CSC 461 Final Project

What is NVIDIA Now?

NVIDIA's GeForce Now is the latest in cloud gaming. In short, cloud gaming is "a method of playing video games using remote servers in data centers" with no local download required (29). Games are "played" on a remote server with all the required gameplay and graphics calculations being done remotely, while all user interactions are still done locally on a user's device and are then sent over the network to be processed at the remote server. With this new form of gaming, several existing issues become even more relevant for a smooth gaming experience. Specifically, latency (poor quality and/or slow graphics) is a significant issue, whether that is from network congestion or from the processing and packaging of data, even more so than it has been with native gaming.

GeForce Now is the brand used by NVIDIA for its cloud gaming service. GeForce Now has 3 tiers, Free, Priority, and the new tier Ultimate. The Free tier being the least powerful and the Ultimate tier being the best in performance and quality. This report focuses on the frontier in cloud gaming offered by the Ultimate tier which includes RTX 4080 gaming rigs.

What is Cloud Gaming?

My interpretation of what cloud gaming is seeking to do, is that it is trying to separate the hardware from the gameplay. Anyone who plays video games is well aware that there are some problems that are the result of having local copies of games. Imagine that in a group of 3 friends there is 1 who plays on a PC, 1 who plays on an Xbox and the other who plays on a PlayStation. With most games, disregarding cross platform enabled games for the moment, they only have native support and so these friends cannot play the game together. With cloud gaming this problem has been eliminated as the data is being processed at a server and only the frames need to display on the screen and any game state data will be sent back, all of which is not hardware specific. This makes more games accessible for users across all systems.

Furthermore, modern games require even more computing power. Realistic graphics, advanced gameplay mechanics, large world, and huge multiplayer arenas are becoming the norm while still requiring the same real time experience. This leads to hardware needing to be powerful enough to support this, which is not always easy or financially viable to achieve, especially for a casual gamer. For example, Ark Survival 2 was just released in beta and many people are unable to run the game due to the extremely high standard of hardware requirements. This is another issue that cloud gaming helps to address. Since none of the gameplay processing is done locally, the need to fulfill hardware requirements fall onto the server's hardware instead of the user's hardware. Not only does this allow users to play the latest, computationally demanding games, but they do not need to be able to afford an expensive PC or console to do so, and instead just require the correct tier of NVIDIA Now.

NVIDIA RTX 4080 Graphics Cards

NVIDIA is known for their graphics cards. Starting from the 1000 series, up to the 4000 series and they are even working on the 6000 series now. The better versions of each are marked with Ti, as in RTX 3090 and RTX 3090 Ti. The earlier GPUs, GTX models, tend to not have tensor or RT cores, while the later series, RTX models and mainly anything after the 1000 series, include tensor and RT cores in their architecture. Intuitively, the 4000 series graphics cards are the best. Specifically, the NVIDIA RTX 4080 and 4090.

So what graphics card is being used in GeForce Now? Currently it is the 4080. Now some of you may be thinking "Why aren't they using the 4090? Isn't that the better model?", well yes, it is currently the best model of graphics card NVIDIA can offer and *ideally* it would be used. There was, however, a big and somewhat ironic problem NVIDIA encountered with the 4090 which is that they are too powerful. The amount of power required for the output of the RTX 4090 exceeded the capacity of the 16-pin power connector resulting in the connector melting and ceasing to function. In 2023, at the peak of this crisis, this caused the average lifespan of the 4090 to be under six months (1).

So, as of now, we have RTX 4080's being used for GeForce Now. Not only that, NVIDIA has incorporated the Ada Lovelace architecture, specifically AD103, into the graphics cards as well "providing massive breakthroughs in speed and power efficiency to tackle demanding creative, design, and engineering workflows" (5). This includes the latest generation RT cores, Tensor cores, and Cuda cores as well as 2 AV1 encoders.

