Impacto de la Entrada/Salida en los Computadores Paralelos*

Sandra Mendez, Dolores Rexachs y Emilio Luque

Departamento de Arquitectura de Computadores y Sistemas Operativos
Universidad Autónoma de Barcelona, España

{sandra.mendez}@caos.uab.es, {dolores.rexachs,emilio.luque}@uab.es

Resumen El aumento de las unidades de procesamiento en los clústers, los avances en velocidad y potencia de las unidades de procesamiento y la creciente complejidad de las aplicaciones científicas demandan mayores exigencias a los sistemas de Entrada/Salida de los computadores paralelos. En este trabajo se propone una metodología para el análisis de E/S en los clústers de computadores, que permita analizar cómo afectan las diferentes configuraciones a la aplicación y usarla para seleccionar la mejor configuración del sistema de E/S. La metodología contempla la caracterización del sistema de E/S a distintos niveles: dispositivo, sistema y aplicación; configuración de diferentes elementos que tienen impacto en las prestaciones y evaluación teniendo en cuenta tanto la aplicación como la arquitectura de E/S.

Keywords: E/S Paralela, Arquitectura de E/S, Librerías de E/S, Almacenamiento Masivo

1. Introducción

El aumento de las unidades de procesamiento en los clústers, los avances tanto en velocidad como en potencia de las unidades de procesamiento y la creciente complejidad de las aplicaciones científicas que utilizan cómputo de altas prestaciones demandan mayores exigencias a los sistemas de Entrada/Salida (E/S) de los computadores paralelos. En muchos casos, debido al *gap* que existe entre las prestaciones del cómputo y el sistema de E/S, éste se vuelve el cuello de botella de los sistemas paralelos. Para poder disminuir el *gap* se deben identificar los factores que influyen en las prestaciones. Esto lleva a plantear las siguientes preguntas: ¿Las aplicaciones deben adaptarse a la E/S? ¿El diseñador del sistema de E/S debe mejorar su rendimiento adaptándose a las aplicaciones de forma transparente? ¿Qué factores de E/S influyen en el rendimiento?

Responder a estas preguntas no es trivial. Las aplicaciones tienen diferentes comportamientos y si bien los programadores o diseñadores pueden realizar las modificaciones necesarias para realizar eficientemente las operaciones de E/S, estas modificaciones son específicas para una aplicación y computador paralelo.

* Este trabajo ha sido subvencionado por el MEC (España), proyecto TIN 2007-64974

Por otro lado, sacar mayores prestaciones al sistema de E/S requiere que el programador sea un experto en los detalles de los sistemas de E/S.

La idea es conocer el sistema de E/S para utilizarlo adecuadamente desde el punto de vista de prestaciones, para esto se debe entender: como funcionan los dispositivos de almacenamiento para determinar que límites imponen a las prestaciones, la conexión con el sistema paralelo, la gestión a través del sistema de fichero y las aplicaciones. Esto puede ayudar a identificar los múltiples e importantes factores que influyen en las prestaciones del sistema de E/S.

Por esta razón, es importante medir el impacto de la E/S en los computadores paralelos. Esto es útil para poder definir qué configuración del sistema de E/S es la más conveniente de acuerdo al tipo de aplicación y arquitectura del computador paralelo. En este trabajo se propone una metodología para el análisis de E/S en los clústers de computadores, que permita examinar cómo afectan las diferentes configuraciones a la aplicación y usarla para seleccionar la mejor configuración del sistema de E/S.

Este trabajo se estructura en las siguientes secciones, en la sección 2 se presentan los trabajos relacionados, en la sección 3 se discuten los componentes del sistema de E/S que permiten establecer el contexto en el que se hace la evaluación, en la sección 4 se introduce la metodología propuesta, en la sección 5 se presentan los experimentos realizados y, finalmente, en la sección 6 las conclusiones y trabajos futuros son reportados.

