

Arquitecturas de GPU

Luis Alejandro Pérez Sarmiento

Departamento de computación
Centro de Investigación y de Estudios Avanzados del IPN

27 de abril de 2022



Cinvestav

1 ¿Qué son las tarjetas de vídeo?

2 Arquitectura de GPU Nvidia

3 GPU vs CPU

Tarjetas de vídeo

Comúnmente llamado tarjetas gráficas, son el componente dedicado al procesamiento de datos relacionados con vídeo eh imagen. Las tarjetas de vídeo obtienen la información del procesador y la transforman en imágenes

Existen 3 tipos

- Integradas en la placa base del PC (Poco común)
- Integradas en el CPU del PC (común)
- Conectadas de forma externa al PC (común)

Integradas en la placa base del PC

Este tipo de tarjetas se consideran descontinuadas, actualmente ya no se producen tarjetas de vídeo que se incorporen en las placas base desde hace 10 años.



Figura: Placas de muestra

Integradas en el CPU del PC

Las tarjetas de vídeo integradas en el CPU son actualmente la opción más común en todo equipo de sobre mesa, estas están ubicadas en una parte del procesador por lo cual tienen de las mayores velocidades de comunicación no obstante tienen limitaciones de potencia y memoria por el tamaño que cuenta el procesador.

