Tema 2: Modelado Analítico

Introducción

Objetivo

Realizar una estimación cuantitativa del rendimiento en tiempo de diseño.

- Diseño de un nuevo sistema (hardware y/o software).
- Planificación de la capacidad y ajuste de sistemas.

Técnica

Construcción, parametrización y evaluación de un modelo matemático del sistema observado.

Modelo matemático de un sistema informático

Construcción teórica que permitirá la representación del sistema en evaluación y que podrá ser utilizado para predecir, operar y controlar su comportamiento en diferentes escenarios.

Variables

- Estado: Determinan en qué situación se encuentra el sistema.
- Entrada: Permiten modificar la evolución del sistema.
- Salida: Describen la respuesta del sistema.

Parámetros: Valores que permites establecer las relaciones entre variables.

Relaciones funcionales: Ecuaciones o inecuaciones que determinan el comportamiento y evolución del sistema.

Teoría de colas

Estudia el comportamiento de los sistemas que se caracterizan por:

- Llegada (aleatoria) de solicitudes de servicio a un sistema.
- Las solicitudes son atendidas en funci ́on de un mecanismo establecido.
- Las que no pueden ser atendidas de inmediato quedan a la espera de recibir el servicio formando una l ínea de espera o cola.

Modelado de redes de colas

Centro de servicio

Conjunto compuesto por servidor(es) y cola de espera.

Tipos:

- Servidor único y una cola de espera.
- Varios servidores y una cola de espera.

- Infinitos servidores sin cola de espera.
    - Los trabajos siempre encuentran un servidor disponible.
    - Se denominan: Centro de retardo o demora.

Red de colas

Una red de colas (**QN**) es una colección de *K* centros de servicio conectados que proporcionan servicio a un conjunto de clientes.

Red abierta: Llegadas y salidas externas. El número de trabajos varía con el tiempo.

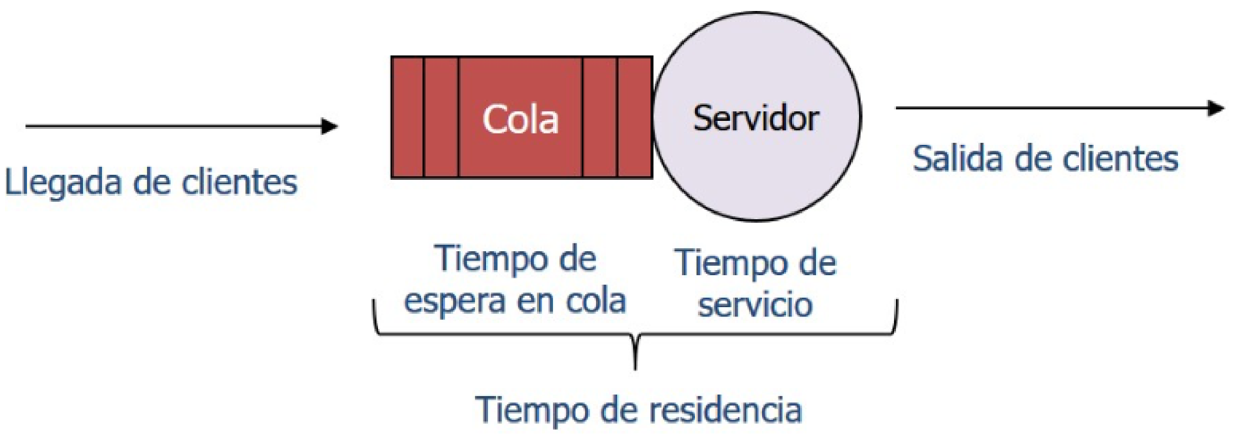
Red cerrada: Número fijo de trabajos en el sistema.

La diferencia entre ambas es la constancia de los trabajos en el sistema. No es lo mismo tener una tasa de llegadas constante que tener un número de trabajos constante en el sistema.

El nivel de **concurrencia** en ambas redes es distinto (en las **redes abiertas no tiene por qué ser constante**, mientras que en **redes cerradas sí que lo es**).

