University of Basel

Lecturer:
Prof. Dr. **Florina M. Ciorba**
Prof. Dr. Erick Lavoie
Prof. Dr. Heiko Schuldt
Prof. Dr. Isabel Wagner

Seminarraum 05.002,
Spiegelgasse 5,
CH-4051 Basel

Assistants:
Ali Ajorian, M.Sc
Jonas Korndörfer, M.Sc
Shiva Parsarad, M.Sc
Yiming Wu, M.Sc

https://vorlesungsverzeichnis.unibas.ch/de/home?id=277318

Foundations of Distributed Systems 2023 (45402-01, 45402-02) **Fall Semester 2023**

Description and Usage of miniHPC

Objectives:
1. Understand the system architecture of the miniHPC cluster
2. Manage job submissions on the miniHPC cluster

Structure:
1. What is miniHPC?
2. How to login to the miniHPC Cluster?
3. What is a batch scheduling system?
4. How to use Slurm to manage job submission?
5. miniHPC usage hands-on.

What is miniHPC?

It is a small HPC cluster that has a peak performance of 28.9 double precision TFLOP/s. The miniHPC has 4 types of nodes (see Table 1), Intel Xeon E5-2640 nodes, Intel Xeon Phi Knights Landing (KNL) 7210 nodes, Intel Xeon Gold 6258 nodes, and AMD EPYC 7742 nodes (see Table 2). The Intel Xeon nodes consist of 22 computing nodes, 1 login node, and 1 node for storage. The Intel Xeon Phi nodes consist of 4 computing nodes. The Intel Xeon Gold 6258 nodes, and the AMD EPYC 7742 nodes are both computing nodes.

Table 1: miniHPC nodes information

Count	CPUs
22	2 Intel Xeon E5-2640
4	1 Intel Xeon Phi 7210
1	2 Intel Xeon Gold 6258
1	2 AMD EPYC 7742

Table 2: CPU information

CPU model	**Count**	**CPU speed (GHz)**	**Cores**	**Threads**	**RAM (GB)**	**Cache**
Intel Xeon E5-2640 v4[a]	44	2.4	10	20	64	L3: 25MB
Intel Xeon Phi 7210[b]	4	1.3	64	256	128	L2: 32MB
Intel Xeon Gold 6258[c]	2	2.7	28	56	1500	L3: 38.5MB
AMD EPYC 7742[d]	2	2.25	64	128	1500	L3: 256MB

[a] http://ark.intel.com/products/92984/Intel-Xeon-Processor-E5-2640-v4-25M-Cache-2_40-GHz

[b] http://ark.intel.com/products/94033/Intel-Xeon-Phi-Processor-7210-16GB-1_30-GHz-64-core

[c] https://ark.intel.com/content/www/us/en/ark/products/199350/intel-xeon-gold-6258r-processor-38-5m-cache-2-70-ghz.html

[d] https://www.gigabyte.com/Enterprise/GPU-Server/G482-Z52-rev-100#Specifications

Node **Intel Xeon Gold 6258R** contains 2 × **NVidia A100-PCIE-40GB GPUs** (https://images.nvidia.com/data-center/a100/a100-datasheet.pdf)

All nodes are interconnected through two different types of interconnection networks. The first network is an Ethernet network with 10 Gbit / s speed, reserved for users and administrators access. The second network is the fastest network, an Intel Omni-Path network with 100 Gbit / s speed, reserved for the high-speed communication between the computing nodes. The topology of this second network interconnects the 30 nodes (24 Xeons E5-2640, 4 KNLs, and 1 Xeon Gold 6258R, 1 AMD EPYC 7742) of the miniHPC cluster via a two-level fat-tree topology.

Table 3: miniHPC interconnection networks

Network fabrics	**Network speed**
Intel Omni-Path	100 Gbit/s
Ethernet	10 Gbit/s

The 30 nodes of the miniHPC cluster are interconnected in a two-level fat-tree topology, illustrated in Figure 1.

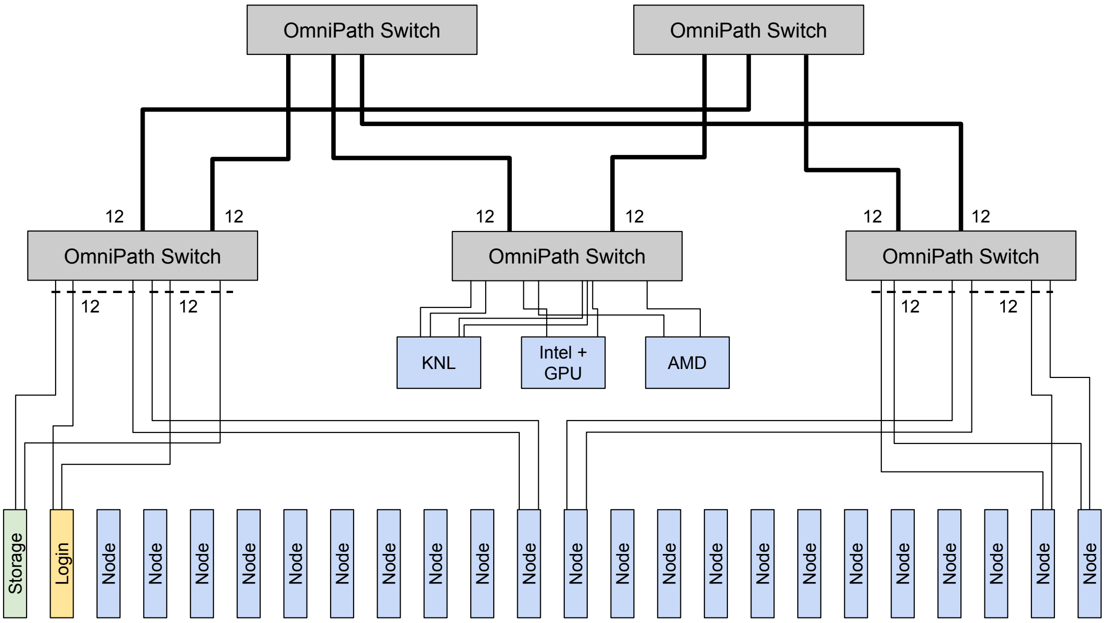

Figure 1: Graphical illustration of the miniHPC two-level fat-tree topology. Number of nodes: 30 (24 Intel Xeon E5-2640, 4 Intel Xeon Phi or KNL, and 1 Xeon Gold 6258R, 1 AMD EPYC 7742). Number of switches: 5. Number of links: 196.

