Un dibujo animado con letras

Descripción generada automáticamente con confianza bajaUniversidad Tecnológica Nacional

Facultad Regional Rosario

Especialidad: Ing. En Sistemas de Información.

Asignatura: Comunicación de datos.

Comisión: 302

Fecha: 19/08/2024

INFORME N°4 – Teoría de colas

|  |  |  |
| --- | --- | --- |
| **Alumno** | **Legajo** | **Firma** |
| DEL POPOLO, Chiara | 49941 |  |
| MONDINO, Juan Cruz | 51922 |  |
| MORENO, Natan | 47962 |  |
| PONTELLI, Juan Martín | 51315 |  |

Contenido

[Glosario 3](#_Toc174978953)

[Introducción 4](#_Toc174978954)

[Elementos en la teoría de colas 8](#_Toc174978955)

[Desarrollo 11](#_Toc174978956)

[Procesos de Poisson 12](#_Toc174978957)

[Conceptos previos para el desarrollo de los Procesos de Poisson: 12](#_Toc174978958)

[Definición de un proceso de Poisson 14](#_Toc174978959)

[Distribución de las llegadas en un proceso de Poisson 16](#_Toc174978960)

[Propiedades de los Procesos de Poisson: 17](#_Toc174978961)

[Notación de Kendall 19](#_Toc174978962)

[Ecuaciones de Little 20](#_Toc174978963)

[Aplicaciones de la Ley de Little 23](#_Toc174978964)

[Cadenas de Markov 24](#_Toc174978965)

[Probabilidades de transición 25](#_Toc174978966)

[Cadenas de Markov de tiempo continuo 25](#_Toc174978967)

[Ecuación de futuro 27](#_Toc174978968)

[Procesos de nacimiento y muerte 27](#_Toc174978969)

[Procesos de nacimiento y muerte en régimen permanente 29](#_Toc174978970)

[Conceptos de tráfico 30](#_Toc174978971)

[Número de unidades 31](#_Toc174978972)

[Parámetros para medir el tráfico 31](#_Toc174978973)

[Tipo de tráfico 32](#_Toc174978974)

[Modelos de tráfico 33](#_Toc174978975)

[Grado de servicio 34](#_Toc174978976)

[Teorema de Burke 35](#_Toc174978977)

[Redes de colas 36](#_Toc174978978)

[Redes en serie 38](#_Toc174978979)

[Redes de Jackson abiertas 38](#_Toc174978980)

[Redes de Jackson cerradas 41](#_Toc174978981)

[Número de servidores 43](#_Toc174978982)

[Colas de un solo servidor 43](#_Toc174978983)

[Colas multiservidor 44](#_Toc174978984)

[Tratamiento en caso de congestión 45](#_Toc174978985)

[Conclusión 46](#_Toc174978986)

[Cuestionario 46](#_Toc174978987)

[Bibliografía 49](#_Toc174978988)

# Glosario

**Proceso estocástico:** Un proceso estocástico es una colección de variables aleatorias indexadas por el tiempo u otra dimensión. Es un modelo matemático que describe la evolución de un sistema que cambia con el tiempo de manera no determinista, es decir, donde la incertidumbre o el azar juegan un papel crucial.

En un proceso estocástico, el estado futuro del sistema depende tanto del estado actual como de un componente aleatorio, lo que significa que, aunque se conozcan las condiciones iniciales y las reglas del proceso, no se puede predecir con certeza el estado exacto del sistema en un momento dado en el futuro; solo se puede describir probabilísticamente.

**Proceso no determinista:** significa que su comportamiento o evolución no está completamente determinado por sus condiciones iniciales. En otras palabras, incluso si se conociese el estado actual del sistema y las reglas que lo gobiernan, no se podría predecir su comportamiento futuro con certeza. El resultado de un proceso no determinista depende en parte del azar o de elementos aleatorios, lo que introduce incertidumbre en las predicciones.

**Distribución de Poisson:** la distribución de Poisson es una distribución de probabilidad discreta que modeliza la frecuencia de eventos determinados durante un intervalo de tiempo fijado a partir de la frecuencia media de aparición de dichos eventos.

**Algoritmo Round Robin:** es un método de planificación de tareas o procesos en el que cada proceso recibe un tiempo fijo, conocido como “quantum”, para ejecutarse. Los procesos se atienden en orden cíclico. Una vez que un proceso consume su “quantum” y no ha finalizado, se coloca al final de la cola y se da paso al siguiente proceso. Este ciclo se repite hasta que todos los procesos han sido completados. Es ampliamente utilizado en sistemas de tiempo compartido para asegurar que todos los procesos reciban un tiempo de ejecución justo.

**Erlang:** Un Erlang (Er) se define como el tiempo en que un recurso está ocupado durante la hora cargada. Su nombre, Erlang, es debido al ingeniero danés A. K. Erlang, creador de la ingeniería de tráfico y la teoría de colas.

**Matriz de transiciones:** Es la matriz formada por el conjunto de probabilidades de pasar de un estado a otro.

**Vector de estado:** Vector formado por el conjunto de las probabilidades de estar en cada uno de los estados de un sistema.

**Régimen permanente:** En un sistema en régimen permanente, las probabilidades de estado se mantienen constantes

# Introducción

La teoría de colas es el estudio matemático de las colas o líneas de espera dentro de un sistema. Esta teoría estudia factores como el tiempo de espera medio en las colas o la capacidad de trabajo del sistema sin que llegue a colapsar. Dentro de las matemáticas, la teoría de colas se engloba en la investigación de operaciones y es un complemento muy importante a la teoría de sistemas y la teoría de control. Se trata así de una teoría que encuentra aplicación en una amplia variedad de situaciones como negocios, comercio, industria, ingenierías, transporte y logística o telecomunicaciones.

La teoría de colas permite modelar sistemas en los que existe una población de agentes (también llamados clientes o usuarios) que demandan ciertos recursos. Dado que puede haber más agentes que recursos, pueden registrarse esperas desde que un agente llega al sistema hasta que el servidor atiende su demanda, o incluso un agente puede ser rechazado del sistema al no haber espacio donde esperar.

En el contexto de la informática o de las tecnologías de la información y la comunicación son paradigmáticas las situaciones de espera dentro de una red de comunicaciones. Así, por ejemplo, los procesos enviados a un servidor en red forman colas de espera mientras no son atendidos; la información solicitada a un servidor Web a través de Internet puede recibirse con demora debido a la congestión en la red o en el servidor; en el caso de telefonía (fija o móvil), una llamada puede no ser realizada si alguna central está saturada.

La teoría de colas aparece a principios del siglo veinte para estudiar los problemas de congestión de tráfico que se presentaban en las comunicaciones telefónicas. Posteriormente esta teoría se ha utilizado para resolver problemas de la vida real, particularmente el tráfico de automóviles, la regulación de semáforos en una ciudad, la determinación de cajeros en bancos, etc. Sin embargo, su aplicación es mucho más amplia, específicamente en la industria de la manufactura y servicios donde hay eventos tanto programados como aleatorios esta teoría es una herramienta importante para la definición de procesos. A continuación se enlistan algunos ejemplos:

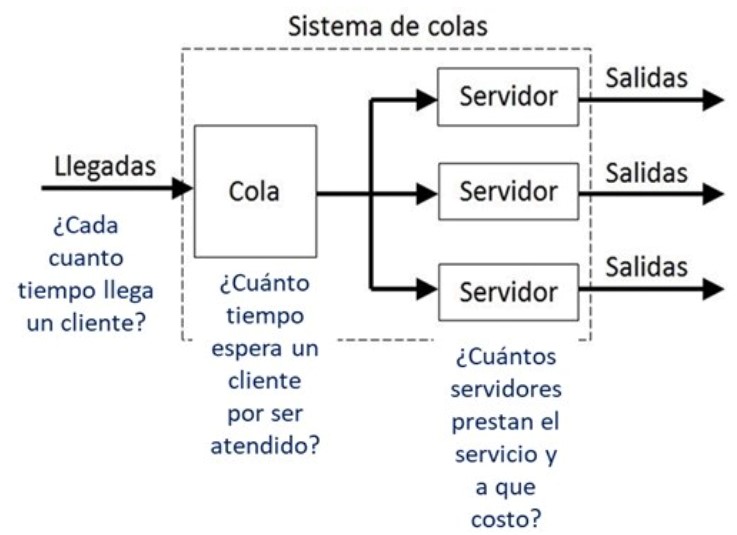
* **Áreas de servicio:** En el área de mantenimiento (en empresas de manufactura, hoteles, etc.) se puede determinar en base a la frecuencia de fallas de los equipos y el tiempo de reparación la cantidad óptima de técnicos en la plantilla para reducir el tiempo muerto y no tener equipos en espera de servicio, esto sin afectar el costo. De forma similar se puede aplicar para la plantilla de las áreas de calidad, sistemas y servicios generales entre otras.
* **Flujo de materiales en líneas de ensamble:** ¿Cuál debe ser la relación de unidades de manejo de material (servidores) y las cargas a mover (clientes)?
* **Almacenes:** Se puede contestar a preguntas típicas de la estructura de almacenes como ¿Cuántos almacenistas y montacarguistas tener en el área de recibo? ¿Con cuántos andenes debe contar el almacén? ¿Cuánto personal para atención en ventanilla es necesario? e inclusive ¿Cuántas ventanillas de despacho deben existir en el almacén?
* **Sistema de transporte:** Este es uno de los usos históricos de la teoría de colas para tratar casos como filas de automóviles en casetas de cobro, barcos esperando servicio de carga, etc.
* **Hospitales:** Así como recientemente se ha empezado ha implementar herramientas de Lean Manufacturing en hospitales también teoría de colas tiene aplicación directa en el análisis de colas que diariamente se generan en sus procesos: determinar número de médicos de guardia, enfermeras, camillas, ambulancias, personal administrativo, equipos de laboratorio, etc.
* **Sistema legislativo:** Un sistema legislativo es una red de colas en que los asuntos a tratar son los clientes.

