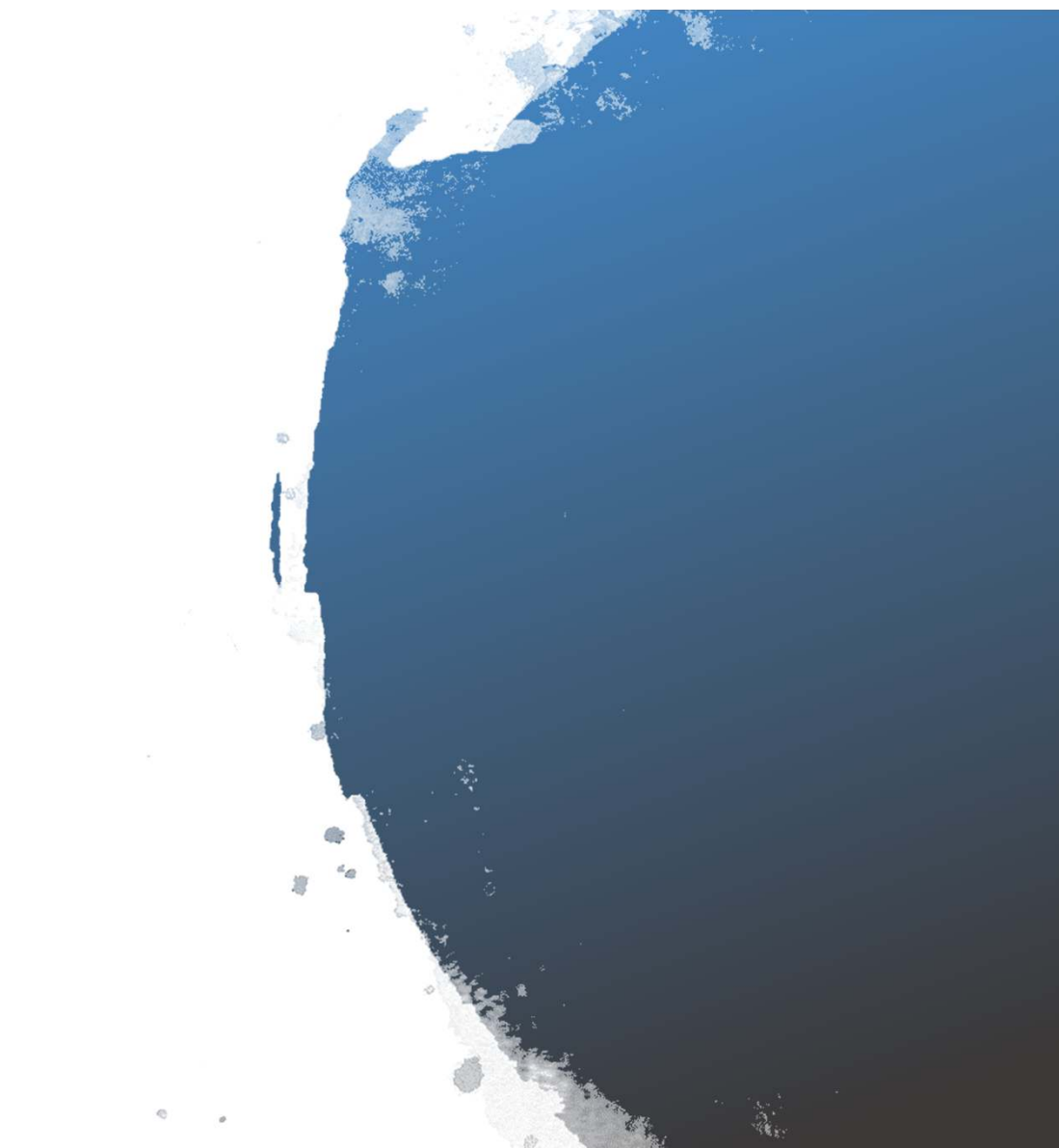


PARAM SHAKTI



# Agenda

Overview to High Performance Computing.

