Architecture Proposal

RACKlette - ETH Zürich
Manuel Burger

October 2019

Abstract

This is the cluster architecture proposal for the RACKlette Team from ETH Zürich for the 2019 Student Cluster Competition at Supercomputing 2019 in Denver, Colorado. We give an overview of the planned hardware and software setup and explain our motivation behind the choosen configuration.

Hardware

Our Cluster is a 4 node system. For the competition the system will be mounted in a 12U rack. All the components are air cooled.

Network

- Mellanox Infiniband 36-port FDR switch, configured with a licence to run 56Gbit/s IB and 40Gbit/s Ethernet over single optical fiber connections to each node
- Additionally the nodes are connected to a 1Gbit/s Ethernet switch for network access, administration and monitoring

Nodes (4x)

- Gigabyte Server enclosures and Gigabyte motherboards
- Dual socket Intel Xeon Platinum 8180 Skylake-SP processors 28 cores running at 2.5GHz base clock with a 205W TDP; in total 8 CPUs in the cluster with 224 cores
- 192GB of DDR4@2666MHz memory; in total 768GB of system memory
- per node 4 Nvidia Tesla V100 32GB GPU accelerators; in total 16 accelerators in the system
- per node 4 Intel Optane SSD DC P4800X with 375GB; in total 16 drives in the system

- per node one Mellanox ConnectX-5 network adapter running 56Gbit/s Infiniband FDR and 40Gbit/s Ethernet
- each node boots from a 240GB Sata SSD
- the master node has an extra 2TB Sata SSD for shared data

Motivation

The system we will bring to the Student Cluster Competition at Supercomputing 2019 in Denver is an iteration of our quite successful setup from our participation at ISC19's Student Cluster Competition in Frankfurt.

The use of established Intel CPUs on the x86 architecture allow us to run almost any modern HPC workload. The use of established and optimized Intel compilers and libraries (such as Intel MKL) allow us to use the CPUs efficiently and at peak performance.

The 16 Nivida Tesla V100 accelerators allow us to further improve our performance on GPU optimized applications. The use of Nvidia acclerators enables the use of Cuda as a widely adopted GPU programming interface, as well as OpenCL code. The accelerators will be especially beneficial for the benchmarks (Linpack and HPCG).

Iterating over our setup from ISC19 we replaced our previous Infiniband EDR switch with an FDR switch and configured it to run IB FDR 56Gbit/s and Ethernet 40Gbit/s over the same optical fiber cable to each node. This allowed us to remove a full Ethernet switch as all the communication and data movement on the cluster now uses the 40Gbit/s Ethernet link. Several tests have shown that most of our applications don't suffer from the lower bandwidth of our Infiniband connection, which went down from 100 to 56 Gbit/s.

16 Intel Optane SSD have been fused together to form a single high-bandwidth and extremely low latency storage system. The storage system reduces I/O bottlenecks for any application and will be very competitive when running the IO500 benchmark.

Our system architecture has been developed in close cooperation with our experienced advisors from the Swiss National Supercomputing Centre (CSCS) and supported by our system component sponsors.

Software

- OS: CentOS Linux release 7.6
- Bright Cluster management
- BeeGFS parallel cluster file system

Motivation

We run a well established Linux distribution for maximal compatibility with any HPC software. On top of our OS we run the Bright Cluster management with a single master node and three compute node configuration. The cluster management allows us to quickly change system parameters, iterate fast over different system configurations and monitor the system. The overhead has been measured to be negligible, however the gains in time and workflow are significant.

We run the BeeGFS parallel cluster file system on our 16 Intel Optane SSDs for optimal storage system performance. Our advisors from CSCS have suggested the use of this file system. The ease of use with its integration into Bright Cluster management and its great performance speak for itself.

Besides that our Intel CPUs and Nvidia accelerators enable us to use many optimized libraries and software developed and tailored to Intel and Nvidia hardware.

Final words

Our system is an iteration over our system used at ISC19 Student Cluster Competition. The improvements should allow us to build on the previous success with this system and we are looking forward to test its performance at the Student Cluster competition at SC19.

Some system components are still under investigation as we are preparing for the competition. Among the most significant potential changes are the GPU configuration and the network setup; however the upper-bound on the energy budget will not exceed the upper-bound on the configuration described above.

Team RACKlette

Manuel Burger, Jinfan Chen, Thore Göbel,
Emir Isman, Valeria Jannelli, Jan Kleine

Contact:
RACKlette Mail: `racklette@lists.inf.ethz.ch`
Web: `https://racklette.ethz.ch`