Balanceo Predictivo y Distribuido del Encaminamiento (PR-DRB)

Carlos Núñez[1], Diego Lugones[1], Daniel Franco[2], Emilio Luque[2]

Departamento de Arquitectura de Ordenadores y Sistemas Operativos, Universitat Autònoma de Barcelona, España.
[1]{carlos.nunez, diego.lugones}@caos.uab.es
[2]{daniel.franco, emilio.luque}@uab.es

Abstract. El desbalance en la carga de comunicaciones puede congestionar la red de interconexión, incrementando la latencia y disminuyendo el throughput, degradando el rendimiento total del sistema paralelo. Las aplicaciones paralelas contienen etapas representativas durante su ejecución las cuales permiten caracterizarlas y obtener un patrón de comunicaciones. Este trabajo presenta el Balanceo Predictivo de Encaminamiento Distribuido (PR-DRB), un nuevo método desarrollado para controlar la congestión en la red basado en la expansión de caminos, la distribución de tráfico y carga efectiva, para mantener una latencia baja. PR-DRB monitorea la latencia de los mensajes en los encaminadores, elige los caminos alternativos a utilizar y registra la información de la congestión en base al patrón de comunicaciones detectado, para luego volver a aplicar la mejor solución cuando dicho patrón se repita. Experimentos de tráfico con congestión fueron llevados a cabo para evaluar el rendimiento del método.

Keywords: Redes de interconexión, encaminamiento predictivo, encaminamiento adaptativo, latencia uniforme, aplicaciones paralelas, computación de alto desempeño.

1 Introducción

El comportamiento de aplicaciones científicas, que se ejecutan en paralelo sobre una red de interconexión de alto rendimiento (HSIN), puede describirse como una colección de procesos asignados implícitamente a cada procesador. El coste de las comunicaciones y su consumo energético es mayor respecto al de los procesadores y debe ser tratada por los mecanismos de control de congestión [1]. Una distribución de carga inapropiada (HotSpot) genera puntos de saturación como si toda la red esté colapsada. Para paliar esta situación se han desarrollado mecanismos de control de congestión que mejoran el throughput utilizando una cantidad apropiada de recursos [1], como las técnicas de encaminamiento adaptativas que dinámicamente manejan los recursos a fin de reducir la congestión. Este trabajo presenta el Balanceo Predictivo de Encaminamiento Distribuido (PR-DRB), una nueva estrategia de encaminamiento adaptativo basada en la utilización de caminos alternativos ante la

This research has been supported by the MEC-MICINN Spain under contract

presencia de congestión a fin de mantener y mejorar el ancho de banda disponible y reducir la latencia global para los patrones de comunicación repetitivos. Esta mejora se realiza a través del uso de información histórica sobre el patrón de comunicaciones que ocasionó la congestión, y servirá como base a toma de decisiones futuras.

PR-DRB se basa en el DRB presentado en [2] pero mejora las prestaciones con el módulo predictivo de monitorización y registro. El modelo propuesto consta de tres fases: monitorización, detección de la congestión y patrón conflictivo, y el control de la congestión. Este trabajo se basa en la repetitividad de etapas fundamentales en aplicaciones paralelas. En la fase de monitorización se registra el valor de latencia de un mensaje y se guarda información del patrón de tráfico que ocasionó la congestión. Cuando el mensaje llega al nodo destino, se notifica al origen de la situación a través de un mensaje de notificación (ACK), y el nodo origen es capaz de proceder a la apertura de caminos alternativos en base a la latencia. Si esta situación ya ha sido analizada previamente, se aplica la mejor solución aprendida y se actualiza su base de datos de históricos apropiadamente.

Cuando se detecta la congestión, el algoritmo DRB va adaptándose a través de la apertura de caminos alternativos, hasta que encuentra un valor de latencia global apropiado para todos los caminos abiertos, y en este proceso DRB invierte un tiempo considerable. Este trabajo propone almacenar la solución óptima encontrada y re aplicarla cuando se detecte una situación similar a la que originó la congestión inicial.

El artículo está organizado de la siguiente manera: La sección 2 presenta los trabajos relacionados, la sección 3 detalla la metodología del PR-DRB. La experimentación se describe en la sección 4, y resultados y trabajos futuros en la sección 5.

