



Graphic Processor Unit (GPU)

Instructor: Ching-Te Chiu





Outline

- Graphic basics
- GPU Evolution
- GPU for PC
- GPU for machine learning



Graphic Processing Unit

GPU (Graphic Processing Unit)

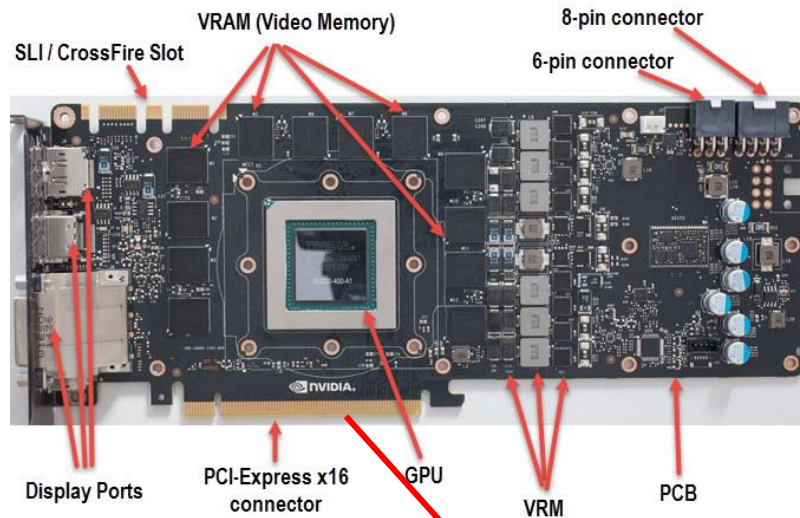
➤ Definition

- **Rapidly** manipulate and alter **memory** to accelerate the creation of **images** in a **frame buffer** intended for output to a **display device**

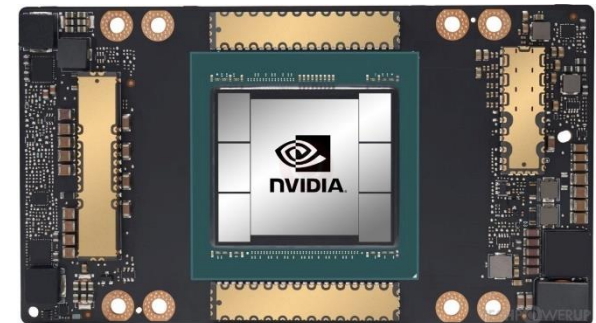
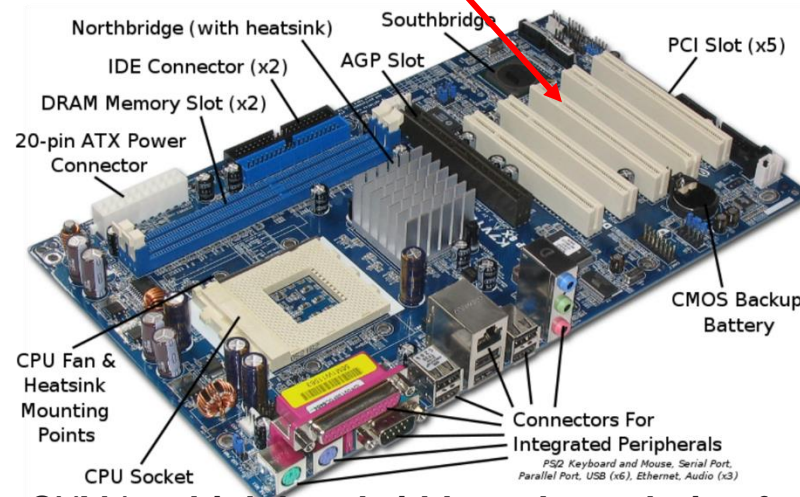
➤ Applications

- Embedded Systems
- Mobile phones
- Personal computers
- Workstations
- Game Consoles

Graphic Card Component



Nvidia A100 PCIE

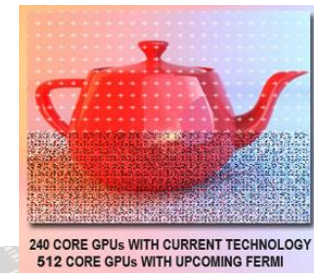
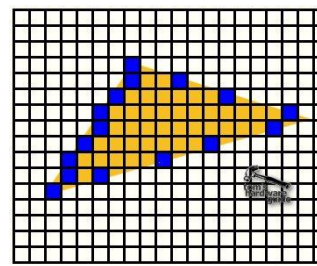
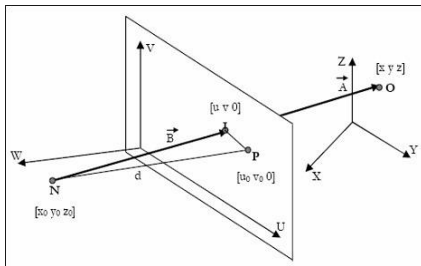


Nvidia A100 SXM4

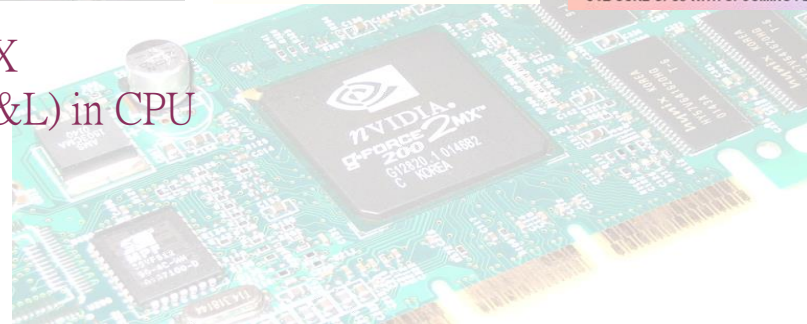
SXM is a high bandwidth socket solution for connecting Nvidia Compute Accelerators to a system.

Image Frame forming Steps

- Image Frame forming steps
 - Transform(座標轉換)
 - Triangle Setup(三角形設定)
 - Fragments
 - Texture/Lighting(光線投影)
 - Rendering(渲染)。

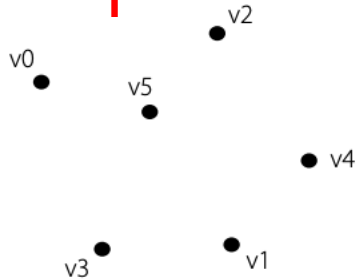


- NVIDIA GeForce 256 and GeForce 2 MX
 - Replace Transform and Lighting(T&L) in CPU



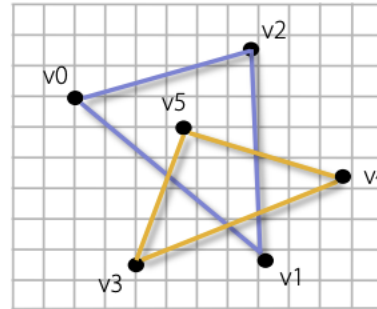
Graphic Workflow (1/3)

Step 1



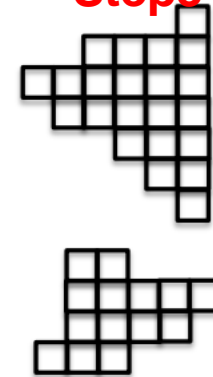
Vertices

Step 2

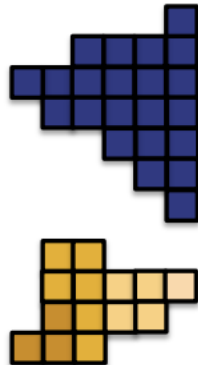


Primitives

Step 3

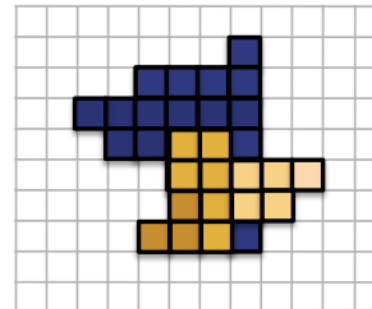


Fragments



Fragments (shaded)

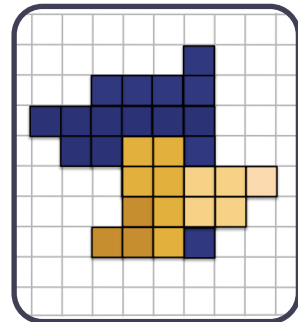
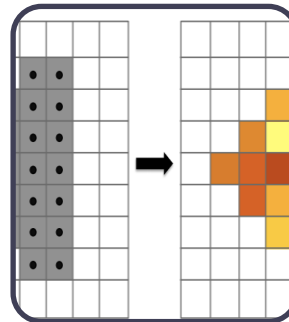
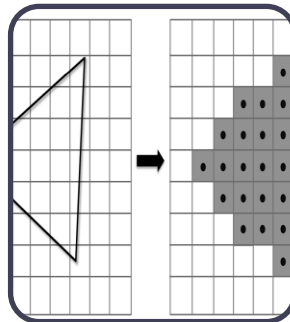
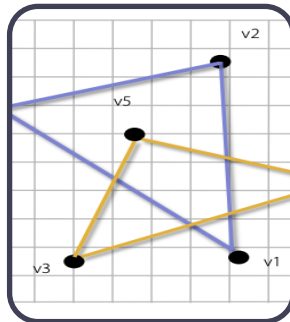
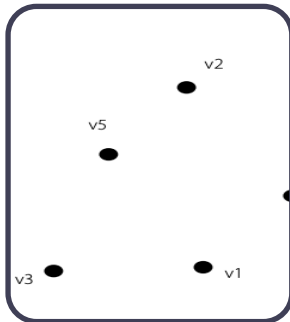
Step 4



Pixels

Step 5

Graphic Workflow (2/3)



Vertex processing

Vertices are transformed into "screen space"

EACH VERTEX IS TRANSFORMED INDEPENDENTLY

Primitive processing (triangles)

Then organized into primitives that are clipped and reduced...

