Desequilibrio de Carga en Redes *k-ary n-cube*

J. Miguel-Alonso[**], J.A. Gregorio[*], V. Puente[*], F. Vallejo[*], R. Beivide* y P. Abad[*]

Resumen-- **Este trabajo estudia el efecto que tiene el bloqueo de cabeza (HOL) en el buffer de inyección sobre el rendimiento de las redes *k-ary n-cube* para valores de *k* por encima de un cierto umbral (aprox. 20). El bloqueo HOL *(Head of Line blocking)* provoca un uso desequilibrado de los canales correspondientes a las dos direcciones de los enlaces bidireccionales y da lugar a una caída del rendimiento de red y a un incremento del retraso de los paquetes. Los resultados de simulación muestran que esta anomalía solamente se presenta en aquellos anillos donde se realizan la mayor parte de las inyecciones de paquetes (normalmente, las del eje X) y la eliminación de este bloqueo permite a la red mantener el *throughput* a su valor máximo de pico incluso una vez alcanzado el punto de saturación.**

Palabras clave—**Redes de Interconexión, desequilibrio de carga, bloqueo de cabeza.**

I. Introducción

El rendimiento de la red de interconexión de un computador paralelo tiene un gran impacto sobre todo el sistema. Las topologías de redes directas más comunes son las denominadas *k-ary n-cube* y abarcan los anillos, mallas y toros. Un elemento central de este tipo de redes es el encaminador (*router*) que inyecta y recoge los paquetes de información desde el elemento de proceso al que se encuentra conectado y redirige los paquetes de otros nodos que no vayan dirigidos a él.

La arquitectura del encaminador tiene un impacto fundamental en el tiempo de ejecución de las aplicaciones que se ejecuten en la máquina paralela [4],[7]. Los objetivos de diseño típicos son mantenerlo simple para reducir el tiempo de ciclo, pero con la máxima funcionalidad posible. Esta simplicidad conduce, por un lado, a situar los buffers de tránsito en la entrada y, por otro, a que tanto éstos como las colas de inyección empleen una política FIFO, pese a los conocidos efectos negativos de la misma. Entre ellos, uno de los más importantes es el bloqueo de cabeza *(Head-of-Line blocking, HOL)*, sobre el que se han realizado numerosos trabajos tratando de reducir sus efectos en las colas de tránsito [5]. Sin embargo, el bloqueo HOL en la inyección, hasta donde conocemos, no ha sido reportado en la literatura y en este artículo se muestra que también tiene un impacto negativo que afecta severamente a la escalabilidad de este tipo de redes. De hecho, el bloqueo de cabeza en la interfaz de red es una de las principales razones por las que el rendimiento cae repentinamente cuando la red sobrepasa el punto de saturación. Mientras incrementamos la carga aplicada, antes de alcanzar el punto de saturación, la red acepta toda la carga ofrecida. Sin embargo, en ocasiones, cuando se alcanza este punto máximo la carga aceptada cae por debajo de los niveles alcanzados hasta el momento de la saturación, en lugar de mantener el nivel máximo alcanzado.

Esta anomalía aparece como un desequilibrio entre los canales que van en direcciones opuestas, aunque sólo en aquellos anillos de las redes *k-ary n-cube* donde se realicen la mayor parte de las inyecciones (típicamente los del eje X) y sólo para tamaños de anillo por encima de un cierto umbral (en torno a 20 nodos por anillo). Este umbral depende de muchas características de la red, tales como la política de evitación de bloqueos, el número de canales virtuales, etc. y por tanto, el número 20 como umbral debe ser tomado como un punto de referencia y no como valor absoluto. Hemos observado variaciones en la forma en que este desequilibrio se materializa para diferentes valores de *k*. En algunos casos, los canales X+ están llenos durante la mayor parte del tiempo de simulación mientras que X- está prácticamente vacío (también puede ocurrir al contrario). En otros casos se producen oscilaciones, pero en ninguno de los analizados el fenómeno desaparece espontáneamente. Para eliminar esta anomalía es necesario eliminar el bloqueo de cabeza de la cola de inyección y, aunque en el proceso de simulación (con tráfico uniforme) la hemos eliminado tirando los paquetes que bloquean la cola, una implementación realista requeriría el uso de colas no-FIFO o la implementación de inyectores separados, uno por dirección.