2 Trabajos Relacionados

El control de congestión se basa en la monitorización, detección y el posterior control. Para evaluar la congestión, generalmente son utilizados la latencia punto a punto [3], el nivel de ocupación de los buffers [4] o el “backpressure” [1], que una vez evaluados permiten a los nodos de cómputo disminuir y controlar la congestión.

Message Throttling [5] es una técnica de corrección que notifica al origen de la congestión a fin de parar o disminuir la inyección de nuevos paquetes hasta mejorarse la situación. Esto controla la utilización de los buffers, pero aumenta la latencia de los mensajes debido a la espera en el nodo origen hasta que se controle la congestión.

Otras técnicas se basan en la administración de buffers, exclusivamente en los encaminadores [4], pero no consiguen un buen rendimiento ya que el nodo origen no es notificado. Existen técnicas de control de congestión basadas en algoritmos de encaminamiento adaptativo [2], [6], [7], que utilizan trayectorias alternativas para inyectar los mensajes. La distribución de trayectorias se realiza en forma dinámica de acuerdo al estado actual de la red y pueden cambiar con el tiempo. Algunas desventajas serían la sobrecarga a la red debido a la monitorización, garantizar la ausencia de interbloqueos y la entrega de mensajes en orden. Esta situación requiere una solución de compromiso entre la velocidad de monitorización y la cantidad de

información a analizarse. Por consiguiente, un algoritmo de encaminamiento eficiente debe ser capaz de obtener el mejor comportamiento ante una situación adversa, a fin de evitar introducir penalizaciones en las comunicaciones.

Estudios del patrón de comunicaciones de las aplicaciones paralelas de HPC que son ejecutadas en las HSIN, demuestran que la mayoría poseen un comportamiento repetitivo y está enmarcado por cómputo y comunicaciones [8].

En las HSIN, el desempeño del encaminamiento depende en gran medida del patrón de comunicaciones utilizado y de la relación con el mapeo de nodos a procesadores de una aplicación. Esto obliga al uso apropiado de recursos en redes HPC donde el costo global de los componentes es prohibitivo [9]. Algunas estrategias de encaminamiento utilizan información de las aplicaciones como ser la tasa de transferencia para determinar mejores rutas que minimicen la latencia, número de flujos por enlace, ancho de banda, deadlocks, etc., pero de manera estática [10].

Para mejorar el rendimiento de las comunicaciones, y aplicaciones, debe hallarse una técnica que combine las soluciones de los algoritmos adaptativos y el patrón de comunicaciones utilizado, a fin de que el encaminamiento y control de congestión apliquen rápidamente las soluciones guardadas minimizando la monitorización

3 PR-DRB. Predecir el comportamiento del tráfico.

PR-DRB busca mejorar el tiempo de respuesta del algoritmo DRB utilizando información histórica de las comunicaciones. PR-DRB usa el concepto de Meta-Caminos como alternativas a las trayectorias iniciales para enviar mensajes. La configuración de los Meta-Caminos define como se crean los caminos alternativos y cuáles de todos ellos son elegidos ante una situación de congestión. Las fases del PR-DRB se muestran en la siguiente figura: **Fig. 1 (a)**. Detección y registro del valor de latencia y del patrón de comunicaciones que ocasionó la congestión. **Fig. 1 (b)**, muestra la Configuración de Meta-Caminos y la **Fig. 1 (c)** muestra caminos alternativos que forman el Camino Multi-Paso (MSP).

La fase de monitorización conlleva la medida del valor de latencia de un mensaje hasta su destino final. La congestión se detecta en los encaminadores intermedios al superar un umbral de latencia máximo, se registra la latencia, el patrón de comunicaciones conflictivo, y el origen/destino involucrados. En el nodo destino, se envía un ACK al origen con la información registrada a lo largo del trayecto.

La cantidad de caminos alternativos a usar viene dada por el valor de latencia registrada, a efectos de distribuir todos los mensajes por estos caminos.