2. Estado del Arte

Existen varios estudios realizados para evaluar las prestaciones de la E/S de los computadores paralelos con intención de caracterizarla y poder mejorar sus prestaciones. La tendencia actual es trabajar tanto en entornos reales como en entornos simulados a patir de trazas. Dado que las prestaciones del sistema de E/S depende del software y hardware estos estudios son realizados para configuraciones de E/S de computadores paralelos específicos.

En [1] se realiza una caracterización de la E/S de las aplicaciones del supercomputador *Jaguar* formado por máquinas *CRAY*. En este trabajo se presenta una herramienta que permite generar trazas de las operaciones de E/S para aplicaciones MPI. A partir de estas trazas se pueden identificar las zonas que pueden ser optimizadas a nivel de E/S. Siguiendo con el estudio en las maquinas CRAY, en [2] se realiza un estudio de las prestaciones centrándose en la sintonización de los parámetros. El mismo equipo en [3] presenta un análisis de las prestaciones a nivel del sistema de almacenamiento, sin hacer cambios en la configuración física, realizan cambios a nivel de sistema de ficheros y evalúan las prestaciones para la aplicación S3D. En [4] se presenta un trabajo donde dan un orden al análisis que realizan, considerando los factores de número de clientes, tamaño de *stripe* y número de *stripe*. Estos estudios se realizaron sobre *CRAYs* para mejorar las prestaciones del sistemas de E/S.

El trabajo presentado en [5] realiza un análisis de prestaciones centrándose en la configuración de la E/S con la intención de demostrar la eficiencia del sistema de E/S del supercomputador Red Storm. En este se presenta un estudio de los

límites teóricos de la arquitectura de E/S y realizan pruebas de prestaciones para *file-per-process* y *shared-file*, sin considerar a la aplicación.

Una evaluación de las prestaciones de la E/S del supercomputador Blue Gene/L en [6] muestra una arquitectura de E/S de archivos escalable.

Por otro lado, una herramienta necesaria para evaluación y cuantificación de los cuellos de botella de la E/S son las herramientas de simulación.

En [7] se presenta una herramienta de simulación que permite modelar y usar una simulación orientada a trazas para evaluar la performance del subsistema interno de E/S paralelo de la arquitectura del computador paralelo Vulcan MPP, que ha permitido hacer una estimación cuantitativa del ratio nodo de computo – nodo de E/S y la potencial escalabilidad de la arquitectura.

En [8] se describe un simulador SIMCAN, que permitirá estudiar el comportamiento de ambientes distribuidos complejos con varios propósitos, la detección de los cuellos de botellas del sistema, cálculo del grado de escalabilidad del sistema y probar la prestaciones de aplicaciones sin usar el sistema real.

3. Sistema de E/S Paralelo

Para poder realizar un análisis de prestaciones del sistema de E/S es necesario concretar que aspectos se consideran. Para este trabajo, se consideran dos nivels a analizar, la Aplicación y la Arquitectura de E/S Paralela.

3.1. Aplicación Científica

Los patrones de acceso de las aplicaciones paralelas pueden ser clasificados en locales y globales. Los patrones locales están determinados por el proceso (*thread*), mostrando como un fichero es accedido por un proceso local. Los patrones globales son sobre la aplicación paralela, representado por los múltiples procesos accediendo a un archivo. En [9] se hace una clasificación de los patrones de acceso en cinco dimensiones para un proceso local. Las cinco dimensiones son espacial (puede ser contiguos o no o una combinación de ambos), tamaño de solicitud (este puede ser pequeño (*small*), mediano (*medium*) y grande (*large*), el tamaño de una solicitud puede ser fija o variable), comportamiento repetitivo (se presenta cuando los bucles o una función con bucles tiene operaciones de E/S), temporal (éstos capturan la regularidad de las ráfagas de E/S en una aplicación, pueden ocurrir tanto periódica como irregularmente) y tipo de operación de E/S (pueden ser de *write only*, *read only* y *read/write*).

