Algoritmos, Estrategias y Análisis de Arquitecturas Orientados al Manejo de Datos Masivos

Rubén Apolloni, Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Cristian Perez-Monte, Fabiana Piccoli, Marcela Printista, Cristian Tissera

LIDIC- Univ. Nacional de San Luís

San Luís, Argentina

{rubenga, mdbarrio, omlopres, ncmiran, mpiccoli, mprinti, ptissera}@unsl.edu.ar

Resumen

En la vida cotidiana, existen problemas cuya solución requiere trabajar con gran cantidad de datos. Algunos de estos problemas incluyen la detección de anomalías en el tráfico en redes, el desarrollo de algoritmos eficientes en la toma de decisiones usando modelos de simulación y el desarrollo de infraestructura para mejorar aspectos de consumo y generación de calor.

En este trabajo se expone distintas líneas de trabajo a seguir teniendo como objetivo desarrollar técnicas de Computación de Alto Desempeño para resolver este tipo de problemas.

Palabras clave: Computación de Alto Desempeño, Datos masivos, Arquitecturas Multicore y Manycore.

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 y la Universidad de Pernambuco, 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 Universidad Nacional de San Luis y el Centro de Informática de la UFPE.

Introducción

El uso masivo de Internet, el surgimiento de nuevas tecnologías y el crecimiento en la velocidad de transmisión de datos originó nuevos conceptos tal como Big Data. Éste [MCJ13] es un conjunto de grandes volúmenes, diversos o no estructurados, complejos, longitudinales o distribuidos de datos, generados desde transacciones en Internet, sensores, instrumentos, vídeos, mails, redes sociales y una variedad de fuentes digitales disponible en la actualidad y también futuras. El conjunto de datos es tan grande y complejo que los medios tradicionales de procesamiento son ineficaces. Por lo cual es un desafío analizar, capturar, recolectar, buscar, compartir, almacenar, transferir, visualizar, etc., cantidades masivas

de información, obtener conocimiento y realizar toda su gestión en un tiempo razonable [N13].

Todo lo expuesto anteriormente, nos lleva a tener la necesidad de utilizar nuevas técnicas y arquitecturas para contribuir a mejorar el procesamiento y los tiempos de respuesta. Las técnicas de computación de altas prestaciones (HPC) permitirán resolver con eficiencia cada uno de los objetivos a plantear.

Uno de los campos de aplicación de Big Data es la detección de anomalías en redes de datos, la cual consiste en la identificación de patrones que se desvían del comportamiento normal del tráfico en una red [BLMP16]. Detectar posibles ataques en la red requiere contar con tecnologías para su clasificación, asociando flujos de datos con las aplicaciones que los generan. Uno de los desafíos actuales es trabajar con un conjunto de datos, los cuales crecen a mayor velocidad que su capacidad de procesamiento. Por ejemplo, utilizar y procesar imágenes para representar el tráfico de red a fin de detectar tráfico anómalo, tiene como ventajas no sólo contar con una herramienta de visualización de tráfico, sino también con las propiedades de las imágenes y su procesamiento: técnicas bien conocidas y naturaleza paralela de las computaciones.

Otro campo de aplicación está relacionado con la recuperación y análisis de grandes volúmenes de datos para la toma de decisiones basados en técnicas de simulación. Este tipo de sistemas normalmente utilizan datos generados en tiempo real provenientes de distintas fuentes, los cuales son usados para desarrollar simulaciones orientadas a reducir la incertidumbre en los escenarios abordados

De acuerdo a 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 del procesador, de la memoria en todos los niveles (tanto a memoria principal y memoria cache) [HP08], de los dispositivos de almacenamiento, y también requiere nuevas soluciones de software, ejemplo de ellos son MapReduce[DG04], Hive[CWR12] e Impala[R13], los cuales 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 los datos, introduciendo nuevos desafíos tanto en las investigaciones como en los desarrollos. Además, estas aplicaciones requieren mayor capacidad de memoria, esperándose un incremento 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.

Con el continuo crecimiento de la Ley de Moore [G65], se observa una constante reducción del tamaño de los transistores, lo que permite diseñar procesadores más potentes, con mayor cantidad de núcleos capaces de empaquetar más datos dentro de una pastilla.

Con los actuales sistemas de computación, 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 multicore. 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 masivos considerando arquitecturas multi y manycore como arquitecturas subyacentes, así como la búsqueda de soluciones a los diferentes problemas que se plantean en la siguiente sección.

