Qüestiió – V. 7, n.º 3 (setembre 1983) pp. 501-510

ANÁLISIS DEL COMPORTAMIENTO DE UN FRAGMENTADOR CONCENTRADOR DE PAQUETES, MODELO DE CUANTO NULO

J. VINYES SANZ, J. RIERA GARCÍA

UNIVERSIDAD POLITÉCNICA DE MADRID

Se analiza el tiempo de respuesta de un sistema de colas con disciplina "Round-Robin" y dos clases de prioridad. Se utiliza la aproximación de cuanto nulo. Los resultados se comparan cuantitativamente con los obtenidos con un modelo de cuanto finito.

A PACKET CONCENTRATOR WITH MESSAGE FRAGMENTATION, NULL QUANTUM ANALYSIS

Keywords: PACKET SWITCHED NETWORKS, MESSAGE CONCENTRATOR, QUEUES, ROUND-ROBIN, PRIORITIES.

1. INTRODUCCION.

En /16/ se analiza una disciplina de fragmentación de mensajes y concentración de los paquetes resultantes, y se obtiene la espera - media condicionada a la longitud del mensaje.

Con objeto de obtener una expresión más compacta que la obtenida en /16/ estudiamos a - continuación el caso límite en que la longitud del cuanto de servicio de los usuarios-2, tiende a cero. El modelo a analizar coincide con el descrito de cuanto finito, salvo que la longitud máxima del cuanto, x_{pm}, tiende a cero.

Utilizamos un método basado en el análisis - de Kleinrock /12/ de un sistema M/G/1 con -- servicio cíclico y cuanto nulo, denominado - procesador compartido (Processor Sharing).

Se define:

- $n_2(x)$ = densidad media de usuarios-2 (que aún no han abandonado el sistema) que ya han recibido x segundos de servicio, i.e. ----- $\int_{x_a}^{x_b} n(x)\,dx$ = número medio de usuarios que han recibido servicio comprendido entre x_a y x_b segundos y que siguen presentes en el sistema.

- $T_2(x)$ = tiempo medio de tránsito (cola+servicio, $T_2(x) = W_2(x)+x$) para un usuario-2 que requiere x segundos de servicio.

Estas magnitudes satisfacen la relación:

$$\frac{d\,T_2(x)}{d\,x} = \frac{n_2(0)}{X_2} \qquad (1)$$

desarrollada /12/ para disciplinas cíclicas - con cuanto nulo y una sola cláse de usuarios, y que es directamente generalizable a varias clases de usuarios. E integrando entre 0 y x:

$$T_2(x) - T_2(0) = \frac{n_2(0)}{X_2}\,x \qquad (2)$$

- J. Vinyes Sanz - J. Riera García
Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid
Ciudad Universitaria - Madrid

- Article rebut el Setembre del 1983.

2. CALCULO DEL TIEMPO MEDIO DE TRANSITO DE LOS USUARIOS-2 CUYO TIEMPO DE SERVICIO TIENDE A CERO, $T_2(0)$.

Utilizamos la técnica del usuario marcado, para obtener el retardo $T_2(0)$, que está compuesto por: $p_1.\bar{x}_1$, el retardo medio causado por un potencial usuario-1 sirviéndose en el instante de llegada del marcado; $p_2.r_p$, el retardo causado por un potencial usuario-2 presente en el servidor en el instante de llegada, esta contribución tiende a cero, por tender a cero el cuanto de servicio; $Q_1.\bar{x}_1$, la espera causada por los usuarios-1, presentes en cola en el instante de llegada; $X_1.T_2(0).\bar{x}_1$, el retardo causado por los $X_1.T_2(0)$ usuarios que por término medio entran en el sistema durante la espera y servicio, $T_2(0)$, del marcado, $Q_2.\bar{x}_p$, el retardo causado por los Q_2, usuarios-2 en cola, que tiende a cero, y por ello además, no existen más componentes de retardo, pues sólo se da un cuanto de servicio.

