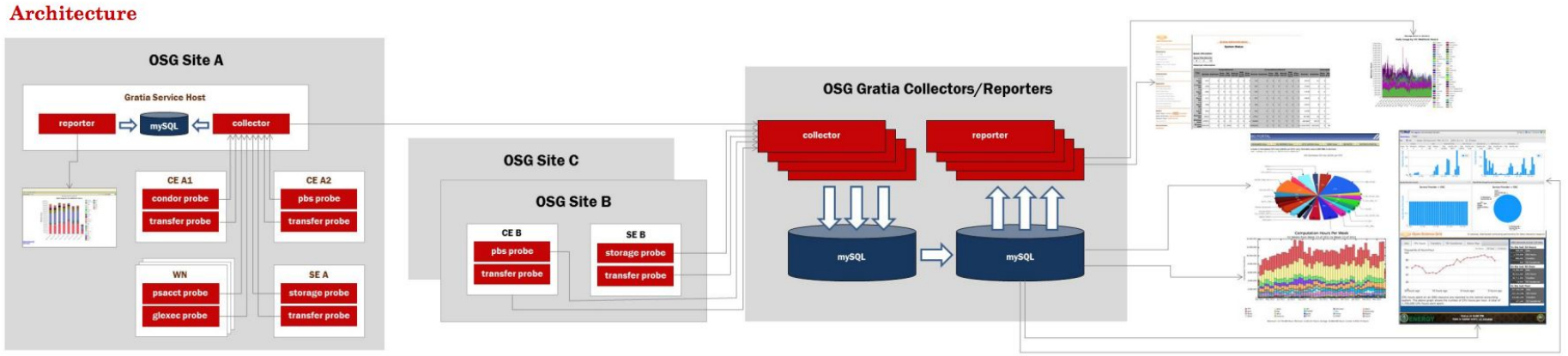
Grid Accounting

...

Tanya Levshina, Kevin Retzke, Juan Mosquera Morales

Part I - Gratia, a beloved relic

Gratia Architecture



Who is using Gratia?

- DOE Annual Report
- WLCG report
- XSEDE report
- OSG ET
- Some site admins, VO admins

Gratia Overview

Gratia Service consists of several subsystems:

- Collector
- Reporter (WEB UI for admins)
- Database (MySQL)
- Email reports generating scripts
- Information is generated by various probes and sent to Gratia collectors via Gratia API. Probes collect information about:
    - Batch and glide-in jobs (condor, lsf, pbs, sge, slurm)
    - Various Metrics (RSV probes)
    - File transfers
    - Storage Usage and Allocation
    - Cloud accounting (OneNebula and AWS deployed at Fermilab)
- GratiaWeb and OSG Display

Gratia supports multiple collectors. It permits hierarchical forwarding & filtering between collectors. This feature is used at UNL and other sites and allows these sites to filter jobs submitted by local users from OSG users.

Implementation Details

- Tomcat 6
- JMS Queue
- Hibernate 4
- Database (MySQL) 5.6
- Java 7 (43, 358 lines of code), JSP (3,560), python (36,832), perl(1,858), shell (6,420)
- Build tools: ANT
- GratiaWeb: CherryPy, google chart, ...
- Packaging: rpm Repository: SVN (http://sourceforge.net/projects/gratia/)
- Ticket tracking: JIRA (https://jira.opensciencegrid.org/browse/GRATIA and https://jira.opensciencegrid.org/browse/GRATIAWeb)

Gratia Stats (I)

OSG collectors since 2005:

- OSG collector (Started at 2004/09):
  - 7.4 M records in SummaryData (jobs and pilot jobs)
  - Stopped removing JobUsageRecord since 2015/01 - after hibernate upgrade Number of records now : ~380M (rate ~1M/day)
  - Database size 962.6 GB
- OSG Transfer collector (Started at 2008/04):
  - 41.9M records in TransferSummary
  - JobUsageRecord (transfer records) - 815M
  - Database size 1005.2 GB
- OSG ITB
  - 75K records in SummaryData (jobs and pilot jobs)
  - 14K records in TransferSummary
  - JobUsageRecord: 2.7M
  - Database size 42.5 GB

Gratia Stats (II)

- 122 sites are reporting 144 batch jobs (Site could have several entry points)
  - Condor probes 114
  - PBS probes 21
  - LSF probes 4
  - SGE probes 5
- 46 campus Grid probes (probe reports actual user jobs):
- Number of SEs reporting gridftp transfers 66
  - gridftp-transfer probes 65
  - dcache-transfer probes 2
  - Very few T2 and T3 sites setup and enabled transfer probe.
- BDII information (Compute Element) ???

