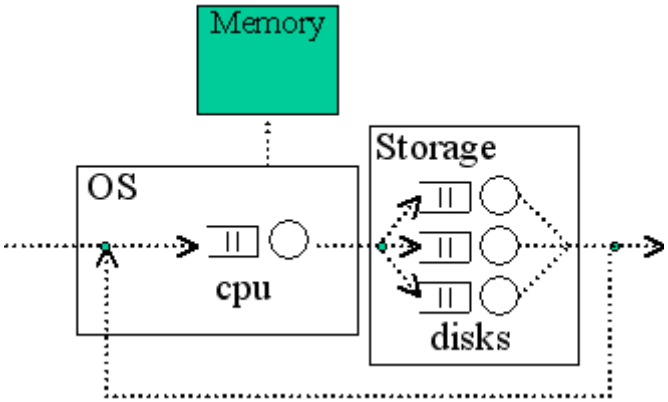
INTRODUCCIÓN AL ANÁLISIS OPERACIONAL

Administradores y diseñadores

¿Cómo modelar el rendimiento de un sistema informático (con teoría de colas)?

CONTENIDO

1. Introducción
   - Estaciones de servicio
2. Redes de colas de espera
   - Abiertas, cerradas y mixtas
3. Variables operacionales
   - Variables básicas y deducidas
4. Leyes operacionales
   - Hipótesis del equilibrio de flujo
   - Ley de Little
   - Ley de utilización
   - Ley del flujo forzado
   - Ley general del tiempo de respuesta



1. INTRODUCCIÓN

Concepto de estación de servicio
Tipos de estaciones de servicio

MODELADO DEL RENDIMIENTO

El rendimiento es uno de los factores clave que deben tenerse en cuenta en el diseño, desarrollo, configuración y puesta a punto de un sistema informático.

El rendimiento de un sistema depende del comportamiento de sus componentes.

Si bien el rendimiento de un sistema existente puede evaluarse mediante mediciones y monitoreo, esto puede no ser posible durante sus etapas de diseño.

- Si no hay un sistema disponible, debería tener al menos una descripción del sistema para hacer un modelo abstracto.

EL MODELO DE UN SISTEMA

Modelo: abstracción del sistema informático real

- Conjunto de recursos relacionados y trabajos que los usan
  - Recursos: procesador, discos, memoria, etc.
  - Trabajos: programas, transacciones, peticiones, accesos, etc.
- Usualmente un recurso solo puede ser usado por un trabajo, el resto tendrá que esperar

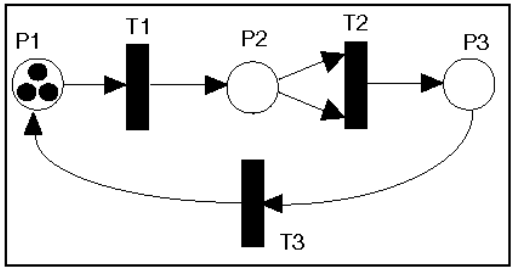
Modelos basados en **redes de colas** (*queueing networks*)

- Introducidos por Jackson en la década de 1950
- Objetivo: cálculo del tiempo de respuesta que experimenta un trabajo procesado por un sistema informático
- Aproximación estadística

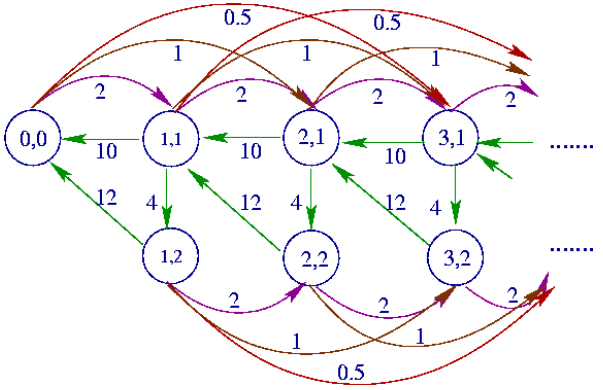
Otros modelos: redes de Petri, álgebra de procesos y cadenas de Markov

EJEMPLOS DE MODELOS CON DISTINTOS FORMALISMOS ESTOCÁSTICOS

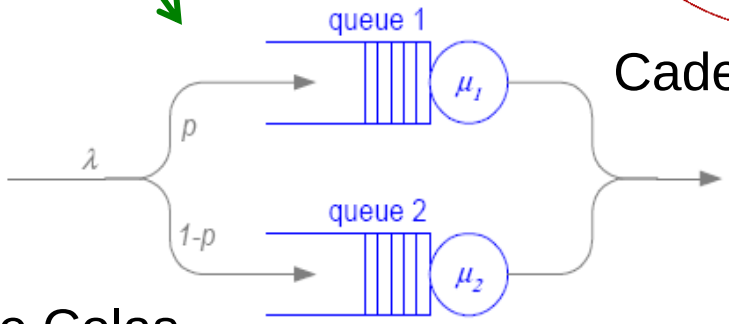
Sistema real



Redes de Petri



Cadenas de Markov



Redes de Colas

MODELADO DEL RENDIMIENTO

Un **formalismo** de modelado es un lenguaje alfanumérico y/o gráfico para la especificación de un modelo de sistema.

Como el **modelo** corresponde a un sistema que aún no existe, siempre hay cierta incertidumbre en su construcción a través de formalismos.

- Se les denominan formalismos estocásticos porque incorporan variables aleatorias respecto al sistema.

Una vez que se construye un modelo, debe analizarse utilizando una **técnica de evaluación**.

- Si el modelo cumple ciertos requisitos, las medidas de rendimiento se pueden calcular directamente con técnicas **analítico-numéricas**.
- En la mayoría de los casos se puede utilizar la **simulación** de eventos discretos.