El resto del artículo esta organizado de la siguiente manera: en la sección 2 se describe la anomalía y el contexto donde aparece. En la sección 3 se presenta su aspecto en anillos *(1-cubes)* y la sección 4 se dedica a describir las causas del desequilibrio y las posibles medidas a tomar para evitarlo. En la sección 5 se estudia el desequilibrio de carga en toros 2D y 3D y en la sección 6 se analiza la aparición del desequilibrio cuando se ejecutan aplicaciones reales. Por último, en la sección 7 se presentan las principales conclusiones del trabajo.

II. Descripción general de la anomalía

Hemos observado que, en las redes *k-ary n-cube* con enlaces bidireccionales, cuando la carga aplicada excede el límite de saturación de la red para el correspondiente patrón de tráfico (lo hemos limitado a tráfico uniforme), la ocupación de los recursos de la red no esta equilibrada en las dos direcciones. Esta anomalía se presenta en cualquiera de las redes mencionadas, y se origina principalmente por la estructura del encaminador de paquetes. La Figura 1 describe la organización de nuestro encaminador básico donde se muestran los módulos usuales: crossbar, buffers, lógica de arbitrio, sincronización, etc.

El diseño del encaminador debe maximizar el uso de los recursos de la red evitando anomalías tales como interbloqueo *(deadlock)*, inanición *(starvation)* y *livelock*. En nuestros experimentos hemos utilizado un encaminador que tiene dos o tres canales virtuales (colas FIFO) por enlace de entrada para soportar encaminamiento completamente adaptativo basado en el

método de la burbuja *(Adaptive Bubble Routing, ABR)* [7]. Cuando se emplea ABR, un subconjunto de los canales virtuales es configurado como una red virtual segura (o de escape) en la que los paquetes nunca se bloquean entre ellos. Los canales virtuales restantes se configuran como una red virtual completamente adaptativa. Los paquetes se mueven bajo dos diferentes políticas. En la red virtual adaptativa, la inyección y el tránsito de paquetes vienen regulados por el control de flujo VCT *(Virtual Cut-Through)* y el encaminamiento adaptativo mínimo. En la red virtual segura, la inyección de los paquetes se regula mediante el mecanismo de la burbuja *(Bubble Flow Control, BFC)*, que evita el interbloqueo de paquetes en topologías basadas en un anillo o en un conjunto de anillos. En el caso de las topologías compuestas por un conjunto de anillos, estos deben ser visitados en "orden de dimensión" (DOR) y los paquetes que pasen de un anillo a otro serán considerados por el anillo receptor como si fuesen nuevas inyecciones. El movimiento de los paquetes en tránsito dentro del anillo seguro se regula mediante VCT, pudiendo moverse libremente desde la red segura a la adaptativa. Sin embargo, en el paso desde la red adaptativa a la red segura la política de regulación debe ser BFC.

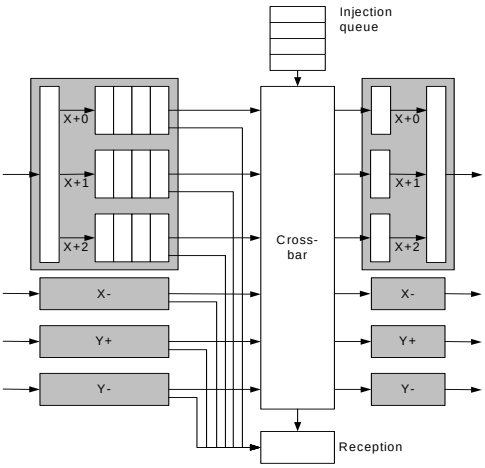

Figura 1. Modelo de encaminador (router) para una k-ary 2-cube. En este caso, cada canal físico es compartido por tres canales virtuales (0 ó escape, 1 y 2). Cada canal virtual tiene su propia cola de entrada para almacenar los paquetes en tránsito.

