Computación de Alto Desempeño y Datos Masivos: Arquitecturas, Modelos y Paradigmas.

Rubén Apolloni, Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Fabiana Píccoli, Marcela Printista, Cristian Tissera

LIDIC- Univ. Nacional de San Luís
San Luís, Argentina
{rubenga, mdbarrio, omlopres, ncmiran, mpiccoli, mprinti, ptissera}@unsl.edu.ar

Resumen

Este trabajo aborda las líneas de trabajo a seguir, con el objetivo de desarrollar e incluir técnicas de Computación de Alto Desempeño en problemas con datos masivos o problemas Big-Data. El desarrollo de la infraestructura, metodologías y herramientas de Computación de Alto Desempeño y su empleo eficiente en la solución de problemas que tratan grandes cantidades de datos, debe considerar la naturaleza de estos.

Palabras clave: Datos Masivos, Computación de Alto Desempeño, Arquitecturas Multicore y Many core

Contexto

Esta propuesta de trabajo se lleva a cabo dentro del proyecto de investigación "Tecnologías Avanzadas aplicadas al Procesamiento de Datos Masivos" y del proyecto binacional CAPG-BA 66/13 entre la Universidad Nacional de San Luis (UNSL) y la Universidad Federal de Pernambuco (UFPE), Recife, Brasil.

El proyecto de investigación se desarrolla en el marco del Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC), de la Facultad de Ciencias Físico, Matemáticas y Naturales de la UNSL y el Centro de Informática de la UFPE.

Introducción

Cuando se habla de "big data" [MCJ13] se hace referencia a un conjunto de grandes volúmenes, diversos, complejos, longitudinales o distribuidos de datos, generados desde transacciones en Internet, sensores, instrumentos, vídeos, mails, redes sociales y desde una variedad de fuentes digitales disponible en la actualidad y futuras. Las técnicas tradicionales para el procesamiento, análisis y obtención de información útil

deben ser redefinidas para formular nuevas metodologías de abordaje. Trabajar con grandes volúmenes de datos implica grandes desafíos, 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 [N13].

Para trabajar con datos a gran escala, es necesario tener en cuenta que la obtención de información útil será a partir de datos no estructurados como texto, audio, imagen y vídeo. Es por ello que se debe considerar la aplicación de nuevas metodologías para el procesamiento eficiente y eficaz de estos grandes volúmenes de datos. Además, como cada uno de los procesos involucrados en la obtención de la información a partir de dichos volúmenes implica un gran número de problemas computacionalmente costosos, el uso de nuevas técnicas y nuevas arquitecturas puede contribuir a mejorar su rendimiento; es por ello que la búsqueda y selección de técnicas de computación de altas prestaciones (HPC) en cada etapa o proceso involucrado permitirá resolver con eficiencia cada uno de los objetivos a plantearse.

Por todo lo expuesto, el procesamiento de grandes volúmenes de datos nos introduce en una nueva era de la computación, debido a que genera mayores demandas al procesador, a la memori en todos los niveles (tanto a memoria principal y memoria cache) [HP08], a los dispositivos de almacenamiento, y también requiere nuevas soluciones de software.

Entre las soluciones de software para este tipo de problemas se desarrollaron algoritmos y modelos de programación distribuidos y paralelos, como MapReduce[DG04], Hive[CWR12] e Impala[R13] que permiten procesar terabytes de información sin necesidad de cambiar las estructuras de datos subyacentes.

Entre los requerimientos de hardware, se encuentra la necesidad de mayor cantidad de almacenamiento para todos los datos, introduciendo nuevos desafíos de investigaciones y desarrollos. Además las aplicaciones requieren mayor capacidad de memoria, esperandose un aumento en la demanda. Otro aspecto a considerar es el consumo de energía, criterio muy importante a tener en cuenta en el diseño y desarrollo de sistemas de computadoras ya que está directamente relacionado con el consumo de energía total de la infraestructura computacional.

En la tecnología actual, la demanda de Computación de Alta Desempeño (HPC) se incrementa rápidamente. Con el continuo crecimiento de la Ley de Moore[G65], se observa una constante reducción del tamaño de los transistores, lo que permite empaquetar mayor cantidad de ellos en una única pastilla. Esto permite diseñar procesadores más potentes y/o incluir

mayor cantidad de núcleos dentro de una pastilla.

El continuo aumento en la cantidad de transistores posibilita el incremento del troughput, entre otras mejoras. Si bien este incremento permite construir procesadores más potentes. También, genera inconvenientes y limitaciones. Para continuar mejorando el desempeño del procesador, en la actualidad, se eligió diseñar procesadores de varios cores, más simples, dentro de una pastilla.

Con los sistemas de computación transformados desde procesadores de un solo núcleo a procesadores con numerosos núcleos, el paralelismo se hace omnipresente a todos los niveles. A nivel micro, el paralelismo es explotado desde los circuitos, el paralelismo a nivel de pipeline e instrucciones sobre procesadores multi-core. A nivel macro, se promueve el paralelismo desde múltiples máquinas en un rack a muchos rack en un centro de datos, hasta llegar a infraestructuras globales basadas en Internet [RR11].

