Identificación de Algoritmos de Cómputo Intensivo para Big Data y su Implementación en Clouds

Maria A. Murazzo[*], Nelson R. Rodriguez[*], Miguel Guevara[&], Fernando G. Tinetti[#]

[*]Departamento e Instituto de Informática – FCEFy N, UNSJ.
[&]Alumno avanzado de la carrera Lic. en Sistemas de Información
[#]III-LIDI, Fac. de Informática, UNLP – Investigador CIC Prov. Bs. As.

marite@unsj-cuim.edu.ar, nelson@iinfo.unsj.edu.ar,
migueljoseguevaratencio@gmail.com, fernando@info.unlp.edu.ar

Resumen

Almacenar, transferir y procesar grandes volúmenes de datos en el área que se ha denominado *Big Data* son un factor determinante y un reto para el Cómputo de Alto Rendimiento (sigla en inglés: HPC de *High Performance Computing)*. Los algoritmos usados para procesar esos datos deben sacar provecho de las ventajas ofrecidas por el cómputo en la Nube (*Cloud*), mediante el uso de algoritmos que permitan agilizar/acelerar el cómputo de o con esos datos.

La conjunción de Big Data y HPC se suele enfocar en la paralelización del procesamiento mediante la distribución de los datos y la delegación del cómputo en los nodos de procesamiento de la plataforma. Estas arquitecturas de cómputo, que para el caso de la memoria distribuida eran tradicionalmente los clusters, se pueden migrar al Cloud. La migración permite montar clusters virtuales (*Cluster as a Service*) logrando un entorno auto-escalable dependiente de la carga de trabajo. Se propone la identificación y evaluación de un conjunto representativo de algoritmos usados en Big Data con énfasis en su implementación en clouds.

Palabras clave: Cloud Computing, Big data, HPC, Cluster Computing

Contexto

Este trabajo se encuadra dentro del área de I/D "Procesamiento Distribuido y Paralelo" y en particular dentro del proyecto de investigación "Evaluación de arquitecturas distribuidas de *commodity* basadas en software libre", el cual ha sido presentado en la última convocatoria de CICITCA, con una duración de dos años y que tiene como unidades ejecutoras al Departamento e Instituto de Informática de la FCEFyN de la UNSJ. El grupo de investigación viene trabajando en proyectos del área desde hace más de 15 años con varias publicaciones y formación de recursos humanos.

Introducción

Debido a la aparición de nuevas tecnologías, dispositivos, medios de comunicación y aplicaciones, la cantidad de datos que se produce en la actualidad aumenta exponencialmente. Se considera que el 90% de los datos existentes se han generado en los últimos dos años, esto está dando lugar a la Era del Exa y Zetta-Byte [1]. Este aumento en la cantidad de datos demanda nuevas estrategias que permitan su almacenamiento, y análisis de manera eficiente; esto conlleva un cambio

de paradigma en las arquitecturas de cómputo, los algoritmos y también los mecanismos de procesamiento.

Frente a esta problemática se ha popularizado el término Big Data [2], el cual se usa para describir grandes conjuntos de datos, los cuales exhiben las propiedades: variabilidad, variedad, valor, volumen, velocidad, y complejidad; todas ellas denotan datos multidimensionales. Estas propiedades hacen que los sistemas de cómputo convencionales sean muchas veces inapropiadas para lograr un procesamiento adecuado [3] [4].

Los sistemas de cómputo tradicionales han evolucionado a sistemas de alto rendimiento (HPC) para llevar a cabo cómputo intensivo y mejorar la velocidad de procesamiento. Estos entornos son los ideales para Big Data debido a la creciente demanda y también creciente complejidad del cómputo necesario [5].

En este aspecto el desafío se centra en cómo se aprovecha al máximo el potencial de la arquitectura física con o por el uso de algoritmos de cómputo intensivo. Para mejorar los tiempos de procesamiento en los algoritmos, se puede optar por la implementación en plataformas con mayor potencia de cálculo como las supercomputadoras, pero los costos son elevados, lo que dificulta su acceso a una gran cantidad de comunidades científicas.

Para resolver los problemas de costo, la computación distribuida es un modelo destinado a resolver problemas de cómputo masivo utilizando un gran número de computadoras organizadas sobre una de comunicaciones. De esta manera es posible compartir recursos (quizás heterogéneos), que pueden ser basados incluso en distintas plataformas, arquitecturas y lenguajes, situados en distintos lugares y pertenecientes a diferentes dominios de administración sobre una red que utiliza estándares abiertos [6].

En función de la problemática para la cual se decide montar una arquitectura distribuida, existen diferentes tipos de sistemas distribuidos: a) de cómputo distribuido, b) de almacenamiento y c) sistemas ubicuos distribuidos. Para el caso de esta línea de investigación se enfocarán los sistemas de cómputo distribuidos, los cuales permiten realizar de manera más eficiente tareas de computación de alto rendimiento basadas en el modelo de memoria distribuida [7]. Ejemplos de arquitecturas distribuidas para cómputo intensivo son los clusters y los clouds [8].

