# Sistemas de colas

Contenidos adaptados del libro «Probabilidad y estadística para informáticos, segunda edición, M. Baron» (Capítulo 7)

# Contenido

1. Objetivos
2. Introducción
   2.1 Definición de sistema de cola
   2.2 Componentes principales de un sistema de cola
   2.3 Parámetros de un sistema de cola
   2.4 Variables aleatorias en un sistema de cola
3. Ley de Little
4. Tipos de sistemas de colas
5. Sistema M/M/1
6. Resumen

# 1. Objetivos

- Modelizar correctamente situaciones de colas de espera bajo modelos Poissonianos (RA6).
- Conocer algunos modelos no Poissonianos, redes y series de colas, y la utilidad de la simulación en la Teoría de Colas (RA6).

# 2. Introducción

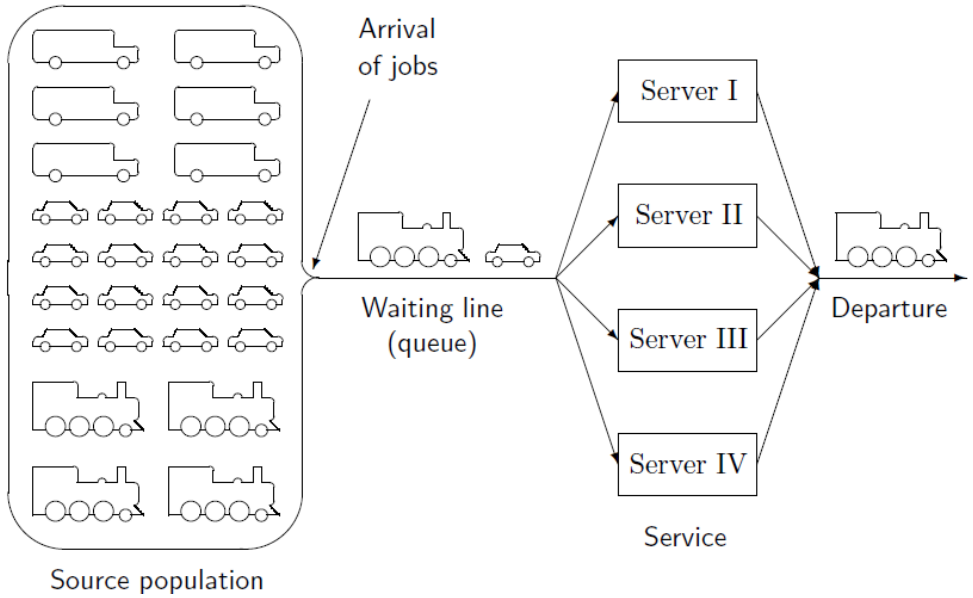
- 2.1 Definición del sistema de cola
- 2.2 Componentes principales de un sistema de cola
- 2.3 Parámetros de un sistema de cola
- 2.4 Variables aleatorias en un sistema de cola

# 2.1 Definición de *sistema de cola*

- Un sistema de cola es una instalación que consiste en uno o varios servidores diseñados para realizar ciertas tareas o procesar ciertos trabajos y una cola de trabajos que esperan ser procesados.
- Los trabajos llegan al sistema, esperan a que esté disponible un servidor, son procesados por este servidor y se van.
- Ejemplos de sistemas de cola son:
  - un ordenador personal o compartido que ejecute tareas enviadas por sus usuarios;
  - un proveedor de servicios de Internet cuyos clientes se conectan a Internet, navegan y se desconectan;
  - una impresora que imprime trabajos enviados desde diferentes ordenadores;
  - un servicio de atención al cliente con uno o varios teleoperadores que respondan a las llamadas de los clientes;
  - una zona de peaje en una carretera, un puesto para comprar comida rápida desde el coche, un cajero automático;
  - un consultorio médico en el que se atiende a pacientes; ...

