Universidad Católica "Nuestra Señora de la Asunción"

Facultad de Ciencias y Tecnología

Ingeniería Informática

Teoría y Aplicaciones de la Informática Dos

Trabajo Práctico de Investigación

Tecnologías de Procesamiento Gráfico: GPU y PPU

Daniel Yaluff

10mo Semestre

2005

Asunción - Paraguay

Introducción

El presente trabajo toca el tema referente al hardware gráfico existente en el mercado actualmente y que posibilita la subsistencia de todo un submundo dentro del mundo de los usuarios de PC: el de los videojuegos.

La ejecución de un videojuego en una PC es probablemente la tarea que estresa más a los componentes del mismo, requiriendo grandes capacidades de almacenamiento en disco duro, mucha memoria RAM y de gran velocidad, y por supuesto, una tarjeta aceleradora 3D que posibilite los cálculos matemáticos en masa que se necesitan.

También se hablará un poco de una nueva tendencia tecnológica en procesamiento gráfico: las PPUs.

Tecnologías de Procesamiento Gráfico: GPU y PPU

GPU

Una Unidad de Procesamiento Gráfico o **GPU** (siglas en inglés de Graphics Processing Unit) es el microprocesador de una tarjeta gráfica de una computadora personal o consola de juegos. Las GPUs modernas son muy eficientes manipulando y mostrando gráficos computarizados, su estructura altamente paralelizada las hace más efectivas que las CPUs típicas para un rango de algoritmos complejos.

Una GPU implementa un número de operaciones primitivas gráficas de un modo que las hace ejecutar mucho más rápidas que dibujando directamente a la pantalla con el procesamiento de la CPU. Las operaciones más comunes de las primeras GPUs incluían gráficos en dos dimensiones, ejecutaban una operación llamada BitBLT, que combina dos mapas de bits en una. También eran capaces de digujar figuras geométicas como rectángulos, triángulos, círculos y arcos. Las GPUs modernas tienen también soporte para gráficos en tres dimensiones, y algunas incluyen funciones relativas a video digital.

Las GPUs actuales son capaces de ejecutar en forma optimizada, cálculos que posibilitan la creación de:

1) **Sombras (shading)**: el proceso de sombrear incluye el cálculo de cómo se verían las caras de un polígono en el caso de que fueran iluminadas por una fuente de luz virtual. El cálculo exacto varía dependiendo no solamente de qué datos están disponibles, sino también de la técnica de sombreado. La sombra es el bloqueo de la luz por algún objeto.

2) **Texturas**: las superficies de los polígonos (la secuencia de caras) pueden contener datos que corresponden no solo a un color, en programas más avanzandos pueden ser impresiones de mapas de bits o algún otro tipo de imagen. Tal imagen se sitúa en una cara, o series de caras. La textura agrega un nuevo grado de personalización acerca de cómo las caras y polígonos se verán después de empezar a ser sombreados, dependiendo del método de sombreado, y cómo la imagen es interpretada durante el mismo. Una imagen (la textura) es adherida a una figura geométrica más simple que se genera en la escena.

3) **Rendering**: es el proceso de generar una imagen de un modelo, por medio de un programa. El modelo es una descripción de objetos en tres dimensiones en un lenguaje estrictamente definido. El "rendereado" es un proceso lento y computacionalmente intensivo.

Cada una de las capacidades anteriormente citadas pueden descomponerse en más capacidades, como lo son: colisiones, efecto de neblina, reflección, transparencia, refracción, iluminación indirecta, profundidad de campos, motion blur, etc.