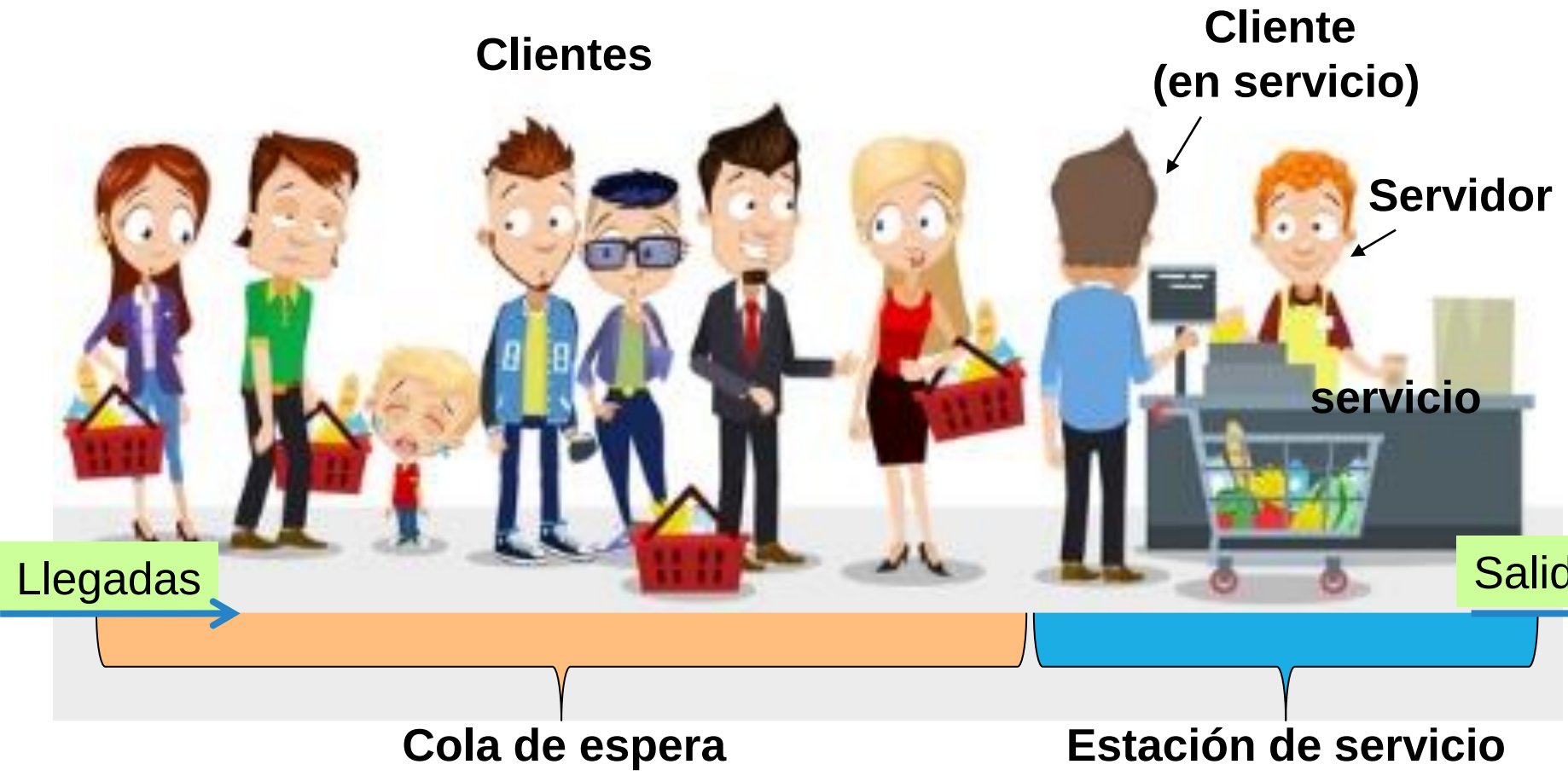
USAREMOS MODELOS BASADOS EN COLAS

EJEMPLO

BLOQUEO EN EL CANAL DE SUEZ

El carguero 'MV Ever Given', de 400 metros de eslora y 59 metros de manga, que transporta 224.000 toneladas de mercancía, lleva encallado en el Canal de Suez desde el pasado martes, interrupiendo el tráfico marítimo

TERMINOLOGÍA EN TEORÍA DE COLAS

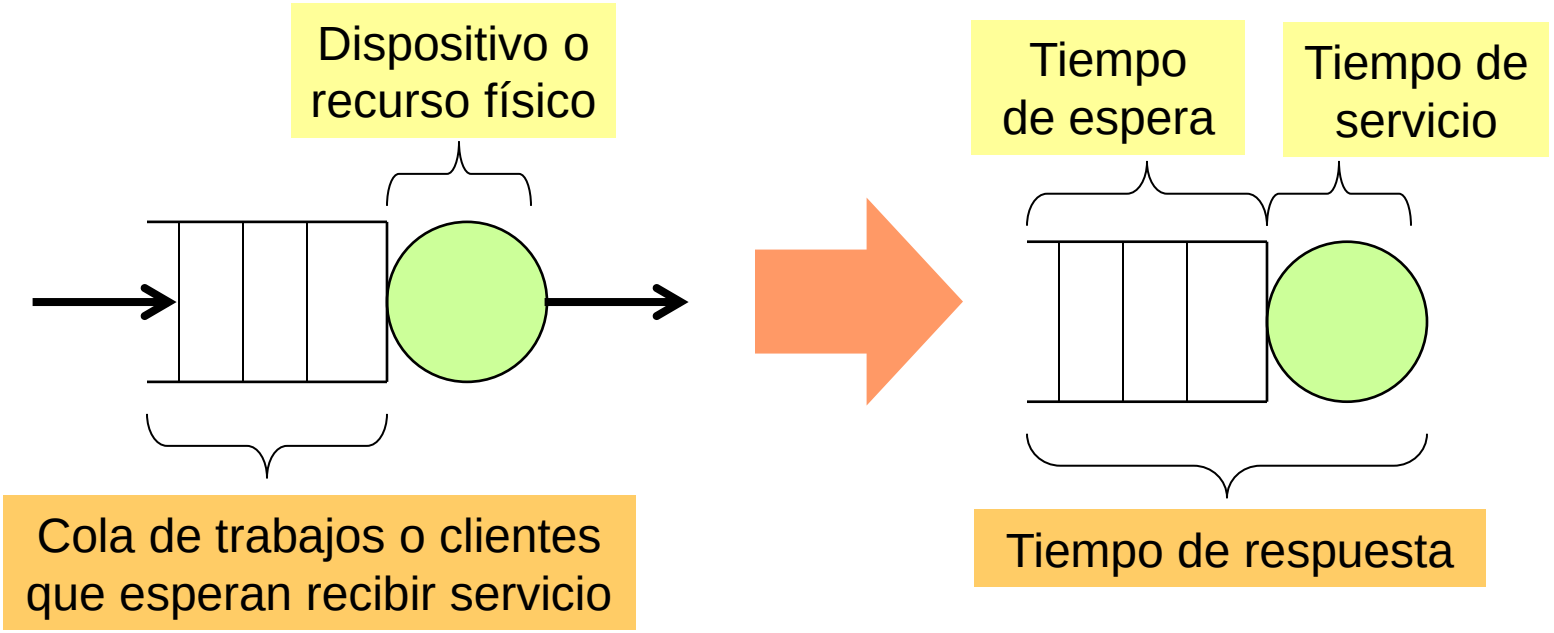


EJEMPLO: VACUNACIÓN COVID-19

CONCEPTO DE ESTACIÓN DE SERVICIO

Estación de servicio (*queue, service station*)

- Objeto abstracto compuesto por un servidor y una cola de espera



VARIABLES TEMPORALES

Tiempo de espera en cola

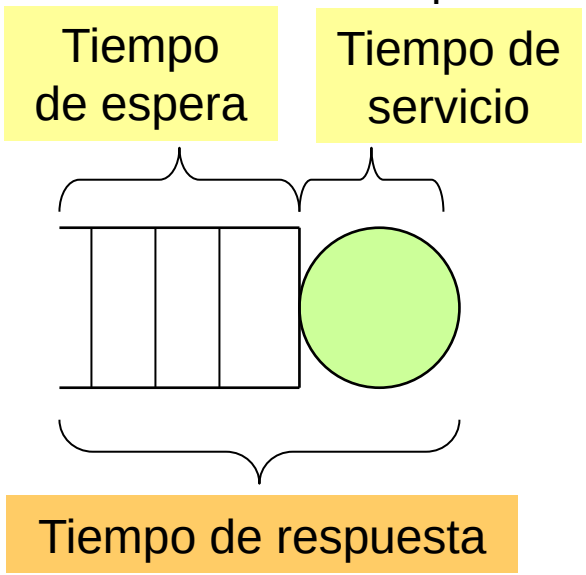
- Tiempo transcurrido desde que un trabajo quiere utilizar de un recurso hasta que realmente empieza a utilizarlo

Tiempo de servicio

- Tiempo transcurrido desde que un trabajo hace uso de un recurso hasta que lo libera