PR-DRB a través de su fase de Configuración de Meta-Caminos usa un esquema de tres pasos, llamado MSP, para encontrar el camino hacia el destino, partiendo del origen hacia un nodo intermedio 1, del nodo intermedio 1 al nodo intermedio 2, y desde éste último hacia el destino. Luego se actualiza la información histórica con los nuevos valores de latencia y de caminos alternativos abiertos para este caso. Así se logra que se registren los mejores caminos para cada par origen-destino específico, bajo una congestión dada, y que más adelante puedan ser aplicadas directamente acelerando y simplificando el proceso de Configuración de Meta-Caminos.

En la fase de Selección de Caminos Multi-Paso el nodo origen inyecta los mensajes en base al MSP calculado en la fase de Configuración de Meta-Caminos.

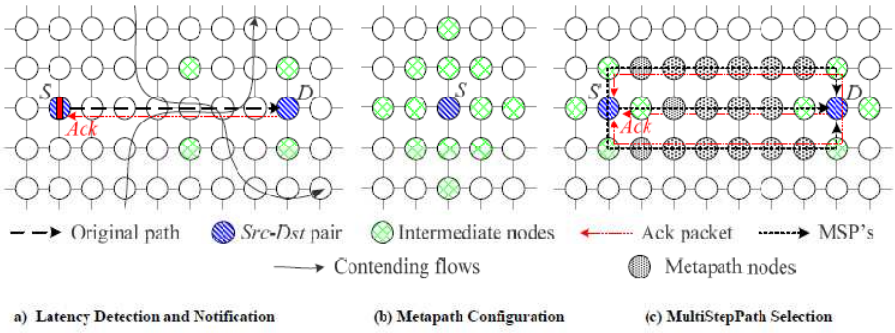


Fig. 1. Fases del Algoritmo PR-DRB.

3.1 Fase de Monitorización y Notificación de la Congestión

Cada encaminador en la red es el encargado de la monitorización del tráfico que lo atraviesa. Además, el encaminador registra la información que más adelante el nodo destino enviará en un ACK al origen. La **Tabla 1** resume la monitorización.

Tabla 1. Fase de Monitorización y Notificación.

Monitorización de Mensajes (Mensaje M, Umbral, MSP)
/* En cada encaminador PR-DRB */
Begin
Por cada mensaje M en cada salto,

1. **If** (Primer Encaminador en el trayecto)
   - Predictivo = FALSE;
2. Acumular Latencia (tiempo de espera en la cola)
   If (Latencia > Umbral) AND NOT (Predictivo)
   - Identificar flujos que intervienen en la congestión
   - Registrar patrón de comunicaciones de la congestión (origen/destino).
   - Predictivo = TRUE.
3. Registrar en el mensaje la latencia acumulada en el enrutador actual.
4. Re enviar mensaje M hacia el siguiente encaminador intermedio o el nodo destino.
5. En el destino, la latencia y el patrón registrados es enviado al origen.

End Monitorización y Notificación

La latencia de contención es el tiempo que un mensaje debe esperar en los buffers del encaminador antes de continuar hacia su destino, debido al bloqueo sometido por otros mensajes que también ocupan el buffer. Esta latencia es incrementada en todos los encaminadores por donde atraviese el mensaje. Ante una contención, el encaminador guarda la información de los nodos origen/destino que estén ocupando el mismo buffer, para determinar flujos que colisionan. El registro del patrón de

comunicaciones solo en el primer encaminador se debe a que la apertura de nuevos caminos alternativos ya disminuirá la congestión para el flujo analizado.

Una vez en el destino se procede a enviar un ACK al origen, así nuevos mensajes inyectados para el mismo destino puedan ser enviados por caminos alternativos. Un ACK tiene mayor prioridad en el encaminamiento, y su tamaño podría considerarse despreciable, ya que solo transfiere información de latencia y de flujos que colisionaron durante una contención. El registro y notificación se lleva a cabo en cada encaminador independientemente con información local, incluyendo a otros mensajes actualmente en los buffers, lo que contribuye a un conocimiento global de la red.

