Dimensionado y Planificación de Redes

Tema 3 – Modelo M/M/1 y extensiones

Ramón Agüero Calvo
ramon.agueroc@unican.es

Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Conceptos básicos

- Modelos de Poisson
  - El número de llegadas en un tiempo determinado sigue una distribución de Poisson

$$\Pr\{k \text{ llegadas en } T\} = P_k(T) = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

    - Implica que el tiempo entre llegadas consecutivas sigue una distribución exponencial negativa de tasa λ
  - La duración de las llamadas sigue una función densidad de probabilidad (fdp) exponencial negativa de media $1/\mu$

$$f_{T_s}(t_s) = \mu e^{-\mu t_s}$$

  - La tasa de llegadas al sistema es constante (proceso estacionario)
- Procesos de nacimiento y muerte

Conceptos básicos

- Relación de Little: $N = W \lambda$
  - N: Número medio de elementos en el sistema (cola, servidores o sistema completo)
  - W: Tiempo medio de permanencia en el sistema
  - λ: Tasa de llegadas al sistema

Modelo M/M/1 – Nodo de comunicaciones

- Se modela el nodo de comunicaciones como un sistema en el que…
  - Las llegadas siguen un proceso de Poisson de intensidad λ
  - La distribución del tiempo de servicio es exponencial, con media $1/\mu$ (t_s)

$$t_s = \frac{1}{\mu} = \frac{L}{C}$$

L: Longitud media de paquete (exp. negativa)
C: Capacidad interfaz (constante)

  - Sólo hay un única interfaz para transmitir los paquetes
  - La cola de espera se supone infinita, así que no hay pérdida
- Se modela el sistema como una sucesión de estados: N = {0, 1, 2…}
  - Cada estado se corresponde con el # de paquetes que hay en el sistema
  - Si n>1 hay algún paquete esperando
- Se puede demostrar que la probabilidad de cada estado es $p_n = \rho^n(1-\rho)$

  donde ρ (tráfico) se puede calcular como $\rho = \frac{\lambda}{\mu} = \lambda \cdot t_s = \frac{\lambda \cdot L}{C}$

Modelo M/M/1 – Nodo de comunicaciones

- A partir de dicho resultado se puede caracterizar el sistema…
  - Número medio de clientes en el sistema

$$\overline{N_t} = \sum_{k=0}^{\infty} k \cdot p_k = \sum_{k=0}^{\infty} k \cdot \rho^k (1-\rho) = \rho(1-\rho) \sum_{k=0}^{\infty} k \cdot \rho^{k-1} = \rho(1-\rho) \frac{d}{d\rho} \left(\sum_{k=0}^{\infty} \rho^k \right) =$$
$$= \rho(1-\rho) \frac{d}{d\rho} \left(\frac{1}{1-\rho} \right) = \frac{\rho}{1-\rho}$$

  - Número medio de clientes en la cola

$$\overline{N_Q} = \sum_{k=1}^{\infty} (k-1) p_k = \sum_{k=1}^{\infty} (k-1) \rho^k (1-\rho) = (1-\rho) \sum_{t=0}^{\infty} t \rho^{t+1} = \frac{\rho^2}{1-\rho}$$

  - Tiempo medio en el sistema y en la cola, aplicando la relación de Little

$$\left.\begin{aligned} \overline{T_t} = \tau = \frac{\overline{N_t}}{\lambda} = \frac{\rho}{1-\rho} \frac{1}{\lambda} = \frac{1}{\mu - \lambda} = \frac{t_s}{1-\rho} \\ \overline{T_Q} = \frac{\overline{N_Q}}{\lambda} = \frac{\rho^2}{1-\rho} \frac{1}{\lambda} = \frac{\rho}{\mu - \lambda} = \frac{t_s \rho}{1-\rho} \end{aligned}\right\} \quad \overline{T_t} - \overline{T_Q} = \frac{1}{\mu - \lambda} - \frac{\rho}{\mu - \lambda} = \frac{1}{\mu} = t_s$$

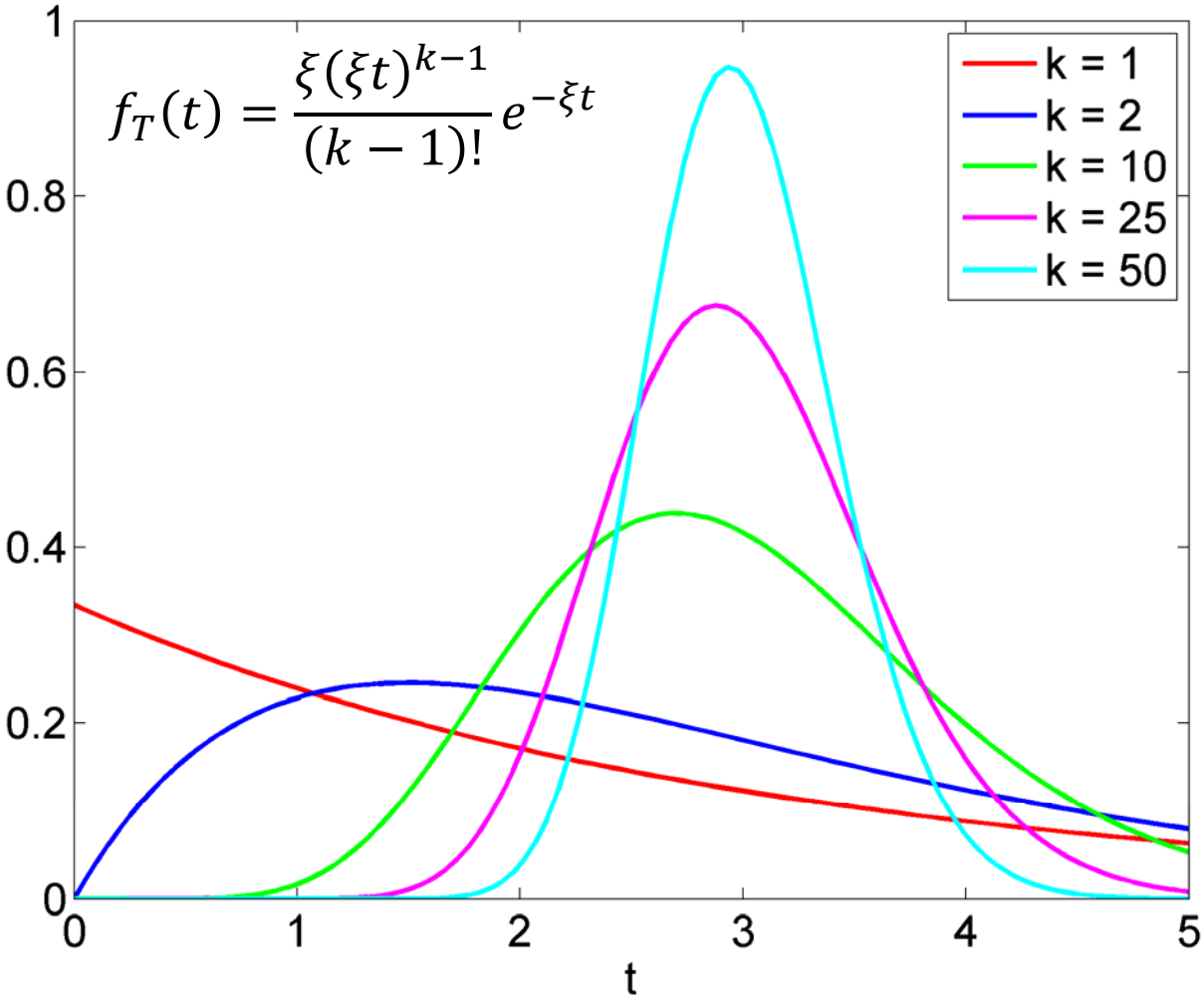
Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Introducción

