Use of NSF Supercomputers

Open Science Grid

Rob Gardner, University of Chicago

OSG Council, Indianapolis, October 3, 2017

Acknoweledgements !!

Frank Wuerthwein
Edgar Fajardo
Mark Neubauer, Dave Lesny & Peter Onyisi
Mats Rynge
Rob Quick

Goal

Standardize ”the inteface” to NSF HPC resources – add them to resource pools used by OSG engaged communities

Identity & doors .. CEs .. Glideins .. Software .. Data .. Network .. Workflow .. Operations ..

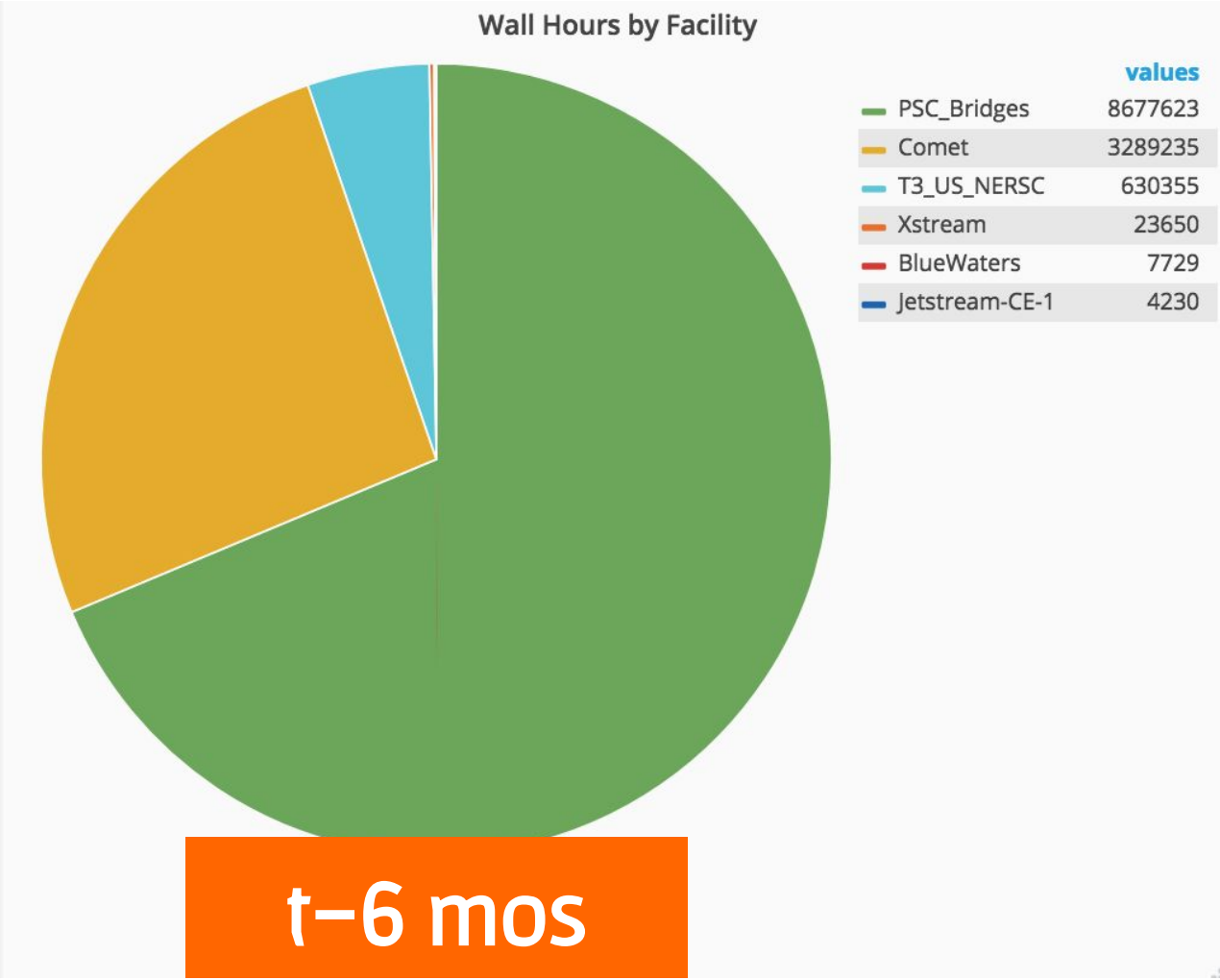
OSG –style ”Science Gateways” c.f. SGCI

General Approach

- Use what is offerred
  - login, MFA, scheduler, platform OS, network
- Minimize footprint at the resource
  - Do as much as possible in OSG managed edge services
- Expand resource pools with NSF HPC transparently without extra work by the VO

Outline for the remainder...