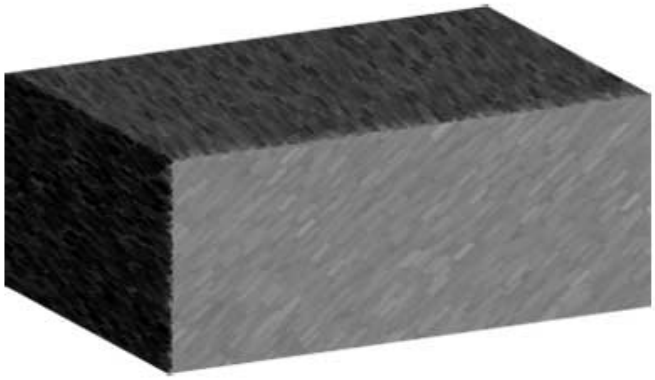


Figura 1 – Sombreado

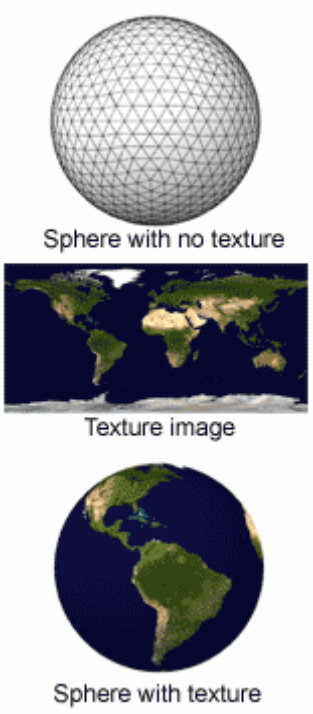


Figura 2 - Texturas

Historia

Las modernas GPUs descienden de chips gráficos monolíticos de la década de los 70s y 80s. Estos chips tenían soporte limitado de BitBLT, si es que tenían, y usualmente no soportaban el dibujado de figuras greométicas. Algunas GPUs podían ejecutar operaciones dentro de una lista de visualizaciones, y podían usar DMA para reducir la carga de la CPU. Un ejemplo es el procesador gráfico ANTIC, usando en la Atari 800 y Atari 5200.

A medida que la tecnología de chips de procesamiento fue evolucionando, se volvió viable poner el soporte de dibujo de formas y BitBLT en la misma placa de video. Estos no eran tan flexibles como los GPUs basados en microprocesadores, pero eran más fáciles de fabricar y de vender. El primer mercado masivo que incluyó esta clase de placas de video fue la de la consola Amiga.

En los comienzos de la década de los 90s, el uso de Microsoft Windows hizo aparecer la necesidad de crear chips de gráficos en 2D rápidos y de alta resolución.

En 1991, **S3 Graphics** introdujo el primer chip acelerador 2D, el **S3 86C911**. El 86C911 creó un número de imitadores; para 1995, cada fabricante de chips gráficos habían adherido aceleración 2D a sus chips.

Con el advenimiento de **DirectX** y **OpenGL**, las GPUs fueron adheridas con capacidades con las cuales se podían programar las sombras. Cada píxel podía ser procesado por un pequeño programa que podía incluir texturas de imágenes adicionales, cada vértice geométrico podía también ser procesado por un pequeño programa antes de ser proyectado en la pantalla. Para el 2003, con la introducción del chip **Nvidia GeForce FX**, vértices y pixeles con sombras podían implementar lazos, procesamientos largos de funciones matemáticas que involucraban números con punto flotante, y rápidamente fueron volviéndose tan flexible como la CPU para operaciones que incluían tratamiento de imágenes.

Hoy, las GPUs paralelas han comenzado a incursionar en procesos computacionales contra la CPU, un subcampo de investigación, llamado **GPGPU** o *General Purpose Computing on GPU*, ha tenido aplicaciones en procesos como exploración petrolífera, procesamiento científico de imágenes e incluso en la determinación de los precios de elementos de stock.

Capacidades actuales de las GPUs

Las modernas GPUs usan la mayoría de sus transistores para hacer cálculos relacionados a gráficos computarizados en 3D. Ellas empezaron acelerando las tareas que requerían un uso intensivo de la memoria, como mapeo de texturas (texture-mapping) y “rendereo” de polígonos, después adhiriendo unidades para acelerar cálculos geométricos tales como mapeo de vértices en diferentes sistemas de coordenadas. Desarrollos recientes en las GPUs incluyen soporte para sombras programables que pueden manipular vértices y

texturas con muchas de las operaciones soportadas por las CPUs, muestreo múltiple (oversampling) y técnicas de interpolación que reducen el aliasing, y formatos que permiten colores con muy alta precisión. Dado que muchas de estas computaciones incluyen cálculos de matrices y vectores, los ingenieros y científicos están estudiando fuertemente cómo usar las GPUs para cálculos que no incluyen gráficos.

En adición a las capacidades de hardware para acelerar los gráficos 3D, las GPUs actuales incluyen aceleración básica de gráficos 2D y capacidades de frame buffer. También, desde 1995 soportan el espacio de colores YUV (pareceido al RGB), primitivas MPEG e iDCT (*invese Discrete Cosine Transform*, usado comúnmente en software que trabaja con diferentes formatos multimedia, como MP3, Vorbis, JPEG).