- Se basa en el método de las fases, y asume una llegada *memoryless* (Poisson)
- Se asume que el servicio (servidor) se constituye de una concatenación de servicios en serie
  - Sólo puede haber una llamada en el sistema completo, por lo que no es posible que haya más de una 'fase' ocupada
- Cada una de las fases tiene un tiempo de servicio, distribuido según una va exponencial negativa de tasa $\xi = k\mu$
- El tiempo de servicio total será la suma de los tiempos de servicio de cada fase
- La suma de k variables aleatorias exponenciales negativas con media $1/k\mu$ da lugar a una variable aleatoria con distribución Erlang-k

Distribución Erlang-k



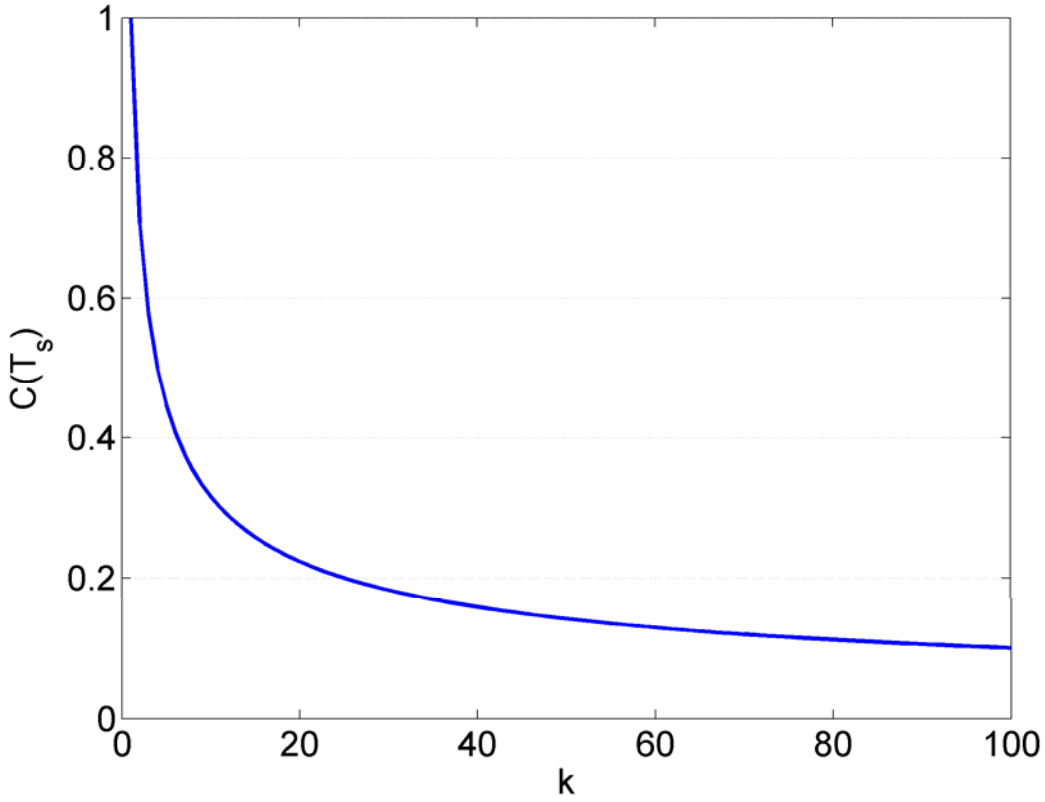
Distribución Erlang-k

- Sus valores característicos son los que aparecen a continuación
  - Se recuerda que la tasa de cada una de las 'fases' es kμ

$$\overline{T_s} = \frac{k}{\xi} = \frac{1}{\mu}$$

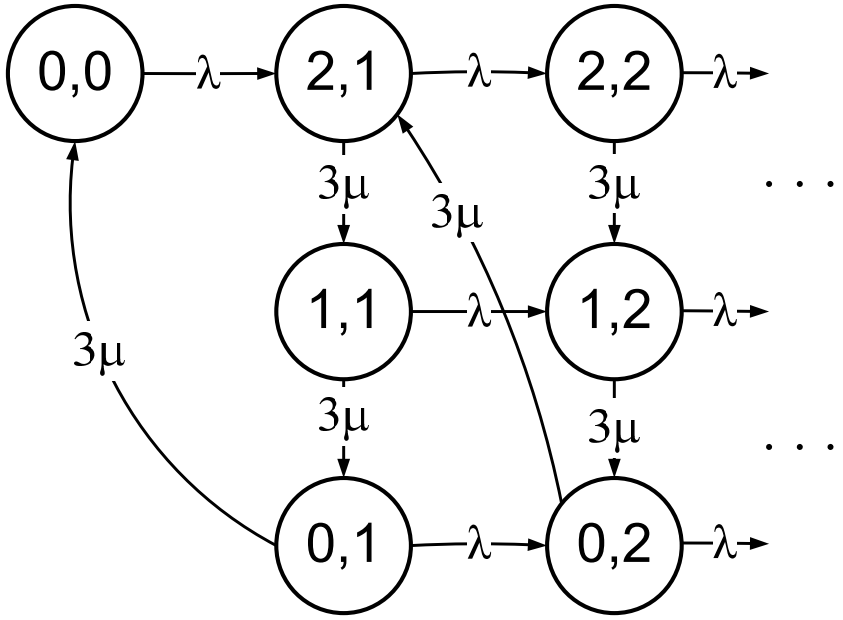
$$\sigma(T_s) = \frac{\sqrt{k}}{\xi} = \frac{1}{\sqrt{k} \cdot \mu}$$

