



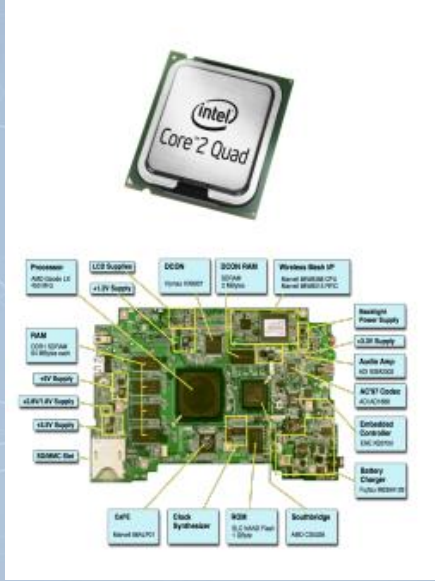
Computer Graphics Hardware

CSU0021: Computer Graphics

Graphics System



Input device



CPU/Memory



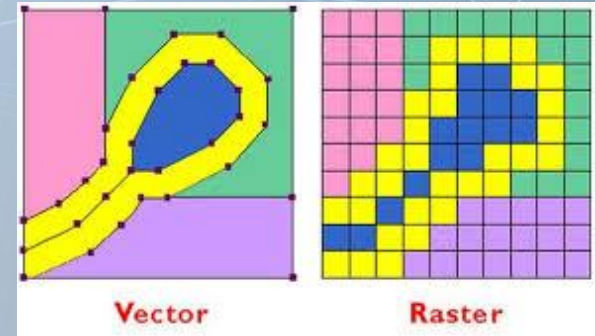
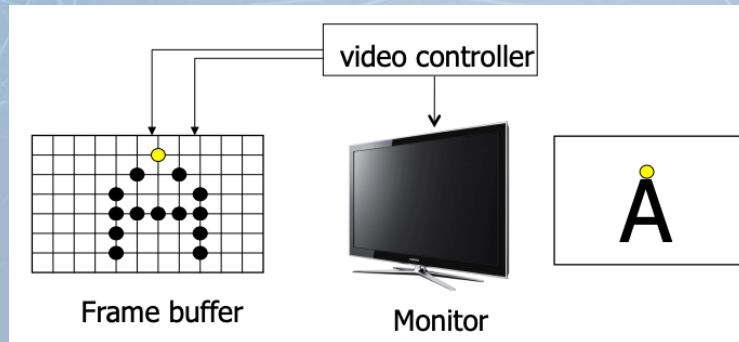
GPU/Memory



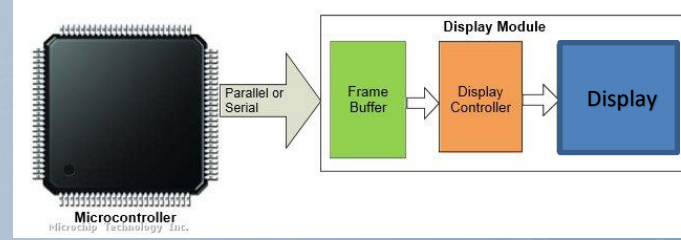
Monitor

Raster Graphics System

- Raster: An array of picture elements
- Based on raster-scan TV technology
- The screen or a picture consists of discrete pixels, and each pixel has a small display area.