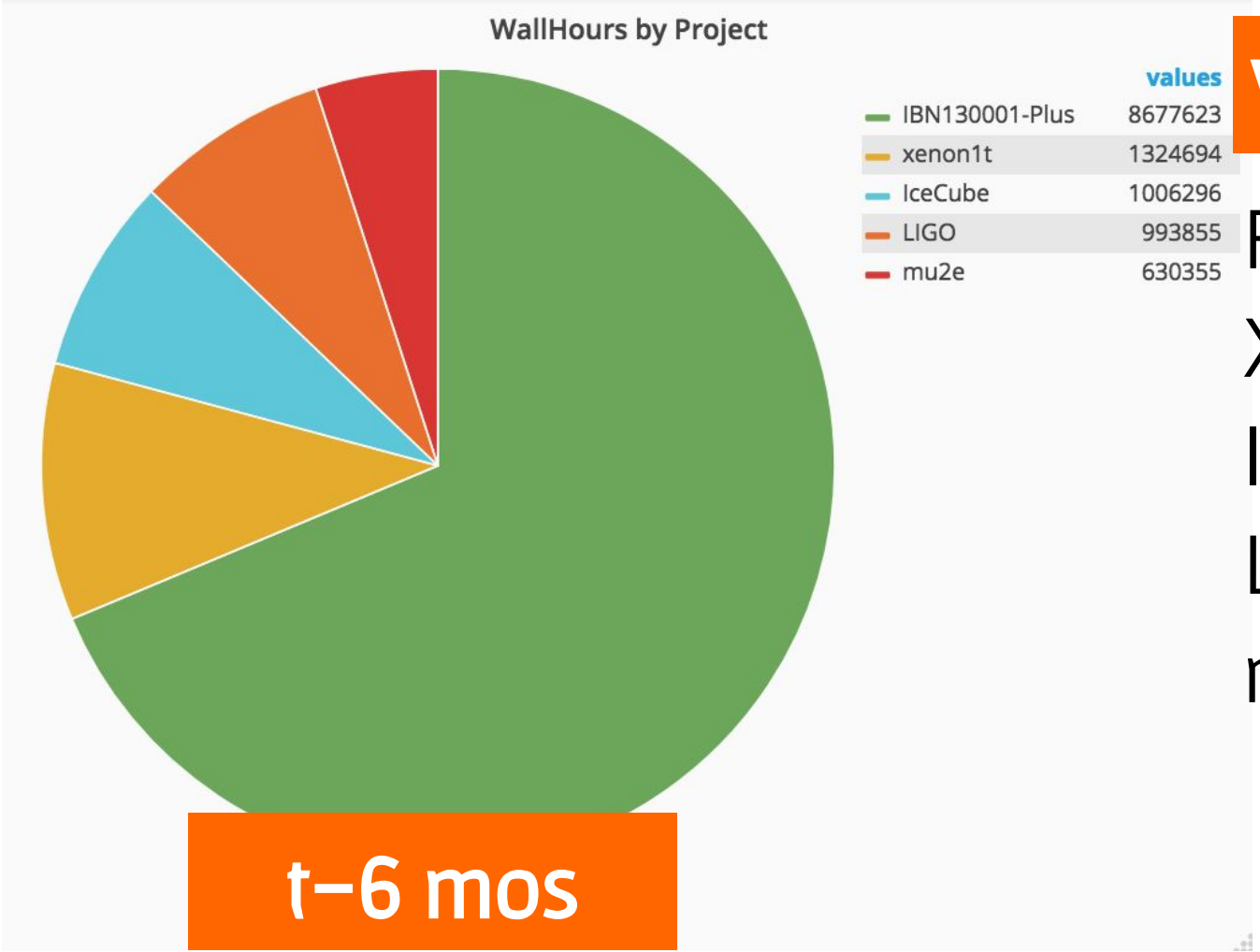
- Survey of efforts
- Common challenges
- Next steps



t-6 mos

Facilities

Bridges
Comet
Cori
Xstream
Blue Waters
Jetstream



t-6 mos

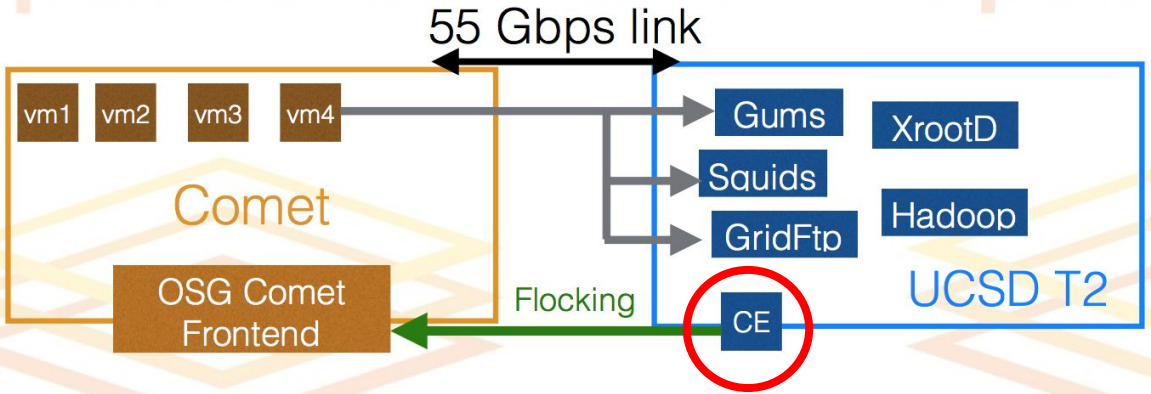
VOs

FuncNeuro
XENON1T
IceCube
LIGO
mu2e

Edgar Fajardo

Where does OSG kick in?

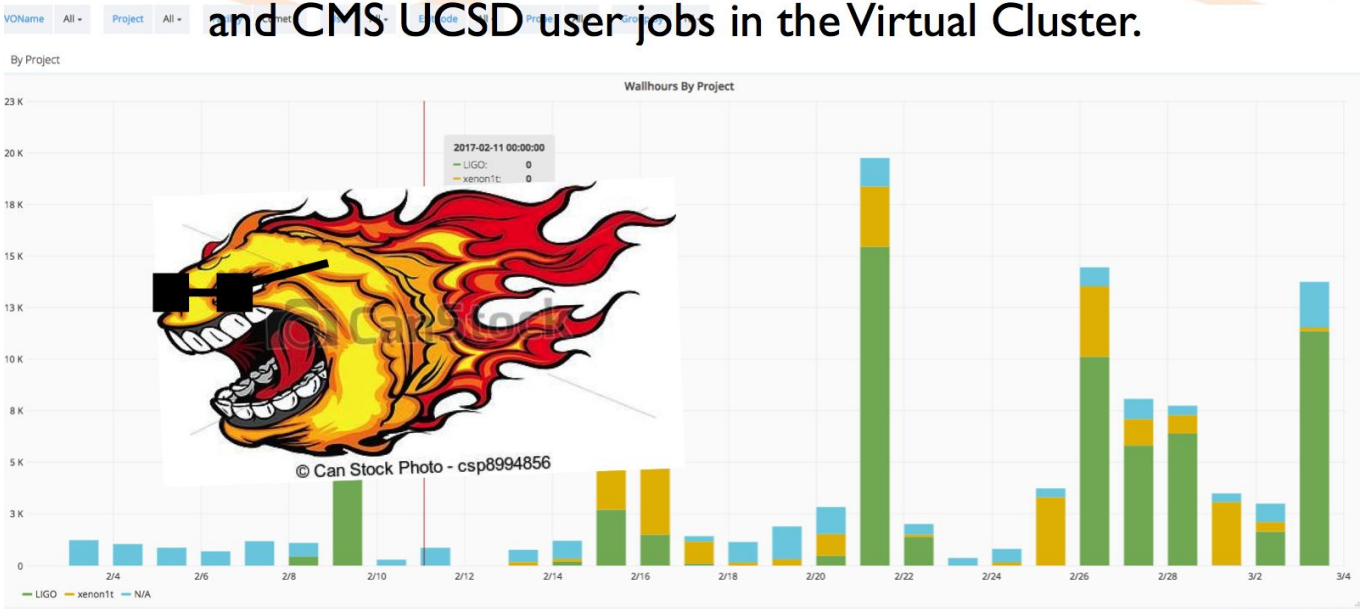
Glideins can get into Comet using the already existing UCSD T2 grid infrastructure