# 2.2 Componentes principales de un sistema de cola

- Llegada de trabajos
  - Por lo general, los trabajos llegan a un sistema de cola en momentos aleatorios
- Cola (línea de espera)
  - Cuando llega un nuevo trabajo, si un servidor está disponible en ese momento, tomará el nuevo trabajo;
  - si todos los servidores están ocupados con otros trabajos, el nuevo trabajo se une a la cola, y espera hasta que se completen todos los trabajos previos.
- Servicio (servidores)
  - Una vez que un servidor está disponible, inmediatamente comienza a procesar el siguiente trabajo asignado
- Salida
  - Cuando se completa el servicio, el trabajo abandona el sistema




- SISTEMA DE COLA = Cola + Servicio

# 2.3 Parámetros de un sistema de cola

| Parámetro | Significado |
|---|---|
| $\lambda_A$ | Tasa de llegadas<br>• Número previsto de llegadas por unidad de tiempo |
| $\lambda_S$ | Tasa de servicio<br>• Número medio de trabajos procesados por un servidor en funcionamiento continuo durante una unidad de tiempo |
| $\mu_A = \frac{1}{\lambda_A}$ | Tiempo medio de llegadas |
| $\mu_S = \frac{1}{\lambda_S}$ | Tiempo medio de servicio |
| $r = \frac{\lambda_A}{\lambda_S} = \frac{\mu_S}{\mu_A}$ | Factor de utilización del sistema |

# 2.4 Variables aleatorias en un sistema de cola

| Variable | Significado |
|---|---|
| $X_s(t)$ | Número de trabajos que reciben servicio en el momento t |
| $X_w(t)$ | Número de trabajos que esperan en la cola en el momento t |
| $X(t) = X_s(t) + X_w(t)$ | Número total de trabajos en el sistema en el momento t |
| $S_k$ | Tiempo de servicio del trabajo k-ésimo |
| $W_k$ | Tiempo de espera del trabajo k-ésimo |
| $R_k = S_k + W_k$ | Tiempo de respuesta, el tiempo total que un trabajo pasa en el sistema desde su llegada hasta la salida<br>• Un sistema es **estacionario** si las distribuciones de $S_k$, $W_k$ y $R_k$ son independientes de k.<br>• En este caso, el índice k se puede omitir. |

# 3. Ley de Little

- La ley de Little establece una relación simple entre el número esperado de trabajos, el tiempo de respuesta esperado y la tasa de llegadas.
- Es válido para cualquier sistema de cola estacionario.
- Ley de Little: $\boldsymbol{\lambda_A E(R) = E(X)}$
  - $\lambda_A$ es la tasa de llegadas
  - E(R) es la esperanza del tiempo de respuesta (tiempo de un trabajo en el sistema)
  - E(X) es la esperanza del número de trabajos en el sistema

