Minerva Town Hall
November 2018

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

Nov 29, 2018
Outline

Welcome and general comments

- 2018 Minerva status
- 2018 accomplishments

2019 Plans for infrastructure and service upgrades

- Storage upgrade
- Compute node and infrastructure upgrades
- OS upgrade, new package rebuild and queuing structure
- Other server infrastructure and service upgrades

Q&A
Welcome and general comments
## Minerva status (2018)

### Accounts

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of new users</td>
<td>360</td>
</tr>
<tr>
<td>Number of active users in 2018</td>
<td>751</td>
</tr>
<tr>
<td>Number of total users</td>
<td>2,095 (722 external users)</td>
</tr>
<tr>
<td>Number of project groups</td>
<td>260</td>
</tr>
</tbody>
</table>

### Compute

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jobs run</td>
<td>21,955,985</td>
</tr>
<tr>
<td>CPU-hours utilized</td>
<td>34,173,228</td>
</tr>
</tbody>
</table>

### Storage

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>High-speed storage used</td>
<td>5.1 PB (77% utilization)</td>
</tr>
<tr>
<td></td>
<td>1,824,776,706 files</td>
</tr>
<tr>
<td>Archival storage used</td>
<td>8.3 PB (16.7 PB total including offsite copies)</td>
</tr>
<tr>
<td></td>
<td>339,122,974 files</td>
</tr>
</tbody>
</table>
Jobs and CPU-hours break down by compute resource:

<table>
<thead>
<tr>
<th>Compute Partition</th>
<th># Jobs</th>
<th>CPU-hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manda</td>
<td>11,416,2793</td>
<td>18,452,872</td>
</tr>
<tr>
<td>Mothra</td>
<td>5,9440642</td>
<td>9,915,929</td>
</tr>
<tr>
<td>BODE</td>
<td>4,592,758</td>
<td>5,769,886</td>
</tr>
<tr>
<td>Hi-memory node</td>
<td>175</td>
<td>27,557</td>
</tr>
<tr>
<td>GPU nodes</td>
<td>1,945</td>
<td>1,005,887</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>20,424,949</strong></td>
<td><strong>31,585,896</strong></td>
</tr>
</tbody>
</table>

Total 5.1 PB used by 1,824,776,706 files from 260 projects.
### Top 10 users - computing

<table>
<thead>
<tr>
<th>PI</th>
<th>Department</th>
<th>Core-hours</th>
<th>Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lisa Edelmann</td>
<td>Genetics and Genomic Sciences</td>
<td>3,251,516</td>
<td>1,173,226</td>
</tr>
<tr>
<td>Andrew Sharp</td>
<td>Genetics and Genomic Sciences</td>
<td>1,447,416</td>
<td>1,370,379</td>
</tr>
<tr>
<td>Bin Zhang</td>
<td>Genetics and Genomic Sciences</td>
<td>911,529</td>
<td>1,285,309</td>
</tr>
<tr>
<td>Eli Stahl</td>
<td>Psychiatry</td>
<td>907,428</td>
<td>796,295</td>
</tr>
<tr>
<td>Andrew McLellan</td>
<td>Genetics and Genomic Sciences</td>
<td>739,872</td>
<td>401,695</td>
</tr>
<tr>
<td>Ke Hao</td>
<td>Genetics and Genomic Sciences</td>
<td>725,530</td>
<td>963,461</td>
</tr>
<tr>
<td>Allison Goate</td>
<td>Neuroscience</td>
<td>895,652</td>
<td>482,459</td>
</tr>
<tr>
<td>Jun Zhu</td>
<td>Institute of Multiscale biology</td>
<td>674,677</td>
<td>724,753</td>
</tr>
<tr>
<td>Panos Roussos</td>
<td>Psychiatry</td>
<td>564,464</td>
<td>680,386</td>
</tr>
<tr>
<td>Eric Schadt</td>
<td>Genetics and Genomic Sciences</td>
<td>502,956</td>
<td>496,388</td>
</tr>
</tbody>
</table>
## Top 10 users - storage