La manipulación de estos patrones de E/S se puede hacer de forma más eficiente con el uso de librerías de E/S. Éstas son interfaces proporcionadas a nivel de software. Las librerías de E/S de bajo nivel soportan los accesos a archivos paralelos e implementan las diferentes técnicas de E/S. Por ejemplo, MPI-IO permite utilizar operaciones colectivas de I/O para accesos discontiguos. Las librerías de alto nivel, también conocidas como librerías de datos científicos, gestionan los datos a un nivel de abstracción del usuario. Por ejemplo: HDF5 permite definir los datos en un formato a nivel de usuario.

3.2. Arquitectura de E/S Paralela

La arquitectura de los computadores paralelos debe proporcionar un balance entre cómputo, ancho de banda, capacidad de memoria y E/S. Las arquitecturas

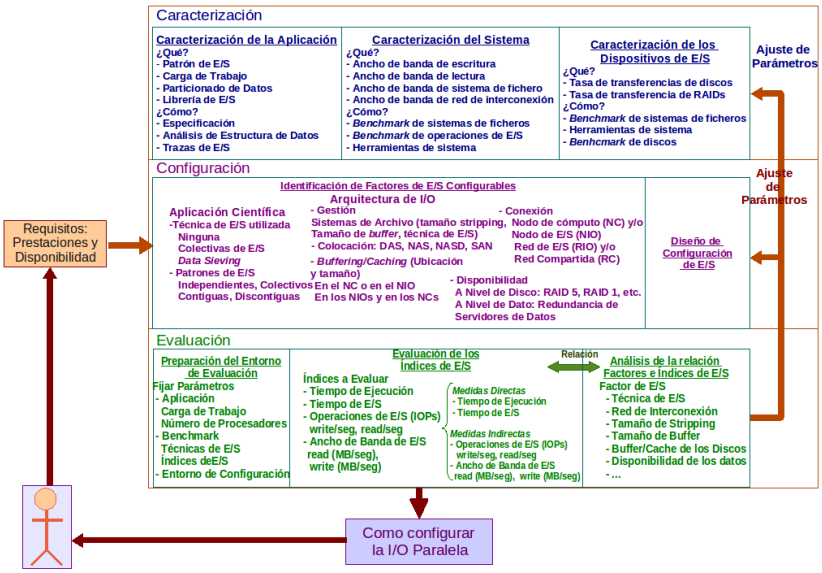

Figura 1. Metodología para Configurar la E/S

para los computadores paralelos se basan en un sistema de E/S interno, acompañado con una colección de nodos de E/S dedicados, cada uno gestionando y proporcionando accesos a un conjunto de discos. Los nodos de E/S están conectados a otros nodos por una red de interconexión que puede se la misma que utilizan los nodos de cómputo o una red específica de E/S.

A continuación se presenta la arquitectura de E/S paralela siguiendo el esquema propuesto por D. Kotz [10]:

- Conexión: Considera la conexión de dispositivos (a nodos de cómputo o a nodos específicos de E/S) y la red de interconexión que conecta los nodos. Existen dos modelos básicos para la conexión. En el primero, cada nodo de cómputo está conectado a su propio disco local o a discos compartidos. En el segundo existen dos clases de nodos: nodo de cómputo y nodo de E/S. Dependiendo de la posición de los nodos de E/S, el subsistema de E/S puede estar integrado con el sistema de cómputo o ser independiente. En cuanto a la red de interconexión, una opción es conectar los nodos de E/S, o incluso lo dispositivos de E/S, directamente a la red principal de interconexión. Otra opción es proporcionar una red dedicada a la E/S.
- Gestión: Una cuestión clave es, ¿Qué procesadores gestionan el acceso a los ficheros? Hay tres soluciones comunes, donde la gestión puede ser: centralizada en un procesador, distribuida entre todos los procesadores y distribuida entre un subconjunto de procesadores que estan dedicados a la E/S, ejemplos de ficheros paralelos son Lustre, GPFS, PVFS, pNFS y un sistema distribuido usado en cluster pequeños o medianos es NFS.
- Ubicación: La posición de los dispositivos de E/S en la topología de la red pueden tener un impacto significativo en el rendimiento del sistema de E/S.

