Modelos clásicos: M/G/1

- Hipótesis:
  - ✓ Tiempos entre llegadas independientes, distribuidos según una exponencial de parámetro λ
  - ✓ Tiempos de servicio independientes, distribuidos según una distribución general F de media $1/\mu$ y varianza σ^2
  - ✓ Un único servidor: $s=1$
  - ✓ Capacidad ilimitada
  - ✓ Disciplina FIFO
- Factor de utilización: $\rho = \dfrac{\lambda}{\mu}$
- Estado estacionario $\Leftrightarrow$ $\rho < 1$
- Fórmula de Pollaczek-Khintchine:

$$L = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)}$$

Modelos clásicos: M/M/1 cerrado

- Hipótesis:
  - ✓ Fuente finita de m clientes
  - ✓ Tiempos de retorno al sistema distribuidos según una exponencial de parámetro λ
  - ✓ Tiempos de servicio independientes, distribuidos según una exponencial de parámetro μ
  - ✓ Un único servidor: $s=1$
  - ✓ Capacidad ilimitada
  - ✓ Disciplina FIFO
- Tasas:
  - ✓ Tasa de retornos λ
  - ✓ Tasa de llegadas

$$\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$$

Modelos clásicos: M/M/1 cerrado

- Distribución estacionaria (existe siempre)

$$p_0 = \left[1 + \sum_{n=1}^{m} \frac{m!\rho^n}{(m-n)!} \right]^{-1}$$

$$p_n = \frac{m!\rho^n}{(m-n)!} p_0 \quad 0 < n \leq m \quad \text{siendo} \quad \rho = \frac{\lambda}{\mu}$$

- Medidas de eficiencia

$$L = m - \frac{1-p_0}{\rho} \qquad L_q = m - \frac{1+\rho}{\rho}(1-p_0)$$

$$W = \frac{L}{(m-L)\lambda} \qquad W_q = \frac{L_q}{(m-L)\lambda} = \frac{1}{\mu}\left[\frac{m}{1-p_0} - \frac{1+\rho}{\rho} \right]$$

- Tasa media de llegadas

$$\lambda_{ef} = (m-L)\lambda$$

Modelos clásicos: M/M/s cerrado

- Hipótesis:
  - ✓ Fuente finita de m clientes
  - ✓ Tiempos de retorno al sistema distribuidos según una exponencial de parámetro λ
  - ✓ s servidores independientes y homogéneos
  - ✓ Tiempos de servicio independientes para cada servidor, distribuidos según una exponencial de parámetro μ
  - ✓ Capacidad ilimitada
  - ✓ Disciplina FIFO
- Tasas:

$$\lambda_n = \begin{cases} (m-n)\lambda & n < m \\ 0 & n \geq m \end{cases}$$

$$\mu_n = \begin{cases} n\mu & 0 \leq n \leq s \\ s\mu & s \leq n \leq m \\ 0 & n > m \end{cases}$$

Modelos clásicos: M/M/s cerrado

- Distribución estacionaria (existe siempre)

$$p_n = \begin{cases} \binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n p_0 & 1 \le n \le s \\ \binom{m}{n}\frac{n!}{s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n p_0 & s \le n \le m \end{cases} \qquad \text{con } p_0 \text{ tal que } \sum_{n=0}^{m} p_n = 1$$

- Medidas de eficiencia: No existen fórmulas sencillas. Se obtiene L a partir de la definición, y el resto mediante las fórmulas de Little
- Tasa media de llegadas

$$\lambda_{ef} = (m - L)\lambda$$

Decisión en los sistemas de colas

- Objetivo: Determinar el nivel de servicio que minimiza el coste total del sistema
- Coste total = Coste servicio + Coste clientes
- Coste servicio: costes por mantener operativo el servicio: Aumenta con la tasa de servicio y con el número de servidores
- Coste clientes
  - ✓ Costes por permanecer en cola
  - ✓ Costes por pérdida de clientes
  - ✓ Costes por dar servicio
- Ambos costes están en conflicto

Optimización de la tasa de servicio

- Costes por unidad de tiempo
  - ✓ C_1 = coste por unidad de μ
  - ✓ $C_1\mu$ = coste por tener una tasa de servicio μ
  - ✓ C_2 = coste por mantener un cliente en el sistema
  - ✓ $C_2\, L(\mu)$ = coste total esperado por mantener los clientes en el sistema
- Función de coste esperado por unidad de tiempo

$$CT(\mu) = C_1\mu + C_2 L(\mu)$$

- Problema a resolver:

$$\min_{\mu} C_1\mu + C_2 L(\mu)$$

Optimización de la tasa de servicio y la capacidad del sistema

- Costes por unidad de tiempo
  - ✓ C_1 y C_2 igual que en el caso anterior
  - ✓ C_3 = coste por unidad de capacidad
  - ✓ $C_3 k$ = coste por tener una capacidad k
  - ✓ C_4 = coste por cada cliente perdido
  - ✓ $C_4 \lambda p_k$ = coste total esperado por clientes perdidos
- Función de coste esperado por unidad de tiempo

$$CT(\mu, k) = C_1\mu + C_2 L(\mu) + C_3 k + C_4 \lambda p_k$$

- Problema a resolver:

$$\min_{\mu, k} C_1\mu + C_2 L(\mu) + C_3 k + C_4 \lambda p_k \qquad k \in \mathbb{N}$$

Optimización del número de servidores

- Costes por unidad de tiempo
  - ✓ C_2 = coste por mantener un cliente en el sistema
  - ✓ $C_2\, L(s)$ = coste total esperado por mantener los clientes en el sistema
  - ✓ C_5 = coste por servidor
  - ✓ $C_5 s$ = coste por tener s servidores
- Función de coste esperado por unidad de tiempo

$$CT(s) = C_5 s + C_2 L(s)$$

- Problema a resolver:

$$\min_s C_5 s + C_2 L(s) \quad s \in \mathbb{N}$$