Balanceo Distribuido del Encaminamiento para el Control de Congestión en Redes Infiniband

Diego Lugones, Daniel Franco, Emilio Luque

Departament d'Arquitectura de Computadors i Sistemes Operatius,
Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona, Spain
Ph. +34-3-5812888
diego.lugones@aomail.uab.es; daniel.franco@uab.es; emilio.luque@uab.es

Abstract

Communications requirements in High Performance Computing (HPC) demand the use of Interconnections Networks to connect processing nodes. Sharing resources in high performance interconnection networks leads to message congestion. Congestion spreading increases latency and reduces network throughput causing important performance degradation. Nowadays most current techniques use message throttling to prevent injection of new messages in network congested region. Message throttling moves contention from switches to sources nodes in order to eliminate congestion, however global latency is highly incremented because of the time that packets must wait in the source node. In this paper, we propose a congestion control mechanism for InfiniBand networks based in an adaptive routing algorithm that perform a communication load balancing over several alternative paths, in order to take load away of the congested network zone, eliminating congestion and maintaining injection rate. Our mechanism's experimentation results show latency, throughput and dynamic behaviour improvement over InfiniBand original congestion control mechanism which is based in message throttling. The proposed mechanism use the management model defined in InfiniBand specs, thus full compatibility is provided.

Keywords: Parallel processing, Communications and networks, Adaptive routing algorithms, Congestion control, InfiniBand networks, Hot-spot avoidance, High Speed Interconnection Networks, Network monitoring, Communication load balancing.

Resumen

El uso de recursos compartidos en las redes de interconexión de alta performance puede provocar situaciones de congestión de mensajes que degradan notablemente las prestaciones, aumentando la latencia de trasporte y disminuyendo la utilización de la red. Hasta el momento las técnicas que intentan solucionar este problema utilizan la regulación de la inyección de mensajes. Esta limitación de la inyección traslada la contención de mensajes desde los conmutadores hacia los nodos fuente, incrementando el valor de la latencia promedio global, pudiendo alcanzar valores muy elevados. En este artículo, proponemos una técnica de control de congestión para redes InfiniBand basada en un mecanismo de encaminamiento adaptativo que distribuye el volumen de comunicaciones entre diversas trayectorias alternativas quitando carga de la zona de congestión, lo que permite eliminarla. La experimentación realizada muestra la mejora obtenida en latencia y *throughput*, respecto al mecanismo de control de congestión original de InfiniBand basado en la regulación de la inyección. El mecanismo propuesto es totalmente compatible y no requiere que se modifique ningún aspecto de la especificación, debido a que se utilizan componentes de gestión definidos en el estándar InfiniBand.

Palabras clave: Procesamiento Paralelo, Comunicaciones y Redes, Algoritmos de encaminamiento adaptativos, Control de congestión, Redes InfiniBand, Evitación de Hot-spots, Redes de interconexión de alta velocidad, Monitorización de la red de interconexión, Balanceo de la carga de comunicaciones.

1 INTRODUCCION

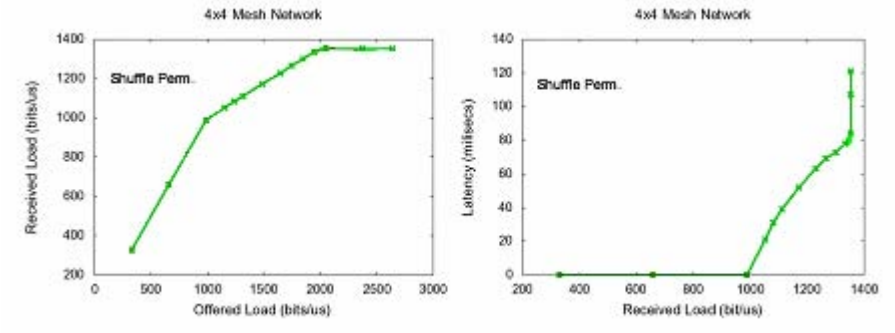
La evolución en el campo de las redes de interconexión para sistemas de cómputo de altas prestaciones *(High Performance Computing, HPC)* ha sido constante en los últimos años. Los avances tecnológicos han permitido una mejora importante en la velocidad de transmisión, incrementando significativamente el ancho de banda de los enlaces. Dichos avances también han tenido gran impacto en la integración de puertos en los conmutadores aumentando la cantidad de conexiones y permitiendo obtener topologías mas complejas y flexibles.

La reciente aparición de redes de interconexión comerciales como InfiniBand, Myrinet, Quadrics,…, etc. con alta velocidad de transmisión de datos, permiten construir las redes de interconexión que necesitan los sistemas de cómputo de altas prestaciones. Estas redes han tenido impacto no sólo en estos sistemas de cómputo, donde magnitudes adecuadas de velocidad de comunicación y tiempo de viaje de mensajes son de extrema importancia, sino también, en otros con menos requerimientos de performance, como redes de sistema (SAN) y clusters de ordenadores [12]. Los nuevos estándares desarrollados y las implementaciones comerciales deben su existencia a la creciente demanda de aplicaciones con grandes requisitos de cómputo.

El principal problema en el diseño de las redes de interconexión radica en manejo inadecuado de la congestión de mensajes en transito [3]. Dicha congestión aparece debido al uso compartido de los recursos de la red de interconexión (enlaces, buffers y conmutadores) y si esta situación no se controla eficientemente, es posible alcanzar la saturación de dichos recursos. Cuando la red no es capaz de manejar el volumen de comunicaciones que recibe en un momento dado, los mensajes en transito deberán competir por los recursos. Esta situación deriva en un aumento en el tiempo de viaje de los mensajes (latencia) y se propaga rápidamente a toda la red teniendo como efecto principal un deterioro global en la performance del sistema.

En la Figura 1 se ilustra la degradación de prestaciones en una red InfiniBand de 16 nodos y puede verse el comportamiento de la latencia de transporte de los mensajes y la carga de tráfico presente en la red. Cuando el volumen de mensajes inyectados presenta cargas bajas y medias de tráfico, la red es capaz de manejar este volumen ofrecido de mensajes y entregarlos a sus destinos; por lo tanto la carga aceptada por los nodos es igual a la carga inyectada (Figura 1 (a)) y la latencia se comporta de manera lineal y acotada (Figura 1 (b)). A medida que el tráfico aumenta y su valor alcanza un nivel determinado se observa que el tráfico recibido es menor al ofrecido, asimismo la latencia de los mensajes aumenta considerablemente. En este punto la red ingresa en el estado de saturación y se observa una degradación importante en sus prestaciones.

