HPC: PARAM Utkarsh Supercomputing Facility
@ C-DAC Bangalore

Enhance your business to next level

Knowledge Park

Electronic City

CPSF,Team
CDAC-Bangalore

परम उत्कर्ष
PARAM Utkarsh

C-DAC PARAM Supercomputing Facility (CPSF) Data Centre

Agenda

- Data Centre
- NSM HPC Architecture
- Software Stack
- Applications
- Scheduler
- Conclusion

C-DAC PARAM Supercomputing Facility (CPSF) Data Centre

DC Area 1800 SqFt (Total Area 7625 SqFt)

Three High density DLC racks cooled with Adiabatic dry-cooler

Two service node racks

Server Racks and Storage racks

Staging area, User terminal Area and Conference room

BMS – Building Management System

- Integrated systems in BMS
- Generator (1+1 redundant)
- UPS (n+1 redundant)
- Precision Air Conditioning – PAC (n+1 redundant)
- Fire Alarm System - FAS (VESDA and NOVEC)

  Very Early Smoke Detection Apparatus – VESDA

  NOVEC- Fire suppresser
- Dry cooler (Adiabatic)
- Temperature and humidity sensors

PARAM Utkarsh Security

Perimeter Firewall (UTM)

Cluster Firewall (HA load balancer and IPS)

