NVIDIA GeForce 30 series

Department： Electrical Engineering

Student ID：108061181

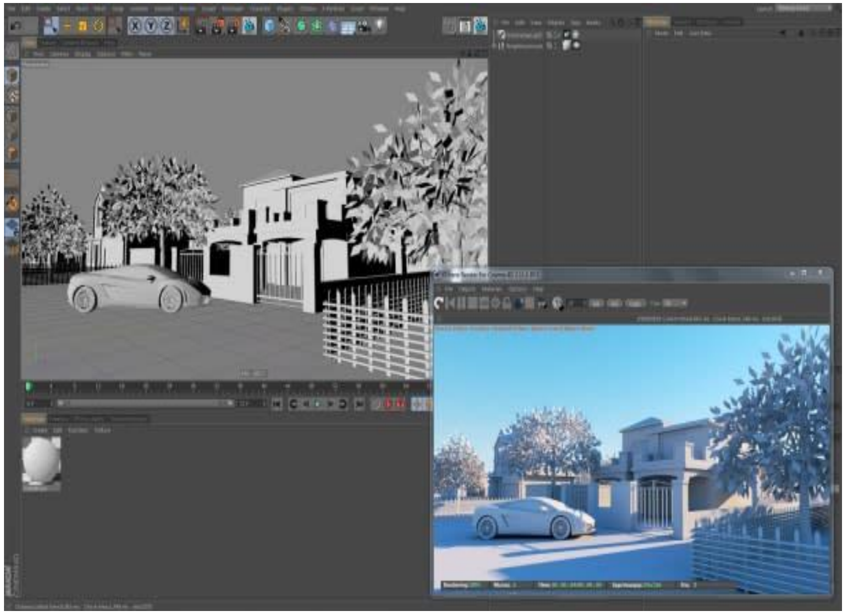
Name ：Ivan Andre Castillo Barahona

Outline

- Application
- Motivation/System Evolution
- Turing(20 series) vs Ampere(30 series) block diagram and specs
- Naming Scheme
- L1 cache
- New gen Tensor, RT Cores and memory
- Industrial Analysis

System application

- **Video editing, 3D graphics rendering, cryptocurrency, machine learning, streaming**



3D rendering program (*CINEMA 4D interface)*

GPUs used in crypto mining

What are NVIDIA GPUs known for?

- **High resolution gaming, ray tracing, DLSS**

C) Far Cry 6 w/o RTx

D) RDR 2 w/o DLSS

Motivation/ System Evolution

- **Motivation? Ans. Need for smooth graphics**

NVIDIA's 1999 Geforce 256

TSMC's 220nm process
17 million transistors
Memory of 32MB
Die size: 139mm²

NVIDIA's Geforce 3090ti

Samsung's 8nm process
28 billion transistors
24GB of G6x memory
Die size: 628mm²

Turing vs Ampere(20 vs 30 series)

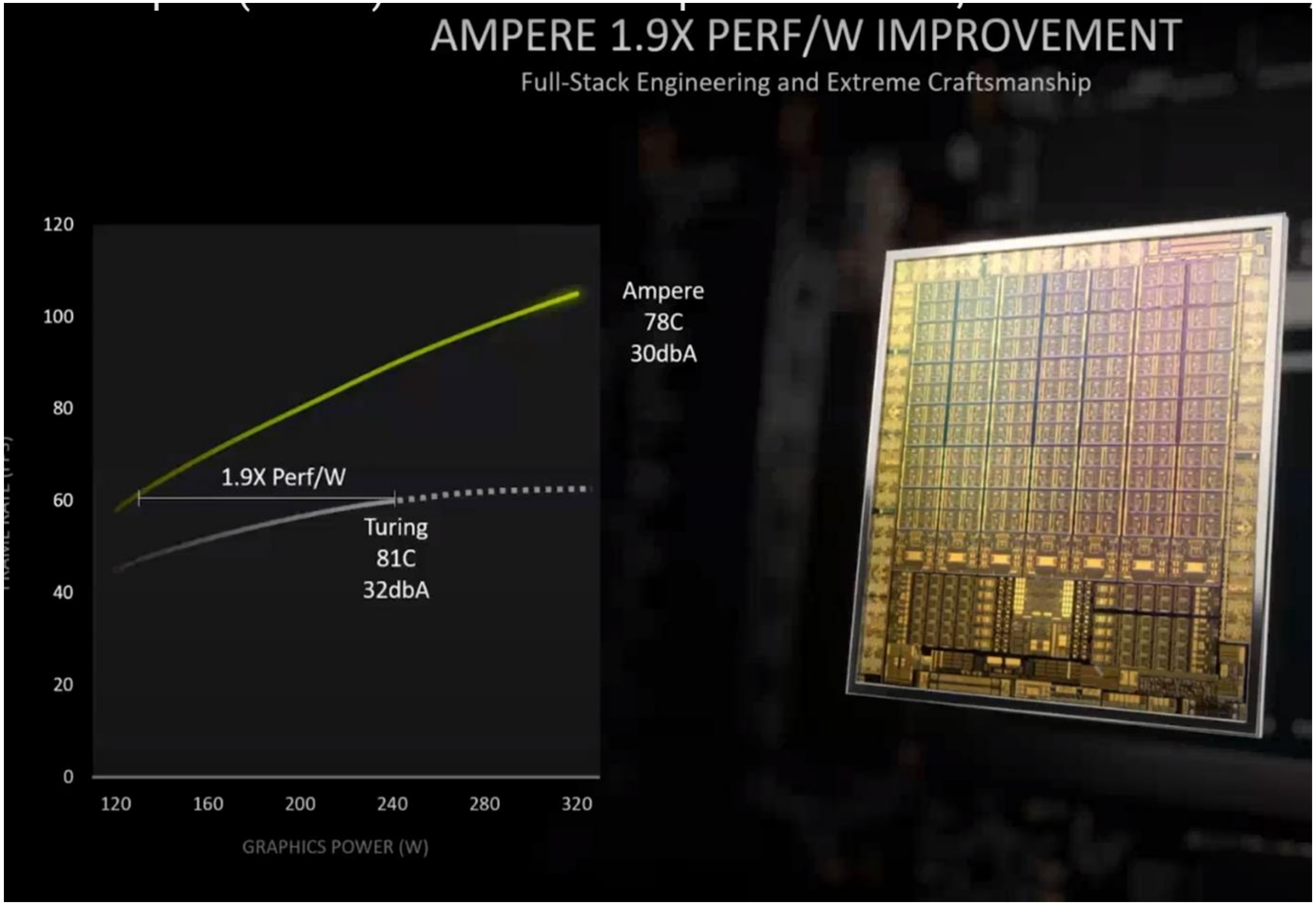
- TSMC 12nm vs Samsung 8nm process
- More transistor density

GPU	GeForce RTX 2080 Super (Founder's Edition)	GeForce RTX 3080 (Founder's Edition)
TGP	250W	320W

TGP(Total graphics power)

- Usually smaller process ➔ smaller power consumption(NOT in this case!)
- What are the effects on power performance per watt?