Las consecuencias derivadas de la congestión son aun mas graves en redes que no permiten el descarte de paquetes, como es el caso de las redes que conforman la mayoría de computadores paralelos de altas prestaciones [3].



(a) **(b)**

Figura 1. Degradación de performance en una red Infiniband con topología malla de 16 nodos. Patrón de trafico *Shuffle permutation*. Longitud de mensaje 4kB.

La solución tradicional al fenómeno de la congestión, radica pues en diseñar la red utilizando una cantidad de recursos mayor a la estrictamente necesaria *(Sobredimensionar la red)* de manera que no exista la necesidad de competir por dichos recursos y evitar la retención de paquetes [8]. Sin embargo esta practica ha quedado obsoleta debido, tanto al aumento en el coste de los componentes de red con respecto al de los procesadores empleados en los computadores paralelos actuales, como al elevado consumo de potencia de dichos componentes debido fundamentalmente al aumento de la velocidad de los enlaces [4]. Es por esta razón que el diseño de redes de interconexión sobredimensionadas no es una solución viable en la construcción de sistemas de cómputo reales.
La otra solución posible al problema de congestión, consiste en el empleo de técnicas *reactivas*, que a diferencia de las soluciones basadas en el sobredimensionamiento *(preventivas)*, utilizan los recursos estrictamente necesarios y los gestionan de forma eficiente. Estas técnicas monitorizan el estado de la red con el motivo de detectar la congestión y eliminarla mediante el uso de algún mecanismo.
Los mecanismos de control de congestión deben cumplir con un conjunto de características de diseño y operación que garanticen un uso eficiente de la red de interconexión y mejoren sus prestaciones de funcionamiento. En primer lugar el mecanismo de control de ofrecer un *tiempo de respuesta* convenientemente acotado para evitar reacciones tardías ante las situaciones de congestión que intentan controlar. Asimismo deben proveer un *balanceo global* de la carga de tráfico, ya que algunos mecanismos pueden proveer una respuesta adecuada a nivel local, pero conllevan un desbalance de carga en la red que conduce a una degradación global de las prestaciones. Por otra parte, ciertas técnicas reaccionan correctamente ante situaciones de congestión, pero *penalizan* el correcto funcionamiento de la red, debido a que generan sobrecarga de mensajes o nuevos fenómenos indeseables (ej. *deadlock, livelock, starvation* [8]) con cargas normales de trafico. El éxito de algunos mecanismos depende en gran parte de las condiciones de tráfico presente en la red, la topología ó el tamaño de los mensajes. Por esta razón el mecanismo de control de congestión debe ser *robusto* y mantener sus prestaciones en una amplia variedad de casos. Por ultimo la técnica utilizada debe ser *eficiente* y *escalable*, con el fin de eliminar el problema completamente y en una amplia variedad de topologías.
En este artículo presentamos el trabajo realizado en el diseño y la aplicación de un algoritmo de encaminamiento (denominado "*Distributed Routing Balancing*", DRB) vigente, factible y realista, citado recientemente en otros artículos y contribuciones, por distintos autores en estudios similares [1] y [4], sobre el emergente estándar InfiniBand cuya utilización esta ganando terreno velozmente en el campo de las redes de interconexión para computadores de alta performance, pero que carece de un control de congestión adecuado. En este sentido, mostraremos la experimentación realizada sobre nuestra propuesta de control de congestión, basada en el balanceo de la carga de comunicaciones y la importante mejora en las prestaciones con respecto al mecanismo de control original que ofrece InfiniBand y que se basa en la regulación de la velocidad de inyección de los mensajes.
El resto del artículo esta organizado de la siguiente manera: En la sección 2 se describen los antecedentes en los mecanismos de control de congestión y la descripción de DRB. La sección 3 presenta las características principales de la arquitectura InfiniBand que hacen posible la aplicación del balanceo distribuido del encaminamiento. En la sección 4 se describe nuestra propuesta de control de congestión. Los resultados de la evaluación de las prestaciones ofrecidas por el mecanismo propuesto para diversas topologías y distribuciones de tráfico, se presentan en la sección 5. Finalmente se presentan las conclusiones extraídas del trabajo realizado.

2 ANTECEDENTES

La capacidad de gestionar un alto número de mensajes sin que se produzca un gran aumento de la latencia es fundamental en las redes de interconexión de altas prestaciones, más aún cuando éstos

soportan aplicaciones en las que la relación de cómputo frente a comunicación es pequeña *(granularidad fina)*, y que por tanto generan un gran tráfico de mensajes entre los nodos de la red. Por este motivo, el control de congestión ha sido objeto de estudio durante los últimos años.
Las técnicas que intentan manejar y solucionar los problemas asociados a la congestión se basan en realizar una *monitorización* del tráfico que circula por la red, o de los recursos que la componen, con el motivo de *detectar* la congestión, *notificar* su existencia y llevar a cabo algún mecanismo para *controlarla y eliminarla*.
Algunas técnicas analizan el tiempo de bloqueo de mensajes (Latencia) [5], o el nivel de recursos ocupados (Canales o *Buffers*) [1][3][4] para determinar la existencia de congestión e informar de la misma al resto (o parte) de los nodos presentes en la red para que ejecuten las acciones que permiten evitar la degradación de performance en la red de interconexión.
Una de las acciones correctivas utilizadas por las técnicas de control, con el fin de eliminar la congestión es la típica regulación de la velocidad de inyección en el nodo fuente *(Message throttling)* [4][9][10] que detiene la inyección de nuevos mensajes, permitiendo que los conmutadores encaminen los mensajes ubicados en la zona congestionada hacia el destino manteniendo acotada la ocupación de los *buffers*. Esta técnica tiene efectos adversos que deterioran las prestaciones de la red, debido a que la latencia de transporte se incrementa considerablemente en virtud de la espera a que se someten los mensajes antes de su inyección.
Otra posibilidad utilizada en la eliminación de la congestión, consiste en el manejo y optimización del uso de los *buffers* en los puertos de los conmutadores *(switchs)* [1][3][4]. Estas soluciones son simples y fáciles de implementar, sin embargo no presentan buenas prestaciones debido principalmente a que no resuelven el problema de congestión a nivel global, sino que intentan manejar el volumen de comunicación mediante la organización de paquetes de forma local.
Por último, las técnicas basadas en el uso de encaminamiento adaptativo *(Adaptive routing)* [1][2][5] también permiten eliminar la congestión debido a que los mensajes son enviados a los destinos correspondientes, teniendo en cuenta el estado de los diferentes caminos posibles. De esta manera si algún puerto de un encaminador que pertenece a una determinada trayectoria se encuentra congestionado, el algoritmo modifica el envío utilizando trayectorias alternativas. Estas técnicas presentan mejores prestaciones pues permiten mantener el nivel de inyección de mensajes y actúan redistribuyendo la carga.
Recientemente, se han propuesto varias técnicas para el control de congestión en redes InfiniBand que intentan mejorar la utilización de la red. Estas técnicas [6][9][10] proponen la regulación de inyección de mensajes como mecanismo para eliminar la congestión. La principal desventaja de este mecanismo, radica en que la congestión se elimina trasladándola congestión desde los conmutadores hacia los nodos fuentes que inyectan los paquetes. De esta forma el comportamiento global de la latencia promedio se incrementa igualmente, pudiendo alcanzar valores muy elevados en presencia de cargas de tráfico adversas. Por otro lado la técnica propuesta en [2][13] permite el uso de trayectorias múltiples pero sólo puede utilizarse con el modelo de comunicaciones de conexión confiable *(Reliable Connection)* y no en los otros tres modelos soportados por Infiniband [1], ya que en estos casos no es posible realizar la detección y la notificación de congestión.