Las más actuales y modernas GPUs vienen en una tarjeta gráfica separada de la placa madre, conectada a la CPU y memoria RAM a través de un bus AGP o PCI Express. Tiene acceso a una memoria RAM en la tarjeta que es normalmente más rápida pero de menor capacidad que la memoria RAM principal. Por otro lado, muchas placas madres tienen una GPU integrada en el chipset Northbridge (encargado de controlar la memoria RAM y la ranura AGP) que usa la memoria principal como frame buffer. Esta es una solución más barata que una GPU independiente, pero reduce drásticamente el desempeño.

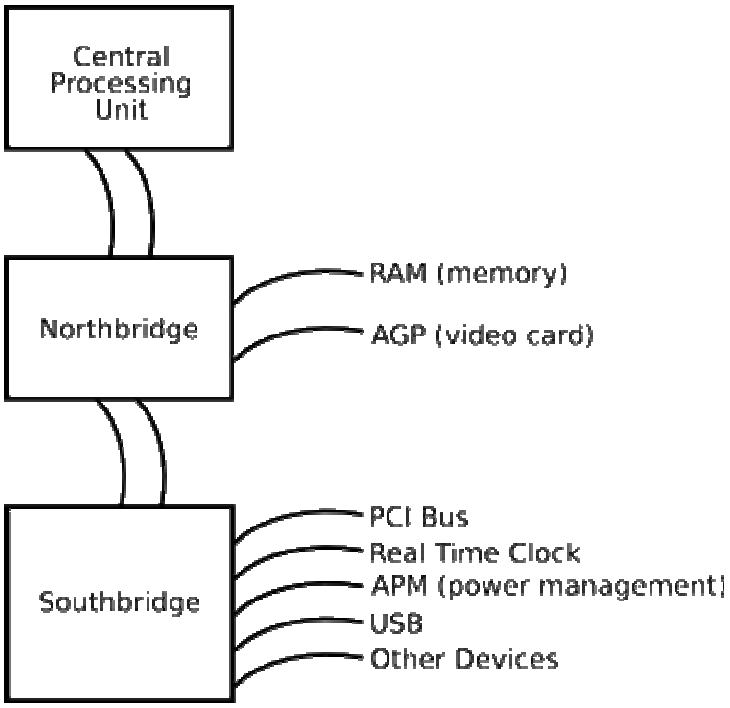


Figura 3 – Arquitectura de una placa madre

Fabricantes de GPUs:

- NVIDIA Corporation
- ATI Technologies
- 3Dlabs
- Martos
- XGI Technology Inc.
- Intel

GPGPU

General-Purpose Computing on Graphics Processing Units o Computación de Propósito General en GPUs, es una tendencia reciente en las ciencias de la computación que usa la GPU para hacer cálculos en vez de utilizar la CPU.

Los programadores de juegos han tenido la necesidad de crear efectos visuales más realísticos, estos han llevado a la creación de sombreado de vértices programable y sombreado de fragmentos. El sombreado de vértices es un programa que corre en la GPU y permite a los programadores especificar cómo los vértices de una figura geométrica se mostrarán en pantalla. El sombreado de fragmentos también corre en la GPU y permite a los programadores especificar cómo cada fragmento será coloreado. Naturalmente, más bits por píxel permiten un mayor rango de colores disponibles, así que las tarjetas gráficas poseen un número incrementado de bits con los que pueden representar cada componente de un color.

Ahora, si combinamos estos programas con aritmética de alta precisión y el hecho de que el poder computacional de los procesadores gráficos ha ido aumentando en una tasa increíblemente alta, será fácil notar porqué las GPUs están siendo vistas como unidades poderosas de procesamiento en áreas ajenas a lo que fue su cometido principal: los gráficos y la visualización.

Las siguientes son algunas áreas donde las GPUs han sido usadas para computación de propósito general:

- Criptografía
- Operaciones en base de datos
- Segmentación
- Procesamiento de efectos de sonido
- Redes neuronales
- Química

No todas las aplicaciones pueden ser transportadas para el procesamiento con GPUs. Algunas aplicaciones simplemente no están preparadas para la arquitectura de los

procesadores gráficos actuales. Más aún, la precisión en una GPU está limitada a 24 bits para tarjetas ATI y a 32 bits para tarjetas NVIDIA. Esta limitación en la precisión disponible en las GPUs previene algunos tipos de computaciones en procesadores gráficos.

