Unidad de Procesamiento Gráfico (GPU).

Graphics Processing Units (GPU).

Autor 1: Miguel Ángel Herrera Solís, Autor 2: David Escamilla Gutiérrez.
Ingeniería de Sistemas y Computación Universidad Tecnológica de Pereira, Pereira, Colombia.
Correo-e: migangel@utp.edu.co, millalml@utp.edu.co.

Resumen—La computación acelerada denominada por GPU, un procesador gráfico, innovación de NVIDIA en 2007, permiten el funcionamiento de centro de datos con eficiencia energética. [4] Los inicios de estos intentos se remonta en la década de los 80, se trataba de pequeños chips controladores (la GPU es descendientes de estos chips gráficos monolíticos). Estas primeras protoGPUs eran bastante básicas, solo cumplían con funciones muy pequeñas, para lo que requiere una unidad aparte en estos tiempos. [1]
Una de las primeras tarjetas gráficas fue el iSBX 275 de Intel, en 1983, insertada en el ordenador Commodore Amiga, uno de los primeros ordenadores en contar con su propia GPU, en 1985. [1]

NVIDIA imparte la necesidad de hacer uso de GPU, en el 2006 se lanza CUDA (Arquitectura Unificada de Dispositivos de Cómputo), con el fin de intensificar el propósito de la GPU. Gracias a CUDA, una herramienta de programación paralela de aparición relativamente reciente, es posible escribir códigos paralelos de propósito general y ejecutarlos en los dispositivos de procesamiento de gráficos (GPU), esto permite aprovechar el poder computacional de las tarjetas gráficas. [5]

Palabras clave—*GPU, NVIDIA, CUDA, Procesamiento paralelo.*

Abstract— The intensive computing named by GPU, a graphic processor, innovation of NVIDIA in 2007, allow the data center functioning with energy efficiency. [4] The beginnings of these attempts goes back in the 80s, it was a question of children chips control (the GPU is progeny of these monolithic graphic chips). These first protoGPUs were quite basic, only they were expiring with very small functions, for what it needs a separate unit in these times. [1]
One of the first graphic cards was the iSBX 275 of Intel, in 1983, inserted in the computer Commodore Amiga, one of the computers first in being provided with its own GPU, in 1985. [1]
NVIDIA gives the need to make use of GPU, in 2006 there is thrown CUDA (Unified Architecture of Devices of Calculation), in order to intensify the intention of the GPU. Thanks to CUDA, a tool of parallel programming of relatively recent appearance, it is possible to write parallel codes of general intention and to execute them in the devices of prosecution of graphs (GPU), this allows to make use of the power computational of the graphic cards. [5]

Key Word —*GPU, NVIDIA, CUDA, Parallel processing.*

I.INTRODUCCIÓN

Debido al poder computacional disponible en las tarjetas gráficas, a fin de solucionar problemas de propósito general nace la idea del GPGPU, (siglas en ingles GPGPU). Inicialmente diseñadas para el procesamiento de gráficos y que ahora permite aprovechar el paralelismo disponible en los actuales dispositivos GPU. [6]

Las actuales unidades de procesamiento de gráficos (GPU) se han convertido en una potente plataforma con un concepto de heterogeneidad en las arquitecturas de computación de múltiples núcleos. Sin embargo, los dominios de aplicación de GPU se limitan actualmente a sistemas específicos. [6]

Los avances más recientes asociados a la tecnología de múltiples núcleos han logrado un aumento de un orden de magnitud en el rendimiento del equipo. [7]

Los ejemplos incluyen las unidades de procesamiento gráfico (GPU), dispositivos de cómputo maduros que mejor abrazan un concepto de funciones heterogéneas de computación de múltiples núcleos. De acuerdo a esto, la GPU está altamente segmentada, lo que indica que posee gran cantidad de unidades funcionales. Estas unidades funcionales se pueden dividir principalmente en dos: aquéllas que procesan vértices, y aquéllas que procesan píxeles. Por tanto, se establecen el vértice y el píxel como las principales unidades que maneja la GPU. Siendo capaz de realizar manipulaciones de gráficos a una velocidad muy superior a lo que pueden hacerlo las CPUs (Central Processing Units o unidades centrales de proceso). [7]