2.1 Balanceo distribuido del encaminamiento (DRB).

La técnica de control de congestión propuesta en este artículo está basada en un mecanismo de balanceo del encaminamiento que intenta uniformizar la carga en todos los enlaces de la red de interconexión. Este mecanismo se conocido como: Balanceo Distribuido del Encaminamiento ("*Distributed Routing Balancing*", DRB por sus siglas en inglés) [5] y se basa en la distribución uniforme de la carga en la red mediante la expansión de caminos. Esta expansión es dinámica y está controlada por el nivel de latencia existente en la red. El método establece nuevos caminos alternativos simultáneos entre cada par fuente y destino con objeto de mantener una latencia baja de

los mensajes. DRB define cómo crear los caminos alternativos para expandir los caminos simples originales y cuándo y cómo usarlos dependiendo del nivel de carga de tráfico en la red de interconexión. Es importante destacar que el mecanismo produce un efecto de balance colectivo (a nivel global), pues esta expansión se produce para todos los pares fuente-destino de la aplicación que interaccionan entre sí.
Conceptualmente, este método mide el estado de la carga en todas las conexiones entre pares fuente destino, al detectar congestión se notifica al/los nodo/s que inyectan mensajes para que configuren nuevas trayectorias posibles y redistribuyan el trafico según su estado de carga. Este funcionamiento se muestra en la figura 2, donde se observa que se detecta la presencia de congestión en los nodos intermedios y se notifica al nodo fuente su existencia mediante un mensaje de reconocimiento. A continuación este nodo determina las trayectorias alternativas que utilizara en el encaminamiento posterior de mensajes. Es importante destacar el concepto en el que el algoritmo basa su funcionamiento, porque puede ser implementado de diversas maneras. Es decir: la detección puede realizarse mediante la medición de la latencia acumulada a través del enlace, o en función de la ocupación de los buffers en los puertos de los conmutadores, etcétera. Mientras que la redistribución de tráfico puede llevarse a cabo de manera aleatoria, o en función de la ocupación de los enlaces, o de la velocidad de éstos en redes no homogéneas.

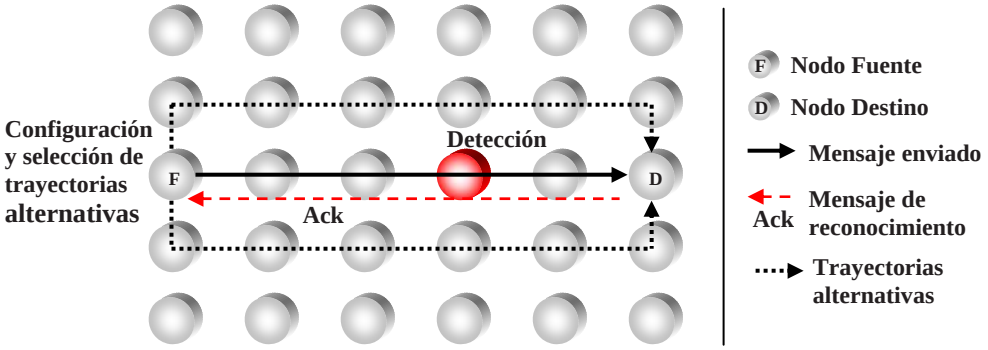


Figura 2. Fases de balanceo distribuido del encaminamiento DRB

Esta versatilidad en la implementación del algoritmo, junto a su buen comportamiento hace que su aplicación a tecnologías de interconexión actuales sea un tema de investigación interesante, tanto para la mejora y evaluación de esta técnica, como para desarrollo de nuevas propuestas sobre las tecnologías utilizadas.

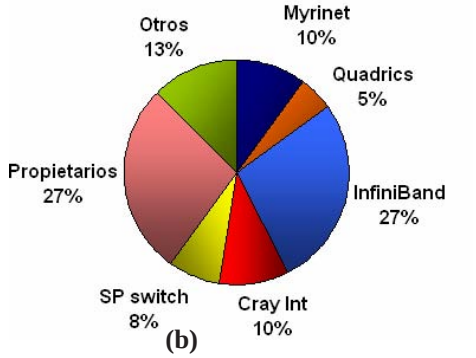
3 LA ARQUITECTURA INFINIBAND (IBA)

