Usando Sistemas Dedicados para Computación Paralela de Propósito General

M. Lopresti, C. Perez-Monte, F. Piccoli
LIDIC- Universidad Nacional de San Luis
Ejército de los Andes 950
Tel: 02652 420823, San Luis, Argentina
{*mpiccoli*}*@unsl.edu.ar*

1. Contexto

Esta propuesta de trabajo se lleva a cabo dentro de la línea de Investigación " Sistemas Paralelos" del proyecto "Nuevas Tecnologías para un tratamiento integral de Datos Multimedia". Este proyecto es desarrollado en el ámbito del Laboratorio de Investigación y Desarrollo en Inteligengia Computacional (LIDIC) de la Universidad Nacional de San Luis.

2. Resumen

Los sistemas diseñados para resolver problemas específicos como los procesadores gráficos(GPU), tienen características (bajo precio en relación a su potencia de cálculo, gran paralelismo, optimización para cálculos en coma flotante) muy atractivas para su uso en aplicaciones de propósito general, en problemas relacionados al ámbito científico, de simulación, ingeniería, entre otros. Esto llevó al desarrollo de herramientas y técnicas para facilitar su utilización y transformarlos en una alternativa válida y casera para resolver la mayor cantidad de problemas.

En este trabajo se presentan las características básicas de las GPU y las distintas líneas de trabajo a seguir. Estas líneas tienen en común la consideración de la GPU como computadora masivamente paralela. Los problemas a tratar están relacionados a las Redes de Computadoras y las Bases de Datos.

3. Introducción

El poder computacional asociado a las tecnologías dedicadas a fines específicos y la posibilidad que ofrecen de mejora constante y bajo costo, han constituido una alternativa válida a las supercomputadoras paralelas. El ejemplo más popular de las tecnologías dedicadas son las GPU (Unidad de Procesamiento Gráfico)[4, 22]. Una tarjeta de video puede proporcionar hasta 50 veces más poder de cómputo que la computadora huésped en algunas aplicaciones[21].

Por muchos años la GPU fue utilizada exclusivamente para acelerar el cálculo de ciertas aplicaciones relacionadas directamente con el procesamiento de imágenes, en aplicaciones como videojuegos o 3D interactivas, por ejemplo. Su buen desempeño en este ámbito, junto a su constante y rápida evolución (comparada con los microprocesadores de propósito general), un número de instrucciones menor, y sin aritmética de doble precisión [20, 22], ha permitido desarrollar un modelo de supercómputo casero en donde, con menos recursos económicos de los requeridos para comprar una PC, es posible resolver cierto tipo de problemas aplicando un modelo de paralelismo masivo sobre una arquitectura de procesadores con varios núcleos, memoria compartida y soporte multihilos.

Existen alternativas para procesamiento en GPU, la más ampliamente utilizada es la tarjeta Nvidia, para la cual se ha desarrollado un kit de programación en C, con un modelo de comunicación de datos y de control de hilos proporcionado por un driver, el cual provee una interfaz GPU-CPU [19]. Este ambiente de desarrollo llamado Compute Unified Device Architecture (CUDA) ha sido diseñado para simplificar el trabajo de sincronización de hilos y la comunicación con la GPU [5, 23] propone un modelo de programación SIMD(Simple Instrucción, Múltiples Datos) con

funcionalidades de procesamiento de vector.

Dadas las ventajas de poder computacional, bajo costo, continua evolución, bandwidth de memoria, flexibilidad y programabilidad de la GPU y a pesar de su limitación y dificultad para resolver cualquier tipo de aplicaciones siguiendo un modelo de programación no usual, se intenta aprovechar la gran potencia de cálculo de las GPU para aplicaciones no relacionadas con los gráficos, en lo que se conoce como GPGPU(General Purpose GPU)[24].

La línea de investigación que se propone seguir pretende evaluar la factibilidad de utilizar la GPU como computadora masivamente paralela para obtener soluciones de alto desempeño a problemas de propósito general, tal como los derivados de la administración de redes de computadoras y de base de datos.

4. Líneas de Investigación y Desarrollos

Utilizar arquiteturas dedicadas, como la GPU, para resolver computacionalmente problemas de naturaleza distinta a la de ellas, implica plantear soluciones paralelas a los problemas considerando el modelo de programación propio de sus interfaces.