Geographical filtering

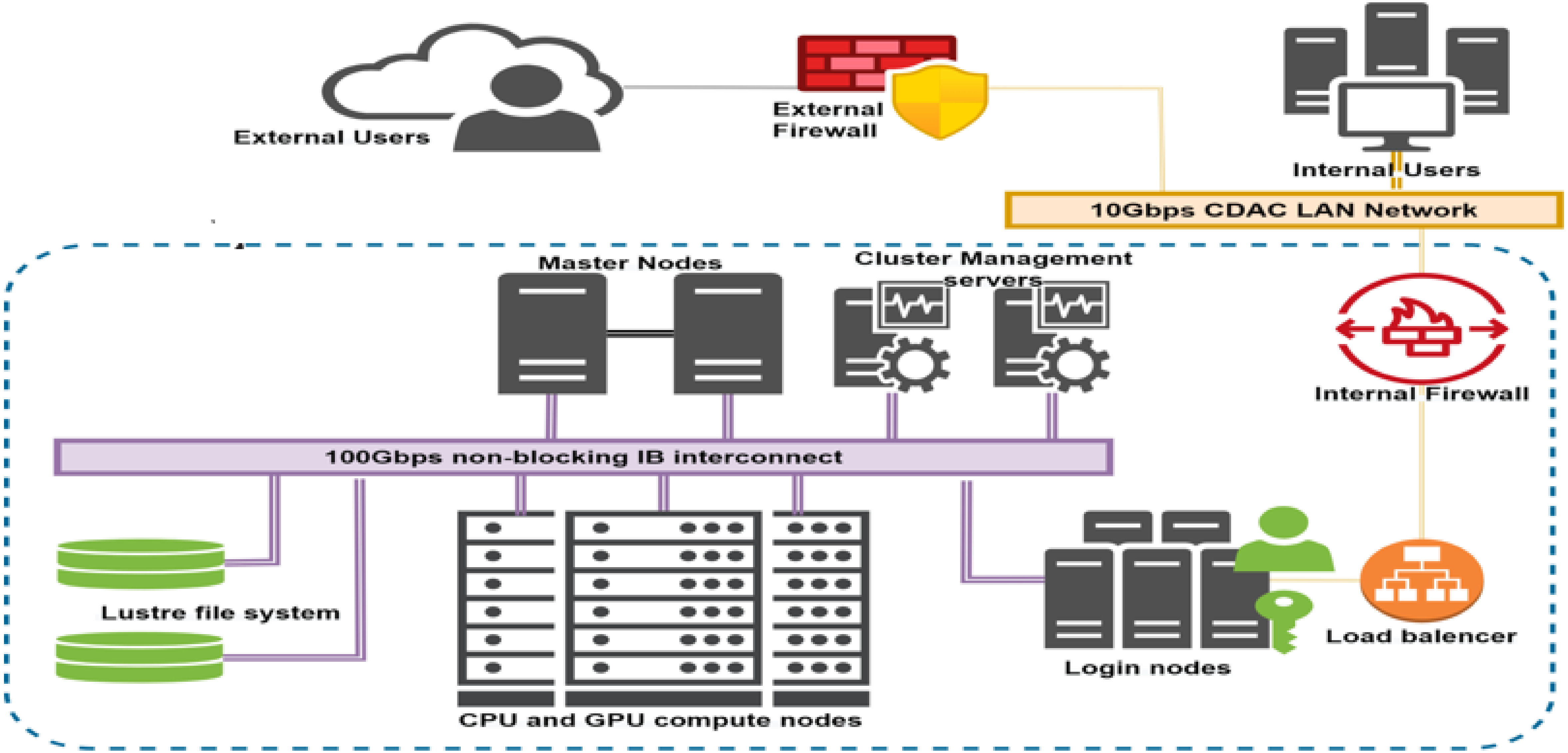
Isolated Network

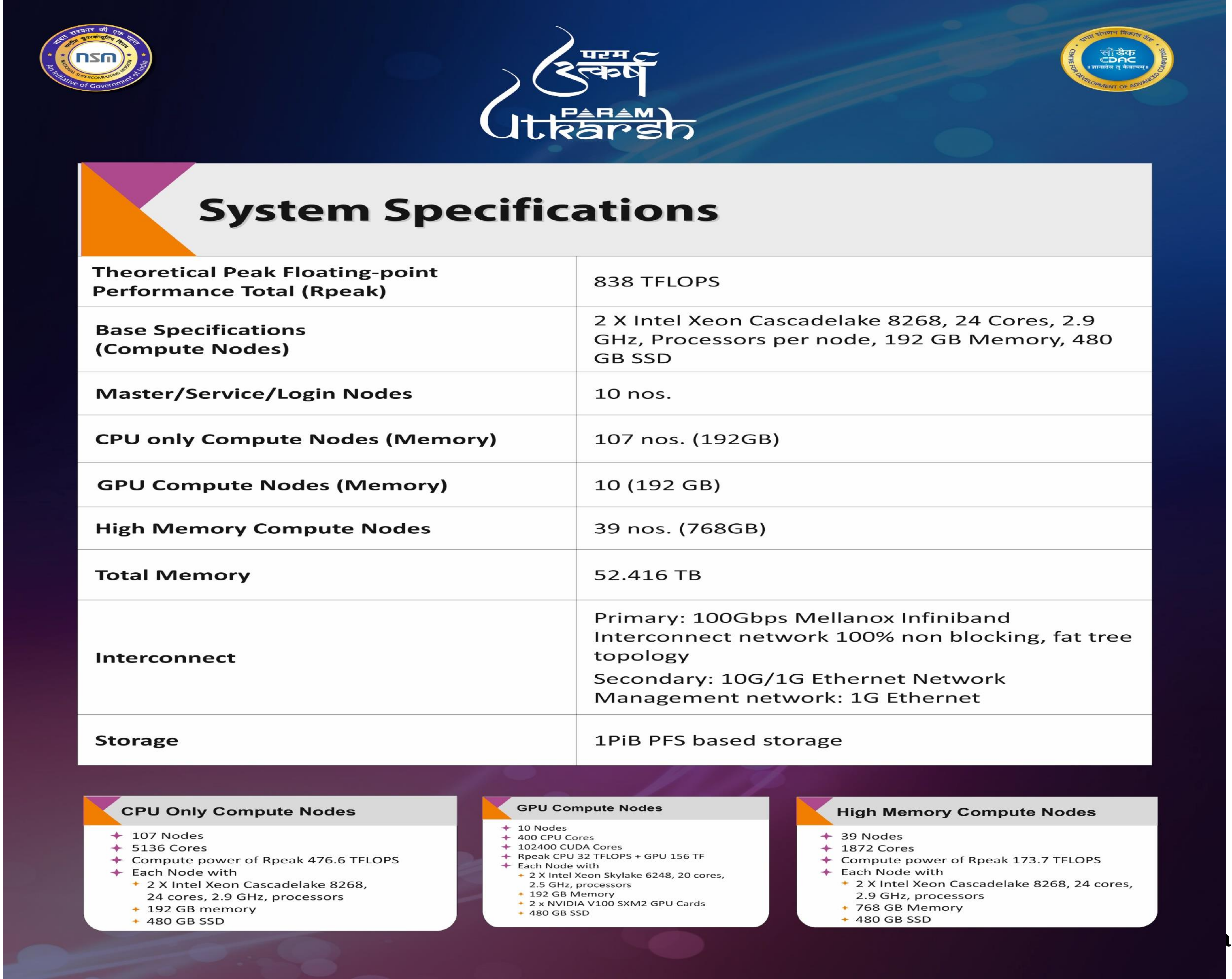
Secured authentication

Physical security with ACS(access control system)

