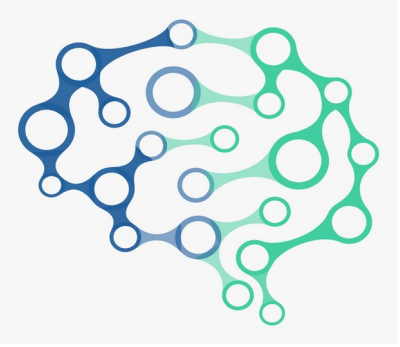
Hardware and Software for AI
GPU

Tushar B. Kute,
http://tusharkute.com



GPU

- A graphics processing unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
- GPUs are used in embedded systems, mobile phones, personal computers, workstations, and game consoles.

GPU

- Modern GPUs are very efficient at manipulating computer graphics and image processing.
- Their highly parallel structure makes them more efficient than general-purpose central processing units (CPUs) for algorithms that process large blocks of data in parallel.
- In a personal computer, a GPU can be present on a video card or embedded on the motherboard. In certain CPUs, they are embedded on the CPU die.

GPU

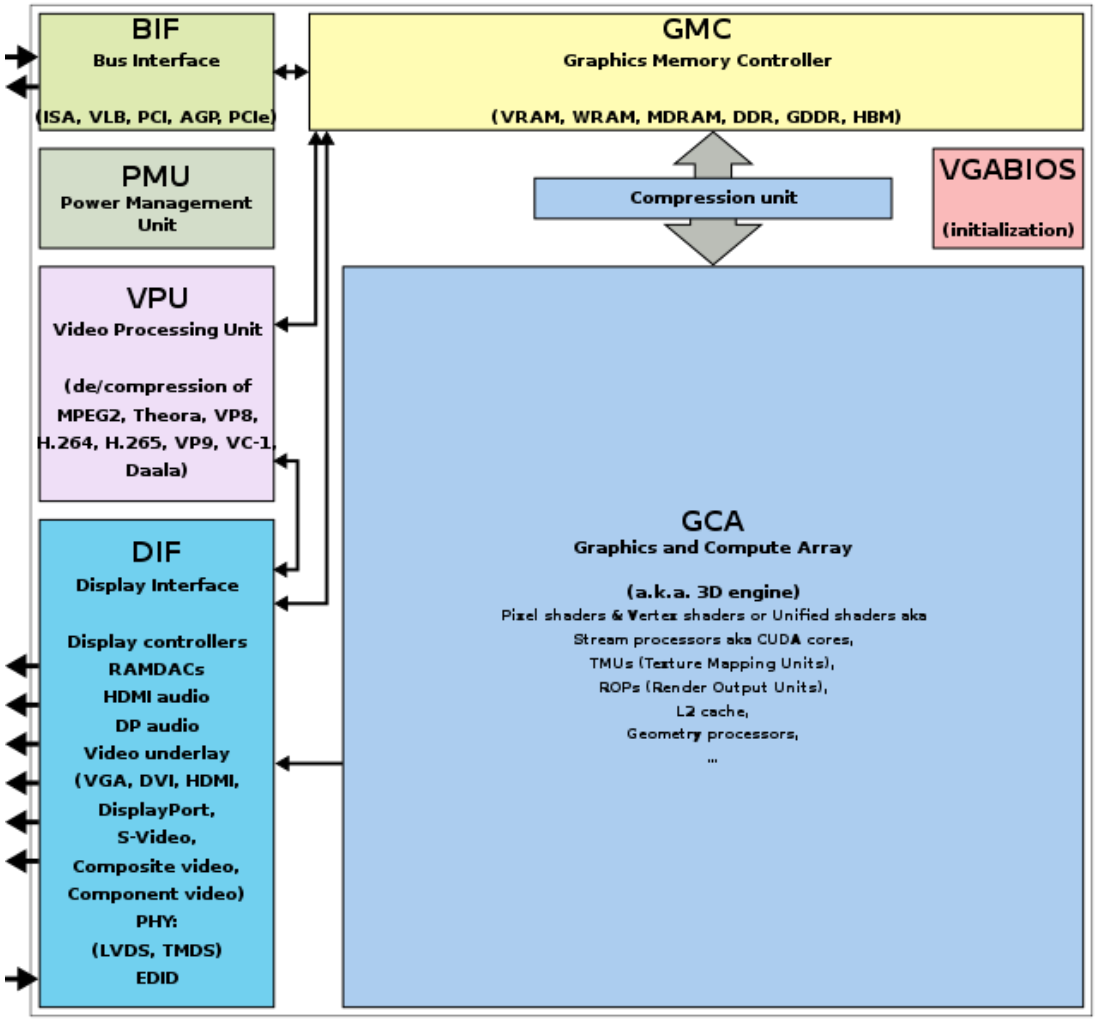
- In the 1970s, the term "GPU" originally stood for graphics processor unit and described a programmable processing unit independently working from the CPU and responsible for graphics manipulation and output.
- Later, in 1994, Sony used the term (now standing for graphics processing unit) in reference to the PlayStation console's Toshiba-designed Sony GPU in 1994.
- The term was popularized by Nvidia in 1999, who marketed the GeForce 256 as "the world's first GPU".

