

Redes de Comunicación

Extensión del modelo M/M/1

Instructor

Dr.-Ing. K.D. HACKBARTH

Versión 11. 10. 2012

© Universidad de Cantabria

- Motivación
- M/G/1
 - El modelo M/Ek/1
 - El modelo M/Hk/1
 - El modelo generalizado M/G/1
- M/G/1 con priorización
 - QoS y SdC
 - SdC con prioridad
 - Modelo NPPQ

- En la asignatura de Redes de Comunicaciones se introdujo el modelo M/M/1
- El modelo M/M/1 se basa en
 - Un proceso de llegada de Poisson
 - Un tiempo de servicio (en el servidor) basado en una fdp exponencial negativa
- En el capítulo de tráfico de fuentes ya se ha discutido que hay muchos sistemas en los que...
 - La fdp del tiempo de servicio no es exponencial negativa e, incluso, el proceso de llegada no tiene una fdp de Poisson
- En este capítulo se generaliza el proceso M/M/1 en procesos M/G/1 con fdp de duración genérica

- En las redes actuales se integran diferentes servicios con carácter de QoS muy diferenciado, desde voz/vídeo hasta tráfico “*best effort*”
- Se requiere, por tanto:
 - Tratamiento diferenciado de los paquetes
 - Se planteará un modelo M/G/1 con sistema de prioridad (colas)
- Aplicación práctica para la planificación de redes privadas virtuales (VPN)

- El modelo M/M/1 se basa en una cadena de nacimiento y muerte
 - Tasa de nacimiento: λ
 - Tasa de muerte: μ
- Al aplicar las condiciones de Markov, se tiene que...
 - El tiempo entre llegadas (t_{ia}) sigue una distribución exponencial negativa: $f_{T_{ia}}(t) = \lambda \exp(-\lambda t)$
 - El tiempo de servicio (t_s) también tiene una distribución exponencial negativa: $f_{T_s}(t) = \mu \exp(-\mu t)$

- En la práctica hay tipos de tráfico en los que la llegada sigue una fdp de *Poisson*, pero la longitud de paquete no se puede modelar con una fdp exponencial negativa
- En el caso exponencial negativa $f_{T_s}(t) = \mu \exp(-\mu t)$
 - $E(T_s) = \sigma(T_s) = 1/\mu$, con lo que: $C(T_s) = 1$
- Para el caso general se puede distinguir
 - $C(T_s) < 1$ que se aproxima con un modelo M/Ek/1
 - $C(T_s) > 1$ que se aproxima con un modelo M/Hk/1

- El modelo M/E_k/1 resulta de una descomposición de un servidor con tasa μ y T_s , con llegadas según una fdp de Erlang-k, en k servidores en serie, con tasa de $\xi = k\mu$, con tiempos de servicio distribuidos (individualmente) según una fdp exponencial negativa
- Resulta:

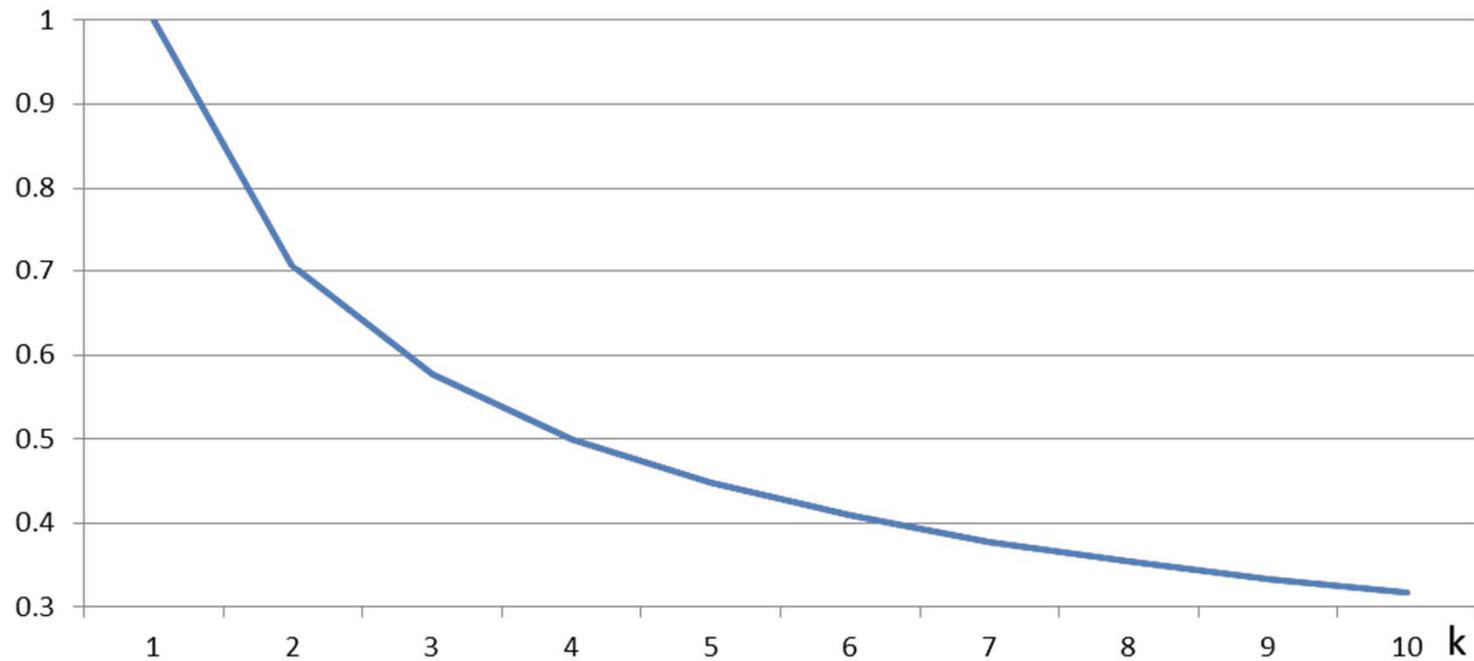
$$f_{T_s}(t) = \frac{\xi(\xi t)^{k-1}}{(k-1)!} \exp(-\xi t)$$

$$E(T_s) = \frac{k}{\xi} \quad \sigma(T_s) = \frac{\sqrt{k}}{\xi} \quad C(T_s) = \frac{1}{\sqrt{k}}$$