Las GPU actualmente disponen de gran cantidad de primitivas, buscando realismo en los efectos. [3]

II. CONTENIDO

Graphics Processing Unit, o en nuestra lengua, la Unidad de Procesamiento Gráfico. Se trata de un procesador que se dedica exclusivamente al procesamiento de gráficos u operaciones de "coma flotante". Lo que hace la GPU es aligerar de trabajo a la CPU, sobre todo a la hora de abrir juegos o aplicaciones con gráficos interactivos 3D. [8]

Pipeline. (Instrucciones de Shader)

Tradicionalmente, los procesadores gráficos funcionan mediante un pipeline de procesamiento formado por etapas muy especializadas en las funciones que desarrollan y que se ejecutan en un orden preestablecido.

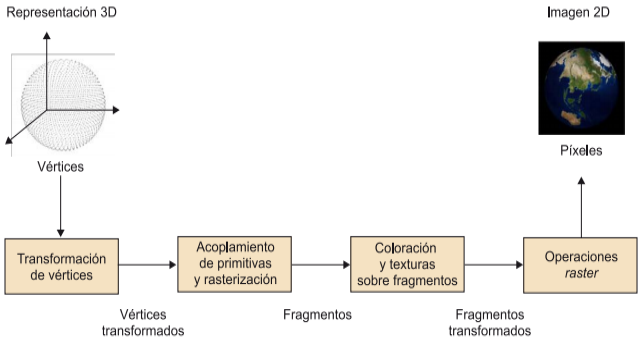

Figura 1. [11]

Aceleración aplicaciones de software mediante GPU.

La computación acelerada por GPU permite asignar a la GPU el trabajo de los aspectos de la aplicación donde la computación es más intensiva, mientras que el resto del código se ejecuta en la CPU. Desde la perspectiva del usuario, las aplicaciones se ejecutan de forma mucho más rápida. [4]

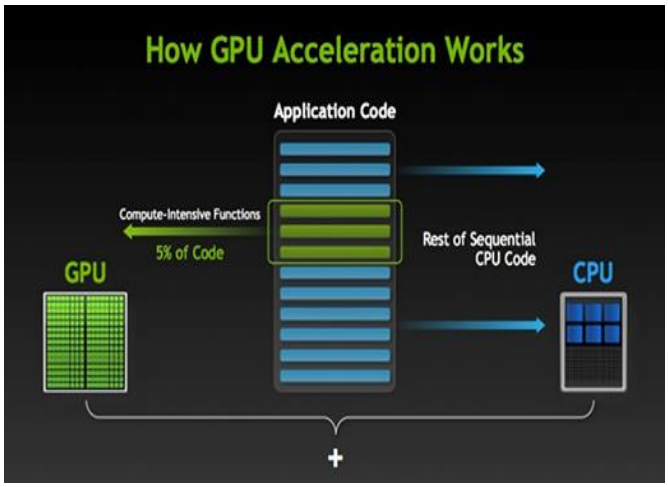

Figura 2. [4]

Rendimiento de la GPU vs. El de la CPU.

Una forma sencilla de comprender la diferencia entre una GPU y una CPU es comparar la forma en que procesan las tareas. Una CPU tiene unos cuantos núcleos optimizados para el procesamiento en serie secuencial, mientras que una GPU cuenta con una arquitectura en paralelo enorme que consiste en miles de núcleos más pequeños y eficaces, y que se diseñaron para resolver varias tareas al mismo tiempo.

Las GPU tienen miles de núcleos para procesar cargas de trabajo en paralelo de forma eficiente.

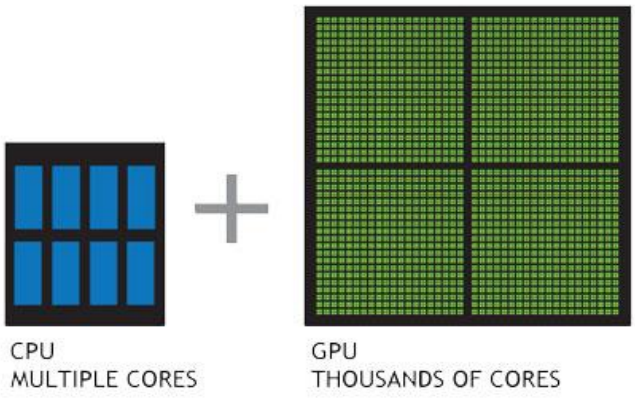

Figura 3. [4]

Tipos de GPU

Podemos distinguir tres tipos de GPU:
Tarjetas dedicadas: son las que mayor potencia proporcionan. Están específicamente diseñadas para esta labor, y se integra a la placa madre a través de un puerto aparte. Tiene una memoria RAM independiente que solo puede ser utilizada por la GPU. No vamos a encontrar de este tipo en terminales Android.

