Chapter 2

Hardware in HPC

2.1 Introduction

Optimization can't be done without a good knowledge of the architecture of the device, machine, or computer. Indeed, nowaday software and API try to take care of most of the cases but the last percent of gain have to be architecture dependent. In this chapter we will describe the most important devices architecures from classical Central Processing Units (CPUs), General Purpose Graphics Processing Units (GPGPUs), Field Programmable Gate Arrays (FPGAs) and Application-specific integrated circuits (ASICs). Then those independants elements are use together in order to build supercomputers. The way they are arranged and the nodes interconnection is something that really matter at large scale.

2.2 Architectures

2.2.1 Classical CPU

The CPU, as we know it today, begin its history with *Texas Instruments Inc* and the first patent describing a CPU is "Computing systems cpu" proposed by *Gary Boone* and published in 1973. This first fonctional processing unit on chip and is based on the Von Neumann Model. Before that vacuum tubes were used instead of transistors making the space used very high.

The Von Neumann model can be extracted for the first time by the physists John Von Neumann and its team in [] The Electric Discrete Variable Automatic Computer (EDVAC) presented in this paper was one of the first binary computer built with vacuum tubes. Von Neumann came in the project and summarized the logical aspects of this machine. This model is presented on Fig. **??**. since, every devices can be represented based on that model with optimizations in term of number of cores, memory accesses and bus.

VECTORIZATIOONNNNN

Intel

We cannot decently present nowadays CPU without presenting the world leader that equiped around 90% of the supercomputers (from November 2017 TOP500 list). Since 2007 Intel adopted a production model called the "Tick Tock", presented on Fig. 2.1.

Since its creation the model was following the same fashion, a new manufacturing technology like shrink of the chip with better engraving on a Tick and a new microarchitecture delivered on a Tock

ARM

Advanced RISC Machines is a specific kind of CPU based on RISC architecture. Indeed, usual CPU implemented a complete Instruction Set Architecture, ISA, to perform complicated op-

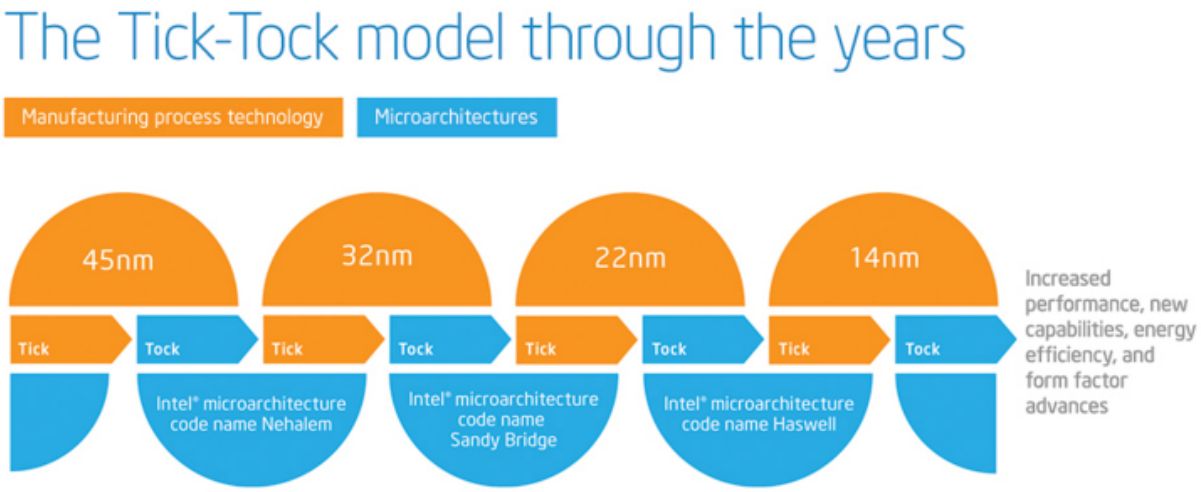

Figure 2.1: Intel Tick-Tock model

erations called CISC, Complex Instruction Set Computers. The downside is the complexity of CISC machines makes them hard to create and they use more energy. The ISA from the RISC is simplier and requires less transistors to operate. Therefore, the energy required and the heat dissipated is less important than for usual CPUs. It would then be easier to create massively parallel processors based on ARM.