- O, en función de μ

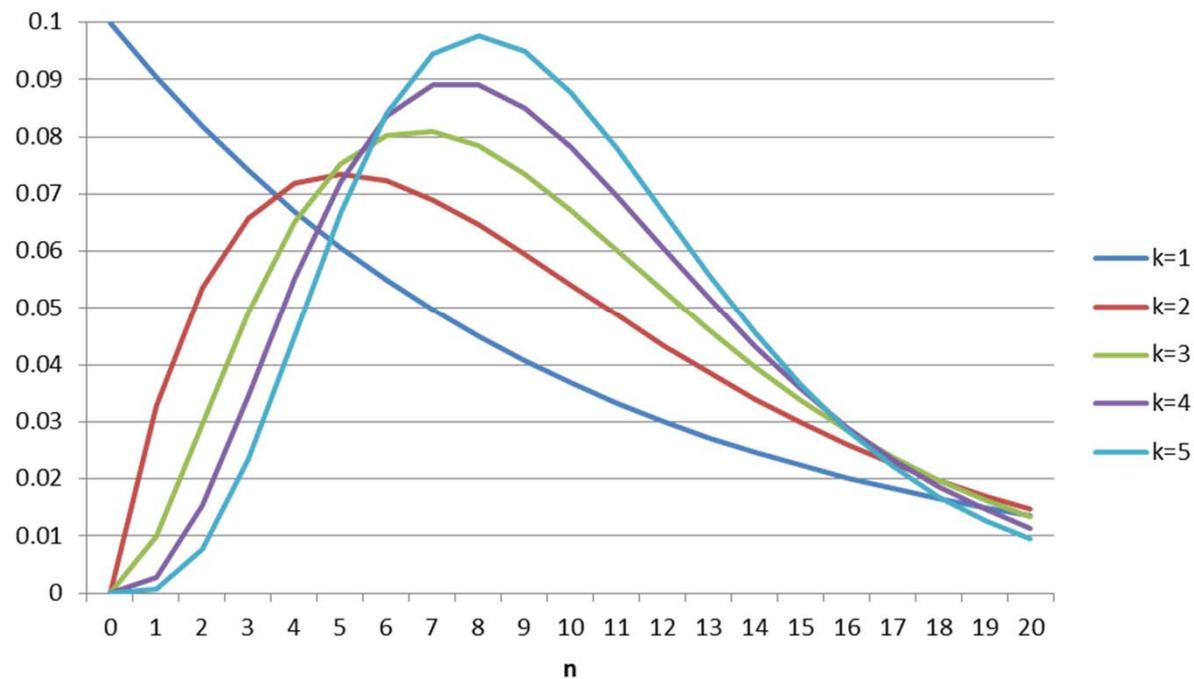
$$E(T_s) = \frac{1}{\mu} \quad \sigma(T_s) = \frac{1}{\sqrt{k} \cdot \mu} \quad C(T_s) = \frac{1}{\sqrt{k}}$$

$C(T_s)$ en función de k

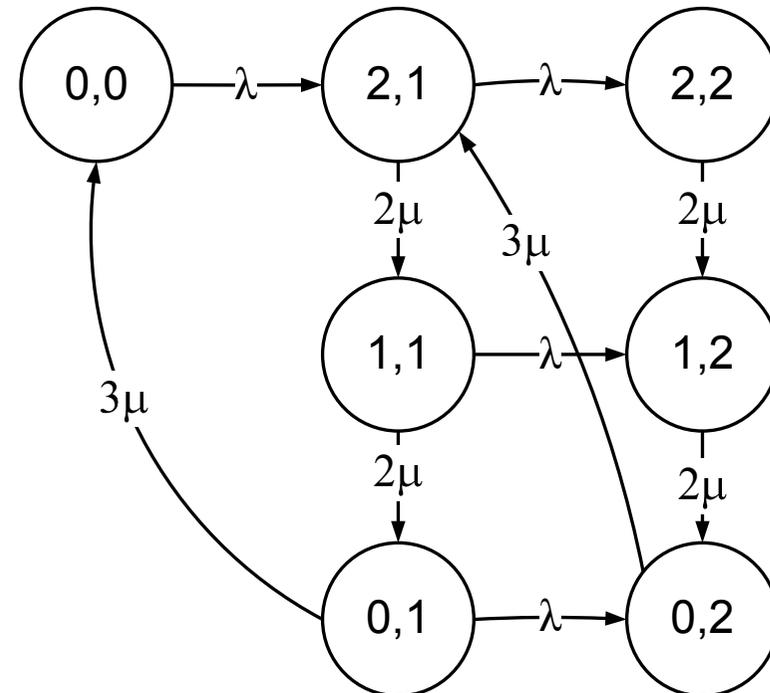


Ejemplo $\mu = 0.1 - k = 1..5$

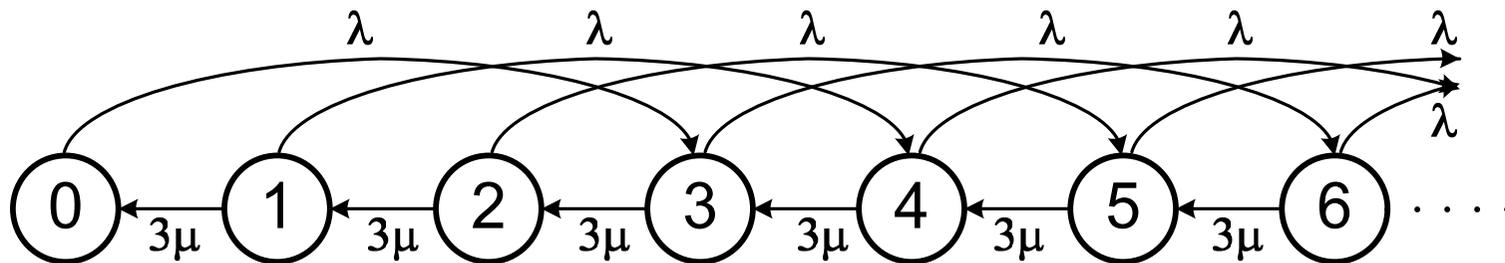
fdp de T_s en función de t y k



- 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 SdC
 - La figura representa un sistema ilustrativo, con $k=3$



- Se transforma en una cadena estrechada, cuya solución en el estado estacionario se deduce a partir de las siguientes ecuaciones
 - $\lambda p_0 = k\mu p_1$
 - $(\lambda + k\mu)p_n = \lambda p_{n-k} + k\mu p_{n+1} \quad n = 1 \dots \infty$ y $p_j = 0$, si $j < 0$



- Se soluciona mediante la transformada Z...

$$P(z) = \sum_{n=0}^{\infty} p_n z^n$$

que resulta...

$$P(z) = \frac{1 - A}{\left(1 - \frac{z}{z_1}\right) \left(1 - \frac{z}{z_2}\right) \cdots \left(1 - \frac{z}{z_k}\right)}$$

con $A = \frac{\lambda}{\mu}$ y z_1, \dots, z_k polos del polinomio $D(z)$

$$D(z) = 1 - \frac{A * (z_1 + z_2 + \cdots + z_k)}{k}$$

- Que permite la descomposición parcial...

$$P(z) = (1 - A) \sum_{i=1}^k \frac{B_i}{1 - \frac{z}{z_i}}$$

$$B_i = \prod_{\substack{j=1 \\ j \neq i}}^k \frac{1}{1 - \frac{z}{z_j}} \Bigg|_{z=z_i}$$

- Con la re-transformación al dominio original resulta:

$$P_j = (1 - A) \sum_{i=1}^k \frac{B_i}{z_i} j$$

- Finalmente se suman las probabilidades resultantes de la descomposición del servidor

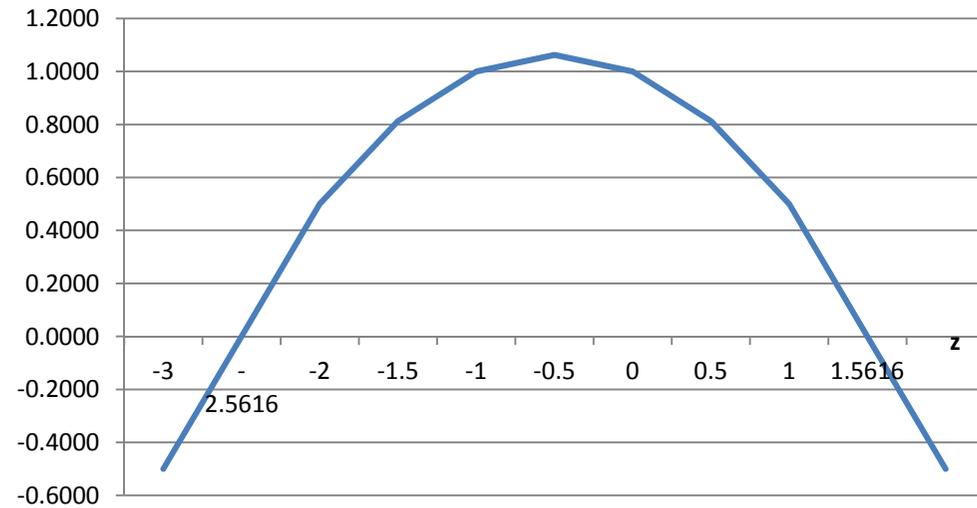
$$p_n = (1 - A) \sum_{j=(n-1)k+1}^k \sum_{i=1}^k B_i \left(\frac{1}{z_i} \right)^j$$

