Implementación de Plataforma HPC Dinámica para la Resolución de Problemas de Alta Complejidad Computacional

Miguel Méndez-Garabetti[1,2], Javier Rosenstein[1], Maria A. Murazzo[3], Ailin Carribero[1], Pedro Orellana[1], Nelson R. Rodriguez[3], Miguel Guevara[3] y Pablo Gomez[3]

[1]Instituto de Investigaciones, Facultad de Informática y Diseño, Universidad Champagnat. Belgrano 721, 5501 Godoy Cruz, Mendoza, Mendoza, Argentina. {mendez-garabettimiguel},{rosensteinjavier},{orellanapedro}@uch.edu.ar, ailin@carribero.com.ar

[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

[3]Departamento de Informática – FCEFy N, UNSJ - CUIM (Complejo Islas Malvinas) Rivadavia, San Juan, Argentina marite@unsj-cuim.edu.ar, nelson@iinfo.unsj.edu.ar, migueljoseguevaratencio@gmail.com, pablo.gomez.allende@gmail.com

RESUMEN

En la actualidad existe gran cantidad de problemas que son resueltos de forma aceptable mediante la utilización de algoritmos secuenciales. Sin embargo, a medida que la complejidad del problema incrementa, las implementaciones secuenciales suelen volverse ineficientes hasta incluso obsoletas. Siendo necesario recurrir a la utilización de herramientas que permitan ofrecer resultados más eficientes. En este tipo de escenarios, la computación paralela se ha convertido en la forma tradicional de resolver gran variedad de problemas de alta complejidad computacional. Si bien existen diferentes niveles de paralelismo, la utilización de esquemas basados en múltiples computadoras es la más tipo. Sin embargo, el acceso a una infraestructura de alto rendimiento no siempre suele estar al alcance de todo tipo de instituciones. Debido a esto, en este trabajo se diseña una implementación de reutilización de equipamiento informático existente, para conformar un clúster HPC. La implementación de esta plataforma será la herramienta fundamental para poder trabajar de forma activa en el área de HPC, posibilitando al desarrollo y evaluación de nuevas técnicas de optimización.

Palabras clave: computo paralelo, HPC, cluster, optimización, metaheurísticas

CONTEXTO

El presente proyecto se desarrolla en el Instituto de Investigaciones de la Facultad de Informática y Diseño de la Universidad Champagnat (Godoy Cruz, Mendoza), en el marco de la Licenciatura en Sistemas de Información; en cooperación con el Departamento de Informática de la Facultad de Ciencias s de la Universi- San Juan).

Este trabajo es parte del proyecto de investigación que dio inicio en agosto de 2016 denominado “Diseño e implementación de una plataforma de computación paralela, dinámica y heterogénea, para el desarrollo de estrategias avanzadas de optimización aplicadas a la resolución de problemas de alta complejidad (FID-002/16)”.

1. INTRODUCCIÓN

El objetivo principal de la presente línea de investigación consiste en el desarrollo y mejora de técnicas de optimización para ser aplicadas en la resolución de problemas de alta complejidad computacional, tales como análisis, procesamiento y visualización de grandes volúmenes de datos. Trabajar con datos masivos, implica un gran desafío debido a la necesidad de explorar un universo de nuevas tecnologías, las cuales no sólo hacen posible la obtención y procesamiento de los datos sino también realizan su gestión en un tiempo razonable [1]. El tratamiento de datos a gran escala posee varias etapas, las cuales deben ser resueltas de manera eficiente y eficaz a fin de obtener información útil que de solución a los problemas. Cada una de estas etapas constituyen en si problemas computacional- mente costosos, para los cuales considerar el uso de nuevas técnicas y arquitecturas contribuirá a mejorar su rendimiento. Es por ello que la búsqueda y selección de técnicas de computación de altas prestaciones (HPC, [2]) en cada una de las etapas o procesos involucrados, permitirá resolver con eficiencia cada uno de sus objetivos.

HPC es la evolución de los sistemas de cómputo convencional, los cuales permiten realizar operaciones de cómputo intensivo y mejorar la velocidad de procesamiento. Esto involucra diferentes tecnologías, tal como los sistemas distribuidos y los sistemas paralelos; incluyendo clúster de computadoras, cloud computing, tarjetas gráficas y computadoras masivamente paralelas. Todos estos entornos son ideales para resolver aplicaciones científicas, computacionalmente costosas con manejo de grandes cantidades de datos, a fin de lograr resultados en menor tiempo.