Integrados gráficos: en esta ocasión, la memoria que se utiliza es la del sistema. Esta es la forma presente en smartphones y tablets. Ahora la GPU está integrada en el procesador.

Existen híbridos, mezclas de ambos tipos. Es decir, tienen una pequeña RAM dedicada, pero también utilizan memoria del sistema.

¿Cómo funciona una GPU?

A diferencia de los procesadores, con pocos núcleos y alta velocidad, las GPU tienen muchos núcleos de procesamiento a velocidades bajas. Están dirigidos a dos funciones diferentes, el procesamiento de vértices y el de píxeles. [8]
El procesamiento de vértices se dedica a obtener información de éstos, previamente calculada por la CPU, y procesar su ordenamiento, espacio y rotación, así como qué segmento del vértice será gráficamente visible, para posteriormente pasar al pixelado. [8]
El procesamiento de píxeles, o dicho más fácil, los gráficos que vemos como tal, es muy complejo y necesita de mucho más procesamiento. En él se aplican todas las capas y efectos necesarios para crear texturas complejas y obtener gráficos lo más realistas posibles. [8]

Una vez procesado todo, se transporta a un monitor digital, en este caso, la pantalla de nuestro smartphone o tablet. [8]

Las aplicaciones que pueden aprovechar mejor las capacidades de las GPU son aquellas que cumplen las dos condiciones siguientes:

- trabajan sobre vectores de datos grandes.
- tienen un paralelismo de grano fino tipo SIMD

Arquitectura de la GPU

GPU está altamente segmentada, lo que indica que posee gran cantidad de unidades funcionales. Estas unidades funcionales se pueden dividir principalmente en dos: aquéllas que procesan vértices, y aquéllas que procesan píxeles. Por tanto, se establecen el vértice y el píxel como las principales unidades que maneja la GPU.
Uno de los principales inconvenientes a la hora de trabajar con GPU es la dificultad para el programador a la hora de transformar programas diseñados para CPU tradicionales en programas que puedan ser ejecutados de manera eficiente en una GPU. Por este motivo, se han desarrollado modelos de programación, de propiedad (CUDA) o abiertos (OpenCL).

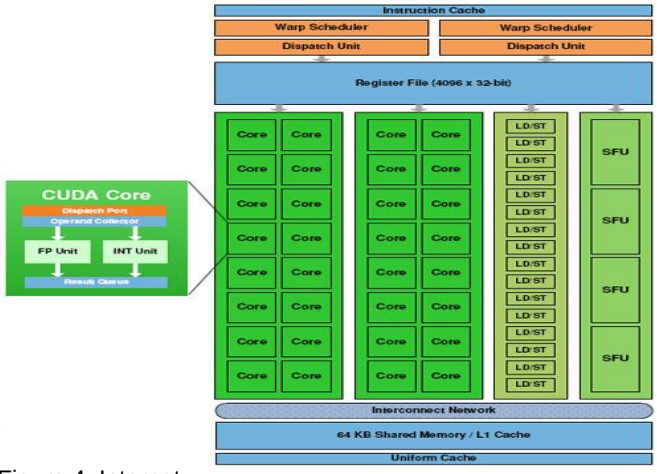

Figura 4. Internet.

SIMD (del inglés Single Instruction, Multiple Data, en español: "una instrucción, múltiples datos") es una técnica empleada para conseguir paralelismo a nivel de datos, útil para implementación y formación de las componentes GPU.

Nvidia (arquitectura G80) arquitectura unificada, vino con el modelo de programación CUDA. Sin diferenciación a nivel de hardware entre las diferentes etapas que forman el pipeline gráfico.