Para un mejor entendimiento se plantea el siguiente caso: existe una población de clientes que requieren un servicio. Estos clientes entran al sistema y se unen a la cola de espera. En determinado momento se selecciona un miembro de la cola para proporcionarle el servicio mediante alguna regla (conocida como disciplina de cola). Terminado el servicio, el cliente sale del sistema de colas.

En general, la cola de espera se forma (y crece) si durante un tiempo la demanda del servicio es mayor que la capacidad del sistema para suministrarlo. Por ejemplo, si un usuario necesita un minuto para ser atendido, pero llegan dos usuarios en menos de un minuto, el segundo usuario tendrá que esperar en cola hasta ser atendido. Estos desequilibrios temporales entre demanda y capacidad suceden dado que tanto el proceso de llegadas de usuarios como el proceso de proporcionar servicio son procesos estocásticos.

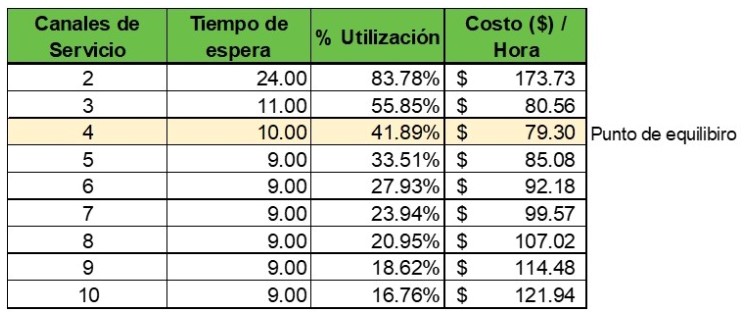
Con problemáticas similares a la anterior en mente, se utiliza la teoría de colas con el objetivo de analizar las prestaciones de un sistema compuesto por uno o más recursos que sirven a una población de clientes.

En cuanto a resultados, se busca obtener el tiempo medio de espera en cola, el tiempo total de estancia en el sistema, el número de recursos necesarios para mantener la probabilidad de rechazo por debajo de un límite, el nivel óptimo de capacidad de un sistema que minimiza un determinado coste, etc.; con el fin de poder analizar cómo impactan estas variables en las prestaciones o rendimientos de un sistema. En otras palabras, el objetivo principal es analizar y optimizar la eficiencia de sistemas en los que hay demanda y asignación de recursos.



Ilustración

El análisis bajo teoría de colas proporciona el número óptimo de servidores para lograr el equilibrio entre costo y tiempo de espera del cliente, y da, mediante la simulación, medidas de rendimiento para la revisión y optimización del sistema como lo son: Promedio de clientes en el sistema, tiempo promedio de espera de los clientes, número de clientes en la cola, % de utilización del servidor, entre otras.



Ilustración

## Elementos en la teoría de colas

**Fuente de entrada o población potencial:** Una característica de la fuente de entrada es su tamaño. El tamaño es el número total de clientes que pueden requerir servicio en determinado momento. Puede suponerse que el tamaño es infinito o finito, habitualmente, se utiliza la suposición de población infinita. Eso implica que la tasa de llegadas se supone constante. Si la población es considerada finita, el tráfico que llegará se irá reduciendo a medida que vayan llegando llamadas al sistema, ya que el número de posibles llamadas se irá reduciendo.

**Cliente:** Es todo individuo de la población potencial que solicita servicio como por ejemplo una lista de trabajo esperando para imprimirse.

**Cola:** Espacio donde los clientes esperan (es decir, están haciendo cola) hasta que empiezan a ser servidos. Se caracteriza por una capacidad o tamaño máximo de la cola, que se puede suponer finita o infinita. Si es finita, es posible que se llene de clientes por lo que nuevos clientes son rechazados.

**Disciplina de la cola:** mecanismo empleado para seleccionar el siguiente miembro que recibirá servicio, una vez que queda un recurso disponible. Por ejemplo, puede ser:

* FIFO (First In, First Out) primero en entrar, primero en salir. En esta disciplina se atiende por orden de llegada. En una cola simple FIFO, los «paquetes» que llegan deben ponerse al final y esperar su turno para ser procesados, la forma de vaciado es tomar el paquete que hace más tiempo que se encuentra en la cola. Todos los paquetes tienen el mismo tratamiento y por ende, no se puede garantizar servicios adicionales como calidad de servicio en ciertos paquetes.
* LIFO (Last In, First Out) también conocida como pila que consiste en atender primero al cliente que ha llegado el último. Es parecido a una pila, donde se apilan elementos y cuando hay que vaciar se toma el último elemento apilado.
* RSS (Random Selection of Service) que selecciona los clientes de manera aleatoria.
* PS (Processor Sharing) sirve a los clientes igualmente. La capacidad de la red se comparte entre los clientes y todos experimentan con eficacia el mismo retraso. Bajo esta disciplina, todos los clientes en la cola comparten equitativamente la capacidad de servicio del servidor. Esto significa que si hay n clientes en el sistema, cada uno recibe 1/n de la tasa de servicio total. Esta disciplina es teórica y se utiliza para modelar situaciones donde los recursos se dividen equitativamente entre múltiples procesos, como en sistemas de tiempo compartido en informática.
* CP (Colas Priority) se usa cuando existe una prioridad dentro de la cola. Dentro de prioridad, pueden ser con o sin interrupción, según si llega un cliente con prioridad se interrumpe el trabajo en curso o no. En este tipo de colas los paquetes que llegan se clasifican por clases de prioridad. Esta clase de prioridad debe estar marcada de alguna forma. Cada clase de prioridad tendría su cola virtual independiente. Cuando toque procesar un paquete se elegirá el paquete de la clase más alta de una cola no vacía. Cada cola trabajará a partir del criterio FIFO para elegir el paquete a procesar. También se pueden hacer disciplinas más complejas, puesto que dentro de cada prioridad se pueden aplicar a su vez disciplinas random, LIFO, o FIFO.
* WFQ (Weighted Fair Queing) es la disciplina donde los paquetes se clasifican en clases y se colocan a su cola correspondiente. Se va dando servicio a cada clase de forma circular e igualitaria usando un algoritmo del tipo Round Robin. En sistema WFQ asigna más capacidad a las colas más llenas, pero, sin dejar de atender a las colas más libres. Ajusta la disciplina de atención a cada cola según la ponderación del servicio de cada clase usando pesos (W) en función de la cantidad de elementos. También puede hacerse en función de la cantidad de servicios solicitados por cada flujo. Este sistema permite procesar distintas necesidades sin penalizar ninguna.

**Recursos o servidores:** elementos encargados de proporcionar el servicio a los clientes. Se caracterizan por un proceso de servicio que típicamente define el tiempo necesario para atender un cliente (por ejemplo, si es un tiempo aleatorio, la variable aleatoria que lo caracteriza). Si hay uno se le conoce como monocanal y si hay más, es multicanal.

**Redes de colas:** Sistema donde existen varias colas y los trabajos fluyen de una a otra. Por ejemplo: las redes de comunicaciones o los sistemas operativos multitarea.

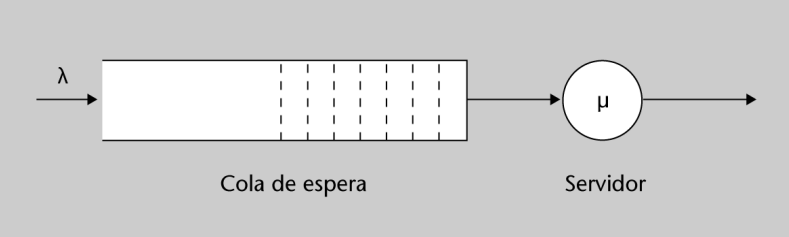
**Capacidad del sistema:** Es el número máximo de clientes que puede haber en el sistema (finito o infinito). Cuando llega una unidad y se encuentra el sistema lleno, ésta se pierde.