Edgar Fajardo

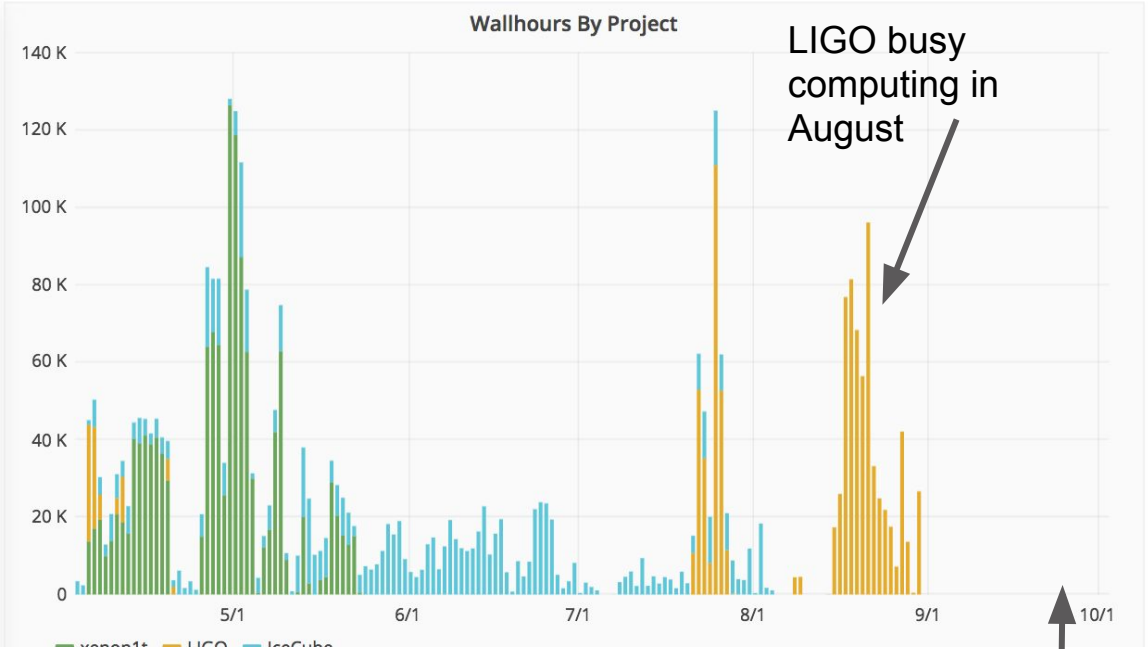
Achievements

- Successfully ran LIGO, Xennon1T, CMS Production and CMS UCSD user jobs in the Virtual Cluster.

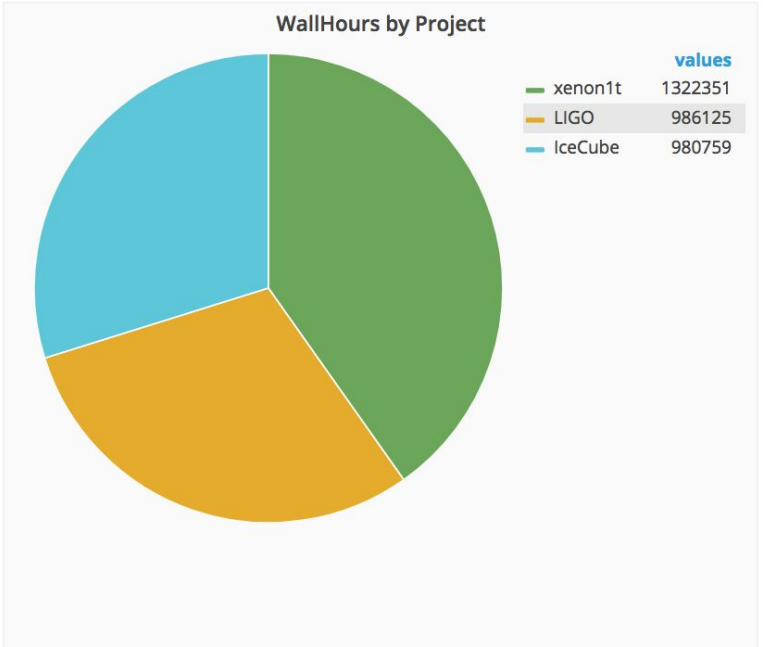


Comet update

By Project



LIGO busy computing in August



Sep 27 latest LIGO result announced

Data Access

- The most standard integration is done for Comet. There we have every node WAN accessible via IPv6, and reached via a regular OSG-CE. We even support the use of StashCache there, but I'm not sure it was used yet by the apps that have run there. CVMFS is of course also available on Comet.
  - I think both LIGO and xenon1t pull in data as needed from the worker nodes. For xenon1t this is done via gridftp, for LIGO via xrdcp, as far as I know.
- This is accomplished at Comet via its special virtual cluster interface. I.e. we effectively have root and can do whatever we want.
- BlueWaters and NERSC also offer the OASIS application environments, but not via CVMFS. BlueWaters for sure does a regular rsynch onto the parallel filesystem. Not 100% sure for NERSC.
- Jetstream offers OASIS, I think, but I'm not sure how.

