NVIDIA GeForce 30 Series

SoC Individual Project Report

By Ivan Andre Castillo Barahona

Student ID:108061181

Major: Electrical Engineering

14/11/2022

Outline

Contents

1. Introduction

GPUs or also known as graphics processing units have been around for quite a while. Since their appearance in the late 90s they have come to stay and have quickly become an essential part of our computers. Just recently due to the growth of the gaming community over the last few years GPUs have not only grown in their technological capabilities, but also in popularity. Nowadays, when building a desktop or buying a new laptop it is a must for us, the consumers, to be familiarize with at least the most important and basic components of our computer. This of course includes the GPU, which as we know will be in charge of the graphics and video rendering in any device.

MSI take on GeForce RTX

Without GPUs a lot of the functions in our computers that we take as given would not be possible, since almost every action requires at the minimum a bit a graphical power.

Even though the main topic of this report is NVIDIA's new GeForce 30 series GPU and the key objective being letting the reader learn all that he/she may need to know about this new NVIDIA's GPU, we will first talk about all of the functions and important application a GPU can have. And by these we will get to know what were the motivation for the creation of this device in the first place.

2.Application

We have said that a GPU is a very important part of our computer and to better understand why we will now discuss its applications. A good GPU will bring a high performance in functions such as:

Video editing (or video encoding), 3D graphics rendering (which is directly related to gaming, films, and digital display art), cryptocurrency (by creating blocks in any cryptocurrency blockchain), machine learning, streaming.

All of these things will be highly benefitted when paired with a fast, efficient and capable GPU.

GPUs used in crypto mining

Now let's introduce the most famous features NVIDIA's GPUs are known for and also focus on their applications. This includes: high resolution gaming, Ray Tracing, and DLSS (deep learning super sampling).

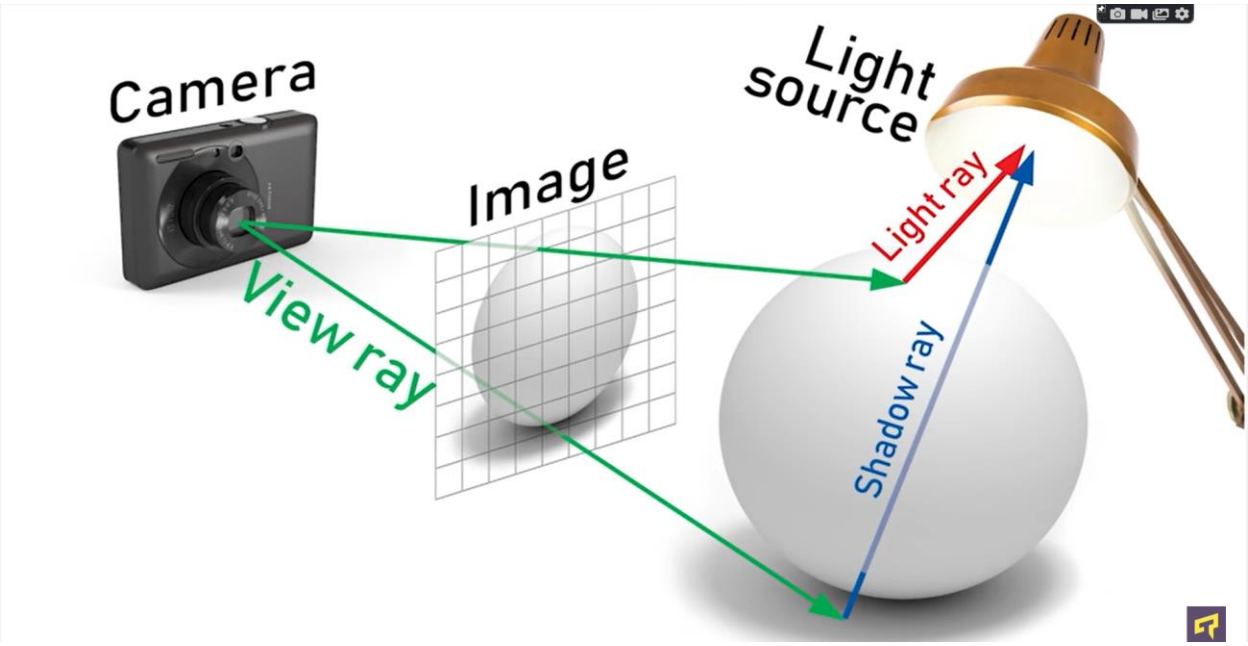
We know Ray tracing is capable of simulating a variety of optical effects, such as reflection, refraction, and shadows. Since RT help us know in a really accurate way how light travels in an image, this will improve a lot the qualities of the shadows and therefore creating more realistic graphics.

And when talking about DLSS we can say it helps increase the frames per second without changing the resolution of an image.

RDR 2 w/o DLSS

3.Ray Tracing and DLSS

When explaining ray tracing there is a useful example to help us understand how it works. Let's say we are in a room and we want to know how light travels from a given point. So, we throw a ray from that given point (let's say from a camera), we will see that this ray bounces in the walls or with other obstacles that are in this room. Seeing how this ray bounces can really help us simulate how light will travel in the room, which will give us an idea on how shadows will realistically look in this same room.



In this picture we can see that by throwing a ray from our camera and by having a light source, we will know our shadow and light ray direction after colliding with our sphere.

