HARDWARE USED FOR TRAINING VIDEO MODELS

Conner Hubbell and Matthew Lessman

TABLE OF CONTENTS

ARTIFICIAL INTELLIGENCE (AI)

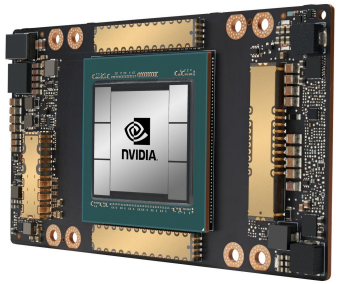
01.

INTRODUCTION

What are GPUs and Video Models???

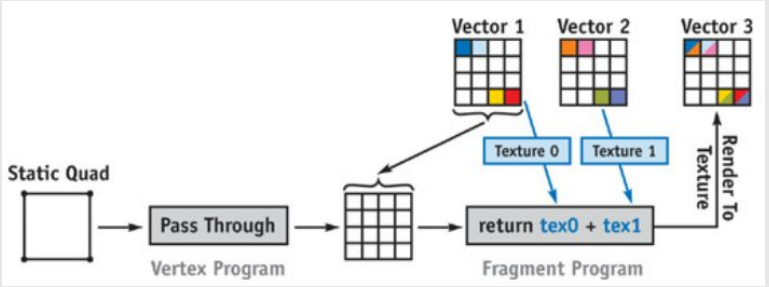
WHAT IS A GPU???

- Electronic circuit
- Creates images/videos through rapid processing
- PC and gaming console display
- Modernly used for parallel processing
  - Quicker computations
  - Used in Machine Learning and Artificial Intelligence



WHAT DOES A GPU DO???

- Quick maths with small cores
  - Partial Differential Equations
  - Linear Algebra



- Vector-Vector Operations
  - 2 vectors with number arrays to create images

- Matrix-Vector Operations
  - Matrix (rectangular array) and vector

VIDEO MODELS - WHAT ARE THEY???

- Video models
  - Use advanced machine learning to understand text and convert it to video

A litter of golden retriever puppies playing in the snow.
their heads pop out of the snow

02.

BACKGROUND

History of Graphics and the GPU

FIRST COMPUTER GRAPHICS SYSTEM

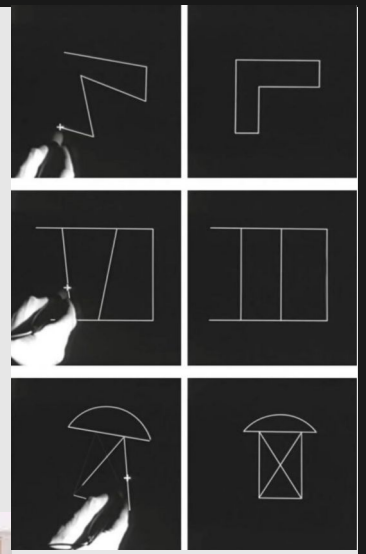
- “Baby” from 1949
- Dot-matrix display
- Demonstrated data stored on a cathode ray tube
- Could remember 2048 bits

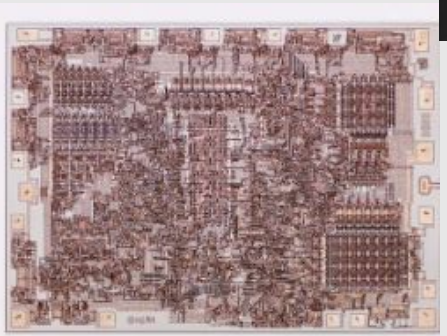
FIRST DIGITAL COMPUTER USED FOR GRAPHICS

- “Whirlwind”
- First to use video displays for output and operate in real time
- Used new core memory for Random Access Memory (RAM)

OTHER BREAKTHROUGHS FOR GRAPHICS

- “Sketchpad” by Ivan Sutherland (1963)
- Term “pixel” (picture element) coined (1965)
- First “tablet,” workstation, and game consoles introduced (1972)
- Personal Computer - PE-8 by Jonathan Titus (1975)
  - Powered by Intel 8008 processor (8-Bit)
- Pixel Planes project (1980-2000)
  - “Genesis of the GPU”
  - Allocated one processor per pixel, allowing simultaneous Image generation





