Unidades de procesamiento gráfico (GPU).

Graphics processing units (GPU).

Autor 1: Johanna Alejandra Jurado Uribe ,Autor 2: Juan Carlos Patiño Hernández, Autor 3: Juan Camilo Olmos Oliveros
Ingeniería de Sistemas y Computación Universidad Tecnológica de Pereira, Pereira, Colombia
Correo-e: alejandra.jurado@utp.edu.co

Resumen— En el presente artículo hablaremos acerca de la unidades de procesamiento gráfico- GPU,de su significado, tipos de GPU, de su funcionamiento,tipo de arquitectura que utiliza ,aplicaciones, procesamiento en paralelo y se realizará un comparativo de la evolución ha tenido la arquitectura desde el año 2012 al año actual .

Uno de los métodos aplicados para resolver cómputos que ejecutan grandes cantidades de instrucciones en la CPU es la técnica del paralelismo – resolver un problema en menor tiempo haciendo división del trabajo – acompañada del uso de GPU. Por consiguiente, la idea de librar a la CPU de procesos específicos reiterativos, da inicio al uso de dispositivos independientes capaces de procesar grandes volúmenes de datos.[9]

Las primeras GPU (principios de los 90) nacen exclusivamente como aceleradores gráficos. Sin embargo, con el paso de los años investigadores intervinieron sus enormes capacidades, dando paso a las GPU de propósito general (GPGPU).[3]

A raíz de la imperiosa necesidad por parte de NVIDIA de hacer uso de la GPU de manera mucho mas intuitiva – y por otro lado unir software y hardware – es que el 2006 se lanza CUDA, siendo así la primera gran solución al uso de GPU de propósito general.[3]

La computación paralela nos permite ejecutar procesos simultáneamente. Existen varias formas de lograr el paralelismo en el procesamiento de los datos; una de ellas, es el uso de un dispositivo (GPU) el cual se controla independientemente del HOST (CPU), con posibilidad de comunicación.[8]

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

Abstract— In this article we will talk about GPUs, their meaning, GPU types, their operation, type of architecture that uses, applications, parallel processing and a comparison of the evolution of the architecture From the year 2012 to the current year.

One of the methods applied to solve computations that execute large amounts of instructions in the CPU is the technique of parallelism - solve a problem in less time doing division of labor - accompanied by the use of GPU. Therefore, the idea of ridding the CPU of specific, repetitive processes leads to the use of independent devices capable of processing large volumes of data.[9]

The first GPUs (early 90s) are born exclusively as graphics accelerators. However, with the passage of the years researchers intervened their enormous capacities, giving way to GPUs of general purpose (GPGPU).[3]

As a result of the imperative need for NVIDIA to make use of the GPU in a much more intuitive way - and on the other hand to link software and hardware - is that in 2006 CUDA is launched, being thus the first great solution to use GPU general purpose.3

Parallel computing allows us to execute processes simultaneously. There are several ways to achieve parallelism in data processing; One of them is the use of a device (GPU) which is controlled independently of the HOST (CPU), with possibility of communication.[8]

Key Word —GPU,CUDA,Parallel processing.

I.INTRODUCCIÓN

Ha crecido el interés por la computación en las unidades de procesamiento de gráficos (GPUs en inglés). El hecho de que las GPUs tengan la capacidad de realizar procesamiento paralelo, a veces de forma limitada, ha suscitado un gran interés entre los investigadores cuyas aplicaciones requieren del uso intensivo de recursos computacionales. Las GPUs, aunque inicialmente diseñadas para el procesamiento de gráficos, están compuestas de por algún tipo de procesadores SIMD (Single Instruction Multiple Data), que permite el procesamiento paralelo de instrucciones aritméticas. De ahí que las GPUs sean capaces 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).[14] Internamente las GPUs están compuestas por

Fecha de Recepción: Noviembre de 2016

una gran cantidad de pequeñas y simples ALUs (ArithmeticLogic Units o unidades aritmético-lógicas) capaces de realizar muchas operaciones de forma simultánea. [14]

Los sistemas informáticos están pasando de realizar el "procesamiento central" en la CPU a realizar "coprocesamiento" repartido entre la CPU y la GPU. Para posibilitar este nuevo paradigma computacional, NVIDIA ha inventado la arquitectura de cálculo paralelo CUDA, que ahora se incluye en las GPUs GeForce,ION,Quadroy Tesla GPUs, lo cual representa una base instalada considerable para los desarrolladores de aplicaciones.[2]