Entre las líneas de investigación y desarrollo que actualmente se siguen se encuentran:

- *Seguridad en Redes de Computadoras*

  La seguridad en sistemas de computación es un tema de gran interes, por su amplitud en las áreas que abarca y su constante evolución. Las redes de computadoras constituyen un ámbito natural para su aplicación[7, 27]. Si bien es cierto que existen numerosas tecnologías para hacer segura una red de computadoras, la Detección de Intrusiones(IDS) constituye una de las más populares[10, 17]. Su objetivo es monitorear las actividades en los sistemas de computación a fin de encontrar violaciones de seguridad.

  Los IDS junto con otras herramientas procuran detectar mediante el monitoreo del tráfico entrante y saliente de un nodo, los intentos de ingresar a la red sin autorización. La mayoría de los IDS, en las redes actuales, basan su seguridad en un conjunto de reglas, las cuales sirven para establecer la autenticidad o no del mensaje. La técnica más común es realizar una búsqueda de *firmas* (*Signatures*) en cada uno de los paquetes circulando en la red [9, 13, 26]. Este proceso implica un alto costo computacional, no sólo por el tiempo involucrado en obtener la solución, sino también por la gran cantidad de recursos del sistema.

  Existen numerosas propuestas para optimizar o reducir el costo de los IDS a través de la aplicación de técnicas de computación de alto desempeño: técnicas de computación paralela o utilización de hardware especializado [2, 6, 12, 25]. Si bien los resultados obtenidos muestran eficiencia, las soluciones son complejas y poco flexibles. Investigaciones recientes están considerando las Unidades de Procesamiento Gráfico(GPUs) como una posibilidad a la hora de acelerar las desiciones de un IDS y de otras aplicaciones en redes [8, 18].

- *Base de Datos*

  Una Base de Datos(BD) es una colección de datos organizados y relacionados entre sí , los cuales son recolectados y analizados, de forma rápida y ordenada, por los sistemas de información de una empresa o negocio en particular [11, 16]. Actualmente, las BD almacenan información no necesariamente estructurada, cualquier conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso constituye una BD.

  Casi todas las bases de datos actuales son Sistema de Gestión de Bases de Datos (DBMS)[11]. Un DBMS establece que una BD no es simplemente un conjunto de archivos, sino además una serie de herramientas para manipular la información almacenada. Es esta característica la que permite, entre otras cosas, hacer consultas y resolverlas obteniendo el conjunto de datos que las satisfagan. Entre las herramientas incormparadas se encuentran los índices (o índice inverso), listas ordenadas de términos (en cualquier formato), los cuales representan a uno o varios datos de la BD.

  Las consultas a bases de datos de gran escala como Google o eBay significan encontrar un

pequeño grupo de entradas específicas entre unas decenas de millones de posibilidades en el menor tiempo posible. Esto implica que un procesador debe resolver consultas de un usuario, o miles, en la misma BD, sobre dominios diferentes y al mismo tiempo, en unos pocos milisegundos. Los índices facilitan dichas consultas; a través de ellos se delimita el acceso a un subconjunto de los datos pertenecientes al universo, y permite dar respuesta a las consultas realizadas. Proveer métodos eficientes para índices y para resolver consultas es primordial, algunos ya lo están intentando [14, 20].

- *Sistema de Información Geográfica*

  Un Sistema de Información Geográfica (GIS) se lo define como un conjunto de equipos informáticos, de programas, de datos geográficos y de técnicas organizadas para recoger, almacenar, actualizar, manipular, analizar y presentar eficientemente todas las formas de información georeferenciada[1, 29]. Los GIS permiten responder preguntas de: Localización: ¿Qué hay en...?; Condición: ¿Dónde se encuentra?; Tendencia: ¿Qué ha cambiado desde...?; Distribución: ¿Qué patrones de distribución espacial existen?; y Modelización: ¿Qué sucede si...?..

  La información geográfica administrada por un GIS contiene una referencia territorial explicita como latitud y longitud o una referencia implícita como domicilio o código postal. Ésta es el elemento diferenciador de un SIG frente a otro tipo de Sistemas de Información. Generalmente la información geográfica es adminstrada en una BD espacial.

