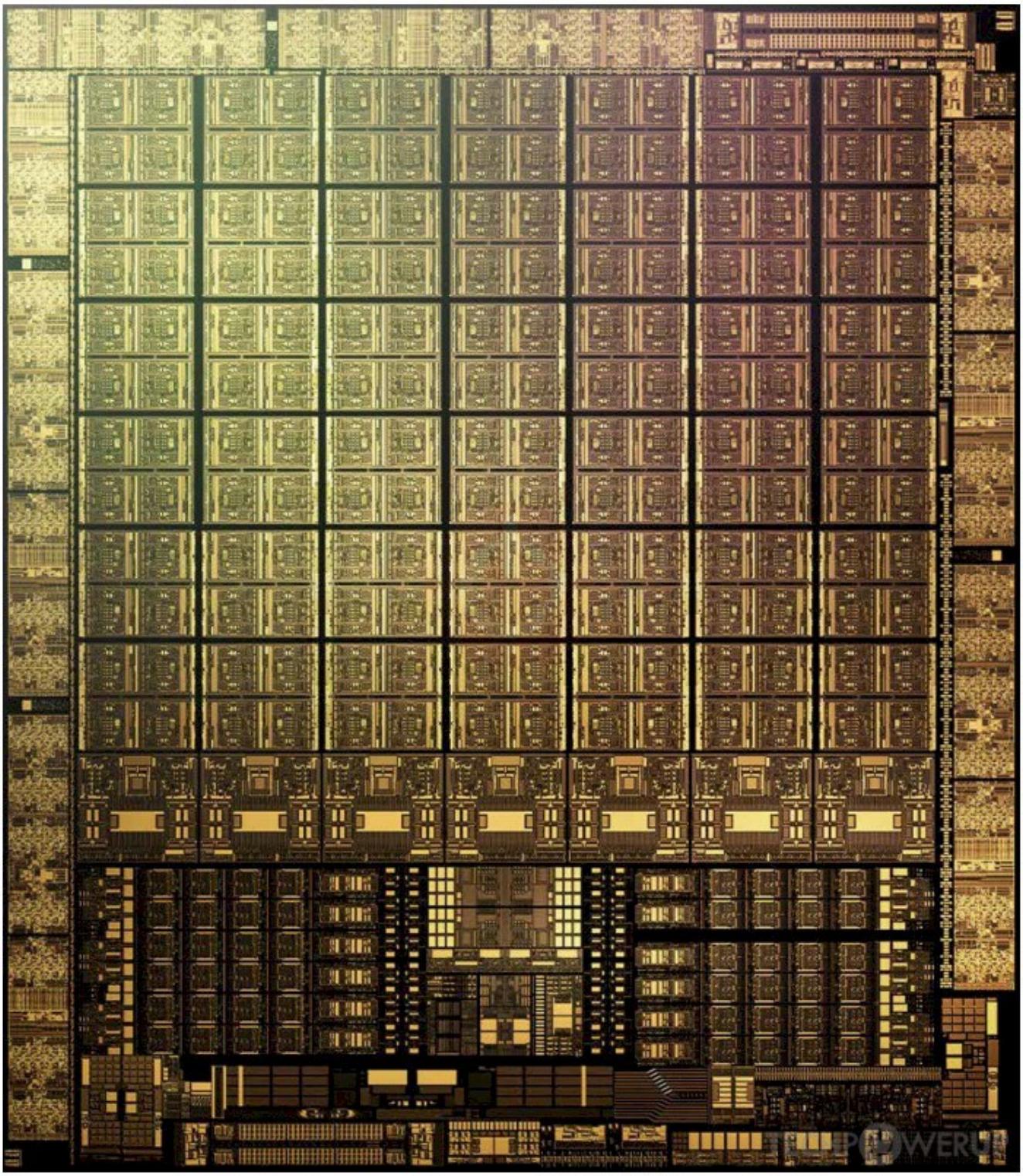
FHW | Tema 8 - Tarjetas Gráficas



Creado por: Jaume Moreno Gallego 23/12/2020

Indice

Tema 1

¿Qué es una tarjeta gráfica?

Normalmente la gente confunde GPU con tarjeta gráfica. Al igual que una CPU no es un ordenador esto se le aplica igual. Una tarjeta gráfica se compone de dos cosas esencialmente "Puede ser no del todo cierto, no encuentro mucha información sobre ello, ya que las CPU ya se encargan de esta tarea. Otra cosa es que la tarjeta gráfica se le puede llamar simplemente gráfica, tiene varios nombres, pero para no liar mejor solo decir gráfica y ya."

- **VPU:** Video Processor Unit o Unidad de Procesamiento de Video.
- **GPU:** Graphics Processor Unit o Unidad de Procesamiento de Gráficos.

La diferencia entre ellas es que la VPU se centra en codificación y decodificación de video y la GPU tiene hardware para rasterización "No creo que sea necesario explicar el proceso de rasterización al igual que lo que continua después" y mapeo de texturas (para gráficos 3D), y cuya arquitectura de memoria está optimizada para manipular imágenes de mapa de bits en la memoria fuera del chip (leer texturas y modificar búferes de fotogramas…). Como luego se verá una GPU siempre va a estar haciendo operaciones matemáticas, en concreto va a hacer matrices.

Y para separar una CPU de una GPU, hay que pensar que si nosotros queremos cortar un árbol, la CPU es la navaja suiza (tiene de todo, pero no es buena en una sola cosa en especifico) y la GPU una motosierra (tiene un hardware especializado en esa tarea)

"En este caso me voy a centrar más en la GPU, ya que la VPU las desconozco mucho y no se como explicarlas, pero es casi lo mismo que va a hacer una CPU."

Especificaciones de una tarjeta gráfica

Normalmente lo primero de lo que se habla de una tarjeta gráfica es de que Arquitectura es (Maxwell, Pascal, Turing, Ampere…) Este concepto ya esta explicado

El motivo por el que se hace es porque la gente confunde que el comparar especificaciones de diferentes arquitecturas, es absurdo. Ya que cada una tiene un funcionamiento diferente. Sin embargo comparar especificaciones dentro de la misma Arquitectura si va a tener sentido.

Que elementos compone una GPU:

Núcleos

- Cuda Cores (**NVIDIA**)

- Stream Processors (**AMD**)

Estos son el nombre de los núcleos y se hablan de cuantos hay. Ya que una tarjeta gráfica no son como las CPU que son 8, 16 núcleos. Aquí se habla de miles de núcleos pequeños trabajando en paralelo.

(Una rama se rompe enseguida, pero muchas de ellas, es muchísimo más complicado de romperlas)

Entonces tener más núcleos significa mayor potencia. "Luego hablo un poco sobre cada núcleo, pero en general siempre es así ponen muchos núcleos para que vaya mejor, dentro del núcleo casi ni lo tocan, solo lo mejoran para el tema de las latencias, luego lo enseño"

Otro valor que se toma en cuenta del núcleo es a que frecuencia funcionan (**core clock=número de ciclos que hace por segundo**) y siempre se va a hablar en Mhz, esto significa que si una gráfica funciona a 1000Mhz hace que su núcleo realiza mil millones de ciclos por segundo, por lo que para medir la potencia de una gráfica sería :

Cuantas operaciones puede hacer en un ciclo y cuantos ciclos realiza por segundo.

Pero eso tiene una unidad que se llama **TFLOPS.**

Los TFLOPS indican el número de operaciones que puede hacer por segundo una gráfica por ejemplo de 11 TFLOPS, significa que hace 11billones de operaciones por segundo.