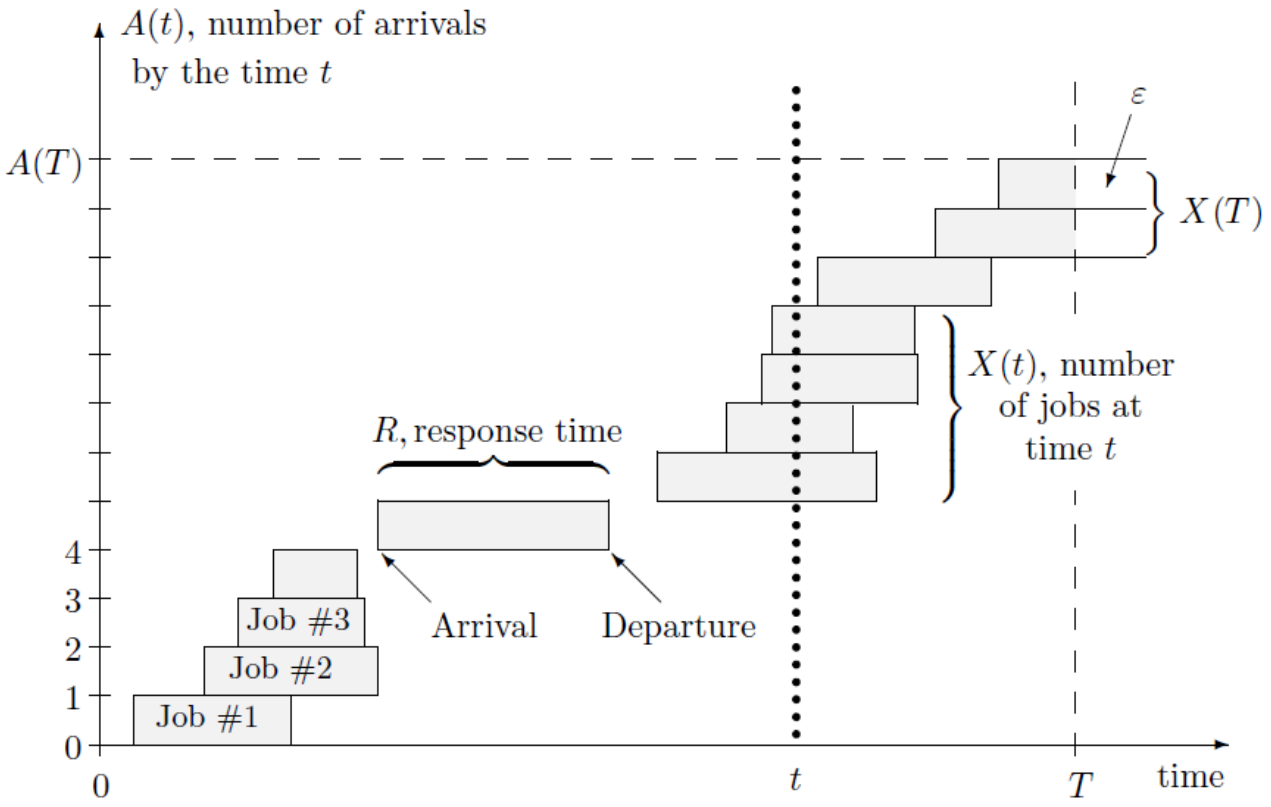



FIGURE 7.2: *Queuing system and the illustration to the Little's Law.*

# 4. Tipos de sistemas de cola

- Un sistema de cola se identifica como A/S/n/C (Notación de Kendall)
- **A** indica la distribución de probabilidad del tiempo entre llegadas
  - El valor habitual es “M” (Markov o Memoryless), que indica Tiempo entre llegadas exponencial (es decir, un proceso de conteo de Poisson).
  - Otros valores posibles son: “D”, “G”, ...
- **S** indica la distribución de probabilidad del tiempos de servicio
  - El valor habitual es “M”, que indica tiempos de servicio exponencial
- **n** es el número de servidores
- **C** es la capacidad de la cola (por defecto, si no se indica, se considera capacidad ilimitada: C = ∞)
  - Es el número máximo de trabajos permitidos en la cola
- Los sistemas de cola más usuales son M/M/1, M/M/k y M/M/∞

# 5. Sistema M/M/1

- Un sistema de cola M/M/1 es un proceso de Markov de tiempo continuo con las siguientes características:
  - A="M": Tiempo entre llegadas exponencial (proceso de llegadas de Poisson)
    - Con tasa de llegadas $\lambda_A$
  - S="M": Tiempo de servicio exponencial
    - Con tasa de servicio $\lambda_S$
  - Los tiempos de llegadas y de servicio son independientes
  - n=1: Un servidor
  - C=∞: Capacidad ilimitada

# 5. Sistema M/M/1
# Distribución estacionaria del número de trabajos

- Cualquier sistema cuya tasa de servicio exceda la tasa de llegadas (es decir, los trabajos pueden ser atendidos más rápido de lo que llegan, por lo que no hay sobrecarga) tiene una distribución estacionaria (estable).
  - Condición: $\lambda_S > \lambda_A \rightarrow r < 1$
  - Si $r \geq 1$ el sistema se sobrecarga
- Es posible calcular la probabilidad de X (el número de trabajos en el sistema) en cualquier momento en que el sistema ha alcanzado el equilibrio
  - $\pi_x = P\{X = x\} = r^x(1 - r)$
  - $E(X) = \frac{r}{1-r}$
  - $Var(X) = \frac{r}{(1-r)^2}$
  - Donde $r = \frac{\lambda_A}{\lambda_S} = \frac{\mu_S}{\mu_A}$

# 5. Sistema M/M/1
# Evaluación del rendimiento del sistema (I)

- Utilización
  - $P\{X > 0\} = 1 - P\{X = 0\} = 1 - (1 - r) = r$
  - $P\{Servidor\ ocupado\} = P\{X_S = 1\} = P\{X > 0\} = r$
    - Si sólo hay un trabajo en el sistema, el trabajo no está en cola, está en el servidor
- Tiempo de espera
  - $W = S_1 + S_2 + \cdots + S_x$
  - $E(W) = E(S_1) + E(S_2) + \cdots + E(S_x) = E(S)E(X) = \frac{1}{\lambda_S} \cdot \frac{r}{1-r} = \mu_S \frac{r}{1-r}$
    - $E(S) = \frac{1}{\lambda_S} = \mu_s$ porque S es exponencial; $E(X) = \frac{r}{1-r}$
- Tiempo de respuesta
  - $E(R) = E(W) + E(S) = \mu_S \frac{r}{1-r} + \mu_S = \frac{1}{\lambda_S(1-r)}$

