

MODELADO ANALÍTICO

1. INTRODUCCIÓN

El modelado analítico construye una representación del sistema real mediante técnicas que son resolubles matemáticamente, dando lugar a un conjunto de ecuaciones.

Las ecuaciones pueden tener solución $\left\{ \begin{array}{l} \text{Exacta} \\ \text{Aproximada} \end{array} \right\}$ Dependerá del método resolución

Ventajas:

- Rápido
- Sencillo

Inconvenientes:

- Baja precisión (Puede necesitar muchas simplificaciones)

Algunos métodos de modelado empleados:

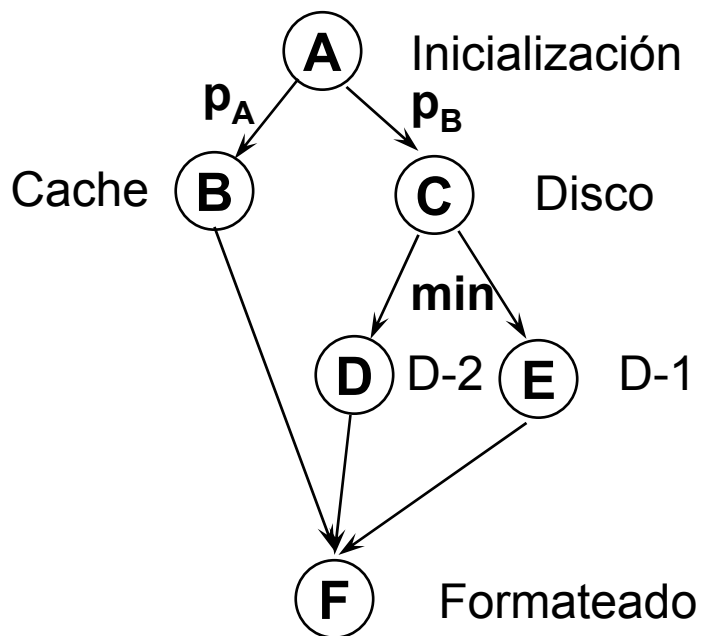
- Grafos
- Modelos de Markov
- Teoría de Colas
- Redes de Petri

MODELADO ANALÍTICO

Grafos

Se emplean para estudiar el comportamiento de programas o procesos. Permiten analizar casos en los que existe concurrencia y/o saltos probabilísticos.

Búsqueda en una base de datos



Dentro de todo el conjunto de grafos, existe un subconjunto que puede analizarse matemáticamente de forma sencilla.

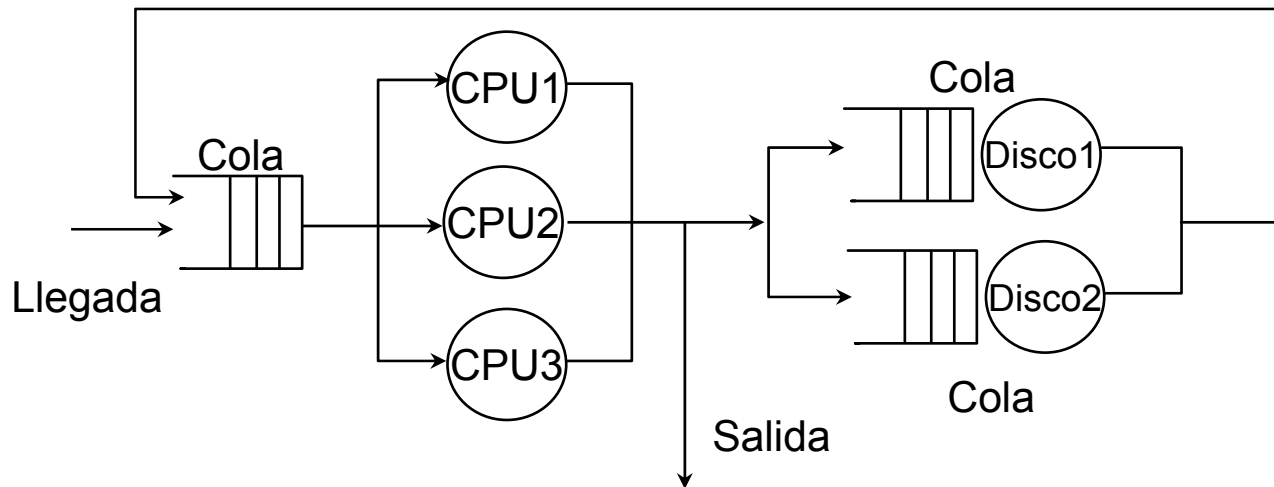
Son útiles para examinar prestaciones desde el punto de vista de programa. No considera la competencia por recursos.

MODELADO ANALÍTICO

Teoría de Colas

La teoría de colas es útil para estudiar las prestaciones globales, a nivel de sistema. Se considera la competencia por recursos. Cada recurso se representa mediante una cola de algún tipo.

Sistema multiprocesador



Existe resolución analítica exacta para algunos tipos de sistemas de colas y aproximadas para algunos más.

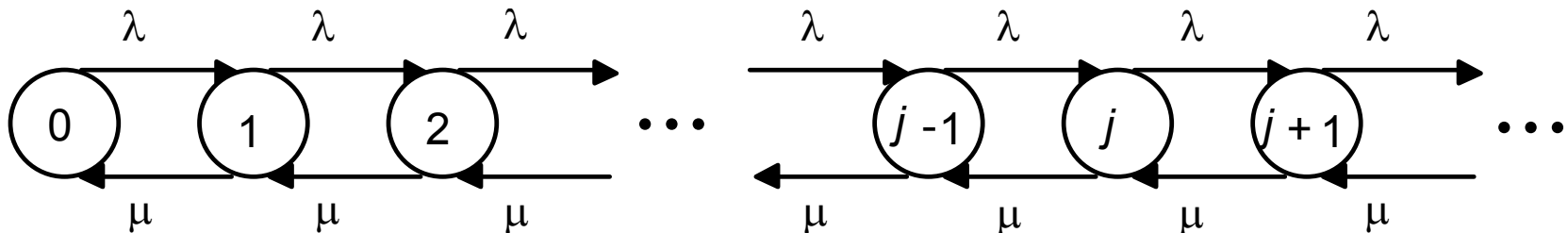
MODELADO ANALÍTICO

Modelos de Markov

Este método permite una representación más complicada de interacción entre componentes.