II.CONTENIDO

Unidad de procesamiento gráfico o **GPU**(Graphics Processor Unit) es un coprocesador dedicado al procesamiento de gráficos u operaciones de coma flotante, para aligerar la carga de trabajo del procesador central en aplicaciones como los videojuegos o aplicaciones 3D interactivas.[1]

Arquitectura de la GPU y funcionamiento

Una GPU está altamente segmentada,es decir,utiliza la arquitectura paralela ,utilizando la taxonomía de Flynn SIMD(Single instruction multiple data) lo que indica que posee gran cantidad de unidades funcionales. [1]

El paralelismo es una forma de computación en la cual varios cálculos pueden realizarse simultáneamente.
Basado en el principio de dividir los problemas grandes para obtener varios problemas pequeños, que son posteriormente solucionados en paralelo. (Divide y vencerás) .[5]

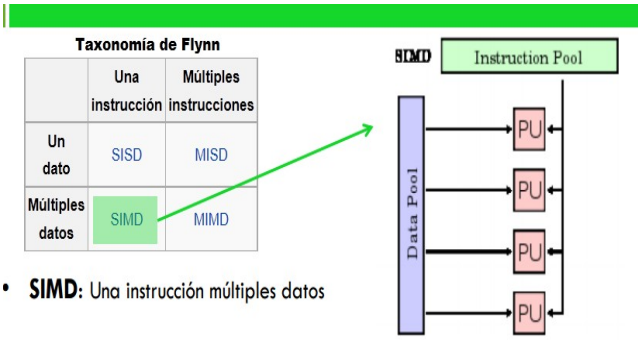


Figura 1. [5]

La GPU se puede considerar como el corazón de la tarjeta gráfica, Junto con la memoria de la tarjeta es el componente al que se refieren las especificaciones, y una buena compenetración entre ambos es determinante para un buen rendimiento.
Se compone, entre otros, del Procesador de Vértices, para hacer cálculos de sombreado, iluminación... y del Procesador de Píxeles, que se dedica exclusivamente a hacer cálculos sobre los píxeles, como efectos de agua. La cantidad y eficiencia de estas unidades determinan la calidad y rapidez con que la tarjeta genera los gráficos. Contiene también unidades para realizar las operaciones ROP(Raster OPeration; el color de cada píxel de la imagen está definido individualmente, se distinguen de las imágenes vectoriales porque no usan ningún tipo de geometría, son en definitiva mapas de bits), mapeado de texturas y otras funciones.[4]

Las características más importantes de una GPU son:

- **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 numero, 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.[4]
- **Frecuencia de funcionamiento:** Se mide en Mhz, al igual que en las CPU y, lógicamente, cuanto mas 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.[4]
- **Velocidad del bus de memoria:**Nos indica a que velocidad se transmite la información por el bus y viene dada en Mhz. Este factor esta mucho más identificado con la memoria que con el procesador, ya que será la primera la que determine la velocidad efectiva del mismo.[4]
- **Bus de conexión:** Como se comento en la sección anterior lo mas normal hoy en día es AGP o PCI-Express en sus distintas variante.[4]
- **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 este implementado en una función, es decir, los chips gráficos no son programables. [4]
- **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.[4]
- **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. Esta íntimamente relacionado con el numero y versión de los shaders.[4]
- **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

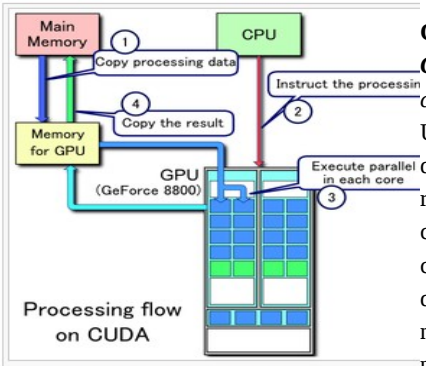
su frecuencia nos dará la máxima que podremos ver. Al tener más de uno nos permite enviar dos señales de video, una al monitor y otra a la TV por ejemplo o a dos monitores distintos.[4]

- **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 numero.[4]
- **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.[4]

Aplicaciones

Las unidades de procesamiento gráfico (GPUs) tienen aplicabilidad en los sistema dispositivo móviles y ordenadores portátiles,de mesa. El cual son desarrollados y mejorados cada día por distintas multinacionales especializada en el desarrollo de unidades de procesamiento gráfico y tecnologías de circuitos integrados para estaciones de trabajo, ordenadores personales y dispositivos móviles,las cuales le realizan cada día mejoras a sus arquitecturas,tanto en el procesamiento de datos (Algoritmos) como en sus estructura ya que contienen más núcleos.[2]

