Intro to GPUs

Stefano Markidis and Sergio Rivas-Gomez

Four Key-Points

1. GPUs are specialized hardware, initially designed for graphics applications and now widely used in HPC applications.
2. GPUs can be either integrated in the processor or have dedicated chip.
3. When using dedicated GPU, we need a CPU that acts as host and provides OS services to the GPU
4. To move data from GPU memory to CPU memory is relatively slow.

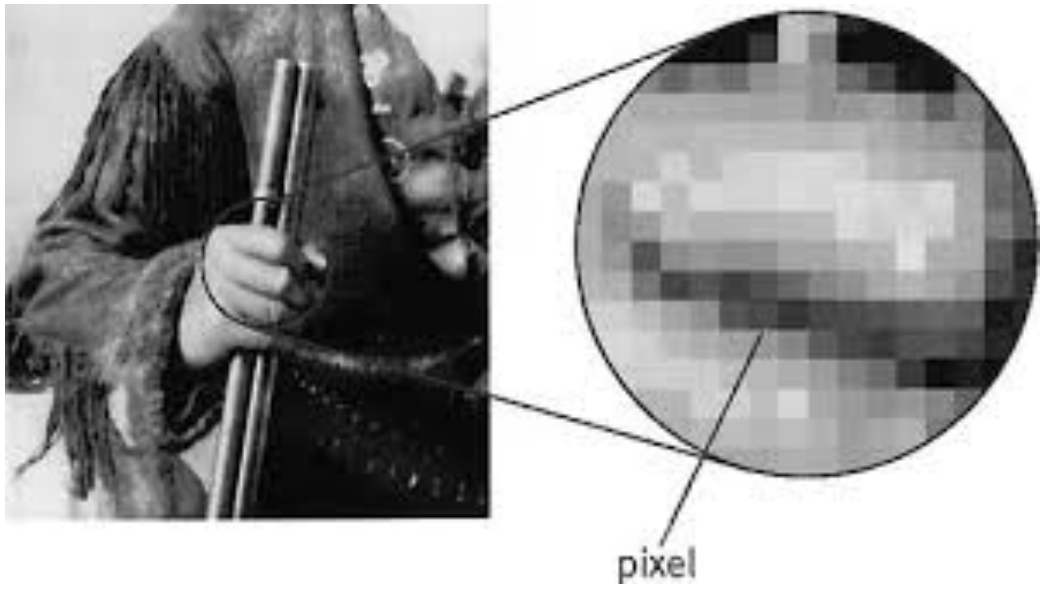
GPUs

GPU = Graphical Processing Unit

= specialized microcircuit to accelerate the creation and manipulation of **images** in **video frame** for **display devices**.

GPU Design Motivation: Process Pixels in Parallel

- **Data parallel workloads**
  - In 1080i and 1080p mode videos, 1920 x 1080 pixels = 2M pixels per video frame → **compute intensive**
- Computation on each pixel is independent from computation on other pixels.
  - No need for **synchronization** or **sophisticated control**



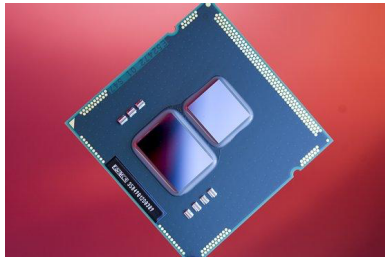
GPUs are more and more present in HPC!

- Why so?
  - **Lots of parallelism** at **low clock** speed → power efficient
  - GPUs compete well in terms of **FLOPS/Watt** vs traditional HPC CPUs
- GPUs are a core technology in many world's **fastest** and **most energy-efficient** supercomputers
  - In the current *Green500*, the top 6 most energy-efficient supercomputers use NVIDIA P100 GPUs

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	61	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	36,288	1,998.0	142	14.110
2	465	**kukai** - ZettaScaler-1.6 GPGPU system, Xeon E5-2650Lv4 14C 1.7GHz, Infiniband FDR, NVIDIA Tesla P100 , ExaScalar Yahoo Japan Corporation Japan	10,080	460.7	33	14.046
3	148	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC National Institute of Advanced Industrial Science and Technology Japan	23,400	961.0	76	12.681
4	305	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Fujitsu Center for Advanced Intelligence Project, RIKEN Japan	11,712	635.1	60	10.603
5	100	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell University of Cambridge United Kingdom	21,240	1,193.0	114	10.428

Where do you find GPUs ?

- **Integrated:** Every laptop has an integrated GPU built into its processor, i.e. Intel HD or Iris Graphics.
- **Dedicated:** A standalone GPU uses its own processor and memory. Most dedicated GPUs are removable. They require more power but also provide higher performance