The other good feature NVIDIA has is DLSS. DLSS uses deep learning to analyze high resolution images and uses AI to learn how to create extra pixels to get a high-resolution image that does not require a high processing power. For example, if our base resolution is 1080p, we can use DLSS to convert it to 4k. Of course, this means that there is only 1080p of native pixels and the rest are all the ones that DLSS created. This new image will require way less processing power than a native 4k image and therefore it will improve the frames per second for the same resolution. We can infer that this new DLSS created pixels will not be 100% accurate but as the AI algorithm improves it will mean more precision and better quality.

If we have an image in 4k but we desire more frames per second then instead of processing all of the native pixels in a 4k resolution, we can use DLSS to decrease the native base resolution and fill in the rest with the ones created by DLSS (in this

example we use 1080p base resolution plus DLSS created pixels which will add up to 4k instead of using a full 4k native, this will use less processing power). And as I said before, the more the DLSS algorithm improve it will start making the images more accurate to the point where the difference between a native image and an image with DLSS in the same resolution will be almost unnoticeable.

4.Motivation and System Evolution

What motivated the creations of GPUs and why do we need them so much? Basically, GPUs are needed since CPU are not optimal for graphics rendering, this created a need for smooth graphics in videos and 3D gaming. This made NVIDIA in 1999 create their first GPU the “GeForce 256”. Using TSMC 220nm process, with around 17 million transistors and with a size of 32mb the “GeForce 256” was the first ever GPU in the market.

Compare that with one of the latest NVIDIA’s GPU the “GeForce 3090” which uses a Samsung 8nm process, has around 28 billion transistors and with a 24GB of G6x memory. We can say in this last 20 years there have been a lot of changes.

5. 20 series (Turing) vs 30 series (Ampere) change in process

One of the most important parts of this report is to compare the differences between GeForce 20 series (Turing architecture) and GeForce 30 series (Ampere architecture). Before talking about their architecture, I wanted to talk about one of the most notable changes, which is the big improvement in the process.

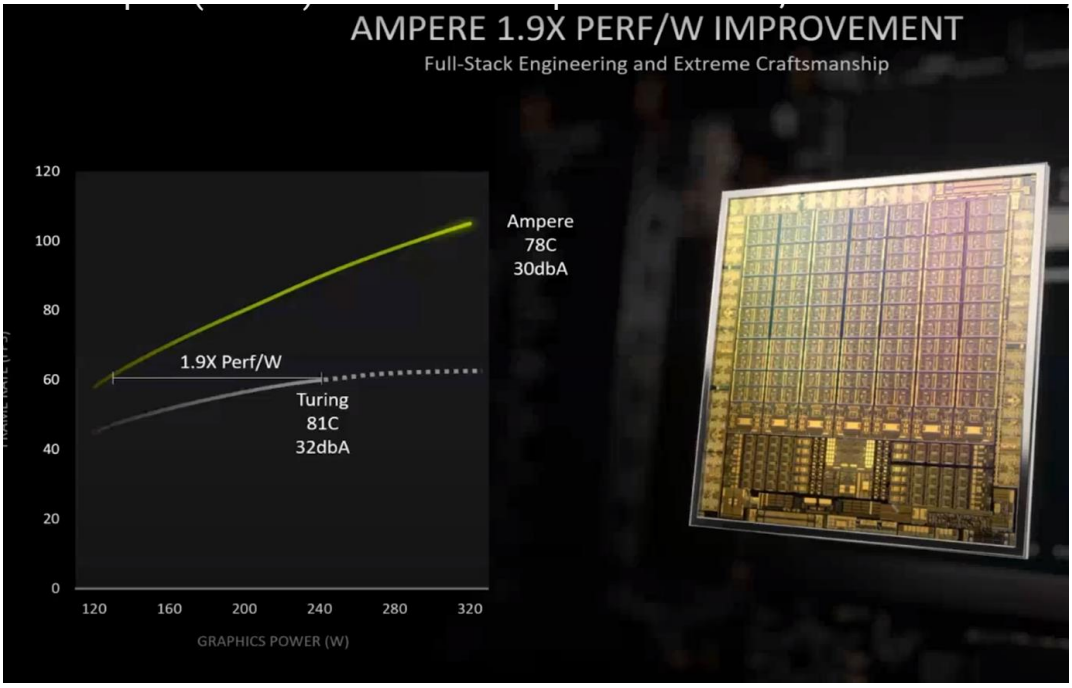
The GeForce 20 series uses TSMC 12nm process, but the 30 series GPU was design with Samsung's 8nm process, which of course this increases the transistor density.

Take a look at the die size of both 20 series and 30 series GPU with specifications. But an important point to look in here is that often when there is a reduction in the process leads to also a reduction on the power consumption. But in this case, it is not like that since the power consumption (TGP) in ampere (30 series architecture) is greater than that of the Turing's (20 series architecture). But actually, this is not a downgrade because the performance per watt is greater with ampere.