This process starts with the **CUDA cores**. These cores, similar to general computation cores, are small and simple although not as powerful relative to a CPU, but importantly specialized for parallel computing and algorithms that can be parallelized. Making them massively important for any GPU. There are often thousands of them on the GPU, for instance the 4080 has 9,728 CUDA cores, which work in parallel to process graphics tasks given by the CPU. In short, they are special types of cores specifically designed to speed up calculations that have to do with graphics processing (24). The massive number of cores indirectly contributes to the amount of processing power that the cards have. Of course, there are other factors to look at. The reason for this is that they offer a boost for "time-intensive workloads, gaming, and deep learning", making them perfect for real-time graphics in a gaming environment, and give NVIDIA the necessary processing power to incorporate neural networks into their graphics cards (24). With a high-level view, the Ada architecture can be broken down into Streaming multiprocessors (SM), each of which can be thought of as a CUDA core. Each CUDA core has floating point and integer units that do the relevant computations, as well as a **Tensor core**. Tensor cores are specially designed to accelerate matrix operations, specifically matrix-multiply-and-accumulate computations, making them exceptional at tensor math and perfect for deep learning applications.

Briefly, **Tensors** are “mathematical objects that can be used to describe physical properties”, for instance scalars and vectors, a scalar being a zero-rank tensor and a vector being a rank one tensor (28). A tensor’s rank corresponds to the number of directions needed to represent it in math; in terms of a computer processing it, it can be thought of the number of columns in an array that are needed to represent it. This means that a Tensor is represented, but is not synonymous with, a matrix. Regular cores are famously inefficient at matrix manipulation, so having a specialized hardware for these types of intensive calculations comes with immense speedups to performance.

NVIDIA’s Tensor cores use **mixed precision**, or transprecision, computation. Unlike multi-precision computing, which is just using processors that can perform computations with different precisions, mixed precision computing “uses different precision levels within a single operation to achieve computational efficiency without sacrificing accuracy” (30). Intuitively, the more bits, or more precision, that are used to represent a value, the longer it takes to perform operations and the more storage space is required when performing those operations. Annoyingly in neural networks (DLSS), the precision correlates to the accuracy of the result. With mixed precision computation “calculations start with half-precision values for rapid matrix math” allowing for the same speeds that can be achieved with half precision computation, but “the machine stores the result at a higher precision” which maintains the quality with the single precision or double precision (30). The Mixed precision technique is said to give a 25 time increase to speed of applications over the traditional double precision methods, all while “shrinking the memory, runtime and power consumption required to run them” (30). Impressively, NVIDIA achieved performance levels of 550 petaflops (PFlop) where 1 PFlop is equal to performing one thousand million million, or 1,000,000,000,000,000, floating point operations per second (30).

Initially implemented in NVIDIAs Hopper architecture, the RTX 40 series graphic cards also use fourth generation Tensor cores with the **FP8 transformer engine** which are said to “deliver 6 times higher performance over FP16” version, which we referenced previously. The concept behind this improvement is the same as above. The difference being that the input has been further reduced to 8-bits instead of 16-bits. Since the input size is smaller, we can use the same logic as before, namely, it allows for all the operations to be done even faster and then be upscaled to the desired precision to maintain quality while reducing the space needed to store intermediary calculations.

With these dramatic speedups, increases to throughput, and the reductions in memory pressure, these new generation tensor cores are perfect for the use in AI technologies. In the case of the RTX 4080 it is **Deep Learning Super Sampling 3 (DLSS 3)**. An important note is that DLSS 3.5 has already been released and on top of AI antialiasing and frame generation, which are present in DLSS 3, it also includes AI ray reconstruction (34). As mentioned in a *The Verge* blog post by Tom Warren, it is going to take time for developers to be able to implement this technology into games, however, all RTX models are currently capable of using DLSS 3.5 (35). Consequently, I have considered this a future application and will be exploring it in more detail later in the Future section of this report.

