Blue Waters
Via
AtlasConnect

OSG All Hands - UCSD/SDSC
March 6-8, 2017

Rob Gardner
University of Chicago

Mark Neubauer
University of Illinois
at Urbana-Champaign

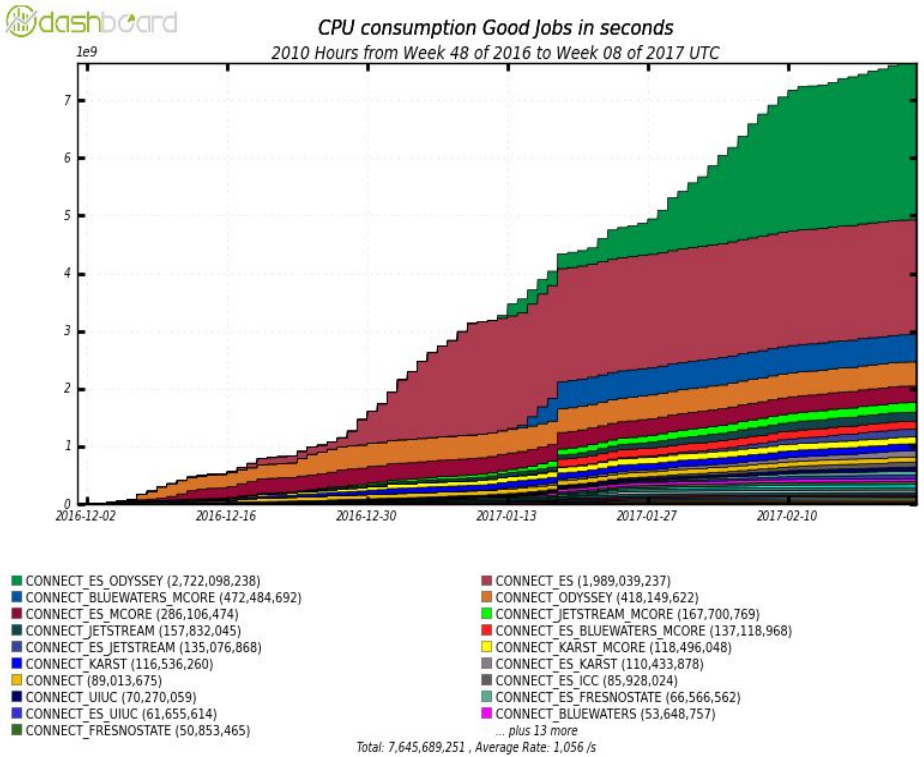
Dave Lesny
University of Illinois
at Urbana-Champaign

Blue Waters via AtlasConnect

- Goal is to make Blue Waters a target site for CONNECT
  - Run all Atlas jobs via PanDA as any other Grid Site
  - No special changes at PanDA side
  - Use stock Atlas Pilot and access to data
  - Use BW as a normal user (no special access)
- Leverage off of current CONNECT and MWT2 services
  - Remote Cluster Connect Facility
  - Remote Cluster Connect AutoPyFactory
  - Stratum-R CVMFS replication service
  - MWT2 Storage Element
  - MWT2 Squids, GUMS, MWT2 CVMFS Repository

CONNECT PanDA queues

- Separate set of PanDA Queues for each CONNECT target site
- Currently nine active target sites
- Last 3 months provided over 2M CPU-Hours



Blue Waters Supercomputer

Blue Waters Supercomputer

- NSF-funded ($200M) Supercomputer @ U. Illinois & NCSA
- Cray XE6/XK7 hybrid machine composed of
  - AMD 6276 "Interlagos" processors
  - NVIDIA GK110 "Kepler" accelerators (XK nodes)
  - Cray Gemini 3D Torus interconnect
- Compute
  - 23k Cray XE nodes
    - ➢ 362k Bulldozer cores; 1.4 PB memory
  - 4.2k Cray XK nodes
    - ➢ 34k Bulldozer cores + 4.2k Kepler accelerators
    - ➢ 135 TB CPU / 25 TB GPU memory
- Storage, I/O
  - Online: 26 PB, aggregate I/O > 1 TB/sec
  - Near-line: 380 PB,1.2 GB cache front-end, 58 GB/s aggregate to tape

