IAHR **AIIH**

XXIV CONGRESO LATINOAMERICANO DE HIDRÁULICA
PUNTA DEL ESTE, URUGUAY, NOVIEMBRE 2010

Cluster FING: Una Plataforma Computacional de Alto Desempeño Aplicable a la Resolución Eficiente de Problemas de Hidráulica

Gerardo Ares[1], Sergio Nesmachnow[1], Pablo Ezzatti[1], Gabriel Usera[2]

[1] Centro de Cálculo, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay
gares@fing.edu.uy, sergion@fing.edu.uy, pezzatti@fing.edu.uy
[2] Instituto de Mecánica de los Fluidos e Ingeniería Ambiental,, Facultad de Ingeniería, Universidad de la República, Uruguay
gusera@fing.edu.uy

RESUMEN:

Los continuos avances en la tecnología de la computación han permitido la resolución de problemas cada vez más complejos. Al mismo tiempo, se han desarrollado infraestructuras de alto desempeño dedicadas al cálculo científico que permiten un mayor poder de cómputo. En los últimos cuatro años la Facultad de Ingeniería de la Universidad de la República ha estado trabajando en la construcción de una plataforma computacional de alto desempeño, esfuerzo que se cristalizó con la implantación del Cluster FING. El cluster ha sido aplicado para la resolución eficiente de problemas hidráulicos con alta demanda de cómputo, en especial, en el área de simulación numérica de fluidos, donde es necesario disponer de un poder de cómputo considerable para llevar a cabo simulaciones precisas de grandes escenarios en tiempos razonables.
Este trabajo presenta la infraestructura de alto desempeño disponible en la Facultad de Ingeniería, un estudio de los componentes a través de la ejecución del benchmark HPCC, y se describen algunas de las aplicaciones de simulación numérica de fluidos que se han beneficiado con esta infraestructura.

ABSTRACT:

The constant evolution of computer technologies has allowed solving increasingly complex problems. At the same time, high performance infrastructures dedicated to scientific computing were developed, making possible to have high computational power. In the last 4 years the Facultad de Ingeniería de la Universidad de la República has been working to build a high performance infrastructure. This effort has crystallized in the implementation of the Cluster FING. This cluster has been applied to resolve hydraulic problems that have high computation requirements, especially in the computational fluid dynamics area where it is necessary to have high computation power to carry accurate simulations of large scenarios in reasonable times.
This paper presents a high performance infrastructure available at the Facultad de Ingenería, a study of the infrastructure components evaluated by the HPCC benchmark, and describes some of the computational fluid dynamics applications that have benefited from this infrastructure

PALABRAS CLAVES:

Computación de alto desempeño, Hidráulica computacional, Cluster.

INTRODUCCIÓN

Las infraestructuras de computación paralela-distribuida se han popularizado masivamente en la última década como una herramienta efectiva para satisfacer el poder de cómputo necesario para la resolución de problemas complejos en diversas áreas de aplicación. En especial, en el contexto de los problemas de hidráulica, la simulación numérica de fluidos, los análisis meteorológicos, etc., es necesario disponer de un poder de cómputo considerable para llevar a cabo simulaciones precisas de grandes escenarios en tiempos razonables. En los últimos años, los avances tecnológicos han propiciado un vertiginoso desarrollo de la capacidad de cálculo de los computadores y una drástica reducción de su costo, creando la oportunidad para abordar un número cada vez mayor de temáticas prácticas mediante la simulación numérica. Esta oportunidad abre nuevas perspectivas en nuestros países, donde la financiación limitada usualmente condiciona el tipo de experimentos necesarios para modelar problemas de gran porte, y del mismo modo, restringe el acceso a recursos computacionales de alto costo. En la Facultad de Ingeniería se comenzó a trabajar con pequeñas plataformas de alto desempeño aplicadas a la mecánica de los fluidos computacional en los inicios de la década de 2000. Posteriormente, un prototipo de cluster de alto desempeño, llamado Cluster Medusa, se implementó en el marco del proyecto "Laboratorio de Simulación Numérica para Flujos de Superficie Libre" (Nesmachnow y Usera, 2008). La evaluación positiva de estas experiencias motivó el desarrollo de una infraestructura de mayor porte para la resolución de este tipo de problemas: el Cluster FING, financiado inicialmente por la Comisión Sectorial de Investigación Científica (CSIC) de la Universidad de la República.
Este trabajo presenta el Cluster FING, y lo identifica como una plataforma computacional de alto desempeño que ha sido aplicada para la resolución eficiente de problemas hidráulicos con altos requisitos de cómputo. Los objetivos principales del proyecto que instaló el cluster fueron: i) brindar una plataforma computacional que permita mejorar sustancialmente la capacidad de abordar en un ámbito nacional problemas de gran porte en varias disciplinas, y en especial problemas que involucren la simulación numérica de fenómenos físicos, y ii) acercar a investigadores de variadas áreas al conocimiento de este tipo de tecnologías y su aplicación como herramienta para resolver problemas que requieran un alto costo computacional. En este artículo se describen las actividades llevadas a cabo en el marco del proyecto, presentando la infraestructura, la evaluación de su desempeño computacional y describiendo aplicaciones que han tomado provecho del cluster.

El resto del artículo se organiza del modo que se describe a continuación. La siguiente sección brinda una descripción detallada de la infraestructura del Cluster FING, analizando sus principales componentes tecnológicos (procesador, memoria y conectividad de red) y las herramientas de software utilizadas para facilitar las tareas de instalación, configuración, análisis de utilización de recursos computacionales, administración del cluster y planificación de tareas a ejecutar en la infraestructura. A continuación se presentan los detalles de la evaluación del desempeño computacional del cluster utilizando un benchmark que contempla los componentes tecnológicos comentados y las principales métricas de desempeño relevantes para la resolución de complejos problemas numéricos como los que surgen en el ámbito de los problemas de hidráulica. Se incluye un estudio comparativo del desempeño frente a dos clusters de menor y mayor porte, de forma de tener un buen panorama del potencial de la infraestructura. Luego se presentan las actividades de formación en el área de computación científica de alto desempeño llevadas a cabo en el marco del proyecto, y se brinda una descripción de dos aplicaciones típicas en el área de la hidráulica que han sido implementadas en el cluster y se han beneficiado de una mejora de desempeño: un modelo numérico de corrientes del Río de la Plata y la simulación numérica de flujos a superficie libre. Por último, las conclusiones del trabajo resumen las actividades presentes y futuras en relación al fortalecimiento del uso de este tipo de arquitecturas por parte de proyectos de investigación en nuestro entorno de trabajo.

EL CLUSTER FING

Un *cluster* es una colección de computadoras que incluye nodos especializados en el cómputo, nodos de administración, nodos de entrada/salida (acceso al sistema de archivos), todos interconectados por una o varias redes de alta velocidad.