Son 4 las forman de acceder a los datos de acuerdo a como estan ubicados los dispositivos de almacenamiento y como los gestiona: DAS (*Direct Attached Storage*), SAN (*Storage Area Network*), NAS (*Network Attached Storage*) y NASD (*NetworkAttached Storage Devices*).

- Buffering/Caching: El buffering y caching es un factor importante en cualquier sistema de E/S para compensar las diferencias de velocidad y la diferente granularidad (bloques o paquetes), y las ráfagas dadas por las características de los dispositivos o la carga (congestión de la red). Es importante su uso (lectura o lectura y escritura) y ubicación en función de como se comportan las aplicaciones.
- Disponibilidad: Un computador paralelo esta formado por varios componentes, usados en paralelo para mejorar la performance. La distribución de los datos sobre varios dispositivos aumenta el rendimiento pero decrementa la fiabilidad. La falla en los discos pueden ser enmascarada con los sistemas RAID. La falla en los nodos de E/S pueden ser enmascarado con la redundancia.

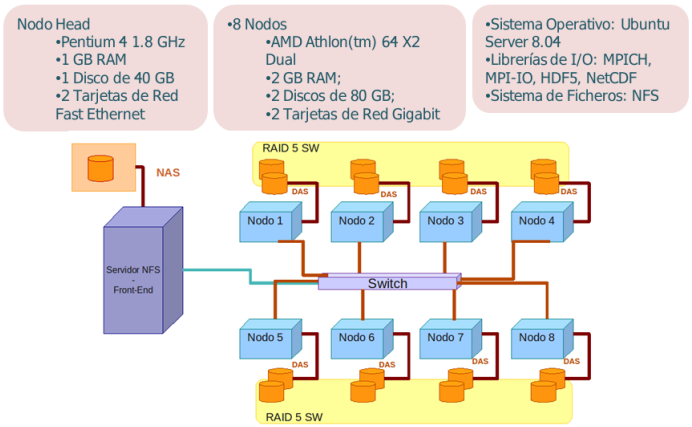

Figura 2. Estructura del Cluster Aohyper

4. Metodología para configura la Configuración de E/S

La metodología propuesta consta de 3 fases : Caracterización, Configuración y Evaluación. Un factor fundamental a considerar son los requisitos del usuario, debido a que será el usuario el que proporcione las restricciones para diseñar el sistema de E/S (Figura 1). La fase de caracterización esta destinada a obtener los factores de la Aplicación, Sistema de Paralelo y Dispositivos de E/S que brinden la información para poder realizar la fase de configuración.

En la fase de configuración se identifican los factores de E/S sobre los que es posible actuar y que se usarán para diseñar el sistema de E/S. Los factores de E/S configurables se extraen de los componentes del sistema de E/S.

De los parámetros identificados se diseña una configuración de E/S coherente.

En la fase de evaluación se analizan las prestaciones para una aplicación y diferentes configuraciones de E/S para determinar qué y cómo los factores están

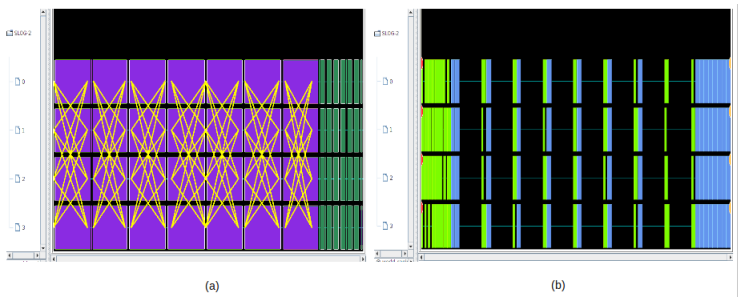

Figura 3. Traza de los *Benchmark* (a) NAS BT-IO para 4 procesos donde el color morado es para escritura y verde para lectura (b) MADBench2 para 4 procesos donde verde es para escritura y celeste para lectura

afectando a la aplicación. Una vez que la aplicación es ejecutada se debe analizar los índices de E/S y relacionarlos con los factores de E/S.