PRIMITIVE 3D GRAPHICS

- Many companies vying for top-spot
  - Poor graphics, memory, and output limited them

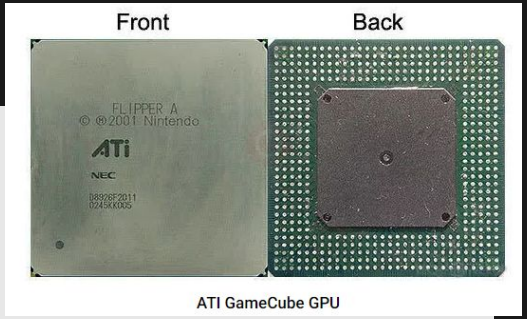
- RCA's "Pixie" video chip (62x128 resolution) - 1976
- TIA "1A" video chip (integral to the Atari 2600, trumped "Pixie") - 1977
- Motorola "MC6845" video address generator - 1978 (monochrome and color display adapter cards in IBM PC in 1981)
- Intel's "82720" graphics chip (8-color data at 256x256 resolution, 512x512 mono)
  - Big step toward graphics evolution - 1983
- ATI and EGA release competing graphics processors in 1987, continued competition with other groups into the 90s
- 1993 - top dogs of graphics start to show, but still offered new competition
  - Nvidia founded in 1993

“THE GAME-CHANGER”

- 3Dfx Voodoo
  - 3D-only chip
  - Essentially rendered 2D obsolete for some companies

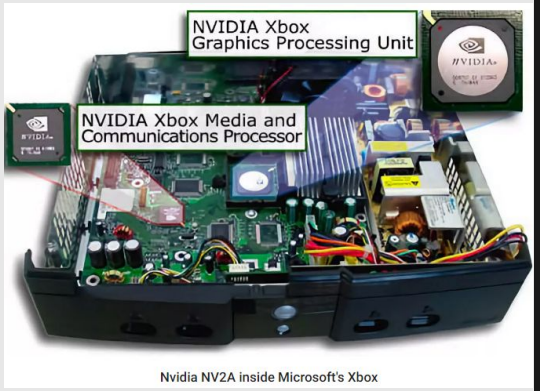
- Led to a plethora of other 3D cards being released
- Estimated that 80-85% of 3D market during Voodoo’s beginnings
- Left all companies in a scramble to figure out the best system for graphics
  - S3 Savage 2000
  - ATI Rage Fury MAXX
  - Nvidia GeForce 256 - project coined the term “GPU”
  - Increased efficiency in 3D image and video generation

NEW ERAS



ATI GameCube GPU

- Early 2000's it was Nvidia versus ATI
- ATI - released Radeon DDR (April, 2000)
  - "Most powerful graphics processor... for desktop PCs"
- Nvidia - released GeForce 2 GTS (GigaTexel Shader)
  - Emphasized details like blending, shading, refraction, waves, etc
- Rapid game and tech developments called for more and more chips to be made, creating constant market competition throughout the era



Nvidia NV2A inside Microsoft's Xbox

STREAM PROCESSING UNITS

- Unified shaders
  - Increased flexibility and efficiency due to all math being handled the same
- Nvidia - DirectX 10, Shader model 4.0
  - Advanced Graphics and increased programmability
- AMD (previously ATI) and its Radeon HD 2000 Series
  - Laid groundwork for GPU advancement with graphics technology
- Increased GPU computing
  - Parallel processing power demand increase (science, data, engineering)
- Expansion - GPUs and parallel computing more available

03.