CCTV surveillance(24/7)

NSM HPC Architecture





परम उत्कर्ष
PARAM Utkarsh

System Specifications

Theoretical Peak Floating-point Performance Total (Rpeak)	838 TFLOPS
Base Specifications (Compute Nodes)	2 X Intel Xeon Cascadelake 8268, 24 Cores, 2.9 GHz, Processors per node, 192 GB Memory, 480 GB SSD
Master/Service/Login Nodes	10 nos.
CPU only Compute Nodes (Memory)	107 nos. (192GB)
GPU Compute Nodes (Memory)	10 (192 GB)
High Memory Compute Nodes	39 nos. (768GB)
Total Memory	52.416 TB
Interconnect	Primary: 100Gbps Mellanox Infiniband Interconnect network 100% non blocking, fat tree topology Secondary: 10G/1G Ethernet Network Management network: 1G Ethernet
Storage	1PiB PFS based storage

CPU Only Compute Nodes

- 107 Nodes
- 5136 Cores
- Compute power of Rpeak 476.6 TFLOPS
- Each Node with
  - 2 X Intel Xeon Cascadelake 8268, 24 cores, 2.9 GHz, processors
  - 192 GB memory
  - 480 GB SSD

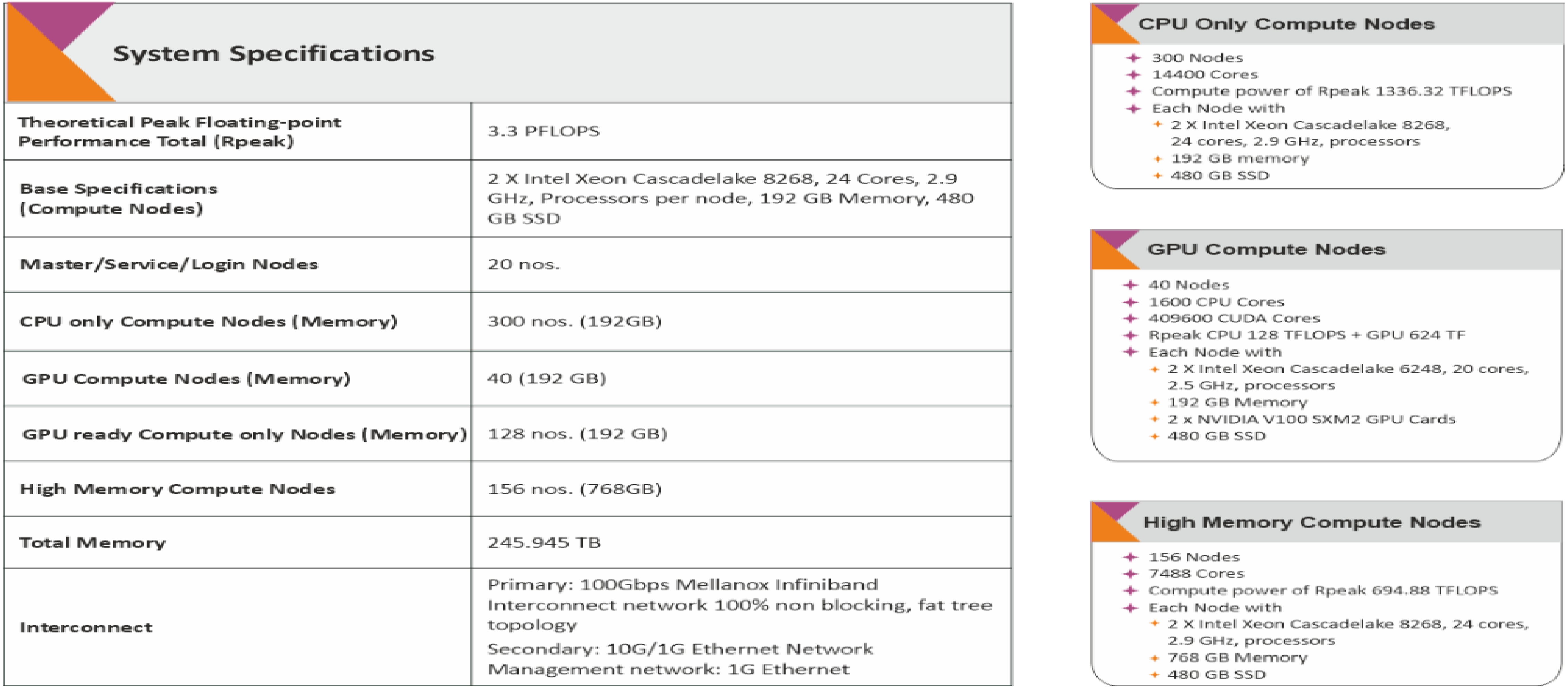
GPU Compute Nodes

- 10 Nodes
- 400 CPU Cores
- 102400 CUDA Cores
- Rpeak CPU 32 TFLOPS + GPU 156 TF
- Each Node with
  - 2 X Intel Xeon Skylake 6248, 20 cores, 2.5 GHz, processors
  - 192 GB Memory
  - 2 x NVIDIA V100 SXM2 GPU Cards
  - 480 GB SSD

High Memory Compute Nodes

- 39 Nodes
- 1872 Cores
- Compute power of Rpeak 173.7 TFLOPS
- Each Node with
  - 2 X Intel Xeon Cascadelake 8268, 24 cores, 2.9 GHz, processors
  - 768 GB Memory
  - 480 GB SSD

PARAM Pravega - System Configuration



System Specifications

Theoretical Peak Floating-point Performance Total (Rpeak)	3.3 PFLOPS
Base Specifications (Compute Nodes)	2 X Intel Xeon Cascadelake 8268, 24 Cores, 2.9 GHz, Processors per node, 192 GB Memory, 480 GB SSD
Master/Service/Login Nodes	20 nos.
CPU only Compute Nodes (Memory)	300 nos. (192GB)
GPU Compute Nodes (Memory)	40 (192 GB)
GPU ready Compute only Nodes (Memory)	128 nos. (192 GB)