Sumando estas contribuciones obtenemos:

$$T_2(0) = p_1 \bar{x}_1 + Q_1 \bar{x}_1 + p_1 T_2(0) \qquad (3)$$

Aplicando la relación de Little, $Q_1 = X_1.W_1$, y tomando el límite, cuando el cuanto tiende a cero (y por tanto $r_p \to 0$) de la expresión de W_1, (8) en /16/, obtenemos:

$$Q_1 \bar{x}_1 = X_1 W_1 \bar{x}_1 = \frac{p_1^2 \bar{x}_1}{1 - p_1} \qquad (4)$$

Y substituyendo en (3), la expresión (4) de $Q_1.\bar{x}_1$, se tiene el retardo buscado, en función de los parámetros de tráfico, p_1 y $\bar{x}_1$:

$$T_2(0) = \frac{1}{(1-p_1)2} \bar{x}_1 \qquad (5)$$

3. CALCULO DE LA DENSIDAD MEDIA $n_2(0)$.

Substituyendo en (2), $T_2(0)$ por su expresión en (5), obtenemos:

$$T_2(x) = \frac{n_2(0)}{X_2} x + \frac{p_1}{(1-p_1)^2} \bar{x}_1 \qquad (6)$$

Por tanto para expresar $T_2(x)$, en función de los parámetros del tráfico, debemos determinar $n_2(0)$. Si somos capaces de obtener una expresión que describa el comportamiento asintótico ($x \to \infty$) de $T_2(x)$, podremos determinar la densidad $n_2(0)$, a partir de la expresión (6) sin más que hacer tender hacia infinito al tiempo de servicio del usuario marcado. En este argumento se basa Kleinrock para su análisis del procesador compartido. Pero en vez de calcular directamente la expresión del comportamiento asintótico de T(x) (Sin subíndice, pues sólo considera una clase de usuarios), propone un sistema cuyo comportamiento asintótico sea idéntico, y tal que ya esté estudiado. En concreto, argumenta que en un procesador compartido un usuario cuyo tiempo de servicio tiende a infinito, recibe el mismo tratamiento que un usuario de la clase no prioritaria, cuya longitud media tiende a infinito, en una cola con dos clases de prioridad y disciplina expulsora y conservadora. Este razonamiento se basa en que cualquier usuario que llegue después del marcado o del usuario no prioritario, respectivamente, recibe en ambos sistemas servicio antes que ellos.

Sin embargo consideramos que tal razonamiento es incompleto, pues existe una pequeña diferencia entre ambos sistemas: en el procesador compartido, el usuario marcado recibe servicio aunque haya otros usuarios presentes, lo contrario ocurre en el sistema equivalente propuesto. Donde el usuario prioritario no recibe servicio mientras haya otro usuario prioritario. Por tanto, igualar ambos comportamientos asintóticos, constituye un planteamiento pesimista.

En consecuencia abordamos el análisis del comportamiento asintótico de nuestro modelo, completando el razonamiento de Kleinrock, proponiendo: 1º. Un modelo pesimista, similar al propuesto por Kleinrock, y 2ºUn modelo optimista, con unos retardos asintóticos T_p y T_o, respectivamente, que acotan al de nuestro modelo, es decir $T_o < T_2(x) < T_p$. Abordamos en primer lugar el análisis del modelo optimista, a continuación el del pesimista, y comprobamos que el comportamiento es idéntico para ambos y por tanto coincide con el de nuestro modelo.

Para ambos análisis utilizamos un sistema con dos clases de usuarios a y b. Los usuarios-a son tales que su longitud media tiende a infinito, que su tasa de llegadas tiende a cero y que su factor de utilización p_a es despreciable, lo que está en correspondencia con un va

lor asintótico cero, para la probabilidad de los usuarios-2 de longitud media infinita. - Los usuarios-b, agrupan a los usuarios-1 y - los usuarios-2 con longitud finita. Se verifica, $p = p_1 + p_2 = p_a + p_b \simeq p_b$.

a) MODELO OPTIMISTA.

La disciplina de servicio es tal que el servidor da un cuanto de servicio alternativamente al usuario que ocupa el primer lugar en la cola de cada clase. Dado que el factor de utilización tiende a cero, es despreciable la probabilidad de que un usuario-a coincida en el sistema con otro de su clase. En este caso es evidente que el usuario-a recibe como poco la mitad de capacidad del servidor, pues sólo lo comparte con el primero de la cola de usuarios-b. El comportamiento recibido es mucho - mejor que el de un usuario marcado de nuestro modelo de partida, en donde (si los usuarios-1, lo permiten), lo comparte cíclicamente con el resto de usuarios-2.