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

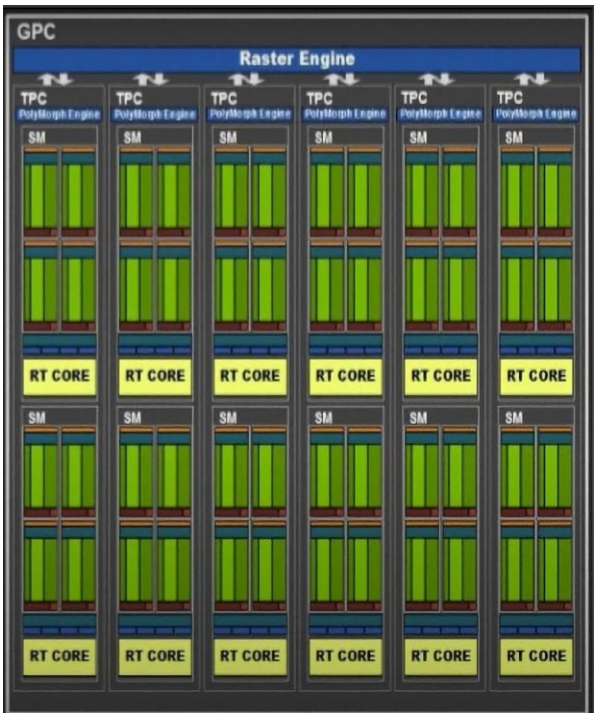
If we use FPS as a performance indicator, we can see that there is a 1.9 performance per watt improvement. Which means that with ampere we can get better FPS with less power.

Also in this picture, it is important to notice that the power consumption does not actually equals and increment in temperature, it is never this simple. As you can see when ampere uses its max power of 320 watts the temperature is around 78c, but Turing reaches the same temperatures at 240 watts.



6. Turing's Block Diagram

Turing' block diagram consists on GPCs (graphics processing cluster) and as we can see we have 6 of them.



If we take a closer look, a GPC has 6 TPC (texture processing cluster) and each TPC contain 2 SM (Streaming multiprocessor). So there is actually 12 SM in the whole GPC, if we look at the bigger picture. In here we can see that the SMs contain CUDA, Tensor and RT cores, and it also contains cache. Which we are going to talk about in a bit.

7. Ampere's GA-102 Block Diagram

For this exact version of the ampere architecture there are 7 GPC. And each GPC architecture is really similar to Turing's GPCs. The big change is in the generation leap between the RT, tensor and CUDA cores.


GPC
Raster Engine
TPC
PolyMorph Engine
SM
RT CORE
SM
L0 i-Cache + Warp Scheduler + Dispatch (32 thread/clk)
Register File (16,384 x 32-bit)
FP32 / INT32
FP32
TENSOR CORE 3rd Gen
LD/ST
SFU
128KB L1 Data Cache / Shared Memory
Tex
RT CORE
2nd Generation

8.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

I know there is a lot of information here but let's focus on the cores. It is smart to not compare the number of CUDA, Tensor, or RT cores directly since all of this are cross generational which means that their improvement is not linear. The difference in these cores has more to do with efficiency instead of their number, which makes a direct counting comparison irrelevant. So, it's not good to compare who has more cores and use this as an improvement indicator.

9.Naming Scheme

Let’s Analyze the ID in a die:

TU: Turing architecture

GA: Ampere architecture

The numbers beside are usually to identify is the product is high, mid or low end.

100: Data center, Professional class

102: High end

104: Mid end

106: Low end

We can see that as it is right now the lower the number the better the components are. Lower number means bigger silicon, which means is closer to the biggest potential of the block diagram we talked before.

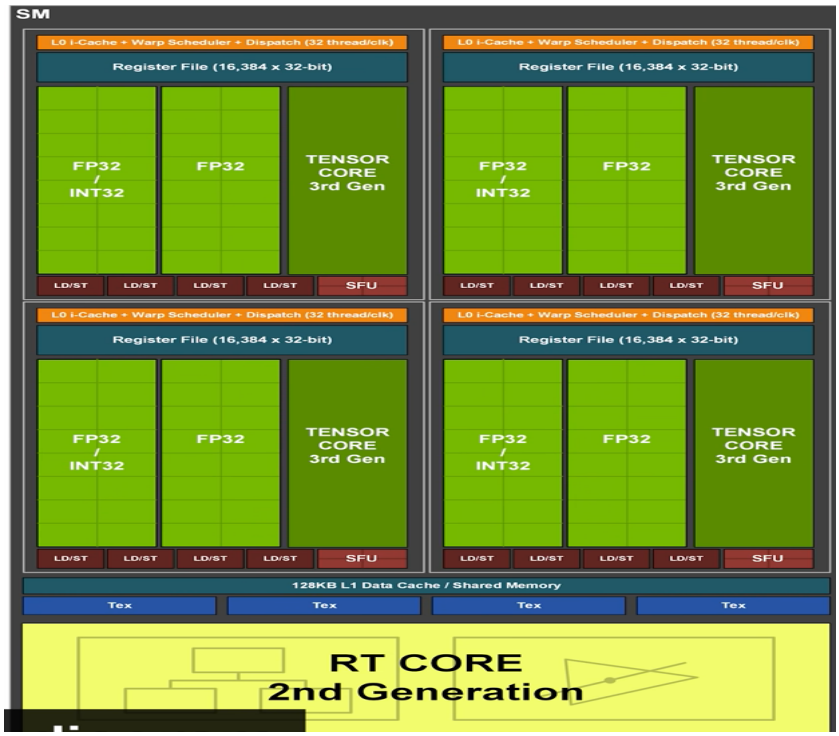
10.L1 Cache

Since for ampere architecture the number of CUDA cores has increase to 128 FP32 operations per cycle execution of the SM, the whole data pipeline needed to be redesign. This meaning that the L1 cache has now double the bandwidth. The L1 data cache now increased from 96kb to 128kb per SM, which is a 33% increase from Turing. Also, it has twice the size partition cache. So, in conclusion the new L1 data cache has a 1.7 increase in performance, which according to NVIDIA will help with does long and complicated shader programs.

