Presenters Names: Juan Guerra, Brian Hert, Joshua Hicks

Unveiling the Power of NVIDIA Ampere Architecture in AI and Deep Learning

From Pixels to Predictive Algorithms

Table of Contents

Section 1: Introduction to NVIDIA

The Genesis of NVIDIA

- Founders' Backgrounds: The company was founded on April 5th, 1993 by three American computer scientists. Jensen had a background in engineering and microprocessors, while Chris and Curtis had experience in computer graphics and chip design. Their combined knowledge and skills provided the foundation for NVIDIA's future success.
- 1993-2022: Journey from the NV1 to the GeForce 40-series, showcasing NVIDIA's relentless innovation in GPU technology.
- Initial focus: Developing GPUs for the gaming industry.
- The leap towards AI and Deep Learning began with the integration of parallel processing capabilities.

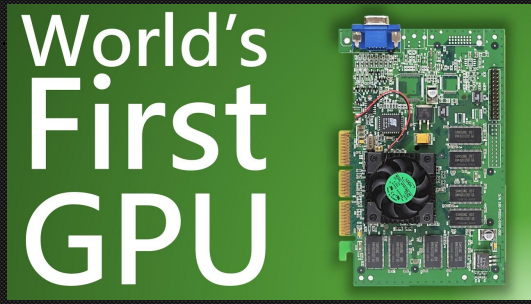
Jensen Huang
Co-Founder, President and CEO

Chris Malachowsky
Co-Founder

Curtis Priem
Co-Founder

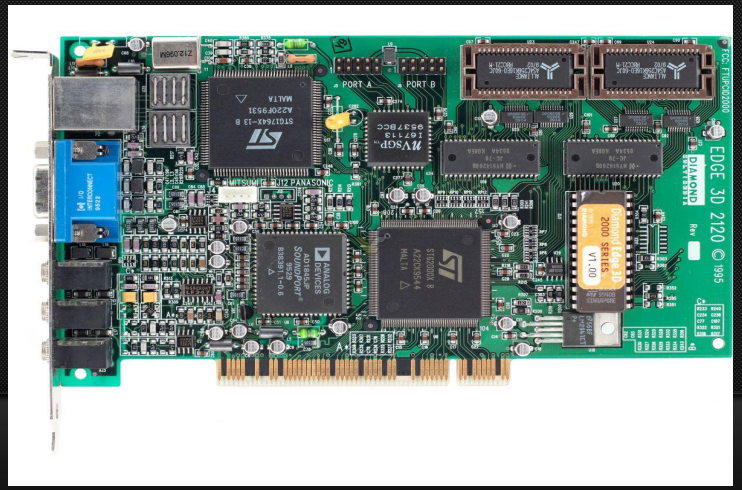
Milestones of Innovation

- CUDA Emergence (2007): Unlocking general-purpose processing on NVIDIA GPUs.
- Volta & Turing Architectures: Leaps toward AI, introducing Tensor cores.
- GPU Evolution: From graphics rendering to AI's backbone.
- Real-world Impact: Speeding up image processing for autonomous driving and facial recognition.
- Deep Learning Advancements: Accelerating neural network training, opening new possibilities in AI.
- Foundation of Artificial Intelligence.
- Transition from slow, inaccurate ML to efficient Deep Learning.
- Case of image/video processing for autonomous driving and facial recognition.
- Parallel processing advantage: Example of image processing.
- Evolution to AI-capable GPUs