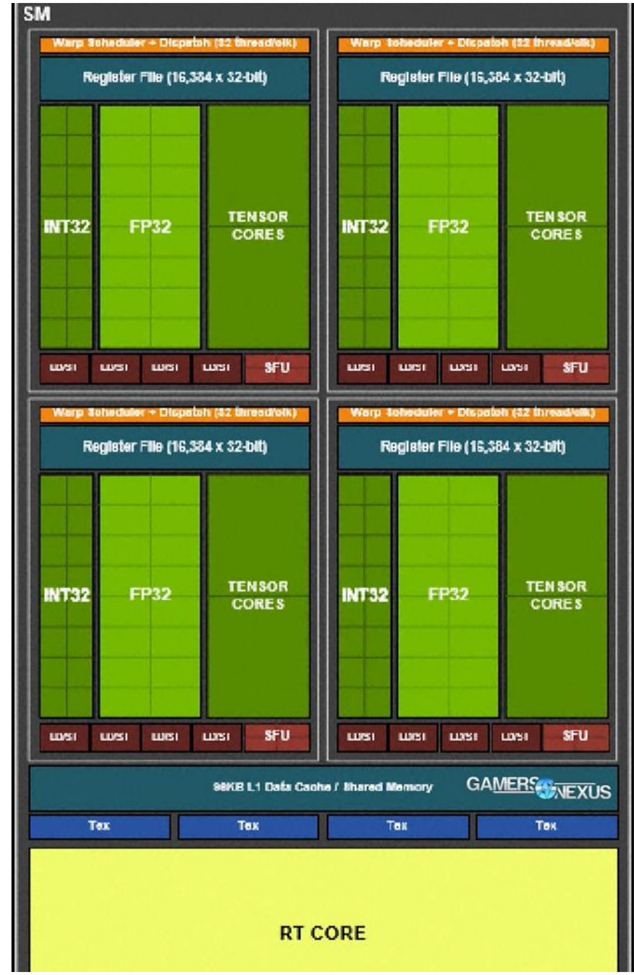
Performance per watt comparison



- Increase in power consumption != increment in temperature

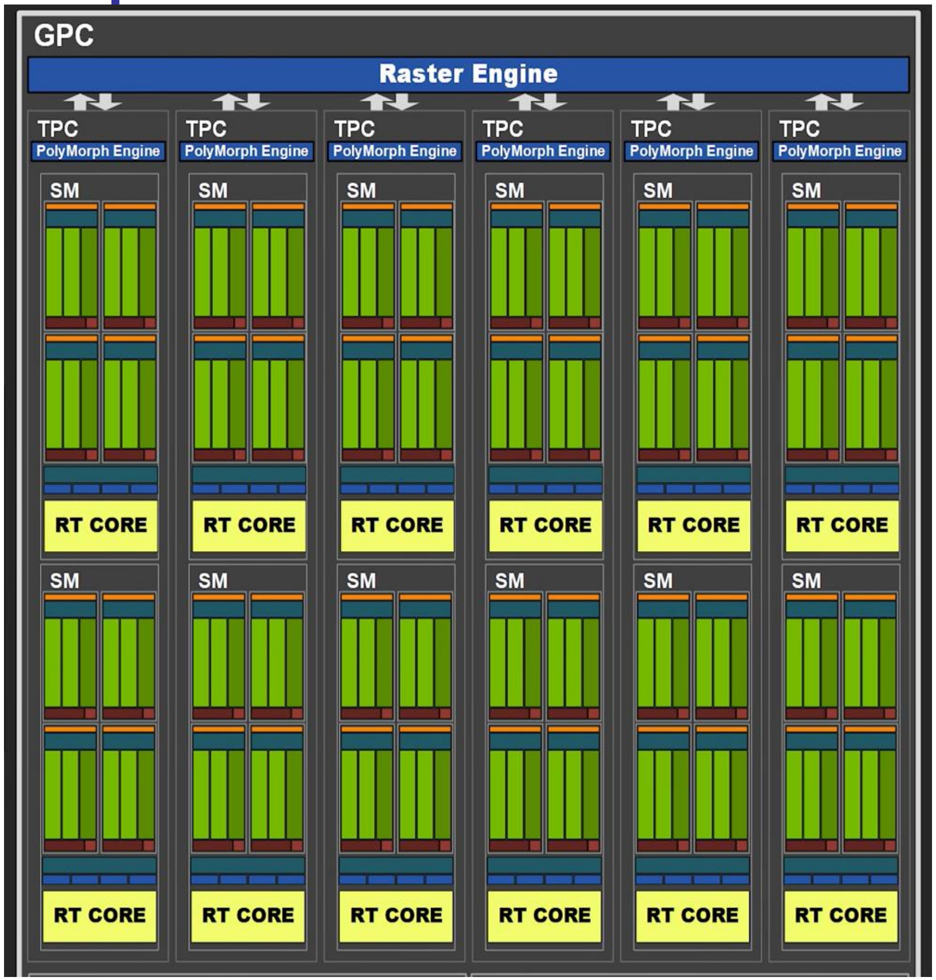
Turing’s block diagram

Turing's GPCs and SMs



Ampere's GA-102 block diagram

Ampere GPCs and SMs



Specs comparison

Table 2. Comparison of GeForce RTX 3080 to GeForce RTX 2080 Super

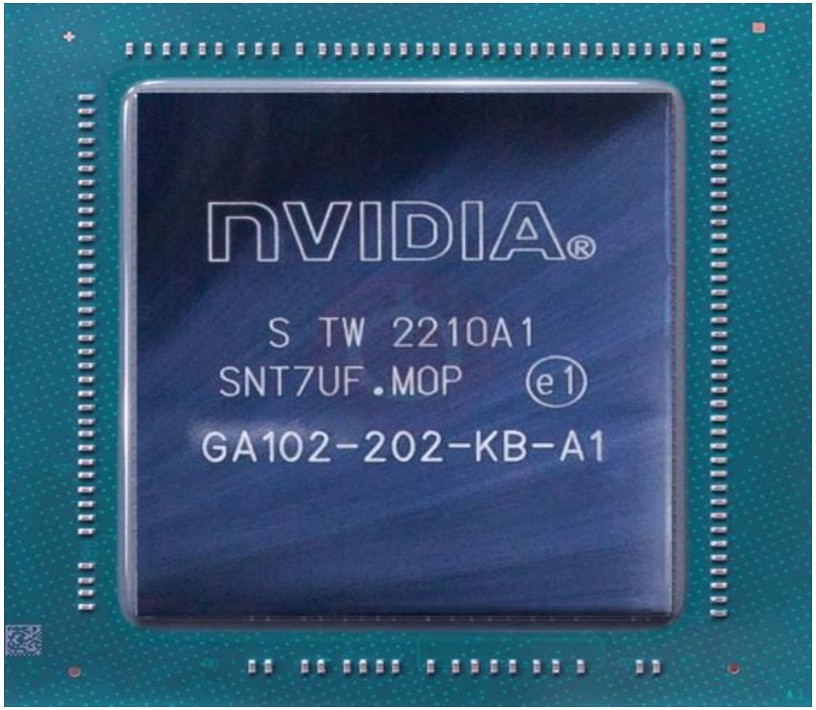
Graphics Card	GeForce RTX 2080 Founders Edition	GeForce RTX 2080 Super Founders Edition	GeForce RTX 3080 10 GB Founders Edition
GPU Codename	TU104	TU104	GA102
GPU Architecture	NVIDIA Turing	NVIDIA Turing	NVIDIA Ampere
GPCs	6	6	6
TPCs	23	24	34
SMs	46	48	68
CUDA Cores / SM	64	64	128
CUDA Cores / GPU	2944	3072	8704
Tensor Cores / SM	8 (2nd Gen)	8 (2nd Gen)	4 (3rd Gen)
Tensor Cores / GPU	368	384 (2nd Gen)	272 (3rd Gen)
RT Cores	46 (1st Gen)	48 (1st Gen)	68 (2nd Gen)
GPU Boost Clock (MHz)	1800	1815	1710
Peak FP32 TFLOPS (non-Tensor)[1]	10.6	11.2	29.8
Peak FP16 TFLOPS (non-Tensor)[1]	21.2	22.3	29.8
Peak BF16 TFLOPS (non-Tensor)[1]	NA	NA	29.8