What is cluster and its Types

Components of HPC

PARAM-Shakti Architecture

Technical Specification of PARAM-Shakti

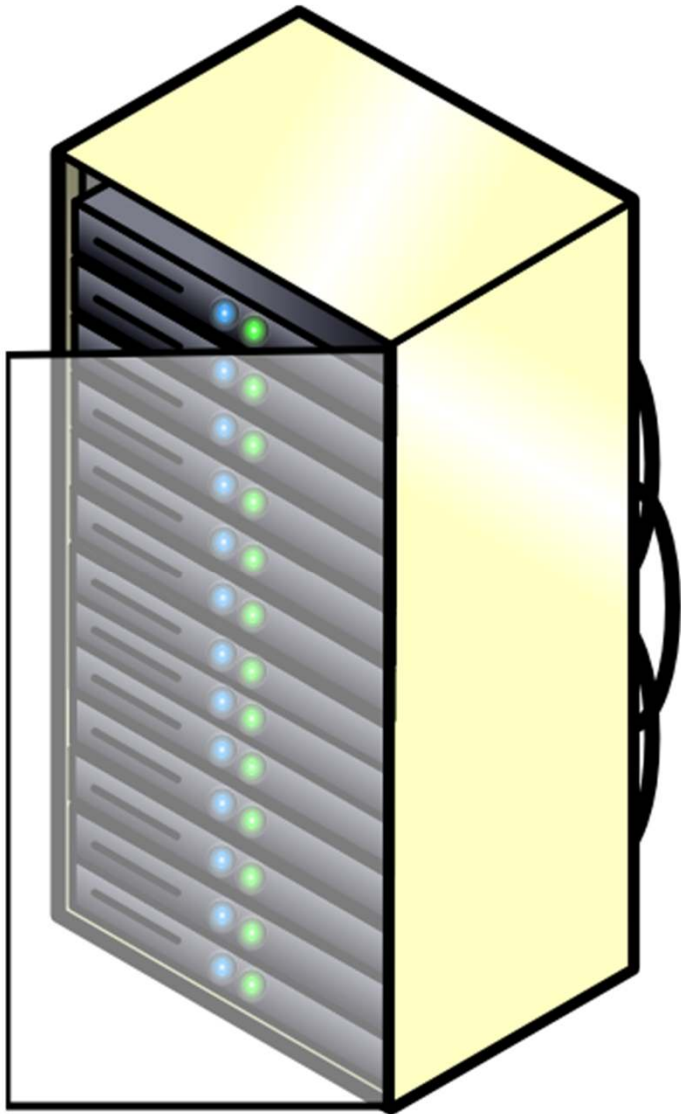
How to access Param-Shakti.

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

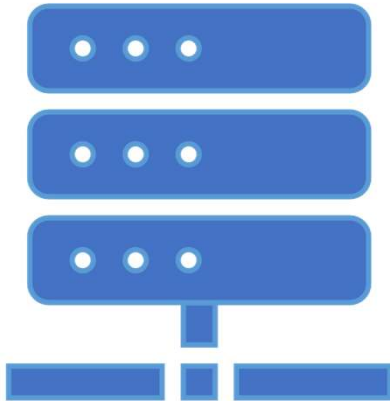
1. Nodes (servers)
2. Parallel File System (Storage & performance)
3. Interconnect (Networking & application execution)
4. Accelerator cards (boosting the performance with many )
5. Optimized compilers and libraries



---

# Components of HPC

---



## Nodes:

- Nodes are the actual 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 different parts of same file at the same time multiple clients can read the file.
- It supports High bandwidth Interconnects such as Mellanox, Omnipath, etc.
- It is a scalable file system.
- It supports HA and it is a POSIX compliant file system.
- ACL can be applied.





# Graphical processing Unit

---



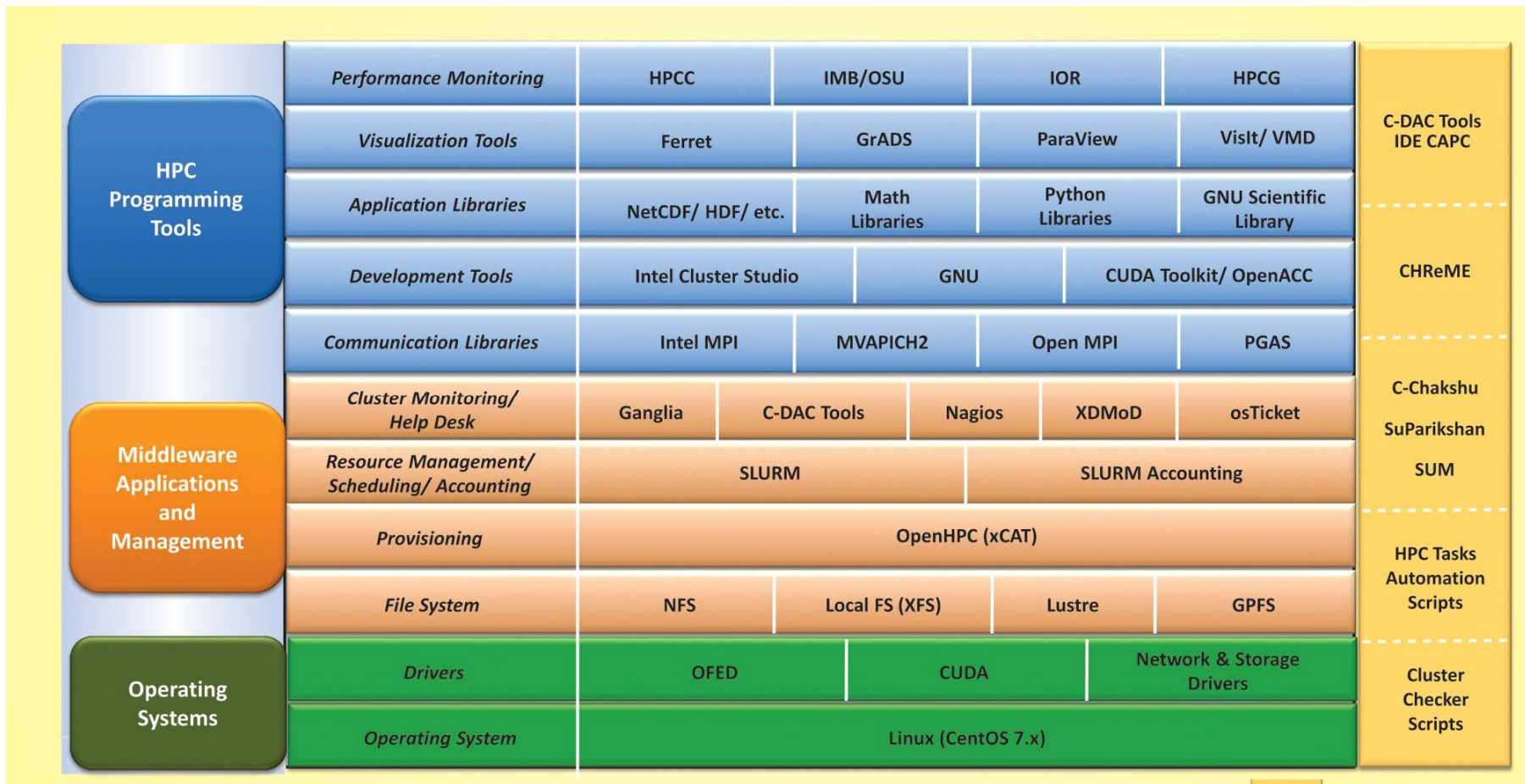
- A co-processor to accelerate general purpose scientific and engineering computing.
- It accelerates applications running on CPU by offloading compute intensive and time consuming portion of the code.
- GPU consists of thousands of smaller cores which together operate to crunch data in the application.

