US ATLAS Distributed Computing Facility Report

Status, feedback, development

Rob Gardner • University of Chicago

ATLAS Sites Jamboree, January 27, 2016

Outline

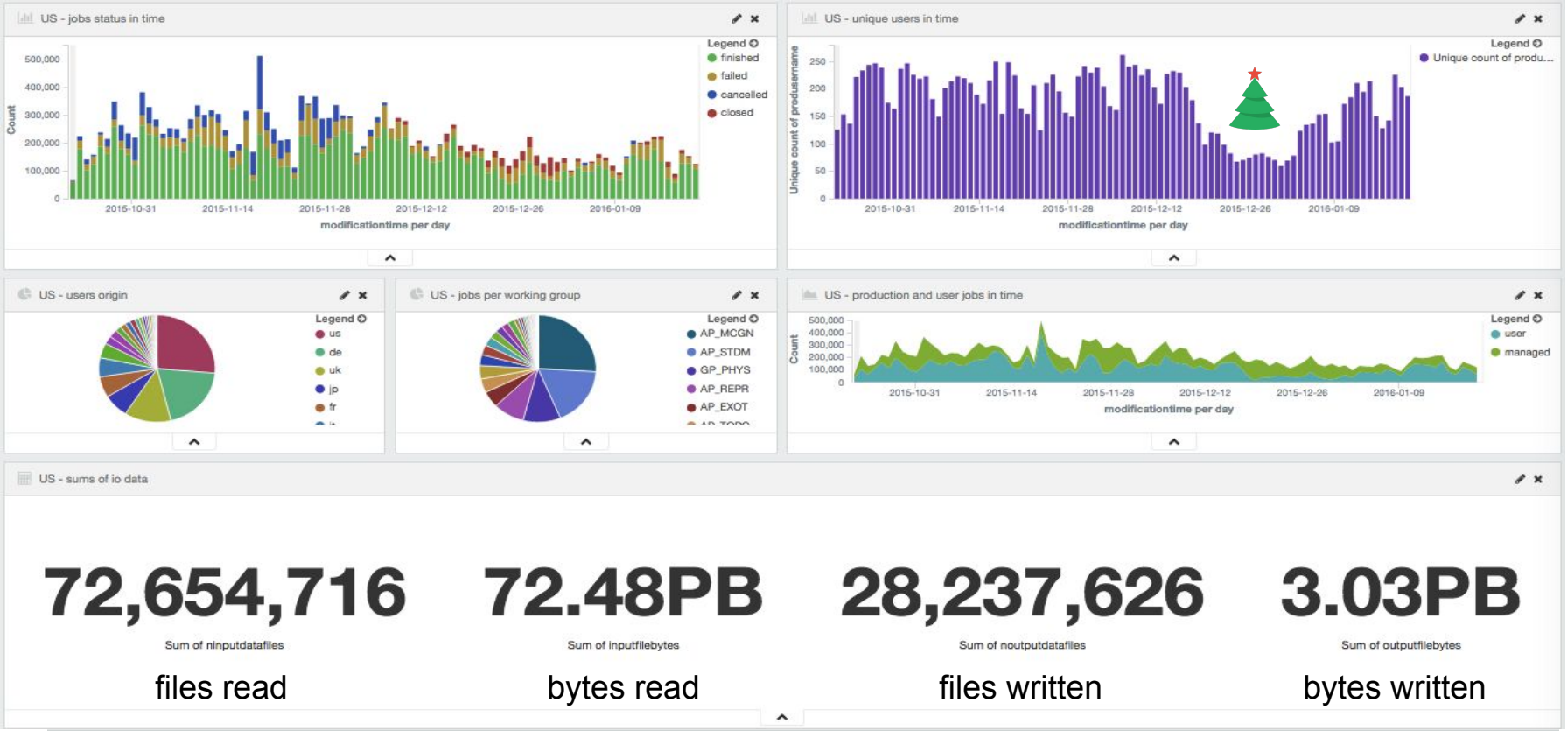
- Current capacity
- Some operational analytics
- Public cloud efforts
- LHCONE peering status
- Operational issues & feedback
- Facility R&D

US ATLAS Capacity

Table 1: Installed capacities as of Dec 2015					Comparison to 2015 and 2016 Pledges					
Center	Total CPU Installed (HS06)	Job slots installed (single logical threads)	Total Disk Installed (TB)	Local Group Disk allocated (TB)	Beyond Pledge CPU HS06 (2015)	Beyond Pledge Job Slots (2015)	Beyond Pledge Disk TB (2015)	Beyond Pledge CPU HS06 (2016)	Beyond Pledge Job Slots (2016)	Beyond Pledge Disk TB (2016)
Tier1	126,381	12948	11000	297	16,381	1,678	2000	-1,619	-166	0
AGLT2	68,506	7020	3712	265	46,506	4,766	1312	43,506	4,458	412
MWT2	123,048	12548	5028	518	90,048	9,183	1428	85,048	8,673	128
NET2	49,710	5096	2600	357	27,710	2,841	200	24,710	2,533	-700
SWT2	62,039	6498	3530	129.6	40,039	4,194	1130	37,039	3,879	230
WT2	42,486	3408	3074	175	20,486	1,643	674	17,486	1,403	-226
USATLAS FACILITY	472,169	47518	28,944	1741.6	241,169	24,304	6,744	206,169	20,780	(156)
USATLAS TIER2	345,788	34570	17,944	1444.6	224,788	22,626	4,744	207,788	20,946	(156)

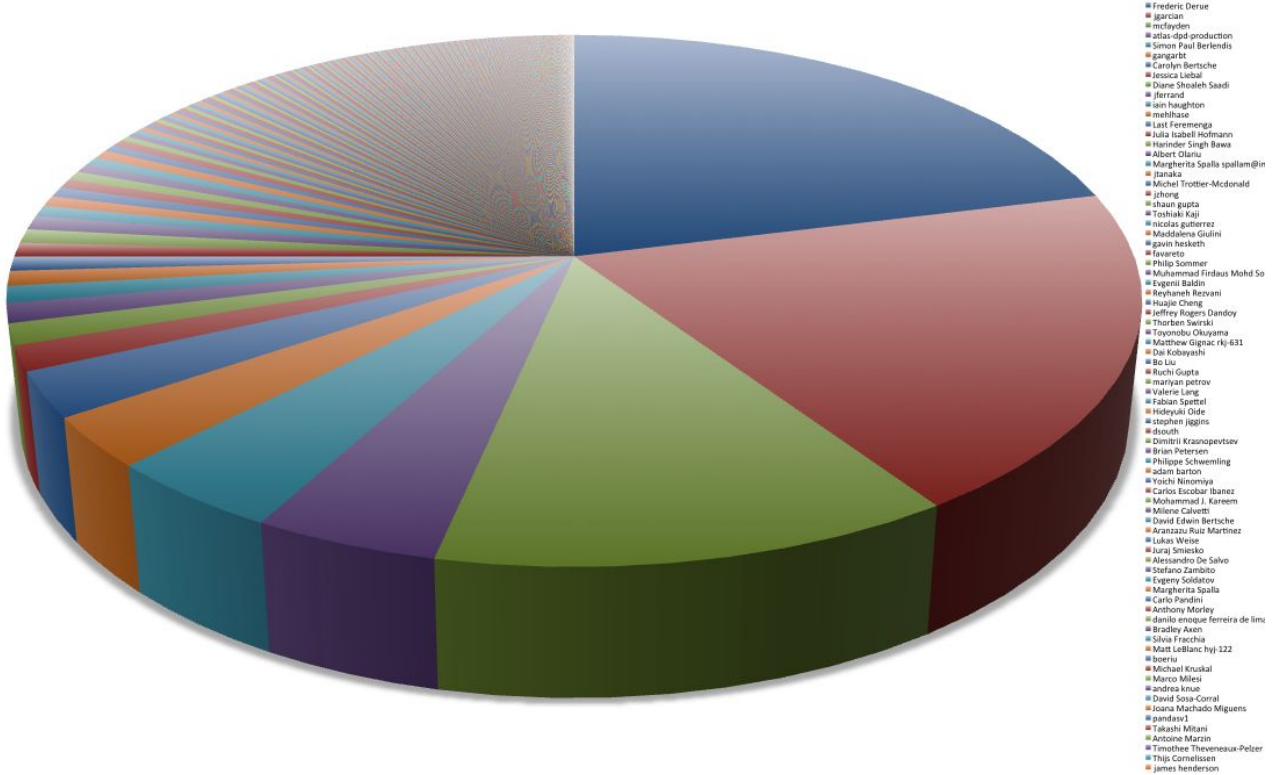
- Additional storage and CPU from FY15 funds coming soon
- Well above pledge for CPU and will be comfortable above for storage (~5PB above after FY16 purchases)
- Planning FY16 purchases (& retirements) now

Last 90 days US production at a glance



ATLAS Users at a US T2 ANALY site

90 days
777 unique users



Running ATLAS Jobs in the Cloud – At Scale and at Low Cost

Joint Project (Amazon AWS, BNL/ATLAS, ESnet) to investigate feasibility (technically and financially) of large-scale usage of commercial cloud resources

- AWS: Provide expertise and guidance to BNL/ATLAS, credits for AWS service investigation and scale-out tests
- BNL: Provide ATLAS-compatible VM image and provisioning infrastructure, incl. demand-driven (i.e. via PanDA server API) VM lifecycle management (create, retire, terminate)
- ESnet: Provide high-performance (up to 100G) network connectivity between AWS facilities and sites connected to R&E networks (general peering and AWS Direct Connect)
- Has made AWS partially waive Egress traffic fee (at level of 15% of total bill)

Cost of AWS/EC2 spot slightly lower than dedicated farm resources at BNL

Running ATLAS Jobs in the Cloud – Experience

“Unlimited” spot resources available on demand

- Had no issues ramping up quickly from 1000 to ~6000 8-core instances (several instance types) in a single (out of 3) AWS region in the U.S.
- Ran at level of 6000 instances for a few days with very low fluctuation (VM instance termination) due to spot overbidding
  - <1% of total running VM instances were terminated by AWS while production jobs were running during a multi-day period
  - Most of the terminations occurred within the first hour after VM instance creation -> no cost to us

Public cloud dvantages: “Unlimited” horizontal scaling in AWS EC2/S3 in terms of network bandwidth between compute and storage. Very high performance Object Store at low cost (when used as temporary storage for intermediate data products)

Running ATLAS Jobs in the Cloud – Potential going forward

Combination of “unlimited” capacity whenever we need it and competitive pricing makes AWS (and presumably other commercial cloud providers) an ideal resource to cover peak demands

- Could think of deploying only components (kind and quantities) & services at our dedicated data centers where cloud providers cannot cope (yet, i.e. technical capabilities and cost)
- Potential to vastly reduce size and scope of our dedicated (and aging) hardware deployment
- Potential to lower computing facility operations cost at improved performance (whenever the collaboration is in desperate need) and availability (e.g. the availability of AWS services is much higher than what WLCG sites provide)
- Potential to vastly increase our flexibility
- Using cloud computing makes us nimble whenever we need specific resources/platforms – temporarily or for long periods.