Peak INT32 TOPS (non-Tensor)[1,3]	10.6	11.2	14.9
Peak FP16 Tensor TFLOPS with FP16 Accumulate[1]	84.8	89.2	119/238[2]
Peak FP16 Tensor TFLOPS with FP32 Accumulate[1]	42.4	44.6	59.5/119[2]
Peak BF16 Tensor TFLOPS with FP32 Accumulate[1]	NA	NA	59.5/119[2]
Peak TF32 Tensor TFLOPS[1]	NA	NA	29.8/59.5[2]
Peak INT8 Tensor TOPS[1]	169.6	178.4	238/476[2]
Peak INT4 Tensor TOPS[1]	339.1	356.8	476/952[2]
Frame Buffer Memory Size and Type	8192 MB GDDR6	8192 MB GDDR6	10240 MB GDDR6X
Memory Interface	256-bit	256-bit	320-bit
Memory Clock (Data Rate)	14 Gbps	15.5 Gbps	19 Gbps
Memory Bandwidth	448 GB/sec	496 GB/sec	760 GB/sec
ROPs	64	64	96
Pixel Fill-rate (Gigapixels/sec)	115.2	116.2	164.2
Texture Units	184	192	272
Texel Fill-rate (Gigatexels/sec)	331.2	348.5	465
L1 Data Cache/Shared Memory	4416 KB	4608 KB	8704 KB

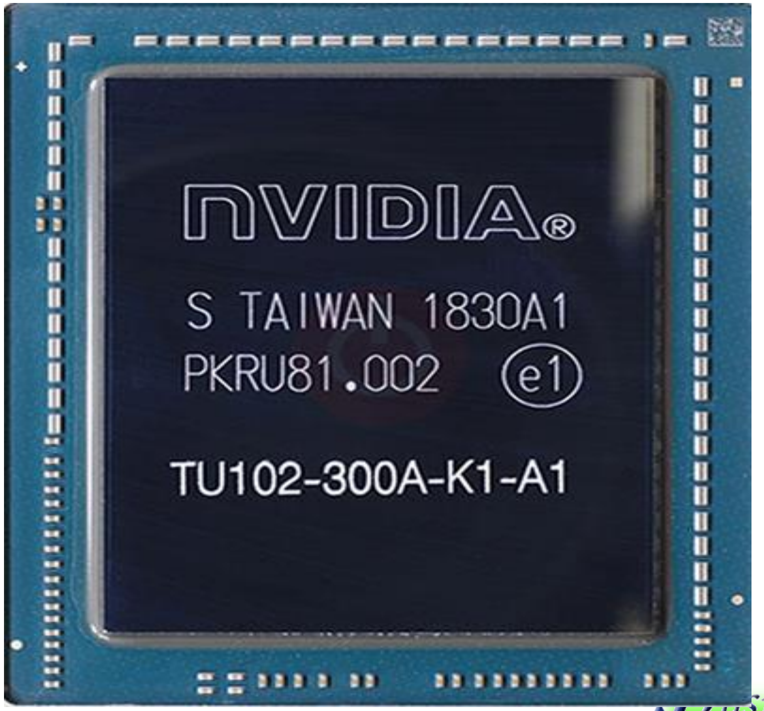
CUDA, RT, and tensor cores are all cross-generational so a direct counting comparison is irrelevant

The improvement are not linear since there is also a change in efficiency.

