Tarjetas gráficas para acelerar el cómputo complejo

Jorge Echevarría *

La búsqueda de mayor rendimiento

A lo largo de la historia de la informática la capacidad de procesamiento ha crecido en forma continua y las que antes eran poderosísimas computadoras se han convertido en objetos obsoletos. Los procesadores de PC han incrementado su rendimiento desde menos de un millón de operaciones por segundo en los primeros microprocesadores 8086 hasta alrededor de 50000 millones de operaciones en los más veloces chips actuales con cuatro núcleos de procesamiento. A partir de la anterior generación de CPU, el Intel Pentium 4 y AMD Athlon 64, surgieron graves problemas de altas temperaturas causados por la elevada frecuencia de operación que llevaron a modificar el diseño de los procesadores. Al no poder incrementarse la frecuencia en la misma medida que venía haciéndose en cada generación a lo largo de 25 años, los fabricantes comenzaron a diseñar procesadores con dos unidades de procesamiento y luego cuatro a las que llamaron núcleos. Estos diseños permiten duplicar y cuadruplicar el poder de proceso pero tienen sus limitaciones. Una gran cantidad de los programas de software escritos actualmente aprovecha sólo un núcleo de procesamiento o dos, por lo que se desaprovechan en gran medida los recursos disponibles. Esto se debe a que en general el código sigue una secuencia lineal de ejecución y no es sencillo dividir las tareas en forma eficiente. Aun así hay aplicaciones que necesitan grandes capacidades de cálculo y que permiten realizar estos cálculos en paralelo cuando el hardware lo permite. Un ejemplo de estas aplicaciones son los programas avanzados de animación en 3D y los programas que encodifican video.

En la búsqueda de mayor procesamiento hay soluciones alternativas a la informática tradicional. Han surgido avances para incrementar el poder de procesamiento con las llamadas computadoras de ADN que intentan aprovechar el paralelismo de la molécula biológica llamada ácido desoxirribonucléico y sus cuatro pares de bases. Otros avances han surgido en el área de la computación cuántica y el estudio de sistemas trinarios. En este ámbito se estudian sistemas que no funcionan en base a 0 y 1 como los binarios que conocemos, sino con estados adicionales que permiten un cálculo en paralelo mucho mayor. El problema más importante para implementar este tipo de tecnologías es que no es factible al menos por ahora llenar con cubetas llenas de líquidos biológicos las oficinas o aislar los sistemas cuánticos para que funcionen a temperaturas cercanas

* Egresado de la Facultad de Ingeniería - UP.

al cero absoluto, a menos de 200 grados centígrados bajo cero. Esto parece indicar que para incrementar en la actualidad el poder de procesamiento paralelo estamos condenados a armar sistemas de computadoras funcionando en paralelo llamados Clusters.

Los Clusters son sistemas de cómputo distribuidos formados por un grupo de computadoras que comparten las tareas de procesamiento y que están interconectadas mediante una red de alta velocidad. El costo de estos sistemas es inferior al de una supercomputadora o HPC y la flexibilidad que poseen es mayor. Los clusters son escalables y se arman utilizando componentes más económicos, comúnmente encontrados en PCs, conectados mediante una rápida red ethernet. Cada cluster del sistema de cómputo distribuido puede incorporar varios CPU y coprocesadores matemáticos costosos para aumentar su rendimiento, pero actualmente existe un tipo especial de coprocesadores matemáticos más económicos que en muchos casos se encuentran ya instalados en una PC hogareña de alto rendimiento, las tarjetas de video poderosas. Para tener una idea del avance del poder de procesamiento, los actuales Intel Core 2 Quad de 4 núcleos de las computadoras caseras más potentes alcanzan los 100 Megaflops o millones de operaciones de coma flotante por segundo. Con diez de estos procesadores se alcanza el mismo poder de cálculo que poseía la primera supercomputadora que logró alcanzar un Teraflop, la ASCI RED, construida en Nuevo Mexico, Estados Unidos, para realizar cálculos sobre explosiones nucleares.