Tarjetas Gráficas

Una tarjeta gráfica o tarjeta de video, es un componente de una computadora que está designada a convertir una representación lógica de una imagen guardada en memoria a una señal que puede ser usada como entrada en algún medio de visualización, normalmente un monitor. También provee funcionalidad para manipular la imagen lógica en memoria. Las tarjetas gráficas son insertadas en una ranura de expansión de la placa madre, pero a veces también vienen incorporadas a la computadora.

Cada vez más, las tarjetas gráficas están dejando de ser simplemente tarjetas en el sentido estricto, se están convirtiendo en una sección de la placa madre dedicada al mismo propósito, a pesar de que tengan un desempeño 3D mucho menor comparadas con las tarjetas gráficas dedicadas debido al uso de chipsets más baratos y la compartición de la memoria del sistema en vez de usar memoria especial (a pesar de este no es siempre el caso); aquellos que requieren desempeño todavía prefieren soluciones no integradas.

Todo lo referente a hardware gráfico potente, normalmente usado para gráficos 3D en juegos, son todavía basados en tarjetas. Sus motores de procesamiento son llamados GPUs. Nuevos productos y tecnologías prometen ofrecer “calidad de Hollywood”, **3dfx** decía que podía generar efectos con calidad de película al promover sus tarjetas *Voodoo 5* con tecnología T-Buffer, permitiendo movimiento difuminado, profundidad de campo y pantalla completa con alisado. **nVidia** hablaba acerca de “el nacimiento de la computación cinemática” cuando introducía su chip *GeForce FX*.

El hardware 3D venía originalmente en una placa que se usaba en conjunción con tarjetas gráficas normales. Las placas adherían gráficos 3D a los 2D de la tarjeta gráfica por medio de un cable. La primera tarjeta gráfica 3D de consumo masivo fue la *Voodoo*, de la ahora extinta **3dfx**.

Las tarjetas 3D usadas en animación son diferentes a las usadas en juegos. La serie “Quadro” de **nVidia**, que puede costar más de 1000 dólares, está designada para renderear animaciones, mientras que la serie *GeForce* que cuesta mucho menos, está diseñada para juegos.

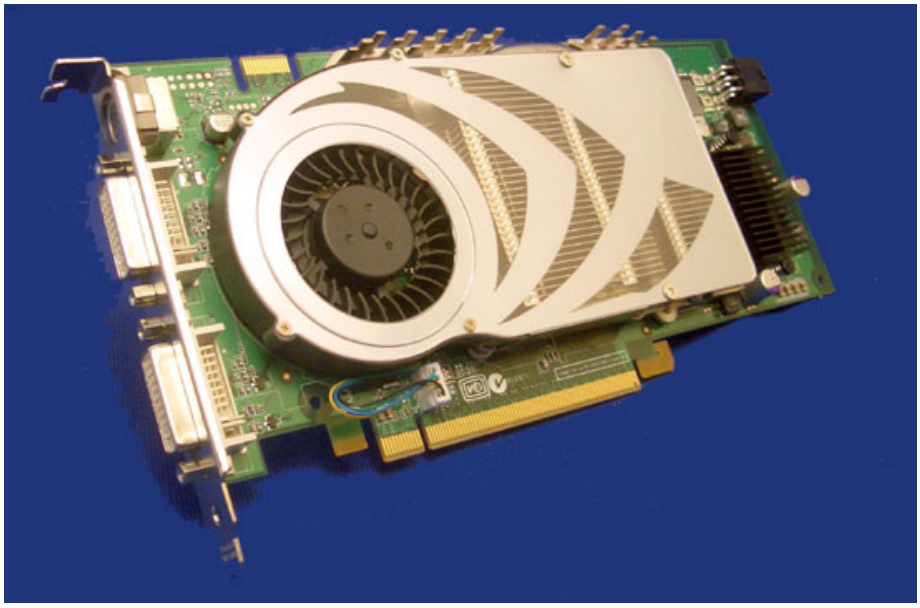
NVIDIA GeForce 6800 Series

El primer modelo, *GeForce 6800 Ultra* fue lanzado en abril del 2004 con la intensión de compensar las limitaciones que tenía la serie *GeForce FX*, en particular poco desempeño en sombreado y excesivo consumo de energía.

Notablemente, la 6800 es de 50 a 80 por ciento más rápida que la *GeForce FX 5950 Ultra*, con menos consumo de energía. Soporta memoria RAM GDDR3 de alta velocidad. Los primeros modelos usaban bus AGP, pero los más nuevos usan bus PCI-Express.