2.2.2 GPGPU

GPUs are based on the SIMD model of the Flynn taxonomy presented previously, *Single Instruction, Multiple Data*. The specific execution model is called SIMT (*Single Instruction, Multiple Thread*). It enables the execution of millions of coordinated threads in a data-parallel mode. Two main companies provide GPGPUs for the HPC world NVIDIA and AMD, we will present them in that order and conclude on the differences.

NVIDIA GPU architecture

The NVIDIA company was fonded in April 1993 in Santa Clara, Carolina, by three persons in which Jensen Huang, the actual CEO. Its name seems to come from *invidia* the latin word for Envy and vision, for the graphics generation.

Known as the pioner in graphics, cryptocurrency, portable devices and now AI, it seems to be even the creator of the name "GPU". It GPU, inspired from visualisation and gaming at a first glance, is available as a dedicated device since the Tesla. The public GPUs can also be use for dedicated computation but does not feature eMMC memory, double precision or special functions/FFT cores.

We will describe here the Kepler architecture, this is the one we worked

As presented in Fig.2.2, NVIDIA GPUs include many *Streaming Multiprocessors* (SM), each of which is composed of many *Streaming Processors* (SP). In the Kepler architecture, the SM new generation is called SMX. Grouped into *blocks*, *threads* execute *kernel* functions synchronously. Threads within a block can cooperate by sharing data on an SMX and synchronizing their execution to coordinate memory accesses; inside a block, the scheduler organizes *warps* of 32

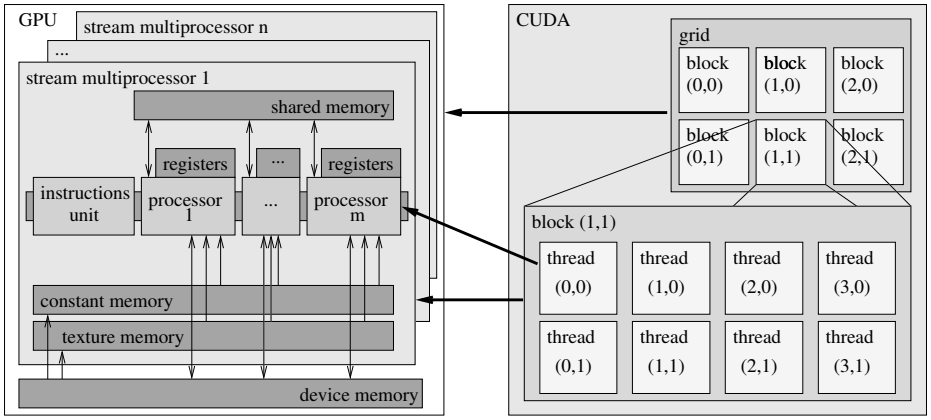

Figure 2.2: NVIDIA GPU and CUDA architecture overview

threads which execute the instructions simultaneously. The blocks are distributed over the GPU SMXs to be executed independently.

Memory, bandwidth and streams:

In order to use data in a CUDA kernel, it has to be first created on the CPU, allocated on the GPU and then transferred from the CPU to the GPU; after the kernel execution, the results have to be transferred back from the GPU to the CPU. GPUs consist of several memory categories, organized hierarchically and differing by size, bandwidth and latency. On the one hand, the device's main memory is relatively large but has a slow access time due to a huge latency. On the other hand, each SMX has a small amount of shared memory and L1 cache, accessible by its SPs, with faster access, and registers organized as an SP-local memory. SMXs also have a constant memory cache and a texture memory cache. Reaching optimal computing efficiency requires considerable effort while programming. Most of the global memory latency can then be hidden by the threads scheduler if there is enough computational effort to be executed while waiting for the global memory access to complete. Another way to hide this latency is to use streams to overlap kernel computation and memory load.

Threads synchronization:

It is also important to note that branching instructions may break the threads synchronous execution inside a warp and thus affect the program efficiency. This is the reason why test-based applications, like combinatorial problems that are inherently irregular, are considered as bad candidates for GPU implementation. Thus we intend to provide a way to regularize their execution, in order to get good acceleration with GPU computation.