$$C(T_s) = \frac{1}{\sqrt{k}}$$



Cadena equivalente

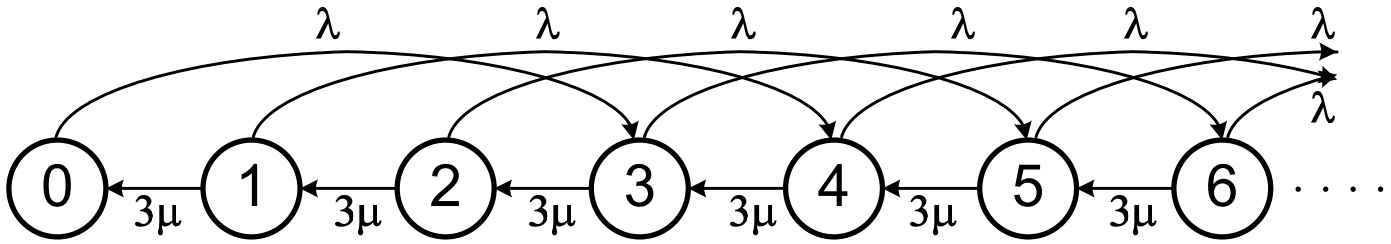
- Su solución exacta se realiza a base una cadena de Markov en dos dimensiones [$X_1(t)$, $X_2(t)$], con…
  - $X_1(t)$: cadena en dirección vertical, que modela la descomposición del servidor
  - $X_2(t)$: cadena en dirección horizontal, que modela el número de paquetes en el sistema



Ejemplo ilustrativo, con k = 3

Cadena equivalente

- Cadena unidemensional: llegada “a ráfagas”



- Ecuaciones de balance de flujo

$$\lambda p_0 = k\mu p_1$$

$$(\lambda + k\mu)p_i = k\mu p_{i+1} \qquad i = 1 \ldots k-1$$

$$(\lambda + k\mu)p_i = \lambda p_{i-k} + k\mu p_{i+1} \qquad i = k \ldots \infty$$

- Solución compleja – ecuaciones diferenciales: transformada Z

Solución

- Número de 'fases' que quedan en el sistema

$$P_i = (1-\rho)\sum_{m=1}^{k} A_m (z_m)^{-i} \qquad i = 1,2 \ldots$$

$$A_m = \prod_{\substack{j=1 \\ j\neq m}}^{k} \frac{1}{1-\frac{z_m}{z_j}}$$

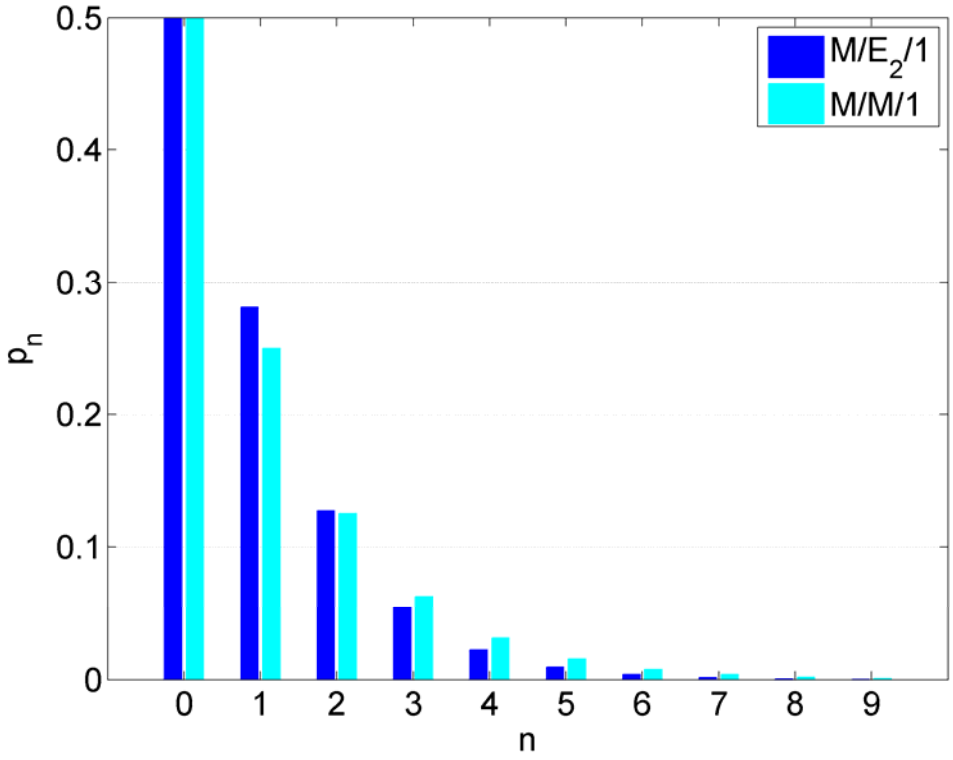
  - Los z_j son las raíces del polinomio $k\mu - \lambda\left(z + z^2 + \cdots + z^k\right)$

- Finalmente, se puede obtener el número total de "clientes" en el sistema

$$p_n = \sum_{i=(n-1)k+1}^{nk} P_i \qquad n = 1,2, \ldots$$

Ejemplo

- Parámetros del modelo
  - λ = 50 paquetes por segundo (proceso de Poisson)
  - Ts = 10 ms (μ = 0.01)
  - A (ρ) = 0.5
  - k = 2



Aplicación – Llegada a ráfagas

- El modelo M/Ek/1 visto anteriormente se puede utilizar para analizar un sistema en el que las llegadas (proceso de Poisson) vienen con k clientes/procesos, etc
  - Por ejemplo se podría modelar situaciones en las que los paquetes generados por la aplicación tuvieran un tamaño grande y tuvieran que ser “troceados”
- Para que la equivalencia fuera total, la tasa de la variable aleatoria del tiempo de servicio debería ser $k\mu$
- El modelo se podría generalizar a situaciones en las que el tamaño del “batch” fuera aleatorio

