Kohinoor-5 Cluster

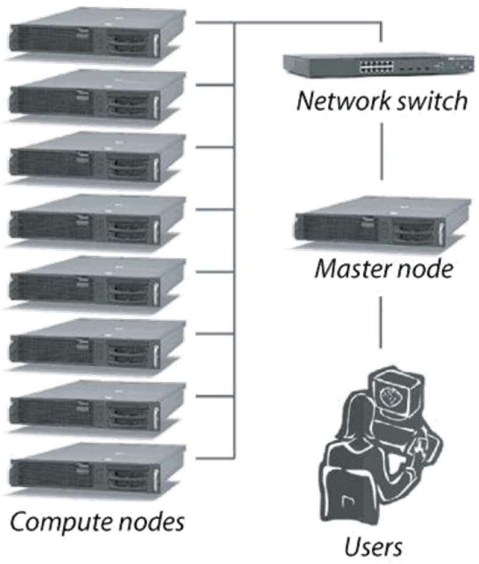
USER MANUAL

VER 2.0

SUBMITTED BY SYSTEM INTEGRATION TEAM MICROPOINT COMPUTERS PVT. LTD | Mumbai

Cluster Architecture Overview

- A Cluster comprises of multiple compute servers
- Connected together over a high speed interconnect.
- A Large mesh is submitted to the Master Server,
- Wherein it is broken into smaller sub-domains and
- Each sub-domain is submitted to a different
- Processor for computing. The JOB Decomposition
- Is done within the application (Cluster Version),
- It is migrated to various Processors using the
- System Middleware and Low latency interconnects
-
- **Key Advantages**
-
- Massive computing power.
- Dedicated Memory to Processor Performance
- Scalability to hundreds of Processors
- Excellent Price to Performance



System Architecture and Configuration

The HPC Cluster Setup is constructed in Two Rack.

Hardware Details	Qty
Exatron Server 2024S-TR - 2U as Login node	1
Exatron Server 2024S-TR - 2U as Master node	1
Exatron Server 2024S-TR - 2U as Compute Node	35
Exatron Server 2024S-TR - 2U as ZFS Storage Node	2
Exatron JBOD 946SE2C-R1K66JBOD as Storage	1
48 Port 10 Gigabit switches for Interconnect Primary Switch	1
48 Port 1G Gigabit switches for IPMI management.	1

Hardware Specifications

Kohinoor 5 Cluster is based on processor AMD EPYC 7643 48-Core Processor The cluster consists of compute nodes connected with 10G interconnect network. The system uses the ZFS file system.

- Total number of nodes: 35
  - Master nodes: 1
  - Login Node: 1

Login Nodes

K5 is an aggregation of a large number of computers connected through networks. The basic purpose of the Login and Master node is to manage and monitor each of the constituent component of K5 from a system's perspective. This involves operations like monitoring the health of the components, the load on the components, the utilization of various sub-components of the computers in K5.

Login Nodes:	
2* AMD EPYC 7313 Cores =16, 3.0 GHz	Total Cores = 32 cores
Memory= 512 GB	Total Memory = 512 GB
HDD = 1.5 TB x 1	

Master Nodes:	
2* AMD EPYC 7313 Cores =16, 3.0 GHz	Total Cores = 32 cores
Memory= 128 GB	Total Memory = 128 GB
HDD = 1 TB x 1	

CPU Compute Nodes

CPU nodes are indeed the work horses of K5. All the CPU intensive activities are carried on these nodes. Users can access these nodes from the login node to run interactive or batch jobs. Some of the nodes have higher memory, which can be exploitedby users in the aforementioned way.

CPU only Compute Nodes: 35	
2* AMD EPYC 7643 Cores = 48, 2.3 GHz	Total Cores = 3360 cores
Memory= 512 GB, DDR4 3200 MHz	Total Memory=17920 GB
SSD = 3.5 TB (local scratch) per node	