**Tratamiento en caso de ocupación:** Se puede tener retención de llamadas, con lo cual hará falta otro intento al recibir señal de congestión. Otra posibilidad es la liberación de llamadas, donde se esperará un tiempo para volver a enviar lo que constituirá otra llamada. Un tercer caso es un sistema de espera, que supone que la llamada entrará en una cola de espera.

**Accesibilidad:** Es el número de salidas que puede tener una cierta entrada. Puede ser completa; por lo tanto, cualquier entrada tiene acceso a cualquier salida, o bien limitada, y, por lo tanto, no todas las salidas libres se pueden conectar a las entradas.

# Desarrollo

Los sistemas de transmisión se pueden modelar a menudo según un esquema como el de la figura siguiente:



Ilustración

Aquí se muestra una fuente generadora de datos y una cola de espera donde se acumularán las unidades de datos esperando a ser atendidas por un servidor. Tanto el proceso de llegada de las unidades al sistema como el proceso de servicio son dos procesos estocásticos. Hay dos parámetros que hay que tener en cuenta: la disciplina con que se generan los datos y la disciplina con que se sirven. El término disciplina hace referencia a cuáles son los parámetros estadísticos de estos elementos.

Se puede considerar que las unidades llegarán independientemente del estado del sistema, por lo que su comportamiento se podrá analizar conjuntamente como un proceso estocástico discreto en tiempo continuo. El tiempo de servicio de las unidades será impredecible y se lo modela también como un proceso estocástico discreto en tiempo continuo, independiente del proceso de llegadas.

Un parámetro que definirá el comportamiento de las llegadas y del mecanismo de servicio será el correspondiente a la media de estas estadísticas. Viene caracterizado por λ y μ, respectivamente.

Velocidades medias:

* λ es la tasa media o velocidad media de llegadas al sistema.
* μ es la velocidad media de servicio.

Se pueden encontrar diferentes posibilidades en función de los valores que tomen λ y μ:

* λ < μ, el sistema es estable y la cola no se llenará.
* λ > μ, el sistema se saturará y se llenará la cola de espera.
* λ = μ, es el límite de estabilidad; se servirán tantos datos como lleguen.

A fin de que la espera no sea muy larga, se intenta que las colas no queden muy ocupadas; eso quiere decir que λ ≤ μ . Aunque para optimizar el coste, habitualmente se dimensiona el sistema de manera que se considera la posibilidad de una cierta congestión siempre que se garantice un mínimo nivel de calidad del servicio.

## Procesos de Poisson

Una de las disciplinas más utilizadas por su simplicidad, propiedades y características generales son los procesos de Poisson. Estas características sólo permiten unos análisis más bien simples, que se ajustan a fuentes de datos en general, pero no válidos para los casos en los que se tenga fuentes de datos más o menos complejos, como los multimedia (audio, vídeo, etc.).

### Conceptos previos para el desarrollo de los Procesos de Poisson:

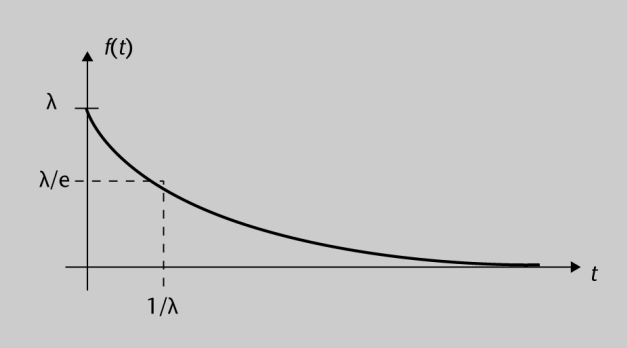
Distribución exponencial. Sistema sin memoria:

Un proceso con estadística exponencial es un proceso aleatorio que tiene una probabilidad distribuida que sigue una función de distribución de tipo exponencial.

Una función de distribución exponencial es una función continua que tiene una función de densidad de probabilidad de tipo exponencial. Eso quiere decir que:



Representación gráfica de la función de distribución exponencial:

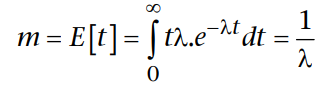


Ilustración

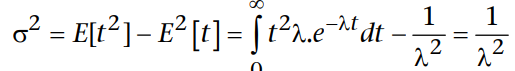
Su función de distribución, obtenida de integrar la función de densidad de probabilidad, es:



A partir de la función de densidad de probabilidad, se obtiene la media, o esperanza, de esta variable aleatoria:



Y la variancia:

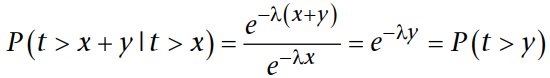


Un sistema con distribución de probabilidad exponencial se lo considera un sistema sin memoria.

Se considera un sistema en un tiempo de observación continuo t, y se observa el estado del sistema en dos instantes de tiempos reales positivos x e y. La distribución de probabilidad de t es sin memoria si la probabilidad en cualquier instante no depende de la probabilidad en instantes anteriores. Matemáticamente se lo expresa:

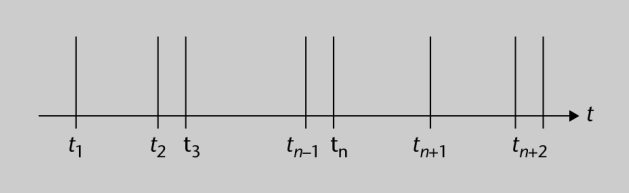


Se comprueba que un sistema definido con la función de distribución exponencial es un sistema sin memoria, ya que cumple esta igualdad:



### Definición de un proceso de Poisson

Con un proceso de llegadas aleatorias de forma continua en el tiempo, tal como se muestra en la figura siguiente:



Ilustración

Para definir un proceso de Poisson, se suponen las siguientes hipótesis:

1. Proceso sin memoria, donde cada llegada sea independiente de cuándo se ha producido la anterior.
2. Población infinita, es decir, que el número de fuentes sea tan grande que se pueda considerar que la tasa media de llegada de unidades no depende del tiempo y, por lo tanto, es una constante de valor λ.
3. Estacionario. La probabilidad de que se produzca una llegada es proporcional al tiempo de observación Δt, es decir, es λΔt.

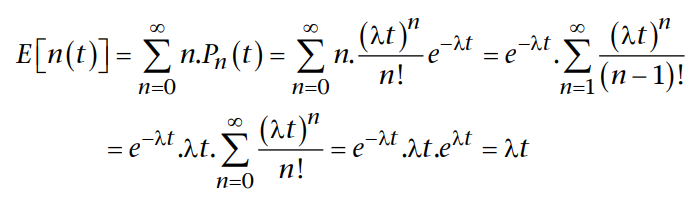
La probabilidad de tener alguna llegada en el instante de tiempo t, que se puede obtener como:



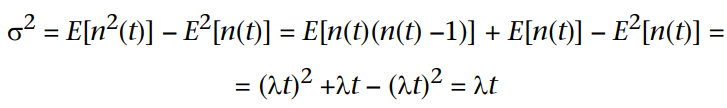
#### Media y variancia

En un proceso de Poisson, la media y variancia toman el mismo valor: λ×t.

Una vez definida la probabilidad de llegada de unidades al sistema, se obtiene el número medio de unidades en un intervalo de tiempo t, que se puede evaluar según la expresión:

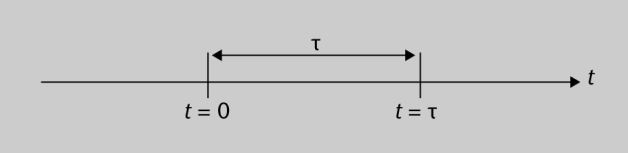


La variancia de las llegadas de un proceso de Poisson se puede calcular como:



### Distribución de las llegadas en un proceso de Poisson

La forma en que se distribuirán en el tiempo las llegadas es la función de distribución del proceso. Con el fin de poder calcularla, se supone un intervalo de tiempo con un origen arbitrario, en cuyo final se produce la llegada de la siguiente unidad.



Ilustración

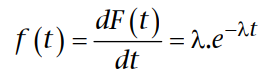
En este caso, hay una llegada para t = 0 y la siguiente se produce para t = τ. No se recibe ninguna unidad en el intervalo de tiempo comprendido en (0, τ), por lo tanto, la probabilidad de no tener ninguna llegada en el intervalo (0, t) es exactamente la probabilidad de que τ sea mayor que t. Es decir:



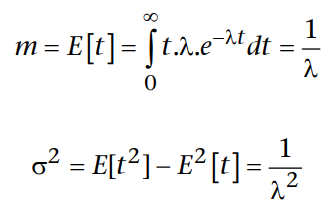
De aquí que la función de distribución sea:



Y derivando, se llega a la función de densidad de probabilidad de tipo exponencial:



A partir de esta función de densidad se obtiene la media de tiempo entre llegadas y la variancia:



Por lo que, en un proceso de Poisson, la duración media entre dos llegadas consecutivas coincide con su desviación típica .

