Minerva Town Hall
April 2019

Patricia Kovatch
Bhupender Thakur, PhD
Francesca Tartaglione, MS
Dansha Jiang, PhD
Eugene Fluder, PhD
Hyung Min Cho, PhD
Lili Gai, PhD

April 1, 2019

Mount
Sinai

Outline

Welcome and general comments

Chimera infrastructure and services

- Chimera architecture
- Storage architecture
- User environment
- LSF details

Future plans and roadmap

2019 Chimera partition installation plan

Important dates:

- **Nov 30, 2018**: Shutdown Demeter
- **Dec 17, 2018**: Retire **2,300 Mothra cores** and **K20 GPU nodes**
- **Feb 11, 2019**: Open Chimera compute + **GPU nodes** + **container support** to friendly users
- **Apr 01, 2019:** Chimera in production
- **Jun 01, 2019:** GPFS 5.x/HIPAA compliant file system in production
- **Jul 01, 2019:** Retire Manda, Mothra, BODE
- **Sep 01, 2019:** Chimera **HIPAA compliant cluster**

Chimera architecture

Compute nodes and infrastructure upgrade: Chimera partition

Specs of the new compute partition (Chimera):

- 12x 42U racks
- **4x login nodes** - Intel Skylake 8168 24C, 2.7GHz - 384 GB memory
- **280 compute nodes** - Intel 8168 24C, 2.7GHz - 192 GB memory
  - 13,440 cores (48 per node (2 sockets/node))
- **4x high memory nodes** - Intel 8168 24C, 2.7GHz - 1.5 TB memory
- **48 V100 GPUs in 12 nodes** - Intel 6142 16C, 2.6GHz - 384 GB memory - 4x V100-16 GB GPU
- 10x gateway nodes

- **New NFS storage** (for users home directories) - 192 TB raw / 160 TB usable RAID6

- Mellanox **EDR Infiniband** fat tree fabric (100Gb/s)