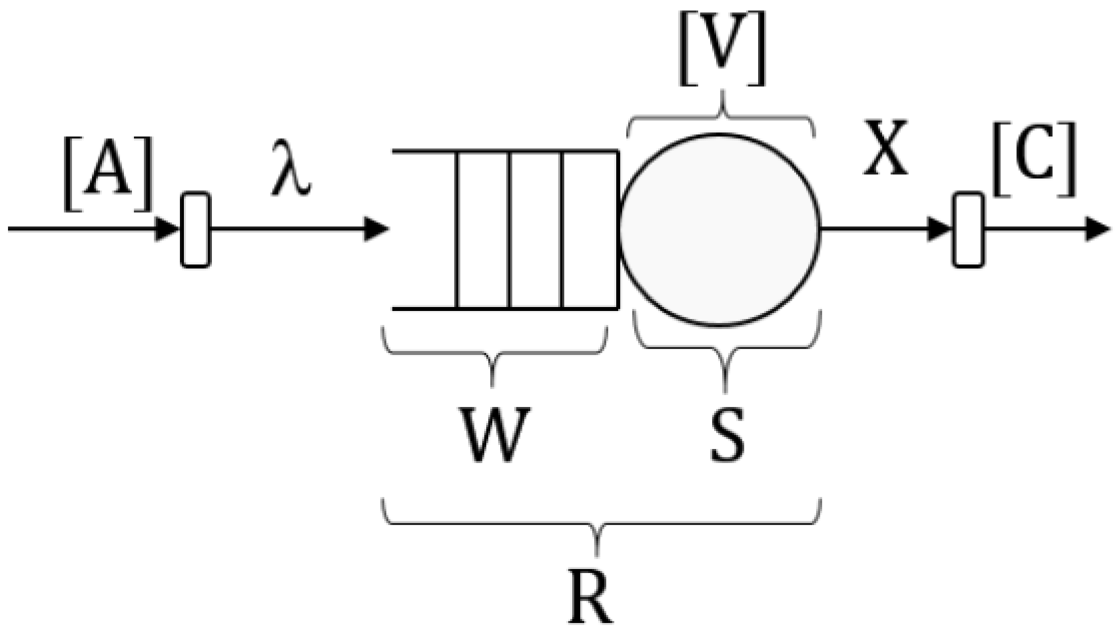
Modelo de servidor central

- Patrón de conexión entre los centros de servicio.
- Las entradas y salidas al sistema son hacia y desde el servidor central.
- Los flujos de trabajos en la red desde y hacia el servidor central.

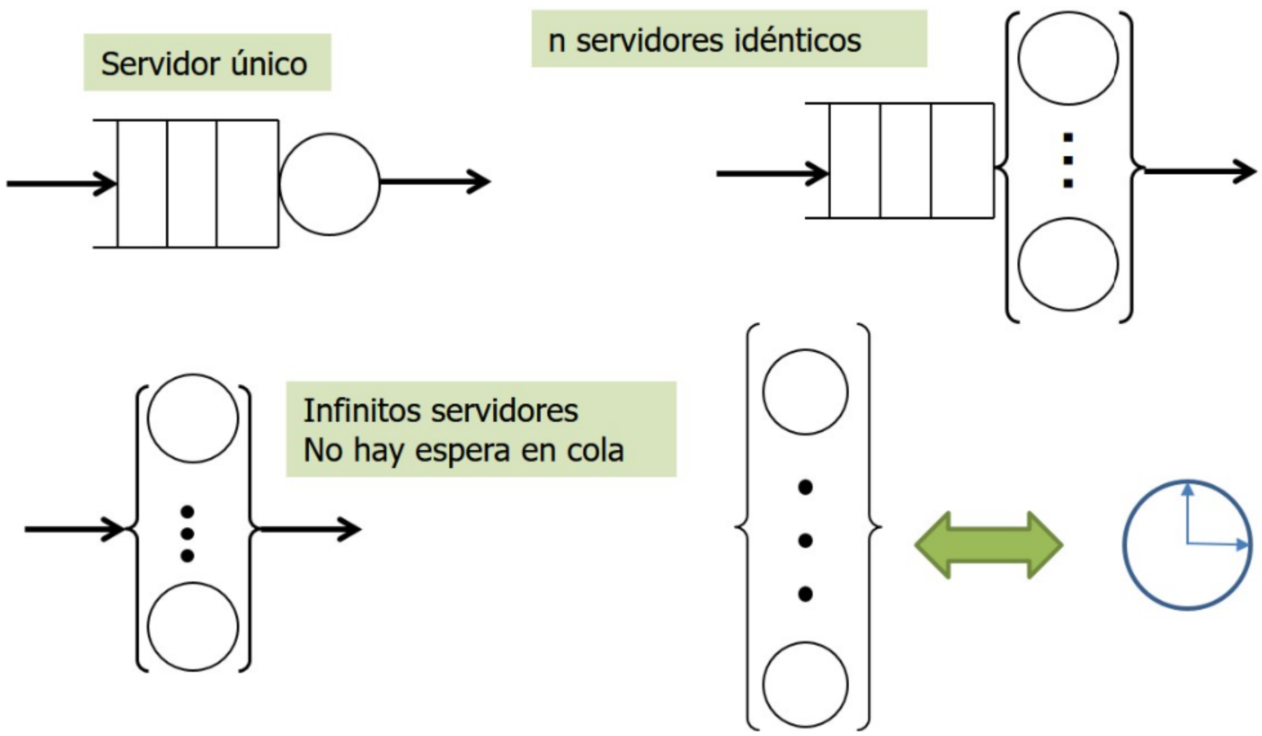


Variables del modelo de servidor central

- [A]: Llegadas.
- [V]: Visitas.
- [C]: Finalizaciones.
- Lambda: Tasa media de llegadas (tiempo de llegada es 1/lambda).
- W: Tiempo de espera en cola.
- S: Tiempo de servicio.
- R: Tiempo de residencia (= W + S).
- X: Productividad o Throughput.