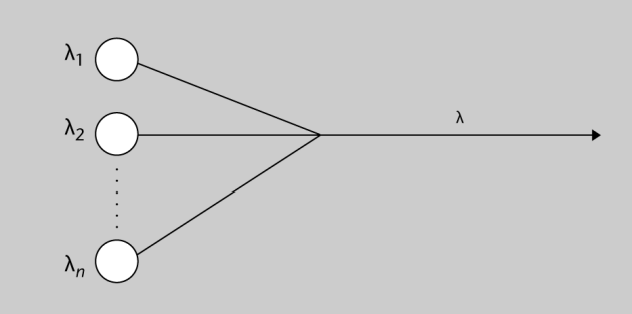
### Propiedades de los Procesos de Poisson:

Las dos propiedades son las siguientes:

**Superposición:**

Se suponen n fuentes de Poisson independientes con tasas de llegada λi. La fuente resultante de la suma de los procesos de Poisson es otro proceso de Poisson con una tasa de llegadas (λ) igual a la suma de las tasas de los procesos.

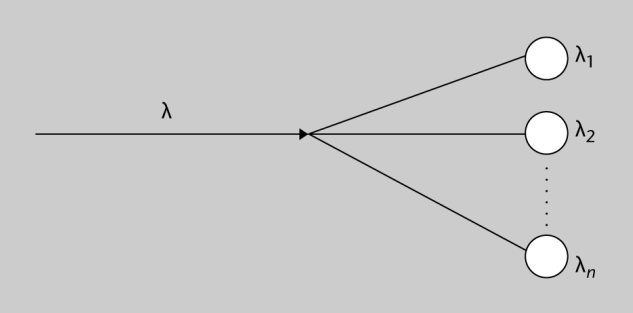




Ilustración

**Descomposición:**

Se supone una fuente generadora que sigue un proceso de Poisson con una tasa de llegadas λ. Si se descompone aleatoriamente este flujo en un conjunto de flujos más pequeños con una probabilidad Pi , los flujos resultantes tendrán una tasa λi = λ×Pi , que serán también de Poisson.



Ilustración

## Notación de Kendall

David Kendall introdujo una notación que permite describir las colas y mostrar las características de las mismas. Esta notación (mejor conocida como Kendall-Lee) sirve para caracterizar un sistema de líneas de espera en el cual todas las llegadas esperan en una sola cola hasta que está libre uno de los servidores paralelos idénticos.

Por lo general, las tasas de llegada (λ) y de servicio (µ) son una incertidumbre, es decir que son de naturaleza estocástica o probabilística. Por ello es que los tiempos de llegada y de servicio deben describirse a través de distribuciones de probabilidad.

La notación de un sistema de líneas de espera es:

A / B / C / D / E / F

En donde:

1. Se sustituye por la letra que denote la distribución de llegada:
   * **GI:** se refiere a una distribución de llegadas que es general e independiente. Esto significa que los tiempos entre llegadas pueden seguir cualquier distribución de probabilidad, sin restricciones específicas, y que las llegadas son independientes entre sí.
   * **D:** se usa para expresar valores determinísticos con tiempo promedio constantes.
   * **M:** indica que existe una distribución de llegada tipo Poisson y que es independiente de la llegada anterior. Es decir, que la distribución es exponencial entre los tiempos de llegadas.
   * **Ek:** indica que los tiempos entre llegadas o servicios siguen una distribución Erlang de k etapas. Esta distribución es una generalización de la distribución exponencial y es útil para modelar tiempos que son más regulares que la distribución exponencial, con menos variabilidad.
2. Se sustituye por la letra que denote la distribución de servicio:
   * **G:** indica una distribución general para los tiempos de servicio, sin asumir una forma específica.
   * **D:** representa un tipo de distribución determinístico.
   * **M:** denota que los tiempos de servicio siguen una distribución exponencial, implicando que el proceso de servicio es de tipo Poisson.
   * **Ek:** representa una distribución Erlang con k etapas, donde el tiempo de servicio es la suma de k tiempos exponenciales idénticos e independientes.
3. Se sustituye por el enteri positivo que denote el número de canales de servicio.
4. Se sustituye por las siglas que iniquen el orden de atención a los clientes:
   * **FCFS:** primeras entradas, primeros servicios (FIFO – First In, First Out).
   * **LCFS:** últimas entradas, primeros servicios.
   * **SIRO:** orden aleatorio.
   * **PR:** con base en prioridades.
   * **GD:** en forma general.
5. Se sustituye por el número máximo de clientes que soporta el sistema en un mismo instante de tiempo.
6. Se sustituye por el número de clientes potenciales del sistema de líneas de espera.

En muchos modelos importantes D / E / F son GD/∞/∞. Entonces estas características generalmente son omitidas. Es común usar la notación de Kendall simplificada, A / B / C, cuando la línea de espera es infinita, la fuente es infinita o la disciplina es en forma general.

## Ecuaciones de Little

La Ley de Little es un principio fundamental en la teoría de colas y se aplica en una variedad de contextos, desde la gestión de procesos empresariales hasta la ingeniería de sistemas. Esta ley establece una relación sencilla pero poderosa entre el número promedio de elementos en un sistema (L), la tasa promedio de llegada de elementos al sistema (λ) y el tiempo promedio que un elemento pasa en el sistema (W). La relación se expresa de la siguiente manera:

Las principales características de operación que interesan en las líneas de espera son:

* El número promedio de unidades en la línea de espera (Lq) o en el sistema (Ls): representa el número promedio de clientes en una cola, el número promedio de trabajos en proceso en una fábrica, o cualquier otra métrica similar.
* El tiempo promedio que cada unidad pasa en la línea de espera (Wq) o en el sistema (Ws): es el tiempo promedio que un elemento pasa en el sistema desde que llega hasta que se va. En el caso de una cola en un supermercado, sería el tiempo promedio que un cliente pasa desde que entra a la cola hasta que termina de pagar.
* Tasa promedio de llegada de elementos al sistema (λ): se refiere a la cantidad de elementos que llegan al sistema por unidad de tiempo. Por ejemplo, en una cola en un supermercado, λ sería el número de clientes que llegan por minuto.

John D.C. Little muestra que estas tres características están relacionadas enforma general y se aplican a diversos modelos de líneas de espera, independientemente.

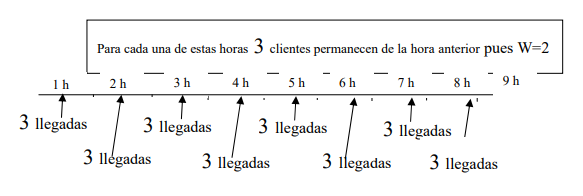
Las ecuaciones son las siguientes:

La primera ecuación general es:

Donde el número promedio de unidades en el sistema (Ls) es igual a la tasa de promedio de llegadas () por el tiempo promedio que una unidad o entidad pasa en el sistema (Ws).

Igualmente, el número promedio de unidades en la cola es igual a la tasa promedio de llegadas por el tiempo promedio que una unidad pasa en la cola.

Para un mejor entendimiento, se supone el siguiente ejemplo: W = 2 horas, λ = 3 clientes/hora, entonces el número medio de clientes en el sistema es 3 x 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):



Ilustración

Otra ecuación general es:

El tiempo promedio en el sistema es igual al tiempo promedio en espera mas el tiempo promedio de servicio.

De donde:

El tiempo promedio de servicio es 1/µ porque µ representa la tasa de servicio, es decir, el número promedio de clientes que un servidor puede atender por unidad de tiempo.

Nótese que para comprender la anterior ecuación, dado que ρ es el porcentaje de tiempo que 1 servidor esté ocupado, el valor c x ρ = λ / µ 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:

La importancia de las ecuaciones de Little es que se aplican a cualquier modelo de espera independientemente de si las llegadas siguen una distribución Poisson o no y de si los tiempos de servicio siguen una distribución exponencial o no. Es extremadamente útil porque proporciona una manera sencilla de comprender y optimizar los sistemas de colas y procesos. Al ajustar las tasas de llegada (λ) o los tiempos de procesamiento (W), las empresas pueden mejorar la eficiencia y reducir los tiempos de espera.

### Aplicaciones de la Ley de Little

* **Gestión de Inventarios**: En una fábrica, la Ley de Little puede ayudar a determinar la cantidad promedio de productos en proceso y el tiempo que un producto pasa en fabricación.
* **Atención al Cliente**: En centros de llamadas o servicios al cliente, puede utilizarse para evaluar el número promedio de clientes en espera y el tiempo de espera promedio.
* **Logística y Transporte**: En la gestión de cadenas de suministro, puede ayudar a calcular el número promedio de productos en tránsito y el tiempo promedio que tardan en llegar a su destino.