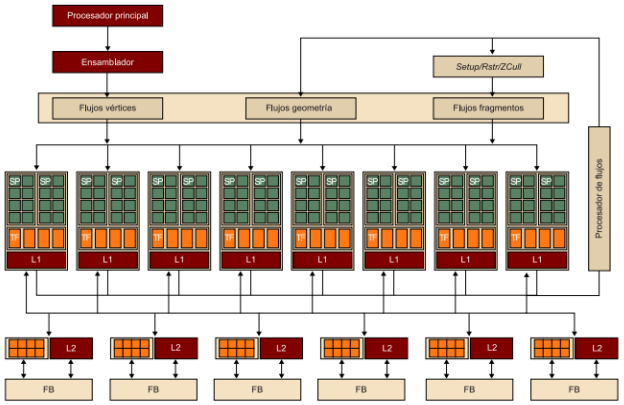

Figura 5. *(Internet) Arquitectura (unificada) de la serie G80 de Nvidia.*

Mediante la arquitectura paralela CUDA Core de NAVIDIA, aprovecha la gran potencia de la GPU (unidad de procesamiento gráfico) para proporcionar un incremento extraordinario del rendimiento del sistema. Donde los sistemas informáticos están pasando de realizar el "procesamiento central" en la CPU a realizar "coprocesamiento" repartido entre la CPU y la GPU. [9]

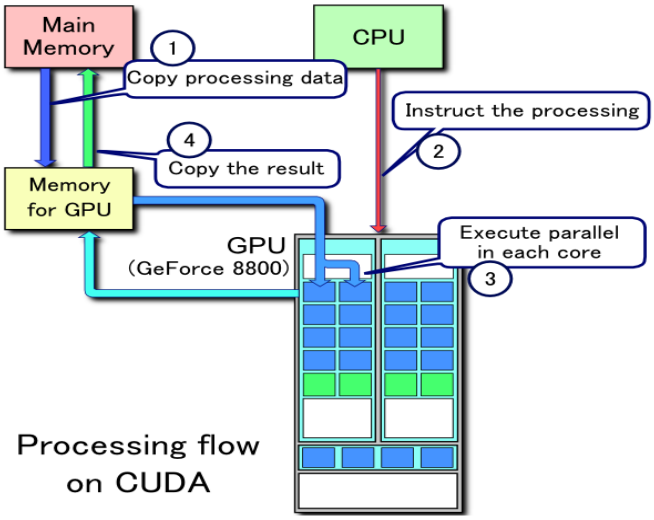

Figura 6. Internet

Concepto de arquitectura unificada.

La arquitectura de la serie GeForce 6 estudiada con anterioridad se podría definir como una arquitectura dividida a nivel de shaders o procesadores programables. Esto quiere decir que dispone de un hardware especializado para ejecutar programas que operan sobre vértices y otro dedicado exclusivamente a la ejecución sobre fragmentos. A pesar de que el hardware dedicado se puede adaptar bastante bien a su función, hay ciertos inconvenientes que hacen que se haya optado por arquitecturas totalmente diferentes a la hora de desarrollar una nueva generación de procesadores gráficos, basados en una arquitectura unificada. [10]
Comparativa de la asignación de procesadores de una GPU en el procesamiento de vértices y de fragmentos en arquitecturas unificadas y no unificadas, para diferentes tipos de aplicaciones. [12]

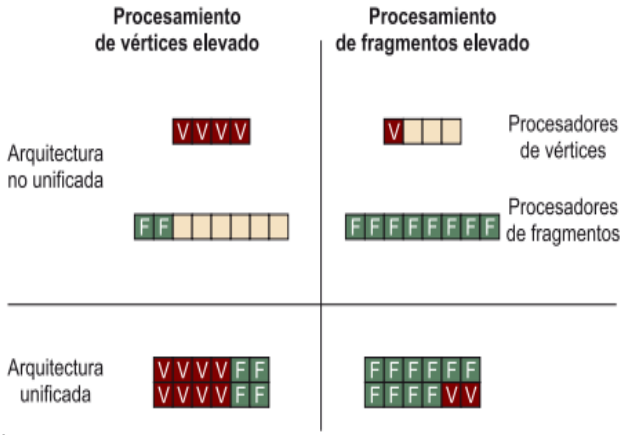

Figura 7. *Internet.*

La arquitectura unificada permite ejecutar más computación simultánea y mejorar la utilización de los recursos. La solución que se desarrolló a partir de la serie G80 de Nvidia. En este tipo de arquitecturas, no existe la división a nivel de hardware entre procesadores de vértices y procesadores de fragmentos. Cualquier unidad de procesamiento que las forma (denominadas también stream processors) es capaz de trabajar tanto a nivel de vértice como a nivel de fragmento, sin estar especializada en un tipo en concreto. [11]
Este tipo de arquitecturas ofrece un potencial mucho mayor para hacer computación de propósito general. [10]

Modelo de memoria.