Naming scheme

- TU: Turing architecture
- GA: Ampere architecture.
- 100: Data center, Professional class
- 102: High end
- 104: Mid end
- 106: Low end





L1 Cache

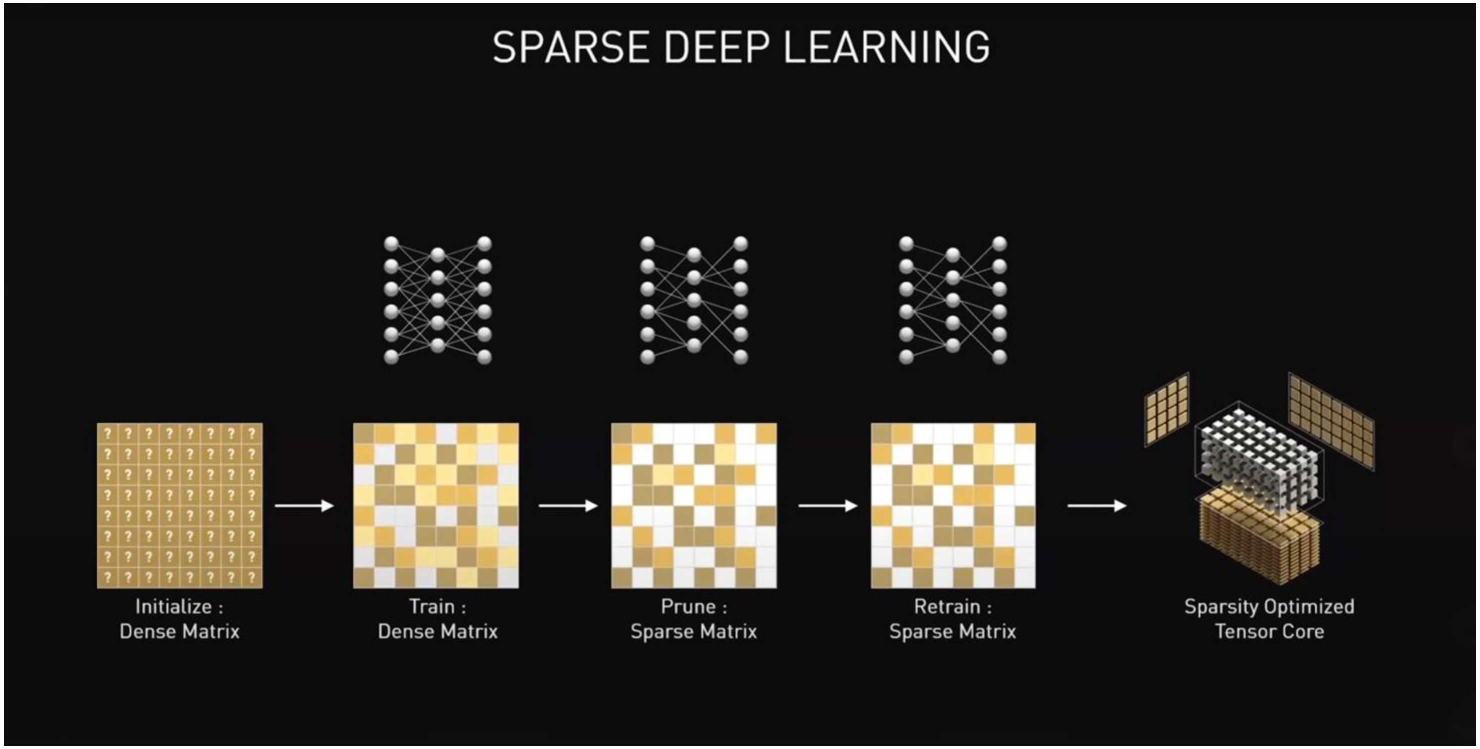
- Higher number of CUDA cores(128 FP32) → redesign data pipeline
- Double bandwidth
- From 98KB to 128KB per SM
- Twice the size partition cache
- 1.7 performance increase

Tensor Cores (2nd gen vs 3rd gen)

- In charge of the AI part
- Really good at linear algebra
- Important for DLLS
- 2nd gen cores use: dense matrices
- 3rd gen cores uses: dense and sparse matrices

Dense and Sparse Matrices

- **Strong connection are kept and weak connections are ignored with the objective of avoiding math that is not needed**

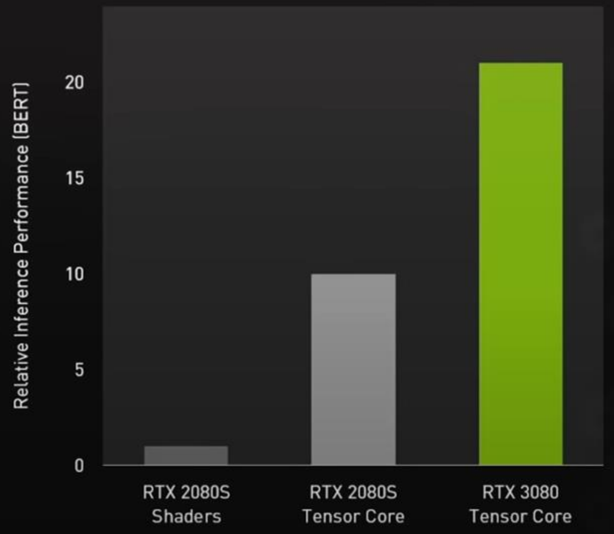


Tensor Cores improvement

- The use of sparse matrices increase the core performance
- Look at FP16 FMA(fused multiply-add) operations