# 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



*One Vision. One Goal... Advanced Computing for Human Advancement...*

# 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 Scheduler (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

---



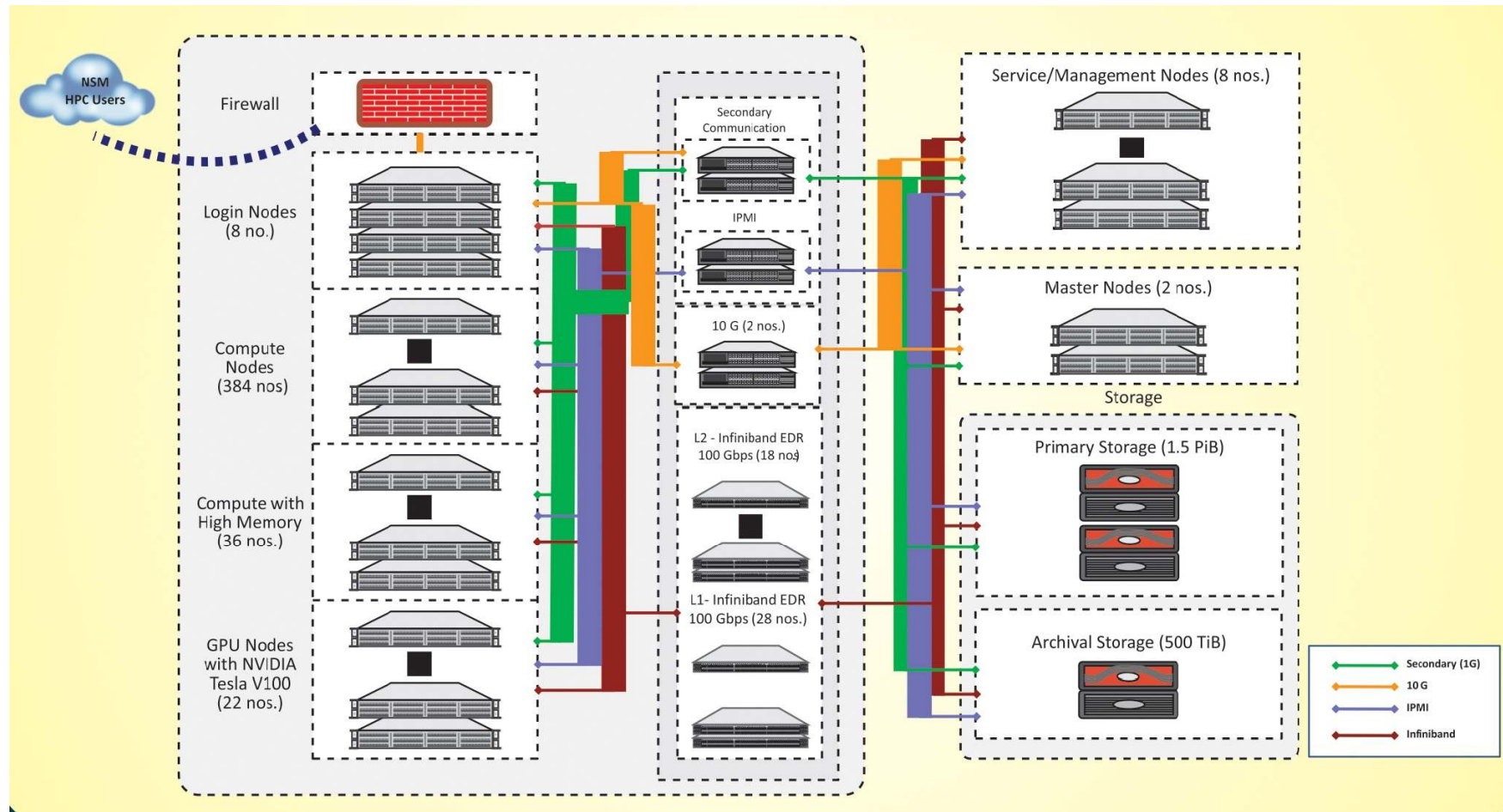
## Login Environment

- 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



*One Vision. One Goal... Advanced Computing for Human Advancement...*

## 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 3<sup>rd</sup> system among Indias top performing HPC

Rank	Site	System	Cores/Processor Sockets/Nodes	Rmax (TFlops)
1	<a href="#">Indian Institute of Tropical Meteorology(IITM), Pune</a>	<a href="#">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.</a> OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	119232/ /3312	3763.9
