OSG Midscale*

Rob Gardner
Benedikt Riedel

OSG Planning Retreat @ University of Wisconsin
November 6-8, 2018

* = perhaps a better name?

CVMFS

- Hosting the origin server for: XENON, SPT, VERITAS, nEXO, modules
- Maintain build machines for both EL6 and EL7
- Maintain software installation for VERTIAS and nEXO
- SPT and XENON mostly install their own software, initial setup done by UChicago

Rucio Test Instances

- Hosting several test instances of rucio for experiments: CMS, LIGO, and IceCube
- Single Postgres DB instances, with different databases for each experiment
- “Rucio” node that runs the daemons for each experiment

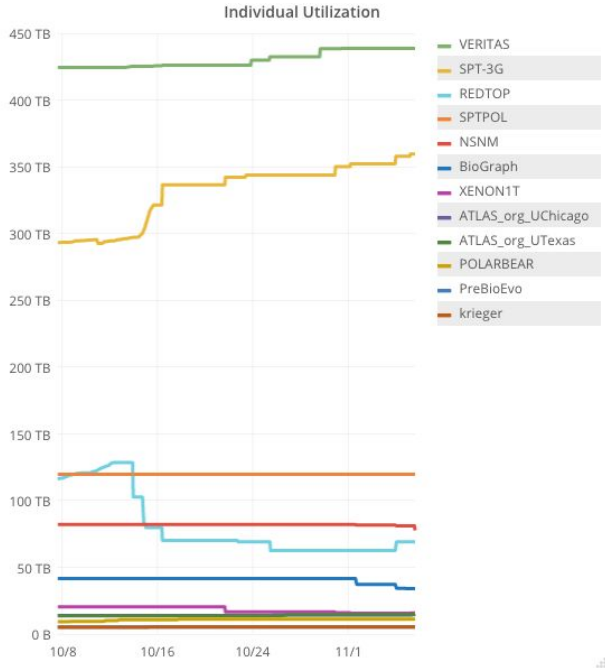
Storage Inventory (2.3 PB)

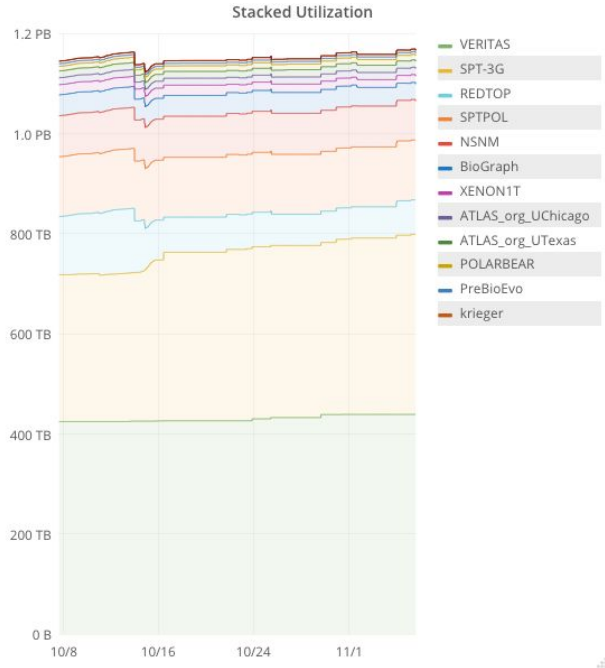
- Stash
  - Largest users: VERITAS (~400 TB) and SPT (~400 TB)
  - OSG Connect users vary from 100s TB to few GB – Space getting tight (overall Stash capacity = 3 PB deployed, ~ 1.2 PB usable)
  - Three gridftp doors (mostly used by SPT and XENON), some users of Ceph S3 interface
  - 38+ servers to that provide storage, interfaces
- dCache
  - Predominantly XENON (~680 TB out of 1.1 PB)
  - Some usage by SPT for 2nd generation experiment data (~3 TB)
- StashCache
  - StashCache origin and cache