<table>
<thead>
<tr>
<th>User</th>
<th>Department</th>
<th>Storage usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bin Zhang</td>
<td>Genetics and Genomic Sciences</td>
<td>817.3 TB</td>
</tr>
<tr>
<td>Robert Sebra</td>
<td>Genetics and Genomic Sciences</td>
<td>546.5 TB</td>
</tr>
<tr>
<td>Lisa Edelmann</td>
<td>Genetics and Genomic Sciences</td>
<td>430.9 TB</td>
</tr>
<tr>
<td>Gabriel Hoffman</td>
<td>Genetics and Genomic Sciences</td>
<td>240.1 TB</td>
</tr>
<tr>
<td>Alison Goate</td>
<td>Neurosciences</td>
<td>238.5 TB</td>
</tr>
<tr>
<td>Eli Stahl</td>
<td>Genetics and Genomic Sciences</td>
<td>218.5 TB</td>
</tr>
<tr>
<td>Milind Mahajan</td>
<td>Genetics and Genomic Sciences</td>
<td>194.8 TB</td>
</tr>
<tr>
<td>Panagiotis Roussos</td>
<td>Psychiatry</td>
<td>161.5 TB</td>
</tr>
<tr>
<td>Eimear Kenny</td>
<td>Genetics and Genomic Sciences</td>
<td>144.5 TB</td>
</tr>
<tr>
<td>Samir Parekh</td>
<td>Oncological Sciences</td>
<td>142.6 TB</td>
</tr>
</tbody>
</table>
2018 Minerva Planned Outages

Light and full PMs scheduled during the year to
- Integrate the new storage/flash into the existing file system.
- Perform OS and software upgrades.
- Upgrade the Infiniband high bandwidth, low latency network.

Allocation accounting system (GOLD) failures
Jobs failed due to the GOLD being unable to handle massive concurrent database queries from LSF. To fix and reduce GOLD failures:
- The GOLD database was optimized.
- The process limit was increased.
- A new expressalloc queue was added to avoid GOLD authentication.
- Other queues were restructured to be able to requeue the failed jobs.

We will investigate further for a more permanent scalable solution.

Achieved 99% uptime in 2018
2018 - Accomplishments

Thank you very much for the feedback from user survey!

Actions we took in 2018 in response to the user survey:

- Upgraded and expanded the available storage.
- Upgraded the OS and rebuilding packages.
- Deployed the Demeter data science cluster for the community.
- Made cloud services available such as Spark and containers (Singularity).
- Updated the documentation and presented 3 tutorial sessions.
- Submitted proposal for NIH S10 grant for BODE2.
- Purchased new compute partition.

Main projects to accomplish in 2019:

- Upgrade GPFS to v5.x.
- Install new compute partition.
- Upgrade OS and rebuild packages.
- Other infrastructure and service upgrades.
- Make Minerva HIPAA-compliant.

Details will be presented in the following slides.
Storage upgrade
2018 - Storage Upgrade

Motivation:
● Accommodate the needs for more storage space.
● Replace hardware that is at End Of Life.

Accomplishments in 2018:
● IBM ESS storage installation and data migration
  Installed 6 PB of IBM Elastic Storage Server to provide additional and faster storage to Minerva users and completed users' data migration to the new pool.
● Upgraded from GPFS 4.1 to 4.2
● Decommissioned IBM GSS (3 PB) and DDN 10K (1.5 PB)
● New storage purchase and installation
  ○ New Flash file system 264 TB (raw) for metadata and small files.
  ○ New ESS file system 4 PB to be added to the existing storage.
● Migrated metadata from old to new Flash
● Decommissioned old Flash
2018 - Storage current status

Orga – GPFS File System ~14 PB
- IBM LE ESS 4 PB
- IBM BE ESS 6 PB
- DDN 12K 3.5 PB
- Performance Tier (Flash) – 264 TB

Fat-Tree InfiniBand Network & Ethernet

Hosted Systems
- 120 Compute
  - 64 AMD Interlagos cores each node
  - 256 GB RAM each
  - MANDA

- 209 Compute
  - 12 Intel Ivy Bridge cores each node
  - 64 GB RAM each
  - MOTHRA

- 207 Compute
  - 12 Intel Haswell cores each node
  - 64 GB RAM each
  - BODE

Logins
Web
TSM Archive
Databases
NFS
Management
Scheduling
Monitoring & Reporting

Sinai Firewall
Genomics Core Facility
All 10Gb
@56Gb each
@40Gb each
@56Gb each
@56Gb each

Sinai Campus
Internet Firewalls
NYSERNET (12) Level3 Cogent
2019 - GPFS upgrade to 5.X

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

Steps/Roadmap:

ETA: 2019 - TBD, we will keep you posted!