National Petascale Computing Facility

Blue Waters Supercomputer

- Policy: "*2% of the available time will be allocated to university projects:* ***(i)*** *faculty whose research and/or education programs would be greatly enhanced by access to Blue Waters and* ***(ii)*** *research and/or education proposals where a commitment of Blue Waters resources will significantly increase the competitiveness of the proposals*"
- Three types of allocations: **Exploratory**, **General**, **Education**
- General proposal (Intended for large-scale research projects)
  - ***March 2016***: submitted a proposal (PI: Neubauer) for ***1M*** node-hours
  - ***May 2016***: Awarded ***50k*** node-hours, ***5TB/50/500TB*** for home/project/scratch managed by Cray's Sonexion Lustre system
    - ❖ "Blue Waters (BW) Project Office has requested an initial award to explore the technical feasibility of your proposal. Your full request is not rejected, but rather deferred for later review pending this initial technical feasibility…"
  - ***August 1, 2016***: Start date for allocation. Met last week with BW technical team and developed plan for initial deployment of services

Blue Waters Supercomputer

Blue Waters Supercomputer

XE Compute Node

AMD 6276 Interlagos Processors	2
Bulldozer Cores*	16
Integer Scheduling Units**	32
Memory / Bulldozer Core	4 GB
Total Node Memory	64 GB
Peak Performance	313.6 GF
Memory Bandwidth	102.4 GB/s

XK Compute Node

AMD 6276 Interlagos Processors	1
Bulldozer Cores*	8
Integer Scheduling Units**	16
Memory / Bulldozer Core	4 GB
Node System Memory	32 GB
GPU Memory	6 GB
Peak CPU Performance	156.8 GF
CPU Memory Bandwidth	51.2 GB/s
CUDA cores	2688
Peak GPU Performance (DP)	1.31 TF
GPU Memory Bandwidth (ECC off)***	250 GB/s

Interconnect

Architecture	3D Torus
Topology	24x24x24
Compute nodes per Gemini	2
Peak Node Injection Bandwidth	9.6 GB/s

Online Storage

Total Usable Storage	26.4 PB
Aggregate I/O Bandwidth	> 1 TB/s

File System	Size (PB)	# of OSTs
home	2.2	144
projects	2.2	144
scratch	22	1440

See: https://bluewaters.ncsa.illinois.edu

Blue Waters Supercomputer

Technical Plans

- BW Scheduler Submission ✓
- Pilot Submission ✓
  - Internal APF/RCCF deployment on edge device
  - External APF/RCCF using gsiftp and OTP (weekly update)
- Software Deployment ✓
  - Shifter container with complete Atlas installation
  - Minimal node with fuse and VM via CVMFS
  - Without fuse, Stratum-R availability of containers ✓
- Deploy RSE ✓
  - Initially use remote StageIn/Out of data via limited NAT
  - Pursue RSE edge device to cache local data
  - Use Globus FTP servers to transfer data in/out
- Configure CONNECT_BLUEWATERS ✓
  - If using RSE, LSM to find data

CONNECT-Blue Waters Development Timeline

- Allocation available to Illinois faculty outside of XSEDE
  - Mark Neubauer, Illinois faculty and MWT2 PI
  - Proposal submitted in March 2016 requesting 1M node hours
  - 50K node hours granted for development in May 2016
  - Allocation starting August 1, 2016 and ending January 31, 2017
- Initial investigation on scope of work began June 2016
- Blue Waters accounts created August 1, 2016
  - Using OTP RSA SecurID, create proxies valid for 11 days
  - Use GSISSH to log into BW Login nodes
  - Learn the layout of BW, PBS scheduler, aprun, etc

Blue Waters Devel Timeline ...