Stash Inventory

Largest projects / users

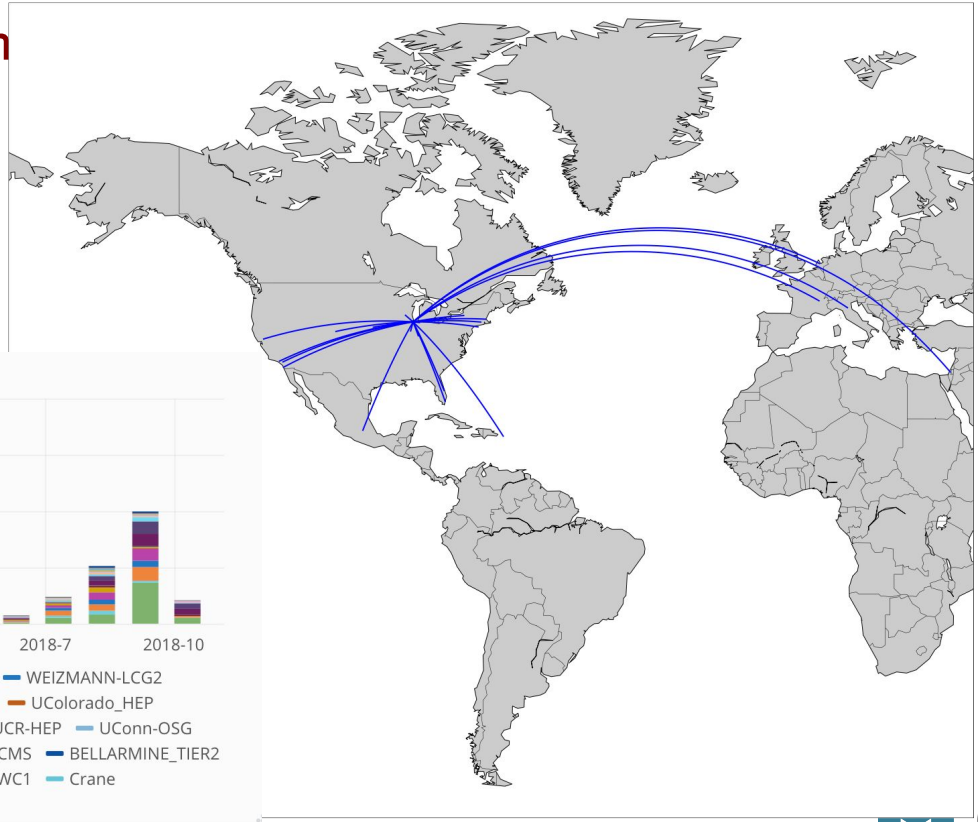
Metric	Current
VERITAS	438.65 TB
SPT-3G	359.57 TB
SPTPOL	119.58 TB
NSNM	77.61 TB
REDTOP	68.95 TB
BioGraph	33.99 TB
XENON1T	15.66 TB
ATLAS_org_UChicago	15.14 TB
ATLAS_org_UTexas	14.52 TB
POLARBEAR	11.27 TB
PreBioEvo	5.31 TB
krieger	5.30 TB
briedel	-

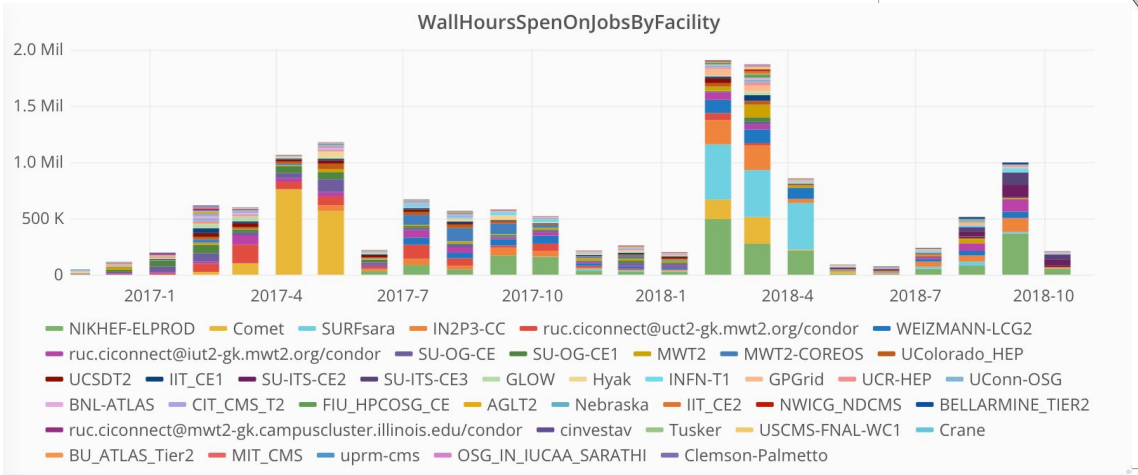




XENON job submission (CI Connect)