3.2 Configuración de los Meta-Caminos

La configuración dinámica de los meta caminos se basa en la información registrada durante la monitorización. El objetivo principal de esta fase es la de determinar, para cada para origen/destino, una cantidad apropiada de caminos alternativos en base al valor de latencia global registrada durante todo el trayecto. En la creación de caminos alternativos se seleccionan nodos intermedios (INs) disjuntos al camino original, y se tienen en cuenta nuevos valores de latencia registrados. La congestión es controlada ampliando el ancho de banda a través de apertura de caminos alternativos. La **Tabla 2** resume la Configuración de Meta-Caminos. La **Fig. 2.** da una visión general del esquema de trabajo del PR-DRB. Durante la etapa 1 de la aplicación paralela, se

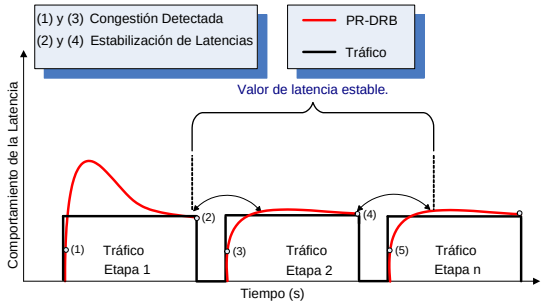


Fig. 2. Procesos del PR-DRB

observa que PR-DRB tiene una latencia elevada. Esto se debe a que el algoritmo está abriendo trayectorias alternativas para mantener el valor de las latencias controlada. Una vez que la cantidad de caminos alternativos para un par fuente/destino se ha estabilizado, entonces el algoritmo predictivo procede a guardar la mejor solución. En la etapa 2 esta solución se utilizará directamente al identificar una situación de congestión similar a la ya encontrada en la etapa 1. Se observa que en las posteriores etapas de la aplicación, la latencia no llega a los picos registrados en la primera, esto se debe a que PR-DRB ha inyectado los mensajes a través de los mejores caminos, evitando de esta manera congestión en trayectorias no óptimas. Así mismo cabe destacar que la zona de trabajo de PR-DRB a lo largo de toda la ejecución está enmarcada en el área de latencias estables.

Tabla 2. Función de Configuración de Meta-Caminos

Configuración de Meta-Caminos.
/* Ejecutado en el nodo origen cada vez que un mensaje ACK es recibido */
Begin
- Recibir el ACK con la Latencia del MSP y flujos involucrados en la congestión.
- Calcular la latencia de los Meta-Caminos.

$$Latencia\ (MP*) = \left(\sum Latencia\ (MSPs)^{-1}\right)^{-1}$$

- **If** $(Latencia\ (MP*) > Umbral)$
  If (Ya existe solución guardada para el MSP para el par origen/destino)
  - Buscar en la Base de Datos el mejor MSP registrado.

  Else
  - Incrementar número de INs para proveer nuevos caminos alternativos.

  End If
  Else If (Latencia (Mp*) < Umbral)
  - Decrementar el número de INs para disminuir el Meta-Camino.
  - Guardar Latencia del (MP *) y los nodos origen/destino involucrados.

  End If

End Configuración de Meta-Caminos.

3.3 Fase de Selección de los Caminos Multi-Paso

Al momento de la inyección de un nuevo mensaje, se tiene en cuenta el valor de latencia acumulada para determinar qué camino alternativo utilizar en cada instante de tiempo, logrando de esta manera distribuir apropiadamente la carga de comunicaciones no solamente con la ampliación del ancho de banda efectivo sino que también se utilizan con mayor frecuencia los mejores caminos. La expansión de caminos es llevada a cabo en forma gradual, pudiendo tomar un tiempo para lograr encontrar el valor apropiado de caminos a utilizar y la distribución de carga para cada uno de ellos. Dado un nodo origen con N caminos alternativos, definimos Lci (i: 1...N) como la latencia registrada para el camino Ci. El camino alternativo Cx será seleccionado para la próxima inyección en base a la probabilidad:

$$\rho(Cx) = \frac{1/L_{Cx}}{\left(\sum_{i=1}^{N} 1/L_{Ci}\right)} \quad (eq.1)$$

Encontrar la mejor combinación de cantidad de caminos/ancho de banda puede llevar un tiempo considerable, y las aperturas intermedias hasta la óptima también introducirán una contención en los encaminadores y el posterior incremento de latencia. PR-DRB va guardando la información de las aperturas y actualizando su base de datos de mejores caminos abiertos y el porcentaje de utilización de cada uno de ellos como un atributo al par origen/destino, a efectos de volver a aplicarlos directamente cuando se detecte la congestión, y así evitar extender los picos de latencia al buscar la mejor solución posible. La **Tabla 3** resume la fase de Selección de Caminos Multi-Paso.