Existen dos modelos predominantes en el mercado

- Intel HD graphics
- AMD Vega

Integradas en el CPU del PC

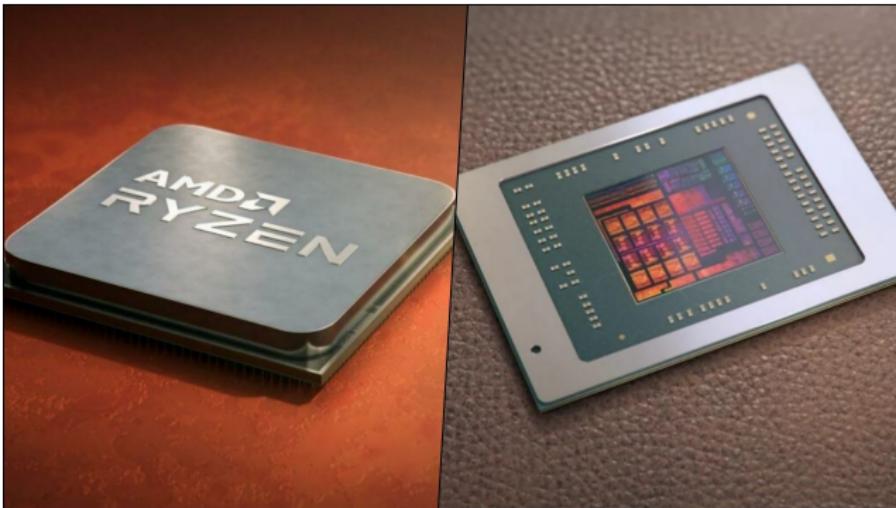


Figura: Estructura de un procesador AMD con una tarjeta de vídeo integrada

Integradas en el CPU del PC

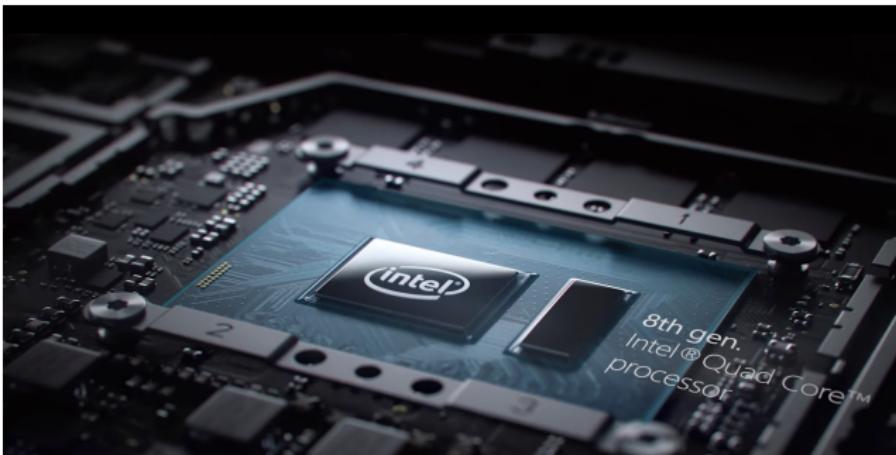


Figura: Estructura de un procesador Intel con una tarjeta de vídeo integrada

Conectadas de forma externa al PC

Conocidas como Unidades Gráficas de Procesamiento GPU, son las tarjetas que van conectadas a la placa madre actualmente por un puerto PCI, existen distintos fabricantes de estas no obstante el mercado esta acaparado por Nvidia y AMD, con distintos modelos. Las tarjetas de vídeo de esta índole cuentan con las siguientes características.

- Una GPU
- RAM
- RAMDAC
- Disipador de calor
- Salidas de video (HDMI, VGA, Displayport, etc.)
- Conector con la placa base

Conectadas de forma externa al PC

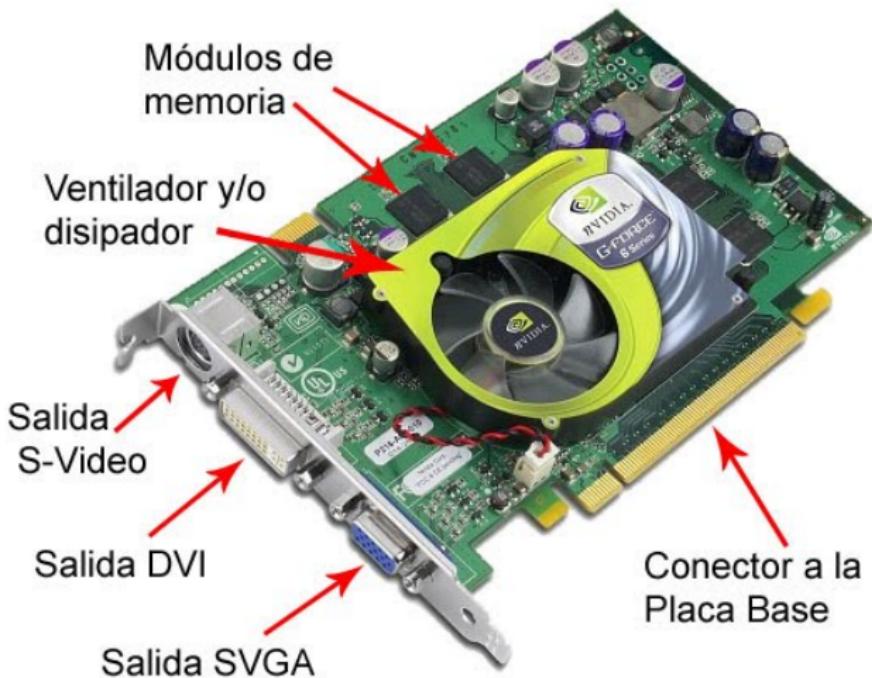


Figura: Componentes de una tarjeta de video

Arquitectura de la GPU

Con el paso de los años Nvidia ha presentado múltiples arquitecturas pero no fue hasta la llegada de la arquitectura Tesla con la cual se inicio a seguir un patrón de diseño. En la siguiente tabla se muestran las arquitecturas más recientes de Nvidia.