- Dedicated login node at UChicago is **sin** resources
- Jobs move through glideinwms to OSG





XENON Storage – Rucio

- Established non-ATLAS Rucio deployment and maintain Rucio instance
- Aggregate storage across 6 different sites and 9 Rucio endpoints in EGI and OSG
- Seamlessly move data between EGI and OSG with Rucio rules
- Overall manage 4.2 PB of allocated space, 2.1 PB currently used
- 6543254 files, 33139 data sets
- UChicago maintains OSG FTS3 instance
  - First user: XENON
  - Current users: XENON, LIGO, IceCube

XENON Storage – OSG

- Stash
  - Up to 60 TB for large reprocessing campaigns
  - Temporary storage for processing output until it gets moved to UChicago RCC
  - Temporary storage for Monte Carlo output
- dCache
  - 1.1 PB storage instance for processing on OSG
  - ~500 TB currently used, extra storage provisioned for next generation experiment
  - **Only US storage site for XENON1T raw data**
  - **Origin for all data processing occurring on the OSG**

XENONnT development

- Preparing for next generation XENON experiment – XENONnT
- Significant changes to processing software and data organization
- OSG developed a REST API for XENON MongoDB “runsDB”
- Close involvement in planning new data processing and monte carlo workflow – Moving data processing to Pegasus-based workflow (Benedikt –> Mats near term)
- Hosting separate rucio and “runsDB” instance for testing

SPT-3G engagement

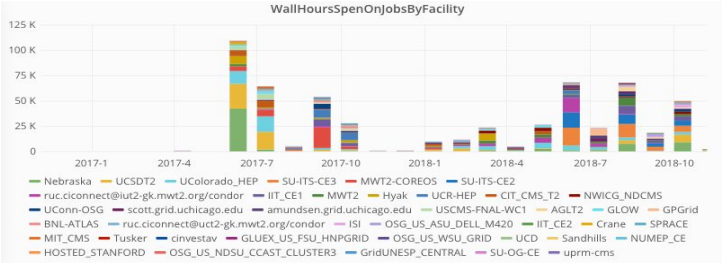
- **First CMB telescope that is using OSG as a primary source of computing** – Usually computing is provided by DOE labs (Planck) or university clusters (BICEP)
- Telescope has had technical issues, fully operational in past 6+ months
- Setup and maintain infrastructure at South Pole
  - UC sysadmin (Judith Stephen) travels to pole to setup and maintain
  - Annually retrieve data that cannot be transferred over satellite
- Partners: Tom Crawford (UC), Nathan Whitehorn (UCLA)