El retardo medio, T_{ao}, experimentado por un - usuario-a está compuesto por: $p_b.r_b$, el retardo causado por el usuario-b atendido, eventualmente, en el instante de llegada, no consideramos un término $p_a.r_a$ pues hemos supuesto que no puede haber más de un usuario en el sistema; $Q_b.\bar{x}_b$, el tiempo de servicio de los usuarios presentes en cola, que abandonan el sistema antes que el usuario-a; $(X_b.T_{ao}.\bar{x}_b)-e$, el tiempo de servicio de los $X_b.T_{ao}$ usuarios-b que por término medio llegan durante el --- tiempo de tránsito T_{ao}, donde e es la fracción de servicio que eventualmente se recibe después que el usuario-a abandone el sistema, y su propio tiempo de servicio $\bar{x}_a$. La suma - de estas contribuciones es:

$$T_{ao}(1 - p_b) = p_b\, r_b + Q_b \,.\, \bar{x}_b - e + \bar{x}_a \qquad (7)$$

Y tomando el límite cuando $\bar{x}_a$ tiende a infinito obtenemos finalmente el retardo del modelo optimista, T_o, cota inferior del retardo $T_2(oo)$.

$$\lim_{x\to oo} T_2(x) > T_o = \lim_{\bar{x}_a = x\to oo} T_{ao} = \frac{x}{1-p_b} = \frac{x}{1-p} \qquad (8)$$

b) MODELO PESIMISTA.

Como ya hemos citado se trata de un sistema con dos clases de prioridad con disciplina - expulsora y conservadora. La clase-a es la - menos prioritaria y por tanto en cuanto haya algún usuario-b se expulsa del servidor al - eventual usuario-a presente, que conserva el servicio recibido hasta ese momento.

El tiempo medio de espera, T_{ap}, para un usuario-a, no prioritario, está dado /12/ por:

$$T_{ap} = \frac{\bar{x}_a(1-(p_a+p_b))+W_o}{(1-(p_a+p_b))(1-p_b)} \qquad (W_o = p_a r_a + p_b r_b) \qquad (9)$$

Tomando el límite de T_{ap} cuando $\bar{x}_a$ tiende a infinito, se tiene la cota superior del retardo $T_2(x)$:

$$\lim_{x\to oo} T_2(x) < T_p = \lim_{\bar{x}_a = x\to oo} T_{ap} = \frac{x}{1-p_b} = \frac{x}{1-p} \qquad (10)$$

En consecuencia, por coincidir las cotas superior e inferior del comportamiento asintótico del retardo condicionado $T_2(x)$, de un - usuario-2, podemos afirmar que:
$\lim_{x\to oo} T_2(x) = x/1-p$. Tomando el límite, - cuando x tiende a infinito, de $T_2(x)$, en la expresión (6), obtenemos:
$\lim_{x\to oo} T_2(x) = n_2(o).x/X_2$. Y ya podemos despejar el valor de la densidad, $n_2(o)$, entre estas dos últimas expresiones, $n_2(o) = X_2/1-p$, que introducido en (6), nos proporciona la - expresión de $T_2(x)$, el retardo medio de tránsito de los usuarios-2 condicionado por el - tiempo de servicio:

$$T_2(x) = \frac{x}{1-p} + \frac{p_1}{(1-p_1)^2} \,.\, \bar{x}_1 \qquad (11)$$

Y la espera media de los usuarios-2, condicionada a su tiempo de servicio es:

$$W_2(x) = T_2(x) - x = \frac{p}{1-p} \,.x + \frac{p_1}{(1-p_1)^2} \,.\bar{x}_1 \qquad (12)$$

Sea $W_e = \bar{x}_1.p1/1-p1$, el tiempo medio de espera, en un sistema M/M/1, sin prioridades, con tasa de llegadas x_1 y tiempo de servicio x_1, y factor de utilización $p_1 = x_1.x_1$. Ahora podemos poner la anterior expresión de $W_2(x)$ como

$$W_2(x) = \frac{p}{1-p} \cdot x + \frac{W_e}{1-p_1} \qquad (13)$$

Se observa que la espera condicionada $W_2(x)$, se compone de: un término independiente del tiempo de servicio x, que es una amplificación de la espera We, por el factor $1/1 - p_1$; y de un término proporcional a x, dependiendo el factor de proporcionalidad del factor de utilización total, p.

La espera media relativa es:

$$\frac{W_2(x)}{x} = \frac{p}{1-p} + \frac{p_1}{(1-p_1)^2} \cdot \frac{\bar{x}_1}{x} \qquad (14)$$

que tiende al cociente p/1-p, a medida que crece el tiempo de servicio x, respecto de, $\bar{x}_1$, el tiempo medio de servicio de los usuarios-1. (Fig. 1).

Simplificando el comportamiento de $W_2(x)/x$, la espera media relativa, consideramos que existen dos zonas: a) zona de coste variable, donde predomina el segundo término de (14), y b) zona de coste fijo, donde predomina el primer término de (14).

