Técnicas de GPGPU en la Identificación de Señales y Espacios Métricos

Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, Fabiana Piccoli
LIDIC- Univ. Nacional de San Luis
San Luis, Argentina
{mdbarrio, omlopres, ncmiran, mpiccoli}@unsl.edu.ar

1. Contexto

Esta propuesta de trabajo se lleva a cabo dentro del proyecto de investigación *"Nuevas Tecnologías para el Tratamiento Integral de Datos Multimedia"* abarcando las líneas de investigación "Computación de Alto Desempeño", "Procesamiento de Información Multimedia" y " Recuperación de Datos e Información Multimedia". Dicho proyecto 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.

2. Resumen

Los sistemas diseñados para resolver problemas específicos como los procesadores gráficos(GPU), tienen características muy atractivas (bajo precio en relación a su potencia de cálculo, gran paralelismo, optimización para cálculos en coma flotante, entre otras) 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 de la GPU y las distintas líneas de trabajo a seguir. Estas líneas tienen en común la consideración de la GPU como computadora masivamente paralela, GPGPU. Los problemas a tratar están relacionados a las Bases de Datos Métricas y a la Identificación y Recuperación de señales, particularmente las señales de audio.

Palabras Claves: Procesamiento de señales, Computación de alta perfomance, Espacios métricos, Recuperación de Información.

3. Introducción

El poder computacional asociado a las tecnologías dedicadas a fines específicos, sus constantes avances y el 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)[1, 14]. Una tarjeta de video puede proporcionar mucho más poder de cómputo que la computadora huésped en algunas aplicaciones[13]. La misma posee una interesante arquitectura de computación de alto rendimiento.

Realizar un mapping de una computación de propósito general en una GPU, implica utilizar el hardware de la tarjeta gráfica para resolver cualquier aplicación, no necesariamente de naturaleza gráfica. Esto se conoce como GPGPU (GPU de propósito general) e implica la utilización de la potencia de cálculo de la GPU para resolver problemas de propósito general.

La programación paralela sobre GPU tiene varias diferencias con la programación paralela en computadoras paralelas típicas, las más relevantes son:

- *El número de unidades de procesamiento*: En computadoras masivamente paralelas, generalmente el número elegido es el mismo que el número de núcleos en el sistema. En GPGPU esto no se tiene en cuenta.
- *Estructura de la memoria CPU-GPU*: El sistema de memoria de la CPU-GPU considera dos espacios de memoria diferentes: la memoria del host (que es compartida por todos

los hilos de la CPU durante la aplicación) y la memoria de la GPU. Como los hilos de la GPU ejecutan en un espacio de memoria separado a los hilos del host en la aplicación, las transferencias de código y datos son necesarios entre el host y el dispositivo en diferentes momentos.

- *Número de hilos paralelos*: La programación de las GPU tiene la posibilidad de iniciar un gran número de hilos con poca sobrecarga. La GPU ofrece mecanismos transparentes y de bajo costo para la creación y administración de hilos.

No todo tipo de problemas pueden ser resueltos en la arquitectura de la GPU, los problemas más adecuados son aquellos que pueden ser implementados con procesamiento de stream y usando memoria limitada. Las aplicaciones más adecuadas son aquellas con cuantioso paralelismo de datos.

Existen diferentes alternativas para procesamiento de aplicaciones de propósito general en la 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, quien provee una interfaz GPU-CPU [12]. Este ambiente de desarrollo llamado Compute Unified Device Architecture (CUDA) propone un modelo de programación SIMD(Simple Instrucción, Múltiples Datos) con funcionalidades de procesamiento de vector.

Las líneas de investigación que se proponen seguir pretende evaluar la factibilidad de utilizar la GPU como computadora masivamente paralela para obtener soluciones de alto desempeño a problemas de propósito general. Los problemas considerados están relacionados a consultas en base de datos métricas y a diferentes técnicas de identificación señales de audio. La utilización de la GPU como computadora paralela se realiza a través de CUDA.

4. Líneas de Investigación y Desarrollo

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:

- *Espacios Métricos:* Una de las principales herramientas que las computadoras ofrecen es almacenar grandes cantidades de datos organizados, siguiendo un determinado esquema o un modelo de datos que facilite su almacenamiento, recuperación y modificación de una forma rápida y ordenada.