En la última década han aparecido una importante cantidad de especificaciones que acompañadas del avance tecnológico adecuado, pretenden desarrollar y estandarizar redes de comunicaciones con las características y requerimientos (conexiones punto a punto, baja latencia, elevado ancho de banda, etc.) necesarios en las redes de interconexión. Puntualmente, la especificación InfiniBand [1] es una nueva y poderosa arquitectura diseñada, no sólo para cubrir con las demandas de performance asociadas con el movimiento de datos en los dispositivos entrada-salida, sino también para conformar los cluster de cómputo de altas prestaciones *(High Performance Computing HPC)*, debido al elevado ancho de banda y la baja latencia de transporte que ofrece.
Los clusters InfiniBand de gran escala están ganado gran popularidad según lo reflejan los rankings de supercomputadores en el *top500* [11] tal como se muestra en la figura 3, donde puede verse como ha evolucionado InfiniBand en el último año, hasta alcanzar la segunda posición entre los sistemas de interconexión usados en los supercomputadores más potentes del mundo (figura 3(a)). La figura 3(b) muestra solamente los cuarenta primeros puestos, y se observa que InfiniBand es el estándar mas utilizado debido a las prestaciones que ofrece. Al mismo tiempo, las topologías directas (mallas, toros, hipercubos…) y la topología fat-tree se han convertido en las más utilizadas

en la interconexión para estos clusters, debido a que permiten múltiples trayectorias disponibles entre un mismo par de nodos. No obstante, incluso en estas topologías, pueden ocurrir situaciones de congestión, que dependen principalmente de la configuración de trayectorias entre nodos y del patrón de comunicación de la aplicación. Para empeorar aun más la situación, la naturaleza determinista del encaminamiento utilizado por defecto en InfiniBand, limita a las aplicaciones del uso eficaz y transparente de trayectorias múltiples que permiten evitar situaciones de congestión.

Interconexión	Junio 2007	Noviembre 2006	Junio 2006
Gigabit Ethernet	41.20 %	42.60 %	51.20 %
Myrinet	9.20 %	15.80 %	17.40 %
InfiniBand	25.60 %	15.60 %	7.20 %
Quadrics	2.20 %	2.80 %	2.80 %

(a)



(b)

Figura 3. Evolución y uso de IBA en el *top500*

InfiniBand define una arquitectura de red que permite interconectar múltiples nodos de procesamiento y dispositivos de entrada-salida, utilizando una red arbitraria con conexiones punto a punto como se muestra en la Figura 4. Los nodos de procesamiento pueden incluir varias CPUs y módulos de memoria, y utilizan uno o varios adaptadores de canal *(Channel Adapters, CAs)* como interfase para conectarse con los conmutadores de la red. La red se estructura en diferentes subredes *(subnets)* que interconectan los nodos de procesamiento a través de varios conmutadores *(switchs)*, las subredes se interconectan entre si mediante el uso de encaminadores *(routers)*.

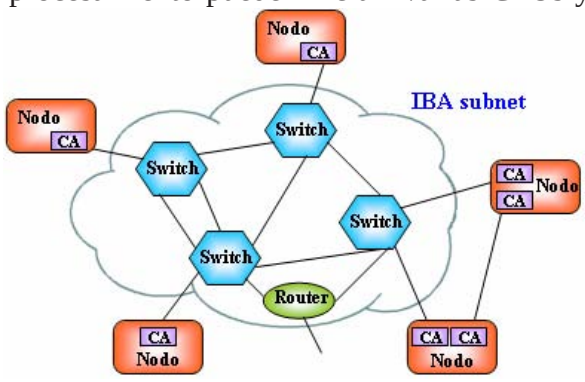


Figura 4. Subred InfiniBand

InfiniBand especifica un protocolo que divide la arquitectura en múltiples capas independientes, la capa física, la capa de enlace, la capa de red, la capa de transporte y las capas superiores. La capa de enlace y la de transporte conforman el corazón de la arquitectura, ya que en estas se crean los paquetes, se establecen las conexiones punto a punto y se realiza la conmutación dentro de la subred. IBA utiliza un mecanismo de canales virtuales (VL) para crear múltiples enlaces virtuales dentro de uno físico, mejorando su utilización.

Dentro de una subred, todos los dispositivos IBA tienen un identificador local de 16 bits *(Local Identificator, LID)* asignado por el gestor de subred *(Subnet Manager, SM)*. Los paquetes tienen incluido este identificador en sus cabeceras y los conmutadores lo utilizan para determinar el puerto de salida correspondiente y realizar el encaminamiento en la subred.

En IBA, las subredes se gestionan de una manera independiente usando un modelo de gestión *(Management Model)* en el que varias entidades se comunican para configurar el funcionamiento y las operaciones de la red. El gestor de subred (SM) es el encargado de descubrir los componentes de la subred y configurarlos (asignándoles los LIDs), y de activar y monitorizar la subred. El SM puede estar presente, tanto en un nodo de procesamiento como en un conmutador y puede implementarse tanto en software como en hardware. Las operaciones mencionadas se realizan utilizando paquetes de gestión que transportan la información entre el gestor de subred *(SM)* y los agentes de gestión *(Subnet Management Agent, SMA)* presentes en todos los dispositivos.

El gestor de subred busca los componentes, les asigna un identificador local y encuentra las trayectorias entre ellos. Asimismo puede configurarse para establecer varias trayectorias entre los

diferentes pares de nodo fuente-destino de la red utilizando una mascara de control *(Local Mask Control, LMC)*, que permite asignar hasta 2^{LMC} identificadores locales a cada puerto. Esta multiplicidad de nombres proporciona la manera de establecer múltiples trayectorias, entre el mismo par de nodos.

Otro componente importante del modelo de gestión es el gestor de control de congestión *(Congestion Control Manager, CCM)*, que provee los mecanismos necesarios para realizar la detección de congestión, generar y enviar notificaciones a los nodos fuente para que lleven a cabo la regulación de la inyección de mensajes.

Mediante el uso de elementos de medición, cada conmutador conoce el estado de ocupación en los buffers de cada canal virtual. La especificación establece un valor umbral de ocupación de los buffers, por encima del cual se activan los mecanismos destinados al control de congestión. Cuando se detecta congestión, el conmutador informa de esta situación marcando los paquetes que están situados en el buffer del canal virtual que ha superado el umbral. Dentro de la cabecera de transporte *(transport header)* de todos los paquetes, existe un bit destinado a tal efecto. Este bit es denominado *Forward Explicit Congestión Notification (FECN).* Una vez que el paquete es marcado, se reenvía por el puerto correspondiente en función la trayectoria especificada.

Si el nodo destino recibe un paquete marcado, el agente de control de congestión de dicho nodo solicita el envío de un mensaje de notificación *(congestión Notification, CN)*, con el objetivo de informar al nodo fuente la existencia de congestión en la trayectoria establecida entre ambos nodos. La notificación se hace efectiva mediante el uso de otro bit en la cabecera de transporte del mensaje CN, conocido como *Backward Explicit congestión Notification (BECN)*. El mensaje de notificación se envía hacia el nodo fuente. El agente de control de congestión analiza el bit *BECN* y responde informando al nodo que disminuya la inyección de mensajes. De esta manera, los puertos congestionados pueden recuperarse liberando los paquetes contenidos en sus buffers. La disminución en la inyección será más restrictiva dependiendo de la cantidad de mensajes de notificación recibidos.

