Integración de Computación Heterogénea con Hadoop para Cloud Computing

Nelson Rodríguez[1], María Murazzo[2], Daniela Villafañe[3], Maximiliano Alves[4], Diego Medel[5]

Departamento e Instituto de Informática - F.C.E.F. y N. - U.N.S.J.

Complejo Islas Malvinas. Cereceto y Meglioli. 5400. Rivadavia. San Juan

0264 – 4234129 Fax: 0264-4234980

[1]nelson@iinfo.unsj.edu.ar [2]marite@unsj-cuim.edu.ar [3]villafane.unsj@hotmail.com
[4]maximilianoalvespinheiro@gmail.com [5]mdiego88@gmail.com

Resumen

La potencia de procesamiento de las supercomputadoras de ayer está ahora disponible en el desktop, aunque la necesidad de mayor poder computacional para resolver los problemas más grandes continúa creciendo. Los sistemas de cluster escalables hoy tienen la promesa de ejecución ilimitada a un costo muy conveniente.

El paralelismo constituye una alternativa real para reducir el tiempo de ejecución de las aplicaciones. Arquitecturas paralelas homogéneas con elementos de proceso de características similares están siendo utilizadas con éxito en distintos ámbitos de la ciencia y de la industria. De forma natural, la ampliación y la potenciación de un sistema homogéneo con nuevos elementos, deriva en un sistema de naturaleza heterogénea en el que los nuevos componentes presentan características diferenciales con los anteriores.

Este modelo de computación presenta nuevos desafíos fundamentalmente para integrarse con tecnologías diversas como la computación distribuida, Internet, HPC, data centers y bases de datos entre otras.

Palabras clave: *Computaciòn heterogénea, Cluster, Hadoop, GPGPU, Cloud Computing*

Contexto

El presente trabajo se encuadra dentro del área de I/D Innovación en Sistemas de Software, y se enmarca dentro del proyecto de investigación Implantación de un ambiente de Cloud Computing para integración de recursos, el cual tiene como unidades ejecutoras al Departamento e Instituto de Informática de la FCEFyN de la UNSJ.

Durante 2011 y 2012 se llevó a cabo el mencionado proyecto que continúa por un año más. Como consecuencia del mismo, se vislumbró la necesidad de contar con la arquitectura de soporte necesarias para montar las distintas capas de la arquitectura IaaS, PaaS y SaaS. En el caso de IaaS, aparece HPC como la alternativa más atractiva para trabajar en ámbitos académicos y de investigación, por todo lo que puede ofrecer, y por supuesto sobre ella montar las otras capas de la arquitectura. Para se ello se analizaron diversas posibilidades, de las cuales la computación heterogénea y Hadoop han resultado las opciones seleccionadas.

Introducción

Con el uso masivo de Internet, el surgimiento de tecnologías asociadas, accesos desde dispositivos variados y nuevos servicios en tiempo real, ha crecido notablemente la demanda de aplicaciones de computadoras. Cloud Computing ha resultado un multiproveedor de servicios, que comparte información, software y recursos abiertos dentro de un ambiente basado en Internet.

El fenómeno comúnmente conocido como Cloud Computing representar un cambio fundamental en el sentido en que los servicios de tecnología de la información (IT) son desarrollados, desplegados, actualizados, escalados, mantenidos, y pagados [1].

En los últimos años, con la popularización del concepto y aplicación de Cloud Computing, surgieron diversas tecnologías para darle soporte. En particular Hadoop, es un modelo de computación opensource para Cloud Computing, y es tenido en cuenta cada vez más por la academia y los círculos industriales [2].

La computación heterogénea se refiere a sistemas que utilizan una variedad de diferentes tipos de unidades de computación. Una unidad de cálculo puede ser un procesador de propósito general (GPP) y un procesador de propósito especial (como DSP, GPU o FPGA). En general, una plataforma de computación heterogénea consiste en procesadores con diferentes arquitecturas de conjuntos de instrucciones.

Las arquitecturas heterogéneas a nivel de nodo han resultado atractivas durante la última década por varias razones comparadas con la CPU tradicional, estas ofrecen alto niveles de performance y son eficientes en consumo de energía y costos [3]. Así pues, es muy frecuente encontrar arquitecturas paralelas donde las características de los elementos que componen el sistema (los procesadores, la memoria, la red,...) pueden ser diferentes.