En particular, en el fragmentador concentrador de paquetes que ha motivado el análisis, sabemos que los mensajes interactivos (clase-1) requieren un servicio medio, $\bar{x}_1$, mucho menor que, $\bar{x}_2$, el de los mensajes masivos (clase-2). Esta característica del tráfico de datos favorece que la zona de coste variable se concentre cerca del origen, puesto que el factor, $\bar{x}_1/x$, es mayor que uno, solo para un servicio requerido, x menor que $\bar{x}_1$, y por tanto mucho menor que $\bar{x}_2$.

Efectuamos a continuación un análisis cuantitativo para ilustrar el comportamiento de $W_2(x)/x$, la espera media relativa. Elegimos como punto de transición, entre la zona de coste variable y la zona de coste fijo, al servicio requerido, x_t, tal que $W_2(x)/x$ sea superior en un diez por ciento al valor asintótico, p/1-p. Introduciendo esta condición en (14):

$$\frac{x_t}{\bar{x}_1} = 10 \cdot \frac{p_1/(1-p_1)^2}{p/(1-p)} = 10\,\frac{p_{r1}(1-p)}{(1-p_{r1}p)^2} \qquad (15)$$

donde la utilización relativa relativa de los usuarios-1, $p_{r1} = p_1/p$, es la relación entre la utilización de los usuarios-1, p_1 y la utilización total, p. Para una relación p_{r1}, dada, el factor $(1-p)/(1-p.p_{r1})$ presenta un máximo en p max=$2-1/p_{r1}$. Se observa (Fig. 1c) que cuanto menor es p_{r1}, menor es la utilización, p, que hace máximo a dicho factor, y que para, $p_{r1}<1/2$, el máximo se produce para p<0. Por tanto, dado que nuestro interés se limita al intervalo $(0 \leq p < 1)$, se tiene:

$$p_{max} = \begin{cases} = 0 & \text{si } p_{r1} \leq 1/2 \\ = 2 - \dfrac{1}{p_{r1}} & \text{si } p_{r1} > 1/2 \end{cases} \qquad (16)$$

En consecuencia el valor máximo del punto de tránsito x_t, viene dado por

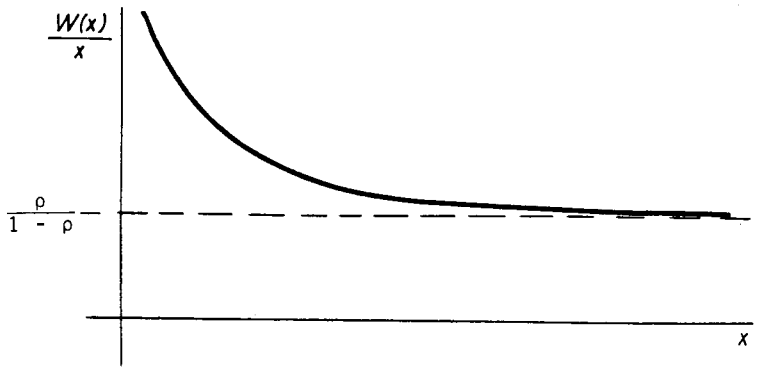
$$\frac{x_t}{\bar{x}_1} = \begin{cases} 10p_{r1} & \text{si } p_{r1} \leq 1/2 \\ \dfrac{2.5}{1-p_{r1}} & \text{si } p_{r1} > 1/2 \end{cases} \qquad (p_{r1} = p_1/p) \qquad (17)$$

En la tabla se muestran los valores máximos del punto de tránsito x_t, correspondientes a diversos valores de p_{r1}, la utilización relativa de los usuarios-1.

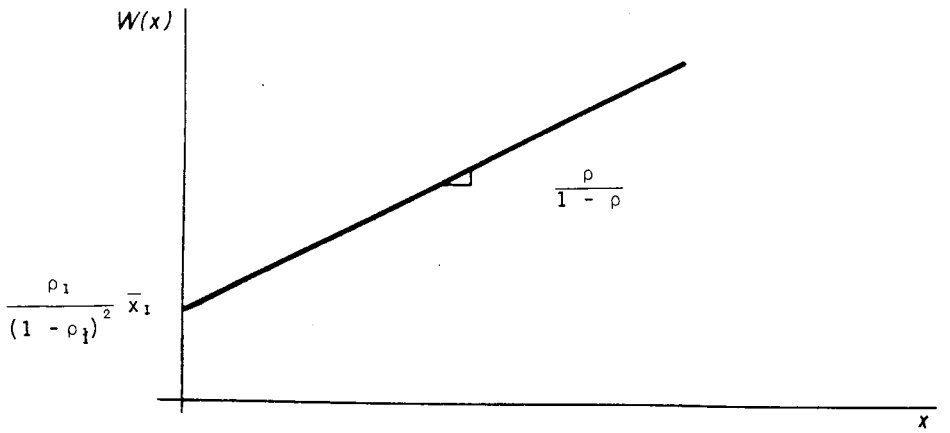
$p_{r1} = \frac{p_1}{p}$	$\frac{x_t}{\bar{x}_{1\,max}}$
0.1	1
0.3	2
0.5	5
0.7	8.5
0.8	12.5
0.9	25