# 5. Sistema M/M/1
# Evaluación del rendimiento del sistema (II)

- Número de trabajos en la cola
  - $X_w = X - X_S$
  - $E(X_w) = E(X) - E(X_S) = \frac{r}{1-r} - r = \frac{r^2}{1-r}$
    - $X_S$ es Bernoulli con p = P{Servidor ocupado}= r; y $E(X_s)=p=r$

- Ley de Little revisada para los sistemas de cola M/M/1
  - $\lambda_A E(R) = E(X)$
  - $\lambda_A E(R) = E(X_S)$
  - $\lambda_A E(R) = E(X_w)$

# 5. Sistema M/M/1
# Ejemplo 7.4 (Transmisión de mensajes)

- A un centro de comunicación llegan mensajes en momentos aleatorios con un promedio de 5 mensajes por minuto.
- Se transmiten a través de un solo canal en el orden en que fueron recibidos.
- En promedio, se tarda 10 segundos en transmitir un mensaje.
- Se cumplen las condiciones de una cola M/M/1.
- Calcular las principales características de rendimiento de este centro.

# 5. Sistema M/M/1
# Ejemplo 7.4 (Solución) (I)

- Tasa de llegadas: $\lambda_A = 5\ min^{-1}$
- Tiempo de servicio esperado: $\mu_S = 10\,\text{sec} = \frac{1}{6} min$
- Factor de utilización: $r = \frac{\lambda_A}{\lambda_S} = \lambda_A \cdot \mu_S = 5 \cdot \frac{1}{6} = \frac{5}{6} = 0.83$
  - Esto también representa la proporción de tiempo cuando el canal está ocupado y la probabilidad de que el tiempo de espera no sea cero.
- Número medio de mensajes almacenados en el sistema en cualquier momento:
  - $E(X) = \frac{r}{1-r} = \frac{\frac{5}{6}}{1-\frac{5}{6}} = 5\ mensajes$

# 5. Sistema M/M/1
# Ejemplo 7.4 (Solución) (II)

- Número medio de mensajes esperando en cola en cualquier momento:
  - $E(X_w) = \frac{r^2}{1-r} = \frac{\left(\frac{5}{6}\right)^2}{1-\frac{5}{6}} = 4.17\ mensajes$
- Cuando un mensaje llega al centro, el promedio de su tiempo de espera hasta que comienza su transmisión es:
  - $E(W) = \mu_S \frac{r}{1-r} = \frac{1}{6} \cdot \frac{\frac{5}{6}}{1-\frac{5}{6}} = \frac{5}{6} = 0.83\ \text{min} = 50\ segundos$
- Promedio del tiempo total desde la llegada de un mensaje hasta el final de su transmisión:
  - $E(R) = \frac{\mu_S}{1-r} = \frac{\frac{1}{6}}{1-\frac{5}{6}} = 1\ min$

# 5. Sistema M/M/1
# Ejemplo 7.5 (Predicciones)

- Continuación del ejemplo 7.4
- Supongamos que el año que viene está previsto aumentar en un 10% la base de clientes del centro de transmisión y, por lo tanto, la intensidad del tráfico entrante también aumentará en un 10 %.
- ¿Cómo afectará esto al rendimiento del centro?

# 5. Sistema M/M/1
# Ejemplo 7.5 (Solución)

- Nueva tasa de llegadas (10% mayor): $\lambda_A = 5 \ + 5 \cdot 0.1 = 5.5 \ min^{-1}$
- Factor de utilización: $r = \frac{5.5}{6} = 0.92$
  - Se acerca peligrosamente a 1 (sobrecrga)
- Otros cálculos:
  - $E(X) = \frac{r}{1-r_2} = 11 \ mensajes$ (antes 5 mensajes)
  - $E(X_w) = \frac{r^2}{1-r_r} = 10.08 \ mensajes$ (antes 4.17 mensajes)
  - $E(W) = \mu_S \frac{r}{1-r} = 110 \ segundos$ (antes 50 segundos)
  - $E(R) = \frac{\mu_S}{1-r} = 2 \ min$ (antes 1 minuto)

- Conclusiones: Se comprueba que el tiempo de respuesta, el tiempo de espera, el número promedio de mensajes almacenados y, por lo tanto, la cantidad promedio requerida de memoria, se duplicará, cuando el número de clientes aumenta en apenas un 10 %.