Esta red de interconexión es muy similar en topología (un toro 3D), control de flujo, mecanismo de evitación de interbloqueos, etc. a la empleada en el super-computador BlueGene/L de IBM [1]. No obstante, hemos observado que el mencionado desequilibrio también aparece empleando diferentes políticas de evitación de interbloqueos, tales como el clásico uso de canales virtuales propuesto por Dally [2].

III. DESEQUILIBRIO DE CARGA EN LOS ANILLOS

Vamos a considerar un anillo *(k-ary 1-cube)* con nodos como los representados en la Figura 1. Cada router se conecta a sus vecinos mediante dos enlaces físicos: X+ (para paquetes moviéndose de izquierda a derecha) y X- (para paquetes moviéndose de derecha a izquierda). La anchura de banda de cada canal físico es un *phit* por ciclo. Adicionalmente, el router se conecta (vía una interfaz) con un elemento de proceso. Para el resto de esta sección consideraremos además las siguientes características de la red:

1. Cada canal físico es compartido por dos canales virtuales. Los denominaremos: X+0, X+1, X-0 y X-1. Los canales "0" son los de escape y los "1" los adaptativos; aunque no es posible adaptarse en un anillo (solamente hay un camino para alcanzar el destino), no se aplica la restricción de la burbuja para la inyección en estos canales.
2. Los paquetes son de 16 *phits*.
3. Cada VC tiene una cola de tránsito de 8 paquetes (128 *phits*) y la capacidad de la cola de inyección también es de 8 paquetes.
4. El patrón de tráfico es uniforme.

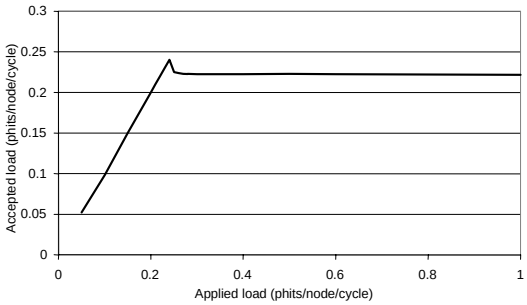

Figura 2. Carga aceptada frente a carga aplicada en una red 30-ary 1-cube. Tráfico uniforme.

Hemos estudiado el rendimiento de esta red inyectando una carga variable (medida en *phits*/ciclo/nodo) de tráfico y determinando la carga realmente soportada por la red. Los datos para un anillo de 30 nodos se muestran en la Figura 2. Toda la carga aplicada es transportada por la red hasta que se alcanza el punto de saturación. Después de ese punto, se observa una caída brusca de la carga aceptada (la magnitud de esta caída también depende de muchos parámetros de diseño). Una causa de esta caída es que cuando se alcanza el punto de saturación, los recursos de la red (las colas) no se emplean de forma equilibrada. Los canales que mueven paquetes hacia la derecha (X+) pueden estar saturados, mientras que las colas en la dirección X- están prácticamente vacías. También puede suceder lo contrario, X- llenos y X+ vacíos.

La Figura 3 muestra la ocupación de las colas para diferentes tamaños de anillos (10, 20, 30 y 100 nodos), todos ellos con las características señaladas anteriormente y sometidas a una elevada carga aplicada (siempre por encima del punto de saturación) de tráfico uniforme. Se ha hecho un muestreo de la ocupación de las colas de un nodo del anillo cada 2Kciclos. En el caso de un anillo de 10 nodos (a) se observa que la ocupación es siempre muy baja a lo largo de todo el tiempo de saturación. La distancia promedio recorrida por los paquetes desde el origen hasta el destino es muy corta y por ello se alcanza la saturación antes de que las colas estén completamente utilizadas.

