Questions and Answers

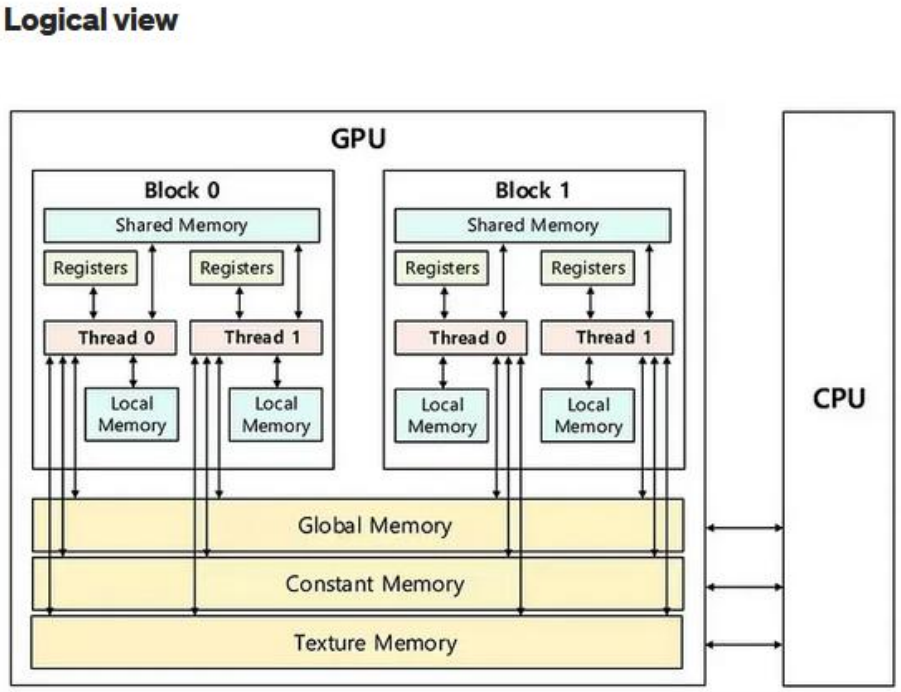
Table of contents

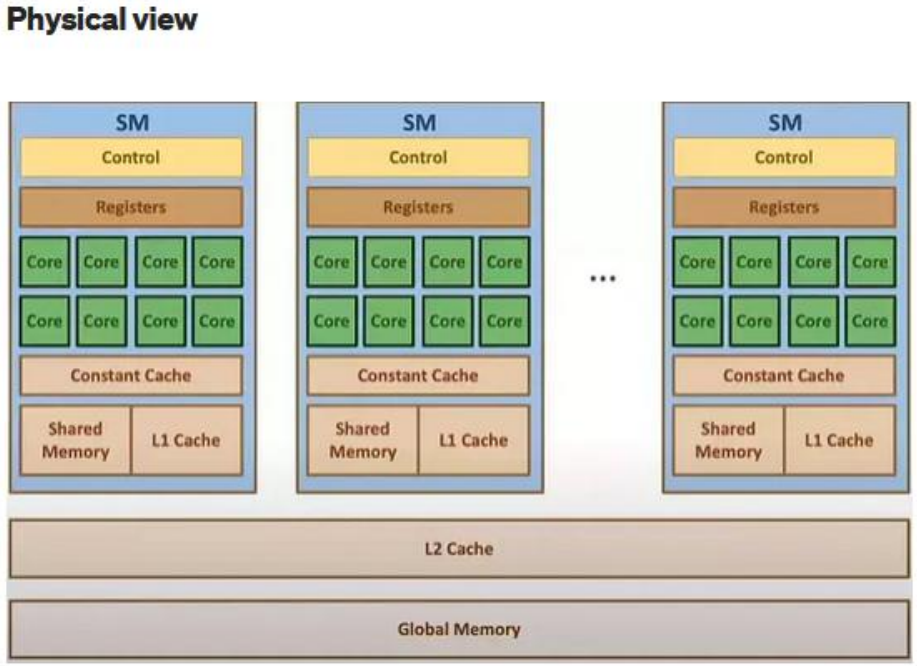
Flash memory

- Electronic non-volatile computer memory storage medium that can be electrically erased and reprogrammed.
- There are two main types of flash memory, NOR flash and NAND flash, with different characteristics.
- Invented at Toshiba in early 1980s, based on EEPROM technology.
- The name comes from the observation that the fast process of erasure remind a camera's flash.
- The NAND type is found mainly in memory cards, USB flash drives, solid-state drives (those produced since 2009), smartphones, and similar products.
- NAND flash memory is erased, written, and read in blocks.
- A key disadvantage of flash memory is that it can endure only a relatively small number of write cycles in a specific block.