Rasterization

Primitives are rasterized into "pixel fragments"

EACH PRIMITIVE IS RASTERIZED INDEPENDENTLY

Fragment processing

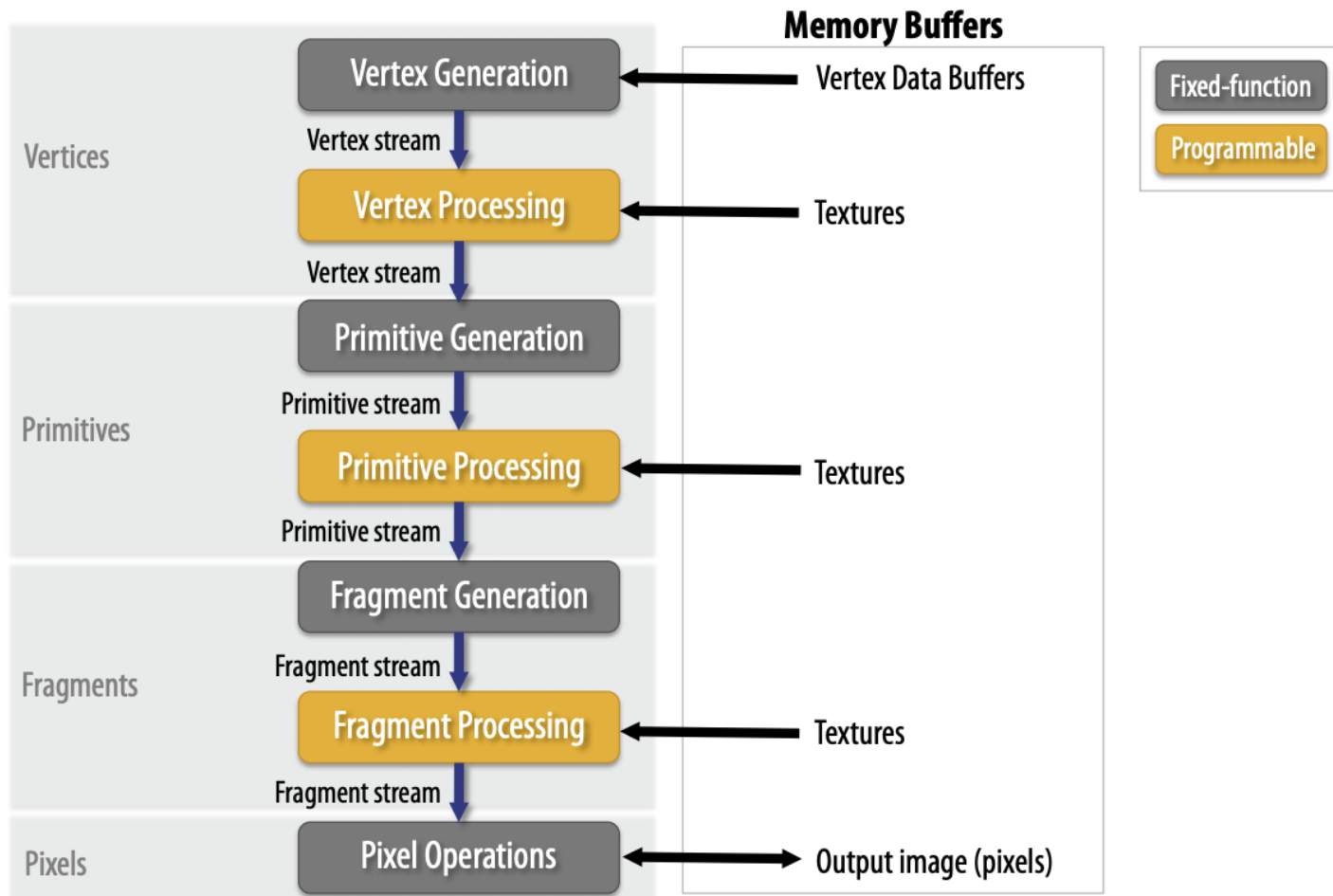
Fragments are shaded to compute a color at each pixel

EACH FRAGMENT IS PROCESSED INDEPENDENTLY

Pixel operations

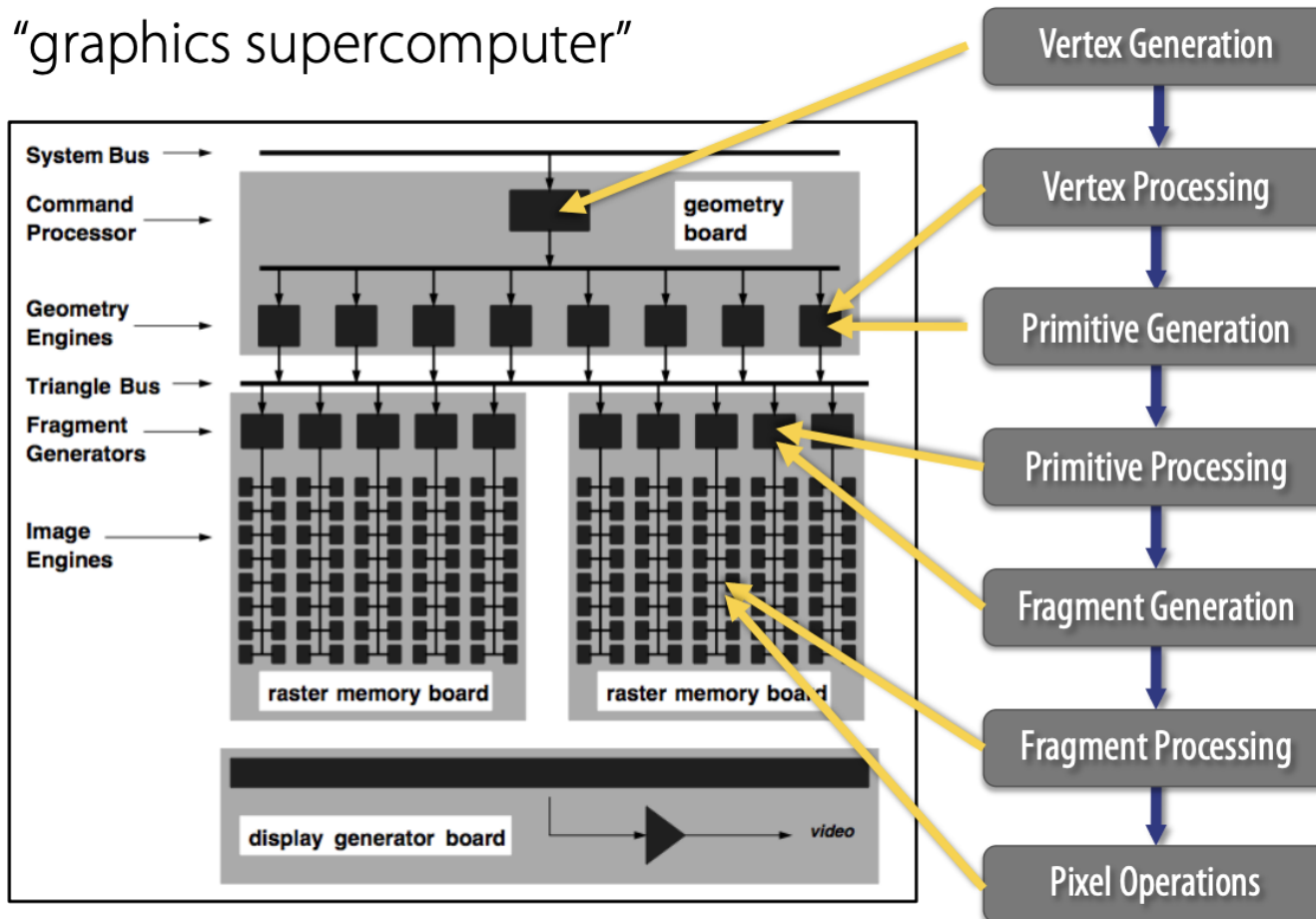
Fragments are **blended** into the **frame buffer** at their **pixel** locations (z-buffer determines visibility)

Graphic workflow (3/3)