Vemos que la zona asintótica se alcanza:rapidamente cuando la utilización relativa de los usuarios-1 es baja, para cinco servicios medios de los usuarios-1, si la utilización es igual para ambas clases de usuarios y para servicios un orden de magnitud superiores a $\bar{x}_1$, si la utilización p_1 es alta.

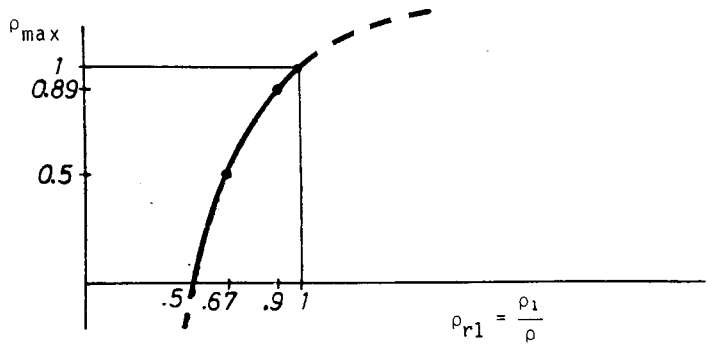
Recordamos que los valores expuestos corresponden al caso peor de una utilización total p=p max. Así por ejemplo si p = 0.8, $p_1 = p_2 = 0.4$, el punto de tránsito es: $x_t/\bar{x}_1 = 2{,}78$, que es casi la mitad del valor máximo correspondiente a $p_{r1} = 0.5$.