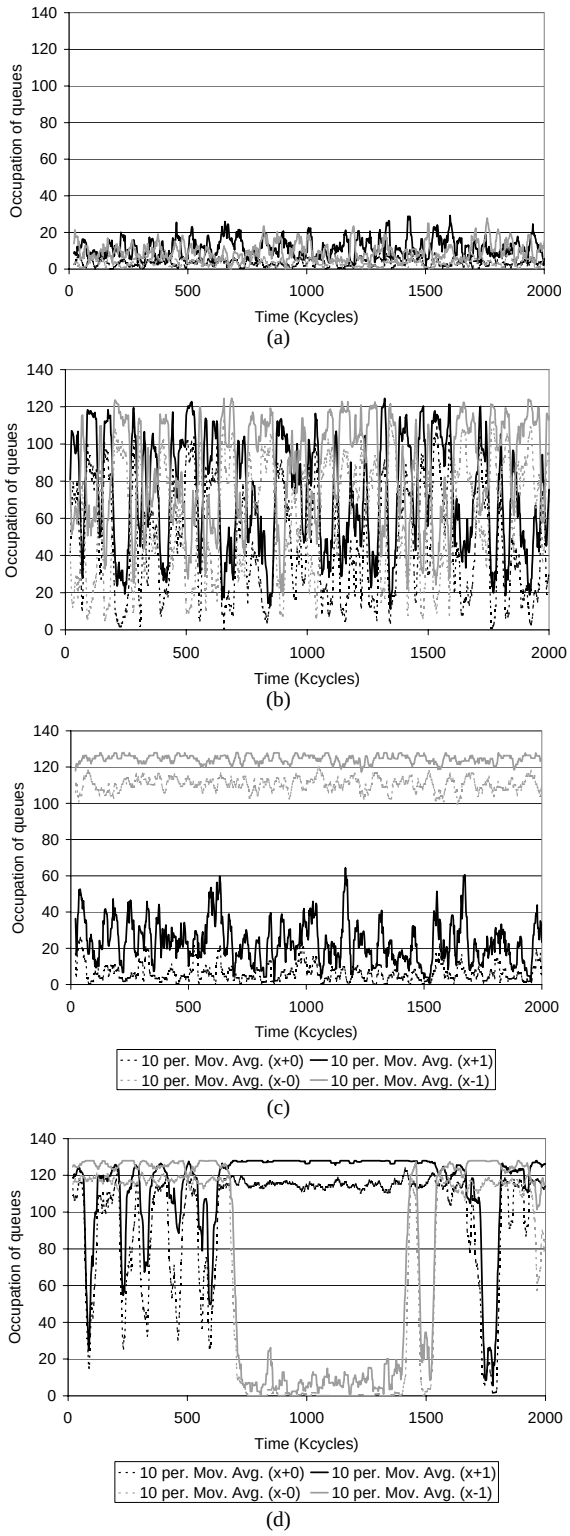

Figura 3 Ocupación de las colas en anillos de 10 (a), 20 (b), 30 (c) y 100 (d) nodos; carga uniforme a máximo nivel. Las medidas fueron tomadas cada 2000 ciclos de simulación; las gráficas representan la media movil. Los valores posibles estan entre 0 (cola totalmente vacia) y 128 (cola llena con 8 paquetes, 128 phits).

Para anillos grandes el comportamiento de la red es completamente diferente. Así, para un anillo de 20 nodos (b) se observa una rápida alternancia del desequilibrio: por un corto periodo de tiempo X+ esta vacío mientras que X- esta lleno, al cabo de unos pocos ciclos se invierte la situación. Para un anillo de 30 nodos (c) esta alternancia no sucede: casi inmediatamente que aparece el desequilibrio, éste se mantiene a lo largo de todo el intervalo de simulación: los canales X- están saturados mientras que los canales X+ tienen una ocupación de alrededor de 30 phits, ¼ de su capacidad.

La gráfica para el anillo de 100 nodos (d) muestra un escenario diferente que, de hecho es una combinación de los dos anteriores: durante 500 Kciclos la utilización de los canales X+ oscila, aunque se mantiene por debajo de la saturación, pero el X- esta completamente saturado; a continuación y durante unos 1000 Kciclos, X+ esta saturado mientras que X- esta mucho menos ocupado. Vemos, por tanto, cómo se puede invertir el desequilibrio.