- We will create a new file system (/sc/ hydra) and provide the system path environment variable so that users can use it in their scripts.
Compute nodes and infrastructure upgrade: Chimera
Compute nodes and infrastructure upgrade: Chimera partition

Specs of the new compute partition (Chimera):

- **12x 42U racks**
- **4x login nodes** - Intel 8168 24C, 2.7GHz - 384 GB memory
- **286 compute nodes** - Intel 8168 24C, 2.7GHz - 192 GB memory
  - 13,728 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)

**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**
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
Compute nodes and infrastructure upgrade: Chimera
OS upgrade, package rebuild and new queue structure
OS upgrade and container support

Currently:
- **Centos 6.9** was deployed on all the compute nodes and login nodes.
- **Centos 7.4/7.5** was deployed on new login nodes: data2, data4.
- **Centos 7.4/7.5** was deployed on GPU node node19-2 and nodes in test-centos7 queue.

In 2019:
- The new pool of compute nodes will be installed with **Centos 7**.
- We will update the rest of the computes partition, i.e., manda, mothra and bode, after the integration of the new compute nodes.
- We will prepare for the **Centos 8** upgrade when it is released.

Containers:
- **Singularity** was implemented on test-Centos7 queue in Minerva.
- This will be part of the new OS stack on Chimera including the GPU nodes.
New package rebuild system

Motivation:
The upgraded OS kernel requires the new software stack.

Package rebuild:

- 2293 packages currently on Minerva, most of which will be compatible on Chimera
- New packages are also being rebuilt on centos7 with up-to-date compilers (158 done)
- Glibc-2.17 available
- New packages/ Lmod module system are available for test in test-centos7 queue

Key features of EasyBuild/Lmod module system:

- EasyBuild fully automates software builds with automatic dependency resolution.
- Lmod directly supports software hierarchy, find all possible modules by module spider, ml etc.
Queue structure changes in Minerva

Queue structure changed in 2018:

- Interactive queue
  - Easier access to interactive sessions when interactive nodes are busy.
- GPU queue
  - Users are now able to specify GPU resources (K20, P100 and V100 in future).
- Expressalloc queue
  - Minimize the effect of gold failure for the submission of millions of short jobs.
- Absolute job priority scheduling
  - APS number calculated according to job and queue priorities, taking user-based fairshare in consideration.

Queue structure changes in 2019:

- Chimera will have its own job scheduler system:
  - Job scheduler: LSF 10.x vs Slurm 19.x?
  - New queue structure
  - New implementation to address gold failure
# Queue structure changes

<table>
<thead>
<tr>
<th>Current queue structure (Manda+Mothra)</th>
<th>Queue structure in Chimera</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Queue</strong></td>
<td><strong>Wall time limit</strong></td>
</tr>
<tr>
<td>premium</td>
<td>6 days</td>
</tr>
<tr>
<td>expressalloc</td>
<td>6 hours</td>
</tr>
<tr>
<td>alloc</td>
<td>6 days</td>
</tr>
<tr>
<td>low</td>
<td>6 days</td>
</tr>
<tr>
<td>interactive</td>
<td>12 hours</td>
</tr>
<tr>
<td>GPU</td>
<td>6 days</td>
</tr>
<tr>
<td>private</td>
<td>unlimited</td>
</tr>
</tbody>
</table>
Login nodes and interactive sessions

New set of login nodes:
- 4 new login nodes: minerva11-minerva14.
- Separate Globus endpoint.
- Data transfer nodes: data2, data4.
- Other login nodes, minerva2&4 and login1&2, will be retired along with their compute partition.

Interactive sessions:
- No interactive nodes will be available in Chimera partition to avoid abusive usage. Interactive sessions will be 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

- 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 next year.
  - Feb. 2019 One training session on “Introduction to Chimera”
  - Spring 2019 Two training sessions
    - Fall 2019 Two training sessions
      - Topics include “Introduction to Minerva” and “LSF job scheduler”
  - Introduction to Scientific Computing BSR1015 is a two credit course with an expanded lab, taught by Anthony Costa, PhD.

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

- Will issue a user survey soon
Other infrastructure and service upgrade
Other infrastructure upgrades

Motivation:
- Server hardware will be End Of Life in 2019.
- Old software stack is unable to support the new service features.

Chimera partition:
- Ldap server
- Job scheduler server choices - need your feedback
  - LSF 10.1
  - Slurm 19.05

Web server
- python 2.7+ to support newer website architectures.

TSM server
- Spectrum protect upgrade to 8.1.X
- Possible tape drive upgrade to LTO7/8, for faster read/write speed and accommodate more data per tape.
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.
Final decision will be made at the end of January 2019.

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

2018

Dec

New nodes delivery and installation

Nodes configuration and customization

Friendly users testing

2019

Jan

Feb

Mar

Apr

May

Jun

Jul

Aug

Sep

Oct

Nov

Dec

Chimera in production

GPFS 5 setup and testing

GPFS 5 in production

old nodes OFF

BODE2?
Question and comments
Thank you!