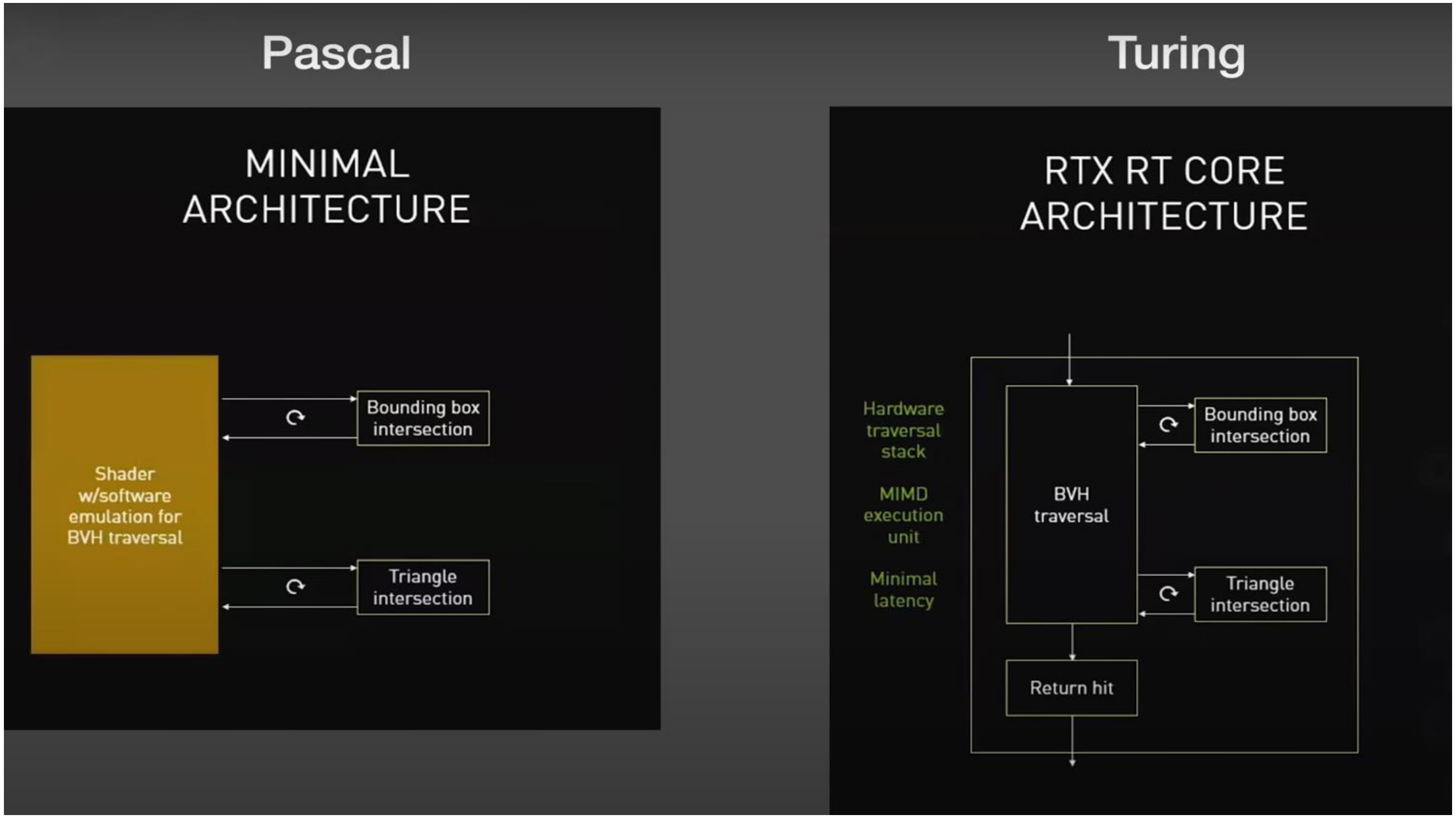
AMPERE 3RD GEN TENSOR CORE

	TU102 SM (RTX 2080S)	GA100 SM (A100)	GA102 SM (RTX 3080)
Tensor Cores per SM	8	4	4
FP16 FMA operations per Tensor Core	64	Dense: 256 Sparse : 512	Dense: 128 Sparse : 256
Total FP16 FMA operations per SM	512	Dense : 1024 Sparse : 2048	Dense : 512 Sparse : 1024



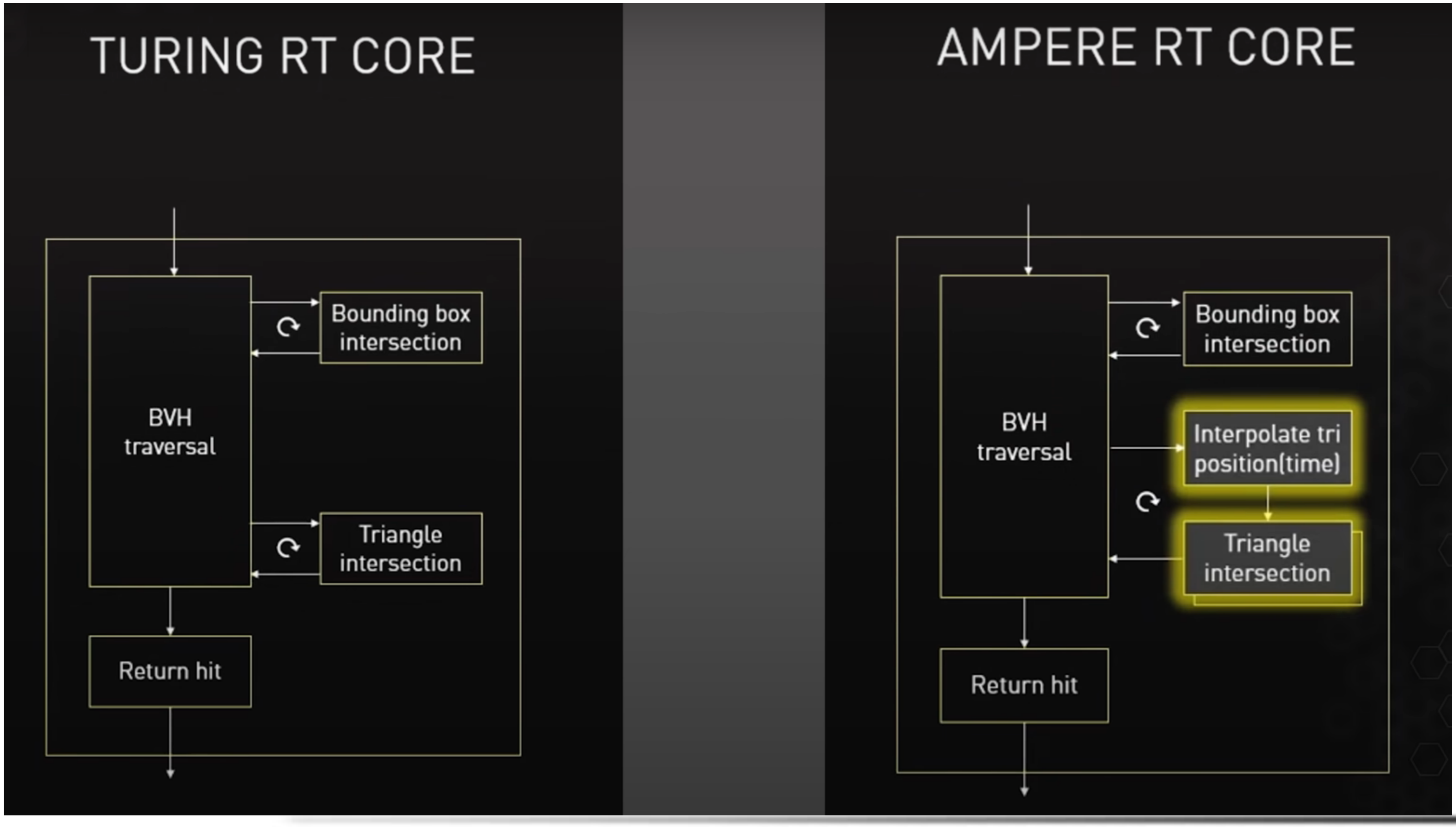
RT Cores(1st gen vs 2nd gen)