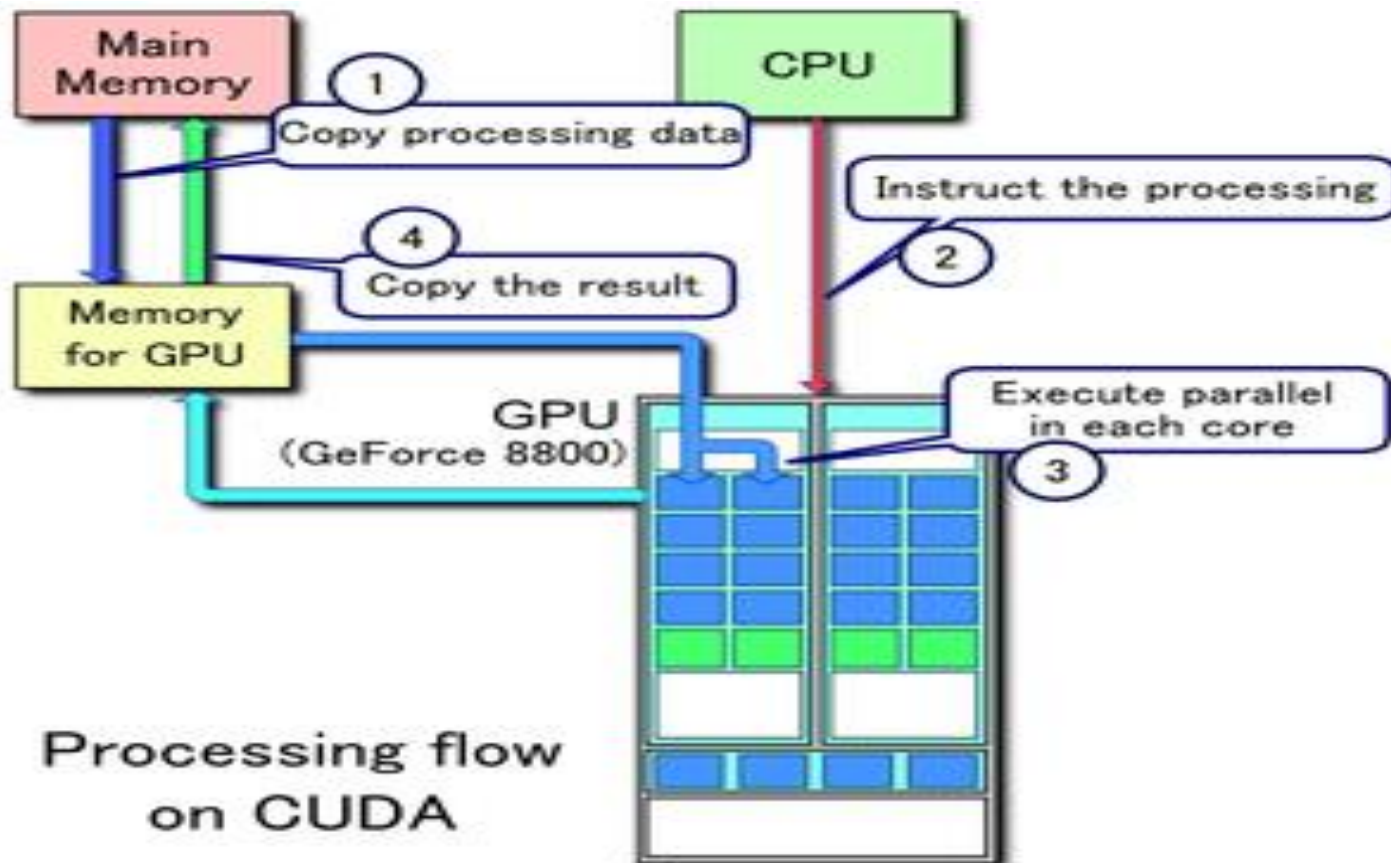


Silicon Graphics Reality Engine (1993)

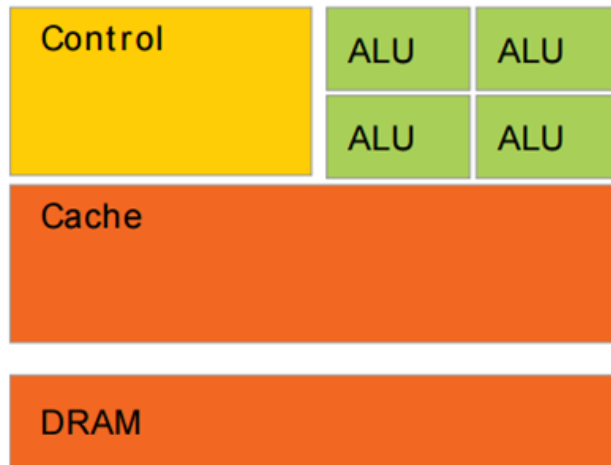
“graphics supercomputer”



Compute Unified Device Architecture (CUDA) Processing Flow



CPU vs GPU Architectures



CPU



GPU

- **Powerful ALU**

- Reduced operation latency

- **Large caches**

- Covert long latency memory accesses to short latency cache accesses

- **Sophisticated Control**

- Branch prediction for reduced branch latency
- Data forwarding for reduced data latency