Commented [JM1]: Insert hyperlink?

For now, let’s consider what is being used most commonly, which is DLSS 3. As mentioned, DLSS 3 provides its worth in two aspects during the graphics render pipeline, super resolution, and frame generation.

With native rendering (without DLSS enabled), each pixel will be sampled for any color, texture, shadow, and any other effect it might have. These will then be calculated on the GPU and sent back to the CPU to be displayed on the screen. If antialiasing is included, which is almost essential in all modern games when realistic, non-blocky graphics are wanted, each pixel will then be sampled multiple times and that could be as high as 64 times per pixel. This, in general, is 64 times more work for the GPU to perform per frame, which roughly corresponds to it taking 64 times longer to display the game, resulting in a lower frame rate overall and higher latency.

What DLSS does is instead of going through all the pixels in the frame, it begins with a downscaled and aliased (has not gone through anti-aliasing) version of the frame which is then processed by a trained deep learning neural network, DLSS 3, generating, the “missing” pixels to upscale the frame to the desired resolution. Some obvious advantages can be seen from this approach.

First, it operates with a frame that has a smaller number of pixels. Now, if the neural network was not there, the resulting frame would have very low quality, but while going through this network, anti-aliasing and artificial intelligence driven super sampling can be applied to eliminate blocking artifacts and produce the missing pixels needed for upscaling the frame to a higher resolution respectively. Also, because anti-aliasing is not needed to be done before upscaling, as it is in native rendering, the time it takes to get the frame ready to be upscaled is decreased as well.

Secondly, often native rendering uses a lot of spatial upscaling techniques. These share some similarities to the super sampling performed with DLSS. One is that they both aim to add more pixels to scale up the resolution from a lower base frame resolution, and they both succeed at doing this. Problematically, and what DLSS aims to fix, is that even if pixels can be added to the frame, details cannot. Spatial upscalers work frame by frame, treating every frame separately and adding more pixels to each frame. For still images this is very effective as the different frames can be stacked on top of each other resulting in better resolution. For games, however, the user is most often moving through the game as well as the movement of lights and shadows etc. are all constantly changing on screen. So, this approach causes a lot of temporal artifacts, presenting as flickering and crawling, as the algorithm can't relate frames in time, but only in space. DLSS, being a temporal method, makes use of motion vectors to improve objects in motion and particle reconstruction, stopping particles from streaking, increasing the temporal stability of each frame that is lost in the spatial methods.

The second function in DLSS 3 is **frame generation.** This is a new addition to DLSS 3 and exclusive to 40 series graphics cards. The basic idea is that by using optical flow accelerators, a hardware component in the graphics cards which compute optical flow fields, and neural networks to predict the motion of pixels from the flow fields, entirely new AI generated frames are generated and inserted between the actual rendered frames (36). The immediate effects of this are a higher frame rate as the generated frames do not have to be rendered by the game itself, resulting in a much smoother gaming experience for the same frame quality. This technique effectively produces frames independent to what is being sent on the network, meaning it is not reacting to a user input but smoothing out the time between rendered frames making it critical to achieving high frame rates when streaming games remotely. It could also be thought of a sort of buffering that we see in video streaming services. In any case, NVIDIA reports about a 2x increase over DLSS 2 and a 4x increase over brute force rendering. Optical flow is trying to move objects between frames and in conjunction with geometric motion vectors, the neural network can create an accurate reconstruction of the movement of an object in a frame.

It is often hard to know when to use each part of the frame, especially in traditional applications. This is where neural networks come in. The neural network will use only the necessary parts of these vector fields and the frames they were computed from to create the new frame. This reconstructed frame is then slotted between the rendered frames. This is effectively injecting a higher frame rate for the user while not requiring additional frames to be rendered and while maintaining accurate geometry and effects.