Details on K20Xm

Describe K20
2688 cores

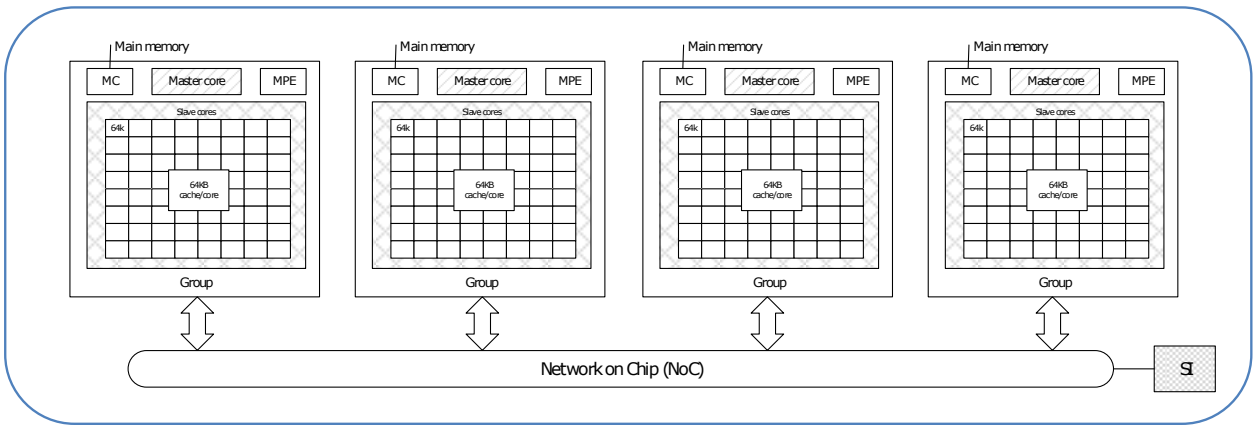

Figure 2.3: Sunway Taihulight node architecture from *Report on the Sunway TaihuLight System*, Jack Dongarra, June 24, 2016.

2.2.3 FPGA and ASICS

2.3 Interconnection and clusters

2.4 Interconnects

2.4.1 TOP500 remarquable clusters

Sunway Taihulight

Sunway Taihulight is the third Chinese supercomputer to be ranked in the first position of the TOP500 list. A recent report from Jack Dongarra, a figure in HPC, decrypt the architecture of this supercomputer[Don16]. The most interessant point is the conception of this machine, completely done in China. The Sunway CPUs were invented and built in China, the Vendor is the Shanghai High Performance IC Design Center.

The SW26010, a many core architecture, features 260 cores based on RISC architecture and a specific conception depicted on Fig.2.3.

Titan

K-Computer

Sequoia

2.5 ROMEO Supercomputer

The ROMEO supercomputer center is the computation center of the Champagne-Ardenne region in France. Hosted since 2002 by the University of Reims Champagne-Ardenne, this so called meso-center (French name for software and hardware architectures) is used for HPC for theoric research and domain science like applied mathematics, physics, biophysics and chemistry.

This project is support by the Champagne-Ardenne region and the CEA (French Alternative Energies and Atomic Energy Commission), aim to host research and production codes of the region for industrial, research and academics purposes.

We are currently working on the third version of ROMEO, installed in 2013. As many of our tests in this study have been done on this machine, we will carefully describe its architecture.

This supercomputer was ranked 151st in the TOP500 and 5th in the GREEN500 list.

2.5.1 ROMEO hardware architecture

ROMEO is a Bull/Atos supercomputer composed of 130 BullX R421 computing nodes.

Each node is composed of two processors Intel Ivy Bridge 8 coeurs @ 2,6 GHz. Each processor have access to 16GB of memory for a total of 32GB per node, the total memory if 4.160TB. Each processor if linked, using PCIe-v3, to an NVIDIA Tesla K20Xm GPGPU. This cluster provide then 260 processors for a total of 2080 CPU cores and 260 GPGPU providing 698880 GPU cores. The computation nodes are interconnected with an Infiniband QDR non-blocking network structured as a FatTree. The Infiniband is a QDR providing 10GB/s.

The storage for users is 57 TB and the cluster also provide 195 GB of Lustre and 88TB of parallel scratch filesystem.

In addition to the 130 computations nodes, the cluster provides a visualization node NVIDIA GRID with two K2 cards and 250GB of DDR3 RAM. The old machine, renamed Clovis, is always available but does not features GPUs.

The supercomputer supports MPI with GPU Aware and GPUDirect.

Bibliography

[Don16] Jack Dongarra. Report on the sunway taihulight system. *PDF). www. netlib. org. Retrieved June*, 20, 2016.