Tiempo de respuesta

- Suma de los dos tiempos anteriores

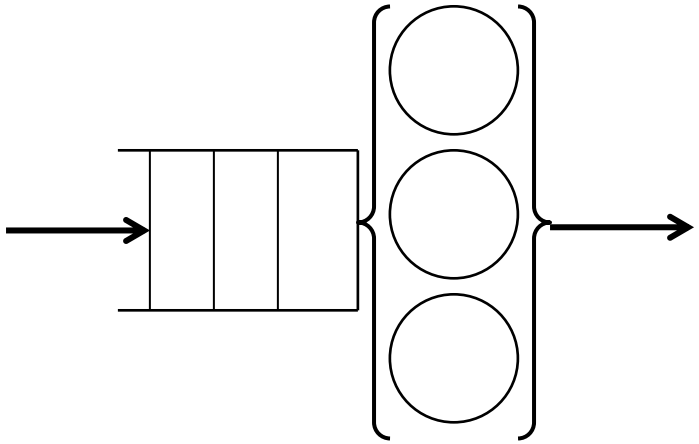


TERMINOLOGÍA DE TEORÍA DE COLAS

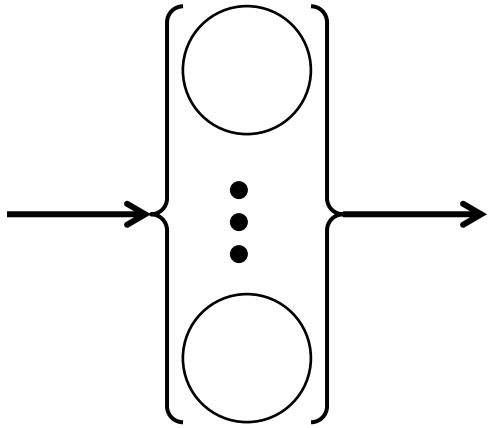
TERMINOLOGÍA DE TEORÍA DE COLAS

ESTACIONES CON MÁS DE UN SERVIDOR

Sirven para atender a más de un trabajo en paralelo



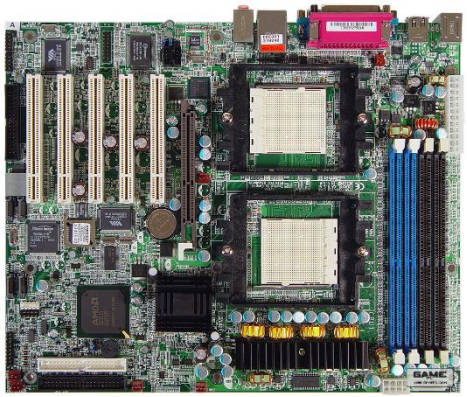
3 servidores idénticos

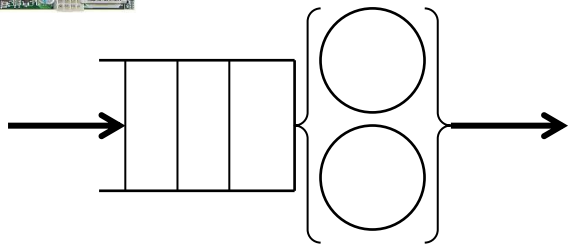


Infinitos servidores: no hay espera en cola

UN PAR DE MODELOS SENCILLOS

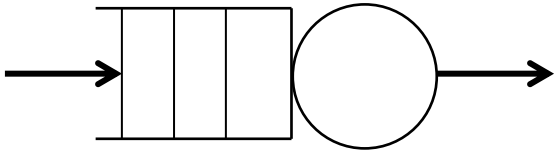
Biprocesador





Tiempo de servicio: instrucciones máquina que se ejecutan dividido por la velocidad de ejecución de cada procesador (MIPS)

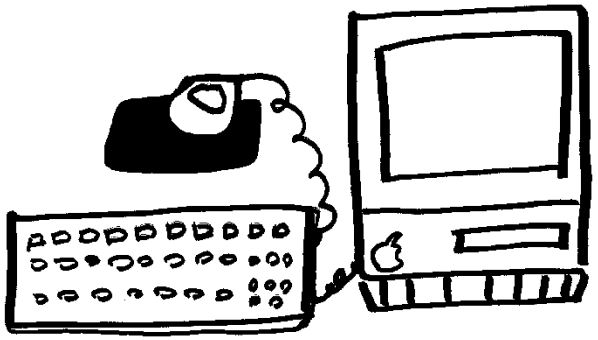
Disco



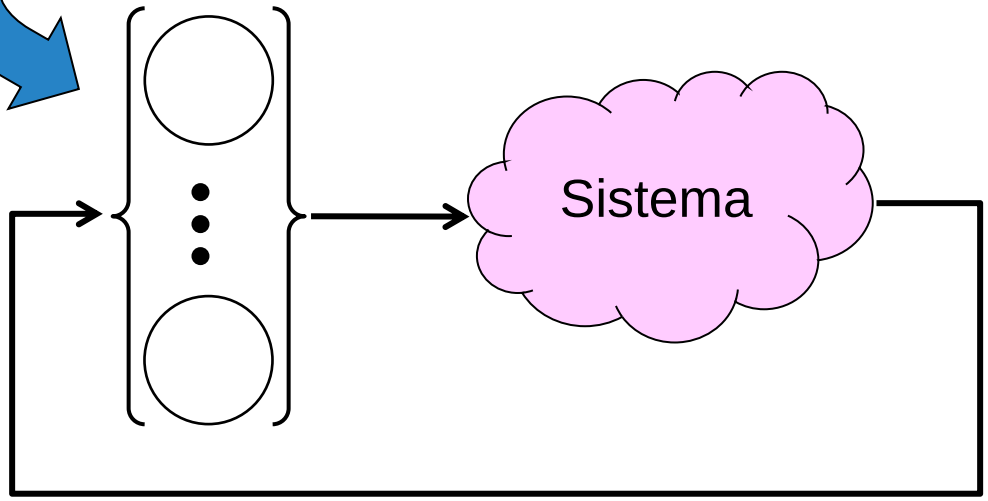
Tiempo de servicio: posicionamento más latencia rotacional más transferencia

EL TIEMPO DE REFLEXIÓN (*THINK TIME*)

Es un parámetro que depende del usuario del sistema informático



No hay espera, sólo servicio = reflexión

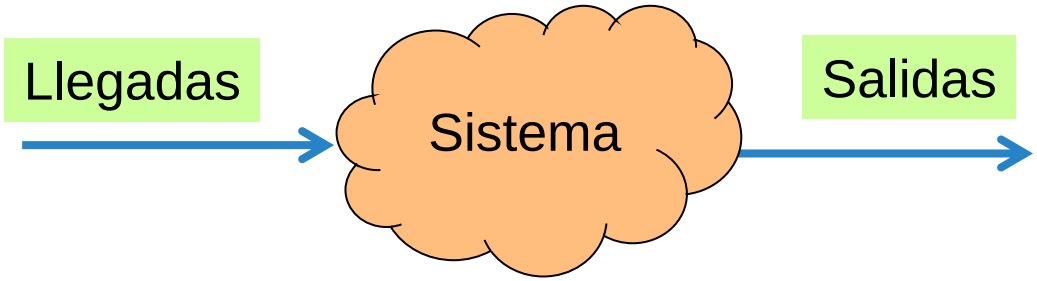


Tiempo de servicio: tiempo que transcurre entre una interacción y el lanzamiento de la siguiente

EL ANÁLISIS OPERACIONAL

Presentado por Denning y Buzen en 1978

Basado en **magnitudes medibles** (*operacionales*) del sistema informático



Leyes operacionales: relaciones entre las magnitudes medibles

Límites optimistas de las prestaciones por medio de cálculos muy sencillos (*back on the envelope calculations*)

ACTIVIDADES

ACTIVIDAD TEMA 4.1 (OBLIGATORIA)

¿Por qué crees que en los hipermercados hay colas "rápidas" para clientes que lleven igual o menor cantidad de un número (relativamente bajo) de productos?

¿Por qué crees que en algunos establecimientos hay cola única para los clientes y varios servidores para la cola única, en vez de usar varias colas, con una por servidor?

Imagina una cola de facturación en el aeropuerto, ¿qué variables principales gobiernan la longitud de la cola?