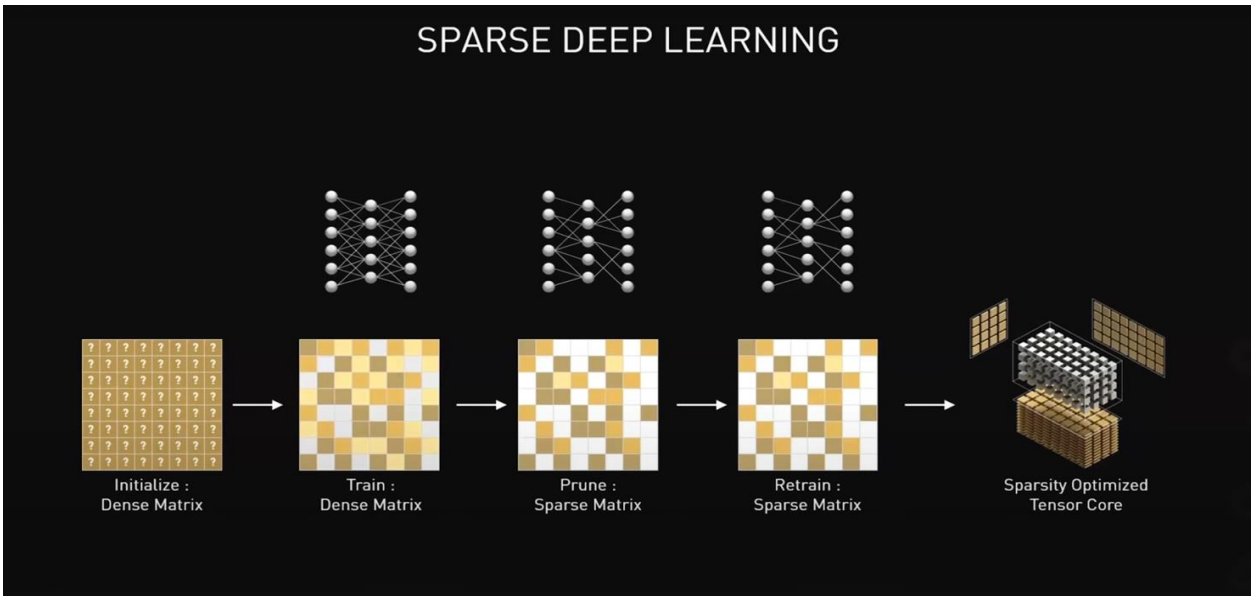
11.Tensor Cores

Now let's talk about the new tensor cores. These cores are in charge of AI part of this GPU. Which means they are really good at making Linear algebra. These cores are mainly the ones that allow us to have really good FPS at a really good resolution (DLSS depends on this cores). So, what's the difference between Turing 2^{nd} gen cores and amperes 3^{rd} gen cores?

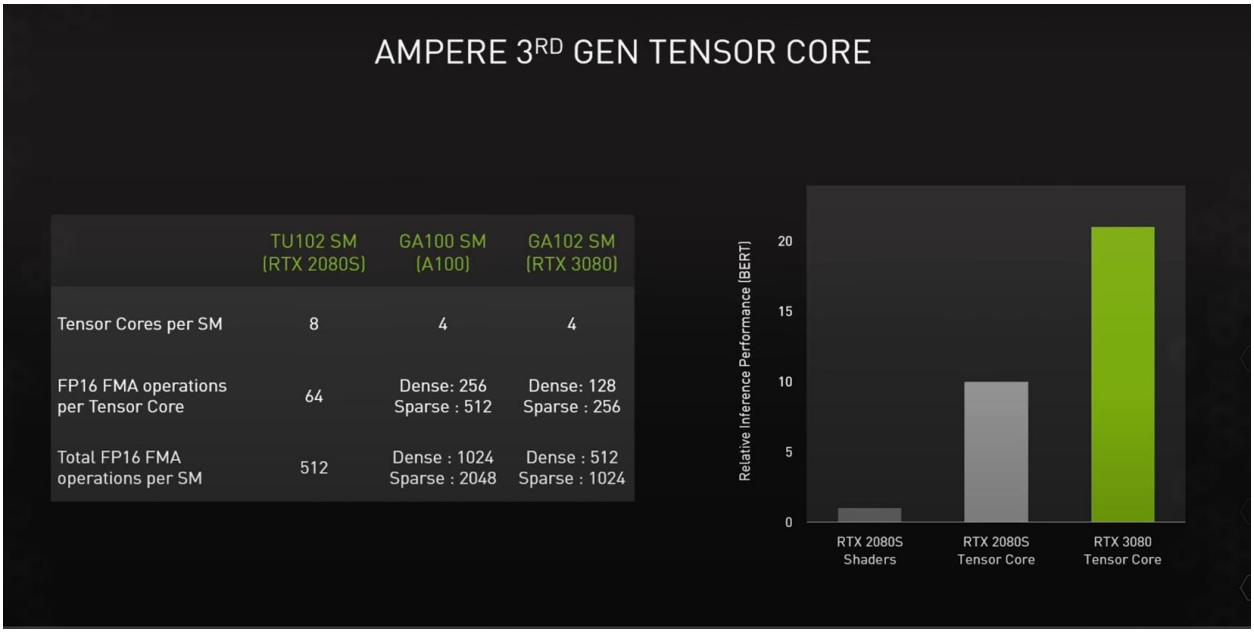
In simple words, 2^{nd} gen tensor cores use only dense matrices for AI application while the 3^{rd} gen uses sparse and dense matrices.