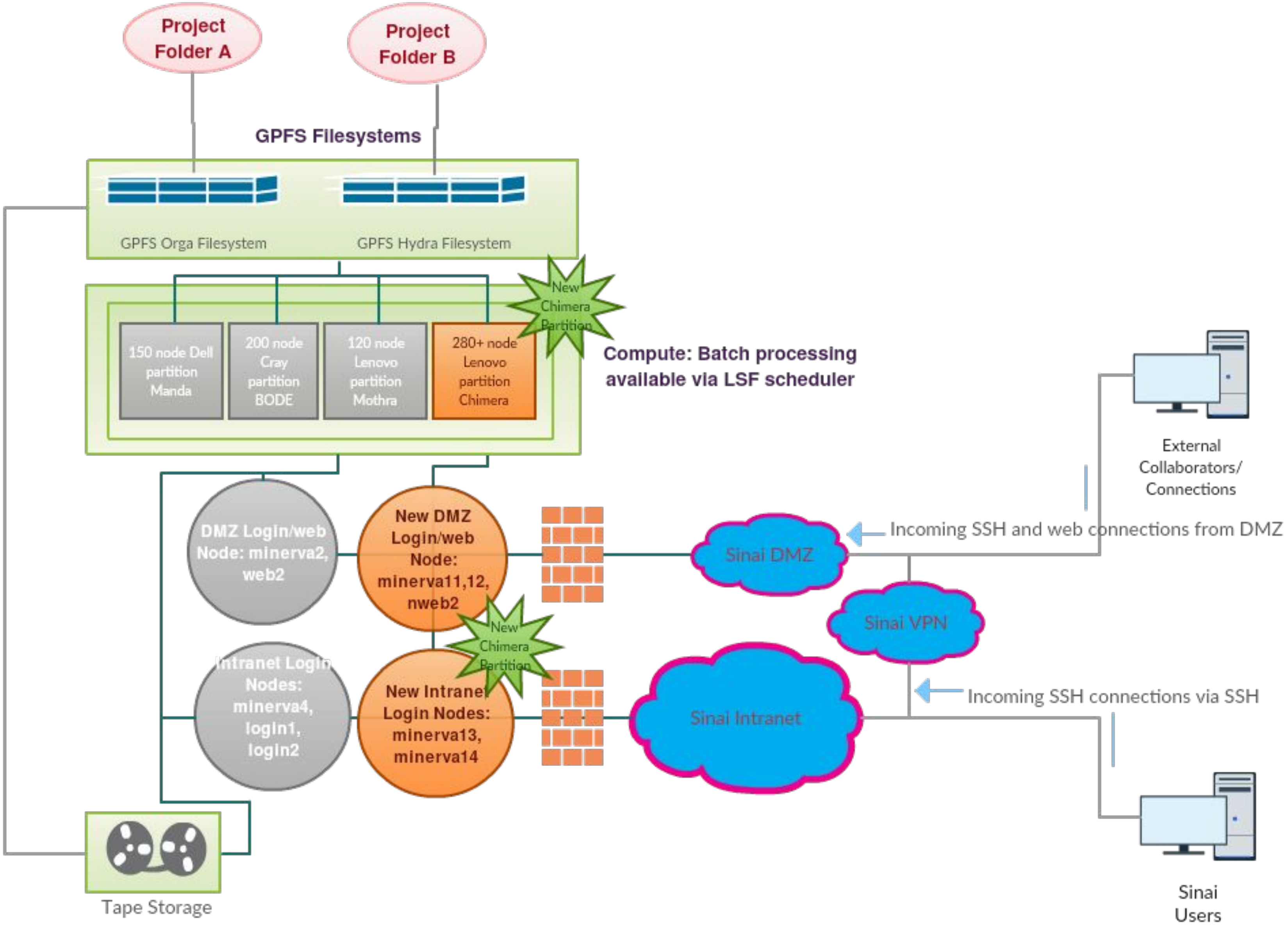
Lenovo

Total system memory (computes + GPU + high mem) = **65.7 TB**

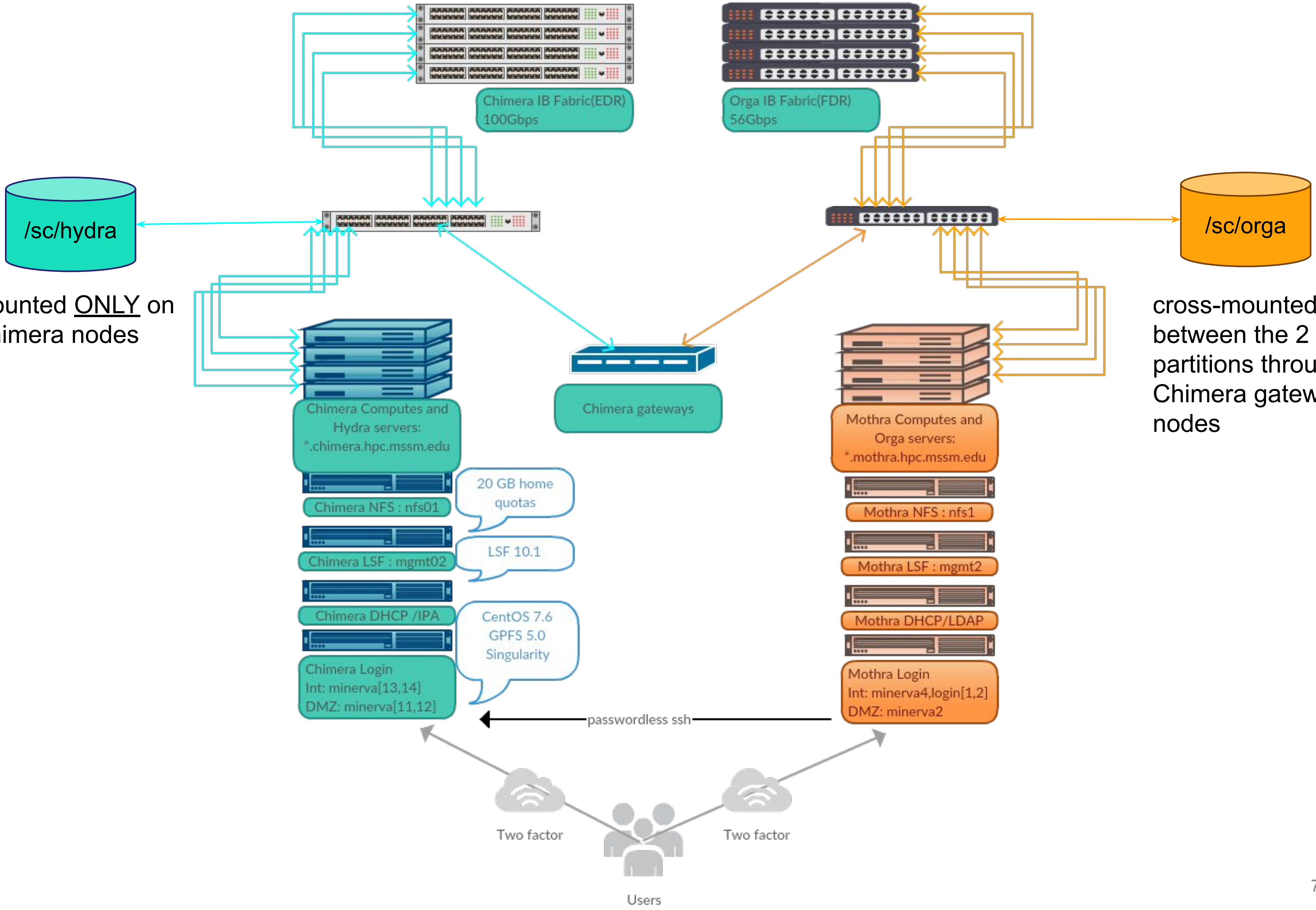
Total number of cores (computes + GPU + high mem) = **14,304 cores**