Es importante destacar que la memoria del procesador principal y del dispositivo son espacios de memoria completamente separados. Esto refleja la realidad por la que los dispositivos son típicamente tarjetas que tienen su propia memoria DRAM. Para ejecutar un kernel en el dispositivo GPU. [12]

Kernel y sus funciones.

Kernel ó núcleo, como el software que constituye una parte fundamental del sistema operativo. Es el principal responsable de facilitar a los distintos programas acceso seguro al hardware de la computadora él es el encargado de gestionar recursos, a través de servicios de llamada al sistema, también se encarga de decidir qué programa podrá hacer uso de un dispositivo de hardware y durante cuánto tiempo, lo que se conoce como multiplexado. Acceder al hardware directamente puede ser realmente complejo, por lo que los núcleos suelen implementar una serie de abstracciones del hardware. Esto permite esconder la complejidad, y proporciona una interfaz limpia y uniforme al hardware subyacente, lo que facilita su uso al programador, sus funciones son las siguientes. [14]

- Administración de la memoria para todos los programas y procesos en ejecución.
- Administración del tiempo de procesador que los programas y procesos en ejecución utilizan.
- La comunicación entre los programas que solicitan recursos y el hardware.
- Gestión de los distintos programas informáticos (tareas) de una máquina.
- Gestión del hardware (memoria, procesador, periférico, forma de almacenamiento, etc.). [11]

Características más importantes de una GPU.

- Escala de integración: Es el tamaño de los transistores y la distancia entre estos dentro del chip. Normalmente se indica en micras y a menor número, mayor cantidad de transistores en el mismo espacio. Al reducir la escala de integración, se puede aumentar la velocidad de un chip y reducir su temperatura. No es un factor determinante a la hora de comprar una tarjeta. [13]

- Frecuencia de funcionamiento: Se mide en Mhz, al igual que en las CPU y, lógicamente, cuanto más mejor, pero claro, siempre con el mismo chip, ya que entre chips distintos no indica mayor rendimiento. Este valor hay que tomarlo con cuidado, ya que para que sirva como referente dos tarjetas solo se pueden diferenciar en la velocidad de funcionamiento de la GPU, si se diferencian en cualquier otra cosa, especialmente la memoria, no sirve de nada. [13]

- Velocidad del bus de memoria: Nos indica a qué velocidad se transmite la información por el bus y viene dada en Mhz. Este factor está mucho más identificado con la memoria que con el procesador, ya que será la primera la que determine la velocidad efectiva del mismo. [13]

- Bus de conexión: Como se comentó en la sección anterior lo más normal hoy en día es AGP o PCI-Express en sus distintas variantes. [13]

- Píxel y Vertex Shaders: Se introdujeron con la familia Geforce3 y permitían que el chip gráfico ofreciera cierta libertad de programación a los desarrolladores de software. Un chip gráfico tiene una serie funciones implementadas, por ejemplo, una función implementada en el chip podría ser que un determinado polígono girase hacia la derecha, sin embargo no puede hacer nada que no esté implementado en una función, es decir, los chips gráficos no son programables. [13]

- Píxel Pipelines, unidades de texturas y demás: Los píxel pipelines son el número de tuberías que trabajan con los píxeles y las unidades de texturas la cantidad de unidades capaces de aplicar una textura a un polígono. [13]

- Generación de software: Esta ha sido una nueva manera de distinguir unos chips de otros aparecida con las GF4MX y consiste en decir que versión de DirectX y OpenGL soportan por hardware las tarjetas. Está íntimamente relacionado con el número y versión de los shaders. [13]

