Evolution of Graphical Processing Units (GPU) and GPU Programming

By Leon Johnson
(963653)

INTRODUCTION

Graphical Processing Unit (GPU)

- A highly **computationally parallel** device used in combination with hardware in computers.
- Used to **solve and accelerate** some of computer science's most complex problems.
    - Graphics
    - General Purpose
- The GPU has become ubiquitous in modern day computer systems
- Developed by some of the largest technology companies in the world
- Platform that has sparked innovation of future hardware

(1)

How has the GPU come this far?

Source: (1) https://www.shacknews.com/article/120126/nvidia-rtx-3090-release-date

Focus of Discussion

What led to the GPUs invention?

How does the GPU work?

- How has hardware architecture evolved?
- How has software architecture evolved?

What impact has the GPU had on the computer science industry?

What impact has the GPU had on other industries?

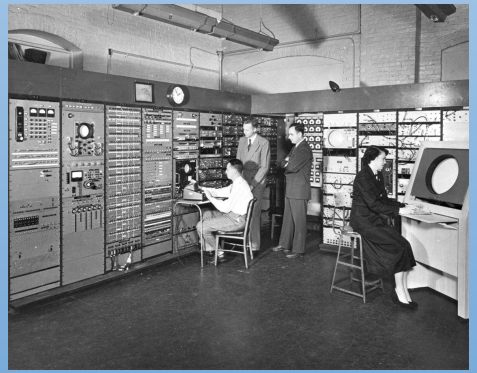
What's next for the GPU?

What led to the GPUs invention?

A historical timeline

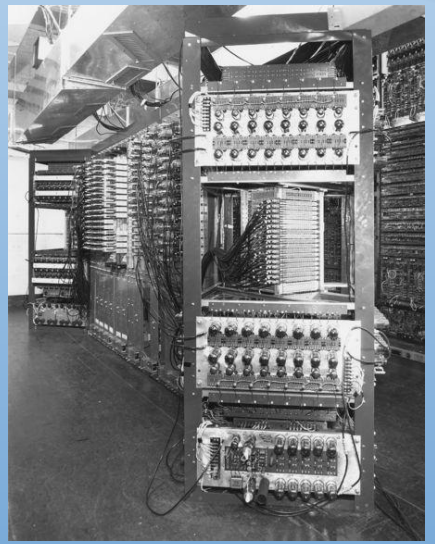
Project Whirlwind (Whirlwind 1) -1951

- A flight simulator developed by MIT during the Cold-War era for the US Navy [1].
- Designed to execute computational cycles in bit-parallel across sixteen units.
- Widely considered to be the first instance of a 3D graphics system.



(1)

Scientists at Whirlwind Test Control



(2)

Whirlwind I Memory Core

Source: (1) https://ethw.org/File:Whirlwind_thf5001.jpg . (2) https://ethw.org/File:Whirlwind_thf2129.jpg

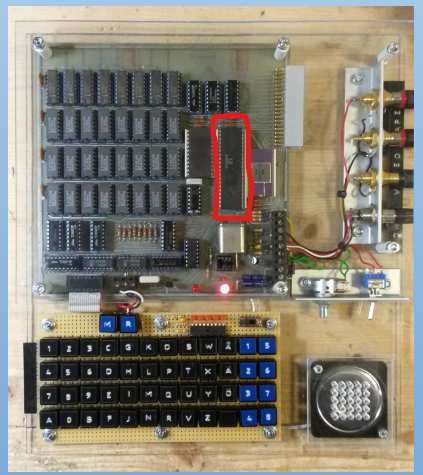
Video Shifters & Video Address Generators – (1970– 1981)

- Emergence of the world's first gaming consoles, era where the foundations of 3D graphics were developed.

- Used as a connection from the CPU to the display, data streams transformed to bitmapped video output values; color, luminosity.

RCA CDP1861 (Pixie) -1975

- Support Chip for the RCA 1802 microprocessor
- Output monochrome bitmapped graphical output at a resolution of 64x128. [2]
- Used in *Telmac 1800* and *Oscom Nano*.



(1)

Telmac 1800, w/ RCACDP1861 (Red Outline)

Source: (1) Original Image: https://en.wikipedia.org/wiki/Telmac_1800#/media/File:Telmac_1800_emolevy.jpg

Television Interface Adapter – 1977

- A “truly unique” video shifter used within *Atari 2600* games console. [3]

- Displayed a range of color palettes (dependant on television signal used).

- Provided input handling from controllers and audio output.

TIA Chip (1)

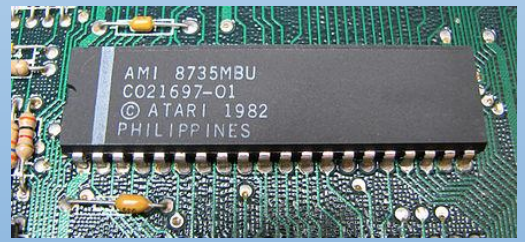
Atari 2600 (2)

Montezuma's Revenge (Atari 2600) (3)

Source: (1) http://www.vgmpf.com/Wiki/index.php?title=File:TIA_-_On_a_2600.jpg (2) https://en.wikipedia.org/wiki/Television_Interface_Adaptor#/media/File:Atari-2600-Wood-4Sw-Set.jpg
(3) https://uk.pcmag.com/gallery/124156/7-forgotten-atari-2600-classics

Alphanumeric Television Interface Controller – 1979

- An integrated circuit chip devoted to generating graphics to a television display.
- Created for use with Atari's 8-bit family of PC's and *Atari 5200.* [4]
- ANTIC generated *"playfield graphics"* but does not process color.
- Passes data to the target *Color Television Interface Adaptor* to output color.