GPU

GPU: Companies

- Many companies have produced GPUs under a number of brand names. In 2009, Intel, Nvidia and AMD/ATI were the market share leaders, with 49.4%, 27.8% and 20.6% market share respectively.
- However, those numbers include Intel's integrated graphics solutions as GPUs. Not counting those, Nvidia and AMD control nearly 100% of the market as of 2018. Their respective market shares are 66% and 33%.
- In addition, Matrox produce GPUs. Modern smartphones also use mostly Adreno GPUs from Qualcomm, PowerVR GPUs from Imagination Technologies and Mali GPUs from ARM.

Architectures



Computational functions

- Modern GPUs use most of their transistors to do calculations related to 3D computer graphics. In addition to the 3D hardware, today's GPUs include basic 2D acceleration and framebuffer capabilities (usually with a VGA compatibility mode).
- Newer cards such as AMD/ATI HD5000-HD7000 even lack 2D acceleration; it has to be emulated by 3D hardware.
- GPUs were initially used to accelerate the memory-intensive work of texture mapping and rendering polygons, later adding units to accelerate geometric calculations such as the rotation and translation of vertices into different coordinate systems.

NVIDIA

- Nvidia is a US technology company based in California, founded in 1993, that designs GPUs for gaming and professional markets, as well as system on a chip units (SoCs) for the mobile computing and automotive markets. Its primary GPU line is GeForce, which is a direct competitor to AMD's Radeon.
- Some of its well-known GPUs include the GeForce RTX 3080, Nvidia Titan V, and the Nvidia RTX A6000. Right now, due to a worldwide shortage of graphic cards, it can be tricky to get your hands on one of its GPUs, especially its GeForce RTX 3070 cards.
- In July 2021, Nvidia switched on what it claimed to be the UK's fastest supercomputer, the Cambridge-1, which contained a number of NVIDIA A100 Tensor Core GPUs.

AMD

- Advanced Micro Devices, also known as AMD, was founded in 1969 and is another US tech company based in California. It develops computer processors and other products for business and consumer markets.
- It originally manufactured semiconductors before spinning off this division in 2008, which was then formed into GlobalFoundries.
- AMD's main products include motherboard chipsets, microprocessors, graphics processors, and embedded processors. Some of its products include the AMD Radeon graphics series and the Ryzen processor range.

GPU for AI

- GPUs are not mandatory for AI/ML workloads. Some customers are actually using CPUs or other accelerators for certain such activities. However, GPUs are extremely efficient for these workloads and will often show the best performance, and here is why:
  - GPUs can perform multiple, simultaneous computations
  - GPUs are highly efficient at these types of calculations
  - GPUs contain massive amounts of cores that can be used efficiently in parallel

Design Factors

- What questions do we want the model to answer?
- What data is relevant in order to answer these questions?
- Where can the data be acquired from?
- Where will we store the data, and how will we move the data around?
- What is the expected dataset size?
- How will we write the model or can we use an existing model?
- How will we validate the model and deploy new models?

Use Cases

- Possibilities abound when it comes to AI/ML use cases, so here is just a sampling:
  - Compliance–Updating the compliance team on new regulations and verifying that the organization abides by these regulations
  - Cyber Security–Making sure that an organization is secure by reviewing multiple video inputs in near real time and constantly scanning for and preventing breaches
  - Fraud Detection in Financial Services–Going over massive sets of financial data to detect and expose possible fraud
  - Conversational AI–Augmenting customer service centers to communicate with customers

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
http://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com