Challenges: Software Distribution

Stampede

- Stratum-R delivers software to Stampede
- Providing support for all the major OSG VOs and the OSG modules

```
login5.stampede(326)$ ll
total 80
drwxrwxr-x  5 usatlas G-815132  4096 Mar  5  2012 atlas.cern.ch
drwxrwxr-x  5 usatlas G-815132  4096 Jan 18  2012 atlas-condb.cern.ch
drwxrwxr-x  9 usatlas G-815132  4096 Jan 13  2014 cernvm-prod.cern.ch
drwxrwxr-x 57 usatlas G-815132 12288 Oct  1 03:44 cms.cern.ch
drwxrwxr-x  7 usatlas G-815132  4096 Aug 12  2014 fermilab.opensciencegrid.org
drwxrwxr-x 12 usatlas G-815132  4096 Mar 31  2014 geant4.cern.ch
drwxrwxr-x 44 usatlas G-815132  4096 Oct  1 14:55 grid.cern.ch
drwxrwxr-x 12 usatlas G-815132  4096 Apr 24  2014 icecube.opensciencegrid.org
drwxrwxr-x  5 usatlas G-815132  4096 Feb 19  2015 minos.opensciencegrid.org
drwxrwxr-x 14 usatlas G-815132  4096 Feb 19  2015 nova.opensciencegrid.org
drwxrwxr-x 32 usatlas G-815132  4096 May 13  2015 oasis.opensciencegrid.org
drwxrwxr-x  8 usatlas G-815132  4096 Aug 18  2015 osg.mwt2.org
drwxrwxr-x  5 usatlas G-815132  4096 Mar 25  2011 sft.cern.ch
drwxrwxr-x 28 usatlas G-815132  4096 Feb  4  2017 singularity.opensciencegrid.org
drwxrwxr-x  7 usatlas G-815132  4096 Oct 31  2016 snoplus.egi.eu
drwxrwxr-x  8 usatlas G-815132  4096 Sep 12  2016 spt.opensciencegrid.org
drwxrwxr-x  5 usatlas G-815132  4096 Mar 29  2017 veritas.opensciencegrid.org
drwxrwxr-x  6 usatlas G-815132  4096 Sep 16  2016 xenon.opensciencegrid.org
```

those are all the repos being replicated to stampede

Challenges: Software Distribution

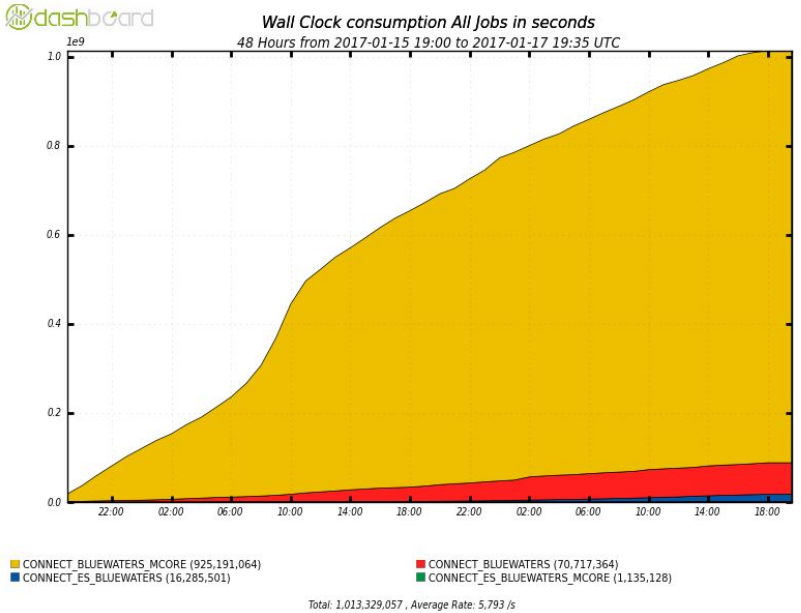
- Stratum-R delivers software to Bluewaters
- IceCube recently added
- Include compat libs needed by LHC exps

```
ddl@h2ologin2:~/cvmfs> ll
total 32
drwxrwxr-x  5 ddl ILL_bafz 4096 Nov 21  2016 atlas.cern.ch
drwxrwxr-x  5 ddl ILL_bafz 4096 Aug 18 09:01 atlas-condb.cern.ch
drwxrwxr-x 12 ddl ILL_bafz 4096 Jul 28 02:17 geant4.cern.ch
drwxrwxr-x 44 ddl ILL_bafz 4096 Jul 20 00:43 grid.cern.ch
drwxrwxr-x 11 ddl ILL_bafz 4096 Sep  1 15:56 icecube.opensciencegrid.org
drwxrwxr-x 32 ddl ILL_bafz 4096 Jan 15  2017 oasis.opensciencegrid.org
drwxrwxr-x  8 ddl ILL_bafz 4096 Jul 23 09:16 osg.mwt2.org
drwxrwxr-x  5 ddl ILL_bafz 4096 Dec  9  2016 sft.cern.ch
```