Source: PC Authority

Source: bit-tech.net

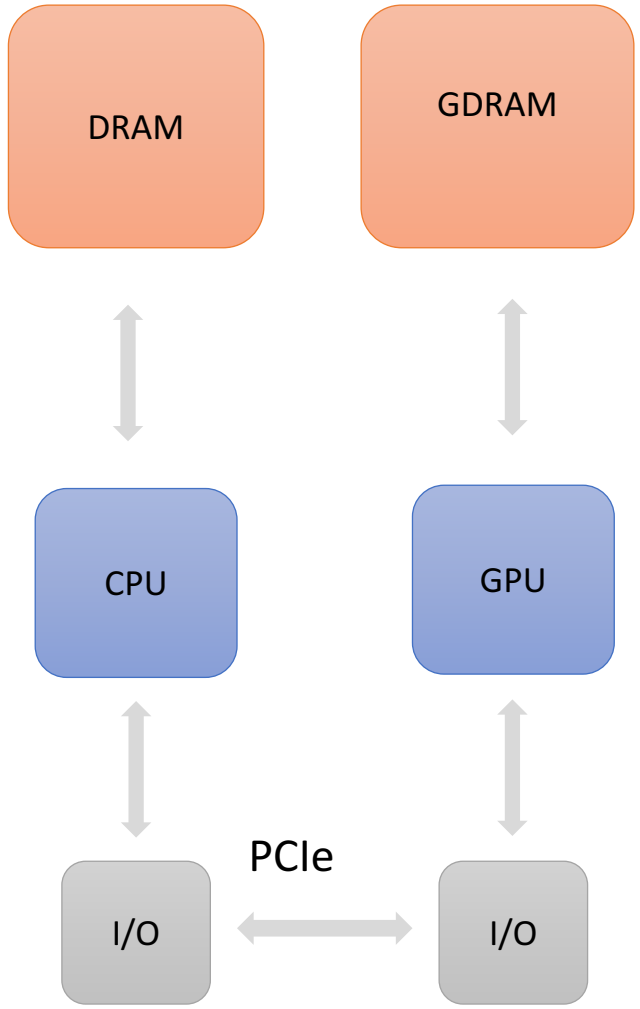
Question: What is the main difference between these two kinds of GPUs?

Vendors of dedicated GPUs

- Some of the most famous GPU vendors are:
  - NVIDIA (https://www.nvidia.se/)
  - AMD (http://www.amd.com/en-us/products/graphics)
  - ASUS (https://www.asus.com/)
- In this course, we will focus on programming **NVIDIA GPUs**.

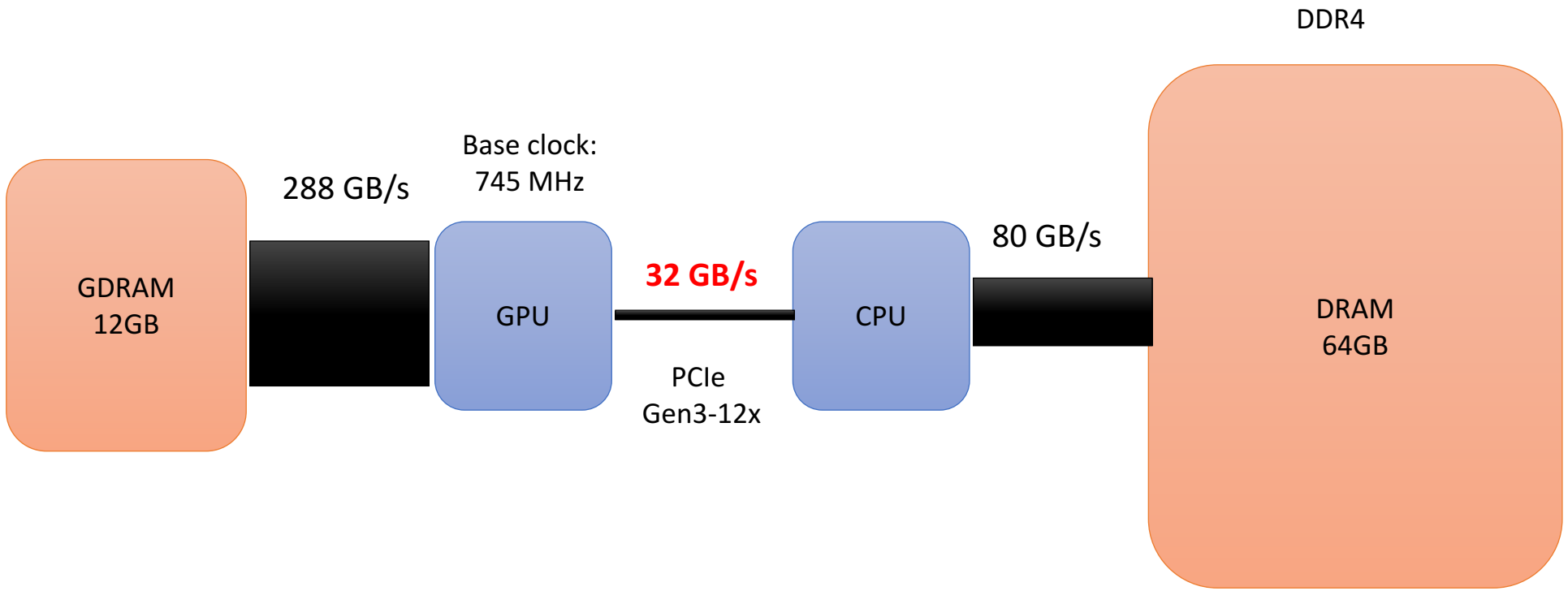
GPUs as Accelerators

- **Problem:** Dedicated GPUs still require OS, IO
- **Solution:** “Hybrid System”
  - CPU provides management and basic services
  - “Accelerators”, (or co-processors) such as GPUs, provide compute power



Weakness of GPUs

Getting data from/to GPU is slow



NVIDIA TESLA K40 = the most common GPU on supercomputers in Nov. 2016 Top500 list

… but not for too long … NVLink

To summarize

1. GPUs are specialized hardware, initially designed for graphics applications and now widely used in many different areas.
2. GPUs can be either integrated in the processor or have dedicated chip. We focus on dedicated GPUs.
3. When using dedicated GPU, we need a CPU that acts as host and provide OS services to the GPU.
4. To move data from GPU memory to CPU memory is “relatively” slow so this might impact the way we program GPUs.