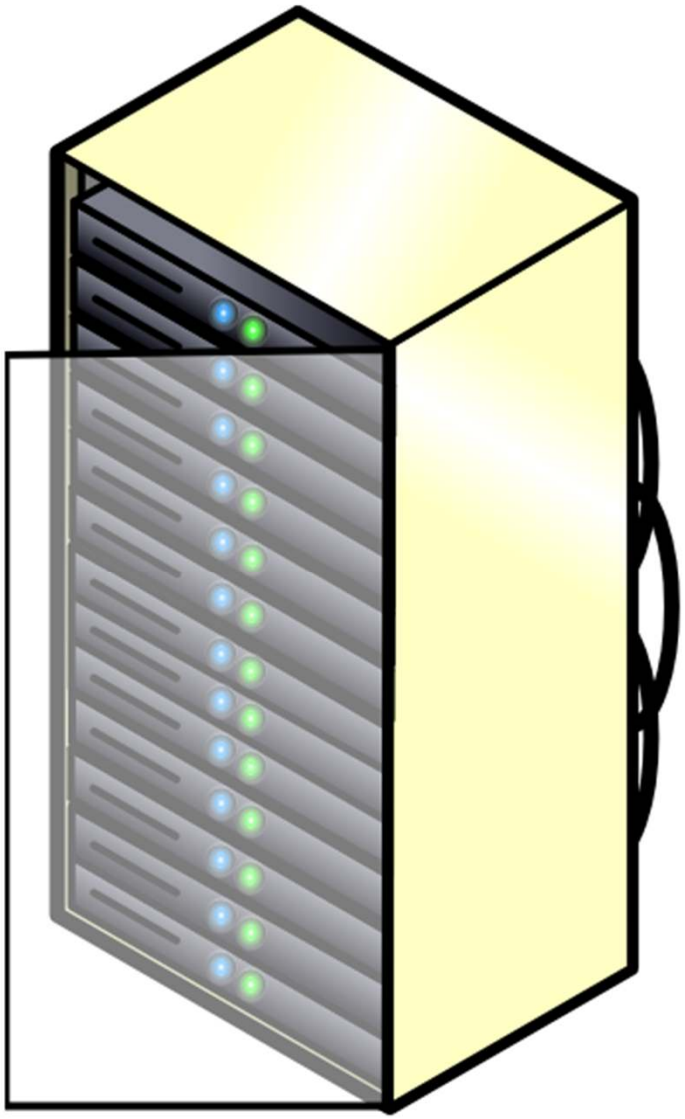
PARAM SHAKTI

Agenda

- Overview to High Performance Computing.
- What is cluster and its Types
- Components of HPC
- PARAM-Shakti Architecture
- Technical Specification of PARAM-Shakti
- How to acceess Param-Shakti.
- C-Chakshu

Cluster Terminology

- Cluster is a group of machines interconnected in a way that they work together as a single system
- Terminology :
  - **Node** – individual machine in a cluster
  - **Head/Master node** – connected to both the private network of the cluster and a public network and are used to access a given cluster.
  - **Compute nodes** – connected to only the private networks of the cluster and are generally used for running jobs assigned to them by the login node(s)
  - Compute nodes can be of different types:
    - CPU only nodes
    - GPU nodes
    - High memory nodes, etc



When one server is not enough

- If the computational task or analysis to complete is daunting for a single server, cluster are used.

Types of HPC cluster

Statefull (Diskfull) Cluster

- Traditional Cluster with OS on each nodes local disk.

Stateless (Diskless) Cluster

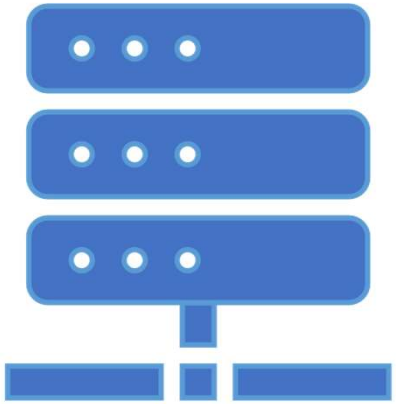
- Nodes are booted using RAMdisk Osimage.

Components of HPC

1. Nodes (servers)
2. Parallel File System (Storage & performance)
3. Interconnect (Networking & application execution)
4. Accelerator cards (boosting the performance with many)
5. Optimized compilers and libraries

lustre

Components of HPC



Nodes:

- Nodes are the actuall server who will take part in computation.
- They have multiple cores and supports Hyperthreading.

Lustre (Parallel File System)

- Lustre is parallel File System where multiple clients can write to the differents part of same file at the same time multiple clients can read the file.
- It supports High bandwith Interconnects such as Mellanox,Omnipath,etc.
- It is a scable file system.
- It supports HA and its is a POSIX complaint file system.
- ACL can be applied.