2	<a href="#">National Centre for Medium Range Weather Forecasting (NCMRWF), Noida</a>	<a href="#">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</a> OEM: Cray Inc., Bidder: Cray Supercomputers India Pvt. Ltd.	83592//2322	2570.4
3	<a href="#">Indian Institute of Technology (IITK), Kharagpur</a>	<a href="#">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)</a> 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

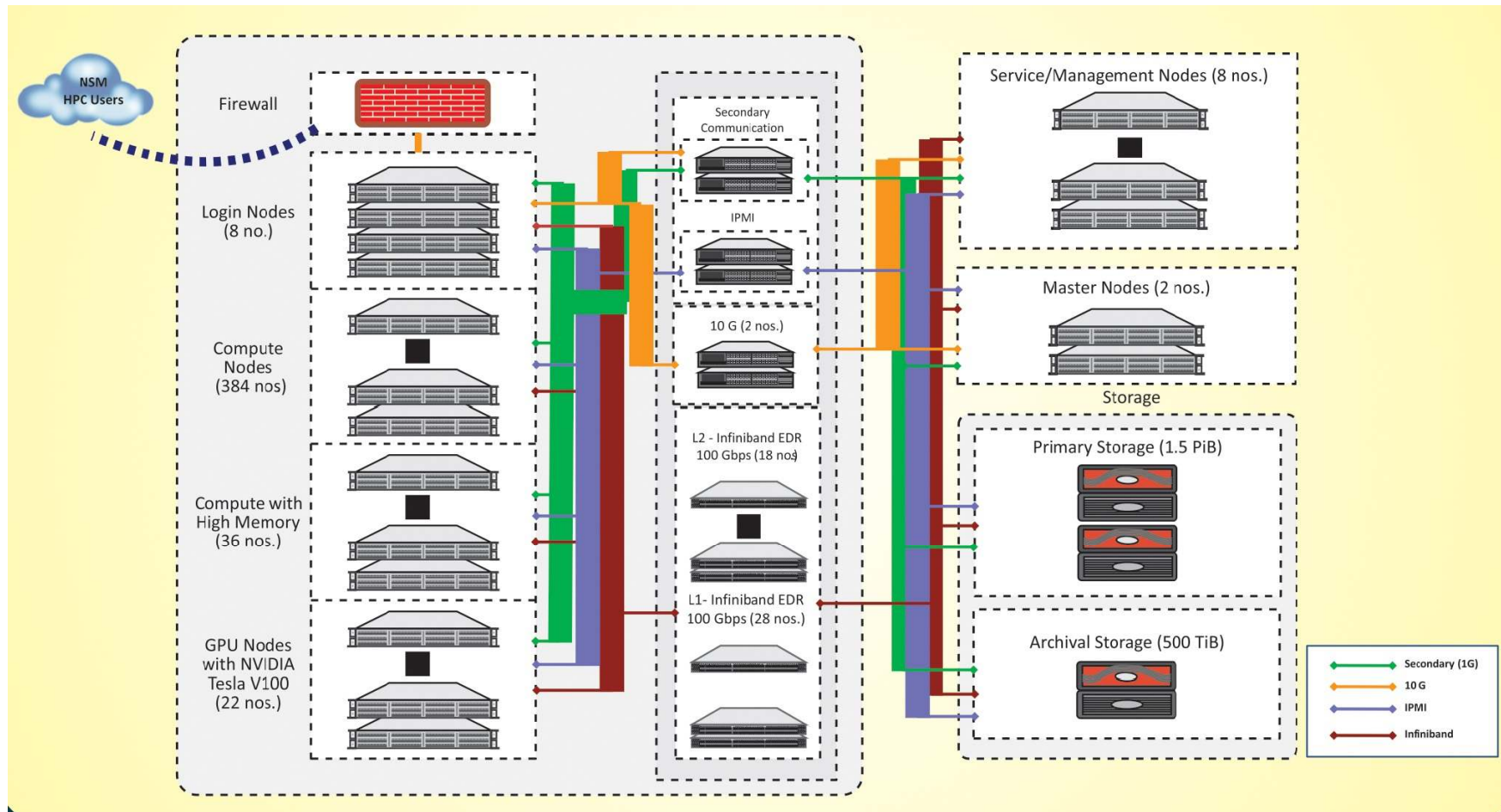
- 36 Nodes
- 1440 Cores
- Compute power of Rpeak 110.59 TFLOPS
- Each Node with
  - 2\* Intel Xeon SKL G-6148, 20 cores, 2.4 GHz, processors
  - 768 GB memory
  - 480 GB SSD