Ejemplo Práctico

Se supone que una tienda en línea quiere analizar la eficiencia de su sistema de entrega. Actualmente, recibe un promedio de 50 pedidos por hora (λ = 50 pedidos/hora) y el tiempo promedio que tarda en procesarse y entregarse un pedido es de 2 horas (W = 2 horas). Usando la Ley de Little:

L = λ × W

L = 50 pedidos/hora × 2 horas

L = 100 pedidos

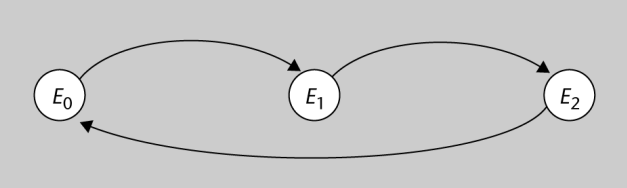
Esto significa que, en promedio, tiene 100 pedidos en el sistema en cualquier momento dado.

## Cadenas de Markov

Las cadenas de Markov son una herramienta que analiza el comportamiento de determinados tipos de procesos estocásticos; por ejemplo, el número de llamadas que llegan a una central telefónica o el número de compradores que llegan a un mostrador.

Un sistema puede cambiar su estado desde el estado actual a otro. El sistema estará en uno u otro estado en función de unas probabilidades. A partir de estas probabilidades, se puede calcular un conjunto de parámetros que permitirán caracterizar el sistema.

Una cadena de Markov representa un sistema que varía su estado a lo largo del tiempo. Cada cambio de estado se llama transición. Una cadena de Markov está formada por un conjunto de estados que se pueden representar gráficamente mediante nodos, enlazados entre ellos mediante arcos o flechas de transición entre unos estados y otros, de forma parecida al diagrama siguiente.



Ilustración

En general, la evolución de un sistema puede depender de todos los estados pasados.

Los procesos de Markov tienen como característica que su comportamiento futuro no depende del pasado, sólo depende del estado presente; son procesos sin memoria.

### Probabilidades de transición

Las probabilidades de transición indican la probabilidad de pasar de un estado al siguiente, y se las indica como pij.



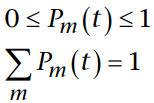
Los valores pij se llaman probabilidades de transición del estado i al estado j. Si estas probabilidades son estacionarias, es decir, no dependen del instante t considerado, es una cadena de Markov homogénea.

### Cadenas de Markov de tiempo continuo

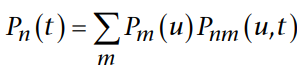
Los sistemas con cambios de estado en instantes de tiempos indefinidos se llaman sistemas de tiempo continuo. Como en los sistemas de comunicación, las llegadas y salidas se producen en instantes de tiempos indeterminados e independientes del tiempo, se trata las cadenas homogéneas de tiempo continuo.

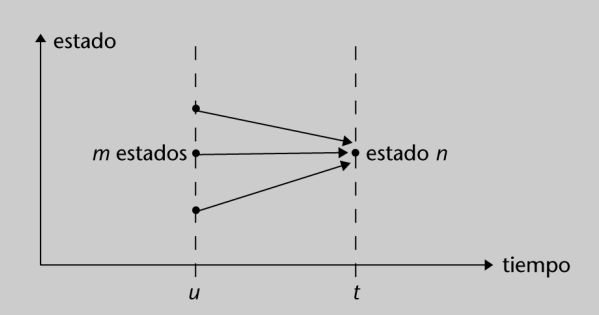
La probabilidad de estar en cada uno de los n estados se puede escribir pues como P0(t), P1(t), P2(t), ..., Pn(t). Este conjunto de probabilidades de los estados de un sistema se llama vector de estado.

Estas probabilidades cumplirán las propiedades:



La probabilidad de estar en el estado n en un instante t se puede descomponer según el conjunto de todos los caminos procedentes de cada uno de los estados anteriores, hasta n. Se deduce que:





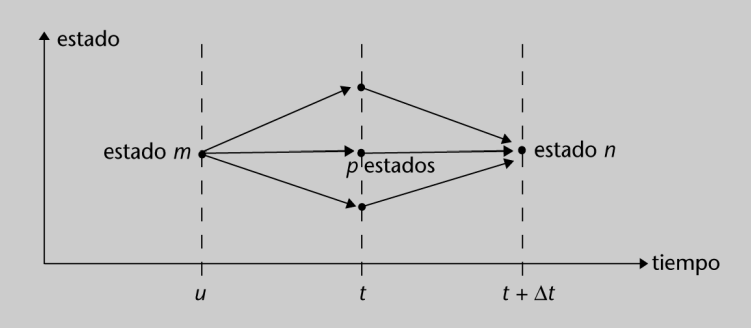
Ilustración

Al conjunto de probabilidades de transición entre estados Pnm(u,t) se lo llama matriz de transiciones.

## Ecuación de futuro

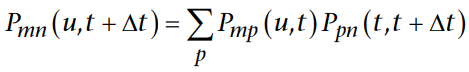
Es una relación entre las probabilidades de transición de los estados de un proceso.

La figura siguiente muestra la evolución del sistema desde un estado m en el instante u, hasta un estado n en el instante t + Δt, pasando por p estados en el instante t.



Ilustración

Se define la ecuación de futuro (ecuación de Chapman-Kolmogorov) como:



## Procesos de nacimiento y muerte

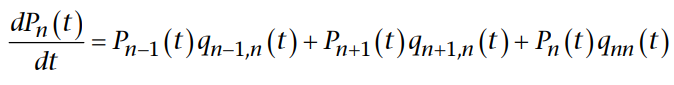
Estos procesos son un caso particular de procesos de Markov, donde las transiciones sólo se pueden realizar entre estados adyacentes.

Los procesos de nacimiento y muerte son el modelo más adecuado para modelar cambios en el tamaño de la población, pero también para analizar las prestaciones de las redes de comunicaciones; por ejemplo, para caracterizar las llamadas que se están cursando en una central de conmutación, o, los paquetes que hay en un router.

Un caso particular de los procesos de Markov podría ser haciendo la restricción de que en el siguiente instante de tiempo únicamente se puede pasar a un estado inmediatamente vecino, es decir, desde el estado En se puede pasar en los estados En+1, En-1 o mantenerse en el estado En. Esto permite definir los procesos de nacimiento (cuando se pasa a un estado superior) y muerte (cuando se pasa a un estado inferior), que constituyen un caso particular de los procesos de Markov.

En este caso, todas las probabilidades de transición serán nulas excepto: Pm,m+1, Pm,m y Pm,m-1.

Así pues, la ecuación de futuro en un proceso de nacimiento y muerte queda reducida a la siguiente expresión:



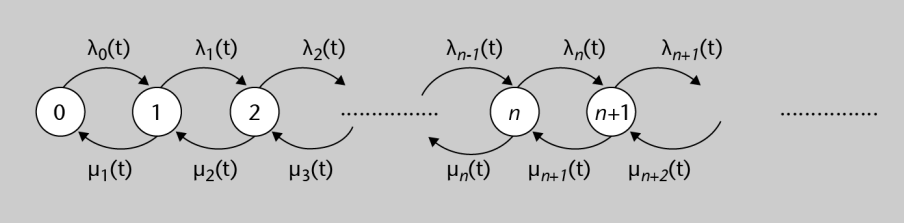
Se define la tasa de nacimiento como:



Y la tasa de muerte como:



Gráficamente, una cadena de Markov de tiempo continuo para procesos de nacimiento y muerte se puede representar:

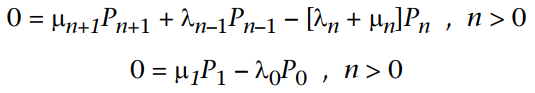


Ilustración

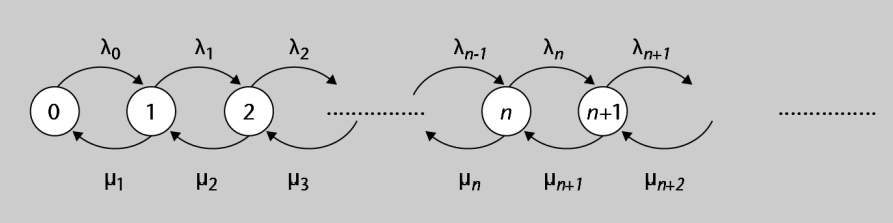
### Procesos de nacimiento y muerte en régimen permanente

Habitualmente, no interesa el régimen transitorio de un sistema, sino su régimen permanente. Se supone que el sistema ha evolucionado suficientemente para que las probabilidades de estado ya no dependan del tiempo y, por lo tanto, no dependan del estado inicial. Las variaciones de las probabilidades (por lo tanto, las derivadas respecto del tiempo) pasan a ser nulas.