(1)

ANTIC microprocessor on an Atari 130XE motherboard

Source: (1) https://en.wikipedia.org/wiki/ANTIC#/media/File:ANTIC_chip_on_an_Atari_130XE_motherboard.jpg

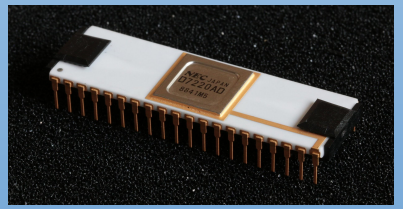
High Performance Graphics Display Controller (NEC 7720) – 1980

- Capable of drawing basic geometry and character bitmaps to a display.

- Developed by NEC Corp. to draw Kanji (Japanese Characters), resolution and sharpness of text was very advanced for the time. [5]

- Chip later licenced by *Intel* and named the 87220.

- Later integrated into the *iSBX 275 Video Graphics Controller Multimode Board* (32KB) output 256x256 (color) or 512x512 (monochrome).



(1)

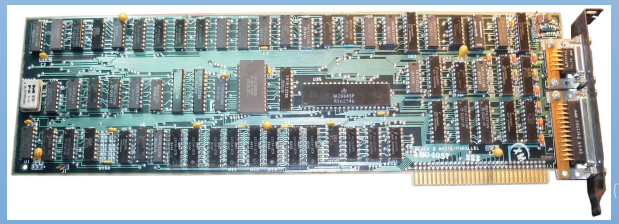
NEC 7720 Chip

Source: (1) https://en.wikipedia.org/wiki/NEC_%C2%B5PD7220#/media/File:NEC_D7220.jpg

IBM's Monochrome Display Adapters & Color Graphics Adapter – 1981

- Used Motorola's MC6845 video address generator. (Handles address code of character buffers and attribute memory)[7]. Job to deliver video values to a CRT assigned to other components.

- MDA and CGA were drivers for the *IBM 5151 Monochrome Display* and *IBM 5153 Color Display* respectively.
- MDA specialised "word processing, spreadsheets, and software development". [8]
- CGA lower resolutions and quality due to outputting color, used in games. [8]



IBM Monochrome Display Adapter (1)

Indiana Jones and the Temple of Doom (CGA) (2)

Source: (1) https://en.wikipedia.org/wiki/IBM_Monochrome_Display_Adapter#/media/File:IBM_PC_Original_Monochrome_Display_and_Parallel_Printer_Adapter.jpg,
(2) https://www.aceinnova.com/en/wp-content/uploads/sites/2/2015/07/indiana-jones-and-the-temple-of-doom.png

The Birth of Array Technology Inc. (ATI) – 1985

- Three Hong Kong nationals residing in Canada founded Array Technology Inc.
- Operating as an OEM for big name clientele such as *IBM* and *Commodore*

(1)

ATI Logo

- ATI released debut product Color Emulation Card (16KB, monochrome, green, amber , white phosphor)
- 1987, released VGA Wonder and EGA Wonder product lines.
- EGA Wonder series 1-4 boasted 256KB of DRAM capable of outputting 16 colors at 640x350.

(2)

ATI VGA Wonder 16

Source: (1) https://www.wikiwand.com/en/ATI_Wonder_series#/MDA/CGA_Cards, (2) https://en.wikipedia.org/wiki/ATI_Technologies#/media/File:ATI-Logo.svg

Graphics Hardware Companies – (1987–1990)

- At this time many new technology companies were founded and product lines shipped.
- Notable companies founded in this era are Realtek, Oak Technology and Trident.
- ATI were big winners in this era capitalising on the booming market of graphics hardware.

(1)

Realtek Semiconductor Corp. Logo

Trident Microsystems Logo (2)

Source: (1) https://en.wikipedia.org/wiki/IBM_Monochrome_Display_Adapter#/media/File:IBM_PC_Original_Monochrome_Display_and_Parallel_Printer_Adapter.jpg,
(2) https://www.aceinnova.com/en/wp-content/uploads/sites/2/2015/07/indiana-jones-and-the-temple-of-doom.png

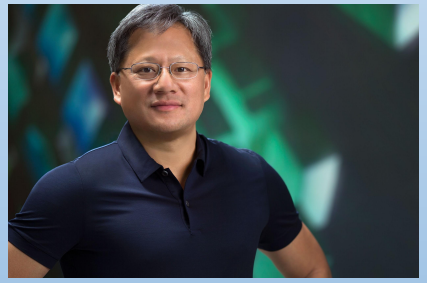
S3’s Porsche 911 (S3 911) – 1991

- Named after the Porsche 911 due to its speed and performance.
- Used to accelerate GUIs in Microsoft's operating systems and architecture.
- “First single chip 2D-accelerator” [9]
- By 1995 all major graphics cards supported 2D acceleration.

(1)

Source: (1) http://www.vgamuseum.info/media/k2/items/cache/49d07d2f2b048709fab28e0845347114_XL.jpg

The Birth of NVIDIA Corporation – 1993



Jensen Huang (1)

Curtis Priem (2)

Chris Malachowsky (3)

- NVIDIA used games as an angle into the graphics industry.