3rd gen tensor cores have been trained to use sparse matrices. This works by assigning “weights” in a matrix and the dropping the ones that have lower “weight” which will make this matrix prone to become a sparse matrix at the cost of some accuracy in actual values. Another way to see it is by creating nodes in a matrix and ignoring the ones that are consider “weak nodes”. After this process is done, our sparse matrix will be compressed and the tensor cores will use this simpler compressed sparse matrix for AI application. This boosts the performance of each tensor core but also reduces a little bit its accuracy. But, this small decrease in precision is worth it when comparing it with the big increment in performance. In conclusion, sparse matrices are nearly as accurate as dense matrices and using them is beneficial for tensor cores efficiency.

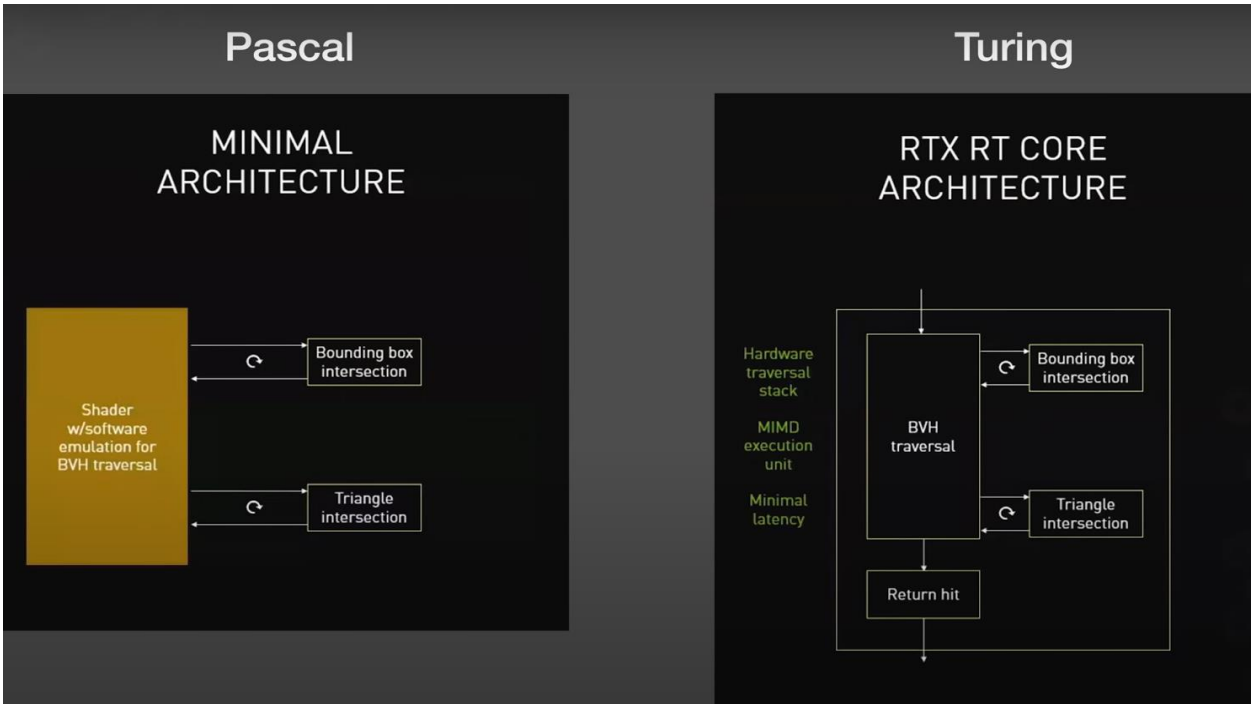


In these graphs we can see that though there are less tensor cores per SM in ampere, each core is way better than the ones in Turing. When performing dense matrix operations, they are the same, but when we get a case where we can use a sparse matrix, that is when the new tensor cores shine, by avoiding math we just don't need to do.

