BLUE WATERS

SUSTAINED PETASCALE COMPUTING

4/20/15

Blue Waters User Monthly Teleconference

NCSA

GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

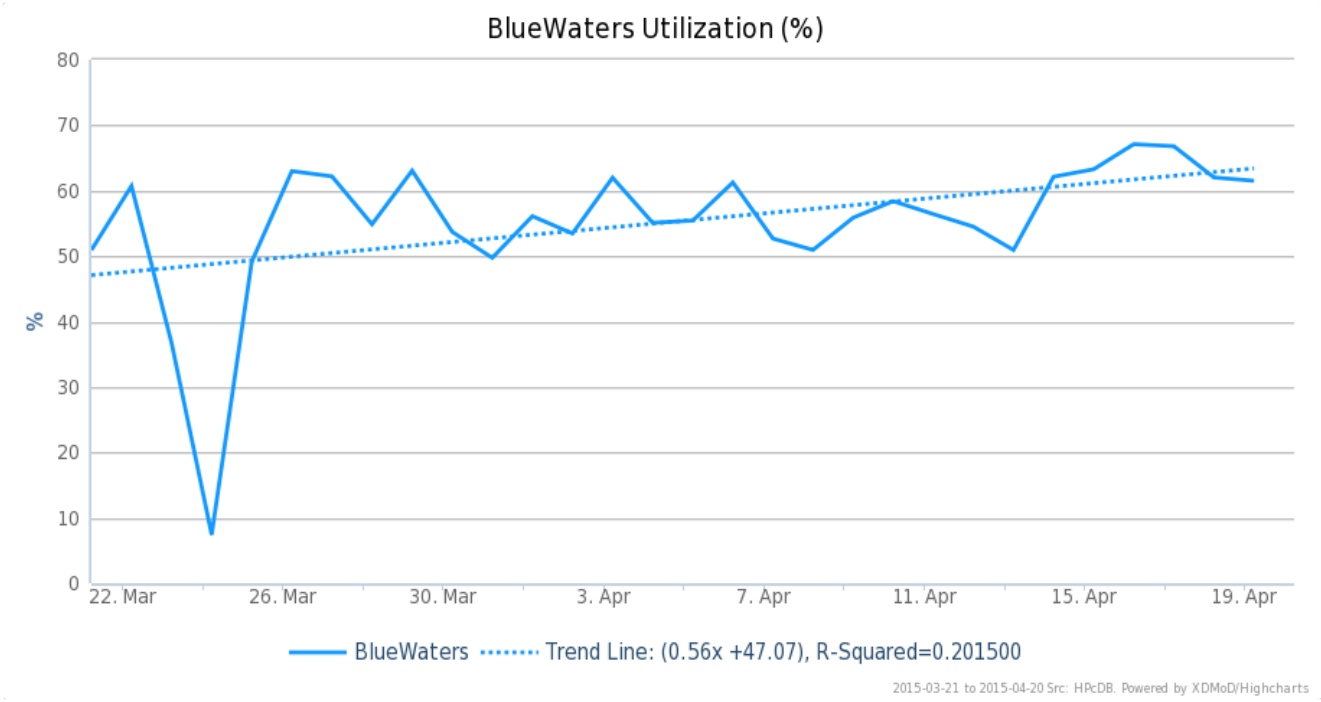
CRAY

Agenda

- Utilization
- Recent events
- Recent changes
- Upcoming changes
- Blue Waters Data Sharing
- 2015 Blue Waters Symposium
- PUBLICATIONS!

System Utilization

- Utilization since last BW User Call (March 16)



- Utilization improving. We are still investigating.

Recent Opportunities

- Discount Period
  - From
    - Friday, April 17th at 12:00AM
    - Midnight May 15th
  - Charge factors reduced by 50% for all queues.

Recent Issues

- Login node responsiveness
  - Improved but not totally resolved.
  - Please continue to report slow login nodes.

- Outages
  - March 24th – high speed network issue due to IO subsystem HA functionality.
  - April 8th – power feed issue.