High Memory Compute Nodes	156 nos. (768GB)
Total Memory	245.945 TB
Interconnect	Primary: 100Gbps Mellanox Infiniband Interconnect network 100% non blocking, fat tree topology Secondary: 10G/1G Ethernet Network Management network: 1G Ethernet

CPU Only Compute Nodes

- 300 Nodes
- 14400 Cores
- Compute power of Rpeak 1336.32 TFLOPS
- Each Node with
  - 2 X Intel Xeon Cascadelake 8268, 24 cores, 2.9 GHz, processors
  - 192 GB memory
  - 480 GB SSD

GPU Compute Nodes

- 40 Nodes
- 1600 CPU Cores
- 409600 CUDA Cores
- Rpeak CPU 128 TFLOPS + GPU 624 TF
- Each Node with
  - 2 X Intel Xeon Cascadelake 6248, 20 cores, 2.5 GHz, processors
  - 192 GB Memory
  - 2 x NVIDIA V100 SXM2 GPU Cards
  - 480 GB SSD

High Memory Compute Nodes

- 156 Nodes
- 7488 Cores
- Compute power of Rpeak 694.88 TFLOPS
- Each Node with
  - 2 X Intel Xeon Cascadelake 8268, 24 cores, 2.9 GHz, processors
  - 768 GB Memory
  - 480 GB SSD

PARAM Utkarsh - System Details

SN	Server	Number
01	Master Node	02
02	Login Nodes	04
03	Management Nodes	03
04	Firewall	01
05	CPU only nodes	75
06	GPU Nodes	10
07	GPU Ready Nodes	32
08	High Memory Nodes	39
	Total Nodes	166

PARAM Pravega- System Details

SN	Server	Number
01	Master Node	02
02	Login Nodes	11
03	Management Nodes	04
04	Firewall	02
05	CPU only nodes	300
06	GPU Nodes	40
07	GPU Ready Node	128
08	High Memory Nodes	156
	Total Nodes	644

PARAM Utkarsh - System Details

Parameter	CPU only(75)	GPU Nodes(10)	GPU Ready(32)	HM Nodes(39)
Processor	2 x Xeon platinum 8268	2 x Xeon G-6248	2 x Xeon platinum 8268	2 x Xeon platinum 8268
Cores	48	40	48	48
Speed	2.9 GHz	2.5 GHz	2.9 GHz	2.9 GHz
Memory	192 GB	192 GB	192 GB	**768 GB**
HDD	480GB SSD	480GB SSD	480GB SSD	480GB SSD
Total cores	**3600**	**400**	**1536**	**1872**
Total Memory	14400 GB	1920 GB	6144 GB	29952 GB
	-	2 x NVIDIA V100	-	-