Graphical processing Unit

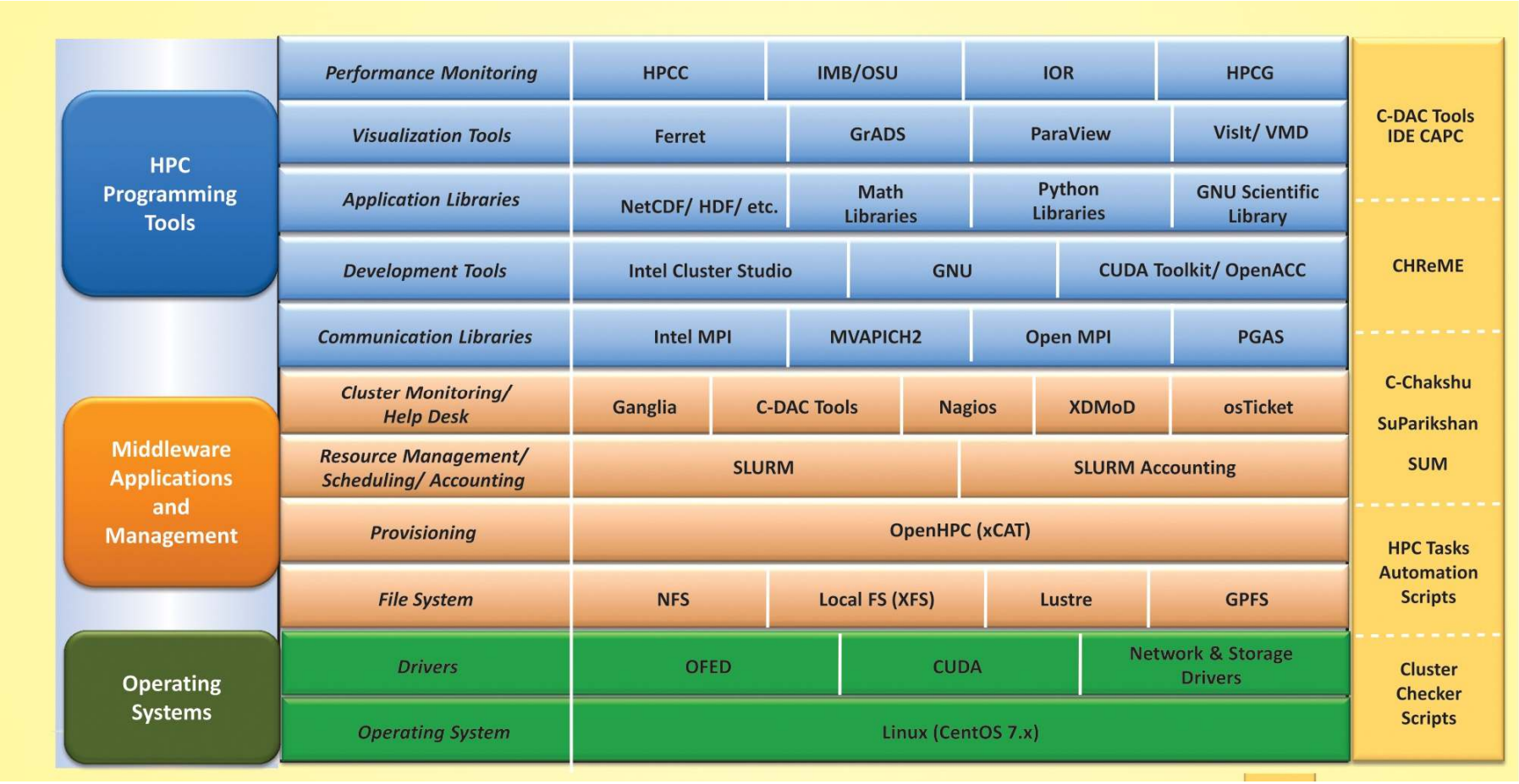
- A co-processor to accelerate general purpose scientific and engineering computing.
- It accelerates applications running on CPU by offloading compute intensive and time consuming portion of the code.
- GPU consists of thousands of smaller cores which together operate to crunch data in the application.

Infiniband

:

- High throughput and low latency technology that interconnects compute nodes and I/O nodes to form a system area Network.
- It uses RDMA (Remote Direct Memory Access) Protocol

Software Stack



Software Components

- **Operating System –**
  - HPC clusters generally are build with Linux operating system as a base OS (Centos7.6)
  - It includes all the device drivers for the H/W connected to each node.

- **Cluster Manager/Orchestrator**
  - Tools in this category builds a centralized architecture where a controller node builds and manages the cluster.
  - xCAT – Is a open source cluster Manager, developed by IBM, Maintained by community, is the widely used tool for HPC as well as cloud clusters.
  - It provides flexibility to handle objects within the cluster with its easy manageable methods
  - It provide methods to deploy nodes with a very light weight stateless images.

- **Resource Manager (SLURM)**

  - As there are a lot of resource within a cluster like : CPU-Cores, Memory banks, GPU accelerator cards managing which becomes a tedious task for a user and a system administrator.

  - Resource manager with in “slurm” tool helps to manage and represent resources to the users in a simplest way.

- **Job Schedular (SLURM)**

  - A scheduler checks the available resources within a cluster and manages which jobs run where and when.
  - Allocating resources to each users for optimal utilization of system resources.
  - Provides multiple algorithm, which provides different ways to initiate jobs on the resources.
  - BACKFILL scheduling is the widely used and the most efficient algorithm.
  - Provider batch jobs as well as Interactive jobs submission methods.