Tarjetas gráficas para el cómputo avanzado

En el mercado de hardware de PC el sector de los gráficos 3D para juegos ha avanzado como ningún otro por la gran competitividad entre las empresas y la selectividad de los usuarios de juegos. En este contexto el poder de procesamiento ha avanzado y superado en gran medida a la capacidad de los procesadores de propósito general o CPUs. Las tarjetas gráficas de alto rendimiento que normalmente se utilizan para jugar han comenzado a utilizarse como coprocesadores para el cálculo avanzado en paralelo. Estas tarjetas poseen un GPU o unidad gráfica de procesamiento con una gran cantidad de unidades aritmético lógicas llamadas shaders programables o núcleos de procesamiento. Los shaders pueden programarse para realizar cálculos complejos en forma paralela. Cuando se utilizan para tareas ajenas al procesamiento gráfico, estos GPU reciben el nombre de GPGPU o GPUs de propósito general. Las más avanzadas tarjetas de la generación más reciente de GPUs de NVIDIA GeForce, llamadas GTX 280, poseen 240 núcleos de procesamiento escalar. Aunque la frecuencia de operación de estos GPU sea de alrededor de 600 MHz en lugar de los 3000Mhz de los procesadores Intel Quad core más avanzados, no hay forma en que los cuatro núcleos puedan equipararse a los 240 de estas tarjetas. La empresa ATI también posee tarjetas con 160 unidades de procesamiento vectoriales. Cada una de estas unidades posee 5 ALUs que pueden realizar tareas en paralelo si ejecutan instrucciones del mismo thread. Estas nuevas tarjetas gráficas procesan en el orden de 1 Teraflop, es decir, una sola de estas tarjetas para PC posee la misma capacidad teórica de cómputo que la ASCI RED, que como

dijimos era la mas poderosa supercomputadora de 1996, ocupaba una gran habitación y consumía 500 KW de energía sin tener en cuenta el consumo de su refrigeración.

Los GPU o unidades de proceso gráfico de estas tarjetas de video ejecutan operaciones complejas de cálculo aplicando cada operación a gran cantidad de datos en paralelo. Además cada uno de los núcleos de procesamiento que poseen funciona como una línea de montaje para maximizar el trabajo realizado en cada ciclo de reloj del GPU. Hay sin embargo ciertas limitaciones en la programación de GPUs (debido a la arquitectura y funcionamiento de las mismas) y en las que el procesador central o CPU mantiene un lugar importante. Pero, si se deben procesar grandes cantidades de datos en paralelo siguiendo el modelo de programación de streams, se consiguen grandes saltos en el rendimiento a un costo muy bajo. Por ejemplo, si deseamos realizar una sola suma de dos números escalares, esta se ejecutará más rápido en un CPU tradicional, pero si queremos sumar un número dado a un vector de 200 componentes, un GPU de este tipo puede hacerlo en una sola pasada. Para visualizar otro ejemplo imaginemos que queremos desenfocar una fotografía utilizando un programa como el Adobe Photoshop. En lugar de que el CPU calcule mediante una función matemática como se va a modificar cada píxel por separado, un GPU puede ser capaz de modificar tantos píxels de la pantalla como shaders tenga, en el caso de la GeForce GTX 280 serían unos 240 a la vez, si el código está bien escrito. Es importante añadir que el software tiene que poder aprovechar los recursos de la tarjeta gráfica, no para procesar gráficos sino para servir como coprocesador matemático que sirve a otro tipo de aplicaciones.

Programación de tarjetas gráficas

Dentro de los actuales lenguajes de programación de alto nivel para GPU se encuentran Open CL, Brook+ y CUDA. Estos lenguajes son similares al lenguaje C con extensiones para programación de streams que son utilizadas para operar con los GPU. Open CL es un lenguaje para el cómputo en paralelo utilizando el CPU y el GPU que fue creado por Apple. Este lenguaje podría ser utilizado en el próximo sistema operativo de Apple, el Snow Leopard, y es soportado por los fabricantes más importantes de procesadores de tarjetas gráficas. Las tarjetas gráficas de AMD también utilizan Brook+. El compilador Brook+ esta incluido en un entorno de desarrollo de alto nivel llamado AMD Stream Computing SDK y hay versiones para los sistemas operativos Windows XP, Vista y Linux. Brook+ fue desarrollado a partir de Brook y es mejorado y mantenido por AMD. Además de poder programar en alto nivel, mediante la capa de abstracción de cómputo de bajo nivel denominada CAL puede accederse directamente al hardware gráfico. Se están desarrollando librerías matemáticas para ampliar la tecnología AMD Stream Computing. Las herramientas para desarrollo de software utilizando AMD Stream Computing pueden descargarse gratuitamente de la página de AMD para aprovechar el poder de cálculo de sus GPUs.