Existe una nueva característica soportada por la *GeForce* serie 6, es la interfase de vínculo escalable (SLI, Scalable Link Interface), que permite a dos tarjetas idénticas conectarse mediante un puente, con el driver ajustando la carga de trabajo a los dos chips en forma dinámica.

GeForce 6800 Ultra - Extreme Edition:



- Reloj del núcleo: 450 MHz
- Reloj de la memoria: 1200MHz
- Memoria: 512 MB GDDR3 con una interfase de 256 bits
- Precio: $600 USD
- Sistema de refrigeración: ventiladores.
- Conectores de energía: 2

Tendencias Tecnologias: PPU

Para finales del año 2005, la empresa **AGEIA** tiene programado el lanzamiento del primer chip del mundo dedicado al procesamiento de cálculos físicos, sus siglas en inglés son PPU (Physics Processing Unit), llamado **physX**.

Tal lanzamiento trae al recuerdo el uso de una ranura especial en placas madres antiguas (aproximadamente 1993), donde se podía insertar un co-procesador matemático, que lo que hacía era tomar el trabajo de tener que hacer cálculos matemáticos, dejando así al procesador principal ejecutar otras instrucciones, automentando así la velocidad de procesamiento de los datos y la ejecución del proceso en cuestión. Vale acotar que las placas madres actuales ya no traen esta ranura especial, porque los procesadores de esta época ya traen incorporado dentro de su circuitería una porción especial para los cálculos matemáticos.

AGEIA promete que el chips physX será capaz de realizar cálculos físicos a una velocidad de 100x más que las más rápidas CPUs existentes actualmente.