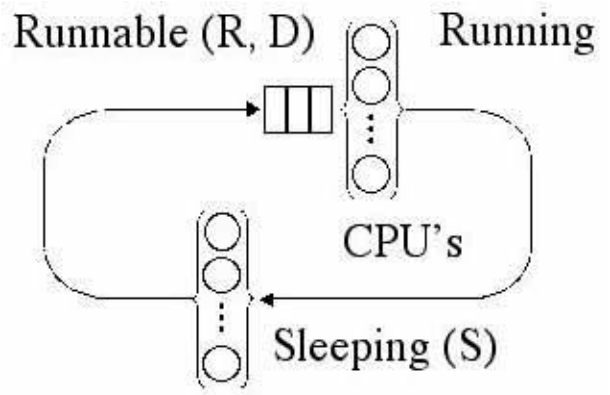
2. REDES DE COLAS

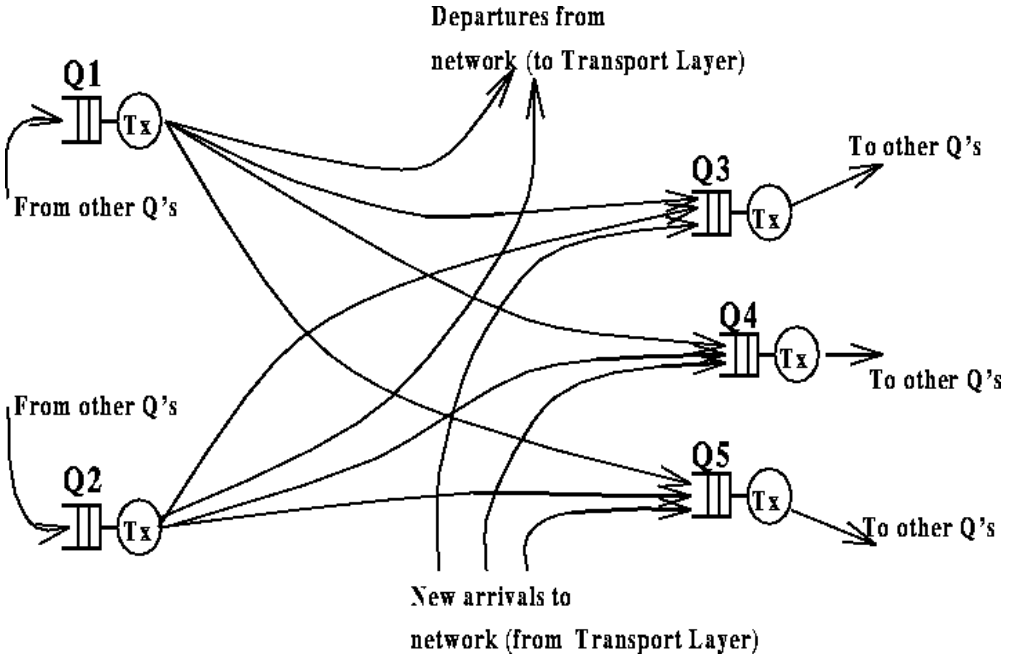
Concepto de red de colas
Modelo del servidor central
Tipos de redes: cerradas, abiertas y mixtas

REDES DE COLAS: CONCEPTO

Conjunto de estaciones de servicio conectadas entre sí

Cada recurso del sistema se representa mediante una estación de servicio





EL MODELO DE SERVIDOR CENTRAL

Representa el comportamiento de los programas en la mayoría de los sistemas informáticos

¿Cuál es este comportamiento?

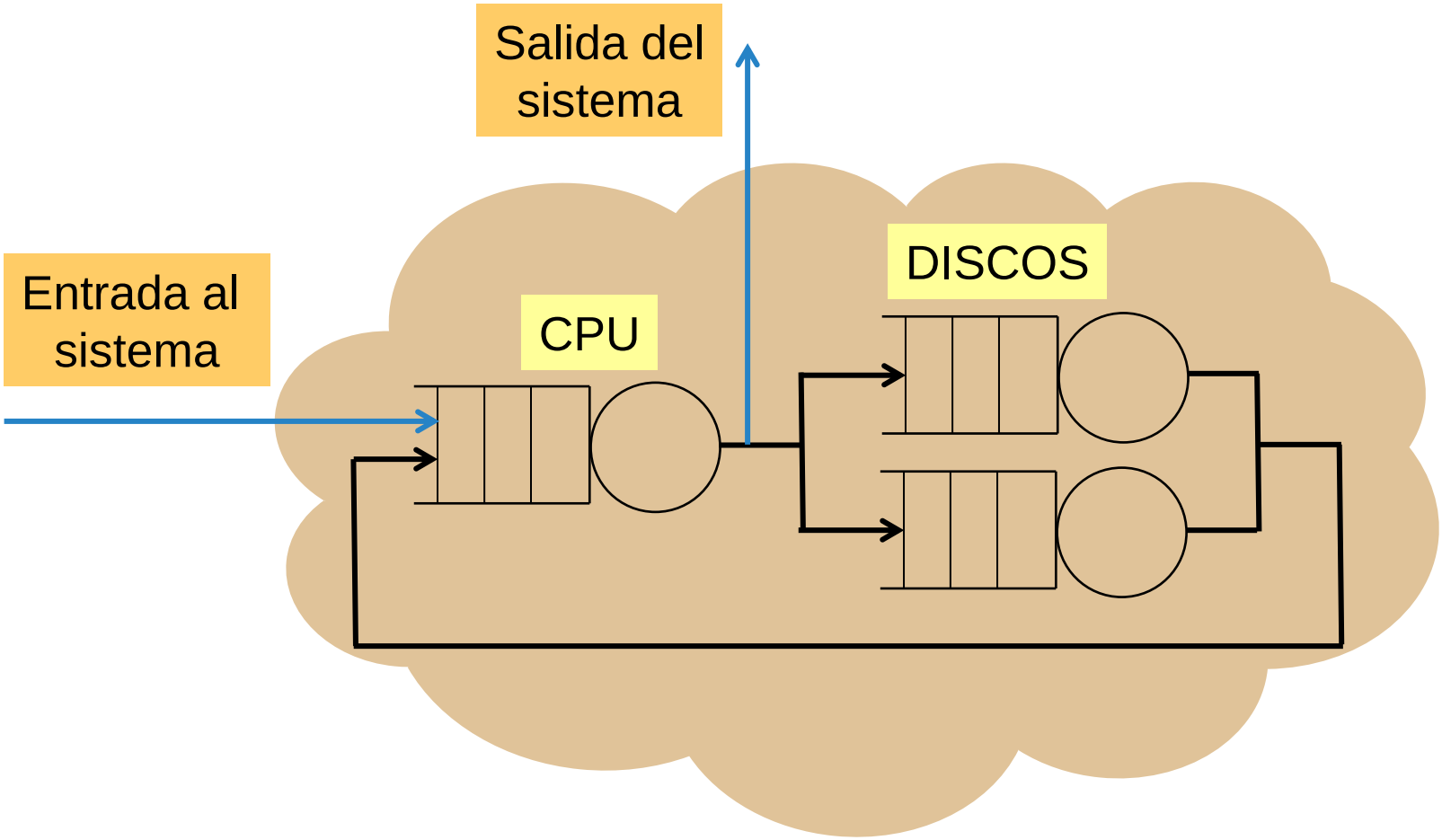
- Un trabajo que llega al sistema comienza utilizando el procesador
- Después de dejar el procesador, el trabajo puede:
  - Terminar (sale del sistema), o bien
  - Realizar un acceso a la unidad de entrada/salida
- Después de una operación con una unidad de entrada/salida, el trabajo vuelve al procesador

Recursos considerados

- Procesador
- Entrada/salida: unidades de disco magnético, óptico, etc.

DIAGRAMA DE CONEXIÓN

Integra tanto los dispositivos como su uso por parte de los trabajos



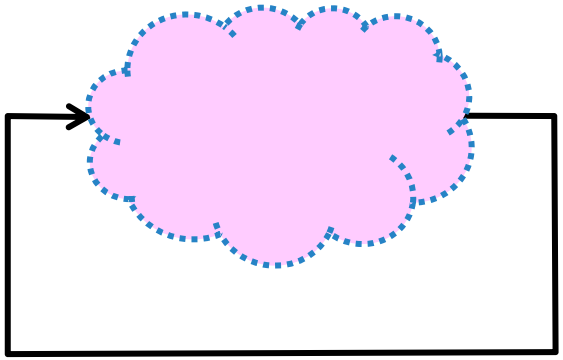
REDES DE COLAS CERRADAS

Sistemas con cargas interactivas y por lotes (*batch*)

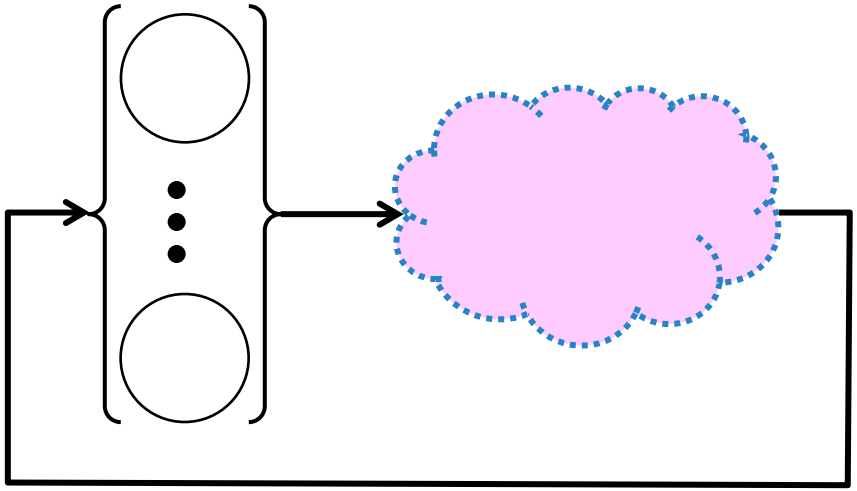
Número constante de trabajos en el sistema (N)