El modelo $M/E_k/1$ (9/10)

Ejemplo (1/2)

Ejemplo $M/E_k/1$			
λ (p/s)	50		
$E(T_s)$ (ms)	10	μ	0.1
A	0.5	k	2
A/k	0.25		

Polinomio $D(z)$

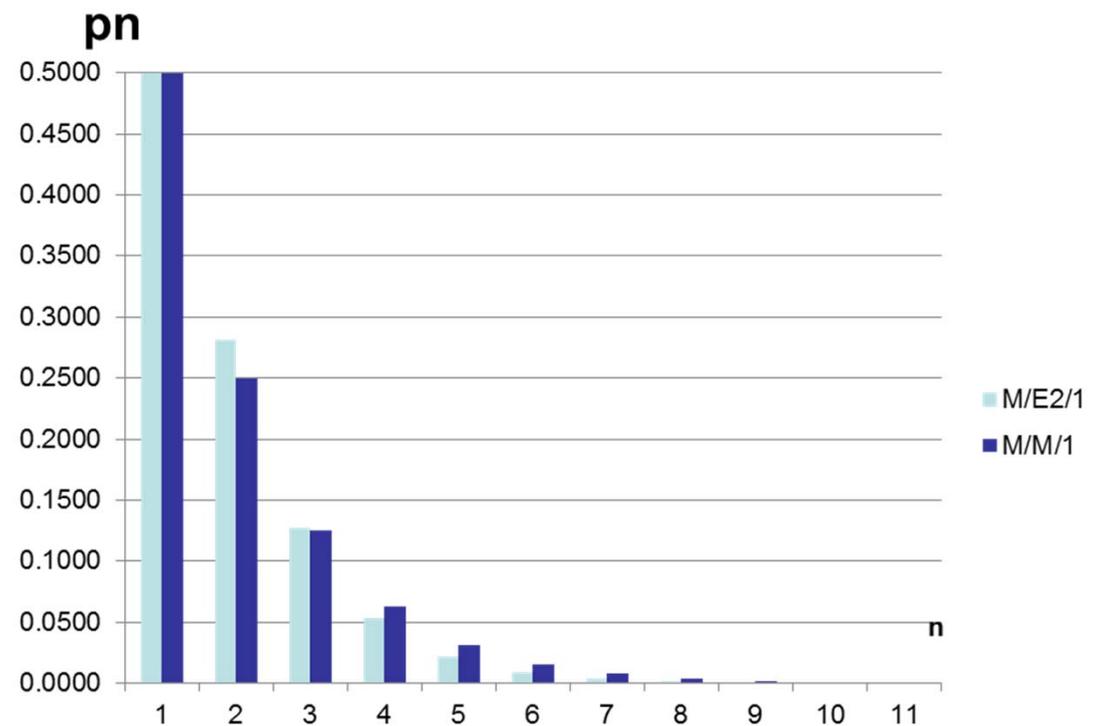


z	-3	-2.5616	-2	-1.5	-1	-0.5	0	0.5	1	1.5616
$D(z)$	-0.5000	0.0000	0.5000	0.8125	1.0000	1.0625	1.0000	0.8125	0.5000	0.0000
B1	0.3787									
B2	0.6213									

El modelo $M/E_k/1$ (10/10)

Ejemplo (2/2)

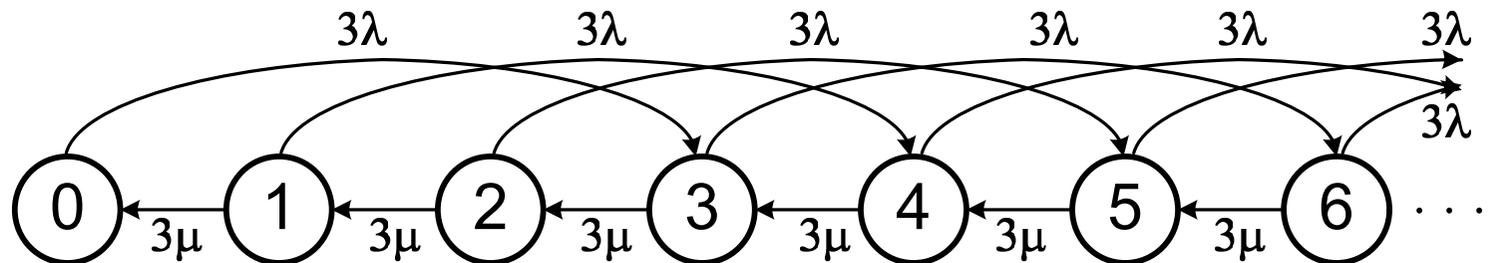
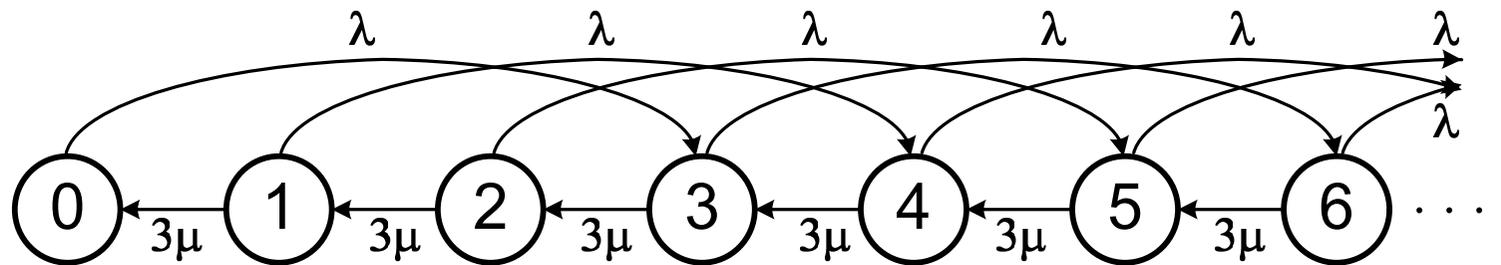
n	p_n M/ E_k /1	p_n M/M/1	Δ [%] con M/M/1
0	0.5000	0.5000	0.00%
1	0.2812	0.2500	12.49%
2	0.1269	0.1250	1.55%
3	0.0538	0.0625	-13.88%
4	0.0223	0.0313	-28.51%
5	0.0092	0.0156	-41.10%
6	0.0038	0.0078	-51.62%
7	0.0016	0.0039	-60.29%
8	0.0006	0.0020	-67.43%
9	0.0003	0.0010	-73.28%
10	0.0001	0.0005	-78.09%



