Teoría de colas III: La cola $M/M/m$

Prof. José Niño Mora

Investigación Operativa, Grado en Estadística y Empresa, 2011/12

Universidad
Carlos III de Madrid
www.uc3m.es

Esquema

- La cola $M/M/m$
- Factor de utilización; estabilidad
- Ecuaciones de balance de flujo
- Cálculo de medidas de rendimiento
- Efectos del tráfico pesado en el rendimiento
- Efecto marginal de añadir un servidor
- Comparación de los sistemas $M/M/1$ y $M/M/m$

La cola $M/M/m$; notación de Kendall

- La primera "M": los tiempos entre llegadas son v.a. i.i.d exponenciales:

$$\tau_1, \tau_2, \ldots \sim \text{Exp}(\lambda).$$

- La segunda "M": los tiempos de servicio de los clientes son v.a. i.i.d. exponenciales:

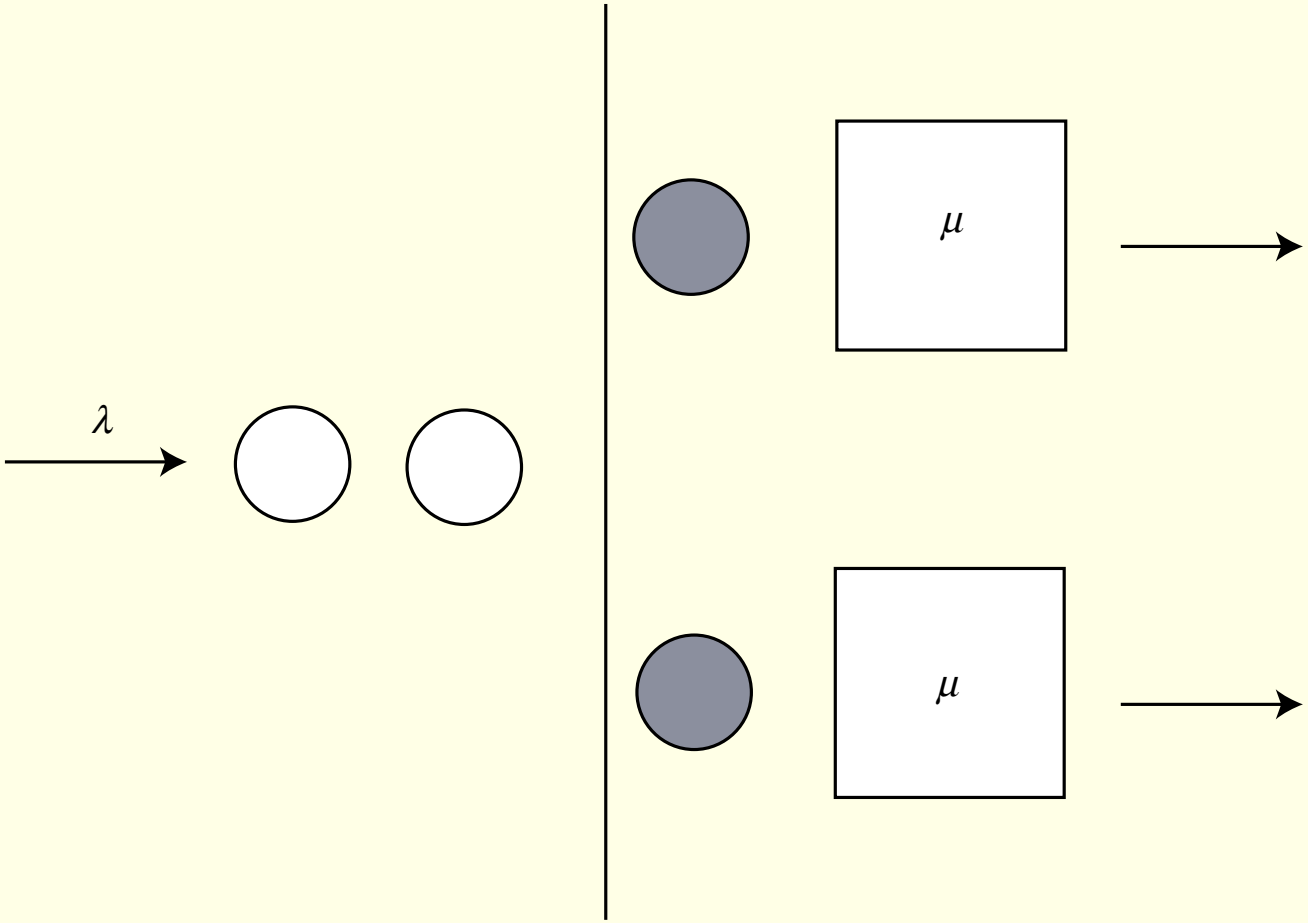
$$\xi_1, \xi_2, \ldots \sim \text{Exp}(\mu)$$

- La "m" final en $M/M/m$: hay m servidores idénticos
- El modelo está dado por tres parámetros:
  - Tasa de llegada: $\lambda > 0$
  - Tasa de servicio: $\mu > 0$

- Número de servidores: $\boxed{m}$

- Política de servicio: *FCFS* (First-Come First-Serve).