- **Main purpose of RT cores is to solve intersection problems (RT calculation)**



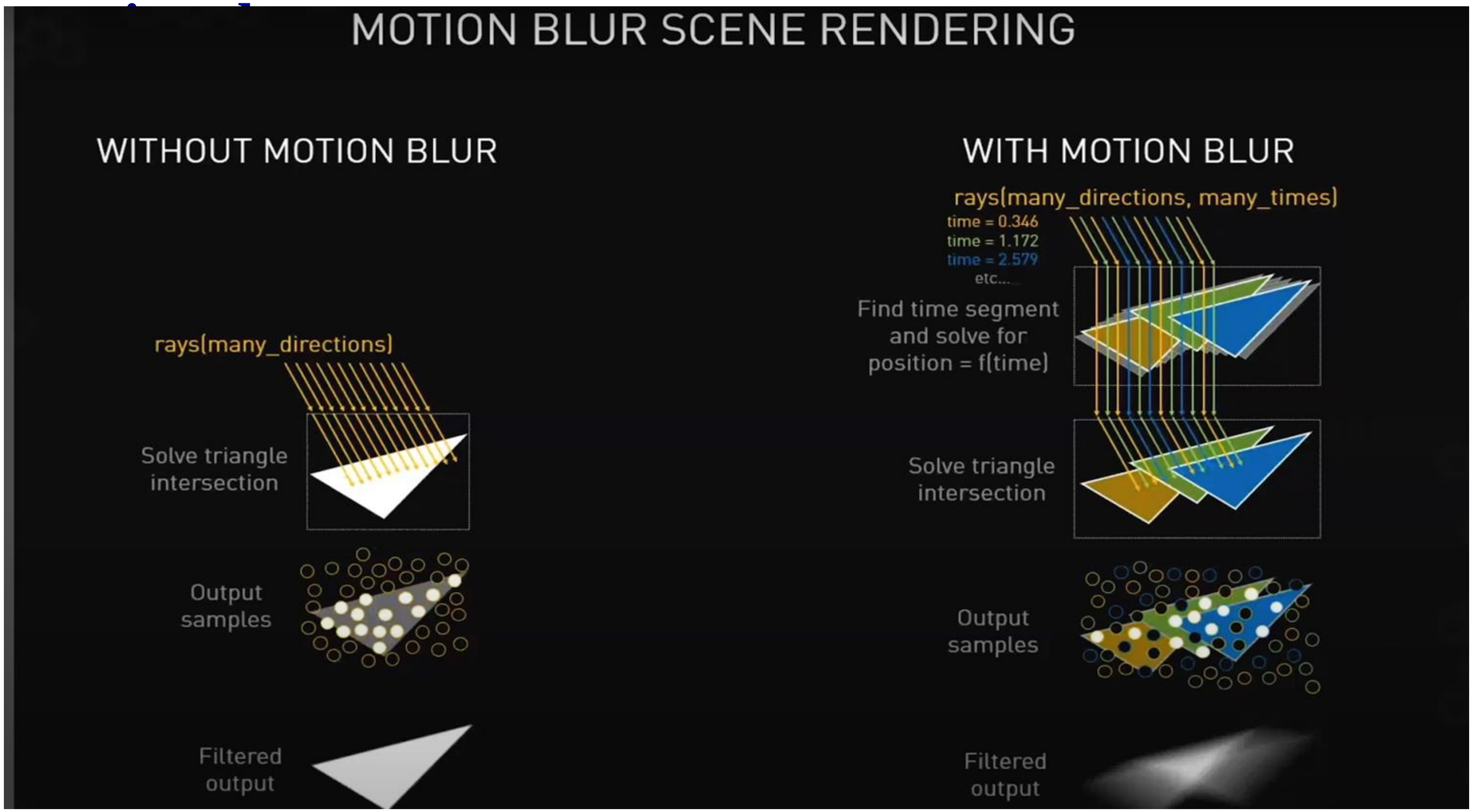
Ampere RT Cores

- Improvement in triangle intersection and new interpolated tri position(time)(for motion blur)



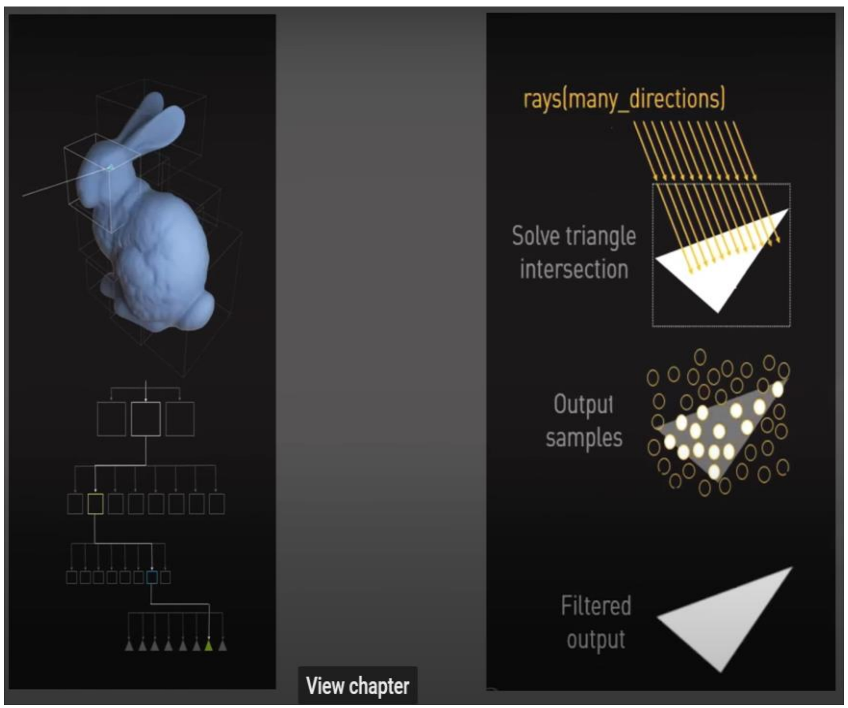
How to implement motion blur?