  Una BD espacial es un caso especial de BD. Ésta se carateríza por optimizar el almacenamiento y las consultas de objetos en el espacio, incluyendo puntos, líneas y polígonos. Un Sistema de administración de Bases de Datos espaciales (SGBDE) consiste en una colección de tipos de datos espaciales, operadores, índices y estrategias de procesamiento[29]. Sus componentes incluyen: modelo de dato espacial, lenguaje de consulta, procesamiento de consultas, organización de archivos e índices y optimización de consultas. Un dato espacial es una variable asociada a una localización del espacio, entre otros. Los datos espaciales refieren a entidades o fenómenos que cumplen con los principios básicos de tener: *Posición absoluta* en un sistema de coordenadas (x,y,z), *Posición relativa* respecto a otros elementos del paisaje, *Una figura geométrica* que las representan, y *Atributos* que lo describen. Las imágenes satelitales son ejemplos de datos espaciales, su procesamiento debe hacerse respecto a un marco de referencia espacial, posiblemente la superficie de la tierra.

  La utilidad de los GIS y las bases de datos espaciales es muy amplio, entre alguno de sus usos están los Mapas de población, Mapas de densidades, Cálculo de distancias, Cartografía. Sus múltiples usos y el incremento de la complejidad de los datos (producto de los nuevos medios de obtención de datos: sensores, satélites, entre otros; y los requerimientos de mayor detalle de análisis) hacen que obtener mejores tiempos de respuesta sea un desafío constante[30].

El común denominador de los tres problemas planteados anteriormente es el rápido crecimiento del volumen de datos a tratar en cada uno y la necesidad de resolverlos en forma rápida, eficiente y precisa. Analizar las ventajas y desventajas de utilizar la tecnología de GPU y su modelo de programación asociado constituye una imnovadora línea de investigación.

5. Resultados obtenidos / esperados

El principal aporte de esta línea de investigación es analizar la factibilidad de utilizar la arquitectura GPU y su modelo de programación asociado para desarrollar soluciones de alto desempeño a problemas de propósito general, estableciendo los límites del modelo GPU-CPU para las soluciones eficientes y eficaces de cada uno de los problemas planteados.

Actualmente se está trabajando en las dos primeras líneas de investigación.

6. Formación de Recursos Humanos

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento

una tesis de maestría y un doctorado en desarrollo; así como también varios trabajos de fin de carrera de la Licenciatura en Ciencias de la Computación.

Acualmente se ha obtenido una beca otorgada por la Facultad de Ciencias Físico, Matemáticas y Naturales de la Universidad Nacional de San Luis, categoría postgrado.

Referencias

[1] O. O. Ayeni, D. N. Saka and G. Ikwuemesi - Developing a multimedia GIS database for tourism industry in NIGERIA

[2] Z. K. Baker and V. K. Prasanna -Time and area eficient pattern matching on FPGAs - In Proceedings of the 2004 ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays (FPGA 04) - Pp 223-232 - New York, NY, USA. 2004.

[3] P. Bakkum, K. Skadron - Accelerating SQL database operations on a GPU with CUDA. Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units. ACM International Conference Proceeding Series; Vol. 425. Pp: 94-103. Pittsburgh, Pennsylvania. ISBN:978-1-60558-935-0. 2010.

[4] Buck, I. - GPU computing with NVIDIA CUDA - SIGGRAPH '07: ACM SIGGRAPH 2007 courses ACM, New York, NY, USA. 2007.

[5] Chen, W. , Hang, H. - H.264/AVC motion estimation implementation on Compute Unified Device Architecture (CUDA) - IEEE International Conference on Multimedia and Expo 2008 - Pp 697:700 - April 2008.

[6] C. Clark, W. Lee, D. Schimmel, D. Contis, M. Kone, and A. Thomas - A hardware platform for network intrusion detection and prevention - In Proceedings of the 3rd Workshop on Network Processors and Aplications (NP3) - 2004.

[7] D. E. Comer - Computer Networks and Internets - ISBN 0136061273 - Prentice Hall, 2008.

[8] D. L. Cook, J. Ioannidis, A. D. Keromytis, and J. Luck - Cryptographics: Secret key cryptography using graphics cards - In Proceedings of RSA Conference, Cryptographer's Track (CT-RSA) - Pp 334-350, 2005.

[9] G. Cretu, A. Stavrou, S. Stolfo, A. Keromytis - Data Sanitization: Improving the Forensic Utility of Anomaly Detection Systems - In the Proceedings of the Third Workshop on Hot Topics in System Dependability - Edinburgh, UK - June 2007

[10] T. Crothers - Implementing Intrusion Detection Systems - Wiley - 2003.

[11] C. J. Date - Database in depth: relational theory for practitioners. O'Reilly Media. ISBN 0596100124, 9780596100124. 2005.