PARAM Pravega- System Details

Parameter	CPU only(300)	GPU Nodes(40)	GPU Ready Nodes (128)	HM Nodes(156)
Processor	2 x Xeon platinum 8268	2 x Xeon G-6248	2 x Xeon platinum 8268	2 x Xeon platinum 8268
Cores	48	40	48	48
Speed	2.9 GHz	2.5 GHz	2.9 GHz	2.9 GHz
Memory	192 GB	192 GB	192 GB	**768 GB**
HDD	480GB SSD	480GB SSD	480GB SSD	480GB SSD
Total cores	**14400**	**1600**	**6144**	**7488**
Total Memory	57600 GB	7680 GB	24576 GB	119808 GB
	-	2 x NVIDIA V100		-

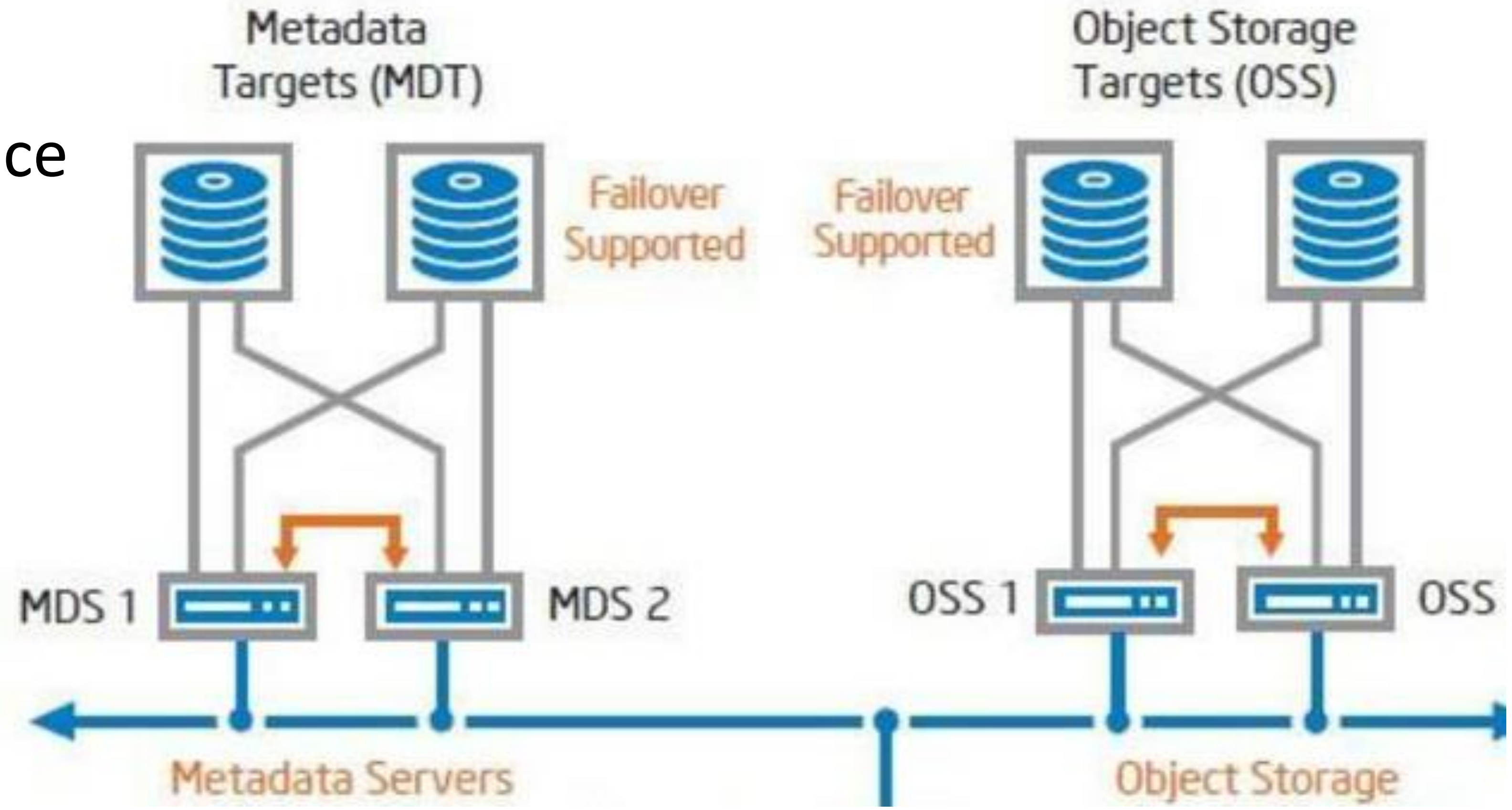
PARAM Utkarsh - Storage Details

Storage Size: 1 PiB Usable capacity

2 Embedded Lustre Parallel File System storage appliance with redundant controller

25 GB/s sustained read and write performance

8 x 100Gbps InfiniBand interconnect.