New Infrastructure Deployment – South Pole

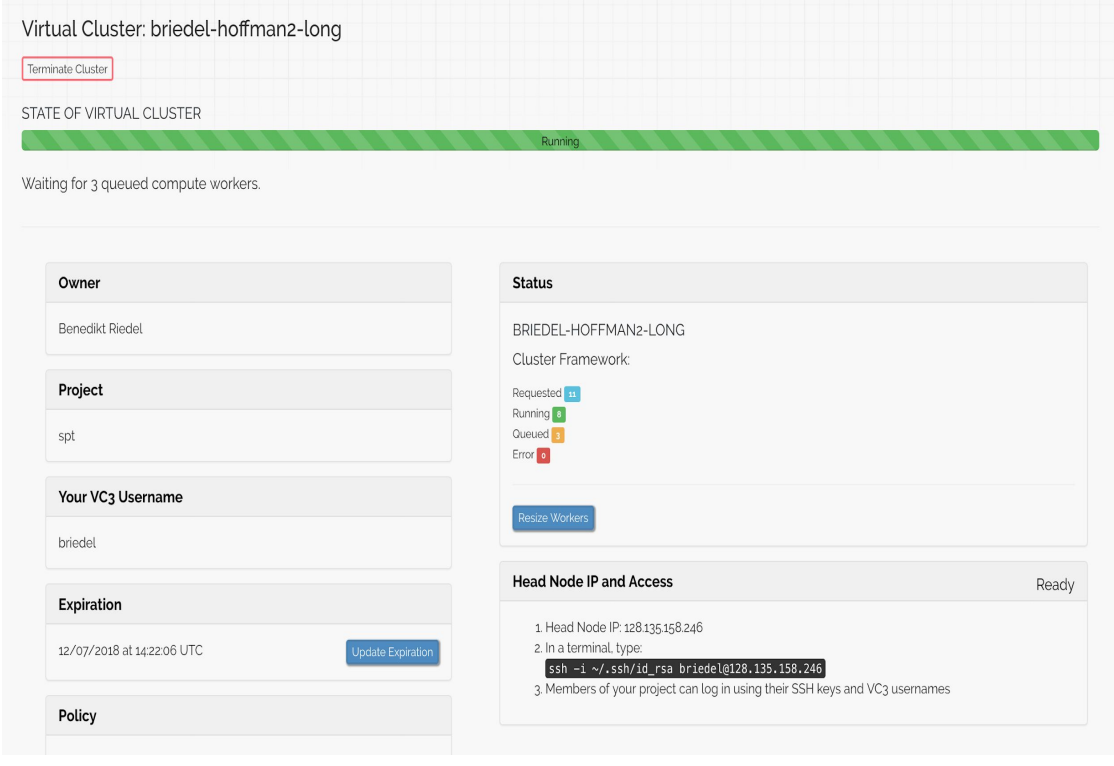
- Internal UChicago EFI MOU (John Carlstrom, PI and spokesman)
- New Hardware in red, managed by Judith
  - 4x Dell R730s:
    - 2x R730 for analysis work (HTCondor pool)
    - 1x R730 as hypervisor, 1x R730 hot spare
  - 2x Dell R330s: Storage controller + backup
  - 2x Dell MD1280s:
    - Primary Copy: ZFS pool, 42x 8 TB, NFS mounted to all R730s
    - Secondary Copy: JBOD, 28x 8 TB
  - 2x UPSes, 6x PDUs
- Old hardware in green – Part of analysis HTCondor pool
- Services: HTCondor, NFS, login, nagios, puppet, software, home dirs, DNS

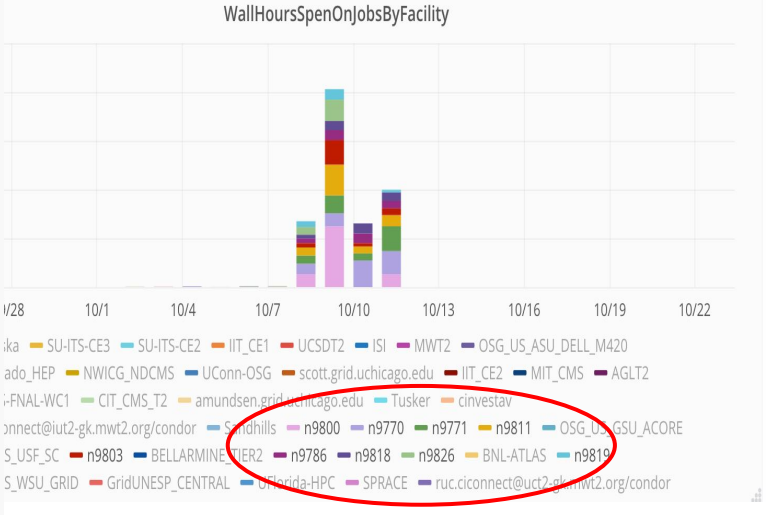
SPT-3G: analysis and data management

- UChicago analysis and data transfer infrastructure
  - Two analysis/OSG submit nodes are setup and maintained
  - Data is **ingested into Stash** and automatically **replicated to NERSC for backup via Globus** – Dedicated VM deployed for this
  - Also replicate to campus research computing storage system (Midway/DALI)
  - Tying dedicated campus resources (UCLA, UChicago) tied into pool using **VC3 provisioned flocking host**
  - Running servers to host SPT VOMS, Trac wiki, websites for data quality



