



OSG GRid ACCounting system :: GRACC

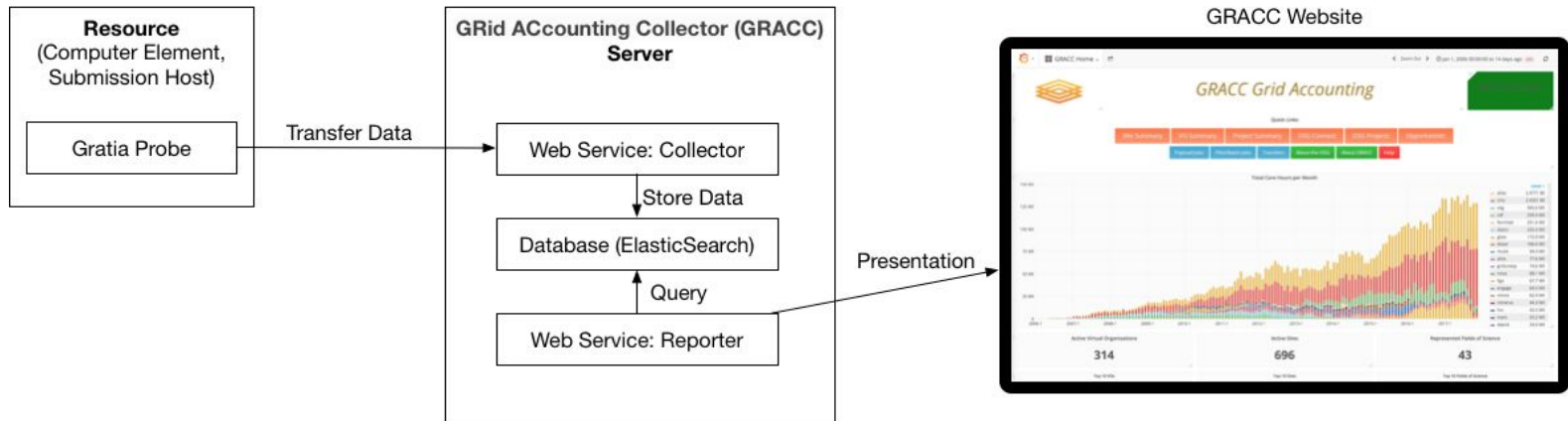
Derek Weitzel, Marian Zvada
Elastic Workshop @FNAL, September 30th, 2019

GRACC - Mapping Jobs to ES

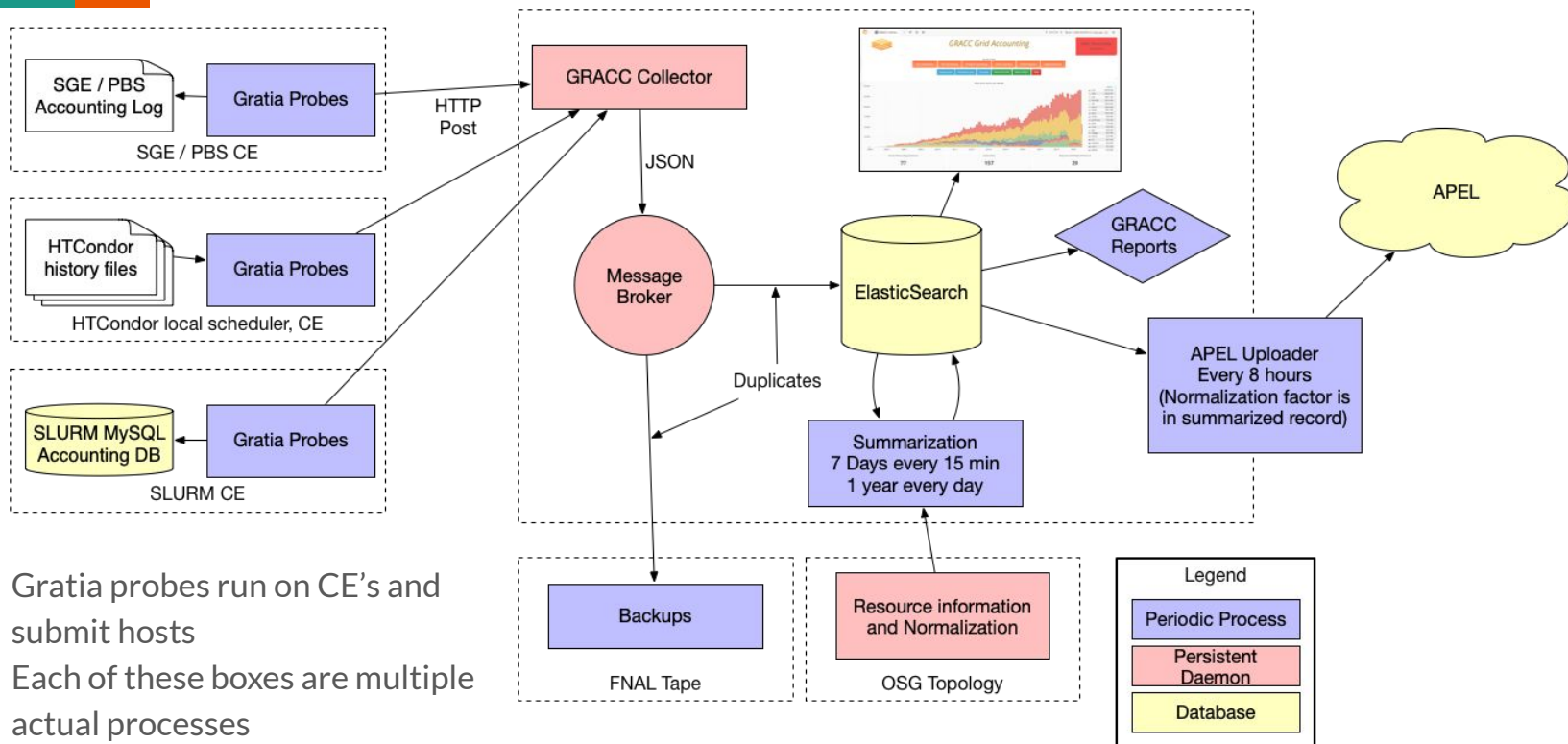
- Each **job** is mapped to a **document** in ES with ~60 attributes each
- GRACC receives 1.2M records a day
- Commodity hardware (and no SSDs)! - ES proved too slow to visualize using raw records over 30+ days.
- Summarized by **bucket**'ing jobs into 1 day periods on specific unique attributes. Summing the usage.
- Enrich the summarized records with outside resource information