# 5. Sistema M/M/1
# Ejercicios propuestos

- Ejercicios 7.8-7.14 del libro
  - Las respuestas de 7.8, 7.9, 7.10, 7.11, 7.12 y 7.14 están disponibles en el libro.

# 6. Simulación de sistemas de colas

- En la práctica, muchos sistemas de cola tienen una estructura bastante compleja.
  - Los trabajos pueden llegar de acuerdo con un proceso que no sea de Poisson; a menudo, la tasa de llegadas cambia durante el día, porque puede haber horas punta.
  - Los tiempos de servicio pueden tener diferentes distribuciones, y no siempre son sin memoria, por lo que la propiedad de Markov puede no estar satisfecha.
  - El número de servidores también puede cambiar durante el día (los servidores adicionales pueden activarse durante las horas punta).
  - Algunos clientes pueden quedar insatisfechos con un largo tiempo de espera y abandonar la cola.
  - ...
- La teoría de colas no cubre todas las situaciones posibles.
- Podemos simular el comportamiento de casi cualquier sistema de colas y estudiar sus propiedades mediante simulación por ordenador de métodos como los conocidos métodos Monte Carlo (ver capítulo 5 del libro de referencia).

# 6. Simulación de sistemas de colas Procedimiento

- Los métodos de Monte Carlo permiten simular y evaluar sistemas complejos de colas
- Siempre y cuando se conozcan las distribuciones del tiempo entre llegadas y del tiempo de servicio, se pueden generar los procesos de llegadas y servicios.
- Para asignar trabajos a servidores, se hace un seguimiento de los servidores que están disponibles cada vez que llega un nuevo trabajo.
- Cuando todos los servidores están ocupados, el nuevo trabajo entrará en una cola.
- A medida que se simula el trabajo de un sistema de cola, se mantienen registros de eventos y variables que interesen.
- Después de un gran número de ejecuciones, se peden promediar los datos registrados para estimar probabilidades y esperanzas de las variables a largo plazo.

# 6. Simulación de sistemas de colas Ejemplo

- Se podría simular el funcionamiento en un día, de 8 am a 10 pm, de un sistema de cola que tiene:
  - cuatro servidores;
  - tiempos de servicio de cada servidor: S1:Gamma(6,0.3), S2:Gamma(10,0.2), S3:Gamma(7,0.7), S4:Gamma(5,1);
  - un proceso de Poisson de llegadas con tasa de 1 llegada cada 4 minutos, independiente del tiempo de servicio;
  - asignación aleatoria de servidores, cuando hay más de un servidor disponible.
  - Suponer que después de 15 minutos de espera, los trabajos se retiran de una cola si su servicio no ha comenzado.
- Para un día ordinario de trabajo de este sistema, podríamos estimar los valores esperados de:
  - el tiempo total que cada servidor está ocupado con trabajos;
  - el número total de trabajos atendidos por cada servidor;
  - el tiempo medio de espera;
  - el tiempo de espera más largo;
  - el número de trabajo retirados;
  - el número de veces que un servidor estaba disponible inmediatamente (es decir, el número de trabajos sin tiempo de espera);
  - el número que quedan en el sistema a las 10 pm.
- Se pueden ver los resultados en el libro, que incluye el código para Matlab y R de la simulación.

# 7. Resumen

- Un sistema de cola es una instalación que consiste en uno o varios servidores que procesan ciertos trabajos que llegan al azar y una cola de trabajos que esperan ser procesados
- Hay varios tipos de sistemas de colas.
  - La notación Kendall se utiliza para identificar diferentes tipos.
  - Los sistemas de cola más usuales son M/M/1, M/M/k y M/M/ ∞.
- Hay fórmulas útiles para evaluar el rendimiento de un sistema de cola, pronosticar su eficiencia futura cuando los parámetros cambian, y ver si sigue funcionando bajo las nuevas condiciones.
- El rendimiento de los sistemas de cola más complicados y avanzados puede evaluarse mediante simulación por ordenador; simular llegadas de trabajos, asignación de servidores y tiempos de servicio y realizar un seguimiento de todas las variables de interés.