"Para obtener ese valor y es con la siguiente operación, CUIDADO esto son solamente los teóricos":

N.º nucleos*coreclock*2= xTFLOPS

"Esto no se muestra ya en las especificaciones de las tarjetas gráficas, más que nada porque la gente sabia que nunca van a alcanzar esos TFLOPS y por otro motivo más que he explicado antes. Esto lo puedes hacer como pregunta para clase. El segundo motivo es porque no podemos comparar diferentes arquitecturas, ya que hemos dicho que no se debe de hacer ya que funcionan de manera diferente."

Aquí puedes ver la diferencia de potencia entre una CPU y una GPU haciendo cálculos reales, y para que compares los datos, si uso esa formula no llega a los que se ven en un caso real.

1280(nucleos) x 1708Mhz x 2= 4372,480 GFlops → Esto son 4,372 TFLOPS

Es por este motivo que no me gusta usar esto para un ejercicio y es mejor usar matrices en un caso real.

Dentro de la frecuencia del núcleo se ven dos cosas:

- Frecuencia Base: A que Mhz va a trabajar cuando no tenga carga o cuando llegue a determinadas temperaturas y se vea obligada a llegar a esta frecuencia.

- Frecuencia boost: A que Mhz va a trabajar cuando este bajo carga.

Conceptos para entender la GPU

"Las siguientes especificaciones normalmente no se ven, pero es bueno saberlo para ver como funciona la tarjeta gráfica, es por eso que no voy a entrar en detalles ni explicar palabra que salen"

- TMU's: Es el motor encargado de procesar las texturas por cada pixel pepeline. Tambien conocidas como Graphical Pipeline, es una pequeña etapa de la GPU que realiza las tareas menos pesadas. En concreto esta unidad, rota y vuelve a clasificar un bitmap según su tamaño para aplicarlo sobre un plano arbitrario de un objeto 3D dado.

Te dejo esta página para que pongas un ejercicio de cuantas TMUs, ROPs, etc tiene una tarjeta gráfica, te doy el de una RTX 2070, pero es para que ellos busquen cuantas TMUs tienen y así ven que si sirve saberlo y que esa especificación existe "https://www.techpowerup.com/gpu-specs/geforce-rtx-2070.c3252 también te va a servir para el siguiente tema, ya lo veras."

-ROPS: Cuando los pixel procesados se almacenan en la memoria de video de la tarjeta gráfica y están listos para ser resueltos en pixeles presentados en pantalla, esta tarea es manejada por una unidad de la GPU llamada ROP.

Un GPU moderna pone un número de ROPs en ejecución, basados en cómo el GPU debe optimizar la salida del pixel sin cuellos de botella, para realizar las tareas finales de la representación. Así como simplemente resuelve y dibuja los pixeles en pantalla, el hardware del ROP también realiza un número de optimizaciones para ahorrar el ancho de banda de la memoria al leer y escribir los pixeles en un framebuffer, tal como compresión del color (incluso ahorrar 1 byte de datos del color por pixel es un ahorro enorme en términos de ancho de banda).

"Este proceso es la rasterización (tambien se asocia al AntiAliasing, que no lo explico porque sería extenderme muchisimo y explicarlo todos e investigar las nuevas cosas que hay), es por eso que es bueno ponerlo y juntarlo con lo anterior, es complicado ponerlo con imágenes este proceso, por lo que dejo un video aquí para que veas lo de la Rasterización y otros conceptos que enseña de manera muy visual, lo puedes enseñar una vez se haya explicado los temás de los que se componen la tarjeta gráfica https://www.youtube.com/watch?v=tbsudki8Sro"

RAM

Ahora pasando de tema, en vez de hablar sobre los núcleos, pasamos a la memoria RAM y normalmente se habla de la cantidad que hay 6, 8 10 GB… Pero el valor que más se mira y luego veremos el porque es la frecuencia de la memoria y de nuevo SIEMPRE se hablan en Mhz (Ahora en las especificaciones también lo ponen como 14Gbps (14000mbps) que realmente serían 14000Mhz, pero la manera correcta siempre ha sido con **Mhz**.) y además también se hablar del ancho del bus.

El ancho del bus sería la cantidad de información que puede enviarse en cada ciclo de memoria.

Esto hay que pensarlo como una carretera, donde el bus sería el número de carriles en la carretera y la frecuencia a la velocidad a la que van los coches. Y lo que interesa es saber cuantos coches han pasado por esa carretera a lo largo del día y con esto nos referimos al ancho de banda que se mide en GB/s.

Y el como se calcula es:

Frecuencia de las memorias*bus/8

Un ejemplo (https://www.pccomponentes.com/gigabyte-geforce-rtx-3070-gaming-oc-8gb-gddr6 ha puesto bien las especificaciones, por eso quiero usar esta como referencia)

14000*256/8= 448000 Mhz → 448GB/s

A mayor ancho de banda, es mejor, ya que este valor significa que tan rápido puede acceder el núcleo a la información que tenemos en memoria, este valor siempre hay que verlo cuando vamos a jugar en

altas resoluciones (8K, 4K, 1440p…), pero si el juego consume 225GB/s y nosotros tenemos 448GB/s nos va a dar igual, porque el rendimiento va a ser el mismo.

Además existen diferentes tipos de memoria, GDDR5, GDDR5X, GDDR6, GDDR6X, HBM…

"No entro en detalles, por ver que ya tienes un tema de memorias RAM. Y la verdad es que me atraganto un poco con este tema ya que no lo tengo muy investigado"

Consumo

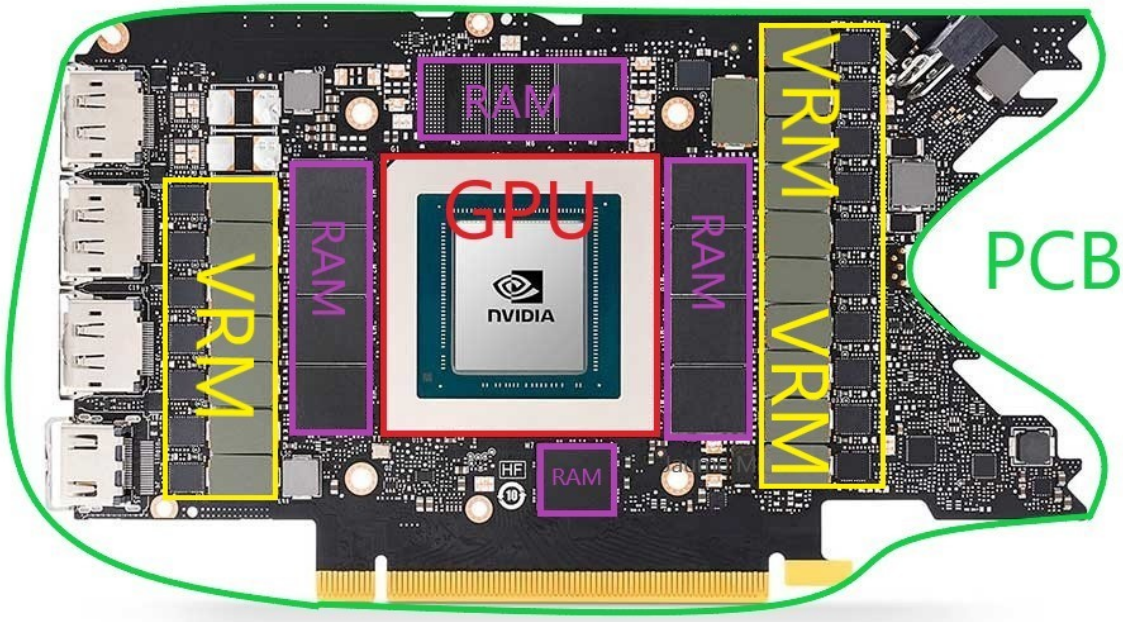
Por ultimo esta el **TDP** que es el calor máximo que va generar el núcleo y es el calor que luego tendremos que disipar y se mide en Watts y no hay que confundirlo con la que consume, aun que normalmente se corresponde con el consumo. Ya que la electricidad que llega a la gráfica por el traspaso de energías, después es lo que vamos a tener que disipar. Lo que pasa es que siempre te dicen 150W de TDP y luego si lo mides son 170W…

VRM "https://hardzone.es/reportajes/que-es/fases-gpu/ te pongo otro enlace si te interesa un poco la electrónica, ya que es importante a mi parecer a nivel avanzado si quieres ver la calidad que tienen."