- Conducted large R&D projects in initial operating phase

NVIDIA Initial Logo (4)

Source: (1) https://en.wikipedia.org/wiki/Jensen_Huang#/media/File:Jen-Hsun_Huang_Headshot_(15313247387).jpg (2) https://nvidianews.nvidia.com/bios/chris-a-malachowsky (3) https://www.crunchbase.com/person/curtis-priem (4)https://1000logos.net/nvidia-logo/

Era of 3D – 1995

- Beginning of the modern era of graphical hardware.

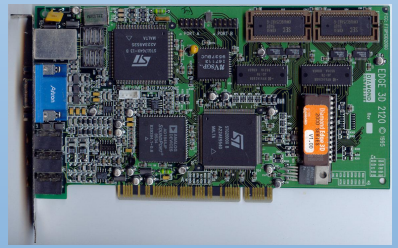
- Large market opportunity with the emergence of 32-bit operating systems and affordability of PCs.

NV1 – 1995

- NVIDIA’s debut product, with the assistance from SGS-Thomson Microelectronics.
- The multimedia PCI card boasted a complete 2D/3D graphical core, with on board VRAM or FPM DRAM.
- Uniquely used quadratic-polygon rendering approaches.
- Eventually sold to Diamond and distributed as Diamond Edge 3D.

(1)

The NV1



(2)

Diamond Edge 3D 2120

(3)

Daytona USA, accelerated by NV1

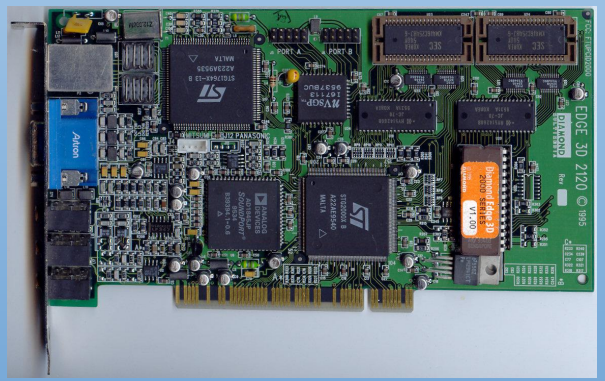
- Microsoft released that their DirectX platform was based on triangle-polygon rendering methods.

Source: (1) https://en.wikipedia.org/wiki/NV1#/media/File:NVIDIA@500nm@Fixed-pipeline@NV1@STG2000X_B_A22FN9609_MALTA_DSC05885_(27662108651).jpg (2) https://en.wikipedia.org/wiki/NV1#/media/File:Dedge3d.jpg
(3) https://vintage3d.org/nv1.php

3DFx Voodoo – 1996

- Launched in Q4 of 1996 Voodoo Graphics was able to release to consumer market, due to cheaper prices of EGA DRAM. (4MB)
- Dedicated 3D accelerator, no 2D acceleration not supported. [9]
- Featured a frame buffer processor and texture mapping unit.
- Accelerated popular games of the time such as Wayne Gretzky's 3D Hockey
- Lead 85% of the market with software developers and consumers by late 1997.

Wayne Gretzky's 3D Hockey


Diamond Monster 3D w/ Voodoo 1 Chipset

Source: (1) https://www.hockeywilderness.com/2014/8/8/5981271/hw-video-game-week-wayne-gretzky-3d-hockey (2) https://en.wikipedia.org/wiki/3dfx_Interactive#/media/File:KL_Diamond_Monster3D_Voodoo_1.jpg

Rendition’s Vérité – 1996

- Industry was seeing more use of 2D GUI acceraters with 3D rasterization boards.
- Rendition capitalized with unifying two chips into a relatively fast, single-board solution.
- Vérité was one of the industry’s first acceptable implementations of the 2D/3D unification
- Equipped with 4MB of EDA DRAM, supporting Microsoft’s DirectX 3D.

(2)

Sierra’s Screamin’ 3D w/ Verite V1000

Source: (1) https://www.sierragamers.com/screamin-3d-1996/

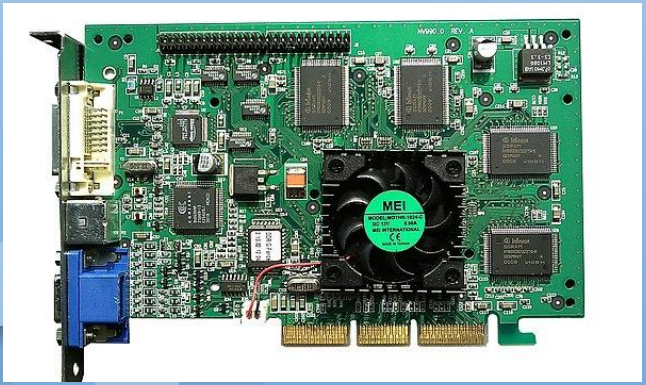
The World's First GPU – 1999

- NVIDIA released their GeForce 256, branded as a "graphical processing unit".

"A single chip processor with integrated transform, lighting, triangle setup/clipping, and rendering engines that is capable of processing a minimum of 10 million polygons per second". [10]

NVIDIA CORPORATION

- Defining features were the ability to perform transform and lighting operations.