Arquitectura	Año de lanzamiento	Litografía	Die (Tamaño del procesador)	Rango de Núcleos
Tesla	2006	90 nm	G80	
Fermi	2010	40 nm	GF100	448-1792
Kepler	2012	28 nm	GK104	2496-3072
Maxwell	2014	28 nm	GM204	1024-4096
Pascal	2016	16 nm	GP102	2048-3584
Turing	2018	12 nm	TU102	2560
Ampere	2020	7 nm	GA102	3584-10752

Tabla: Arquitecturas de Nvidia en el paso de los años

Antes de la era Tesla

Al comienzo las GPU de Nvidia estaban correlacionados a los estados lógicos de la API de renderizado, esto generaba grandes problemas de cuello de botella si no se equilibraba cada parte de forma correcta.

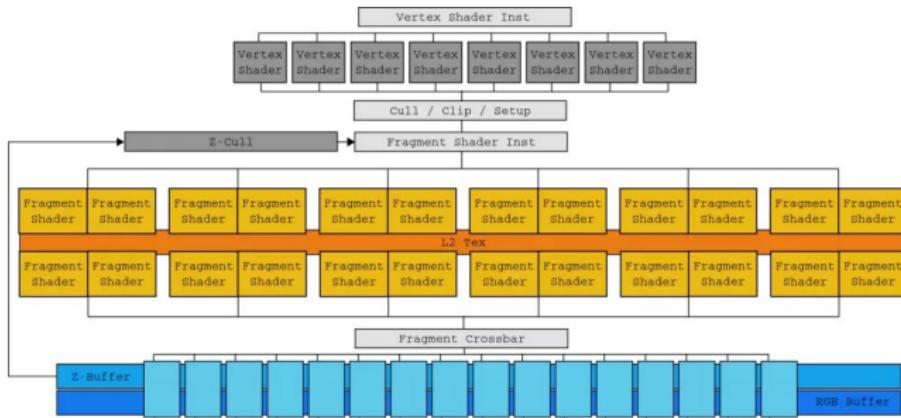


Figura: Arquitectura del die G71

Arquitectura Tesla

La arquitectura tesla fue la primera en ser una arquitectura unificada en ella ya no existía la distinción entre capas todo era un solo kernel, estaba hecha a 90nm.

Contenía

- Stream Multiprocessor (SM).
- Una unidad de instrucciones de subprocessos múltiples (MT).
- Dos unidades de funciones especiales (SFU) con 4 núcleos CUDA cada una.
- El uso de warps, paquetes de 32 hilos para los SM.
- Un archivo de registro (RF) de 4Kb.
- Una memoria compartida de 16kib.
- Una memoria de instrucciones cache.

Arquitectura Tesla

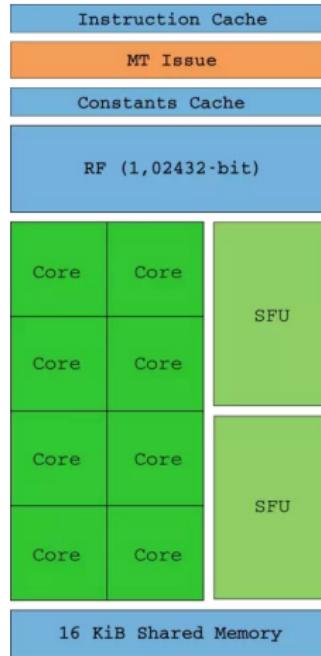


Figura: Arquitectura Tesla

Arquitectura Tesla

- SM Se encargan de unificar toda la arquitectura.
- MT Se encarga de habilitar y deshabilitar los hilos de cada warp dado los punteros.
- SFU es la unidad encargada de ayudar con los cálculos mayores a 32 bits.
- RF es donde se almacena los estados de los hilos y los subprocesos.
- Memoria compartida, donde se almacena las variables compartidas.
- Memoria de instrucciones, se almacena las instrucciones a realizar.

Arquitectura Fermi

Fermi siguió los mismos paso que su antecesor, pero al estar hecho a 40nm, esta arquitectura logró cuadruplicar toda la potencia.