- **Hundreds of Cores**

- **Thousands of Threads**

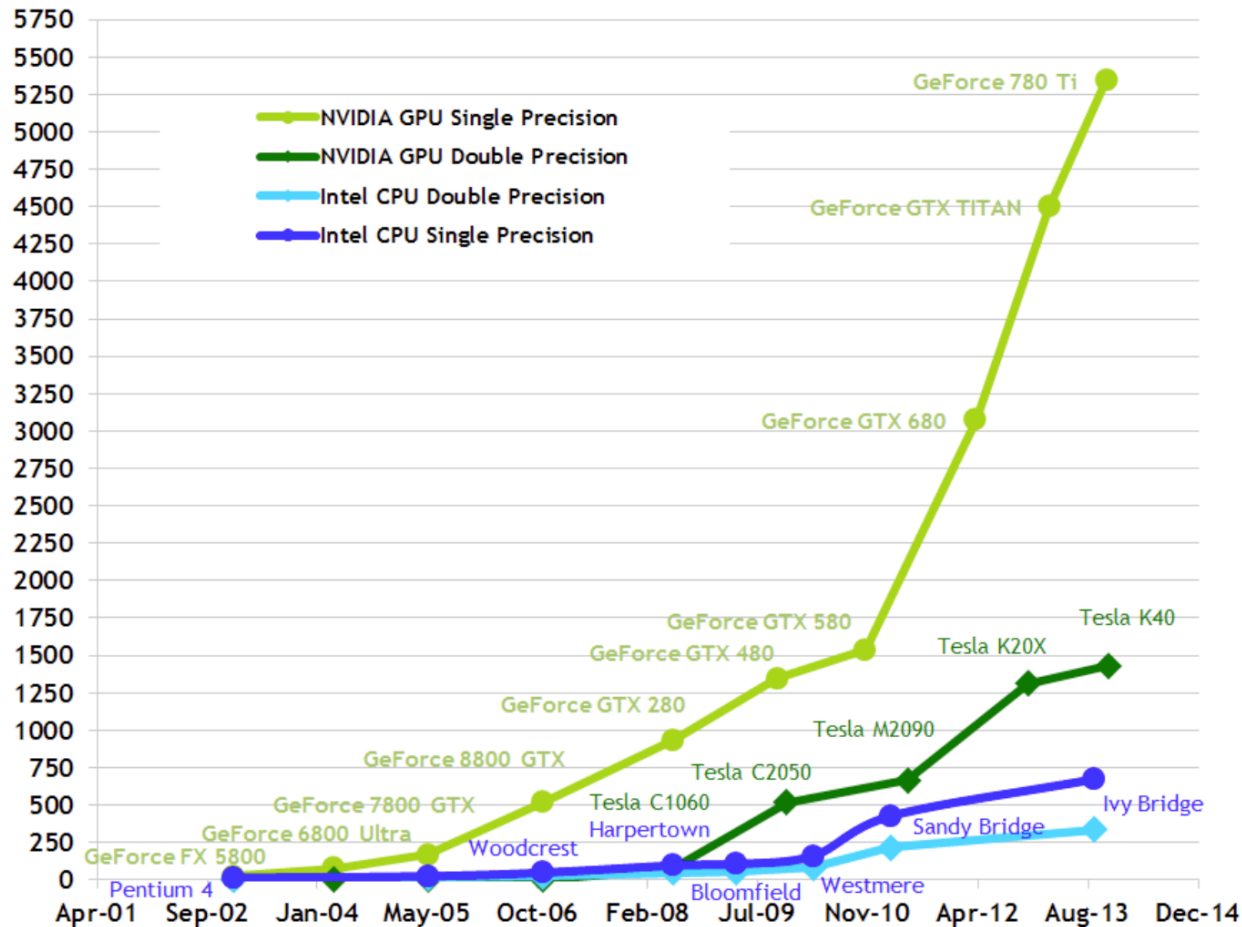
- **Single Process Execution**

- **Compute Unified Device Architecture (CUDA)**

- Parallel computing

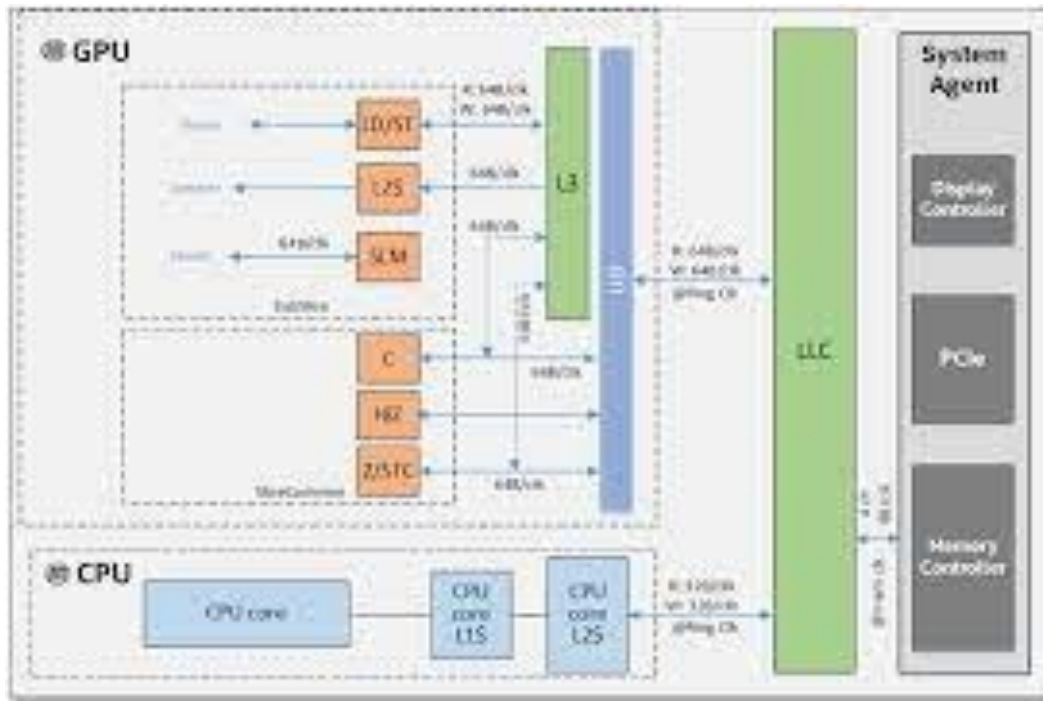
CPU vs GPU GFLOP/s Comparison

Theoretical GFLOP/s



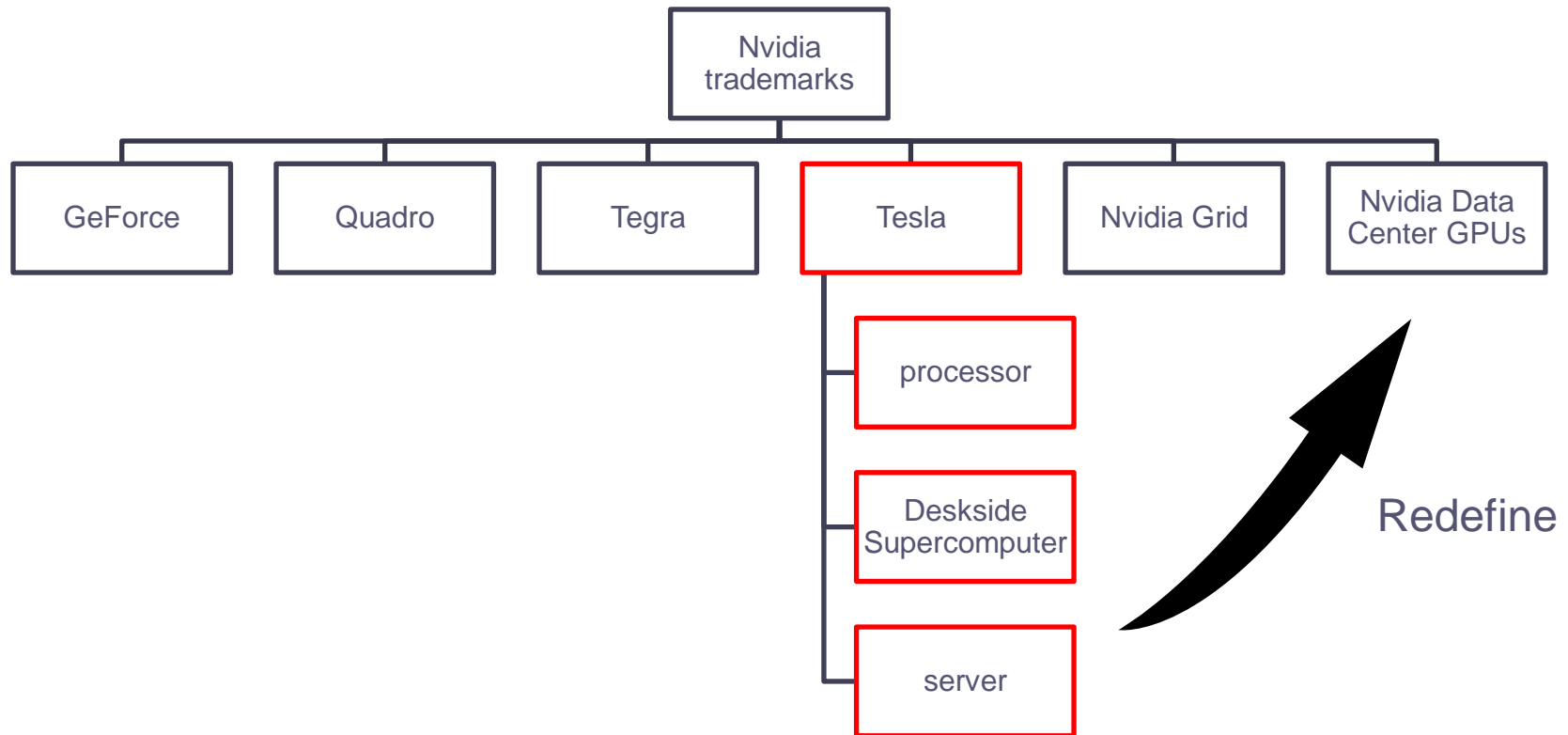
GPU for PC

- GPU integrates with CPU in SoC
- Intel acquires notebook chip maker Chips and Technologies at 1997
- AMD acquires ATI (graphic ard company) at 2006



Intel SoC integrates CPU and GPU gen9

Nvidia Trademarks Hierarchy



What is Nvidia Tesla GPU

- **Tesla** - GPU microarchitecture developed by Nvidia, and released in 2006.
- **Nvidia Tesla** - targeted at **stream processing** or general-purpose graphics processing units (**GPGPU**)
- The Nvidia Tesla product - competed with
 - AMD's Radeon Instinct
 - Intel Xeon Phi lines of deep learning and GPU cards.
- Nvidia **retired the Tesla brand in May 2020**, reportedly because of potential confusion with the brand of cars.
- Its new GPUs are branded **Nvidia Data Center GPUs**, as in the Ampere A100 GPU.

Tesla V100s 2019
GPU Computing Solutions for HPC
PCIe



Nvidia Graphics Processor Comparison



Graphics Processor	Curie	Tesla	Fermi	Kepler	Maxwell
GPU Name	G70	G80	GF100	GK104	GM200
Codename	NV47	NV50	NVC0	NVE4	NV120
Architecture	Curie	Tesla	Fermi	Kepler	Maxwell 2.0
Foundry	TSMC	TSMC	TSMC	TSMC	TSMC
Process Size	110 nm	90 nm	40 nm	28 nm	28 nm
Transistors	302 million	681 million	3,100 million	3,540 million	8,000 million
Density	906.9K / mm ²	1.4M / mm ²	5.9M / mm ²	12.0M / mm ²	13.3M / mm ²
Die Size	333 mm ²	484 mm ²	529 mm ²	294 mm ²	601 mm ²
Released	Jun 22nd, 2005	Nov 8th, 2006	Mar 26th, 2010	Mar 22nd, 2012	Mar 17th, 2015



Why GPU is suitable for Machine Learning

- Machine Learning

- Big Data for high accuracy
- Large Training Dataset
- Layers of convolutional computations (Multiplication/Add)

- GPU has the following three features

- High Bandwidth (number of data per second accessed from memory) from DRAM
 - 100GB/s
- Parallel
 - Thousands of cores
- Faster memory access from cache
 - 40TB/s