5. Experimentación

Los experimentos se realizaron siguiendo la metodología propuesta, como aplicación a analizar se eligen los *Benchmark NAS-BT-I/O* [11] y *MADBench2* [12] que son kernels de aplicaciones científicas. Los experimentos se realizaron en un clúster formado por 8 nodos de cómputo dual core y un nodo de E/S que también cumplía el rol de nodo *front-end*. En la figura 2 se muestra la estructura del clúster y las descripción técnica. Debido a que BT-IO y MADBench2 requieren un número cuadrado de procesadores, las pruebas se hicieron para 4 y 9 procesos.

5.1. Caracterización de los Benchmarks

En la Figura 3 se presentan las trazas de ambos *benchmark* para 4 procesos en ésta se pueden observar los distintos patrones de E/S. NAS BT-IO es una extensión del benchmark NAS BT, incluye los requisitos de E/S de BT. BTIO presenta un patrón de particionado tridiagonal en una arreglo de tres dimensiones en un número cuadrado de procesos. Se uso como workload la clase B que requiere un espacio de almacenamiento de 1,5 GB y usa la librería MPI-IO.

MADBench2 es un derivado del código de análisis de datos MADspec. Como parte de sus cálculos, el *MADspec* realiza operaciones sobre una matriz que no cabe en memoria , requeriendo sucesivas escrituras y lecturas de una gran cantidad de datos contiguos de archivos compartidos o individuales. Las funciones S y W realizan escrituras, y las funciones W y C realizan escrituras. Los valores para los parámetros son: número de PIX 5000, BIN 8, GANG 1, tamaño de bloque 4096, RMOD 1 y WMOD 1.

5.2. Caracterización del Sistema Paralelo y los Dispositivos de Almacenamiento

Se caracterizaron los discos de cada nodo a nivel local con los *benchmark* *Bonnie++* e *Iozone* muy utilizados en la literatura. Para tamaño de archivos de 4GB, Los valores medios para la tasa de trasferencia para Bonnie++ fueron para escritura de 55 MB/seg y para lectura de 63 MB/seg, en tanto que Iozone reporto para escritura 53 MB/seg y lectura de 63 MB/seg. Para el sistema de fichero NFS se realizaron mediciones para un archivo de 512 MB evaluando el tiempo de transferencia para realizar la copia del archivo por NFS. Esto se hizo desde un nodo de cómputo al nodo *front-end* reportando 8,3 MB/seg, desde un nodo de cómputo en RAID 5 por software que reporto 4,6 MB/seg y a nivel local en un nodo de cómputo reportando 57,2 MB/seg.

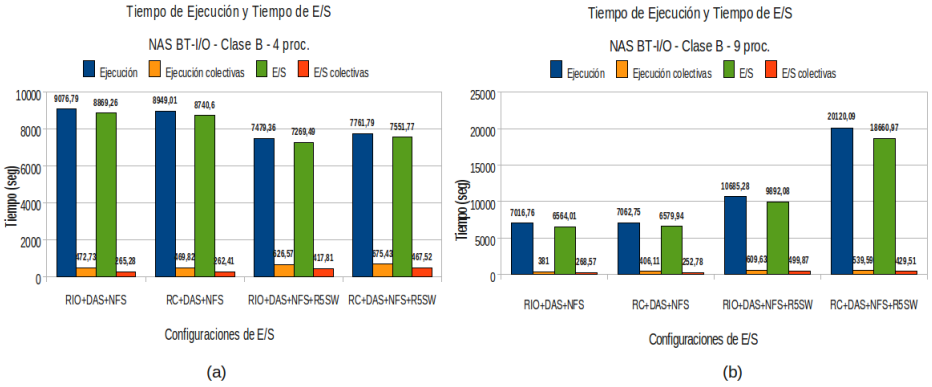

Figura 4. Tiempos de ejecución y de E/S del NAS BTIO con NFS y DAS

5.3. Configuración

Los factores de E/S identificados como configurables a nivel de arquitectura de E/S fueron el sistema de fichero NFS, red de interconexión (red de E/S (RIO) y red compartida (RC)), disponibilidad disco único o RAID 5 por software (R5SW), técnica de E/S (con y sin colectivas de E/S).