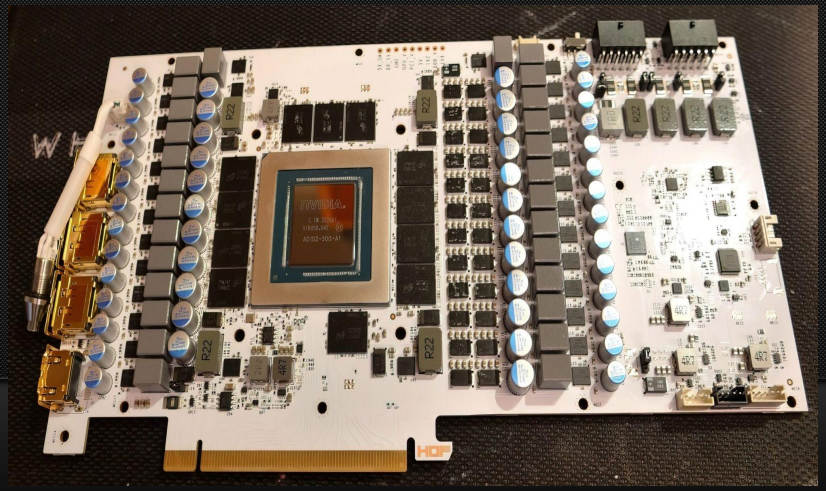


nVIDIA

Section 2: The Evolution of NVIDIA's GPU Architecture



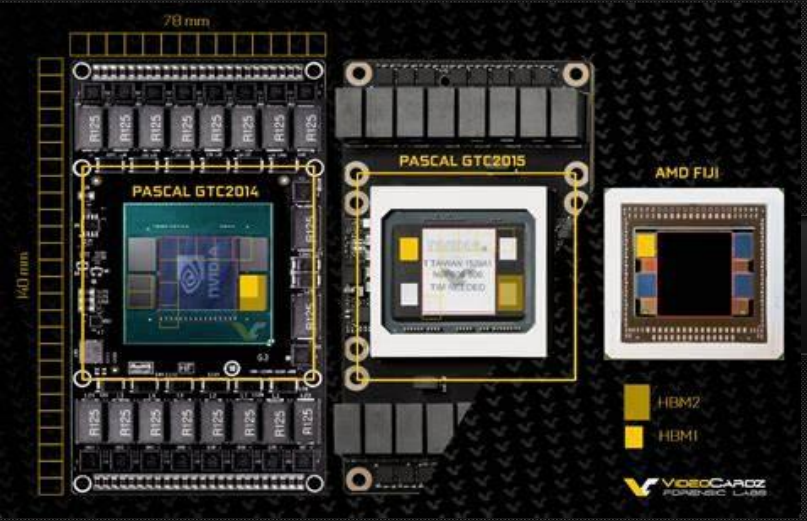
->

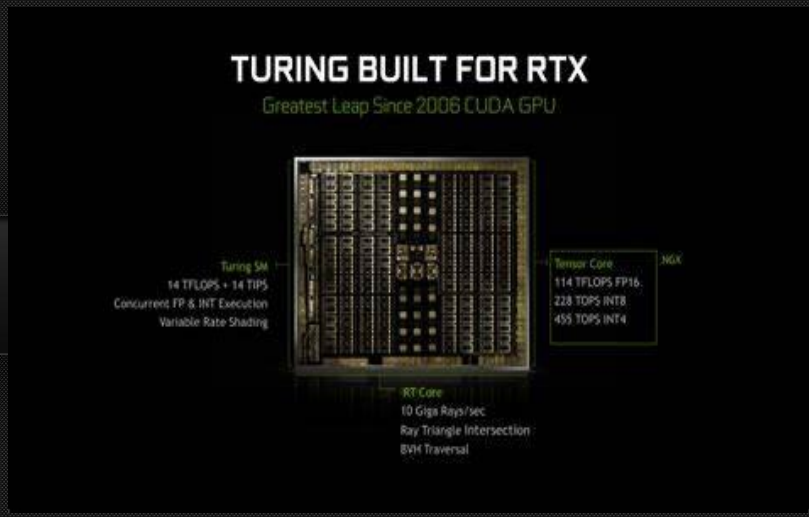


nVIDIA

From Pascal to Turing

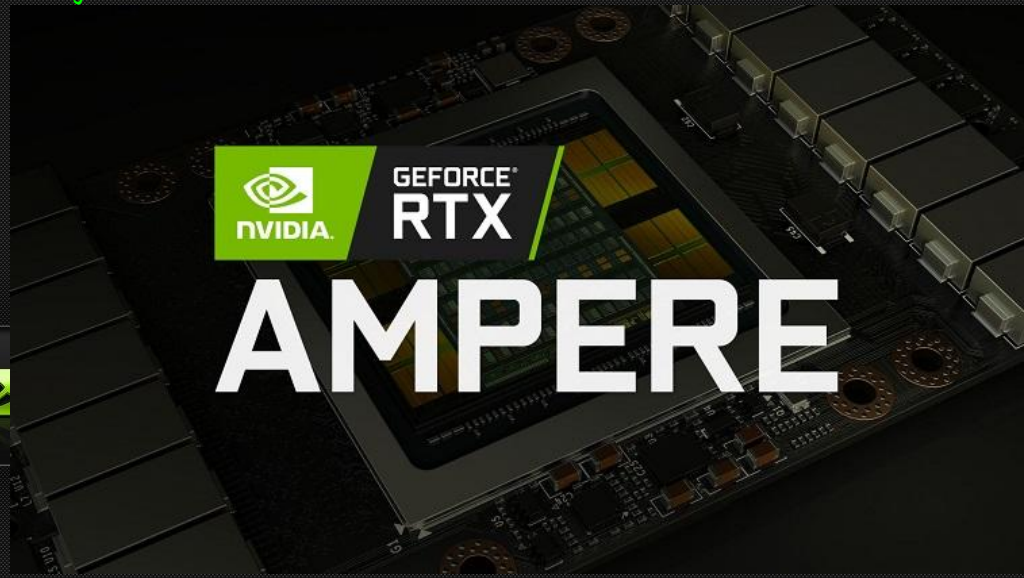
- Introduction of Tensor cores in Volta architecture (2017) marking NVIDIA's foray into AI processing.
- Turing (2019) further enhanced Tensor cores and introduced Ray Tracing cores for photorealistic rendering.





Embracing the Ampere Architecture

- New Die Size and Transistor Count: Transition from Tesla V100's 815mm on 14nm process with 21.1 billion transistors to Ampere's 826mm on TSMC's 7nm process with 54.2 billion transistors.