Accessing the cluster

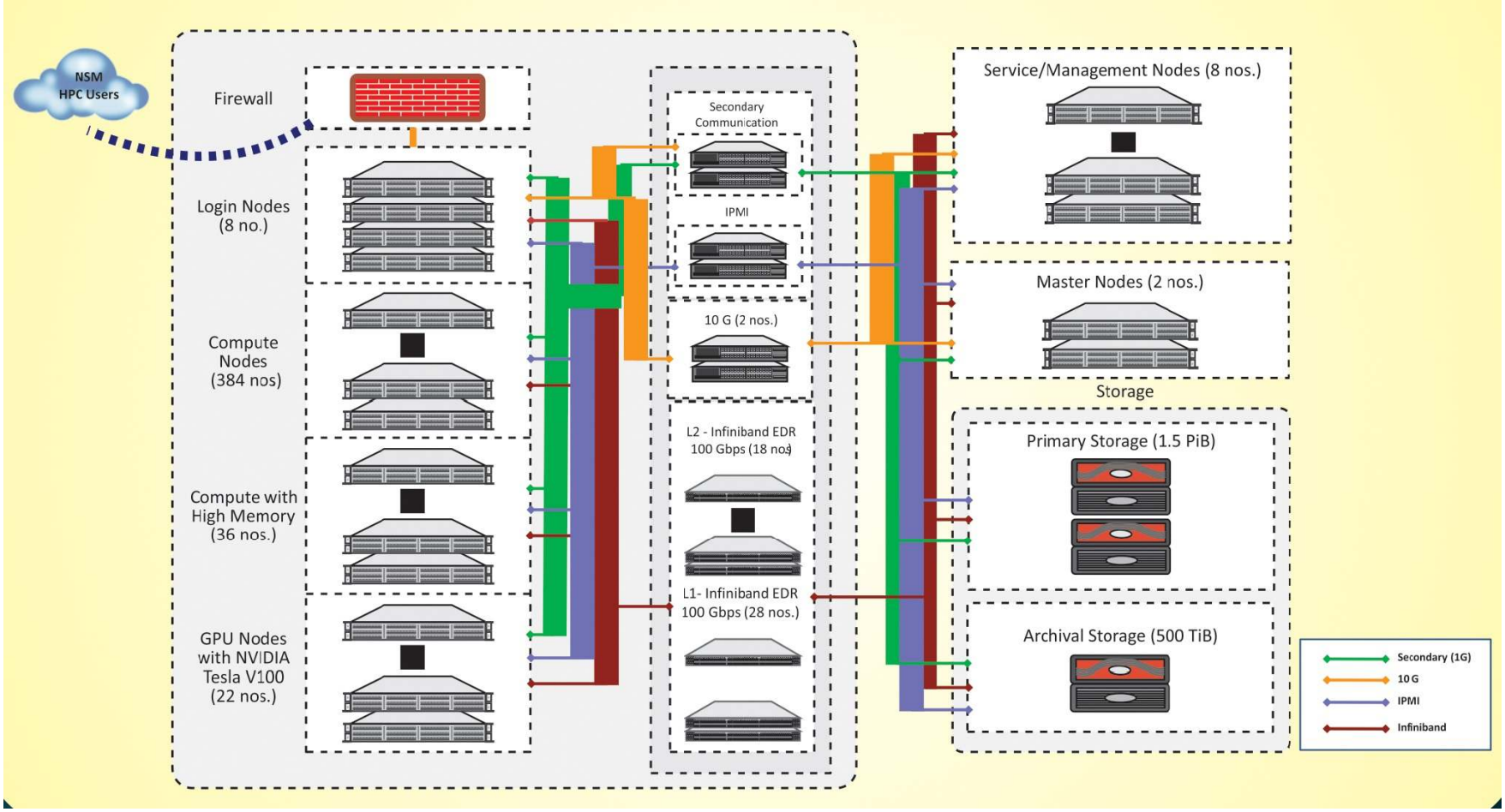
Login Environment

- The cluster can be accessed through 8 general login nodes.
- The login nodes is primary gateway to the rest of the cluster.
- User can perform all its functions on login node.
- All libraries, compilers, preinstalled applications, user installed application are available over login nodes.

Remote Login

- You may access login node through ssh.
- Using SSH in Windows (Putty,Moab-xterm,etc).
- Using SSH in Linux via terminal (ssh –p 4422 example-user@paramshakti.iitkgp.ac.in).
- For example, to connect to the PARAM Shakti Login Node, with the username.

PARAM SHAKTI ARCHITECTURE DIAGRAM



PARAM Shakti Performance Specifications(IIT-KGP)

- Rpeak: **1.66** PFLOPS.
- Rmax: **850** TFLOPS (CPU only Nodes) + **200** TFLOPS (GPU Nodes).
- Total Nodes:420 CPU only nodes + 22 GPU nodes.
- Total Cores:2,42,080.
- Total Memory: 105.6 TB.
- Storage: 1.5PiB PFS + 500 TiB Archival.

Listed as top 3rd system among Indias top performing HPC

Rank	Site	System	Cores/Processor Sockets/Nodes	Rmax (TFlops)
1	Indian Institute of Tropical Meteorology(IITM), Pune	Cray XC-40 class system with 3315 CPU-only (Intel Xeon Broadwell E5-2695 v4 CPU) nodes with Cray Linux environment as OS, and connected by Cray Aries interconnect. OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	119232/ /3312	3763.9
2	National Centre for Medium Range Weather Forecasting (NCMRWF), Noida	Cray XC-40 class system with 2322 CPU-only (Intel Xeon Broadwell E5-2695 v4 CPU) nodes with Cray Linux environment as OS, and connected by Cray Aries interconnect OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	83592//2322	2570.4
3	Indian Institute of Technology (IITK), Kharagpur	The supercomputer PARAM Shakti is based on a heterogeneous and hybrid configuration of Intel Xeon Skylake(6148, 20C, 2.4Ghz) processors, and NVIDIA Tesla V100. The system was designed and implemented by HPC Technologies team, Centre for Development of Advanced Computing (C-DAC) with total peak computing capacity of 1.66 (CPU+GPU) PFLOPS performance. The system uses the Lustre parallel file system (primary, storage 1.5 PiB usable with 50 GB/Sec write throughput. Archival Storage 500TiB based on GPFS)| OEM: Atos India Pvt Ltd., Bidder: Atos India Pvt Ltd.	17280/2/432	935

Technical Specifications

CPU only Compute Nodes

- 384 Nodes
- 15360 Cores
- Compute power of Rpeak 1179 TFLOPS
- Each Node with
  - 2* Intel Xeon SKL G-6148, 20 cores, 2.4 GHz, processors
  - 192 GB memory
  - 480 GB SSD