El diseño de un cluster de alto desempeño requiere el estudio de las diferentes arquitecturas computacionales disponibles, en conjunto con el análisis de las clases de problemas que se desean abordar, para, de esa forma, lograr minimizar el costo monetario de la instalación. Considerando las variadas características de las áreas de investigación en la Facultad de Ingeniería de la Universidad de la República, el Cluster FING fue ideado para ser una plataforma orientada a actividades multidisciplinarias. Para ello, se buscó lograr un equilibrio entre la inversión en nodos de cómputo y en memoria física asignada a cada nodo y, por lo tanto, permitir cubrir aplicaciones que requieran un alto uso de memoria, así como aplicaciones que requieran mayor cantidad de unidades de cómputo.
El poder de procesamiento es el elemento primordial en un cluster dedicado al cálculo intensivo. El análisis de las opciones de procesadores se concentró en varios aspectos: el número de nodos de ejecución (que condiciona el espacio físico de instalación, el consumo energético y el costo efectivo de adquisición, operación y mantenimiento), la cantidad de unidades de procesamiento por nodo, la velocidad de reloj y registros manejados por el procesador (que condicionan el desempeño individual de cada nodo de cómputo, y afectan el rendimiento del cluster en su conjunto), el tamaño de la memoria física y virtual (que determina la capacidad de manejo de datos) y el diseño y la velocidad del bus del sistema (que permiten evaluar la velocidad de procesamiento efectivo de la arquitectura, tratando de evitar cuellos de botella en las transferencias de datos). En ese sentido, se relevaron y evaluaron las arquitecturas de mayor difusión para clusters de gran porte según el ranking de supercomputadores TOP500 de junio de 2008 (TOP500, 2008). Se estudiaron las prestaciones reales de los procesadores evaluados, analizando los resultados de la ejecución de benchmarks estándar sobre clusters de alto desempeño, disponibles públicamente. Las aplicaciones de resolución de problemas de hidráulica, tanto como otras áreas, realizan una gran cantidad de operaciones sobre números enteros y reales (representados mediante punto flotante en una computadora), de esa forma, se orientó la evaluación al desempeño de los procesadores en operaciones con números enteros y de punto flotante y número variable de núcleos de procesamiento. A su vez, es importante tener un buen nivel de velocidad de acceso a la información por parte de los procesadores, para no tener cuellos de botella importantes que limiten la capacidad de cómputo. En ese sentido, se evaluó la velocidad de acceso a memoria y a disco, y manejo de datos a través de la red de intercomunicación. Estas evaluaciones fueron hechas a través de la ejecución del benchmark HPC Challenge (HPCC, 2010) que brinda un conjunto de test que cubren, de buena forma, las necesidades planteadas.

A seguir se introducirá la arquitectura del Cluster FING realizando una revisión de procesadores, memoria, red de interconexión, y herramientas de administración del cluster.

Procesador

Al momento de realizar los estudios preliminares en junio de 2008 el 99,6% de los supercomputadores del TOP500 estaba basado en procesadores de arquitectura escalar. A su vez, el 82,2% del total se basaba en la arquitectura x86-64, el 13,6% en Power, el 3,2% en Intel Itanium, y el resto de los procesadores estaba (Cray, NEC, IA-32) por debajo del 1% de uso. Dentro de la arquitectura x86-64, el 71,0% correspondía a la familia de procesadores Intel EM64T y el 11,1% a la familia AMD X86_64.

A su vez, en los últimos cinco años los procesadores escalares comenzaron a utilizar la tecnología multi-núcleo (varios núcleos de procesamiento por procesador) que brinda dos grandes ventajas: la

reducción en el consumo de energía y los costos asociados son menores en relación al poder de cómputo obtenido. Es por esto interesante tener la mayor cantidad posible de núcleos en cada procesador.

A partir de la información recolectada se decidió estructurar una propuesta sobre la arquitectura x86-64 en base a la tecnología Quad-core (cuatro núcleos por procesador, cantidad máxima de núcleos disponibles al momento de la formulación del proyecto). Se evaluaron propuestas con modelos AMD Opteron e Intel Harpertown y finalmente se optó por una propuesta basada en los procesadores Intel E5430 descripta por Gepner (Gepner et al., 2008).

Los procesadores Intel E5430 pertenecen a la serie 5000 de procesadores Intel que ha sido diseñada para el uso en ambientes de alto desempeño y buscando maximizar el aprovechamiento de la energía. Los procesadores Intel E5430 pertenecen a la tecnología Harpertown, los cuales se interconectan a un chipset Intel 5000. Estos chipset se dividen en dos componentes: Northbridge (Controlador de memoria) y Southbridge (Controlador de dispositivos de Entrada/Salida como discos, red). El componente Northbridge se concentra en la interconexión con la memoria RAM, y provee hasta cuatro canales de acceso a esta . En particular, el procesador E5430 posee cuatro núcleos de ejecución, con una velocidad del reloj de 2.66GHz, 12MB de L2 cache y un front-side-bus de 1333MHz. El componente Southbridge permite el acceso a dispositivos de Entrada/Salida (I/O) como red y disco.

Como se menciono anteriormente, es necesario poder realizar un estudio de la cantidad de operaciones de punto flotante que se realizan por segundo para la evaluación del poder de cómputo de la arquitectura. El valor establecido en la comunidad científica es el FLOPS, acrónimo de FLoat point Operations Per Second, que representa la cantidad de operaciones de punto flotante que un procesador es capaz de realizar por segundo. De esa forma, el estudio teórico de operaciones de punto flotante presentado por Gepner et al. (2008) sobre esta serie de procesadores Intel, se define como la multiplicación de la frecuencia del procesador por la cantidad de operaciones de punto flotante por ciclo del reloj y por la cantidad de núcleos por procesador. Tomando en cuenta que la frecuencia del procesador Intel E5430 es de 2.6 GHz, que cada núcleo genera hasta cuatro operaciones de punto flotante por ciclo de reloj, y que se tienen cuatro núcleos por procesador, el pico teórico de operaciones de punto flotante por segundo es de 42,6 GFlops.

Memoria

Conjuntamente con el poder de cómputo que brindan los procesadores es necesario definir una buena relación de memoria RAM asignada para cada unidad de procesamiento (núcleo). El costo de la memoria se torna bastante significativo en configuraciones grandes, por lo se debe establecer una relación núcleo/memoria que permita cubrir las necesidades de la mayoría de las aplicaciones que ejecutarán sobre la arquitectura. De esa forma, se estableció asignar 1 GByte de memoria RAM por núcleo ya que con esto se abarcaban todas las aplicaciones que se utilizarían en el tiempo cercano . De todas formas, los servidores adquiridos permiten incrementar la memoria hasta un máximo de 4GByte para aplicaciones que puedan requerir mayores dimensiones de memoria.