  Cualquier conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso constituye una base de datos tradicional, en las cuales se busca sólo a partir de una clave.

  Actualmente ha surgido la necesidad de crear bases de datos de información no estructurada, como por ejemplo imágenes, texto, sonido y video. Muchas aplicaciones computacionales necesitan buscar eficientemente información sobre estas bases de datos. Para realizar consultas en este tipo de bases de datos se han creado nuevos modelos y algoritmos de búsqueda más generales que los correspondientes a bases de datos tradicionales [8].

  A estas nuevas bases de datos se las denomina base de batos métricas y las consultas se realizan según búsquedas por similitud o proximidad. La búsqueda por proximidad es la búsqueda en bases de datos de elementos similares o cercanos al elemento consultado. La similitud es modelada con una función de distancia, la cual satisface las siguientes propiedades. Si X denota el universo de objetos válidos. Un subconjunto finito de él U, de tamaño n, es el conjunto de objetos o base de datos en donde se realizan las búsquedas. La función $d : X * X \longrightarrow \mathbb{R}^+$ denota la medida de distancia entre los objetos. Esta función de distancia debe cumplir las propiedades de positividad $\forall x, y \in X, d(x, y) \geq 0$, simetría $(\forall x, y \in X, d(x, y) = d(y, x))$, reflexividad $(\forall x \in X, d(x, x) = 0)$, desigualdad triangular $\forall x, y, z \in X, d(x, y) \leq d(x, y) + (d(y, z)$ y en la mayoría de los casos la positividad es estricta $(\forall x, y \in X, x \neq y \Rightarrow d(x, y) > 0)$. El conjunto de elementos u objetos sobre los que se define la función de distancia es lla-

mado *espacio métrico* [8].

Mientras más chica sea la distancia entre los objetos más similares son dichos objetos. Existen tres tipos básicos de consultas por proximidad en espacios métricos:

- Consulta por rango$(q, r)_d$: recupera los elementos que están a lo más a distancia r de q.
- Consulta por vecino más cercano $NN(q)$: recupera el vecino más cercano a q en U.
- Consulta por k vecinos más cercanos $NN_k(q)$: recupera los k vecinos más cercanos a q en U.

Si tenemos una base de datos de cardinalidad n, todas estas consultas pueden ser resueltas ejecutando n evaluaciones de distancia. Generalmente el número de evaluaciones de distancias ejecutadas es la medida de complejidad de los algoritmos. El desafío es diseñar un algoritmo de indexación eficiente que reduzca el número de las evaluaciones de distancia [8, 9, 10, 11, 15, 20, 21, 22, **?**].

Los algoritmos de indexación permiten construir a priori un índice, una estructura de datos capaz de ahorrar computaciones de distancias al responder consultas por proximidad.

En general todos los algoritmos de indexación particionan el conjunto X en subconjuntos X_i. Se construye un índice para permitir determinar una lista de subconjuntos X_i de candidatos potenciales a contener objetos relevantes para la consulta.

Las técnicas de computación de alto desempeño nos permitiran acelerar no sólo el proceso de indexación sino también la obtención de las respuestas a las consultas.

- *Identificación de Señales:* Las bases de datos y repositorios de información multimedia (audio, imagen, video y texto) no pueden ser trabajadas tan eficientemente como en las bases de datos tradicionales debido a que la información multimedia debe ser recuperada por similitud; mientras que en las bases de datos tradicionales esta búsqueda es exacta.

  Un modelo estándar de búsqueda en este tipo de bases de datos consiste en utilizar una medida de (dis)similitud entre los objetos almacenados. Esta medida de distancia entre objetos debería modelar esencialmente el comportamiento de una persona al comparar dos objetos de esa naturaleza. Dos objetos iguales perceptualmente deberían recibir distancias pequeñas; mientras que dos objetos perceptualmente distintos deberían recibir distancias grandes [8]. Esto nos daría un mecanismo eficaz de recuperación; sin embargo, este mecanismo, no escala cuando la base de datos crece [2].

