Final Architecture Proposal

SC19 Student Cluster Competition

Jiaao He, Shengqi Chen, Liyan Zheng,
Kezhao Huang, Chenggang Zhao, Chen Zhang,
Jidong Zhai
Tsinghua University
September 25, 2019

1 Hardware Configuration

The main idea of our hardware is to combine Intel® Cascade Lake CPUs and NVIDIA® GPU accelerators. We will place 8 GPUs per node on 2 nodes, while other nodes have no GPUs. Let's call them "GPU nodes" and "CPU nodes" respectively.

We will have 6 nodes in total, which consists of 2 GPU nodes and 4 CPU nodes. Each node is equipped with two Intel® Xeon® Platinum 8280 Processor. There are 16 GPUs in total placed on 2 nodes. For network, we use a combinition of 25Gb Ethernet and Infiniband HDR100.

Here is a more detailed list of our hardware. Among those lines we add some comments on why we believe the choice is appropriate for the challenge.

- **Chassis**:
  - GPU node: 2× Supermicro SuperServer 4029GP-TRT
    This model has PCI-E x16 ports and space sufficient for 8 GPUs and HDR100 IB Card.
  - CPU node: 4× Supermicro SuperServer 2029GP-TR
    To minimize potential problems and controlling cost, we also choose Supermicro for CPU nodes.
- **CPU**: 2× Intel® Xeon® Platinum 8280 Processor per node (12 total)
  We consider that 8280 is power-efficient and of high performance. Among all Cascade Lake CPUs, it maximizes the ratio of total frequency to power, and has advanced features such as AVX512 instruction set.
- **Memory**: 12× RDIMM DDR4 2933 MT/s 16GB per node (hexa-channel, 384GB per node, 2304GB total)
- **Power Supply**: 2000W Redundant Power Supplies per node
- **Accelerator**: 8× NVIDIA® Tesla® V100 per GPU node (16 total)
  This can accelerate LINPACK and HPCG benchmarks, plus the mystery application if possible.
- **Storage**:
  - Local storage: Intel® SSD DC S3610 100GB per node
    This is used to store the OS and other software installed, and also some small temporary data.
  - Data disk: Intel® SSD DC P4618 (6.4TB) and P3608 (4TB) on head node
    This is used to store large datasets, including the input files and results for the challenge. We will also use this high-speed disk to run the IO500 benchmark.
- **Network**:

  - Infiniband: Mellanox® HDR 100Gb/s InfiniBand adapter per node
    This can accelerate the communication of MPI programs, namely almost all applications we will run.
  - Ethernet: 25Gbps adapter per CPU node and 10Gbps adapter per GPU node
    We use Ethernet for performance monitoring and controlling (e.g. SSH & file transfer inside the local network).
- **Others**:
  - 42U rack cabin
  - PDU and LCD display (provided by SCC official)

It will be a challenge to control the power of 16 GPUs, since the maximum possible power consumption is 250W×16=4000W. We will lower the power limit as well as the computation frequency of GPU to control the power usage.

Nonetheless, it is possible that some details of the hardware configuration are prone to change due to our future knowledge to the applications and/or further needs. For example, we might remove a CPU node, adjust the number of GPUs, or move some GPU accelerators to CPU nodes. However, The main idea, Intel® Cascade Lake CPUs with NVIDIA® accelerators, will not change.

2 Software Used

- **Operating System**:
  - Linux kernel: `4.9.0-7-amd64`
  - Distribution: Debian GNU/Linux 9.8
- File system: `ZFS`
  With periodical snapshots, we can inspect history versions of a file, and can roll it back if necessary. This can also help us recover from misoperation.
- Optional package manager: `Spack`
  This supports multiple versions, configurations and compilers. Softwares of different versions can peacefully coexist, which enable us to conveniently install multiple versions of compilers and libraries. For each application, we can try out different version combinations of compilers and libraries, to get a optimal solution.
- **Compilers**:
  - Intel C/C++/Fortran compiler 17.0.7 / 18.0.3 / 19.0.1
  - `gcc` version 6.3.0 / 8.2.0
- **Communication Libraries**:
  - Intel MPI 2017.4.239 / 2018.3.222 / 2019.1.144
  - Mellanox® HPC-X™ 1.3
  - OpenMPI 1.10.7 / 3.1.2
- NVIDIA driver 418.40.04
- CUDA 8.0.61 / 9.0.176 / 9.2.88 / 10.0.130 / 10.1

Besides, we also install or compile other libraries specially for some applications, e.g. `HDF5` for SST and `METIS` for VPIC.