Red de interconexión

La tecnología de interconexión de los nodos es un factor crítico en el desempeño de las aplicaciones ejecutando sobre un computador con memoria distribuida. Para el diseño del Cluster FING se evaluaron las tecnologías de interconexión más populares para un cluster de alto desempeño: Gigabit Ethernet, Myrinet, SCI, Quadrics e InfiniBand. En junio de 2008, el 56,6% de los servidores del TOP500 (TOP500, 2008) estaban interconectados con una red Gigabit Ethernet, el 24,2% con Infiniband, mientras que el resto está por debajo del 8% (Myrinet 2,4% y Quadrics 1%). Dado esto se relevaron las tecnologías de Infiniband y Gigabit Ethernet. Si bien Infiniband es una tecnología que permite un mayor desempeño frente a Gigabit Ethernet su alto costo (sobre todo a

nivel de switches), y, a su vez, dada la buena aceptación de Gigabit Ethernet a nivel de cluster de alto desempeño, se decidió utilizar esta última tecnología. El cluster fue provisto con un switch Dell PowerConnect 5424 de 24 puertos Gigabit Ethernet con un cableado categoría 6.

Arquitectura del cluster

La configuración inicial del Cluster FING, implementada en diciembre de 2008, consistió de ocho servidores de cómputo Dell PowerEdge multinúcleo (cada nodo contiene dos procesadores Xeon EM64T E5430, con cuatro núcleos a 2.66 GHz cada uno), y un nodo administrativo con las mismas características que, a su vez, brinda acceso a un arreglo de discos a través del sistema de archivo Network File System (NFS). Este nodo también puede funcionar como nodo de cómputo, totalizando 72 núcleos de procesamiento. Como se mencionó anteriormente, los nodos se interconectaron con una red Gigabit Ethernet a través de un switch Dell PowerConnect 5424.

En la configuración de 64 unidades de cómputo el Cluster FING permite un pico teórico máximo de 681,6 GFlops y de 766,8 GFlops en la configuración de 72 unidades de cómputo.

Sistema operativo y herramienta administrativa

El sistema operativo seleccionado para el cluster fue Linux, debido a su alta difusión en la comunidad científica, la cantidad de compiladores disponibles y bibliotecas de libre acceso, así como la solidez que ha presentado para el tipo de uso que tiene el cluster. Una prueba de la importancia de los sistemas operativos Linux en el ámbito de la computación de alta performance se puede ver al estudiar los sistemas operativos utilizados por los principales cluster en la lista del TOP500 de junio de 2008 (TOP500, 2008): 85,4% de los clusters utilizanban Linux, y el 80% empleaban una distribución de uso libre. Dentro de las distribuciones libres se evaluaron Fedora, Debian y CentOS. Dado que las tres distribuciones proveían las bibliotecas y el software de desarrollo necesario para el cluster, no existió una argumentación que predominara sobre las demás, por lo que finalmente se decidió optar por CentOS 5 debido a contar con una mayor experiencia en esta distribución por parte de los administradores del cluster.

A su vez, se analizó la disponibilidad de herramientas semiautomáticas de instalación y administración de clusters de alto desempeño, que engloban aplicaciones para simplificar la instalación, monitoreo y administración de los recursos de cómputo y las comunicaciones. Las herramientas evaluadas fueron OSCAR (Open Source Cluster Application Resources) (OSCAR, 2010) y ROCKS (ROCKS, 2010), seleccionándose OSCAR al contarse con una experiencia previa satisfactoria en la instalación del cluster Medusa (Nesmachnow y Salsano, 2007).

OSCAR es una compilación autoinstalable que provee el software para instalar y administrar un cluster: bibliotecas y compiladores para programación paralela y distribuida herramientas de administración, y herramientas de monitoreo. El componente de instalación permite crear una imagen de un nodo, y luego instalar y configurar automáticamente los restantes nodos del cluster. Varios componentes adicionales ofrecen un entorno integrado para la administración del cluster, utilizando una base de datos para almacenar el software y su configuración. En cuanto a las herramientas de monitoreo, OSCAR incorpora la herramienta Ganglia Monitoring System (GMS, 2010) que permite visualizar el uso de los recursos computacionales (procesador, memoria, red, disco) a nivel global y a nivel de nodo. La herramienta brinda la información a través de una interface web con gráficos actualizados en tiempo real, permitiendo el acceso a la información en forma sencilla y clara. Ganglia se basa en la configuración de agentes que ejecutan en los nodos y reportan a un nivel central el estado en tiempo real del cluster. A su vez, OSCAR provee el software para planificación de tareas TORQUE (Terascale Open-Source Resource and QUEue manager) (TORQUE, 2010). Este software permite la planificación de las tareas a ser ejecutadas en el cluster. Basándose en un sistema de colas de prioridad de tareas, en donde cada tarea es asignada a una cola según algunos criterios preestablecido en el Cluster FING. Posteriormente, el planificador de trabajos decide el momento en que cada tarea comienza su ejecución. De esta forma, se logra utilizar el cluster de forma ordenada y equitativa entre los distintos grupos de investigadores.

Actualmente, TORQUE ha sido configurado y está operativo.

De esta forma fue presentada la arquitectura del Cluster FING, describiendo sus componentes principales. En la siguiente sección se presenta una evaluación de su desempeño computacional.

EVALUACIÓN DEL DESEMPEÑO COMPUTACIONAL DEL CLUSTER FING

En esta sección se realiza una evaluación del desempeño computacional del Cluster FING. Inicialmente, se presenta un benchmark utilizado para la evaluación. Seguidamente, se realiza un análisis de los resultados obtenidos. Finalmente, se mencionan las actualizaciones recientes realizadas en el equipamiento.

La motivación del estudio del desempeño de un cluster surge para tener una medida conocida del poder de cómputo del mismo, así como también para realizar ajustes en los parámetros del sistema (sistema operativo, redes, componentes de hardware) de forma de obtener el mayor rendimiento posible en la ejecución de aplicaciones en las áreas temáticas involucradas.