A finales del siglo XX, en la década de los 90, con la consolidación de estándares para la programación en arquitecturas de paso de mensaje como PVM y MPI, la programación de sistemas heterogéneos se convierte en un problema tan sencillo o complicado como pueda serlo la programación de arquitecturas homogéneas. Sin embargo, aunque la portabilidad de los códigos está garantizada, no ocurre lo mismo con el rendimiento observado en las aplicaciones.

Aparecen nuevas situaciones y cuellos de botella que no pueden ser resueltos mediante la aplicación directa de los modelos y las técnicas conocidas para el caso homogéneo. Se hace necesario adaptar los métodos conocidos y en muchos casos diseñar nuevas estrategias comenzando desde cero. Aparece un conjunto importante de problemas abiertos que están siendo intensamente estudiados y analizados [4].

Dentro del ámbito de la investigación científica, siempre ha existido la necesidad de tener computadoras con grandes capacidades de procesamiento (del orden de los Teraflops) y grandes capacidades de almacenamiento (del orden de los Petabytes). El aporte que ofrecen estos tipos de computadoras es de crucial importancia debido a que permiten realizar números de cálculos imponentes que se desarrollan a nivel de investigación como son las simulaciones en producción industrial, computología examinando bugs en grandes programas, explorando propiedades ferromagnéticas en la física, analizar dinámica molecular a nivel químico, etc. El área informática que lleva a cabo este tipo de procesamiento se llama HPC (High Performance Computing).

La dificultad de la aplicación de esta área a los ambientes de investigación radica en el costo del equipamiento que se necesita para obtener esas capacidades de procesamiento y almacenamiento, antes mencionadas, sobre todo en la formación de recursos humanos en el ambiente académico. Para poder solucionar este problema, se han desarrollado algunas áreas dentro de HPC que ofrecen una solución con costos accesibles y de excelentes resultados tanto para almacenamiento como procesamiento, estos son la Computación Heterogénea y Hadoop.

El concepto de Computación Heterogénea se aplica, en ocasiones, a sistemas compuestos por diferentes tipos de PCs y

máquinas con múltiples procesadores conectados mediante redes. Debido a las diferencias entre las máquinas que forman el sistema, es probable que las velocidades de cómputo de los procesadores sean distintas y los tiempos de transferencia de datos también pueden ser diferentes en las comunicaciones entre cada par de procesadores. La naturaleza de esta red es inherentemente dinámica y depende de qué máquinas se utilicen en cada momento para resolver un problema y cuales sean sus características (capacidades de cómputo, memoria, comunicaciones, etc.). La programación dependiente de la arquitectura de la máquina supone una dificultad adicional en este tipo de sistemas. Otra de las desventajas que aparecen en los sistemas heterogéneos se debe al comportamiento de los tiempos de ejecución de las aplicaciones.

Algunos autores definen La informática híbrida, la cual representa en la actualidad la intersección de tres paradigmas ampliamente usados para infraestructura de computación: (1) HPC (tradicional) centrada en el propietario; (2) la computación Grid (compartición de recursos); (3) Cloud Computing (provisión de recursos y servicios bajo demanda) [5]. A esto, para la propuesta presentada, se le debe agregar la capacidad computacional de GPU.

Cada paradigma está caracterizado por un conjunto de atributos de la generación de recursos hasta la infraestructura y de las aplicaciones ejecutando en esa infraestructura.

Los objetivos de este trabajo se centran en el desarrollo de una plataforma de HPC en Cluster que integre varios recursos de forma de poder analizar, evaluar y estudiar diversos aspectos y medidas como performance e interoperabilidad entre otros.