Cada nodo del modelo de Markov representa un posible estado del sistema, el estado futuro depende sólo del estado presente y no del pasado.

Cola simple como modelo de Markov



Este es un modelo de Markov sencillo, cada estado sólo puede cambiar en un solo paso. Los modelos pueden ser más complicados.

Los modelos de Markov pueden representarse mediante ecuaciones.

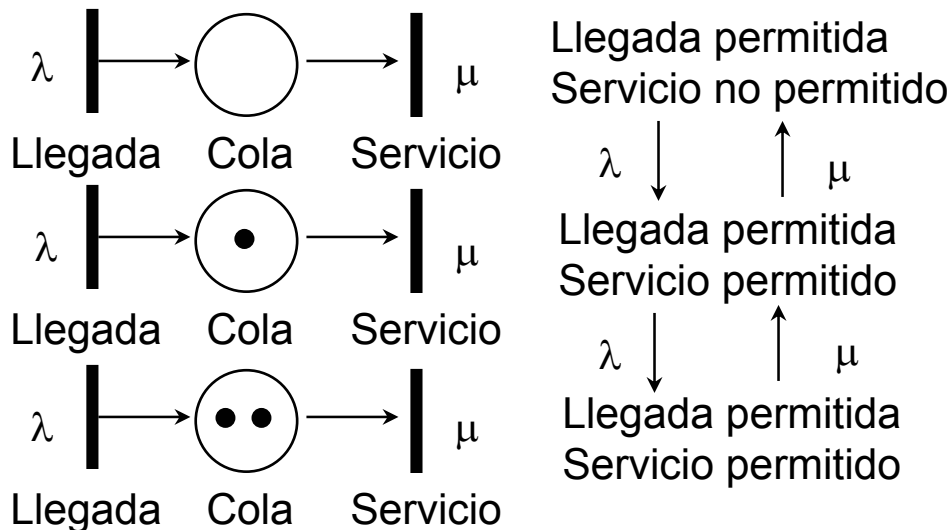
MODELADO ANALÍTICO

Redes de Petri

Cuando los estados que toma el sistema son elevados o la complejidad es alta, se utilizan las redes de Petri.

Las redes de Petri están formadas por *places*, transiciones, arcos y *tokens*. Los *tokens* residen en los *places* y se mueven de un *place* a otro según indiquen los arcos a través de las transiciones.

Cola simple como red de Petri



Para este tipo de modelos existen herramientas software que permiten analizar y resolver las redes de Petri.

2. INTRODUCCIÓN A LA TEORÍA DE COLAS

Se elige como método de modelado analítico las redes de colas porque:

- Son adecuadas para estudiar prestaciones a nivel global de un sistema computador.
- En un sistema computador, las tareas comparten recursos. Cuando una tarea usa el recurso, las demás esperan en una cola

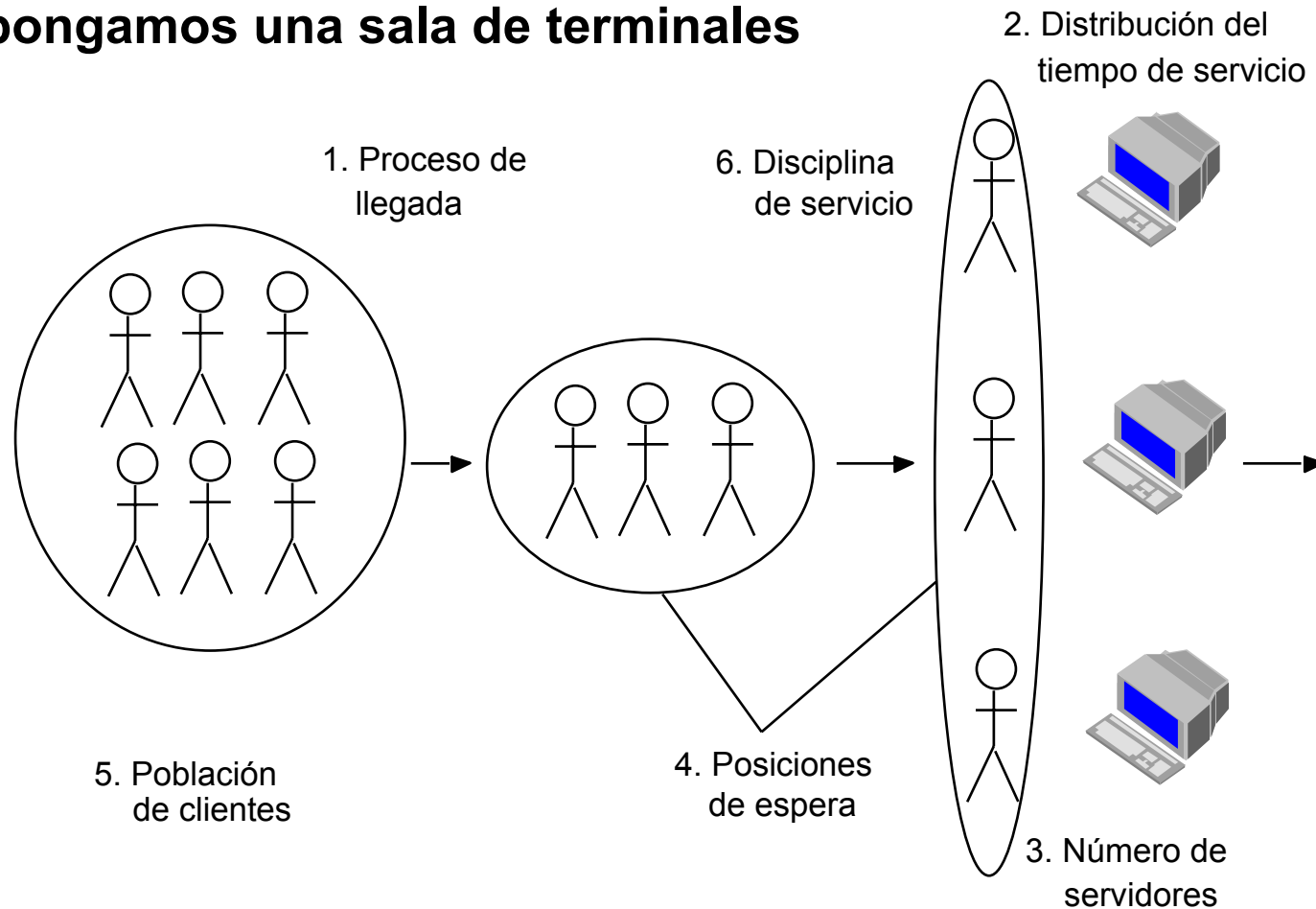
Se verá en este apartado:

- Notación en las colas
- Parámetros
- Propiedades de los procesos estocásticos

MODELADO ANALÍTICO

Notación en las colas

Supongamos una sala de terminales



MODELADO ANALÍTICO

1. *Proceso de llegada.*- Es el tiempo que transcurre entre dos llegadas consecutivas. Se considera una variable aleatoria independiente e idénticamente distribuida (IID).
2. *Distribución del tiempo de servicio.*- Es el tiempo que cada usuario pasa en la terminal. Se considera una variable aleatoria independiente e idénticamente distribuida (IID).
3. *Número de servidores.*- Es el número de terminales idénticas disponibles. Diferentes clases de terminales se representarían mediante colas diferentes.
4. *Capacidad del sistema.*- Número máximo de usuarios que pueden estar en el sistema. Incluye tanto a los usuarios que están esperando en la cola como a los que están recibiendo servicio.

MODELADO ANALÍTICO

5. *Tamaño de la población.*- Es el número total de potenciales usuarios del sistema. La población suele ser finita en la mayoría de los sistemas, aunque en la práctica por simplicidad de cálculo se considera infinita.

6. *Disciplina de servicio.*- Es el orden en el que se sirven los usuarios cuando esperan en la cola.

Existen unos elementos que se conocen como servidores infinitos (IS), también llamados centros de retardo (delay centers).

En estos centros se supone que el usuario no tiene que esperar en la cola para recibir el servicio, por tanto todo el tiempo que pasa en el centro de servicio es tiempo recibiendo servicio.

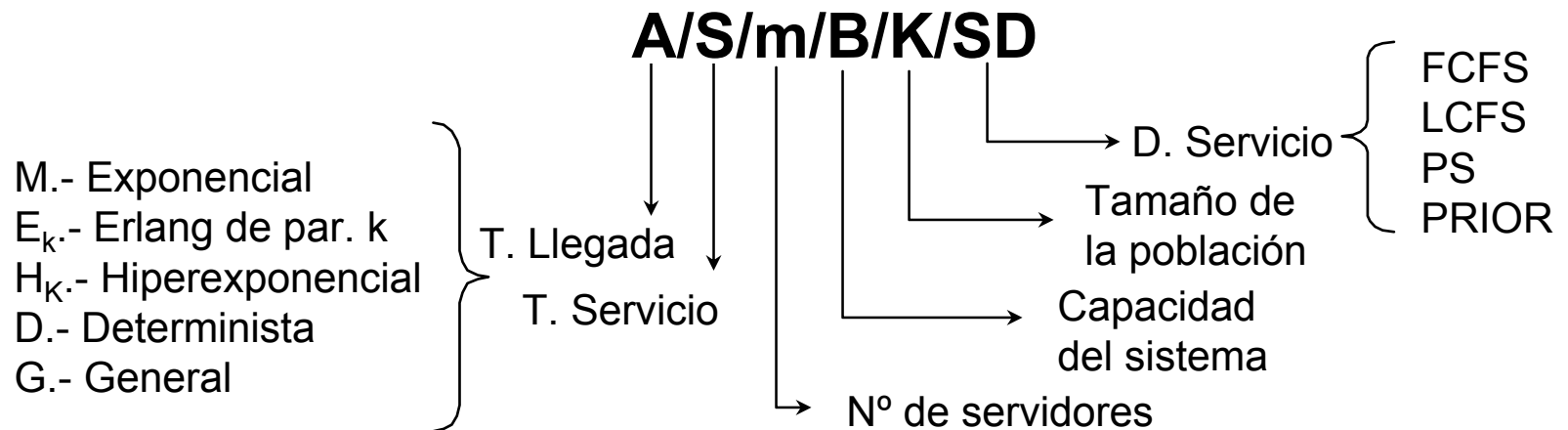
Un ejemplo de este tipo de centro son las terminales directamente conectadas a un sistema multiusuario (centauro y terminales X).

Cada terminal conecta con un proceso que la atiende en exclusiva.