Podemos mencionar unas de las empresas pioneras en el desarrollo de GPUs,la cual desarrollo un procesamiento paralelo llamado CUDA que es una arquitectura de cálculo paralelo.[2]



Ejemplo de flujo de procesamiento CUDA
1. Se copian los datos de la memoria principal a la memoria de la GPU
2. La CPU encarga el proceso a la GPU
3. La GPU lo ejecuta en paralelo en cada núcleo
4. Se copia el resultado de la memoria de la GPU a memoria principal

Figura 3.[7]

CUDA: Son las siglas de ***C****ompute****U****nified****D****evice****A****r****chitecture***(Arquitectura Unificada de Dispositivos de Cómputo) que hace referencia tanto a un compilador como a un conjunto de herramientas de desarrollo creadas por nVidia que permiten a los programadores usar una variación del lenguaje de programación C para codificar algoritmos en GPU de nVidia.[7]

Por medio de wrappers se puede usar Python,Fortran y Java en vez de C/C++ y en el futuro también se añadirá FORTRAN, OpenGL y Direct3D.[7]

CUDA intenta explotar las ventajas de las GPU frente a las CPU de propósito general utilizando el paralelismo que ofrecen sus múltiples núcleos, que permiten el lanzamiento de un altísimo número de hilos simultáneos. Por ello, si una aplicación está diseñada utilizando numerosos hilos que realizan tareas independientes (que es lo que hacen las GPU al procesar gráficos, su tarea natural), una GPU podrá ofrecer un gran rendimiento en campos que podrían ir desde la biología computacional a la criptografía.[7]

Comparativo de la arquitectura de CUDA en el año 2012 al la arquitectura actual 2016.

Realizando un comparativo de la arquitectura del año 2012 enfocado a la parte de la clasificación de los algoritmos se puede observar que los algoritmos de quicksort, merge sort y radix sort (o*rdenación rápida,ordenamiento por mezcla,*y ordenamiento de raíz) no se le ha realizado modificaciones.

Evolución de la arquitectura GPU

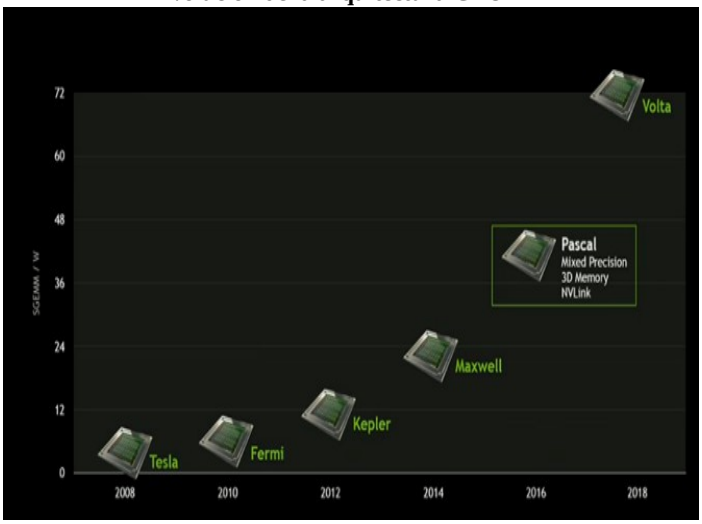


Figura 2. [15]

*Nvidia en el 2012 lanzó la arquitectura Kepler, la arquitectura de alta computación (HPC) .

GPU dinámica con Kepler: Simplifica la programación en la GPU ya que facilita la aceleración de bucles anidados paralelos, lo que significa que una GPU puede iniciar nuevos subprocesos de forma dinámica por sí misma, sin necesidad de volver a la CPU.[12]
Hyper-Q de Kepler: Reduce el tiempo de inactividad de la CPU al permitir que múltiples núcleos de ésta utilicen una misma GPU Kepler, lo que mejora drásticamente la programabilidad y la eficiencia.[12]

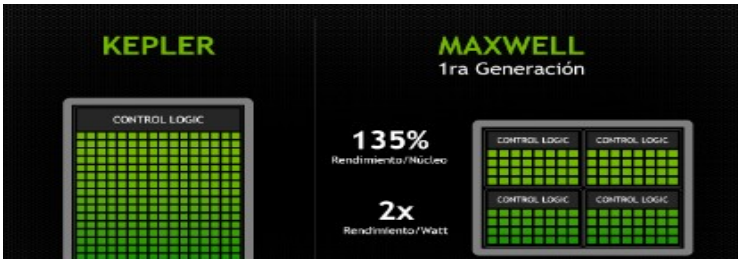


Figura 4.[16]