- Modify Connect to support GSISSH, BW PBS, etc
  - First CONNECT Glidein ran on August 31, 2016

```
################################################################################
#####                                                                      #####
#####          Job is running within a Remote Cluster Connect Factory      #####
#####                                                                      #####
##### Date: Wed Aug 31 09:25:04 CDT 2016                                   #####
##### RCCF: bluewaters                                                     #####
##### User: ddl                                                            #####
##### Host: nid00176                                                       #####
#####                                                                      #####
################################################################################

nid00176

################################################################################
################################## Job Complete ################################
################################################################################
```

Blue Waters Devel Timeline ....

- Initial Shifter/docker work revealed bugs in Shifter
  - Problem with network access
  - Cray did not resolve the problem until December 2016
- Lack of fuse created CVMFS access problems
  - Stratum-R GO replication became operational in December 2016
- First Shifter/Docker/CVMFS job ran December 15, 2016
- Local Site Mover (LSM) from MWT2 adopted for BW
  - Used pCache to reduce the number of files in stagein
  - Records transfer data to Elastic Search

Blue Waters Devel Timeline .....

- First PanDA pilot ran January 4, 2017
- First Hammer Cloud job ran January 6, 2017
- Initial PanDA jobs failed
  - Some issues with GO replication (fixed)
  - Some missing RPMs in Docker Image (fixed)
- Began a production run from January 13-17, 2017
  - Standard and Event Service production jobs
  - Used up to 48 glideins, with 16 nodes per glidein, 16 cores per node
  - 12K cores used all 50K hours in 1 week

CONNECTing Blue Waters to PanDA

- Platform
  - BW OS is Cray Linux Environment (CLE)/SUSE Linux 3.x
  - Container support via Shifter/Docker
  - Built Image on Docker Hub:
    - SL6 with all need utilities
    - HEP OSlibs meta rpm
    - Many other compatibility rpms not listed in HEP OSlibs
- ATLAS releases and other repositories
  - "Stratum-R": full disk replica of ATLAS releases sync'd daily
  - Replication to local disk at Blue Waters via Globus Online
  - Symlink in Docker image points /cvmfs to replicas on disk

Blue Waters Storage

- Very Large Lustre Scratch disk, 14P total
  - Quota: 500TB and 100M files
  - Holds Stratum-R files, pCache and scratch space for all jobs
  - Needed to get quota raised to 200M files (using 150M)
  - Needed whitelisting of space to avoid 30 day cleanup
  - Globus Online endpoint: ncsa#Bluewaters

Blue Waters CVMFS Replicas

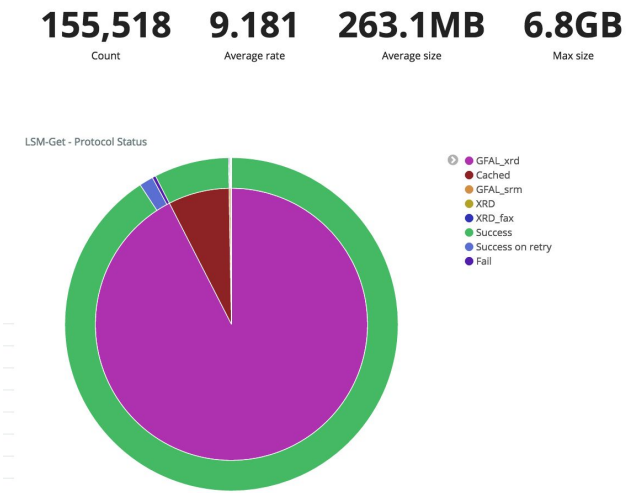
- CVMFS client is not available (fuse not allowed)
  - Solution: replicate all needed repositories with Stratum-R
- Stratum-R: Provides local disk copies of all repositories
  - Servers at UChicago, Illinois, Indiana
  - Serves files via “rsync” (over x509 protected stunnel) and Globus Online
- Use Globus Online to copy/sync repositories to BW
  - 11 day proxy created with OTP on GO to access the BW endpoint
  - GO does not replicate symlinks or permissions
  - After GO transfer completes, a symlink creation pushed via GSISSH
  - Automatically scheduled via crons from the Connect Facility