MODELADO ANALÍTICO

Notación Kendall

Cada cola se representa por una tupla de 6 elementos de la forma:

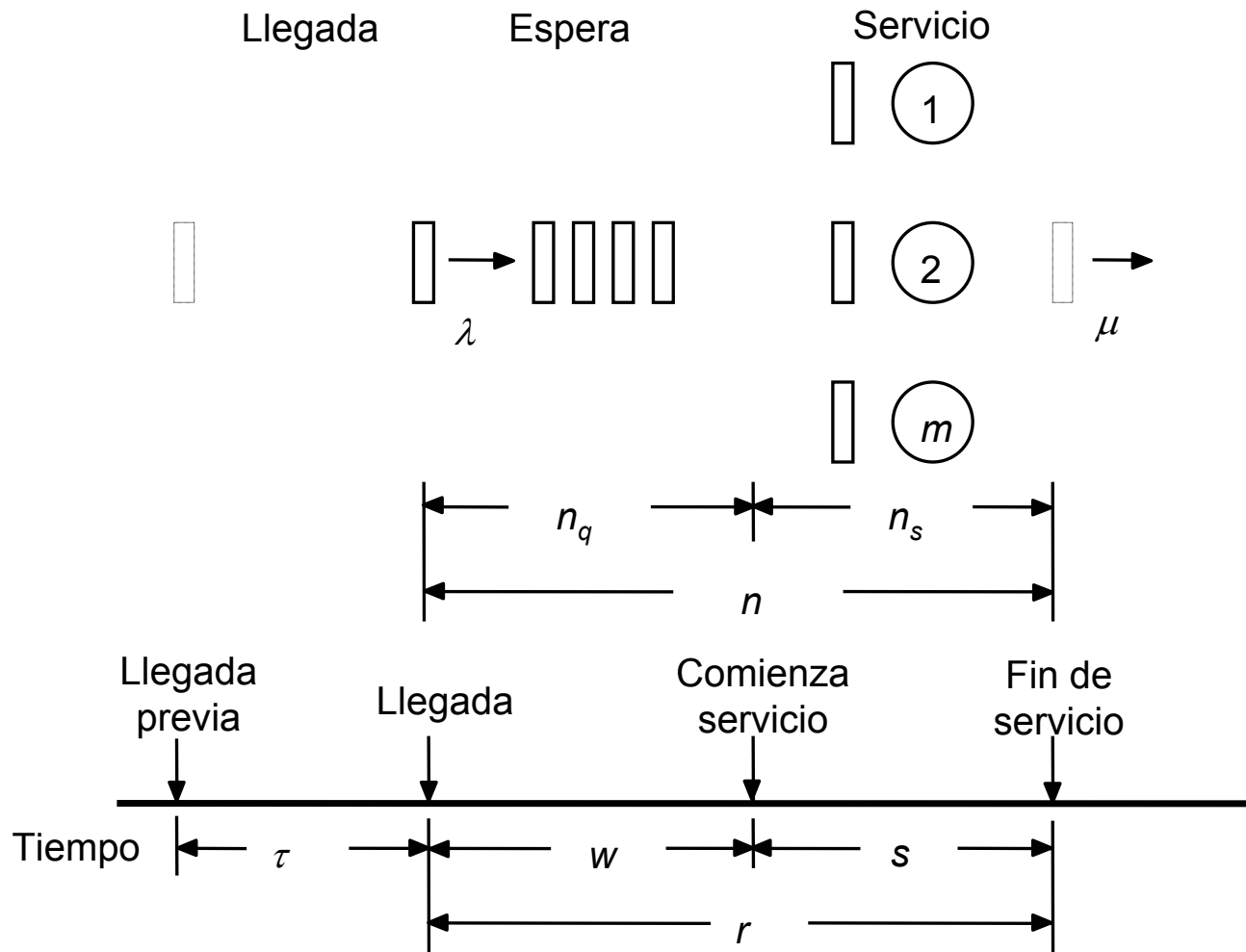


Considerando: capacidad infinita, población infinita y disciplina FIFO, la notación se simplifica a:

A/S/m

MODELADO ANALÍTICO

Parámetros en todas las colas



MODELADO ANALÍTICO

Los parámetros de las colas son variables aleatorias que pueden seguir algún tipo de distribución estadística.

τ .- Tiempo entre dos llegadas consecutivas.

λ .- Razón o cadencia media de llegadas. $\lambda = 1/E[\tau]$

s .- Tiempo recibiendo servicio. Tiempo de servicio.

r .- Tiempo en el sistema. Tiempo de respuesta.

w .- Tiempo que transcurre hasta que empieza a recibir servicio.
Tiempo de espera.

μ .- Razón o cadencia media de servicio “por servidor”. $\mu = 1/E[s]$
Para m servidores será $m\mu$.

n .- Número de trabajos en el sistema.

n_q .- Número de trabajos esperando para recibir servicio.

n_s .- Número de trabajos recibiendo servicio.

MODELADO ANALÍTICO

Relaciones entre variables:

1.- Condición de estabilidad.

$$\lambda < m \cdot \mu$$

El sistema es inestable si la cantidad de tareas que le llegan es superior a la que puede atender. No es aplicable a sistemas con capacidad finita.

2.- Número de elementos en el sistema.

$$n = n_q + n_s$$

El número de elementos o tareas en el sistema es siempre igual a la suma de las tareas que están esperando para recibir servicio y las tareas que están recibiendo servicio.

MODELADO ANALÍTICO

3.- Relación entre nº de elementos y el tiempo. (Ley de Little)

Si no se pierden tareas (por capacidad limitada del sistema), se cumple la relación:

Nº de elementos sistema = Razón de llegada \times Tiempo de respuesta

$$n = \lambda \cdot r$$

Esta relación es aplicable a cualquier parte del sistema donde se cumpla la conservación de tareas, así:

Nº de elementos en cola = Razón de llegada \times Tiempo de espera

$$n_q = \lambda \cdot w$$

4.- Distribución de tiempos en el sistema

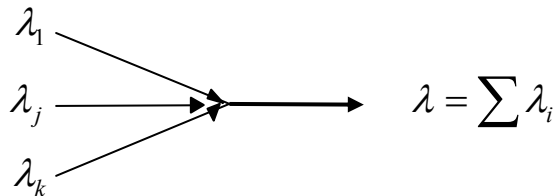
$$r = w + s$$

El tiempo de respuesta es la suma del tiempo de espera (hasta recibir servicio) más el tiempo de servicio.