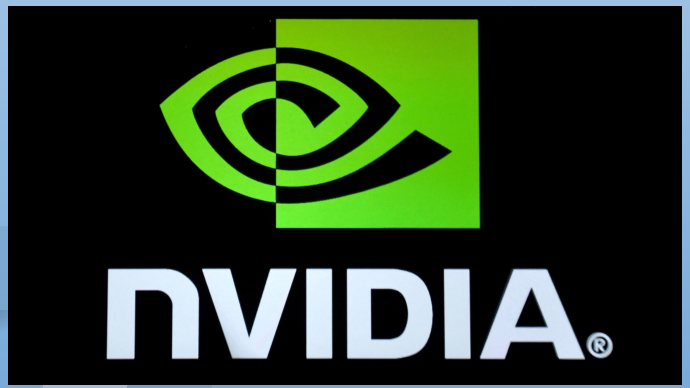
GeForce 256 (2)

NVIDIA Grass Demo (GeForce 256) (2)

Source: (1) https://en.wikipedia.org/wiki/GeForce_256#/media/File:VisionTek_GeForce_256.jpg (2) Snapshot from Demo Grass:: https://www.nvidia.com/en-gb/geforce/community/demos/

The Great War: NVIDIA vs ATI – 2000's

- By the early 2000 NVIDIA and ATI had established themselves as front runners of the industry.
- Battling on grounds of graphics software and hardware architectures
- More and more new features being developed to GPU devices, such as; (vertex blending, shadow volumes, refraction)



NVIDIA Logo (1)

VS

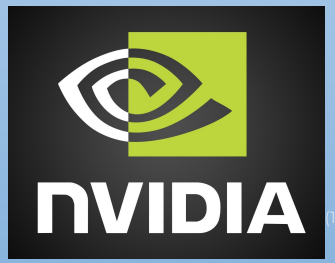
ATI Logo (2)

Source: (1) https://www.pcmag.com/news/is-the-rtx-3060-coming-nvidia-tips-jan-12-geforce-event (2) https://en.wikipedia.org/wiki/ATI_Technologies#/media/File:ATI-Logo.svg

General Purpose GPU – (2006 – Present)

- GPUs previously used to solve graphical problems, i.e Rendering.

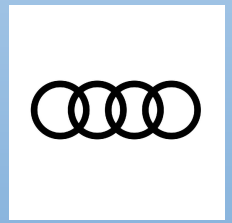
- Major graphical companies such as NVIDIA and ATI (later AMD) saw that the GPU's parallel properties could be applied to other fields.



(1)

NVIDIA Logo

+



(2)

Audi Logo

(3)

OpenCL Logo

(4)

NVIDIA CUDA Logo

Source: (1) https://windowsreport.com/nvidia-kernel-mode-driver-error/ (2) https://www.audi-mediacenter.com/en/photos/detail/audi-logo-new-corporate-design-1282 (3) https://en.wikipedia.org/wiki/OpenCL#/media/File:OpenCL_Logo.svg (4) http://www.cgchannel.com/2019/11/nvidia-drops-macos-support-for-cuda/

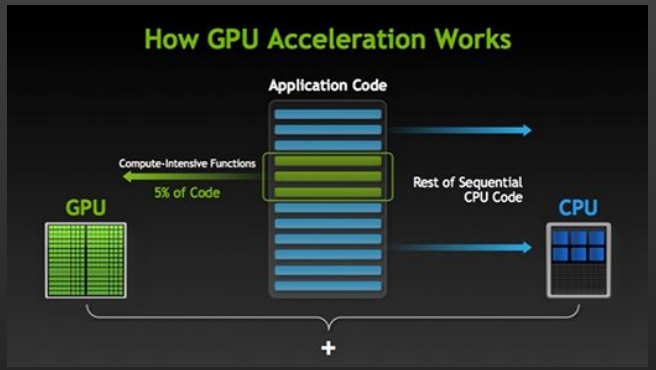
How does the GPU work?

Hardware Overview

- GPU used in combination with CPU (heterogeneous computing)
- GPUs have huge number of cores compared to the CPU.
- Computationally expensive code executed in parallel by GPU, sequential application code handled by CPU.

(1)

CPU vs GPU Architectural Comparison



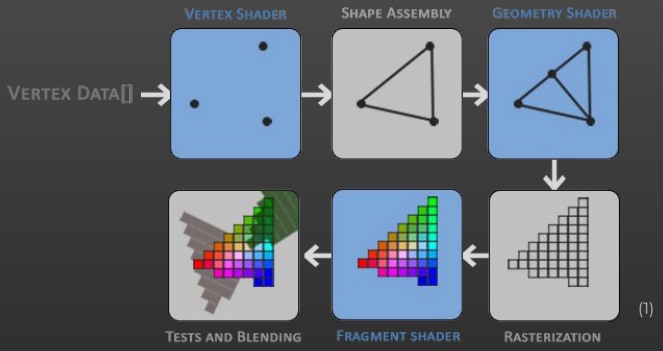
(2)

Application Acceleration w/ GPU

Source: (1)(2) https://nyu-cds.github.io/python-gpu/01-introduction/

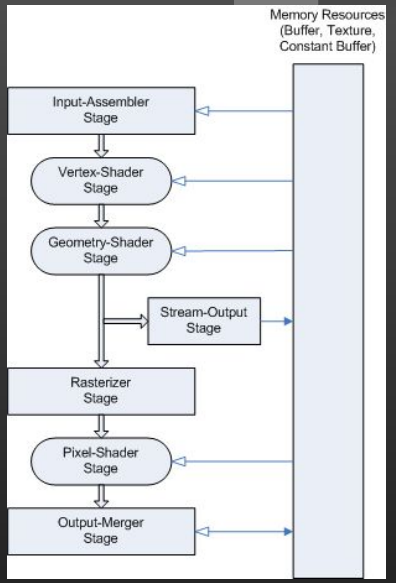
Software Overview - Graphical Pipeline

- Conceptual model outlining the protocol a graphics program issues to render a 3D scene onto a 2D display.