- Performance: 19.5 teraflops of FP32 performance, 6,912 CUDA cores, 40GB of memory, and 1.6TB/s of memory bandwidth.
- Partnership with Red Hat OpenShift simplifies GPU usage for data science workflows, illustrating real-world application.
- Sparse INT8 Performance: Peak performance of a single A100 is 1250 TFLOPS.
- Significant architectural improvements:
- Updated PCIe Host Interface to PCIe 4.0.
- Support for GDDR6 memory.
- Enhanced memory bandwidth and cache size2.
- CUDA parallel computing platform.
- Application in real-world AI/ML problems.



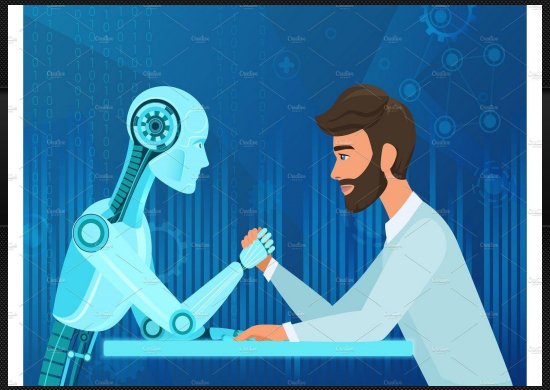
Section 3: The Surge of Artificial Intelligence

