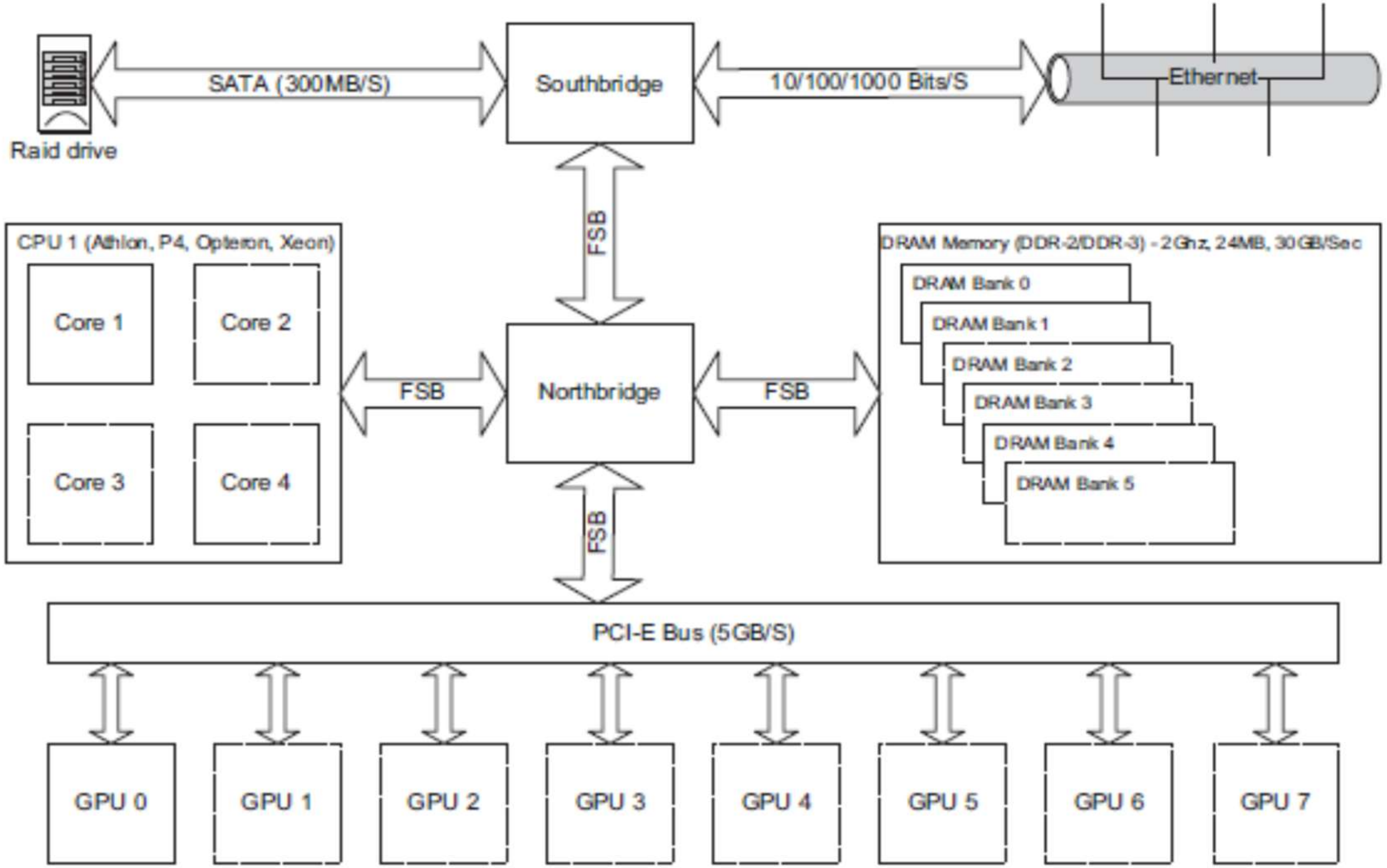
Core 2 series Architecture

Core 2 series Architecture

- All GPU devices are connected to the processor via the PCI-E bus.
- To get data from the processor, we need to go through the Northbridge device over the slow FSB (front-side bus).
- The FSB can run up to 1600 MHz clock rate.
- This is typically one-third of the clock rate of a fast processor.
- Memory is accessed through the Northbridge.
- Peripherals through the Northbridge and Southbridge chipset.

Core 2 series Architecture



Core 2 series Architecture

- The Northbridge deals with all the high-speed components like memory, CPU, PCI-E bus connections, etc.
- The Southbridge chip deals with the slower devices such as hard disks, USB, keyboard, network connections, etc.
- PCI-E-Peripheral Communications Interconnect Express is a bus.
  - it's based on guaranteed bandwidth.
- In the old PCI system each component could use the full bandwidth of the bus, but only one device at a time.
- The more cards you add, the less available bandwidth each card would receive.