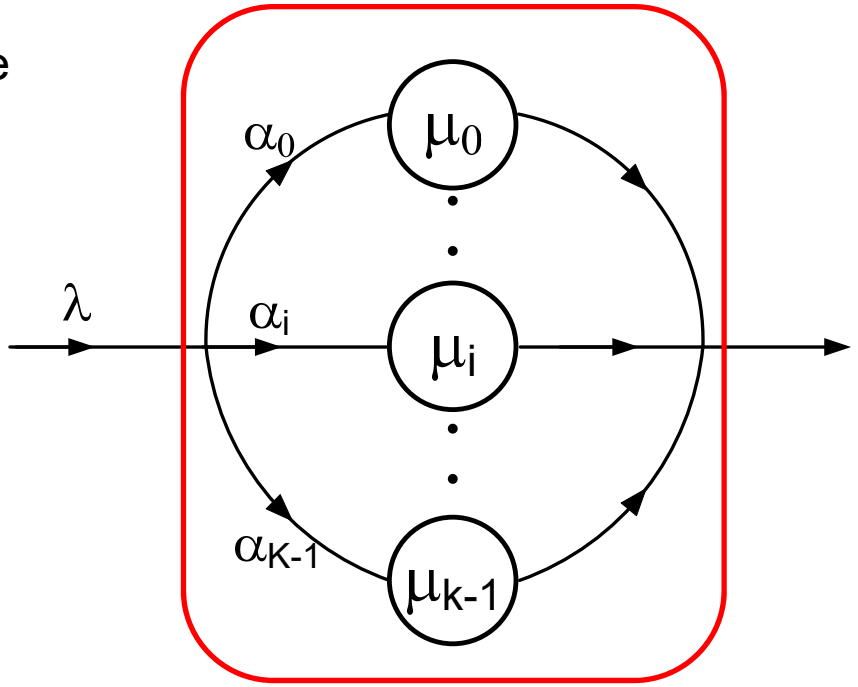
Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Introducción

- Se asume que el servidor (servicio) se compone de k entidades
  - Cada entidad tiene un tiempo de servicio modelado con una variable aleatoria exponencial negativa de tasa μ_i
- El cliente, al llegar al servicio, elige la rama *i-ésima* con probabilidad α_i
  - Se cumple que $\sum_{i=0}^{k-1} \alpha_i = 1$
- El tiempo de servicio *global* se corresponde con una variable aleatoria hiperexponencial

$$f_{T_s} = \sum_{i=0}^{k} \alpha_i \mu_i e^{-\mu_i t}$$



Servicio 'completo'

Distribución hiper-exponencial

- Valores característicos

$$\overline{T_s} = \sum_{i=0}^{k-1} \frac{\alpha_i}{\mu_i} \qquad \sigma(T_s) = \sqrt{2\sum_{i=0}^{k-1} \frac{\alpha_i}{\mu_i^2} - \left(\sum_{i=0}^{k-1} \frac{\alpha_i}{\mu_i}\right)^2}$$

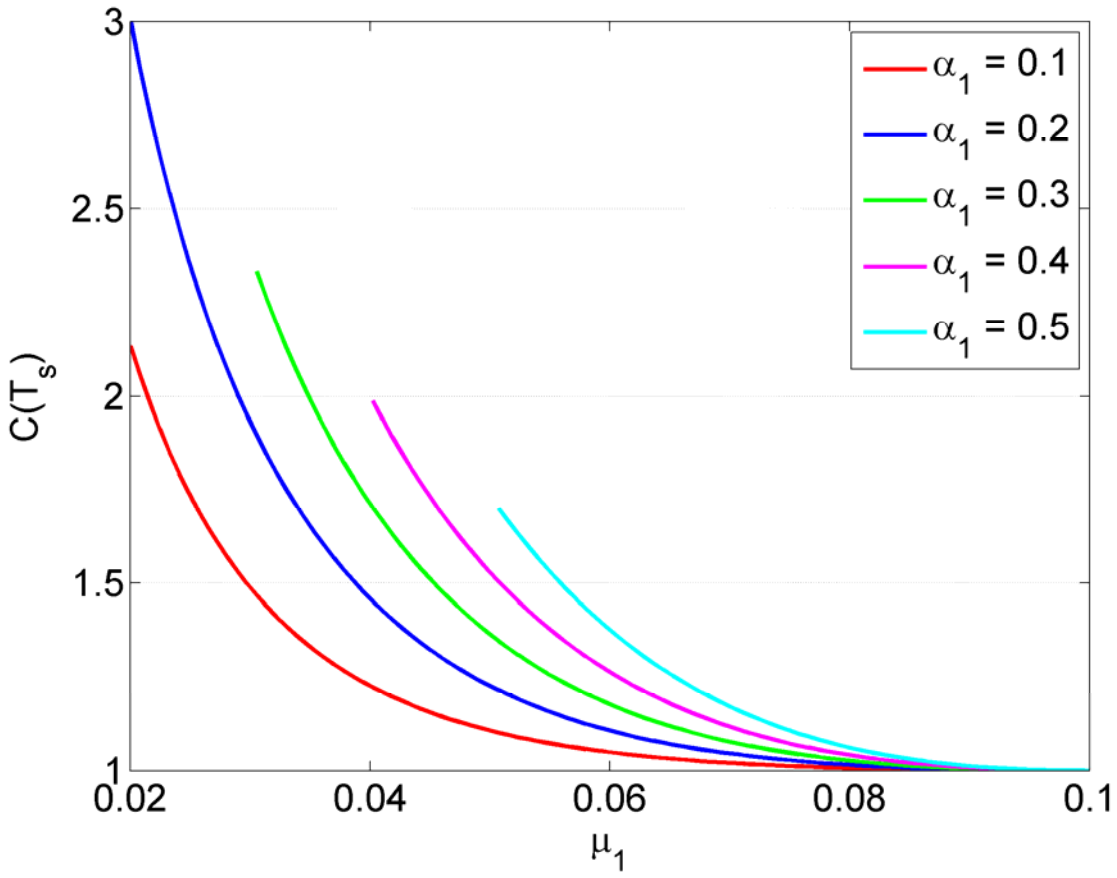
- Coeficiente de dispersión

$$C(T_s) = \sqrt{\frac{2\sum_{i=0}^{k-1}\left(\frac{\alpha_i}{\mu_i^2}\right)}{\left(\sum_{i=0}^{k-1}\frac{\alpha_i}{\mu_i}\right)^2} - 1}$$