## 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



# PARAM SHAKTI ARCHITECTURE DIAGRAM

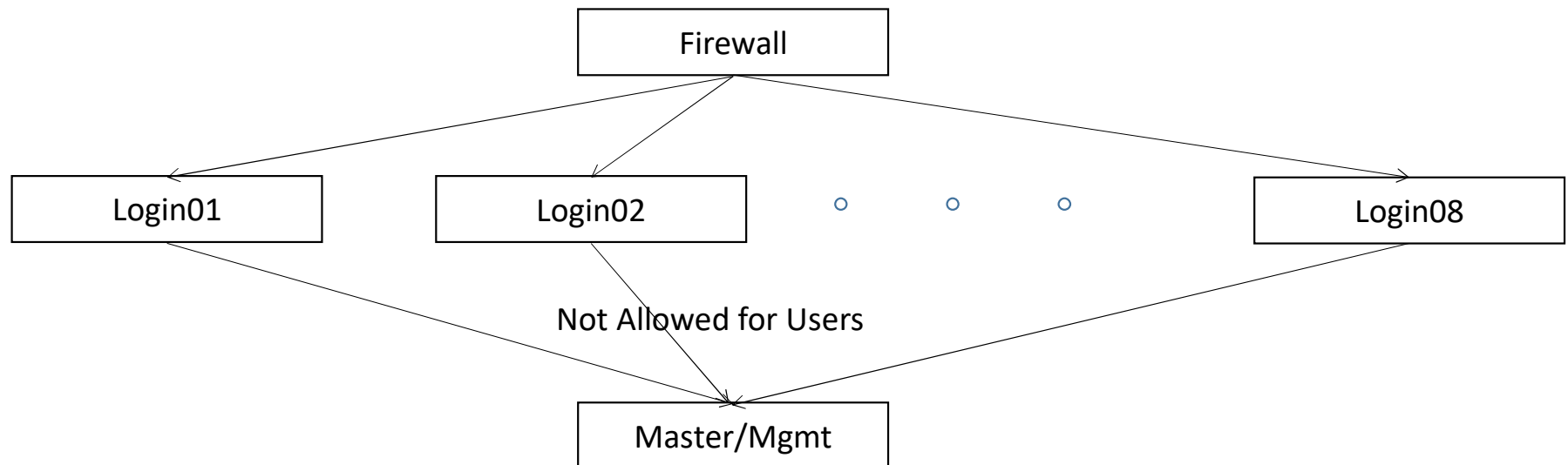


*One Vision. One Goal... Advanced Computing for Human Advancement...*

# 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>

The screenshot shows the Param Shreshtha Support Center website. At the top, the logo "PARAM SHRESHTA" is displayed. Below the logo is a navigation bar with four links: "Support Center Home", "Open a Ticket", "Check Ticket Status", and "Check FAQs". The main content area features a "Welcome to the Support Center" heading, followed by a paragraph explaining the support ticket system. To the right of the text are two buttons: "Open a New Ticket" (blue) and "Check Ticket Status" (green). Below the text, there is a note: "Please login through your LDAP account".