Según [9] cloud es un modelo de prestación de servicios informáticos cuya principal orientación es la escalabilidad. Esto es, que desde el punto de vista de los usuarios, los servicios son elásticos, o sea, pueden crecer o recuperar su tamaño original de manera rápida y sencilla. Esta orientación permite que los usuarios que acceden a los servicios, perciban que todo funciona de manera simple y rápida, dando como resultado una experiencia más gratificante.

Gracias a estas características, cloud se ha convertido en un enorme repositorio de recursos computacionales, lo cual es una buena posibilidad para construir una plataforma para las aplicaciones que necesitan una gran cantidad de recursos. Esta capacidad del cloud se debe principalmente a la habilidad de escalado elástico de recursos en función de las necesidades de las aplicaciones y el presupuesto del usuario. Esta es una tecnología centrada en ofrecer cualquier recurso (bases de datos, red, procesador, etc.) y ofrecerlo como un servicio (AaaS, Anything as a Service) bajo demanda, inclusive el cómputo [10].

Sin embargo la disponibilidad de un gran número de recursos que ofrece el cloud, también está disponibles en plataformas paralelas, las cuales ofrecen performance por sobre escalabilidad y despliegue inmediato de recursos, sin la necesidad de la virtualización

que produce una degradación del desempeño. A pesar de este inconveniente cada vez más las aplicaciones de HPC están migrando al cloud debido los aspectos económicos, que permiten contar con una plataforma configurable a las necesidades de las aplicaciones, mediante la implementación de *High Performance Computing as a Service (HPCaaS)* [11].

El paradigma cloud es atractivo para las aplicaciones HPC debido a que ofrece disponibilidad inmediata de recursos, un stack de software que puede seleccionar el dueño de la aplicación, elasticidad de los recursos, abstracción del hardware mediante la virtualización, lo cual provee portabilidad, entre las más importantes. Estos beneficios hacen que usar cloud como plataforma para los algoritmos de cómputo intensivo sea una alternativa atractiva desde el punto de vista de la performance y la eficiencia.

En los ambientes distribuidos, tales como cluster y cloud, un modelo de programación adecuado es MapReduce. MapReduce es parte de Hadoop [14], y se utiliza para desarrollar aplicaciones escalables y tolerantes a fallas en el cloud. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos

La principal motivación del modelo de programación MapReduce es la delegación del cómputo intensivo en cluster físicos o virtuales (Cluster as a Service) que, mediante un sistema de ficheros distribuido, reparte la carga de trabajo, optimizando tiempo y recursos. Asimismo, facilita un patrón de desarrollo paralelo para simplificar la implementación de algoritmos de cómputo intensivo en entornos distribuidos. Este modelo puede dividir un espacio grande de problema en espacios pequeños y paralelizar la ejecución de tareas más pequeñas en estos sub espacios [12] [13].

Líneas de Investigación, Desarrollo e Innovación

En función de lo explicado anteriormente, se consideran muy necesarias las siguientes líneas de investigación:

1. Identificar algoritmos necesarios, útiles, o al menos referenciados en big data. Caracterizar estos algoritmos al menos por área de aplicación o por tipo de procesamiento.
2. Implementar al menos algunos algoritmos en clouds. Esta implementación inicialmente puede considerar un cloud como la combinación de un cluster (en hardware y procesamiento) al que se tiene acceso vía una interfaz REST (*Representational State Transfe*r).
3. Agrupar/clasificar en bibliotecas los posibles algoritmos, tanto implementados como identificados. Evaluar su utilidad o área de aplicación.
4. Analizar factores que determinarán la relación costo/rendimiento y limitaciones, tanto de los algoritmos como de la o las bibliotecas propuestas.

Resultados y Objetivos

Resultados Obtenidos

Se han realizado varias publicaciones en esta línea de investigación, entre las que se destacan [15] [16] [17] [18] [19] [20] [21] [22] [23] [24] [25].

Como base para para comenzar con los trabajos de investigación se ha instalado una infraestructura cloud privada basada en OpenStack sobre la cual se está configurando Hadoop. Además, se cuenta con un cluster de seis nodos con Hadoop instalado para comenzar a realizar las investigaciones.

Como mínimo, se espera contar con una caracterización específicamente de rendimiento de los algoritmos en clouds, mostrando su rendimiento post-migración a este tipo de plataforma. Por otro lado, se espera caracterizar lo más objetivamente posible el tráfico hacia y desde el cloud para utilizar los algoritmos disponibles.

Objetivos

El principal objetivo de esta línea de investigación es tratar de identificar, proponer, implementar y determinar el rendimiento (al menos de manera experimental) de un conjunto de algoritmos de cómputo intensivo que se considere de importancia para aplicaciones de Big Data y que se orienten a ser usados en arquitecturas distribuidas.

Para realizar esta investigación, se tomaran como base las arquitecturas cluster y cloud, y sus posibles combinaciones, tales como Cluster as a Service.

Formación de Recursos Humanos

El equipo de trabajo está compuesto por los dos (2) docentes-investigadores de la F.C.E.F.y N. de la U.N.S.J. y un (1) docente-investigador de la Facultad de Informática de la U.N.L.P. y tres alumnos avanzados, dos de la Licenciatura en Ciencias de la Computación, y otro de la Licenciatura en Sistemas de Información.