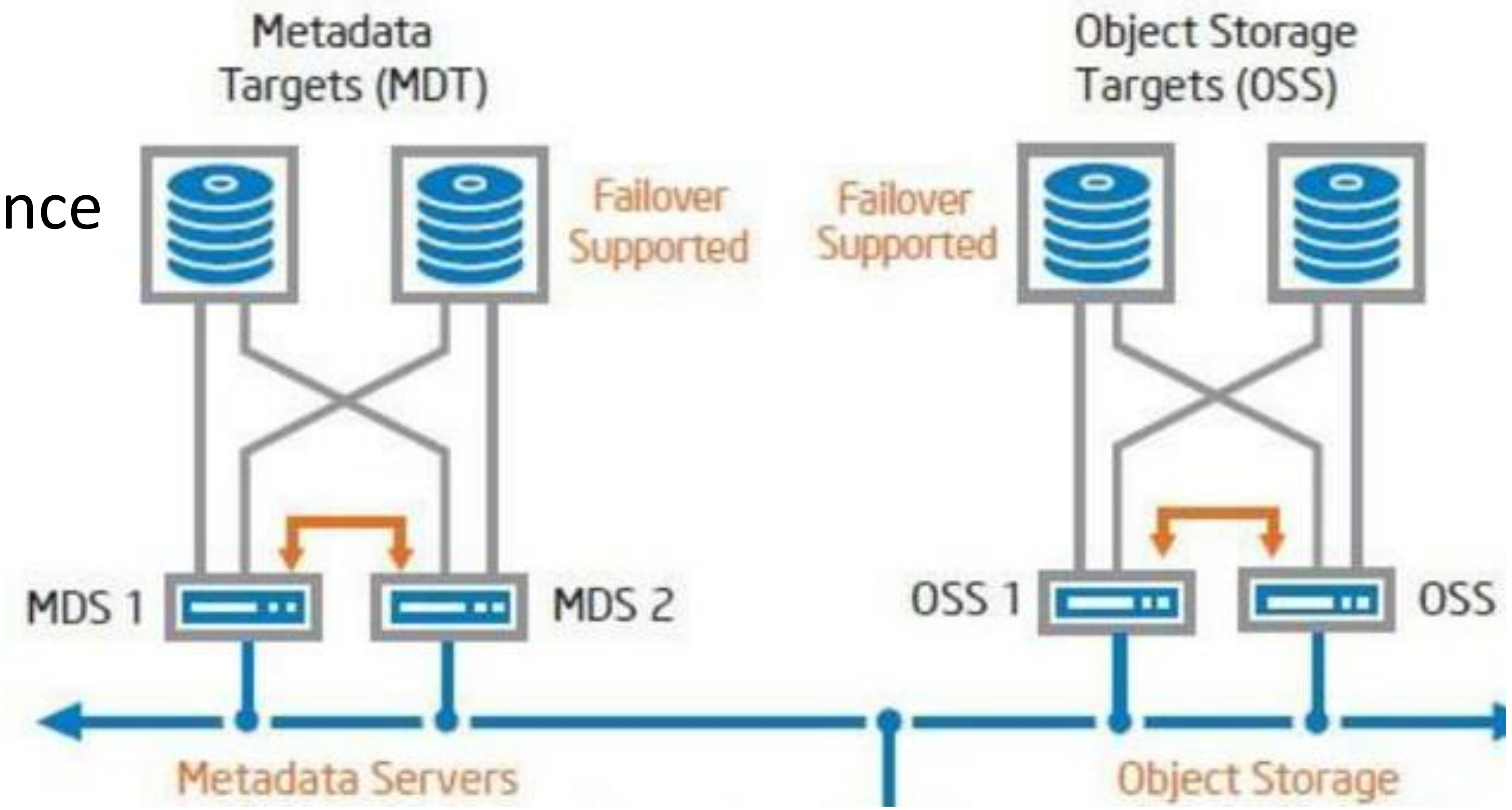
PARAM Pravega - Storage Details

Storage Size: 4 PiB Usable capacity

2 Embedded Lustre Parallel File System storage