5.3 GRAPHICS PROCESSING UNITS

GPU is designed to lessen the work of the CPU and produce faster video and graphics. GPU can be thought as an extension of CPU with thousands of cores. A GPU is extensively used in a PC on a video card or motherboard, mobile phones, display adapters, workstations and game consoles. They are mainly used for offloading computation intensive application. This is also known as a **visual processing unit (VPU).**

A Graphics Processing Unit (GPU) is a single-chip processor primarily used to manage and boost the performance of video and graphics. It is a dedicated parallel processor for accelerating graphical and deeper computations.

Differences between CPU and GPU

GPU	**CPU**
They facilitate highly parallel operations.	This supports serial execution of programs.
This has more number of cores(in thousands).	This has less number of cores.
They need special faster interfaces to facilitate faster data transfers.	No such special interface are required.
They have deeper pipelines	They have comparatively shallow.

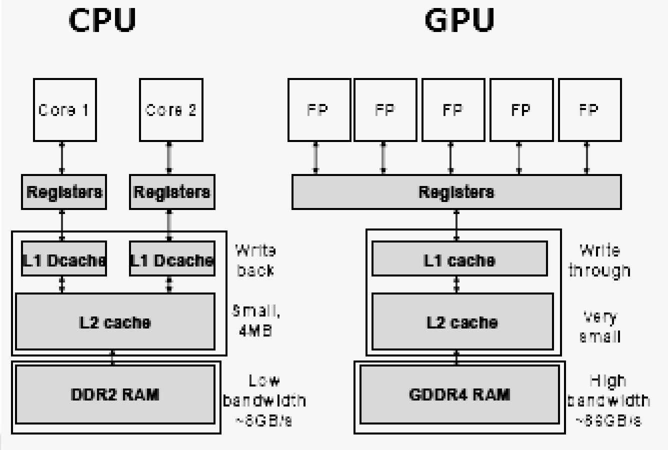

Fig 1: CPU vs GPU architecture

Source: Miles J. Murdocca and Vincent P. Heuring, — "Computer Architecture and Organization: An Integrated approach"

GPU features

The following are prominent features of GPU:

- 2-D or 3-D graphics
- Digital output to flat panel display monitors
- Texture mapping
- Application support for high-intensity graphics software such as AutoCAD
- Rendering polygons
- Support for YUV color space
- Hardware overlays
- MPEG decoding

Development of GPU

- The first GPU was developed by NVidia in 1999 and named as GeForce 256.
- This GPU model could process 10 million polygons per second and had more than 22 million transistors.
- This is a single-chip processor with integrated transform, drawing and BitBLT support, lighting effects, triangle setup /clipping and rendering engines.

- The GPU is connected to the CPU and is completely separate from the motherboard.
- The RAM is connected through the Accelerated Graphics Port (AGP) or the PCI express bus.
- Sometimes, GPUs are integrated into the north bridge on the motherboard and use the main memory as a digital storage area, but these GPUs are slower and have poorer performance.
- The accelerated memory in GPU is used for mapping vertices and can also supports programmable shade implementing textures, mathematical vertices and accurate color formats.
- Applications such as Computer-Aided Design (CAD) can process over 200 billion operations per second and deliver up to 17 million polygons per second.
- The main configurations of GPU processor are: Graphics coprocessor which is independent of CPU and Graphics accelerator that is based on commands from CPU.

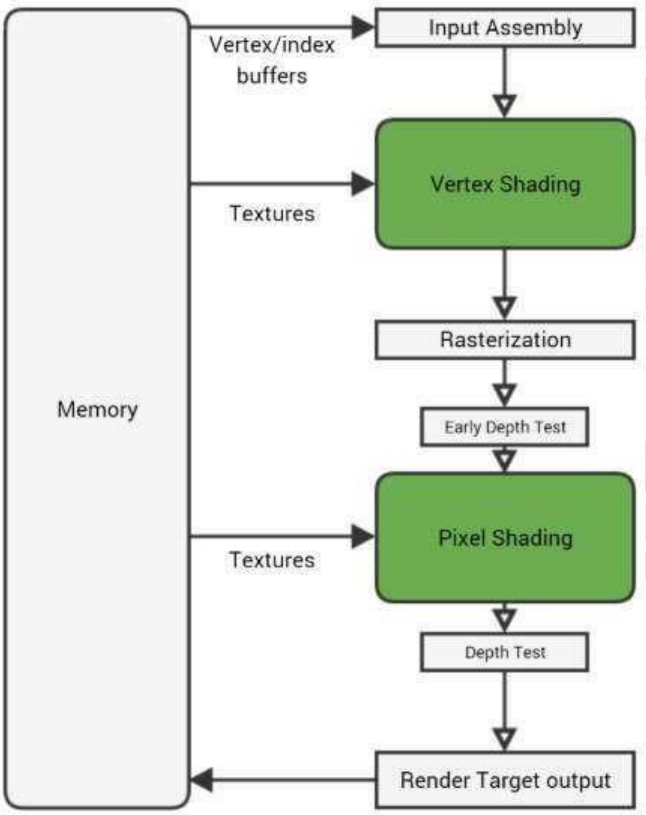

Fig 2: GPU Pipeline

Source: Miles J. Murdocca and Vincent P. Heuring, — "Computer Architecture and Organization: An Integrated approach"

Input Assembler stage

- This stage is the communication bridge between the CPU and GPU.
- It receives commands from the CPU and also pulls geometry information from system memory.
- It outputs a stream of vertices in object space with all their associated information.

Vertex Processing

- This processes vertices performing operations like transformation, skinning and lighting.
- A vertex shade takes a single input vertex and produces a single output vertex.

Pixel Processing

- Each pixel provided by triangle setup is fed into pixel processing as a set of attributes which are used to compute the final color for this pixel.
- The computations taking place here include texture mapping and math operations

Output Merger Stage

- The output-merger stage combines various types of output data to generate the final pipeline result.