  La representación de los objetos de forma estable y persistente a diversas degradaciones, tanto naturales como ataques maliciosos se denomina huella digital de la señal. Idealmente la huella digital debe ser una invariante de la señal; aquellas características intrínsecas, no alteradas por su constante manipulación.

  En [3, 4, 5], se discute un método muy eficaz para identificar sin error secuencias de audio sujetas a degradaciones severas; el método de identificación es secuencial. La identificación implica determinar la representación de la señal y su búsqueda en una base de datos de audio. La aplicación de computación de alto desempeño permitió mejorar los resultados citados haciéndolos eficientes para su cálculo en tiempo real [16, 17, 18, 19]. Actualmente se está trabajando en lograr mayores optimizaciones.

- *Identificación de Colecciones de Audio:* El uso de firmas de audio, representaciones sucintas de una secuencia, ha permitido comparar de manera eficiente dos segmentos de audio cuando ambos son del mismo tamaño, o cuando uno está contenido en el otro. Si una de las instancias se ve afectada por una distorsión temporal, se deben alinear los dos audios de la misma forma que se alinean las cadenas de ADN en la búsqueda de semejanzas entre genomas. Para realizar alineamiento de audio, se ha propuesto recientemente en [6, 7] una técnica que permite comparar huellas digitales de sonidos con distorsión temporal, utilizándose un esquema de alineamiento no local basado en la búsqueda de todos los *q-gramas* (subcadenas de largo q) de una secuencia de audio A, contra to-

dos los *q-gramas* de la secuencia de audio *B*. Este procedimiento se utiliza para empatar dos interpretaciones de audio.

Se seleccionan los formantes comunes a una colección que aparezcan varias veces con tamaños aproximados en toda la colección. Se define a un formante como un segmento común a un número significativo de secuencias de audio de una colección. En una colección de audio se podría, en principio, extraer el conjunto de cadenas comunes a k secuencias, para diferentes k. Estos formarían una base que puede describir las partes significativas de la colección de audio. Una vez establecida la base de la colección, el contenido de una secuencia podría ser descrito en términos de elementos de esta base; como una secuencia ahora de formantes en lugar de una secuencia de caracteres. Para lograr este objetivo, se pretende desarrollar una técnica robusta al ruido, la cual permita, a través de computación de alto desempeño en un sistema CPU-GPU, analizar secuencias de audio con distorsiones temporales. Se quiere además poder calcular una alineación entre los audio para determinar las mejores correspondencias. Cuando en una colección grande se tengan factorizados todos los formantes de una representación robusta al ruido será posible darle un tratamiento textual, más cercano a la recuperación de información, que permitiría en principio aplicar una gran cantidad de herramientas que no se han aplicado tradicionalmente a las colecciones de audio.

5. Resultados obtenidos / esperados

En las dos primeras líneas de investigación se está se están evaluando y comparando los rendimientos de las soluciones secuenciales y paralelas, no sólo respecto a los parámetros de evaluación típicos de los sistemas de alto desempeño, sino también en la confiabilidad de las respuestas. En el caso de los espacios métricos, se están desarrollando y evaluando distintos tipos de consultas y métodos de indexación, lograndose en algunos casos tiempos constantes. Respecto a la identificación de señales de audio, se han desarrollado algoritmos para la determinación de la huella digital con muy buenos resultados[16, 17, 18, 19].

Considerando la última línea de investigación, se está en el proceso de diseño y análsis de distintos algoritmos para la transformación de audio y la extracción de características utilizando técnicas GPGPU. Actualmente se está analizando el proceso de compresión / descompresión de audio en la GPU a fin de reducir los costos en las tranferencias entre las memorias de la CPU y la GPU.

6. Formación de Recursos Humanos

Los resultados esperados respecto a la formación de recursos humanos son hasta el momento una tesis de maestría y dos tesis de doctorado en desarrollo.

Además la finalización de beca de postgrado tipo II otorgada por el Consejo Nacional de Investigaciones Científicas (CONICET), obtenida el 01/04/11.

Referencias

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

[2] Bustos Cárdenas, B.E. - Index Structures for Similarity Search in Multimedia Databases - PhD thesis, Universitat Konstanz, Fachbereich Informatik, Germany - Octubre 2006.

