Services in Support of Multi-Institutional Science

2019 Joint HSF/OSG/WLCG Workshop
March 18-22 2019

Open Science Grid

Pascal Paschos, Ph.D.
University of Chicago

Outline

- Services at Scale
- OSG Connect, *-Connect Services
- Hosted CE Services
- Data Management Service: Rucio
- StashCache Data Federation
- VC3: Virtual Clusters for Community Computation
- SLATE
- Stratum-R
- GRACC
- Expertise as a service
- Summary

Services at scale

- Multi-institutional and multi-disciplinary academic research is complex and stratified creating barriers for integrated service deployment and support
    - Policies and infrastructure vary from site to site
    - Lack of strategic vision and fear of complex solutions
    - Research is collaborative but often siloed in
- Services that enable and support research must be engaging, innovative, adaptable, elegant and have strategic depth in vision and capacity, while respecting the individuality and priorities of local institutions
- Open Science Grid and it's partners can bring infrastructure supported science to academic research and enable coordination between hubs of scientific research at scale: from a single user to single user engagement to large collaborative projects

OSG Connect, *-Connect

- Help Desk and Collaboration Support
- Software Support
  - OASIS (OSG Application Software Installation Service via CVMFS)
  - Singularity containers
- Data Management: Storage and Transfer
  - HTCondor, StashCache, GridFTP, Globus
- Collaborative or Individual Projects
  - Consultation services
  - XSEDE user access of OSG infrastructure through the OSG-XD login
- Submission instances deployed for specific **collaborations**
  - XENON, SPT-3G
  - CMS Connect, ATLAS Connect

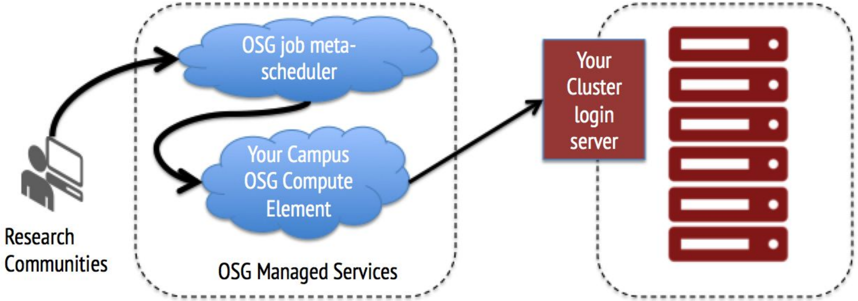
Hosted CE services

- An OSG Compute Element (CE) allows factory submitted pilot jobs from the grid to land and execute on a local cluster. A gateway installed software authenticates and authorizes incoming pilot job containers which then run on the local resource.
- Setting up a CE at the front end of a local cluster is *not* an easy process
- Academic Clusters do not necessarily use HTCondor as a scheduler and local IT teams are unlikely to adopt a scheduler they are not familiar with

.

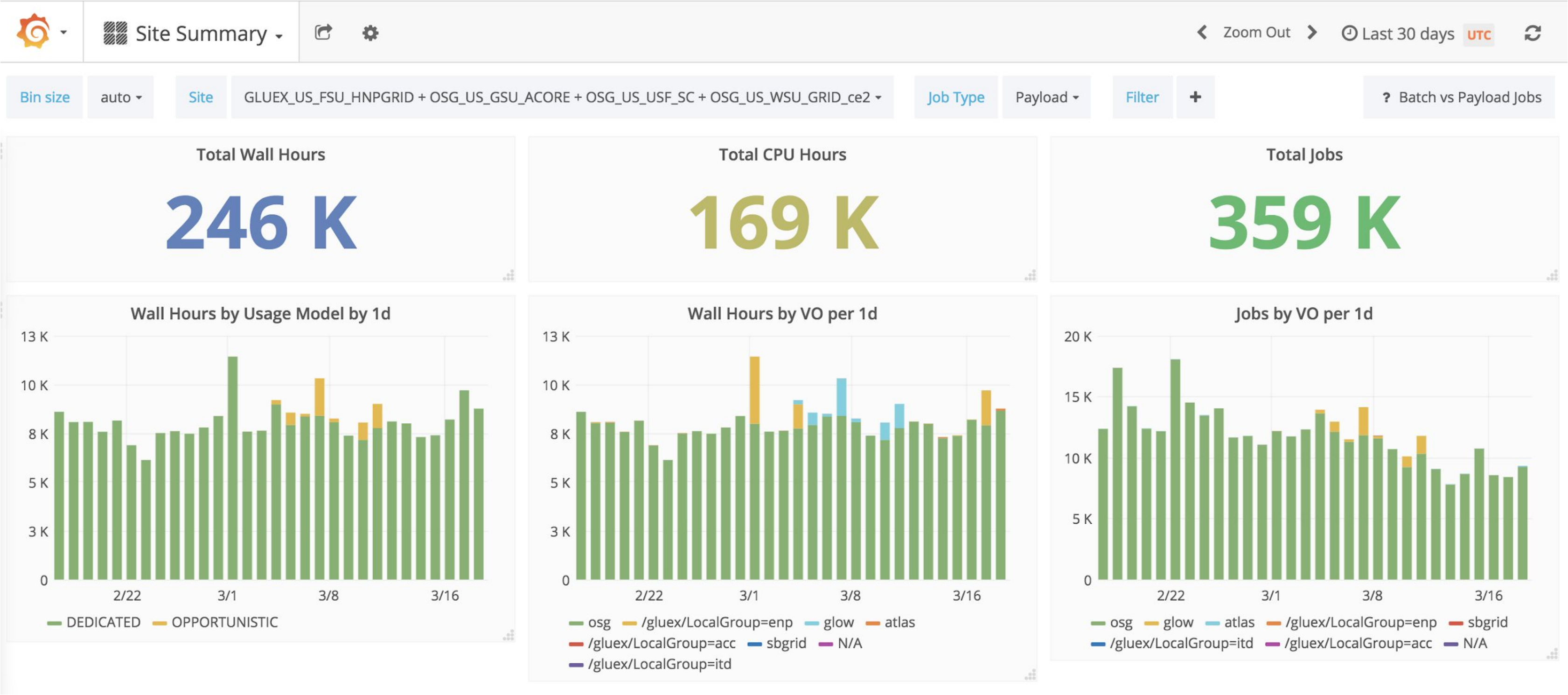
Hosted CE services

- To facilitate onboarding communities of researchers to OSG without heavy local admin involvement a facility can request an OSG Hosted Compute Element
- Local cluster installs HTCondorCE which runs condor_job_router to transform jobs for the local batch system
- The Hosted CE then submits pilot jobs to the local cluster over ssh
- There are at present about 20 Hosted CE services at UC connecting OSG to institutions and facilities