Taking a step back and looking at both super resolution and frame generation together, 7 out of every 8 pixels on the screen are generated using neural rendering from DLSS 3. The only pixels going through a somewhat traditional rendering process are the ones present in the initial raw, downscaled frames, from there super resolution upscaling is used to create and fill the “missing” pixels of that frame to achieve the necessary resolution, and frame generation is used to generate an entirely new full resolution frame which is inserted before the next rendered frame. Basically, 2 full resolution, anti-aliased, and artifact free frames are created for the computational price of rendering a quarter of 1 frame.

We have seen how DLSS can elevate the graphics in a game. It also fixes another major problem that games are facing now with the advancements in CPU processing speed falling behind GPUs, this is referred to as CPU boundedness. Graphics cards have been improved to the point that the bottleneck for a graphics rendering, depending on the implementation, is caused more often by the job delegation from the CPU to the GPU then the computation on the GPU side itself. Although the CPU isn’t responsible for processing graphics for modern games, the GPU, expert at processing visual data, is not capable of issuing tasks by itself. So, the CPU is still responsible for passing off the tasks for the GPU to compute. DLSS 3 generates frames independent of the CPU so a higher frame rate can be reached even in CPU bounded situations while not affecting the load put on the CPU.

This dual neural network approach hinges on the powerful matrix processing of the tensor cores. For every frame there are 2 neural networks being run simultaneously, one for super resolution and the other for frame generation, making the computation requirements very high. Tensor cores are essential for incorporating neural networks into the RTX 4080, and subsequently responsible for the massive increases in quality and higher throughput of frames caused by the AI intervention.

The next component of the CUDA core are **Ray Tracing cores**, or RT cores. Much like Tensor cores, RT cores are highly specialized hardware, but instead of matrix operations they “accelerate ray-traced graphics, allowing them to be rendered in real time” by doing ray tracing math very fast (37). In a nutshell, ray tracing is simulating the reverse of how light is reaching your eyes. Normally what happens is that light is coming from the sun, or any light source, and bouncing everywhere with a small proportion of the light rays reaching your eyes, this being what you see. Instead, ray tracing sends out a finite number of rays originating at your eyes (the camera) and through calculations, traces them as they bounce around the scene. If a ray hits a light source, then it will be rendered to the screen. This requires extremely high computation abilities, and this is where RT cores become crucial to providing users with photorealistic frames during gameplay. RT cores add more circuits to CUDA cores that are included in the render pipeline (37). CUDA cores will delegate ray tracing tasks to the RT core and then use the result to render the scene on screen (37). The processing of RT cores hinge on looking for ray intersections within NVIDIAs Bounding Volume Hierarchy (BVH) and then applying a machine learning denoiser to produce the result fit for display. Briefly, BVH is a way to organize the geometry in the scene to make it less computationally intense for the GPU to process (37). Similar to giving computation specific tasks to Tensor cores to speed up matrix calculations, RT cores are given ray tracing tasks to complete once again speeding up processing time of each frame.

The components talked about up until now have to do with processing the data and turning it into a form fit for sending to the users, but the resulting output cannot be sent as is. The size of each frame is much too large to be sent over a network, much less for a real-time gaming experience. For this, video compression is used to package the video into a size that can be sent over the network. As a result, NVIDIA has adopted the most recent compression codec, **NVENC AV1**. The RTX 4080 cards have 2 built in AV1 encoders which are used to compress video quickly and efficiently, lending itself to significantly reducing the time every frame spends at the NVIDIA servers.

The NVENC AV1 codec, created by *Alliance for Open Media consortium,* is seemingly the cutting edge of video compression. Achieving about “30% compression gains over its predecessor VP9”, as stated by Jingning Han et al, it is not hard to see why it has gained recognition and is being used by NVIDIA (1). AV1 functions similarly to its predecessor codecs, namely H.264, H.265, and VP9. The process follows the same steps, partitioning, prediction, transformation, quantization, then entropy encoding, all of which are present in previous versions. Where AV1 pulls ahead is the number of options that can be performed in each step. For example, for intra-prediction in VP9, there are 8 directional modes that can be used, whereas in AV1 there are 56 directional modes that can be chosen from. There is a lot to unpack with AV1, but the two main takeaways that I want to highlight come from the increased maximum block size and the increase in options that are present at each step.