La evaluación de un cluster creados para fines de investigación y que, a su vez, permita ser utilizado en una ambiente multipropósito (no dedicado) por diferentes disciplinas, requiere un estudio de prácticamente todos los componentes del mismo (cantidad de operaciones, acceso a red, memoria, etc.). En este caso, se utilizó el benchmark HPC Challenge Benchmark. Este benchmark fue diseñado y es mantenido por investigadores de la Universidad de Tennnessee, los cuales se basan en distintos benchmarks que permiten evaluar diferentes componentes de un cluster, todos disponibles libremente bajo la licencia BSD. El benchmark se compone de los siguientes siete test descriptos a continuación:

- HPL (Linpack) es un benchmark que evalúa el número de operaciones de punto flotante de doble precisión en la resolución de un sistema lineal de ecuaciones a través de la descomposición LU.
- DGEMM evalúa el número de operaciones de punto flotante en la multiplicación de matrices de reales de doble precisión. Se proveen dos test SDGEMM (SingleDGEMM) que permite ejecutar en un único procesador y EPDGEMM (StarDGEMM) que permite ejecutar en varios.
- STREAM es un programa que evalúa el ancho de banda para el acceso a memoria RAM. Utiliza cuatro operaciones de evaluación: copia (Copy), escalado (Scale), suma (Add) y tríada (Triad). A su vez, se proveen de dos test para las cuatro operaciones SSTREAM (SingleSTREAM) que permite la ejecución en un único procesador y EPSTREAM (StarSTREAM) que permite la ejecución en varios procesadores.
- PTRANS evalúa la comunicación entre dos procesadores en equipos distintos, dando una buena medida de la capacidad de comunicación en la red.
- RandomAccess evalúa el número de actualización de memoria en números enteros. Se proveen dos test SRandomAccess y EPRandomAccess (StarRandomAccess).
- FFT evalúa el número de operaciones de punto flotante al ejecutar transformadas de Fourier discretas y uni-dimensionales sobre aritmética compleja de punto flotante doble. HPCC provee tres variants: SFFT (SingleFFT), EPFFT(StartFFT) y MPIFFT (MPI-FFT).
- Communication and Latency evalúa el ancho de banda y la latencia para varios patrones estándar de comunicación sobre mensajes de largo variable, utilizando dos test que consideran topología de anillo ordenado y de anillo aleatorio.

Información más detallada sobre el HPCC se presenta en el sitio web http://icl.cs.utk.edu/hpcc. Un resumen de su contenido se encuentra en el análisis del desempeño computacional del Cluster Medusa, realizado por Nesmachnow y Salsano (2007).

Resultados Experimentales

Esta sección comenta la motivación de la estrategia seleccionada para la ejecución del benchmark HPCC sobre el Cluster FING y a continuación presenta y analiza los resultados de eficiencia computacional obtenidos en el estudio de desempeño.

Los problemas de simulación numérica de flujos son abordados a través de la resolución numérica de ecuaciones diferenciales que, finalmente, derivan en la resolución de sistemas de ecuaciones que involucran matrices de gran dimensión. Estos sistemas requieren de un elevado número de recursos computacionales para su resolución, por lo que suelen aplicarse técnicas de computación paralela y de alto desempeño para obtener resultados eficientes en tiempos razonables. Las alternativas más utilizadas involucran la resolución con múltiples procesos utilizando el paradigma de memoria compartida y la aplicación de la estrategia de división de dominio que implica asignar submatrices a diferentes unidades de procesamiento (Golub et al., 1996). Como consecuencia, las métricas más relevantes en el estudio de rendimiento de las aplicaciones paralelas se concentran en evaluar el aporte de la programación paralela al incrementarse el tamaño del problema (*escalabilidad*) y analizar la capacidad de mejorar el desempeño para un problema de tamaño fijo (*paralelicibidad*).

El benchmark HPCC se basa en la resolución de un sistema lineal de ecuaciones de tamaño *N*. La matriz del sistema se crea dinámicamente de forma aleatoria y su tamaño es un parámetro configurable al comienzo de la ejecución. Considerando los objetivos mencionados, se realizó un plan de ejecuciones para analizar las métricas relevantes (escalabilidad y paralelicibidad). Para el estudio de la escalabilidad se realizaron ejecuciones del benchmark en 1, 2, 4 y 8 nodos con configuraciones de tamaño de matriz *N*=30000, *N*=43000, *N*=60000 y *N*=82500, valores seleccionados para maximizar el uso de la memoria RAM en cada configuración. El tamaño ocupado por la matriz en memoria se calcula como la cantidad de elementos de la matriz por el número de bytes necesarios para representar su valor. Dado que el espacio utilizado por la representación de punto flotante de doble precisión es 8 bytes, para maximizar el uso de la memoria se debe utilizar un problema de tamaño *N* igual a la raíz cuadrada la cantidad de elementos que se puedan representar en la memoria disponible. Por ejemplo, en la configuración de cuatro nodos se dispone de un total de 32 GB de memoria RAM, que permite almacenar 4×10^9 elementos (una matriz de 63245×63245). Dado que es necesario elegir un tamaño de problema un poco menor para dejar lugar de memoria al sistema operativo. De esa forma, el *N* seleccionado para cuatro nodos fue 60000, y así mismo se procedió para los otros casos.

Los resultados de las ejecuciones del benchmark HPCC para evaluar la escalabilidad del Cluster FING se presentan en las Tablas 1, 2 y 3.

Tabla 1.- HPCC (HPL,DGEMM,PTRANS,RandomAccess,FFT)

Cantidad de nodos	**HPL** [GFlops/s]	**SDGEMM** [GFlops/s]	**EPDGEMM** [GFlops/s]	**PTRANS** [GByte/s]	**SRandom Access** [GUPS/s]	**EPRandom Access** [GUPS/s]	**SFFT** [GFlops/s]	**EPFFT** [GFlops/s]	**MPIFFT** [GFlops/s]
1	70.9	79.1	77.8	0.6	0.01	0.01	1.12	1.09	0.80
2	122.7	78.1	78.2	0.3	0.01	0.01	0.97	1.04	0.70
4	203.6	78.5	78.5	0.4	0.01	0.01	0.97	1.04	0.98
8	397.1	78.7	78.6	0.6	0.01	0.01	1.13	1.01	1.36

Tabla 2.- HPCC (STREAM)

Cantidad de nodos	**SSTREAM (Copy)** [GByte/s]	**SSTREAM (Scale)** [GByte/s]	**SSTREAM (Add)** [GByte/s]	**SSTREAM (Triad)** [GByte/s]	**EPSTREAM (Copy)** [GByte/s]	**EPSTREAM (Scale)** [GByte/s]	**EPSTREAM (Add)** [GByte/s]	**EPSTREAM (Triad)** [GByte/s]
1	3.38	3.38	3.59	3.58	3.38	3.38	3.59	3.58
2	3.38	3.39	3.58	3.58	3.38	3.39	3.58	3.58
4	3.38	3.38	3.57	3.57	3.38	3.39	3.58	3.58
8	3.38	3.39	3.58	3.58	3.39	3.39	3.59	3.59

Tabla 3.- HPCC (Communication and Latency)

Cantidad de nodos	Communication (anillo aleatorio) [GByte/s]	Communication (anillo ordenado) [GByte/s]	Latency (anillo aleatorio) [µs]	Latency (anillo ordenado) [µs]
1	N/C	N/C	N/C	N/C
2	0.108	0.107	30.22	33.69
4	0.081	0.082	31.54	30.49
8	0.078	0.085	36.32	36.81

Para la evaluación de la paralelicibidad del algoritmo propuesto por el benchmark HPCC se realizaron las ejecuciones en 1, 2, 4 y 8 nodos con un mismo tamaño de problema. El tamaño seleccionado de la matriz fue de *N*=30000 (máximo tamaño recomendado para la ejecución en un nodo). Los resultados obtenidos en las ejecuciones se reportan en la Tabla 4.