- Una llegada a ráfagas ocurre cuando los paquetes que genera la aplicación son muy largos y se dividen en varios *trozos (chunks)*
 - Por ejemplo un fichero de 10 kBytes puede causar una ráfaga de 10 paquetes, cada una de un 1 kByte
- Aparece un modelo de tipo G/D/1, donde G modela la ráfaga y D la longitud de paquete, típicamente constante
- Se puede aproximar con el modelo de la cadena estrechada visto en el modelo M/Ek/1

- Modelo bulk arrival
 - Las llegadas de los paquetes originales, con tasa λ , siguen una distribución de *Poisson*
 - La tasa de los paquetes (chunks) en los que se dividen es $\xi = k \lambda$, siendo insignificante la distancia temporal entre ellos (*bulk*)
 - Tanto la longitud de los paquetes originales (L), como la de los divididos $L_b = L/k$, da lugar a un tiempo de servicio T_s con distribución exponencial negativa

- Comparación modelo estrechado M/Ek/1 con Bulk/M/1

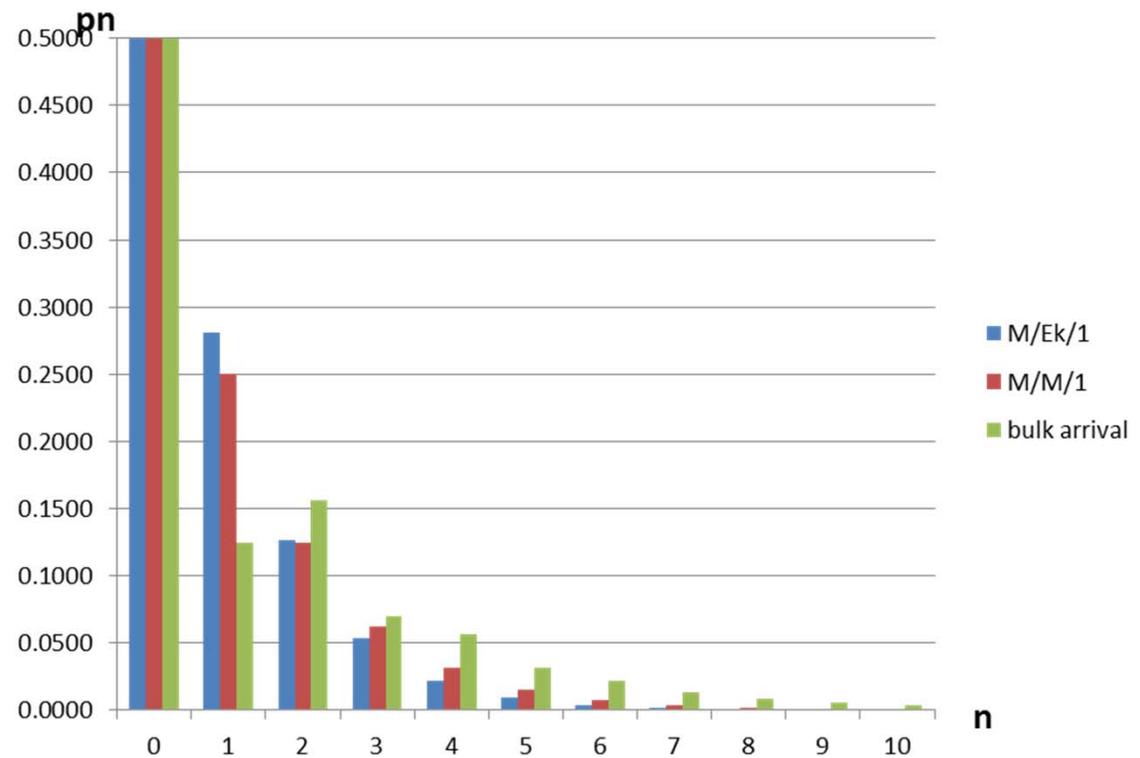


- Se modela con la cadena de Markov estrechada, en la que los estados son los originales, con lo que $P_n = p_n$ y...

$$P_j = (1 - A) \sum_{i=1}^k B_i \left(\frac{1}{z_i} \right)^j \quad p_n = P_j \text{ con } n = j$$

Original	Bulk
λ Poisson	$\lambda_b = k \cdot \lambda$ bulk-k
L	$L_b = L/k$
$t_s = L \cdot 8/v_s$, $\mu = 1/t_s$	$t_{b_s} = t_s/k$, $\mu_b = k \cdot \mu$
$A = \lambda \cdot t_s$	$A_b = \lambda_b \cdot t_{b_s} = A$

p_n M/Ek/1	p_n M/M/1	p_n bulk arrival
0.5000	0.5	0.5000
0.2812	0.2500	0.1250
0.1269	0.1250	0.1562
0.0538	0.0625	0.0703
0.0223	0.0313	0.0566
0.0092	0.0156	0.0317
0.0038	0.0078	0.0221
0.0016	0.0039	0.0135
0.0006	0.0020	0.0089
0.0003	0.0010	0.0056
0.0001	0.0005	0.0036

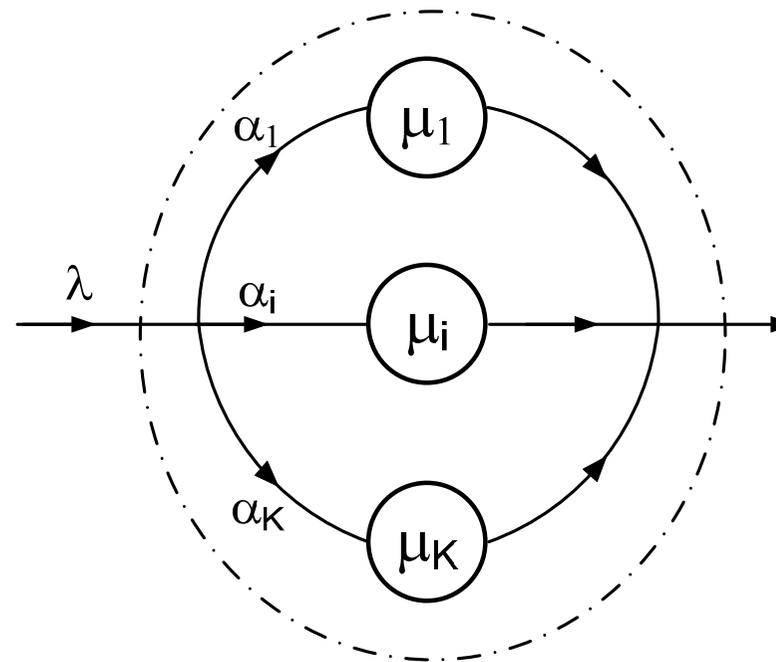


- El modelo M/H_k/1 resulta de una descomposición de un servidor con tasa μ y un tiempo de servicio (T_s) distribuida según una fdp hiper-exponencial-k en k servidores en paralelo con...

$$\cdot \mu = \sum_{k=1 \dots k} \alpha_k \mu_k$$

$$\cdot \sum_{k=1 \dots k} \alpha_k$$

$$\cdot 0 < \alpha_k < 1$$



$$f_{T_S}(t) = \sum_{i=1..k} \alpha_i \mu_i e^{-\mu_i t} \quad \sum_{i=1..k} \alpha_i = 1$$

$$E(T_S) = \sum_{i=1..k} \frac{\alpha_i}{\mu_i}$$

$$\sigma(T_S) = \sqrt{2 \sum_{i=1..k} \left(\frac{\alpha_i}{\mu_i^2} \right) - \left(\sum_{i=1..k} \frac{\alpha_i}{\mu_i} \right)^2}$$