Peak performance (computes + GPU + high mem, CPU only) = **1.2 PFlops/s**

Chimera architecture


Project Folder A
Project Folder B
GPFS Filesystems
GPFS Orga Filesystem
GPFS Hydra Filesystem
New Chimera Partition
150 node Dell partition Manda
200 node Cray partition BODE
120 node Lenovo partition Mothra
280+ node Lenovo partition Chimera
Compute: Batch processing available via LSF scheduler
External Collaborators/ Connections
DMZ Login/web Node: minerva2, web2
New DMZ Login/web Node: minerva11,12, nweb2
Incoming SSH and web connections from DMZ
Sinai DMZ
Sinai VPN
New Chimera Partition
Intranet Login Nodes: minerva4, login1, login2
New Intranet Login Nodes: minerva13, minerva14
Sinai Intranet
Incoming SSH connections via SSH
Tape Storage
Sinai Users

Chimera architecture



Chimera IB Fabric(EDR) 100Gbps

Orga IB Fabric(FDR) 56Gbps

/sc/hydra

/sc/orga

Chimera gateways

Chimera Computes and Hydra servers: *.chimera.hpc.mssm.edu

Chimera NFS : nfs01

Chimera LSF : mgmt02

Chimera DHCP /IPA

Chimera Login
Int: minerva[13,14]
DMZ: minerva[11,12]

20 GB home quotas

LSF 10.1

CentOS 7.6
GPFS 5.0
Singularity

Mothra Computes and Orga servers: *.mothra.hpc.mssm.edu

Mothra NFS : nfs1

Mothra LSF : mgmt2

Mothra DHCP/LDAP

Mothra Login
Int: minerva4,login[1,2]
DMZ: minerva2

passwordless ssh

Two factor

Two factor

Users

mounted ONLY on Chimera nodes

cross-mounted between the 2 partitions through Chimera gateway nodes

Storage architecture - GPFS

GPFS upgrade to 5.X - hydra

IBM
Spectrum
Scale

Motivations:

- New features
    - autoBuildGPL
    - file system maintenance mode
    - estimate an offline mmfsck
    - mmcachectl deeper look into pagepool
    - new commands to display system health
    - **file audit logging**
    - **security compliance to NIST guidelines for encryption**
    - etc…
- Performance enhancements
- Network improvements (all Infiniband EDR fabric - 100Gb/s)

Steps:

▶ Created a new file system (**/sc/hydra**) with GPFS 5.0.2 and mounted on all chimera nodes

▶ /sc/hydra will have the same structure as /sc/orga (work, projects, scratch directories)

▶ We will provide the system path environment variable so that users can use it in their scripts.

Hydra file system

Now:

- only 1 ESS pair as part of the hydra file system for a total of 4PB

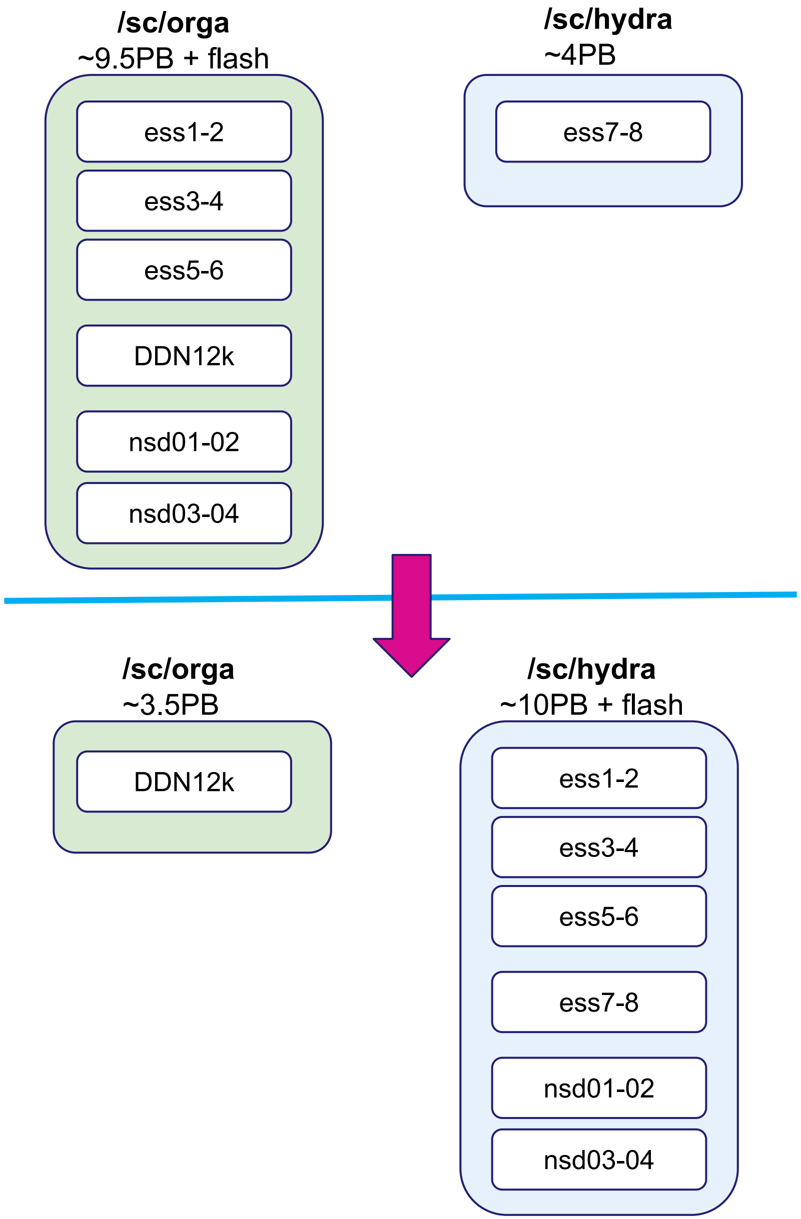
Plan:

- migrate orga directories to hydra
- migrate the remaining ESS blocks and the Flash in orga and integrate them into hydra
- upgrade them to GPFS 5
- **<u>/sc/hydra becomes the primary storage</u>**

We will <u>NOT</u> integrate DDN12k (out of support by end of 2019)

<u>Estimated migration timeline:</u>

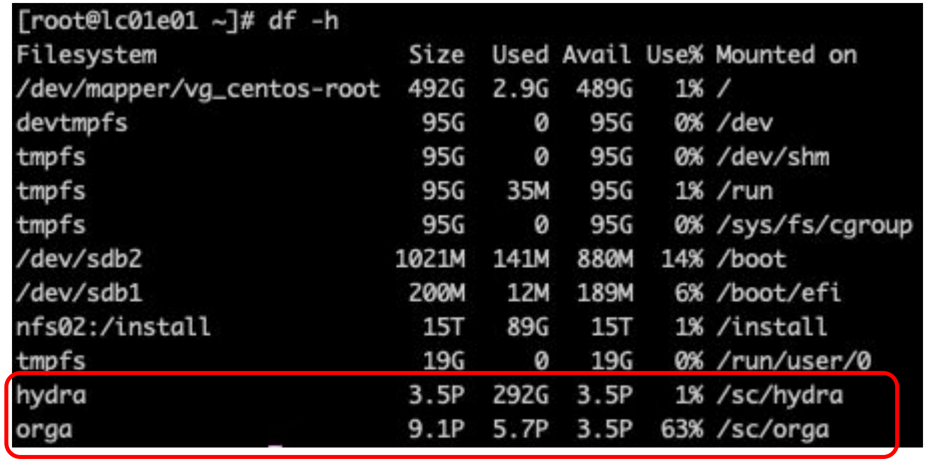
- 3-4 months (April - August)
- We will ask your cooperation and we will keep you posted on the progress



Hydra file system

Right now:

- users can access boh hydra and orga from chimera nodes
- orga is mounted through the old Infiniband fabric and has still GPFS 4.2, if you see performance issues, let us know



```
[root@lc01e01 ~]# df -h
Filesystem                Size  Used Avail Use% Mounted on
/dev/mapper/vg_centos-root 492G  2.9G  489G   1% /
devtmpfs                   95G     0   95G   0% /dev
tmpfs                      95G     0   95G   0% /dev/shm
tmpfs                      95G   35M   95G   1% /run
tmpfs                      95G     0   95G   0% /sys/fs/cgroup
/dev/sdb2                1021M  141M  880M  14% /boot
/dev/sdb1                 200M   12M  189M   6% /boot/efi
nfs02:/install             15T   89G   15T   1% /install
tmpfs                      19G     0   19G   0% /run/user/0
hydra                     3.5P  292G  3.5P   1% /sc/hydra
orga                      9.1P  5.7P  3.5P  63% /sc/orga
```

Eventually, when everything has been migrated (directories and storage servers), **we will remove orga (ETA: end of 2019)**.

Use /sc/hydra as default storage location!

User environment - login, account migration with freeIPA

New login nodes

New set of login nodes:

- 4 new login nodes: **minerva[11-14]**, which points to the login node **li03c[01-04]**.
  - **minerva[13-14] (or li03c[03-04]) are internal login nodes**

    - currently available via Minerva and MSSM campus.
  - **minerva[11-12] (or li03c[01-02]) are external login nodes**

    - will be available when DMZ network is setup.
- Separate **Globus endpoint** will be configured.
- Data transfer nodes: **data2, data4**.
  - will be included in Chimera partition.
- Other login nodes, minerva2&4 and login1&2, will be retired along with their compute partition.

We will update the changes of login method after the migration period.

Updates will be announced on our HPC website as well as the weekly newsletter.

freeIPA and account migration

The authentication for the Chimera partition is pointing to the new freeIPA server.

freeIPA instead of LDAP:

- More user friendly administration (GUI, CLI)
- Stronger security standards and more powerful account management (password expire/reset)

For external users and users who use HPC password (+vldap +yldap):

A new set of HPC passwords will be deployed.

Password setup will be announced when the new external login nodes are configured.

Login method

During migration period, you can assess Chimera by:

1. All users including external users

   Hop from Minerva internal login nodes (passwordless).

```
[jiangd03@minerva4 ~]$ ssh li03c03
[jiangd03@minerva4 ~]$ ssh li03c04
```

2. Sinai users

   Login from campus (two factor authentication), please choose any of the following combination:

Login method	Login hosts	Password Components
user1 user1+vkrb	@chimera.hpc.mssm.edu @minerva13.hpc.mssm.edu @minerva14.hpc.mssm.edu	Sinai Password + 6 Digit Symantec VIP token code
user1+ykrb		Sinai Password + YubiKey Button Push

Note: Load balancer **Round-robin** is configured for **chimera.hpc.mssm.edu.** It will distribute client connections across a group of login nodes.

User environment - NFS home directory, compute node, software packages

New NFS storage

New NFS storage (/hpc):

- For users home directories and applications.
- Mounted to all chimera nodes @100Gb/s.
- Storage available: 160TB usable RAID6.
- User quota is increased to 20G.

To facilitate the transition and move data over the new storage:

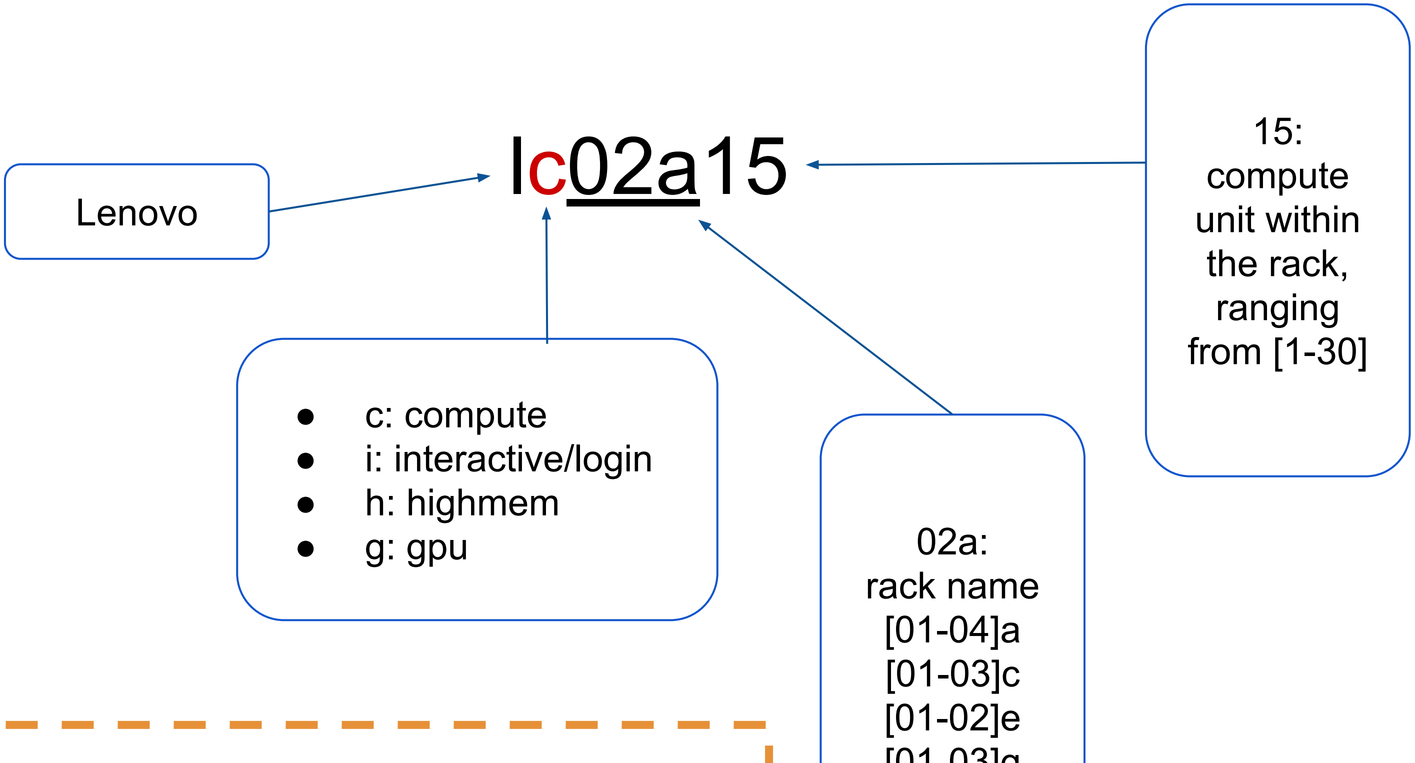
- The old /hpc is mounted on the login nodes as **/hpc-old/users in READ-ONLY mode**.
- A link (~/userid-old-home) is created in user home dir pointing to their old home directory.

```
[jiangd03@li03c03 ~]$ ls -la jiangd03-old-home
lrwxrwxrwx 1 root root 23 Mar 26 10:53 jiangd03-old-home -> /hpc-old/users/jiangd03
```

- **Note:** /hpc-old is not mounted on compute nodes, so job will fail if it uses ~/userid-old-home.

We urge users to move their data over the new storage as the old NFS will be out of support in July 2019.

Compute node naming scheme



lc02a15

Lenovo

- c: compute
- i: interactive/login
- h: highmem
- g: gpu

02a:
rack name
[01-04]a
[01-03]c
[01-02]e
[01-03]g

15:
compute
unit within
the rack,
ranging
from [1-30]

Quiz time

- li03c01
- lg03a04
- lc01g05
- lh03c01

OS Upgrade And New Software Environment

OS: **Centos 7.6** was deployed in Chimera

Containers:

Singularity was installed on login node and a couple of nodes as resources, to submit job through LSF:

```
bsub -q premium -n 1 -W 00:10 -R singularity “singularity run hello.simg”
```

Package Upgrade:

Glibc-2.17 available. If higher version needed, go with a container

Key packages of latest version are being built under centos7.6

Such as gcc/8.3.0, openmpi/4.0.1, intel/2019, Python/3.7.3, R/3.5.3, Rstudio/1.1.463

Lmod Software Environment Module system implemented:

Lmod directly supports software hierarchy

module spider and **module keyword** to find specified modules

ml convenient tool

Version precedence; Autocompletion

module save: Lmod provides a simple way to store the currently loaded modules and restore them later through named collections

LSF 10.1 and new queue structure

Job scheduling system

New Job scheduling system in Chimera:

- New job scheduling server.
- LSF upgrade to v10.1.
- New job ID serials.
- Job temporary dir configured to /local/JOBS instead of /tmp.
- Gold is not implemented. LSF flag “*-P acc_xxx*” is no longer needed.