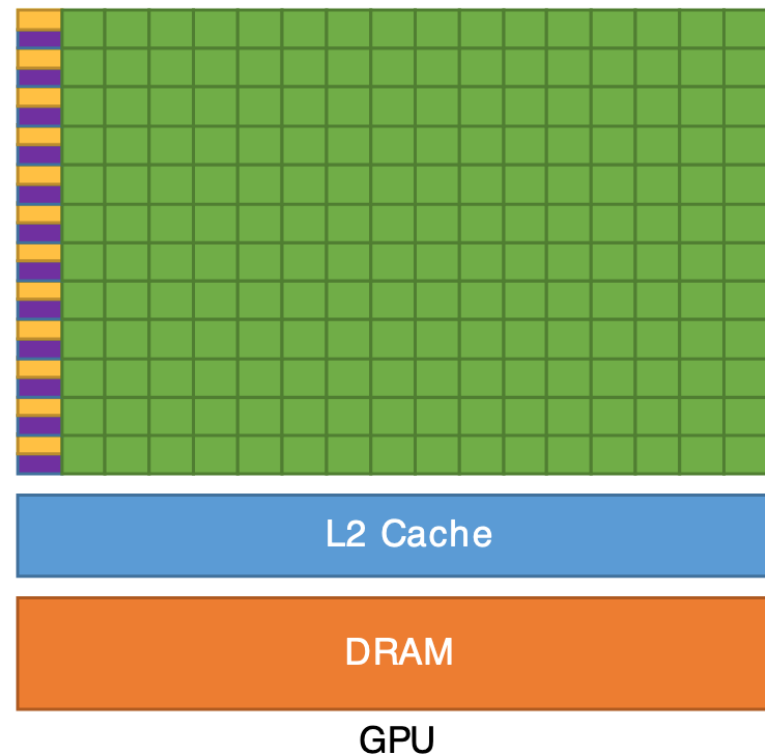
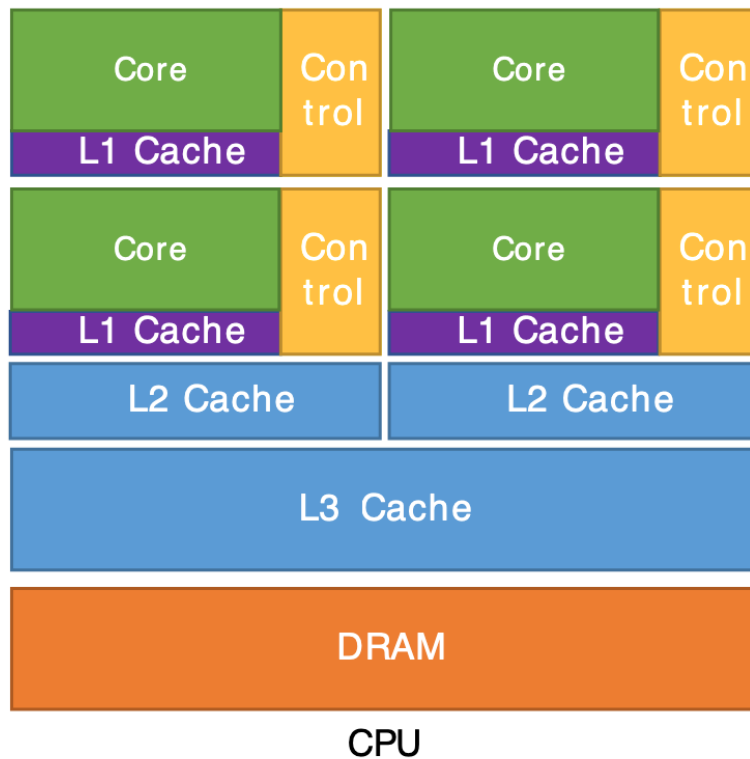
Why GPU is suitable for Machine Learning

■ CPU

- Sequential Complex computation (Weather forecast)

■ GPU

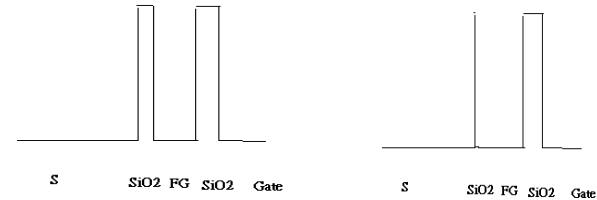
- Parallel simple computation (Yolo data training)



Comparison memory of CPU and GPU

	CPU Intel 8 core Sandy Bridge CPU	GPU Nvidia GK 110 GPU	CPU AMD Firestorm
Registers (capacity/ Bandwidth)	4kB (5TB/s)	4MB (40TB/s)	-
Cache L1 (capacity/ Bandwidth)	512KB (1TB/s)	1MB Constant mem(13TB/s) 1MB Shared mem(1TB/s)	192KB(l) +128KB(d)
Cache L2 (capacity/ Bandwidth)	2MB (1TB/s)	1.5MB (500GB/s)	8MB
Cache L3 (capacity/ Bandwidth)	8MB (500 GB/s)	-	-
Main Memory (capacity/ Bandwidth)	10GB (20GB/s)	4GB (150GB/s)	64GB (400GB/s)
Access data from memory	Fewer data/low bandwidth	Large amount of data/high bandwidth	High bandwidth 19

Fowler-Nordheim



■ Fowler-Nordheim(FN)

- It directly applies **high voltage on both sides of the insulating layer** to form a **high-strength electric field** to help electrons enter and exit the **floating gate through the oxide layer** channel.

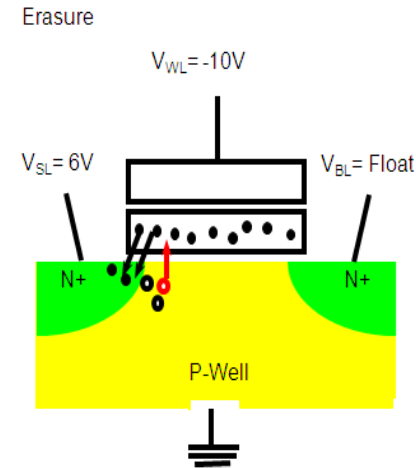
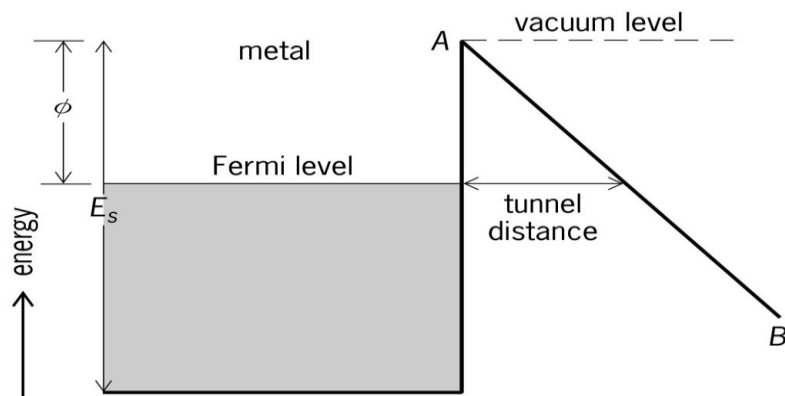
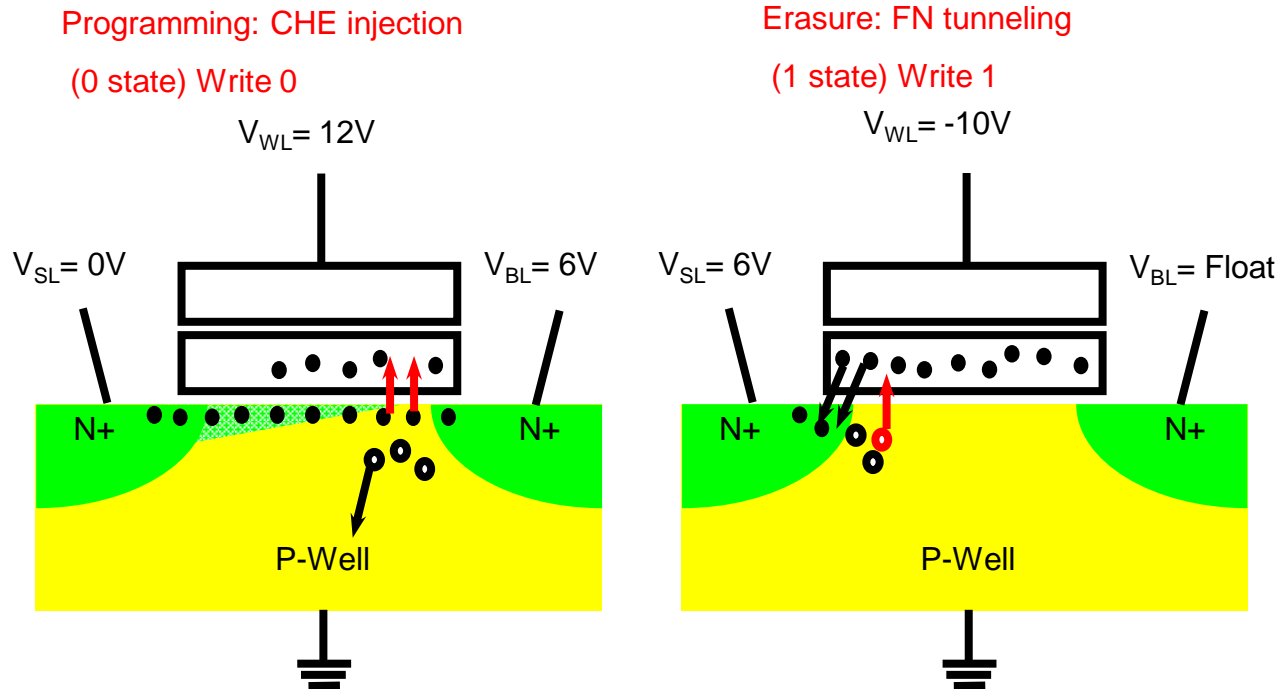