Tiempo de reflexión (Z, *think time*)

Objetivo: cálculo del tiempo de respuesta y de la productividad



Sistema batch



Sistema interactivo

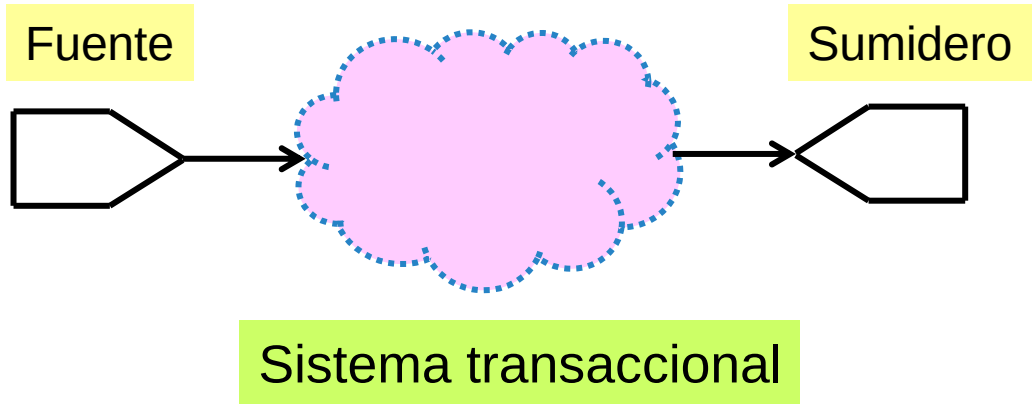
REDES DE COLAS ABIERTAS

Sistemas con cargas transaccionales

Se parte de una tasa de llegada de trabajos conocida (λ)

El número de trabajos en el sistema varía con el tiempo

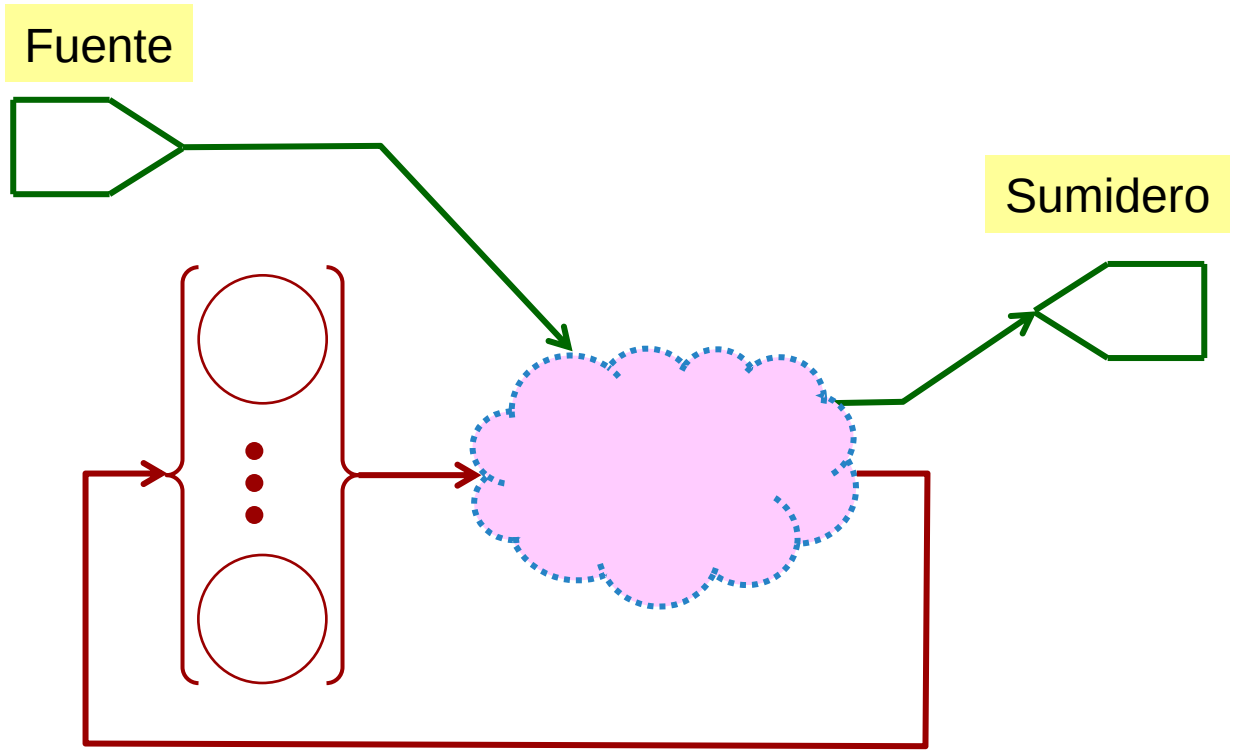
Objetivo: cálculo del tiempo de respuesta y del número de trabajos en el sistema



REDES DE COLAS MIXTAS

Más de un tipo de carga que hace uso del sistema

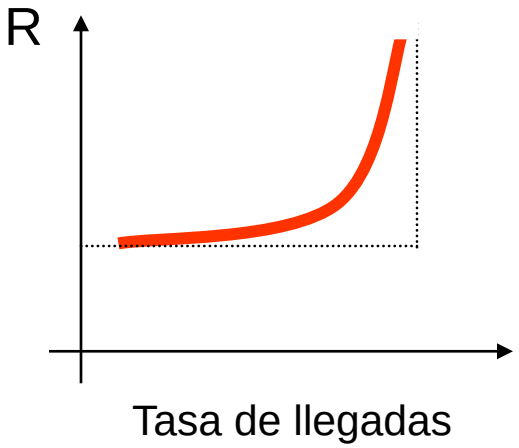
Ejemplo: sistema con carga interactiva y transaccional