Durante 2015 se rindió una tesina de grado y se están realizando dos (2) más que se rendirán en 2016.

Durante el año 2015 se realizó un (1) Trabajo Integrador de la Especialización en Redes y Seguridad (UNLP). Para el año 2016 se espera realizar una (1) tesis de la Maestría en Redes (UNLP) sobre esta temática. Además, un miembro del grupo se encuentra realizando el Doctorado en Ciencias de la Informática (UNSJ) cuya tesis aborda la linea de investigación aquí presentada.

Por otro lado también se prevé la divulgación de varios temas investigados por medio de cursos de postgrado y actualización o publicaciones de divulgación.

Referencias

[1] Acín, Bird, Boccali, Cancio, Collier, Corney, Fuhrmann, *Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms.* In Journal of Physics: Conference Series (Vol. 664, No. 4, p. 042009). IOP Publishing. 2015.

[2] Nyikes, Rajnai, Z. *Big data, as part of the critical infrastructure.* In Intelligent Systems and Informatics (SISY), 2015 IEEE 13th International Symposium on (pp. 217-222). IEEE. 2015

[3] Katal, A., Wazid, M., & Goudar, R. H. (2013, August). *Big data: issues, challenges, tools and good practices.* In Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.

[4] Chen, Mao, Liu, Y. (2014). *Big data: a survey.* Mobile Networks and Applications, 19(2), 171-209.

[5] Kashyap, Fewings, Davies, Morris, Green, Guest. *Big Data at HPC Wales*. arXiv preprint arXiv:1506.08907. 2015.

[6] Kahanwal, Singh. *The distributed computing paradigms: P2P, grid, cluster, cloud, and jungle*. arXiv preprint arXiv:1311.3070. 2013.

[7] Tanembaum, Van Steen. *Sistemas Distribuidos, Principios y Prácticas.* Pearson Prentice Hall. 2008.

[8] Antonopoulos, Gillam. *Cloud Computing; Principles, Systems and Applications.* Editorial Springer Science & Business Media. 2010.

[9] Mell, Grance. *The NIST definition of cloud computing*. NIST Special Publication 800 – 145. 2011.

[10] Chee, Franklin. *Cloud computing: technologies and strategies of the ubiquitous data center*. CRC Press. 2010.

[11] Petcu. *On Autonomic HPC Clouds*. Proceedings of the Second International Workshop on Sustainable Ultrascale

Computing Systems (NESUS 2015), Krakow, Poland. 2015.
[12] Malik, Sangwan. (2015). *Mapreduce Algorithms Optimizes the Potential of Big Data.* International Journal of Computer Science and Mobile Computing, Vol.4 Issue.6, June – 2015.
[13] Carrera, Geyer. *Modeling the Performance of MapReduce Applications for the Cloud.* Latin American Journal of Computing Faculty of Systems Engineering National Polytechnic School Quito-Ecuador, 2(2). 2015.
[14] Hadoop. Welcome to Apache Hadoop. http://hadoop.apache.org.
[15] Murazzo, Rodriguez, Chavez, Valenzuela, Martin, Guevara. *Despliegue de una arquitectura de Cloud Computing híbrida Open Source.* CONAIISI 2014.
[16] Rodriguez, Valenzuela, Murazzo, Martin, Chavez, Villafañe, González. *Cloud Computing con herramientas libres para evaluación de modelos de despliegue híbrido.* WICC 2014.
[17] Rodriguez, Murazzo, Medel, Fernandez, Gonzalez. *Evaluación de costos de comunicación en arquitecturas para computación heterogénea aplicadas a computación científica.* WICC 2014.
[18] Rodriguez, Valenzuela, Murazzo, Chavez, Martin, Villafañe, Gonzalez. *Análisis de los parámetros de performan-ce y escalabilidad para Clouds híbridos.* SABTIC 2014.
[19] Rodriguez, Murazzo, Chavez, Guevara. *Arquitectura de Cloud Computing híbrida basada en tecnología Open Source.* CACIC 2014.
[20] Martín, Chavez, Murazzo, Rodriguez, Valenzuela. *MongoDB en ambiente Cloud Híbrido con OpenStack.* WICC 2015.
[21] Rodriguez, Valenzuela, Murazzo, Chavez, Martin, Villafañe. *Estudio y Análisis de estrategias de seguridad para Cloud Computing híbridos.* WICC 2015.
[22] Murazzo, Tinetti, Rodriguez, Guevara. *Infraestrctura de Cloud Computing.* WICC 2015.
[23] Murazzo, Tinetti, Rodriguez. *Despliegue de una Infraestructura Cloud Privada de Código Abierto.* III JCC & Big Data 2015.
[24] Rodriguez, Murazzo, Medel Chavez, Martin, Valenzuela. *Análisis de mejora en la escalabilidad de las infraestructuras de cloud computing.* III JCC & Big Data 2015.
[25] Murazzo, Rodriguez. *Evaluación del Impacto de Migración al Cloud.* SABTIC 2015.