Chapter 3

Runtimes, Software, API and Benchmarks

3.1 Introduction

After presenting the rules of HPC and the hardware that compose the cluster we need to introduice ways to target this supercomputer. Several options are present in the language, the multi-processing API, the distribution and the accelerators code. This chapter details the most important software options for HPC programming and include the choices we made for our applications.

Then it presents the software used to benchmark the supercomputers. We present here the most famous, the TOP500, GRAPH500 and GREEN500 to give their advantages and weaknesses.

Parler des pitfalls, load balancing, concurrence, ... or when the case.

3.2 Sofware/API

In this section we present the main runtimes, API and programming language use in HPC and in this study in particular. The considered language will be C/C++, the most present in HPC world.

3.2.1 Parallel programming

PThreads

The POSIX threads API is an execution model available in most of the languages. It allows the user to define threads that will execute concurrently on the processor ressources using shared/private memory. PThreads is the low level handling of threads and the user need to handle concurrency with mutex, conditions variables and synchronization "by hand". This makes the PThreads hard to use in complex applications and used only for very fine-grained control over the threads management. Fine-grain Coarse-grain applications

OpenMP

Open Multi-Processing, OpenMP[1] [Cha08, Sup17], is an API for multi-processing shared memory like UMA and CC-NUMA. It is available in C/C++ and Fortran. The user is provided with pragmas and functions to declare parallel loop and regions in the code. In this model the main thread, the first one before forks, command the fork-join operations.

The last versions of OpenMP also allow the user to target accelerators. During compilation the user specify on which processor or accelerator the code will be executed in parallel.

[1] http://www.openmp.org

Ohters

Many other tools handle parallel programming for processors. Cilk for C/C++ base, like OpenMP, on Fork-Join paradm. Threads Building Blocks, TBB, from Intel.

3.2.2 Distributed programming

In the cluster once the code have been developped locally and using the multiple cores available, the new step is to distribute it all over the nodes of the cluster. This step requires the processes to access NoRMA memory from a node to another. Several runtime are possible for this purpose.

MPI

The Message Passing Interface, MPI, is the most and widely spread runtime for distributed computing [Gro14, Gro15]. Several implementations exists from Intel MPI[2] (IMPI), MVAPICH[3] by the Ohio State University and OpenMP[4] combining several MPI work like Los Alamos MPI (LA-MPI). Those implementation follow the MPI standards 1.0, 2.0 or the latest, 3.0. Who define standards ??

Some MPI implementation offer a support for accelerators targeting directly their memory through the network without multiple copies on host memory.

Charm++

Charm++[5] is another API for distributed programming developped by the University of Illinois Urbana-Champaign. It is asynchronous messages paradigm driven. In contrary of runtime like MPI that are synchronous but can handle asynchronous, charm++ is natively asynchronous. It is based on *chare object* that can be activated in response to messages from other *chare objects* with actions and callbacks. The repartition of data to processors is completely done by the API, the user just have to define correctly the partition of the program. Charm++ also provide a GPU manager implementing data movement, asynchronous kernel launch, callbacks, etc.

A perfect example can be the hydrodynamics N-body simulation code Charm++ N-body Gravity Solver, ChaNGa [JWG+10], implemented with charm++ and GPU support.

Legion

The Legion[6] is a distributed runtime support but Stanford University, Los Alamos National Laboratory and NVIDIA. This runtime is data-centered targeting distributed heterogeneous architectures. Data-centered runtime focus to keep the data dependency and locality moving the tasks to the data and moving data only if requested. In this runtime the user define data organization, partitions, priviledges and coherence. Many aspect of the distribution and parallelization are then handle by the runtime itself.

3.2.3 Other tools

HPX ? Others ?