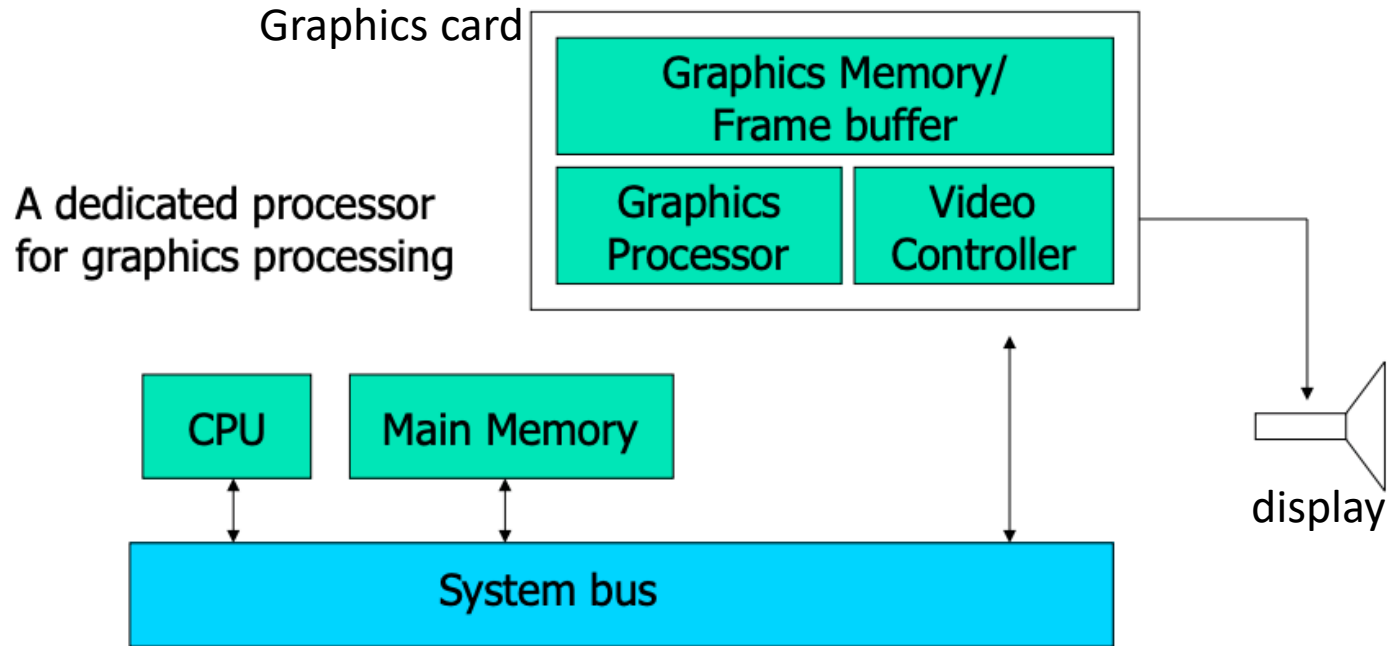


Frame Buffer



- Frame buffer: the memory to hold the pixel properties (color, alpha, depth, stencil mask, etc)
- Properties of a frame buffer that affect the graphics performance
 - Size: screen resolution
 - Depth: color level
 - 1 bit/pixel: black and white
 - 8bits/pixel: 256 levels gray or color pallet index
 - 24bits/pixel: 16 million colors
 - Speed: refresh rate

Graphics Acceleration (card)



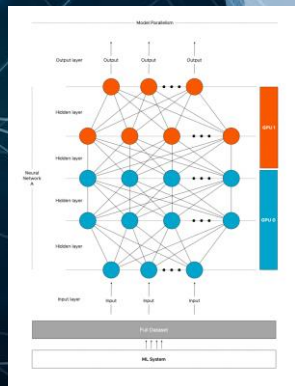
Graphics Accelerator (card)



What do GPUs (graphics card) do?



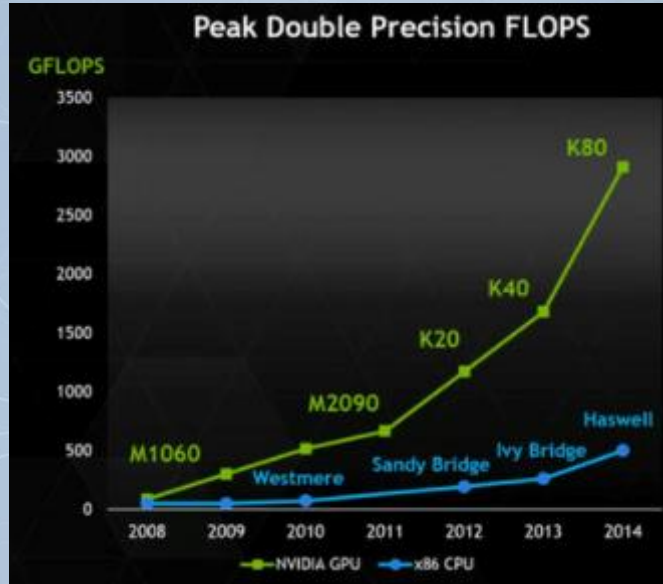
- GPUs are massively parallel processors
 - Process geometry/pixels and produce images to be displayed on the screen
 - Can also be used to perform general purpose computation (via CUDA/ OpenGL)