ZFS Storage Nodes

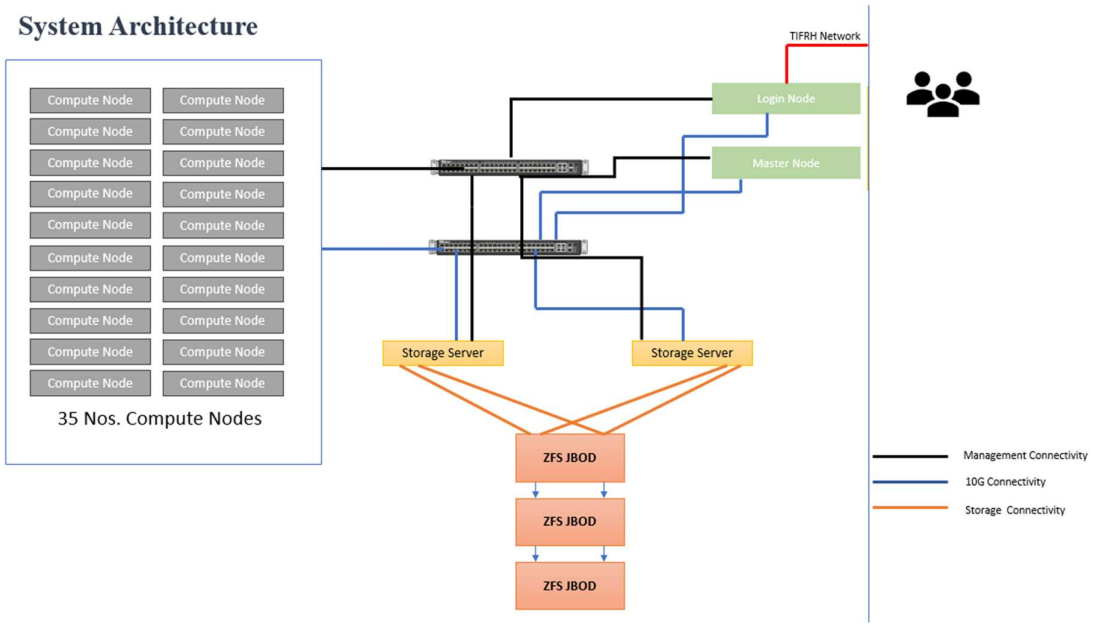
Typically, the purpose of the service node is to provide Security, Management, monitoring and other services to the cluster.

Storage

- Based on ZFS parallel file system
- Total useable capacity 860 TiB
- /home is 380TiB and /backup is 380TiB

Operating System

- Operating system on Kohinoor 5 is Rocky Linux 8.8 x86_64



Network infrastructure

A robust network infrastructure is essential to implement the basic functionalities of a cluster. These functionalities are:

a) Management functionalities i.e. to monitor, trouble shoot, start, stop various components of the cluster, etc. (Network/ portion of Network which implements this functionality is referred to as Management fabric).

b) Ensuring fast read/ write access to the storage (Network/ portion of Network which implements this functionality is referred to as storage fabric).

c) Ensuring fast I/O operations like connecting to other clusters, connecting the cluster to various users on the campus LAN, etc. (Network/ portion of Network which implements this functionality is referred to as I/O Fabric).

d) Ensuring High-Bandwidth, Low-latency communication amongst processors to for achieving high-scalability (Network/ portion of Network which implements this functionality is referred to as Message Passing Fabric)

Technically, ALL the aforementioned functionalities can be implemented in a single network. From the perspectives of requirements, optimal performance and economic suitability, the aforementioned functionalities are implemented using two different networks based on different technologies, as mentioned next

Primary Interconnection Network

Computing nodes of Kohinoor 5 are interconnected by a high-bandwidth, low-latency interconnect network.

10G Interconnect

10G is a high-performance communication architecture. This communication architecture offers low communication latency, low power consumption and a high throughput. All CPU nodes are connected via 10G interconnect network.

Secondary Interconnection Network Gigabit Ethernet: 1 Gbps

Gigabit Ethernet is the interconnection network that is most commonly available. For Gigabit Ethernet, no additional modules or libraries are required. The Open MPI, MPICH implementations will work over Gigabit Ethernet.

ZFS

ZFS Architecture

ZFS is a local file system and logical volume manager created by Sun Microsystems Inc. to direct and control the placement, storage and retrieval of data in enterprise-class computing systems.