En el caso de los procesos de nacimiento y muerte en régimen permanente, el sistema de ecuaciones queda reducido a:



Gráficamente, el proceso de nacimiento y muerte en régimen permanente se puede representar:

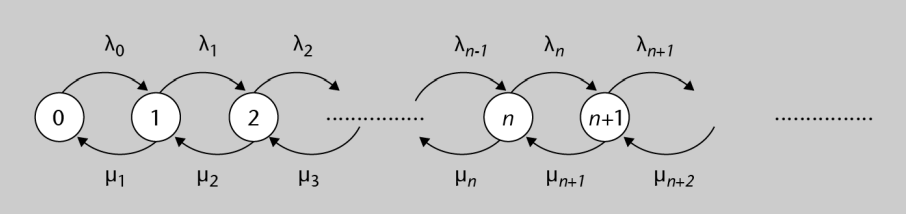


Ilustración

## Conceptos de tráfico

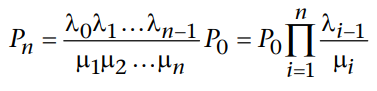
A partir de un sistema de nacimiento y muerte en régimen permanente, como el que se trató en el apartado anterior, se definen diferentes conceptos referentes al tráfico.

Suponiendo unas tasas o velocidades de nacimiento: λ0, λ1, λ2, ..., λn y unas tasas de muerte: μ0, μ1, μ2, ..., μn tal como muestra la figura siguiente:



Ilustración

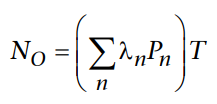
La probabilidad de que el sistema esté en un estado n venga dada por la expresión:



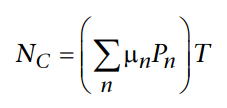
### Número de unidades

Mediante este sistema se puede calcular el número medio esperado de unidades ofrecidas y servidas, es decir, el número de unidades que han llegado al sistema y el número de unidades que han finalizado en un cierto intervalo de tiempo T.

El número de unidades ofrecidas se lo calcula a partir del producto de los nacimientos en cada estado por la probabilidad de estar en este estado y por el intervalo T:



El número de unidades servidas (cursadas) se calcula a partir de las muertes en cada estado multiplicadas por la probabilidad de estar en este estado y por el intervalo T:



Hay que tener en cuenta que el número de unidades ofrecidas no coincidirá con el número de unidades cursadas, ya que se perderán las unidades que encuentren los servidores y la cola (si hay) llenos, es decir, las unidades perdidas.

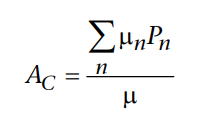
### Parámetros para medir el tráfico

La hora cargada (HC) es el periodo de una hora del día en que el tráfico es más elevado. De la definición de Erlang se puede deducir que la intensidad de tráfico de un recurso, como máximo, puede llegar a valer la unidad (1 Er = 60 minutos de tráfico en un circuito en la HC). El tráfico de un Erlang correspondería a un único recurso utilizado continuamente, o bien, a dos canales utilizados en un 50%.

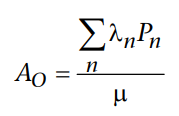
### Tipo de tráfico

La intensidad de tráfico, o simplemente tráfico, es una unidad de medida de la ocupación de un determinado recurso por unidad de tiempo. Se puede obtener del cociente entre el tiempo de ocupación respecto al tiempo de observación. Es una magnitud sin dimensiones que habitualmente se expresa en Erlangs.

Para calcular el tráfico cursado y, por lo tanto, la cantidad de unidades servidas con éxito, es necesario tener en cuenta las unidades servidas por unidad de tiempo y la duración media de las unidades que hay que servir (1/μ):



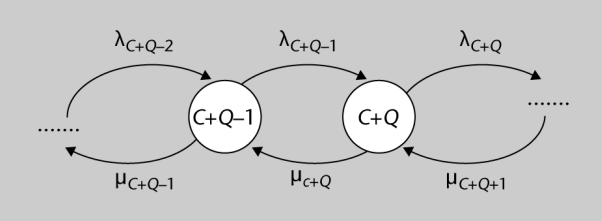
El tráfico ofrecido y, por lo tanto, lo que llega al sistema se pueden calcular de forma similar a partir de las unidades ofrecidas:



El tráfico ofrecido sería el que se habría cursado si el sistema hubiera podido absorberlo todo, lo que no suele ocurrir nunca. Hay unas llamadas que se pierden en el primer intento por la limitación del sistema. Estas llamadas forman el tráfico perdido:

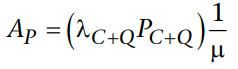


Si se supone que el sistema puede servir C unidades simultáneamente y tiene una cola de espera de Q unidades, se empezarán a perder unidades cuando se encuentren llenos los C servidores y la cola. El tráfico perdido corresponderá, pues, al estado C + Q del proceso.



Ilustración

Se podrá calcular como:



### Modelos de tráfico

En esta tabla se compara algunos de los modelos de tráfico. Los más utilizados son los de Erlang B, Erlang B extendido y Erlang C:



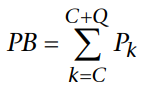
Ilustración

## Grado de servicio

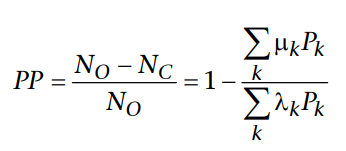
El grado de servicio (GoS) es un parámetro que se utiliza para medir la calidad del servicio, y se calcula como el cociente entre el número de unidades perdidas y el número de unidades ofrecidas.

En este punto se definen los principales conceptos para evaluar el grado de servicio o calidad del servicio. Este concepto va directamente asociado a la probabilidad de bloqueo.

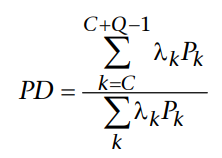
La probabilidad de bloqueo de un sistema es la probabilidad de que una unidad no se pueda servir porque todos los servidores se encuentran ocupados, debido a que la capacidad (K) del sistema es limitada.



La probabilidad de pérdida de un sistema equivale a la parte de las unidades ofrecidas al sistema que se pierden por encontrar el sistema lleno.



La probabilidad de espera o demora de un sistema es la parte de unidades ofrecidas al sistema que, sin perderse, se tienen que esperar a ser servidas porque el sistema está ocupado.



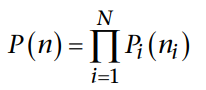
La probabilidad de servicio inmediato es la probabilidad de que la unidad se sirva inmediatamente y, por lo tanto, de que no se pierda ni se tenga que esperar.



## Teorema de Burke

La salida de una cola del tipo M/M/1, M/M/c o M/M/∞, con una tasa de llegadas λ es un proceso de Poisson con tasa λ. En cualquier instante de tiempo t, el número de unidades que hay en el sistema es independente de las salidas que ha habido antes de este instante. Se podría decir que el sistema es reversible, es decir, que las propiedades del sistema no cambian si se invierte el flujo del tiempo.

Según el teorema de Burke, para un sistema de colas M/M/c/∞, si la capacidad de las colas es infinita, se puede estudiar cada una de ellas por separado; por lo tanto, la serie estará formada por N colas independientes. La probabilidad de que en un instante haya n1 unidades en la cola 1, n2 unidades en la cola 2 ... y nN unidades en la cola N es:



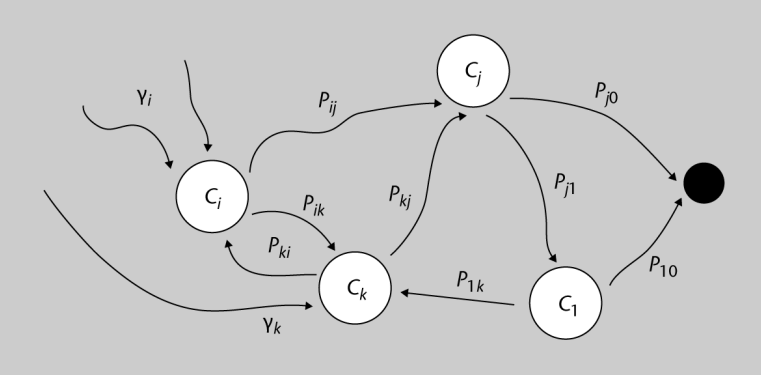
## Redes de colas

Una red de colas es un conjunto de nodos interconectados por medio de caminos. Cada uno de estos nodos está formado por un sistema de colas con uno o más servidores. Estas colas están conectadas con líneas que operan de forma asíncrona y concurrente, es decir, no hay sincronismo entre entradas y salidas, y actúan simultáneamente.

Las colas pueden estar conectadas entre ellas en serie o en tándem, donde el tráfico saliente de una cola es el tráfico entrante de la siguiente; pero también se pueden encontrar bifurcaciones y fusiones de tráfico donde se divide o se unen diversos flujos de tráfico. Ejemplos de redes de colas son las redes de ordenadores, las líneas de producción en una fábrica, el tráfico de vehículos en una ciudad, etc.

