UNIVERSIDAD DE CANTABRIA

DPTO. DE ELECTRÓNICA Y COMPUTADORES.

IMPACTO DEL SUBSISTEMA DE COMUNICACIÓN EN EL RENDIMIENTO DE LOS COMPUTADORES PARALELOS: DESDE EL HARDWARE HASTA LAS APLICACIONES.

Presentada por:

Valentin Puente Varona

Dirigida por:

Ramón Beivide Palacio.

SANTANDER, OCTUBRE DE 1999

Capítulo 7

Conclusiones y Perspectivas Futuras

7.1 Conclusiones y Principales Aportaciones del Trabajo.

A continuación pasaremos a comentar las principales conclusiones y aportaciones que se desprenden del trabajo expuesto en esta memoria. Entre las más importantes podemos citar las siguientes:

1. Acerca de la metodología de evaluación.
   A lo largo de todo el trabajo hemos propuesto y empleado una metodología compleja que ha permitido conocer con una precisión razonable cuáles son las restricciones que impone el *hardware* en el rendimiento de la red de interconexión. Por otro lado, partiendo del diseño *hardware* de la red hemos podido valorar cuál es el impacto que tiene este subsistema en el redimiendo global de los computadores paralelos. Por ello, podemos considerar que las conclusiones extraídas de los estudios realizados en este marco son bastante próximas a las que obtendríamos del análisis de un hipotético sistema real. Para ser más concretos, respecto a este punto podemos concluir lo siguiente:

1.(a). Acerca de las restricciones del *Hardware* subyacente.

A lo largo de todo el trabajo expuesto, ha quedado clara la necesidad de incorporar este tipo de aspectos si se pretende analizar de forma justa el rendimiento de la red de interconexión. A este respecto, aparecen multitud de ejemplos que muestran cómo encaminadores con un buen comportamiento desde el punto de vista estructural pero de elevada complejidad, pierden sus ventajas al tener en cuenta el coste *hardware*. Por lo tanto, consideramos imprescindible saber, al menos de forma aproximada, cómo son las características de bajo nivel de las implementaciones e incorporar estos datos en la simulaciones para garantizar que los resultados relativos al rendimiento real de la red sean fiables. Como hemos expuesto, la metodología empleada permite saber, de forma bastante precisa, cuál es la complejidad de los encaminadores y, aunque existen multitud de factores que pueden alterar en mayor o menor medida sus resultados, la aproximación que nos aporta es suficientemente válida.

1.(b). Acerca de la demanda que ejercen las cargas reales sobre la red de interconexión.

En cada uno de los encaminadores o redes estudiadas, hemos considerado necesario evaluar la influencia real de la red de interconexión en las aplicaciones paralelas. Para alcanzar este objetivo hemos desarrollado una infraestructura de simulación que permite considerar las características *hardware* de los encaminadores en el contexto de los sistemas ccNUMA. Sobre este sistema es posible analizar una variedad de aplicaciones numéricas que permiten abarcar un campo de aplicación importante de los computadores paralelos. Los resultados muestran que, aunque no necesariamente todas las ganancias observadas en tráfico sintético se trasladan al caso de las aplicaciones reales, la tendencia observada sí es bastante similar. En el caso concreto del sistema empleado, se ha constatado que es preciso ser extremadamente cuidadoso a la hora de proponer nuevas arquitecturas para la red, ya que para que las aplicaciones mejoren su rendimiento es necesario balancear adecuadamente la relación latencia-*throughput*. Se ha demostrado, en este tipo de sistemas, cómo ambas figuras de mérito de la red son cruciales.

2. Acerca de la adaptatividad: ventajas e inconvenientes.
   Hemos propuesto un nuevo mecanismo de evitación de interbloqueo en redes toroidales con encaminamiento adaptativo. Los resultados obtenidos para esta propuesta en el marco de trabajo establecido, muestran que es posible emplear este tipo de encaminamiento sin incrementar de forma acusada el coste del encaminador. De hecho, hemos comprobado como el coste *hardware* de nuestra propuesta es menor que el obtenido para encaminadores adaptativos basados en otros mecanismos de evitación que requieren un mayor número de canales virtuales. La solución aportada llega a acercarse incluso al coste de algunos encaminadores deterministas. Por lo tanto, el empleo de la adaptatividad con este mecanismo permite ampliar considerablemente el margen de trabajo de la red sin afectar a la latencia en cargas bajas. Posteriormente, hemos analizado de modo comparativo su rendimiento en el caso de cargas reales en el contexto de un sistema ccNUMA. Si bien las mejoras logradas no son elevadas, en la mayoría de los casos se logra aumentar el rendimiento de las aplicaciones consideradas. Aunque el tipo de sistemas empleado en este análisis es muy sensible a la latencia, nuestra propuesta logra mejorar el tiempo de ejecución de las aplicaciones mientras que otros encaminadores adaptativos basados en otras propuestas implican una bajada de rendimiento considerable. Estos resultados refuerzan el hecho de que es necesario plantear arquitecturas para los encaminadores en las que la relación entre la latencia y el *throughput* sea equilibrada.