Tabla 3. Fase de Selección de Caminos Multi-Paso.

Selección de Caminos Multi-Paso
/* Ejecutado en el nodo origen antes de inyectar un nuevo mensaje */
Begin
– Construir la función de distribución acumulativa y normalización de ancho de banda de los MSP.
– Seleccionar un MSP utilizando la función de distribución acumulativa.
– Inyectar el mensaje en la red
– Armar encabezados con información de destinos intermedios y final.
– Concatenar los encabezados.
– Inyectar el mensaje con el formato PR-DRB.
End Selección de Caminos Multi-Paso.

3.4 Integración de todas las fases.

El esquema general de funcionamiento de PR-DRB se muestra en la **Fig. 3**. Cuando se genera por primera vez desde el origen un mensaje para un origen/destino específico, este se inyecta directamente a la red, debido a que aún no se tienen estadísticas sobre la situación de la red. A partir de este punto el mensaje va hacia su destino atravesando múltiples encaminadores intermedios, donde se evalúa el valor de la latencia sufrida en cada encaminador y se determina si supera un umbral máximo a efectos de analizar el patrón de tráfico que está ocasionando dicha congestión, y de esta manera guardar información que lo identifique. En el nodo destino, se procede a enviar al nodo fuente la información registrada a través de la inyección de un paquete especial de notificación (ACK), el cual contiene la latencia acumulada en todos los encaminadores por donde atravesó el mensaje así como la información del patrón de comunicaciones que causó la congestión. Cuando el mensaje arriba al nodo origen, la fase de configuración de Meta-Caminos se lleva a cabo y se guardan los valores de latencia así como del flujo que ocasionó la congestión, junto con el par origen/destino del mensaje. Con esta información actualizada, la fase de configuración de Meta-Caminos puede analizar la apertura o cierre de nuevos caminos alternativos en caso que la latencia global de la red aún no esté en una situación estable. Cuando se requiera inyectar un nuevo mensaje en la red, la fase de Selección de Caminos Multi-Paso toma el control y procede a seleccionar una de las trayectorias alternativas disponibles. Como las aplicaciones paralelas presentan repetitividad en las fases de su ejecución, los patrones de comunicación tienden a repetirse en el tiempo. Para este caso concreto, la fase de Configuración de Meta-Caminos puede simplificarse a la tarea de buscar el mejor camino registrado durante la primera etapa de la ejecución de la aplicación paralela, con lo cual se aplica directamente la mejor solución encontrada y se ahorra considerablemente en tiempo de cálculo y comunicaciones. Aplicaciones que requieran mantener el orden de los paquetes pueden usar el *algoritmo de ventanas deslizantes*, como es utilizado en [6].

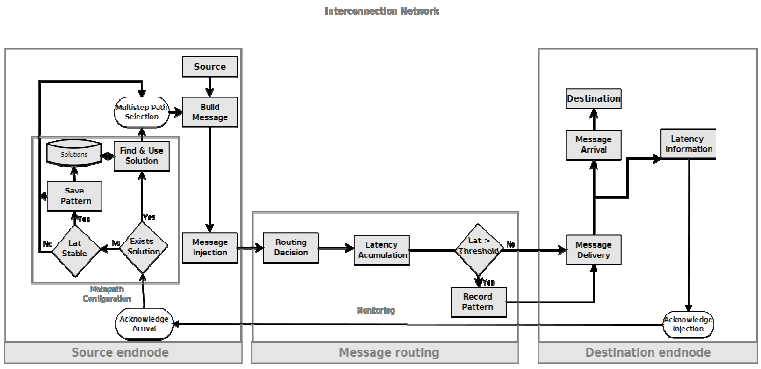


Fig. 3 Algoritmo PR-DRB con todas las fases integradas

4 Experimentación.

Se ha analizado el comportamiento de PR-DRB frente al DRB original [2], que ya ha sido comparado satisfactoriamente con otros de la literatura, para verificar la ganancia que se puede obtener con el módulo predictivo. Se analiza cómo la *latencia* mejora al aplicar las políticas predictivas, así como el *throughput* que no se ve perjudicado por dicha mejora, y se presenta el mapa de latencias.