Hasta el momento el lenguaje más desarrollado y flexible es CUDA. La tecnología CUDA de Nvidia es un entorno de desarrollo en lenguaje C para PC que permite a los programadores y desarrolladores escribir software para resolver problemas computacionales complejos en un tiempo muy reducido. La tecnología esta disponible

para Microsoft Windows XP, Vista, Linux y Mac OS X. Todas las tarjetas gráficas de la serie 8 en adelante de Nvidia en la línea Gforce, Quadro y Tesla soportan CUDA. El lenguaje CUDA utilizado en tarjetas Nvidia permite programar en lenguaje C con extensiones. Parte del código se ejecuta en el CPU y cuando se necesita potencia de cálculo paralelo se deriva el código para que sea ejecutado en el GPU por medio de una llamada. Luego la secuencia de ejecución es devuelta al CPU. De esta manera se maximiza el potencial de proceso de la computadora. Las herramientas de desarrollo están constituidas por tres componentes clave. Un controlador de video con soporte para CUDA, el controlador de CUDA propiamente dicho y numerosos códigos de ejemplo. El SDK o kit de desarrollo de software contiene entre otras cosas un compilador C de Nvidia, librerías de programación, el manual, un controlador de runtime y un debugger para el GPU. La librería de programación de CUDA incluye programas de álgebra lineal y transformaciones de Furier. Toda la documentación, ejemplos, librerías, el entorno de desarrollo y el compilador se encuentran en la página de Nvidia en forma gratuita. Lo único que se requiere para comenzar es poseer una tarjeta gráfica compatible, instalar el controlador de la misma y bajar el kit de desarrollo CUDA de la Web. Como decíamos, el modelo de programación tiene algunas limitaciones, por ejemplo, las funciones recursivas deben convertirse a loops porque no están soportadas, los threads deben ser ejecutados en grupos para obtener buen rendimiento y el pasaje de datos entre el CPU y el GPU puede convertirse en un cuello de botella según la implementación. Aún así, si se utiliza correctamente se obtienen grandes beneficios.

La apuesta de CUDA es fuerte ya que es difícil que los programadores se adapten a nuevas formas de trabajar y a nuevas limitaciones en cuanto a cómo y qué tipo de código se puede generar. Aún así hay más de 80 millones de tarjetas gráficas en oficinas y hogares habilitadas para utilizar esta tecnología y en muchos casos el salto de rendimiento es enorme. El lenguaje C es ampliamente utilizado y sólo se tienen que aprender a manejar las nuevas extensiones que lo extienden. Una gran cantidad de aplicaciones comerciales y científicas han adoptado la tecnología CUDA y ahora también están comenzando a surgir aplicaciones de consumo que la aprovechan.

Implementaciones de la tecnología CUDA

Dentro de las aplicaciones actuales se encuentra Badaboom de Elemental Technologies. Badaboom es un programa de trascodificación de video que convierte archivos de video a otros formatos. Por ejemplo, el programa puede convertir un archivo de video de alta definición para poder reproducirse en un iPod u otro dispositivo portátil. La trascodificación de video puede ser una de las tareas que más tiempo requiere en el cómputo casero. Convertir una película de dos horas, por ejemplo, puede requerir seis o más horas cuando se utiliza la CPU de la computadora. Sin embargo, con Badaboom y utilizando la GPU, el proceso de conversión puede ser hasta 18 veces más rápido que con los métodos tradicionales, realizando el trabajo en pocos minutos y, al mismo tiempo, liberando la CPU para manejar otras tareas como el correo electrónico y la navegación Web.

La empresa TechniScan Medical Systems desarrolló una tecnología llamada UltraSoundCT asistida por GPUs que utiliza ultrasonidos para escanear las mamas para estudios complementarios a las mamografias. La computadora utiliza cuatro GeForce 8 para producir cortes coronales del tejido. También puede generarse una imagen 3D. Utlilizando el cómputo de las tarjetas gráficas toma sólo 20 minutos el renderizado de las imágenes comparado con varias horas que se requerirían en un CPU.

El grupo de investigaciones ASTRA de la Universidad de Antwerp desarrolla métodos de tomografía computada. La tomografía es una técnica utilizada en escaners médicos para crear imágenes tridimensionales de los órganos internos de los pacientes, basadas en un gran número de fotos de rayos X adquiridas en un rango de ángulos diferentes. ASTRA desarrolla técnicas de reconstrucción para obtener mayor calidad de la que se obtiene normalmente. Para las reconstrucciones se necesita un enorme poder de cálculo. Uno de los métodos es utilizar clusters de cientos de PCs para distribuir el trabajo en paralelo, lo cual es bastante caro y ocupa mucho espacio físico. En vez de implementar esa solución ASTRA armó una sola PC con cuatro tarjetas gráficas que poseen dos GPU cada una. Estas ocho GPU en una sola PC tienen el poder de cálculo de 350 CPUs para realizar las reconstrucciones. Con sólo 4000 Euros realizan las reconstrucciones con el mismo rendimiento que con la supercomputadora de la universidad que tuvo un costo de 3,5 millones de Euros.