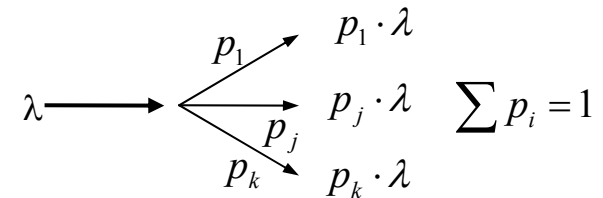
MODELADO ANALÍTICO

Procesos de Poisson

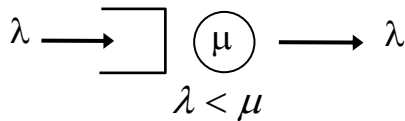
Si una variable aleatoria (tiempo entre llegadas) es independiente y está distribuida exponencialmente, se denomina proceso de Poisson y cumple las propiedades:



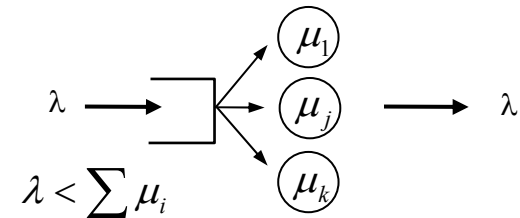
(a) La mezcla de flujos de Poisson es un flujo de Poisson.



(b) Un flujo de Poisson se puede dividir en flujos de Poisson.



(c) La partida de una cola M/M/1 es un proceso de Poisson.



(d) La partida de una cola M/M/m es un proceso de Poisson.

3. REDES DE COLAS

Las colas se combinan para dar un servicio complejo a una tarea. Esta combinación recibe el nombre de red de colas.

- No existe una notación generalizada para las redes de colas.
- No todas las redes de colas tienen solución exacta o aproximada.

Existen tres tipos de redes:

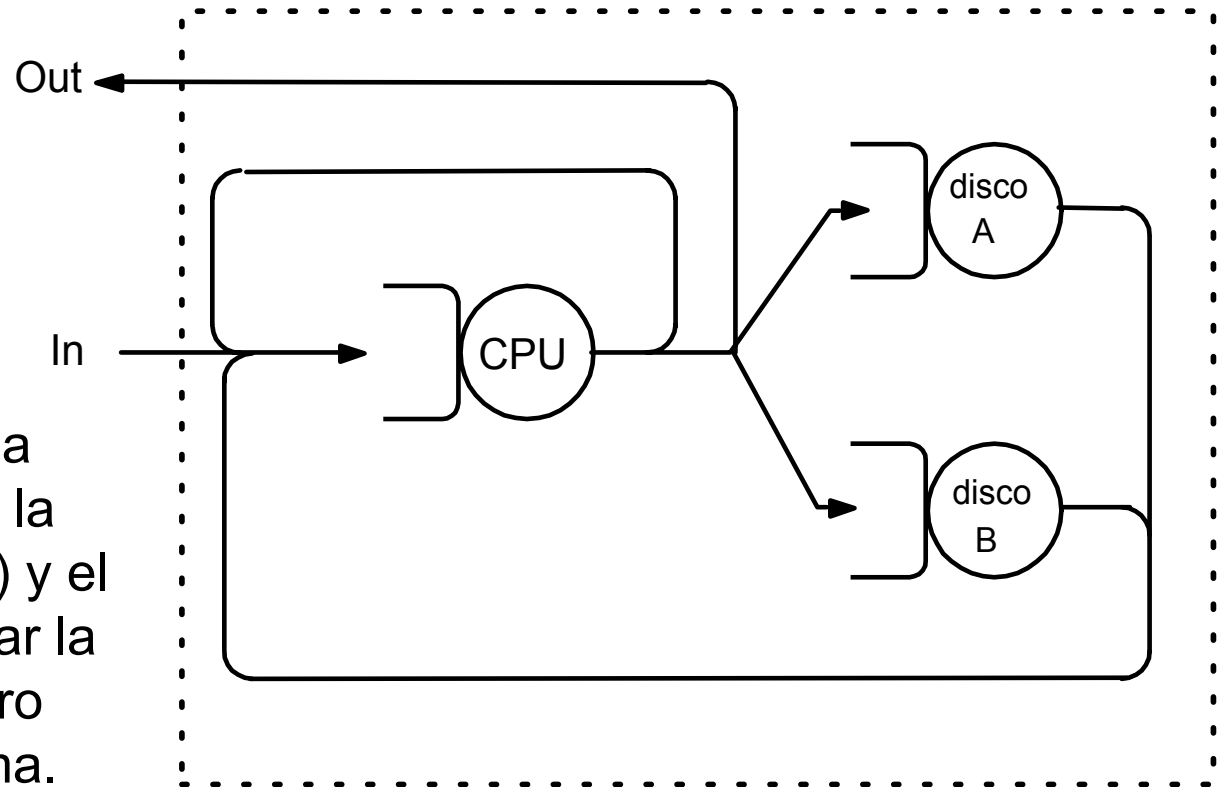
- Redes abiertas
- Redes cerradas
- Redes Mixtas

MODELADO ANALÍTICO

Red de colas Abierta

Una red de colas abierta tiene comunicación con el exterior

Se supone conocida la productividad (igual a la cadencia de llegadas) y el objetivo es caracterizar la distribución del número de tareas en el sistema.