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

- Su solución exacta se realiza a base una cadena de *Markov* expandida, similar a la solución presentada para el sistema M/Ek/1
- Depende de $2k-2$ parámetros, con...

$$\cdot \alpha_i, \mu_i \quad i = 1..k$$

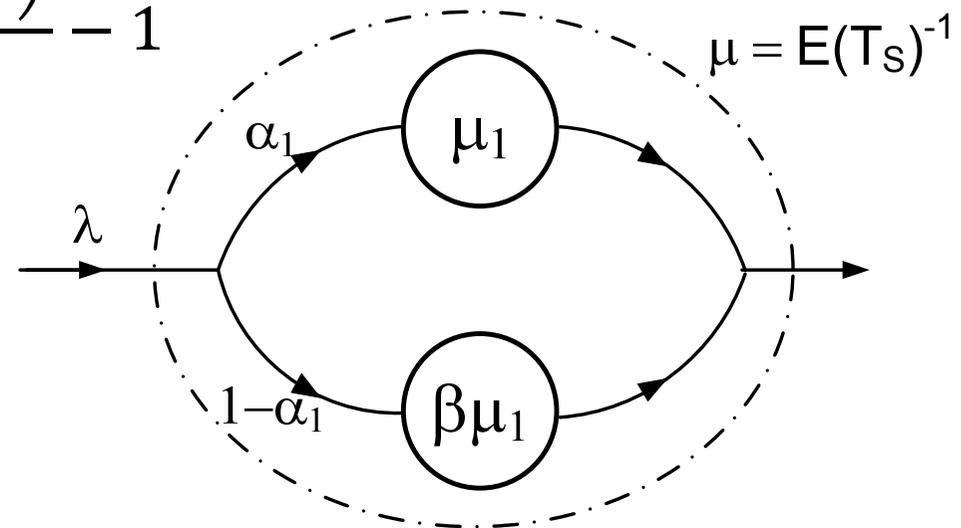
$$\cdot \sum_{i=1..k} \alpha_i = 1, \quad 0 < \alpha_i < 1 \quad E(T_s) = \frac{1}{\mu}$$

- Para el caso particular de $k = 2$ resulta...

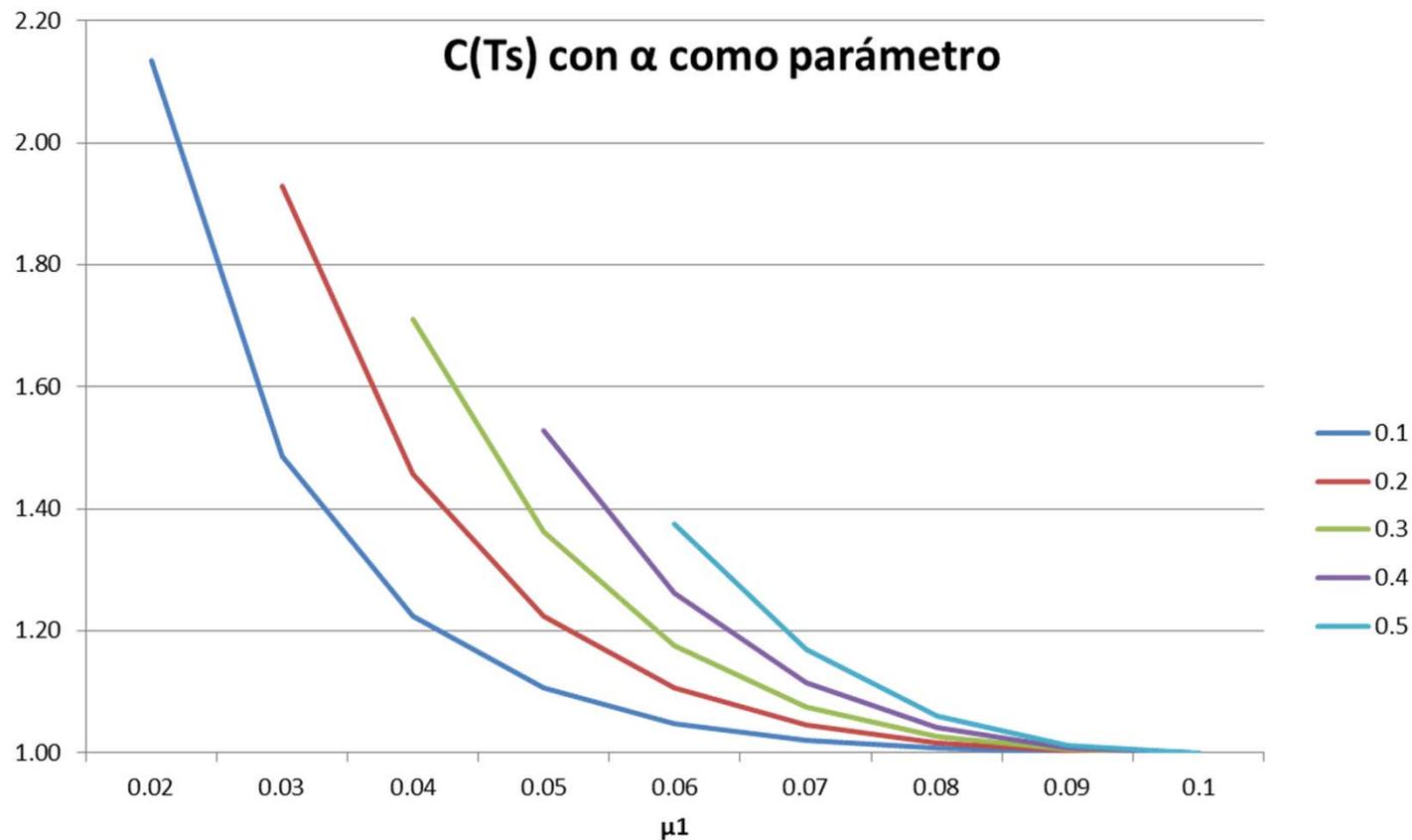
$$\alpha_2 = 1 - \alpha_1 \quad \text{y} \quad \mu_2 = \beta\mu_1$$

$$E(T_s) = \frac{\alpha_1}{\mu_1} + \frac{1 - \alpha_1}{\beta\mu_1}$$

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



- Evolución de $C(T_s)$ en función de α , con $E(T_s)=10$ ms y $k=2$



- Se ha visto que hay sistemas que se pueden solucionar con un proceso estocástico en dos dimensiones $[X_1(t), X_2(t)]$, describiendo la propia cadena y el proceso de servidor
- Aplicando un método de estados finitos para el proceso del servidor, éste se puede modelar con una cadena de *Markov* en dos dimensiones como se ha visto en el ejemplo de la $M/E_k/1$ o $M/H_k/1$
 - En dicho ejemplo se hallaba la solución completa de la fdp de la cadena, pero con una evaluación individual en función de los parámetros A y k en la $M/E_k/1$

- Este método se puede generalizar...
 - Método de variable suplementaria
 - Solución complicada y particular para cada caso; no hay fórmulas genéricas
- Se utiliza el método de la cadena de Markov incluida para obtener expresiones genéricas para los parámetros principales del SdC
 - $E(n)$
 - $E(u)$
 - T_w
 - τ

- El resultado viene dado por las tres ecuaciones de *Pollaczek-Khinchin*, que se expresan como sigue...