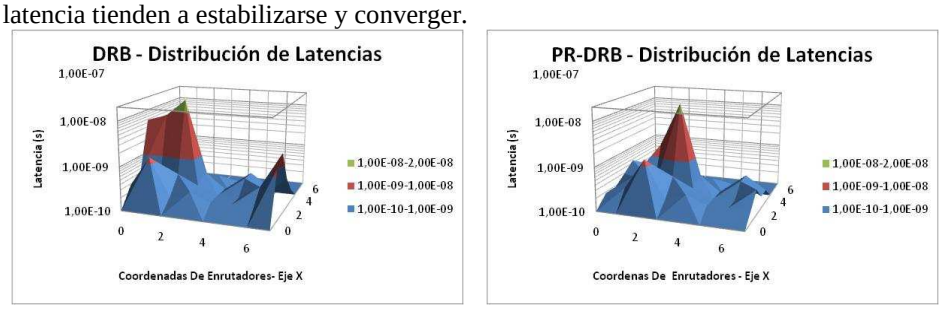
El tráfico es a ráfagas y con situaciones de HotSpot, para evaluar el comportamiento transitorio y distribución de tráfico en situaciones extremas de carga. Esta configuración establece nodos específicos como destino a fin de conseguir áreas saturadas. Se inyecta tráfico en otras áreas a efectos de que los caminos alternativos iniciales encontrados no sean óptimos. El modelo fue implementado utilizando la herramienta de simulación y modelado OPNET modeler [11].

Simulaciones fueron llevadas a cabo en una red de 64 nodos con una topología Toro 2D. El control de flujo fue el Virtual Cut-through, tamaño de paquete de 1024 bytes y ancho de banda de los enlaces a 2 Gbps.

4.1 Análisis de HotSpot.

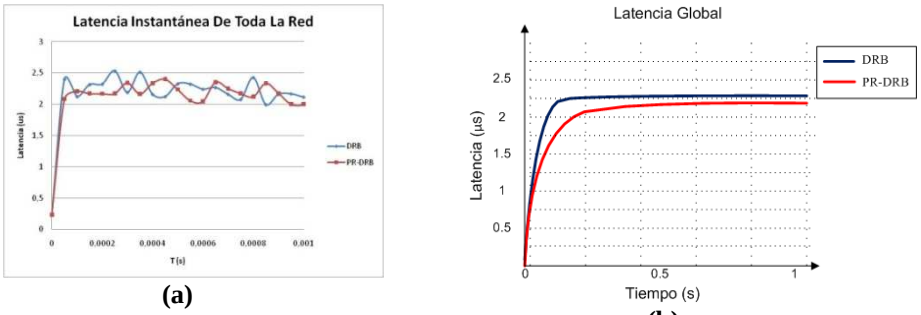
Fig. 4 (a) y (b) muestran la superficie de latencias de la red, para el algoritmo DRB y la propuesta PR-DRB. Se muestra la latencia de contención promedio a través del mapa de latencias, donde cada punto (x,y) representa la latencia de un encaminador. La **Fig. 4 (a)** corresponde al DRB, donde se observa un pico en el comportamiento de latencias en la zona congestionada y que la distribución de carga entre los encaminadores (x,y) 0,1, 6,2 y 6,4 son elevadas debido a que el DRB utiliza estas trayectorias como caminos alternativos. En la **Fig. 4 (b)** se observa el resultado del PR-DRB, donde se puede apreciar que el pico de latencia es inferior al del DRB. También puede observarse que la distribución de carga es mejor que en el caso de DRB, debido a que PR-DRB ha aplicado directamente las mejores soluciones y evitó

sobrecargar encaminadores hasta encontrar la solución. En **la Fig. 5(a)** se muestra la latencia durante los primeros instantes de la ejecución a efectos de analizar las consecuencias de la inyección de tráfico y reacción de los algoritmos. En promedio el algoritmo PR-DRB tiene mejor desempeño, ya que llega a mejores valores de latencia global y en menor tiempo hasta estabilizarse, y las mejoras en la latencia no han introducido ninguna penalización en el throughput. La latencia a lo largo de toda la simulación puede observarse en la **Fig. 5 (b).** El algoritmo DRB sufre un incremento al inicio, debido a las aperturas de trayectorias alternativas hasta el instante de tiempo 0.5, donde prácticamente los valores convergen. PR-DRB por el contrario exhibe un comportamiento mejor en su etapa inicial ya que ha evitado las trayectorias innecesarias y ha aplicado las mejores soluciones encontradas cuando vuelve a repetirse el patrón conflictivo, con lo que el valor promedio de ganancia de latencia puede reflejarse entre el tiempo 0 y 0.5. A partir de este instante, ambos valores de latencia tienden a estabilizarse y converger.