Running ATLAS Jobs in the Cloud – Matching Workloads

But all these wonderful things are not compatible with and/or applicable to our current processing model

- Most of our compute-intensive jobs run for 6-24 hours making spot VM instance terminations likely at probability of up to 80%
- Potentially a huge waste of resources we would have to pay for
- In a previous run we've observed 10-20% VM terminations (out of 2500 VM instances) with 2-hour jobs -> the shorter the job the better

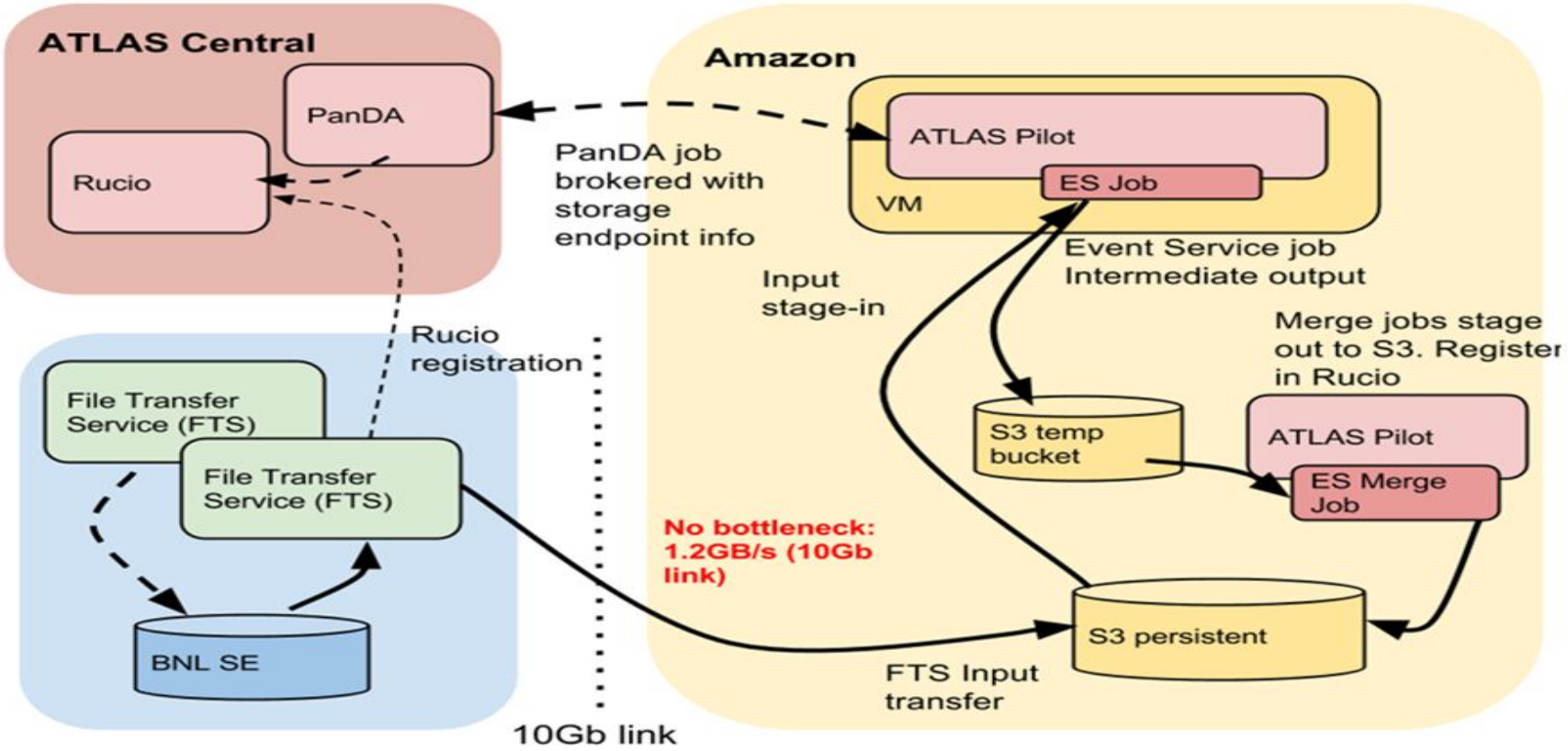
➢ ATLAS needs to match volatile and opportunistic resources with workload profile that suits the characteristics of a volatile resource

➢ Minimize loss due to resource becoming unavailable at any point in time

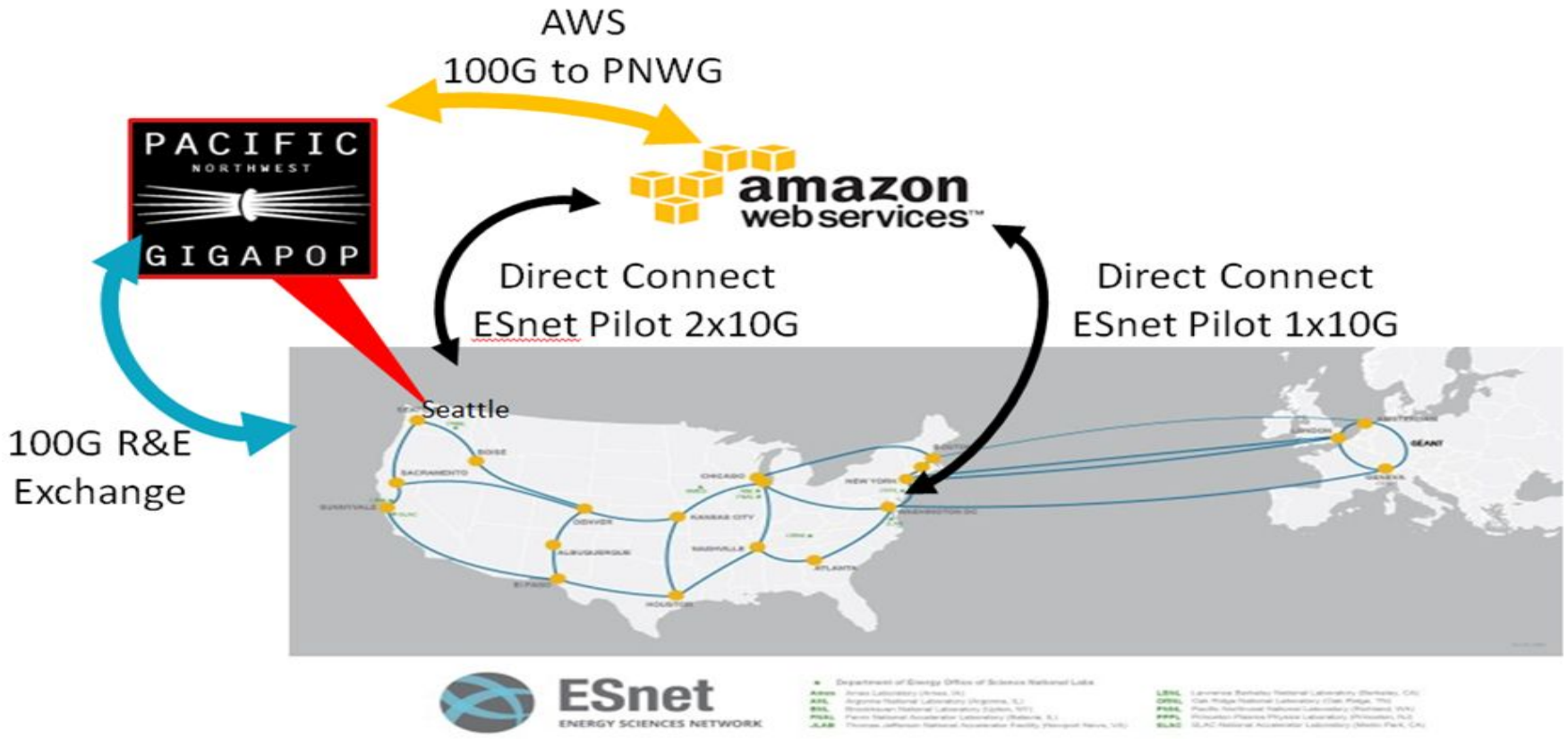
➢ The Event Server comes with all features that perfectly fit the characteristics of the AWS spot market

- Fine-grained processing at the event level – if we lose a VM we lose no more than a single event
- Supports parallel processing at high degree – can "grab" and utilize as many CPU resources as the provider can offer
- Utilizes high performance Object Store technology – this is the storage technology cloud providers have been focusing on

Event Server based Simulation on AWS



ATLAS ⇔ AWS Facilities Networking



US ATLAS LHCONE Sites

ESnet-LHCONE sites

- BNL, Michigan (AGLT2), Chicago and UIUC (MWT2), SLAC (WT2)

Internet2-LHCONE peered Tier2s:

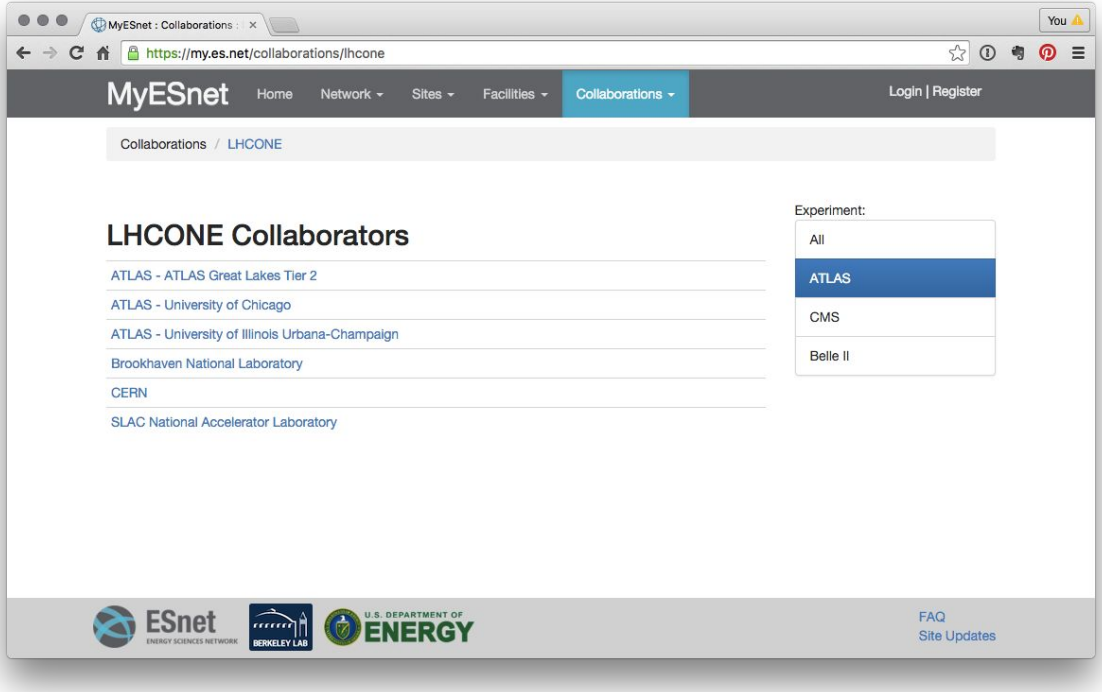
- Indiana U (MWT2)
- UTA, OU (SWT2)