AI: The New Electric

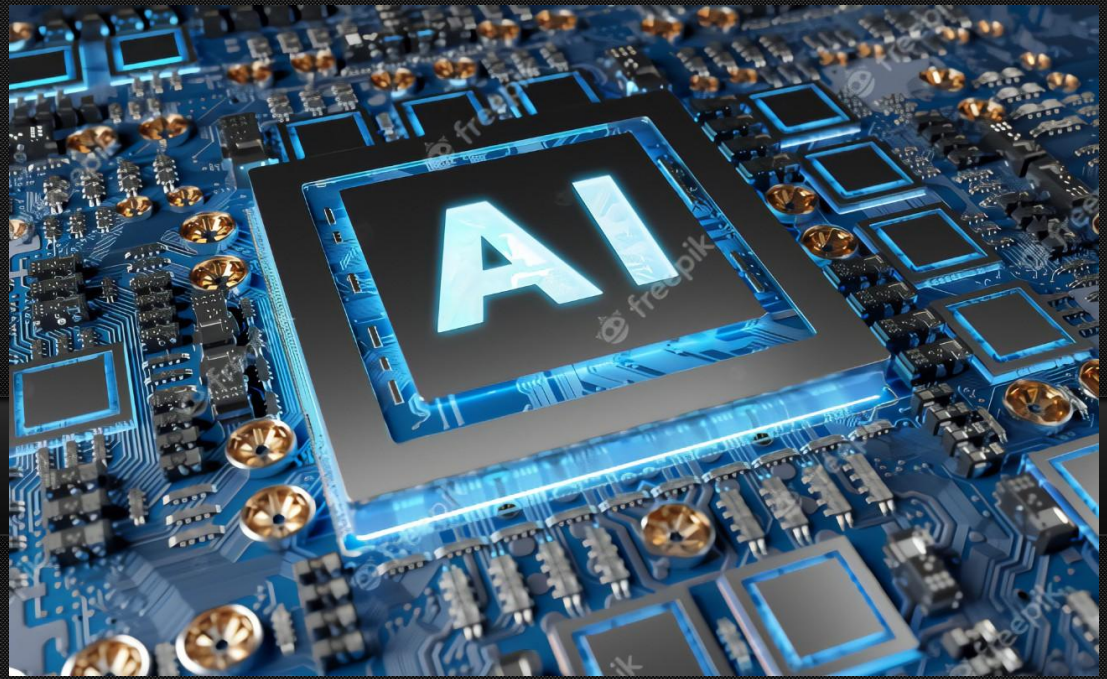
- TensorFloat-32: A new number format optimized for Tensor Cores, replacing FP32 in some workloads for better performance without code changes.
- Speedup: 6x speedup in AI training, 7x speedup in inference compared to V100.
- Fine-Grained Structured Sparsity: A new concept that improves compute performance of deep neural networks by compressing matrix math.
- Exponential Growth: AI's rapid expansion across industries.
- Ubiquity: Transitioning from a tool to a foundational utility akin to electricity.
- Sector Transformation: Intelligent workflows revolutionizing mundane tasks.
- Ampere's Role: Fueling AI's growth, becoming the conduit for AI as a utility.

Challenges in AI Computation

- There are many computational challenges posed by complex AI models and these demand a need for powerful computing solutions.
- Integration with cloud computing and high-performance computing.
- Scalability and performance in large-scale AI/ML projects.
- Continuous innovation and partnerships for efficient AI application development.

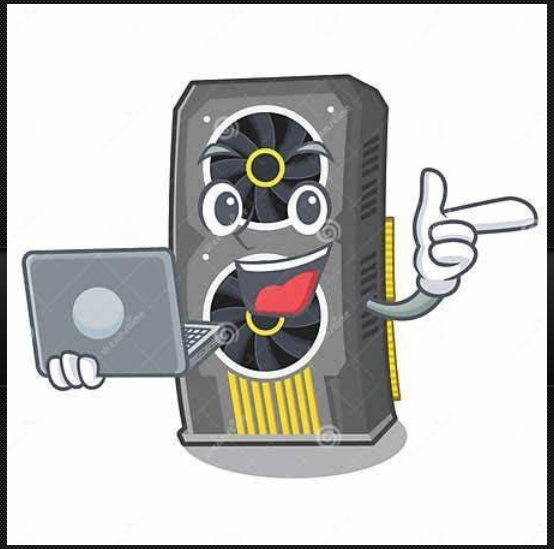


Section 4: The Essential Usage of GPUs in AI



Why GPUs?