*En el 2014 Nvidia lanzó otra versión de la arquitectura GPU la cual la denomino MaxWell,fue la es la arquitectura de GPU ,todo en ella gira en torno a la luz y está diseñada para ser el motor de la próxima generación de juegos de PC,la cual tiene buena potencia y rendimiento.[11]

Las GPUs Maxwell introduce niveles de rendimiento de juego y eficiencia energética en las últimas GPUs de la serie GeForce GTX 900. La GeForce® GTX 980 es la tarjeta gráfica más rápida del mundo, mientras que la GTX 970 ofrece el mayor rendimiento de su categoría. Ambas duplican el rendimiento de la anterior generación de tarjetas con menos consumo de potencia y mínimas emisiones de ruido y calor.[11]

VIDEOCARDZ.com NVIDIA GeForce GTX 980 Specifications

	GeForce GTX 980	GeForce GTX 770	GeForce GTX 780 Ti
Architecture	28nm Maxwell	28nm Kepler	28nm Kepler
GPU Codename	GM204-400	GK104-425	GK110-425
Die Size	398 mm^2	294 mm^2	581 mm^2
L2 Cache	2 MB	512 kB	1.5 MB
Transistors	5.2b	3.54b	7.08b
CUDA Cores	2048	1536	2880
TMUs	128	128	240
ROPs	64	32	48
Base Clock	1126 MHz	1046 MHz	875 MHz
Boost Clock	1216 MHz	1085 MHz	928 MHz
Memory Clock	1750 MHz	1750 MHz	1750 MHz
Memory	4GB GDDR5	2GB GDDR5	3GB GDDR5
Memory Bus	256-bit	256-bit	384-bit
Bandwidth	224 GB/s	224 GB/s	336 GB/s
FP Performance (SP)	4.6 TFLOPs	3.2 TFLOPs	5.1 TFLOPs
Pixel Fillrate	72.1 GP/s	33.5 GP/s	53.3 GP/s
Texture Fillrate	144 GT/s	134 GT/s	213 GT/s
Power Connectors	6pin + 6pin	6pin + 8pin	6pin + 8pin
Thermal Design Power	165W	230W	250W

Figura 5.Comparativo con especificaciones entre Kepler y Maxwell [18]

* En el 2016 Nvidia lanzó otra versión de la arquitectura GPU la cual la denomino **Pascal** ,es la versión actual. Es la arquitectura gráfica más avanzada del mundo. Las tarjetas gráficas basadas en Pascal proporcionan mucho más rendimiento y eficiencia energética gracias a un diseño basado en transistores FinFET y funciones DirectX™ 12 para brindar las experiencias de juego más brillantes con la máxima velocidad, fluidez de imagen y eficiencia energética. La GeForce GTX 1080, la GPU estrella de la gama Pascal, incluye también tecnologías GDDR5X (G5X) de elevado ancho de banda para asegurar sesiones de gaming memorables.[6]

Samsung y Apple para sus equipos utiliza en este momento la arquitectura basada en Pascal.

Figura 5.[17]

Como podemos analizar y realizar el comparativo de la evolución de la arquitectura de la GPU durante estos 4 años ,ha venido realizando versiones y mejores a la arquitectura,a la arquitectura Kepler le realizó un versiona miento con la arquitectura MaxWell a la cual se le mejoró la velocidad y consume menos energía,se sacó una nueva versión debido a que diseño de Kepler y durante una operación de cálculo, muchos núcleos CUDA estaban a veces inactivos. [11]Rompieron su lógica de núcleo monolítico, para dar lugar a una serie de componentes independientes, cada uno de los cuales controla un pequeño número de núcleos CUDA. Como resultado, ahora, cuando un procesador individual está en reposo, la lógica de control puede apagarlo.[11]

Y en el estado actual de la arquitectura Pascal de 16nm,fue un versiona-miento realizó a la arquitectura MaxWell, a la cual se le mejoró el rendimiento ,su capacidad para **ejecutar operaciones de Precisión Mixta.** A lo que se refiere es que el GPU podrá realizar operaciones de punto flotante de precisión media (FP16) al doble de velocidad que operaciones punto flotante de simple precisión (FP32). Está mejora es importante debido a que las arquitecturas anteriores Kepler y MaxWell solo pueden ejecutarlas a la misma velocidad. [6]Otra de los atributos especiales de la arquitectura también será **el uso de memoria con diseño tridimensional (3D)**. De esta manera no solo se podrá usar hasta 32GB de RAM sino también **el ancho de banda de memoria a 750 GB/s.**[6]Finalmente, la última pieza revelada hasta ahora es la **interconexión NVLink** que jugará un papel importante en cómo se comunican los GPUs y la posibilidad de **mejorar el escalamiento con hasta 64 GPUs.**[6]