PanDA Queues setup

Gardner, Lesny, Neubauer

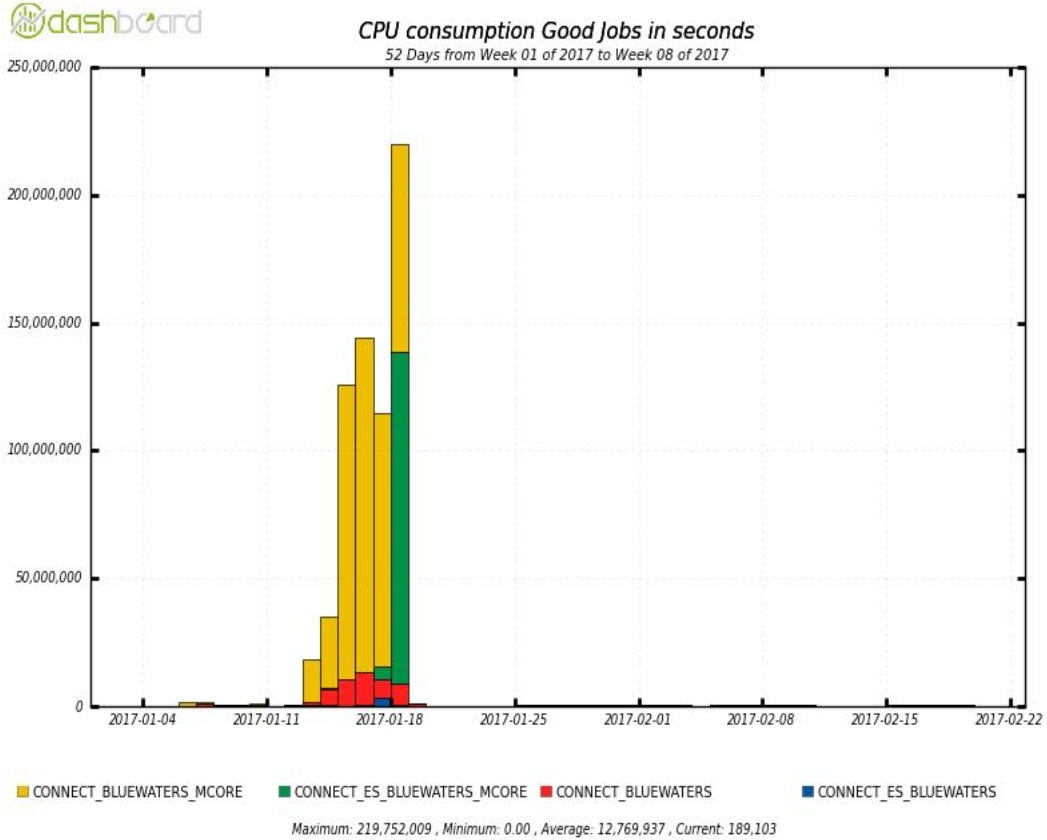
- 4 Panda (**general**) Production Queues
  - CONNECT_BLUEWATERS
  - CONNECT_BLUEWATERS_MCORE
  - CONNECT_ES_BLUEWATERS
  - CONNECT_ES_BLUEWATERS_MCORE
  - **No restriction on tasks or releases**
- Each queue configured for BW
  - LSM transfer
  - Standard: 36H guaranteed
  - ES: 4H guaranteed up to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



PanDA CPU provided by Blue Waters

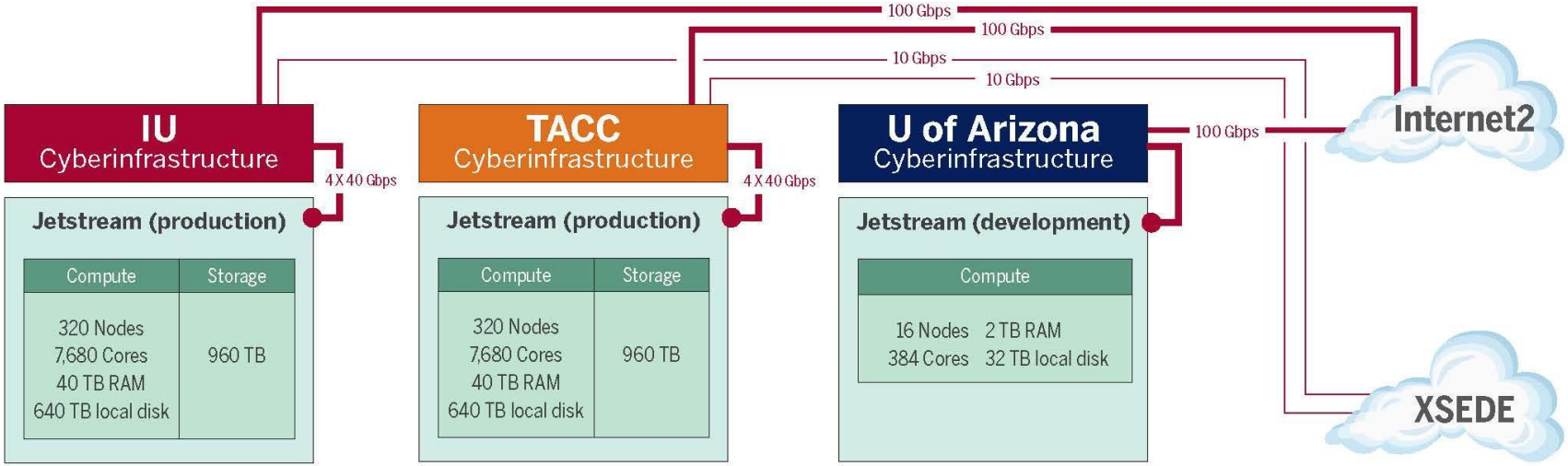
Blue Waters

Gardner, Lesny, Neubauer



ATLAS EXPERIMENT

Jetstream System Overview



jetstream
https://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604

Quick

OSG and Jetstream

- Running on a few cores consistently Sinc
  May.
- Most effort has been in how to how to
  efficiently expand/contract the size of th
  pool.
- This is very close, two tasks left.
  - Update webhooks code to provide unique instance names.
  - Plug in webhooks to scale to the number of instances based on idle nodes.

Jetstream
https://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604

Quick

Edgar Fajardo

OSG on Jetstream

Initial configuration attempts to follow **standard OSG model**.

- Glidein submission to an HTCondor-CE
- Local HTCondor Pool
  - Schedd + Central Manager running on same VM as CE
- Other supporting services: Squid, etc.

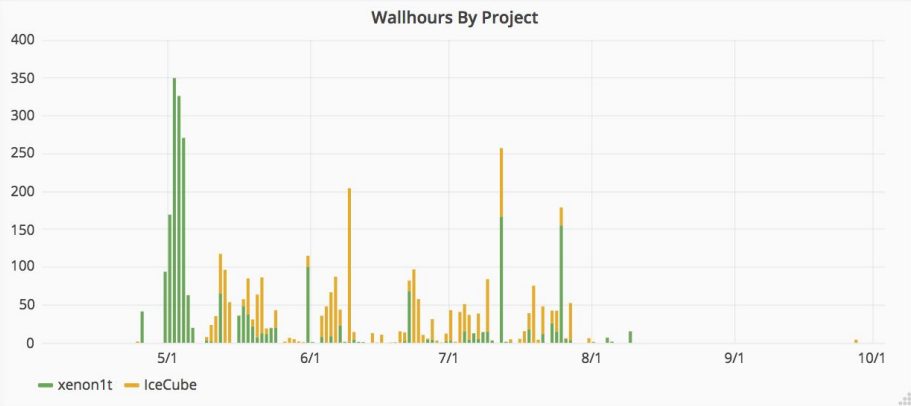
Developing **bootstrapping script(s)** to automate image builds and configuration, which should help facilitate long-term/shared management of site.