Tabla 4.- Comparativo de escalabilidad de HPCC en Cluster FING

Test	Unidades	1 nodo	2 nodos	4 nodos	8 nodos
HPL	GFlops/s	70.9	112.2	145.3	230.5
SDGEMM	GFlops/s	79.1	77.3	76.1	76.8
EPDGEMM	GFlops/s	77.8	77.2	76.2	76.2
PTRANS	GByte/s	0.65	0.32	0.41	0.66
SRandomAccess	GUPS/s	0.01	0.01	0.02	0.02
EPRandomAccess	GUPS/s	0.01	0.01	0.02	0.02
SFFT	GFlops/s	1.12	1.04	1.18	1.17
EPFFT	GFlops/s	1.09	1.10	1.16	1.14
MPIFFT	GFlops/s	0.80	0.71	0.97	1.27
SSTREAM (Copy)	GByte/s	3.38	3.38	3.38	3.38
SSTREAM (Scale)	GByte/s	3.38	3.39	3.39	3.39
SSTREAM (Add)	GByte/s	3.59	3.58	3.58	3.58
SSTREAM (Triad)	GByte/s	3.58	3.58	3.58	3.58
EPSTREAM (Copy)	GByte/s	3.38	3.38	3.38	3.39
EPSTREAM (Scale)	GByte/s	3.38	3.39	3.39	3.39
EPSTREAM (Add)	GByte/s	3.59	3.58	3.59	3.59
EPSTREAM (Triad)	GByte/s	3.58	3.58	3.58	3.59
Communication (anillo aleatorio)	GByte/s	N/C	0.107	0.081	0.078
Communication (anillo ordenado)	GByte/s	N/C	0.107	0.083	0.089
Latency (anillo aleatorio)	µs	N/C	31.23	30.93	35.84
Latency (anillo ordenado)	µs	N/C	29.11	28.80	37.50

Análisis de los resultados

El análisis de los resultados en la Tabla 1 permite concluir que la infraestructura mostró un muy buen comportamiento de escalabilidad al resolver el problema, obteniéndose un valor de 5,6 (397,1/70,9), analizando el resultado del test HPL). Este valor indica un resultado promisorio para aplicaciones que aborden la resolución mediante la división de dominio, usuales en problemas de hidrodinámica y de simulación numérica de fluidos. A su vez, en la configuración con 8 nodos (64 unidades de cómputo) se logra una eficiencia computacional que corresponde al 57% del pico teórico. La diferencia con el pico teórico se debe, en gran medida, a la memoria disponible en el cluster (1 GB por núcleo de ejecución), que no permite escalar el problema a dimensiones que permitan tomar mayor partido de la estrategia de resolución en paralelo. La relación se puede visualizar en la Figura 1. De todas formas, hay un rendimiento estable al escalar de 4 a 8 nodos de cómputo. En cuanto a los test sobre el acceso a la memoria (RandomAccess, STREAM), el rendimiento se mantiene estable en todas las configuraciones, indicando que al escalar el problema no se modifica la velocidad de acceso a la memoria, que se mantiene constante. Por otro lado, la

evaluación de la red de comunicaciones (Communication and Latency) muestra una degradación de 19% al escalar de 2 a 8 nodos, valor ínfimo al considerar que el problema fue escalado en un 378%.

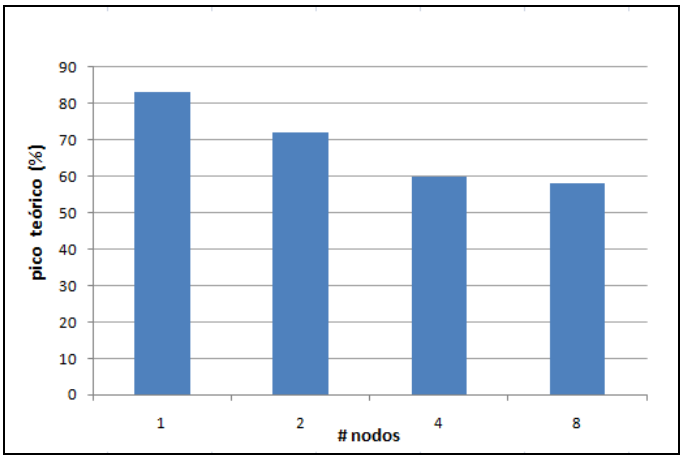

Figura 1.- Rendimiento del Cluster FING en relación al número de nodos

Con respecto a la paralelicibidad del problema, evaluada en los resultados de la Tabla 4, se observa que se obtiene un incremento de eficiencia del orden de 3,2 (230,5 /70,9) según el test HPL al utilizar los 8 nodos de cómputo. Los accesos a la memoria, evaluados por los test RandomAccess y STREAM, se mantienen estables como en los test de escalabilidad. Por último, los test que evalúan la red de comunicaciones (Communication and Latency) muestran una degradación del 10% al utilizar 8 nodos, valor muy pequeño en relación con el incremento de eficiencia de 3,2 en la resolución del problema.

Comparación con otras plataformas de alto desempeño

Complementando la evaluación de escalabilidad y paralelicibidad, se realizó un análisis comparativo de la eficiencia del Cluster FING con otros clusters similares. El objetivo de la comparación fue obtener un punto de comparación para validar las evaluaciones realizadas y, a su vez, brindar un marco que muestre las capacidades y falencias de la infraestructura actual para tomar decisiones en el momento de realizar una actualización de equipamiento. Los clusters elegidos para la comparación fueron el Cluster Medusa y el cluster “Hewlett Packard BL280cG6” cuyos resultados están disponible en la página web del sitio HPCC. El Cluster Medusa se compone de 12 unidades de procesamiento AMD Opteron 175 a 2.2GHz con un total de 12GB de memoria RAM e interconectados a través de una red de 100 Megabit Ethernet. El Cluster HP se compone de 64 unidades de procesamiento Intel Nehalem X5550 a 2.6GHz con un total de 192GB de memoria RAM e interconectados a través de una red Infiniband QDR. La comparación frente al Cluster Medusa permite evaluar comparativamente la eficiencia de dos infraestructuras disponibles en Facultad de Ingeniería, mientras que la comparación con el Cluster HP permite evaluar el Cluster FING contra un cluster con tecnologías superior en los tres componenetes principales (procesador, memoria y red de interconexión). Los valores de las ejecuciones se reportan en la Tabla 5.
El análisis de los resultados obtenidos permite concluir que se dispone de una mejora sustancial en el poder de cómputo, acceso a memoria y capacidad de interconexión entre los nodos con respecto al Cluster Medusa. El análisis comparativo con el Cluster HP reporta una menor performance de los test de acceso a la memoria y en la red de interconexión. Sin embargo, se debe tener en cuenta que el problema resuelto por el Cluster HP es de tamaño N=139008, es de entre dos y tres veces mayor en orden de magnitud, como consecuencia de la menor cantidad de memoria disponible en el Cluster FING (64GB frente a 192GB en el Cluster HP). Si bien el Cluster FING presenta mejores valores para el test DGEMM, debe considerarse que el valor es calculado en base a cada proceso.
En el Cluster FING fueron utilizados para la resolución ocho procesos, en donde cada uno fue asignado a un nodo de cómputo, mientras que en el Cluster HP se asignó un proceso por núcleo.