[12] S. Dharmapurikar and J. Lockwood - Fast and scalable pattern matching for content Filtering - In Proceedings of the 2005 ACM symposium on Architecture for networking and communications systems (ANCS 05), Pp 183-192, New York, NY, USA, 2005. ctober 2004.

[13] V. Frias-Martinez, S. Stolfo, A. Keromytis - Behavior-Profile Clustering for False Alert Reduction in Anomaly Detection Sensors - In the Proceedings of the Annual Computer Security Applications Conference (ACSAC) - 2008.

[14] N.K. Govindaraju, N. Raghuvanshi, M. Henson, D. Tuft and Dinesh Manoch - A cache-efficient sorting algorithm for database and data mining computations using graphics processors - 2005.

[15] B.He, K. Yang, R. Fang, M. Lu, N. Govindaraju, Q. Luo, P. Sander - Relational joins on graphics processors. International Conference on Management of Data. Proceedings of the 2008 ACM SIGMOD international conference on Management. Vancouver, Canada Pp: 511-524. ISBN:978-1-60558-102-6. 2008.

[16] J.A. Hoffer, M. Prescott, H. Topi - Modern Database Management. Prentice Hall. ISBN 0136003915, 9780136003915. 2008.

[17] P. Inella, O. McMillan - An Introduction to Intrusion Detection Systems - SecurityFocus.com

[18] N. Jacob and C. Brodley - O?oading IDS computation to the GPU - In Proceedings of the 22nd Annual Computer Security Aplications Conference on Annual Computer Security Aplications Conference (ACSAC 06), Pp 371-380, Washington, DC, USA, 2006.

[19] Joselli, M., Zamith, M., Clua, E., Montenegro, A., Conci, A., Leal-Toledo, R. Valente, L., Feijo, B., dórnellas, M., Pozzer, C - Automatic Dynamic Task Distribution between CPU and GPU for Real-Time Systems - . 11th IEEE International Conference on Computational Science and Engineering, 2008 (CSE '08) - Pp 48:55 - July 2008.

[20] Lieberman, M.D. and Sankaranarayanan, J. and Samet, H. - A Fast Similarity Join Algorithm Using Graphics Processing Units - ICDE 2008. IEEE 24th International Conference on Data Engineering 2008 - Pp 1111:1120 - April 2008.

[21] Lloyd, D., Boyd, C., Govindaraju, N. - Fast computation of general Fourier Transforms on GPUS - IEEE International Conference on Multimedia and Expo - Pp 5:8 - April 2008.

[22] Luebke, D., Humphreys, G. - How GPUs Work Computer - vol 40, N 2 - Pp 96:100 - ISSN 0018-9162 - Feb. 2007.

[23] Luebke, D. - CUDA: Scalable parallel programming for high-performance scientific computing - 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro. ISBI 2008. Pp 836:838 - May 2008.

[24] J.D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A.E. Lefohn and T. Purcell - Survey of General-Purpose Computation on Graphics Hardware - Eurographics 2005, State of the Art Reports. Pp. 21-51. September 2005

[25] V. Paxson, R. Sommer, and N. Weaver. An architecture for exploiting multi-core processors to parallelize network intrusion prevention. In Proceedings of the IEEE Sarno? Symposium, May 2007.

[26] Y. Song, A. D. Keromytis and S. J. Stolfo - Spectrogram: A Mixture-of-Markov-Chains Model for Anomaly Detection in Web Traffic - In the Proceedings of the 16th Annual Network & Distributed System Security Symposium (NDSS). San Diego, CA, USA. February 2009.

[27] A. Tanenbaum - Computer networks - ISBN 8120321758 - Prentice-Hall - 2007

[28] G. Vasiliadis, S. Antonatos, M.Polychronakis, E. Markatos and S. Ioannidis - Gnort: High Performance Network Intrusion Detection Using Graphics Processors - RAID - Pp 116:134 - 2008.

[29] Shashi Shekhar and Sanjay Chawla - Spatial Databases: A Tour - Prentice Hall, 2003 (ISBN 013-017480-7)

[30] Y. Wu, Y. Ge, W. Yan, X. Li - Improving the performance of spatial raster analysis in GIS using GPU. Geoinformatics 2007: Geospatial Information Technology and Applications. Edited by Gong, Peng; Liu, Yongxue. Proceedings of the SPIE, Volume 6754, pp. 67540P. 2007.