- C(Ts) ≥ 1 (Desigualdad de Cauchy-Schwarz)

Variación del C(Ts)

- Ejemplo ilustrativo, con…
  - K = 2
  - E(Ts) = 10 ms



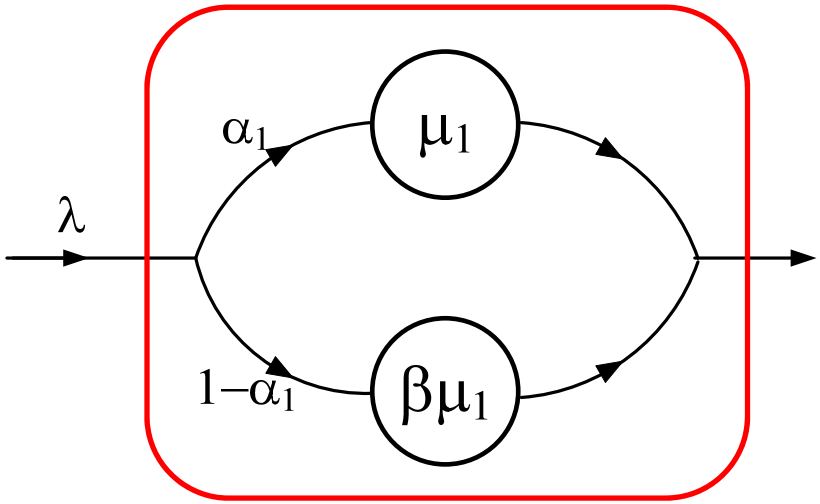
Ejemplo (k=2)

- Si k = 2 ($\alpha_2 = 1 - \alpha_1$), y asumimos que $\mu_2 = \beta\ \mu_1$, resulta que…

$$\overline{T_s} = \frac{1}{\mu_1}\left(\alpha_1 + \frac{1-\alpha_1}{\beta}\right)$$

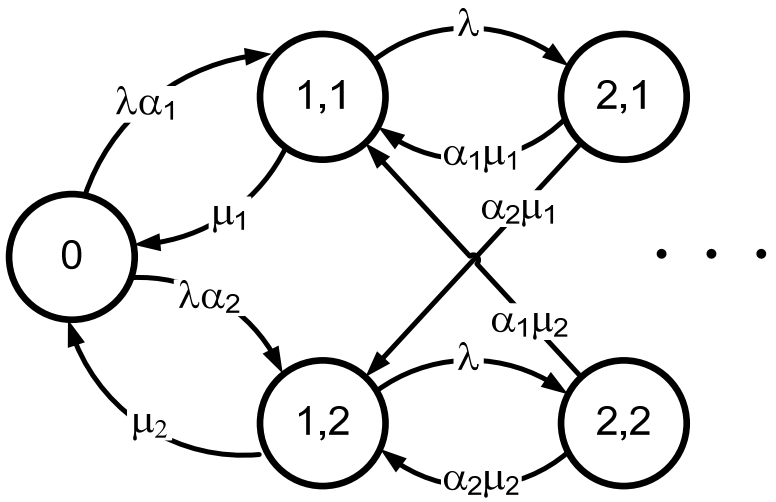
$$\sigma(T_s) = \frac{1}{\mu_1}\sqrt{2\left(\alpha_1 + \frac{1-\alpha_1}{\beta^2}\right) - \left(\alpha_1 + \frac{1-\alpha_1}{\beta}\right)^2}$$

$$C(T_s) = \sqrt{\frac{2\left(\alpha_1 + \frac{1-\alpha_1}{\beta^2}\right)}{\left(\alpha_1 + \frac{1-\alpha_1}{\beta}\right)^2} - 1}$$



Servicio 'completo' – E(Ts) = μ^{-1}

Ejemplo (k=2) – Cadena equivalente



- Ecuaciones de balance de flujo

$$p_0\lambda = p_{1,1}\mu_1 + p_{1,2}\mu_2$$
$$p_{i,1}(\lambda + \mu_1) = p_{i-1,1}\lambda + p_{i+1,1}\alpha_1\mu_1 + p_{i+1,2}\alpha_1\mu_2$$
$$p_{i,2}(\lambda + \mu_2) = p_{i-1,2}\lambda + p_{i+1,1}\alpha_2\mu_1 + p_{i+1,2}\alpha_2\mu_2$$

  - De manera genérica

$$p_{i,j}(\lambda + \mu_j) = p_{i-1,j}\lambda + \alpha_j(p_{i+1,1}\mu_1 + p_{i+1,2}\mu_2)$$

Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Introducción

- El modelo M/G/1 considera…
  - Proceso de llegadas según un proceso de Poisson
  - Tiempo de servicio distribuido según una variable aleatoria genérica
  - Un servidor
  - Capacidad de espera infinita
- Los casos M/Ek/1 y M/Hk/1 son particularizaciones y se ha visto que su solución “exacta” es compleja
- Para analizar el rendimiento de este tipo de sistemas se utiliza el método de las cadenas de Markov *embebidas*
- Se obtienen los valores medios estudiados para el caso del M/M/1: número medio de unidades en el sistema y en la cola de espera
  - A partir de ellos se obtiene los tiempos medios
- Se sigue cumpliendo que $p_0 = 1 - \rho$ y que el número medio de servidores ocupados (factor de utilización) es ρ

Fórmula Pollaczek-Khintchine (PK)

- Considerar un usuario que llega a un sistema; ¿cómo se puede determinar el retraso que tendrá que soportar?
- Dicho retraso tiene dos posibles contribuciones
  - Los clientes que pudieran estar esperando
  - El cliente que pudiera estar en el servidor
- Si hay N_Q clientes esperando, el retraso *total (valor medio)* que ocasionan será $N_Q \cdot E(T_S)$
- El retardo ocasionado por el cliente que se encuentra en el servidor será E(tiempo servicio residual | servidor ocupado)
- En total se tendrá que…

$$T_Q = N_Q E(T_s) + \Pr\{\text{servidor ocupado}\} \cdot E(\text{tiempo servicio residual} \mid \text{servidor ocupado})$$

Fórmula Pollaczek-Khintchine (PK)