Tabla 5.- Comparativo HPCC (Cluster FING-Medusa-HP)

Test	Unidades	Cluster FING	Cluster Medusa	Cluster HP
HPL	GFlops/s	395.6	18.4	613.8
SDGEMM	GFlops/s	77.2	3.8	10.3
EPDGEMM	GFlops/s	77.7	4.0	10.3
PTRANS	GByte/s	0.60	0.03	13.83
SRandomAccess	GUPS/s	0.01	0.01	0.03
EPRandomAccess	GUPS/s	0.01	0.01	0.02
SFFT	GFlops/s	1.13	0.52	2.02
EPFFT	GFlops/s	1.01	0.45	1.45
MPIFFT	GFlops/s	1.36	0.09	37.04
SSTREAM (Copy)	GByte/s	3.38	2.29	9.77
SSTREAM (Scale)	GByte/s	3.39	2.32	10.54
SSTREAM (Add)	GByte/s	3.58	2.77	11.21
SSTREAM (Triad)	GByte/s	3.58	2.55	11.00
EPSTREAM (Copy)	GByte/s	3.39	1.24	4.96
EPSTREAM (Scale)	GByte/s	3.39	1.24	3.37
EPSTREAM (Add)	GByte/s	3.59	1.42	3.70
EPSTREAM (Triad)	GByte/s	3.59	1.43	3.71
Communication (anillo aleatorio)	GByte/s	0.078	0.003	0.299
Communication (anillo ordenado)	GByte/s	0.085	0.003	1.466
Latency (anillo aleatorio)	µs	36.32	87.89	1.31
Latency (anillo ordenado)	µs	36.81	94.37	0.92

Considerando este aspecto, el valor normalizado para el Cluster FING es 9,7 frente al 10,3 logrado por el Cluster HP. El valor 9,7 obtenido para el Cluster FING indica un muy buen desempeño, ya que es en el orden del 91,4% del pico teórico. El test STREAM muestra la mejora en el acceso a la memoria por parte de la tecnología Nehalem frente a la Harpertown. En la tecnología Nehalem los controladores de memoria fueron migrados a cada procesador, dejando de lado la tecnología presentada en la sección del procesador Intel E5430. Finalmente, los test de comunicaciones y latencia permiten visualizar diferencias en la red de interconexión. Esto se debe al uso de una tecnología de mayor desempeño en el Cluster HP (Infiniband) frente a la utilizada en el Cluster FING (Gigabit Ethernet).

Situación actual del cluster

Sobre comienzos del 2010, el Cluster FING ha sido actualizado con la incorporación de cuatro nodos de procesamiento con tecnología Intel Nehlaem. Cada uno de estos nodos contiene dos procesadores Intel Nehalem E5520 de 2.27 GHz. De esta forma, el pico teórico de Gigaflops por segundo fue incrementado a 972,2 GFlops**.** Cada uno de estos cuatro nodos tiene un total de 24 GB de memoria RAM, por lo que la relación núcleo/memoria es de 3 GB por núcleo, permitiendo cubrir las necesidades de recursos de memoria para algunos de los proyectos de investigación que así lo requerían. A su vez, se ha incorporado recientemente un nodo que consta de cuatro NVIDIA Tesla C1060 (Lindholm et al., 2008) que permite el desarrollo en la investigación sobre la nueva tecnología paralela GPGPU (General Purpose Graphical Processor Unit). Cada Tesla C1060 contiene 240 procesadores de precisión simple y 30 procesadores de precisión doble, permitiendo un pico teórico de 82 GFlops en cada tarjeta. Este nodo contiene 2 procesadores Intel Nehalem E5530 de 2.4 GHz con 48 GB de memoria RAM. De esta forma, el nodo Tesla tiene un pico teórico de 387,6 GFlops, compuestos por los 310,8 GFlops de la componente GPU y 76,8 GFlops que brindan los procesadores Intel. Entonces, el pico teórico actual del Cluster Fing es de **1,368 TeraFlops**. Información actualizada y en constante desarrollo del cluster se presenta en detalle en su sitio web (http://www.fing.edu.uy/cluster).

APLICACIONES DEL CLUSTER FING

Desde su instalación y configuración, el Cluster FING ha sido utilizado como herramienta de cómputo por diferentes grupos de investigación dentro de la Facultad de Ingeniería. En las siguientes subsecciones se describen algunas de las aplicaciones de simulación numérica de flujos para las cuales se ha utilizado el cluster.

Modelos numéricos del Río de la Plata

Se implementó un modelo tridimensional baroclínico (Fossati et al, 2009), para representar todas las características del flujo en la zona del Río de la Plata y su frente marina, basado en un modelo numérico MOHID que utiliza una aproximación a través de volúmenes finitos para discretizar el dominio del problema, Ferziger y Peric (Ferziger, 2002) presenta una cobertura exhaustiva del tema. MOHID fue desarrollado por investigadores del Instituto Superior Técnico de la Universidad Técnica de Lisboa, Portugal. La aproximación basada en volúmenes finitos permite usar coordenadas verticales genéricas, dependiendo del proceso principal del área a estudiar o investigar. Una de las principales características de MOHID es que implementa la metodología de modelos encajados. A través de esta estrategia es posible anidar grillas de resolución espacial de manera creciente, forzando los modelos locales con resultados de aplicaciones de mayor escala. De esta forma, el modelo permite estudiar áreas cada vez más cercanas a la región de interés, a partir del pasaje de las condiciones de borde del modelo "padre" hacia el modelo anidado ("hijo"), aumentando en general la resolución. Una particularidad importante del modelo es que permite utilizar diferentes pasos de tiempo en los diferentes modelos encajados, restringiendo únicamente que los pasos de tiempo del modelo padre sean múltiplos de los pasos de tiempo de los modelos hijos. Al utilizar la estrategia de modelos encajados es posible la ejecución en paralelo de dichos modelos. El paradigma de paralelismo utilizado por el framework MOHID es el de pipeline, implementado en forma distribuida a través del paradigma de pasaje de mensajes utilizando el estándar MPI (Message Passing Interface). Es importante hacer notar que la opción de emplear la versión paralela, obliga a que los pasos de tiempos de los diferentes modelos sean iguales (entre el padre y el hijo) para que no ocurran incoherencias en los valores que estén simulando los diferentes procesos. Además el modelo permite la utilización de paralelismo en memoria compartida mediante la utilización de directivas de computación openMP (open MultiProcessing). En la Figura 3 se presenta un ejemplo de anidamiento de grillas para aumentar la relación en las zonas de Montevideo y Punta del Este.