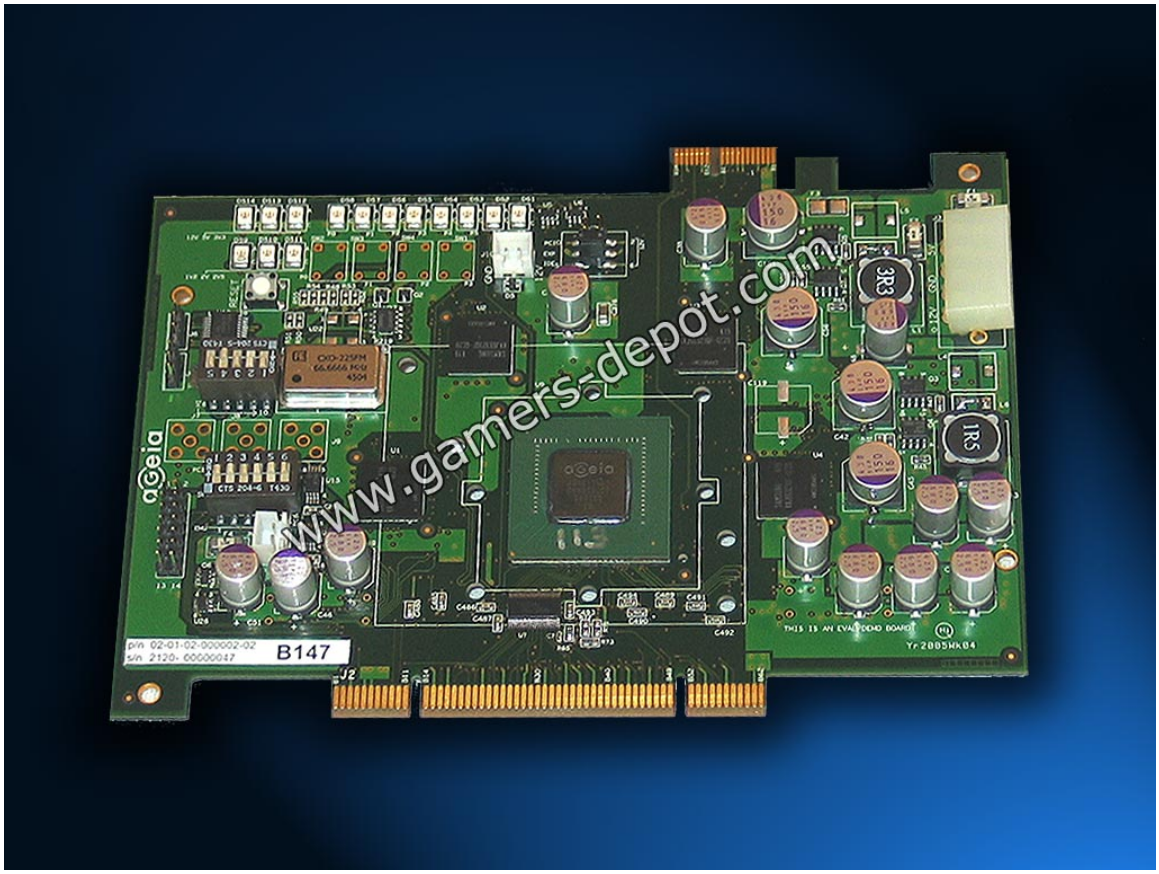


Figura 4 - physX

El procesaro physX es el primer procesador físico del mundo, es completamente una nueva categoría de tecnología que promete transformar todo lo relacionado a video juegos. Extrayendo de la CPU y la GPU todo lo referente al cálculo físico, physX completa el triángulo de la función del juego, entornos gráficos e interactivos en tiempo real, balanceando la carga de estos procesos y permitiendo realismo increíble en los juegos del mañana.

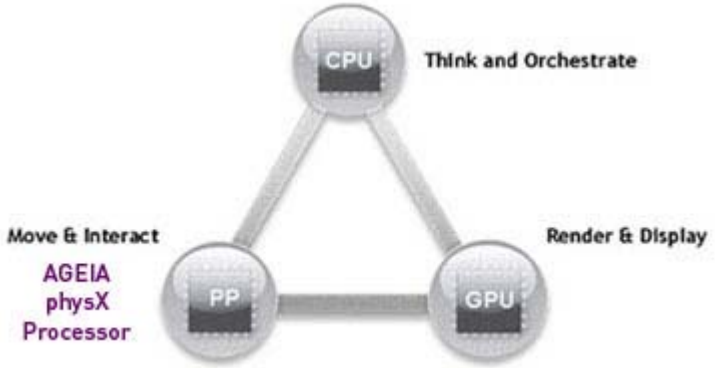


Figura 5 – Triángulo de los videojuegos.

El mercado potencial de compradores de physX son aquellas empresas y personas particulares que se dedican a crear juegos para PC o consolas, y por supuesto, a aquellas

personas que utilizan la PC como una estación de juegos. También apunta a usuarios de programas como 3D-Studio y otros de simulación gráfica.

La arquitectura del procesador physX ha sido diseñada para permitir una aceleración radical de:

- Propiedades de la materia: son características físicas como la densidad, fricción y robustez. Los diseñadores pueden crear superficies resbaladizas en las cuales es difícil caminar, objetos de madera que se curvan y se quiebran, superficies de caucho que rebotan, superficies de metal que resisten curvaturas pero que se abollan por extrema fuerza, y piedras que se rompen cuando son golpeadas con mucha fuerza.
- Movimientos de cuerpos rígidos y detección de colisión: son tecnologías de simulación que proveen movimiento Newtoniano creíble a los objetos dentro del juego. La confianza de un jugador por un juego se desgasta cuando objetos pasan a través de otros.
- Conexiones y resortes: son herramientas para modelar mecanismos complejos, yendo más allá de las canastas, entrando en el reino de los vehículos, movimientos de personajes, puertas y elevadores, y la habilidad del jugador de alzar y manipular objetos en el mundo.

Especificaciones:

- Memoria: 128 MB GDDR3
- Instalación: PCI y PCI-Express
- Número de transistores: 125 millones
- Consumo de energía: 25 watts
- Fecha de lanzamiento: octubre del 2005
- Precio aproximado: 200 dólares.

Algunas empresas que están apostando tempranamente a physX:

Anexos

Sh Embedded Metaprogramming Language
Sh es un lenguaje de metaprogramación para GPUs programbles, desarrollado en la Universidad de Waterloo. Un lenguaje de alto nivel permite programar GPUs con sintaxis y construcciones familiares, sin preocuparse de los detalles del hardware. Ofrece la conveniente sintaxis de C++ y evita la carga que significa el manejo de registros y otros asuntos de bajo nivel al programador. Esto permite que los programas para GPUs sean escritos más rápidamente y con fácil portabilidad. Sh es un proyecto open-source.

Brook para GPUs: Stream Computing para hardware gráfico
Brook para GPUs un sistema de computación de propósito general sobre hardware gráfico programable. Brook extiende C para incluir simples construcciones de datos paralelos, permitiendo el uso de las GPUs como un co-procesador de streaming.