$$(PK1) \quad P(z) = (1 - A)(1 - z) \frac{f_{T_s}(s) \Big|_{s=\lambda(1-z)}}{f_{T_s}(s) \Big|_{s=\lambda-z} - z}$$

$$(PK2) \quad E(q) = A + \lambda^2 E(T_s^2) / [2(1 - A)]$$

$$(PK3) \quad f_{T_w}(s) = \frac{s - A}{s - \lambda + \lambda f_{T_s}(s)}$$

con $p_0 = 1 - A$ y $p_w = A$

- A partir de las expresiones anteriores se puede obtener...
 - Las ecuaciones de valores medios espaciales...

$$E(v) = A$$

$$E(n) = \frac{A}{1-A} \left[1 - \frac{A}{2} (1 - C(T_s)^2) \right]$$

$$E(u) = E(n) - E(v)$$

- y, aplicando la fórmula de Little, se obtienen (como en un M/M/1) los valores medios temporales...

$$\tau = \frac{E(n)}{\lambda}$$
$$t_w = \frac{E(u)}{\lambda}$$

- Notar que para el cálculo de los valores medios del SdC se requiere conocer solamente el valor medio y la desviación típica del tiempo de permanencia de los paquetes en el servidor con...

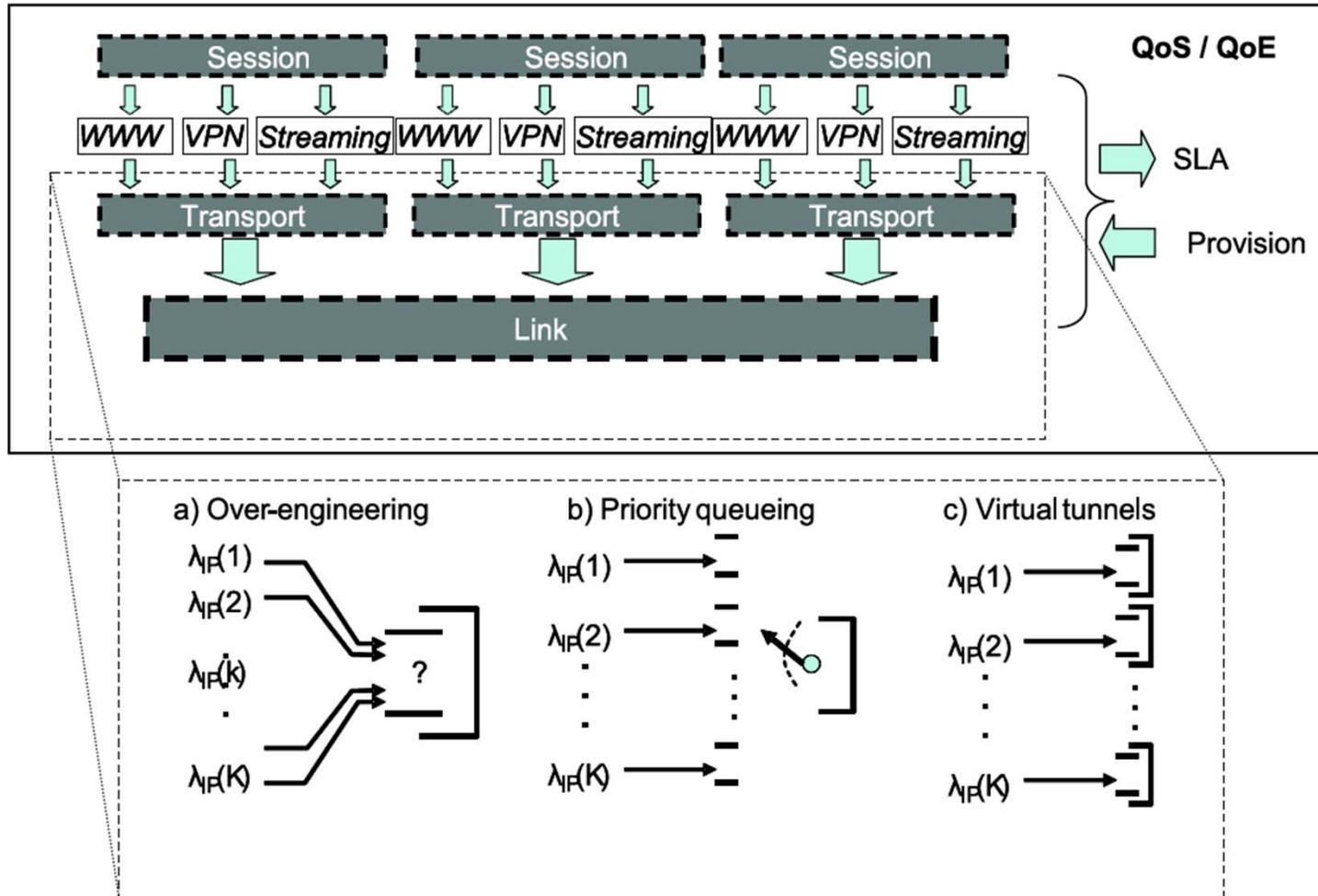
$$t_s = \frac{8 \cdot E(L)}{v_s} \quad y \quad \sigma(t_s) = \frac{8 \cdot \sigma(L)}{v_s} \quad y \quad C(L) = C(T_s)$$

- basta promediar la longitud de paquete durante un intervalo lo suficientemente largo para realizar los cálculos de rendimiento de un SdC M/G/1
- o calcular $E(L)$ y $\sigma(L)$ a partir de las características de servicios ofrecidos

- Ejemplo, con $T_s = 10\text{ms}$ $A=0.5$; $\mu=0.1$
 - Para Erlang-k: $k=2$ $\mu_1=\mu_2 = \mu/2$
 - Para Hk: $k=2$ $\mu_1=0.05$ $\alpha_1=0.3$ $\beta=3.5$

tipo SdC	ρ_0	ρ_w	$C(T_s)$	$E(n)$	τ
M/E2/1	0.5	0.5	0.707	0.875	17.500
M/M/1	0.5	0.5	1.000	1.000	20.000
M/H2/1	0.5	0.5	1.363	1.214	24.286

- En redes con diferentes servicios se deben considerar los parámetros de QoS de manera particular para cada clase de servicio
- Nos limitamos a un SdC M/G/1 y consideramos un único parámetro QoS, el retardo medio
- Se puede hablar (en la actualidad) de cuatro clases de servicios:
 - Real time
 - Streaming
 - Data
 - Best effort
- En ingeniería de tráfico se pueden considerar tres esquemas:
 - **Traffic separation**: Colas y túneles separados para el tráfico de cada clase
 - **Over-Engineering**: Colas y capacidad común, dimensionando a partir del parámetro QoS más restrictivo
 - **Priority Queing**: Colas separadas y capacidades en común



Fuente: Alberto E. Garcia · Laura Rodriguez · Klaus D. Hackbarth "Cost models for QoS-differentiated interconnecting and wholesale access services in future generation networks". Springer Science+Business Media, LLC 2011

- Los SdC con prioridad tienen importancia en la transición del best effort actual hacia el Internet de la siguiente generación (NGI)
- Asumimos que los paquetes contienen una marca que indica su prioridad.
 - En IPv4 existe un campo denominado TOS que se usa para implementar el servicio con prioridad DiffServ
- Cuando se integran voz y datos, los paquetes de voz deben cumplir unos restricciones de retardo que pueden conseguirse mediante SdC con prioridad
- Una alternativa sería dividir el ancho de banda total en dos partes y reservar para cada tipo de servicio su capacidad propia (con su cola correspondiente)