Core 2 series Architecture

- PCI-E solved this problem by the introduction of PCI-E lanes.
  - These are high-speed serial links that can be combined together to form X1, X2, X4, X8, or X16 links.
- We have a 5 GB/s full-duplex bus, meaning we get the same upload and download speed, at the same time.
- we can transfer 5 GB/ s to the card, while at the same time receiving 5 GB/s from the card.
- this does not mean we can transfer 10 GB/s to the card if we're not receiving any data (i.e., the bandwidth is not cumulative).

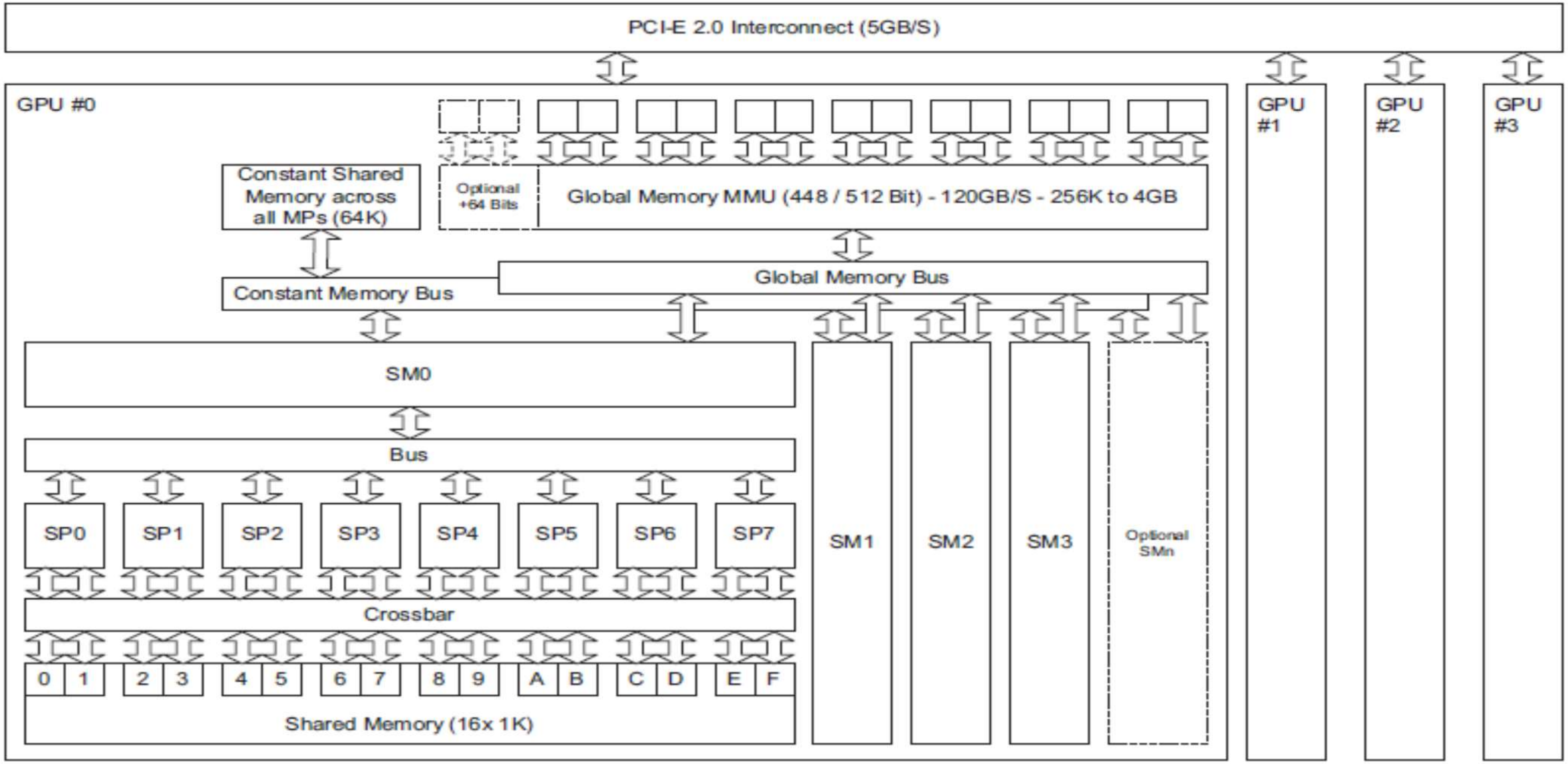
Core 2 series Architecture

- In MPI the latency can be considerable if the Ethernet connections are attached to the Southbridge instead of the PCI-E bus.
- Dedicated high-speed interconnects like InfiniBand or 10 Gigabit Ethernet cards are often used on the PCI-E bus.
- Nehalem architecture brought us QPI (Quick Path Interconnect), which was actually a huge advance over the FSB (Front Side Bus).
- QPI is a high-speed interconnect that can be used to talk to other devices or CPUs.