High Memory Compute Nodes

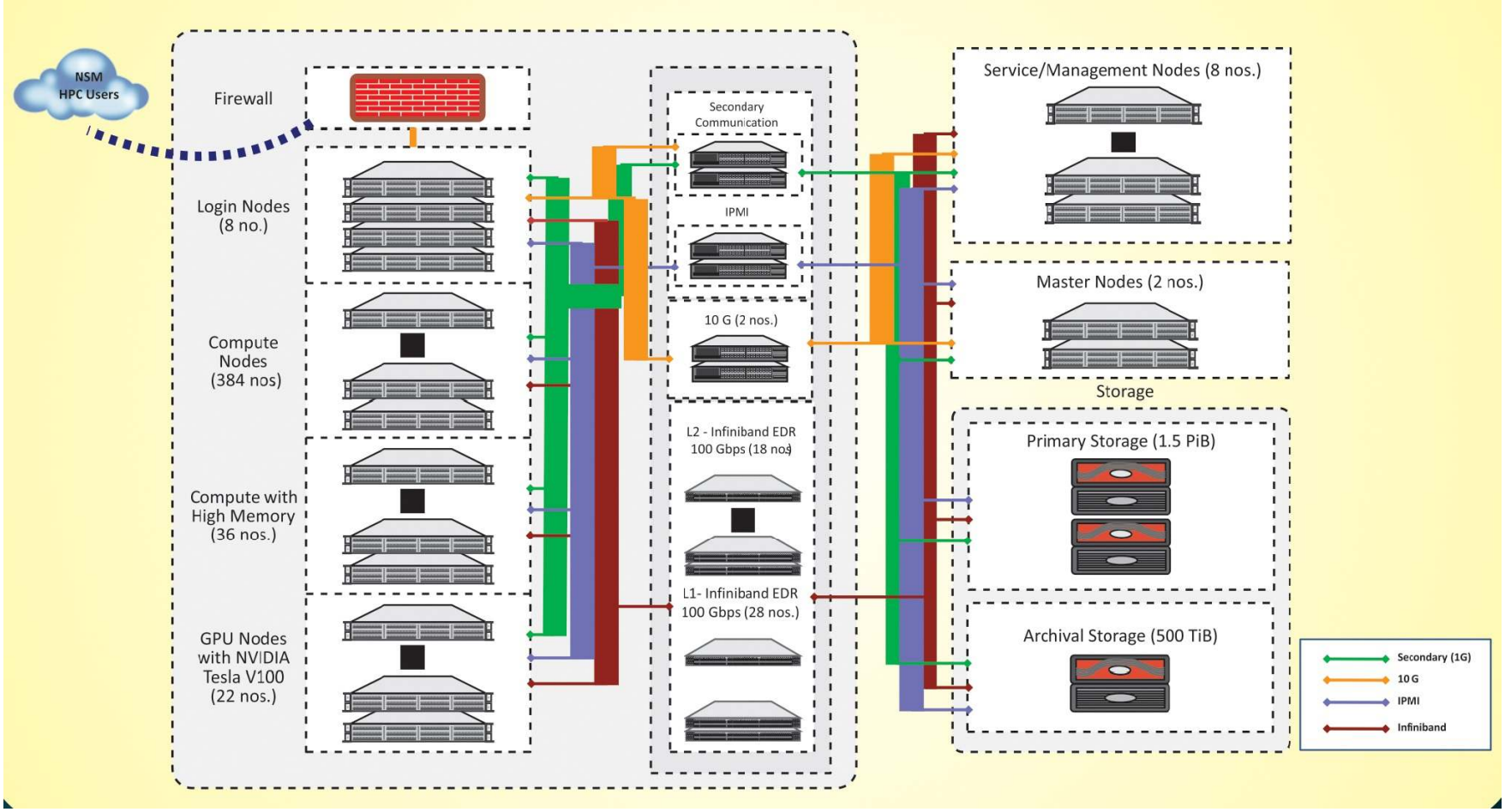
- 36 Nodes
- 1440 Cores
- Compute power of Rpeak 110.59 TFLOPS
- Each Node with
  - 2* Intel Xeon SKL G-6148, 20 cores, 2.4 GHz, processors
  - 768 GB memory
  - 480 GB SSD

GPU Compute Nodes

- 22 Nodes
- 880 CPU cores
- 225280 GPU Cores
- Compute power of Rpeak 67 TFLOPS + 308 TF
- Each Node with
  - 2* Intel Xeon SKL G-6148, 20 cores, 2.4 GHz, processors
  - 192 GB Memory
  - 480 GB SSD
  - 2xNvidia V100 SXM2 GPU cards each with 5120 CUDA cores