- Requirements:
  - Single OSG service account has access to the login server
  - Cluster scheduler server and nodes have outbound IP connectivity
  - Supported operating system: CentOS/RHEL 6.x, 7.x
  - Cluster login should have common batch scheduler deployed (e.g. SLURM, PBS, HTCondor)

GRACC reporting of select Hosted-CEs



Site Summary

Zoom Out

Last 30 days UTC

Bin size: auto

Site: GLUEX_US_FSU_HNPGRID + OSG_US_GSU_ACORE + OSG_US_USF_SC + OSG_US_WSU_GRID_ce2

Job Type: Payload

Filter +

? Batch vs Payload Jobs

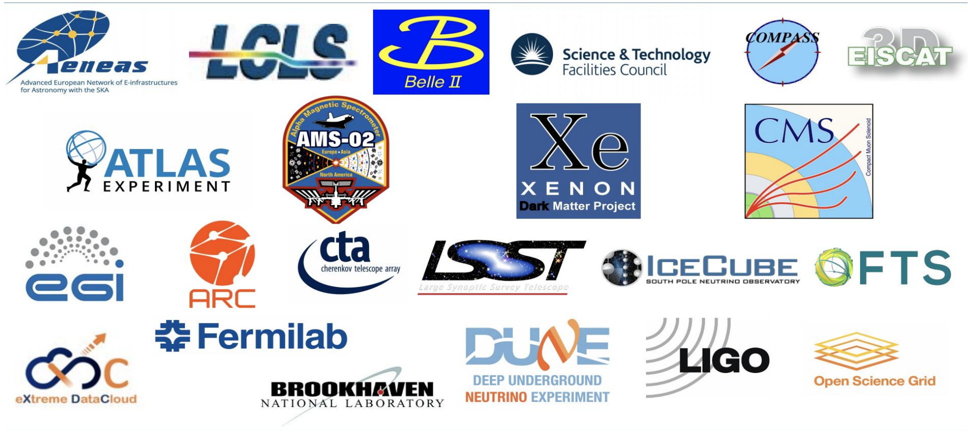
Total Wall Hours: 246 K

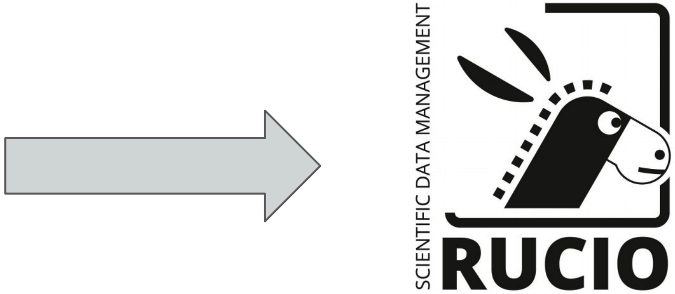
Total CPU Hours: 169 K

Total Jobs: 359 K

Data Management - Working with a Community

- For XENON1T, needed to work with storage providers from EU, US, and Israel
- Needed data management capabilities for the long term
- Important to adopt/join an open source community


Aeneas
Advanced European Network of E-infrastructures for Astronomy with the SKA
LCLS
Belle II
Science & Technology Facilities Council
COMPASS
EISCAT 3D
ATLAS EXPERIMENT
AMS-02
Xe XENON Dark Matter Project
CMS
egi
ARC
cta cherenkov telescope array
LSST Large Synoptic Survey Telescope
IceCube South Pole Neutrino Observatory
FTS
XDC eXtreme DataCloud
Fermilab
Brookhaven National Laboratory
DUNE Deep Underground Neutrino Experiment
LIGO
Open Science Grid