Eventualmente, la congestión desaparece y la inyección debe recuperarse. Esta tarea se realiza utilizando un temporizador. Cada vez que transcurre un intervalo de tiempo sin que se hayan recibido mensajes de notificación, se recupera el nivel normal de inyección. Como se ha mencionado previamente, la principal desventaja de este mecanismo, radica en que la congestión se elimina trasladando la congestión desde los conmutadores, hacia los nodos fuentes que inyectan los paquetes. De esta forma el comportamiento global de la latencia promedio se incrementa igualmente, pudiendo alcanzar valores muy elevados en presencia de cargas de tráfico adversas.

Teniendo en cuentas estas razones, proponemos la aplicación del balanceo distribuido del encaminamiento en redes InfiniBand, utilizando en forma conjunta las características de establecimiento de trayectorias múltiples (ofrecido por el SM) y los mecanismos de monitorización de recursos y notificación mediante mensajes de reconocimiento (ofrecidos por el CCM), para aplicar un mecanismo de control de congestión eficiente, utilizando un concepto que permite un alto grado de utilización de los enlaces y un bajo valor de latencia de transporte en los mensajes, debido al uso de trayectorias alternativas que permiten la distribución de la carga de trafico, y no a la regulación de inyección propuesta en el mecanismo original.

4 BALANCEO DISTRIBUIDO DEL ENCAMINAMIENTO EN REDES INFINIBAND

La aplicación del balanceo distribuido del encaminamiento se lleva a cabo dentro del contexto establecido por la arquitectura de InfiniBand, teniendo en cuenta las definiciones descriptas en la especificación y utilizando sus características. De esta manera no se requiere modificación alguna y se mantiene la compatibilidad con el estándar IBA.

La *detección* de la congestión se lleva acabo en los canales virtuales de cada puerto, monitorizando la ocupación de los mismos con respecto a un umbral relativo al tamaño del buffer. El valor del umbral es establecido por el CCM y los diversos valores posibles, varían entre los números 0 y 15; donde el valor 0 indica que ningún paquete ha de marcarse en este puerto, y el valor 15 especifica un umbral muy restrictivo, ver Figura 5. Cuando el mecanismo detecta congestión dentro del conmutador, informa al nodo destino poniendo a *uno* el bit denominado FECN dentro de la cabecera de transporte *(Base Transport Header, BTH)* presente en todos los paquetes del protocolo. A continuación el mensaje se transmite hacia el nodo destino, según la trayectoria especificada inicialmente.

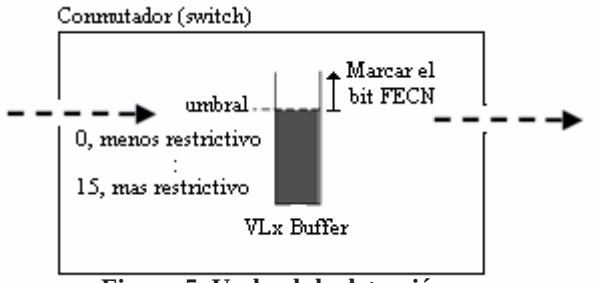


Figura 5. Umbral de detección

La *notificación* de la congestión se realiza una vez que el paquete con el bit de notificación FECN marcado alcanza el adaptador de canal *(CA)* del nodo destino, donde el agente de control de congestión *(CCA)* determina el envío de un mensaje de respuesta al nodo fuente, mediante un paquete de notificación de congestión, como se muestra en la Figura 6. En la cabecera de este paquete, existe un bit denominado BECN que informa explícitamente esta situación. Los nodos fuente que inyectan paquetes en zonas congestionadas de la red de interconexión comenzarán eventualmente a recibir notificaciones desde los nodos destino. En este momento, el agente de control de congestión (CCA) del nodo fuente activa el mecanismo que *configura y selecciona* los caminos alternativos. La apertura de las trayectorias se hace de forma gradual en función de la distribución de notificaciones recibidas (tal que las trayectorias alternativas son seleccionadas de forma inversamente proporcional a dicha distribución, es decir, los caminos menos ocupados son los más utilizados.) y las trayectorias son seleccionadas de manera que su largo no involucre un tiempo de transmisión demasiado grande, por lo que deben ser lo mas cortas posible. De esta manera, se consigue que el problema producido por la contención en los buffers no se traslade hacia una perdida de prestaciones provocada por un tiempo de viaje elevado.

Figura 6. Notificación

Para que la selección de trayectorias múltiples sea posible, el SM asigna un identificador local *(LID)* a cada puerto de la red, como se ha mencionado en la sección 3. El formato de este identificador, situado en las cabeceras de los paquetes, puede verse en la Figura 7. Cuando se reciben los paquetes dentro de un conmutador, los 8 bits menos significativos *(LMC)* son ignorados, de esta manera es posible modificar el valor de esta máscara para asignar a los puertos de los CA varios LIDs (es decir, varios nombres). De esta manera el SM puede establecer varias trayectorias para el mismo nodo, en virtud de la multiplicidad de nombres.

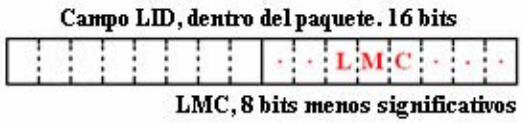


Figura 7. LID y LMC

En la fase inicial de configuración de la red, una vez que se recolecta toda la información de la misma, el gestor de subred entra en la fase de construcción de trayectorias. La especificación no define un algoritmo para cumplir con esta finalidad, por este motivo se ha implementado un mecanismo que genera caminos múltiples y selecciona los disjuntos, para cada par de nodos de la red. El mecanismo utilizado es el típico algoritmo de búsqueda en profundidad *(Depth-first search, DFS)*, con el agregado de una función que selecciona los caminos disjuntos. De esta manera, se configuran los caminos alternativos que serán utilizados cuando aparezca la congestión.