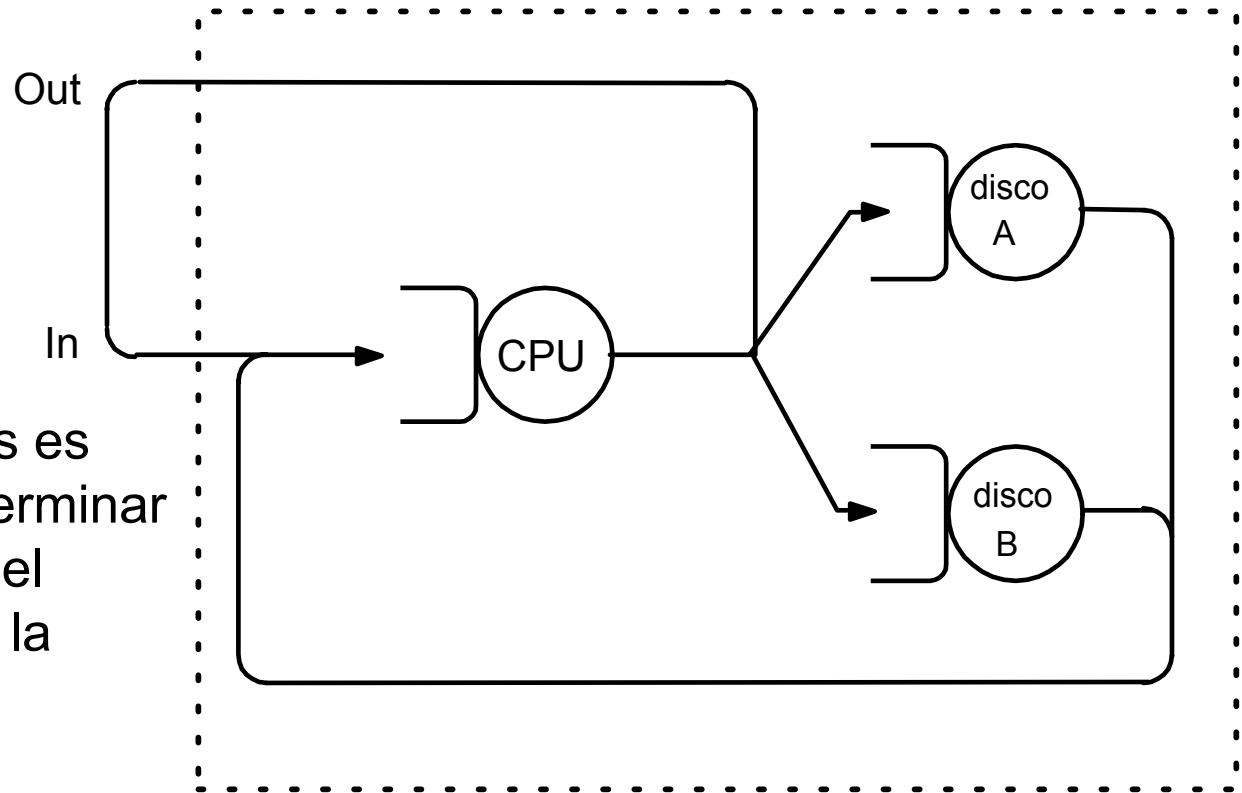


MODELADO ANALÍTICO

Red de colas Cerrada

Una red de colas cerrada no tiene comunicación con el exterior

El número de tareas es fijo. Se trata de determinar la productividad en el enlace ficticio entre la entrada y la salida

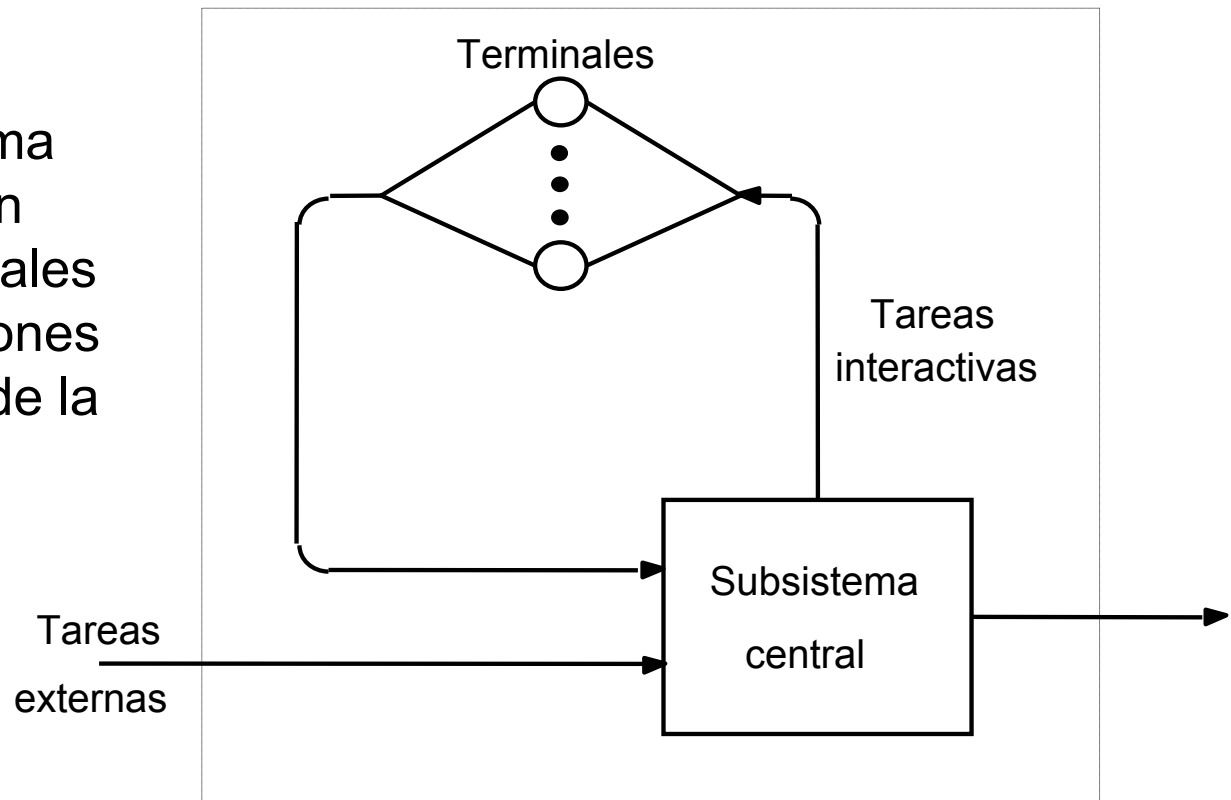


MODELADO ANALÍTICO

Red de colas Mixta

Esta red de colas será abierta para algunas tareas y cerrada para otras.