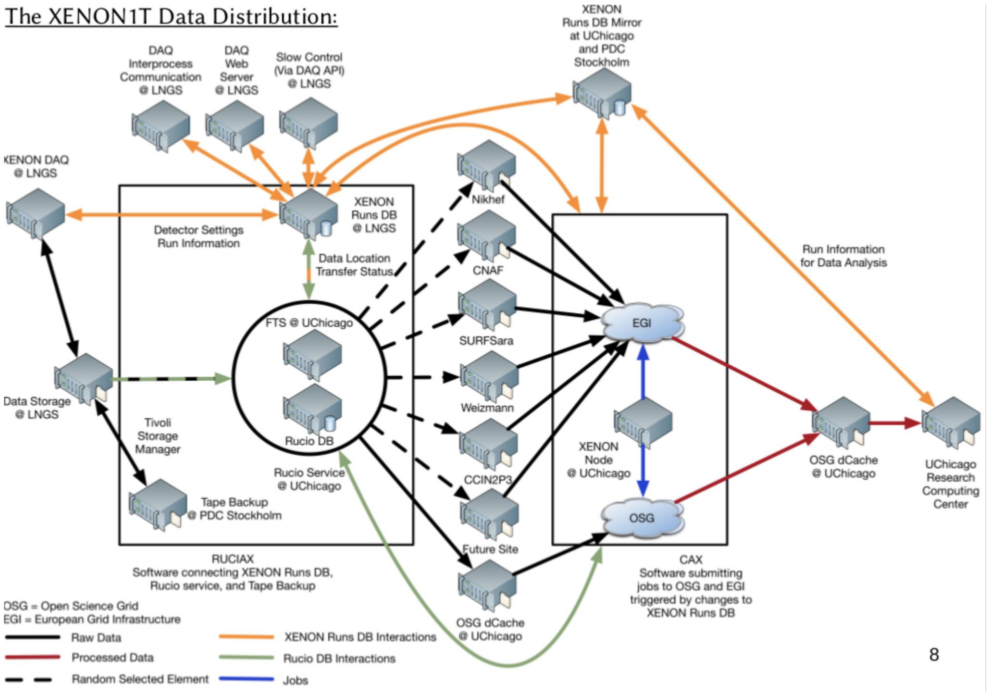


Data Management Service: Rucio

- A scalable data management solution for multi-disciplinary research (e.g., HEP, astronomy, biology).
- Open source, developed for ATLAS experiment at CERN but driven by a community of scientific collaborations
- Works for OSG, EGI, and, cloud storage providers
- Provides namespace (file catalog), reliable transfer service (FTS), storage endpoints, consistency service
- Subscription model for data placement with declarative "rules"
- Adopted by data driven communities for their data management and replication services
- Deployed or evaluation instances for Xenon, CMS, IceCube, LIGO, LSST, FIFE etc

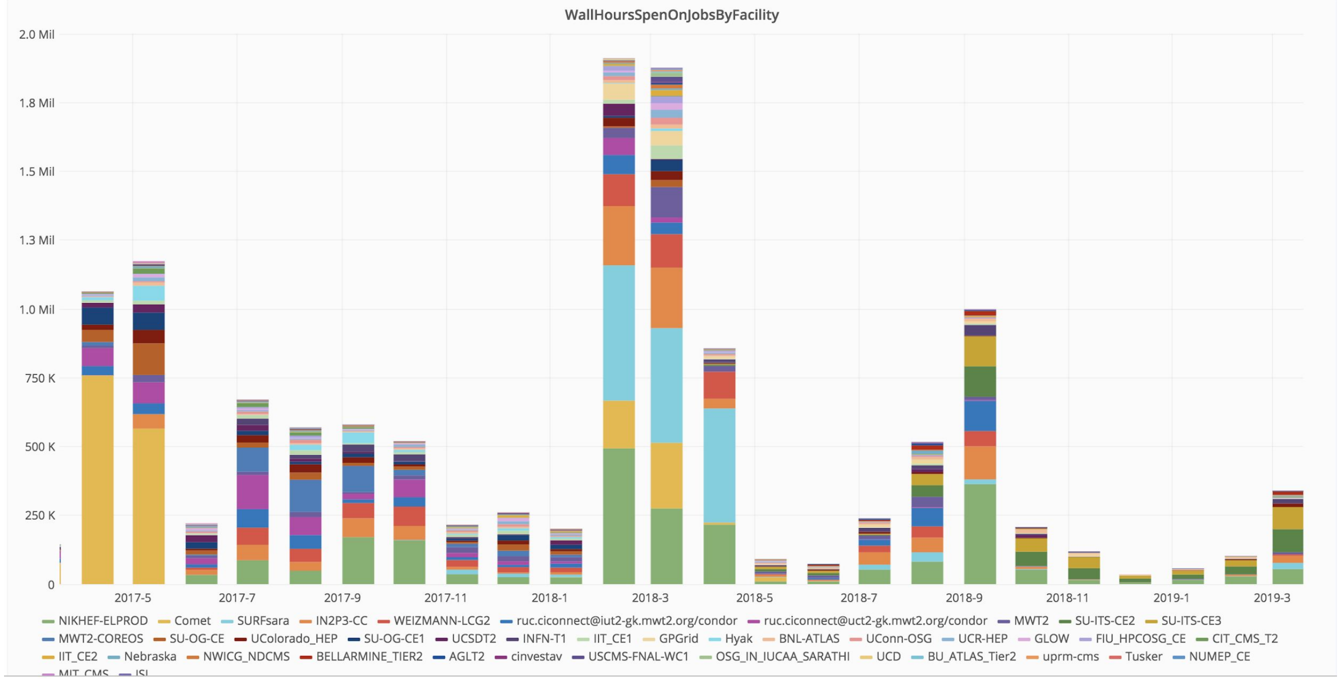
Production Rucio for the Xenon Experiment

- EGI storage distributes data during large data collection periods at the INFN Gran Sasso National Laboratory
- Jobs launched from the xenon node at UChicago. Jobs move through glideinWMS to OSG.
- If lands at EGI, pulls data from EGI; Same with OSG
- Storage and compute pool is expandable
- Rucio/FTS automates the data movement


The XENON1T Data Distribution:
DAQ Interprocess Communication @ LNGS
DAQ Web Server @ LNGS
Slow Control (Via DAQ API) @ LNGS
XENON Runs DB Mirror at UChicago and PDC Stockholm
XENON DAQ @ LNGS
Detector Settings Run Information
XENON Runs DB @ LNGS
Data Location Transfer Status
FTS @ UChicago
Rucio DB
Rucio Service @ UChicago
Data Storage @ LNGS
Tivoli Storage Manager
Tape Backup @ PDC Stockholm
RUCIAX Software connecting XENON Runs DB, Rucio service, and Tape Backup
Nikhef
CNAF
SURFSara
Weizmann
CCIN2P3
Future Site
OSG dCache @ UChicago
EGI
XENON Node @ UChicago
OSG
CAX Software submitting jobs to OSG and EGI triggered by changes to XENON Runs DB
Run Information for Data Analysis
OSG dCache @ UChicago
UChicago Research Computing Center
OSG = Open Science Grid
EGI = European Grid Infrastructure
Raw Data
Processed Data
Random Selected Element
XENON Runs DB Interactions
Rucio DB Interactions
Jobs
8

