Final Architecture Proposal

Shanghai Jiao Tong University

October 7, 2019

1 Hardware Configuration

With suggestions from our vendor partner, the Inspur Group, we've decided to use their NF5280M5 server for SC19 SCC. The model we choose supports 2 CPUs and at most 4 GPUs. Though we are not going to use 4 GPUs for each of our server, those PCIe slots can enable us to add more solid state drives to our system, which can be beneficial for our IO-500 benchmark and applications with parallel IO access pattern. Additionally, we've tried this model in previous cluster competitions, and we are fairly familiar with it.

Our choice for CPU is Intel Xeon Gold 6240. The Gold 6240 is based on the Cascade Lake microarchitecture and is manufactured on a 14nm process. As a major advantage over Skylake, the former microarchitecture, Cascade Lake supports 2933Hz DRAM, which brings about a larger memory bandwidth. In an HPCG benchmark, we observed roughly 10GB/s more bandwidth when using 2933Hz DRAM, comparing to that with 2666Hz DRAM, with little increase in power consumption. We would equip each server board (containing two 6240 CPU) with 12 $\times$ 32GB 2933Hz DDR4 RAMs to fully exploit this benefit in the competition.

To conclude, we are going to use the following hardware configuration for the Student Cluster Competition. Our power budget will be later discussed in Section 3.

- Inspur NF5280M5 server $\times$ 8
- 32GB 2933Hz DDR4 RAM $\times$ 12 each server
- Intel Xeon Gold 6240 $\times$ 2 each server
- Mellanox EDR 100GB $\times$ 1 each server
- Mellanox EDR Switch $\times$ 1
- Ethernet Switch 1000M $\times$ 1
- NVIDIA Tesla V100 GPU $\times$ 8

2 Software Configuration

The team is experienced in working with CentOS 7.6 and Intel Parallel Studio, which is a seasoned software stack extensively used in various supercomputer clusters, including Pi, the supercomputer platform in Shanghai Jiao Tong University. We are testing out benchmarks and applications for SC19 on this software platform, and we will also use it on the spot.

For filesystems, we used to set up BeeGFS for small clusters, a parallel filesystem with quite good performance for IO intensive applications.

For package and environment management, we would use Environment-Modules and Spack. Currently we've got no plan to use schedulers.

Our software configuration can be concluded into the following list.

- CentOS 7.6
- Intel Parallel Studio 2019-update4
- Environment-Modules for environment management
- Spack for additionally package management
- BeeGFS for parallel filesystem

3 Power Budget

To estimate the total power consumption of our target cluster, we utilized the HPL and HPCG benchmarks, which can represent a large class of HPC applications. When running benchmarks on a single node, we used the *ipmitool* to read out the on-board power sensors. The results are shown in Figure 1 and 2.

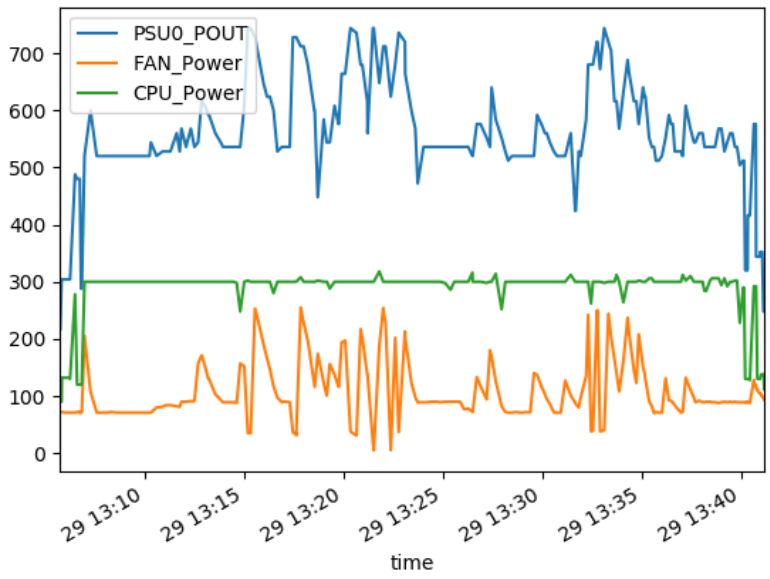

Figure 1: HPL Power Consumption

The result shows that the CPUs consume about 300W at peak. Also we see that the power consumption of the fans (using auto mode) can range from less than 100W to nearly 300W, which introduced large variability into the total power consumption. Hence, we will have to control the fan speeds by manual mode.

With manual mode, the power consumption of the fans can be constantly controlled to 50-100W. Also, with Intel RAPL, we can further setup a power limit for CPUs. In total, we would expect that the power consumption of a single node can be controlled to about 300-400W, which should be feasible, allowing us to setup a 6-to-8-node cluster, and we've still got a 600W budget for GPUs and outsets.

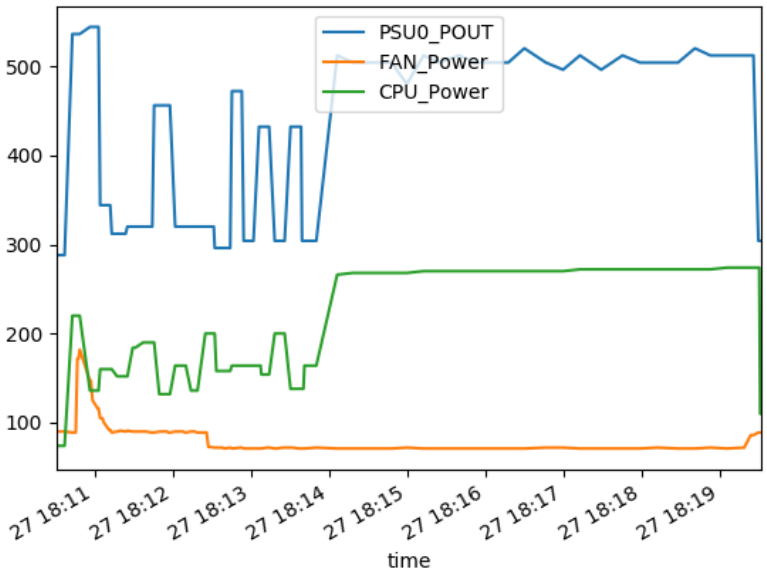

Figure 2: HPCG Power Consumption

Although a single Tesla V100 can consume 250W at most, for most applications, a lower power limit will enable a good enough performance, according to our previous observation. Also considering that our server board supports 4 GPUs, we can use just part of the cluster for GPU workloads, putting the rest servers to sleep states to save power.

Our actual hardware configuration may subject to change according to power consumption conditions measured on the spot. To sum up, we think the hardware configuration mentioned in Section 1 has offered us great flexibility and we are able to manage this configuration within the 3000W power limit.