La cola $M/M/2$



La cola $M/M/m$

- Importante en aplicaciones

- Modelo de **procesamiento en paralelo**: múltiples servidores

- Ej: centro de atención telefónica (call center)

- Ej: ordenador con múltiples CPUs

Factor de utilización; estabilidad

- El factor de utilización del sistema $M/M/m$ es:

$$\rho = \frac{\text{tasa media de llegada de trabajo}}{\text{capacidad del sistema}} = \frac{\lambda/\mu}{m} = \frac{\lambda}{m\mu}$$

- El sistema es estable si

$$\rho < 1$$

Cómo calcular $\bar{L}$

- Recordemos que $\bar{L} = \mathsf{E}[L]$

- Probabilidad en equilibrio, o a largo plazo, de que haya n clientes en el sistema:

$$p_n = \mathsf{P}\{L = n\}, \quad n = 0, 1, 2, \ldots,$$

- Plan para calcular el # medio de clientes en el sistema, $\bar{L}$:

  1. Calcular p_n, para $n = 0, 1, 2, \ldots$

  2. Calcular $\bar{L}$ mediante la relación: $\bar{L} = \sum_{n=0}^{\infty} n p_n$

Ecuaciones de balance del flujo

- ¿Cómo calcular las probabilidades p_n?

- Consideramos el **diagrama de tasas de transición entre estados**. (Nota: el estado es el # de clientes en el sistema)

- $L(t)$ es un **proceso de nacimiento-muerte (N-M)**: sólo hay transiciones entre estados contiguos. Para $n = 0, 1, \ldots$,
  - Tasa de flujo de n a $n+1$: λp_n
  - Tasa de flujo de $n+1$ a n: $\min(n+1, m)\mu p_{n+1}$

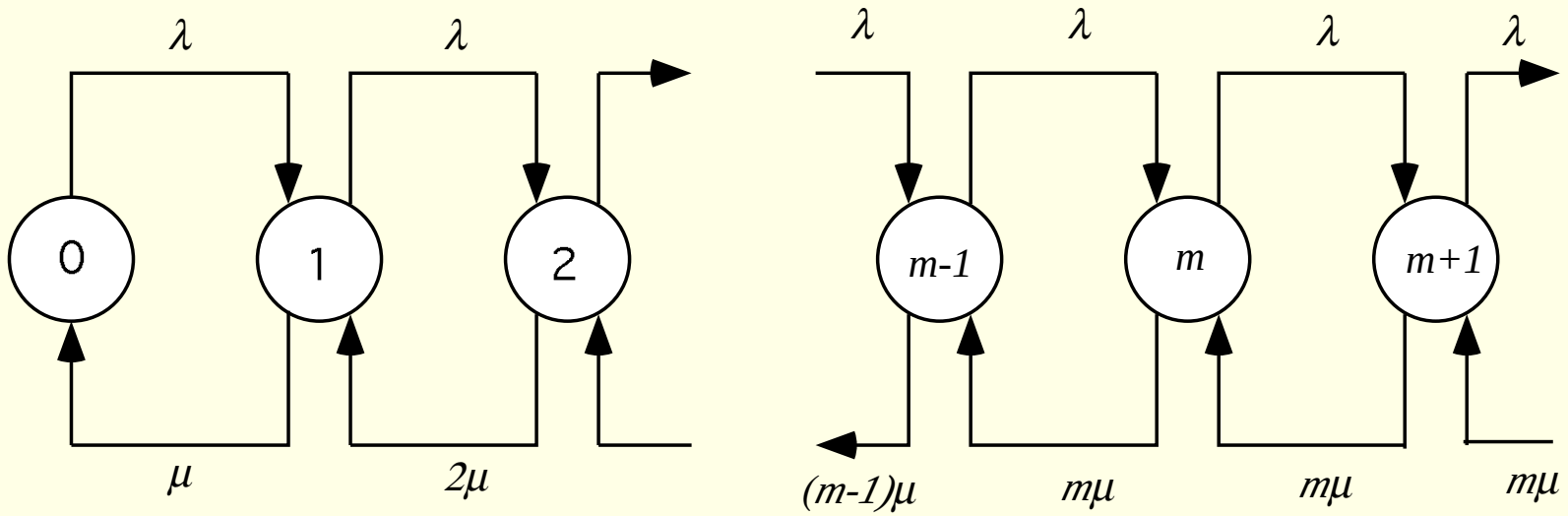
- **Ecuaciones de balance del flujo**: En equilibrio,