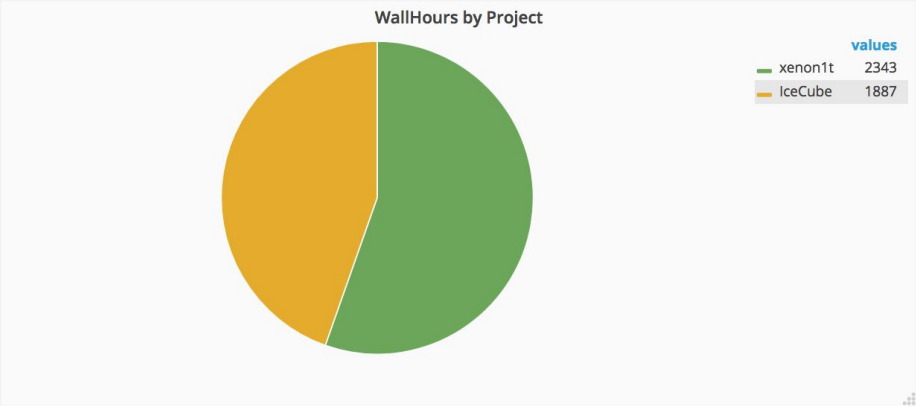
Some **cloud-related configuration issues**:

- Public/private network interfaces.
- Multiple public/private hostnames per network interface; e.g., Openstack's Nova (compute) and Neutron (networking) services do not share consistent hostnames by default.

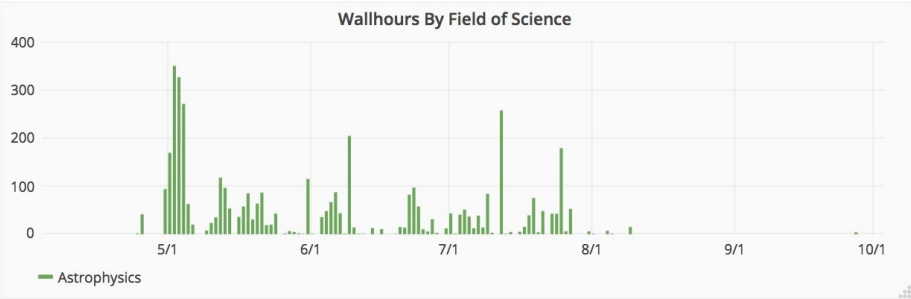
Unknown: How to advertise size of available pool?