Ordenamiento de alta performance con GPUs
Se desarrolló un algoritmo de ordenamiento bitónico con eficiencia de caché para GPUs. El algoritmo utiliza hardware programable y texture mapping para ordenar datos de punto flotante IEEE de 32-bit incluyendo punteros, y ha sido utilizado para realizar data mining y queries de bases de datos relacionales. Los resultados indicaron un significativo aumento de performance sobre los algoritmos basados en CPU.

Rápidas operaciones de bases de datos utilizando procesadores gráficos
Se probaron operaciones como selecciones de conjuntos, agregaciones, queries semi lineares, que son componentes esenciales de una base de datos, datawarehouse. Se utilizó una GeForce FX 5900 y los resultados indicaron que los procesadores gráficos constituyen efectivos coprocesadores en operaciones de bases de datos.

BionicFX utiliza la GPU como poderoso procesador de efectos de audio
BionicFX anuncia una revolucionaria tecnología para producción de música que convierte las tarjetas de video NVIDIA en procesadores de efectos de audio. Audio Video Exchange(AVEX) convierte audio digital en datos gráficos, y luego realiza cálculos de efecto usando la arquitectura 3D de la GPU. Las últimas tarjetas de NVIDIA son capaces de más de 40 gigaflops de "power processing" comparado a los menos de 6 gigaflops de los procesadores Intel y AMD. AVEX representa un gran logro tecnológico que permite a los aficionados a la música o profesionales correr efectos de calidad de "estudio" a altas tasas de muestreo en sus PCs de escritorio.

Otros procesadores dedicados

Sonido	Video
nVidia nForce 2 y 3	ATI Imagen para PDAs
GMUK1 de Creative	
Cirrus Logia CM 5XXX Series	
Via Envy Z4	

La utilización de una PPU permitirá ejecutar más rápido los siguientes algoritmos:

- **Análisis de elementos finitos**: es una técnica numérica para una gran variedad de problemas. Fue desarrollado en los años cuarenta para usar en análisis estructural. En esta aplicación, el objeto es representado por un modelo similar geométrico, consistiendo de múltiples, simplificadas representaciones de regiones discretas (por ejemplo, elementos finitos). Ecuaciones de equilibrio, en conjunción con consideraciones físicas aplicables tales como compatibilidad y relaciones constitutivas, son aplicadas a cada elemento, permitiendo la creación de un sistema de ecuaciones simultáneas. El sistema de ecuaciones es resuelta para valores desconocidos usando las técnicas de algebra lineal.

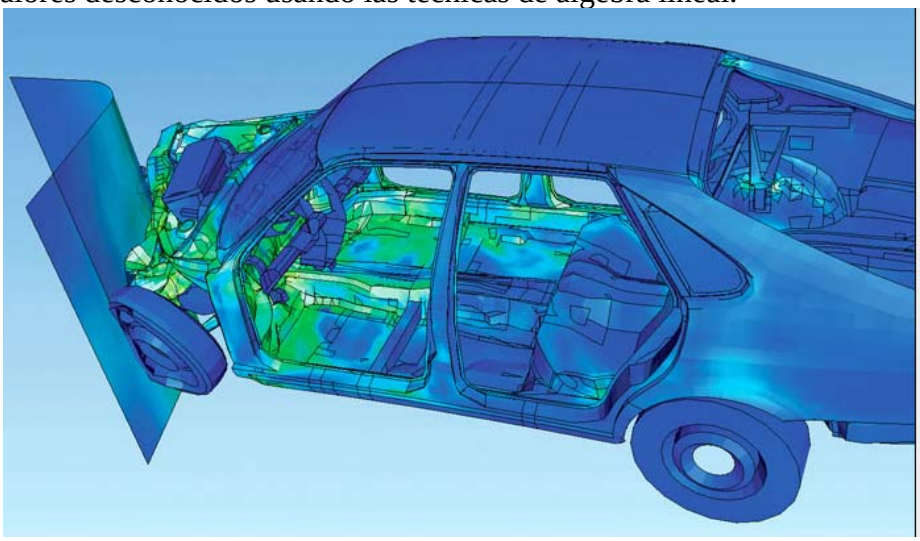


Figura 6 – Visualización de cómo un automóvil se deforma en un choque asimétrico usando análisis de elementos finitos.

- **Movimientos de fluidos**: es el estudio macroscópico del comportamiento físico de los fluidos. Los fluidos son específicamente líquidos y gases, aunque algunos otros materiales y sistemas pueden ser descriptos de una manera similar. La solución de un problema de movimiento de fluidos envuelve cálculos de varias propiedades del fluido, tales como velocidad, presión, densidad y temperatura, son funciones del tiempo y espacio.

