

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

1/18/16

Blue Waters User Monthly Teleconference



GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

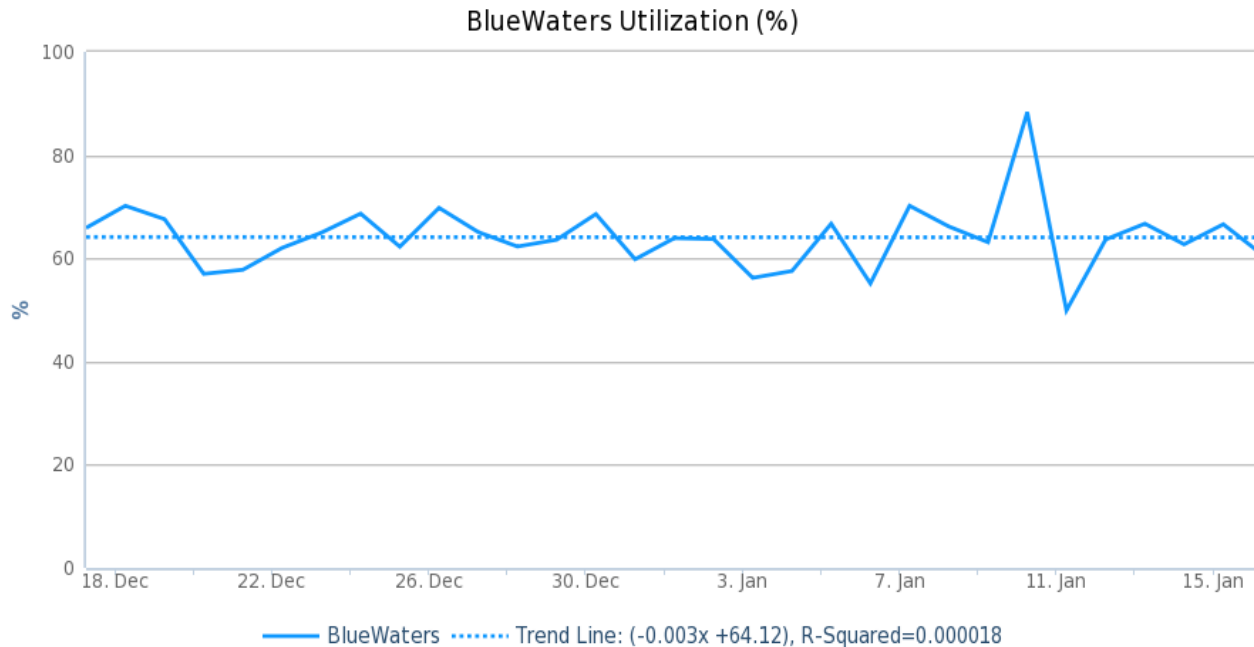
CRAY®

Agenda

- Utilization
- Maintenance, Upgrades and Updates
- Topology Aware Scheduling Analysis
- Recent Events
- Upcoming Opportunities
- PUBLICATIONS!

System Utilization

- Utilization since last BW User Call (December 18)



2015-12-17 to 2016-01-16 Src: HPCDB. Powered by XDMoD/Highcharts

- Consistent utilization. Still opportunities for backfill etc.

Maintenance, Upgrades and Updates

- Cray Linux Environment upgrade 5.2UP04
- CUDA 7.0 (CUDA 7.5 coming)
 - C++11 support, Thrust 1.8. Runtime Compilation library (nvrtc), [Release Notes](#)
- Programming environment defaults PE15.12
- Compilers
 - CCE 8.4.2 – switch -hgnu is now default.
 - GNU 4.9.0 for CUDA 7 support). 5.2.0 is available.
 - PGI 15.10.0 - 15 and 14 are incompatible.

Maintenance, Upgrades and Updates

- CRAY-MPICH 7.3.0

- MPICH_MEMORY_REPORT – provide memory use summary or per rank:

```
# MPICH_MEMORY: Max memory allocated by malloc: 52176000 bytes by rank 32
# MPICH_MEMORY: Min memory allocated by malloc: 52169968 bytes by rank 13
# MPICH_MEMORY: Max memory allocated by mmap: 10762304 bytes by rank 0
# MPICH_MEMORY: Min memory allocated by mmap: 10756160 bytes by rank 1
# MPICH_MEMORY: Max memory allocated by shmget: 82859480 bytes by rank 0
# MPICH_MEMORY: Min memory allocated by shmget: 0 bytes by rank 1
```

- `cray_cb_write_lock_mode` - Specifies the file locking mode: single lock shared by all MPI ranks for writes only.