- Por la relación de Little se sabe que $N_Q = T_Q \cdot \lambda$
- Además, la probabilidad de que el servidor esté ocupado coincide con el factor de utilización, y $\rho = \lambda \cdot E(Ts)$, con lo que tendremos que…

$$T_Q = \frac{\rho \cdot E(\text{tiempo servicio residual} \mid \text{servidor ocupado})}{1-\rho}$$

- Solo queda por conocer E(tiempo servicio residual | servidor ocupado); por la *renewal theory* se obtiene que…

$$E(\text{tiempo servicio residual} \mid \text{servidor ocupado}) = \frac{E(T_s^2)}{2E(T_s)} = \frac{1+C(T_s)^2}{2}E(T_s)$$

- Para obtener finalmente que…

$$\boxed{T_Q = \frac{1+C(T_s)^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E(T_s)}$$

Fórmula de Pollaczek-Khintchine (PK)

Fórmula Pollaczek-Khintchine (PK)

- Solo se requiere conocer los momentos de orden 1 y 2 de la variable aleatoria tiempo de servicio para calcular el número medio de unidades en el sistema de espera
- Particularización de la fórmula de PK

Sistema	Tiempo de servicio (va)	C(Ts)	T_Q
M/M/1	Exponencial negativa	1	$\frac{\rho}{1-\rho} \cdot E(T_s)$
M/Ek/1	Erlang-k	$\frac{1}{\sqrt{k}}$	$\frac{k+1}{2k} \cdot \frac{\rho}{1-\rho} \cdot E(T_s)$
M/D/1	Determinista	0	$\frac{1}{2} \cdot \frac{\rho}{1-\rho} \cdot E(T_s)$

- Se observa que…

$$T_Q = \frac{1+C(T_s)^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E(T_s) = \frac{1+C(T_s)^2}{2} (T_Q)_{MM1}$$

Caracterización de un sistema M/G/1

- A partir de la fórmula PK, se pueden obtener todos los parámetros de rendimiento del sistema…

Sistema	Tiempo de permanencia	Número de unidades (paquetes) Little: $\mathbf{N = T \cdot \lambda}$
Servidor	$T_S = E(T_s)$	$N_S = \rho$
Cola	$T_Q = \frac{1 + C(T_s)^2}{2} \cdot \frac{\rho}{1-\rho} \cdot E(T_s)$	$N_Q = \frac{1 + C(T_s)^2}{2} \cdot \frac{\rho^2}{1-\rho}$
Total	$T_t = \left[1 - \frac{\rho}{2}(1 - C(T_s)^2)\right] \cdot \frac{E(T_s)}{1-\rho}$	$N_t = \left[1 - \frac{\rho}{2}(1 - C(T_s)^2)\right] \cdot \frac{\rho}{1-\rho}$

Contenido

- Repaso
- Modelo M/Ek/1
- Modelo M/Hk/1
- Modelo M/G/1
- Sistemas con prioridad

Motivación

- En sistemas de espera pura el QoS se puede “asociar” al retardo medio total (tiempo de permanencia en el sistema)
- ¿Qué sucede cuando hay varios tipos de servicio?
  - Se tienen que considerar los parámetros para cada tipo de servicio de manera individual
- Posibilidades de diseño
  - Túneles virtuales (*virtual tunnels*): hay una cola y una capacidad “reservada” para cada servicio – varios M/M/1 independientes entre sí
  - Sobredimensionado (*over-engineering*): se utiliza el criterio de QoS más restrictivo para dimensionar el sistema y todos los servicios comparten el *buffer* de espera y la capacidad
  - Prioridad (*priority queing*): todos los servicios comparten la misma capacidad, pero se establecen diferentes *buffer* de espera y prioridades entre ellos

Planteamiento

- Se cuenta con un sistema de un único servidor que gestiona varios tipos de servicio
- Cada uno de esos servicios está caracterizado por…
  - Llegadas: proceso de Poisson de tasa λ_k
  - Tiempo de servicio: variable aleatoria T_s (genérica)
- Cuando el servidor está vacío selecciona "paquetes" (nodo de comunicaciones) por orden de prioridad
  - Siempre drenará los paquetes del buffer *i-1* antes de comenzar con los del buffer *i*
- Se asume además que los paquetes no pueden "desplazar" a otros cuando están siendo servidos (a pesar de que sean de mayor prioridad): *nonpreemptive system*
- La utilización del servidor para los paquetes del servicio k-ésimo será $\rho_k = \lambda_k \cdot E[(T_s)_k]$
- Se tiene que cumplir que el factor de utilización total $\rho = \sum_{i=0}^{K} \rho_i < 1$ (condición de estabilidad)

Ejemplo con dos servicios $(T_Q)_1$

- Sean dos servicios (i = 1,2), con 1 siendo más prioritario que 2
- ¿Cuál es el tiempo medio de espera para los "paquetes" del servicio 1?
- Se pueden distinguir dos contribuciones
  - Paquetes de tipo 1 que estén esperando: $(N_Q)_1$
  - Tiempo que le queda al paquete que esté en el servidor (puede ser de tipo 1 ó 2) – tiempo de servicio residual
- Tendremos, por tanto…

$$(T_Q)_1 = (N_Q)_1 E[(T_s)_1] + \Pr\{\text{servidor ocupado}\}\, E(t_s \text{ residual} \mid \text{servidor ocupado})$$

- Teniendo en cuenta que $(N_Q)_1 = \lambda_1 \cdot (T_Q)_1$, $\lambda_1 \cdot E[(T_S)_1] = \rho_1$, y Pr{servidor ocupado} = ρ, se llega a…

$$(T_Q)_1 = \frac{\rho \cdot E(t_s \text{ residual} \mid \text{servidor ocupado})}{1 - \rho_1}$$

Ejemplo con dos servicios $(T_Q)_1$

- En este caso, la esperanza del tiempo de servicio residual tendrá que considerar que el servicio activo puede ser de tipo 1 ó 2

$$E(t_s \text{ residual} \mid \text{servidor ocupado}) = \Xi =$$
$$= E\left((t_s)_1 \text{ residual} \middle| \text{servicio 1 ON}\right) \cdot \Pr\{1 \text{ ON} \mid \text{ servidor ocupado}\} +$$
$$+ E\left((t_s)_2 \text{ residual} \middle| \text{servicio 2 ON}\right) \cdot \Pr\{2 \text{ ON} \mid \text{ servidor ocupado}\}$$

donde…

$$\Pr\{j \text{ ON} \mid \text{ servidor ocupado}\} = \frac{\rho_j}{\rho}$$

$$E\left((t_s)_j \text{ residual} \mid \text{servicio } j \text{ ON}\right) = \frac{E\left[(T_s)_j^2\right]}{2E\left[(T_s)_j\right]}$$ *(Renewal theory)*