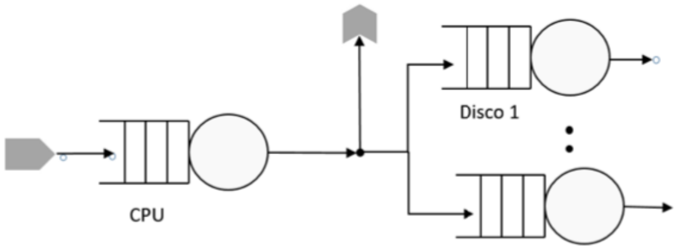


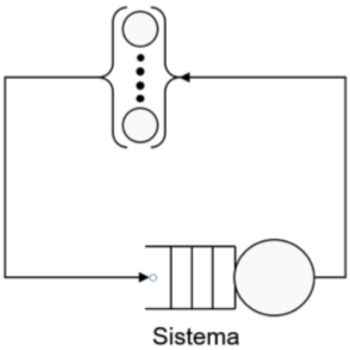
Tipos de centro de servicio

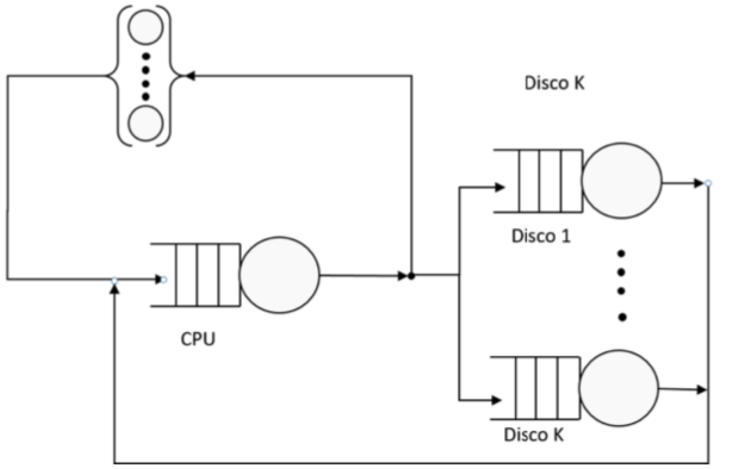


Para nosotros el más común será el de servidor único.

Tipos de red de colas







Modelo de redes de colas

Debemos describir 2 categorías en los parámetros del modelo.

1. Descripción de los centros de servicio
   - Tiempo de servicio.
   - Número de servidores.
   - Capacidad (Número máximo de clientes en el centro de servicio): Trabajaremos con capacidad infinita.
   - Disciplina de servicio: FCFS (First Come Firt Serve)
2. Descripción de los clientes (carga de trabajo a procesar)
   - Intensidad. Tasa a la que llegan los clientes al sistema.
   - Demanda de servicio. Tiempo total de servicio proporcionado por un recurso a una determinada clase de trabajos (Media).

Modelado analítico

Objetivo: Representar el comportamiento de los componentes del sistema y de la carga a procesar.

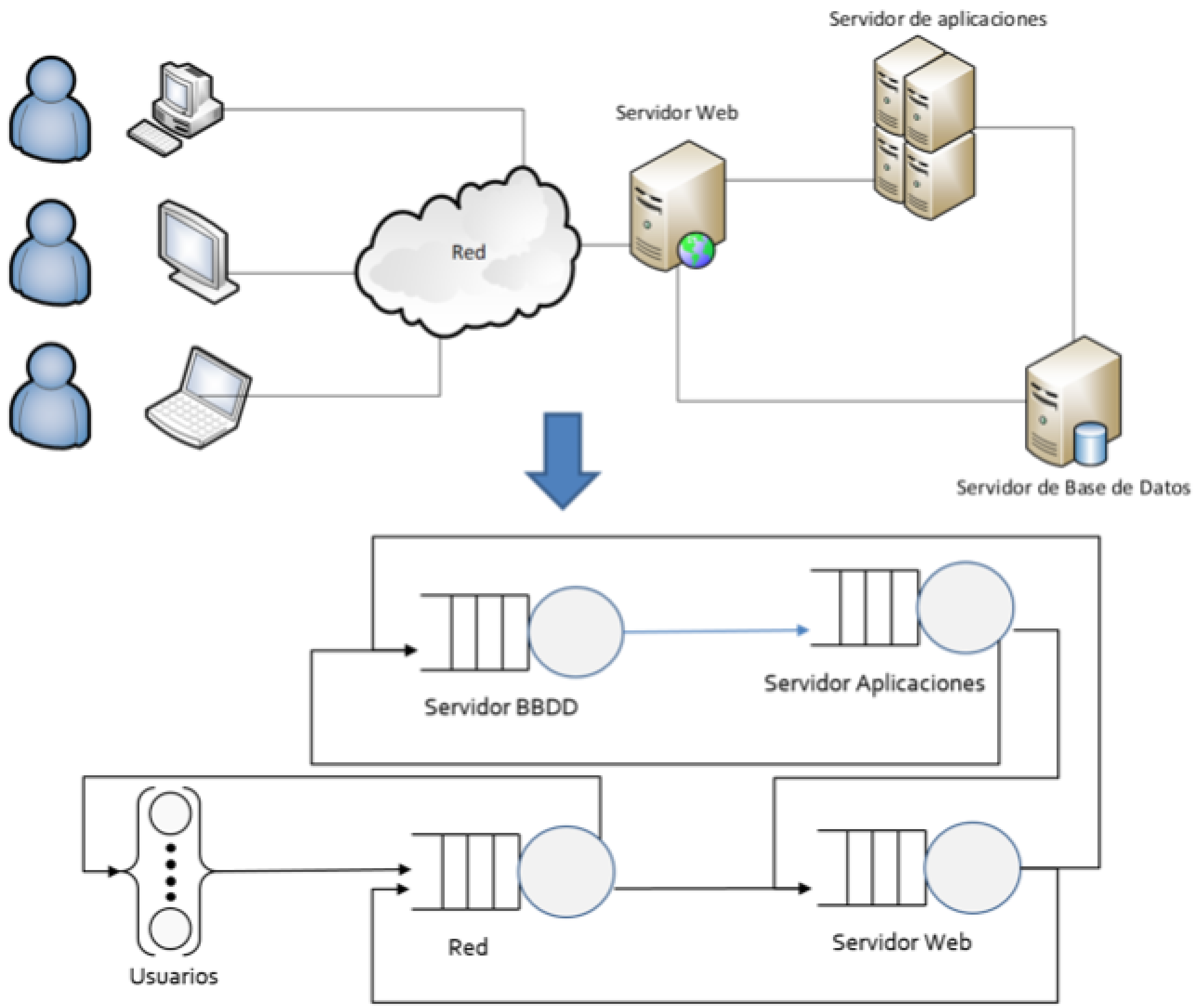
Sistema: Determinado por el nivel de abstracción al que se representa el sistema y los aspectos a considerar (Sistema de computación, de información, ...).