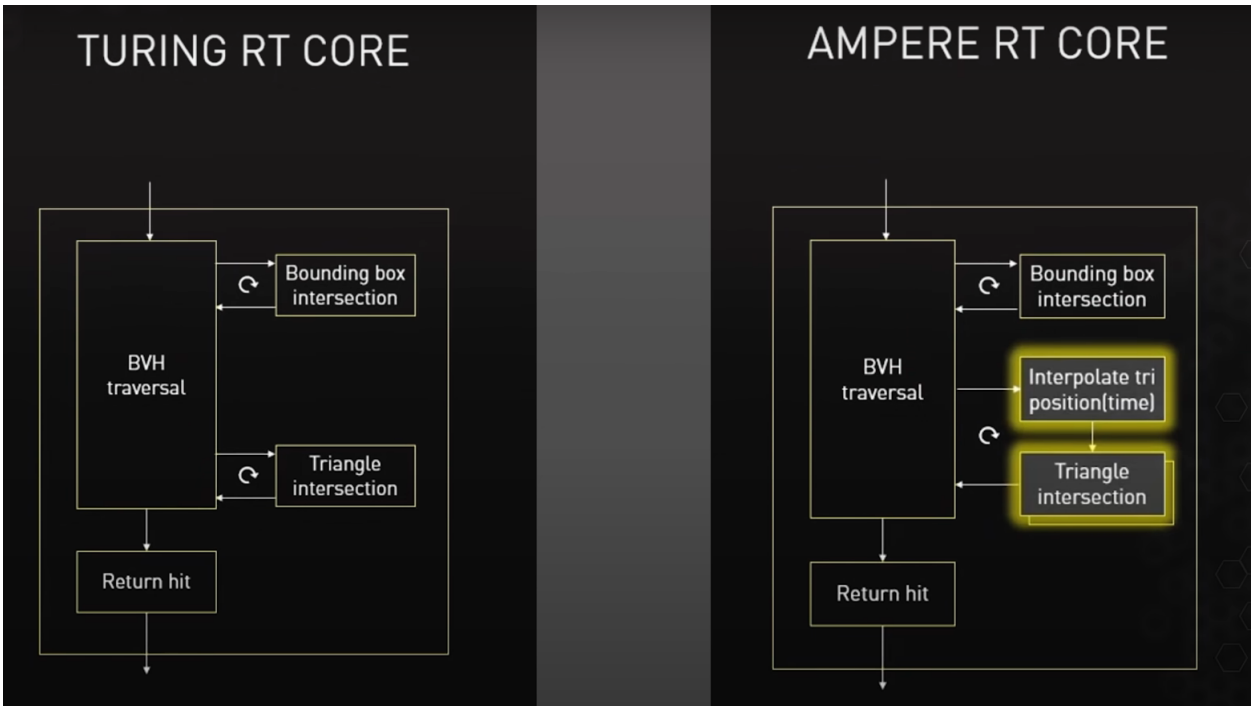


12. RT Cores

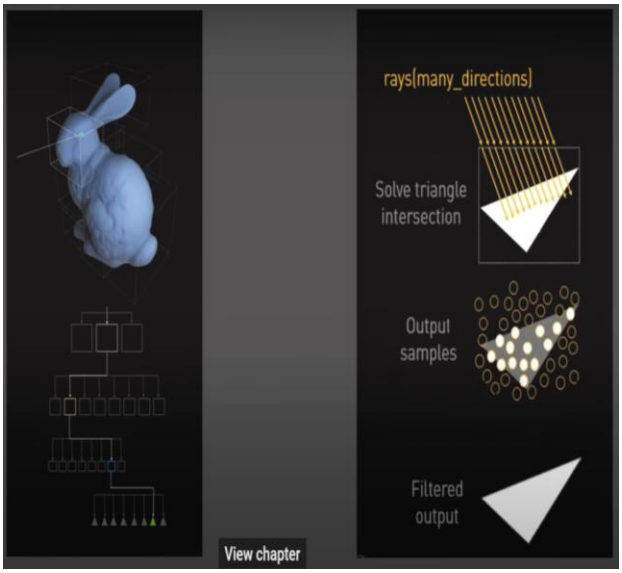
So, For Turing's RT cores their main purpose was to solve intersection problems, RT cores most important job is to be an independent hardware unit that handle all the ray tracing calculations that come from the shaders core. This shader core does not do any RT calculations. The rt core inside has 2 math units: bounding boxes intersection, triangle intersection. The return hit is used to return results to the shader. The big box named as BVH (Bounding volume hierarchy) traversal is actually an MIMD (multiple instruction multiple data) execution unit. This BHV are basically used to eliminate potential intersections in a scene by removing geometrical objects which are not intersected by the current ray.

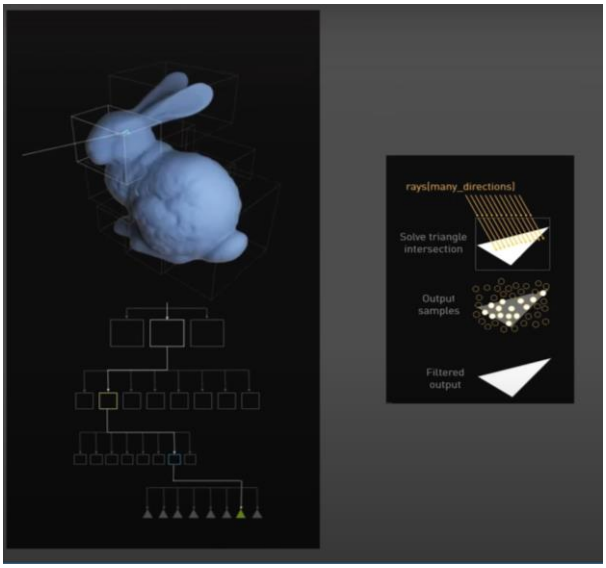


Ampere 2nd gen RT cores have some improvement in the triangle intersection unit and created a new interpolated tri position, which is used primarily for motion blur.