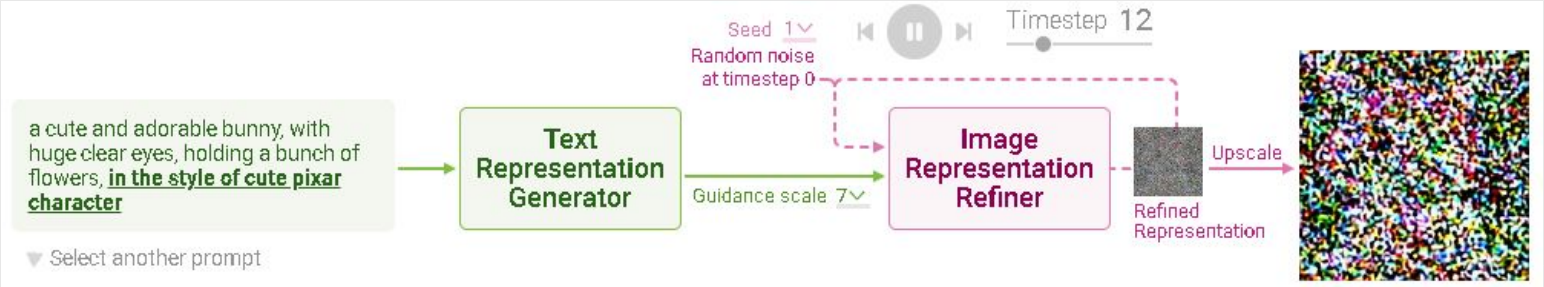
CURRENT

Current GPUs and Video Models

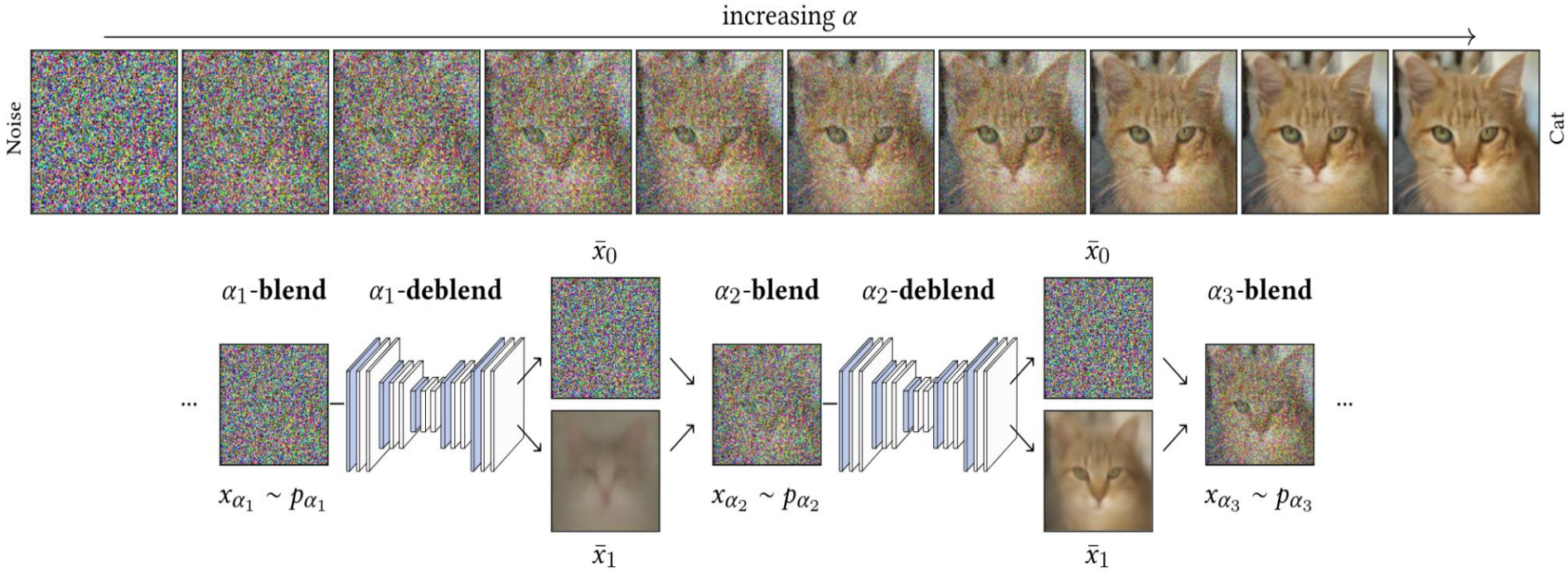
IMAGE MODELS

HOW IMAGES ARE GENERATED

- Text prompt as input from the user is tokenized and attached to starting image (noise)
- Image is sent through the neural network with the prompt attached and model tries to “denoise” the image.
- Loop through the process of taking the resulting denoised image and adding back most of the noise to get a more refined image.