Clientes: Los trabajos (transacciones, jobs, interacciones, ...) a procesar por el sistema.

Servicio: El proporcionado por los diferentes subsistemas (memoria, procesador, ES, ...).

Cola: Abstracción de un buffer.

Representación gráfica de las conexiones entre dispositivos



Sistemas de computación

Centros de servicio: Conjunto de subsistemas individuales conectados que representan recursos del sistema.

Tipos de recurso por tasa de servicio:

- Intependientes de la carga.
- Dependientes de la carga.
- Centro de demora.

Disciplina del proceso de servicio: **FCFS**, Cola con prioridad, Round Robin, ...

Modelado de la carga. Clientes: Trabajos a ser ejecutados por el sistema. Compiten por los recursos del mismo.

Intensidad de carga:

- Transacciones. Clase abierta.
  - Tasa de llegadas.
- Batch y terminales. Clase cerrada.
  - Batch. El número de peticiones concurrentes que están en ejecución.
  - Interactiva. Número de terminales o máquinas clientes y por el tiempo medio de pensar.

Demanda de servicio.

Clases de clientes: Una clase de trabajos, Múltiples clases de trabajos (se simplifica con la media).

Análisis de colas

Análisis operacional: La elaboración y resolución del modelo se lleva a cabo usando variables operacionales, variables medidas en un intervalo de observación.

- Proporciona una visión de alto nivel del comportamiento, en términos de valores medios, del sistema.
- Hipótesis que pueden verificarse mediante medidas directas (operacionales).

Análisis estocástico La elaboración y resolución del modelo se lleva a cabo teniendo en cuenta que los tiempos entre llegadas de los trabajos y los tiempos de servicio se ajustan a ciertas distribuciones de probabilidad.

- Permite proprocionar respuestas más detalladas
- El comportamiento del sistema se específica en función de la distribución del número de trabajos en el sistema en el instante de tiempo t.

Condiciones

- Balance de flujo de trabajos: Para un período de observación suficientemente grande el número de trabajos que llegan a un sistema ha de ser igual al número de trabajos que salen.
- One-step behavior: Ni movimientos simultáneos de trabajos entre servidores, ni entradas y/o salidas simultáneas al sistema.
- Homogeneidad: La tasa media de llegadas al sistema y la tasa media de salidas del sistema son independientes del estado del sistema. El enrutamiento de los trabajos es independiente del estado del sistema.
- Exclusividad: Un trabajo no puede estar en dos o más dispositivos al mismo tiempo. Cuándo un trabajo está recibiendo servicio, tiene uso exclusivo del servidor.
- No bloqueo: Los dispositivos no pueden ser bloqueados por otros.
- Independencia: Los trabajos no puede interaccionar de ninguna forma. No se permite la sincronización de trabajos.

Consideraciones

Una red de colas es *separable* o *product-form* si cumple las condiciones anteriores y además cumple la condición de balance de flujo de trabajo para cada centro de servicio.

- Facilita la resolución de las redes de colas.
- Se puede evaluar la solución para cada centro de servicio de forma aislada.

Centros independientes de la carga o fixed-capacity service centers: La tasa de finalizaciones independiente de la longitud de la cola (homogeneidad).

Centros dependientes de la carga: La tasa de finalizaciones dependiente de la longitud de la cola (homogeneidad débil).

Modelo de red de colas

Variables operacionales: Variables cuyos valores se pueden obtener mediante medida directa del sistema durante un intervalo de observación finito.

- T (segundos). Es el tiempo, intervalo de observación o medida del sistema.
- K. Número de recursos (centros de servicio) en el sistema.
- A. Número de llegadas de peticiones al sistema en el tiempo T.
- A_k. Número de solicitudes de servicio al recurso *k* en el tiempo T.
- C. Número de peticiones completadas por el sistema en el tiempo T.
- C_k. Número de finalizaciones de servicio en el recurso *k* en el de tiempo T.
- V_k. Número de visitas al servidor k en el intervale de tiempo T.
- B_k. Tiempo en el que el recurso k ha estado ocupado durante el tiempo T.