Tasa de flujo de n a $n+1$ = Tasa de flujo de $n+1$ a n

es decir,

$$\lambda p_n = \min(n+1, m)\mu p_{n+1}, \quad n = 0, 1, 2, \ldots$$

$M/M/m$: diagrama de tasas de transición



Ecuaciones de balance del flujo (EBF)

Ecuaciones de balance del flujo (EBF)

- Reformulamos las EBF como

$$p_n = \frac{n+1}{m\rho} p_{n+1}, \quad n = 0, \ldots, m-1$$

$$p_{n+1} = \rho p_n, \quad n \geq m$$

Ecuaciones de balance del flujo (EBF)

- Reformulamos las EBF como

$$p_n = \frac{n+1}{m\rho} p_{n+1}, \quad n = 0, \ldots, m-1$$

$$p_{n+1} = \rho p_n, \quad n \geq m$$

- De donde obtenemos las p_n en función de p_m:

$$p_{m-k} = \frac{(m-k+1)\cdots m}{(m\rho)^k} p_m$$

$$= \frac{m!}{(m-k)!(m\,\rho)^k} p_m, \quad k = 1, \ldots, m$$

$$p_{m+k} = \rho^k p_m, \quad k = 1, 2, \ldots$$

Cálculo de p_m

Cálculo de p_m

- Para calcular p_m, sustituimos en

$$\sum_{k=1}^{m} p_{m-k} + \sum_{k=0}^{\infty} p_{m+k} = \sum_{n=0}^{\infty} p_n = 1,$$

Cálculo de p_m

- Para calcular p_m, sustituimos en

$$\sum_{k=1}^{m} p_{m-k} + \sum_{k=0}^{\infty} p_{m+k} = \sum_{n=0}^{\infty} p_n = 1,$$

obteniendo:

$$p_m \left(\sum_{k=1}^{m} \frac{m!}{(m-k)!(m\rho)^k} + \frac{1}{1-\rho} \right) = 1$$

Cálculo de p_m

- Para calcular p_m, sustituimos en

$$\sum_{k=1}^{m} p_{m-k} + \sum_{k=0}^{\infty} p_{m+k} = \sum_{n=0}^{\infty} p_n = 1,$$

obteniendo:

$$p_m \left(\sum_{k=1}^{m} \frac{m!}{(m-k)!(m\rho)^k} + \frac{1}{1-\rho} \right) = 1$$

Por tanto:

$$p_m = \frac{1}{\displaystyle\sum_{k=1}^{m} \frac{m!}{(m-k)!(m\rho)^k} + \frac{1}{1-\rho}}$$

Cálculo del # medio en el sistema

Cálculo del # medio en el sistema

- Calculamos

$$\bar{L} = \sum_{n=0}^{\infty} n\, p_n = \sum_{k=1}^{m} (m-k) p_{m-k} + \sum_{k=0}^{\infty} (m+k) p_{m+k}$$

$$= \sum_{k=1}^{m-1} \frac{(m-k)m!}{(m-k)!(m\rho)^k}\, p_m + \sum_{k=0}^{\infty} (m+k)\rho^k p_m$$

$$= p_m \left[\sum_{k=1}^{m-1} \frac{m!}{(m-k-1)!(m\rho)^k} + \frac{m(1-\rho)+\rho}{(1-\rho)^2} \right]$$

Cálculo del # medio en servicio

• Aplicamos la ley de Little al "sistema" formado por los clientes en servicio: obtenemos

$$\bar{B} = \lambda \frac{1}{\mu} = m\rho$$

Cálculo del # medio en cola

- Como $\bar{L} = \bar{Q} + \bar{B} = \bar{Q} + m\rho$, tenemos que:

$$\bar{Q} = \bar{L} - m\rho$$

Cálculo del tiempo medio en el sistema

- Por la ley de Little,

$$\boxed{\bar{L} = \lambda \bar{S},}$$

por tanto, el **tiempo medio por cliente en el sistema** es:

$$\boxed{\bar{S} = \frac{\bar{L}}{\lambda}}$$

Cálculo del tiempo medio en cola (espera)