En ningún caso, de los experimentos realizados con anillos de tamaño superior a 20 nodos, hemos observado la desaparición espontánea de la anomalía.

IV. Causas del desequilibrio y fomas de evitarlo

La causa que genera este desequilibrio es el bloqueo de cabeza de los paquetes en la cola de inyección del encaminador. Como se muestra en la Figura 1, los paquetes inyectados en la red desde un elemento de proceso (la única fuente/sumidero de tráfico) tienen que atravesar la cola de inyección asociada a la interfaz de red. Supongamos que el paquete situado en la cabecera de la cola de inyección tiene que ser encaminado hacia X+ (porque así lo indique el algoritmo de encaminamiento) y este canal esté saturado. El paquete debe esperar y también hará esperar (probablemente de forma innecesaria) al siguiente paquete de la cola, aunque tenga que ser enviado hacia X-. Bajo tráfico uniforme, la probabilidad de que se presente esta situación es la misma para X+ bloqueando a X- que para X- bloqueando a X+ y, por lo tanto, el patrón de tráfico no es el causante de esta anomalía.

Sin embargo, tan pronto como se produzca un cierto desequilibrio (que es algo que podemos esperar con tráfico uniforme), existe un efecto de realimentación positiva que provoca la inestabilidad. El paquete en la cabecera de la cola de inyección desea entrar en el canal saturado y debe esperar (forzando a los restantes paquetes de la cola a esperar también). Mientras tanto, el otro canal (menos ocupado) no recibe tráfico adicional y por tanto su carga decrecerá a medida que sus paquetes van siendo consumidos. Tan pronto como el canal saturado tiene espacio para otro paquete, el que estaba esperando lo usará y, por tanto, el canal se saturará de nuevo. Así, un canal permanece muy ocupado mientras el otro está casi vacío. Los paquetes viajando a través de éste último, cuando finalmente sean inyectados, rápidamente alcanzarán su destino. Por el contrario, los del canal muy ocupado atravesarán lentamente la secuencia de todas las colas llenas. El efecto del desequilibrio causado por el bloqueo HOL se acrecentará.

Si la causa de la anomalía es el bloqueo de cabeza, la solución para eliminar la anomalía será evitar dicho bloqueo. Por ejemplo, si se utiliza una política no-FIFO algunos paquetes pueden adelantar a otros que estén bloqueados [5]. Obviamente, esta solución incrementa la complejidad del encaminador. Otra posible solución es disponer de más de una cola de inyección en la interfaz de red, por ejemplo, una por dirección. Se puede realizar una decisión de encaminamiento preliminar en la

interfaz colocando el paquete en su correspondiente cola. Todos los paquetes almacenados en una de estas colas seguirán la misma dirección y por lo tanto no habrá bloqueo HOL. Esta solución, además de incrementar la complejidad de la interfaz, requiere colas de inyección suficientemente largas ya que, en caso contrario, el bloqueo HOL reaparecería en los niveles superiores (dentro del elemento de proceso) cuando se llenasen las colas.

Una solución trivial, empleada en este análisis de simulación, consiste en tirar los paquetes que causan el bloqueo (el que se encuentra en la cabecera de la cola de inyección, que no puede ser inyectado). Desde un punto de vista práctico, esta no es una solución válida porque forzaría a los niveles altos de los protocolos de comunicación (o incluso las aplicaciones) a tratar la recuperación de los paquetes perdidos y reduciría drásticamente el rendimiento. Sin embargo, es perfectamente aplicable cuando tratamos con procesos de simulación y cargas sintéticas.