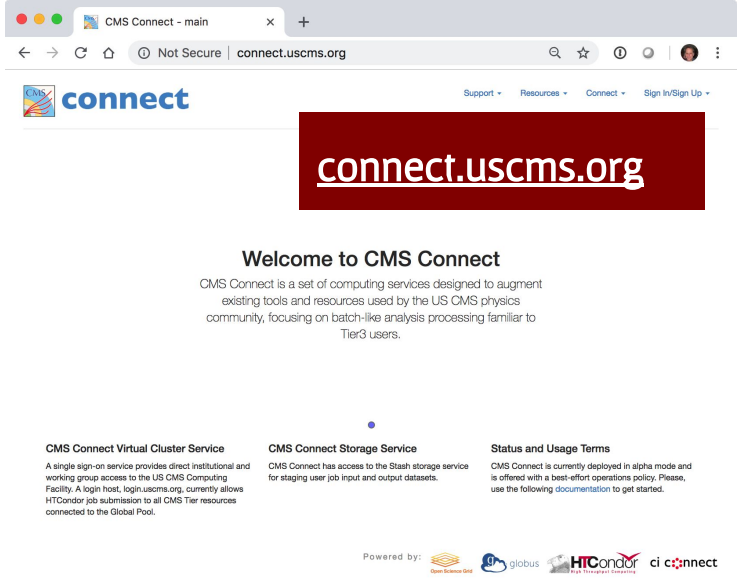
VC3 provisioned flocking host





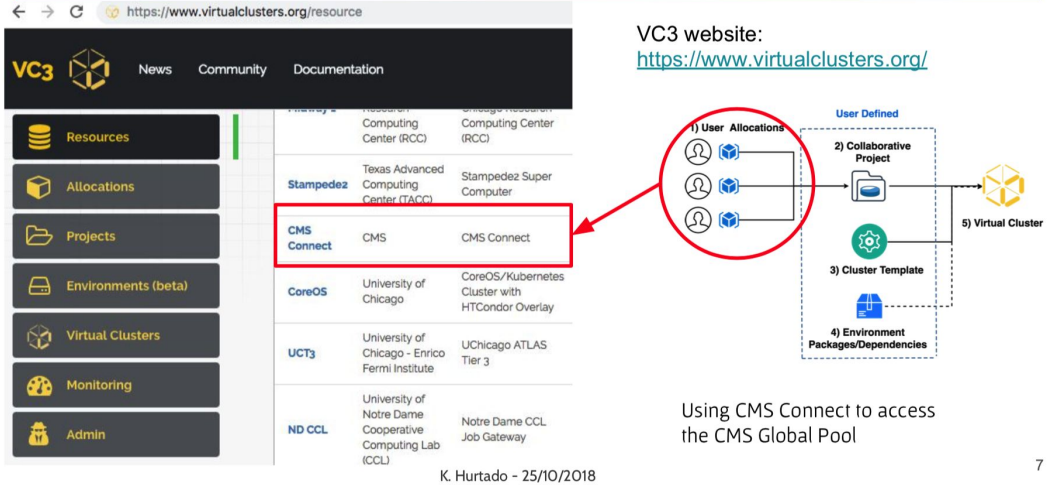
CMS Connect Services

Submit host for CRAB-alternative analysis platform. Recently extended to provision **Spark & Tier3 queue** over Notre Dame campus cluster using **VC3**