a) Espera condicionada relativa



b) Espera condicionada



c) Utilizacion global para x_t máximo

Fig. 1. Espera condicionada

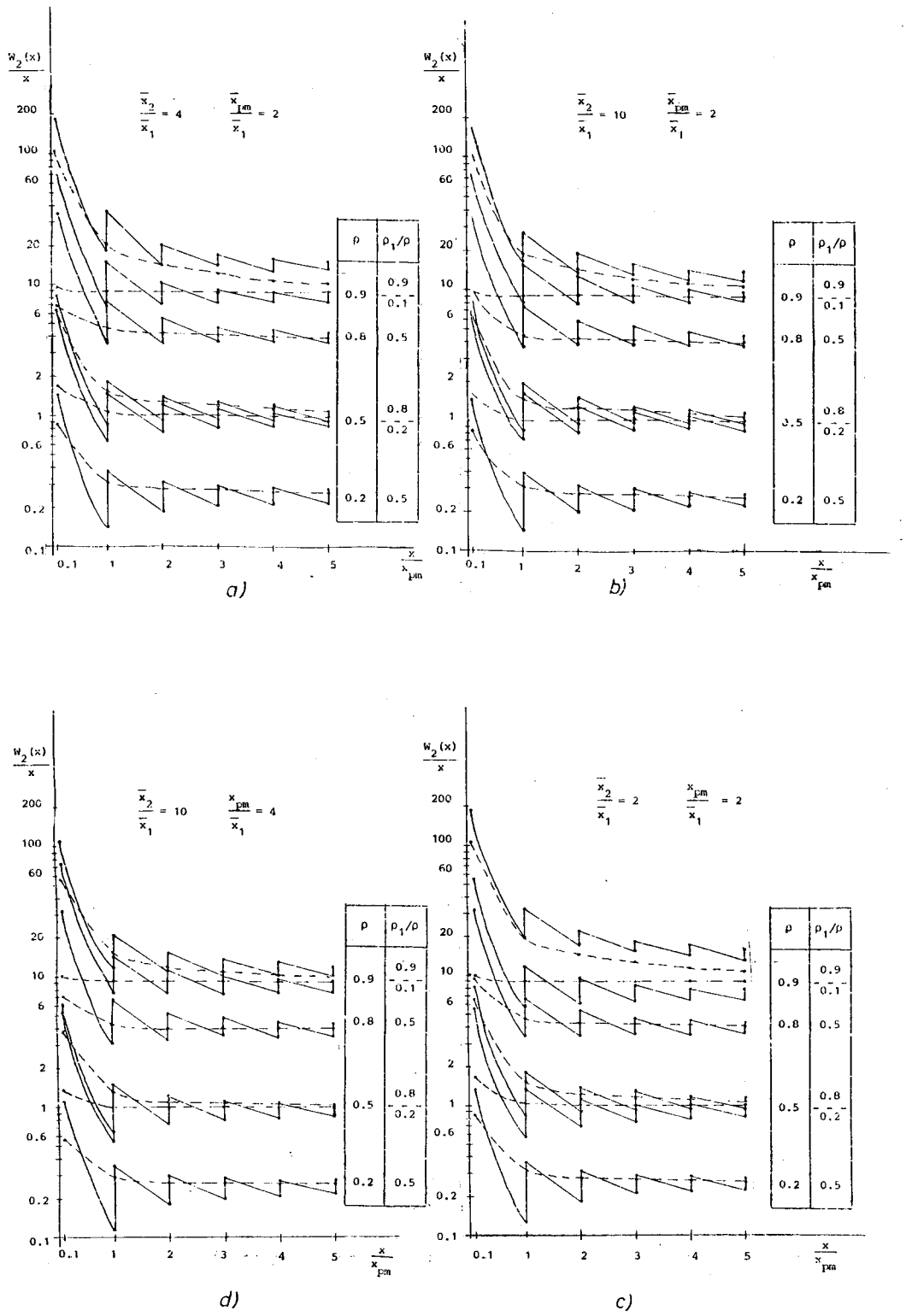
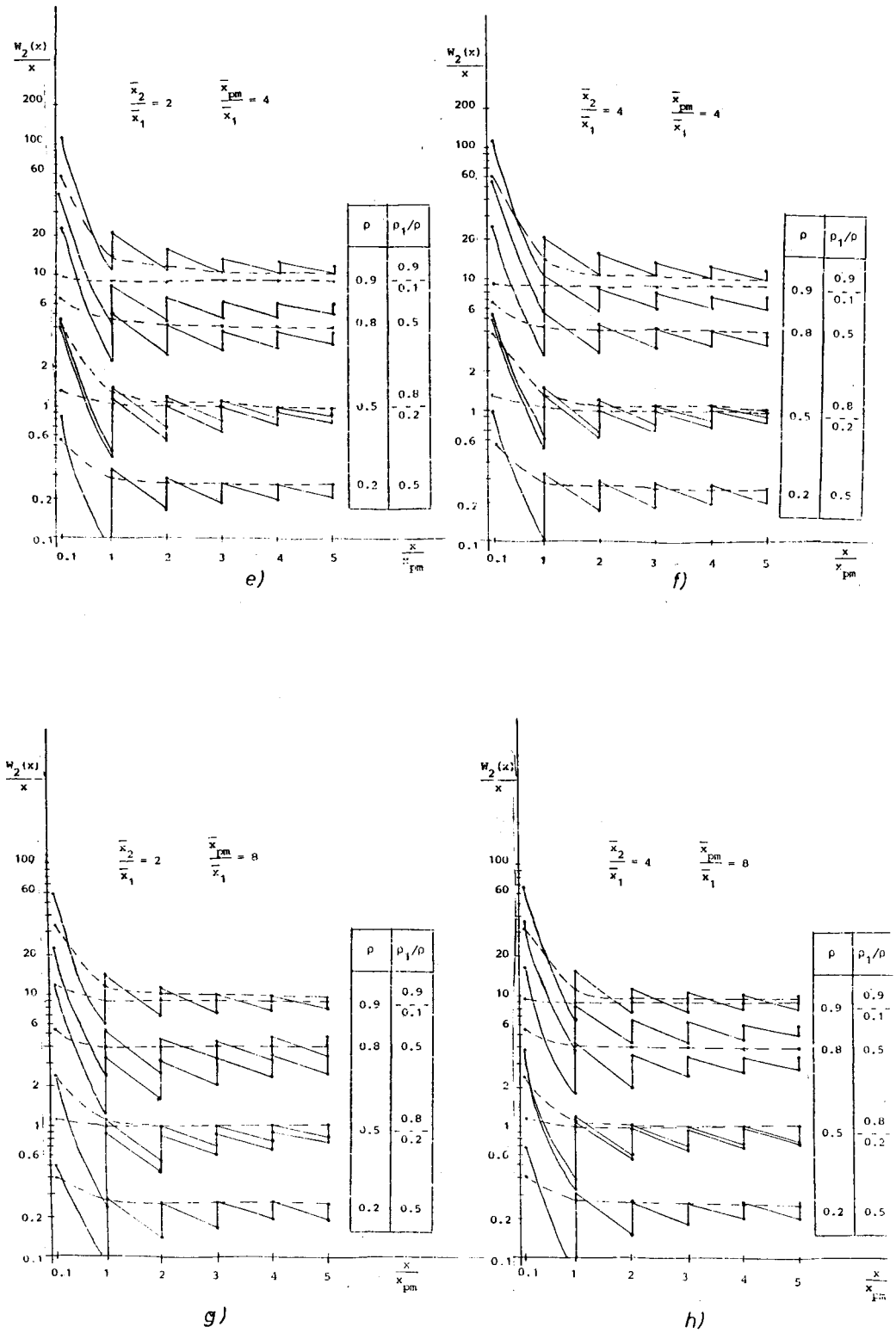
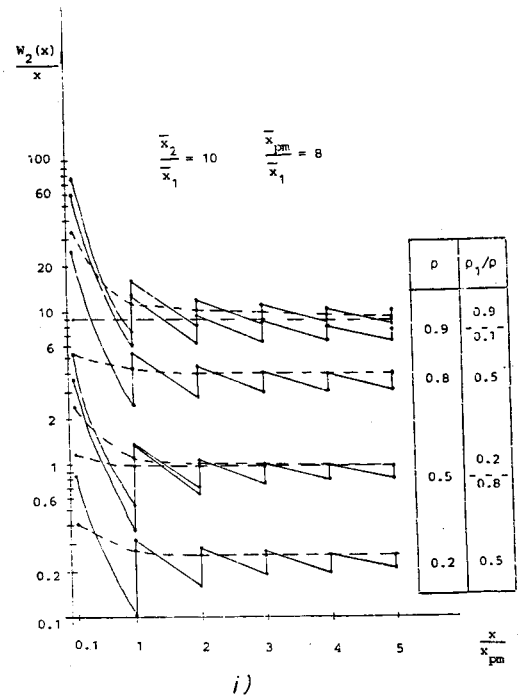