A nivel de aplicación se ejecuto la aplicación con y sin colectivas de E/S para el NAS BT-IO y el tipo de acceso a los archivos para *MADBench2*. En total se probaron los *benchmarks* para 4 configuraciones de la arquitectura y dos de aplicación obteniendo un total de 8 configuraciones para cada *benchmark*.

5.4. Evaluación NAS BT-IO

La ejecución del NAS BT-IO se hizo sobre 8 variación en la configuración, en la Figura 4 se presentan el tiempo de ejecuciòn y el tiempo dedicado a la operaciones de E/S para cada una de las configuraciones. Vemos que el tiempo dedicado a la E/S domina. Los resultados muestran que el uso de librerías que permiten el uso de operaciones colectivas de E/S mejora las prestaciones de forma significativa pasando los tiempos ejecución para todas las configuraciones de E/S del 90 % al 30 %.

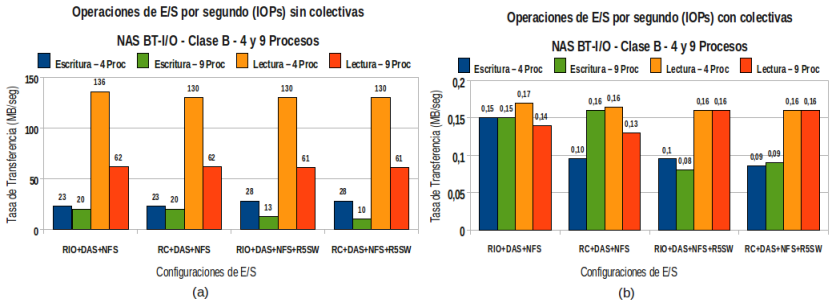

Figura 5. Operaciones de E/S para el NAS BT-E/S

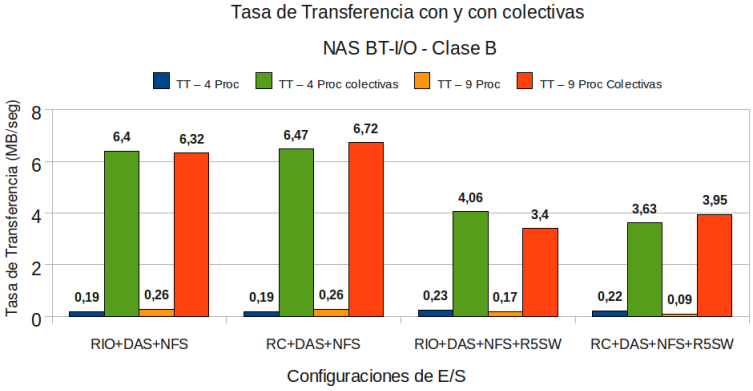

Figura 6. Tasa de Transferencia para el NAS BT-E/S con y sin colectivas de E/S

Partiendo del comportamiento de la aplicación vemos que se realizan muchas operaciones de E/S (Figura 5) lo que hace que el disco penalice mucho, sin embargo,cuando se utilizan operaciones colectivas uniendo peticiones de diferentes procesos para acceder al disco, se logra mejorar el impacto del disco, reduciendo el número de accesos y el tamaño del acceso (Figura 6). Desde el punto de vista de utilización de RAID, vemos que cuando no se usan colectivas, para muchas solicitudes de acceso pequeñas es mejor utilizar un RAID 5, sin embargo, cuando se realizan accesos colectivos, el RAID 5 no ayuda mejorar las prestaciones, sólo mejora la disponibilidad. En este caso también influye tener una red dedicada a la E/S.