### Frequently Asked Questions

[General FAQ](#)

[Environment](#)

[Job Submission](#)

[Applications](#)

[ML / DL](#)

[Visualization](#)

[Best Practices](#)

[Hardware Specifications](#)

[Help](#)

[Create New Ticket](#)



## 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

[How to get account on HPC cluster ?](#)

[How do I Access the HPC Cluster ?](#)

[What if, Error : Disk quota exceeded ?](#)

[I get "Disk quota exceeded" error message when trying to remove files. What can I do?](#)

[What if, SCP not functional ?](#)

[What if, Error : Out of memory / segmentation fault ?](#)

[What if, Error :ERROR : Bad Interpreter ?](#)

[Can I run MS Windows applications on HPC?](#)

[How to Access Internet on HPC Cluster ?](#)

[How much of the file/space quota I have used ?](#)

[My account expired ! What should I do? Is my data gone forever?](#)



## General FAQ



Search



Search\_Your\_Queries

### Frequently Asked Questions

[General FAQ](#)

[Environment](#)

[Job Submission](#)

[Applications](#)

[ML / DL](#)

[Visualization](#)

[Best Practices](#)

[Hardware Specifications](#)

[Help](#)

[Create New Ticket](#)

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`

### Table of contents

[How to get account on HPC cluster ?](#)

[How do I Access the HPC Cluster ?](#)

[What if, Error : Disk quota exceeded ?](#)

[I get "Disk quota exceeded" error message when trying to remove files. What can I do?](#)

[What if, SCP not functional ?](#)

[What if, Error : Out of memory / segmentation fault ?](#)

[What if, Error :ERROR : Bad Interpreter ?](#)

[Can I run MS Windows applications on HPC?](#)

[How to Access Internet on HPC Cluster ?](#)

[How much of the file/space quota I have used ?](#)

[Mv account expired ! What](#)

[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

1. C-Chakshu Overview
2. Salient Features
3. Roles and their Purpose
4. What is User Portal ?
5. How to access? / Demo

---

## 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.



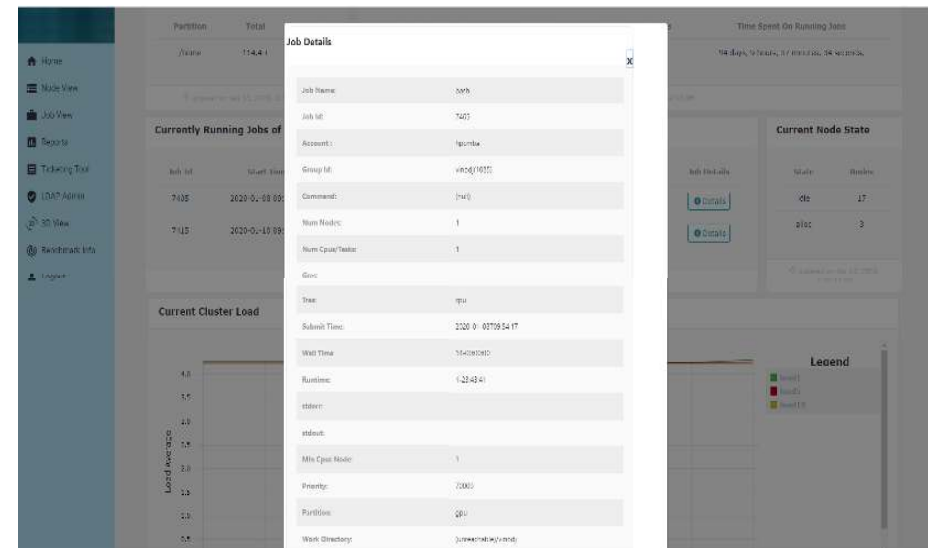
# What is User Portal?

- 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.



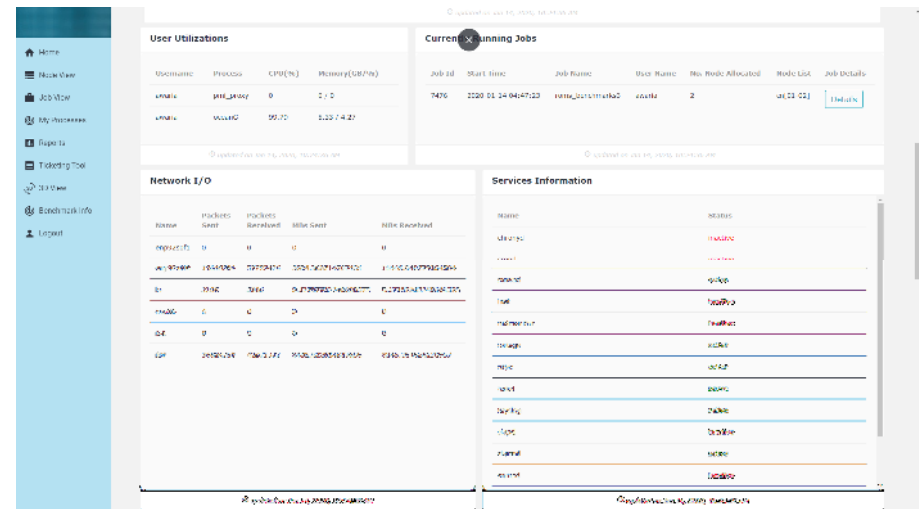
# What is User Portal?

- 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.