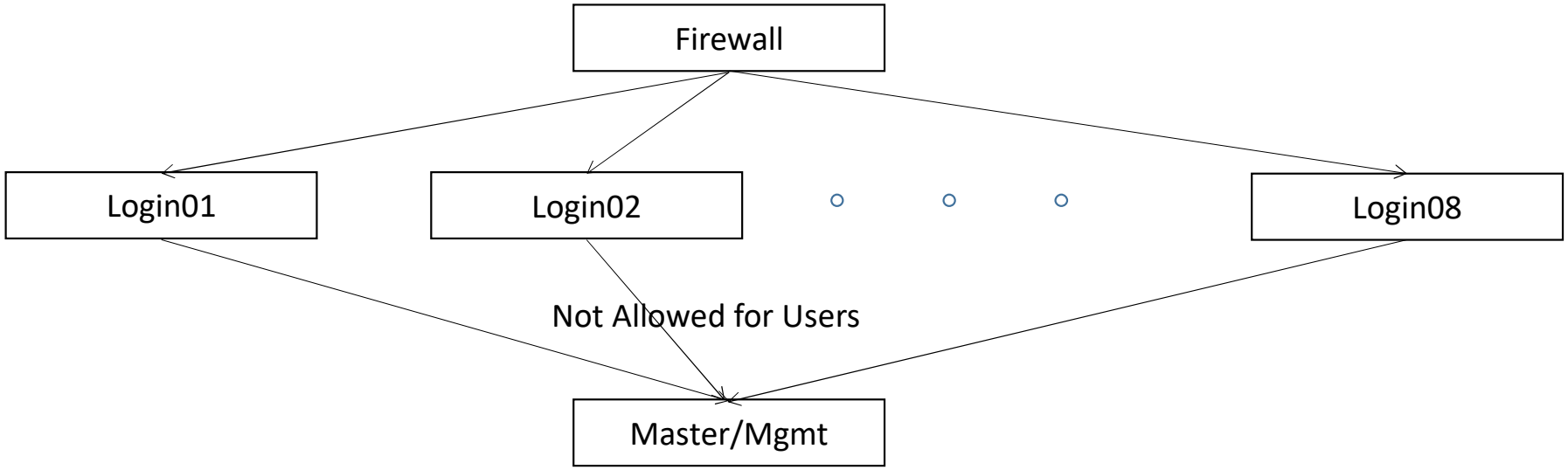

परम शक्ति
PARAM
shakti

PARAM SHAKTI ARCHITECTURE DIAGRAM



Access Policy

- **Access to Login nodes are in Round-Robin Mode.**
- **Users are not allowed to access Master/Management Nodes**



How to Access PARAM Shakti ?

- **If you are using windows you can access via(SSH Clients):**
  - MobaXterm
  - Putty, etc
- **Within IITKGP Campus:**

  ssh username@paramshakti.iitkgp.ac.in
- **Outside IITKGP Campus:**

  ssh username@paramshakti.iitkgp.ac.in -p 4422

Ticketing Tool

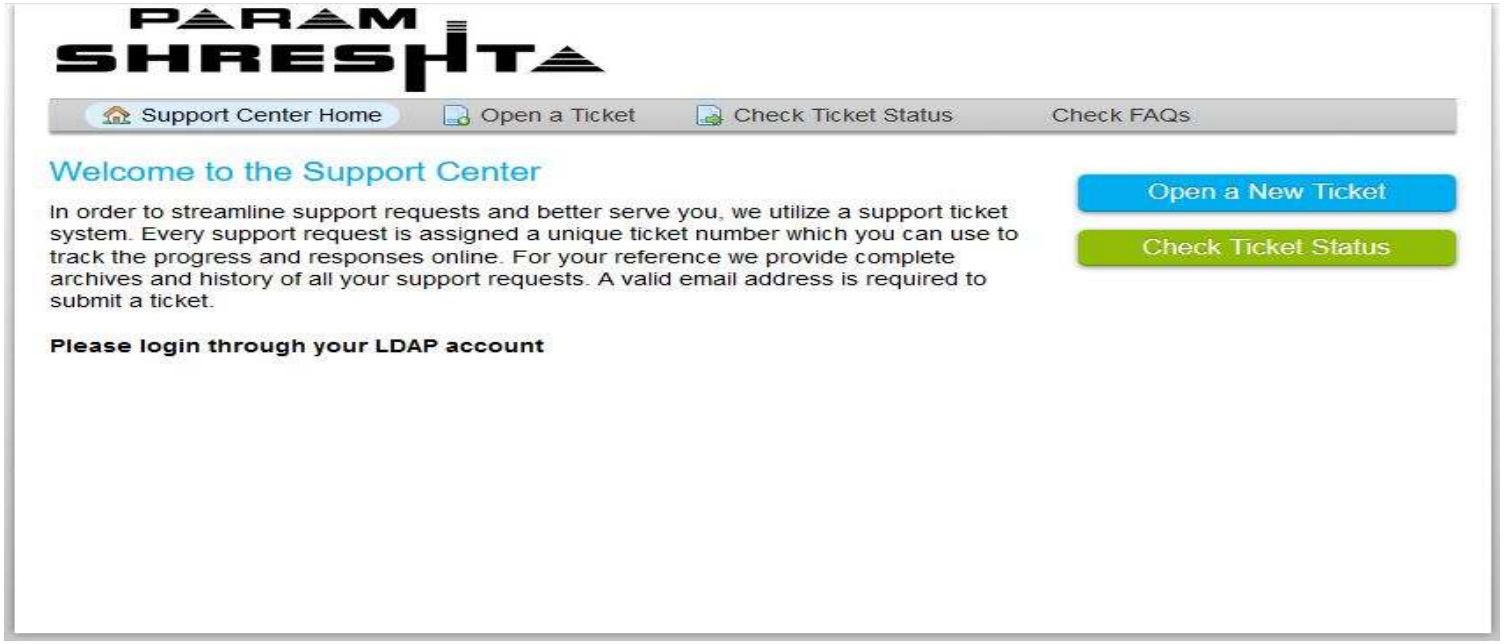
- **A Support Portal is created for Assisting the Users.**

  https://paramshakti.iitkgp.ac.in/support

Upcoming Mechanism for Accessing the Support Portal.

- **Users have to find their solutions in the FAQ First.**

  https://paramshakti.iitkgp.ac.in



National Supercomputing Mission

Centre for Development of Advanced Computing

General FAQ

How to get account on HPC cluster ?