5.5. Evaluación MADBench2

En la Figura 7 se presentan el ancho de banda de las operaciones de lectura y en la Figura 8 los IOPs. Debido al patrón de E/S que presenta esta aplicación las resultados de tasa de transferencia son muy parecidos a los obtenidos por NAS BT-IO cuando usa colectivas de E/S. Con las configuraciones con RAID 5 se logra mayor tasa de transferencia para las lecturas. En el caso de las escrituras se logra un mayor ancho de banda con las configuraciones sin RAID. Los IOPs permiten ver como a mayor operaciones por segundo mayor ancho de banda se obtiene. Los tiempos de E/S (Figura 9) muestran que la aplicación se beneficia

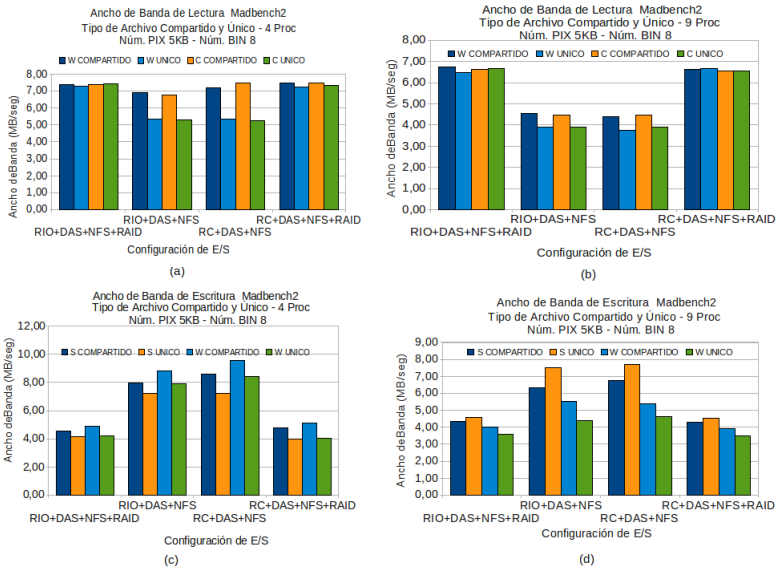

Figura 7. Ancho de Banda de las operaciones de lectura y escritura de Madbench2

del acceso compartido, en el acceso único, si bien cada procesador accede a un archivo exclusivo este acceso es a través de NFS y la comunicación penaliza los tiempos de E/S.

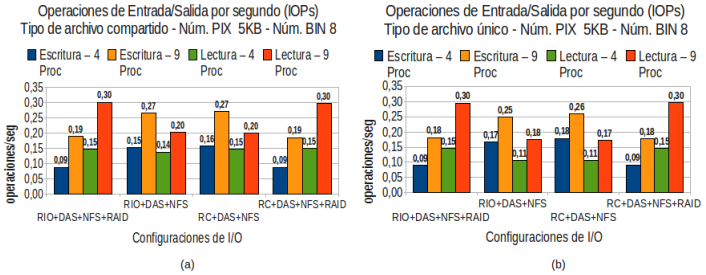

Figura 8. Lecturas y Escrituras por segundo para MADBench 2 en 4 y 9 procesadores

6. Conclusiones y Trabajos Futuros

Se ha propuesto una Metodología para la Configuración del sistema de E/S para computadores paralelos que ha permitido establecer establecer pautas para el análisis y la configuración de la E/S y evaluar la influencia de la configuración en las prestaciones. La metodología contempla la caracterización del sistema de E/S a distintos niveles: dispositivo, sistema y aplicación; configuración de diferentes elementos que tienen impacto en las prestaciones y evaluación teniendo en cuenta tanto la aplicación como la arquitectura de E/S. Este análisis de diferentes configuraciones de E/S es útil para la identificación de los factores de E/S que influyen en las prestaciones y es un trabajo inicial para llegar a largo plazo a definir un modelo para la configuración de E/S paralela. Para poder