(1)

OpenGL Pipeline Overview



(2)

Direct3D 10 Pipeline

Source: (1) https://learnopengl.com/Getting-started/Hello-Triangle (2) https://docs.microsoft.com/en-us/windows/win32/direct3d10/d3d10-graphics-programming-guide-pipeline-stages

Software Overview – API's

Special languages created to program the GPU and harness the parallelism.

Most popular within the industry for rendering; DirectX , OpenGL , Vulkan,

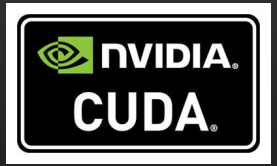
Have accompanying shader languages to program pipelines.

(1)

(2)

(3)

- General purpose APIs: CUDA, OpenCL



(4)

(5)

Source: (1) https://en.wikipedia.org/wiki/DirectX#/media/File:Microsoft-DirectX-Logo-wordmark.svg (2) (3) https://www.khronos.org/legal/trademarks/ (4 https://en.wikipedia.org/wiki/CUDA#/media/File:Nvidia_CUDA_Logo.jpg)
(5) https://en.wikipedia.org/wiki/OpenCL#/media/File:OpenCL_Logo.svg

How does the GPU work?

NVIDIA’s Implementation

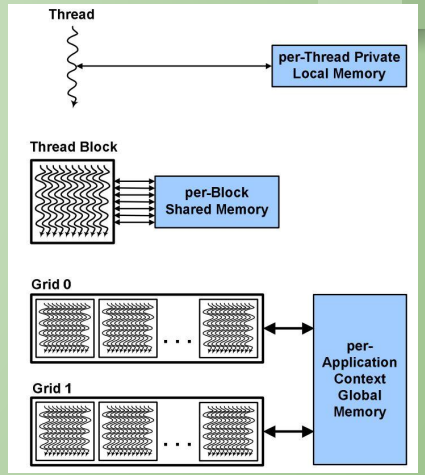
CUDA (Compute Unified Device Architecture)

- NVIDIA's software and hardware architecture, enables NVIDIA GPUs to be programmed by languages such as C, C++ and OpenCL.

- CUDA programs are instantiated across a grid of parallel thread blocks, consisting of parallel threads.

- Threads within a block read and write to the same shared memory block.

  - Grids output results into global memory space after kernel wide global synchronization calls.

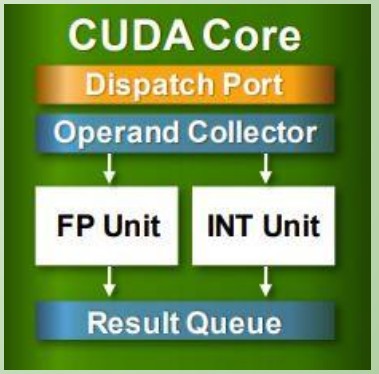


CUDA Thread, Thread Block, Grid Hierarchy (1)

Source: (1) [11]

CUDA Cores

- CUDA's architecture is extended by CUDA cores, which are contained within *Streaming Multiprocessor* units.
- Execute integer and floating point arithmetic instructions.
- Have independent logic, move and compare units.



(1)

CUDA Core Architecture

Source: (1) [11]

Streaming Multiprocessors

- SM's consist of large amounts of CUDA cores, have internal schedulers, registers, and L1 cache memory shared across all CUDA cores.

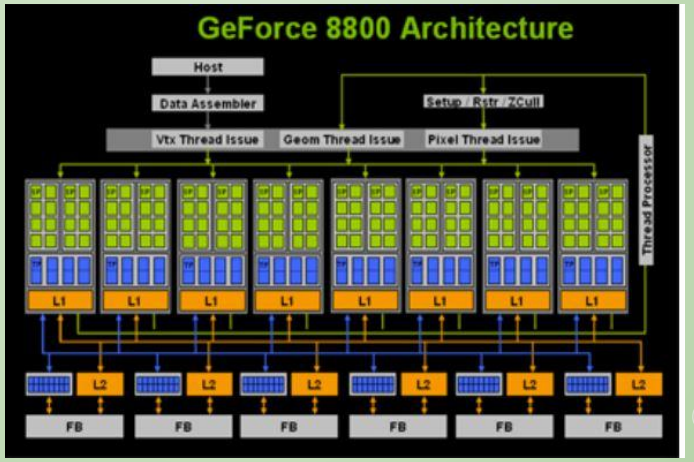
- Component that perform the actual computation.