appliance with redundant controller

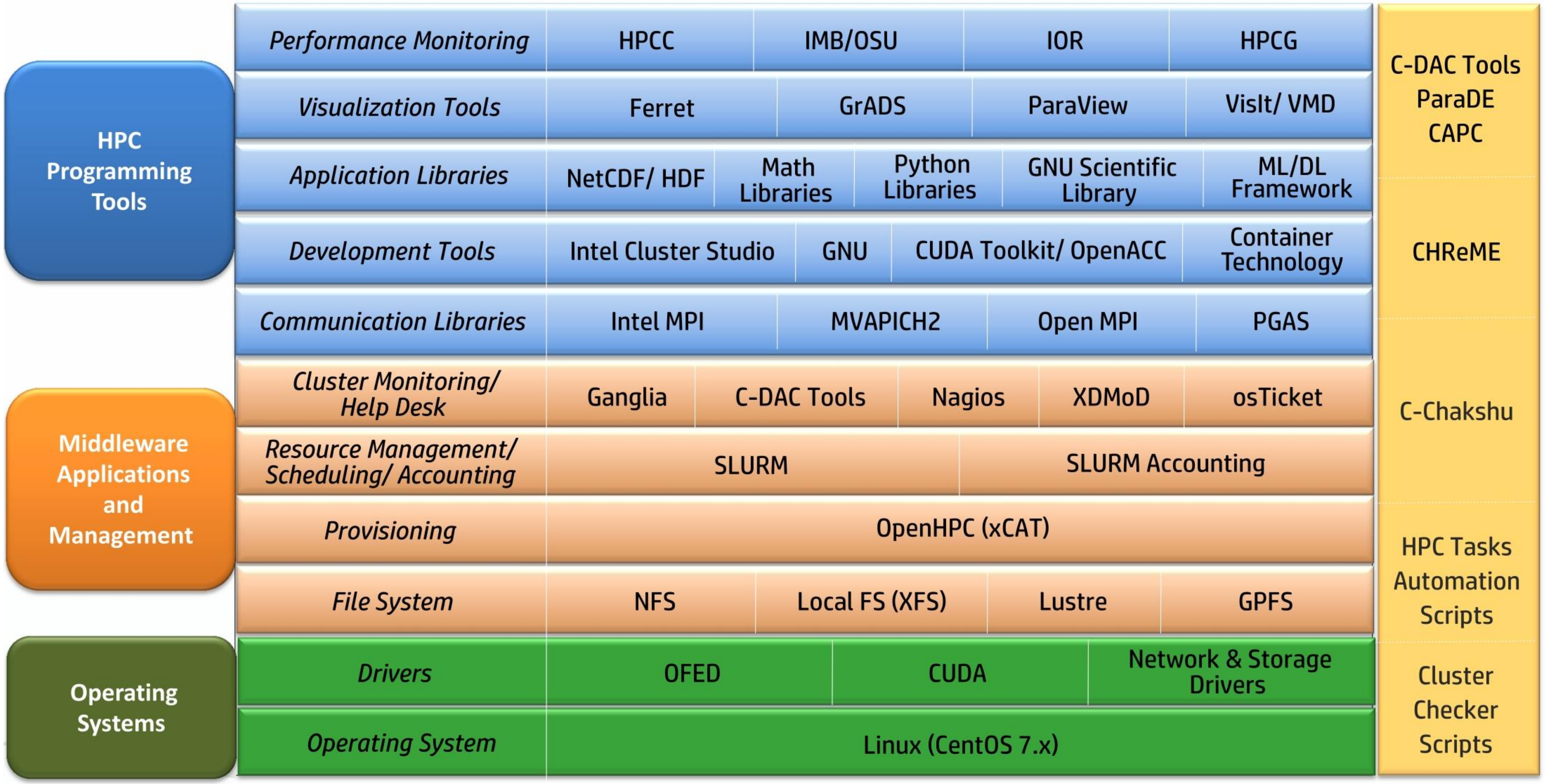
100 GB/s sustained read and write performance

8 x 100Gbps InfiniBand interconnect.