Ahora usan más el termino TGP y TBP.

-**TGP**: Coge el consumo de la GPU y la PCB, esto no incluye ni la iluminación ni la refrigeración que incorporen.

- **TBP:** Coge todo el consumo, PCB, GPU, refrigeración, el pico de máximo consumo…



"Esta es la PCB de las RTX, sinceramente un muy buen trabajo de ingeniera, puedes si quieres hablar de sus sistema de refrigeración, ya que es muy interesante como lo pensaron. SPOILER es peor que la de las ensambladoras, aun que eso siempre ha sido así, pero al menos han innovado."

Como dato curioso, se puede ver cuanto consumo podría alcanzar una tarjeta gráfica, aun que casi nunca van a llegar a consumir esta cantidad, por razones obvias de temperatura y estabilidad "No he me explayo mucho, porque me dijiste que no querías que hablara de la electrónica".

Para saber ese consumo hay que fijar primero en la cantidad de conectores que tiene y su numero.

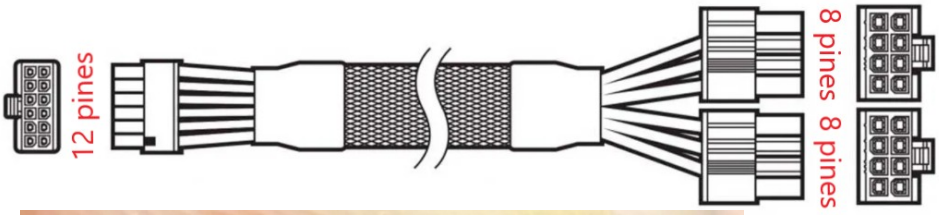
1 conector de 6 pines consume: 75W

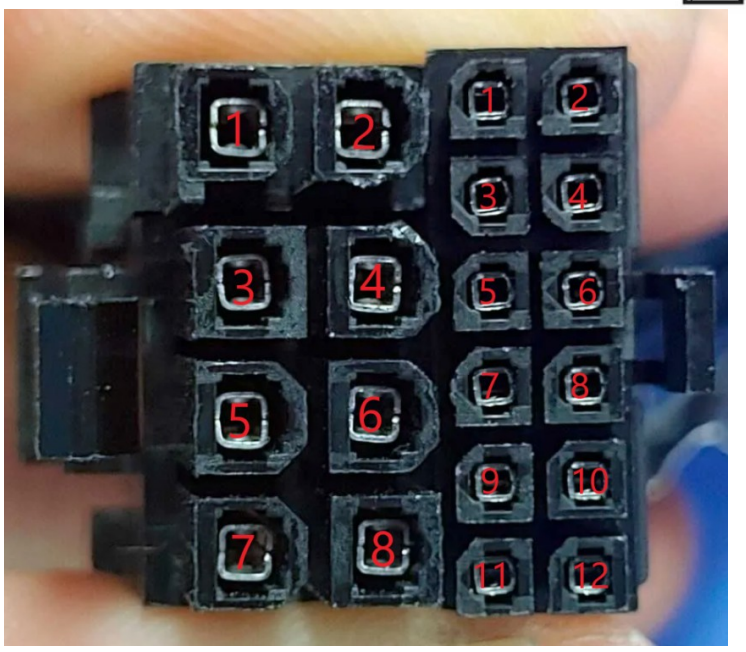
1 conector de 8 pines consume: 150W

Puerto PCIe consume: 75W

¿Entonces si una gráfica tiene dos conectores de 8 pines cual seria su consumo? "Quiero ver si pican y dicen 300W, ya que realmente seria 375W, ya que sin la conexión PCIe directamente ni se conecta a la placa base"

Si te preguntan cuantos pines puede tener un conector de la gráfica y con eso me refiero a un solo conector, no es solo 6 o 8 pines. NVIDIA hizo un gran trabajo y creo uno de 12 pines, que es más pequeño y da el mismo suministro que dos de 8 pines. El Fail por decirlo así, es que te dan un cable minúsculo y que estéticamente queda muy mal y es por eso que solo se ve en el modelo de referencia de la marca.





Tema 2

Arquitectura de NVIDIA

En este tema voy a explicar como funciona la arquitectura de NVIDIA, intentándolo explicar como los CCD y CCX de AMD, ya que tiene alguna similitud.

NVIDIA a lo largo de los años ha ido sacando diferentes arquitecturas para mejorar el rendimiento de las tarjetas gráficas. En este caso solo voy a hablar de dos arquitecturas que tiene NVIDIA y luego solo me centrare en una, así se podrá ver las diferencias que tienen de una generación a otra.

De las arquitecturas que voy a hablar son Turing(Serie 20 y 16) y la continuación de Turing que es Ampere (serie 30).

Es importante saber que cuando se habla de la serie 30 se refieren a la serie 3000. Esto se dice así porque los ingleses dicen mucho “30 80” en vez de “3080”, de igual forma es correcto decir 3000 o 30, ya que al final es lo mismo, esto solo se le aplica hasta la Pascal, que es la serie 1000 o 10.

“También puedes ver que a la serie le llama generación, dentro de la comunidad hay gente que lo dice de diferentes maneras. Yo en mi caso muchas veces me lio y en vez de decir serie, digo generación, de igual modo no hay problema porque uno son números y otros son letras. Pero para llevar un mejor orden y así todos seguimos un guion a la arquitectura le llamamos generación y a los números serie. Porque dentro de una generación hay diferentes serie, mejor ejemplo es este: Generación Turing serie 20 o 16. Si lo ves mejor de otra forma comentamelo y así todos lo tenemos más claro.”

CHIP Ampere

Al igual que una CPU le llamamos procesador, a la GPU le llaman chip (gráficos). “Ahora entro en más detalles de porque siempre sale como chip porque es un poco difícil de explicarlo” Esto es debido a que NVIDIA, AMD, Intel, etc. Hacen un proceso en el que crean diferentes lineas de procesador, entonces nos podemos encontrar un chip de gama muy alta en una tarjeta de gama alta. Y el porque lo hacen es sencillo, el proceso de litografiá, es decir las obleas que se crean para luego tener una GPU, CPU, etc es un proceso que tarda mucho normalmente sobre unos 5-6 meses (Lo podemos ver ahora con todos los problemas de stock que hay, aun que la pandemia también tiene algo que ver), entonces las compañías no son tontas, si un chip sale defectuoso no lo desechan, sino que lo asocian a otra cosa. Lo pongo un poco con un ejemplo.

Si nosotros tenemos 3 procesador diferentes, gama baja, media y alta. El de gama alta tiene 16 núcleos el de gama media tiene 8 y el de baja tiene 4.

Cuando uno de gama baja sale defectuoso y solo funcionan 10 núcleos no van a tirar ese procesador, sino que desactivan 2 núcleos más para que se convierta en uno de 8 y entonces ya tenemos un procesador de “gama alta” en uno de gama media.

Esta es la siguiente tabla de mejor a peor gráficas en Ampere y también las de Turing, ya que Ampere no ha sacado toda su gama de tarjetas gráficas.

El como se identifica en NVIDIA gama alta, media, baja, etc. Es de más a menos números.

