Multi-core Programming
Assignment 4

Ali Gholami

Department of Computer Engineering & Information Technology
Amirkabir University of Technology

https://aligholamee.github.io
aligholami7596@gmail.com

Abstract

Keywords.

1 GPU Characteristics

Each system has different parts. Some parts of a system could be one or more CPU, GPU, RAM, HDD and SSD. These parts come together differently as generations develop in hardware architecture. This integration confines the performance of a system. The following questions will point out some of the considerations we take into account while designing these kind of systems. Please refer to the book *CUDA Programming: A Developer's Guide to Parallel Computing with GPUs* for more information on these questions. Explain and justify your answers.

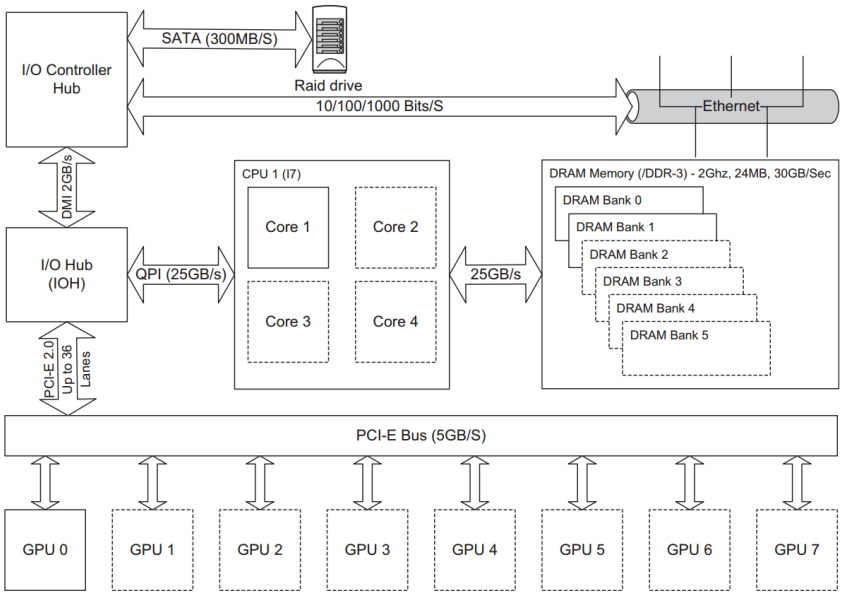

Figure 1.1: Nehalem/X58 System

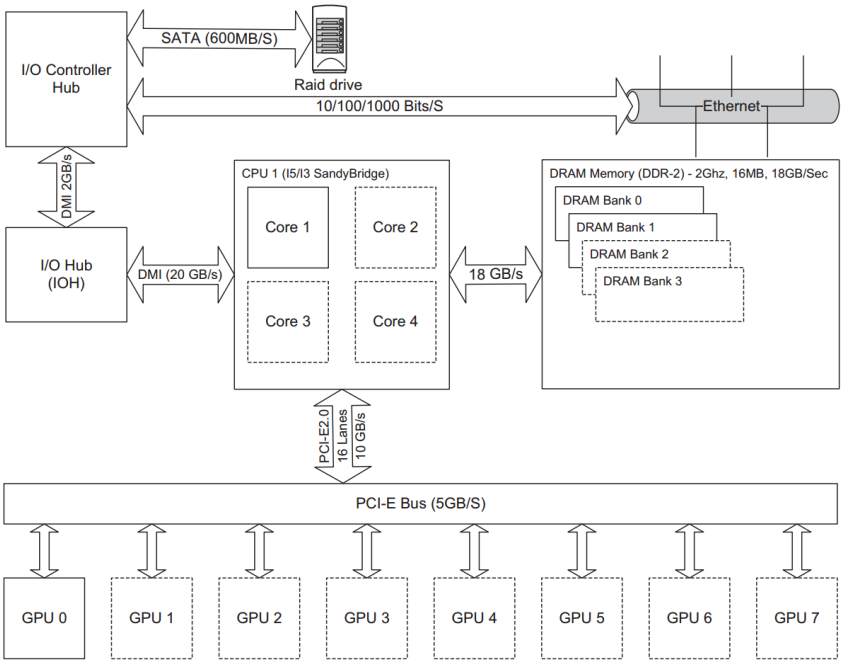

Figure 1.2: Sandybridge Design

- What is the difference between peripheral connection in these architectures?

  One of the most noticeable improvements of Sandybridge was the support for the SATA-3 standard, which supports 600 MB/s transfer rates. This, combined with SSDs, allows for considerable input/output (I/O) performance with loading and saving data. While in the Nehalem/X58 architecture peripheral transfer rate is only 300 MB/s. One significant advantage of the AMD chipsets over the Intel ones is the support for up to six SATA (Serial ATA) 6 GB/s ports. SATA3 can very quickly overload the bandwidth of Southbridge when using multiple SSDs (solid state drives). A PCI-E bus solution may be a better one, but it obviously requires additional costs.

- Does peripheral connection speed matters? Explain.

  If you consider that the slowest component in any system usually limits the overall throughput, this is something that needs some consideration. For example, SATA3 can very quickly overload the bandwidth of Southbridge when using multiple SSDs (solid state drives). As another fact, If using MPI (Message Passing Interface), which is commonly used in clusters, the latency for this arrangement can be considerable if the Ethernet connections are attached to the Southbridge instead of the PCI-E bus. Consequently, dedicated high-speed interconnects like InfiniBand or 10 Gigabit Ethernet cards are often used on the PCI-E bus.

- What kind of port is used to connect to the GPU? What are the attributes of this port?

  PCI-E is used. PCI-E (Peripheral Communications Interconnect Express) is an interesting bus as, unlike its predecessor, PCI (Peripheral Component Interconnect), 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. Thus, the more cards you added, the less available bandwidth each card would receive. 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. Most GPUs now use at least the PCI-E 2.0, X16 specification, as shown in Figure 3.1. With this setup, we have a 5 GB/s full-duplex bus, meaning we get the same upload and download speed, at the same time. Thus, we can transfer 5 GB/ s to the card, while at the same time receiving 5 GB/s from the card. However, 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).

- According to the given architectures in figure 1.1 and figure 1.2, how many GPUs can be connected in each system? What about their bandwidth? Explain.

  In both cases, 8 GPUs can be connected to the system via PCI-E bus. however, the big downside of socket 1155 Sandybridge design: It supports only 16 PCI-E lanes, limiting the PCI-E bandwidth to 16 GB/s theoretical, 10 GB/s actual bandwidth. Thus each GPU is able to use $\frac{10}{8} = 1.25$ GB/s of bandwidth available. On the other hand, the *Nehalem/X58* provides $\frac{16}{8} = 2$ GB/s of bandwidth.