Características

- Los Stream Multiprocessor todavía usaba 32 hilos.
- Es posible programar medios warps por los 16 núcleos CUDA.
- Cuatro unidades de funciones especiales (SFU) con 8 núcleos CUDA cada una.
- Se cuadruplicó la memoria compartida.
- Se añadió un motor polymorph para la obtención de vectores.
- Se añadieron características de C++ para la programación en CUDA.



Arquitectura Fermi

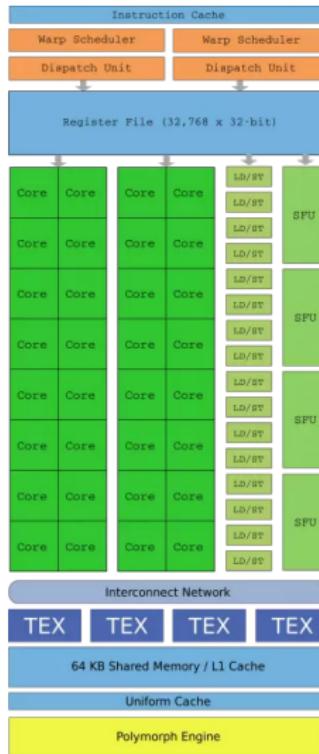


Figura: Arquitectura Fermi

Arquitectura Fermi Die GF100

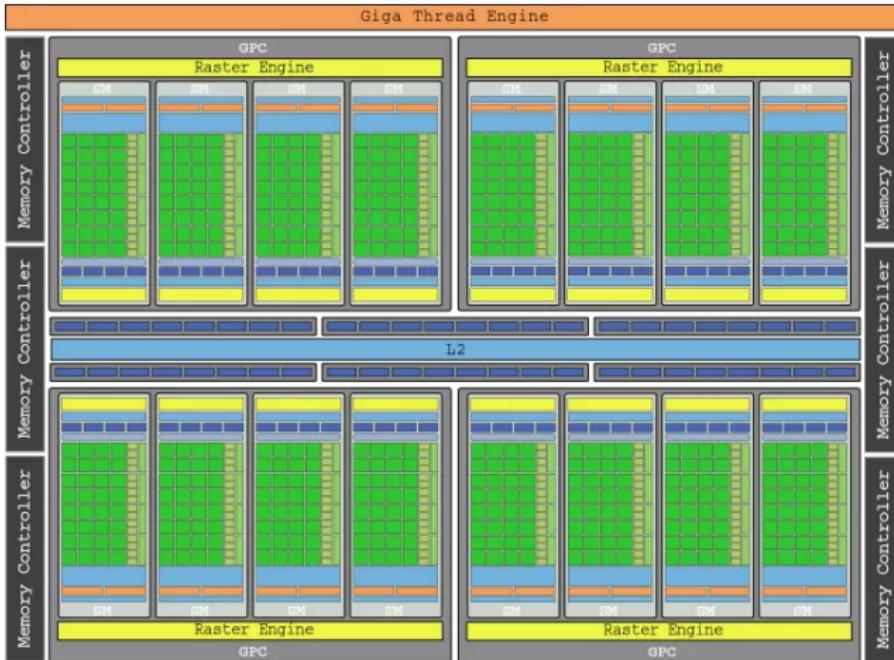


Figura: Die GF100

Arquitectura Kepler

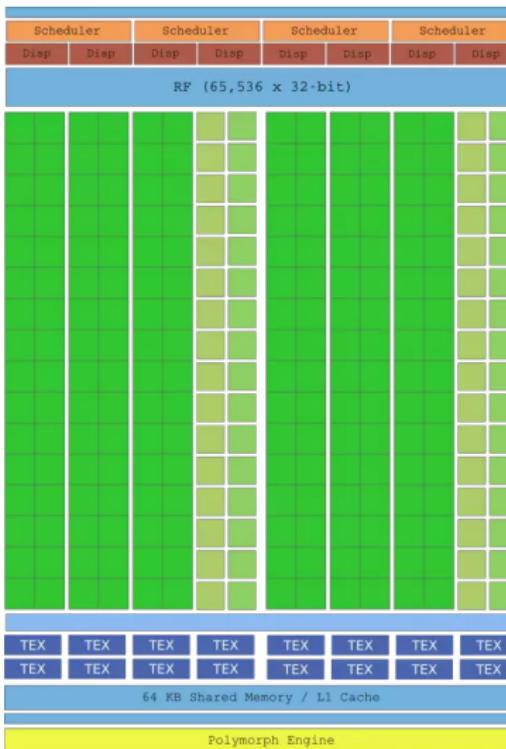


Figura: Arquitectura kepler

Arquitectura Kepler Die GK104