Fig. 2. Esperas condicionales de mensajes no prioritarios. En trazo continuo, análisis cuanto finito; en trazo discontinuo análisis cuanto nulo.



Continuación Figura 2.



i)

Continuación Figura 2.

4. ANALISIS DE RESULTADOS.

Acabamos de comprobar que según el modelo de cuanto nulo $W_2(x)/x$, la espera condicionada relativa, tiende asintoticamente hacia una constante, lo cual está de acuerdo con nuestros objetivos para el FCP.

Es de esperar que con el modelo de cuanto finito se observe un comportamiento similar de $W_2(x)/x$, (con $W_2(x)$ dado por (23) y (24),/16/. Pero la complejidad de la expresión (23) impide una estimación rápida de su comportamiento. En consecuencia, hemos evaluado $W_2(x)/x$, para diversos casos particulares, con los siguientes objetivos:

1º) Observar el comportamiento de $W_2(x)/x$, según el modelo de cuanto finito.

2º) Comparar cuantitativamente los resultados de $W_2(x)/x$ obtenidos con ambos modelos.

En las figuras 2(a-l) se muestran en abscisas la espera condicionada relativa, $W_2(x)/x$, y en ordenadas el tiempo de servicio, x. Las curvas de trazo contínuo corresponden al modelo de cuanto finito, mientras que las de trazo discontínuo son del modelo de cuanto nulo. Cada gráfica se ha obtenido para unos valores fijos de $\bar{x}_2/\bar{x}_1$ y $\bar{x}_1/x_{pm}$ es decir, las relaciones entre el tiempo medio de servicio de los usuarios-2 y el de los usuarios-1, y entre este y la duración máxima de un cuanto. Las diferentes curvas de cada gráfico se han construido tomando como parámetros p y p_1/p, (con $p = p_1+p_2$).

Se ha evaluado la espera condicionada para diferentes servicios $\bar{x}_1$, como muestra se discuten las gráficas correspondientes a un servicio típico de los usuarios-1, $\bar{x}_1 = 10/24$, (p.ej. un mensaje de mil bits sobre una línea de 2400 bit/s), dado que las curvas presentan un comportamiento similar para diferentes servicios $\bar{x}_1$.