Flash memory is divided into 128 pages of 4 kB each. Each page is divided into eight rows, each consisting of 128 four-byte words. Flash can only be erased a page at a time, by setting all bits to ones. Writes can only flip ones to zeros, never zeros to ones. Writes can only be done to a single four-byte word or an entire row at once.

Memory in GPU RAM versus VRAM





- Local memory resides in the card's DRAM. You should use shared memory to minimize local memory usage.
- Shared memory resides on the GPU - access is much faster than other types of memory.
- The CPU has access to global memory, but not to shared memory.

https://giahuy04.medium.com/memory-types-in-gpu-6373b7a0ca47

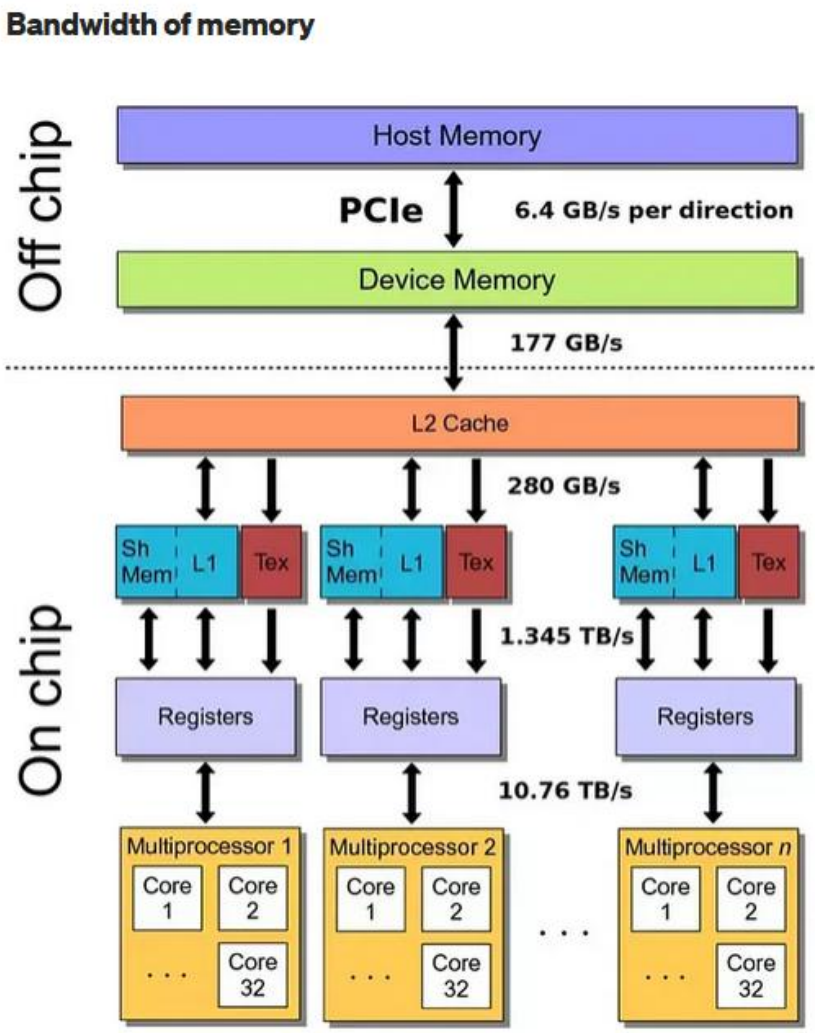
- System RAM is meant for CPUs and VRAM (Video RAM) is meant for GPUs, and these two different processors have very different needs.
- System RAM has very low latency (which is good) but has comparatively low memory bandwidth.
- VRAM has extremely high memory bandwidth with much higher latency.
- CPUs need low latency more than they need high bandwidth, and vice versa for GPUs.