Intensidad de carga

Tasa de llegadas: Mide las llegadas al sistema de los clientes o peticiones por unidad de tiempo.

$$\lambda = \frac{A}{T}$$

Balance de flujo de trabajo: Bajo la hipótesis del equilibrio de flujo, y para un T suficientemente grande se verifica que A es casi igual a C. Dividiendo ambos valores por el valor del período de medida T, se obtiene,

$$\lambda = X$$

Notas

- Un sistema está en estado estacionario si Lambda = X.
- Intervalo entre llegadas: 1/Lambda.

Medidas básicas de rendimiento

Productividad: Mide la cantidad de trabajo realizado por un elemento por unidad de tiempo.

- Productividad del sistema, throughput, X.

$$X = \frac{C}{T}$$

- Productividad de cada centro de servicio, throughput local, X_k.

$$X_k = \frac{C_k}{T}$$

- Productividad relativa. Relación de proporcionalidad que se da entre las finalizaciones globales y locales cuando el sistema está en estado estacionario.

$$(C_k \propto C)$$

Ley de flujo forzado

- Ratio de visitas V_k: Constante de proporcionalidad que indica la relación entre el número de operaciones locales y globales ejecutadas durante el periodo de observación T.

Partiendo del concepto de productividad relativa y de la definición de ratio de visitas V_k, se puede expresar la relación anterior cómo $C_k = V_k C$, y dividiendo ambos términos por T se obtiene,

$$X_k = \frac{C_k}{T} = \frac{V_k C}{T} = V_k X \text{ para } k = 1, \ldots, K$$

- Ley del flujo forzado. $X_k = V_k X$. Establece la relación entre la productividad del sistema y la de cada uno de sus dispositivos.
- Tiempo de servicio. Es el tiempo medio necesario para dar servicio a una petición en el centro de servicio k. Tiempo medio durante el que un cliente ocupa el servidor.

$$S_k = \frac{B_k}{C_k}$$

- Tasa de servicio. Es la inversa del tiempo de servicio.

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

- Demanda de servicio. Cantidad de servicio que cada petición a nivel de sistema requiere de cada uno de sus recursos.

$$D_k = \frac{\textit{tiempo medio dispositivo ocupado}}{\textit{número finalizaciones sistema}} = \frac{B_k}{C}$$

multiplicando y dividiendo por C_k se obtiene,

$$D_k = \frac{B_k}{C} = \frac{B_k}{C_k} \times \frac{C_k}{C} = S_k.V_k$$

- Utilización. Es la proporción del tiempo de ocupación del servidor en relación al periodo de observación T, y su expresión es:

$$U_k = \frac{B_k}{T}$$

```
- Notación. En teoría de colas lo habitual es denotar la utilización por
*p*.
- Observación. Teniendo en cuenta que B<sub>k</sub> <= T se tendrá que 0
<= U<sub>k</sub> <= 1.
```

Ley de la utilización: La utilización de un recurso es igual al producto del *throughput* por el tiempo medio de servicio $U_k = X_k S_k$.

$$U_k = \frac{B_k}{T} = \frac{C_k}{T} \times \frac{B_k}{C_k} = X_k.S_k$$

- Relaciones en utilización:

- En equilibrio de flujo, $A \approx C$, por lo que, $\lambda \approx X$, y por lo tanto, $U = \lambda S$
- Un recurso se dice que está **saturado** cuando ha estado ocupado durante todo el intervalo de observación ($U = 1$). En estas condiciones,

$$U = 1 = XS \Rightarrow X = \frac{1}{S}$$

- A partir de las leyes de la utilización y del flujo forzado se tiene que:

$$U_k = X_k S_k = X V_k S_k = X D_k$$

es decir, $U_k \propto D_k$, lo que permite identificar el *dispositivo cuello de botella*.
- De la definición de la tasa de servicio $\mu = 1/S$, la ley de la utilización $U = \lambda S$, se puede escribir como:

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

(en algunos textos se conoce como *intensidad de tráfico*).

- Tiempo de residencia. Es el tiempo medio que una petición consume en el centro de servicio. Es la suma del tiempo medio de espera en la cola W_k con el tiempo medio de servicio S_k

$$R_k = W_k + S_k$$

- Tiempo de respuesta. El tiempo medio de respuesta es la suma de los tiempos medios de residencia en cada uno de los centros k de la red de colas.

$$R = \sum_{k=1}^{K} V_k R_k$$

Nota. En esta definición R_k es el tiempo consumido en una visita. El modelo está parametrizado en términos de V_k y S_k.

Ley de Little

La longitud media de la cola *Q* es igual a la tasa de llegadas media *Lamba* por el tiempo consumido en el centro cola, el tiempo medio de residencia *R*.

$$Q = \lambda \times R$$

- La media hace referencia a la *media en el tiempo*
- Teniendo en cuenta la condición de equilibrio de flujo, $\lambda \simeq X$, la ley de Little se puede escribir como $Q = XR$ o $Q_k = X_k R_k$ si queremos hacer referencia a un centro k concreto.
- Aplicable a cualquier nivel de detalle del sistema, desde el nivel de recurso aislado, sin incluir la cola del recurso, hasta el sistema completo.