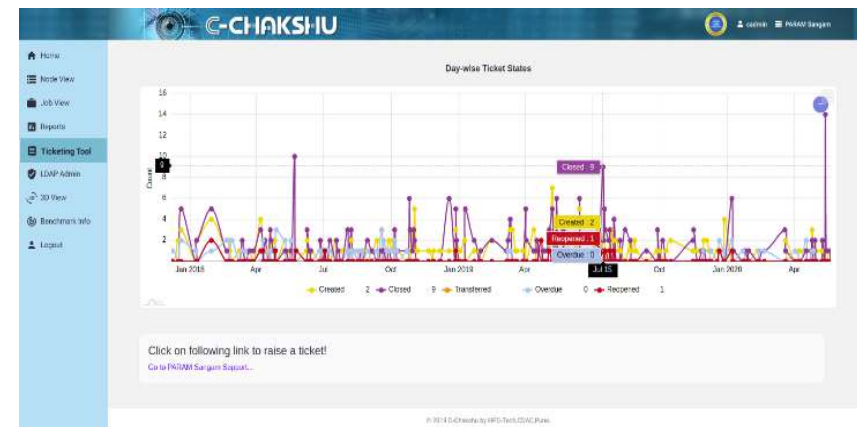
# What is User Portal?

- 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

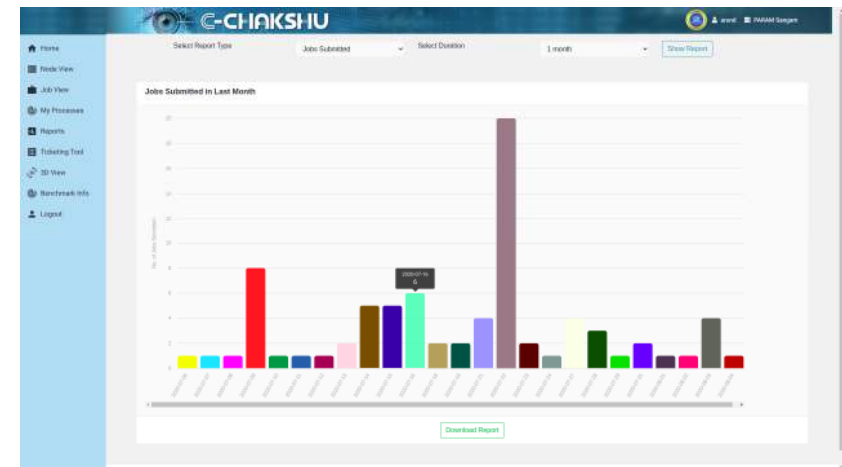


The screenshot displays the 'C-CHAKSHU' user portal interface. The main content area is titled 'Allocated Nodes With Running Processes'. It shows a table for node 'cn04' with columns for PID, User, Process, Status, Max Throughput, CPU(%) usage, and Memory(MB) usage. Below this, a table for node 'cn05' is partially visible.

Allocated Nodes With Running Processes						
cn04						
PID	User	Process	Status	Max Throughput	CPU(%)	Memory(MB)
173333	anvita	lsyncd	running	2	99.9	1.27
173332	anvita	rsyncd	running	2	99.9	1.13
173285	anvita	lsyncd	running	2	100.0	0.22
173313	anvita	rsyncd	running	2	99.9	1.11
173249	anvita	lsyncd	running	2	100.0	0.11
173258	anvita	rsyncd	running	2	99.8	0.33
173254	anvita	rsyncd	running	2	99.8	0.13
173289	anvita	rsyncd	running	2	99.8	0.21
173275	anvita	rsyncd	running	2	99.8	0.12
173272	anvita	rsyncd	running	2	99.9	0.11
173263	anvita	rsyncd	running	2	99.7	0.11
173264	anvita	rsyncd	running	2	99.7	0.11
cn05						
PID	User	Process	Status	Max Throughput	CPU(%)	Memory(MB)
188138	anvita	rsyncd	running	2	100	0.51

## What is User Portal?

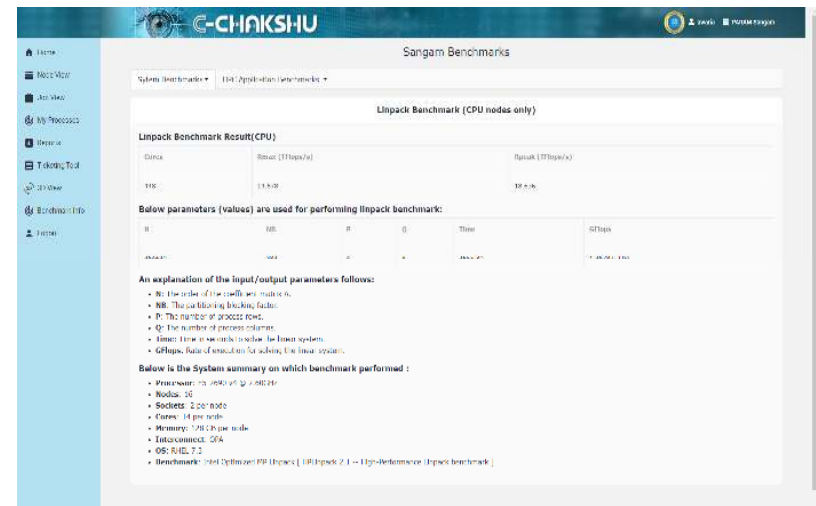
- Ticketing system for raising and following up on issue related to HPC system/application.
- Monitor user specific processes running on different nodes in single window.
- User specific reporting facility for usage analysis.
- System and Application benchmarks