3090 es la tope de gama, pero si existe 3090 TI, el TI significa **Titanium** que es como decir la tope tope de gama, tambien existen las Titan pero eso ya va enfocado a otro sector en la anterior generación (Turing) se llaman RTX Titan.

Diferentes gamas de mayor a menor (Solo las **confirmadas**):

Ampere (Serie 30)	**Turing (Serie 20)**
3090	RTX Titan
3080	RTX 2080 TI
3070	RTX 2080 Super
3060 TI	RTX 2080
N/A	RTX 2070 TI
N/A	RTX 2070 Super
N/A	RTX 2070
N/A	2060 Super
N/A	2060

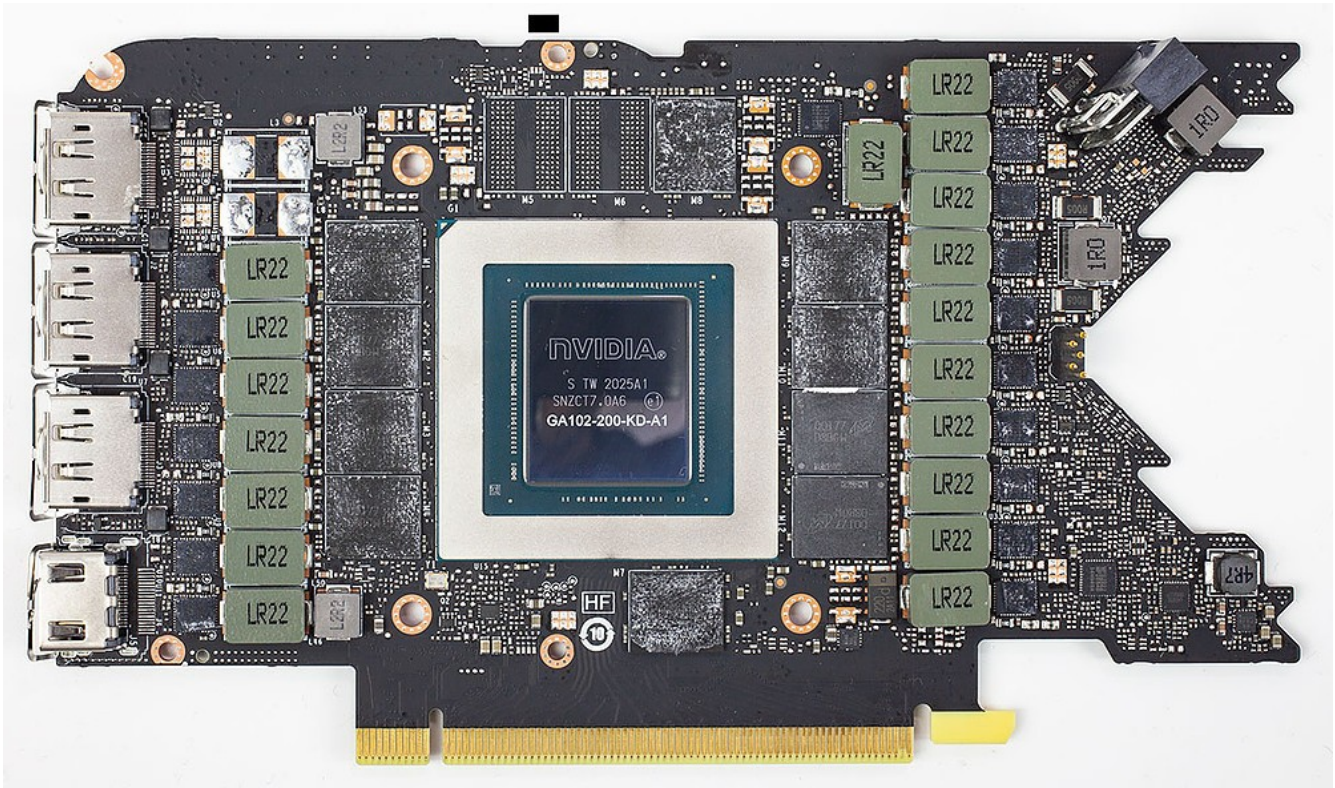
NVIDIA también sacan generaciones con la nomenclatura Max-Q, esto simplemente son gráficas recortadas para que en un portátil no se queme. De igual manera Turing es una versión muy eficiente y por eso nos encontramos por ejemplo en un portátil: 2070Max-Q y 2070. Es importante saber que si nos tenemos que decidir por la potencia y rendimiento que tienen, siempre va a ser mejor la que **NO** tenga el **Max-Q.**

En el caso de NVIDIA y hablando solo de **Ampere** es así (algunas tarjetas gráficas aun no están a la venta y son filtraciones) "Dentro de un mismo chip NVIDIA le pone una variante para el tema de logistica saber cual va con cual":

Serie Geforce 3000	**CHIP GA**
3090 **Confirmado**	GA102-300-A1
3080TI **Filtración**	GA102-250-KD-A1
3080 **Confirmado**	GA102-200-KD-A1
3070 TI **Filtración**	GA103-XXX-XX
3070 **Confirmado**	GA104-300-A1
3060 TI **Confirmado**	GA104-200-A1
3050 **Filtración**	GA107-300-A1

"Si quieres como ejercicio puedes poner que busquen los chips la anterior generación, porque de la nueva hay pocos modelos, en turing el chip se llama TU100 (TU102, TU104...)"

Hay otra manera de ver cual es su chip y es desmontando la gráfica (PCB= "placa base") y ver el chip.



"Puedes poner como ejercicio que por esta foto te digan que chip es. En este caso es el GA102, la variación 200"

Explorando la arquitectura

Así es como se ven una chip de NVIDIA y hay algunos elementos que ya hemos visto como el PCIe 4.0, la cache, etc. Pero en lo que me voy a centrar es en los SM.