http://bit.ly/cms-connect-vc3

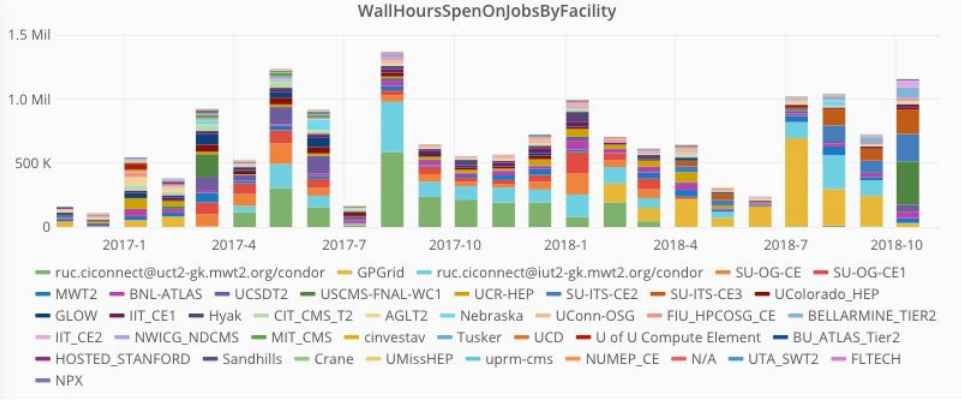
Creating a Spark Cluster – Step 1



K. Hurtado - 25/10/2018

VERTIAS – dedicated submit & storage

- A. Nepomuk Otte, GaTech engagement
  - Requests for storage and more compute
  - Complaints in 2017 about 50% higher compute time on OSG led us to deploy specialized submit & storage
- Job submission infrastructure
  - Submits jobs through CI Connect submit node (largest user)
  - Pegasus workflow
- Stash storage – Largest user (~400 TB)
- Preparation for submitting to GaTech as well as OSG



LIGO

- Attending Peter's weekly meetings with OSG staff regarding issues with StashCache, sites, workflows, etc.
- Discussion on technical issues, e.g. Singularity, and job failures due to StashCache, e.g. no close-by cache
- LIGO setup its own Rucio instance (using our FTS) and running initial tests to move data between sites
- Deploying GaTech campus PACE HPC for LIGO

OSG–Georgia Tech LIGO engagement

History

- Georgia Tech's PACE team deployed an OSG cluster in 2016 to run computations for the LIGO project.
- Back then, OSG/LIGO integration was partially experimental. This cluster eventually stopped receiving jobs due to the lack of dedicated PACE personnel to keep this system updated and operational.
- The virtual machines used in this proof–of–concept implementation failed to achieve the performance and reliability required for production runs.

OSG–Georgia Tech LIGO engagement

The primary objective of this project is to restore OSG/LIGO services on the cluster, making this resource available to local and external researchers who are members of the LIGO scientific collaboration.

An equally important goal is to build a comprehensive knowledgebase that will enable PACE team to maintain this resource in the long term. This includes detailed journal of system changes, links to relevant documentation, and training of a PACE team member (name TBD) who will be tasked with maintaining this cluster.

OSG-Georgia Tech LIGO engagement

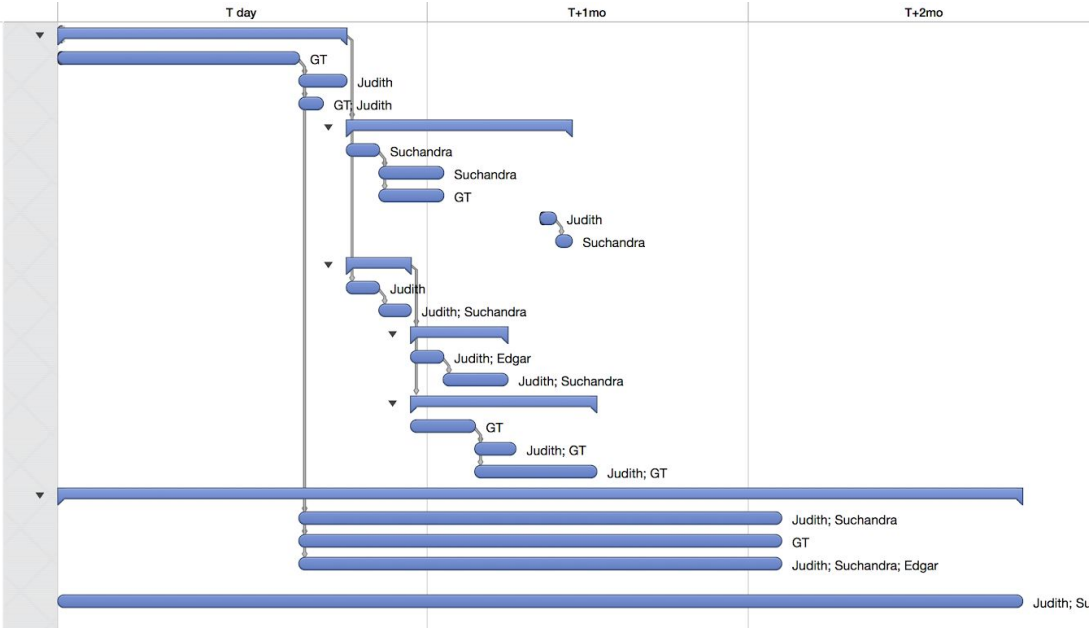
- Weekly meetings with GaTech sys admins
  - Judith, Edgar, Benedikt attending
- Judith (UC) doing the core systems administration
  - Deploying worker nodes, HTCondor, HTCondorCE, job submission host, StashCache instance for LIGO, LIGO shibboleth-webserver
  - Setup definitions in GaTech provisioning framework (SALT)
- Edgar (UCSD) doing the glideinWMS integration
  - Testing LIGO workloads with James Clark
- Everything set up services are either close to production or in testing – LIGO production jobs are arriving at GaTech, fully-tested authenticated StashCache last missing piece
- Most remaining items are site-specific and optimizations, e.g. admin training, network settings, edge cases, etc.