- Como $\bar{S} = \bar{W} + 1/\mu$, tenemos que:

$$\bar{W} = \bar{S} - \frac{1}{\mu}$$

Cálculo del tiempo medio en cola (espera)

- Como $\bar{S} = \bar{W} + 1/\mu$, tenemos que:

$$\bar{W} = \bar{S} - \frac{1}{\mu}$$

- Otro argumento, basado en la ley de Little:

$$\bar{Q} = \lambda \bar{W}$$

por tanto: $\bar{W} = \bar{Q}/\lambda$

Tráfico pesado: cuando $\rho \approx 1$

- Cuando el factor de utilización de un sistema de colas está próximo a la unidad ($\rho \approx 1$) decimos que el sistema está en un régimen de **tráfico pesado**

Tráfico pesado: cuando $\rho \approx 1$

- Cuando el factor de utilización de un sistema de colas está próximo a la unidad ($\rho \approx 1$) decimos que el sistema está en un régimen de **tráfico pesado**

- ¿Qué ocurre en ese caso con las medidas de rendimiento $\bar{L}, \bar{Q}, \bar{S}, \bar{W}$?

Tráfico pesado: cuando $\rho \approx 1$

- Cuando el factor de utilización de un sistema de colas está próximo a la unidad ($\rho \approx 1$) decimos que el sistema está en un régimen de **tráfico pesado**

- ¿Qué ocurre en ese caso con las medidas de rendimiento $\bar{L}, \bar{Q}, \bar{S}, \bar{W}$?

- La congestión media aumenta muy deprisa según ρ se aproxima a la unidad

Efecto marginal de añadir un servidor

- Consideremos una cola $M/M/1$ con parámetros λ, μ, y utilización $\rho = \lambda/\mu < 1$

Efecto marginal de añadir un servidor

- Consideremos una cola $M/M/1$ con parámetros λ, μ, y utilización $\rho = \lambda/\mu < 1$
- Su congestión media es: $\bar{L} = \rho/(1-\rho)$

Efecto marginal de añadir un servidor

- Consideremos una cola $M/M/1$ con parámetros λ, μ, y utilización $\rho = \lambda/\mu < 1$
- Su congestión media es: $\bar{L} = \rho/(1-\rho)$
- Si añadimos un segundo servidor con velocidad μ, obtenemos una cola $M/M/2$ con parámetros λ, μ, y utilización $\hat{\rho} = \lambda/(2\mu) = \rho/2$; sea $\hat{L}$ su congestión media
- ¿Cuál es la disminución relativa en la congestión media?

$$\frac{\bar{L}-\hat{L}}{\bar{L}} = \frac{1}{2} - \frac{1}{2}\frac{4-8\rho+\rho^2}{(2-\rho)(\rho+2)} : \begin{cases} > \frac{1}{2} & \text{si } \rho > 4-2\sqrt{3} \approx 0.54 \\ < \frac{1}{2} & \text{si } \rho < 4-2\sqrt{3} \end{cases}$$

Efecto marginal de añadir un servidor

- ¿Cuál es la disminución relativa en la congestión media?

$$\frac{\bar{L}-\hat{L}}{\bar{L}} = \frac{1}{2} - \frac{1}{2}\frac{4-8\rho+\rho^2}{(2-\rho)(\rho+2)} : \begin{cases} > \frac{1}{2} & \text{si } \rho > 4-2\sqrt{3} \approx 0.54 \\ < \frac{1}{2} & \text{si } \rho < 4-2\sqrt{3} \end{cases}$$

Efecto marginal de añadir un servidor

- ¿Cuál es la disminución relativa en la congestión media?

$$\frac{\bar{L}-\hat{L}}{\bar{L}} = \frac{1}{2} - \frac{1}{2}\frac{4-8\rho+\rho^2}{(2-\rho)(\rho+2)} : \begin{cases} > \frac{1}{2} & \text{si } \rho > 4-2\sqrt{3} \approx 0.54 \\ < \frac{1}{2} & \text{si } \rho < 4-2\sqrt{3} \end{cases}$$

- En palabras: Si $\rho > 0.54$, añadir un servidor reduce la congestión media en más del 50%; si $\rho < 0.54$, la congestión media se reduce en menos del 50 %

- Además, según la cola $M/M/1$ se aproxima a tráfico pesado, la reducción de congestión media resultante de

añadir un servidor se aproxima al 100%:

$$\lim_{\rho \nearrow 1} \frac{\bar{L} - \hat{L}}{\bar{L}} = \lim_{\rho \nearrow 1} \frac{1}{2} - \frac{1}{2} \frac{4 - 8\rho + \rho^2}{(2 - \rho)(\rho + 2)} = 1$$

Relación entre las colas $M/M/1$ y $M/M/m$

- Consideremos una cola $M/M/m$ con parámetros λ, μ: $\rho = \lambda/(m\mu) < 1$

- Ej: Sistema informático con $m \geq 2$ procesadores idénticos en paralelo con velocidad de procesamiento μ

- Consideremos un sistema $M/M/1$ correspondiente, con parámetros λ, $m\mu$

- Ej: Sistema informático con 1 procesador con velocidad de procesamiento $m\mu$, i.e. m veces más rápido

- Pregunta: ¿Cuál de los dos sistemas es "mejor", i.e. presenta menor congestión y tiempo medio de respuesta?