Niveles del sistema

- Nivel 1. Recurso sencillo sin incluir cola.
  - Población. Trabajos presentes en el servidor (Una petición el U_k(%) del tiempo y cero peticiones el 1 - U_k(%) del tiempo). *Q es equivalente a U*.
  - Tasa de llegadas. Tasa con la que el recurso satisface las peticiones (X_k).
  - Tiempo medio de servicio para las peticiones en el recurso. *R_k equivalente a S_k*.
  - Ley de Little. U_k equilavente a $X_k S_k$ = *Lambda*S. Se obtiene la ley de la utilización como un caso particular de la Ley de Little.
- Nivel 2. Centro de servicio.
  - Población. Trabajos presentes en la cola y en el servidor.
  - Tasa de llegadas. Tasa con la que llegan las peticiones al recurso (*$Lamba_k$*).
  - Tiempo de residencia. Tiempo medio que la petición consume por cada visita al recurso (incluye el tiempo en cola y en servicio). *R_k= W_k + S_k*.
  - Ley de Little. *Q = Lambda(W+S) = LambdaW + U*.
- Nivel 3. Subsistema central, sin los terminales.
  - Población. Clientes presentes en el subsistema central, aquellos usuarios que no están pensando.
  - Tasa de llegadas. Tasa de interacciones entre el subsistema central y los terminales.
  - Tiempo de residencia. La definición habitual de tiempo de respuesta.
  - Ley de Little. *Q = LambdaR*

Análisis de Rendimiento

Elementos para la evaluación del modelo

Variables operacionales: Valores medios de cantidades que se pueden medir directamente.

Leyes operacionales: Relaciones entre variables operacionales. Permiten la caracterización del comportamiento del sistema.

Estimación del rendimiento del sistema.

- Soportado por las medias de valores observados: Buena aproximación. El grado de simplificación es bastante notable.
- Soportado por procesos aleatorios o estocásticos. Modelado más preciso del comportamiento.

Configuraciones

Centros:

- Un centro de Servicio.
  - Servidor único. Análisis a nivel de sistema o a nivel de componente aislado.
  - Multiservidor. Análisis de sistemas multiprocesador, p.ej.
- Red de colas. Múltiples centros de servicio. Análisis detallado del rendimiento, incorpora las interacciones entre los elementos.

Número de clientes: No limitado (**Cola abierta**) vs limitado (**Cola cerrada**).

Caracterización de la carga: Una clase de clientes vs múltiples clases de clientes.

Un centro de servicio. Servidor único

Cola Abierta

Condiciones

- Número limitado de clientes.
- Condición de estabilidad (se evitan colas infinitas):
- Tasa de llegadas de transacciones independiente de la carga del sistema.
- Distribuciones exponenciales
  - Tiempo entre llegadas: El tiempo que transcurre hasta la próxima llegada es independiente del instante en el que se produjo la última.
  - Tiempos de servicio: El tiempo residual que le queda a un cliente para finalizar su servicio es independiente del tiempo que ya lleva en servicio.
- Centro de servicio independiente de carga.
- Una clase de clientes.

Tiempo de residencia en cola: *R = S(Q+1)*

- Longitud media de la cola Q. La espera estimada será de QS segundos.
- Tiempo medio de servicio S. La espera estimada será de S segundos.

Aplicando la Ley de Little, *Q = Lambda * R*, a la expresión anterior tenemos

$$R = \frac{S}{1 - \lambda S}$$

y por la ley de la utilización *U = Lambda * S*, se tiene

$$R = \frac{S}{1 - U}$$

Partiendo de la expresión del tiempo de residencia, multiplicando por *Lambda* ambos lados de la ecuación, se obtiene:

$$R = \frac{S}{1 - U} \quad Q = \frac{U}{1 - U}$$

El tiempo de espera en cola estimado viene dado por *Q * S*,

$$W = QS = \frac{US}{1-U}$$

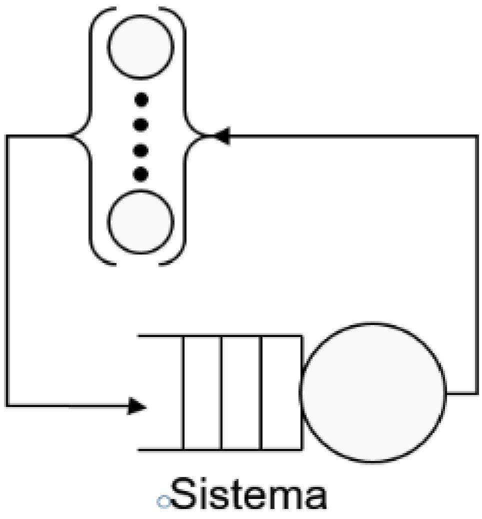
lo que permitirá calcular el número medio de trabajos esperando en cola,

$$L = \lambda W = \frac{U^2}{1-U}$$

Cola cerrada

Condiciones

- Número limitado de clientes fijo y finito (N).
- Clientes en una de dos situaciones
  - Estado tiempo de pensar (Z): Esperando a incorporarse a la cola.
  - En el centro de servicio.
- Centro de servicio independiente de carga.
- Distribución exponencial para los tiempos de servicio.
- Una clase de clientes.



El tiempo de respuesta R y la productividad X dependen del número de clientes N (Efecto retroalimentación)

- A medida que el centro está más ocupado, la tasa a que se va ocupando se reduce.
- Si los N clientes están en el sistema, no hay nuevas llegadas.