Tier2 sites pending LHCONE

- Boston U (NET2)
- Langston U (SWT2)

Pending “Tier3” requests:

- Duke University
- University of Texas Advanced Computing Center (TACC)

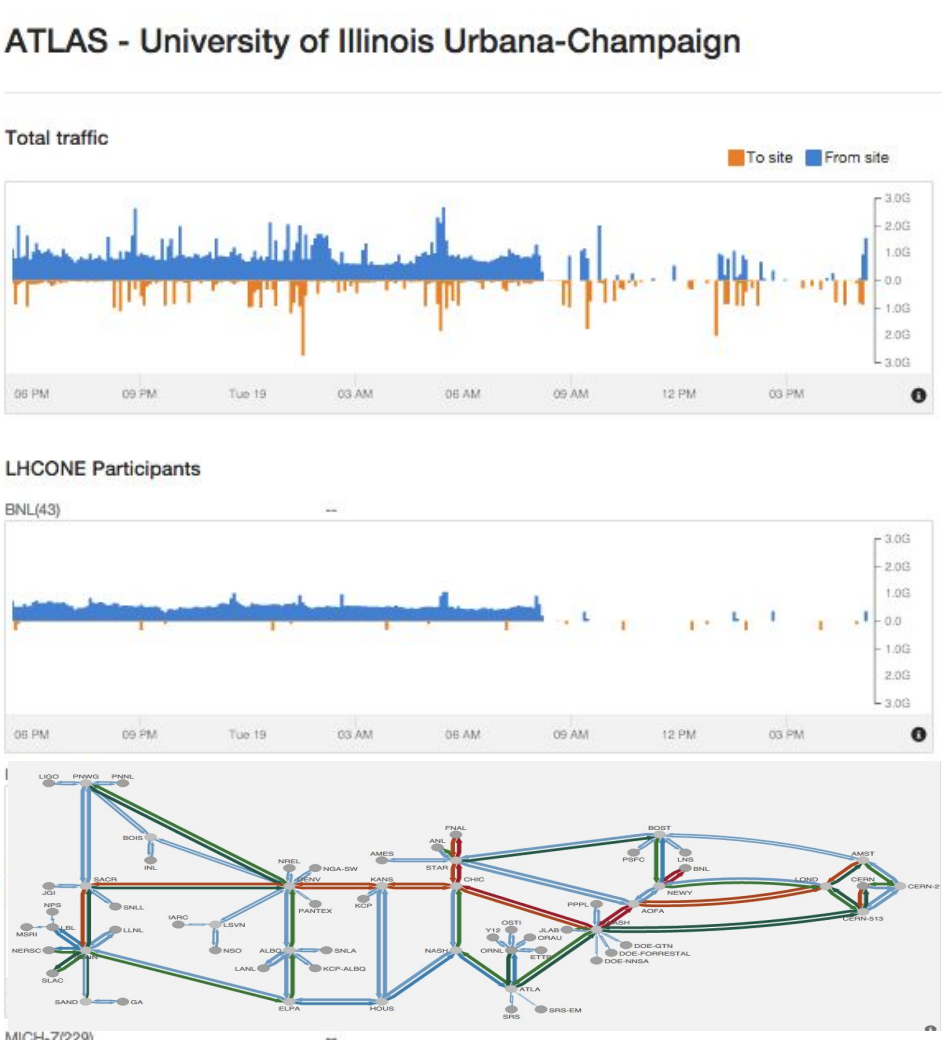


Traffic Monitoring

As part of the ESnet VRF service we have aggregate traffic monitoring between LHCONE peered sites.

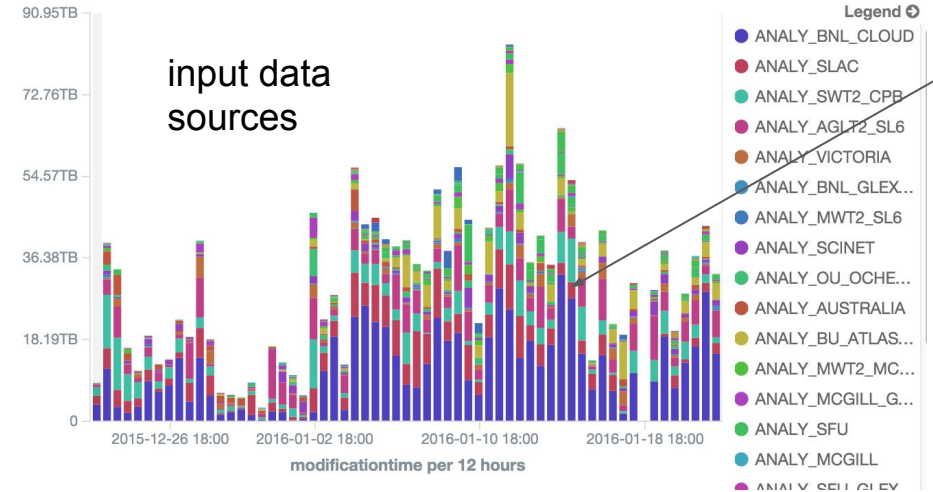
In process of migrating all US ATLAS sites to use the ESnet VRF service.

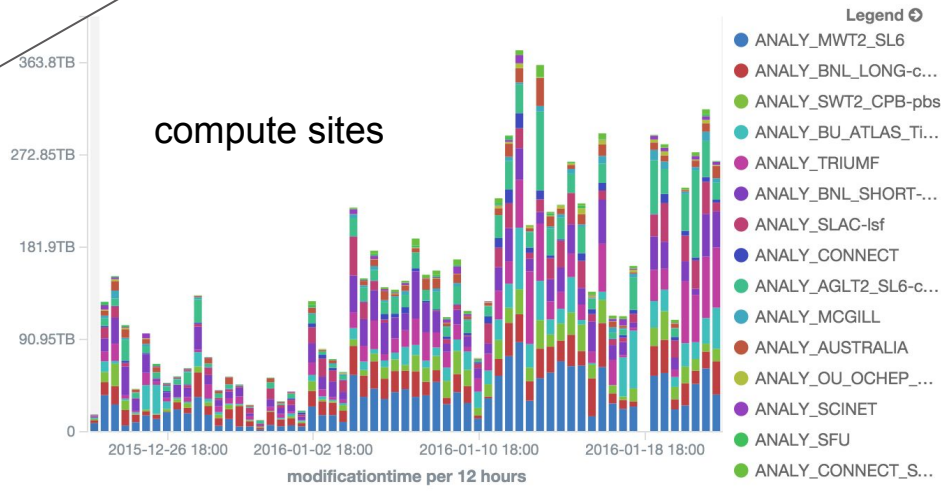
Will work with developer at LBL to forward this data to the analytics platform, providing a reliable measure of traffic capacities and protocols used over all LHCONE links

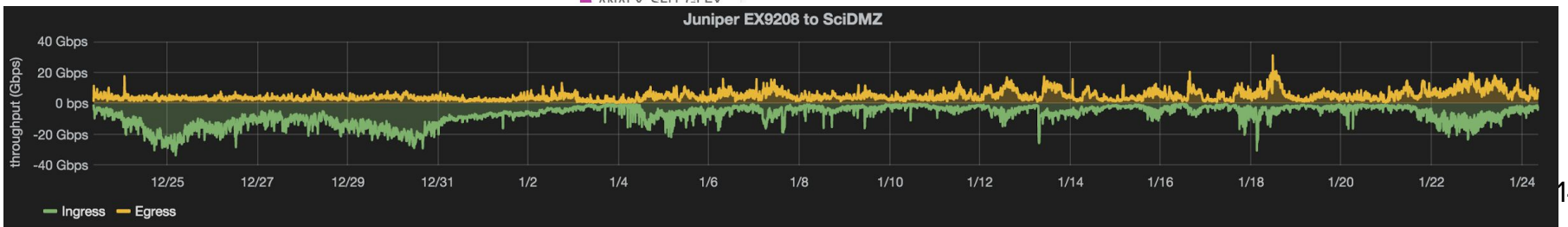


FAX and Overflow bandwidth in the US

to stress test caching proxy server at MWT2, overflow rates to MWT2 significantly increased. Even so average rates still modest (largest 12 h average delivered by BNL 1.55 Gbps).

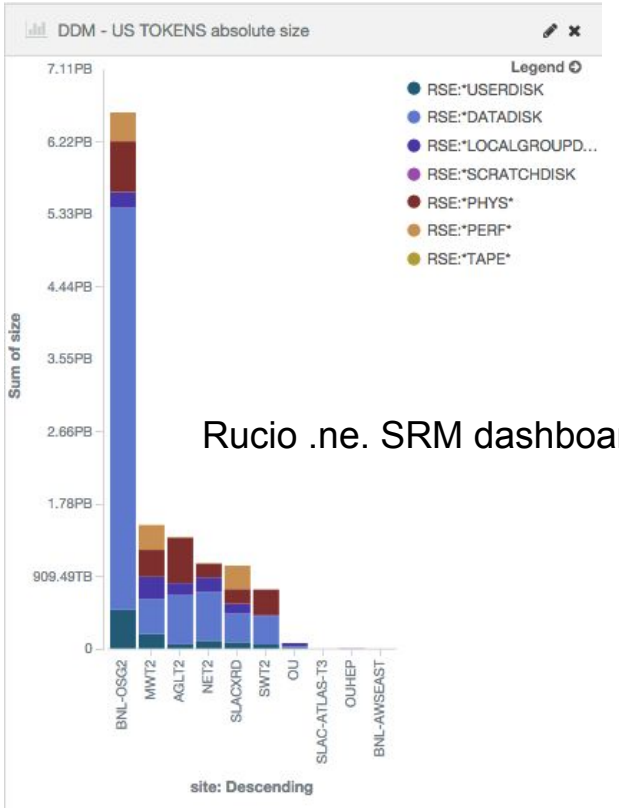






Operational Issues & Feedback

- Space management by ATLAS receives the most discussion in our bi-weekly Integration & Operations meetings
- Storage consistency (Rucio, SRM, local), dark data deletion and transparency of process still a concern