- Evolved from simple video scan controllers, to special purpose processors that implement a simple pipeline with fixed graphics functionality, to complex many-core architectures that contain several deep parallel pipelines
 - Example: Nvidia Tesla V100 has 5120 cores and 21.1 billions transistors
 - Nowadays, a graphics card can easily have more than 4 GB of video memory

The diagram illustrates the NVIDIA HGX H100 architecture, showing a 3x3 grid of compute nodes. Each node contains three GPCs (Graphics Processing Clusters), each with 10 SMs (Streaming Multiprocessors). The nodes are interconnected via NVLink (Non-Volatile Link) technology. The top and bottom edges of the node grid are connected to a High-Speed Hub. The left and right edges are connected to Memory Controllers. The entire system is connected to a PCI Express 3.0 Host Interface. The diagram is color-coded: orange for the host interface, blue for memory controllers, green for NVLink, and grey for the high-speed hub.

CPU/GPU Performance Gap



Why are GPU's so fast?

- Entertainment industry has driven the economy of these chips
 - Recently, deep learning has driven these economy, too
- Moore's Law
- Simplified design (stream processing)
- Single-chip designs

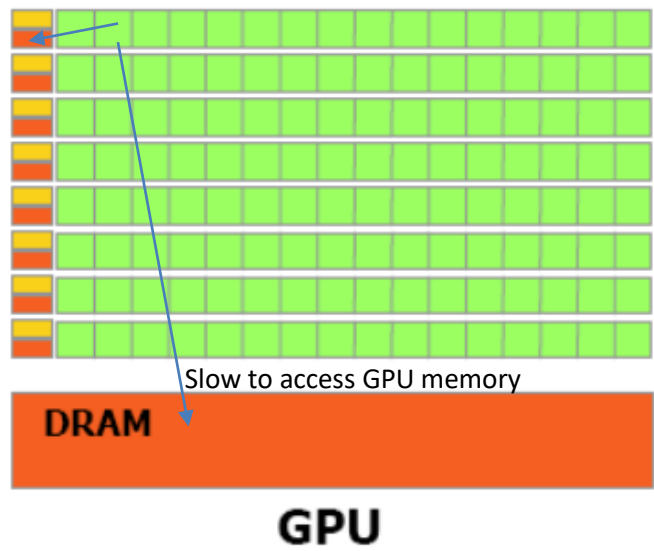
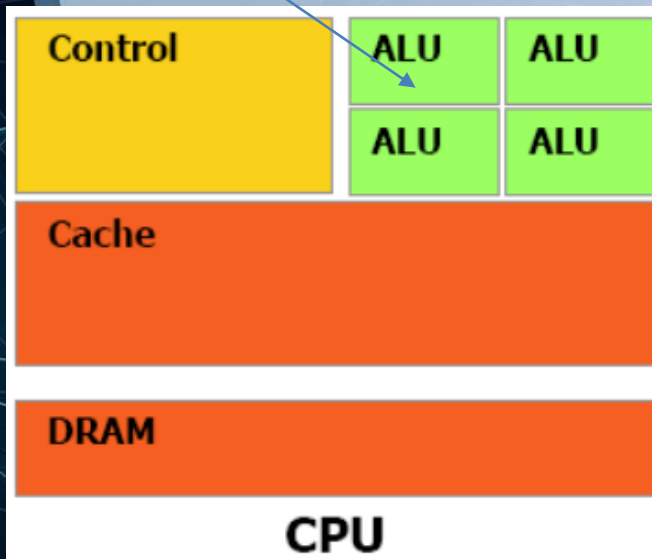
Modern GPU has more ALU's

More powerful computing unit

access in the block cache very efficiently (block cache is small)

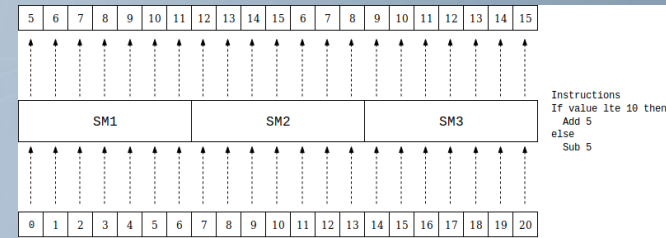
A lot of computing unit (green), but weak