En ausencia de congestión, las trayectorias utilizadas para encaminar los paquetes dentro de la red de interconexión deben ser de largo mínimo con el objeto de ofrecer una baja latencia de transporte,

ya que en este caso, el valor de latencia esta principalmente determinado por la velocidad de los enlaces. La arquitectura especifica que cada agente de control de congestión debe contener un contador que permite medir el tiempo transcurrido desde que el último paquete de notificación ha arribado al nodo. La duración en la que este contador expira, es un parámetro que se establece mediante el gestor de control de congestión. El mecanismo propuesto en este artículo utiliza este contador para contraer las trayectorias en ausencia de congestión. Cada vez que el contador expira y no hayan llegado al nodo destino paquetes de notificación, el camino conformado por las múltiples trayectorias disjuntas se *contrae* gradualmente hasta recuperar la trayectoria original.
Mediante el uso de estas técnicas, se proporciona a la arquitectura InfiniBand, un mecanismo de control de congestión que mejora notablemente los resultados derivados del empleo del mecanismo especificado en el estándar, como se muestra en la sección siguiente.

5 EXPERIMENTACION Y EVALUACIÓN

En esta sección, se evalúa el comportamiento del mecanismo propuesto DRB y su mejora sobre la técnica de control de congestión que ofrece InfiniBand. El modelado de la arquitectura InfiniBand y las técnicas bajo estudio se realizan utilizando la herramienta de simulación estándar OPNET Modeler [8]. Esta herramienta provee un simulador DES *(Discrete Event Simulator)* [8], y ofrece un entorno de modelado jerárquico con técnicas de programación orientada a objetos que permite determinar el comportamiento de los componentes de red en base a la creación de maquinas de estados finitos *(Finite State Models)*. Las métricas más significativas en el estudio de las prestaciones en las redes de interconexión son: la latencia de transporte, que representa el tiempo requerido para entregar un mensaje desde su generación, incluyendo el tiempo en que se almacena en el buffer de salida del nodo fuente y el rendimiento *(throughput)* que representa el tráfico máximo aceptado por la red. Este tráfico se mide en bits/μs y la latencia en milisegundos. Las métricas descriptas presentan el comportamiento global de la red en valores promedio. Por esta razón es también necesario el estudio de la respuesta temporal de los mecanismos analizados con el fin de evaluar parámetros dinámicos como el tiempo de respuesta, la sobrecarga de mensajes en la red, etcétera. Por este motivo, también se ha evaluado la distribución de carga en los enlaces para comprobar las mejoras conseguidas con DRB sobre las características dinámicas de funcionamiento de la red. Los modelos utilizados en la simulación de los nodos de procesamiento y los conmutadores describen el funcionamiento de las capas del protocolo a nivel físico, de enlace y de transporte. También se han modelado los elementos de gestión de subred y de control de congestión que permiten la aplicación de *DRB*. Cada nodo contiene un adaptador de canal, y los conmutadores contienen varios puertos físicos con tres canales virtuales cada uno y el crossbar que permite la interconexión de los puertos de entrada con los puertos de salida. La metodología de evaluación desarrolla varios puntos. El primero consiste en evaluar DRB para un conjunto de redes de interconexión de diversos tamaños (toros y mallas) y de patrones de comunicación. La experimentación se realiza de forma exhaustiva y se enfoca en la respuesta en latencia y *throughput* a patrones de comunicación persistentes tomados de aplicaciones numéricas (*"Butterfly"*, *"Perfect Shuffle"* y *"Matrix Transpose"*) [8]. El segundo punto, consiste en la evaluación de la respuesta de la red de interconexión respecto a un patrón de comunicaciones que provoca la aparición de un "*hot-spot*" con el que se evalúa la respuesta dinámica de la red a través de la utilización de los enlaces congestionados. Debido a la similitud en los resultados obtenidos en todos los casos y por razones de tamaño, solo se presenta un subconjunto representativo de resultados. En la Figura 8 se muestran las prestaciones obtenidas en una red InfiniBand de 64 nodos conectados en una topología toro bidimensional. Los resultados son similares para todos los patrones utilizados. DRB ofrece mejores resultados que la técnica de control IBA y se observa que la diferencia entre las curvas para cada patrón se incrementa a medida que lo hace la carga de tráfico de la red. Se puede observar que a cargas bajas (un ancho de banda menor que 400 bits/μs), las técnicas propuestas se comportan de

forma similar. Esto es importante porque implica que DRB no cambia el comportamiento de la red cuando no es necesario, de manera que no introduce ninguna sobrecarga *(overhead)* con demandas bajas de inyección.

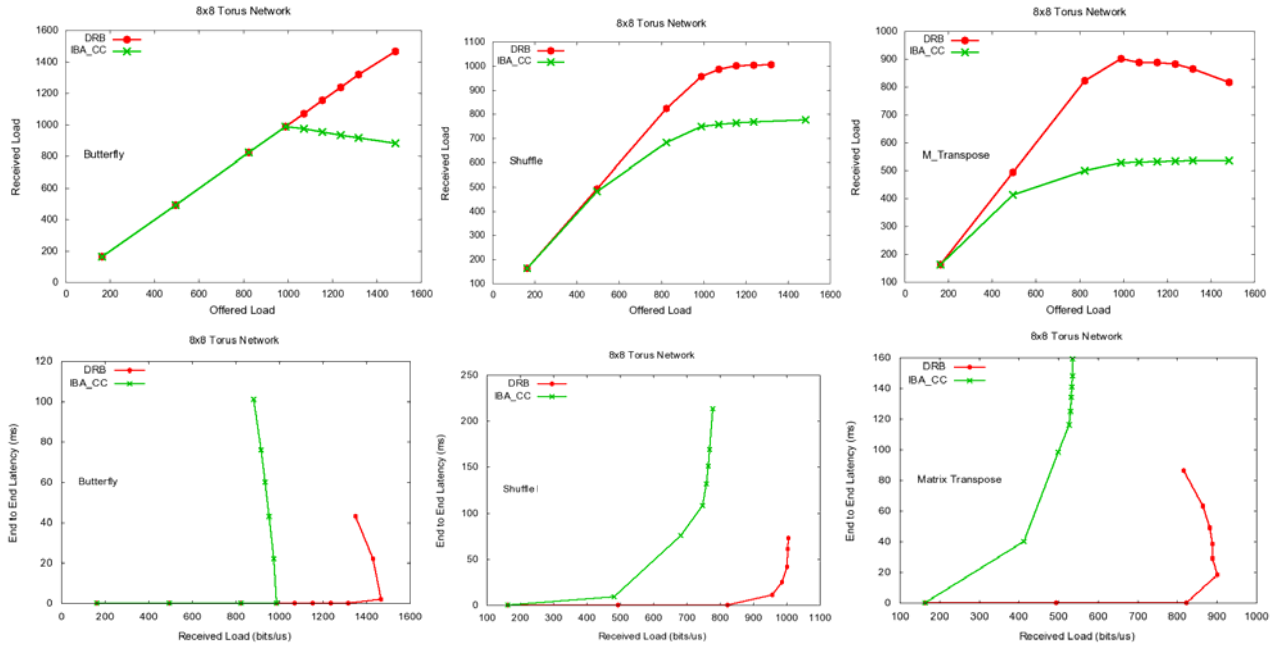


Figura 8. Prestaciones en una red Infiniband de 64 nodos con topología toro. *Throughput y latencia*