Figure 2 shows pictures taken from the miniHPC cluster to illustrate how the cluster looks in real life. The left side of Figure 2 shows the front part of miniHPC exposing the switches and nodes. The right side of Figure 2 shows the back part of miniHPC exposing the connections of the cluster.

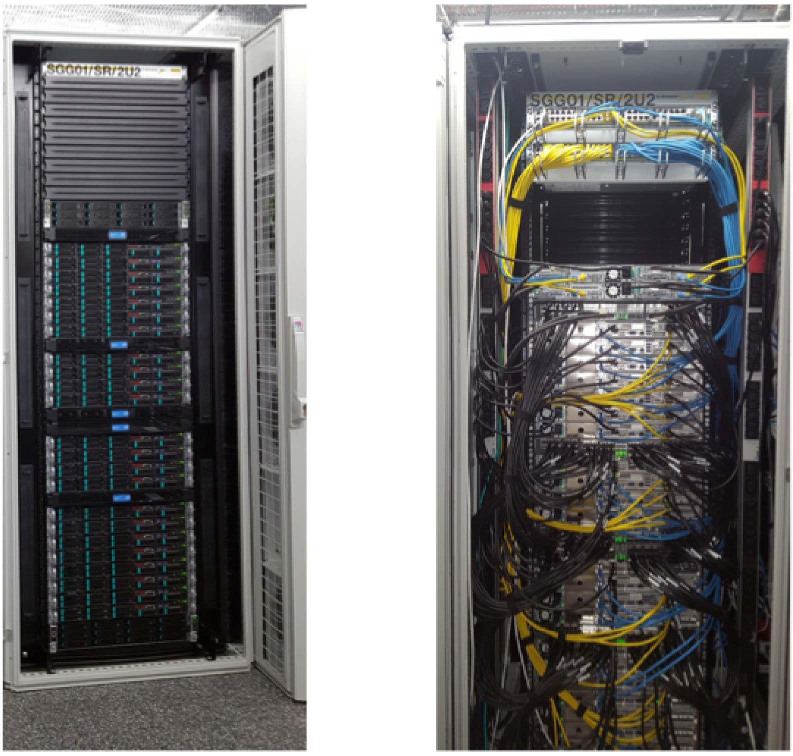

Figure 2: Pictures of the front side (left, switches and nodes) and the back side (right, connections) of miniHPC.

How to login to the miniHPC cluster?

Prerequisites:

a. Internet access

b. ssh client software
   - For Mac and Linux machines, one can just use the terminal
   - For Windows machine, one can install PuTTY or MobaXterm

Login Procedure:

Type into your ssh client
- ssh -X UserName@cl-login.dmi.unibas.ch
- It will ask for a password, you must use the same password as your university email password

After successfully login, you will be at your home directory on the miniHPC login node.

Note that miniHPC is a shared computing facility with limited resources. Therefore, we employ a batch scheduling software to deploy the execution of user jobs (e.g., your applications) on it.

What is a batch scheduling system?

The batch scheduling system (usually called job scheduler) is a single point to interface with the cluster, to which users can request one or more of the cluster resources to execute their programs (jobs). The job scheduler allocates resources to different users based on an allocation policy defined by the cluster administrators. There are many job schedulers, and on miniHPC we use Slurm.

How to use Slurm to manage job submission?

There are 3 basic commands:

Command	**Format**	**Description**
sbatch	sbatch [options] [scriptfile | [script args]]	To submit jobs to the scheduler
scancel	scancel [-u user] [Job_ID] . . .	To cancel running or queued jobs
squeue	squeue [-u user] . . .	To display the status of jobs

miniHPC usage hands-on:

a. To copy files: These commands run only on your local machine
- To miniHPC: *scp certain_file username@cl-login.dmi.unibas.ch:destination_path*
- From miniHPC: *scp username@cl-login.dmi.unibas.ch:certain_file destination_path*

b. To login
- *ssh –X UserName@cl-login.dmi.unibas.ch*
- Once logged in, you will be redirected to your home directory

c. To use Intel tools including the Intel compiler and Intel MPI
- Type *ml intel*
- Compiling with Intel compiler has to be done on the login node using *icc* or *mpiicc* commands

d. To generate a job script
- You can use the following tool `https://courses.cs.unibas.ch/cluster/index.html`
- To run on a separate node and avoid interference from other jobs, check the exclusive box.
- Copy the generated script into a new file jobscriptname.sh on miniHPC (see step a above)

e. To submit a job
- *sbatch jobscriptname.sh*
- The system will assign an ID to your submitted job

f. To check job status
- *squeue*
- possible statuses *R* for running, *PD* for pending, *CA* for canceled, and *CG* for completed.

g. To delete a job
- *scancel Job_ID*

h. To check the output of your submitted job
- Upon completion of your job, you will find a new file, *slurm-Job_ID.out* that contains the output and any errors