A row = A block





The GPU devotes more transistors to data processing

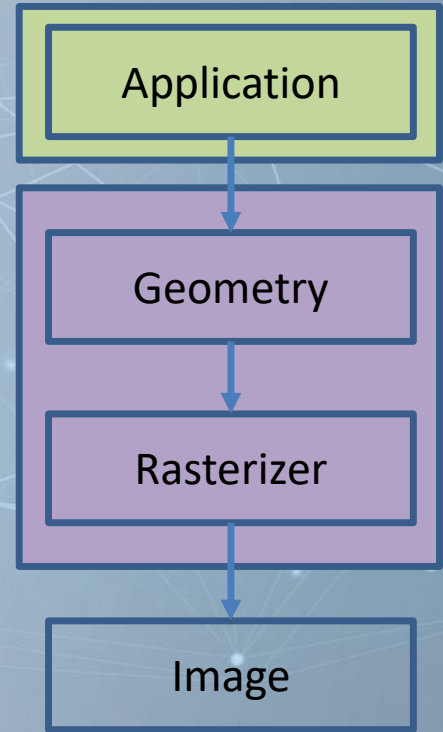
A Specialize Processor



- Very efficient for
 - Fast parallel floating-point processing
 - Single instruction multiple data operations
 - High computation per memory access
- Not as efficient for
 - Double precision
 - Branching-intensive operations
 - Random access, memory-intensive operations

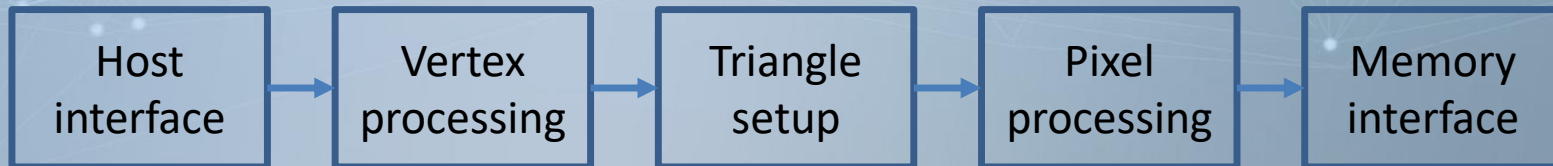
The Rendering Pipeline

- The basic construction – three conceptual stage
- Each stage is a pipeline and runs in parallel
- Graphics performance is determined by the slowest stage
- Modern graphics system:
 - Software: 
 - Hardware: 



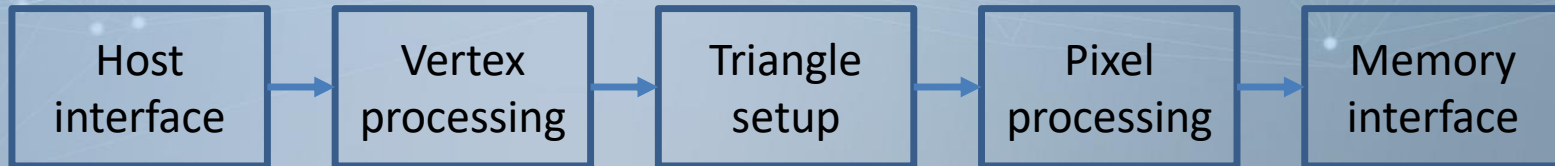
The Rendering Pipeline

- The process to generate two-dimensional images from given virtual cameras and 3D objects
- The pipeline stages implement various core graphics rendering algorithms
- Why should you know the pipeline?
 - Necessary for programming GPUs
 - Understand various graphics algorithms
 - Analyze performance bottleneck



Rendering Pipeline

- Host interface: move data from CPU to GPU
- Vertex processing: transform vertex from object to screen space
- Triangle setup: rasterization
- Pixel processing: color pixels
- Memory interface: produce final image



The Quest for Realism



Fracture



Soft Shadows



Detailed Characters



Rich Environments



Indirect Lighting



Subsurface Scatter



Ambient Occlusion



Turbulence



Participating Media



Simulations



Fluids