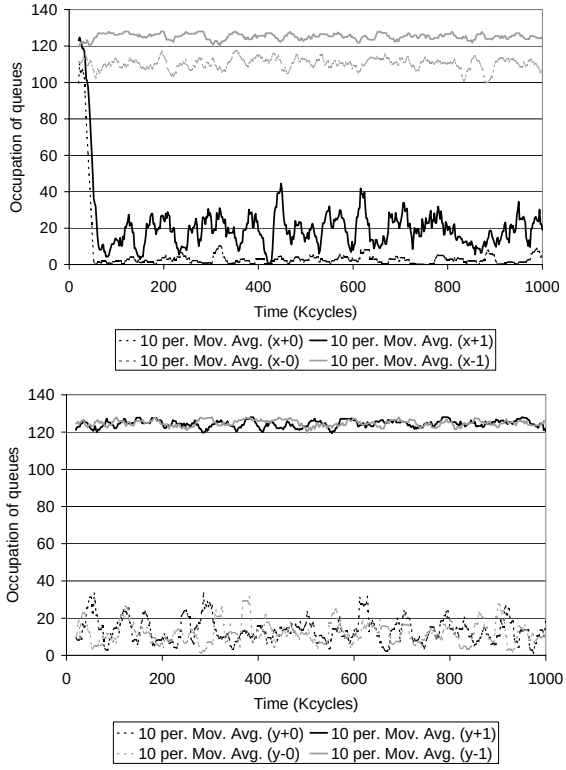

Figura 4 Ocupación de las colas en una red toroidal 30x30 para los anillos-X (arriba) y los anillos-Y (abajo) bajo carga uniforme máxima. Las medidas fueron tomadas cada 2000 ciclos de simulación; las gráficas representan la media movil. Los valores posibles estan entre 0 (cola totalmente vacia) y 128 (cola llena con 8 paquetes, 128 *phits*).

V. Estudio con toros 2D y 3D

Un toro 2D *(k-ary 2–cube)* esta formado por un conjunto de anillos en la dimensión X (anillos-X) y otro conjunto en la dimensión Y (anillos-Y). En un estudio de simulación de esta red hemos observado que los anillos-X tienen un comportamiento prácticamente idéntico al de los anillos analizados en la sección 3 para el caso de las redes 1D. Sin embargo, no aparece desequilibrio en los anillos-Y. La razón se encuentra en la política seguida para hacer avanzar los paquetes. En el caso de los canales adaptativos, no es imprescindible inyectar en un anillo-X, pero hay una cierta preferencia a hacerlo porque se comprueba la disponibilidad de los anillos-X antes de hacerlo con los anillos-Y. Los paquetes en el canal de escape avanzan siguiendo un encaminamiento en orden dimensional (DOR), por tanto bajo saturación (cuando los canales de escape son más utilizados) la mayor parte de los paquetes necesariamente deben ser inyectados en un anillo-X. Además, los canales de escape en los anillos-Y tienen más inyectores porque aceptan tráfico de los anillos-X y por tanto, la probabilidad de que suceda un bloqueo HOL es menor que en los anillos-X.

La Figura 4 representa la ocupación de las colas en un anillo-X y en un anillo-Y de un toro 30x30. A medida que avanzamos, el gráfico del anillo-X se asemeja al del anillo de la red unidimensional (1D) de 30 nodos. Sin embargo, las curvas para el anillo-Y muestran un comportamiento más razonable, pero diferente: los canales adaptativos Y+1 e Y-1 están altamente cargados (pero con carga similar entre ellos) y los canales de escape (Y+0, Y-0) están siendo mucho menos utilizados, y ambos en similar medida.