First, with a larger block size it allows for better allocation of processing time. In places with large similarities among neighboring pixels, there is no need to use small block sizes to represent it. For example, if there is a frame that consists of mostly clear blue skies but with one person, it is inefficient to have small block sizes for the sky, which would require the allocation of more processing time, as it is the same color and would need the same, or at least very similar, types of prediction, transformation, and quantization across all the pixels. Including more pixels in each block would allow for the algorithm to spend more time partitioning and predicting in areas such as the person which would have a bigger discrepancy between neighboring pixels and would require more block partitioning with different options for each block in each step in the process.

Secondly, with more options at each step of compression, a more precise representation of each pixel can be achieved. That could be from partitioning each block in a way to better represent the object, blending the motion vectors of neighboring blocks when predicting to better represent the motion of the frame, or having more reference frames buffered that can be compared to. Having this fine-grained specificity at each point in the frame helps reduce the degradation in frame quality present in all lossy compression algorithms, resulting in a higher quality when decoded at the user's device.

In a blog post by NVIDIA on the use of AV1 encoding, they compare the reconstruction quality of AV1 compared to H.264 by plotting the PSNR score. AV1 provides a higher PSNR score throughout the entire bitrate range, and importantly it has a better PSNR score at low data rates. The PSNR value for AV1 at a bit rate of 4 Mbps is roughly the same as H.264 at almost double that bit rate. What this means is that if AV1 is being used, not only can a user get better quality reconstruction, but they can get better reconstruction for a lower bit rate. Or conversely, they can achieve an even higher quality for the same bitrates. This could be the difference between being able to play a game at 60 fps and not being able to play the game at all. Intuitively, cloud gaming has some minimum requirements on the network connection a user has. No matter how fast the processing is on the server end, if the user's bandwidth can't keep up with the downlink throughput, then the framerate and quality will suffer and may cause it to be unplayable. This makes sense when thinking of the **transmission delay** and **network congestion**. The fewer bits that need to be put through the links will also decrease the time it takes to push all the bits into a link, decreasing **transmission delay** and the network latency and round-trip time will follow. If we look at all the users who are using the platform, at peak hours NVIDIA is reporting there to be more than 200 million users around the world, each one sending and receiving data. In the modern era of the internet, **congestion** has become a big problem. Cloud gaming is no exception to this, resulting in NVIDIA reporting a high number of packets lost. Being able to send smaller payloads will inherently increase the amount of data that can be sent while keeping the network resources that NVIDIA has access to from becoming totally congested. With an uncongested network, less packets will be dropped, meaning less packets being sent more than once, reducing the load on the network, and reducing the time it takes users to get the next frame that needs to be displayed.

As stated before, the RTX 4080 includes **2 AV1 encoders** in its hardware. With multiple encoders, video streams can be encoded in parallel. Being able to encode using multiple encoders also helps with the number of users sending data since multiple streams can be encoded at the same time. This also allows the GPU to always be working by having “different frames from different streams … scheduled across multiple” encoders (21).

Lastly, included as a software component within DLSS 3, NVIDIA Reflex aims to optimize the system latency of each user. The motivation behind this is reducing the system latency means reducing the time before input data can be sent to the server. This is especially important when thinking about the added time for cloud gaming. Not only does the users local system need to register input, but it then needs to send it to the cloud server, have it processed there and then sent back. Minimizing the time before it is sent through the network is critical to reducing the total latency that the user feels. NVIDIA also makes a distinction between the notion of frames per second and latency as a measure for system performance. Frame rate is a measure of system throughput but is not the only thing affecting the latency that the user experiences. Both are important overall, but only caring about frames per second is neglecting some parts that could be optimized to provide an even better experience. For cloud gaming, everything needs to be optimized to produce the desired gaming experience. NVIDIA Reflex gives players the option to measure and optimize system latency for their device.