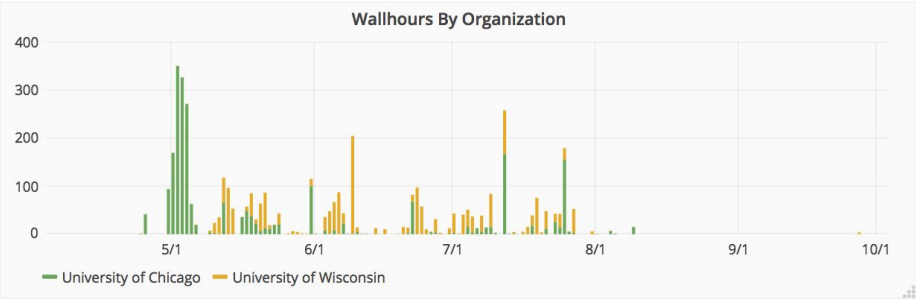
By Project





By Field Of Science





JetStream via CONNECT

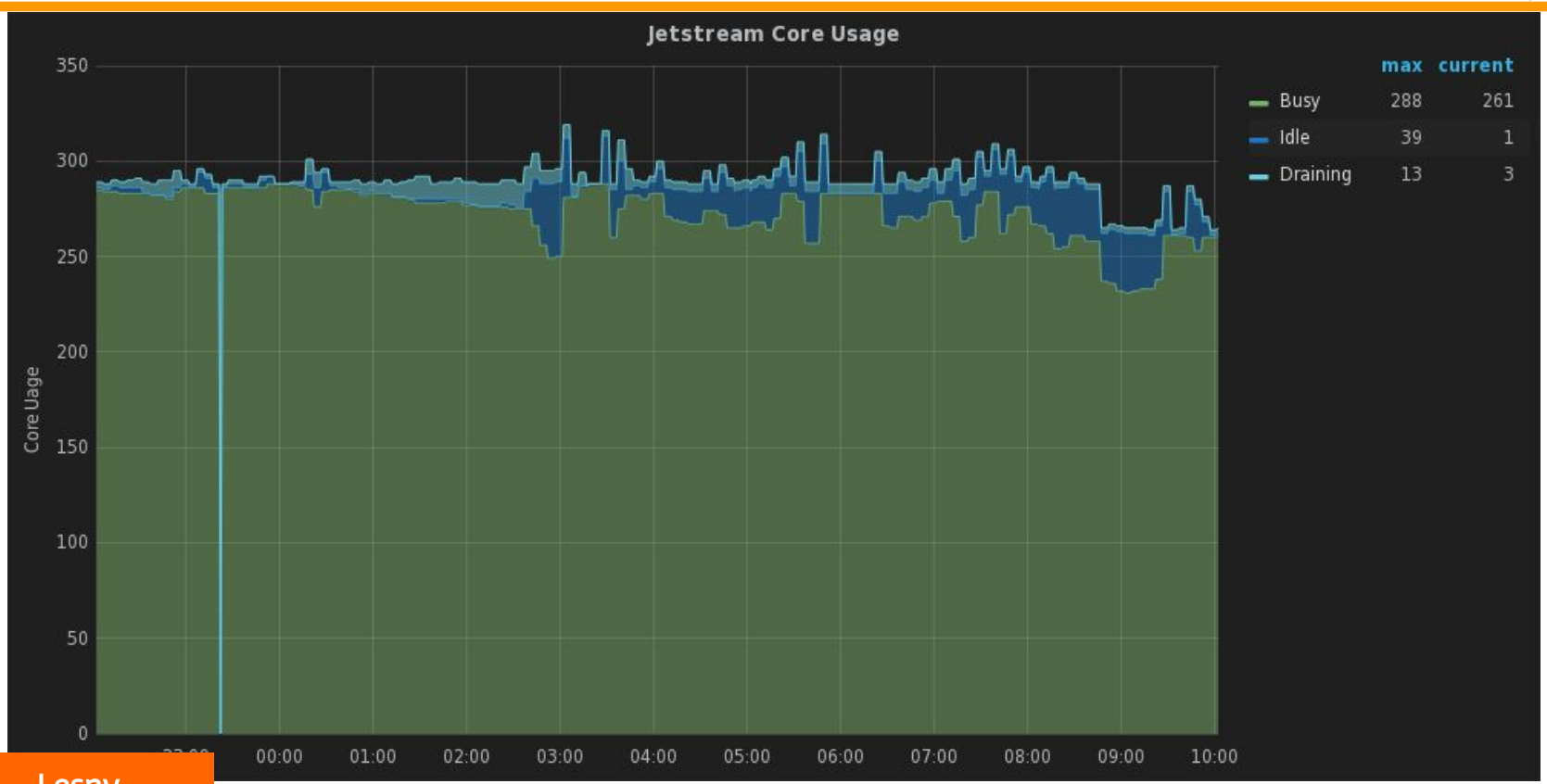
Lesny, Onyisi

- Jetstream is just another target site for CONNECT
  - VMs reside in a Condor pool with SCHEDD on utatlas tier3 login node
- CONNECT submits SSH Glideins into this pool
  - Each glidein requests the whole VM (24 cores, 48GB memory)
  - Allows Connect to do its own scheduling, matchmaking, classads
  - PortableCVMFS brought into the VM (which has fuse)
  - Docker image has all other Atlas dependencies
- PanDA access via CONNECT AutoPyFactory
  - CONNECT_JETSTREAM, CONNECT_JETSTREAM_MCORE
  - CONNECT_ES_JETSTREAM, CONNECT_ES_JETSTREAM_MCORE

JetStream Cores via CONNECT

Jetstream

Lesny, Onyisi



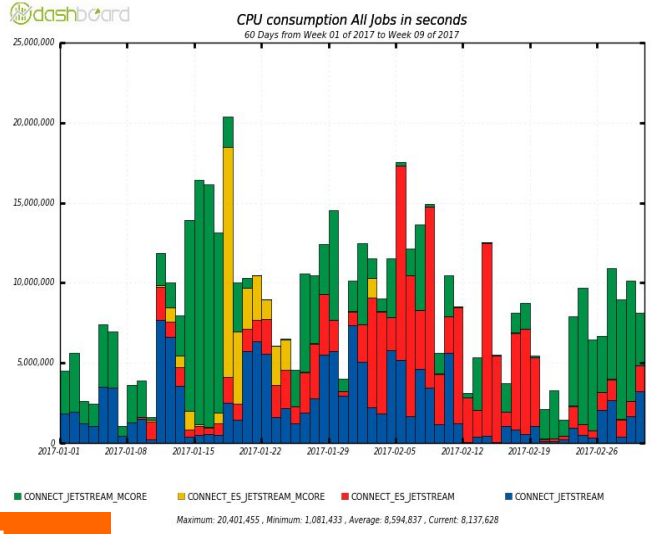
Lesny

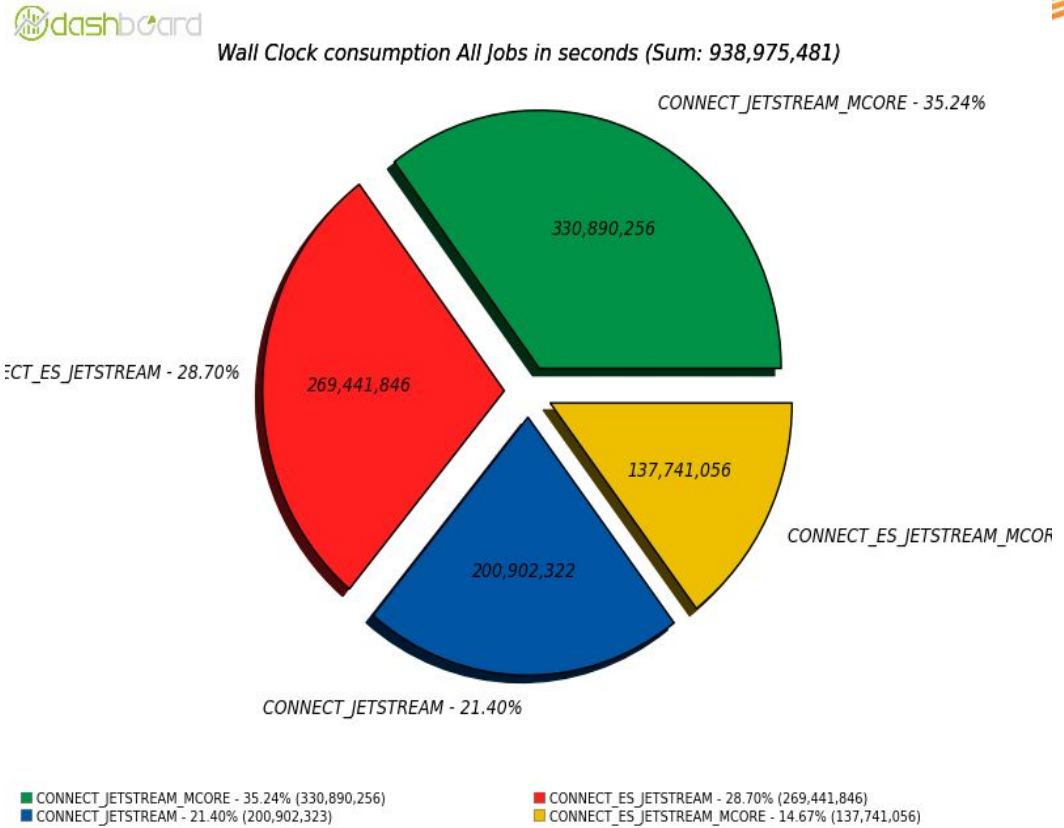
JetStream PanDA (January 1, 2017 to March 6, 2017)