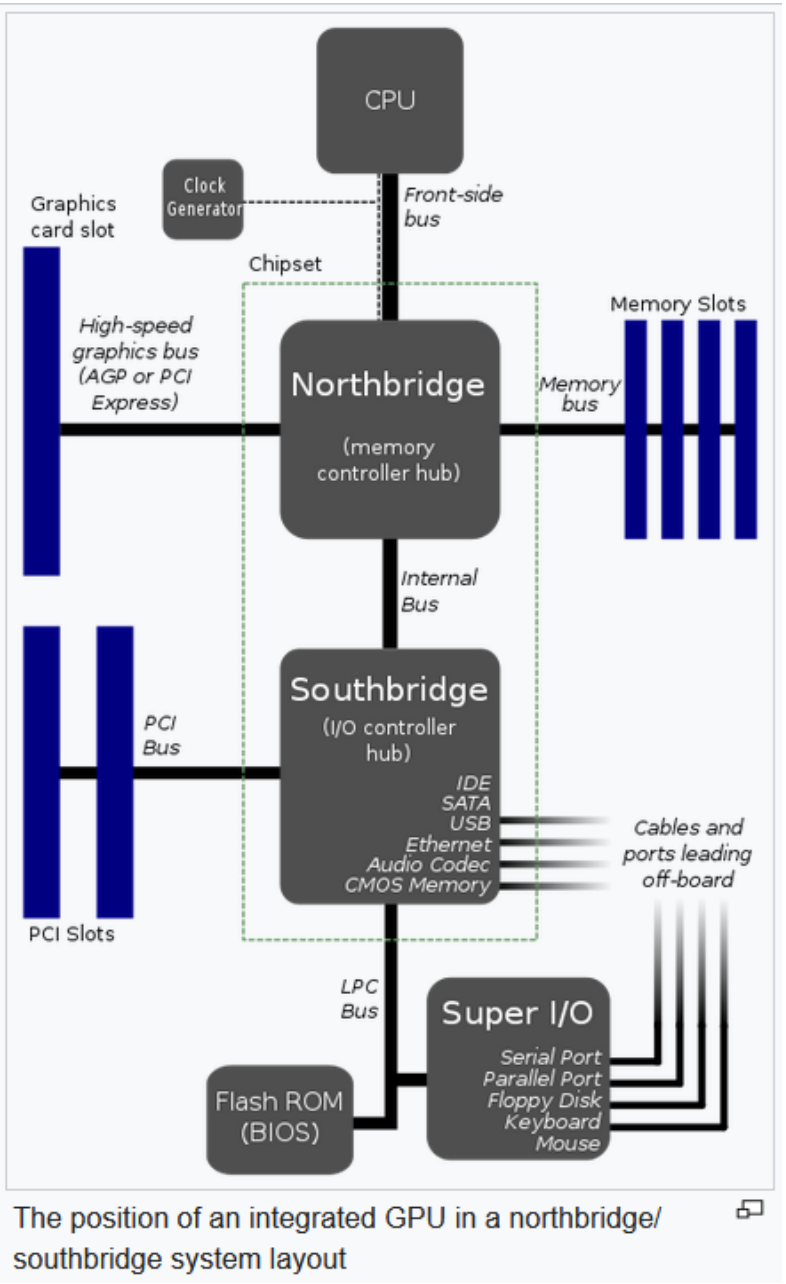
Memory in GPU

The bandwidth, and more importantly, latency between the GPU and RAM over the PCIe bus is an order of magnitude worse than between the GPU and VRAM.

CPU can use a *part* of VRAM (part mapped into the PCI aperture, usually 256MB) directly as RAM, but it will be slower than regular RAM because PCIe is a bottleneck.

Many integrated GPUs use system RAM, because they do not even have their own.





The position of an integrated GPU in a northbridge/ southbridge system layout

https://giahuy04.medium.com/memory-types-in-gpu-6373b7a0ca47

https://en.wikipedia.org/wiki/Graphics_processing_unit

Memory in GPU

https://news.ycombinator.com/item?id=30860259

- Typically CPU and GPU communicate over the PCI Express bus. From the perspective of software running on the CPU, that communication is typically in the form of memory-mapped IO. The GPU has registers and memory mapped into the CPU address space using PCIe. A write to a particular address generates a message on the PCIe bus that's received by the GPU and produces a write to a GPU register or GPU memory.
- The GPU also has access to system memory through the PCIe bus. Typically, the CPU will construct buffers in memory with data (textures, vertices), commands, and GPU code. It will then store the buffer address in a GPU register and ring some sort of doorbell by writing to another GPU register. The GPU (specifically, the GPU command processor) will then read the buffers from system memory and start executing the commands. Those commands can include, for example, loading GPU shader programs into shader memory and triggering the shaders to execute those shaders.