Según [3], un clúster HPC es una colección de estaciones de trabajo autónomas o PCs, interconectadas entre sí por una red de alta velocidad que trabajan conjuntamente como un solo recurso informático integrado. Partiendo de esta definición podemos entender que un clúster posee tanto, componentes hardware: nodos o PCs y elementos de comunicación; y componentes software: como los sistemas operativos y las aplicaciones que conforman el entorno de programación [4],[5],[6]. Existen diferentes tipos de implementaciones HPC, aunque dentro de las soluciones que priman el aspecto económico, aquella que ofrece el menor costo es sin duda la que consiste en "reutilizar" el equipamiento existente. Es importante remarcar que con reutilización nos referimos a dar una segunda utilidad o función a una computadora que ya tiene un rol asignado. Por ejemplo, un conjunto de computadoras de un laboratorio de una universidad, el cual es utilizado para realizar operaciones de cálculo científico en la franja horaria que el mismo se encuentra ocioso. Si bien este tipo de soluciones no es el escenario ideal y habitual, presente en los grandes centros de HPC, suele ser una solución muy utilizada por instituciones que no pueden justificar el costo de adquirir un clúster HPC dedicado, ya sea por falta de presupuesto o por un bajo volumen de trabajo (e.g. generalmente esto ocurre en aquellos grupos de investigación que se están iniciando en el uso o aplicación de HPC). Si bien la reutilización de equipamiento informático es la solución más viable económicamente, trae consigo diversos retos y complicaciones, tanto desde el punto de vista logístico, administrativo, político, como así también técnico.

Construir una arquitectura de bajo costo y reutilizable, que sea eficiente en tiempo y uso de recursos, no es tarea sencilla. La complejidad de una arquitectura distribuida está dada debido a la diversidad de componentes que la misma incluye como son: servidores, almacenamiento, arquitecturas de software en capas, variedad de middlewares y redes, que deben trabajar conjuntamente de la forma más óptima

posible y que además deben ser configurados adecuadamente para ser utilizados bajo demanda.
Una vez que la plataforma de cálculo se encuentre implementada, será utilizada para diseñar y evaluar el rendimiento de nuevas técnicas de optimización, las cuales operen exclusivamente en ambientes paralelos. Como técnicas de optimización nos referimos al uso de metaheurísticas paralelas y sus diferentes esquemas de colaboración.
Según [7], las metaheurísticas son estrategias inteligentes, de propósito general, que tienen como objetivo diseñar y mejorar procedimientos heurísticos para resolver problemas de alta complejidad. Si bien las metaheurísticas implementadas de manera tradicional pueden resolver de forma eficiente gran cantidad de problemas, en ciertos casos es necesario recurrir a estrategias avanzadas. Como el caso de la hibridación de metaheurísticas utilizando esquemas paralelos, permitiendo de esta forma sumar las bondades y minimizar sus deficiencias, en pos de obtener resultados de mayor calidad.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

El presente proyecto está compuesto por dos etapas o fases de trabajo. Una de ellas corresponde al diseño e implementación técnica de la plataforma HPC, la cual será la base necesaria para poder realizar las actividades propias de la segunda etapa. Esta última corresponde a la línea de investigación (i.e., línea I+D) propiamente dicha. Una breve descripción de cada una de las etapas se describe a continuación:

1. La primer parte consiste en la implementación de un clúster que permita a cada equipamiento de uso administrativo y/o educativo de la institución, formar parte del conjunto de nodos de cálculo del clúster. Para ello se deberá determinar una adecuada configuración que permita operar con equipamiento completamente heterogéneo y de forma dinámica, sin afectar el fin principal que cada terminal tiene definido.

2. La segunda etapa consiste en el desarrollo y mejora de nuevas técnicas de optimización existentes. Los desarrollos consistirán en el estudio de diferentes metaheurísticas y sus diferentes combinaciones, con el objetivo de incrementar su robustez y capacidad de rendimiento. Dichas técnicas deberán ser evaluadas mediante su aplicación en diferentes problemas de alto costo computacional, que permitan evidenciar las capacidades de dichas técnicas.