Blue Waters Glideins

- Local Scheduler: PBS
  - Requires multiple nodes reservation per job: Currently requesting 16
  - Each node 32 cores, 64 GB, no swap => use only 16 cores to avoid OOM
- GSISSH based Glidein (Connect Factory)
  - Authorization: One Time Password creates proxy good for 11 days
  - Glidein requests 16 nodes and runs one HTCondor overlay per node
  - Requests Shifter usage with a Docker Image from Docker Hub
  - HTC overlay creates 16 partitionable slots with 16 cores per slot
  - Connect AutoPyFactory injects pilots into these slots which run on BW
  - Glidein life is 48 hours and will run consecutive Atlas jobs in the slots
  - Need a mix of standard and Event Service jobs to minimise idle cores

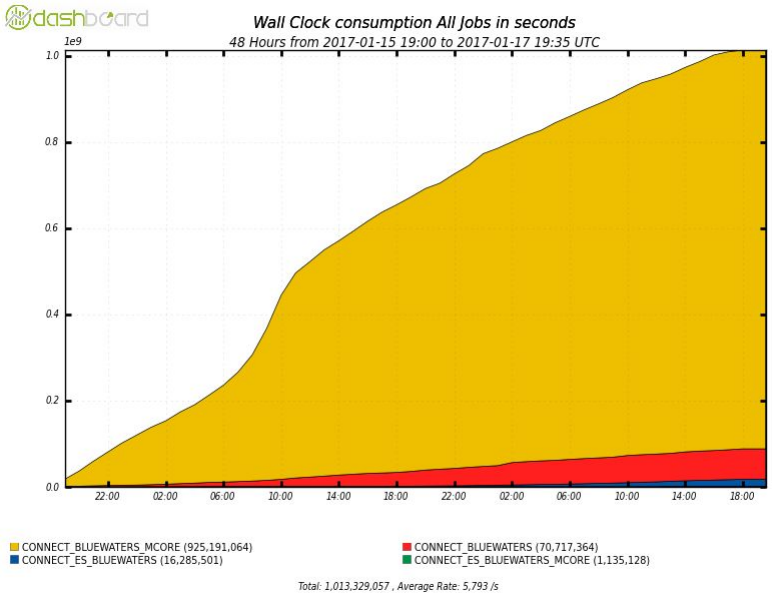
Blue Waters Data Transfer

- BW nodes have limited access to WAN
  - Number of ports available to outside is restriction
  - Ports needed for HTC overlay and stagein/out of data
- "Local Site Mover" (`lsm-get,lsm-put`)
  - Using MWT2 SE as storage endpoint
  - Transfer utility is gfal-copy, root://, srm:// or Xrootd; retries with simple backoff and protocols change on failure; pCache (WN cache) used by lsm-get to help reduce stagein of duplicate files
  - I/O metrics logged to Elastic Search