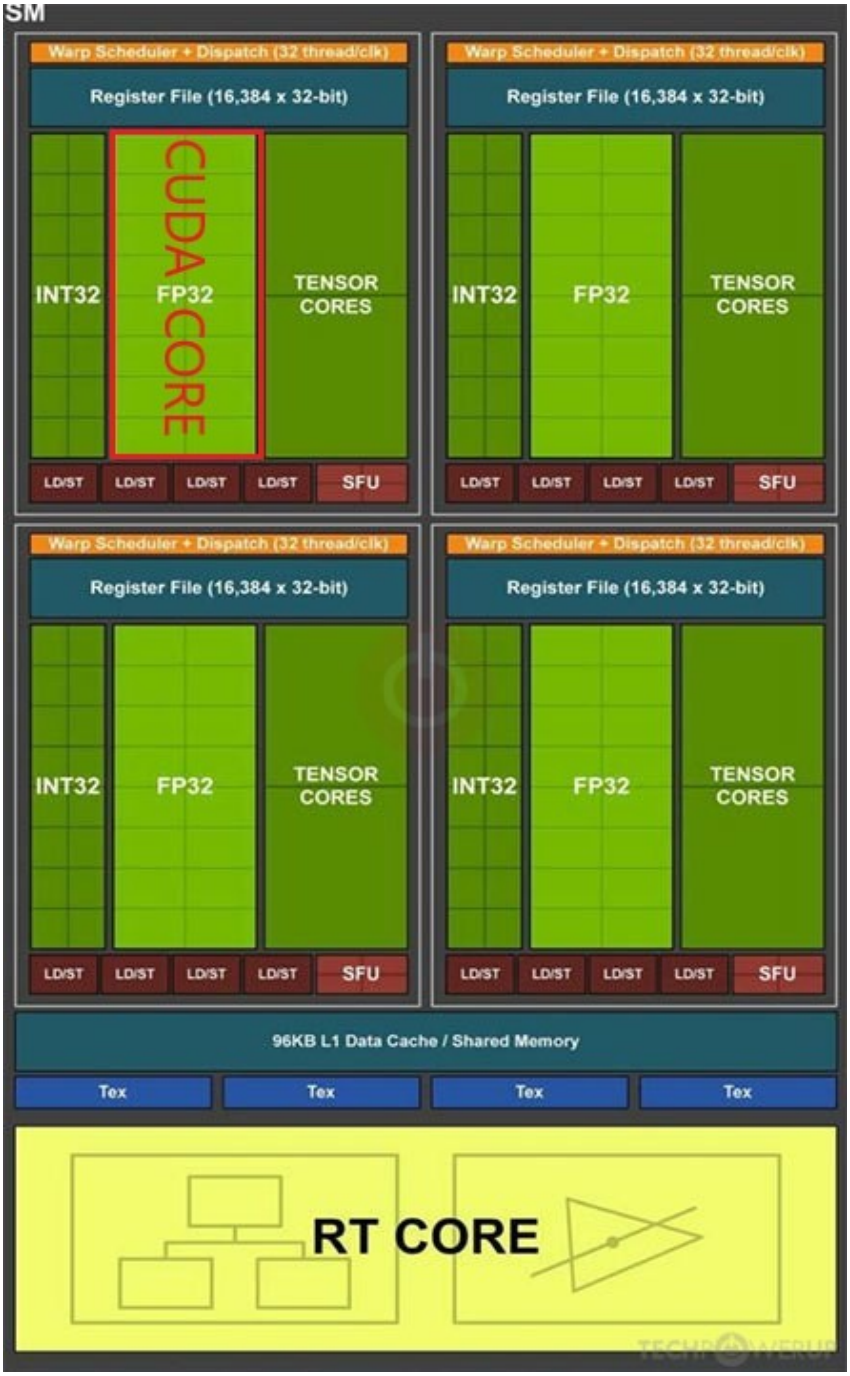
"Aquí explico un poco el como funciona la gráfica como en los AMD los CCD y CCX"

SM (Streaming multiprocesos)

“Voy a empezar poniendo una imagen de Turing para que vean la diferencia interna que hay entre dos generaciones”

Los SM son como los grupos de trabajo que es algo parecido a los núcleos de un procesador.

TURING →



Entonces en un SM nos encontramos unidades INT32 (Es la unidad que hace operaciones con números enteros), FP32 (Es la unidad que hace operaciones con decimales), tensor cores…

Para NVIDIA una unidad FP32 es un CUDA CORE es decir un núcleo que se encarga de hacer cálculos matemáticos.

Por lo tanto 1 SM son 4 Data Pad "así es como lo llama NVIDIA"

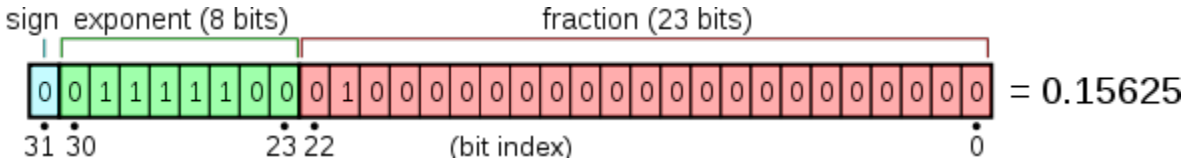
Dentro de cada Data PAD hay 16 CUDA CORES. “No he querido editar esa imagen porque sería poner 32 números, ya que en INT32 también se cuentan, lo que pasa es que son como por separado. Entonces son 16 int32 y 16 FP32. Se ponen por separado por las operaciones que hacen cada uno.”

NVIDIA es muy lista y para poder incrementar el máximo las tarjetas gráficas, se consigue haciendo que cada SM haga el mayor numero de instrucciones por ciclo posibles, es decir, que puede hacer el máximo cosas al mismo tiempo (Cuantos más cálculos puedas hacer en paralelo más potente es la GPU).

Hablando un poco más sobre los INT y FP. Vienen porque los motores gráficos utilizan dos tipos de datos:

- **Integer (Int):** 1 2 3 4 5 6 7 8 9… (Números Naturales)

- **Float (FP):** 1,2345678 2,87634587 3,786345876… (Números en coma flotante)

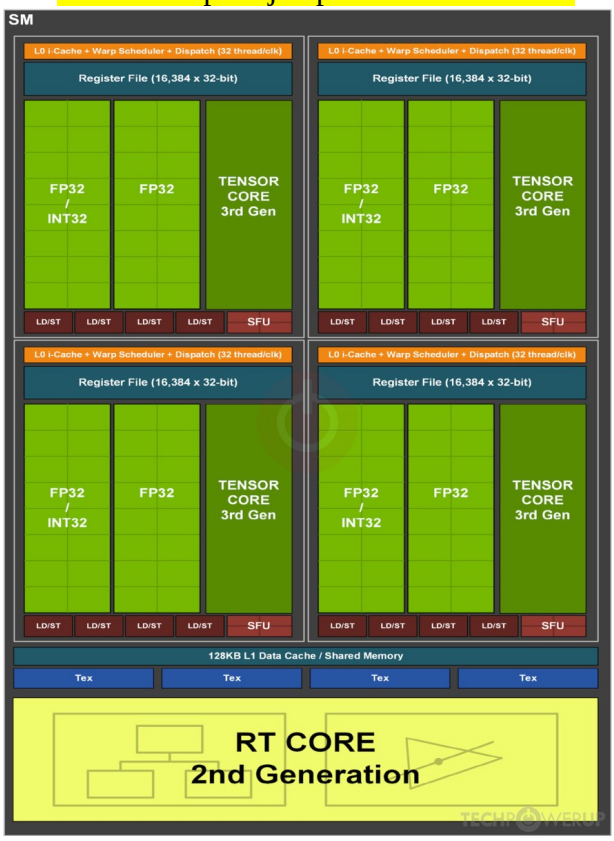


Esto viene así porque en las antiguas generaciones los núcleos CUDA alternaban entre INT y FP. Pero NVIDIA los dividió haciendo que hayan 16 núcleos Int32 (el 32 es porque son 32 bits) y 16 núcleos FP32 y lo mejor de todo es que trabajan de forma paralela.

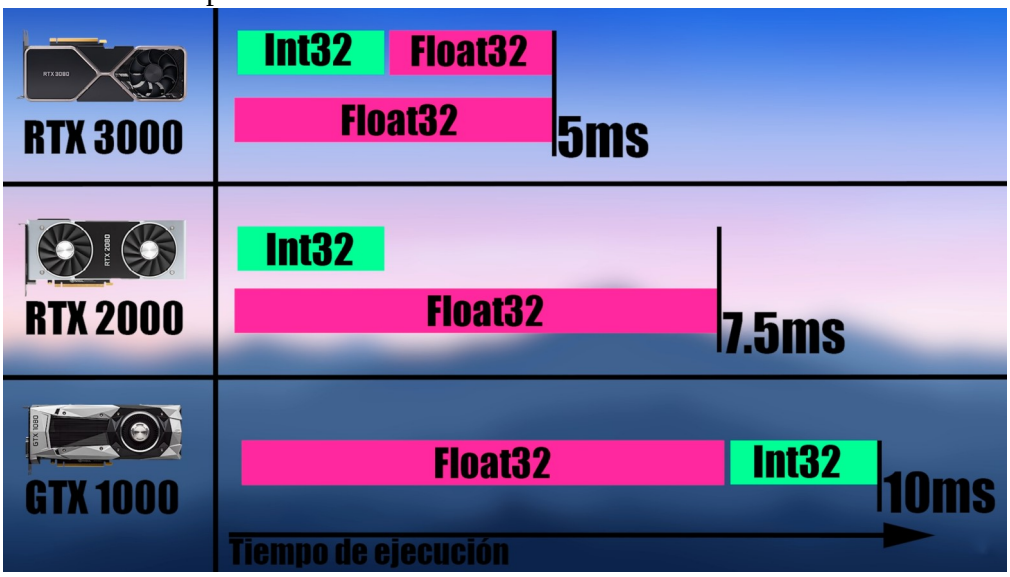
Ahora con Ampere NVIDIA a extendido este concepto.