Ejemplo: un sistema multiusuario con un conjunto de terminales dedicadas y peticiones externas a través de la red. (Ej. centauro)



MODELADO ANALÍTICO

Red de tipo producto o BCMP

Este tipo de redes de colas tiene resolución matemática exacta.

Se llaman de tipo producto porque su distribución de probabilidades sigue la expresión:

$$P(n_1, n_2, \dots, n_M) = \frac{1}{G(N)} \prod_{i=1}^M f_i(n_i)$$

Son redes de este tipo las que cumplen:

- Disciplinas de servicio: FCFS, PS, IS ó LCFS-PR
- Las tareas no cambian de tipo mientras reciben servicio.
- Distribución de servicio exponencial para FCFS.
- El tiempo de servicio sólo depende del estado de la cola.
- En redes abiertas, las llegadas están distribuidas exponencialmente. No se permiten llegadas en grupo.

4. LEYES OPERACIONALES

Los problemas de análisis de prestaciones de los computadores pueden resolverse empleando relaciones sencillas que no requieren hipótesis sobre distribuciones. Estas relaciones se llaman Leyes Operacionales.

Considerando el sistema como una caja negra, durante un tiempo T podemos medir:

- Número de tareas que han llegado al sistema: A_i
- Número de tareas que han recibido servicio: C_i
- Intervalo de tiempo durante el cual el sistema está ocupado, B_i

A partir de estas cuatro medidas se pueden definir las siguientes relaciones:

MODELADO ANALÍTICO

Cadencia de llegada: $\lambda_i = \frac{A_i}{T}$ $\frac{\text{Nº de llegadas}}{\text{Periodo de medida}}$

Productividad: $X_i = \frac{C_i}{T}$ $\frac{\text{Nº tareas completadas}}{\text{Periodo de medida}}$

Utilización: $U_i = \frac{B_i}{T}$ $\frac{\text{Tiempo de ocupación}}{\text{Periodo de medida}}$

Tiempo medio de servicio: $S_i = \frac{B_i}{C_i}$ $\frac{\text{Tiempo de ocupación}}{\text{Nº tareas completadas}}$

MODELADO ANALÍTICO

1º Ley de Utilización

Permite establecer una relación alternativa para la utilización:

$$U_i = \frac{B_i}{T} = \frac{C_i}{T} \times \frac{B_i}{C_i} \Rightarrow U_i = X_i \cdot S_i$$

Ejemplo: Consideremos un *gateway*, donde los paquetes llegan a razón de 125 paquetes por segundo (pps) y el *gateway* emplea un promedio de 2 milisegundos en atenderlos.

Productividad X_i = razón de salida = razón de llegada = 125 pps

Tiempo de servicio S_i = 0.002 segundos

Utilización $U_i = X_i S_i = 125 \times 0.002 = 0.25 = 25\%$

Los resultados en este caso son válidos sin necesidad de realizar suposiciones sobre la distribución de las variables.

MODELADO ANALÍTICO

2ª Ley de flujo forzado

Relaciona la productividad del sistema con las productividades de los dispositivos individuales.

Consideramos:

- Existe balance o equilibrio de flujo. $A_i = C_i$
- Se define la *razón de visitas*, V_i , como la relación entre el número de tareas que entran y el número de veces que las tareas pasan por un dispositivo concreto, es decir:

$$V_i = \frac{C_i}{C_0} \quad \frac{\text{Nº tareas por el dispositivo}}{\text{Nº tareas del exterior}}$$

- Se define la productividad global del sistema:

$$X = \frac{C_0}{T} \quad \frac{\text{Nº total de tareas completadas}}{\text{Tiempo total}}$$

MODELADO ANALÍTICO

Teniendo en cuenta lo anterior, la productividad de un dispositivo i puede escribirse:

$$X_i = \frac{C_i}{T} = \frac{C_i}{C_0} \times \frac{C_0}{T} \Rightarrow X_i = V_i \cdot X$$

Teniendo en cuenta la relación anterior en la ley de utilización:

$$U_i = X_i \cdot S_i$$

$$U_i = X \cdot V_i \cdot S_i$$

$$U_i = X \cdot D_i$$

Al término:

Se le denomina demanda total de servicio del dispositivo i para una tarea.

$$V_i \cdot S_i = D_i$$

El elemento con mayor D_i será el más utilizado y se denomina “*cuello de botella*” (bottleneck).

MODELADO ANALÍTICO

Problema: En un sistema de tiempo compartido, se registra el siguiente perfil de los programas de usuario. Cada programa requiere 5 segundos de tiempo de CPU y hace 80 peticiones de E/S al disco A y 100 peticiones E/S al disco B. El tiempo promedio de reflexión de los usuarios fue de 18 segundos. De las especificaciones de los dispositivos sabemos que el disco A emplea 50 milisegundos para satisfacer una petición de E/S y el disco B emplea 30 milisegundos por petición. Con 17 terminales activas, se observó que la productividad del disco A es de 15.70 peticiones E/S por segundo. Queremos calcular la productividad del sistema y la utilización de los dispositivos.

Nota: cada paso por un dispositivo de E/S implica un paso por la CPU.

MODELADO ANALÍTICO

3ª Ley de Little

Permite establecer una relación entre elementos y tiempo. La única consideración es que exista flujo equilibrado de tareas:

$$A_i = C_i$$

Nº medio de tareas en el dispositivo = Cadencia de llegada \times Tiempo medio en el dispositivo

$$Q_i = X_i \cdot R_i$$

Si el flujo está equilibrado:

$$\lambda_i = X_i$$

Puede aplicarse a todo el sistema o a cualquier componente que cumpla la condición anterior.