New features in queue:

- Absolute job priority scheduling is on all queues
  - APS number calculated according to job and queue priorities, taking user-based fairshare in consideration.
  - APS factore is adjustable reflecting user bonus.

To submit a job (more detail will be given in the training session):

```
# interactive session
[jiangd03@li03c03 ~]$ bsub -q interactive -n 1 -W 00:10 -Is /bin/bash
# example job submission
[jiangd03@li03c03 ~]$ bsub -q normal -n 1 -W 00:10 echo “Hello World”
```

Queue structure in Chimera

Queue structure in Chimera			
Queue	**priority/APS**	**Wall time limit**	**available resources**
interactive		12 hours	4 nodes+1 GPU node
normal	100/APS	3 days	77 nodes
premium	200/APS	6 days	200 nodes + 2 high-men
express		12 hours	280 nodes
long	100/APS	2 weeks	2 dedicated highmen nodes (96 cores)
GPU	100/APS	6 days	48 V100
private		unlimited	private nodes

Interactive sessions

Interactive sessions:

- No interactive nodes is configured in Chimera partition to avoid abusive usage. Interactive sessions is available via job scheduler in the **interactive queue**.
- Nodes in interactive queues will have outside network access, i.e., **data transfer** will be available in the interactive sessions.
- **Interactive GPU** will be available for job testing.
- Interactive1&2 and Interactive5&6 will be retired along with their compute partition.

Documentation and training

Documentation and training

- **For most recent announcement and updates**
  - Join our mail-list: hpcusers@mssm.edu
  - Follow us on Twitter @mssmhpc
  - Minerva user group meetings will be scheduled as needed.

- **Different training sessions will be offered this year**
  - **April 9th 2019** One training session on “Introduction to Chimera”
    Where and when: Icahn Building L3-82 @ 2:00pm ~ 2:30pm
  - **Fall 2019** Two training sessions
    Topics include “Introduction to Minerva” and “LSF job scheduler”

- **Documentation update on the website (https://hpc.mssm.edu/)**
  - We are periodically refreshing the website.
  - We will provide additional training material (including slides) online.

Future roadmap

Future plans

- GUI server for Rstudio, jupyter
- HIPAA compliant cluster
- /sc/orga migration to /sc/hydra (ESS storage servers and data)
- New TSM server
- New Web server
- New Globus endpoint

Big Omics Data Engine 2 NIH S10 Proposal

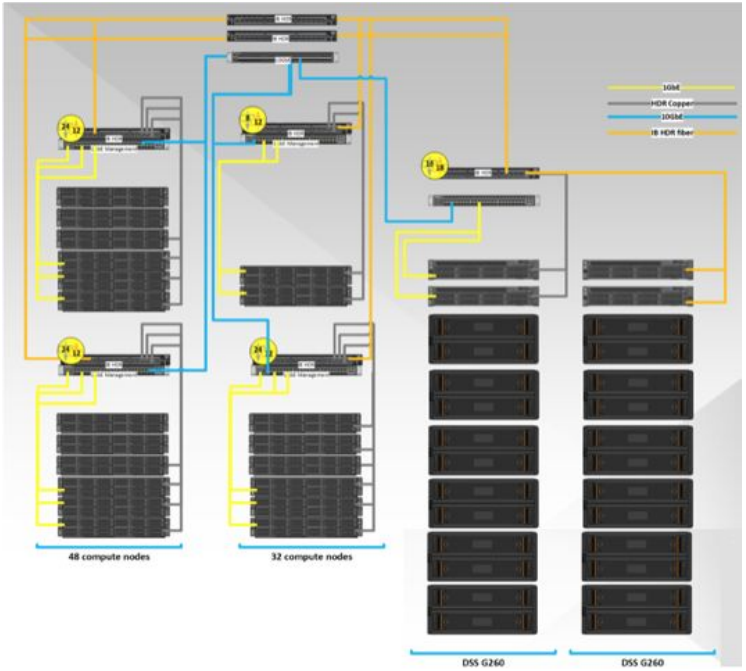
BODE2

- **11 PB** of Lenovo DSS high performance storage
- **3,200 compute cores** (80 nodes with 40 Intel Cascade Lake cores and 192 GB memory each)
- Mellanox **EDR Infiniband** fat tree fabric (100Gb/s)