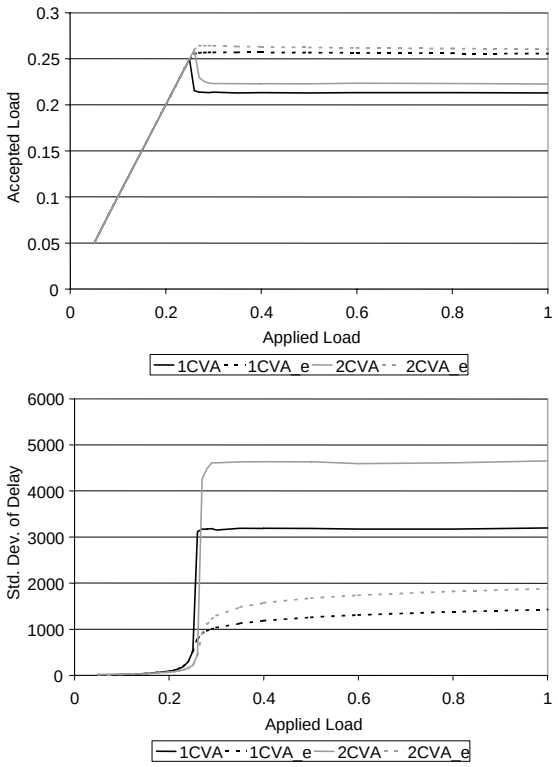

Figura 5 Arriba: Carga aceptada frente a carga aplicada en una red toroidal 30x30 con 1 y 2 VCs (además del canal de escape) y con/sin evitar el bloqueo HOL. Abajo: Desviación estandar del retraso promedio de los paquetes en esta red.

Lo que sucede con el rendimiento de la red se muestra en la Figura 5 (arriba) donde se ha representado la carga aceptada frente a la carga ofrecida para cuatro variantes del toro 30x30 con y sin emplear ningún mecanismo de evitación del bloqueo HOL. Se observa que el empleo de dos canales adaptativos (2CVA) mejora el rendimiento respecto al empleo de uno solo (1CVA). Sin embargo, el desequilibrio de los anillos-X causa caídas del rendimiento en ambos casos, aunque en el caso

"2CVA" esto sucede cuando se aplica una carga ligeramente superior.

Puede observarse que la eliminación del bloqueo HOL permite alcanzar un mayor nivel de carga aceptada (próximo al máximo teórico) y, lo que es más importante, mantener ese nivel bajo cargas aplicadas por encima del punto de saturación.

Por otro lado, es aun más destacable el impacto que tiene este desequilibrio sobre el retraso promedio (y su desviación estándar) experimentado por los paquetes cuando atraviesan la red. La Figura 5 (abajo) muestra que evitando el bloqueo HOL, la desviación estándar del retraso de los paquetes cae hasta alrededor de 1/3 de su valor original. Se comprende fácilmente que la existencia de un camino saturado junto a otro casi vacío causa una gran dispersión del número de ciclos que los paquetes requieren para alcanzar sus respectivos destinos: mientras unos paquetes experimentan una red prácticamente vacía, otros tienen que competir para pasar a través de una colección de canales saturados.

El desequilibrio también aparece en los anillos-X en los toros 3D cuando se sobrepasa el umbral señalado (en torno a los 20 nodos). Una consecuencia obvia es que este fenómeno solo puede ser observado en redes muy grandes (alrededor de 8000 nodos). A modo de ejemplo, la red toroidal previa de 30x30 (900 nodos) sufre de este efecto, mientras que una red 3D similar (10x10x10, ó 1000 nodos) no tiene esta anomalía. Además de las mejores características topológicas (en términos de anchura de banda de la bisección, distancia promedio, etc.) esta puede ser otra razón para hacer el toro 3D una topología mas apropiada.

VI. Desequilibrio en aplicaciones reales

En esta sección se comprueba que el desequilibrio no es una consecuencia de emplear una carga sintética uniforme. Empleando un simulador de sistemas multiprocesadores [6] integrado con el simulador de redes de interconexión SICOSYS [8], hemos llevado a cabo una simulación conducida por ejecución de la aplicación Radix (parte de los *benchmarks* SPLASH-2) sobre un anillo de 30 nodos. La Figura 6 muestra los resultados de la medida, cada 5000 ciclos, de la ocupación de las colas de tránsito de los canales virtuales. La longitud de los paquetes es de 20 phits y la capacidad de cada cola de 4 paquetes.

Esta figura muestra los resultados para una aplicación caracterizada por su patrón de tráfico uniforme, como es Radix. Entre los ciclos 1M y 3M el canal X-1 (adaptativo) esta mucho más ocupado que el X+1, siendo este efecto claramente visible en torno al ciclo 2,5M. Además, en torno a los 6Mciclos se observa el fenómeno de alternancia previamente descrito para los anillos. EL desequilibrio no puede ser atribuido a las características de la aplicación (los nodos ejecutando Radix no envían más datos en una dirección que en otra) porque para los periodos de tiempo representados en la figura la aplicación intercambia "llaves" de manera altamente uniforme y aleatoria. Por tanto, este comportamiento confirma nuestra hipótesis acerca de la ocurrencia de esta anomalía no solo con cargas sintéticas, sino también con aplicaciones reales.