GRACC Big Picture

- **Gratia probe:** A piece of software that collects accounting data from the computer on which it's running, and transmits it to a Gratia server.
- **GRACC server:** A server that collects Gratia accounting data from one or more sites and can share it with users via a web page. The GRACC server is hosted by the OSG.
- **Reporter:** A web service running on the GRACC server. Users can connect to the reporter via a web browser to explore the Gratia data.
- **Collector:** A web service running on the GRACC server that collects data from one or more Gratia probes. Users do not directly interact with the collector.



GRACC components architecture



- Gratia probes run on CE's and submit hosts
- Each of these boxes are multiple actual processes

GRACC Collector



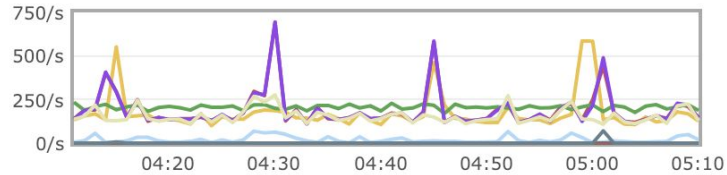
- Program that listens for HTTP **POSTs** from gratia probes.
- Parses a semi-XML format from the POST into JSON
- Places the records onto the message bus for ingestion into ES

Message Bus

Message bus is utilized by GRACC, Network Monitoring, StashCache federation accounting

- Hosted on commercial provider: CloudAMQP
- Monitored through Grafana alerts, and CloudAMQP alerts

Message rates last hour ?



ES Ingestion



- We use Logstash receive from the message bus and insert into ES
- Network ingestion uses custom ingester, and constantly a source of trouble
 - Very difficult to write a correct message bus to ES ingester
 - Many error conditions
 - Correctly confirming to message bus when ingested

Elastic



- Elasticsearch 5.6.5 (really old)
- Read-only ES interface with 2 layers of security
 - NGINX proxy that only allows GET requests, no POST or PUTS...
 - [Read Only Rest](#) instance
- Backups
 - HDFS daily snapshots
- Grafana (4.6.3)
- Kibana (5.6.5)

Interfaces



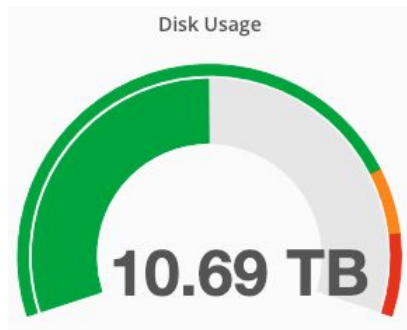
- **Grafana (prod)**
 - Dashboards made for/by stakeholders
- **Kibana - Debug**
 - Used primarily for debug and early prototyping
- **Email Reports**
 - Periodic status updates
 - Queries the Read Only interface with custom query