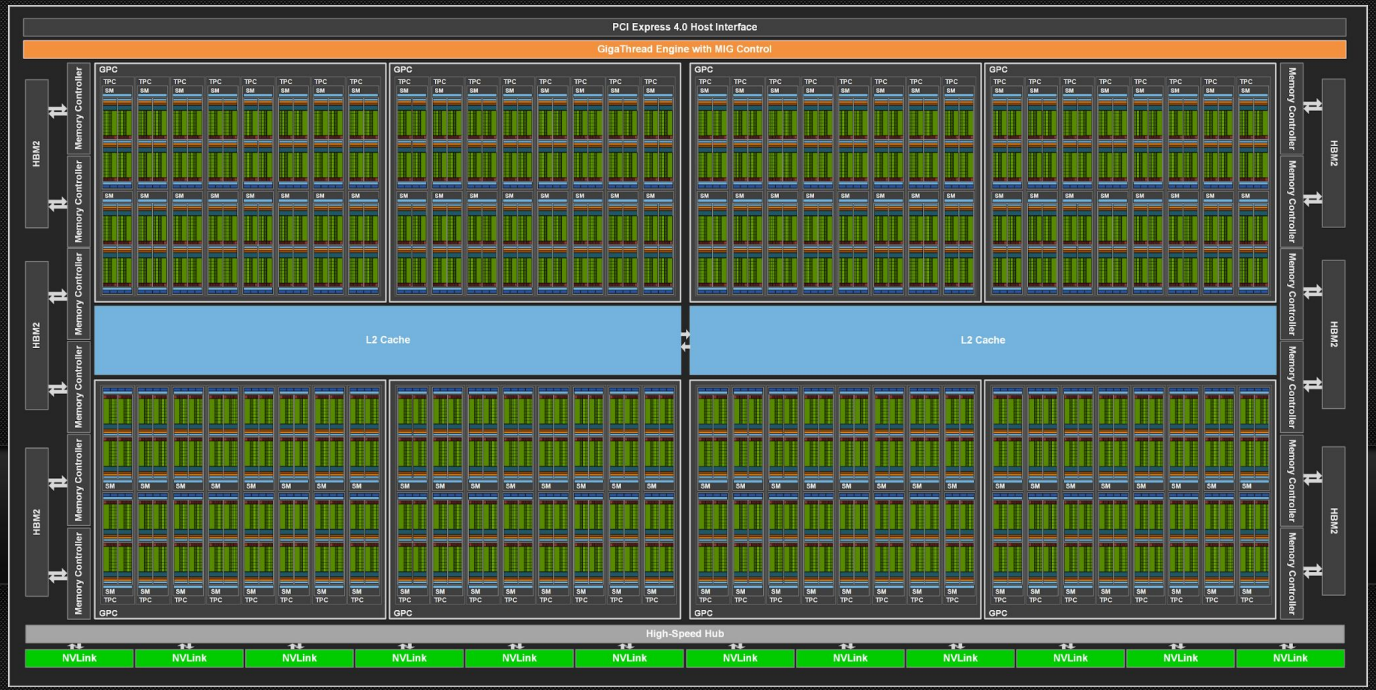
- Massive parallel processing capabilities of GPUs making them a preferred choice for AI computations.



NVIDIA's GPU Dominance

- NVIDIA's Prowess: In the dominion of GPUs, NVIDIA stands tall. Its GPUs have become the linchpin in AI computations, offering a blend of performance, efficiency, and reliability that is unparalleled.