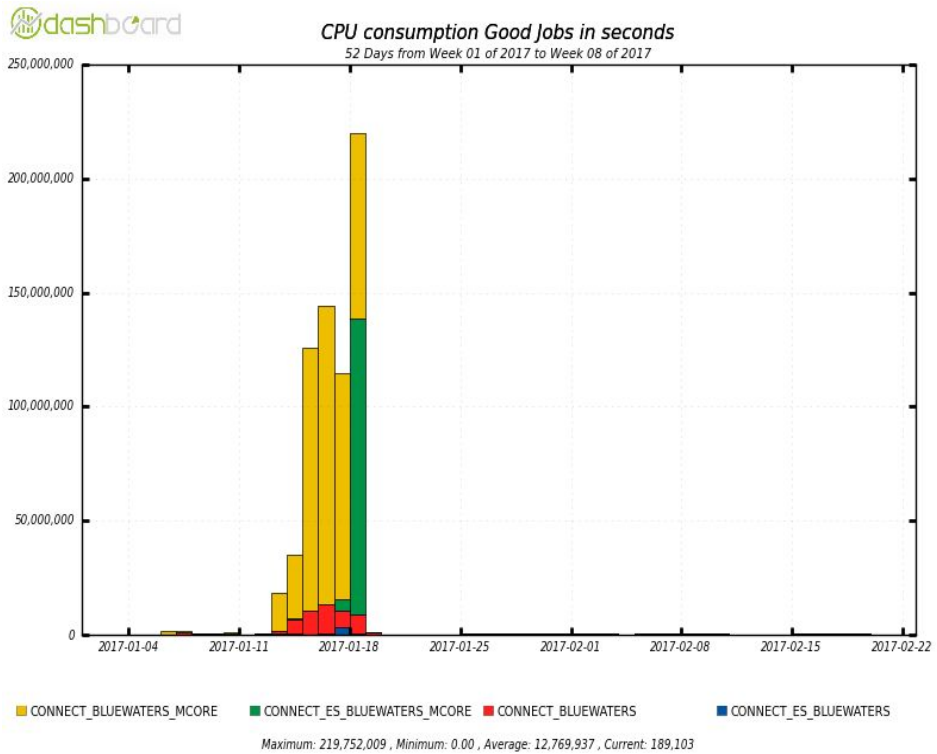


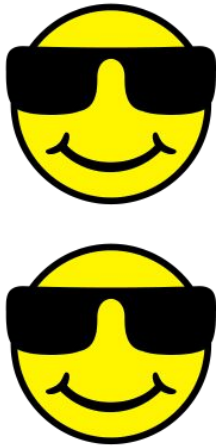
Panda Queues setup

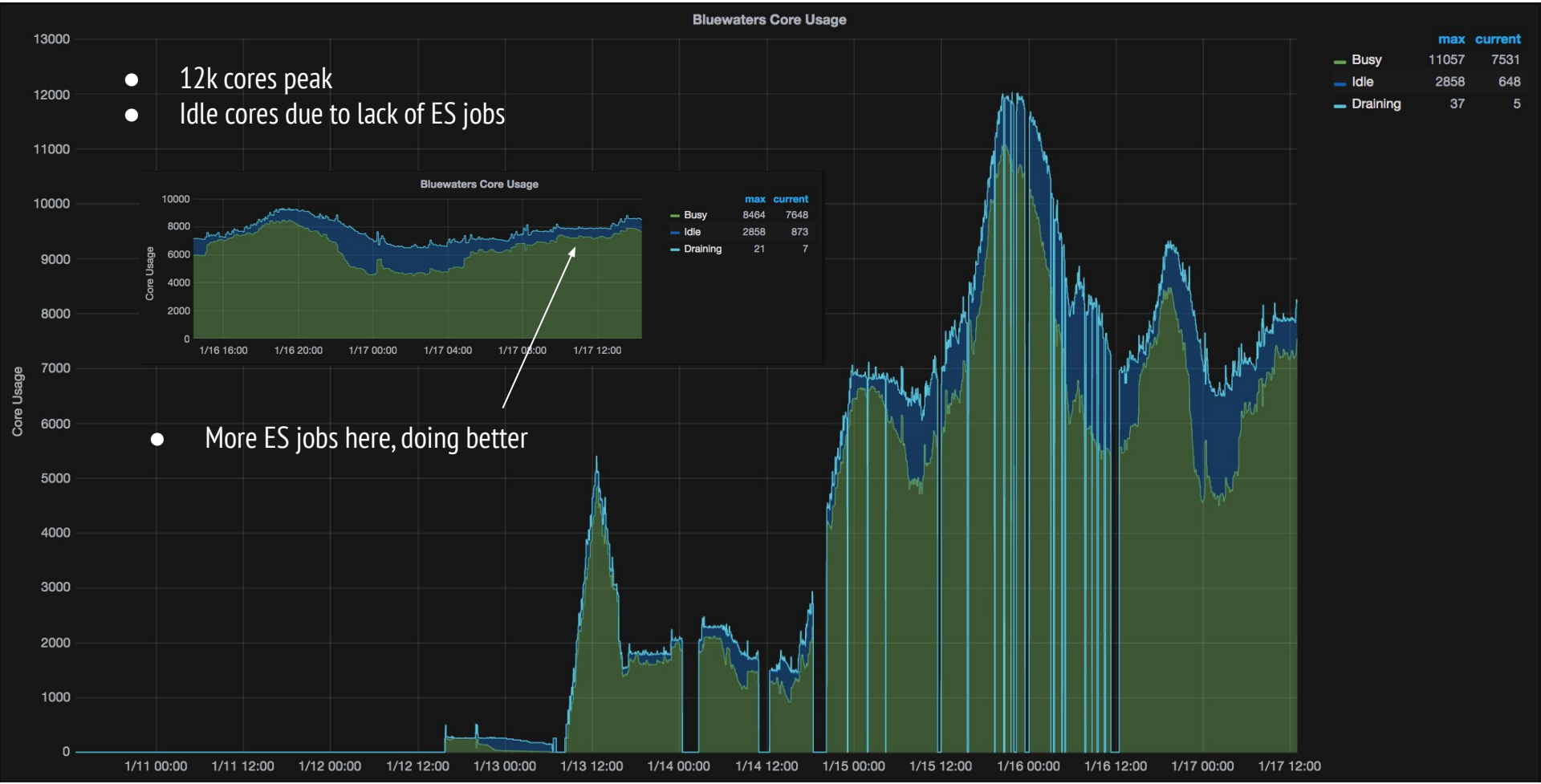
- 4 Panda (**general**) Production Queues
  - CONNECT_BLUEWATERS
  - CONNECT_BLUEWATERS_MCORE
  - CONNECT_ES_BLUEWATERS
  - CONNECT_ES_BLUEWATERS_MCORE
  - **No restriction on tasks or releases**
- Each Q configured for BW
  - LSM transfer
  - Standard: 36H guaranteed
  - ES: 4H guaranteed up to 36H max
  - 4H jobs fill in scheduling holes



PanDA CPU provided by Blue Waters






Bluewaters Core Usage
max current
Busy 11057 7531
Idle 2858 648
Draining 37 5
Core Usage
13000
12000
11000
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1/11 00:00
1/11 12:00
1/12 00:00
1/12 12:00
1/13 00:00
1/13 12:00
1/14 00:00
1/14 12:00
1/15 00:00
1/15 12:00
1/16 00:00
1/16 12:00
1/17 00:00
1/17 12:00
12k cores peak
Idle cores due to lack of ES jobs
Bluewaters Core Usage
max current
Busy 8464 7648
Idle 2858 873
Draining 21 7
Core Usage
10000
8000
6000
4000
2000
0
1/16 16:00
1/16 20:00
1/17 00:00
1/17 04:00
1/17 08:00
1/17 12:00
More ES jobs here, doing better

Future work

- Submit a second proposal by March 15, 2017
  - Request 1M node hours
  - Use Illinois allocation assigned to Mark Neubauer
- Change Glideins to use backfill and preemption
  - Backfill are small node allocations requesting under 6 hours
  - Preemption allows jobs to be killed at any time
  - Uses ½ of the node hours
  - Perfect for Event Service jobs
- Should get nearly 1 year run with 12K cores