Job management with HTCondor & workflow pipeline tools

XENON OSG/EGI/XSEDE Usage

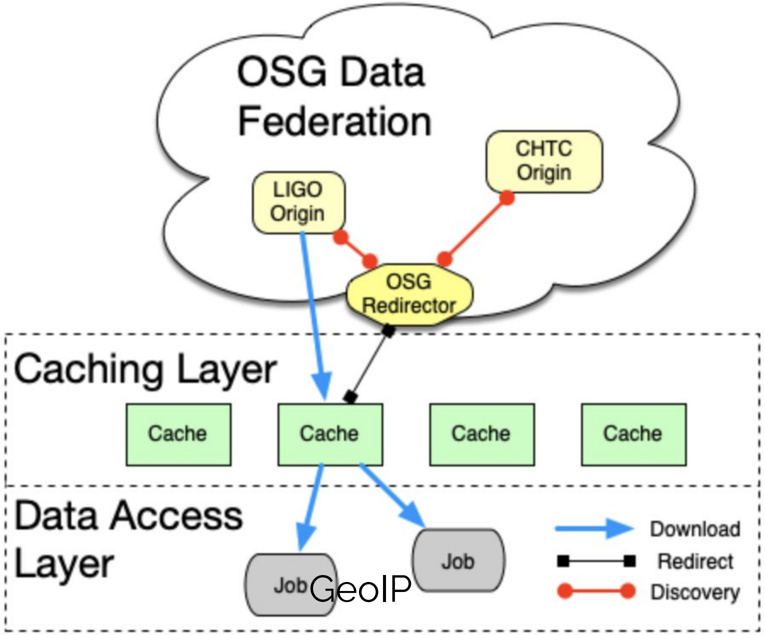


- Combined resources from shared OSG cyberinfrastructure, XSEDE allocations, and allocations on European grid sites
- **Up to** 2M hours per month!

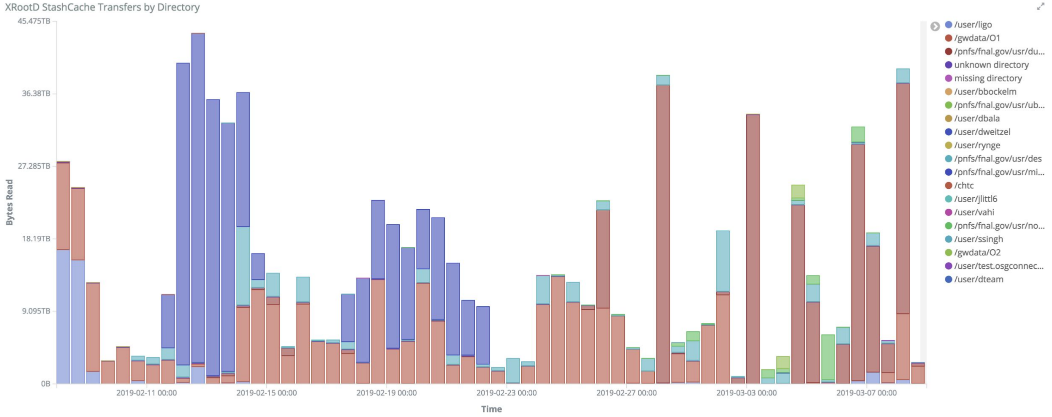
StashCache Data Federation Service

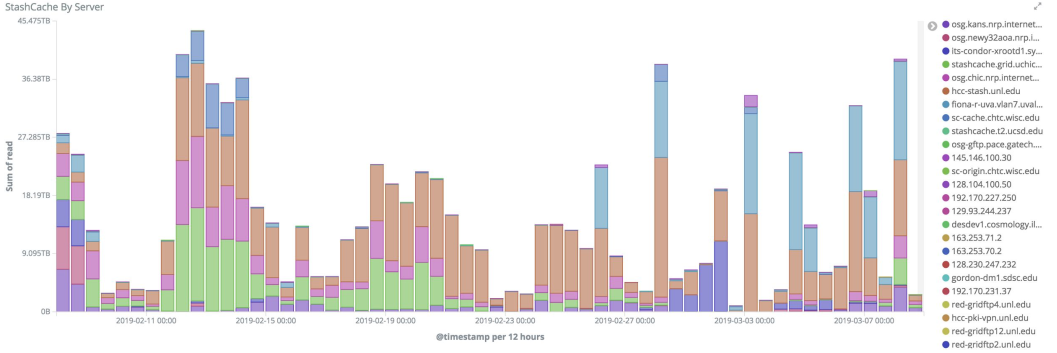
StashCache data federation enables the scalable distribution of research data to large volume of jobs that need them negating the need to prestage data or deploy expensive local storage infrastructure.

- **Origin**: Authoritative copy of the data.
- **Cache**: Transfers data for jobs. A number of caches are operationalized across the OSG aiming for proximity to a compute site which can operate it's own cache.
- **Redirector**: Centrally run service by OSG which instructs the cache to access the proper origin for data.



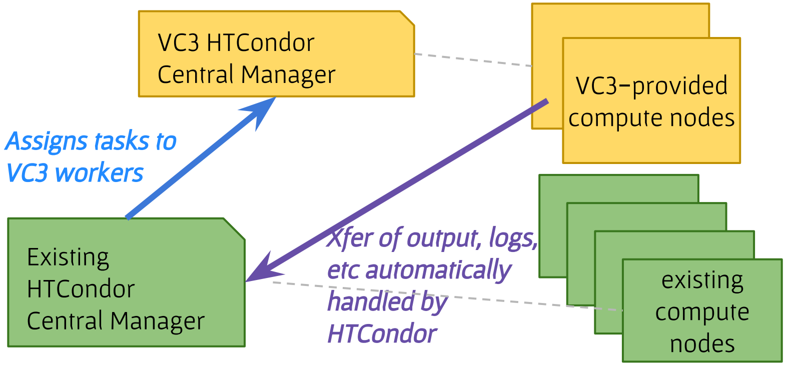
StashCache - over 800 TB read last month