Líneas de investigación y desarrollo

El rendimiento conseguido con un multiprocesador o multicomputador paralelo suele ser más elevado que el obtenido en un entorno de máquinas heterogéneas; los usuarios pueden verse obligados a aceptar en algunos casos una reducción del rendimiento de sus aplicaciones a favor de una gran reducción en el coste del sistema. Una de las razones para que esta situación se produzca, es que la mayoría de los programas paralelos han sido desarrollados bajo la hipótesis de trabajar sobre una arquitectura homogénea [6]. La computación Heterogénea se encuentra dentro del contexto de HPC y es una solución potente, debido a que la computadora que lidera el TOP 500 de las maquinas con mayor capacidad de procesamiento del mundo utiliza esta tecnología [6]. Lo que se pretende es combinar la escalabilidad de los clúster con MPI y la gran capacidad de procesamiento de las placas GPGPU[7] con CUDA[8] que son dos modelos prácticamente ortogonales pero que combinados pueden alcanzar niveles colosales de procesamiento. Este modelo sirve para procesar grandes cantidades de cómputo, los cuales en la mayoría de las ocasiones va acompañado de una gran cantidad de datos, para los cuales la computación heterogénea no brinda un soporte muy estable, sino que su mayor preocupación es la cantidad de instrucciones por segundo que se pueden ejecutar.

HADOOP [9] segunda solución que se planteó ante la problemática citada anteriormente, es un framework que da soporte a aplicaciones paralelas del tipo MAP-REDUCE [10]. La misma ofrece no solo un motor de procesamiento, sino un sistema de archivos distribuido llamado HDFS (Hadoop Distributed File System), que fue diseñado para brindar un soporte de almacenamiento fiable sobre un gran número de máquinas en un gran clúster. Este FS (File System) maneja tolerancia a fallos y balanceo de carga, lo que lo convierte en una poderosa herramienta para el manejo de grandes volúmenes de datos.

La adopción de procesamiento de las GPUs heterogéneas en algunos de los sitios más importantes del mundo HPC indica que este paradigma se está moviendo más allá de la fase experimental, y las GPU son cada vez más confiables. Tanto como el hardware de la

GPU y las tecnologías de software avancen, a medida que más estudiantes universitarios y otras personas puedan aprender cómo aprovechar las GPUs y tarjetas gráficas, IDC cree que las GPU tendrá un papel cada vez más importante en el mercado mundial de HPC, como complemento de los procesadores x86 dentro del ecosistema de alto rendimiento. En cuanto a la combinación de las tecnologías computación Heterogénea y HADOOP puede ser un aporte de gran aceptación en el ambiente científico y académico, ya que ofrece un soporte a problemáticas investigativas de áreas científicas.

La arquitectura resultante permitirá realizar análisis de escalabilidad y rendimiento. A pesar de que ambos conceptos se confunden son muy diferentes. Caso testigo como el ocurrido con la migración de servidores del sitio web Amazon.com minorista web, que pasó de 300 transacciones por segundo (TPS) a tan sólo 3 TPS cada uno después de mudarse a una arquitectura más escalable. La ventaja es que, si bien todos los servidores web pueden tener un menor rendimiento individual, el sistema en su conjunto se convirtió en mucho más escalable y nuevos servidores web se podría añadir el infinito.

En un futuro Aspectos de eficiencia, extensibilidad, disponibilidad, reusabilidad y personalización son sin duda otros problemas a resolver. Además por la naturaleza distribuida de los datos, resulta dificultoso mantener la consistencia y coherencia de datos, contar con una representación homogénea de los mismos (en una plataforma heterogénea), eficiencia en la sincronización entre las plataformas (distribuida y paralela).

La programación y la paralelizaciòn de algoritmos para procesamiento heterogéneo, es un problema muy difícil de resolver, con lo cual la verificación y validación de programas de este tipo, parece por ahora imposible de lograr.

Resultados y Objetivos

Resultados Obtenidos

Los trabajos publicados más relevantes de mencionan en [12 a 20], los cuales se enmarcan fundamentalmente en el área de Cloud Computing. También se llevaron a cabo trabajos de divulgación.

Además se ha aprobado cuatro (4) tesinas sobre soluciones a distintos problemas que presenta Cloud Computing.

Resultados Esperados (Objetivos)

El objetivo del grupo de investigación es la construcción de una plataforma para Cloud Computing basada en Cluster que integre la computación de alta performance con servicios como Hadoop.

Se espera que una vez realizada dicha plataforma, se puedan realizar estudios sobre aspectos como la eficiencia, balance de cargar e interoperabilidad, y proponer las soluciones necesarias para resolver las fallas que ocurran, tratando de favorecer el uso de los estándares abiertos.