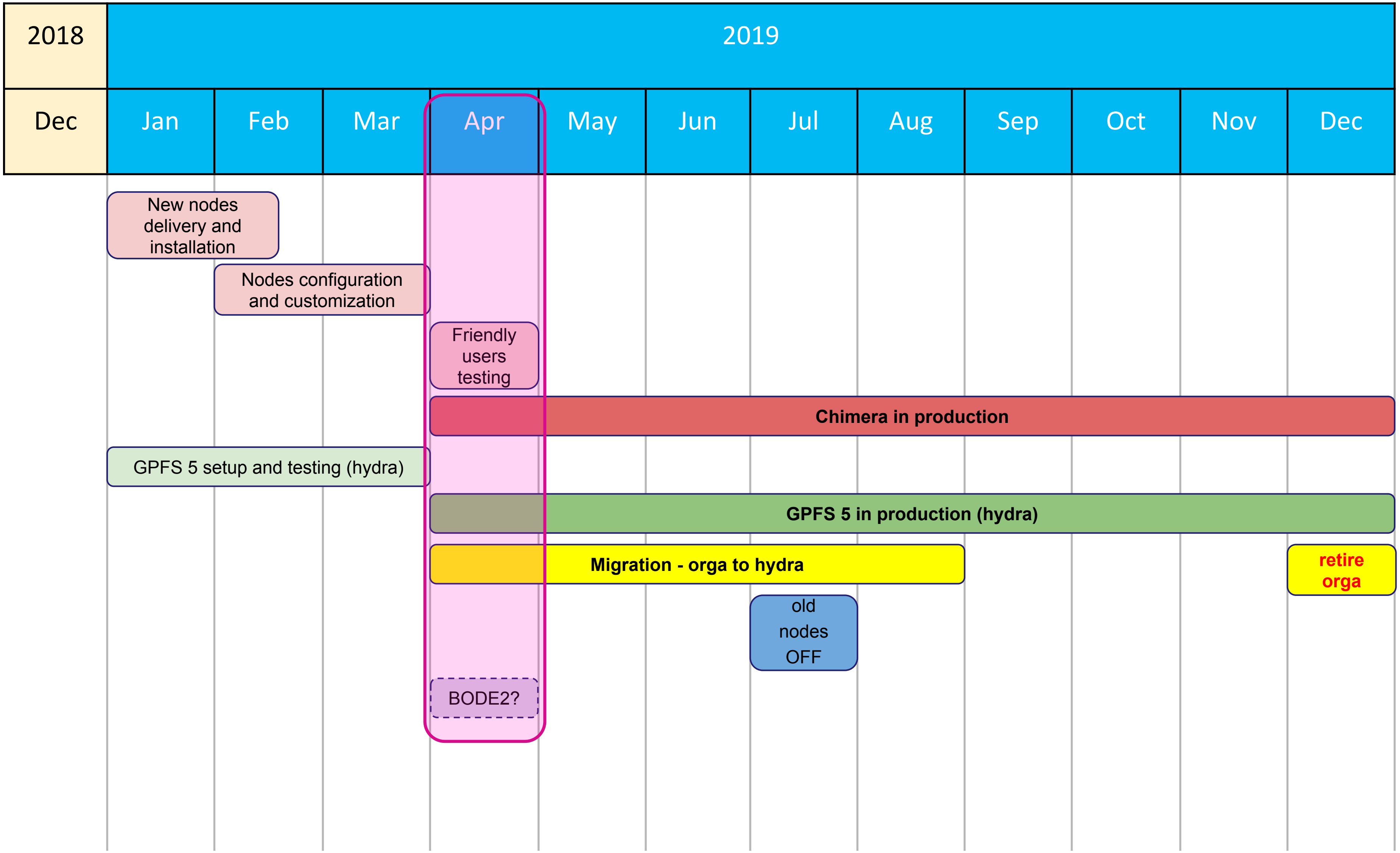
Thank you very much for your input while preparing the proposal!

We got a score of **18** from the reviewers.
We are waiting for the final decision to be made.

If awarded, we will have 1 year to deploy the system.



HPC Roadmap - updated



2018

2019

Dec

Jan

Feb

Mar

Apr

May

Jun

Jul

Aug

Sep

Oct

Nov

Dec

New nodes delivery and installation

Nodes configuration and customization

Friendly users testing

Chimera in production

GPFS 5 setup and testing (hydra)

GPFS 5 in production (hydra)

Migration - orga to hydra

retire orga

old nodes OFF

BODE2?

Question and comments

Thank you!