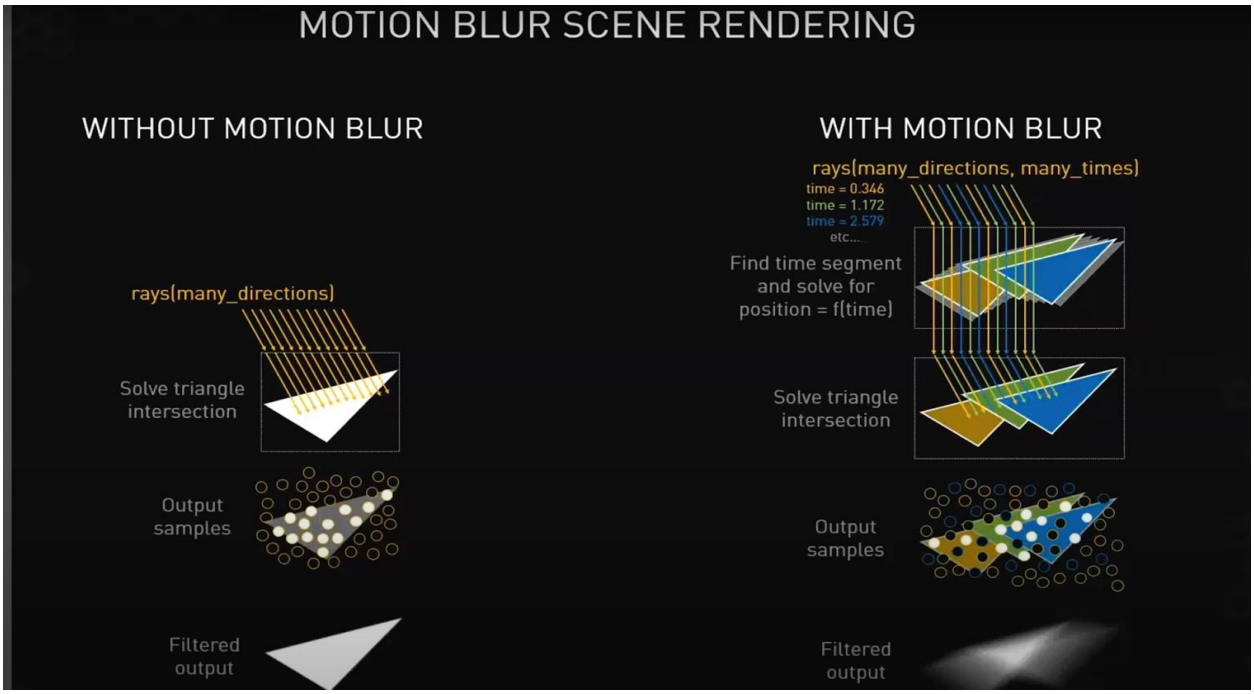
Now, the reason for the improvement of the triangle intersection unit is that ideally in a rt core, the bounding box and triangle intersection should be working in parallel, but while testing these two units in the 1st gen RT core, they discovered that the rates of the triangle intersection unit were too small, creating a bottleneck in the flow of work. This problem of course was solved in this new gen by improving the triangles intersection rates.


rays(many_directions)
Solve triangle intersection
Output samples
Filtered output
View chapter


rays(many_directions)
Solve triangle intersection
Output samples
Filtered output

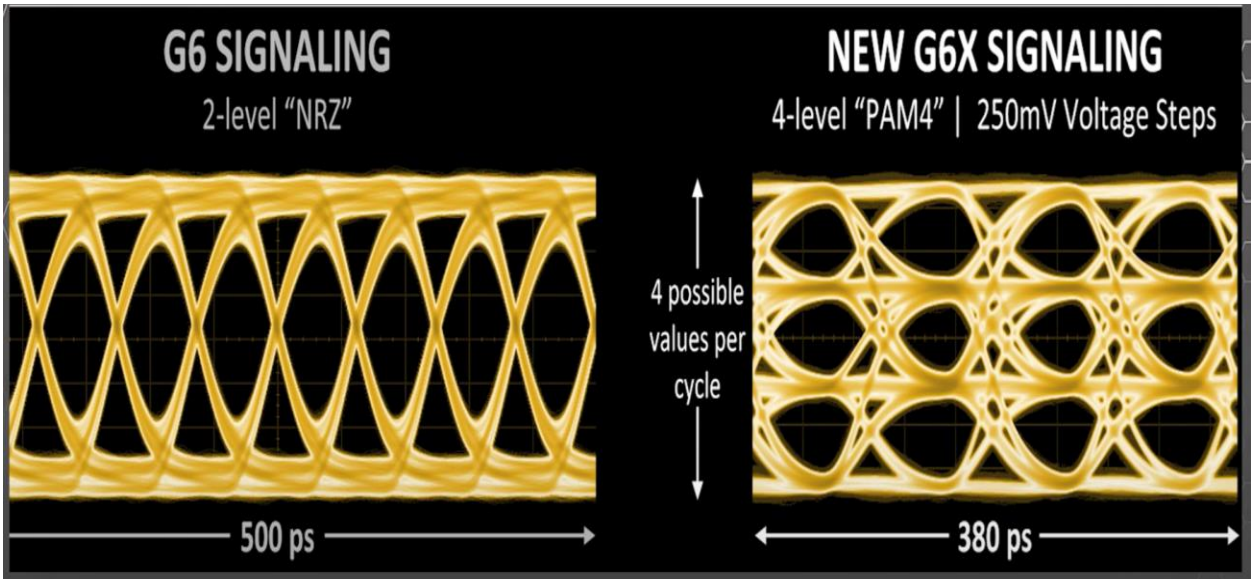
13. Motion Blur

How was motion blur implemented? By interpolating the triangle position in a given point in time. We know that ray tracing usually works with a fixed position, but for motion blur we do not know the exact position of this, and this is where this unit shines. How it works is that now the ray has a time variable that helps detect the position. It is fair to say that there is more than one ray for a given scene and each ray will hit a different triangle, of course, based on its function of time).



14.GDDR6x

Let's talk about the new memory GDDR6x: for this memory it is really needed to have a high power to unlock its full potential of 90 gigabits per second memory bandwidth. That why the increment in power for the ampere architecture is really suited for this memory. Compare to the older version, there is a reduction of size of the memory box, with a 22% improvement of the memory bandwidth, this is because of the new "pam4" which allows 4 cycles at the time. PAM4 sends 1 of the 4 different voltage levels, in 250 mV voltage steps, every clock cycle. This allows the new memory to transmit twice as much data.