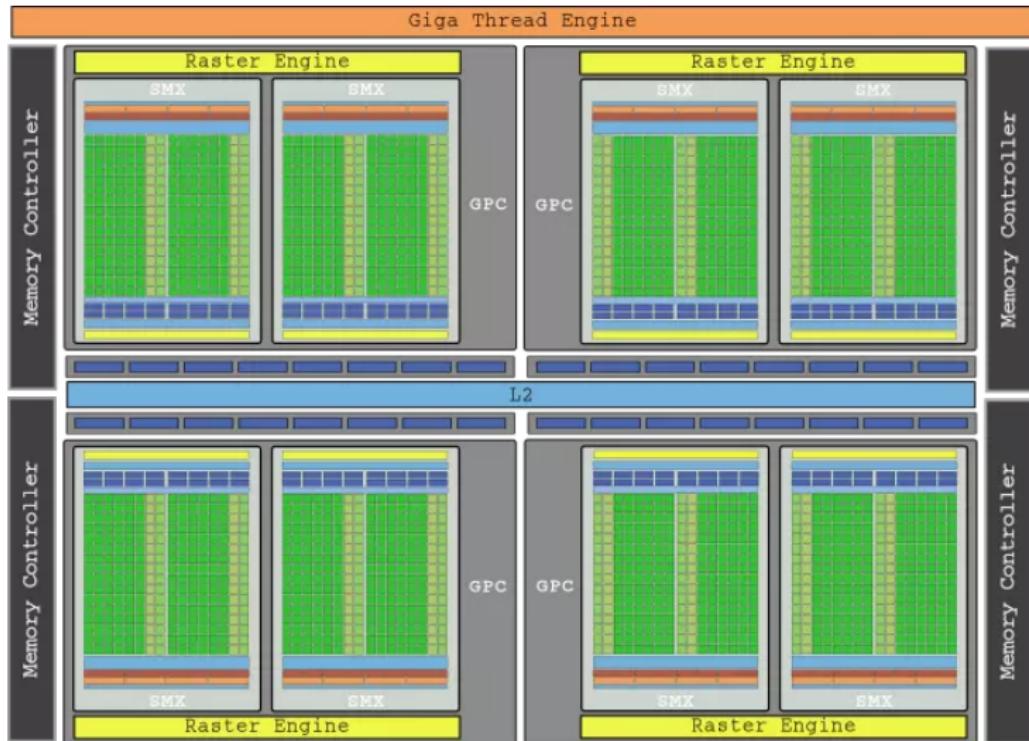


Figura: Die GK104

Arquitectura Maxwell

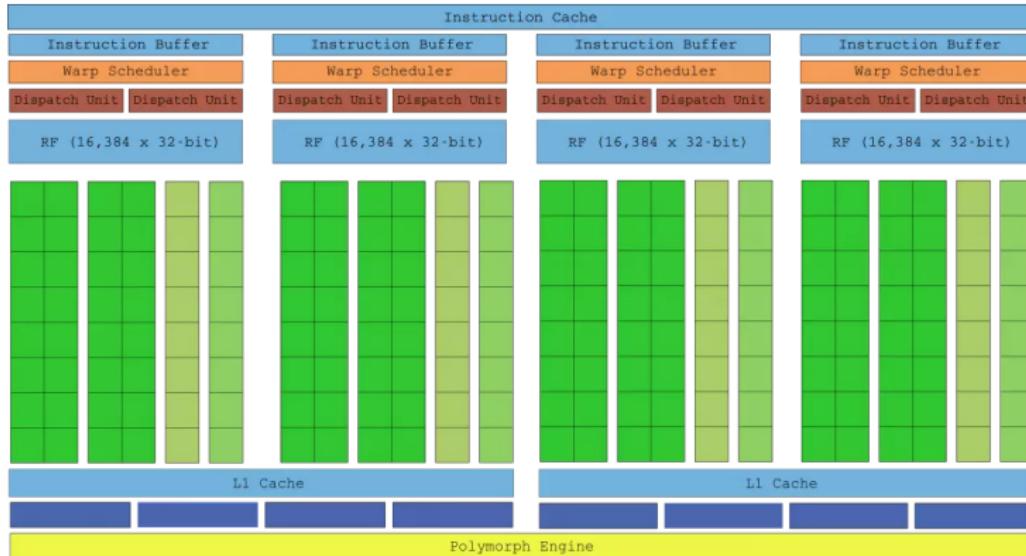


Figura: Arquitectura Maxwell

Arquitectura Maxwell Die GM200

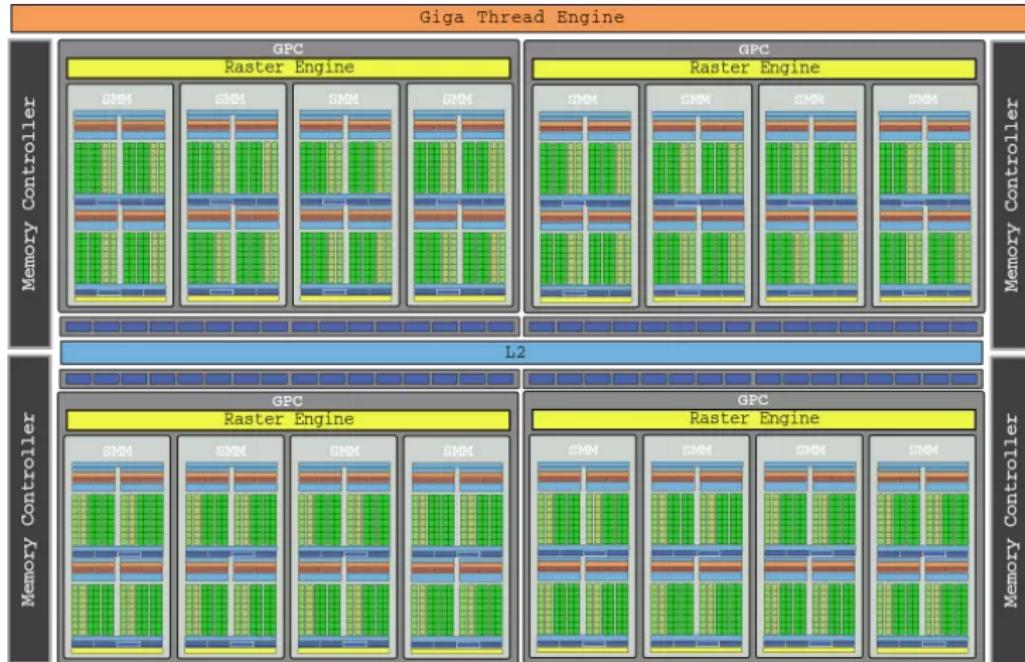


Figura: Die GM200

Arquitectura Pascal Die GP102

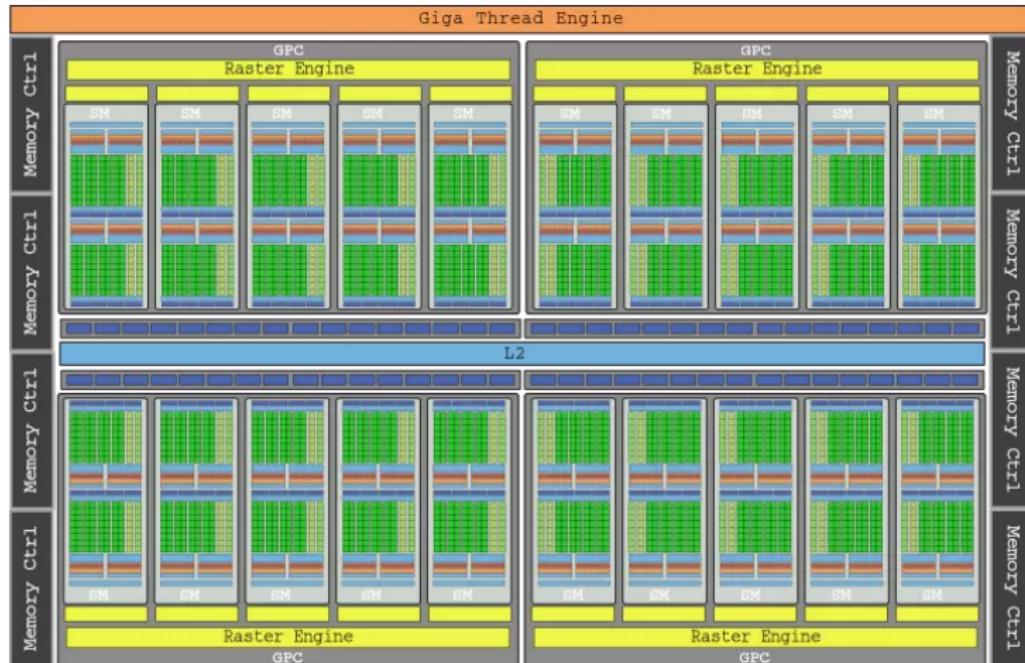


Figura: Die GP102