[2] https://software.intel.com/en-us/intel-mpi-library
[3] http://mvapich.cse.ohio-state.edu/
[4] http://www.open-mpi.org
[5] http://charmplusplus.org/
[6] http://legion.stanford.edu/

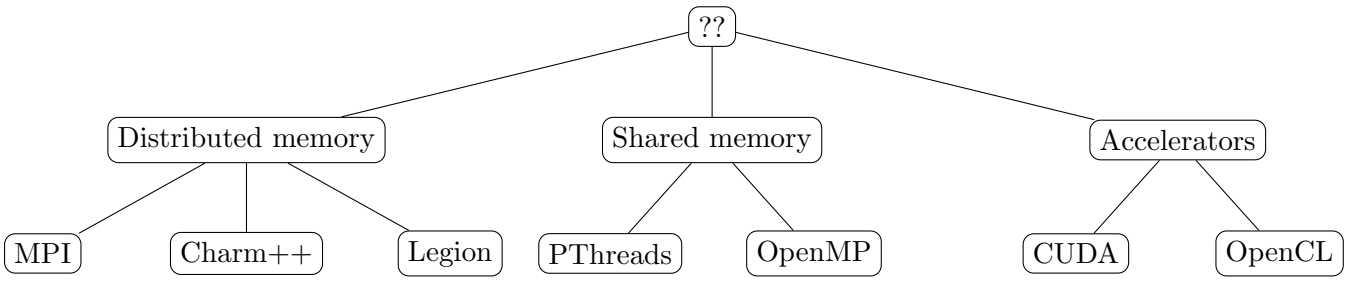

Figure 3.1: MIMD memory models

3.2.4 Accelerators

CUDA

The Compute Device Unified Architecture is the API develop in C/C++ Fortran by NVIDIA to target its GPGPUs. The API provide high and low level functions. The driver API allows a fine grain control over the executions.

The CUDA compiler is called NVdia C Compiler, NVCC. It converts the device code into Parallel Thread eXecution, PTX, and rely to the C++ host compiler for host code. PTX is a pseudo assembly language translated by the GPU in binary code that is then execute. As the ISA is pretty simple it able the user to work directly in assembly for very fine grain optimizations.

OpenCL

Section for profiling?

The runtimes, libraries and API we describe can be use in combination. The usual one is MPI for distribution, OpenMP and CUDA to target GPUs.

3.3 Benchmark

This section regroup a bunch of the most famous nowadays benchmarks for HPC.

3.3.1 TOP500

The most famous benchmark is certainly the TOP500[7]. It gives the ranking of the 500 most powerful, known, supercomputers of the world as its name indicates. Since 1993 the organization assembles and maintains this list updated twice a year in June and November.

This benchmark is based on the LINPACK[DMS+94] a benchmark introduced by Jack J. Dongarra. This benchmark rely on solving dense system of linear equations. As specified in this document this benchmark is just one of the tools to define the performance of a supercomputer. It reflects "the performance of a dedicated system for solving a dense system of linear equations". This kind of benchmark is very regular in computation giving high results for FLOPS.

3.3.2 GREEN500

In conjunction of the TOP500, the GREEN500 focus on the energy consumption of supercomputers. The scale is based on FLOPS per watts [FC07].

3.3.3 GRAPH500

The GRAPH500 benchmark focus on irregular memory accesses, and communications. It will be detailed in Part. II Chapter II in our benchmark suite.

[7] http://www.top500.org

3.4 Conclusion

Bibliography

[Cha08] Barbara Chapman. *Using OpenMP : portable shared memory parallel programming.* MIT Press, Cambridge, Mass, 2008.

[DMS+94] Jack J Dongarra, Hans W Meuer, Erich Strohmaier, et al. Top500 supercomputer sites, 1994.

[FC07] Wu-chun Feng and Kirk Cameron. The green500 list: Encouraging sustainable supercomputing. *Computer*, 40(12), 2007.

[Gro14] William Gropp. *Using MPI : portable parallel programming with the Message-Passing-Interface.* The MIT Press, Cambridge, MA, 2014.

[Gro15] William Gropp. *Using advanced MPI : modern features of the Message-Passing-Interface.* The MIT Press, Cambridge, MA, 2015.

[JWG+10] Pritish Jetley, Lukasz Wesolowski, Filippo Gioachin, Laxmikant V Kalé, and Thomas R Quinn. Scaling hierarchical n-body simulations on gpu clusters. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11. IEEE Computer Society, 2010.

[Sup17] Bronis Supinski. *Scaling OpenMP for Exascale Performance and Portability : 13th International Workshop on OpenMP, IWOMP 2017, Stony Brook, NY, USA, September 20-22, 2017, Proceedings.* Springer International Publishing, Cham, 2017.