Section 5: The NVIDIA Ampere Architecture: A Leap Forward

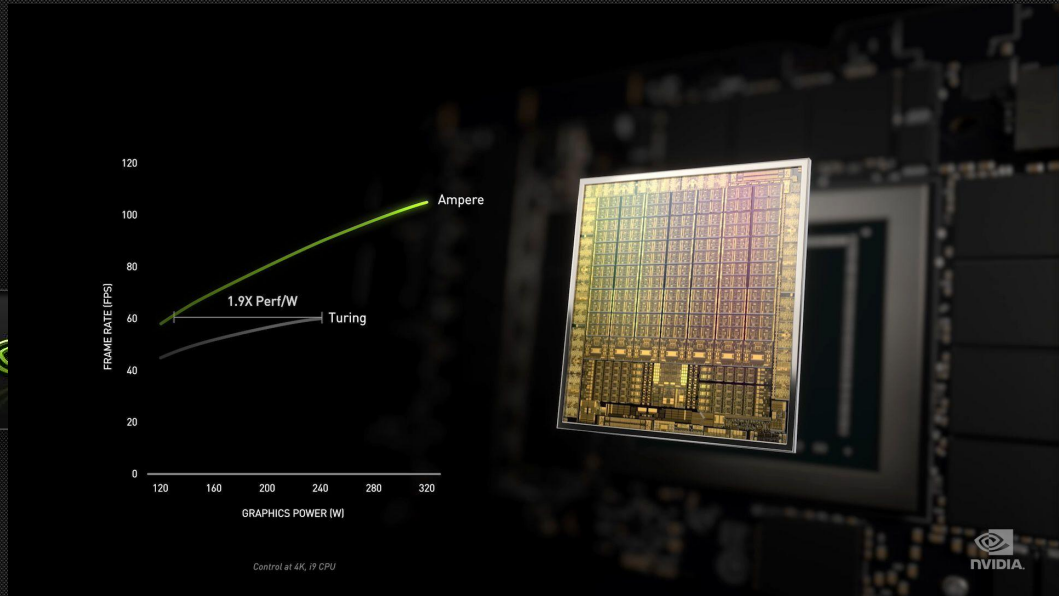


Ampere: A Synthesis of Power and Efficiency

- Maximizing Tensor Core Performance and Efficiency for Deep Learning Applications1.
- Mixed Precision Training with Tensor Cores: Enhancing performance while maintaining accuracy.
- The Marvel of Tensor Cores: Over 2x Speedup in Matrix Multiplication.
- Efficiency through Memory Bandwidth: Ensuring continuous operation of Tensor Cores by maximizing memory bandwidth.

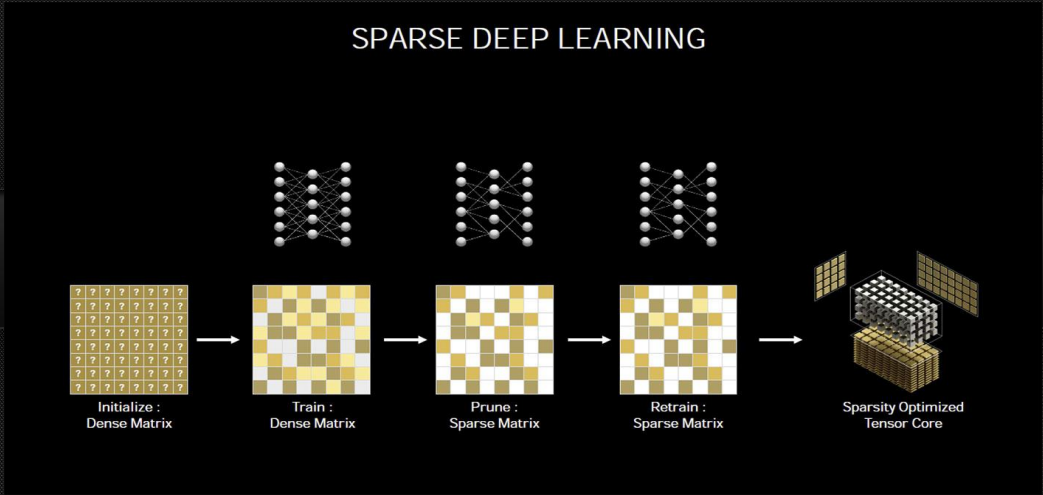
Unifying AI Training and Inference

- Ampere's unified architecture delivering a 20x performance boost from its predecessor.