The ZFS file system and volume manager is characterized by data integrity, high scalability and built-in storage features such as:

Replication - the process of making a replica (a copy) of something.
Deduplication - a process that eliminates redundant copies of data and reduces storage overhead.
Compression - a reduction in the number of bits needed to represent data.
Snapshots - a set of reference markers for data at a particular point in time.
Clones - an identical copy of something.
Data protection - the process of safeguarding important information from corruption and/or loss.

ZFS Mount point in K-5

```
[root@vault1 ~]#  zpool list
NAME     SIZE  ALLOC   FREE  CKPOINT  EXPANDSZ   FRAG    CAP  DEDUP    HEALTH  ALTROOT
backup   500T  18.6T   482T        -         -     0%     3%  1.00x    ONLINE  -
home     500T  3.17T   497T        -         -     0%     0%  1.00x    ONLINE  -
[root@vault1 ~]#
```

Rack Overview

RACK	Rack 1	RACK	Rack 2	RACK	Rack 3
42	Blank	42	Blank	42	Blank
41	Blank	41	10 G Switch	41	Blank
40	Blank	40	Blank	40	Blank
39	Blank	39	Management Switch	39	Blank
38	Blank	38	Blank	38	Blank
37	Blank	37	Blank	37	Blank
36	Blank	36	Blank	36	Blank
35	Compute Node-16	35	Compute Node-19	35	Compute Node-35
34		34		34	
33	Compute Node-15	33	Compute Node-18	33	Compute Node-34
32		32		32	
31	Compute Node-14	31	Compute Node-17	31	Compute Node-33
30		30		30	
29	Compute Node-13	29	Login Node	29	Compute Node-32
28		28		28	
27	Compute Node-12	27	Blank	27	Compute Node-31
26		26	Master Node	26	
25	Compute Node-11	25		25	Compute Node-30
24		24	17" LCD Console & KVM	24	
23	Blank	23	Blank	23	Blank
22	Compute Node-10	22	Storage Node-2	22	Compute Node-29
21		21		21	
20	Compute Node-9	20	Blank	20	Compute Node-28
19		19	Storage Node-1	19	
18	Compute Node-8	18		18	Compute Node-27
17		17	Blank	17	
16	Compute Node-7	16	Storage JBOD-3	16	Compute Node-26
15		15		15	
14	Compute Node-6	14		14	Compute Node-25
13		13		13	
12	Blank	12	Blank	12	Blank
11	Compute Node-5	11	Storage JBOD-2	11	Compute Node-24
10		10		10	
9	Compute Node-4	9		9	Compute Node-23
8		8		8	
7	Compute Node-3	7	Blank	7	Compute Node-22
6		6	Storage JBOD-1	6	
5	Compute Node-2	5		5	Compute Node-21
4		4		4	
3	Compute Node-1	3		3	Compute Node-20
2		2	Blank	2	
1	Blank	1	Blank	1	Blank

Accessing the cluster

The cluster can be accessed through master nodes, which allows users to login.

- You may access master node through ssh.
- The Login node is primary gateway to the rest of the cluster, which has a job scheduler (called PBS). You may submit jobs to the queue and they will run when the required resources are available.
- Please do not run programs directly on master node. Login node is use to submit jobs, transfer data and to compile source code. (If your compilation takes more than a fewminutes, you should submit the compilation job into the queue to be run on the cluster.)
- By default, two directories are available (i.e. /home and /scratch). These directories are available on login and master node as well as the other nodes on the cluster. /scratch is for temporary data storage, generally used to store data required for running jobs.

Remote Access

Using SSH in Windows

To access K5 you need to "ssh" the master server. PuTTY is the most popular open source "ssh" client application for Windows, you can download it from (http://www.putty.org/). Once installed, find the PuTTY application shortcut in your Start Menu, desktop. On clicking the PuTTY icon The PuTTY Configuration dialog should appear.Locate the "Host Name or IP Address" input Field in the PuTTY Configuration screen. Enterthe user name along with IP address or Hostname with which you wish to connect.

(e.g. [username]@ k5login.tifrh.res) Enter your

password when prompted, and press Enter.