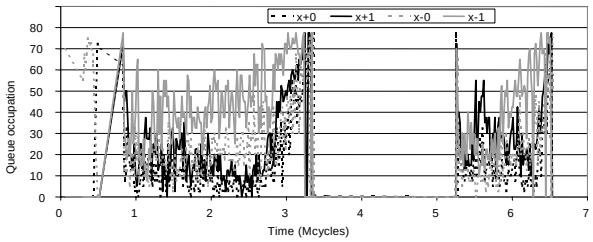

Figura 6 Ocupación de las colas cuando se ejecuta la aplicación paralela Radix sobre un anillo de 30 nodos con 512K llaves. Datos tomados de una simulación conducida por ejecución combinando RSIM y SICOSYS.

VII. Conclusiones

Hemos identificado y analizado el desequilibrio de carga que aparece en anillos bidirecionales de redes *k-ary n-cube* cuando se aplica tráfico uniforme. Es causado por el bloqueo de cabeza *(HOL blocking)* en las colas de inyección. Comprobamos que la eliminación de este bloqueo permite mantener los niveles de *throughput* a los valores de pico alcanzados en la saturación, así como reducir la desviación estándar de la latencia de red hasta un 30%. Este desequilibrio aparece en redes 1D, 2D y 3D, bajo diferentes técnicas de evitación de interbloqueo de paquetes *(deadlock),* pero solamente para tamaños por encima de 20 nodos y en aquellos anillos donde más paquetes son inyectados (normalmente, los anillos de la dimensión X). Por último, hemos comprobado que la anomalía también está presente con tráfico generado por aplicaciones reales.

Agradecimientos

Este trabajo ha sido parcialmente financiado por Ministerio de Ciencia y Tecnología, (TIC2001-0591-C02-01 y TIC2001-0591-C02-02), y por la Diputación Foral de Guipúzcoa (OF-758/2003).

Referencias

[1] NR Adiga, GS Almasi, Y Aridor, M Bae, Rajkishore Barik, et al., "An Overview of the BlueGene/L Supercomputer", Proc. of SuperComputing 2002, Baltimore, Nov. 16-22, 2002.

[2] W. J. Dally and C. L. Seitz, "Deadlock-free message routing in multiprocessor interconnec-tion networks", IEEE Transactions on Computers, vol. 36, no.5, pp. 547-553, May 1987.

[3] J. Duato. "A Necessary and Sufficient Condition for Deadlock-Free Routing in Cut-Through and Store-and-Forward Networks". IEEE Trans. on Parallel and Distributed Systems, vol. 7, no. 8, pp. 841-854, 1996.

[4] J. Duato, S. Yalamanchili, and L. Ni, Interconnection networks. "An engineering approach", IEEE Computer Society, 2003.

[5] G. L. Frazier and Y. Tamir, "Dynamically-Allocated Multi-Queue Buffers for VLSI Com-munication Switches". IEEE Trans. on Computers, vol. 41, no.6, pp. 725-737, June 1992.

[6] V.S.Pai, P. Ranganathan, and S.V.Adve, "RSIM: An Execution-Driven Simulator for ILP-Based Shared-Memory Multiprocessors and Uniprocessors". IEEE TCCA New., Oct. 1997

[7] V. Puente, C. Izu, R. Beivide, J.A. Gregorio, F. Vallejo and J.M. Prellezo, "The Adaptive Bubble Router", J. of Parallel and Distributed Computing. Vol 61, no. 9, September 2001

[8] V. Puente, J.A. Gregorio, R. Beivide, "SICOSYS: An Integrated Framework for studying Interconnection Network in Multiprocessor Systems", Euromicro Workshop on Parallel and Distributed Processing, 2002