परम उत्कर्ष
PARAM Utkarsh

Software Stack



HPC Programming Tools	*Performance Monitoring*	HPCC	IMB/OSU	IOR	HPCG		
	Visualization Tools	Ferret	GrADS	ParaView	VisIt/ VMD		C-DAC Tools ParaDE CAPC
	Application Libraries	NetCDF/ HDF	Math Libraries	Python Libraries	GNU Scientific Library	ML/DL Framework	
	Development Tools	Intel Cluster Studio	GNU	CUDA Toolkit/ OpenACC	Container Technology		CHReME
	Communication Libraries	Intel MPI	MVAPICH2	Open MPI	PGAS		
Middleware Applications and Management	*Cluster Monitoring/ Help Desk*	Ganglia	C-DAC Tools	Nagios	XDMoD	osTicket	C-Chakshu
	Resource Management/ Scheduling/ Accounting	SLURM	SLURM Accounting				
	Provisioning	OpenHPC (xCAT)					
	File System	NFS	Local FS (XFS)	Lustre	GPFS		HPC Tasks Automation Scripts
Operating Systems	*Drivers*	OFED	CUDA	Network & Storage Drivers			Cluster Checker Scripts
	Operating System	Linux (CentOS 7.x)					

🟨 **C-DAC Components**

NSM Clusters –Applications, Tools, Programming Models- AI & HPC

HPC Applications		
	Bio-informatics	MUMmer, HMMER, MEME, PHYLIP, mpiBLAST, ClustalW
	Molecular Dynamics	NAMD (CPU & GPU), LAMMPS(CPU & GPU), GROMACS
	Material Modeling, Quantum Chemistry	Quantum-Espresso, Abinit, CP2K, NWChem,
	CFD	OpenFOAM, FDS, SU2
	Weather, Ocean, Climate	WRF, RegCM, MOM, ROMS
	Disaster Management	ANUGA Hydro

Visualization Programs	GrADS, ParaView, VisIt, VMD
Dependency Libraries	NetCDF, PNETCDF, Jasper, HDF5, Tcl, Boost, FFTW
Programming Models	MPI, OpenMP, OpenACC, CUDA, PGAS, Pthreads

Installed additional applications, libraries, tools as per requirements from users

AI/ ML/ DL Tools/ Technologies	**DL Frame work:** TensorFlow , keras, theano, pytorch, scikit-learn,scipy, cuDNN
	Data Science: Numpy , RAPIDS
	Distributed DL Framework: TensorFlow with Horovod
	Container Technology: enroot
	JupyterHub: DL application development platforms and web based IDE

Job Scheduler - SLURM

- When you login to HPC cluster, you land on **Login Nodes**
    - Login nodes are not meant to run jobs
    - These are used to submit jobs to **Compute Nodes.** You can setup your files and data on Login node. If compilation/installation of an application is required, user must do it on login node.

- To submit job on the cluster, you need to write a scheduler job script

- SLURM - Simple Linux Utility for Resource Management

- It is a workload manager that provides a framework for job queues, allocation of compute nodes, and the start and execution of jobs.

SLURM

How to set Environment ?

- ➢ By default no application is set in your environment. User must explicitly set required ones

- ➢ **module** is the utility (also command name) to enable use of applications / libraries / compilers available on the HPC cluster.
- ➢ Module structure on Cluster
  - apps/<application name>/version : Applications available on the cluste
  - compiler/<compiler name>/version : Compilers available on the cluster
  - lib/<library name>/version : Available libraries
  - ….

SLURM

- Some Important commands :
  - **module avail** To see the available software installed on HPC system
    - list of precompiled applications
    - different compilers and libraries (compilers include GNU, Intel, PGI)
  - **module list** Shows the currently loaded modules in your shell
  - **module load <Name of the module>**
    - module load compiler/intel/2018.2.199 (to set Intel compilers version 2018 in your environment)
    - module load apps/namd/2.12/impi2018/cpu (to set NAMD app version 2.12 in your environment)

SLURM

- Some Important commands :
  - **module unload <Name of the module>**
  - **module purge** To clear all the loaded modules

Note : If you want the corresponding environment to be loaded into your shell by default, then you can set the environment via .bashrc file.

Caution : Try to avoid loading of too many modules or setup of unwanted variables via .bashrc.

SLURM

- **sbatch <script>** To submit the job on HPC cluster

- **squeue** To see the status of all jobs submitted on the cluster

  squeue -u <user name> To see status of user’s jobs only. Also shows job-id.

SLURM: Useful Commands

- **sinfo** Provides the basic information about the resources on HPC cluster such as
    - Partitions/queue such as for cpu / gpu / high memory nodes
    - Number of nodes for each type and their numbering/names
    - State of the nodes

scancel <job ID> To delete the submitted jobs

- **scontrol hold <Job ID>**
- **scontrol release <Job ID>**
- **scontrol show job <Job ID>**
- **srun** To get resources in interactive mode

    -
      ```
      srun --nodes=1 --ntasks-per- node=1 --time=00:05:00 --pty bash -i
      ```

Thank you

utkarsh-support@cdac.in