Cuando la inyección de trafico en la red se incrementa, con cargas entre 400 y 800 bits/µs, el incremento de la latencia obtenido con DRB es notablemente inferior al que se obtiene con la técnica InfiniBand, esto es debido a que, en DRB, los nodos fuente comienzan a utilizar caminos alternativos para el envío de paquetes, mientras que con IBA_CC los paquetes esperan en el nodo fuente. Con cargas elevadas en la red, para valores mayores a 800 bits/µs en la inyección, DRB utiliza el mayor número de caminos alternativos permitidos en la configuración (en este caso cuatro), resultando en valores de latencia menores respecto a la técnica InfiniBand. Al mismo tiempo que estas latencias se reducen, el *throughput* conseguido se mejora y se observa un incremento notable en la utilización de la red. Este aspecto puede verse en las graficas presentadas en primera fila de la Figura 8 donde se muestra la carga aceptada como función de la carga aplicada. La curva correspondiente a DRB representa una carga mayor que la correspondiente a IBA_CC, donde la red se satura antes y, por tanto, otorga valores más bajos de carga aceptada. Las ganancias con respecto al método original están entre el 35 y el 50%, según el patrón utilizado.

En la Figura 9 se muestran los resultados obtenidos para una red de 32 nodos conectados en una topología malla (o grid). Esta topología, por sus características físicas ofrece menor cantidad de caminos alternativos y por tanto el *throughput* producido es ligeramente peor que en el caso del toro.

Al igual que en el caso anterior, el comportamiento de DRB y el de IBA_CC son similares con cargas muy bajas, pero DRB es mejor en la zona de carga máxima, en la que presenta menores latencias y ofrece mejores prestaciones. Esto significa que DRB es capaz de soportar cargas mayores y se demuestra que, ante condiciones extremas, ofrece mejores prestaciones que el otro método debido a la distribución de caminos utilizada. Esto puede observarse en las dos topologías estudiadas donde el balanceo distribuido del encaminamiento ofrece ganancias entre 200 y 300% en la parte plana de la curva de latencia.

Estas ganancias se deben principalmente a que DRB aprovecha las trayectorias alternativas para el envío de mensajes, mientras que con la técnica IBA_CC deben esperar en el nodo fuente.

Según se ha mencionado anteriormente, hemos diseñado un experimento donde el patrón de comunicación definido provoca la aparición del "*hot-spot*", donde varios mensajes compiten por los

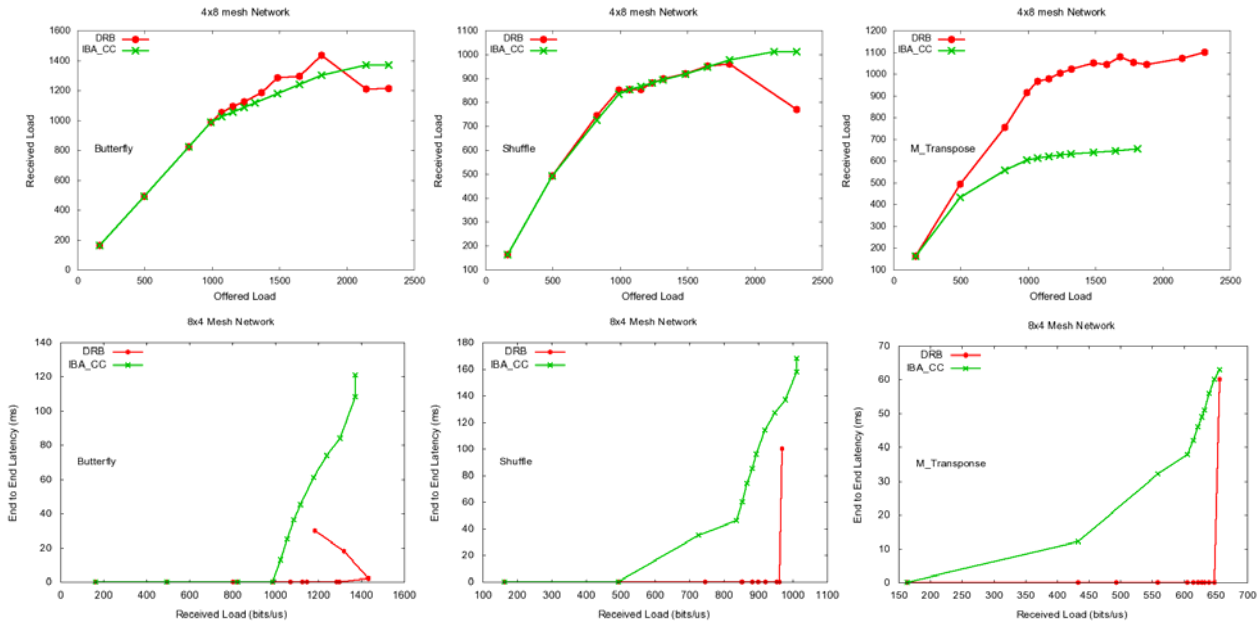


Figura 9. Prestaciones en una red Infiniband de 32 nodos con topología malla. *Throughput y latencia*

recursos sobre un camino común. Este patrón permite analizar y comparar las dos técnicas bajo condiciones extremas de carga aplicada. En la Figura 10 puede observarse el tráfico presente en los enlaces de la red, donde se genera repentinamente una gran carga localizada en una zona concreta de la red.