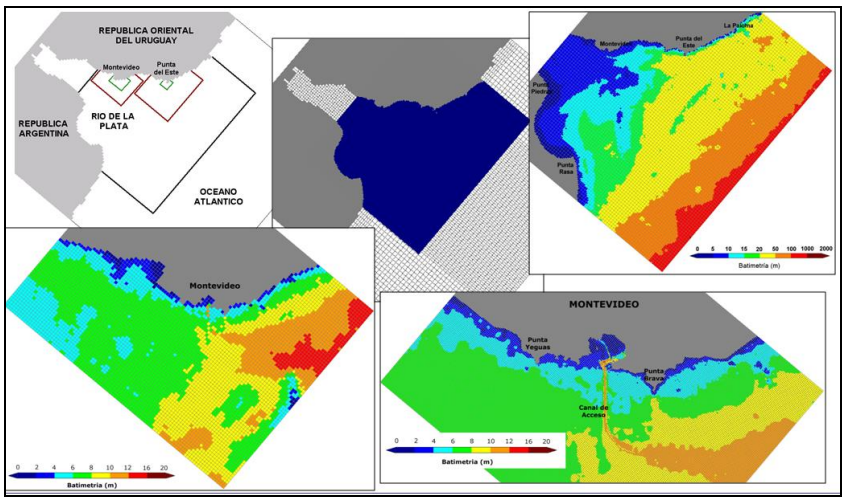

Figura 3.- Grillas anidadas para modelar el Río de la Plata

Los estudios preliminares del desempeño computacional del modelo en el Cluster FING mostraron que el beneficio al explotar el paralelismo a nivel de memoria compartida solo aporta beneficios marginales en el tiempo de cómputo. Mientras que al utilizar el paralelismo a través de la utilización de memoria distribuida mostró que en escenarios donde los distintos modelos de la jerarquía están balanceados (en cuanto al costo computacional) se alcanzan casi un beneficio lineal, trabajando con cuatro dominios anidados (máximo permitido). A su vez, un uso combinado de ambas técnicas permite obtener un beneficio muy alentador. Esto, agregado a la aceleración debido al poder de cómputo de cada procesador, implica disminuciones de tiempos muy significativos, permitiendo abordar problemas de dominios de mayor escala y utilizando grillas de mayor precisión.

Simulación numérica de flujos a superficie libre

Se experimentó con el modelo caffa3d.MB (Usera et al., 2008). El modelo caffa3d.MB es una implementación del método de volúmenes finitos aplicado a la simulación numérica de flujos viscosos o turbulentos, tridimensionales y con transporte de escalares. caffa3d.MB utiliza un interpolador lineal mejorado que permite construir las aproximaciones discretas a las ecuaciones de movimiento del fluido, y representa la geometría del dominio de cálculo mediante mallas curvilíneas estructuradas por bloques. Las grillas estructuradas por bloques permiten describir dominios con geometrías complejas, para los cuales es difícil construir una malla completamente estructurada (de bloque único) con buenas propiedades. La interfaz entre bloques se maneja de forma implícita, al igual que el resto del dominio, evitándose de esta manera una degradación de la convergencia y de la eficiencia computacional del método. Las celdas de una interfaz entre dos bloques pueden asociarse “una a una” o en la estrategia de “muchas a una”, definiendo bloques que permiten refinar localmente la malla. Complementariamente, el modelo permite incorporar en el dominio de cálculo piezas rígidas en movimiento y estudiar su interacción con el fluido, utilizando interfaces deslizantes entre bloques.

Para mejorar la eficiencia computacional del modelo caffa3d.MB y permitir la simulación de escenarios realistas, se incorporó el uso de la computación paralela en multiprocesadores de memoria compartida, aprovechando la descomposición del dominio en bloques y mediante la introducción de directivas de compilación openMP (open Multi-Processing). Aunque el esquema de openMP carece de la flexibilidad del paradigma de pasaje de mensajes (que permitirían realizar una implementación distribuida), provee un método conceptualmente simple y de eficiencia comprobada en sistemas simétricos (SMP) (Silberschatz et al., 2008). La técnica de descomposición de dominio puede adaptarse de modo sencillo al esquema de openMP en el caso de grillas estructuradas que propone el modelo caffa3d.MB. La estructura de datos diseñada para soportar grillas estructuradas se utiliza como base para distribuir las tareas de cómputo a diferentes procesadores en un computador simétrico. El esquema de paralelismo contempla ejecutar en un único procesador las tareas asociadas a las interfaces entre bloques, para evitar inconvenientes en el acceso simultáneo a datos en la memoria compartida.

Para la validación del modelo se ha experimentado con dos casos de estudio relacionados con la rotura de una presa en un escenario sin obstáculos, y en un escenario donde el agua choca con un obstáculo cilíndrico. En el primer caso, una región de agua inicialmente en reposo y en contacto con el aire, comienza a derramarse al eliminarse la barrera que la contiene. El escenario tiene una relativa simplicidad geométrica, pero entraña dificultades en la simulación debido a la desaceleración abrupta de la masa de agua al golpear con la pared opuesta. La Figura 4 muestra la evolución del sistema simulado en diversos instantes de tiempo.

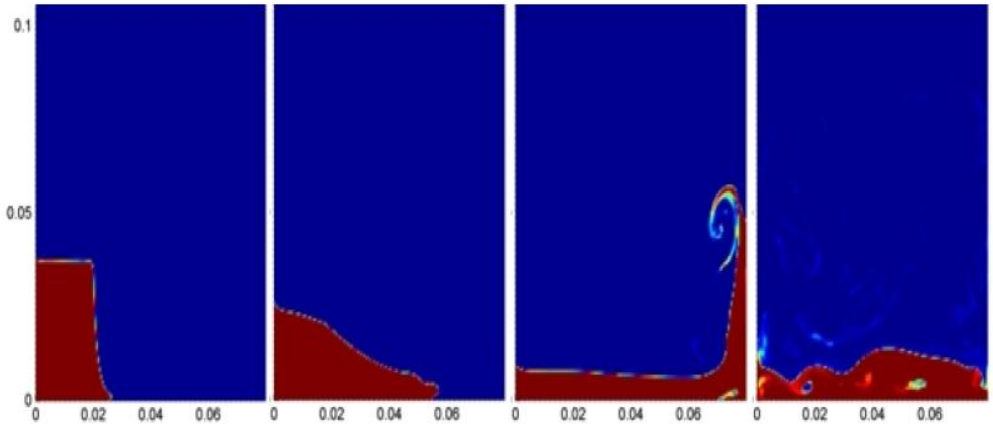

Figura 4.- Evolución del agua (zona roja) al derramarse, desplazar el aire (zona azul) y golpear con la pared opuesta

La aceleración obtenida utilizando las técnicas descriptas ha sido del orden de 4 a 6 utilizando ocho nodos de cómputo. Este resultado corrobora el valor 5,6 obtenido en la ejecución del benchmark HPCC descripto en la sección anterior.