Jetstream

Lesny, Onyisi

- Total: 261K cpus hours
- Using 12 24-core VMs
- Evenly split over all Qs





Neubauer

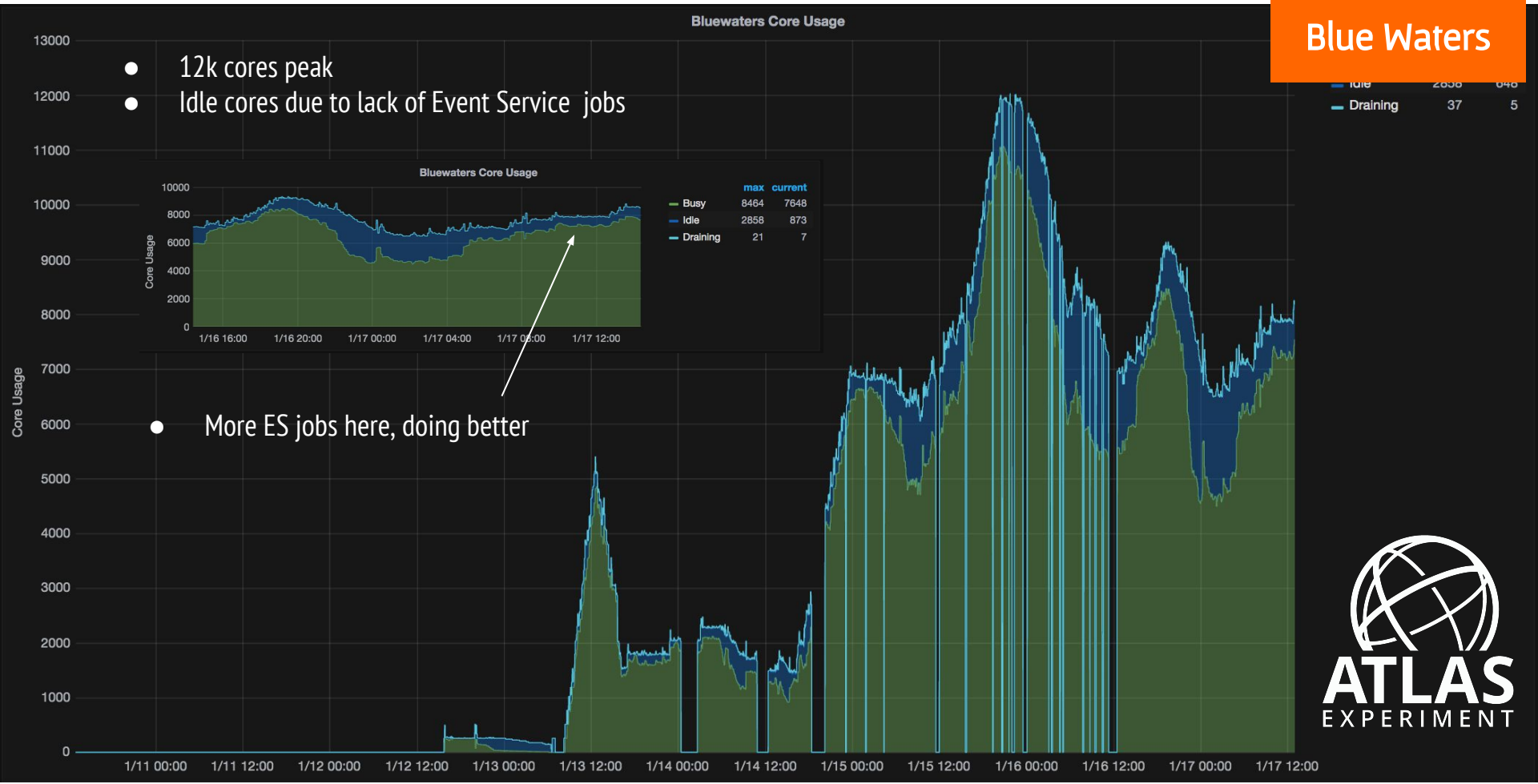
Summary

- Our goal is to standardize interfaces to NSF supercomputers & OSG HTC for existing VOs
  - Overlay scheduling (using the OSG CE)
    - Hosted CEs
  - Software delivery (either containers or CVMFS modules)
  - Data delivery (StashCache)
- Near term: focus on Stampede2
  - Discussing with TACC a 2FA equivalent (key+subnet)
  - Hosted CE w/ extensions to individual logins for accounting for hosted HTCondorCE-Bosco

extra

some details

Blue Waters



ATLAS EXPERIMENT

Blue Waters Gliedeins

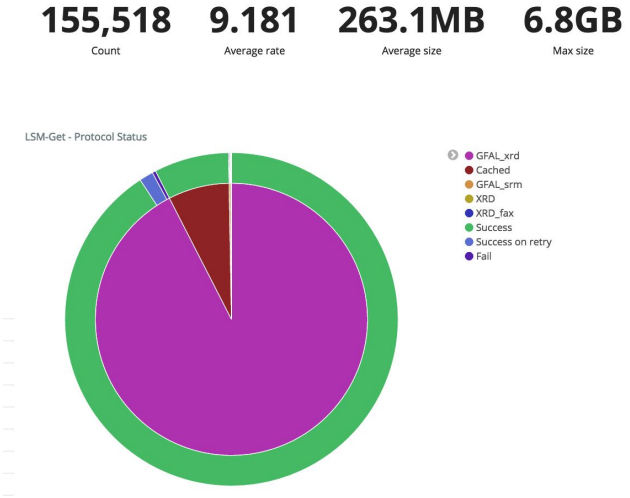
- Local Scheduler: PBS
  - Requires multiple nodes reservation per job: Currently requesting 16
  - Each node 32 cores, 64 GB, no swap => use only 16 cores to avoid OOM
- GSISSH based Glidein (Connect Factory)
  - Authorization: One Time Password creates proxy good for 11 days
  - Glidein requests 16 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

Blue Waters

Gardner, Lesny, Neubauer

- 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



Blue Waters Gliedeins

- 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