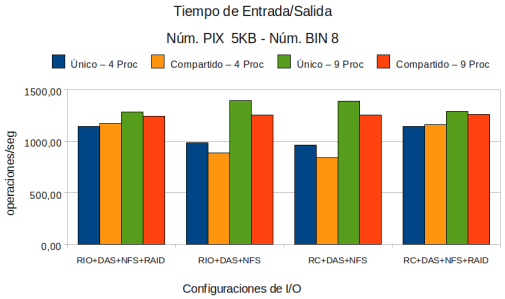

Figura 9. Tiempo de E/S para MADBench2 para 4 y 9 procesadores

validar y extender el modelo se necesita evaluar diferentes configuraciones y por esta razón se debe incluir herramientas de simulación que permita modelar la arquitectura de E/S, las librerías de E/S y los patrones de E/S de la aplicación científica.

Referencias

1. Roth, P. C.: Characterizing the I/O Behavior of Scientific Applications on the Cray XT. In: PDSW '07: Proceedings of the 2nd International Workshop on Petascale Data Storage, pp. 50–55. ACM, NY, USA (2007)
2. Yu, W., Oral, S., Vetter, J., Barrett, R.: Efficiency Evaluation of Cray XT Parallel IO Stack. (2007)
3. Yu, W., Oral, H. S., Canon, R. S., Vetter, J. S., Sankaran, R.: Empirical Analysis of a Large-Scale Hierarchical Storage System. In: Euro-Par '08, pp. 130–140. Springer-Verlag, Heidelberg (2008)
4. Fahey, M., Larkin, J., Adams, J.: I/O Performance on a Massively Parallel Cray XT3/XT4. In: Parallel and Distributed Processing.IPDPS , pp. 1 -12. IEEE (2008)
5. Laros, J.H., Lee, W., Klundt, R., Kelly, S., Tomkins, J. L., Kellogg, B. R.: Red Storm IO Performance Analysis. In: CLUSTER '07: Proceedings of the 2007 IEEE Int.Conf., pp. 50–57. Washington, DC, USA (2007)
6. Yu, H.,Sahoo, R.K.,Howson, C., Almasi, G., Castanos, J.G., Gupta, M., Moreira, J.E., Parker, J.J., Engelsiepen, T.E., Ross, R.B., Thakur, R., Latham, R., Gropp, W.D.: High Performance File I/O for the Blue Gene/L Supercomputer. In: High-Performance Computer Architecture , pp. 11–15. IEEE (2006)
7. Baylor, S. J., Benveniste, C. and Hsu, Y.: Performance Evaluation of a Massively Parallel I/O Subsystem. SIGARCH Comput. Archit. News. 22, 5–10 (1994)
8. Núnez, A., Fernández, J., Garcia, J. D., Prada, L., Carretero, J.: SIMCAN: A SIMulator Framework for Computer Architectures and Storage Networks. In: Simutools '08: Proceedings of the 1st Int.Conf. on Simul.Tools., pp. 1–8. ICST, Belgium (2008)
9. Byna, S., Yong, C., Xian-He, S., Thakur, R., Gropp, W.: Parallel I/O Prefetching Using MPI File Caching and I/O Signatures. In: High Performance Computing, Networking, Storage and Analysis. SC 2008, pp.15–21. IEEE (2008)
10. Kotz, D.: Introduction to Multiprocessor I/O Architecture. In: Input/Output in Parallel and Distributed Computer Systems, pp. 97–123. Kluwer Acad.Pub.(1996)
11. Wong, P., Van Der Wijngaart, Rob. F.: NAS Parallel Benchmarks I/O Version 2.4. Technical report, NASA Advanced Supercomputing (NAS) Division (2003)
12. Borrill, J., Oliker, L., Shalf, J., Shan, H.: Investigation of Leading HPC I/O Performance Using a Scientific-Application Derived Benchmark. In: SC '07: Proceedings of Conf. on Supercomputing, pp. 1–12. ACM, New York, USA (2007)