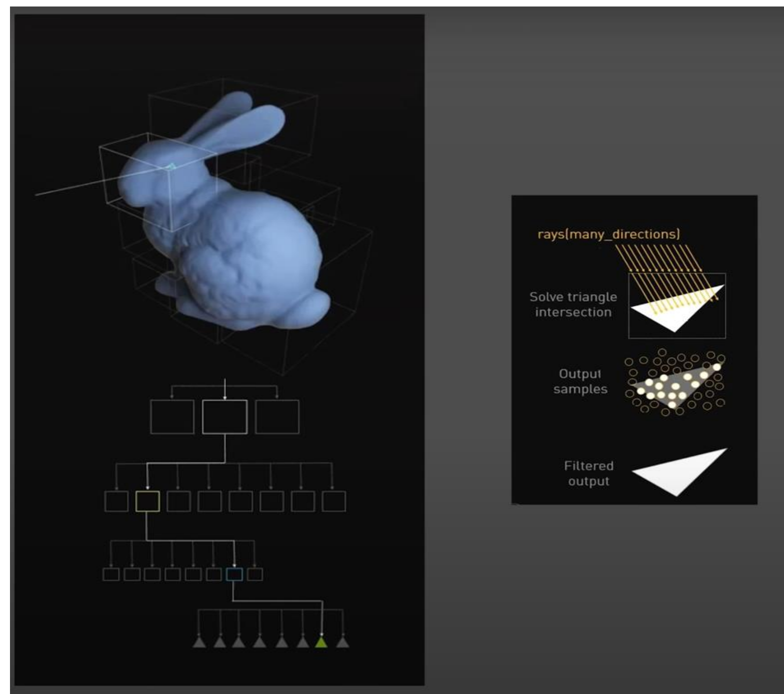
- Now ray tracing has a time variable which helps detect position. Each ray touches a different version of the



Why improve triangle intersection unit?

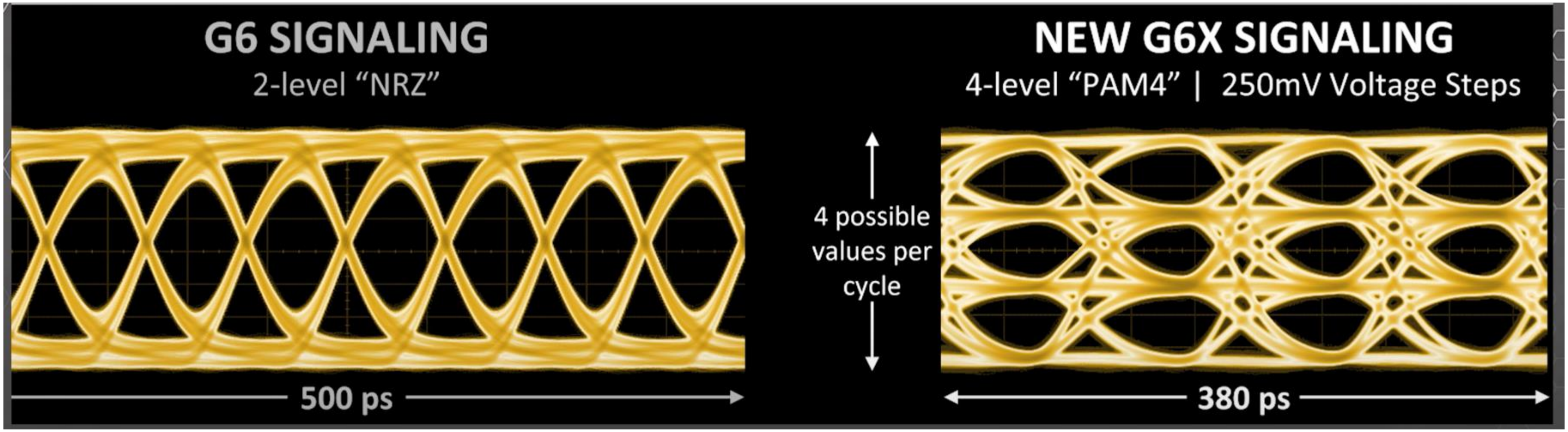
- **Ideally in a RT core the bounding box and triangle intersection should be working in parallel, but the rates of the triangle intersection were too small, affecting the work flow negatively**





New GDDR6x

- Reduction in memory box
- 22% improvement in memory bandwidth
- Pam4 allows 4 cycles at the time, allowing twice as much data to be transmitted.



List of Related Companies(market value

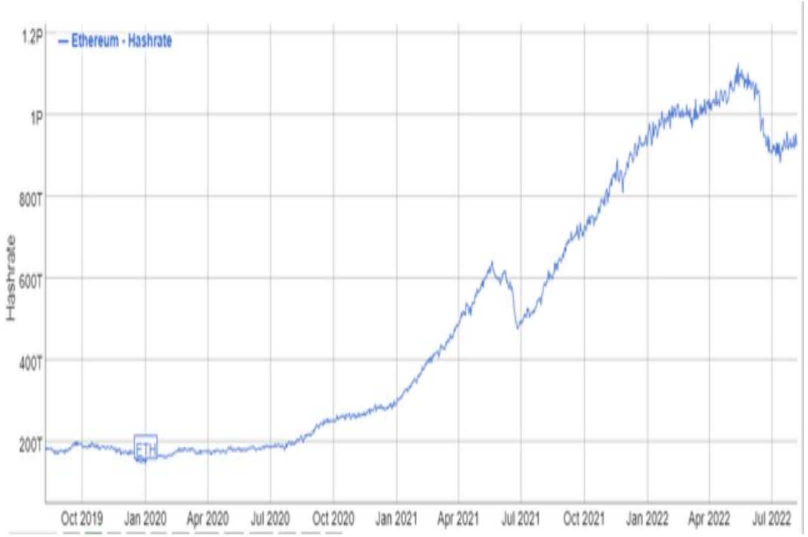
- Nvidia (361 billion)
- AMD (136 billion)
- Intel (130 billion)
- ASUS (6.16 billion)
- Apple (2.6 trillion)
- GIGABYTE (1.9 billion)
- ZOTAC (326 million)
- EVGA (120 billion)
- Sapphire(26 billion)