Ejemplo con dos servicios $(T_Q)_1$

- Quedaría entonces…

$$\rho \cdot \Xi = \rho \left\{ \frac{E[(T_s)_1^2]}{2E[(T_s)_1]} \cdot \frac{\rho_1}{\rho} + \frac{E[(T_s)_2^2]}{2E[(T_s)_2]} \cdot \frac{\rho_2}{\rho} \right\} = \frac{1}{2} \{ \lambda_1 \cdot E[(T_s)_1^2] + \lambda_2 \cdot E[(T_s)_2^2] \}$$

- Con lo que se llegaría a la siguiente expresión…

$$(T_Q)_1 = \frac{\frac{1}{2} \{ \lambda_1 \cdot E[(T_s)_1^2] + \lambda_2 \cdot E[(T_s)_2^2] \}}{1 - \rho_1}$$

Ejemplo con dos servicios $(T_Q)_2$

- Cuando llega un servicio de tipo 2 al sistema, su tiempo de espera tendrá cuatro contribuciones
  - El tiempo correspondiente a los paquetes de tipo 1 que estuvieran esperando: $(N_Q)_1 E[(T_s)_1]$
  - Ídem para los paquetes de tipo 2 que estuvieran esperando: $(N_Q)_2 E[(T_s)_2]$
  - Tiempo que le queda al paquete que esté en el servidor (puede ser de tipo 1 ó 2) – tiempo de servicio residual (igual que para los servicios de tipo 1): $\rho \cdot \Xi$
  - El tiempo correspondiente a los servicios de los paquetes de tipo 1 que pudieran llegar mientras que el servicio objeto de estudio está esperando: $\lambda_1 (T_Q)_2 E[(T_s)_1]$, ya que el número de paquetes que llegarían en dicho periodo es $\lambda_1 (T_Q)_2$

Ejemplo con dos servicios $(T_Q)_2$

- Se tiene por tanto, que…

$$(T_Q)_2 = (N_Q)_1 E[(T_s)_1] + (N_Q)_2 E[(T_s)_2] + \rho \cdot \Xi + \lambda_1 (T_Q)_2 E[(T_s)_1] =$$

$$= (T_Q)_1 \cdot \rho_1 + (T_Q)_2 \cdot \rho_2 + \rho \cdot \Xi + \rho_1 (T_Q)_2 = (T_Q)_2 [\rho_1 + \rho_2] + \rho_1 \frac{\rho \cdot \Xi}{1 - \rho_1} + \rho \cdot \Xi$$

- Con un poco de álgebra se llega a…

$$(T_Q)_2 [1 - \rho_1 - \rho_2] = \frac{\rho \cdot \Xi}{(1 - \rho_1)} \rightarrow (T_Q)_2 = \frac{\frac{1}{2}\{\lambda_1 \cdot E[(T_s)_1^2] + \lambda_2 \cdot E[(T_s)_2^2]\}}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Generalización para K servicios

- La fórmula anterior se puede generalizar para el caso en el que haya K servicios (i=1…K), ordenados por prioridad (de mayor a menor)

$$(T_Q)_i = \frac{\frac{1}{2}\sum_{i=1}^{K} \lambda_i \cdot E[(T_s)_i^2]}{(1-\sigma_{i-1})(1-\sigma_i)}$$

donde $\sigma_i = \sum_{m=1}^{i} \rho_m$

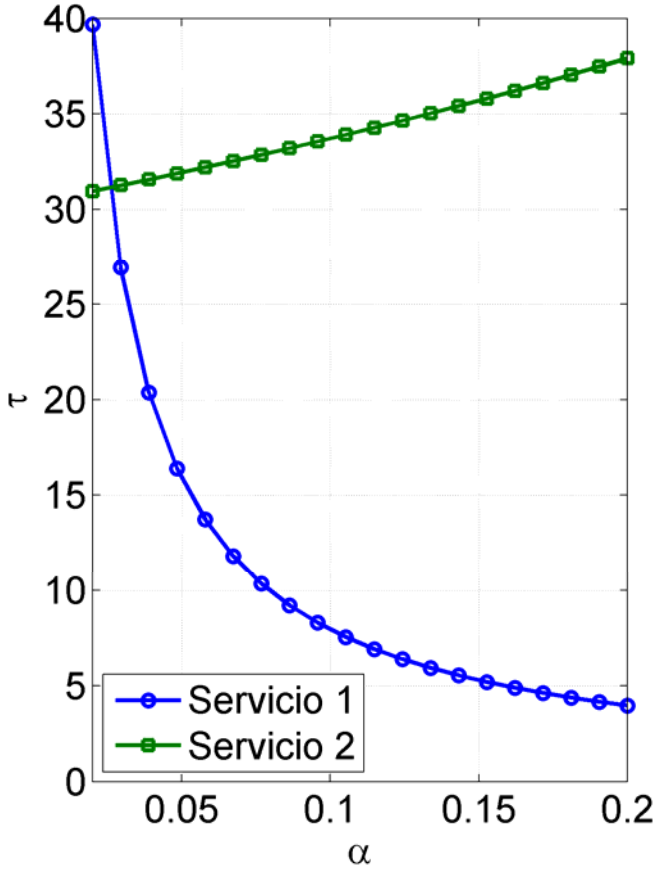
- Notar que el numerador se puede expresar como…