(a) **(b)**

Fig. 4. Mapa de Latencias



(a) **(b)**

Fig. 5. Latencia de toda la red, instantánea (a) y promedio (b)

5 Conclusiones y Trabajos Futuros

Este trabajo ha propuesto el Balanceo Predictivo y Distribuido del Encaminamiento PR-DRB, para las HSIN. Esta estrategia pretende utilizar caminos alternativos ante la presencia de congestión, para mejorar valores de latencias y disponibilidad de ancho de banda, en el menor tiempo posible y dinámicamente. Las aplicaciones paralelas

presentan un comportamiento repetitivo, y PR-DRB es capaz de aprender y re aplicar las mejores soluciones encontradas cuando se detecte el mismo patrón. PR-DRB ha sido diseñado con HSIN para clusters en mente, donde la latencia debe ser uniforme bajo cualquier carga de tráfico. Experimentos muestran que la propuesta aumenta la ganancia de latencia, sin afectar el throughput. Como continuación de este trabajo se pretende predecir y evitar una futura congestión analizando la tendencia de latencias y re aplicando las soluciones guardadas. También se propone caracterizar y obtener patrones de aplicaciones paralelas para hacer al PR-DRB application aware.

6 Referencias.

1. Baydal, E., Lopez, P., Duato, J.: A Family of Mechanisms for Congestion Control in Wormhole Networks. IEEE Trans. Parallel Distrib. Syst. 16(9), 772-784 (2005)
2. Franco, D., Garcés, Luque, E.: A new method to make communication latency uniform: distributed routing balancing., pp.210-219 (1999)
3. Lugones, D., Franco, D., Luque, E.: Dynamic and Distributed Multipath Routing Policy for High-Speed Cluster Networks., pp.396-403 (2009)
4. Garcia, P. J., Quiles, F. J., Flich, J., Duato, J., Johnson, I., Naven, F.: RECN-DD: A Memory-Efficient Congestion Management Technique for Advanced Switching. Parallel Processing, International Conference on 0, 23-32 (2006)
5. Yan, S., Min, G., Awan, I.: An Enhanced Congestion Control Mechanism in InfiniBand Networks for High Performance Computing Systems. Advanced Information Networking and Applications, International Conference on 1, 845-850 (2006)
6. Singh, A., Dally, W., Towles, B., Gupta, A.: Globally Adaptive Load-Balanced Routing on Tori. IEEE Comput. Archit. Lett. 3(1), 2 (2004)
7. Glass, C., Ni, L.: The turn model for adaptive routing. SIGARCH Comput. Archit. News 20(2), 278-287 (1992)
8. Wong, A., Rexachs, D., Luque, E.: Parallel application signature. Cluster Computing and Workshops, 2009. CLUSTER '09. IEEE International Conference on 1, 1-4 (2009)
9. Rodriguez, G., Beivide, R., Minkenberg, C., Labarta, J., Valero, M.: Exploring pattern-aware routing in generalized fat tree networks., pp.276-285 (2009)
10. Kinsy, M., Cho, M., Wen, T., Suh, E., Dijk, M., Devadas, S.: Application-aware deadlock-free oblivious routing., pp.208-219 (2009)
11. OPNET Technologies: Opnet Modeler Accelerating Network R&D. (2008)
12. Wu, X., Sun, X.-H.: Performance Modeling for Interconnection Networks. High-Performance Computing in the Asia-Pacific Region, International Conference on 1, 380 (2000)
13. Sherwood, T., Perelman, E., Calder, B.: Basic Block Distribution Analysis to Find Periodic Behavior and Simulation Points in Applications., pp.3-14 (2001)