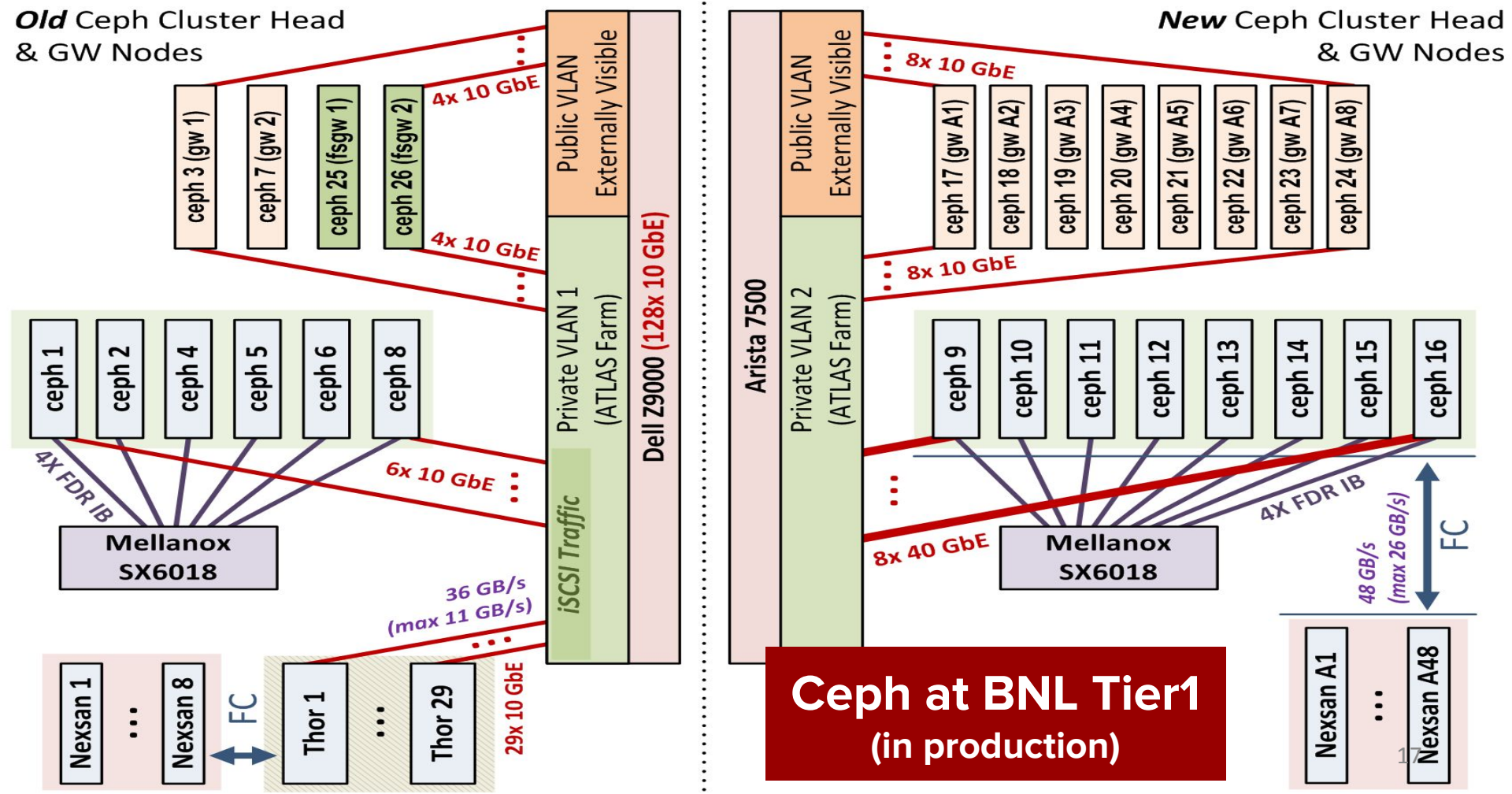
Rucio .ne. SRM dashboard?

Facility R&D

- Ceph storage
- Xrootd cache
- (distributed) virtualized data centers

Two Ceph clusters deployed in RACF as of 2015Q4

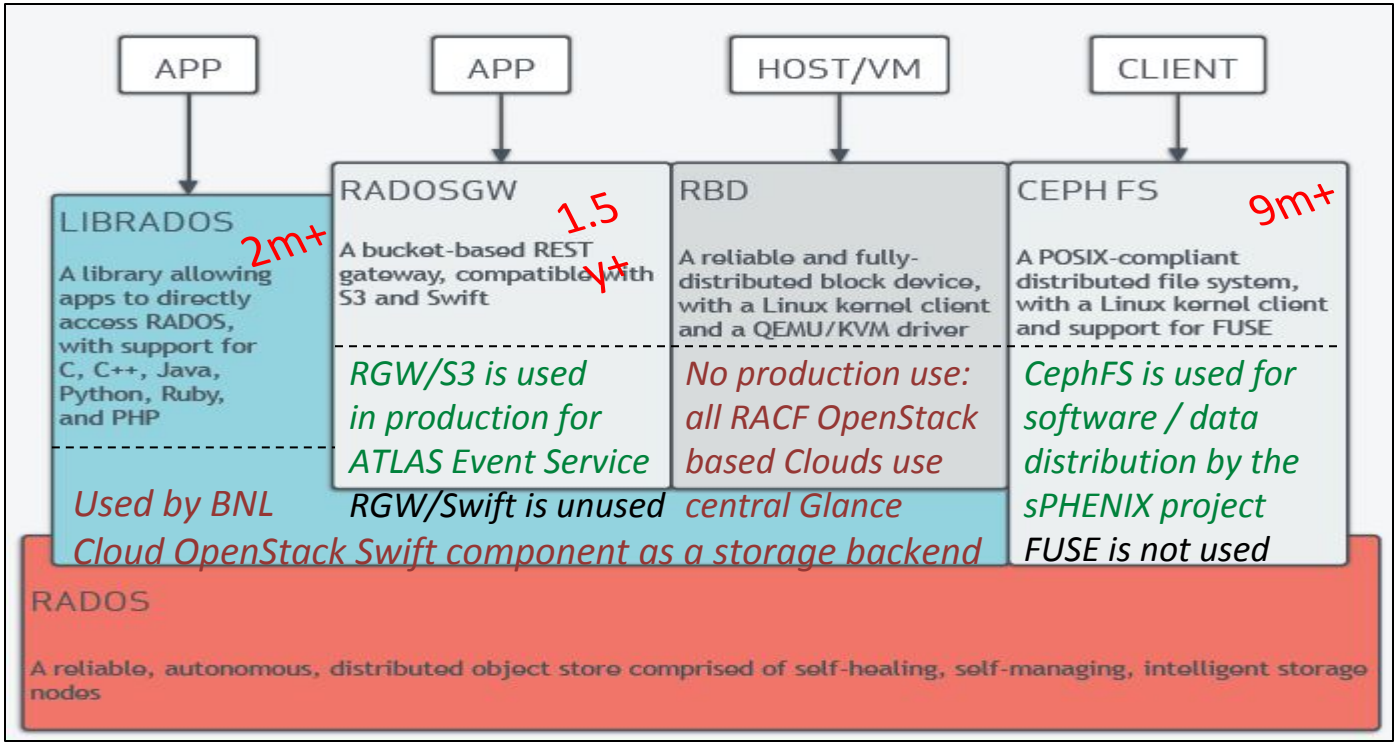
0.6 PB + 0.4 PB usable capacity split



Ceph at BNL Tier1
(in production)

Michael Ernst
BNL

Ceph Components in US ATLAS

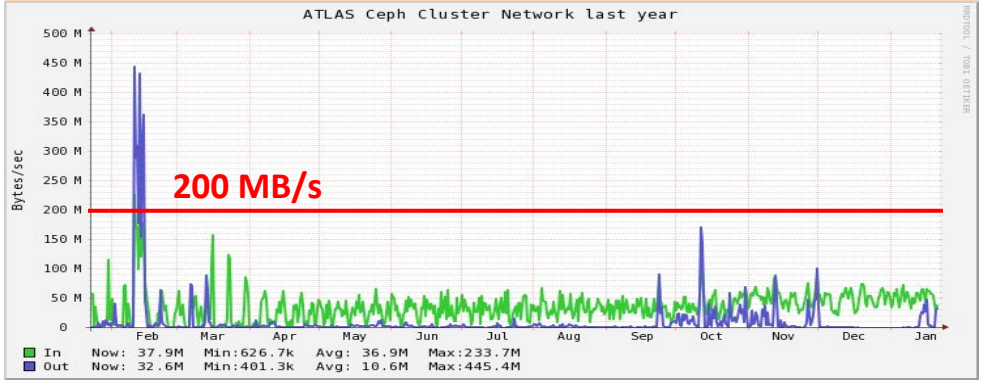


The BNL Cloud instance is the first user of Ceph@BNL that utilizes low level/low overhead object store API of Ceph directly

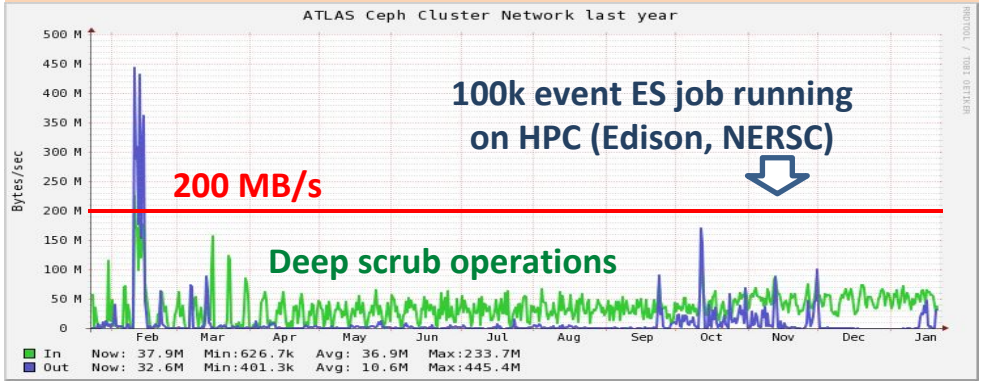
Production Experience with ATLAS Event Service

- 34M objects in PGMAP of the production cluster after 11 months of use by ATLAS ES
- 21M objects in a single atlas_pilot_bucket (no hard limit set)
- 8% of the available capacity of the cluster is used for ATLAS ES related data so far
- 18% of the maximum I/O capacity to the RadosGW/S3 clients (≈900 MB/s) of the current production cluster is used by ATLAS ES so far
- The ATLAS ES load have increased a factor of 2x since Oct 2015 and can further increase by factor 5 while still staying well within the existing capabilities