Se pueden encontrar dos tipos de redes de colas:

* **Cerradas:** Los flujos ni entran ni salen del sistema, por lo tanto, continúan circulando por el interior del sistema indefinidamente. El número de unidades se mantiene constante, ya que no se puede identificar un inicio y un final.
* **Abiertas:** Cada flujo entra en el sistema por un punto en un momento dado y, después de pasar por una o más colas, sale del sistema. No se puede considerar el número de unidades constante. Pueden ser:
* Acíclicas: un flujo nunca puede volver a la misma cola.
* Cíclicas: hay bucles en la red.



Ilustración

Parámetros de una red de colas:

* Las colas se representan mediante N nodos conectados por caminos.
* El nodo i da servicio con distribución exponencial, con una tasa de servicio µi.
* Las unidades que llegan al sistema desde el exterior a un nodo i llegan con una distribución de Poisson con una tasa Ƴi.

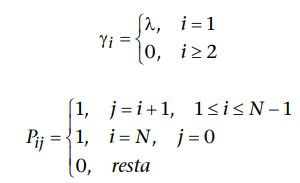
Una vez se ha servido una unidad en un nodo, pasa al nodo siguiente con una probabilidad fija o sale del sistema.

La probabilidad de ir del nodo i, al nodo j (j ≠ i) será:

La probabilidad de salir del sistema es:

### Redes en serie

La estructura de red más simple es la de las redes en serie, que son las que cumplen:



Las unidades entran por el nodo 1 y salen por el nodo N después de pasar por cada uno de los nodos.

### Redes de Jackson abiertas

El teorema de Jackson enuncia que, en una red de colas con las condiciones siguientes, cada nodo es un sistema independiente con entrada de Poisson; cada nodo se puede analizar por separado del resto utilizando un modelo M/M/1 o M/M/c, y los resultados se pueden combinar estadísticamente.

El teorema se basa en los siguientes supuestos:

* Los N nodos tienen servicio exponencial con tasa de servicio μi.
* Las unidades que llegan del exterior a un nodo y tienen distribución de Poisson con una tasa γi.
* La probabilidad de ir del nodo i al nodo j (j ≠ i) es Pij > 0, y la de abandonar el sistema es Pi0.

Por lo tanto, como el flujo total de entrada a un nodo j tiene que ser igual al que sale, se obtienen las ecuaciones de tráfico:



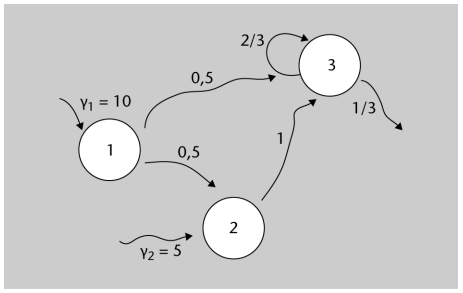
Para que el sistema no se sature, hará falta que se cumpla la condición siguiente para cada nodo i (ci , número de servidores en el nodo i):



La resolución de este sistema de N ecuaciones lineales permitirá obtener las tasas de llegada en cada uno de los N nodos, λi . Entonces, el teorema de Jackson dice que, en estado estacionario, la distribución del número de unidades en cada nodo es:



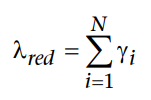
Donde Pi (ni) es la probabilidad de que haya ni clientes en el nodo i, calculada según los modelos M/M/c.



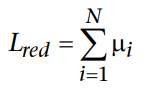
Ilustración

Además, se obtienen los siguientes parámetros de análisis de la red:

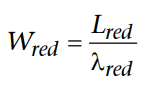
a) La tasa de entrada en la red es la suma de las tasas de llegada desde el exterior a cada uno de los nodos:



b) El número medio de unidades en la red será la suma de los números medios de unidades en cada uno de los N nodos:



c) El tiempo medio de permanencia en el sistema es el tiempo medio que pasa desde que una unidad entra en la red hasta que sale:



### Redes de Jackson cerradas

El teorema de Jackson para las redes cerradas se basa en los siguientes supuestos:

* La red de colas consta de N nodos, todos y cada uno de ellos con servicio exponencial independiente con tasa de servicio μi.
* El número de unidades que hay en el sistema es constante, de valor M.
* No hay entradas ni salidas, por lo tanto, γi = 0 y Pi0 = 0.
* Una vez se ha servido una unidad en un nodo, pasa al nodo siguiente con una probabilidad fija. La probabilidad de ir del nodo i al nodo j (i ≠ j) será Pij.
* Cada nodo se comporta como una cola con un modelo M/M/c.

Como el flujo total de entrada a un nodo tiene que ser igual al flujo total de salida del nodo, las ecuaciones de tráfico en este caso son:

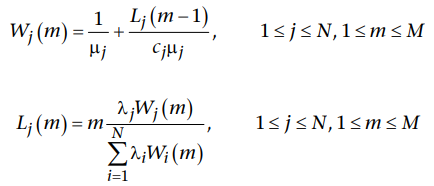


Estas N ecuaciones forman un sistema lineal indeterminado con un grado de libertad, que se resuelven para encontrar las tasas de llegada relativas a cada nodo, λi. Para resolverlo se supone, por ejemplo, que λ1 = 1.

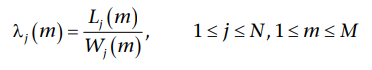
Mediante el análisis del valor medio (MVA), se resuelve el sistema de ecuaciones y calcular los siguientes parámetros del sistema con M unidades:

* Li (M): número medio de unidades en el nodo i.
* Wi (M): tiempo medio de permanencia en el nodo i.
* λi (M): tasa real de llegadas/salidas en el nodo i.

Es un algoritmo iterativo que va calculando Li (m), Wi (m) para los diferentes valores crecientes de m, a partir de m = 0. Se obtiene:

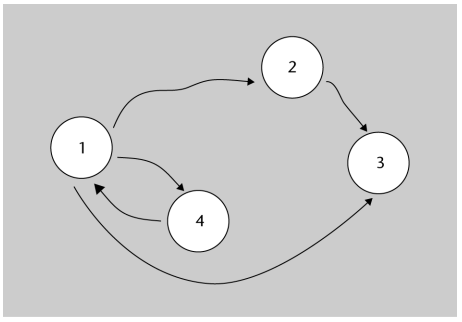


Aplicando la fórmula de Little:



Teniendo en cuenta que inicialmente:



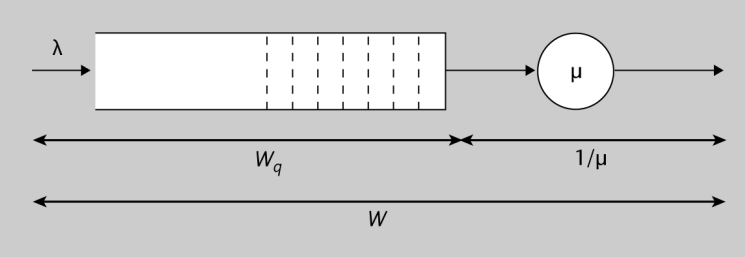


Ilustración

## Número de servidores

### Colas de un solo servidor

El modelo de colas de un solo servidor es el más sencillo. Se trata de un servidor que ofrece servicio a las unidades que le llegan. Las unidades de una cierta población llegan al sistema para ser servidas; si el servidor está vacío, pasan a ser servidas automáticamente, y, si no, pasan a una cola de espera.



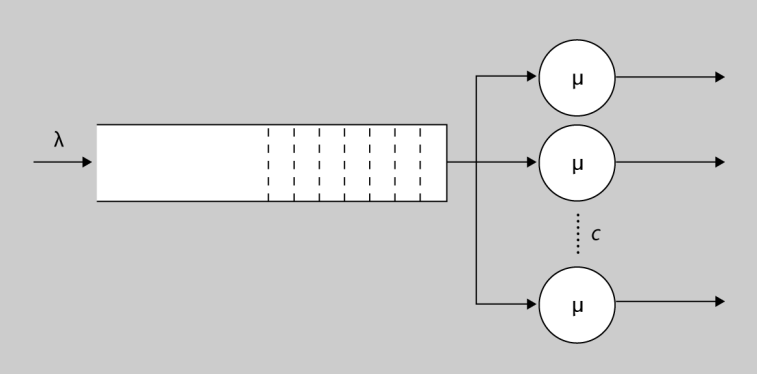
Ilustración

Las unidades llegan al sistema con una tasa de llegada λ (unidades medias/segundo). En cualquier momento habrá un cierto número de unidades esperando en la cola para ser servidas. El tiempo medio de espera de una unidad en la cola se lo llama Wq. El servidor atenderá a las unidades con un tiempo medio de servicio de valor 1/μ, por lo tanto, el tiempo medio que estará una unidad en el sistema será W, que se obtendrá de sumar el tiempo de espera en la cola y el tiempo de servicio.

Si la cola es infinita, nunca se perderá ninguna unidad; sin embargo, si la cola sólo permite un número finito de unidades, cuando esté llena, las unidades que vayan llegando se perderán.