- Ramdac: Este dispositivo empezó estando fuera del propio chip pero poco a poco se ha ido introduciendo dentro de la GPU y aumentando su número, siendo lo más normal que ahora mismo la tarjeta disponga de 2 internos. Su función es transformar la información tratada en la tarjeta para que sea entendida por el monitor, en función de su frecuencia nos dará la máxima que podremos ver. Al tener más de uno nos permite enviar dos señales de video, una al monitor y otra a la TV por ejemplo o a dos monitores distintos. [13]

- Bits de color: Actualmente no se usa como referente, porque todas las tarjetas usan 8 bits por color, pero los chips de nueva generación tienen previsto aumentar este número. [13]

- T&L: Esta es la unidad de “transformación y luces” y apareció con la primera GeForce. Antes de la aparición de esta unidad la aceleradora de video solamente se encargaba de calcular y renderizar triángulos, todo lo demás lo realizaba la CPU, incluido el cálculo de transformaciones e iluminación. A modo resumido diremos que la CPU calcula un objeto, un humanoide.[13]

Aplicaciones.

Las unidades de procesamiento gráfico (GPUs) su aplicabilidad está directamente hacia los dispositivos móviles, computadores o cualquier dispositivo que lo requiera. Uno de los ejemplos más claro de aplicabilidad, lo vemos a diario en los videojuegos que requieren alta gama para funcionalidad, también se puede apreciar en animaciones que requieren una alta dimensión de gráficos, a partir de una estructura determinista en su definición.

Figura 8. [1]

III. CONCLUSIÓN

Las GPUs han comenzado a establecerse como nuevas arquitecturas paralelas. En un principio, su utilización para problemas de propósito general estaba discutida. Sin embargo, el incremento en el desarrollo de aplicaciones sobre GPU y el nacimiento de CUDA para facilitar la adaptación de las mismas a las GPUs ha hecho que se beneficien de la gran aceleración que puede alcanzarse con ellas. Reducir los tiempos de procesamiento con las arquitecturas CPUs del mercado de hoy nos insume un costo mayor comparado con las arquitecturas GPUs, sin mencionar la heterogeneidad de las primeras que requieren de una programación cuidadosa en varios aspectos para lograr la ganancia esperada de su uso. Si bien el gran cuello de botella que se encuentra al trabajar con GPUs está dado por la comunicación de ésta con la CPU, se ha podido sortear este inconveniente dividiendo el tamaño del problema en porciones más pequeñas, lo que permite reducir los datos enviados a la GPU.

- CUDA es una tecnología que permite obtener grandes rendimientos para problemas con un alto paralelismo.
- Hay que tener claro su funcionamiento para saber si es adecuado y obtener el mayor rendimiento posible.
- Programar en CUDA es fácil, pero no lo es obtener rendimiento.

REFERENCIAS

[1]. https://hipertextual.com/archivo/2013/12/hardware-gpu-grafica/

[2]. http://www.taringa.net/posts/info/6075966/Que-es-un-GPU-Unidad-de-Procesamiento-Grafico.html

[3]. http://procesadoresmartphones.wikidot.com/unidad-de-procesamiento-grafico

[4]. http://la.nvidia.com/object/what-is-gpu-computing-la.html

[5]. https://es.wikipedia.org/wiki/CUDA

[6]. http://sedici.unlp.edu.ar/bitstream/handle/10915/41317/Documento_completo.pdf?sequence=1

[7]. https://es.wikipedia.org/wiki/Unidad_de_procesamiento_gr%C3%A1fico

[8]. http://www.androidpit.es/que-es-como-funciona-gpu

[9]. http://www.nvidia.es/object/cuda-parallel-computing-es.html

[10]. https://www.exabyteinformatica.com/uoc/Informatica/Arquitecturas_de_computadores_avanzadas/Arquitecturas_de_computadores_avanzadas_(Modulo_5).pdf

[11]. https://www.exabyteinformatica.com/uoc/Informatica/Arquitecturas_de_computadores_avanzadas/Arquitecturas_de_computadores_avanzadas_(Modulo_5).pdf

[12]. http://www.linguee.com/english-spanish/translation/gpu+architecture.html

[13]. http://sabia.tic.udc.es/gc/Contenidos%20adicionales/trabajos/Hardware/tarjetas%20graficas/gpu.html

[14]. https://es.wikipedia.org/wiki/N%C3%BAcleo_(inform%C3%A