Ceph internal cluster network activity over the last 12 months



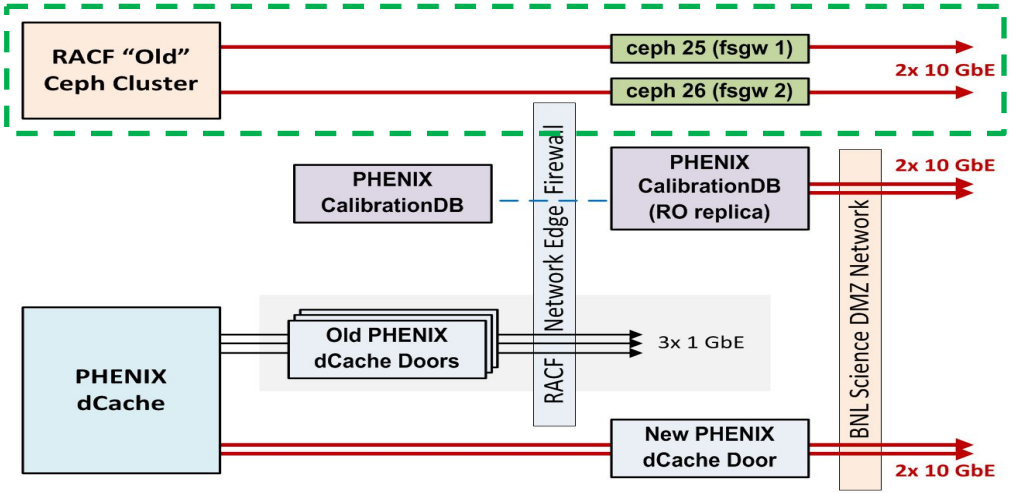
Ceph internal cluster network activity over the last 6 months

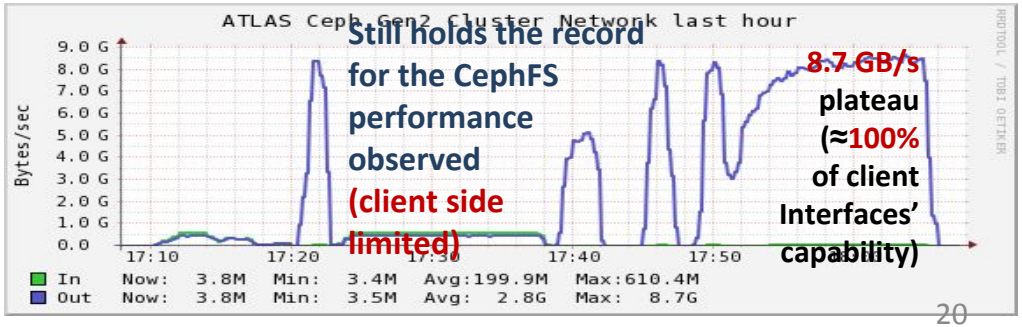


ES originated Internal traffic within the Ceph cluster through the RadosGW/S3 (localized spikes up to 160 MB/s)

Production Experience with CephFS/GridFTP

- New data access infrastructure enabling Production on OSG Opportunistic Resources includes a group of CephFS / GridFTP gateways
  - Globus GridFTP server version 7.x
  - OSD striping is custom tuned for user directories in CephFS for maximum performance by using the *xattr* mechanism
- Available for production use since Aug 2015
  - Up to 300 TB of usable space (with factor of 3x replication protection)
  - Capable of serving data through CephFS at the level of 8.7 GB/s





Ceph at MWT2

- Initial purchase: (14) Dell R730xd
  - (2) E5-2650 v3 @ 2.3GHz; 96 GB RAM
  - (12) 6TB disks; 400GB NVMe SSD for journals
  - 10Gbps SFP+
- Recent additions: (20) Dell R730xd
  - (2) E5-2650 v3 @ 2.3GHz, 96GB RAM
  - (14) 8TB disks; 2x200GB SATA SSD
  - 10Gbps SFP+
- ~3 PB total raw disk space
- BeStMan SRM + GridFTP plugins lifted from EMI
- XRootD / FAX
- Testing GaneshaFS for NSF access
- Evaluating local group disk from dCache to Ceph

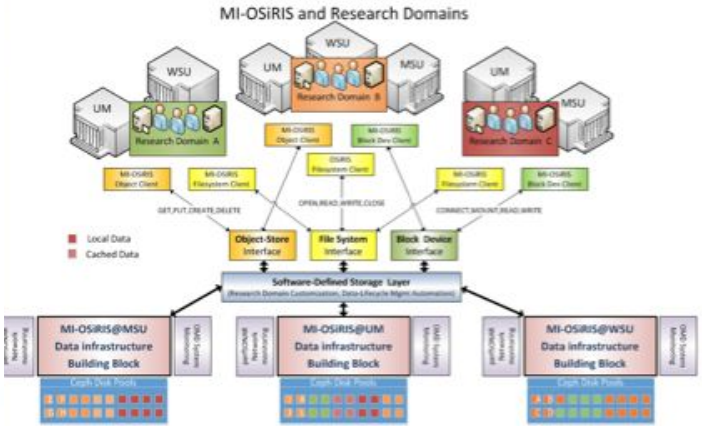
Storage R&D: OSiRIS "building blocks"

OSiRIS (NSF CC*DNI DIBBs)

- OSiRIS (Open Storage Research Infrastructure) is a NSF CC*DNI DIBBs grant awarded to the University of Michigan, Michigan State, Wayne State and Indiana University.
  - See https://nsf.gov/awardsearch/showAward?AWD_ID=1541335
  - One of 4 funded nationally by this program
  - Approximately $5M (~$1M/year)
- OSiRIS will provide a distributed, multi-institutional storage infrastructure that lets researchers write, manage, and share data from their own computing facility locations.
- The goal is transparent, high-performance access to the same storage infrastructure from well-connected computing locations on any participating campus.
  - It includes network discovery, monitoring and management(SDN) tools as well as the creative use of Ceph features.
  - The project will provide data sharing, archiving, security and life-cycle management, implemented and maintained with a single distributed service.

Shawn McKee
University of Michigan / AGLT2

Logical View of OSiRIS



Key: software defined storage interface specific to multiple research domains

Xrootd Cache R&D at MWT2

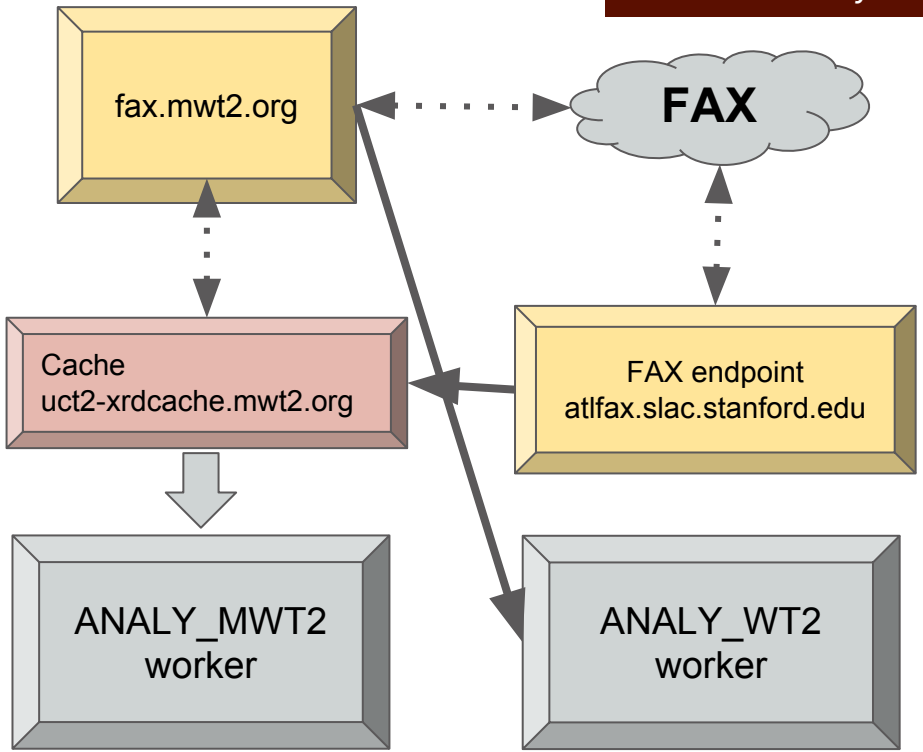
Ilija Vukotic
Wei Yang
Andy Hanushevsky
Lincoln Bryant

Caching machine:

- 26TB of RAID6 HDD
- 318GB SSDs
- 10 Gb NIC
- 12GB RAM

Test Configuration:

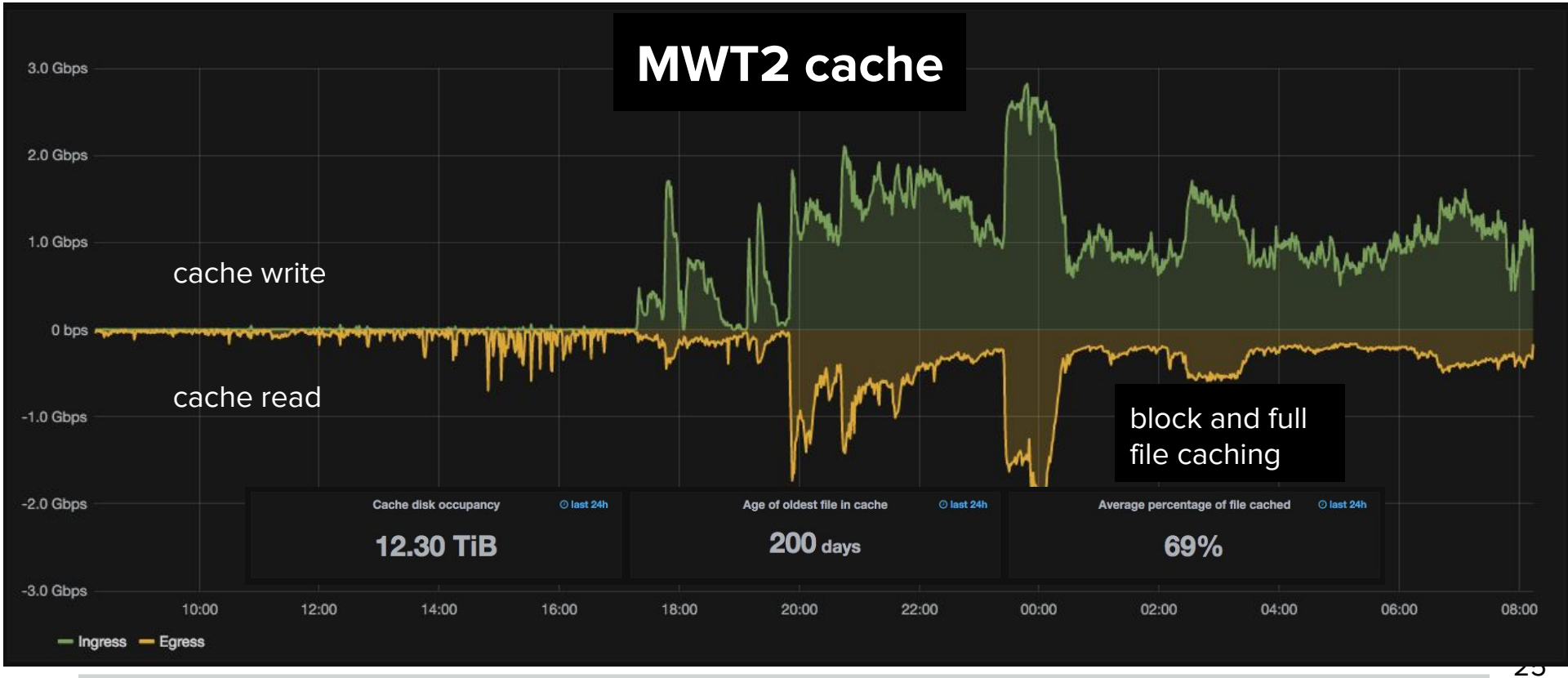
- Changed translation service so remote accesses use MWT2 caching server. When cache missed, uses FAX endpoint to get the data.
- Overflow jobs running at other sites still use our FAX endpoint.
- Failover jobs work as before.