Using SSH in Mac or Linux

Both Mac and Linux systems provide a built-in SSH client, so there is no need to install any additional package. Open the terminal, connect to an SSH server by typing the following command:

```
ssh [username]@[hostname]
```

For example, to connect to the K5 Master Node, with the username

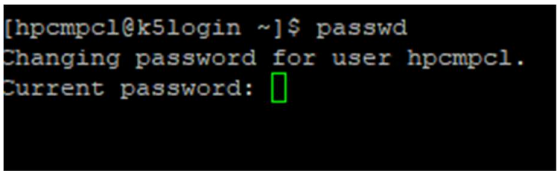
```
hpcmpcl: ssh hpcmpcl@ k5login.tifrh.res  (outside access)
```

You will be prompted for a password, and then will be connected to the server.

Password

How to change the user password?

Use the **passwd** command to change the password for the user from login node.



Transferring files between local machine and HPC cluster

Users need to have the data and application related to their project/research work on K5.

To store the data special directories have been made available to the users with name “scratch and home” the path to this directory is “/scratch” and “/home”. Whereas these directories are common to all the users, a user will get his own directory with their username in /scratch/ as well as /home/ directories where they can store their data.

```
/home/<username>/: ! This directory is generally used by the user to
install applications.
```

```
/scratch/users/<username>/: ! This directory is user to store the user
datarelated to the project/research.
```

However, there is limit to the storage provided to the users, the limits have been defined according to quota over these directories, all users will be allotted same quota by default. When a user wishes to transfer data from their local system (laptop/desktop) to HPC system, they can use various methods and tools.

A user using ‘Windows’ operating system will get methods and tools that are native to Microsoft windows and tools that could be installed on your Microsoft windows machine. Linux operating system users do not require any tool. They can just use “scp” command on their terminal, as mentioned below.

Users are advised to keep a copy of their data with themselves, once the project/research work is completed by transferring the data in from K5 to their local system (laptop/desktop). The command shown below can be used for effecting file transfers (In all the tools):

```
Scp –r <path to the local data directory>   <your username>@<IP of
k5-login>:<path to directory on HPC where to save the data>
```

Example:

Same Command could be used to transfer data from HPC system to your local system (laptop/desktop).

```
Scp  –r  /dir/dir/file  hpcmpcl@<cluster IP/Name>:/home/hpcmpcl
```

Example:

```
Scp –r  <path to directory on HPC>   <your username>@<IP of local
system>:<path to the local data directory>
```

```
Scp  –r  /home/hpcmpcl  hpcmpcl@<local system IP/Name>:/dir/dir/file
```

Note: The Local system (laptop/desktop) should be connected to the network with which it can access the HPC system.

To reiterate,

Copying Directory/File from local machine to K5:

To copy a local directory from your Linux system (say Wrf-2.0) to your home directory in your K5 HPC account, the procedure is:

1. From terminal go to the parent directory using cd command.

   user@mylaptop:~$cd ~/MyData/

2. Under parent directory type ls <& press Enter key>, & notice Wrf-2.0 is there.
   user@mylaptop: ~$ls Files TempFiles-0.5 Wrf-2.0
3. Begin copy by typing:

   user@mylaptop:~$ scp -r Wrf-2.0 (username)@ k5login.tifrh.res

   < you will be prompted for password ; enter your password >

4. Now login to your account as: user@mylaptop:~$ ssh (your username)@ k5login.tifrh.res< you will be prompted for password ; enter password > [user1@login ~]$
5. ls command, you should see Wrf-2.0 directory.
6. While copying from K5 to your local machine, follow the same steps

By interchanging source and destination in the scp command. Refer to the generic copyingdescribed earlier.

Tools

MobaXterm (Windows installable application):

It is a third party freely available tool which can be used to access the HPC system andtransfer file to K5 system through your local systems (laptop/desktop).

Link to download this tool : https://mobaxterm.mobatek.net/download-home-edition.html

Command Prompt (Windows native application):

This is a native tool for Windows machine which can be used to transfer data from K5 system through your local systems (laptop/desktop).

PowerShell (Windows native application):

This is a This is a native tool for Windows machine which could be used to transfer datafrom K5 system through your local systems (laptop/desktop).

WinSCP (Windows installable application):

This popular tool is freely available and is used very often to transfer data from Windowsmachine to Linux machine. This tool is GUI based which makes it very user-friendly.