Maintenance, Upgrades and Updates

- Moab (9.0) and Torque (6.0 + patches) updated
 - New machine to host services.
 - Improved iteration time.
 - Fixes for some stability issues.
- New Lustre client - 2.5.2.

Changes

- Change in PBS_JOBID suffix ...
 - JOBID.bw
- Maximum wallclock time increased to 48 hours.
 - Check your job wall clock accuracy; discount for 75% accuracy or better.

Recent Events

- 1/4 - HSN issue during weekly warm-swap maintenance period. A compute blade dropped during the network quiesce. Changes were made to limit impact.

LAPACK Survey

The LAPACK Team would like your opinion and input on dense linear algebra software packages that you use, in particular the libraries LAPACK, ScaLAPACK, PLASMA and MAGMA. The purpose of this survey is to improve these libraries to benefit you, the users. There are many possible improvements that could be made, and this survey will help the team prioritize them.

The survey has six parts and it's easy to skip a part if it's not relevant.

1. A general section about your applications and their needs.
2. Specific questions about your LAPACK uses, if any.
3. Specific questions about your ScaLAPACK uses, if any.
4. Specific questions about your PLASMA uses, if any.
5. Specific questions about your MAGMA uses, if any.
6. An open section for any additional comment.

You will find the survey here: <https://www.surveymonkey.com/r/2016DenseLinearAlgebra>

Please send any questions you may have to dongarra@icl.utk.edu.

For more information on LAPACK please visit <http://www.netlib.org/lapack/>

iSTEM Survey

Topology-aware Scheduling Investigation

- Topology-aware job placement was disabled from started Nov 6 to Dec 6.
 - Was done without an announcement to minimize changes in user behavior.
- We are interested in comparing performance differences for comparable jobs run during the test period. We need your help.
- If interested please provide:
 - the name of the application
 - job ids for the jobs referenced,
 - a one-line description of the computation being performed,
 - measure of performance (e.g job run-times or time per iteration) during and after the test period.
- In return,
 - 100% credit for the compared
 - **Deadline Jan 15, 2016**

Note: Jobs which don't rely on inter-node communication would not have been impacted by different placement, and thus, are excluded from this solicitation. This would include single node jobs, bundled single node jobs.

Upcoming Changes

- File System Upgrade 12/2015-01/2016 (DELAYED)
 - Move to declustered RAID: GridRAID
 - Move to Hierarchical Storage Management (HSM).
 - Working to minimize inconvenience

File System Upgrade 11/2015 – 01/2016

- Upgrade to Lustre 2.5.1+ server
- Move to Declustered RAID
 - **Fewer OSTs** while keeping same total capacity.
- Internally
 - Data organized better – faster rebuild times
 - Support for Hierarchical Storage Management (HSM) for Lustre to Nearline.
- Externally
 - Better performance.
 - Ease of use from HSM.

Advanced User Workshop (Rescheduled)


- At NCSA.
- The new date will be in March 14-18, 2016.
- Agenda to appear soon.

Annual Blue Waters Symposium

- June 13-15, 2016
- [Sunriver Resort](#) near Bend, Oregon
- Information about past events
 - [2015](#)
 - [2014](#)
- Serves as PRAC PI meeting.



Request for Science Successes

- We need to be current on products that result from time on Blue Waters such as:
 - Publications, Preprints (e.g. [arXiv.org](https://arxiv.org) ), Presentations.
 - Very interested in data product sharing.
- Appreciate updates sooner than annual reports.
 - Send to gbauer@illinois.edu
- NSF PRAC teams send information to PoCs.
- See the [Share Results](#) section of the portal as well.
- **Be sure to include [proper acknowledgment](#)**
 - Blue Waters - National Science Foundation (ACI 1238993)
 - NSF PRAC – OCI award number

Discount charging

Charge factors for completed jobs that meet one or more of the following criteria below are discounted by 25% for each of the following opportunities with a resulting maximum discount of 69% when compounded.

1. job backfills available nodes
 2. submit pre-emptible job
 3. use flexible wall clock time
 4. job wall clock accuracy of 75% or better
- For more information see the [July 9th blog entry Charge Factor Discounts for jobs on Blue Waters.](#)

Review of Best Practices

- Improper use of login nodes
 - Use compute nodes for all production workloads.
- Avoid excessive calling of job scheduling commands
 - Unintentional denial of service may result otherwise.
- Unbundling of Jobs
 - Independent jobs bundled to 32 nodes or less best for backfill etc.
- Small files usage
 - For application small file IO use projects then scratch.
 - Tar up files before transferring to Nearline.
 - Use directory hierarchies.
 - Avoid many writers to same directory.