“Este pertenece a una de uso no comercial GA100.”

“Este sería por ejemplo el de una 3090.”



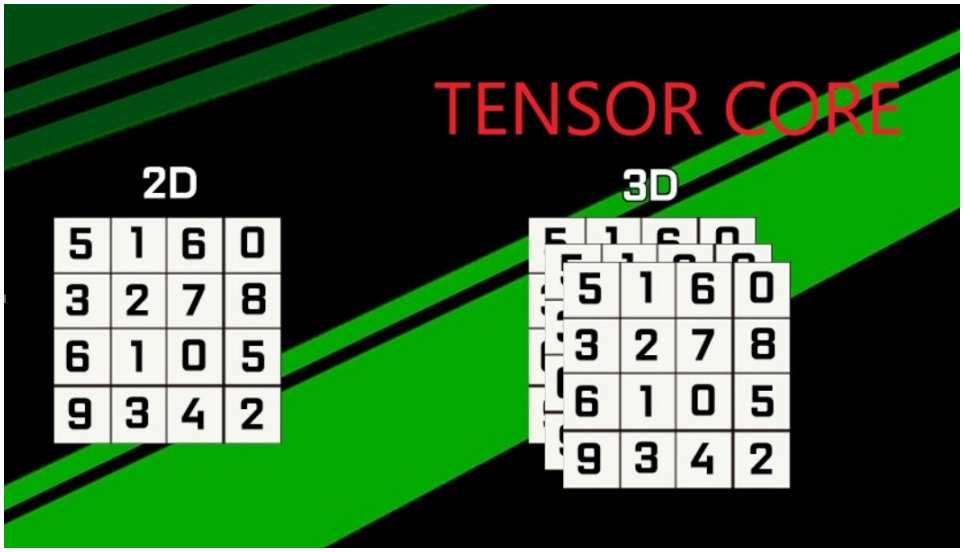
Ahora no nos encontramos en cada Data PAD que los INT32 pueden hacer tanto operaciones INT32 como FP32, si te preguntas porque hacen esto. El motivo es muy simple y tiene todo el sentido del mundo, en los juegos se usa alrededor del 20%-30% de los cálculos en INT32, entonces una vez se terminen los cálculos de INT32, esos núcleos se quedan ociosos, por lo que la solución más fácil es que también puedan hacer otro tipo de cálculos.

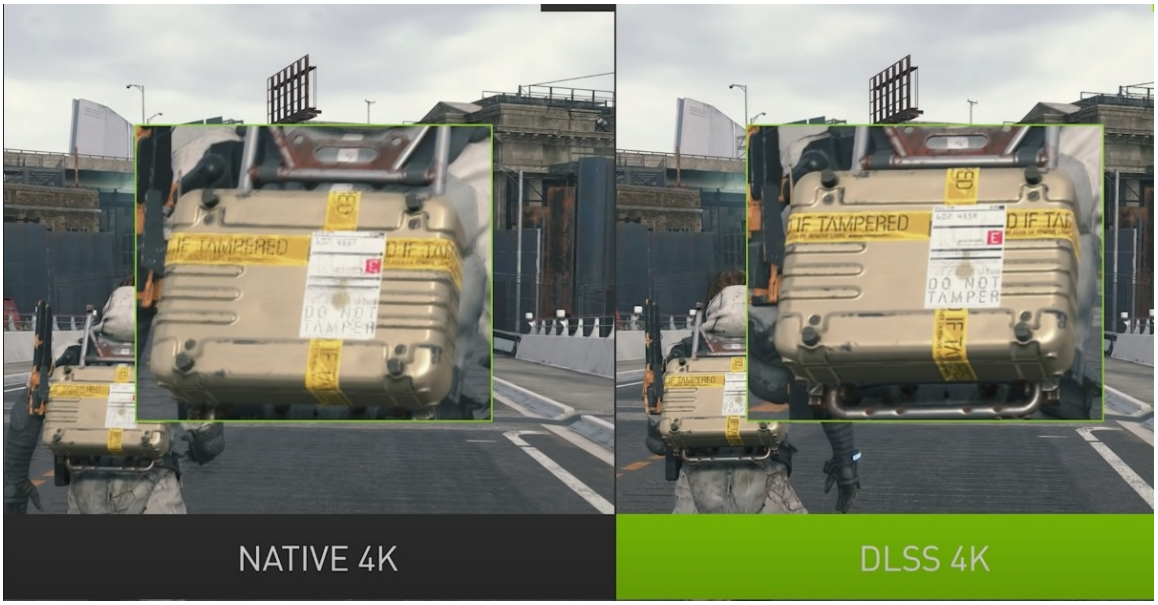


Tensor Cores

Si recordamos bien lo que es un tarjeta gráfica, siempre va a estar haciendo matrices, pero no hemos dicho que tipos de matrices. En general una tarjeta gráfica siempre va a hacer matrices en 2d es decir (x – y) pero y si queremos hacer que una IA calcule por ejemplo los pixeles que podrían tener en una escena en una resolución menor. "Me refiero a que si jugamos a 1080p lo reescale a 1440p. Esta es una tecnología que implementa NVIDIA que es el DLSS, hay muchas más como NVIDIA RTX Voice… No entro en detalles porque todo eso simplemente es decir que calculo quieres hacer en el tensor core, se podría decir que es una IPU"

Aquí es donde entra los Tensor Cores, el nombre Tensor viene porque hace cálculos de matrices en 3d (x - y – z) a una velocidad muy rápida, ya que son núcleos especializados en este tipo de tarea.





"Video muy interesante de porque son buenas las IA en las GPU https://www.youtube.com/watch?v=C_wSHKG8_fg"

LD/ST y SFU

Las LD/ST son muy sencilla sirven para cargar o almacenar datos.

Las SFU sirven para otro tipo de cálculos.

“No lo explico mucho, porque estas dos cosas es todo programación muy avanzada para mi. Esto es lo que usan los programadores para decir más o menos lo que tiene que ir haciendo. Te dejo aquí el manual de desarrollador por si quieres verlo

https://docs.nvidia.com/cuda/cuda-c-programming-guide/

https://docs.nvidia.com/cuda/cuda-math-api/group__CUDA__MATH__INTRINSIC__SINGLE.html (Esto es SFU)”

RT CORES

Los RT CORES, RT que viene de ray tracing en español trazado e rayos, es una nueva técnica que se usa que emplea diferentes cálculos para intentar imitar la luz real. Lo que hace ray tracing es solo calcular unos cuantos rayos de luz, ya que si cogiera todos sería imposible, ya que son infinitos.

Entonces lo que nos tiene que quedar claro es que esto núcleos son de nuevo hardware especializado para la tarea de calculo de trazado de rayos. No sirven para nada más.

“No quiero comentar más tecnologías que hay, porque son muchas y ya seria basándome mucho en los videojuegos y sería cambiar de tema, yo si quieres puedo comentar cosas en clase sobre esas tecnologías y que es lo que hacen. Es por eso que he decidido no incluirlas en este trabajo, como por ejemplo lo que se esta usando en consolas, que es que la tarjeta gráfica lea directamente desde el disco duro y un gran etc. Espero que te haya gustado y que se comprenda todo.”