DIFFUSION MODEL



SORA VIDEO MODEL

- Uses “patches” which are video version of tokens for an LLM
- Able to train on any resolution video or photos instead of cropping them all to the same base resolution
- Has a separate “captioner model” to caption the video with text before being used to train the text-to-video sora model
- Can be prompted with images and videos as well as the text-to-video option

SORA'S CAPABILITIES

3D CONSISTENCY

OBJECT PERMANENCE

LIMITATIONS

UNREALISTIC PHYSICS

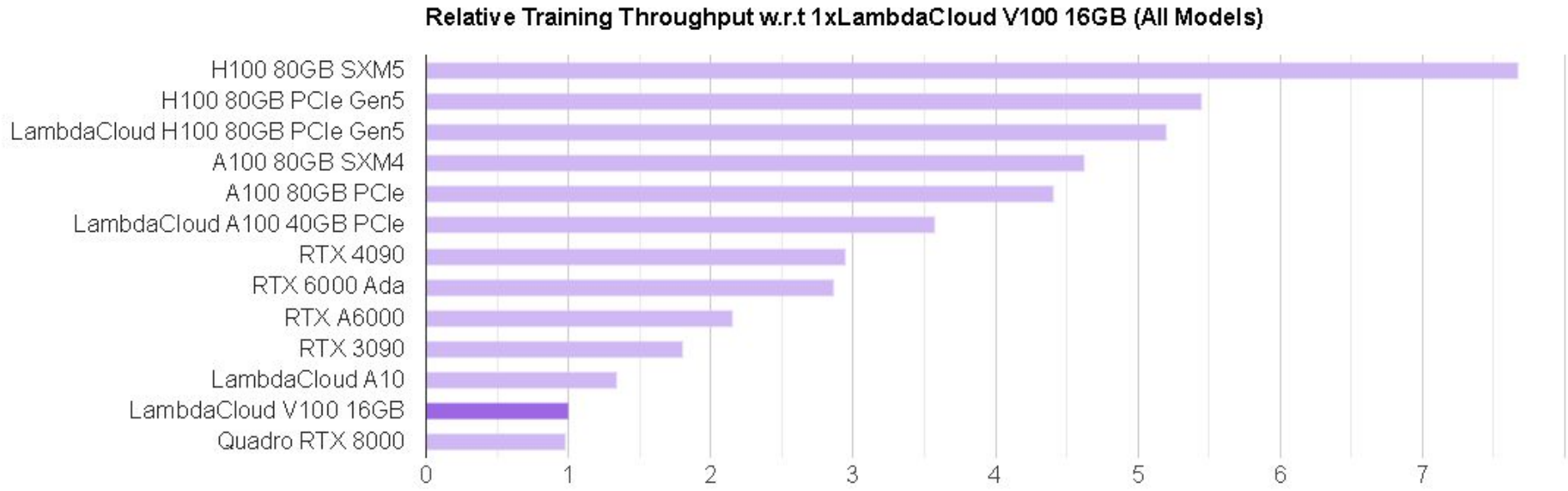
WEIRD HANDS / FACES

GAPS IN RESEARCH

- Always evolving and changing because of the insane demand for AI
- Most companies keep trade secrets close to stay ahead
- Enormous cost of entry to compete keeps smaller companies from competing