SM Architecture (Fermi)

Source: (1) [11]

NVIDIA G80 Architecture (2006)

- Revolutionary update to GPU architecture, large revisions to pipeline model and dataflow.
- Altered sequential flow of execution to be more loop oriented [12]



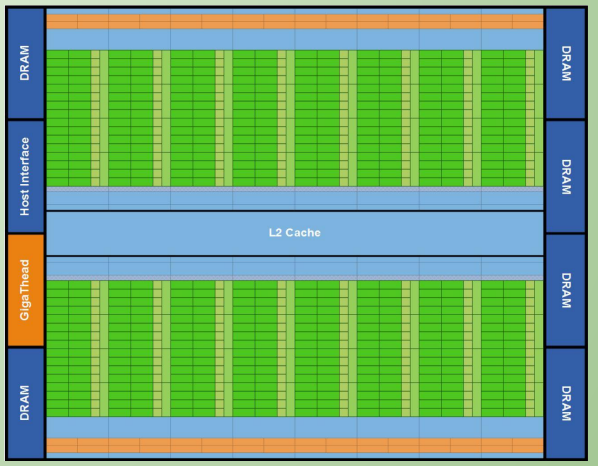
(1)

GeForce 8800 Architecture

Source: (1) [12]

NVIDIA Fermi Architecture (2010)

- Greatest architectural design for GPUs at the time of release.
- Fermi featured 3 billion transistors with 512 CUDA cores, batches of 32, within 16 SM units.
- 384-bit memory interface (6 partitions of 64-bit memory) with 6GB of GGDDR5 DRAM.
- GPU and CPU are connected via PCI-Express bus



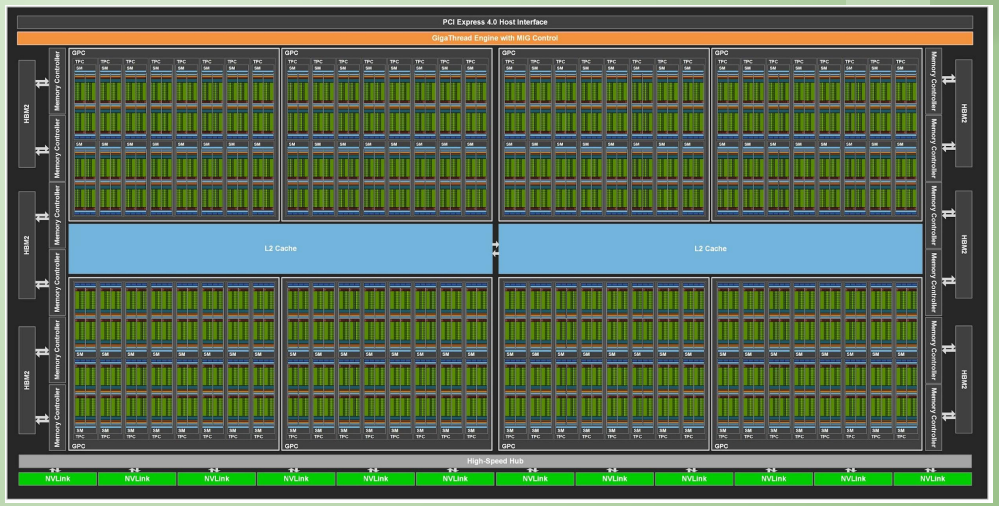
(1)

Fermi Architectural Design

Source: (1) [12]

NVIDIA Ampere Architecture (2020)

- Used as “data centers” to train complex deep learning models and accelerating AI training.
- Massive parallelization, contains 128 SM’s, (SM’s also have refined internal architecture)
- 40GB of HBM (High Bandwidth Memory).



(1)

Fermi Architectural Design

Source: (1) [12]

What impact has the GPU had on the computer science industry?

Rendering **Timeline**

- Ray Casting (1968)
- Gouraud (Per Vertex) Shading (1971)
- Phong (Per Fragment) Shading (1973)
- Blinn Shading (1977)
- Ray Tracing (1980)
    - Radiosity (1984)
        - Transform, clipping , lighting (1993)
        - Directional Lighting (1993)
        - Z-culling (1993)
            - Ambient Occlusion (1994)
            - Precomputed Radiance Transfer (2002)

Rendering

Examples

Tomb Raider (1996)

(1)

VS

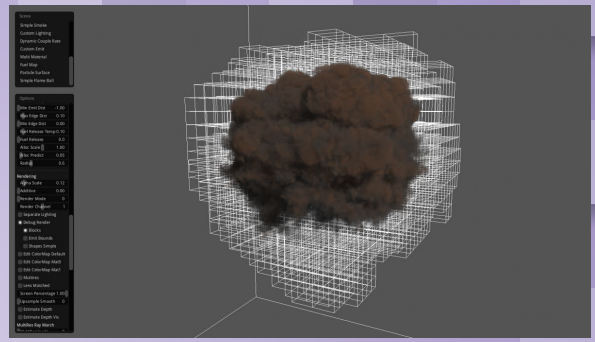
Shadow of the Tomb Raider: The Nightmare (2013)

(2)

Source: (1)https://www.esquire.com/uk/culture/news/a14554/play-tomb-raider-web-browser/ (2) https://square-enix-games.com/en_GB/news/path-home-shadow-tomb-raider-dlc-now-available

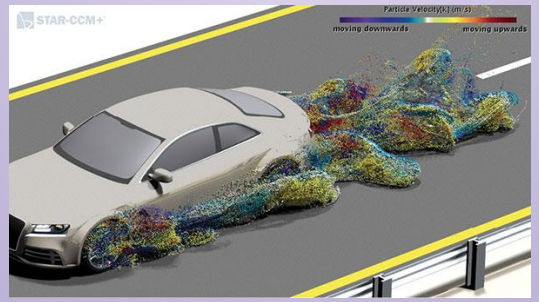
Simulations

- **Computational Fluid Dynamics**: Fire, Water, Gas
- **Particle Analysis**
- **Aerospace:** Flight simulations, wind tunnels etc.