Current Efforts

- Tanya (20% partially paid by OSG)
- Kevin (30% Fermi)
- Carl (OSG Software)
- Juan (100% OSG)
- Database group and GCO operation (3% reported?? charged to OSG)

Issues

- Upgrade to hibernate 4 broke housecleaning. Cannot delete JobUsageRecord doing bulk deletion. Surviving by truncating JobUsageRecord_Xml table. Will not be able to do it for long. (The remedy: find bulk deletion patch and rebuild hibernate)
- Upgrade to MySql 5.6 has demonstrated “bad” design choices (missing index on auxiliary tables)
- Started to collect Storage and Tape data for AAF project, AWS VM with charges and other new info for HEP Cloud but unable to summarize these because there is no Summary Table for these records.
- Any changes in schema - major nightmare
- You cannot expect that software designed in 2004 will be really adequate in 2016!

Part II - Next Generation Accounting

Generic requirements

- All historical summary data (job records and transfer) needs to be preserved
- Data should be archived.
- It should be possible to extract the historical data in suitable format in order to upload to the future accounting service.
- Gratia probes should not be changed drastically and new service should be able to deal with older version of probes.
- Should be able to use WebUI to extract the accounting information.

OSG Requirements

OSG:

- Daily summary of job wall duration, cpu usage? per site, VO, Project, DN (with role), user, exit code for Batch and BatchPilot resources
- Daily transfer summary (size and wall duration) per storage site, vo, user, direction, exit code.
- Ability to aggregate at different levels (site versus cluster versus CE) and rename site aggregation.

Fermilab requirements

- AAF Project is using Gratia to report charges to customers that based on several quantities:
  - The total accumulated amount of tape storage used at the end of the year
  - The amount of tape media that needs to be acquired for the year
  - The amount of tape-drive hours used
  - Daily transfer summary (size and wall duration) per storage type (enstore and dcache), vo, user, direction, exit code.
- HEPCloud Project is extending the current Fermilab Computing Facility to transparently run on a variety of resources including commercial clouds. It is using Gratia to get historical information about instantiated VMs (wall duration, cpu usage, charges) per vo, instance type, availability zone.
- FIFE Project is using gratia accounting to facilitate experiments preparation for SPPM meetings, get historical information about users efficiency, success rate, data transfer and dCache pool usage. In order to provide this information
  - Daily summary of job wall duration, cpu usage? per site, VO, user, exit code for Batch and BatchPilot resources
  - Daily transfer summary (size and wall duration) per storage site, vo, user, direction, exit code.
  - Collect daily dcache pools utilization per vo

“Other” requirements

- Nebraska
  All internal HCC usage accounting is done using Gratia. Reports to stakeholders and users often include gratia produced graphs. In order to do so the following information is needed:
  - daily summary of job wall duration per site, VO
  - use the *-running probes to show utilization graphs
- CMS
  - According to Dave Mason CMS Tier-1 doesn’t need accounting information but we need to report to WLCG and the only source is Gratia.

Evaluation of Open Source Grid Accounting

- XDMoD
- EGI Accounting (APEL)

see details in google doc.

Look like similar relics as dear old Gratia.

Elasticsearch (ask Kevin about technical details....)

“We were not pioneers ourselves, but we journeyed over old trails that were new to us, and with hearts open. Who shall distinguish?”
J. Monroe Thorington

- UChicago
- CERN
- dCACHE
- CMS DAQ
- ALICE Tier-2 the Torino INFN CC

details in google doc.

Proof of Concept

A minimal collector program was written to accept forwarded bundles of records from an existing collector and index them into Elasticsearch. The collector splits the bundle into individual XML records, and converts the XML into a “flattened” JSON structure. Several instances of this collector were run, with JobUsageRecord forwarding enabled from the following collectors:

1. OSG Main collector (gratia-main-osg.fnal.gov)
2. OSG ITB (gratia-itb-osg.fnal.gov)
3. HEP Cloud/AWS (fermicloud054.fnal.gov)

No changes were made to any probes or existing collectors.

In addition, logstash was used to migrate the complete MasterSummaryData table (7.4 million records) from the main OSG database to Elasticsearch, with foreign key fields denormalized (e.g. VO, Site).