[3] Camarena-Ibarrola, A., Chavez E. - A robust entropy based audio-fingerprint - IEEE International Conference on Multimedia and Expo 2006.

[4] Camarena-Ibarrola A., Chávez E. - On Musical Performances Identification Entropy and String Matching - MICAI 2006

[5] Camarena-Ibarrola J.A. - Análisis digital de la señal de voz - PhD thesis Borrador - Universidad Michoacana de San Nicolás de Hidalgo, México - Agosto 2007.

[6] Camarena-Ibarrola, A., Chavez, E. - Real time tracking of musical performances. - In MICAI, volume To appear - 2010.

[7] Camarena-Ibarrola, A., Chavez, E. - Online Music Tracking with Global Alignment - International Journal of Machine Learning and Cybernetics. - Springer. To appear.

[8] Chávez E., Navarro G., Baeza Yates R.A., Marroquín J.L. - Searching in metric spaces - ACM Comput. Surv. Vol 33 N3 - Pp 273:321 - 2001.

[9] Chávez, E. and Navarro, G. - A compact space decomposition for effective metric indexing. - Pattern Recognition Letters 26(9) - Pp 1363:1376 - 2005.

[10] Dohnal, V., Gennaro, C., Savino, C. and Zezula, P. - D-index: Distance searching index for metric data sets. - Multimedia Tools and Applications (MTAP), - 21(1): 9:33 - 2003.

[11] Figueroa Mora, K. - Indexación Efectiva de Espacios Métricos usando Permutaciones. - PhD thesis - Universidad de Chile, Santiago, Chile - 2007. - Director: Dr. G. Navarro y Dr. E. Chávez.

[12] 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.

[13] 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.

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

[15] Mamede, M. - Recursive lists of clusters: A dynamic data structure for range queries in metric spaces. In Proc. 20th Intl. Symp. on Computer and Information Sciences (ISCIS'05), LNCS 3733 - pages 843:853 - 2005.

[16] Miranda N., Piccoli F., Chávez E. - Using GPU to Speed Up the Process of Audio Identification - I2TS 2010 - 9th International Information and Telecommunication Technologies Symposium. IEEE - R10. ISBN 978-85-89264-11-2. - Rio de Janeiro, Brazil - December 2010.

[17] Miranda N., Piccoli F., Chávez E. Camarena Ibarrola A. - Finding Audio Fingerprinter Using GPU - IX Congreso Argentino de Mecánica Computacional - XXXI Congreso Ibérico Latinoamericano de Métodos Computacionales en Ingeniería (Mecom - Ciilamce 2010) - ISSN 1666-6070.- Pp 3107:3126 - Noviembre 2010 - Buenos Aires, Argentina.

[18] Miranda N., Piccoli F., Chávez E., Camarena Ibarrola A. - Fast GPU Audio Identification - 16vo Congreso Argentino de Ciencias de la Computación (CACIC 2010) - ISBN 978-950-9474-49-9. - Pp 229:242 - Univ de Moron, Buenos Aires, Argentina - Octubre 2010.

[19] Miranda, N., Piccoli, F., Chavez, Edgar - Considering a Pure GPU Model for an Audio Fingerprinting System. - XIX Congreso sobre Métodos Numéricos y sus Aplicaciones (ENIEF 2011). - ISSN 1666-6070 Vol XXX. - Pp 3033-3044. - Noviembre 2011 - Rosario Santa Fe, Argentina.

[20] Reyes, N. - Indices dinámicos para espacios métricos de alta dimensionalidad. - Master's thesis - Universidad Nacional de San Luis - San Luis, Argentina - 2002. - Director: Dr. G. Navarro.

[21] Samet, H. - Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling). - Morgan Kaufmann Publishers Inc. - San Francisco, CA, USA - 2005.

[22] Zezula, P., Amato, G., Dohnal, V. and Batko, M. - Similarity Search: The Metric Space Approach (Advances in Database Systems). - Springer-Verlag New York, Inc. - Secaucus, NJ, USA - 2005. - XVIII, 220 p., Hardcover - ISBN: 0-387-29146-6.