GPU	GeForce GTX 980 (Maxwell)	GeForce GTX 1080 (Pascal)
SMs	16	20
CUDA Cores	2048	2560
Base Clock	1126 MHz	1607 MHz
GPU Boost Clock	1216 MHz	1733 MHz
GFLOPs	4981[1]	8873[1]
Texture Units	128	160
Texel fill-rate	155.6 Gigatexels/sec	277.3 Gigatexels/sec
Memory Clock (Data Rate)	7,000 MHz	10,000 MHz
Memory Bandwidth	224 GB/sec	320 GB/sec
ROPs	64	64
L2 Cache Size	2048 KB	2048 KB
TDP	165 Watts	180 Watts
Transistors	5.2 billion	7.2 billion
Die Size	398 mm^2	314 mm^2
Manufacturing Process	28 nm	16 nm

Figura 6.Comparativo con especificaciones entre Maxwell y Pascal[19]

III.CONCLUSIÓN

Los modelos actuales de GPU suelen tener una media docena de procesadores de vértices (que ejecutan vertex shaders), y hasta dos o tres veces más procesadores de fragmentos o píxeles (que ejecuta fragment shaders). De este modo, hoy en día en las GPU de más potencia, muy baja en comparación con lo ofrecido por las CPU se traduce en una potencia de cálculo mucho mayor gracias a su arquitectura en paralelo. [10]

Gracias a millones de GPUs CUDA vendidas hasta la fecha, miles de desarrolladores, científicos e investigadores están encontrando innumerables aplicaciones prácticas para esta tecnología en campos como el procesamiento de vídeo e imágenes, la biología y la química computacional, la simulación de la dinámica de fluidos, la reconstrucción de

imágenes de TC, el análisis sísmico o el trazado de rayos, entre otras.[2]

Con la arquitectura Pascal se entregará mayor rendimiento en configuraciones de multi-GPU y hasta 5 veces más con una sola tarjeta.

IV. TRABAJO FUTURO

La nueva GPU más rápida que tendrá como nombre **Tesla P100,** estará disponible para los servidores en 2017.[13]

El chip se fabricó usando el proceso de FinFET 16 nanómetros. Además, tiene memoria HBM2 y garantiza un ancho de banda de 256 GBps, es decir, es el doble de rápido que su predecesor, el HBM.[13]

Una nueva interfaz NVLink puede transferir datos a 160 GBps, cinco veces más rápido que PCI-Express. Sin embargo, sigue habiendo dudas sobre como encajaran con estos servidores NVLink. IBM ha confirmado que su arquitectura Power apoyará NVLink, pero que los servidores con chips de Intel utilizan PCI-Express para conectar las GPU con las placas base. [13].El nuevo chip supondrá mejoras importantes en cuanto a sistemas de aprendizaje profundo, velocidad y clasificación de datos en servidores y superordenadores.[13]

V. REFERENCIAS

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

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

[3] http://www.nanocell.cl/introduccion-a-cuda/

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

[5]http://www.dtic.ua.es/jgpu11/material/sesion1_jgpu11.pdf

[6]http://www.nvidia.es/graphics-cards/geforce/pascal/

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

[8]http://sedici.unlp.edu.ar/bitstream/handle/10915/52345/Documento_completo__.pdf?sequence=1

[9] http://dis.um.es/~domingo/09/ERBASE/minicursoespanol.pdf

[10]http://repositorio.uchile.cl/handle/2250/103150

[11]http://www.nvidia.es/object/maxwell-gpu-architecture-es.html

[12]http://www.nvidia.es/object/nvidia-kepler-es.html

[13]http://www.nvidia.es/object/tesla-high-performance-computing-es.html

[14] http://www.congresomaeb2012.uclm.es/papers/paper_52.pdf

[15]http://videocardz.com/specials/roadmaps

[16]https://devblogs.nvidia.com/parallelforall/5-chingas-you-should-know-about-new-maxwell-gpu-architecture/

[17]http://www.nvidia.es/graphics-cards/geforce/pascal/gtx-1080/

[18]Nvidia-GeForce-GTX-980-vs-Nvidia-GeForce-GTX-770-vs-Nvidia-GeForce-GTX-780-Ti1.jpg

[19]gtx-1080-vs-980-100661286-orig.png