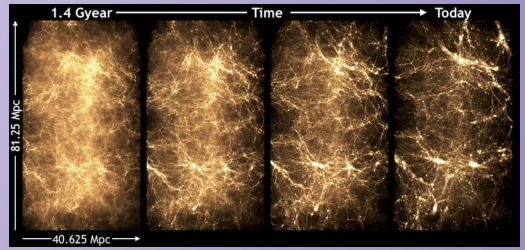
(1)

NVIDIA Flow Framework (Fluid Dynamics)



(2)

Lagrangian Mass Particle Simulation



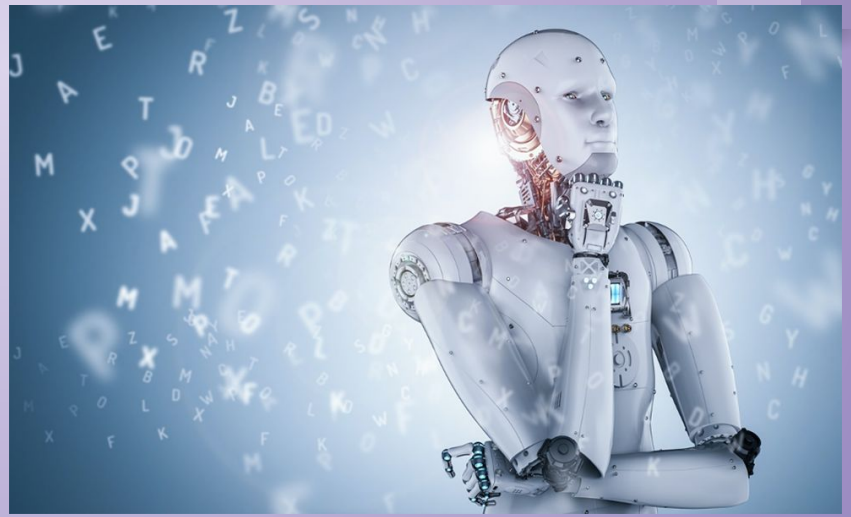
(3)

GPU-Accelerated Cosmological Analysis on the Titan Supercomputer (NVIDIA)

Source: (1)https://developer.nvidia.com/nvidia-flow (2) https://www.digitalengineering247.com/article/the-expanding-role-of-gpus-in-simulation (3) https://developer.nvidia.com/blog/gpu-accelerated-cosmological-analysis-titan-supercomputer/

Artificial Intelligence and Deep Learning

- Extremely large and complex training models and data can be processed in parallel.
- GPUs large bandwidth drastically speeds up processes.



Source: (1)https://developer.nvidia.com/nvidia-flow (2) https://www.digitalengineering247.com/article/the-expanding-role-of-gpus-in-simulation (3) https://developer.nvidia.com/blog/gpu-accelerated-cosmological-analysis-titan-supercomputer/

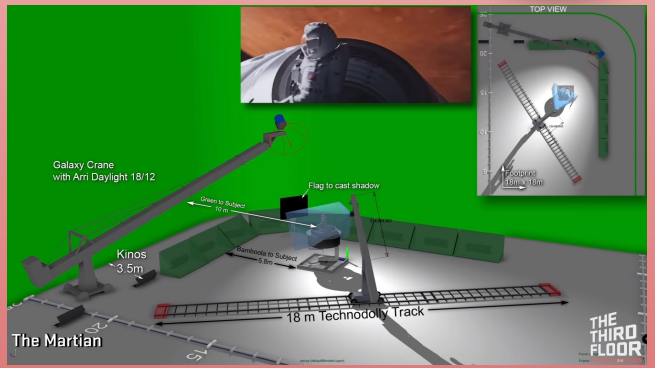
What impact has the GPU had on other industries?

Film

- Computer Generated Imagery

- Virtual Production (Pre, Post, Full)

CGI Render of Sonic The Hedgehog [1]



Snippet From: Virtual Production: A New Era of Filmmaking | Unreal Engine [2]

Source: (1) https://www.menshealth.com/entertainment/a27333297/creepy-cgi-characters/ (2) Snippet From: https://www.youtube.com/watch?v=voVPO8w9-x4

Automotive

- Navigation Optimization
- Self Driving Cars
- Enhanced/Accelerated Dashboard Screens

Tesla Interior (NVIDIA Pascal GPU) (1)

Source: (1) https://electrek.co/2019/10/04/tesla-bezel-less-display-technology-holographic/

Finance (Crypto-currencies)

- Crypto-Mining
- Real-Time Stock Market Processing
- Finance App Acceleration

(1)

Crypto-mining Rig (NVIDIA GeForce RTX 3080)

Source: (1) https://videocardz.com/newz/this-geforce-rtx-3080-ethereum-mining-rig-now-makes-20k-per-month

Personal Contributions

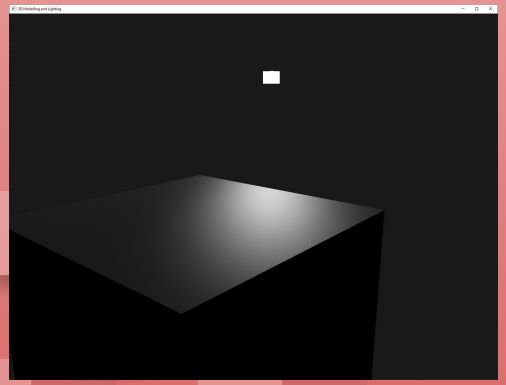
- Gouraud, Phong and Blinn Phong Shader and Renderer (OpenGL, C++)