Ley de tiempo de respuesta interactivo: Cada usuario, por término medio, *T/(R+Z)* peticiones en el intervalo de tiempo T

$$X = \frac{numero\ peticiones}{intervalo\ tiempo} = \frac{N[T/(R+Z)]}{T} = \frac{N}{R+Z}$$

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

Modelos de redes de colas. Una clase de trabajos

Análisis de redes abiertas

Condiciones

- Distribución exponencial para tiempos de llegada y de servicio.
- Todos los centros de servicio independiente de carga.
- Una clase de clientes.
- Red de colas separable (product-form)

Entradas

- Intensidad de carga: *Lambda*.
- Demandas de Servicio: V_k, S_k

Evaluación del modelo

- Métricas de rendimiento: $$X(\lambda), X_k(\lambda), U_k(\lambda), R((\lambda), R_k(\lambda), Q_k(\lambda)$$

Métricas de rendimiento

- Throughput del sistema (estamos asumiendo equilibrio de flujos): $X = Lambda$.
- El throughput del dispositivo k, X_k, por la Ley del flujo forzado: $X_k = Lambda * V_k$.
- Utilización del dispositivo k, U_k, por la Ley de la Utilización: $U_k = X_k * S_k = X * V_k * S_k = Lambda * D_k$.
- Número de trabajos en el dispositivo k, Q_k, aplicando la Ley de Little será: $$Q_k = X_k R_k = X_k S_k (1 + Q_k) = U_k (1 + Q_k)$$

  o

  $$Q_k = \frac{U_k}{1 - U_k}$$
- Tiempo de residencia de cada dispositivo R_k(*Lambda*), sustituyendo $Q_k = U_k / (1 - U_k)$ en la ecuación $R_k = S_k + (1 + Q_k)$, se obtiene,

$$R_k = \frac{S_k}{1 - U_k}$$

- En los centros de demora el tiempo de residencia es igual al tiempo de servicio, $R_k = S_k$ por lo tanto, $Q_k = R_k * X_k = S_k * X * V_k = X * D_k = U_k$.
- El número de trabajos en el sistema

$$Q = \sum_{i=1}^{K} Q_k$$

- El tiempo de respuesta del sistema

$$R = \sum_{i=1}^{K} R_k V_k$$

Algoritmo de redes abiertas

$$\begin{aligned}
&\textbf{input}\quad\textbf{:}\\
&\qquad\qquad \lambda = \text{ tasa de llegadas al sistema}\\
&\qquad\qquad K = \text{ n}^{\text{o}} \text{ de dispositivos - centros de servicio}\\
&\qquad\qquad S_k = \text{ tiempo de servicio por visita en el i-simo dispositivo}\\
&\qquad\qquad V_k = \text{ número de visitas en el i-simo dispositivo}\\
&\textbf{output:}\\
&\qquad\qquad Q_k(\lambda) = \text{ número medio de trabajos en el dispositivo i-simo}\\
&\qquad\qquad R_k(\lambda) = \text{ tiempo de residencia en el i-simo dispositivo}\\
&\qquad\qquad U_k(\lambda) = \text{ utilización del dispositivo i-simo}\\
&\qquad\qquad R(\lambda) = \text{ tiempo de respuesta del sistema}\\
&\qquad\qquad Q(\lambda) = \text{ número medio de trabajos en el sistema}
\end{aligned}$$

$$\begin{aligned}
&\text{Para cada dispositivo se calcula}\\
&\ \text{Demanda de los dispositivos: } D_k = V_k S_k\\
&\ \text{Utilización de los dispositivos: } U_k = X D_k\\
&\ \text{Throughput de los dispositivos: } U_k = X D_k
\end{aligned}$$

$$\begin{aligned}
&\text{Demanda de los dispositivos: } D_k = V_k S_k\\
&\ \text{Utilización de los dispositivos: } U_k = X D_k\\
&\ \text{Throughput de los dispositivos: } U_k = X D_k
\end{aligned}$$

$$\text{Tiempos de respuesta de cada dispositivo:}$$

$$R_k = \begin{cases} \frac{S_k}{1-U_k} & \text{centros de capacidad fija}\\ S_k & \text{centros de demora}\end{cases}$$

$$\text{Número de trabajos en cada dispositivo (longitud de la cola):}$$

$$Q_k = \begin{cases} \frac{U_k}{1-U_k} & \text{centros independientes de carga o de capacidad fija}\\ U_k & \text{centros de demora}\end{cases}$$

$$\text{Tiempo de respuesta del sistema: } R = \sum_{i=1}^{K} R_k V_k$$

$$\text{Número de trabajos en el sistema: } Q = \sum_{i=1}^{K} Q_k$$

Análisis de redes cerradas

Condiciones

- Distribución exponencial para tiempos de servicio.
- Todos los centros de servicio independientes de carga.
- Una clase de clientes.
- Red de colas separable (product-form).
- Nodo de demora. Sustituye a los nodos fuente y sumidero.

Entradas

- Intensidad de la carga: N (número de usuarios o trabajos).
- Demandas de Servicio: V_k, S_k.
- Tiempo de pensar: Z.

Evaluación del modelo

- Métricas de rendimiento:

$$X(N), X_k(N), U_k(N), R(N), R_k(N)$$

Análisis del Valor Medio. Solución exacta.