3. Acerca de la toma de decisiones relativas a la implementación de los encaminadores.
   Hemos comprobado de qué manera pueden afectar las decisiones tomadas a la hora de implementar el encaminador. Hemos evaluado cómo influyen este tipo de cuestiones en el caso del encaminador adaptativo propuesto. Hemos analizado de qué manera pueden afectar tres diferentes sistemas de arbitrio en el rendimiento del encaminador. En este estudio se ha introducido un nuevo sistema de arbitraje de extraordinaria sencillez y lo hemos comparado con un sistema empleado comúnmente y con otro de elevado rendimiento desde el punto de vista funcional pero de alta complejidad. La principal conclusión extraída es que esta clase de decisiones afecta sensiblemente al rendimiento de la red en términos estructurales y puede llegar a influir considerablemente en términos tecnológicos. En función de la complejidad de la implementación, las diferencias en el tiempo de ciclo de los encaminadores son acusadas. Hemos constatado cómo este factor es uno de los que más puede afectar al rendimiento del sistema. En el caso concreto de los sistemas ccNUMA, se ha comprobado para las alternativas consideradas, cómo este tipo de decisiones puede afectar en más de un 30% al tiempo de ejecución de algunas aplicaciones. Bajo este punto de vista, ha quedado claro que, a la hora de implementar un encaminador, es preciso ser extremadamente cuidadoso sobre la complejidad que se incorpore en su diseño.

4. Acerca de la influencia de la topología de la red en su rendimiento.
   Hemos estudiado de qué manera puede afectar la topología de la red al rendimiento de las aplicaciones paralelas. En este punto hemos analizado topologías comunes como la malla y el toro. Además, hemos incorporado en el análisis una topología nunca analizada en un contexto realista: la *Midimew.* En este caso, hemos logrado proponer soluciones eficaces contra el interbloqueo y la complejidad de encaminamiento, que en el pasado, desaconsejaban su empleo. Hemos estudiado de qué manera afecta la propia topología a la complejidad de la implementación *hardware* de los encaminadores y la conclusión final es que empleando mecanismos de evitación de interbloqueo como los propuestos en este trabajo, el impacto es muy reducido, por no decir despreciable. Por último, hemos analizado el rendimiento de cada topología bajo condiciones de carga real en un sistema ccNUMA. La conclusión extraída es que la nueva red propuesta logra mejorar, de forma apreciable, el rendimiento del sistema.

5. Acerca de la organización interna de los encaminadores y del impacto de la evolución de las tecnologías de implementación.
   Hemos considerado de qué manera pueden afectar las mejoras en la tecnología VLSI en el diseño del encaminador. En este caso, hemos implementado encaminadores que poseen estructuras complejas de realizar con tecnologías desfasadas. En concreto, hemos analizado diversos mecanismos de evitación de *HLB* basados en el empleo de múltiples canales virtuales o almacenamiento en la salida. En este estudio, hemos implementado dos encaminadores con bufferes multipuerto en la salida. Para que la implementación de estos elementos no tenga un coste desmesurado hemos recurrido al empleo de una tecnología de implementación reciente. Los resultados muestran como es posible mejorar el rendimiento de la red de interconexión de forma considerable al emplear estas estructuras. Hemos considerado mas organizaciones para los espacios de almacenamiento del encaminador y la conclusión definitiva es que el encaminador adaptativo propuesto con buffers en la salida alcanza un rendimiento superior a cualquiera de los demás. Los resultados muestran como los mecanismos de evitación del *HLB* pierden eficacia al emplearlos en encaminadores deterministas. Sin embargo, el uso de mecanismos de encaminamiento adaptativo sí permite explotar adecuadamente la evitación del *HLB.*