Diagram of the energy-level scheme for field emission from a metal at absolute zero temperature

Write and Erase

Program(1 to 0) with CHE (channel hot electron) injection
or Fowler Nordheim (FN) electron tunneling
Erase(0 to 1): with FN (Fowler-Nordheim) tunneling





Nvidia GPU Microarchitecture list

1999 - World's First GPU: GeForce 256

2001 - First Programmable GPU: GeForce3

2004 - GeForce 6 (HDR), Scalable Link Interface

2006 - CUDA Architecture Announced

Year	Arch	Main Technology	Consumer Series	Data Center Series	Die	Fab Process		Enthusiast Consumer Card	Enthusiast Server Card
2006	Tesla	CUDA	GeForce 8	Tesla S00,C1000,S1000, Quadro Plex	G80	90 nm	CMOS	GTX8800	S870
2010	Fermi	FP64 ECC	GeForce 400	Tesla C2000,M2000,S2000	GF100	40 nm		GTX 480	S2050
2012	Kepler	Dynamic Parallelism	GeForce 600	GRID: K1,K2,K340,K520 Tesla: K10,K20	GK104	28 nm		GTX 860	GRID K520 Tesla K10
2014	Maxwell	Higher Pref/Watt	GeForce 900	Tesla M	GM204	28 nm		GTX 980 Ti	Tesla M60
2016	Pascal	Unified Memory Stacked DRAM NVLINK	GeForce 10	Tesla P	GP102	16 nm	FinFET	GTX 1080 Ti	Tesla P40 Tesla P100
2017	Volta*	Tensor core	Titan V	Tesla V100	GV100	12 nm	FinFET	Titan V	Tesla V100
2018	Turing	RTX	GeForce 20	Tesla T4	TU102	12 nm	FinFET	RTX 2080 Ti	Tesla T4
2020	Ampere	MIG	GeForce 30	Tesla A100	GA102	8 nm		RTX 3080	Tesla A100

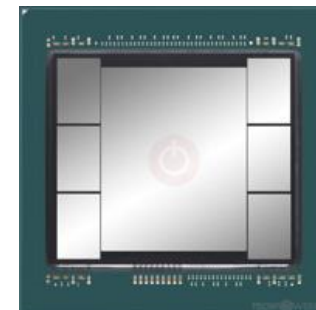
*Quadro GV100 is the first application of Tensor core designed for deep learning

RTX:Ray Tracing

MIG: Multi-Instance GPU

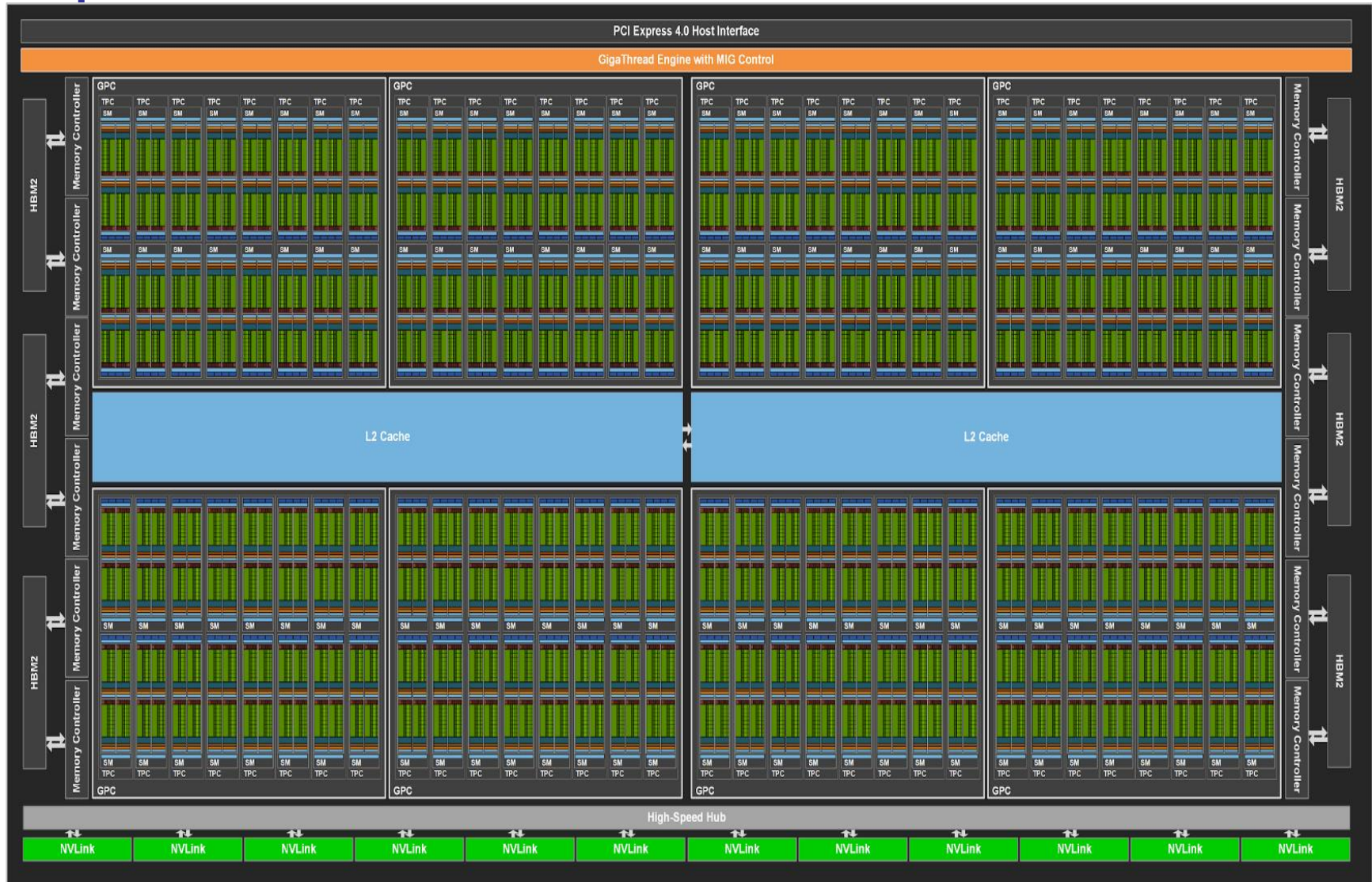


Nvidia Graphics Processor Comparison (cont'd)



Graphics Processor	Pascal	Volta	Turing	Ampere
GPU Name	GP100	GV100	TU102	GA100
Codename	NV130	NV140	NV162	NV170
Architecture	Pascal	Volta	Turing	Ampere
Foundry	TSMC	TSMC	TSMC	TSMC
Process Size	16 nm	12 nm	12 nm	7 nm
Transistors	15,300 million	21,100 million	18,600 million	54,200 million
Density	25.1M / mm ²	25.9M / mm ²	24.7M / mm ²	65.6M / mm ²
Die Size	610 mm ²	815 mm ²	754 mm ²	826 mm ²
Released	Apr 5th, 2016	Jun 21st, 2017	Aug 13th, 2018	May 14th, 2020

2020 Ampere Microarchitecture-GA100



GA100 GPU with 128 SM includes

- Multiple GPU processing clusters (GPCs)
- Texture processing clusters (TPCs)
- Streaming multiprocessors (SMs)
- Second generation High bandwidth Memory (HBM2) controllers

2020 Ampere Microarchitecture-GA100

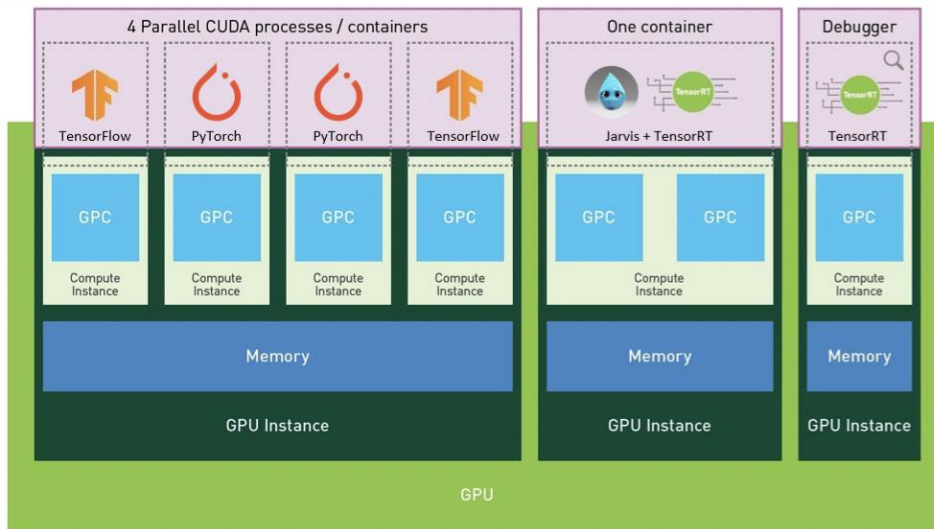
- The full implementation of the GA100 GPU includes the following units:
 - 8 GPCs (GPU processing clusters)
 - 8 TPCs/GPC (texture processing clusters)
 - 2 SMs/TPC (streaming multiprocessors)
 - 16 SMs/GPC, 128 SMs per full GPU
 - 64 FP32 CUDA Cores/SM,
 - 8192 FP32 CUDA Cores per full GPU
 - 4 Third-generation Tensor Cores/SM,
 - 512 Third-generation Tensor Cores per full GPU
 - 6 HBM2 stacks
 - 12- 512-bit Memory Controllers
 - A100 7nm TSMC Process, RTX30(real-time raytracing) 8nm Samsung



