Graphics Processing Unit

- **A graphics processing unit** (GPU), is similar CPU
- Designed specifically for performing the complex mathematical and geometric calculations that are necessary for graphics rendering.

Graphics Processing Unit

- A graphics processing unit (GPU) is a computer chip that performs rapid mathematical calculations, primarily for the purpose of rendering images.
- occasionally called **visual processing unit (VPU)**
- GPU is able to render images more quickly than a CPU because of its parallel processing architecture
- Nvidia introduced the first GPU, the GeForce 256, in 1999
- Others include AMD, Intel and ARM.
- In 2012, Nvidia released a virtualized GPU, which offloads graphics processing from the server CPU in a virtual desktop infrastructure.

Graphics Processing Unit

- GPUs are used in
  - Embedded Systems
  - Mobile phones
  - Personal computers
  - Workstations
  - Game consoles

GPU Vs CPU

- A GPU is tailored for highly parallel operation while a CPU executes programs serially.
- For this reason, GPUs have many parallel execution units and higher transistor counts, while CPUs have few execution units and higher clock speeds
- A GPU is for the most part deterministic in its operation
- GPUs have much deeper pipelines (several thousand stages vs 10-20 for CPUs)
- GPUs have significantly faster and more advanced memory interfaces as they need to shift around a lot more data than CPUs

High-end CPU-GPU Comparison

	Xeon 8180M	**Titan V**
Cores	28	5120 (+ 640)
Active threads	2 per core	32 per core
Frequency	2.5 (3.8) GHz	1.2 (1.45) GHz
Peak performance (SP)	4.1 TFlop/s	13.8 TFlop/s
Peak mem. bandwidth	119 GB/s	653 GB/s
Maximum power	205 W	250 W
Launch price	$13,000	$3000

Release dates

Xeon: Q3’17

Titan V: Q4’17

What are GPU's Growth?

- Entertainment Industry has driven the economy of these chips?

– people age 15-35 buy $15B in video games / year

- Moore's Law ++
- Simplified design (stream processing)
- Single-chip designs

GPU

- Very Efficient For
  - Fast Parallel Floating Point Processing
  - Single Instruction Multiple Data Operations
  - High Computation per Memory Access
- Not Efficient For
  - Double Precision
  - Logical Operations on Integer Data
  - Branching-Intensive Operations
  - Random Access, Memory-Intensive Operations