Recent Changes

- Thursday, April 9th
  - Disabled SSLv3 support for Globus.
    - 500-globus_gsi_gssapi: SSLv3 handshake problems: Couldn't do ssl handshake
    - 500-OpenSSL Error: s3_srvr.c:956: in library: SSL routines, function SSL3_GET_CLIENT_HELLO: wrong version
  - TLS is now the supported method. Please update installations of Globus Connect and Globus Toolkit.

Recent Changes

- Monday April 20th.
  - XK (GPU node) reconfiguration on the high-speed network (HSN). Will have 4228 nodes.
    - Was - 15x6x24
    - By tomorrow - 23x4x24
    - Fewer Y links and more XZ links will provide better aggregate bandwidth.
    - No XE gap jobs that can cause interference.
    - No changes to job scripts needed unless you are using the #PBS –l geometry= and the Y values > 4.

Recent Changes

- Monday April 20th.
  - RSIP gateway port limit
    - Network gateway software patch (per scale testing)
    - Before – 22 ports per node
    - After – several thousand ports per node
    - Impacted use of coupled cluster mode (CCM) workflows at scale.

Upcoming Changes

- CUDA 6.5
  - Installed on our test and development system.
  - Initial tests for functionality are good.
  - Looks like a recompilation is required.
- Modules
  - Continue to clean out unused and old modules.
  - Removing deprecated modules: old xt-xyz replaced with cray-xyz.

Improving Job Turn-around time

Improvement of job wall clock accuracy

Wall clock accuracy is defined as actual elapsed wall clock time divided by requested wall clock time. Providing accurate wall clock times makes it easier for the scheduler.

Job backfills available nodes that are draining for other jobs

Backfilling improves job turn-around time and utilization and happens most often for jobs requesting less than several hundred nodes and wall clock times less than 6 hours. Backfill opportunities vary based on system load and are not guaranteed.

Unbundling Jobs

Prior to topology aware scheduling (TAS), bundling was recommended. With TAS for workloads that don't require large, convex shapes we don't recommend bundling of jobs.

Improving Job Turn-around time

Flexible wall clock time

Flexible wall clock times can improve job turn-around time and utilization. By providing a minimum wall clock time for a job to start as well as a maximum time, the scheduler can start a job early and try to keep the job running by extending the working wall clock time in increments of 30 minutes for as long as the nodes remain available. Wall clock accuracy will be based on the extended wall clock time which is the mintime plus the total of the 30 minute extensions given to the job. The syntax to enable flexible wall clock time in a job is

```
#PBS -l minwclimit=[mintime]  -l walltime=[maxtime]
```

Job Preemption

Job preemption is possible in any queue and is no longer default for the low queue as was announced previously. Job preemption will not occur before 4 hours of wall clock time is reached. Preemption is based on relative job priority and can be used in conjunction with flexible wall clock. Job starts when specified walltime is available. The PBS syntax for preemption specification is

```
#PBS -l flags=preemptee
```

Signal handling (COMING SOON)

To complement the use of the above options it is possible to have a signal sent from the scheduler to the application that can be used to alert the application that it will run out of wallclock time or be preempted soon. At the moment this feature has issues and is not available. We are working with Adaptive on this issue.

Blue Waters Data Sharing Service

- Prototype service for data sharing.
- https://bluewaters.ncsa.illinois.edu/data-sharing
- Sharing of data sets from on-line or Nearline / projects/sciteam/xyz/share using Globus Online or web services (http).
- Metadata includes contact information, description, website, size and count, DOI*.

2015 Blue Waters Symposium

- May 10-14, 2015
- PI meeting for NSF PRAC teams.
- Sunriver Resort in Sunriver, OR

- Recreational opportunities: hiking, biking, fly fishing, rafting, canoeing, rock climbing, …

GPU Hackathon

- At NCSA this week. Look for ORNL and CSCS later this year.

- https://www.olcf.ornl.gov/training-event/2015-gpu-hackathons/

PGI NVIDIA

GPU [Hackathon]

April 20 - 24 NCSA

July 6 -10 CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre

October 19 - 23 OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

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 ), 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

Future Topics?

- Please send us your suggestions on topics for future teleconferences / webinars