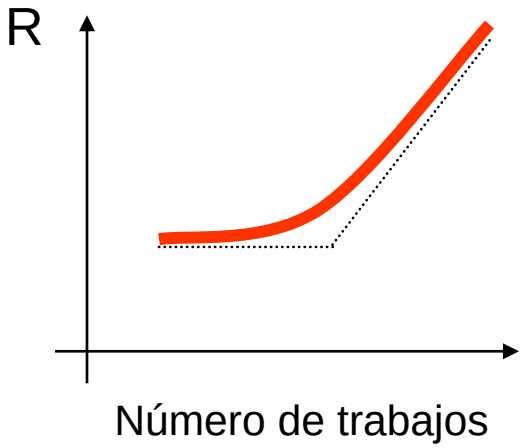
TIEMPO MEDIO DE RESPUESTA

Se mide desde que el trabajo entra al sistema hasta que lo abandona

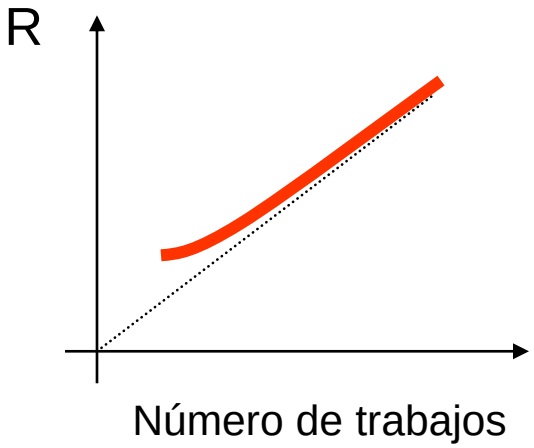
Transaccional



Interactivo: usuarios



Lotes: planificado



3. VARIABLES OPERACIONALES

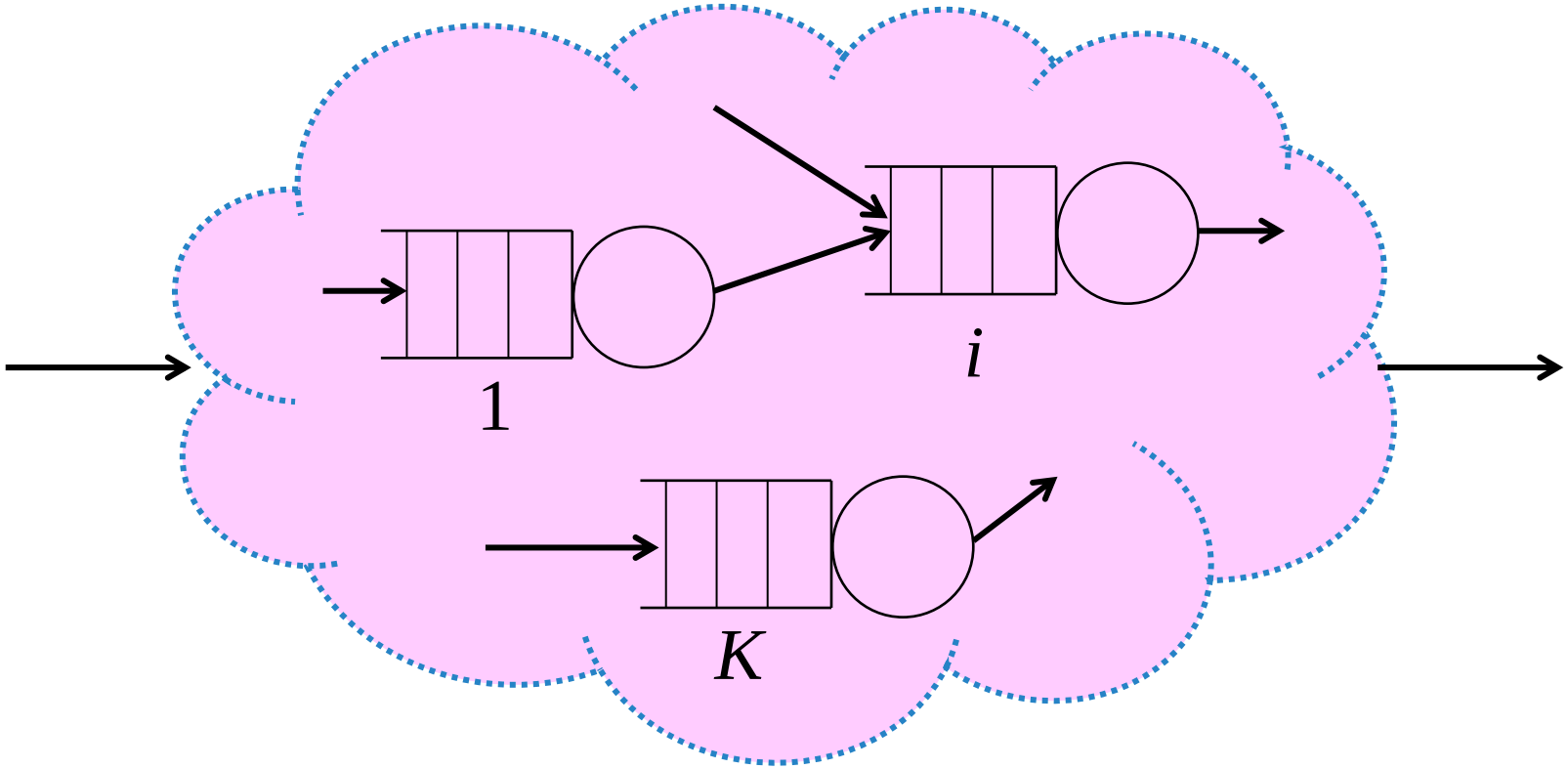
Variables básicas: directamente medibles

Variables deducidas

VARIABLES: SISTEMA VS. ESTACIÓN

El sistema contiene K recursos o dispositivos

El exterior se indica como el dispositivo cero (0)



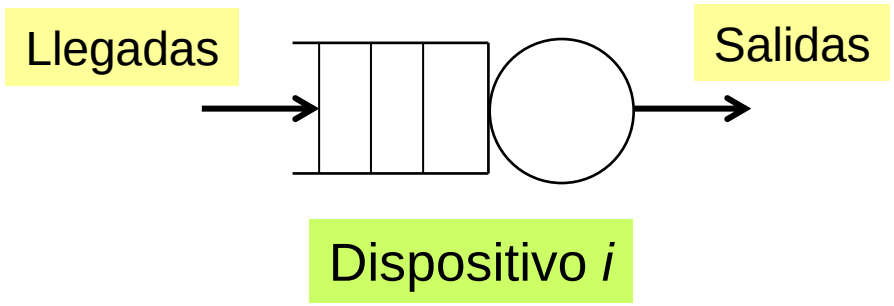
LAS VARIABLES BÁSICAS

Variable temporal

- T Duración del periodo de medida (*time*)

Variables relacionadas con el dispositivo *i*

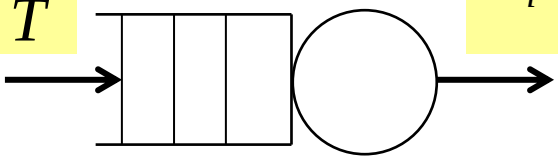
- A_i Número de trabajos que llegan (*arrivals*)
- C_i Número de trabajos que se van (*completions*)
- B_i Tiempo de ocupación (*busy time*)



LAS VARIABLES DEDUCIDAS

▪ U_i	Utilización (*utilization*)	Adimensional
▪ λ_i	Tasa de llegadas (*arrival rate*)	Trabajos/tiempo
▪ X_i	Productividad (*throughput*)	Trabajos/tiempo
▪ S_i	Tiempo de servicio (*service time*)	Tiempo
▪ V_i	Razón de visita (*visit ratio*)	Adimensional
▪ D_i	Demanda de servicio (*service demand*)	Tiempo

$$\lambda_i = \frac{A_i}{T}$$



$$X_i = \frac{C_i}{T}$$

$$S_i = \frac{B_i}{C_i}$$

$$U_i = \frac{B_i}{T}$$

$$V_i = \frac{C_i}{C_0}$$

$$D_i = V_i \times S_i$$

ALGUNOS DETALLES IMPORTANTES

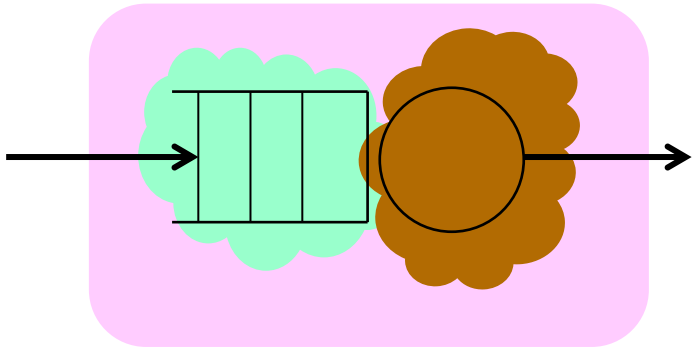
- Las variables deducidas **son valores medios**
- La **utilización** de un dispositivo está **entre 0 y 1**
- El **tiempo de servicio** es el tiempo que un trabajo pasa **en el servidor** del dispositivo
- La razón de visita (V_i) indica las veces que un trabajo visita un determinado dispositivo por cada trabajo
- La demanda de servicio (D_i) no tiene en cuenta la posible espera en cola. Representa la carga que un trabajo provoca en el sistema

OTRAS VARIABLES DE UNA ESTACIÓN

- R_i Tiempo de respuesta (*response time*)
- W_i Tiempo de espera en cola (*waiting time*)
- N_i Trabajos en toda la estación (cola más servidor)
- Q_i Trabajos en cola de espera (*waiting customers*)

Dimensión temporal:

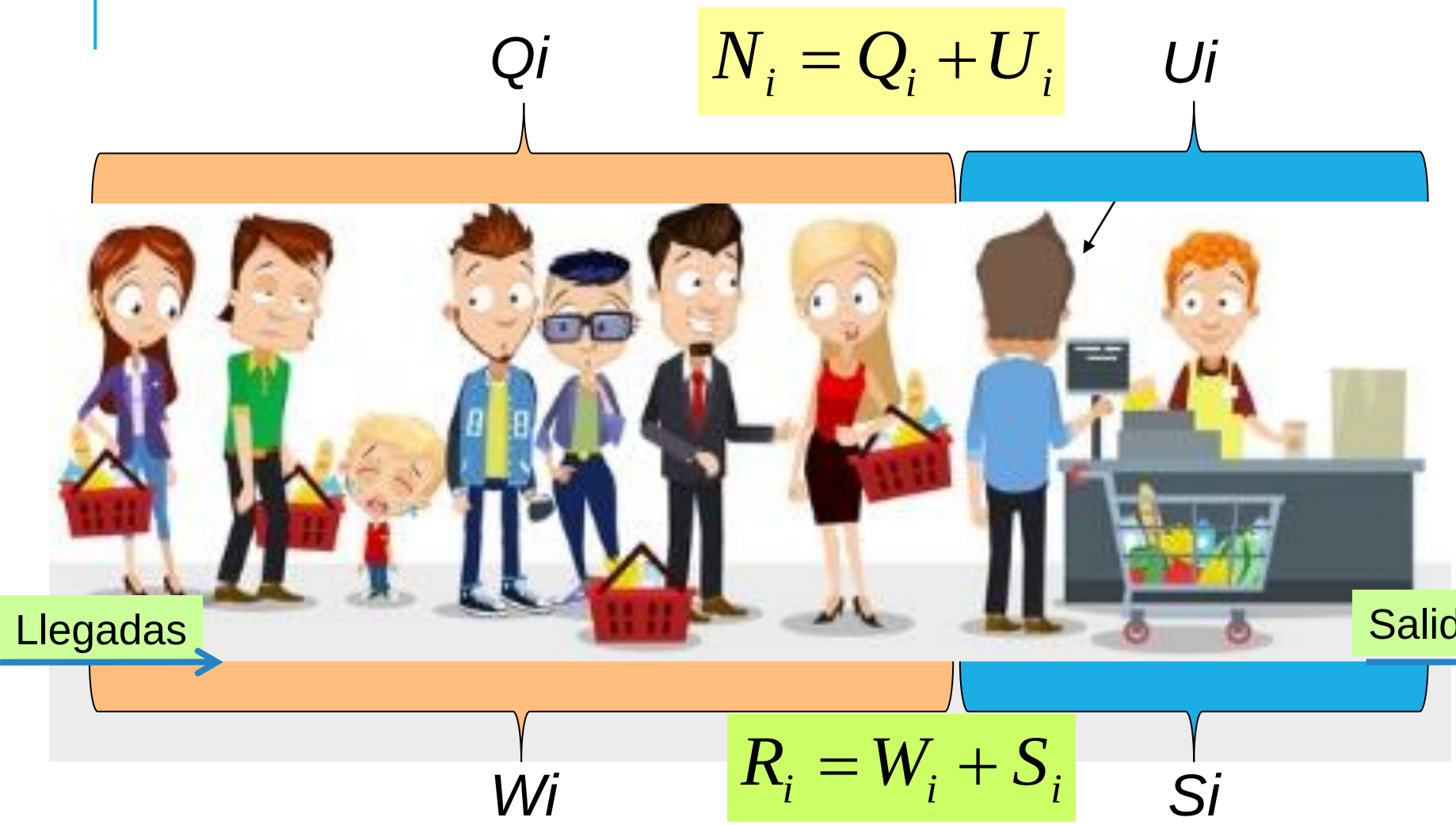
$$R_i = W_i + S_i$$



Dimensión espacial:

$$N_i = Q_i + U_i$$

TERMINOLOGÍA EN TEORÍA DE COLAS



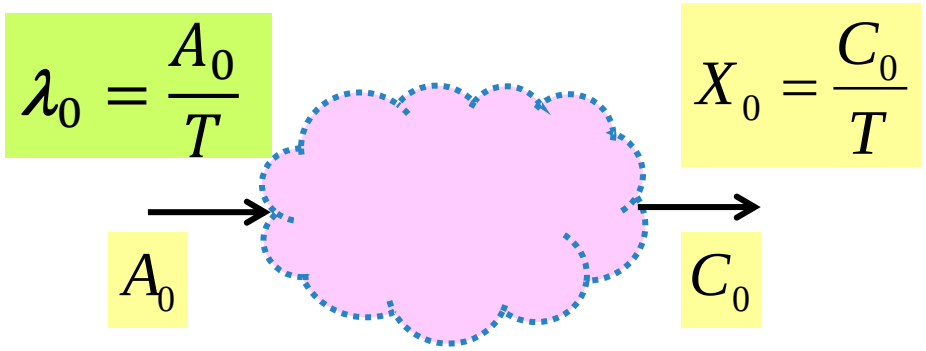
LAS VARIABLES DEL SISTEMA

Variables básicas

- A_0 Número de trabajos que llegan (*arrivals*)
- C_0 Número de trabajos que se van (*completions*)

Variables deducidas

- λ_0 Tasa de llegadas (*arrival rate*)
- X_0 Productividad (*throughput*)



ACTIVIDADES

ACTIVIDAD TEMA 4.2 (VOLUNTARIA)

Si aplicamos el modelado de colas a un hospital, que asumimos que es una cola de servicio, ¿qué pretenden decir en los medios de comunicación que el hospital se va a colapsar con una pandemia? Expresa el colapso con variables de una cola de las que hemos visto.

Si el servidor de una cola está utilizado en un % inferior al 100%, ¿significa que no hay cola de espera? Razona porqué no es así.

Puesto que los ordenadores pueden ejecutar programas en paralelo, ¿la teoría de colas no sirve para modelarlos?

4. LEYES OPERACIONALES

Relaciones entre las variables operacionales

LEYES OPERACIONALES

El valor de las variables operacionales **depende del intervalo de observación T**

- **pero las relaciones entre las variables operacionales se mantienen para cualquier intervalo de observación**

Estas relaciones se denominan leyes operacionales porque son de aplicación universal

- no dependen de suposiciones sobre distribuciones estadísticas del tiempo de servicio o del tiempo entre llegadas

HIPÓTESIS DEL EQUILIBRIO DE FLUJO

El equilibrio de flujo de trabajos

- Supone que el sistema trabaja en estado estable (estacionario, no transitorio)
- El sistema cumple el supuesto de equilibrio de flujo si para cada dispositivo:
  - La tasa de llegada coincide con la tasa de salida ($\lambda_i = X_i$), o bien,
  - El número de trabajos que llegan coincide con el que sale ($A_i = C_i$)
- Aproximación aceptable: para intervalos de observación suficientemente largos

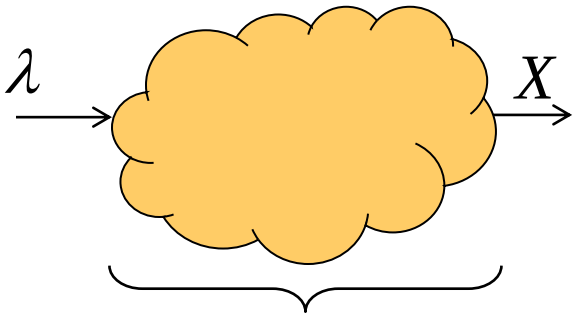
$$\left| \frac{A_i - C_i}{C_i} \right| \cong 0$$

$$\text{Si } A_i = C_i \Rightarrow \lambda_i = X_i$$

LEY DE LITTLE (1961)

Parte del cumplimiento del supuesto de equilibrio de flujo

Relaciona el número de trabajos en el sistema con el tiempo de permanencia y su productividad o tasa de llegada



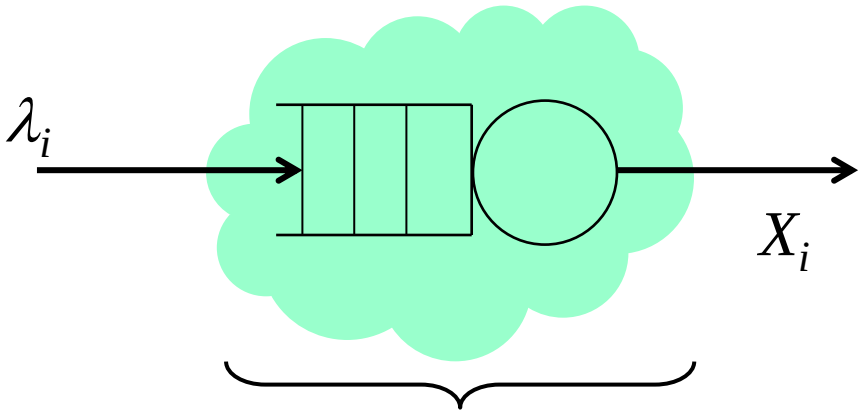
$$N = \lambda R = XR$$

R = tiempo de permanencia
N = número de trabajos

Esta ley puede ser aplicada a diferentes niveles del sistema

¿CÓMO APLICAR LA LEY DE LITTLE?