15. Related Companies

- **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)**

Just for reference here is a list of the biggest GPUs companies in the world. Note that the number beside them is its market value of the whole company, not only their GPUs. But if you want to see this list in terms of their revenue only in their GPUs, then is already in order (NVIDIA being the best all the way to Sapphire being the worst in this list)

16.SWOT

Strength:

Cool features such as: RT, DLSS

Intellectual rights property (IPs)

Strong financial position (due to in recent years people have been staying at home and the gaming community has grown stronger)

Weakness:

Extremely expensive

according to people AMD 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:

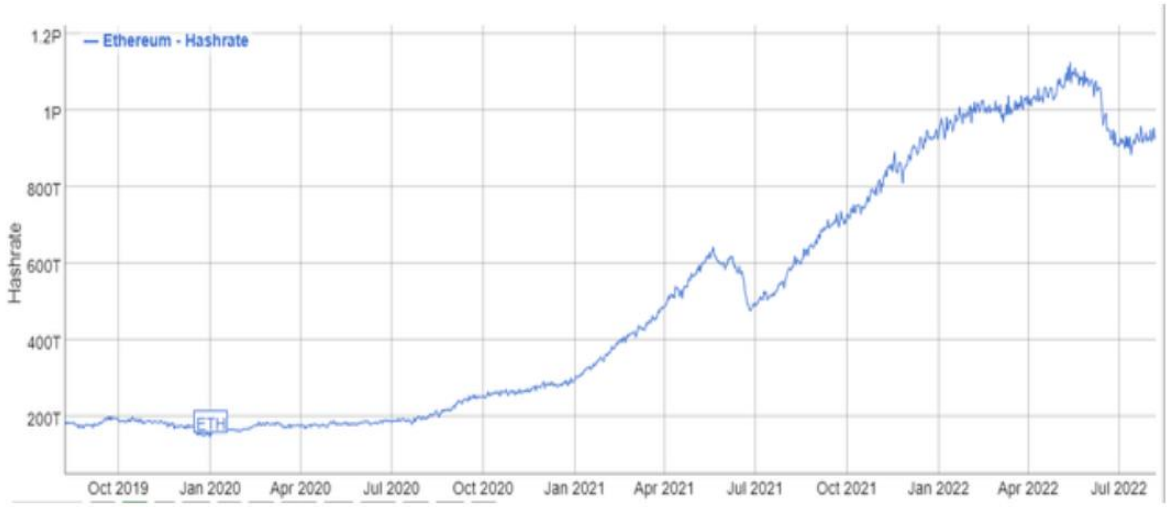
Not enough skilled workers in the market

Intense competition

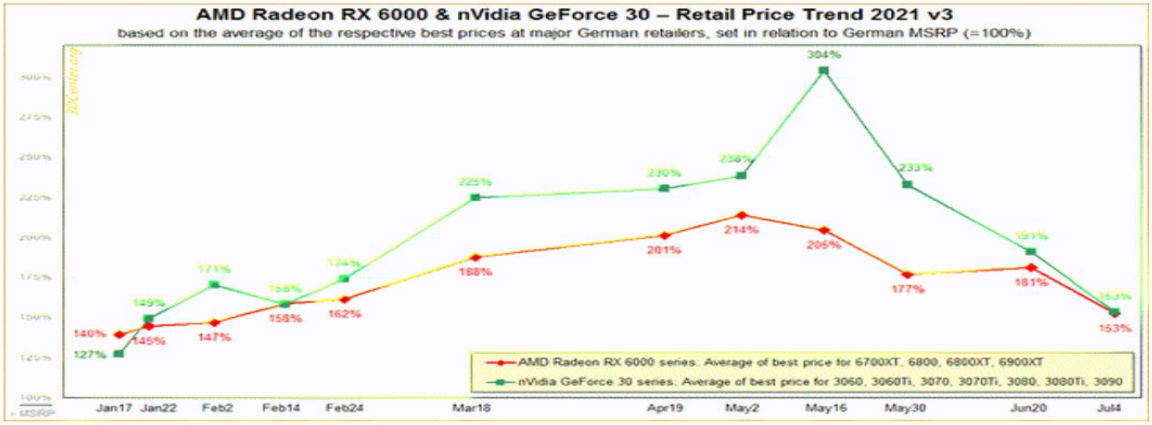
17.Future trends

To analyze the future of this company we must first look at its current situation and try our best to predict what will happen. So, first things we see are:

*Increase in hash rate in network means more GPU in use.



*The demand of mining has driven GPU prices to the moon.



More GPUs are in demand, plus their prices are really high. It seems that now NVIDIA is in a really good situation.

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

*Nvidia's revenues, margins, and profits are all going to take a dive (of course this by no means indicate that NVIDIA will go bankrupt, but we can at least expect their growth to be affected).

19.Conclusion/Knowledge gained

Doing this report and presentation made me realize how important GPUs are in nowadays. By seeing their applications and functions I saw that they are basically use in almost every action in our computers. But definitely one of the most exciting and fun parts was learning all of the ways the new generation improves when compare to their predecessor. The incorporations of new technologies such as ray tracing and DLSS, or the use of sparse matrices in tensor cores are just a few examples of all of the exciting ideas engineers think of when brainstorming how to make that “generation leap” that everybody is waiting for. From now on I will always try to be up to date with the new inventions people have created

20.References

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