Third-Generation Tensor Cores

- Enhanced performance and efficiency by taking advantage of fine-grained sparsity in network weights6.
- Enhanced Tensor Core Instructions: TF32, BF16, and FP64 for versatile performance.
- Sparsity Feature: Doubling the throughput of Tensor Core operations.



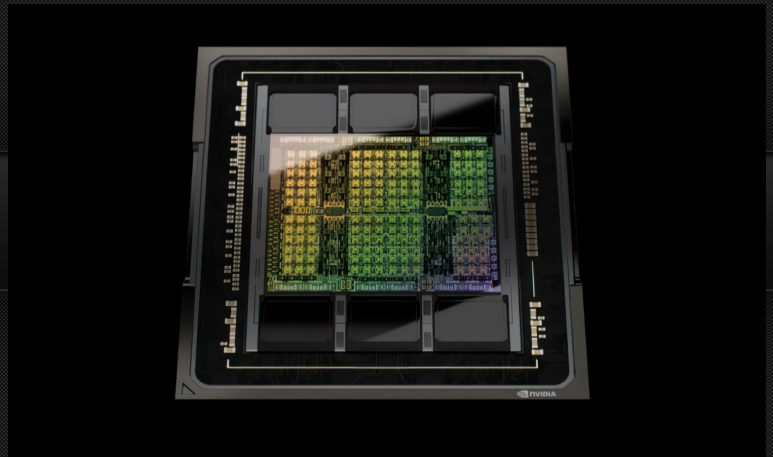
Section 6: Envisioning the Future: Hopper and Beyond

The Road Ahead

- Future Innovations: The Ampere architecture had not only set a new precedent but also illuminated the path for future innovations, such as the current Lovelace architecture. It's the harbinger of an era where AI and Deep Learning are bound only by the limits of human imagination.

The Horizon - Hopper Next

- Upcoming Architecture: As we stand on the cusp of a new computational era, NVIDIA has already charted the course ahead with the confirmation of their next architecture coming in 2024. Little information is known about this new architecture at this time, other than rumors naming it "Blackwell" but this will be another leap forward for AI development.



Thank You!!

Works Cited

- Clark, D. (2023, August 21). How Nvidia built a competitive moat around A.I. chips. The New York Times. https://www.nytimes.com/2023/08/21/technology/nvidia-ai-chips-gpu.html
- Dettmers, T. (2023, January 31). The best gpus for Deep Learning in 2023 - an in-depth analysis. Which GPU(s) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning. https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/
- Kinghorn, D. (2023, October 5). Hardware recommendations for machine learning / AI. Puget Systems. https://www.pugetsystems.com/solutions/scientific-computing-workstations/machine-learning-ai/hardware-recommendations/
- Krashinsky, R., Giroux, O., Jones, S., Stam, N., & Ramaswamy, S. (2023, May 24). Nvidia ampere architecture in-depth. NVIDIA Technical Blog. https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/
- Mujtaba, H. (2023, May 29). Nvidia reaffirms hopper-next gpus to launch in 2024, another giant leap for HPC & AI. Wccftech. https://wccftech.com/nvidia-reaffirms-hopper-next-gpus-to-launch-in-2024-another-giant-leap-for-hpc-ai/
- Reznik, A., Nelson, T., Abdo, K., & Xu, C. (2023, September 20). Why GPUs are essential for AI and high-performance computing. Red Hat Developer. https://developers.redhat.com/articles/2022/11/21/why-gpus-are-essential-computing#how_gpus_work
- Willings, A. (2023, March 25). Nvidia GPUs through the ages: The history of Nvidia's graphics cards. Pocket. https://www.pocket-lint.com/nvidia-gpu-history/
- XD, E. (2023, October 9). Best GPU for Deep Learning - Top 9 gpus for DL & AI (2023). ByteXD. https://bytexd.com/hardware/best-gpu-for-deep-learning/