Líneas de Investigación, Desarrollo e Innovación

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

- Detectar anomalías en redes, consiste en la identificación de patrones que se desvían del comportamiento normal de tráfico. Con el fin de descubrir comportamientos anormales, se deben utilizar modelos de tráfico precisos y estables para describir un comportamiento de tráfico libre de anomalías. Este es un paso crítico en su detección, ya que un modelo de tráfico incorrecto o inestable causaría un alto número de falsas alarmas.
  Modelar el tráfico de red es realizado mediante imágenes, debido a que facilita la comprensión de las características tanto a gran como a pequeña escala de los datos, permitiendo revelar propiedades no sólo relacionada a los datos en sí, sino a la forma en la cual fueron recolectados.
  En esta línea, el objetivo es detectar posibles anomalías en el tráfico de una red haciendo uso de una combinación de técnicas de análisis de tráfico de red, procesamiento de imágenes y HPC.
- Algoritmos y Estrategias de recuperación de datos para soporte en la toma de decisiones: Actualmente existe una tendencia a desarrollar sistemas de soporte a la toma de decisiones basados en técnicas de simulación. Estos sistemas generalmente usan datos generados en tiempo real por diferentes tipos de sensores o dispositivos móviles para realizar la simulación del sistema a modelar e incluso estos datos son utilizados para corregir posibles desviaciones en la ejecución en curso. Ejemplos de estos sistemas son, la monitorización de individuos para el desarrollo de estrategias de evacuación, mitigar el impacto de enfermedades infecciosas [CT13, FC16] o el estudio de cuencas de ríos [AG16] con la finalidad de lanzar alertas tempranas ante inundaciones. En estos casos, además de contar con un modelo de simulación, es importante desarrollar algoritmos y estrategias de alto desempeño que nos permitan trabajar con grandes volúmenes de datos heterogéneos, con el objetivo de que puedan ser procesados por los sistemas de simulación para la toma de decisiones.
- El análisis de las arquitecturas de procesadores y de las jerarquías de memoria es importante para determinar el desempeño de un sistema HPC, y aún más dado que los volúmenes de datos actuales requieren mayor capacidad de memoria. Una manera de abordar los problemas de densidad, consumo, desempeño y escalabilidad de las tecnologías de memorias y almacenamientos tradicionales, es empleando las Memorias No Volátiles (NVM). A pesar de que se avizoran nuevas tecnologías NVM, también se introducen nuevos desafíos a ser abordados, tales como limitada durabilidad y alta latencia de las escrituras. Otro de los aspectos es considerar el manejo de la cache de último nivel (LCC), la cual es compartida por todos los núcleos del procesador y presenta inconvenientes cuando el número de núcleos aumenta: la

contención producida por las aplicaciones que la comparten se incrementa, el rendimiento de estos sistemas estará influenciado por la eficiencia del manejo de esta cache.

Todas las líneas de investigación mencionadas 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

[AG16] A. Gaudiani, E. Luque, P. Garcia, M. Naiouf, A. De Giusti. "*Optimización y computación paralela aplicadas a mejorar la predicción de un simulador de cauce de ríos*", XXII Congreso Argentino de Ciencias de la Computación. CACIC 2016. Pp. 179-188. Octubre 2016, San Luis, Argentina

[BLMP16] Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Fabiana Píccoli. "*Un enfoque para la detección de anomalías en el tráfico de red usando imágenes y técnicas de Computación de Alto Desempeño*". XXII Congreso Argentino De Ciencias de la Computación. CACIC 2016. Pp. 1166-1175. Octubre2016, San Luis, Argentina

[CT13] P.C. Tissera, M. Printista, E. Luque. "*Simulating behaviors to face up an emergency evacuation*", International Journal of Soft Computing and Software Engineering-JSCSE. Volumen 3. Pp. 857-863. 2013

[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.

[FC16] F. Casares, P.C.Tissera, F. Piccoli. "*A parallel proposal for SEIR model using Cellular Automata*". XXII Congreso Argentino de Ciencias de la Computación. CACIC 2016. Pp. 208-219. Octubre2016, San Luis, Argentina.

[G65] G. E. Moore. "*Cramming More Components onto Integrated Circuits. Electronics*". Proceedings of the IEEE, Volume 38, No. 8, pp 114-117. April 19, 1965.

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

[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.

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

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