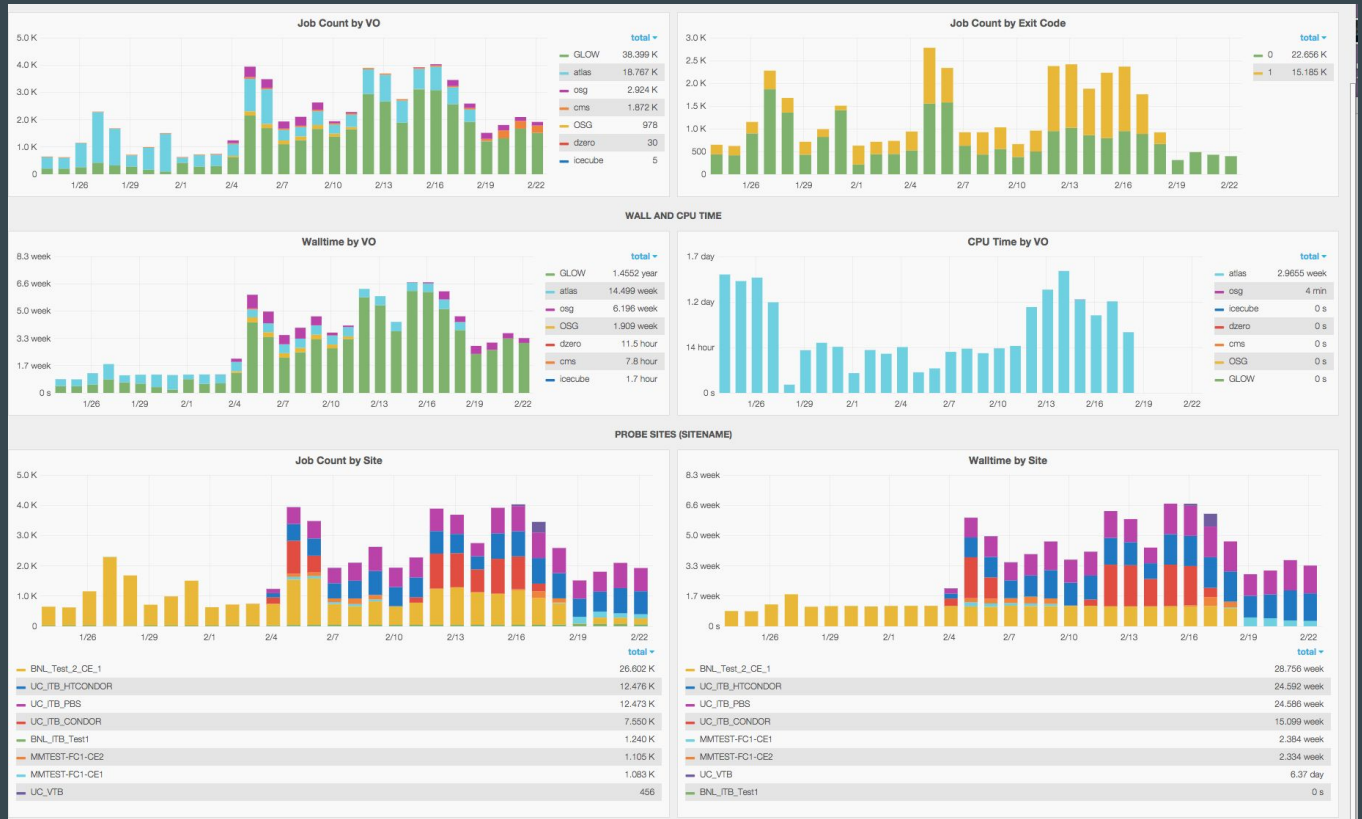
Elasticsearch Development Cluster

- Three-node cluster of VMs running on Fermicloud, each with 4 cores, 8 GB RAM, and 250 GB disk.
- All access through proxy server running on fifemondata.fnal.gov, requires authorized certificate to perform write or admin operations.
- Serving 15K requests/hr
- Usage stats as of 2/22: 168 indices, 83M documents, 38 GB
  - Gratia OSG MasterSummaryData (7.4M records & 3 GB)
  - Gratia OSG JobUsageRecord (1.5M records & 1 GB per day)
  - Gratia ITB JobUsageRecord (200K records & 117 MB)
  - Gratia AWS JobUsageRecord (2.5M records & 1.2 GB)
  - Fifebatch monitoring:
    - Job attributes (2.5M records &1.5 GB per month)
    - Slot attributes (5M records & 2 GB per day)
  - Misc logs and test data