A continuación se mencionan otras aplicaciones de la simulación numérica de fluidos donde se ha utilizado la infraestructura:

- Simulación numérica de la capa límite turbulenta sobre bloques de edificios: se analizaron las estructuras de flujo que gobiernan la dispersión contaminante en los 'desfiladeros urbanos' formados entre los bloques de edificios, en función de la orientación del viento, para estudiar la dispersión de contaminantes atmosféricos.
- Pronósticos climáticos intersectoriales con modelos numéricos de atmósfera: se realizaron simulaciones atmosféricas globales, pronósticos de temperatura de superficie de mar globales, y pronósticos de precipitaciones.
- Simulación de energía eléctrica aplicado a la optimización de los recursos energéticos hidráulicos del país. El uso de técnicas de programación paralela y del Cluster FING ha permitido reducir significativamente los tiempos de simulación.

Conjuntamente, el Cluster FING ha sido aplicado en proyectos de investigación en múltiples áreas de la ingeniería, como ejemplo: simulación numérica de fluctuaciones ciclo a ciclo en motores Otto, planificación de entornos de cómputo heterogéneos con algoritmos evolutivos, paralelismo aplicado a algoritmos cuánticos de búsqueda, búsqueda masiva de grafos de gran orden con grado y diámetro fijo.

Actividades de formación

En el desarrollo del objetivo de acercar a los investigadores de diferentes áreas al uso de este tipo de tecnologías se realizaron las siguientes actividades de formación:

- Cursos de nivelación donde se presentó a los investigadores de diferentes áreas herramientas básicas para la programación de aplicaciones.
- Curso de posgrado "Computación de Alta Performance" dictado por el Instituto de Computación de la Facultad de Ingeniería.
- Realización del primer "Seminario Multidisciplinario de Computación Científica de Alto Desempeño" donde se presentaron tutoriales de la programación paralela sobre memoria compartida y distribuida, y se brindaron charlas sobre aplicaciones que se desarrollaron sobre la infraestructura.

CONLUSIONES Y TRABAJO FUTURO

El Cluster FING ha cumplido los objetivos para los cuales fue implementado. La Facultad de Ingeniería dispone de un cluster dedicado enteramente a la investigación científica multidisciplinaria. Este cluster está compuesto con nodos de cómputo de alto desempeño de última tecnología, interconectados mediante una red de alto rendimiento, para la ejecución eficiente de modelos y simulaciones numéricas. Conjuntamente, se han brindado cursos, charlas y seminarios a investigadores de diferentes áreas que han permitido acercarlos a este tipo de tecnología y fortalecer los vínculos entre los investigadores de las diferentes áreas.

El proyecto se marcó dos políticas principales de funcionamiento: i) mantener el equipamiento actualizado, esto debido a la dificultad que puede implicar migrar importantes modelos a este tipo de arquitecturas y los constantes avances de las nuevas tecnologías de hardware, es necesario brindar perspectivas a los investigadores que van a disponer de un equipamiento equivalente al actual y ii) auto-sustentarse, de forma de no depender de ningún organismo para cumplir la primera política. Estas dos políticas han sido llevadas adelante en forma exitosa, como así lo indica la incorporación de equipamiento a principios de 2010 que potenció al cluster permitiendo alcanzar un pico teórico de 1,3 TeraFlops.

La infraestructura ha permitido mejorar el desempeño en la resolución de problemas abordados por diferentes áreas de investigación, en particular problemas de simulaciones numéricas de flujo. A su vez, la disposición de un mayor poder de cómputo ha sido un motivador para el abordaje de problemas más complejos

REFERENCIAS

Ferziger, J. H. and Peric, M. (2002). *Computational Methods for Fluids Dynamics, 3rd Edition. Springer.*

Fossati, M. and Fernández, M. and Piedra-Cueva, I. (2009). *Implementation of a 3D Lagrangian Model for evaluating submarine outfalls in the Rio de la Plata coastal area*. *En 33rd IAHR Congress - Water Engineering for a Sustainable Environment*, Canada.

Gepner, P. and Fraser, D. and Kowalik, M. (2008). *Second generation Quad-Core Intel Xeon processors bring 45nm technology and a new level of performance to HPC applications. Proceedings of the 8th international conference on Computational Science, Part I, pp. 417-426.* Krakow, Poland.

GMS (2010). *Ganglia Monitoring System. Disponible en* http://ganglia.sourceforge.net. *Consultado en Julio de 2010.*

Golub, G. and Van Loan, C. (1996).*Matrix Computations, 3rd Edition. The Johns Hopkins University Press.*

HPCC (2010). *High Performance Computing Challange. Disponible en* http://icl.cs.utk.edu/hpcc. *Consultado en Julio de 2010.*

Linholm, E. and Nickolls, J. and Oberman, S. and Montrym, J. (2008). *NVIDIA Tesla: A Unified Graphics and Computing Architecture.IEEE Micro Vol.28, pp. 39-55.* EUA.

MPI (1993).*Message Passing Interface Forum. MPI: A Message Passing Interface. In Proc. of Supercomputing '93, pages 878–883. IEEE Computer Society Press*.

Nesmachnow, S. and Salsano, E. (2007). *Evaluación del desempeño computacional del cluster Medusa*. *Reporte Técnico RT 07-12, InCo, Universidad de la República*, Uruguay.

OSCAR (2010). *Open Source Cluster Application Resource. Disponible en* http://svn.oscar.openclustergroup.org/trac/oscar. *Consultado en Julio de 2010.*

ROCKS (2010).*Rocks Cluster Distribution. Disponible en* http://rocksclusters.org. *Consultado en Julio de 2010.*

Silberschatz, A. and Galvin, P.B. and Gagne, G. (2008). *Operation System Concepts, 8th Edition. John Wiley & Sons, Inc.*

Usera G., Vernet, A., Ferré, J. (2008). *A Parallel Block-Structured Finite Volume Method for Flows in Complex Geometry with Sliding Interfaces. Flow Turbulence and Combustion Journal 8, pp. 471-495.*

TORQUE (2010). *Torque Resource Manager. Disponible en* http://www.clusterresources.com/products/torque-resource-manager.php. *Consultado en Julio de 2010.*

TOP500 (2010) *Top500 Supercomputing Sites. Disponible en* www.top500.org. *Consultado en Julio de 2010.*