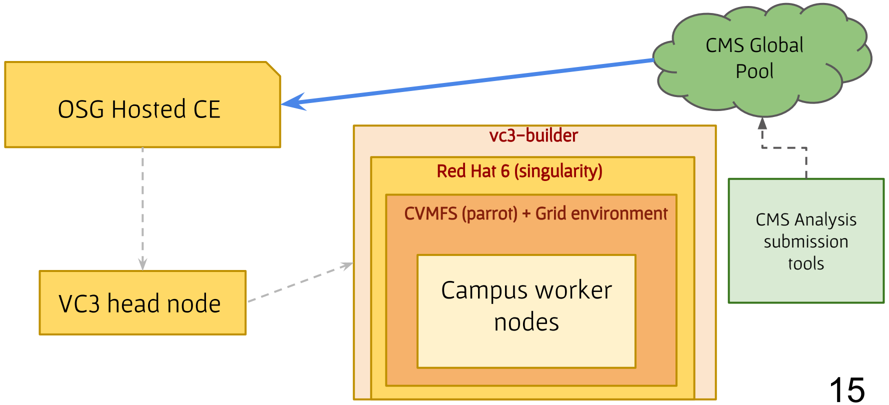
Virtual Clusters for Collaborations

- VC3 (virtualclusters.org) is a platform for provisioning cluster frameworks for collaborative science teams
- Collaborations who have existing HTCondor pools can extend them by adding more worker nodes via VC3 (manually at present)
- Add XSEDE resources, Open Science Grid, and campus HPC clusters
- End-users can transparently use additional resources
- South Pole Telescope (SPT-3G), XENON1T Analysis Framework, IceCube Simulation Framework early users


VC3 HTCondor Central Manager
VC3-provided compute nodes
Assigns tasks to VC3 workers
Existing HTCondor Central Manager
Xfer of output, logs, etc automatically handled by HTCondor
existing compute nodes

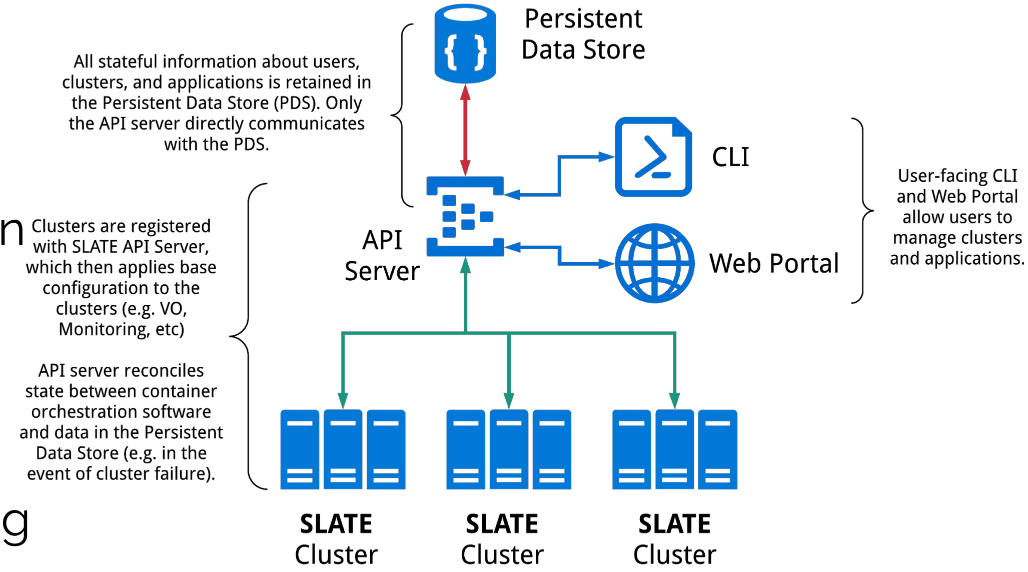
VC3 Deployment example

- Tier-3 campus deployment via VC3 which allows the provisioning at user-level - no root access - of:
  - The CERN File System (CVMFS) (via parrot virtual filesystem)
  - The OSG grid environment on the worker nodes (via CVMFS)
  - Customized Operating Systems (via singularity)
  - Integrated OSG Compute Element (CE) with the VC3 submit host



SLATE: Services Layer At The Edge

- A DevOps friendly Kubernetes based distributed service orchestration platform
- **Federation software & model** for independent Kubernetes clusters
- Site autonomous, allows for user controlled Kubernetes configuration
- Single entrypoint using institutional identity with UNIX-like permissions model
- Scale resources on demand
- Applications are curated in a catalog
- An infrastructure, and software



All stateful information about users, clusters, and applications is retained in the Persistent Data Store (PDS). Only the API server directly communicates with the PDS.

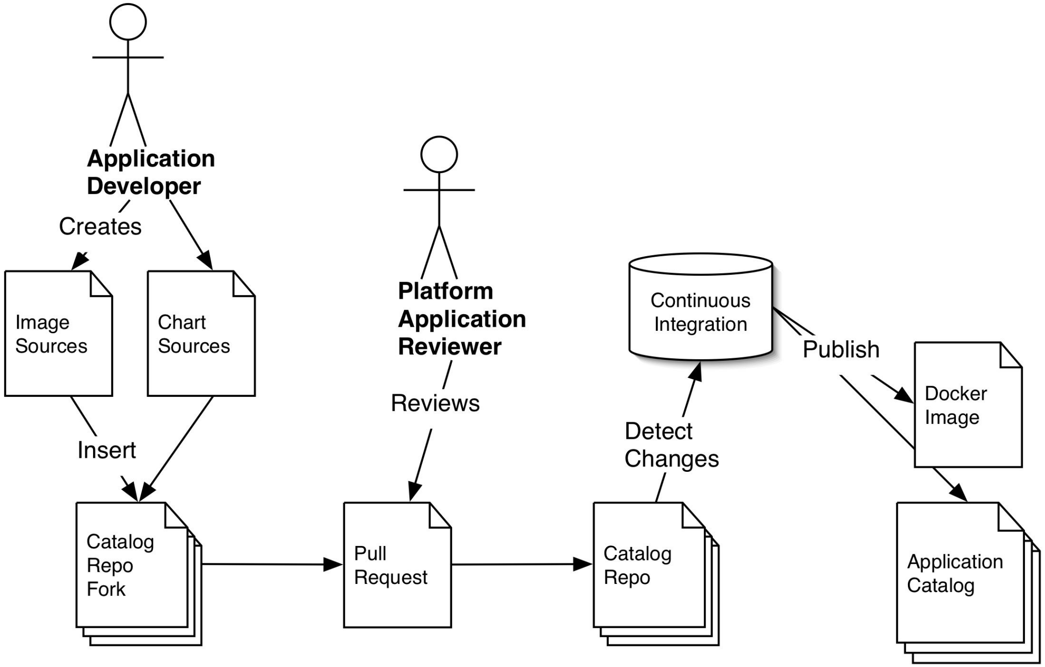
User-facing CLI and Web Portal allow users to manage clusters and applications.