La presente propuesta tiene como objetivo aplicar técnicas HPC en las etapas del proceso de obtención de información a partir de datos masivo considerando arquitecturas multi y many core como arquitecturas subyacentes.

Líneas de Investigación, Desarrollo e Innovación

Mejorar el trabajo con BigData implica considerar diferentes áreas, estas constituyen sendas líneas de investigación. Para lograrlo nos planteamos las siguientes líneas de investigación, ellas son:

- Modelos y paradigmas de computación de alto desempeño: la programación paralela involucra muchos aspectos, los cuales no se presentan en la programación convencional. El diseño de un sistema paralelo tiene que considerar entre otras cosas, el tipo de arquitectura sobre la cual se va a ejecutar el programa, las necesidades de tiempo y espacio requeridas por la aplicación; las técnicas y estructuras de programación paralela adecuadas para implementarla; y la forma de coordinar y comunicar las diferentes unidades computacionales dedicadas a resolver conjuntamente el problema. Además, en la última década las arquitecturas paralelas han evolucionado drásticamente (clusters de pc, procesadores multicore, procesadores manycore) existiendo un desafío adicional para las aplicaciones paralelas que consiste en explotar tales arquitecturas a su máximo potencial.

- Algoritmos y Estrategias de alta performance en grandes volúmenes de datos: un punto a tener en cuenta cuando lo que se desea es rendimiento, es el origen de los datos de un problema. Con el auge de Internet los datos son generados

automáticamente en forma distribuida (redes de sensores "wireless", observaciones meteorológicas, centros de observaciones satelitales, etc). Otro inconveniente es el relacionado a la heterogeneidad de modos/formatos en la que se encuentra disponible la información y su administración eficiente.

- El análisis de las arquitecturas de procesadores y de las jerarquías de memoria juega un papel fundamental en determinar el desempeño global de un sistema HPC, y cobra mayor importancia dado que los volúmenes de datos cada vez requieren mayor capacidad de memoria y se espera que la demanda continúen aumentando. Una manera de abordar los problemas de densidad, consumo, desempeño y escalabilidad de las tecnologías de memorias y almacenamientos tradicionales, es empleando Memorias No Volátiles (NVM), en la actualidad se están desarrollando numerosas tecnologías NVM. A pesar de que se avizoran numerosos beneficios, también introducen nuevos desafíos, tales como, limitada durabilidad y alta latencia de las escrituras, que necesitan ser abordados. Otro de los aspecto que se abordará, es el manejo de la cache de último nivel (LCC), la cual es compartida por todo los núcleos del procesador, presenta el inconveniente que a medida que el número de núcleos aumenta, la contención causada por las aplicaciones que comparten la LLC se incrementa, por lo que el rendimiento de estos sistemas estará muy influenciado por la eficiencia con la que se maneja la cache compartida.

En las líneas, las investigaciones tienen en cuenta la portabilidad de los desarrollos a pesar de las características propias de cada uno de los datos no estructurados.

Resultados y Objetivos

Como objetivos de las líneas de investigación nos planteamos facilitar el desarrollo de soluciones paralelas portables, de costo predecible y bajo consumo, capaces de explotar las ventajas de modernos ambientes de HPC a través de herramientas y "frameworks de computación" de alto nivel. Para ello será necesario proponer nuevas metodologías a ser aplicadas en cada una de las fases del tratamiento de datos masivos.

Formación de Recursos Humanos

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento el desarrollo de 6 tesis doctorales y 4 tesis de maestría. Además se están ejecutando varias tesinas de grado.

Referencias

[G65] G. E. Moore. Cramming More Components onto Integrated Circuits. Electronics 38, No. 8, pp 114-117 (April 19, 1965).

[CWR12] E. Capriolo , D. Wampler , J. Rutherglen. Programming Hive: Data Warehouse and Query Language for Hadoop. O'Reilly Media. 2012.

[DG04] J. Dean and S. Ghemawat: MapReduce: Simplified Data Processing on Large Clusters. Proc. Sixth Symposium on Operating System Design and Implementation, 2004.

[HP08] J. L. Hennesy and D. A. Patterson. Computer Organization & Design - The Hardware/Software Interface. Morgan Kaufmann, 4th edition, 2008.

[HW11] G. Hager, G. Wellein. Introduction to High Performance Computing for Scientists and Engineers. Chapman & Hall/CRC Computational Science. 2011.

[MCJ13] V. Mayer-Schönberger, K. Cukier. A.I. Jurado. Big data: La revolución de los datos masivos. Turner. 2013.

[N13] J. Needham. Disruptive Possibilities: How Big Data Changes Everything. Kindle Edition. O'Reilly Media Inc. 2013.

[OPS01] S. Orlando, R. Perego, and F. Silvestri. Design of a Parallel and Distributed WEB Search Engine. In Proceedings of Parallel Computing (ParCo) 2001 conference. Imperial College Press, September 2001.

[R13] J.Rusell: Cloudera Impala. O'Reilly Media, Inc.2013.

[RR11] T. Rauber, G. Runger. Parallel Programming for multicore and Cluster Systems. Springer. 2011.

[W09] T. White, D. Cutting: Hadoop-The Definitive Guide. 2009 by O'Reilly Media 2009.