3. RESULTADOS ESPERADOS

Como resultados esperados, los mismos pueden dividirse en dos grandes grupos:

1. *Plataforma HPC*, se espera obtener como producto final un esquema de configuración de equipos de red que permita utilizar cualquier equipamiento informático de la Universidad Champagant para conformar uno o más clusters HPC utilizando principalmente: MPI [8], OpenMP [9], C y C++ [10]. Inicialmente se trabajará solamente con el equipa- miento informático de dos laboratorios de informática, los cuales poseen las siguientes características:

   a. **Laboratorio 1**: 7 PCs con procesadores AMD Athlon 64x2 4600 con 2Gb RAM.

   b. **Laboratorio 2**: 18 PCs con procesadores i7 4770 con 8GB de memoria RAM y ATI Radeon HD6450 1Gb.

También está previsto trabajar con ecosistemas distribuidos sobre cluster, tales como Hadoop [11] y Spark [12], los cuales usan el paradigma MapReduce [13] para el procesamiento de datos masivos. Además, se prevé el estudio de bases de datos avanzadas (NoSQL [14] y NewSQL [15]) para el almacenamiento de los grandes volúmenes de datos.

2. *Técnicas de optimización*, en líneas generales se espera identificar las estrategias de colaboración y las distintas hibridaciones que ofrecen más eficiencia tras ser aplicados a problemas de alto costo computacional [16].

4. FORMACIÓN DE RECURSOS HUMANOS

La línea de I+D presentada está vinculada con el desarrollo de una tesina de grado, por parte del estudiante de la Licenciatura en Sistemas de Información de la Universidad Champagnat, Pedro Orellana. Dicha tesina se centra en el análisis e implementación de alternativas para la plataforma de HPC. Por parte de la misma institución, también se cuenta con la reciente participación de la estudiante de segundo año de la carrera: Ailin Carribero, quien pertenece al programa de Iniciación la Investigación Científica de la UCH. Dicho programa invita a participar en tareas de investigación a estudiantes desde los primeros años de la carrera.

Por parte del Departamento de Informática de la FCEFyN de la UNSJ se cuenta con la participación de dos estudiantes de grado, en instancia de tesis, ellos son: Miguel Guevara y Pablo Gomez, ambos de la carrera Licenciatura en Sistemas de Información.

Finalmente, una vez que la plataforma se encuentre implementada, la misma será utilizada como recurso para el dictado de talleres de computación paralela tanto para estudiantes de la universidad, como así también para aquellos interesados externos a la UCH.

5. BIBLIOGRAFÍA

1. Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co..
2. Hager, G., & Wellein, G. (2010). Introduction to high performance computing for scientists and engineers. CRC Press.
3. Jin, H., Buyya, R., & Baker, M. (n.d.). Cluster Computing Tools, Applications, and Australian Initiatives for Low Cost Supercomputing.
4. Gropp, Lusk and Sterling, (2003). "Beowulf Cluster Computing with Linux". The MIT Press, second edition,
5. Brown. "What's a Beowulf?". Engineering a Beowulf-style Compute Cluster. Physics Department. Duke University. http://www.phy.duke.edu/~rgb/brahma//beowulf_book/node9.html, (2003). Accedido en marzo de 2017.
6. Ferreira, Kettmann, Thomasch, Silcocks, Chen, Daunois, Ihamo, Harada, Hill, Bernocchi and Ford. (2001). "Linux HPC Cluster Installation". IBM Redbooks, first edition (ISBN: 9780738422787).
7. Glover F. (1986). Future paths for integer programming and artificial intelligence. Computers & Operations Research, 13(5):533–549.
8. Message P Forum. (1994). Mpi: a Message-Passing Interface Standard. Technical Report. University of Tennessee, Knoxville, TN, USA.
9. Dagum L. and Menon R. (1998). OpenMP: An Industry-Standard API for Shared-Memory Programming. IEEE Comput. Sci. Eng. 5, 46-55.

DOI=http://dx.doi.org/10.1109/99.660313

10. Schildt H. (2000). C/C++ Programmer's Reference (2nd ed.). McGraw-Hill, Inc., New York, NY, USA.
11. White, T. (2012). Hadoop: The definitive guide. O'Reilly Media, Inc.
12. Spark, A. (2016). Apache spark: A fast and general engine for large-scale data processing. http://spark. apache. org
13. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
14. Hecht, R., & Jablonski, S. (2011). NoSQL evaluation: A use case oriented survey. In Cloud and Service Computing (CSC), 2011 International Conference on (pp. 336-341). IEEE.
15. Stonebraker, M. (2012). Newsql: An alternative to nosql and old sql for new oltp apps. Communications of the ACM. Retrieved, 07-06.
16. Engineering and Physical Sciences Research Council, “International Review of Research Using HPC in the UK”, Engineering and Physical Sciences Research Council, (ISBN 1-904425-54-2), (2005).