Figura 7 – Gráfico del video de demostración

- **Detección de colisiones**: en simulaciones físicas, video juegos y geometría computarizada, la detección de colisiones incluye algoritmos para chequear la intersección entre dos cuerpos sólidos, para calcular trayectorias, tiempos de impacto y puntos de impacto en una simulación física.

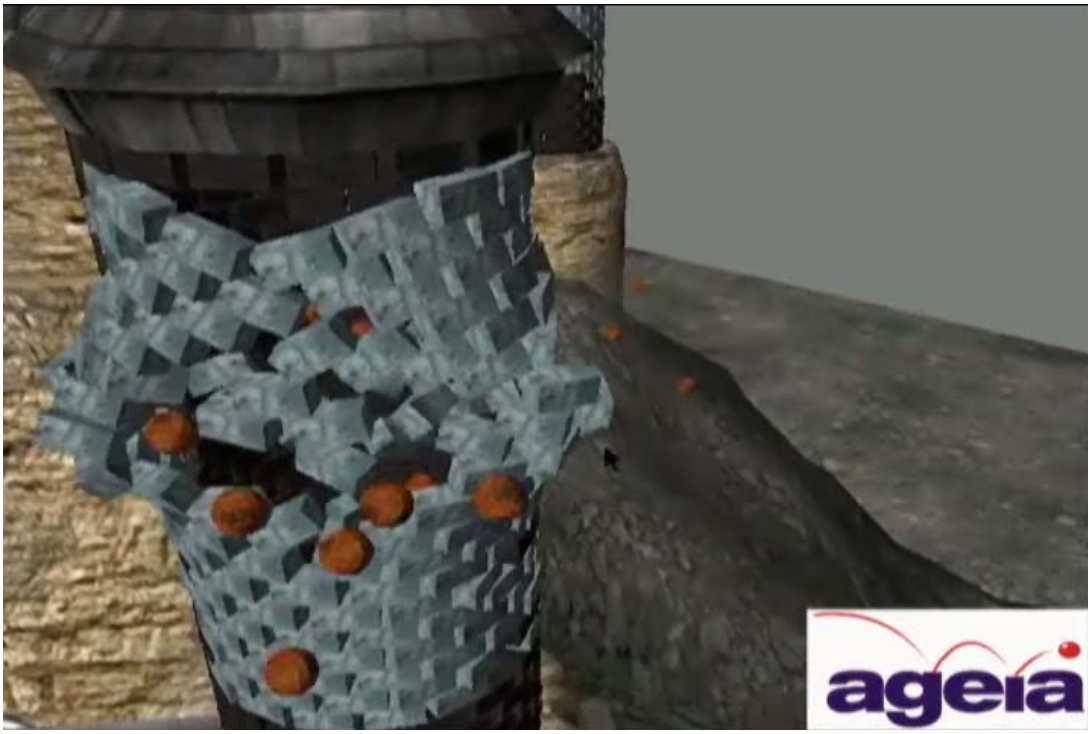


Figura 8 – Gráfico del video de demostración.

Conclusión

La capacidad de procesamiento gráfico de las GPUs actuales está creciendo en forma constante, pero mucho más rápida que la de las CPUs, llegando hasta el punto de en cierto modo “reemplazar” la CPU por la GPU para ciertos cálculos matemáticos. Esto se debe a que las GPUs tienen mucha precisión y ejecutan con extrema rapidez ciertos tipos de cálculos que incluyen punto flotante, también se debe a que poseen su propia memoria RAM de gran ancho de banda que permiten que los datos fluyan mucho más rápido entre la GPU y su memoria RAM (gráfica) que entre la CPU y la memoria principal.

El futuro de la PPU es completamente incierto, AGEIA promete que revolucionará tanto como cuando las GPUs agarraron el mercado masivo de los gamers. Otros piensan que será solo algo pasajero, que a medida que la CPU vaya aumentando en capacidad de cómputo, será posible tener una PPU dentro de la misma CPU, dejando a physX completamente de lado. Particularmente pienso que AGEIA será absorbida por NVIDIA Corp. o por ATI Technologies, y que estos incluirán en sus GPUs sus propias PPUs. Sólo el futuro nos dirá la verdad.

Bibliografía

- www.ageia.com
- www.wikipedia.com
- www.webopedia.com
- www.nvidia.com
- www.ati.com
- www.tomshardware.com
- www.anandtech.com
- www.gpgpu.org