# What is User Portal?

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



The screenshot displays the 'Sangam Benchmarks' user portal. The main content area shows 'Linpack Benchmark (CPU nodes only)' results. The results table indicates a score of 11.808 MFlops/s. Below the table, there is a section for 'Below parameters (values) are used for performing lmpack benchmark:' with fields for N, NB, P, Q, Rtime, and Gflops. An explanatory text block follows, detailing the meaning of these parameters and providing a system summary for the benchmark performed.

Linpack Benchmark Result(CPU)					
Score	Units (MFlops/s)	Units (MFlops/s)			
11.808		11.808			

**Below parameters (values) are used for performing lmpack benchmark:**

N	NB	P	Q	Rtime	Gflops
1024	1024	1	1	0.000000	11.808

**An explanation of the input/output parameters follows:**

- N: The order of the square matrix.
- NB: The job blocking factor.
- P: The number of process rows.
- Q: The number of process columns.
- Rtime: Time in seconds to solve the linear system.
- Gflops: Rate of execution for solving the linear system.

**Below is the System summary on which benchmark performed :**

- Processor: 10 x AMD EPYC 7513
- Nodes: 10
- Sockets: 1 per node
- Cores: 16 per node
- Memory: 128 GB per node
- Interconnect: GbE
- OS: RHEL 7.7
- Benchmark: Intel Optimized MP (lmpack [ 10] lmpack 2 ] -- High-Performance Linpack benchmark)

# What is User Portal?

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



# How to access? / Demo

**C-CHAKSHU** Dashboard Overview:

- Live Nodes Status:** 0 Jobs
- Currently Running Jobs:** 0 Jobs
- Jobs Submitted Till Date:** 1,122 Jobs
- Last Week CPU Usage:** 42.810 %

**System Information:**

Kernel Version	Architecture	Resource Manager	CPU Cores	CPU Slices	GPU Cores	GPU Slices
3.10.0-851.12.2.el7.x86_64	x86_64	kubernetes	28	28	8	384

**Current Job Statistics Till Date:**

Total Job Submitted	Submitted Jobs	Total Allocated Nodes	Time Spent on Running Jobs
1222	361	508	37 days, 1 hour, 57 minutes, 30 seconds

**Allocated Nodes:**

Node ID	Resource State	Temperature	CPU Usage
cn01	Alloc	48.0°C	100.0%
cn02	Alloc	48.0°C	100.0%
cn03	Alloc	48.0°C	100.0%
cn04	Alloc	48.0°C	100.0%
cn05	Alloc	48.0°C	100.0%
cn06	Alloc	52.0°C	92.0%
cn07	Alloc	48.0°C	100.0%
cn08	Alloc	48.0°C	100.0%
cn09	Alloc	48.0°C	100.0%
cn10	Alloc	48.0°C	100.0%
cn11	Alloc	48.0°C	100.0%
cn12	Alloc	48.0°C	100.0%
cn13	Alloc	48.0°C	100.0%
cn14	Alloc	48.0°C	100.0%
cn15	Alloc	48.0°C	100.0%

**Allot Nodes Utilization (CPU)**

Bar chart showing CPU utilization for nodes cn01 through cn07. The y-axis represents Utilization in % (0 to 100). The x-axis represents nodes. A tooltip for node cn03 shows: **cn03: 48.0%**. The legend indicates: **Allocated Nodes** (purple), **Idle** (green), and **Others** (red).

**Job History Table:**

Job ID	Job Name	State	Start Time	Health	Submitted	Node List
AD001	arkon	Completed	2019-11-27 09:00:00	1	100%	
AD002	arkon	Completed	2019-11-27 09:01:00	1	100%	
AD004	arkon	Completed	2019-11-27 13:23:31	1	100%	
AD009	arkon	Completed	2019-11-28 18:20:03	1	100%	
AD012	arkon	Completed	2019-11-28 18:23:44	1	100%	
AD128	arkon	Completed	2019-11-28 14:25:50	1	100%	
AD021	arkon	Completed	2019-11-28 17:38:23	1	100%	
AD042	arkon	Completed	2019-11-28 18:41:09	1	100%	
AD058	arkon	Completed	2019-11-27 17:28:18	1	100%	
AD066	arkon	Completed	2019-11-28 18:21:34	1	100%	
AD026	arkon	Completed	2019-11-28 18:42:25	1	100%	
AD008	arkon	Completed	2019-11-28 18:36:04	1	100%	
AD006	arkon	Completed	2019-11-28 18:34:34	1	100%	



  
**Thank You.**