OSG-Georgia Tech LIGO engagement

- Headnodes
  - 4x Relion 2940s
  - osg-login1.pace.gatech.edu: HTCondor submit host
  - osg-sched.pace.gatech.edu: HTCondor central manager, HTCondor-CE, Frontier Squid
  - osg-gftp.pace.gatech.edu: Stashcache, GridFTP
  - osg-shibboleth.pace.gatech.edu: Shibboleth webserver
- Worker pool
  - 35x Relion OCP1930es
  - 2x Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz
  - 128 GB RAM
  - HTCondor worker node

OSG–Georgia Tech LIGO engagement

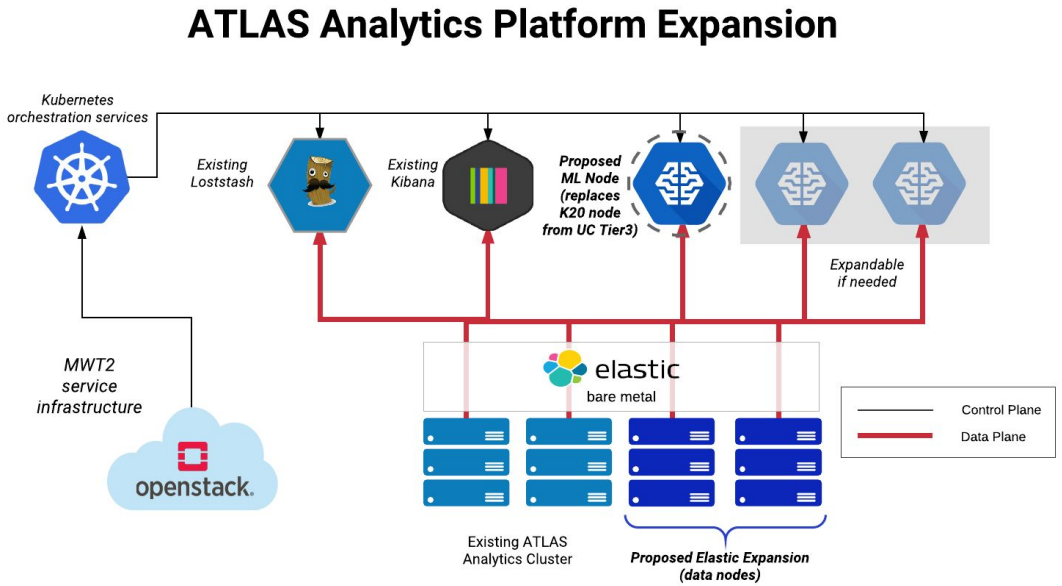
Title	Start	End	Duration	Assigned
1) Cluster Setup	**T day**	T+17d	18d	
• 1.1) Re-image Compute nodes with RHEL7	T day	T+14d	15d	GT
• 1.2) Install Condor on compute nodes	T+15d	T+17d	3d	Judith
• 1.3) Install Condor Collector	T+15d	T+16d	1.5d	GT; Judith
2) CE Setup	**T+18d**	T+31d	14d	
• 2.1) Provision Hosted CE	T+18d	T+19d	2d	Suchandra
• 2.2) Setup and Test Hosted CE	T+20d	T+23d	4d	Suchandra
• 2.3) Provision OSG CE	T+20d	T+23d	4d	GT
• 2.4) Setup OSG CE	T+30d	T+30d	1d	Judith
• 2.5) Transition to OSG CE	T+31d	T+31d	1d	Suchandra
3) CVMFS Setup	**T+18d**	T+21d	4d	
• 3.1) Provision Frontier Squid CE	T+18d	T+19d	2d	Judith
• 3.2) Setup CVMFS on compute nodes	T+20d	T+21d	2d	Judith; Suchandra
4) StashCache Cache Setup	**T+22d**	T+27d	6d	
• 4.1) Provision StashCache Cache	T+22d	T+23d	2d	Judith; Edgar
• 4.2) Setup StashCache cache	T+24d	T+27d	4d	Judith; Suchandra
5) LIGO Webservice (Shibboleth)	**T+22d**	T+33d	11.5d	
• 5.1) Setup NFS gateway	T+22d	T+25d	4d	GT
• 5.2) Setup Apache configuration	T+26d	T+28d	2.5d	Judith; GT
• 5.3) Setup Shibboleth integration	T+26d	T+33d	7.5d	Judith; GT
6) CFM, Documentation and training	**T day**	T+59d	60d	
• 6.1) Prepare Salt configurations	T+15d	T+44d	30d	Judith; Suchandra
• 6.2) Deploy Salt configurations and test changes	T+15d	T+44d	30d	GT
• 6.3) Journal of changes and documentation	T+15d	T+44d	30d	Judith; Suchandra; Edgar
• 6.4) Training of PACE staff member	T day	T+59d	60d	Judith; Suchandra; Edgar