Un trabajo similar se realizó en el Hospital General de Massachussetts donde utilizan rayos X para formar una imagen en tiempo real mediante un proceso llamado Tomosíntesis digital. Para ello anteriormente se utilizaba un sistema computacional de 35 PC formando un cluster. Actualmente el mismo trabajo es realizado mediante tarjetas gráficas 100 veces más poderosas para esas tareas que los CPU.

Dentro del ámbito de la investigación médica, hay un programa de cómputo distribuido llamado Folding@home de la Universidad Stanford. Este programa permite a cualquier persona en el mundo con una computadora bajar un software cliente a la PC que, al igual que un protector de pantalla, cuando la PC está ociosa, utiliza los recursos de ésta para trabajar. Hay software cliente para CUDA y para AMD Stream Computing. Este software realiza cálculos científicos que permiten estudiar el plegamiento de proteínas y sus uniones para luego devolver el resultado al servidor del programa en Internet y tomar nuevos datos para continuar los cálculos. Los biólogos simulan el plegamiento de proteínas para entender cómo se pliegan y descubrir lo que sucede si no lo hacen correctamente. Se cree que enfermedades como el Alzheimer, la fibrosis cística, BSE (la enfermedad de las Vacas Locas), una forma hereditaria de enfisema, y muchos tipos de cáncer son resultado del mal plegamiento de las proteínas. El cliente Folding@home es un programa gratuito que corre en segundo plano en la PC, lo que permite que la gente que lo utiliza tenga un impacto real en la investigación de una cura para estas enfermedades. El uso de los GPU de las tarjetas gráficas en este caso llega a ser 30 veces más rápido que con los software clientes que utilizan procesadores centrales o CPUs. Con estas técnicas se pueden buscar respuestas a preguntas que previamente eran imposibles de hacer debido a la falta de capacidad de cómputo.

Otras aplicaciones que utilizan esta tecnología realizan simulaciones y animación de fluidos de baja viscosidad, microscopía holográfica digital en tiempo real, estudios de la dinámica molecular del ADN, sistemas de reconocimiento facial en tiempo real, aceleración de cálculos de densidad funcional con GPUs, investigación climática, análisis financiero, exploración de gas y petróleo, etc. En algunos casos se consiguen rendimientos 100 veces mayores que con los métodos tradicionales.

Por el momento los desarrollos más importantes de estas tecnologías se encuentran en el ámbito universitario. Para incrementar el uso de CUDA la empresa Nvidia apoya a las universidades con donaciones económicas, equipos y colaboración para montar un cluster si estas dictan cursos e investigan con esta tecnología en sus laboratorios. Pero no es sólo Nvidia la que está desarrollando estas tecnologías. AMD está avanzando en el desarrollo del entorno de trabajo y sus librerías para sus tarjetas gráficas y el fabricante de CPU Intel está desarrollando su primer GPGPU llamado Larabee. Este GPU se utilizará como tarjeta gráfica y coprocesador del CPU por su arquitectura altamente paralela constituida por decenas de procesadores Pentium modificados con nuevas unidades de ejecución para operaciones de grandes conjuntos de datos. La futura tarjeta gráfica de Intel podría estar en el mercado en el 2010 y al poseer una arquitectura X86 podría ser aún más fácil de programar que las actuales. Aunque el producto real de Intel esté todavía lejos del mercado y su desempeño y características sean especulaciones, su desarrollo indica una clara dirección en cuanto al futuro de los sistemas de cómputo complejo.

Conclusión

El uso de las nuevas tarjetas gráficas para acelerar el procesamiento de aplicaciones utilizándolas como coprocesadores es una realidad y está en sus primeras etapas. Las principales empresas fabricantes de GPU, Nvidia y AMD, ya tienen tanto el hardware como el software disponible para que los programadores puedan aprovecharlo. La tecnología está siendo utilizada y brinda soluciones a campos muy diferentes. Incluso el líder en el mercado de los CPU, Intel, esta apostando fuerte a esta nueva tecnología que permite obtener grandes saltos en rendimiento en las aplicaciones de cómputo complejo que pueden programarse utilizando GPUs. La adopción de la misma dependerá de la flexibilidad de los lenguajes y de los programadores. Si esta tecnología triunfa las futuras PC utilizaran su CPU a altas frecuencias para ejecutar el código lineal y los GPU masivamente paralelos para cálculos en conjunto, lo que brindará al usuario nuevas experiencias y posibilidades.

Este artículo fue escrito el 9 de septiembre de 2008

www.nvidia.com/theforcewithin

http://www.badaboomit.com

http://folding.stanford.edu

http://fastra.ua.ac.be/en/index.html

http://www.nvidia.com/object/cuda_home.html#

http://ati.amd.com/technology/streamcomputing/index.html