6. Acerca de la mejor alternativa para el subsistema de comunicación.
   A la vista del estudio de los diversos aspectos relacionados con la red de interconexión que se ha realizado en esta tesis, es posible extraer varias conclusiones en este sentido. En primer lugar, a la hora de abordar estudios de determinados subsistemas de los computadores y más en concreto en el caso que nos ocupa, es necesario ser conscientes y evaluar con una precisión adecuada cuáles son los costes *hardware* de las propuestas arquitectónicas que se

puedan realizar. Además, es preciso tomar en cuenta los continuos avances en la tecnología VLSI: las tecnologías más recientes permiten abordar problemas desde puntos de vista que no necesariamente tenían que ser adecuados en tecnologías previas. En cuanto al rendimiento, se ha observado la importancia de las dos figuras de mérito de la red de interconexión: latencia y *throughput*. Ambas son importantes para mejorar el rendimiento del sistema total. En esta línea, hemos comprobado como la complejidad de los encaminadores puede influir de forma acusada en el rendimiento por lo que es preciso proponer soluciones que mejoren la productividad de la red, pero sin complicar en exceso los routers. A la vista de los resultados alcanzados, la propuesta definitiva estaría constituida por un encaminador adaptativo con el mecanismo de evitación de interbloqueos propuesto y con un planificador interno para el encaminador lo más sencillo posible. Los espacios de almacenamiento en ese encaminador habrían de estar situados, sin lugar a dudas, en los puertos de salida, siendo preciso para ello emplear tecnologías de implementación evolucionadas. Por ultimo, sería muy recomendable emplear una topología para la red de interconexión como la *Midimew*. El aunar todas estas características mejoraría considerablemente el rendimiento del susbsistema de comunicación y por tanto el sistema global se vería considerablemente favorecido.

Para ir finalizando, cabría destacar como aportaciones inéditas y originales de esta tesis las que se citan a continuación:

a. Propuesta de un nuevo mecanismo de conmutación que, de forma simple, permite evitar interbloqueos con encaminamiento adaptativo.

b. Propuesta de un nuevo mecanismo de arbitraje distribuido que permite reducir el tiempo de ciclo de los routers.

c. Aplicación realista de ambas propuestas en redes topológicamente ricas para las cuales no existían realizaciones previas.

d. Propuesta de nuevas soluciones estructurales en el diseño de los routers basadas en la ubicación y gestión de bufferes.

e. Empleo de una metodología exhaustiva de evaluación que parte del diseño a nivel *hardware* y llega hasta el nivel de aplicación, permitiendo cuantificaciones fiables y robustas.

Versiones preliminares relacionadas con varias partes de este trabajo se han expuesto en diversos foros. Algunas de ellas son:

- V. Puente, J.A. Gregorio,J. M. Prellezo, R.Beivide, J. Duato, and C. Izu, ``Adaptive Bubble Router: a Design to Balance Latency and Throughput in Networks for Parallel Computers", in Proc. of International Conference on Parallel Computing (ICPP'99), Sept. 1999.
- V. Puente, J.A. Gregorio, C. Izu, R. Beivide and F. Vallejo, " Low-level Router Design and its Impact on Supercomputer System Performance", International Conference on Supercomputing (ICS'99), June 1999.
- V. Puente, J.A. Gregorio, C. Izu and R. Beivide , "Impact of the Head-of-Line Blocking on Parallel Computer Networks: Hardware to Applications", Europar'99, Sept. 1999.
- V. Puente, C. Izu, J.A. Gregorio, R. Beivide, J. M. Prellezo and F. Vallejo, "Rearranging Links to Improve the Performance of Parallel Computers: The Case of Midimew Networks", En proceso de revisión.
- V. Puente, B. Torón, J.A. Gregorio, R. Beivide, "Plataformas Paralelas Basadas en Redes Myrinet", VII Jornadas de Paralelismo, Septiembre de 1997.
- J.M. Prellezo, V. Puente, J.A. Gregorio y R. Beivide., "SICOSYS: Un Simulador de Redes de Interconexión para Computadores Paralelos", VIII Jornadas de Paralelismo, Septiembre de 1998.
- V. Puente, J.M. Prellezo, C. Izu, J.A. Gregorio and R. Beivide, "A Case Study of Trace-driven Simulation for Analyzing Interconnection Networks: cc-NUMAs with ILP Processors", Aceptado en el 8th Euromicro Workshop on Parallel and Distributed Processing, January 2000.
- V. Puente, J.A. Gregorio, C. Izu, R. Beivide, J.M. Prellezo, J. Vinuesa and F. Vallejo, "Effects of Hardware Design Decisions on the Performance of Real Applications", IX Jornadas de Paralelismo, Septiembre de 1999.
- J.Vinuesa, R. Menendez, V. Puente, B. Toron, "Parallel Paradigms Applied in a Fluid-Dynamic Problem to Model a Glass Manufacturing Process". 3rd International Meeting on Vector and Parallel Processing (VECPAR) June 1998.