Para cada par de curvas (cuanto finito - cuanto nulo) asociado a unos parámetros dados se ha medido su distancia en función de su separación relativa (en %) para un tiempo de servicio de cinco cuantos ($x/x_{pm}= 5$). Es decir, la distancia desde el punto de la curva con cuanto nulo (punto de referencia) hasta la cresta (o valle) de la curva de cuanto finito dividida por el valor de la cresta (o valle).

En síntesis se observa /14/ que para servicios alejados del origen, el modelo de cuanto nulo es en general conservador (pesimista) excepto para la situación extrema, p=0.9, $p_1/p=0.9$, donde es optimista como mucho en un 29% (para $\bar{x}_2/\bar{x}_1=2$ y $x_{pm}/\bar{x}_1=2$). Y en general es tanto más pesimista cuanto mayor es el cuanto x_{pm}, respecto a $\bar{x}_1$, con un máximo de pesimismo de un 148% para una situación extrema con $x_{pm}/\bar{x}_1=8$, y $\bar{x}_2/x_{pm}=1/4$, que está bastante alejada de las condiciones típicas en un FCP.

En consecuencia, concluimos que el modelo de cuanto nulo constituye una aproximación aceptable para el análisis de la disciplina de fragmentación-concentración de paquetes propuesta. Cuya espera,condicionada a la longitud del mensaje, se expresa en una forma muy compacta que permite hablar de una "tasa de retardo" por bit a transmitir, dada por el cociente p/1-p.

5. BIBLIOGRAFIA.

/1/ I. ADIRI, "Computer Time-Sharing Queues with Priorities", JACM, Vol. 16 nº 4, Octubre 1969, pg. 631-645.

/2/ I. ADIRI, B. AVI-ITZHAK, "A Time-Sharing Queue", Management Science, Vol. 15 Nº11, Julio 1969, pg. 639-657.

/3/ J. BABA, "A Generalized Multi-entrace -- Time Sharing Priority Queue", JACM, Vol. 22, nº 2, Abril 1975, pg. 232-247.

/4/ P. BOCKER, V. THOMANEK, "Current and Future Problems in Packet Switching Concepts" Conference Record ICC 79, Junio 1979, pg. 20.7.1-20.7.5.

/5/ A. COBHAM, "Priority Assignement in Waiting Line Problems", Operations Research, Vol. 2, 1954, pg. 70-76.

/6/ E. COFFMAN, L. KLEINROCK, "Feedback Queueing Models for T-S Systems", JACM, Vol.15, nº 4, Oct. 1968, pg. 549-576.

/7/ J. CHAMMAS, "Response Times over Packet Switched Networks-some Performance --- Issues", ICCC82, London, September 1982, pg. 993-998.

/8/ G. FAYOLLE y otros, "Sharing a Processor Among Many Job Classes", JACM, Vol. 27 - nº 3, Julio 1980, pg. 519-552.

/9/ H. HEACOX, "Analysis of two T-S Queueing Models", JACM, Vol. 19, nº 1, Enero 1972, pg. 70-91.

/10/ L. KLEINROCK, "Time-Shared Systems: A Theoretical Treatment", JACM, Vol. 14, nº 2, Abril 1967, pg. 242-261.

/11/ L. KLEINROCK,"Queueing Systems"- Vol.I Theory, J. Wiley, New-York, 1975.

/12/ L. KLEINROCK, "Queueing Systems", Vol. II, Computer Applications, J. Wiley, -- New-York 1976.

/13/ H. MIYAHARA y otros "Delay and Trough-put Evaluation of Switching Methods in Communications Networks", IEEE Trans--- actions on Communic. Vol. COM-26, nº 3, Marzo 1978, pg. 337-344.

/14/ J. VINYES, "Contribución al análisis y modelado de Redes de Conmutación de Paquetes", Tesis Doctoral. ETSI Telecomunicación, Universidad Politécnica de Madrid, Octubre 1980.

/15/ J. VINYES, J.B. RIERA, "Análisis del -- Comportamiento de un Sistema de Tiempo Compartido con Prioridades". CIL-83. -- Barcelona, Junio 1983.

/16/ J. VINYES, J.B. RIERA, "Análisis del -- Comportamiento de un Fragmentador-Con-- centrador de Paquetes. Modelo de Cuanto Finito". QUESTIIO, V.7 Nº 3.

/17/ H. ZIMMERMANN, "High Level Standardization: Technical and Political Issues", Proceedings ICC76, Agosto 1976, pg. 373-376.