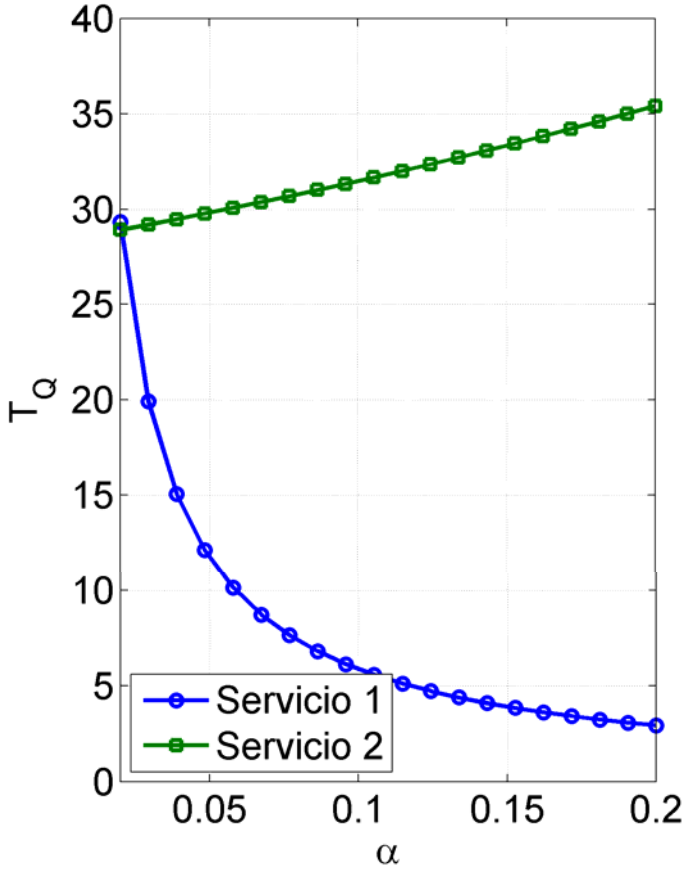
$$\frac{1}{2}\sum_{i=1}^{K} \lambda_i \cdot E\left[(T_s)_i^2\right] = \frac{\lambda}{2}\sum_{i=1}^{K} \frac{\lambda_i}{\lambda} \cdot E\left[(T_s)_i^2\right]$$

Ejemplo

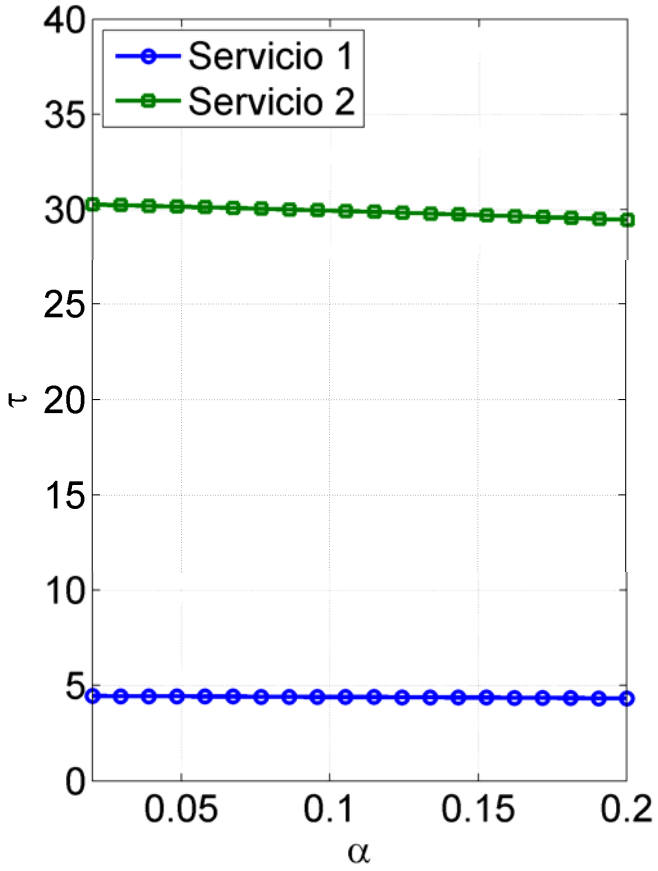
- A un nodo de comunicaciones llegan paquetes de dos tipos de servicios (en ambos casos según un proceso de Poisson)
  - VoIP (1) – Longitud fija de 53 Bytes
  - Datos (2) – Longitud media de 512 Bytes, con desviación típica de 1024 Bytes
- La capacidad del enlace es de 2048 kbps
- Se pretende que el tráfico total (factor de utilización – ρ) sea inferior a 0.85
- Se asume que el tráfico de voz es $\alpha \cdot \rho$, y se "calcula" el tráfico de datos, para mantener $\rho = 0.85$
- Se utilizan dos esquemas: túneles virtuales (cada servicio tiene un sistema M/G/1 diferente) y un esquema de prioridad (VoIP más prioridad que Datos)

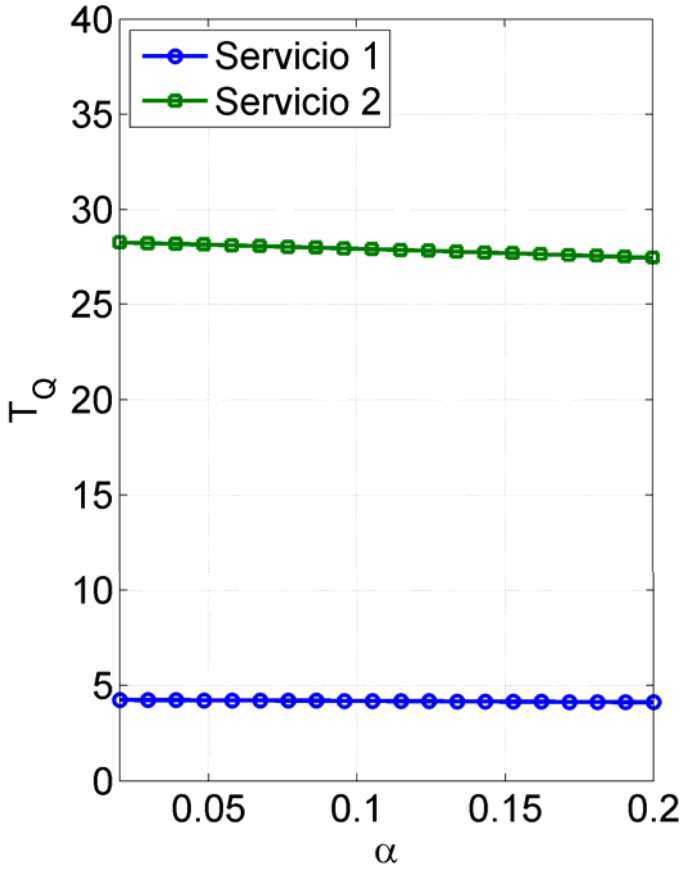
Resultados – Túneles virtuales





Resultados – Sistema con prioridad





Comparativa

- Se representa el cociente entre el rendimiento (QoS) del sistema con prioridad y el que emplea túneles virtuales