Clusters are registered with SLATE API Server, which then applies base configuration to the clusters (e.g. VO, Monitoring, etc)

API server reconciles state between container orchestration software and data in the Persistent Data Store (e.g. in the event of cluster failure).

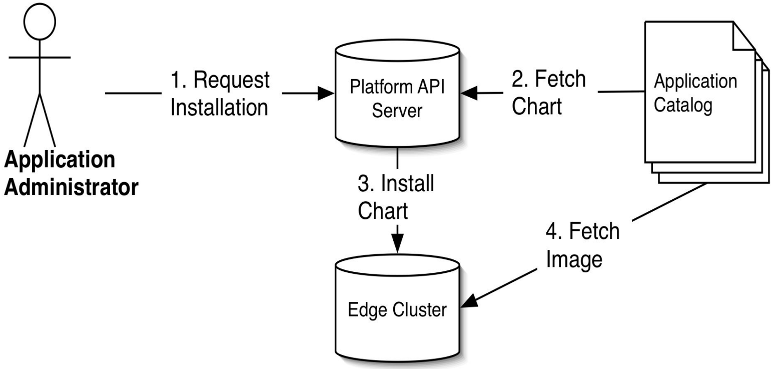
Security, Policy on Federated Edge Systems

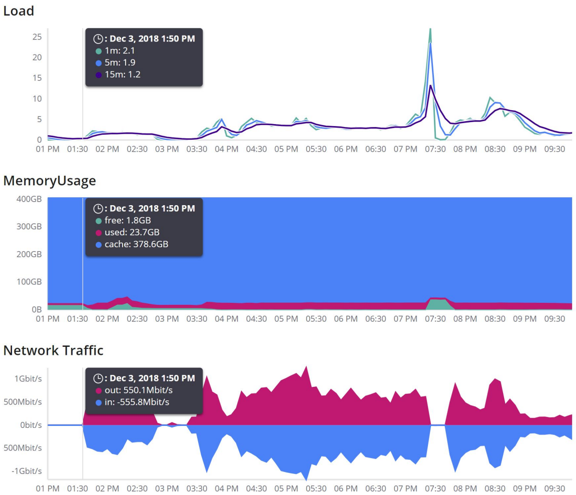
- Application developers create Helm charts and Docker images
- The SLATE team is developing a trust model for acceptance of these containerized "apps"
- Curated application catalog provides point of review for widespread deployment



SLATE Deployment

- **Trusted group model**: experts ("app admins") from a stakeholder or a service organization can deploy and operate services across multiple sites
- Early deployments for ATLAS XCache - considered for WLCG "Lightweight Sites" (c.f. Thurs session)
- Working with OSG & PRP to **containerize & curate** services to more easily integrate campus resources
  - frontier-squid
  - HTCondor services: CE, submit
  - StashCache
- Other "apps" e.g. JupyterHub useful for training events