- En el caso más sencillo se distinguen dos prioridades
 - E.g. paquetes para servicios de empresas y paquetes para servicio de Internet de clientes residenciales o paquetes de voz y paquetes de datos
- Cada una de las dos corrientes de paquetes tiene asociada un conjunto de variables (índice $k=1$ indica paquetes de prioridad e índice $k=2$ los que no tienen prioridad)
 - λ_k tasa de llegada de paquetes, según un proceso de Poisson
 - L_k longitud de paquetes, con $E(L_k)$ valor medio y $\sigma(L_k)$ desviación típica (ambos en octetos)
- La velocidad del servidor es v_s [kbps] y se asume colas de longitud infinita para ambos sistemas

- En el modelo de prioridad se sirven primero los paquetes del servicio con prioridad, mientras que los paquetes del otro servicio se sirven solamente cuando la cola con prioridad esté vacía
- Se asume además que un paquete de prioridad baja que se encuentra en el servidor termine su servicio, aunque llegue un paquete con prioridad alta
- Este aspecto se denomina en inglés “*non-preemptive priority queuing*” (NPPQ)
- La alternativa (si hubiera un paquete de no-prioridad en el servidor al llegar un paquete de prioridad aquel se devuelve a su cola) se denomina PPQ y no se considera (aplicación limitada)

L byte / BW kbps	64	128	384	2048
256	32.000	16.000	5.333	1.000
512	64.000	32.000	10.667	2.000
768	96.000	48.000	16.000	3.000
1024	128.000	64.000	21.333	4.000

- El valor medio del tiempo de espera de cada servicio se calcula (modelo NPPQ) como

$$t_w(1) = \frac{\frac{1}{2} (\lambda_1 E_2(T_s(1)) + \lambda_2 E_2(T_s(2)))}{1 - A_1}$$

$$t_w(2) = \frac{\frac{1}{2} (\lambda_1 E_2(T_s(1)) + \lambda_2 E_2(T_s(2)))}{(1 - A_1)(1 - A_1 - A_2)}$$

con $E_2[T_s(k)]$ segundo momento de $T_s(k)$ en el origen, que se calcula como sigue...

$$E_2(T_s(k)) = V(T_s(k)) + (E(T_s(k)))^2$$

- El retardo completo resulta

$$\tau_k = t_w(k) + t_s(k) \quad k = 1, 2$$

- La fórmula se puede generalizar para $k = 1, \dots, K$

$$t_w(k) = \frac{1}{2} \sum_{k=1}^K \frac{\lambda_k E_2[T_s(k)]}{(1 - \sum_{j=1}^{k-1} A_j)(1 - \sum_{j=1}^k A_j)}$$

SdC con prioridad (6/14)

Modelo NPPQ (3/11) – Ejemplo (1/9)

- Una empresa conecta su infraestructura de telecomunicación (Voz y Datos) con un IUR de 2048 kbps a la red IP de un operador nacional
- Por el IUR se transmiten dos tipos de paquetes, cuyas características se indican en la tabla
- Se pretende que la ocupación por el tráfico total de ambos servicios en el IUR no supera un valor de $A_t = 0.85$, pero éste se distribuye de forma diferente dependiendo de la hora
- Además se asume que el tiempo de interllegada sigue una fdp exponencial negativa

Parámetro	K=1	K=2
Servicio	VoIP	Datos
E(L) (octetos)	53	512
$\sigma(L)$ (octetos)	0	1024

- Para un tráfico del servicio de voz $A_1 = \alpha \cdot A_t$ con $\alpha=0.05$ y un esquema NPPQ, se calcula (para cada servicio)
 - La tasa de llegada de paquetes λ_k [p/s],
 - La duración media $t_s(k)$ de la transmisión de un paquete por el IUR, y su desviación típica $\sigma_s(k)$.
 - El tráfico de cada servicio, a partir de los resultados λ_k y $t_s(k)$

Parámetro	K=1	K=2
λ	205.2830	403.7500
t_s (ms)	0.2070	2.0000
$\sigma(t_s)$	0.0000	4.0000
Tráfico A	0.0425	0.8075

SdC con prioridad (8/14)

Modelo NPPQ (5/11) – Ejemplo (3/9)

- Se asume que el gestor de la red de la empresa divide la capacidad de IUR en dos partes (túneles virtuales), reservando cada una para el uso exclusivo de un servicio
 - La división se realiza en relación con el tráfico de cada servicio.
 - Se calculan los valores medios espaciales (n , u , v) y temporales (t_w , τ) para cada servicio

Parámetro	K=1	K=2
v_s	102.4000	1945.6000
t_s (ms)	4.1406	2.1053
$\sigma(ts)$	0.0000	4.2105
Tráfico A	0.8500	0.8500
n	3.2583	12.8917
u	2.4083	12.0417
v	0.8500	0.8500
t_w (ms)	11.7318	29.8246
τ (ms)	15.8724	31.9298

- Ahora el gestor comparte el ancho de banda del IUR entre ambos servicios, e implementa un esquema de dos colas separadas (NPPQ)
- Se calcula el t_w y τ para cada servicio, y se comparan los resultados de las variables temporales con los resultados del modelo anterior (túneles virtuales)

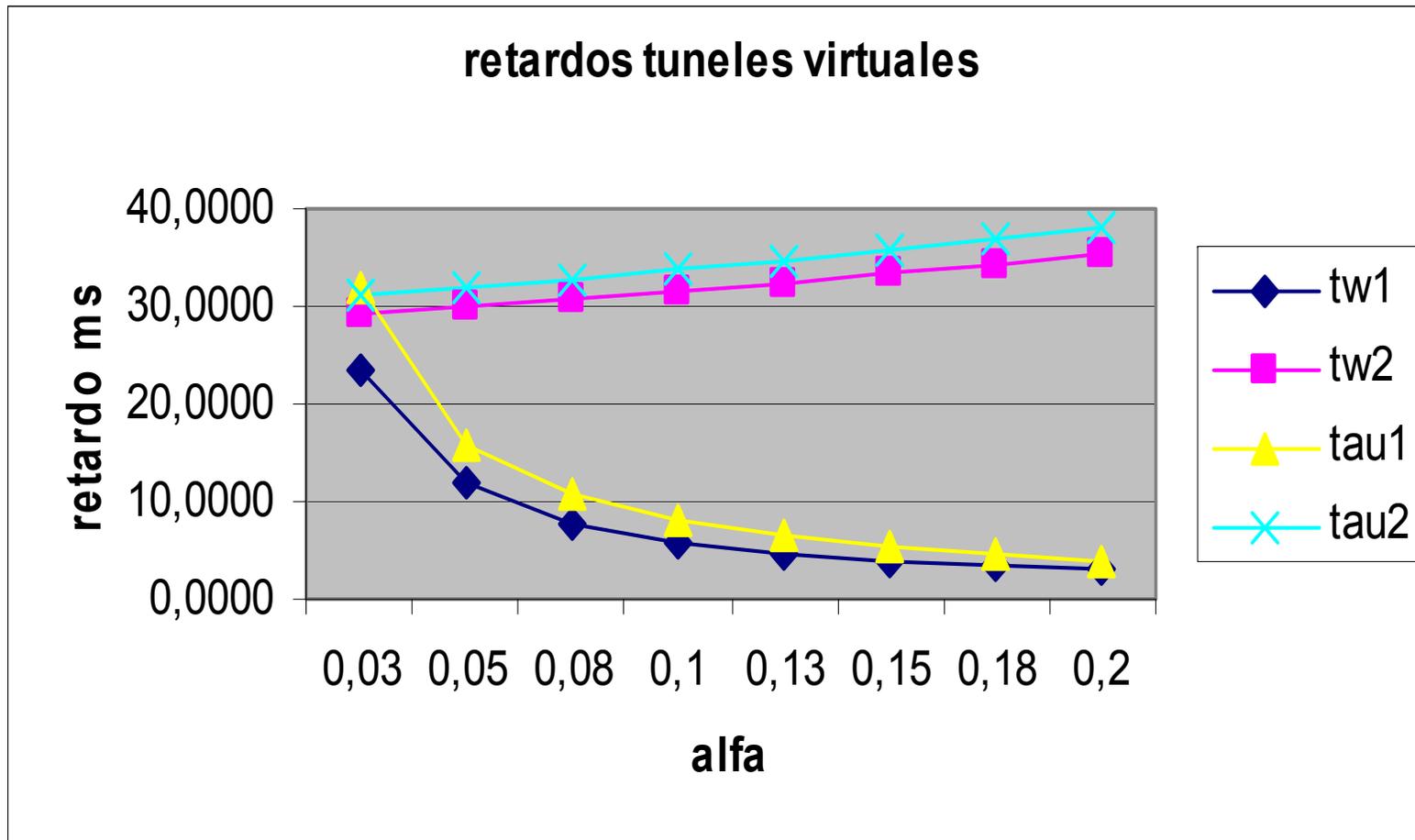
Numerador Común	4.4E-06	0.0040	0.0040
Denominador	0.9575	0.1436	
t_w (ms)	4.221	28.142	
τ (ms)	4.428	30.142	