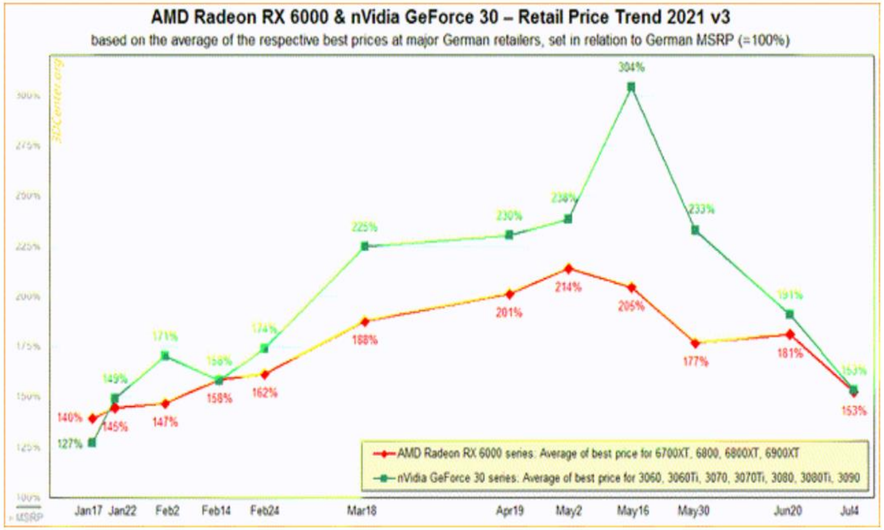
SWOT

- Strength:
- Cool features such as: RT, DLSS
- Intellectual rights property
- Strong financial position (due to in recent years people have been staying at home)
- Weakness:
- Extremely expensive
- AMD according to people has better low to mid-range GPUs
- High employee turnover ratio (too many employees leave)
- Opportunities:
- The rise in interest in the gaming sector (PS5, Xbox need high end chipsets)
- Cryptocurrency
- Threat:
- Not enough skilled workers in the market
- Intense competition

Future trends part 1

- Increment in gaming community
- Crypto-mining makes GPU be in demand and have high prices
- Increase in hash rate in network means more GPU in





Future trends part 2

- Ethereum switch away from GPU-based mining will permanently remove over $10 billion in demand for GPUs.(switch to coin ownership)

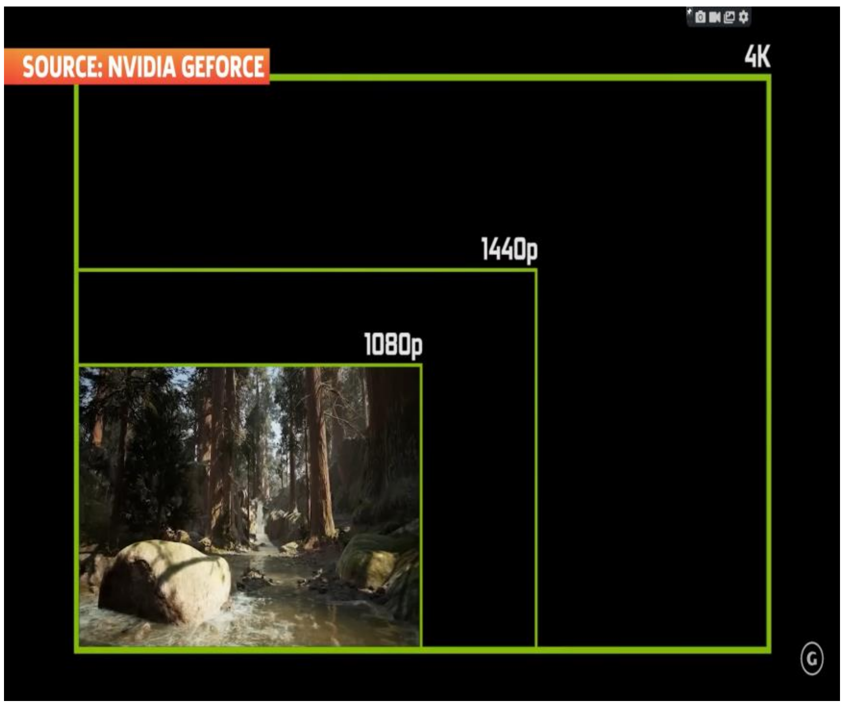
- Some studies estimate Nvidia's revenues, margins, and profits are all going to take a dive.

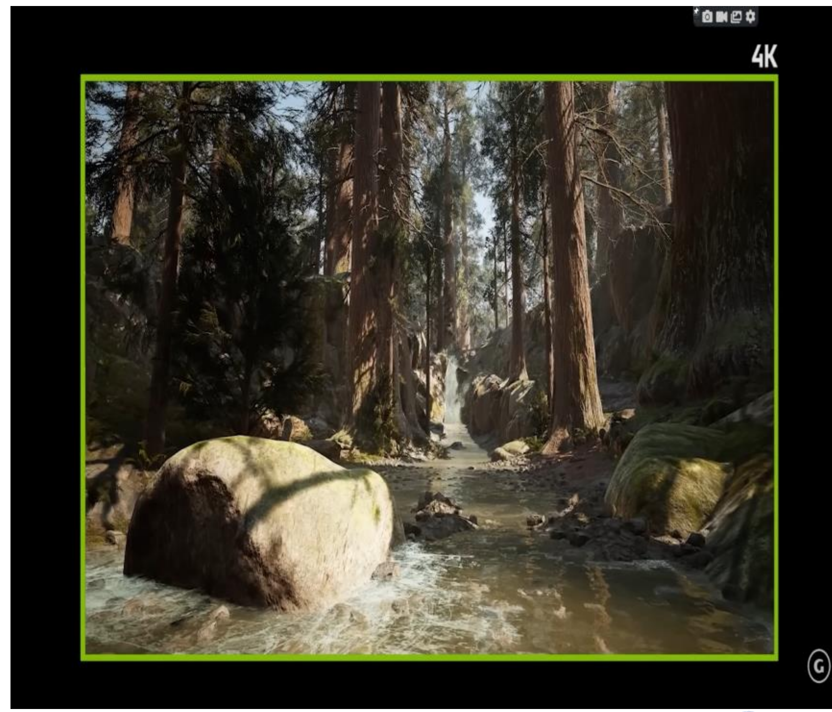
References(most info from whitepaper)

- https://history-computer.com/largest-gpu-companies-in-the-world-and-what-they-do/
- https://www.digitaltrends.com/computing/what-is-ray-tracing/
- https://zh.wikipedia.org/wiki/NVIDIA_GeForce_256
- https://blog.paperspace.com/understanding-tensor-cores/
- https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf
- https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units#GeForce_30_series
- https://seekingalpha.com/article/4536320-nvidia-10b-gpu-demand-may-be-gone-permanently

Explaining DLSS(optional)

- By analyzing a high resolution image, it uses AI to create frames that don't required high processing power.
- #FPS(native resolution 1080p) close to #FPS(dlss on 4k)





Another way to use DLSS(optional)

- Reduce the amount of native frames processed in a fixed resolution. #FPS(DLSS 4k) > #FPS(Native 4k)
- Same quality with lower base resolution