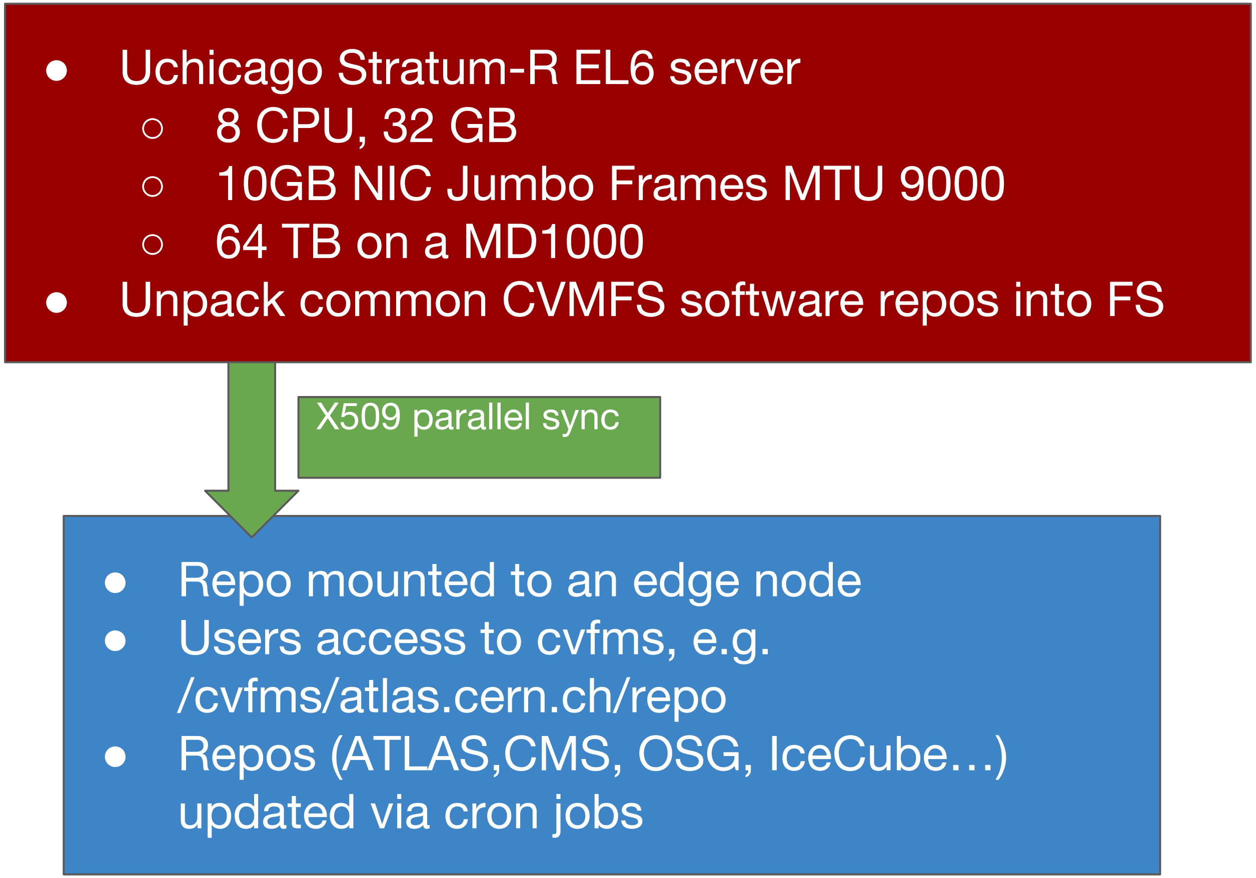
Stratum-R overview

- Allow worker nodes to access CVMFS repos on sites without CVMFS client
  - Repos unpacked by server and stored as Linux FS at target site
  - Particularly useful for HPC sites where CVMFS client impossible
- Problem: Direct replication of CVMFS repos to Linux FS not feasible
  - Large repo size: Limited by network latency, load on server
- Solution: Build Stratum-1 server, add local client to use server as source
  - Parallel rsync to unpack repos into FS (or container images) Second rsync from Stratum R to HPC edge node
  - FS mounted on all worker nodes, symlink between repo area and cvmfs
  - Voila! User can log into remote site and use repos as normal

Credit: Marc Weinberg

Stratum-R deployment Example (US-ATLAS on Stampede2)

THE UNIVERSITY OF CHICAGO

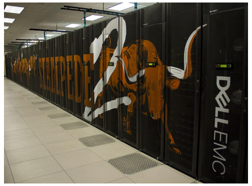


- Uchicago Stratum-R EL6 server
  - 8 CPU, 32 GB
  - 10GB NIC Jumbo Frames MTU 9000
  - 64 TB on a MD1000
- Unpack common CVMFS software repos into FS

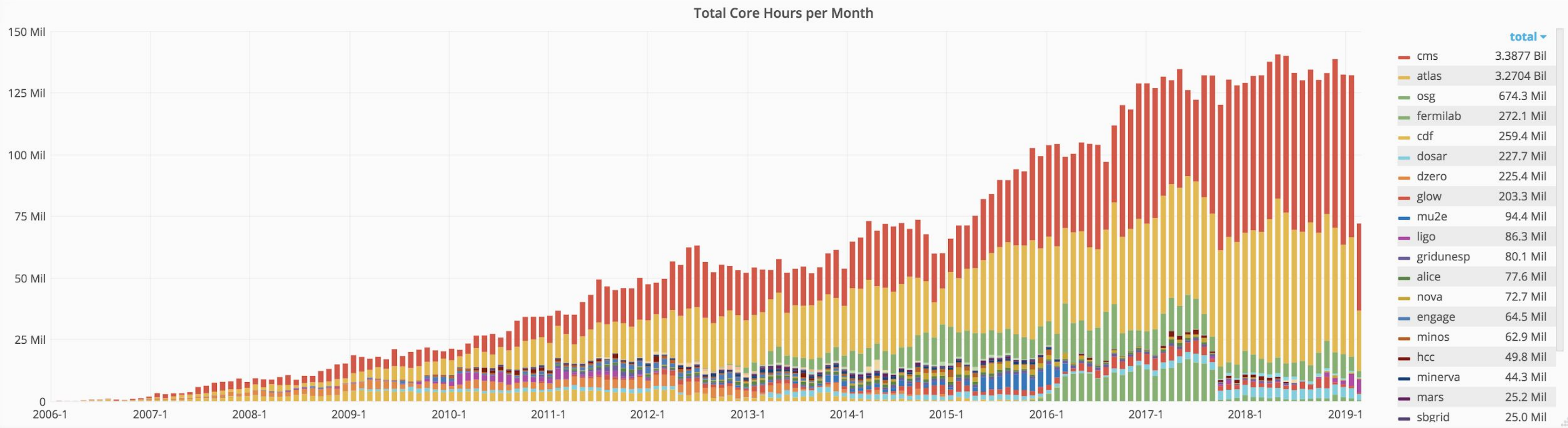
X509 parallel sync

- Repo mounted to an edge node
- Users access to cvfms, e.g. /cvfms/atlas.cern.ch/repo
- Repos (ATLAS,CMS, OSG, IceCube…) updated via cron jobs

TACC
TEXAS ADVANCED COMPUTING CENTER



GRACC Grid Accounting Service


Total Core Hours per Month

	total
cms	3.3877 Bil
atlas	3.2704 Bil
osg	674.3 Mil
fermilab	272.1 Mil
cdf	259.4 Mil
dosar	227.7 Mil
dzero	225.4 Mil
glow	203.3 Mil
mu2e	94.4 Mil
ligo	86.3 Mil
gridunesp	80.1 Mil
alice	77.6 Mil
nova	72.7 Mil
engage	64.5 Mil
minos	62.9 Mil
hcc	49.8 Mil
minerva	44.3 Mil
mars	25.2 Mil
sbgrid	25.0 Mil

Active Virtual Organizations	Active Sites	Represented Fields of Science
56	2018	24