- Get 'User Account Creation Form'
- Fill the relevant details.
- Get the signatures of your Head of the Department and the 'approving authority'.
- You will receive an Email in your official Email ID intimating the creation of your account

Table of contents

If your application is hybrid :

`export I_MPI_FALLBACK="0"` [Do not switch to other available network]

`export I_MPI_FABRICS="shm:ofa"`

`export I_MPI_FABRICS="shm:dapl"` (if using DAPL)

For OpenMP :

`export I_MPI_FABRICS="shm:shm"`

To check which fabric is currently used, you can set the I_MPI_DEBUG environment variable to 2:

`mpirun –np n -genv I_MPI_DEBUG=2` your_command/command_path ; where "n" => number of processes.

For Ex. : `mpirun -np 48 -genv I_MPI_DEBUG=2 myprog`

You can also specify above variables in your mpirun command :

`mpirun –np n -genv I_MPI_FALLBACK=0 -genv I_MPI_FABRICS="shm:ofa"` your_command/command_path

For Ex. : `mpirun –np 48 -genv I_MPI_FALLBACK=0 -genv I_MPI_FABRICS="shm:ofa" myprog`

Cannot find your queries ? Click here to create a ticket.

Next
Environment →

Monitoring Tools

Ganglia

- Ganglia is a scalable distributed monitoring system for high-performance computing systems, clusters and networks.

- It is based on a hierarchical design targeted at federations of clusters

https://paramshakti.iitkgp.ac.in/ganglia

C-CHAKSHU

MULTI CLUSTER MONITORING PLATFORM

Contents

C-Chakshu

C-Chakshu is a Multi cluster Monitoring tool developed as per special requirements of HPC system administrators and users under NSM project.

Salient Features

- Centralized multi cluster monitoring
- Application performance monitoring and analysis
- Integrated Ticketing system for better user support
- History of HPC usage statistics and reporting
- Live and quick infrastructure visibility via graphs
- Real time 3D HPC system rack view
- Job accounting and analysis
- HPC infrastructure health monitoring

Roles and their Purpose

1. **Admin**
   - To monitor various metrics of HPC system.
   - To manage multiple HPC systems under NSM.
   - To manage the users reliably.
2. **User**
   - To check job status, resources used and available, application monitoring.
   - To use ticketing system for better user support.

What is User Portal?

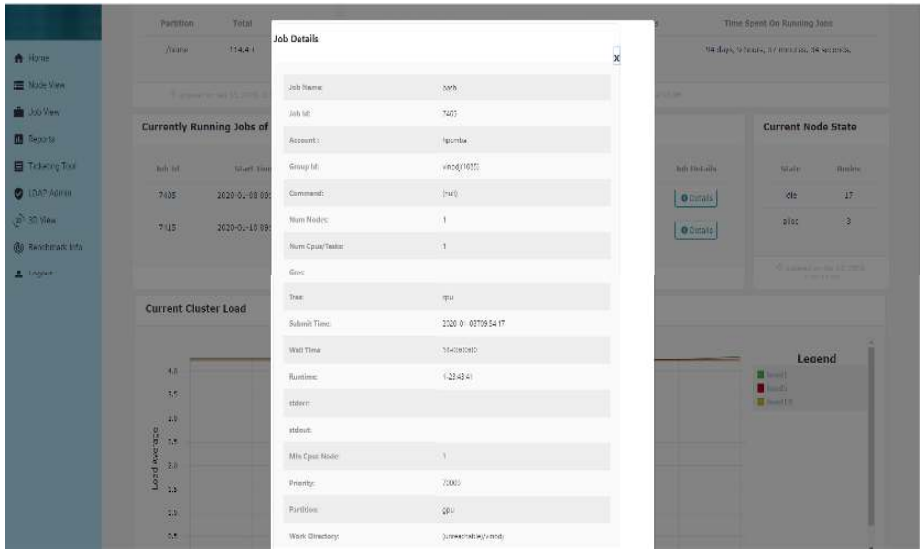
- Exclusively for the HPC application users
- Information about HPC system and available resources
- User specific Job queue list with detailed information about the job.
- Job-wise allocated nodes information.
- Specific node monitoring for exact utilization details.

What is User Portal?

- Exclusively for the HPC application users
- Information about HPC system and available resources
- User specific Job queue list with detailed information about the job.
- Job-wise allocated nodes information.
- Specific node monitoring for exact utilization details.

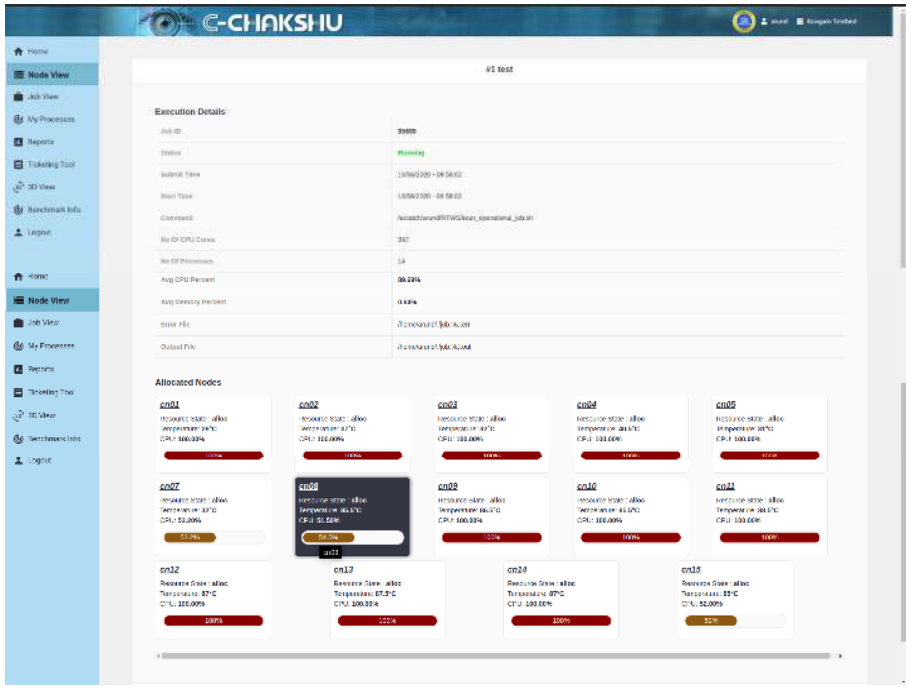
What is User Portal?