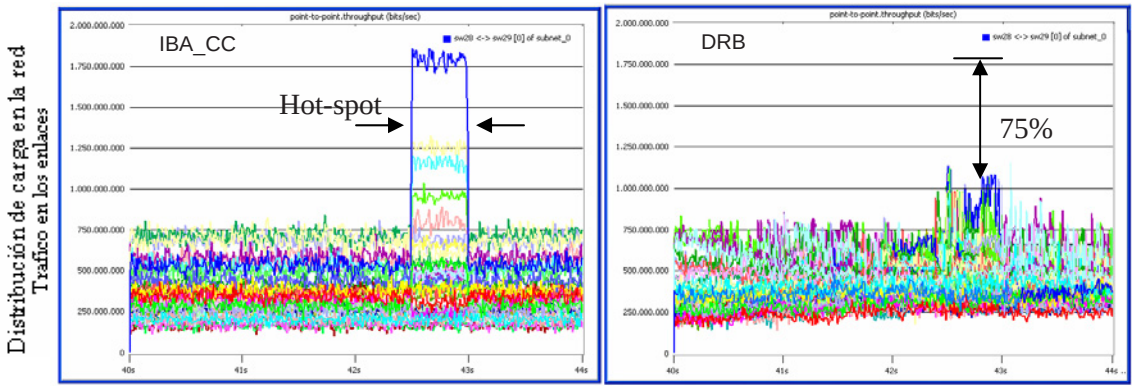


Figura 10. Distribución de carga en los enlaces de la red. Respuesta de los mecanismo al patrón hot-spot

En este análisis puede verse cómo la aplicación del algoritmo DRB mejora los resultados frente a la técnica de control de congestión utilizada por IBA en aproximadamente un *75%*. Cuando se utiliza DRB se eliminan efectivamente los picos de comunicación que se traducen en valores elevados de latencia, debido a que los paquetes deben esperar en los buffers a los que se conectan estos enlaces. DRB ofrece mejores resultados con patrones *"hot- spot"*, que presentan una gran concentración de carga local, ya que es capaz de distribuir el exceso de carga entre los diversos enlaces de la red y balancear de manera eficiente el volumen de comunicación de toda la red de interconexión.

6 CONCLUSIONES

En este artículo se ha propuesto, diseñado y evaluado un nuevo esquema de control de congestión para redes InfiniBand. El mecanismo propuesto elimina la degradación de las prestaciones de la red de interconexión, provocada por la elevada demanda de recursos en una determinada zona de la red. La eliminación del fenómeno de congestión se realiza distribuyendo el tráfico a través de trayectorias alternativas, lo que permite quitar carga en la zona congestionada. A diferencia de las técnicas que utilizan la regulación de mensajes (como la definida por IBA), el balanceo distribuido

del encaminamiento permite mantener la velocidad de inyección de mensajes lo que deriva en un aumento importante de la utilización de la red. Según se ha visto en la experimentación, este aumento está entre el *35* y el *50%* para el *throughput* y entre el *200* y el *300%* para la latencia. El mecanismo basa su funcionamiento en el uso conjunto de dos componentes independientes definidos en el modelo de gestión que especifica la arquitectura IBA. En primer lugar se configura el gestor de subred *(SM)* para que realice el descubrimiento de los componentes y asigne los identificadores locales y las mascaras que posibilitan el establecimiento de varias trayectorias alternativas entre el mismo par de nodos. Por otro lado se utilizan las capacidades de detección y notificación que ofrece el gestor de control de congestión *(CCM)*, y se desactiva la regulación de mensajes. De esta manera cuando se reciben mensajes de notificación, el mecanismo responde balanceando la carga en los enlaces de la red, mediante la selección de trayectorias alternativas. Debido a que ambos componentes de gestión están definidos en InfiniBand, nuestra propuesta es totalmente compatible y no requiere que se modifique ningún aspecto de la especificación. En el futuro tenemos planeado experimentar con diferentes patrones de tráfico y topologías más complejas, a fin de evaluar exhaustivamente a DRB y mejorar sus características intentando refinar la técnica de marcado de paquetes y el mecanismo de selección de trayectorias, teniendo en cuenta la sobrecarga provocada por los mensajes de notificación.

REFERENCIAS

[1] A Singh, WJ Dally, B Towles, AK Gupta 'Globally Adaptive Load-Balanced Routing on Tori', Computer Architecture Letters, IEEE, 2004.

[2] A. Vishnu, M. Koop, A. Moody, A. R. Mamidala, S. Narravula, D. K. Panda, "Hot-Spot Avoidance With Multi-Pathing Over InfiniBand: An MPI Perspective," ccgrid, pp. 479-486, 7º IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07), 2007

[3] Baydal, E., 'A Family of Mechanisms for Congestion Control in Wormhole Networks', IEEE Trans. Parallel Distrib. Syst. 16(9), pp.772--784. 2005

[4] Duato, J. Johnson, I. Flich, J. Naven, F. Garcia, P. Nachiondo, T. "A new scalable and cost-effective congestion management strategy for lossless multistage interconnection networks" High-Performance Computer Architecture, HPCA-11. 2005

[5] Franco, D.; Garcés, I. & Luque, E., 'Avoiding Communication Hot-Spots in Interconnection Networks', in 'HICSS '99: Proceedings of the Thirty-second Annual Hawaii International Conference on System Sciences-Volume 8', IEEE Computer Society. 1999

[6] G. Pfister et al. "Solving Hot Spot Contention Using Infiniband Architecture Congestion Control". *Ion HPI-DC*, 2005.

[7] 'InfiniBand Architecture Specification' (v. 1.2), InfiniBand Trade Association. Disponible en: http://www.InfiniBandta.com/, 2004

[8] '*Opnet Modeler Accelerating Network R & D'* OPNET Technologies, Inc.,at http://opnet.com

[9] Santos, J.R.; Turner, Y.; Janakiraman G. 'End-to-end congestion control for infiniband' Infocom 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE, 2003

[10] Shihang Yan, Geyong Min, Irfan Awan, "An Enhanced Congestion Control Mechanism in InfiniBand Networks for High Performance Computing Systems," aina, pp. 845-850, 20th International Conference on Advanced Information Networking and Applications. AINA 2006

[11] *'Top500 Supercomputers Site'*, at http://www.top500.org. Consultado el 19/07/2007.

[12] William Dally, Brian towles. Principles and practices of interconnection networks. Morgan Kaufmann publishers. 2004.

[13] Xuan-Yi Lin; Yeh-Ching Chung; Tai-Yi Huang**. '**A multiple LID routing scheme for fat-tree-based InfiniBand networks'**.** Parallel and Distributed Processing Symposium, Proceedings. 18th International, pp. 26-30, 2004.