One important place of optimization for NVIDIA Reflex is the render queue. Traditionally, frames are queued up by the CPU in a render queue which are then processed by the GPU. What NVIDIA have done is get rid of the render queue by syncing up the CPU and the GPU, eliminating the need for frames to spend time queueing and waiting to be processed. This means that the GPU will be as close to full utilization as possible while reducing the backpressure on the CPU. Ultimately, this will result in higher fps and lower latency in GeForce Now.

Future of GeForce Now

NVIDIA sees potential in the cloud gaming space and will continue to devote significant resources towards future developments. As mentioned, DLSS 3.5 is already available in NVIDIAs modern graphics cards, but with limited implementation in today's games it is not commonplace yet. One example of this already being in use is in Cyberpunk 2077. DLSS 3.5 introduces ray reconstruction, a technique that enhances the quality of ray tracing using AI. In some ways it is similar to frame generation, it uses trained neural networks that "generates higher quality pixels between already sampled rays" instead of generating frames and weaving them in between sampled frames like in frame generation (38). It operates at the same time as super sampling in the last stage of the lighting pipeline, effectively replacing hand-tuned denoisers. The result from using ray reconstruction would be a cleaner final frame quality with less ghosting as it can use both temporal and spatial data to retain high frequency information for ray construction and frame upscaling. There has also been talk about the next generation of the Lovelace architecture, Ada Lovelace Next and Blackwell being potential names, being released in 2025. Although not much is known about the specifics of the architecture, given the success of the current Lovelace architecture it points to even greater improvement for NVIDIAs graphics cards and therefore GeForce Now.

My Thoughts

I have had some experience using the Free tier of GeForce Now. It worked well and the latency and quality, while certainly differentiable from playing locally, were functional and more than worth playing for free. I noticed a small but not insignificant lag time between my input, and it being registered on screen. In my experience, it compares to about a 100ms lag time in a local copy of the game. Considering that it is a free service, and it provides a playable and visually coherent result from data streamed to and from a remote server, it is exceptional for any casual user. From secondhand experience, mainly video and blog reviews from others, the Ultimate tier of GeForce Now provides an identical if not a superior experience to that of high-end gaming PCs. Another problem that arose as the user base expanded was the wait times to reserve a gaming rig. Users must queue for the Free tier gaming rigs, which in my experience was on average 5 to 10 minutes for the Free tier, but there have been many reports of it being an hour or more. This is obviously not ideal, and it does take away from the user accessibility of the platform. However, the Priority and Ultimate tiers do not suffer from this problem and have their dedicated queues which guarantee low or no waiting before being allocated a gaming rig.

Summary

Throughout this report, it has been established that latency and quality are both obstacles faced by cloud gaming. Putting together the components of the hardware of the GeForce Now Ultimate servers, throughout all areas, from user to server and back, the use of AI, advanced compression algorithms, and system latency optimization have been used to provide a smooth and responsive experience. The users input response times and network capabilities are measured and adjusted to provide an optimized throughput from the user’s device to NVIDIAs servers. Then, at the servers, 4th generation tensor cores coupled with 3rd generation RT cores in the latest RTX 4080 graphics cards quickly process the data and use super sampling and frame generation incorporated in DLSS's neural networks to produce upscaled high resolution frames as well as artificially generated frames to provide CPU independent, high frame rates. Then, AV1 is used to compress and encode frames with high quality reconstruction which are sent back to the user and displayed on their screen. Although not solved, the problems with latency, quality, and user accessibility are managed in a way that produces a comparable gaming experience to playing with a local game copy with a more manageable investment for the average user.