- Ahora se asume que la parte del tráfico del servicio con prioridad ($k=1$) crece, pero el tráfico del otro servicio se reduce proporcionalmente, de manera que el tráfico total mantiene su valor de 0.85
 - Se realiza una variación del parámetro α indicado inicialmente desde 0.025 hasta 0.2, con pasos de 0.025
 - Los resultados $[t_w(k)](\alpha)$ y $[t(k)](\alpha)$ se representan gráficamente, tanto para el modelo de túneles virtuales como para el NPPQ

Tunnel

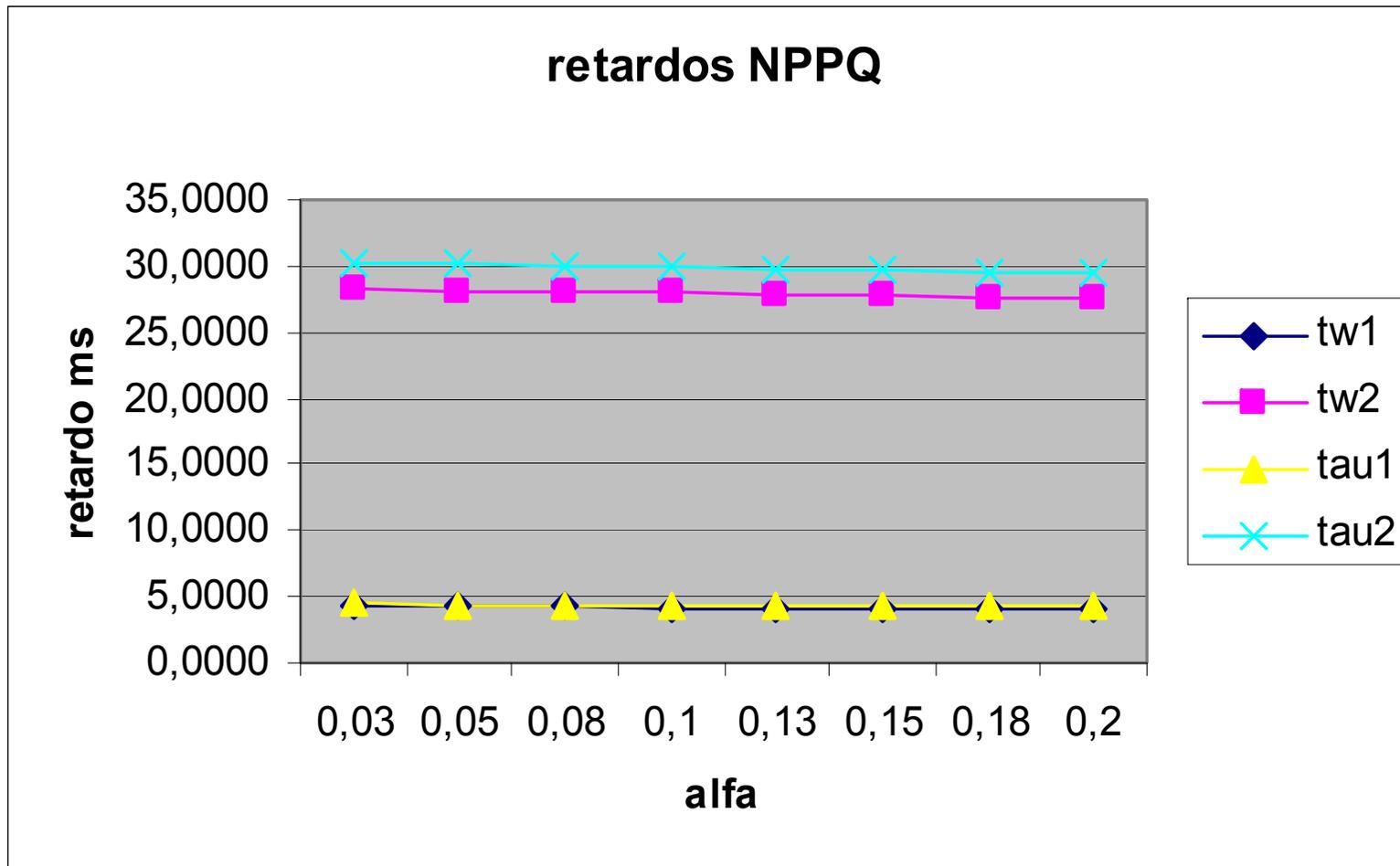
NPPQ

α	t_w (1)	t_w (2)	τ (1)	τ (2)	t_w (1)	t_w (2)	τ (1)	τ (2)
0.025	23.4635	29.0598	31.7448	31.1111	4.2360	28.2398	4.4430	30.2398
0.050	11.7318	29.8246	15.8724	31.9298	4.2213	28.1420	4.4283	30.1420
0.075	7.8212	30.6306	10.5816	32.7928	4.2060	28.0399	4.4130	30.0399
0.100	5.8659	31.4815	7.9362	33.7037	4.1899	27.9330	4.3970	29.9330
0.125	4.6927	32.3810	6.3490	34.6667	4.1731	27.8210	4.3802	29.8210
0.150	3.9106	33.3333	5.2908	35.6863	4.1555	27.7035	4.3626	29.7035
0.175	3.3519	34.3434	4.5350	36.7677	4.1370	27.5802	4.3441	29.5802
0.200	2.9329	35.4167	3.9681	37.9167	4.1176	27.4506	4.3246	29.4506



SdC con prioridad (13/14)

Modelo NPPQ (10/11) – Ejemplo (8/9)



- Finalmente se calculan los retardos relativos $[t_w(k)]^{NPPQ}/[t_w(k)]^{VT}$ y $[\tau(k)]^{NPPQ}/[\tau(k)]^{VT}$, en función del parámetro α y se representan los resultados gráficamente