Link for this tool is : https://winscp.net/eng/download.php

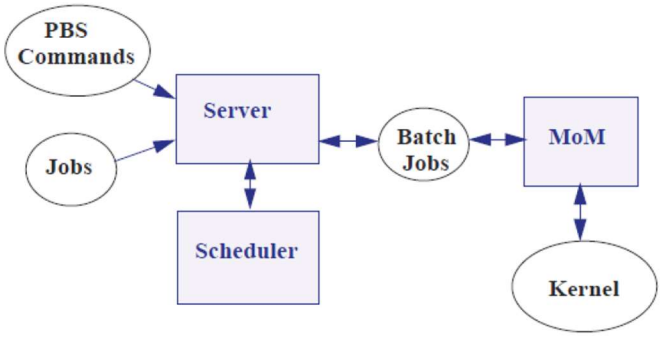
Job Scheduler – PBS

PBS Professional is workload manager and job scheduler for high-performance computing (HPC) environments, used to improve productivity, optimize resource utilization & efficiency, and simplify the process of cluster workload management.

PBS keeps track of which hardware and licenses are available, and all waiting and running tasks. PBS matches the requirements of each of your tasks to the right hardware, licenses, and time slot, and makes sure that tasks are run according to the site's policy.

PBS pro is configured in High availability mode. If primary PBS server fails then jobs will be moved to secondary PBS server.

PBS components:



Jobs are submitted to the PBS server. The scheduler chooses where and when to run the jobs, and the server sends the jobs to MoM. PBS com-mands communicate with the server.

PBS Commands

PBS provides a set of commands that you can use to submit, monitor, alter, and delete jobs. The PBS commands can be installed on any supported platform, with or without the other PBS components.

Some PBS commands can be run by any PBS user, while some require administrator or operator privilege. Some commands provide extended features for administrators and operators.

PBS Job

A PBS job is a task, in the form of a shell script, cmd batch file, python script, etc. describing the commands and/or applications you want to run. You hand your task off to PBS, where it becomes a PBS job.

Server

The PBS server manages jobs for the PBS complex. PBS commands talk to the PBS server, jobs are submitted to the server, and the server queues the jobs and sends them to execution hosts.

Scheduler

The scheduler runs jobs according to the policy specified by the site administrator. The scheduler matches each job's requirements with available resources, and prioritizes jobs and allocates resources according to policy.

MoM

MoM manages jobs once they are sent to the execution host. One MoM manages the jobs on each execution host. MoM stages files in, runs any prologue, starts each job, monitors the job, stages files out and returns output to the job submitter, runs any epilogue, and cleans up after the job. MoM can also run any execution host hooks.

Queue configuration

Sr. NO	Queue Name	Max run	Max ncpu	Max wall time	Max queue
1	cpuq	384 core is for all user in cluster	96	3 days	192 cores
2	dftq	96 core is for all user in cluster	96	3 days	N/A

PBS Commands

qsub	Submit a job
qstat	Show status of batch jobs
qdel	Delete a job
qalter	Alter a job's attributes
qhold	Put the job on hold
qmove	Move a job to different queue or server
qrerun	Terminate an executing job and return it to a queue
qselect	Obtain a list of jobs that meet certain criteria
pbsnodes	Obtain a detailed listing of all the hosts
Qmgr	Provides an administrator interface to query and configure batch system parameters

Submitting a serial Job:

[hpcmpcl@k5login ~]$vim submit.sh

#!/bin/bash

#These commands set up the Grid Environment for your job:

#PBS -N ExampleJob

#PBS -l select=1,

#PBS -q select queue (i.e. run, short,long)

#print the time and date

date

#wait 10 seconds

sleep 10

#print the time and date again

Date

Submitting Parallel Job

```
#!/bin/bash
#PBS -N "Job Name"
#PBS -q workq
#PBS -o $PBS_JOBID.out
#PBS -e $PBS_JOBID.err
#PBS -l select=2:nprocs=32:mpiprocs=32
#PBS -S /bin/bash -V
export I_MPI_FABRICS=shm:tmi
cd $PBS_O_WORKDIR
cat $PBS_NODEFILE > pbs_nodes
cd $PBS_O_WORKDIR
echo Working directory is $PBS_O_WORKDIR
NPROCS=`wc -l < $PBS_NODEFILE`
NNODES=`uniq $PBS_NODEFILE | wc -l`
mpirun -hostfile $PBS_NODEFILE -np ${NPROCS} ./hi
```