Arquitectura Turing

El cambio más drástico en los últimos 10 años, se volvió al diseño de capas y se añadió tecnología Ray Tracing, hardware dedicado para el trazado de rayos.



Figura: Arquitectura Turing

Arquitectura Turing



Figura: SM Turing

GPU vs CPU

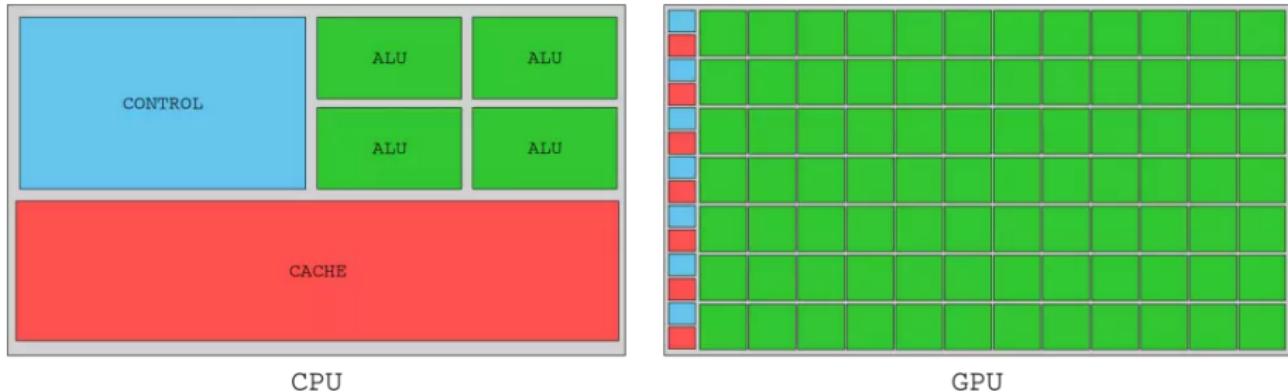


Figura: Comparativa entre GPU y CPU