Caching algorithm features

- Partial file (block based) caching, on-demand
- Only downloads the requested fixed-size blocks of a file (we set it to 1MB)
  - Serves them to client from memory.
  - Does not begin prefetching until a read request is actually received; a check is made if data required to fulfill the request exists on disk; if it doesn't, the required blocks get queued for download.
- Downloaded blocks are stored with the same filename as the original file. The cached file is sparse.
- For each downloaded file there is a metadata file containing:
  - Position of blocks already cached
  - Number of times file was accessed
  - Times of first and last access
  - Bytes requested

Cache response after loading up overflow



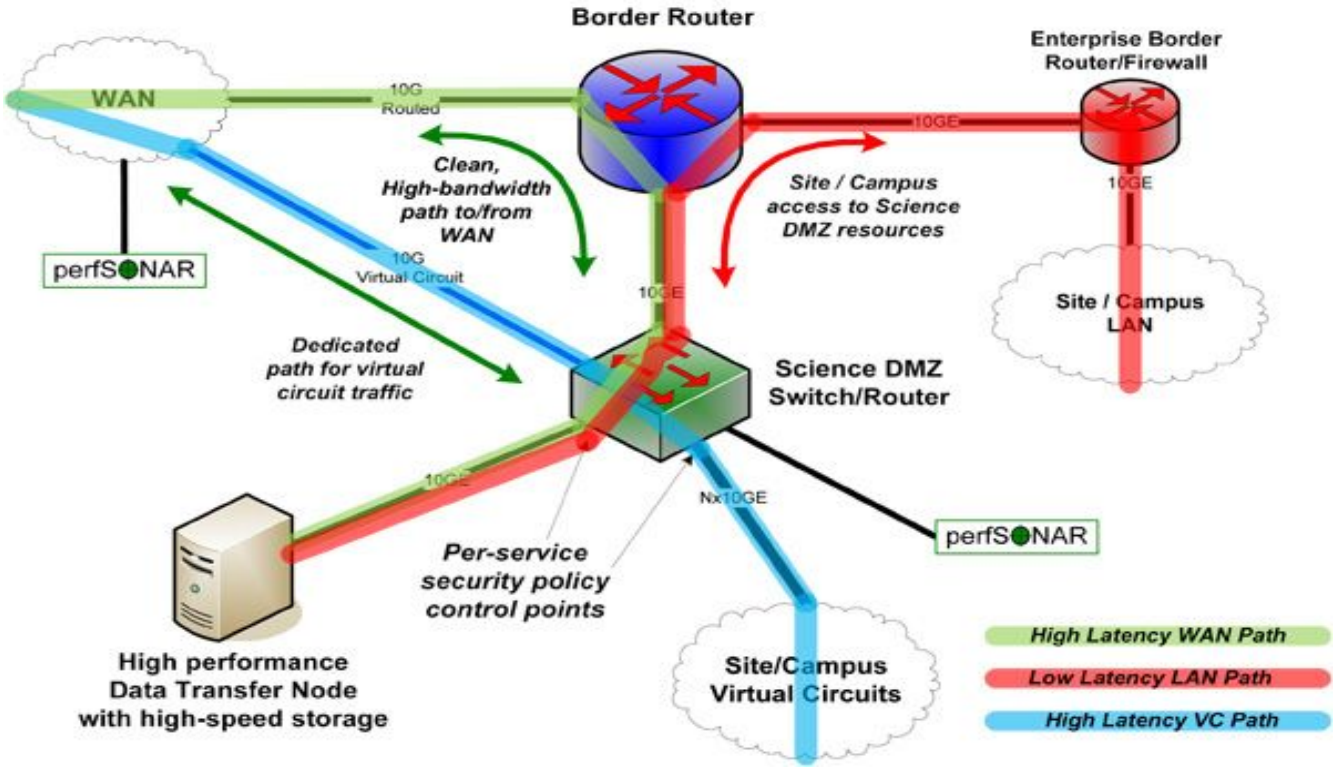
Virtualized Data Centers (“Cloudy Tier2”)

- Reduce ATLAS IT footprint and ops burden
  - Centralize deployment, operations, monitoring
- Exploring virtualized data center technologies
  - Joyent, Mesos, Kubernetes and containerizing services
- “Blue sky” goal
  - virtualize (connecting components of) entire US facility
  - ubiquitous “CI substrate”

Key: Ubiquitous & Easy “CI Substrate”

- Pioneer a new phase of advanced cyberinfrastructure deployment in the U.S., allowing sites to flexibly evolve and sustain both on-premise and commercial cloud-based infrastructure
- Hosted services, such as a FAX cache, Frontier-squid, etc., could be centrally deployed onto Tier2 “CI substrates” within a trusted CI zones and remotely operated, upgraded, and optimized for performance.

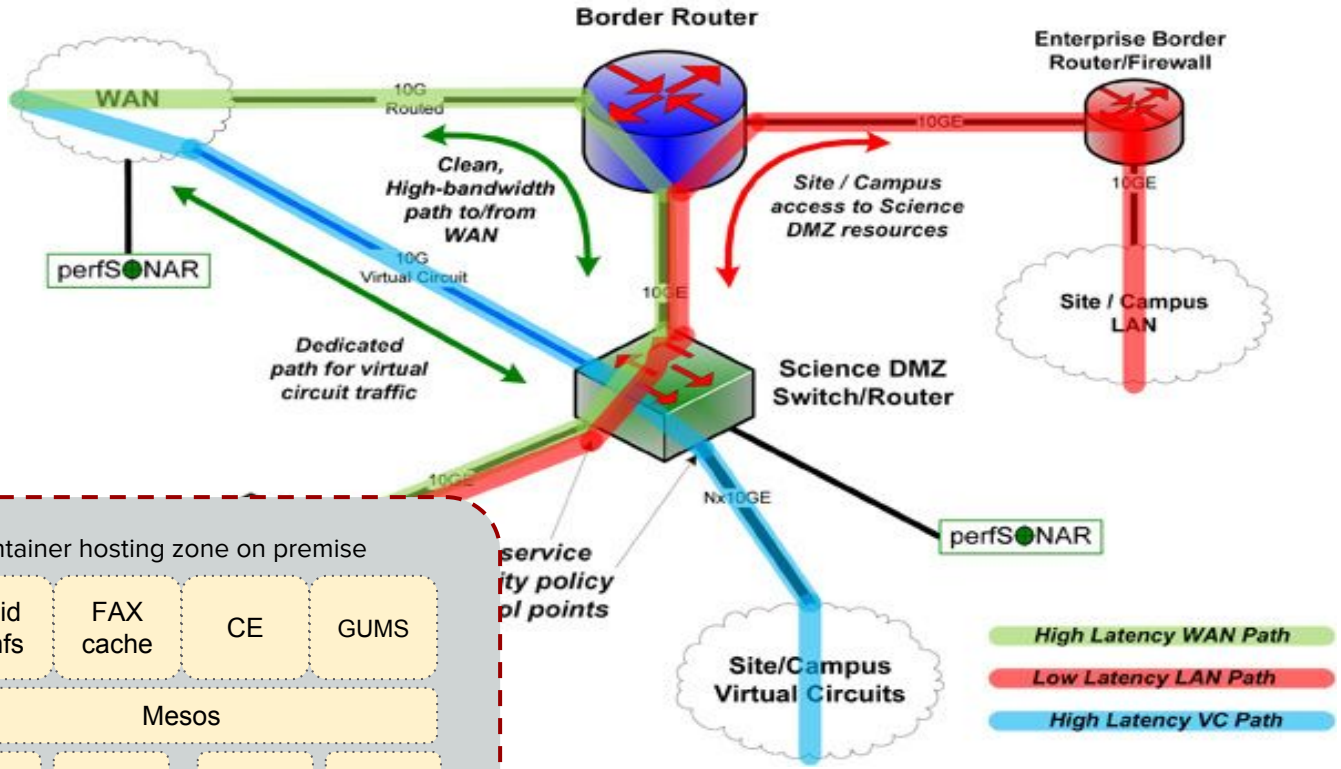
Canonical SciDMZ



SciDMZ with CI Substrate

US ATLAS Faclity central ops console:

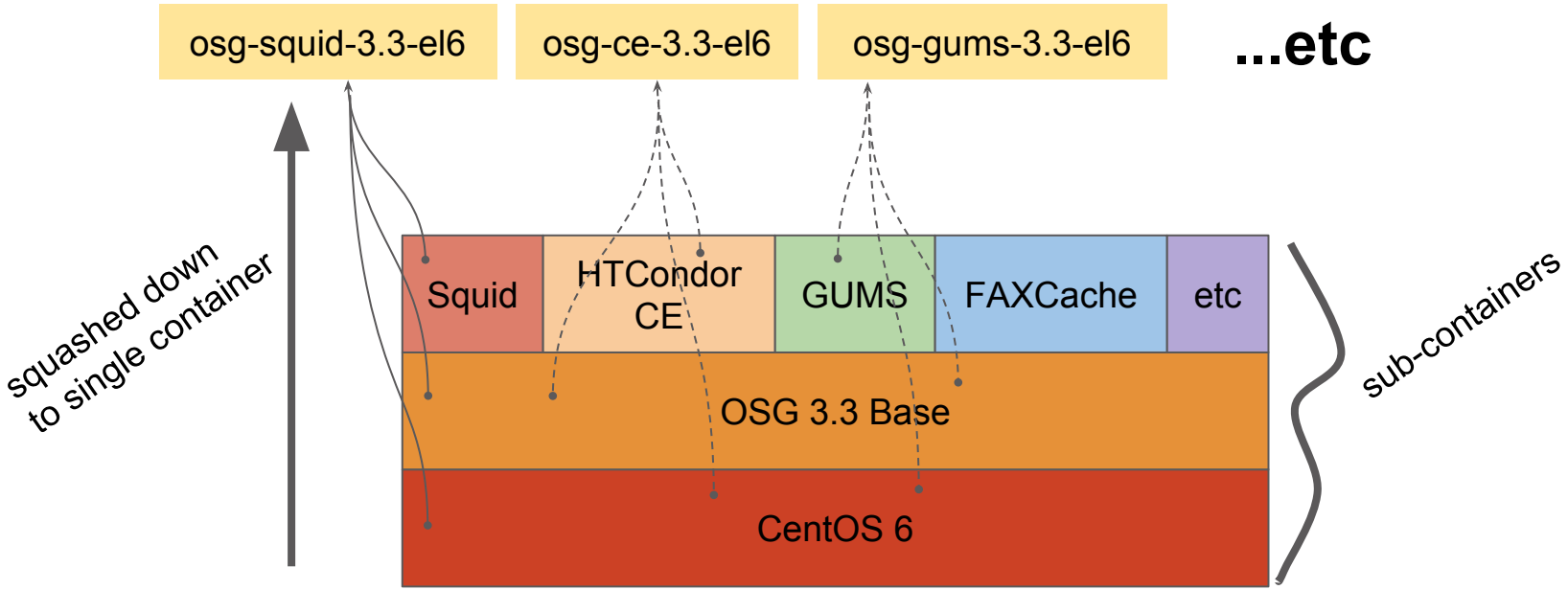
$ dcos package install front-squid.3.2 --sites MWT2 ALGT2 WT2



Edge container hosting zone on premise

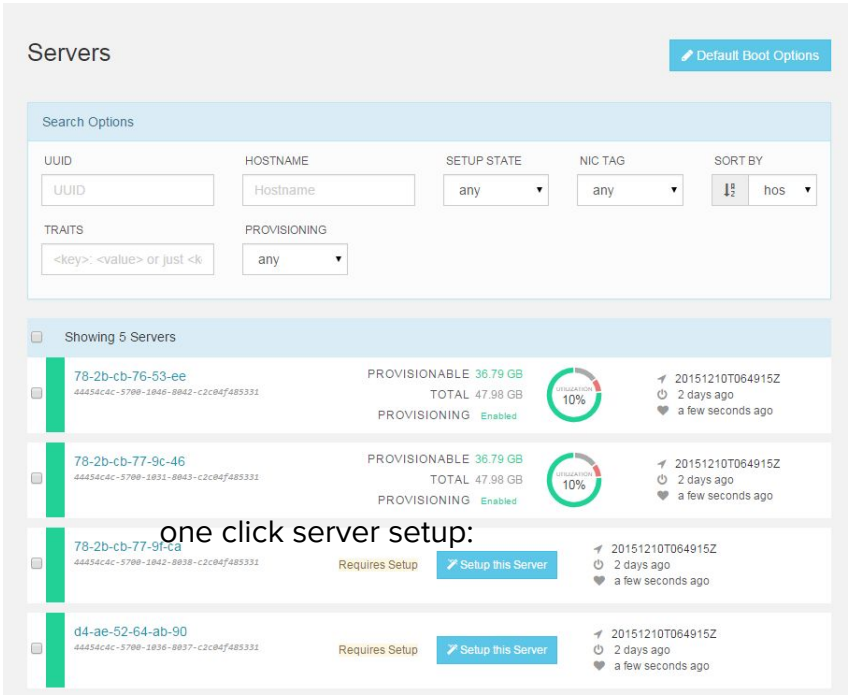
squid cvmfs	FAX cache	CE	GUMS
Mesos			
node	node	node	node

Key: Containerizing Services

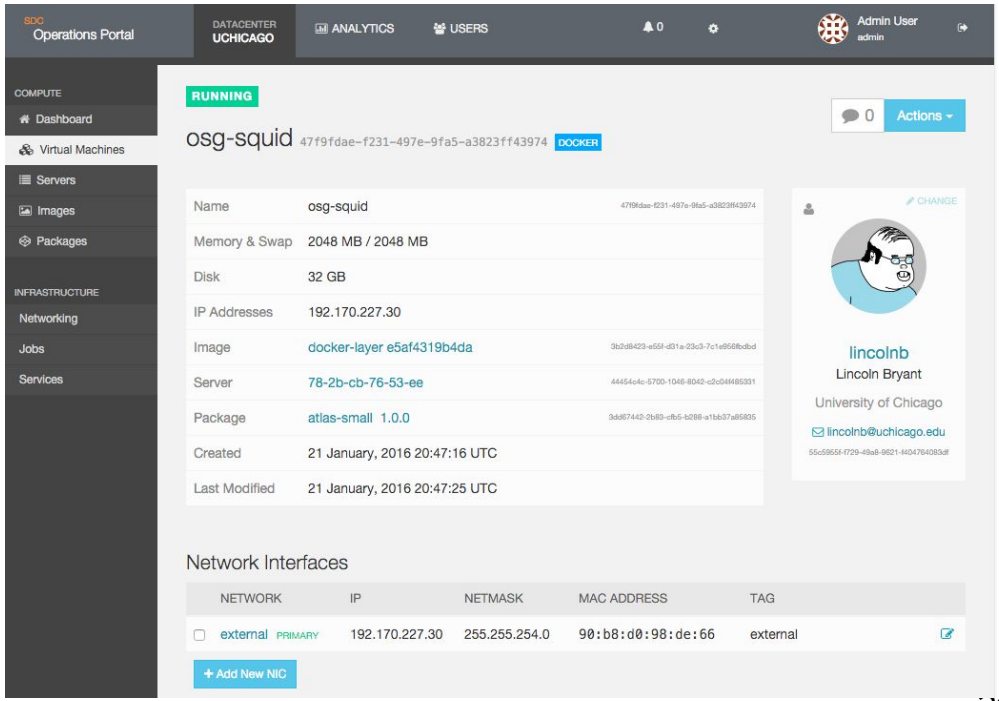


Frontier-Squid Containerized

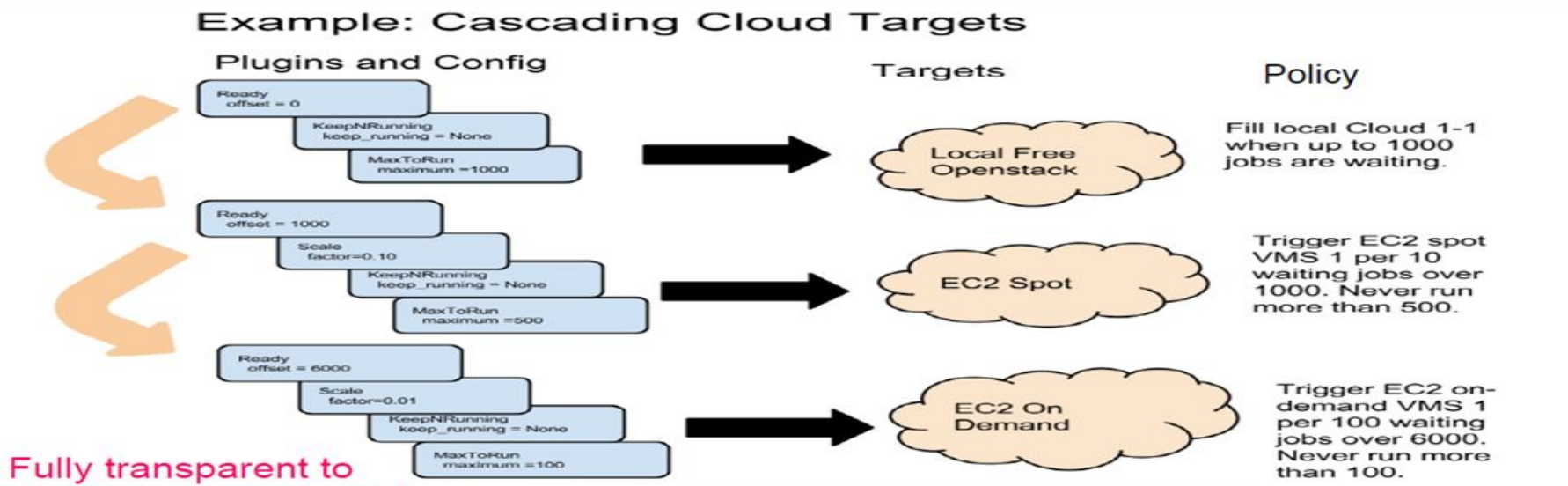
- And deployed in a hybrid cloud:



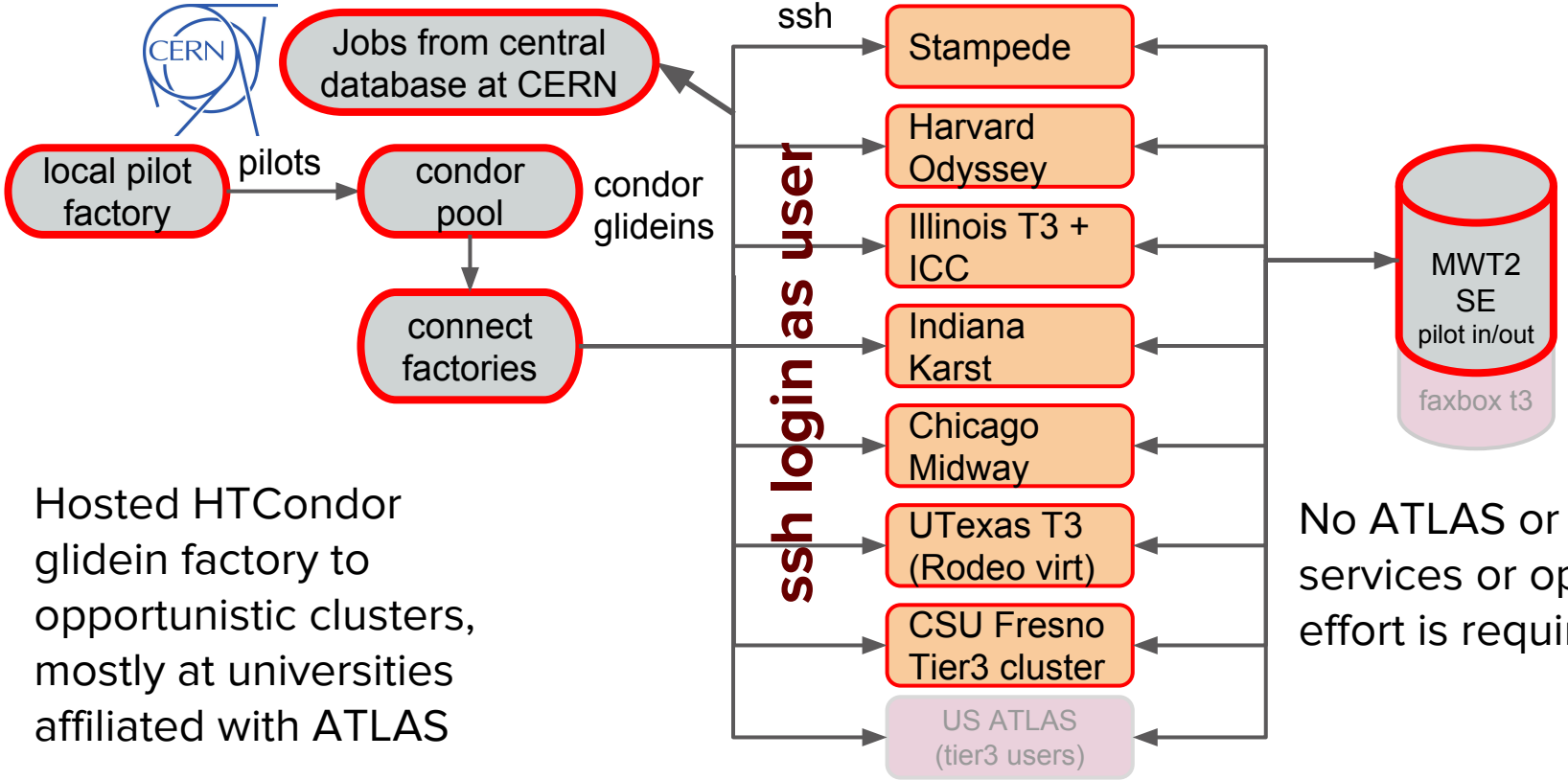
one click server setup:



Key: Policy based Cloud Scheduler



ATLAS Connect: Glideins via SSH



Hosted HTCondor glidein factory to opportunistic clusters, mostly at universities affiliated with ATLAS

No ATLAS or OSG services or operational effort is required

Additional slides

RACF Ceph Clusters: Building Blocks

First gen. head nodes, first and second gen. gateways

Dell PowerEdge R420 (1U) x18

2x 1 TB HDDs in RAID-1 + 1 hot spare
50 GB RAM + 1x 250 GB SSD (up to 10 OSDs)
1x 40 GbE + 1x IPoIB/4X FDR IB (56 Gbps) – *Head nodes*
2x 10 GbE – *Gateways*

Second gen. head nodes

Dell PowerEdge R720XD (2U) x8

8x 4 TB HDDs in RAID-10 + 2 hot spares
128 GB RAM + 2x 250 GB SSDs (up to 24 OSDs)
1x 40 GbE + 1x IPoIB/4X FDR IB (56 Gbps) +
12x 4 Gbps FC ports

Storage backend (retired ATLAS dCache HW RAID disk arrays)

iSCSI export nodes

SUN Thor servers (Thors)

48x 1 TB HDDs under ZFS
8 GB RAM
1x 10 GbE
4x 4 Gbps FC (no longer used)

FC attached storage arrays

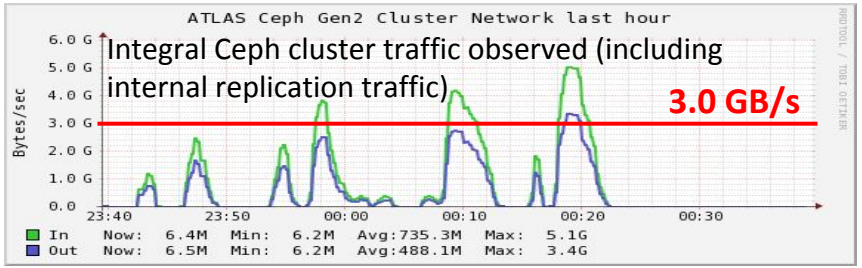
Nexsan SATABeast arrays (Thors)

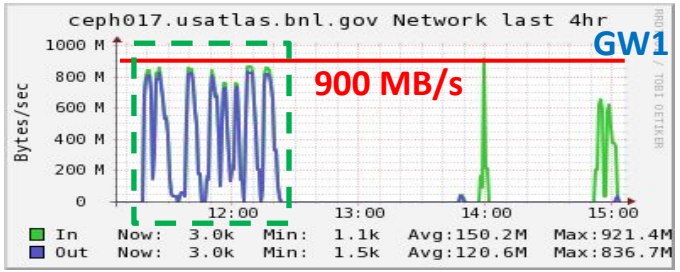
40x 1 TB HDDs in
HW RAID-6 + 2 hot spares
2x 4 Gbps FC (no longer used)

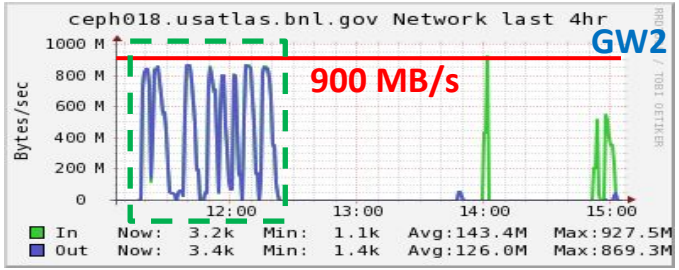
x56

OpenStack Swift Integration with Ceph

- BNL Cloud is an OpenStack based computing resource
  - In Oct 2015 the need for providing the BNL Cloud installation users with centralized object storage system was realized and the OpenStack Swift front end was chosen as a user frontend
- Ceph object store layer as a stand-alone storage backend for OpenStack Swift was successfully demonstrated with
  - Ceph v9.2.0 (Infernalis release)
  - OpenStack Swift v2.3.0 (Kilo release)
  - The third party swift-ceph-backend RADOS API wrapper
  - Two gateway machines with 20 Gbps between the *new* Ceph cluster and the BNL Cloud installation (GW1 and GW2)
- This Swift/Ceph storage system is now production-ready
- Aggregate write performance of 1.7 GB/s in multithreaded tests with 2x 10 GbE attached Swift client node pushing up to 2.0 GB/s, ~85% utilization of line rate (multi-GB files, 64 concurrent threads using both Swift/Ceph gateways)







Ceph Object Server Backend for OpenStack Swift: https://github.com/openstack/swift-ceph-backend

Xrootd development (proxy server)

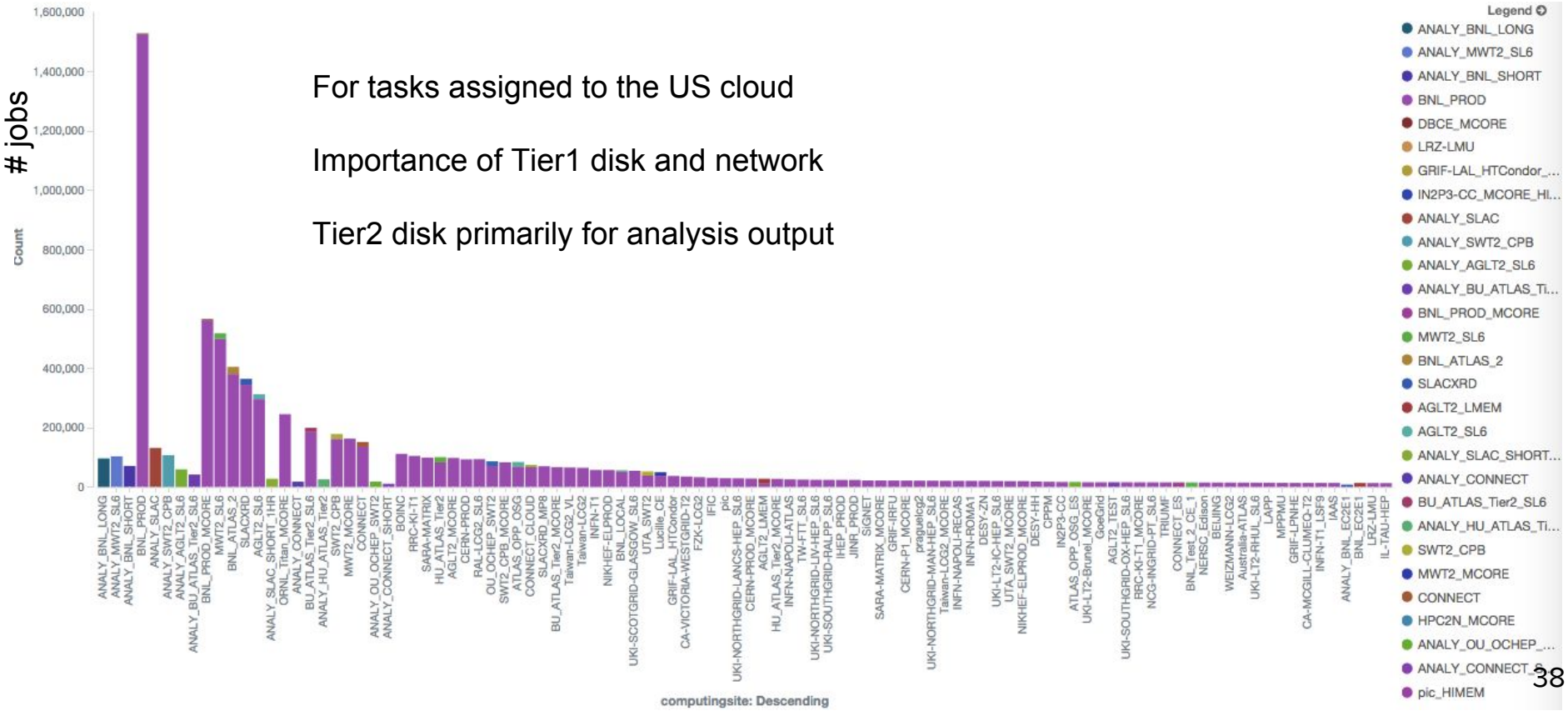
Disk Caching Proxy Future

- **Additional improvements are on the way**
  - Use async I/O to handle remote reads
    - This should reduce latency under load
  - Move to global buffers as opposed to per file buffers
    - Better memory utilization
  - Open remote file only when needed
    - Avoids delays when file is already in the cache
- Improvements to appear by the end of 2Q16
  - Remote file open may be delayed due to complexity

Proxy Base Code Fixes

- **Various fixes specific to proxy server**
  - Multiple event loops (default is 3)
    - Removes CPU bottleneck under heavy I/O
  - Avoid using Ffs code path when not needed
    - Stability improvement for large proxied clusters (e.g. FAX)
  - Properly manage deletion of object when I/O is pending
    - Resolved certain edge cases that would lead to a crash
- The above further resolved stability issues
  - These were specific to the proxy and hit under heavy load

Destination SE & computing sites



Software Defined Networks

- Nascent effort to explore space of SDN context of ATLAS computing and US ATLAS Facility
- Exploring Open vSwitch for testing at AGLT2, MWT2, NET2
- Need to understand use-cases
- Long term effort required