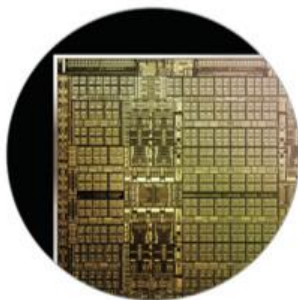
SM in GA100 GPU



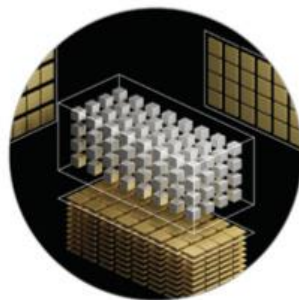
2020 Ampere Microarchitecture-GA100 (cont'd)



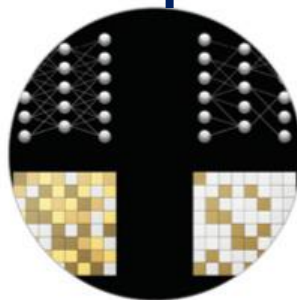
- A Compute Instance is defined as including one Sys Pipe with up to 7 GPCs within a GPU Instance



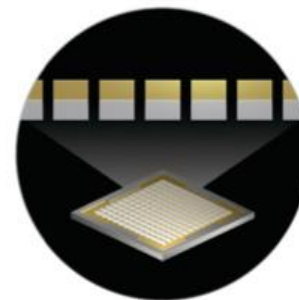
54 BILLION XTORS



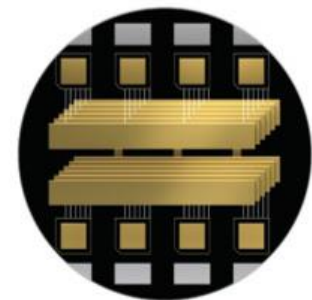
3rd GEN
TENSOR CORES



SPARSITY
ACCELERATION



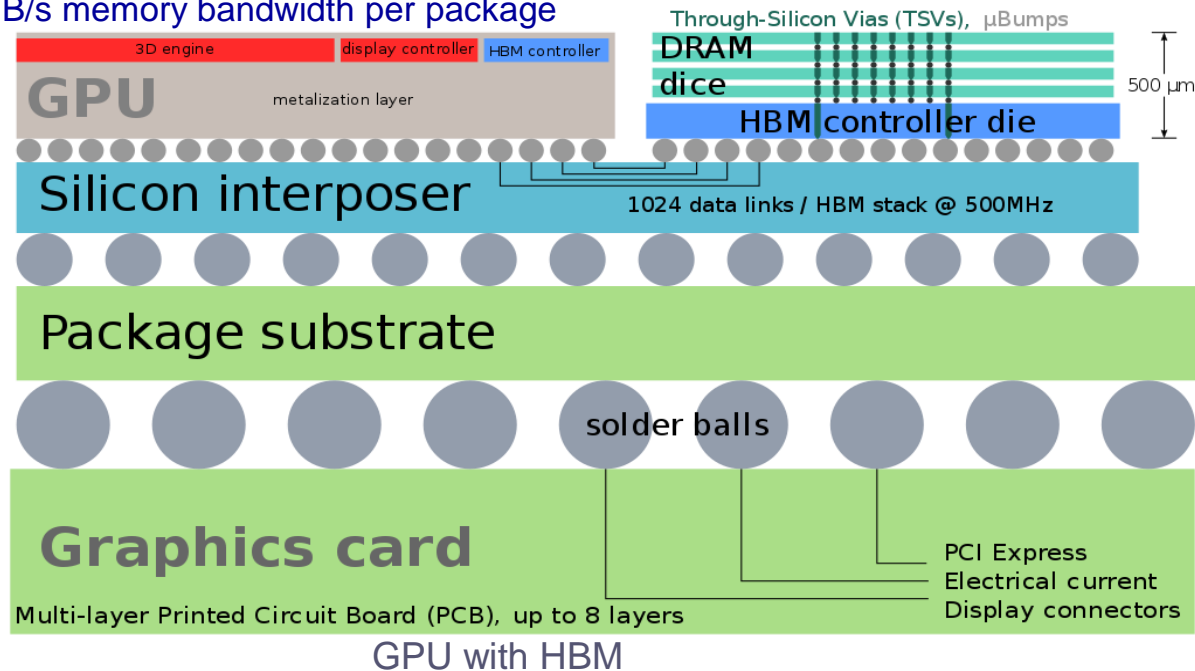
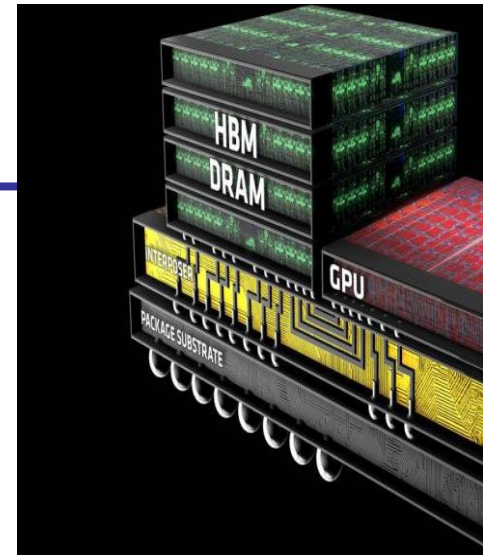
MIG



3rd GEN
NVLINK & NVSWITCH

High Bandwidth Memory (HBM)

- HMB: high-speed memory interface for 3D stacked SDRAM
 - Applications: GPU/network devices/AI accelerator
 - Stacking DRAM dies with an optional base die through a substrate (silicon interposer)
 - HBM2
 - Eight DRAM dies/ stack
 - 2GT/s (giga transfer/second)
 - 1024 bit width
 - 256GB/s memory bandwidth per package

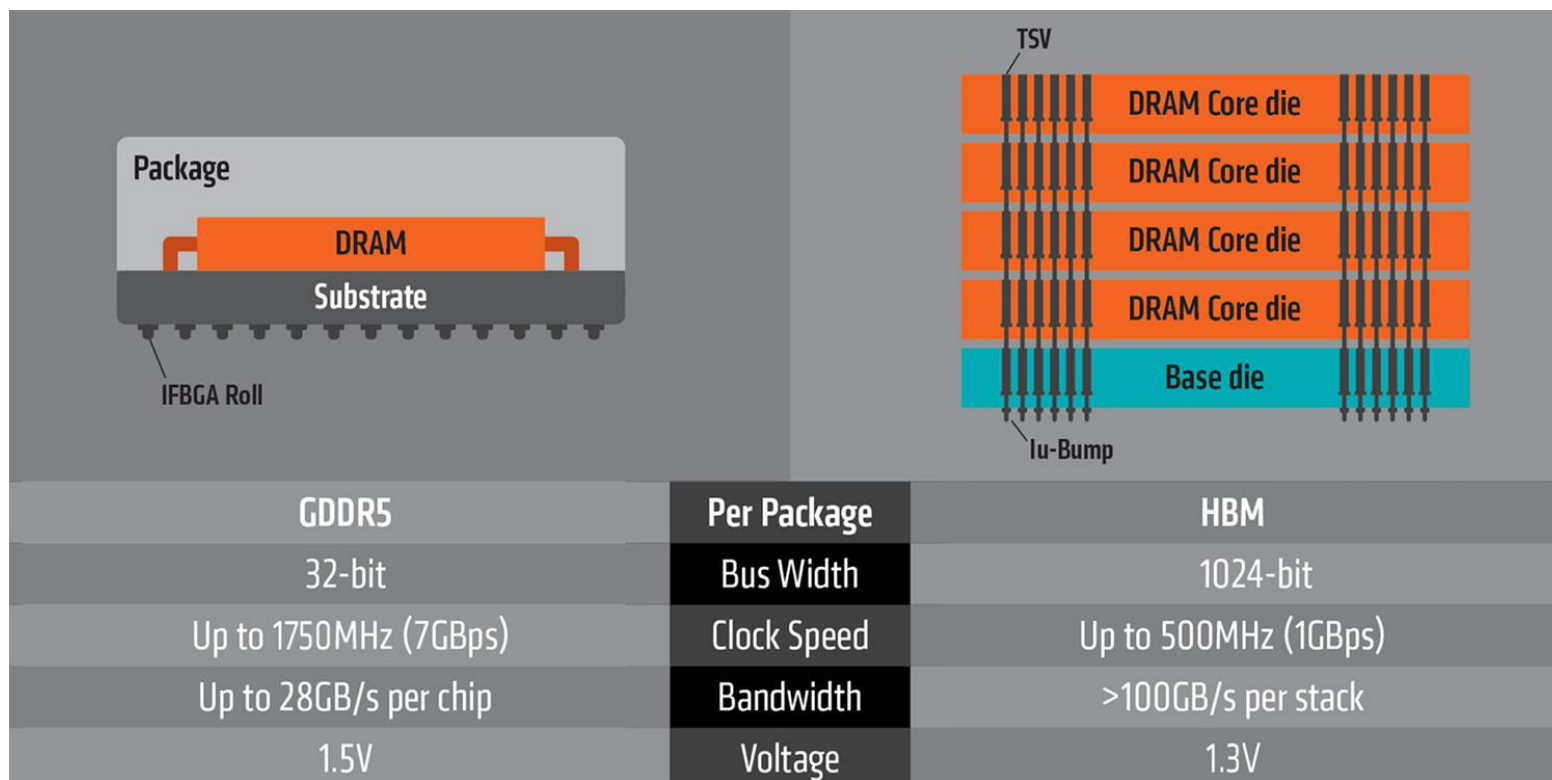


Comparison of GDDR and High Bandwidth Memory

■ HMB:

- Applications: GPU/network devices/AI accelerator
- Stacking DRAM dies with an optional base die through a substrate (silicon interposer)

■ HBM2















Nvidia Competitors or Alternative

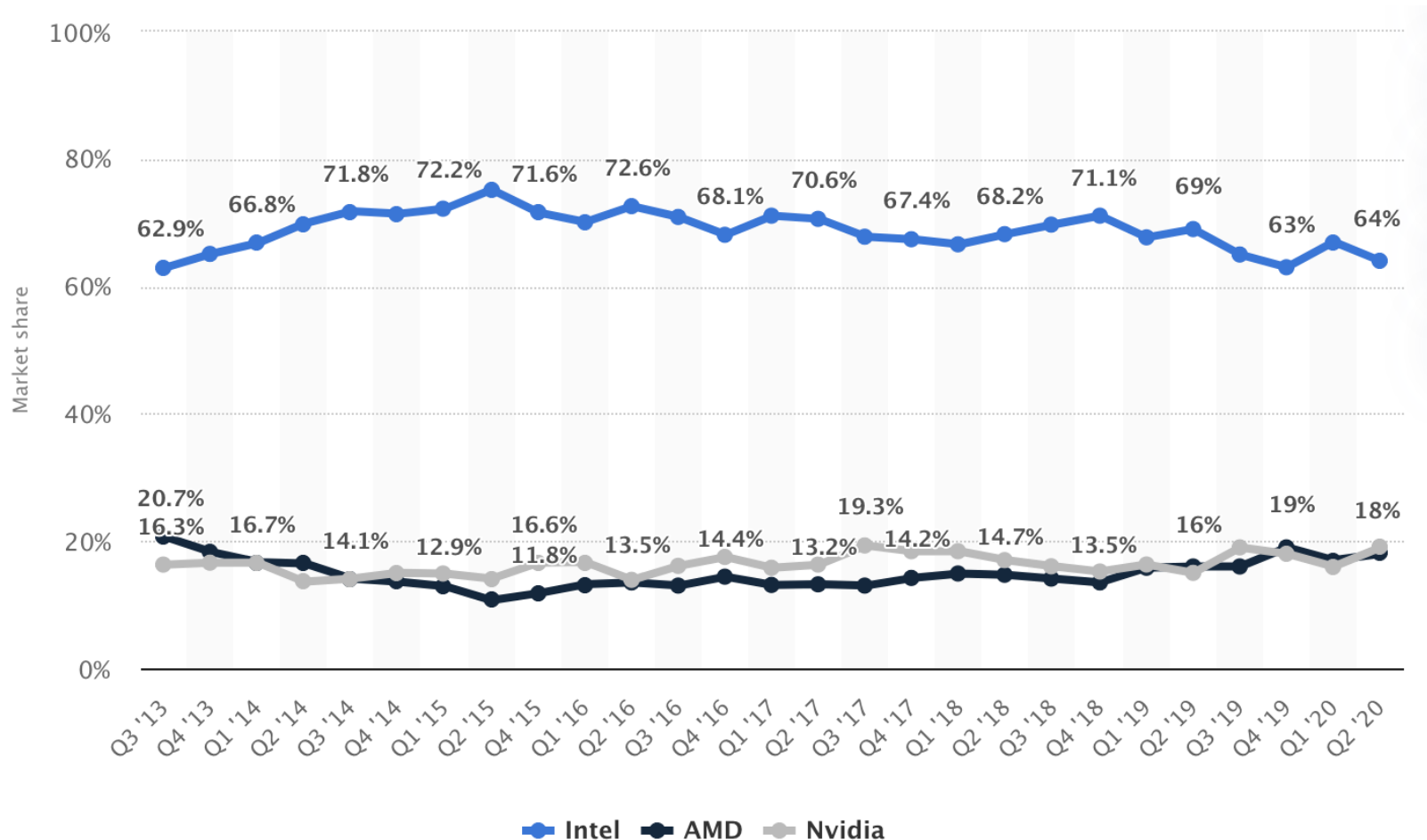
■ Below are the top 4 Nvidia competitors:

1. AMD (Advanced Micro Devices)
2. Intel
3. Marvell technology group
4. Qualcomm

TOP COMPETITORS OR ALTERNATIVES

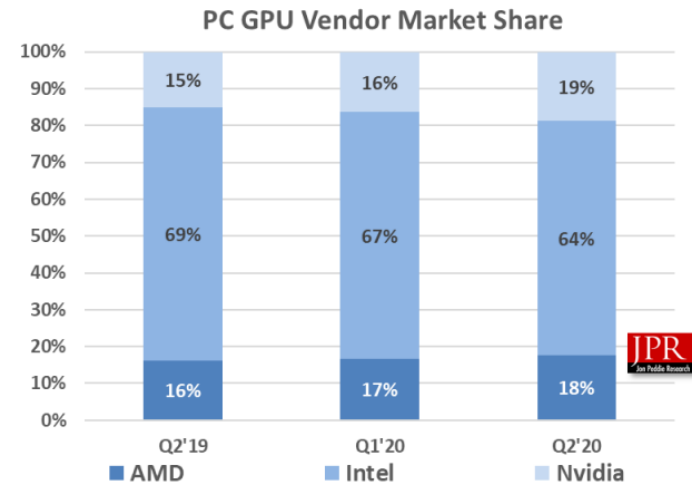
RANK	COMPANY	CEO	CEO RATING	EMPLOYEES	FUNDING	REVENUE
	 NVIDIA	 Jen-Hsun Huang President & CEO	63/100	13,277	\$42M	\$10.9B
1	 AMD	 Lisa Su President & CEO	80/100	10,100	\$1.8B	\$7.6B
2	 intel	 Robert H. Swan CEO	58/100	107,400	\$3B	\$79B
3	 XILINX	 Victor Peng President & CEO	76/100	4,433	\$750M	\$3B
4	 Ambarella	 Fermi Wang CEO	71/100	750	\$36M	\$229.9M
5	 BROADCOM	 Hock E. Tan President & CEO	63/100	10,650	\$78.1M	\$8.5B

PC graphics processing unit (GPU) shipment share worldwide from 3rd quarter 2013 to 2nd quarter 2020, by vendor



GPU Market Share forecast

- The microprocessor and GPU market was valued at USD 83.1 billion in 2019 and is projected to reach USD **112.7 billion** by 2025
- it is projected to grow at a compound annual growth rate (CAGR) of 7.3% between 2020 and 2025.





GPU in Cloud Data Center Provider

Share of Compute Instance Types with Dedicated Accelerators Offered by the Top Four Public Clouds

(Alibaba Cloud, Amazon Web Services, Google Cloud & Microsoft Azure)

Company	Accelerator	March 2019	April 2019	May 2019
NVIDIA	GPU	97.0%	97.3%	97.4%
AMD	GPU	1.2%	1.1%	1.0%
Xilinx	FPGA	1.1%	1.0%	1.0%
Intel	FPGA	0.6%	0.6%	0.6%
Total Types	All	1,852	1,990	2,003

Source: Liftr Cloud Insights, June 2019

Reference

- <http://www.nvidia.com.tw/page/products.html>
- <http://zh.wikipedia.org/wiki/NVIDIA>
- <http://www.anandtech.com/show/2911>
- http://mag.udn.com/mag/digital/storypage.jsp?f_ART_ID=297019
- <http://www.anandtech.com/show/5762/nvidia-plots-mobile-soc-gpu-performance-surpassing-xbox-360-by-2014>
- <http://chinese.engadget.com/2011/05/26/android-3-0-main-processor-introducing-nvidia-tegra-2/>
- <http://www.ciol.com/Semicon/Biz-Watch/News-Reports/Fabless-IC-suppliers-ranking/162275/0/>
- <http://www.techbang.com/posts/7491-41-core-tegra-3-5-speed-performance-nda-11-9-1400-ban>