- Exclusively for the HPC application users
- Information about HPC system and available resources
- User specific Job queue list with detailed information about the job.
- Job-wise allocated nodes information.
- Specific node monitoring for exact utilization details.



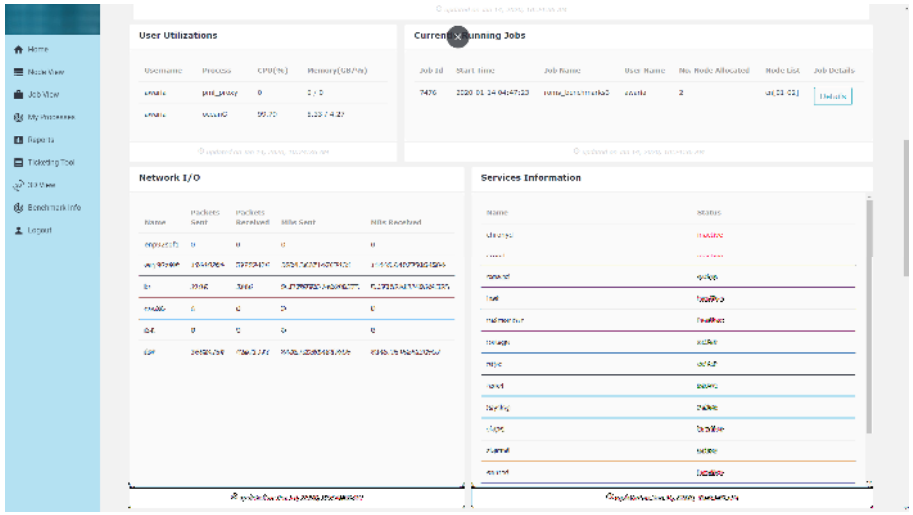
What is User Portal?

- Exclusively for the HPC application users
- Information about HPC system and available resources
- User specific Job queue list with detailed information about the job.
- Job-wise allocated nodes information.
- Specific node monitoring for exact utilization details.



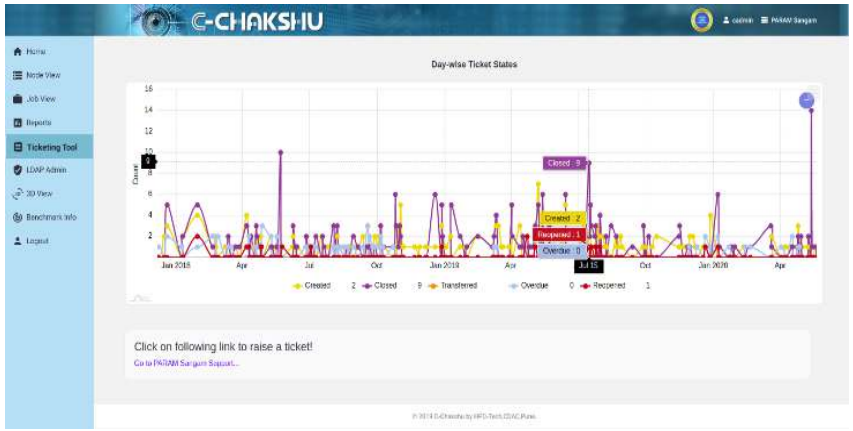
What is User Portal?

- Information about HPC system and available resources
- User specific Job queue list with detailed information about the job.
- Job-wise allocated nodes information.
- Specific node monitoring for exact utilization details.



What is User Portal?

- Ticketing system for raising and following up on issue related to HPC system/application.
- Monitor user specific processes running on different nodes in single window.
- User specific reporting facility for usage analysis.
- System and Application benchmarks

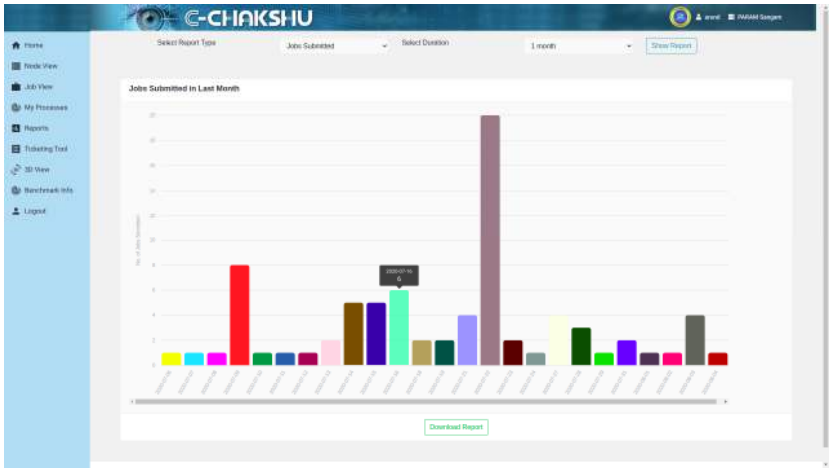


What is User Portal?