NVIDIA a lo largo de los años

Gamas generales

GT: Se refieres a los modelos de entrada, son gráficas de baja potencia y bajo consumo y siempre van a estar por debajo de las GTX

GTX: Se refieren a gráficas con un rendimiento bueno y que se desarrollan mucho mejor en los videojuegos respecto a las GT

RTX: Es lo mismo que GTX, solo que implementan la tecnología de trazado de rayos (ray tracing)

Serie 700 , 900 y 1000 de NVIDIA

"Aquí te voy a explicar la comparativa de las generaciones de una a otra de manera breve, sin entrar mucho en detalles"

Serie 700 vs 900

Cual es la principal diferencia que se ve entre la serie 700 (Kepler) versus la serie 900 (maxwell).

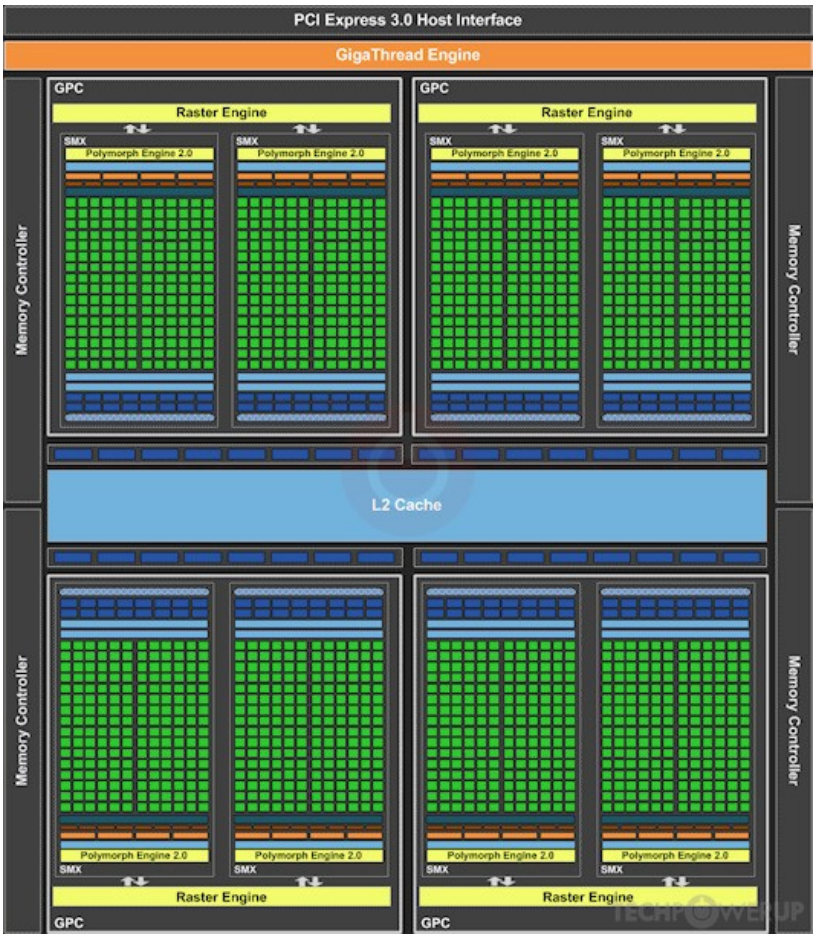
Las latencias, el consumo y las frecuencias de la arquitectura maxwell es lo que hacen que sea un mejor partido a la hora de hacer cálculos y es por eso que son mejores que la serie 700. ¿Pero es verdad que son mejores?

Si comparamos la 760 vs la 960, podemos ver que no es así, hay en juegos donde se mueve mejor la 960 que la 760. Esto pasa porque los juegos más modernos siempre va a salir ganando la 960. Pero en potencia bruta gana la 760 y esto es porque tiene mucho más núcleos que la 960. Ya que en las gráficas siempre es lo mismo. Más núcleos, más rendimiento.

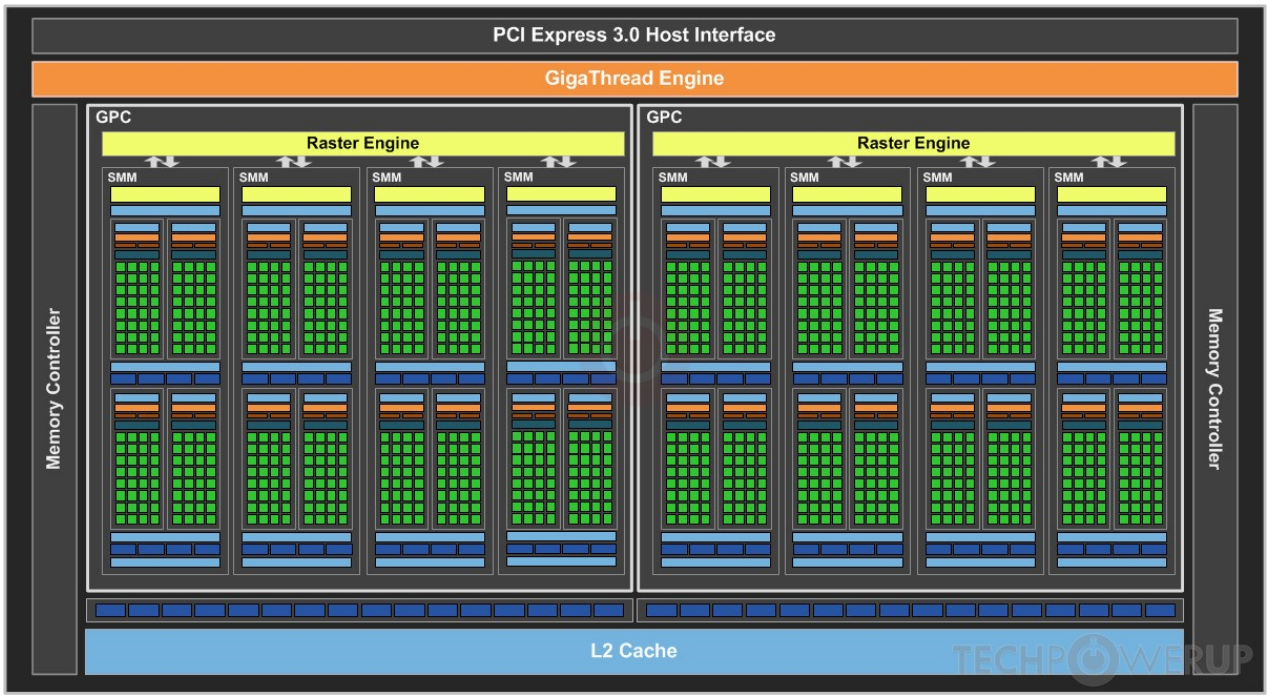
Como se aprecia en las siguientes imágenes, maxwell lo único que ha hecho es empezar a lo que después veremos en pascal, que es agrupar diferentes núcleos en un data pad. En vez de solo tener un sm y un data pad. En maxwell se ve que son por cada SM 2 data pads.

Mi opinión sobre la 960 es que vale la pena gastar 30€ por el simple hecho, de pensar a futuro. Si nosotros no tenemos el ultimo driver para jugar un juego, significa que esa gráfica ya no nos sirve. Por desgracia esto siempre pasa y por eso los usuarios siempre vamos pensando a futuro, en el caso tuyo es mejor gastar 30€ porque sabes que durara más el soporte que el de la 760. Pero en rendimiento son "iguales". Maxwell era el primer paso a pascal. Que es donde ya dieron el salto.