- Phong Shader (DirectX12, C++)

- 3D Rendering and Simulation of Fluid (DirectX12, C++) , TBC

```glsl
struct Material {
    vec3 ambient;
    vec3 diffuse;
    vec3 specular;
    float shininess;
};


in VERTEX_SHADER_OUTPUT
{
    vec3 normal;
    vec3 fragment_position;
} fragment_shader_input;

uniform vec3 view_pos;
uniform Material material;
uniform Light light;

void main()
{
    //Ambient Component
    vec3 ambient = material.ambient * light.ambient;

    //Diffuse Component
    vec3 norm = normalize(fragment_shader_input.normal);
    vec3 light_direction = normalize(light.position - fragment_shader_input.fragment_position);

    float diff = max(dot(norm, light_direction), 0.0);
    vec3 diffuse =  (diff * material.diffuse) * light.diffuse;

    //Specular Component
    vec3 view_direction = normalize(view_pos - fragment_shader_input.fragment_position);
    vec3 reflect_direction = reflect(-light_direction, norm);
    float spec = pow(max(dot(view_direction, reflect_direction), 0.0), material.shininess);
    vec3 specular = (spec * material.specular) * light.specular;

    //Combination
    vec3 result = ambient + diffuse + specular;
    fragment_colour = vec4(result, 1.0);
}
```

Phong Pixel Shader (GLSL)



Blinn-Phong 3D Render

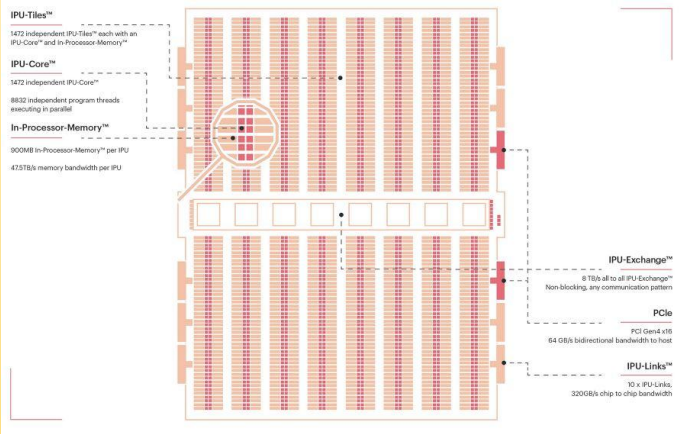
What’s next for the GPU?

- More and More Cores!

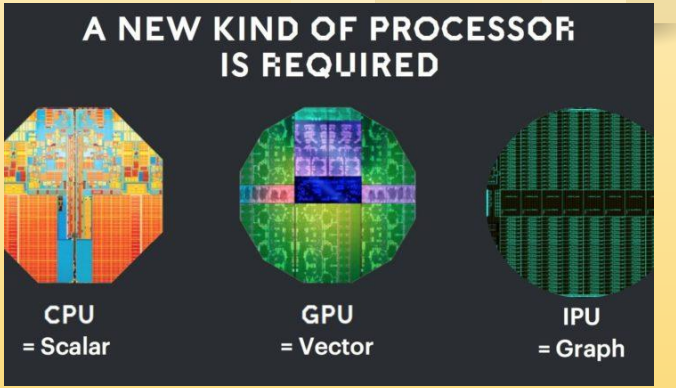
- Revisions to pipeline and data structures (more control to GPU programmer)

New Hardware?

- Intelligence Processing Unit (IPU) by Graphcore



Colossus™ MK2 GC200 IPU (1)



Processor Conceptual Comparison (2)

- Optimized for machine learning, deep learning and AI.

Source: (1)https://www.graphcore.ai/products/ipu (2) https://www.nextbigfuture.com/2019/02/graphcore-chips-could-speed-up-ai-by-100-times.html

Bibliography

[1] T. M. S. Kent C. Redmond, Project Whirlwind : The History of a Pioneer Computer, Digital Press, 1980.
[2] RCA, RCA 1800 Microprocessors, Design Ideas Book, RCA, 1975.
[3] I. B. Nick Montfort,Racing the Beam: The Atari Video Computer System.MIT Press, 2009
[4] Atari Inc., Data processing system with programmable graphics generator, 1979, U.S Patents.
[5] (Online) NEC μPD7220 Graphics Display Controller: The first graphics processor chip, https://www.electronicdesign.com/technologies/embedded-revolution/article/21122304/jon-peddie-research-vol-1-no-1-nec-pd7220-graphics-display-controller-the-first-graphics-processor-chip
[6] Intel Corporation, iSBX 275 Video Graphics Controller Multimodule Board Reference Manual, 1982
[7] D. C. Lili Zhao, “Research and design of CRT controller based on CPLD,” 2012
[8] E. D. Larry Press, “IBM PC,” 2003
[9] L Kelty, P Beckett, L Zalcman, “Desktop simulation,” 1999.
[10] https://www.nvidia.com/en-us/about-nvidia/corporate-timeline/
[11] NVIDIA Corporation, NVIDIA’s Next Generation CUDA TM Compute Architecture: Fermi, 2009
[12] Technical Brief - NVIDIA GeForce 8800 GPU Architecture Overview,” 2006

Final Thoughts...

Thank you for listening!

Any Questions?