GPU HARDWARE

- GPU hardware is radically different than CPU hardware.
- The GPU hardware consists of a number of key blocks:
  - Memory (global, constant, shared)
  - Streaming multiprocessors (SMs)
  - Streaming processors (SPs)
- GPU is really an array of SMs, each of which has N cores
  - 8 in G80 and GT200, 32–48 in Fermi, 8 plus in Kepler

GPU HARDWARE



: Block diagram of a GPU (G80/GT200) card.

GPU HARDWARE

- A GPU device consists of one or more SMs.
- Add more SMs to the device and you make the GPU able to process more tasks at the same time, or the same task quicker, if you have enough parallelism in the task.
- NVIDIA hardware will increase in performance by growing a combination of the number of SMs and number of cores per SM.
- There are multiple SPs in each SM. There are 8 SPs shown here; in Fermi this grows to 32–48 SPs and in Kepler to 192.

GPU HARDWARE

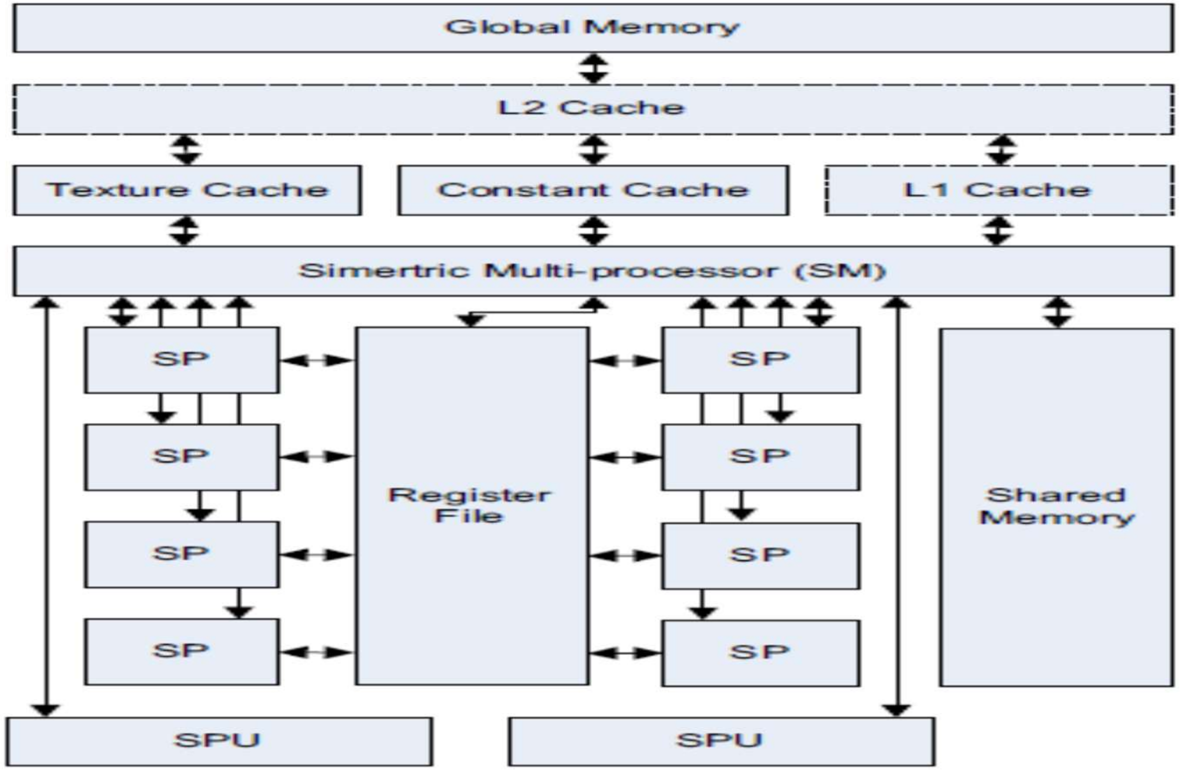


Fig: Inside an SM.

GPU HARDWARE

- Each SM has access to a register file
  - like a chunk of memory that runs at the same speed as the SP units
  - so there is effectively zero wait time on this memory.
- It is used for storing the registers in use within the threads running on an SP.
- There is shared memory block accessible only to the individual SM; this can be used as a program-managed cache.
- Each SM has a separate bus into the texture memory, constant memory, and global memory spaces.
- Texture memory is a special view onto the global memory, which is useful for data where there is interpolation,
  - Ex: 2D or 3D lookup tables.

GPU HARDWARE

- Constant memory is used for read-only data and is cached on all hardware revisions.
- Like texture memory, constant memory is simply a view into the main global memory.
- Global memory is supplied via GDDR (Graphic Double Data Rate) on the graphics card.
- This is a high-performance version of DDR(Double DataRate) memory.
- Memory bus width can be up to 512 bits wide, giving a bandwidth of 5 to 10 times more than found on CPUs, up to 190 GB/s with the Fermi hardware.

GPU HARDWARE

- Each SM also has two or more special-purpose units (SPUs)
- It perform special hardware instructions, such as the high-speed 24-bit sin/cosine/exponent operations.
- Double-precision units are also present on GT200 and Fermi hardware