Sources

1. https://www.tomshardware.com/news/rtx-4090-16-pin-connector-melted-after-one-year-of-usage
2. https://arxiv.org/pdf/2008.06091.pdf
3. https://www.pcmag.com/news/us-to-block-nvidia-from-shipping-more-geforce-rtx-4090-gpus-to-china
4. https://www.pcworld.com/article/2010870/no-geforce-rtx-4090-ti-after-all-nvidia-has-other-plans.html
5. https://www.nvidia.com/en-us/design-visualization/rtx-4000/
6. https://www.nvidia.com/en-us/data-center/tensor-cores/
7. https://images.nvidia.com/aem-dam/en-zz/Solutions/technologies/NVIDIA-ADA-GPU-PROVIZ-Architecture-Whitepaper_1.1.pdf
8. https://www.nvidia.com/en-us/technologies/ada-architecture/
9. https://www.nvidia.com/en-us/design-visualization/rtx-4000/
10. https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4080/
11. https://www.youtube.com/watch?v=qubPzBcYCTw
12. https://www.nvidia.com/en-us/geforce/news/reflex-low-latency-platform/
13. https://www.nvidia.com/en-us/geforce-now/
14. https://clouddosage.com/geforce-now-av1/
15. https://nvidianews.nvidia.com/news/nvidia-brings-rtx-4080-to-geforce-now
16. https://www.nvidia.com/en-us/geforce/technologies/reflex/
17. https://www.pcmag.com/news/us-to-block-nvidia-from-shipping-more-geforce-rtx-4090-gpus-to-china
18. https://www.tomshardware.com/news/rtx-4090-16-pin-connector-melted-after-one-year-of-usage
19. https://www.pcworld.com/article/2010870/no-geforce-rtx-4090-ti-after-all-nvidia-has-other-plans.html
20. https://www.nvidia.com/en-us/geforce/ada-lovelace-architecture/
21. https://developer.nvidia.com/blog/improving-video-quality-and-performance-with-av1-and-nvidia-ada-lovelace-architecture/
22. https://developer.nvidia.com/blog/av1-encoding-and-fruc-video-performance-boosts-and-higher-fidelity-on-the-nvidia-ada-architecture/
23. https://www.yololiv.com/blog/h265-vs-h264-whats-the-difference-which-is-better/
24. https://www.acecloudhosting.com/blog/nvidia-cuda-cores-explained/
25. https://shiyan.medium.com/some-cuda-concepts-explained-12ecc390d10f
26. https://wccftech.com/roundup/nvidia-geforce-rtx-4080/
27. https://www.pny.com/professional/software-solutions/nvidia-ada-lovelace
28. https://www.doitpoms.ac.uk/tlplib/tensors/what_is_tensor.php
29. https://www.digitaltrends.com/gaming/what-is-cloud-gaming-explained/

30. https://blogs.nvidia.com/blog/whats-the-difference-between-single-double-multi-and-mixed-precision-computing/
31. https://developer.nvidia.com/blog/programming-tensor-cores-cuda-9/
32. https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/
33. https://www.youtube.com/embed/udZzk8aEW98?origin=https%3A%2F%2Fwww.nvidia.com&hl=en_US&mute=1&autoplay=1&loop=0&controls=1&enablejsapi=1
34. https://www.nvidia.com/en-us/geforce/technologies/dlss/
35. https://www.theverge.com/2023/8/22/23841148/nvidia-dlss-3-5-ray-reconstruction-ray-tracing-quality
36. https://www.youtube.com/watch?v=pSiczcJgY1s

37. https://www.titancomputers.com/What-Are-RT-Cores-in-Nvidia-GPUs-s/1208.htm

38. https://www.nvidia.com/en-us/geforce/news/dlss-3-5-available-september-21/

39. https://www.tomshardware.com/news/nvidia-ada-lovelace-successor-in-2025