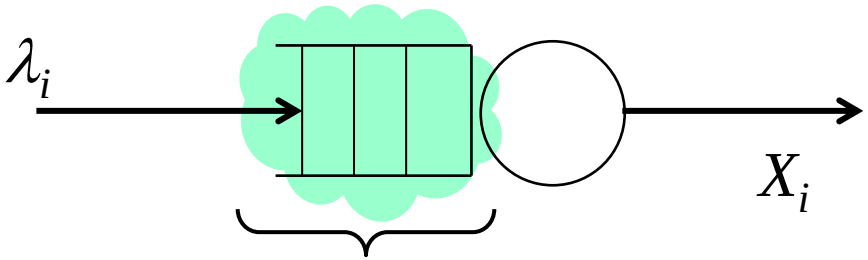
Aplicación a toda una estación de servicio



$$N_i = \lambda_i R_i = X_i R_i$$

Tiempo de respuesta: R_i
Trabajos en la estación: N_i

Aplicación a la cola de una estación de servicio



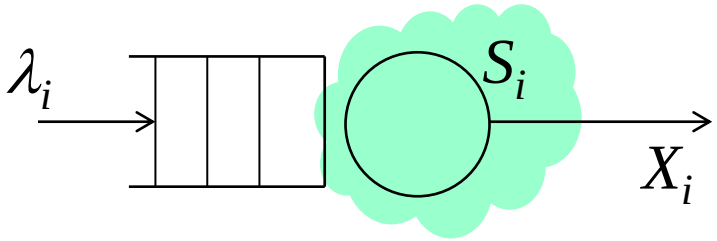
$$Q_i = \lambda_i W_i = X_i W_i$$

Tiempo de espera en cola: W_i
Trabajos en la cola: Q_i

LEY DE LA UTILIZACIÓN

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

Caso particular de la ley de Little aplicada al servidor de una estación



$$U_i = \lambda_i S_i = X_i S_i$$

EJEMPLO DE APLICACIÓN I

Como consecuencia de unas medidas sobre un sistema informático, se obtuvo que el número medio de trabajos en un disco (en espera y en servicio) fue de 1.2 y su productividad de 25 trabajos/s. Su tiempo medio de servicio (posicionamiento más latencia más transferencia) fue de 30 ms.

- Cálculo del tiempo de respuesta:

$$N_i = X_i R_i \Rightarrow R_i = \frac{N_i}{X_i} = \frac{1.2}{25} = 0.048\,\text{s} = 48\,\text{ms}$$

- Cálculo de la utilización:

$$U_i = X_i S_i = 25 \times 0.03 = 0.75$$

EJEMPLO DE APLICACIÓN I (CONTINUACIÓN)

El tiempo de respuesta (48 ms) es mayor que el tiempo de servicio (30 ms) a pesar de que la utilización no llega al 100%. Esto es debido a que el disco puede estar vacío o bien puede que lleguen trabajos cuando ya hay alguno en servicio.

- Cálculo del número de trabajos en la cola de espera

$$N_i - U_i = 1.2 - 0.75 = 0.45 \text{ trabajos}$$

- Cálculo del tiempo de espera en cola

$$\frac{N_i - U_i}{X_i} = \frac{1.2 - 0.75}{25} = 0.018\,\text{s} = 18\,\text{ms}$$

LEY DEL FLUJO FORZADO

Los flujos (productividades) a diferentes niveles del sistema tienen que ser proporcionales

- Relaciona la productividad del sistema con la de los dispositivos

$$V_i = \frac{C_i}{C_0} \Rightarrow C_i = C_0 V_i \Rightarrow \frac{C_i}{T} = \frac{C_0}{T} V_i$$

$$\Rightarrow X_i = X_0 V_i$$

- Las utilizaciones también son proporcionales a la productividad del sistema:

$$U_i = X_i S_i = X_0 V_i S_i = X_0 D_i$$

EJEMPLO DE APLICACIÓN II

En una instalación informática cada trabajo realiza una media de 5 accesos a una unidad de disco, la cual tiene una productividad de 20 accesos/s. ¿Cuál es la productividad del sistema informático?

$$X_i = X_0 V_i \Rightarrow X_0 = \frac{X_i}{V_i} = \frac{20}{5} = 4 \text{ trabajos/s}$$

Si la utilización del disco es del 40%, ¿cuál es su tiempo de servicio? ¿Y su demanda de servicio?

$$U_i = X_i S_i \Rightarrow S_i = \frac{U_i}{X_i} = \frac{0.4}{20} = 0.02 \text{ s}$$

$$D_i = V_i S_i = 5 \times 0.02 = 0.1 \text{ s}$$

LEY GENERAL DEL TIEMPO DE RESPUESTA

Es independiente del tipo de sistema (abierto o cerrado)

Solo se consideran las razones de visita y los tiempos de respuesta de cada estación

En general,

$$R \neq R_1 + R_2 + \ldots + R_K = \sum_{i=1}^{K} R_i$$

En particular,

$$R = V_1 \times R_1 + V_2 \times R_2 + \ldots + V_K \times R_K = \sum_{i=1}^{K} V_i \times R_i$$

EJEMPLO DE APLICACIÓN III

Un sistema informático dispone de dos dispositivos, 1 y 2, con los siguientes parámetros:

$$V_1 = 30; \quad R_1 = 3\,\text{ms}$$
$$V_2 = 12; \quad R_2 = 5\,\text{ms}$$

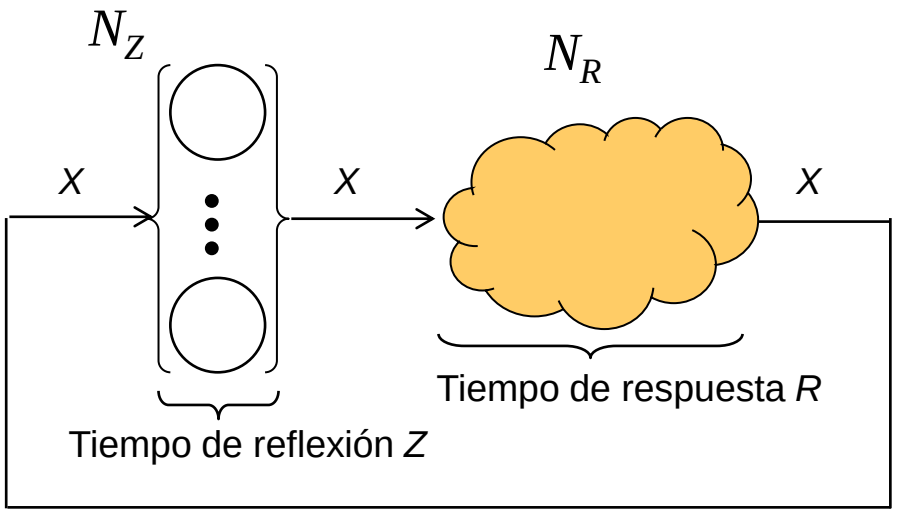
¿Cuál es su tiempo de respuesta?

$$R = \sum_{i=1}^{2} V_i \times R_i = V_1 \times R_1 + V_2 \times R_2 = 30 \times 3 + 12 \times 5 = 150\,\text{ms}$$

$$\text{Nótese que } R \neq \sum_{i=1}^{2} R_i = R_1 + R_2 = 3 + 5 = 8\,\text{ms}$$

LEY DEL TIEMPO DE RESPUESTA INTERACTIVA

Se obtiene mediante la aplicación de la ley de Little a un sistema informático cuando la carga es interactiva (Z>0) o *batch* (Z=0)



$$N_Z = XZ;\ N_R = XR$$

$$N = N_Z + N_R = XZ + XR$$

$$= X(Z + R)$$

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

EJEMPLO DE APLICACIÓN IV

Un sistema informático interactivo dispone de 30 usuarios activos (pensando o trabajando). El tiempo de reflexión es de 20 segundos y su productividad de 1 interacción/s. ¿Cuál es su tiempo de respuesta?

$$R = \left(\frac{N}{X}\right) - Z = \left(\frac{30}{1}\right) - 20 = 10 \text{ s}$$

Si se quiere conseguir un tiempo de respuesta de 2 s, ¿qué productividad debería tener el sistema?

$$X = \frac{N}{R+Z} = \frac{30}{2+20} = 1.37 \text{ interacciones/s}$$