IceCube

- One of the largest users of GPUs across OSG
- New computing manager: Benedikt Riedel (currently OSG staff at UChicago)
- Using rucio instance hosted by UChicago to move data to DESY dCache, will move to a private instance in future
- "Nice-to-have":
  - Easier ways to use OSG glideinwms to extend pool:
    - Targeting sites or higher priority at sites with IceCube affiliation (AGLT2, SWT2) through OSG
    - Better interoperability and communication with EGI
  - Work closer with supercomputing centers, e.g. agree to support a common job submission API, remote job submission at MFA sites (Stampede2 uses SSH key+IP)
  - Better knowledge sharing with large experiments, e.g. data transfer to/from supercomputers
  - Temporary storage for intermediate outputs

Analytics infrastructure (UC+USATLAS Ops; tbd SAND)



- UChicago Elasticsearch instances hosts both ATLAS and OSG (& SAND) monitoring data
- Total of three head nodes and ten data nodes
- Two servers for logstash instances
-

Network data

	documents	size
esnet	138243434	10861694803
packetloss	788962668	109180081979
throughput	4756305	931962600
meta	398912	566252863
owd	1327024436	287458509572
retransmits	5213794	912109360
status	114759	11044505
trace	114591259	62580023345
stashcache	11768274	2003542425
x1t	352443664	91152621191
xrd	31910045	12916769463
total	2775427550	578574612106
		538.8395974 GB

2.7 Billion documents taking 538 GB of space x 2 as we have two copies.

Midscale Service Catalog

- Will endeavor to create list of services provided by our lab and coordinate/track with Jeff as appropriate.
- OSG midscale services are hosted on a mix of infrastructure, all provided by UChicago or the experiments, and managed by our group.
- http://bit.ly/maniac-osg-services

Additional OSG Services

- OSG Flock Host – Moved gwms flocking target from IU to UChicago
- xd-login – Moved xd-login host from IU to UChicago
- Seven login servers:
  - Three for general public – One EL6 and two EL7
  - Four for specific use cases – fsurf, Duke CI-Connect, UChicago CI-Connect, CMS Connect

Additional OSG Services

- Two training hosts
- Hosted CEs – Currently hosting 22 hosted CEs
- StashCache – 6 servers
- FTS/Rucio – 3 FTS servers
- Stratum–R – Server w/ 50 TB disk array to replicate CVMFS at TACC, BlueWaters
- Misc. – 6 additional servers for fsurf, website, local HTCondor pool, etc.

DevOps for the midscale

- As these experiments are somewhat more flexible, and have a need (little manpower), service-oriented DevOps can be a tool
- SLATE, VC3 projects would like to work with OSG on containerization
  - https://docs.google.com/presentation/d/1KT2xdUmI4wDDMr0Xp7ueShkmRC4L_ojZTAFPMNHQykY/edit?ts=5be1cbbe#slide=id.g45ba6a3f7e_0_0