Formación de Recursos Humanos

El proyecto marco sobre el que se realizan las investigaciones ha tenido como objetivo la interoperabilidad en Cloud Computing, a partir del mismo se han realizado publicaciones y trabajos de divulgación donde participan alumnos, becarios e investigadores, además cuatro tesina de licenciatura, una de tecnicatura en programación y otra de programador universitario. Por otro lado hay 2 (dos) tesinas de tecnicatura en desarrollo y 3 (tres) tesinas de licenciatura, y se espera realizar alguna tesis de maestría y aumentar el número de publicaciones. Se encuentran desarrollando sus proyecto dos becarios de iniciación a la investigación, que se enfocan en las líneas de investigación presentadas. Por otro lado también se prevé la divulgación de varios temas investigados por medio de cursos de postgrado y actualización.

Referencias

[1] Marston, Li, Bandyopadhyay, Zhang, Ghalsasi. “Cloud computing — The business perspective”. Decision Support Systems 51 (2011) 176-180. Elsevier. 2011.

[2] Lu, Hai-shan, Ting-ting.”Research on Hadoop Cloud Computing Model and its Applications”.2012. Third International Conference on Networking and Distributed Computing

[3] Brodtkorba,, Dykena, Hagena, Hjelmervika, Storaasl.. “State-of-the-art in heterogeneous computing”. Scientific Programming 18 (2010) 1–33.

[4][6] Moreno de Antonio. “Computación paralela y entornos heterogéneos”. Soportes Audiovisuales e informáticos. Serie Tesis Doctorales. Servicio de Publicaciones. Universidad de la laguna. Curso 2004/05. Ciencias y tecnologías/23. ISBN.: 84-7756-662-3.

[5] Mateescua, Gentzsch, Ribbens. “Hybrid Computing—Where HPC meets grid and Cloud Computing”. Future Generation Computer Systems 27 (2011) 440–453

[7] Olexandr Isayev. “Computación Heterogénea: Nuevo Paradigma Para La Era Exaescala”. Paralelizados.com. Comunidad de usuarios de HPC. GPUScience. IDC-Exascale-Executive-Brief_Nov2011. Noviembre 23, 2011 ·

[8] [9] Programación paralela facilitada. NVIDIA Corporation.http://la.nvidia.com/object/cuda_home_new_la.html. 2013.

[10] Hadoop. Welcome to Apache Hadoop. http://hadoop.apache.org/

[11] J. D.Sanjay Ghemawat. “Google. MapReduce: Simplified Data Processing on Large Clusters”.OSDI'04: Sixth Symposium on Operating System Design and Implementation,San Francisco, CA, December, 2004.http://research.google.com/archive/mapreduce.html

[12] Murazzo, Rodríguez. "Mobile Cloud Computing". WICC 2010. Calafate. Mayo 2010.

[13] Murazzo, Millán, Rodríguez, Segura, Villafañe. Desarrollo de aplicaciones para Cloud Computing. CACIC 2010. Morón. Oct. 2010.

[14] Murazzo, Rodríguez, Millán, Segura y Villafañe."Plataformas Educativas Implementadas Con Cloud Computing". XVI CACIC 2010, Workshop de Tecnologías Informáticas Aplicadas a la Educación. Morón. Oct. 2010.

[15] Murazzo, Rodríguez. "Una propuesta para el desarrollo de aplicaciones para Mobile Cloud Computing". Congreso Internacional de Computación y Telecomunicaciones – COMTEL 2010, Lima, Perú. Oct. 2010.

[16] Rodríguez, Murazzo, Ene. "Cloud Computing". WICC 2009. San Juan. Nov 2009.

[17] Rodríguez, Chávez, Martín, Murazzo, Valenzuela. “Interoperabilidad en Cloud Computing”. WICC 2011. Rosario 2011.

[18] Chávez, Martín, Rodríguez, Murazzo, Valenzuela Metodología AGIL para el desarrollo SaaS. WICC 2012. Posadas. 2012.

[19] Rodríguez, Valenzuela, Chávez, Martín, Murazzo, Villafañe. “Ambiente de desarrollo para lengua de señas basado en Cloud”. WICC 2012. Posadas. 2012.

[20] Rodríguez, Villafañe, Murazzo, Gallardo, Tarrachano. "GAE, una estrategia para complementar SaaS y PaaS a traves de la Web". 2do SABTIC. Tres de Maio, Brasil. Agosto 2012.