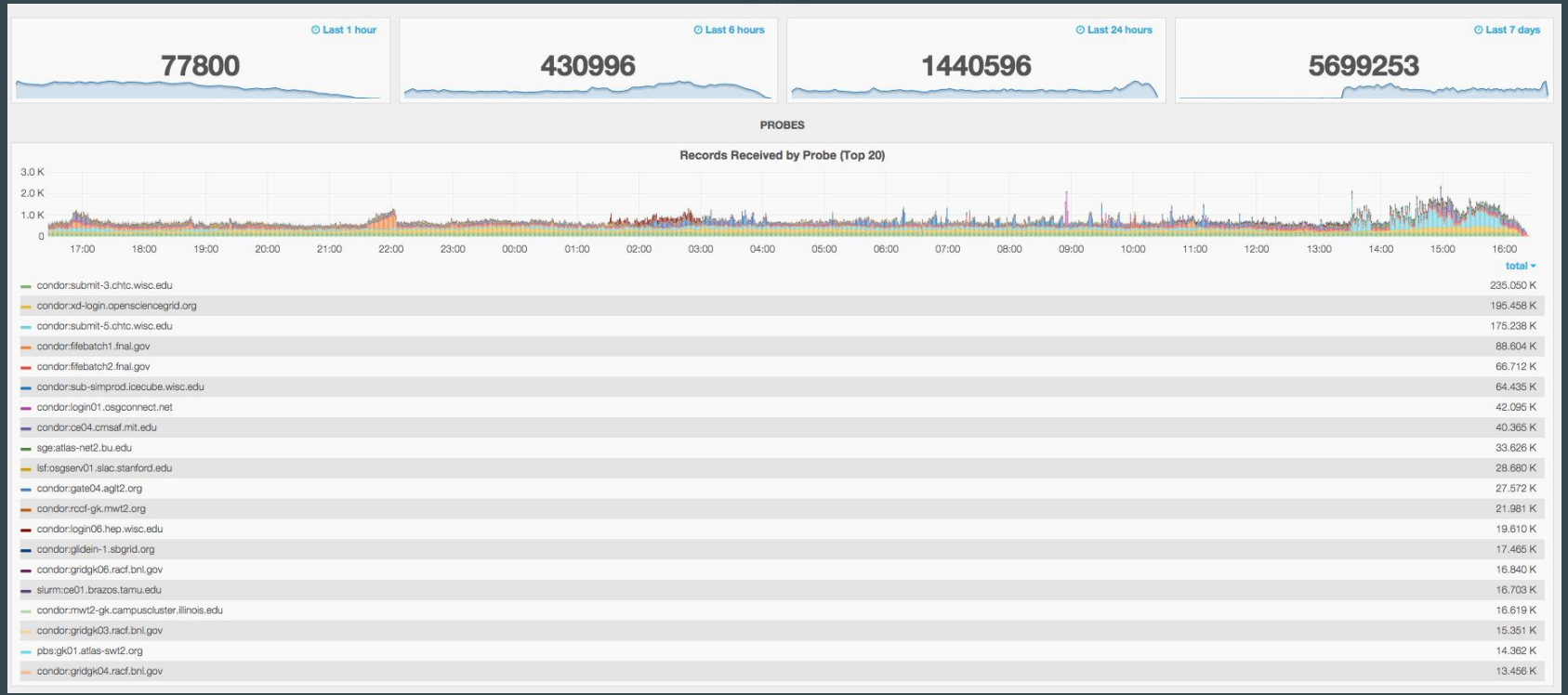
Grafana/FIFEMON

The Grafana dashboard framework is currently used to power Fifemon. It can graph data and generate tables from several sources: Graphite, InfluxDB, Elasticsearch, KairosDB, OpenTSDB, Prometheus, and Amazon CloudWatch.

Example dashboards were created in Grafana displaying the Gratia data from Elasticsearch: OSG JobUsageRecord:



Probe Monitoring Example



Grafana/Elasticsearch vs GratiaWeb/MySQL

The data from MasterSummaryData (about 7 million rows) was migrated to elasticsearch using logstash.

MySQL vs Elasticsearch:

The response time using elasticsearch in comparison with MySQL decreased from minutes to seconds for large amounts of data (10 years).