Relación entre las colas $M/M/1$ y $M/M/m$

- Sea $\bar{L}^{M/M/m}$ el # medio en el sistema para la cola $M/M/m$, y sea $\bar{L}^{M/M/1} = \rho/(1-\rho)$ para la cola $M/M/1$ correspondiente

Relación entre las colas $M/M/1$ y $M/M/m$

- Sea $\bar{L}^{M/M/m}$ el # medio en el sistema para la cola $M/M/m$, y sea $\bar{L}^{M/M/1} = \rho/(1-\rho)$ para la cola $M/M/1$ correspondiente

- En el caso $m = 2$, tenemos que:

$$\bar{L}^{M/M/2} - \bar{L}^{M/M/1} = \frac{\rho}{\rho + 1} > 0$$

es decir, el sistema $M/M/1$ es más eficiente que el $M/M/2$, para un mismo nivel de utilización. ¿Por qué?

Relación entre las colas $M/M/1$ y $M/M/2$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/2} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{1-\rho}{1+\rho}$$

Relación entre las colas $M/M/1$ y $M/M/2$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/2} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{1-\rho}{1+\rho}$$

- El deterioro relativo es cada vez menos significativo según el sistema se aproxima a un régimen de tráfico pesado:

$$\lim_{\rho \nearrow 1} \frac{\bar{L}^{M/M/2} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \lim_{\rho \nearrow 1} \frac{1-\rho}{1+\rho} = 0$$

Relación entre las colas $M/M/1$ y $M/M/2$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/2} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{1-\rho}{1+\rho}$$

- El deterioro relativo es cada vez menos significativo según el sistema se aproxima a un régimen de tráfico pesado:

$$\lim_{\rho \nearrow 1} \frac{\bar{L}^{M/M/2} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \lim_{\rho \nearrow 1} \frac{1-\rho}{1+\rho} = 0$$

- Así, se considera que, en tráfico pesado, ambos sistemas son "equivalentes"

Relación entre las colas $M/M/1$ y $M/M/3$

- En el caso $m = 3$, tenemos que:

$$\bar{L}^{M/M/3} - \bar{L}^{M/M/1} = 2\frac{\rho\,(3\rho+2)}{3\rho^2+4\rho+2} > \frac{\rho}{\rho+1}$$

es decir, el sistema $M/M/1$ es más eficiente que el $M/M/2$, que a su vez es más eficiente que el $M/M/3$, para un mismo nivel de utilización. ¿Por qué?

Relación entre las colas $M/M/1$ y $M/M/3$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/3} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{2\,(3\rho + 2)}{3\rho^2 + 4\rho + 2}\,(1 - \rho)$$

Relación entre las colas $M/M/1$ y $M/M/3$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/3} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{2\,(3\rho+2)}{3\rho^2+4\rho+2}\,(1-\rho)$$

- El deterioro relativo es cada vez menos significativo según el sistema se aproxima a un régimen de tráfico pesado:

$$\lim_{\rho \nearrow 1} \frac{\bar{L}^{M/M/3} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \lim_{\rho \nearrow 1} \frac{2\,(3\rho+2)}{3\rho^2+4\rho+2}\,(1-\rho) = 0$$

Relación entre las colas $M/M/1$ y $M/M/3$

- Calculamos el deterioro relativo en la congestión media:

$$\frac{\bar{L}^{M/M/3} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \frac{2\,(3\rho+2)}{3\rho^2+4\rho+2}\,(1-\rho)$$

- El deterioro relativo es cada vez menos significativo según el sistema se aproxima a un régimen de tráfico pesado:

$$\lim_{\rho \nearrow 1} \frac{\bar{L}^{M/M/3} - \bar{L}^{M/M/1}}{\bar{L}^{M/M/1}} = \lim_{\rho \nearrow 1} \frac{2\,(3\rho+2)}{3\rho^2+4\rho+2}\,(1-\rho) = 0$$

- Así, se considera que, en tráfico pesado, ambos sistemas son "equivalentes"