### Colas multiservidor

El modelo de un servidor se puede generalizar fácilmente en el caso de tener múltiples servidores compartiendo una cola común. Si se supone que los servidores son idénticos, cuando una unidad llega al sistema se sirve de cualquiera de los servidores que esté libre. Cuando todos los servidores están llenos, las unidades que llegan se empiezan a almacenar en la cola, a la espera de que un servidor esté libre y puedan ser servidas con una determinada disciplina de servicio.



Ilustración

Las características habituales de una cola multiservidor son las mismas que para una única cola: población infinita, cola infinita y servicio de tipo FIFO.

## Tratamiento en caso de congestión

Según cómo actúe el sistema ante la llegada de unidades cuando el sistema está congestionado, se puede hablar de los siguientes sistemas.

a) **Sistemas con pérdidas:**

LCC (lost calls cleared): cuando el sistema está congestionado (todos los servidores se encuentran ocupados) señaliza su entrada indicando que está ocupado y lo obliga a reintentar la comunicación al cabo de un rato. Cualquier usuario puede reintentar la comunicación. Es el sistema que utilizan las centrales de conmutación en Europa.

LCH (lost calls held): cuando una llamada es bloqueada, señaliza al usuario que reintenta la llamada sin espera. Es el sistema que utilizan las centrales de conmutación en Norteamérica.

b) **Sistema de espera:**

LCD (lost calls delayed): Cuando el sistema está congestionado, no avisa a la fuente de que el sistema está ocupado de forma inmediata, sino que espera en un sistema de colas un tiempo, hasta que algún servidor esté disponible. En este caso, un parámetro importante es el tiempo de espera en función de la carga de tráfico.

c) **Sistema con reintento:**

LCR (lost calls retried): Cuando una entrada queda bloqueada, el sistema va haciendo reintentos hasta que la llamada se sirve. Es un modelo derivado del modelo LCC.

En general, todos los routers utilizan un sistema de pérdidas. De todas maneras, en routers modernos se utiliza el concepto de sistema de espera en determinadas partes.

# Conclusión

En conclusión, la teoría de colas es una herramienta matemática esencial para modelar y optimizar sistemas en los que existe una demanda de recursos que puede superar la capacidad disponible, generando tiempos de espera. Su aplicación se extiende a diversas áreas como las telecomunicaciones, el transporte, la manufactura, y los servicios de salud, entre otros. Al analizar factores como el tiempo medio de espera, el número de recursos necesarios, y las tasas de llegada y servicio, se pueden tomar decisiones informadas para mejorar la eficiencia del sistema. La teoría de colas también permite entender y gestionar redes de colas complejas, como las que se encuentran en sistemas de comunicación y redes de computadoras, proporcionando modelos para prever y controlar la congestión y asegurar un rendimiento óptimo. En resumen, esta teoría es vital para diseñar sistemas que equilibran el costo con la calidad del servicio, minimizando tiempos de espera y maximizando la eficiencia.

# Cuestionario

**¿Qué es la teoría de colas y cuál es su objetivo principal?**

La teoría de colas es el estudio matemático de las colas o líneas de espera dentro de un sistema. Su objetivo principal es analizar y optimizar la eficiencia de sistemas en los que hay demanda y asignación de recursos, calculando variables como el tiempo medio de espera en cola, el tiempo total de estancia en el sistema, y el número de recursos necesarios para mantener la probabilidad de rechazo por debajo de un límite.

**¿Qué factores se consideran en la teoría de colas para modelar un sistema?**

Los factores considerados incluyen:

1. **Fuente de entrada o población potencial**: Tamaño de la población que puede requerir servicio.
2. **Cliente**: Individuo que solicita servicio.
3. **Cola**: Espacio donde los clientes esperan su turno para ser atendidos.
4. **Disciplina de la cola**: Regla que determina el orden en que los clientes son atendidos.
5. **Recursos o servidores**: Elementos que proporcionan el servicio.
6. **Capacidad del sistema**: Número máximo de clientes que puede haber en el sistema.
7. **Tratamiento en caso de ocupación**: Cómo se manejan las llamadas o solicitudes cuando el sistema está ocupado.

**¿Qué significa que el sistema sea estable, y cómo se relacionan las velocidades de llegada (λ) y servicio (μ) en este contexto?**

Un sistema es estable cuando la tasa de llegada de clientes (λ) es menor que la tasa de servicio (μ), es decir, λ < μ. Esto significa que el sistema puede atender a los clientes a medida que llegan, evitando que la cola crezca indefinidamente.

**¿Qué es una red de colas y cuáles son los dos tipos principales?**

Una red de colas es un conjunto de nodos interconectados por caminos, donde cada nodo representa un sistema de colas con uno o más servidores. Los dos tipos principales de redes de colas son:

* **Redes cerradas**: Donde los flujos ni entran ni salen del sistema, manteniendo constante el número de unidades.
* **Redes abiertas**: Donde los flujos entran y salen del sistema, permitiendo que el número de unidades varíe.

**Describe brevemente los modelos de colas de un solo servidor y multiservidor.**

* **Colas de un solo servidor**: Un servidor único que atiende a las unidades que llegan al sistema. Si el servidor está ocupado, las unidades esperan en una cola.
* **Colas multiservidor**: Varias unidades servidoras que comparten una cola común. Cuando todos los servidores están ocupados, las unidades que llegan esperan en la cola hasta que un servidor esté libre.

**¿Qué es la disciplina FIFO en el contexto de colas?**

La disciplina FIFO (First In, First Out) es una regla donde el primer cliente en llegar es el primero en ser atendido. Todos los clientes son tratados por orden de llegada, sin preferencias.

**Menciona tres aplicaciones prácticas de la teoría de colas.**

1. **Gestión de tráfico vehicular**: Optimización de semáforos y control de flujos de tráfico.
2. **Servicios de atención al cliente**: Determinación del número de cajeros en un banco para minimizar el tiempo de espera.
3. **Redes de comunicación**: Análisis de colas en servidores web para mejorar la eficiencia del servicio.

**¿Qué implica un sistema LCC (Lost Calls Cleared) en el contexto del tratamiento en caso de congestión?**

En un sistema LCC, cuando todos los servidores están ocupados, se señaliza al usuario que el sistema está ocupado, obligando a reintentar la comunicación después de un rato. Este sistema es común en las centrales de conmutación en Europa.

**¿Qué caracteriza a un sistema de colas bajo el teorema de Jackson?**

El teorema de Jackson caracteriza redes de colas en las que cada nodo es un sistema independiente con entrada de Poisson y servicio exponencial. Permite analizar cada nodo por separado y combinar los resultados estadísticamente, aplicando modelos M/M/1 o M/M/c.

**¿Qué es el modelo de reintento LCR (Lost Calls Retried) en un sistema de colas?**

En el modelo LCR, cuando una entrada queda bloqueada debido a la congestión, el sistema realiza reintentos hasta que la llamada o solicitud sea servida. Es un modelo derivado del sistema LCC.

# Bibliografía

* <https://es.wikipedia.org/wiki/Teor%C3%ADa_de_colas#:~:text=La%20teor%C3%ADa%20de%20colas%20sirve,productiva%20en%20la%20ingenier%C3%ADa%20industrial>.
* <https://es.linkedin.com/pulse/aplicaciones-y-an%C3%A1lisis-de-modelos-teor%C3%ADa-colas-en-espera-gonz%C3%A1lez#:~:text=La%20notaci%C3%B3n%20de%20Kendall%20es,es%20el%20n%C3%BAmero%20de%20servidores>.
* <https://www.eumed.net/libros-gratis/2011b/969/notacion%20de%20kendall%20lee.html#:~:text=3.5.5%20Notaci%C3%B3n%20de%20Kendall%20Lee&text=Designa%20el%20hecho%20de%20que,con%20media%20y%20varianza%20conocida>.
* <https://es.scribd.com/document/248880104/Notacion-de-Kendall-y-Ecuaciones-de-Flujo#:~:text=La%20notaci%C3%B3n%20de%20Kendall%2Dlee,al%20servicio%2C%20y%20as%C3%AD%20sucesivamente>.
* <https://openaccess.uoc.edu/bitstream/10609/148227/2/Modulo2_AnalisisMdianteTeoriaDeColas.pdf>
* <file:///D:/NO%20BORRAR/Nueva%20carpeta%20(2)/ojsadministrador,+009.pdf>
* <https://pascua.iit.comillas.edu/aramos/simio/transpa/t_qt_ac.pdf>
* <https://economipedia.com/definiciones/teoria-de-colas.html>
* Gross, D.; Shortle, J. F.; Thompson, J. M.; Harris, C. M. (2008). Fundamentals of queueing theory (4.ª ed.). Nueva York: John Wiley and Sons.
* Análisis mediante teoría de colas - Enric López Rocafiguera, Pere Barberán - FUOC, 2019 Av. Tibidabo, 39-43, 08035 Barcelona.