7.2 Perspectivas y Líneas de Trabajo Abiertas.

A la vista de las conclusiones extraídas del desarrollo de este trabajo, a continuación pasaremos a detallar cuáles son las perspectivas de trabajo a medio-largo plazo en orden de complejidad y tiempo.

1. Como se ha comprobado a lo largo de los diferentes capítulos del trabajo, uno de los factores más críticos del sistema, cuando tomamos en cuenta las restricciones *hardware*, es el tiempo de ciclo del encaminador. Un aspecto directamente relacionado con esta figura de mérito es la segmentación del encaminador. Es por ello interesante analizar de forma autocontenida cuál es el número de etapas adecuado para explotar al máximo el rendimiento de la red. De acuerdo con esto, el primer objetivo a perseguir es analizar, desde las implementaciones *hardware* hasta las propias aplicaciones, cuál es el número de etapas más apropiado. Para realizar esta tarea, es necesario determinar, de forma precisa, cual es el *overhead* que introduce la propia segmentación en la latencia del encaminador y en qué medida las mejoras de productividad, introducidas por la reducción en el tiempo de ciclo, pueden mejorar el rendimiento del sistema completo.
2. En todo el trabajo presentado no se ha prestado demasiada atención a aspectos relacionados con el interconexionado físico de los encaminadores. Bajo determinadas condiciones es posible que las restricciones que puede llegar a imponer la distribución espacial de la propia red de interconexión pueden ser más graves que las impuestas por la estructura de los encaminadores. En este sentido sería muy interesante estudiar cómo este tipo de factores puede influir sobre el rendimiento de la red de interconexión.
3. Pese a haber introducido una metodología completa en el análisis de la red, contando con cargas reales, solo hemos analizado una clase de sistemas. Sería interesante ampliar el espectro, analizando el comportamiento de aplicaciones y sistemas de paso de mensajes en lugar de sistemas de memoria compartida. En la misma línea, sería conveniente abordar el estudio de sistemas más grandes que los considerados en este trabajo, así como tamaños de problema más realistas. En esta tarea será preciso relajar o plantear soluciones que reduzcan el coste computacional del sistema a simular. Algunas alternativas podrían pasa por intentar paralelizar el propio simulador o relajar el cociente coste computacional-precisión.
4. Sería de alto interés analizar el comportamiento de la red de interconexión bajo aplicaciones no numéricas. Es conocido que actualmente el rango de utilización de los sistemas paralelos tiende a acercarse cada vez mas a problemas de *data mining*, análisis de riesgos y bases de datos en general, entre otros. Por ello, consideramos oportuno conocer la demanda de estas aplicaciones hacia la red de interconexión. Otro aspecto interesante en la línea de cargas de

trabajo realistas, es saber en qué grado es posible modelar su comportamiento en términos de estructuras matemáticas apoyadas en la teoría de Caos. Actualmente existen algunos investigadores en el área de arquitectura de computadores que están comenzando a plantear evaluaciones de rendimiento de distintos subsistemas de los computadores desde el punto de vista de estas teorías. Desde un punto de vista personal, este tipo de ideas resulta altamente atractivo dado que incorporan una componente teórica muy interesante a un problema tan complejo como la evaluación de los computadores paralelos.

5. Por último, desde nuestro punto de vista parece oportuno abordar el estudio de la red de interconexión de forma unificada con respecto al resto de la jerarquía de memoria del sistema. Bajo esta aproximación es posible tener en cuenta las peculiaridades de toda la jerarquía de memoria a la hora de proponer nuevas soluciones para la red de interconexión, logrando así adecuarla a las necesidades reales del sistema. De la misma forma, es posible plantear modificaciones en el sistema superior atendiendo a las características de la red de interconexión.