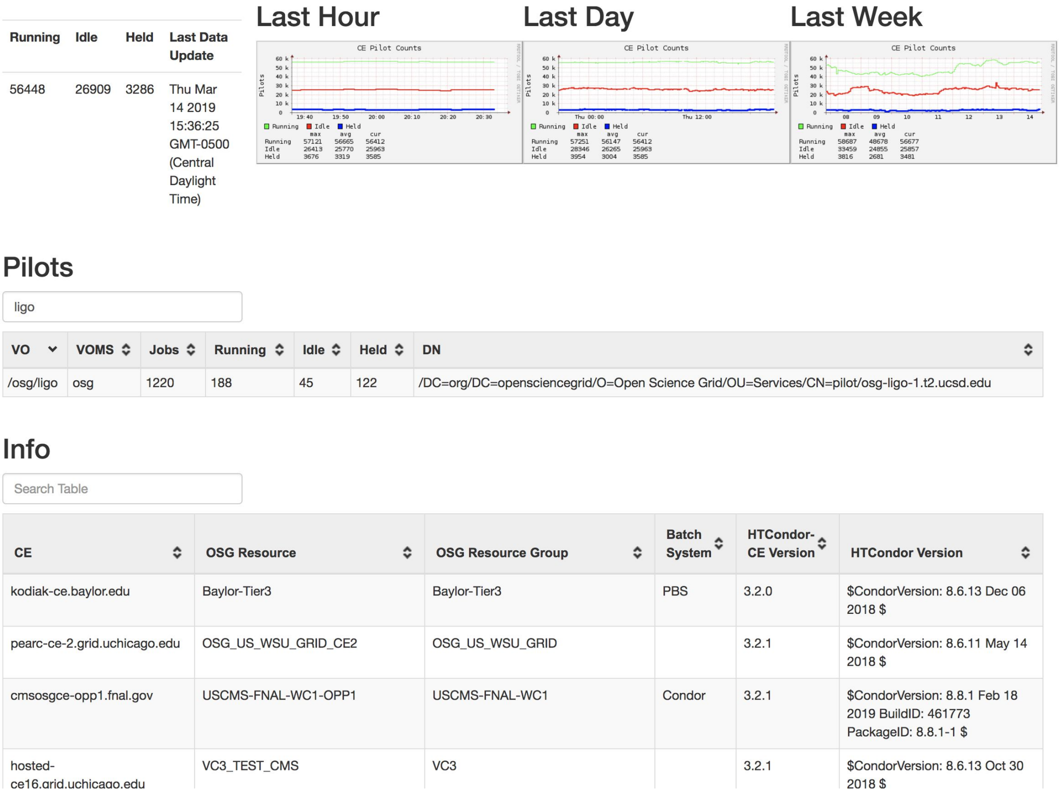
GRACC Grid Accounting Service

- Provides a view of grid utilization metrics which are collected, indexed and displayed over a Grafana interface
- Allows for inspecting summary statistics at scale, from individual users to projects, to facilities and collaborations
- Enhances multi-institutional engagements by allowing for a data driven evaluation of Grid resource utilization, performance, comparison and added-value in leveraging a shared resource
- Enables research domain specific metrics to be captured and repurposed in support of proposals and strategic planning by facilities, administrators and researchers

Central Collector Service

The central collector services allows for a real time monitoring of the status of pilot jobs on the HTCondor-CEs in the OSG

Contains information about site queues - number cores, max memory, walltime, etc



Running	Idle	Held	Last Data Update
56448	26909	3286	Thu Mar 14 2019 15:36:25 GMT-0500 (Central Daylight Time)

Last Hour

CE Pilot Counts

	max	avg	cur
Running	57121	56665	56412
Idle	26413	25770	25963
Held	3676	3319	3585

Last Day

CE Pilot Counts

	max	avg	cur
Running	57251	56147	56412
Idle	28346	26265	25963
Held	3954	3004	3585

Last Week

CE Pilot Counts

	max	avg	cur
Running	58687	48678	56677
Idle	33459	24855	25857
Held	3816	2681	3481

Pilots

ligo

VO	VOMS	Jobs	Running	Idle	Held	DN
/osg/ligo	osg	1220	188	45	122	/DC=org/DC=opensciencegrid/O=Open Science Grid/OU=Services/CN=pilot/osg-ligo-1.t2.ucsd.edu

Info

Search Table

CE	OSG Resource	OSG Resource Group	Batch System	HTCondor-CE Version	HTCondor Version
kodiak-ce.baylor.edu	Baylor-Tier3	Baylor-Tier3	PBS	3.2.0	$CondorVersion: 8.6.13 Dec 06 2018 $
pearc-ce-2.grid.uchicago.edu	OSG_US_WSU_GRID_CE2	OSG_US_WSU_GRID		3.2.1	$CondorVersion: 8.6.11 May 14 2018 $
cmsosgce-opp1.fnal.gov	USCMS-FNAL-WC1-OPP1	USCMS-FNAL-WC1	Condor	3.2.1	$CondorVersion: 8.8.1 Feb 18 2019 BuildID: 461773 PackageID: 8.8.1-1 $
hosted-ce16.grid.uchicago.edu	VC3_TEST_CMS	VC3		3.2.1	$CondorVersion: 8.6.13 Oct 30 2018 $

Expertise as a service

- Availability of technical expertise safeguards the successful deployment and operationalization of other services
- Training of researchers in expertise development improves efficiency and coordination between end-points and centrally managed services
- Research and the path to discovery is not enabled by just a single technological solution but by architecting a synergy of service elements
- Example: South Pole Telescope (SPT) experiment

Example: SPT (South Pole Telescope)

- OSG staff setup and maintain infrastructure for the collaboration at the South Pole
- Provision and maintain Stash node, software repo, manage data transfers, online analysis and storage capacity
- These services come together in enabling the success of this multi-institutional project and streamline the operational robustness of the computational and data management workflow

Summary

- Services presented today
    - Aim in advancing shared infrastructure deployment and enable collaborations
    - Remove the visibility of layers of complexity, promote and empower academic engagement and foster the potential of growth in a restrictive funding landscape
- Seek improvement and increased adoption rates through success stories