CURRENT GPUS

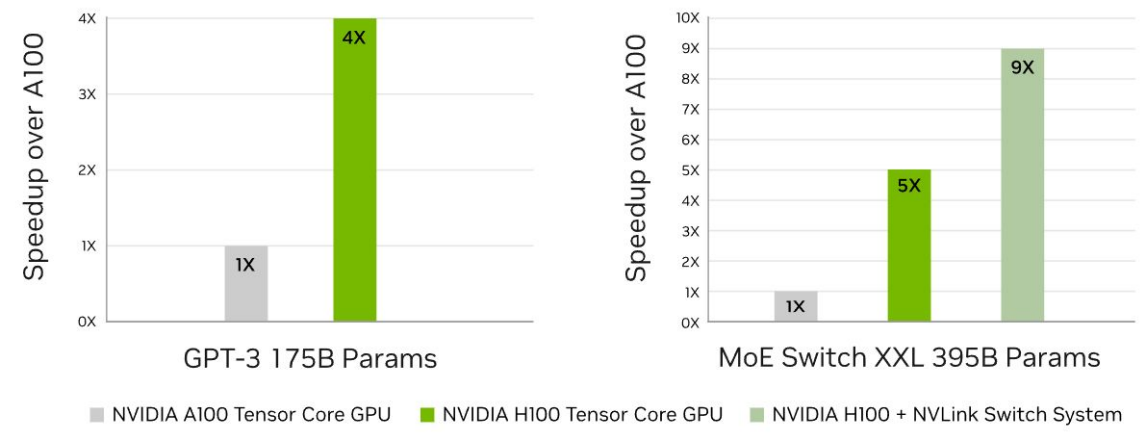
- GPUs perform technical calculations faster and with greater energy efficiency than CPUs
- GPU performance has increased roughly 7,000 times since 2003
- NVLink interconnects allow for massive scaling with very little overhead



TRAINING SPEEDS

NVIDIA H100

- Up to 4x faster AI Training on GPT-3 model
- 350W TDP
- Used by every major AI company



NVIDIA ANNOUNCES BLACKWELL

- Brand new architecture built using TSMC 4NM process
- Flagship is 2 B200 GPU dies connected with super fast 10 terabytes per second (TB/s) chip-to-chip interconnect

04.

FUTURE IMPLICATIONS

What's next???

A LOOK INTO THE FUTURE

- Advancements in Artificial Intelligence
  - Adding sound to videos, more realistic
  - AI learning about the physical world
- Scientific Research
  - Simulations, data analysis, molecular modeling, climate modeling, etc
- VR and AR
  - As GPU improves, as will VR and AR implications and applications
  - Gaming, healthcare, training simulations, etc
- Autonomous Vehicles
  - Sensor fusion, object detection, planning the path
  - Eventually safer travel
- Real-time video analytics
  - Hand-in-hand with autonomous cars
  - Other autonomous vehicles, facial recognition, higher quality streaming

05.

CONCLUSIONS

Takeaways

CONCLUSION

- GPU product turnover has been high since the 1990s
  - Trend will continue
  - Moore's Law - double transistors every 2 years
- Video or reality???
  - Future, or even current videos could trick people
- AI Surge creates a larger demand for GPUs
  - Creates more turnover - desire to be better
- Future implications
  - With better GPUs, other real-world applications will thrive

06.

REFERENCES

REFERENCES

- OpenAI Sora video model: Introducing Sora — OpenAI's text-to-video model
- https://arstechnica.com/information-technology/2024/02/openai-collapses-media-reality-with-sora-a-photorealistic-ai-video-generator/
- https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html
- https://openai.com/research/video-generation-models-as-world-simulators
- https://www.techspot.com/article/650-history-of-the-gpu/
- https://books.google.com/books?hl=en&lr=&id=lfKkEAAAQBAJ&oi=fnd&pg=PR6&dq=history+of+gpus&ots=1TCuhWiJ1N&sig=Pn0cY73OpbvPd1luTaT7FT23tzs#v=onepage&q=history%20of%20gpus&f=false
- https://developer.nvidia.com/gpugems/gpugems2/part-vi-simulation-and-numerical-algorithms/chapter-44-gpu-framework-solving#:~:text=Built%20upon%20efficient%20GPU%20representations,vector%20and%20matrix%2Dvector%20operations.
- https://www.tweaktown.com/news/97140/openai-sora-video-tool-large-scale-deployment-uses-720-000-nvidia-h100-gpus-worth-21-6-billion/index.html