Algoritmo iterativo

Calcula el tiempo de respuesta en cada centro a partir del número de peticiones en cada centro obtenido en la iteración anterior. Utiliza la siguiente ecuación (no operacional)

$$R_k(N) = S_k[1 + Q_k(N-1)]$$

$Q_k(N-1)$ es el número medio de trabajos en el dispositivo k con $N-1$ trabajos en la red

Calcula el tiempo de respuesta del sistema a partir de los tiempos de respuesta obtenidos para cada centro.

Calcula el *throughput* del sistema aplicando la ley del tiempo de respuesta.

Calcula el nuevo número de peticiones en cada centro aplicando la ley de Little.

$$
\begin{array}{l}
\textbf{input} \quad \textbf{:} \\
\qquad N = \text{ número de usuarios} \\
\qquad Z = \text{ tiempo de pensar} \\
\qquad K = \\
\qquad \text{ n}^{\text{o}} \text{ de dispositivos - centros de servicio sin incluir terminales} \\
\qquad S_k = \text{ tiempo de servicio por visita en el i-simo dispositivo} \\
\qquad V_k = \text{ número de visitas en el i-simo dispositivo} \\
\textbf{output:} \\
\qquad X = \text{ throughput del sistema} \\
\qquad Q_k = \text{ número medio de trabajos en el dispositivo i-simo} \\
\qquad R_k = \text{ tiempo de respuesta en el dispositivo i-simo} \\
\qquad R = \text{ tiempo de respuesta del sistema} \\
\qquad U_k = \text{ utilización del dispositivo i-simo}
\end{array}
$$

inicialization: **for** i=*1* **to** K **do** $Q_k(0) = 0$;

iteration;

for n=*1* **to** N **do**

begin

  for i=*1* **to** K **do**

  begin

$$R_k(n) = \begin{cases} S_k(1 + Q_k) & \text{centros de capacidad fija} \\ S_k & \text{centros de demora} \end{cases}$$

  end

$$R(n) = \sum_{i=1}^{K} R_k(n) V_k$$

$$X(n) = \frac{n}{Z + R(n)}$$

  for i=*1* **to** K **do**

  begin

$$X_k(n) = X(n) V_k$$

$$Q_k(n) = X_k(n) R_k(n) = X(n) V_k R_k(n)$$

$$U_k(n) = X(n) S_k V_k$$

  end

end

Análisis del Valor Medio. Solución aproximada.

Aproximación de Schweitzer

- Evitar la recursión en valores grandes de N
- Estima la longitud de las colas con N trabajos.
  Con los valores calculados se pueden recalcular las longitudes de las colas
- Supuesto: *Si el número de trabajos en la red se incrementa el número de trabajos en cola en cada dispositivo se incrementa de forma proporcional.*

$$\frac{Q_k(N)}{N} = a_k \text{ (constante) } \forall N$$

en particular esto implica que:

$$\frac{Q_k(N-1)}{N-1} = \frac{Q_k(N)}{N}$$

que es equivalente a:

$$Q_k(N-1) = \frac{N-1}{N} Q_k(N)$$

- Cada iteración
  - Comienza con unos valores para el número de trabajos $Q_k(N)$
  - Finaliza recalculando nuevos valores para $Q_k(N)$
- El criterio de parada es un criterio de convergencia. Diferencia entre los valores al inicio y al final de la iteración
- Elección de valores iniciales para las longitudes de las colas
  - No deberían afectar al resultado final
  - Puede afectar al número de iteraciones

Criterio de inicialización propuesto
Comenzar con todas las longitudes de colas iguales.

$$Q_k = \frac{N}{K}$$

Teniendo en cuenta la ecuación

$$Q_k(N-1) = \frac{N-1}{N} Q_k(N)$$

las ecuaciones MVA se reescriben como,

$$R_k(N) = \begin{cases} S_k(1 + \frac{N-1}{N} Q_k(N)) & \text{centros de capacidad fija} \\ S_k & \text{centros de demora} \end{cases}$$

$$X(N) = \frac{N}{Z + \sum V_k R_k(N)}$$

$$Q_k(N) = X(N) V_k R_k(N)$$

input :

$N =$ número de usuarios
$Z =$ tiempo de pensar
$K =$
nº de dispositivos - centros de servicio sin incluir terminales
$S_k =$ tiempo de servicio por visita en el i-simo dispositivo
$V_k =$ número de visitas en el i-simo dispositivo
$\epsilon =$ error máximo admisible en la longitud de la cola

output:

$X =$ throughput del sistema
$Q_k =$ número medio de trabajos en el dispositivo i-simo
$R_k =$ tiempo de respuesta en el i-simo dispositivo
$R =$ tiempo de respuesta del sistema
$U_k =$ utilización del dispositivo 1-simo

inicialization;
$X = 0$;
for *i=1* **to** K **do** $Q_k = N/K$;

iteration;

while $max_k|Q_k - XR_kV_k| > \epsilon$ **do**
begin
  for *i=1* **to** K **do**
  begin

$$R_k = \begin{cases} S_k(1 + \frac{N-1}{N}Q_k) & \text{centros de capacidad fija} \\ S_k & \text{centros de demora} \end{cases}$$

  end

$$R = \sum_{i=1}^{K} R_kV_k$$

$$X = \frac{N}{Z+R}$$

  for *i=1* **to** K **do**
  begin

$$Q_k = XV_kR_k$$

$$X_k = XV_k$$

$$U_k = XS_kV_k$$

  end
end