- Ticketing system for raising and following up on issue related to HPC system/application.
- Monitor user specific processes running on different nodes in single window.
- User specific reporting facility for usage analysis.
- System and Application benchmarks

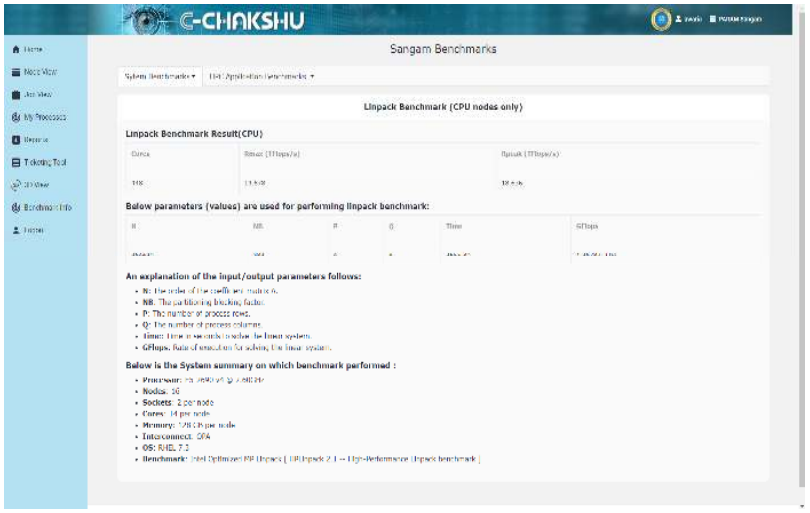
What is User Portal?

- Ticketing system for raising and following up on issue related to HPC system/application.
- Monitor user specific processes running on different nodes in single window.
- User specific reporting facility for usage analysis.
- System and Application benchmarks



What is User Portal?

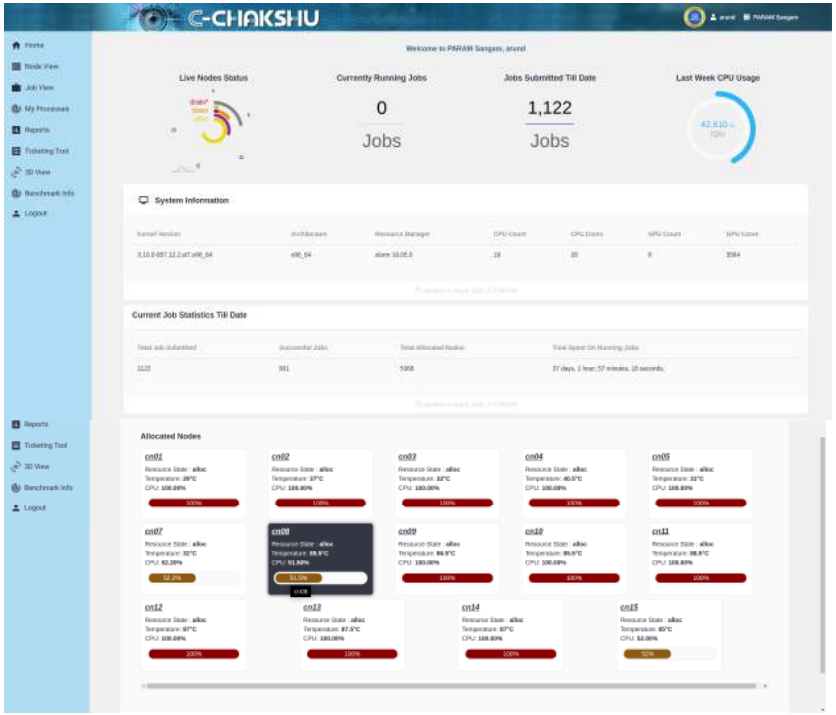
- User specific reporting facility for usage analysis.
- System and Application benchmarks

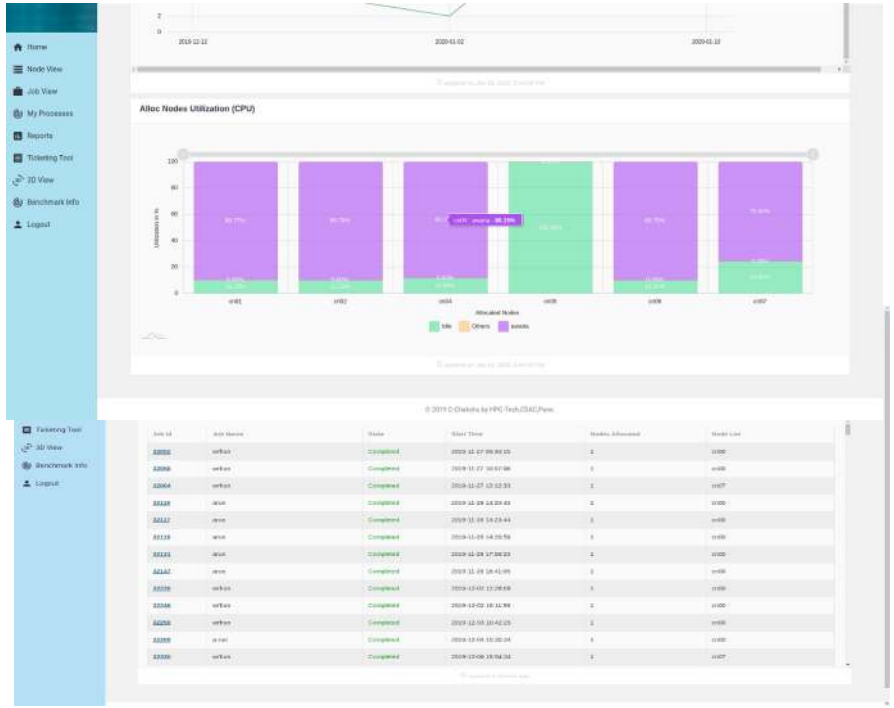


What is User Portal?

- Visualizing HPC system in 3D Rack View model with performance and health information.

How to access? / Demo





Thank You.