MODELADO ANALÍTICO

4ª Ley general del tiempo de respuesta

El tiempo de respuesta total de un sistema es:

$$R = \sum_{i=1}^M R_i \cdot V_i$$

5ª Ley del tiempo de respuesta interactivo

En un sistema interactivo (terminales conectadas a un servidor) se cumple:

- El usuario envía una petición y el sistema tarda R en responder.
- El usuario piensa durante un tiempo Z (tiempo de reflexión) antes de enviar otra petición. Tiempo por petición ($R + Z$)
- Cada usuario hace $\frac{T}{R + Z}$ peticiones.
- Existen N usuarios (Número de terminales)

MODELADO ANALÍTICO

La productividad del sistema será:

$$X = \frac{\text{Nº total de peticiones}}{\text{Tiempo total}}$$

$$X = \frac{N \cdot [T / (R + Z)]}{T} = \frac{N}{(R + Z)}$$

Despejando el tiempo de respuesta:

$$R = \left(\frac{N}{X} \right) - Z$$

MODELADO ANALÍTICO

Análisis de cuellos de botella

El “*cuello de botella*” es el elemento con mayor demanda, es decir, mayor utilización. Este elemento limita el rendimiento del sistema.

Se pueden establecer unos límites asintóticos tanto para la productividad como para el tiempo de respuesta del sistema.

Si tenemos M dispositivos:

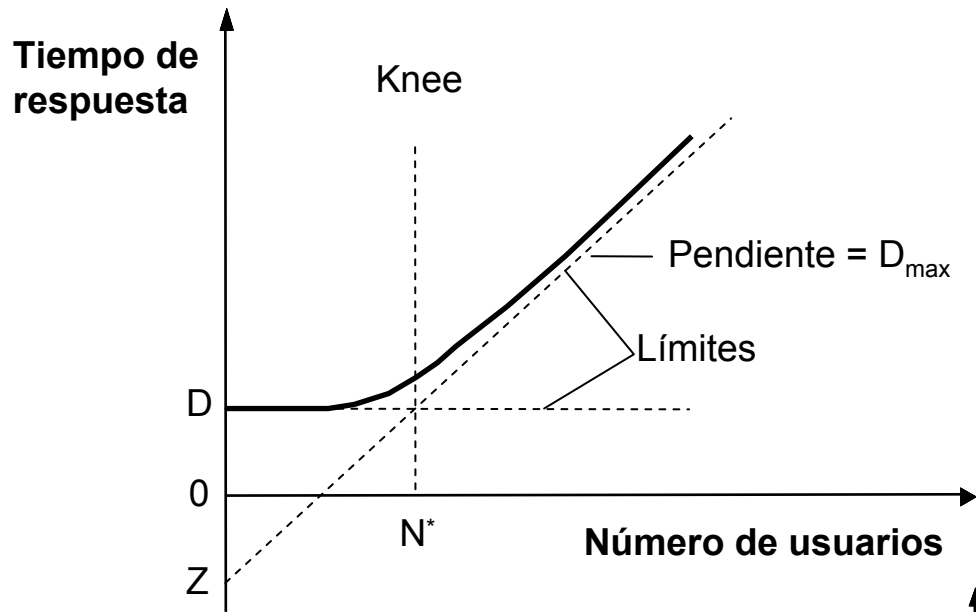
$$D_b = D_{max} \text{ entre las } D_1, D_2, \dots, D_M \qquad D = \sum_{i=1}^M D_i$$

Tendremos como límites:

$$R(N) \geq \max\{D, N \cdot D_{max} - Z\}$$

$$X(N) \leq \min\left\{\frac{1}{D_{max}}, \frac{N}{(D + Z)}\right\}$$

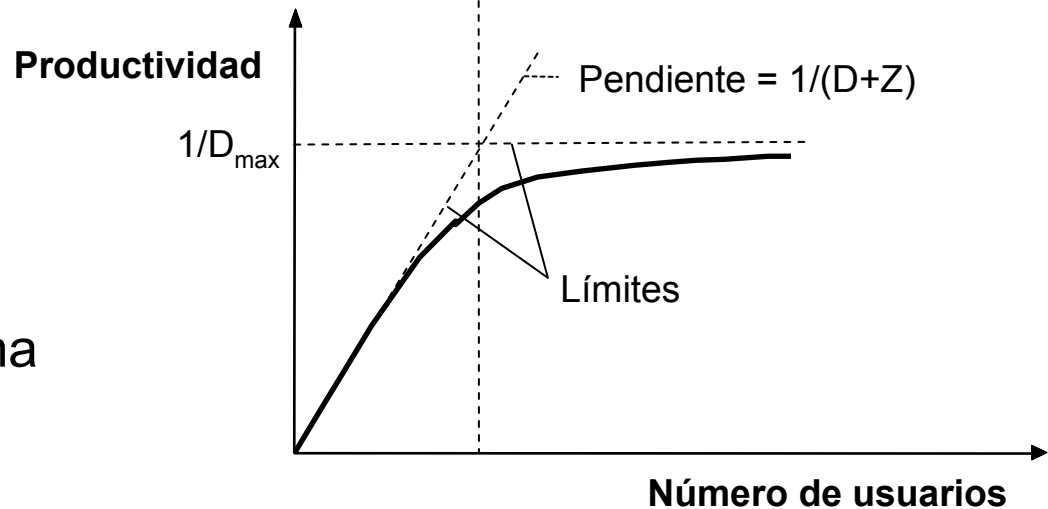
MODELADO ANALÍTICO



Límite asintótico para el tiempo de respuesta del problema propuesto

N^* .- nº de terminales (usuarios) a partir del cual comienza a saturarse el sistema

Límite asintótico para la productividad del problema propuesto



ANEXO. LIMITACIONES DE LA TEORÍA DE COLAS

La teoría de colas no puede resolver problemas que consideren:

- Tiempos de servicio no exponenciales
- Llegadas dependientes de la carga
- Llegadas en bloque
- Saltos entre colas
- Bloqueos
- Posesión simultánea de recursos
- Eliminación de las colas
- Transitorios
- Etc.