700→


PCI Express 3.0 Host Interface
GigaThread Engine
GPC
Raster Engine
SMX
Polymorph Engine 2.0
Memory Controller
L2 Cache

900→

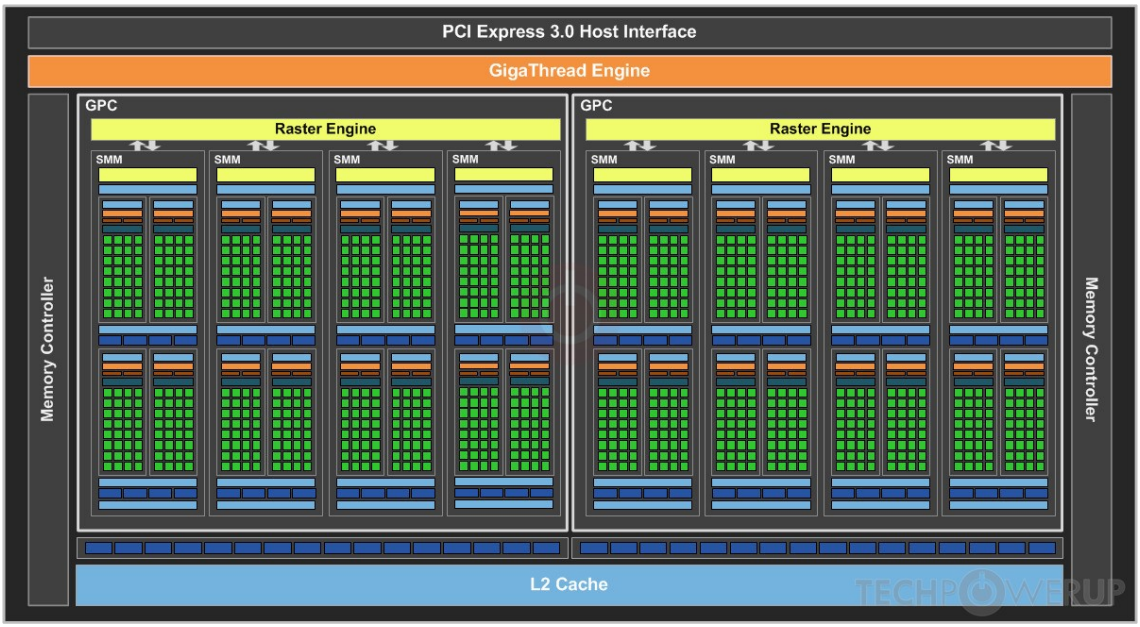

PCI Express 3.0 Host Interface
GigaThread Engine
GPC
Raster Engine
SMM
Memory Controller
L2 Cache

Serie 900 vs 1000

La serie 1000 fue un gran salto de generación y esto se debe a que pasaron a un proceso de fabricación más pequeño. Tanto la serie 700, como la 900 estaban fabricadas a 28nm. Pascal esta fabrica en 16nm.

Y el porque son mejores estas tarjetas gráfica lo dice el mismo tamaño. Donde antes cambia un núcleo ahora caben 2 núcleos. Y entonces pasamos de 1024 núcleos de la 960 a 1280 núcleos de la 1060 6GB (digo 6GB porque tienen dos modelos, la de 3GB y la de 6GB.)

900→



1000→ “Esta serie usaban casi todos el mismo chip, es por eso que es tan grande. (1060 6GB - 1080)”

Pascal vs Turing

¿Porque Turing es mejor que Pascal?

De nuevo esto nos lleva a los núcleos y las latencias internas que tienen, pero también a nuevos tipos de núcleos, es cierto que tiene muchos más núcleos "Tiene truco, porque la RAM que llevan es GDDR6, lo que permite un mayor ancho de banda, lo que hace que los juegos modernos puedan aprovecharlo y hacer que vaya mucho mejor, es por eso que siempre hay que pensar a futuro."

NVIDIA entiende como núcleo CUDA el que hace operaciones fp32 y las int32 las deja de lado, ya que en los videojuegos son solo un 20-30%. "Yo creo que aquí es un poco cagada de parte de NVIDIA, ya que no encuentro ningún nombre para int32 , por lo que no se si ellos ahora dicen que int32 y fp32 es un cuda core. Porque si nosotros pensamos solo como shading units, entonces si entran int32 y fp32, porque es el conjunto de esas dos. Si te lia con el comentario de arriba. **Deja como CUDA CORE int32 y FP32 y así no liamos.** Así de paso ven la diferencia entre Pascal y turing."

Por lo que el principal problema que vemos es que el núcleo de pascal hace operaciones tanto int32 como fp32, pero va alternando entre uno y otro. Por lo que se desperdicia tiempo en calcular operaciones (Esto se ve antes del tema).

NVIDIA lo soluciono creando un nuevo tipo de SM que divide int32 de FP32 y además mete hardware especializado para la IA y para el ray tracing, salvo los tensor cores, el ray tracing es solo un extra innecesario. Ya que la IA si se ve que el rendimiento es mucho mejor que sin ella. Pero el del ray tracing es penoso en la primera generación, en muchos casos es injugable, esto también es debido a la poca RAM de la que disponen, pero porque no pensaron bien lo que hacían al principio.

"He puesto en Turing el chip tope de gama, para que no se piensen que tiene menos núcleos y no hayan diferencias, por si quieres coger las mismas imágenes y demás. Tambien se puede ver el NVLink, no lo he explicado porque hoy en día no se aprovecha, casi ningún juego lo implementa y los que lo implementan va fatal, es malgasto de dinero para un usuario común. En el apartado profesional no tengo ni idea de si tiene uso. Pero lo único que hacen es que trabajen simultáneamente entre ellas y repartan la carga entre las dos gráficas."

En resumen, más núcleos y mejores núcleos (nuevos tipos de núcleos que aprovechan tecnologías que optimizan los videojuegos). NVIDIA siempre ha sido la mejor porque siempre incluye mejores tecnologías que AMD.

Pascal→

Turing→

Conclusiones

Dejo aquí una conclusión de porque las gráficas no son más potentes porque mejoren el núcleo internamente. Si miras los núcleos de diferentes generaciones, pongo como ejemplo mi 1060 de 6GB y una de nueva generación, vas a ver como más núcleos = mejor rendimiento. Es cierto que hay ahora hardware especializado, etc. Pero yo hablo de solamente los Cuda Cores, el rendimiento al final siempre va siempre por más núcleos mejor, a pesar de que ampere haya mejorado el núcleo y hacer que en INT32 pueda usarse FP32, no afecta en nada, eso es solamente el tiempo de ejecución, no a la potencia bruta que tiene.

Mi 1060 tiene 1280 núcleos, y la 3060 Ti tiene 4864 núcleos.

Un mejor caso es la 3070 vs las 2080 TI. La 3070 en potencia es mucho mejor que la 2080 TI, en benchmarks se ve otra cosa, pero por un motivo muy importante y que muchas veces nadie lo dice, la RAM que tienen. La 2080 TI tiene 11GB de RAM y la 3070 solo 8GB. Esa RAM extra hace que no se limite la gráfica, ya que no puede conseguir lo suficiente, es por eso que NVIDIA va a sacar nuevas versiones de estas, que sería la 3070TI con 16GB de RAM y la 3080TI con 20GB de RAM (son expeculaciones, me refiero a que luego pueden ser 1-2GB de RAM menos).

Una 3070 tiene 5888 núcleos y una 2080 TI tiene 4352 núcleos.

Ahora mismo por los problemas de fabricación por la pandemia, los precios han subido, pero esto se le aplica a todos los productos de la informática. Porque cuando habían cerrado las fabricas no pasaba nada, ya que ya tenían stock almacenado, pero es ahora cuando ven que no pueden y si pueden sería a un precio exageradamente alto. En el caso de NVIDIA le pillo justo cuando empezaba a fabricar.

Licencia