GRACC technical specs

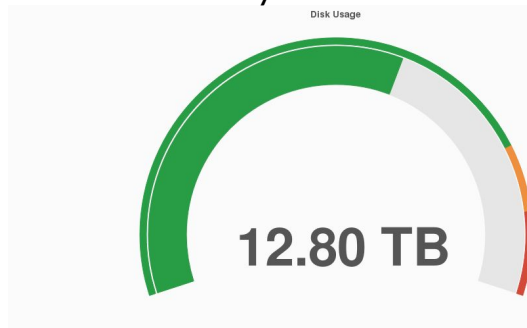
Hardware hosted on OpenStack platform

- ElasticSearch cluster (ELK), CEPH storage
- 1 VM Front-End (64GB RAM, 2TB data volume)
- 5 VMs data nodes (32GB RAM, 5TB data volume)
- With this allocated volume size we're good for another ~3 years

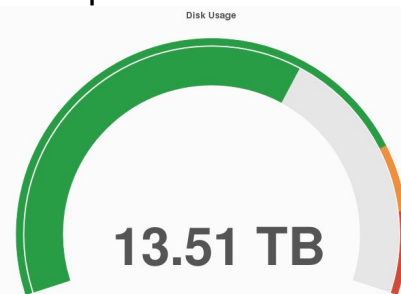
End of Jan 2019



End of July 2019



End of Sep 2019











GRACC

Monitoring

- check_mk with automated notifications

Deployment

- fully puppetized
- docker containers (not for everything)

Local site red, gracc-data5.opensciencegrid.org			
State	Service	Icons	Status detail
OK	Filesystem /	 	OK - 5.1% used (8.08 of 159.99 GB), trend: -13.04 kB / 24 hours
OK	Filesystem /data	 	OK - 39.5% used (1.93 of 4.88 TB), trend: +4.60 GB / 24 hours
Local site red, gracc.opensciencegrid.org			
State	Service	Icons	Status detail
OK	Filesystem /	 	OK - 54.9% used (137.16 of 249.99 GB), trend: -372.48 MB / 24 hours
OK	Filesystem /data	 	OK - 1.14% used (22.74 GB of 1.95 TB), trend: +396.41 MB / 24 hours

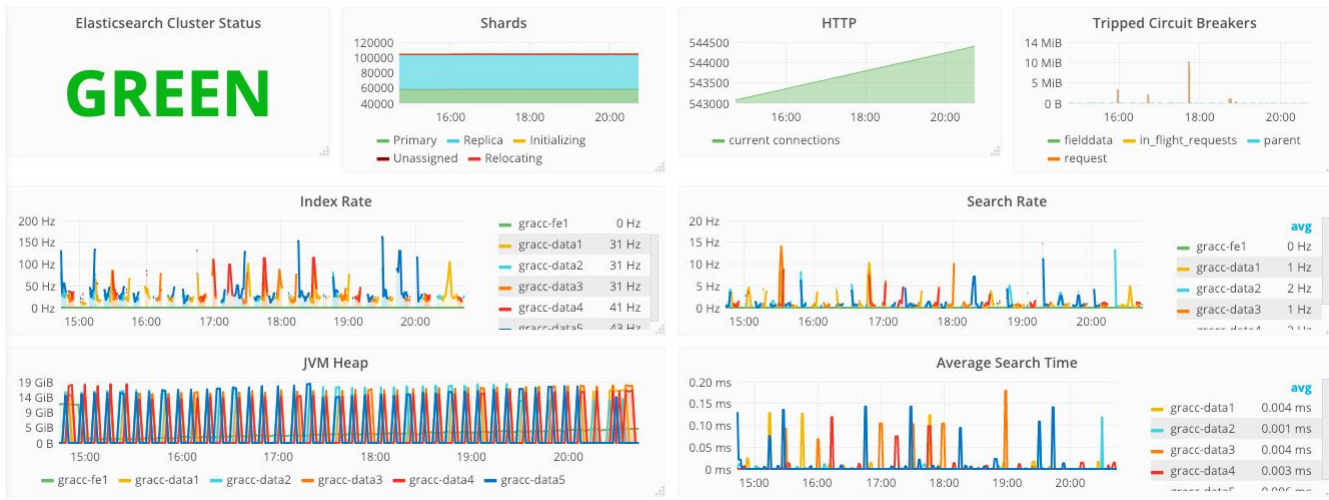
gracc.opensciencegrid.org	systemd: elasticsearch-ro.service	100.00%
gracc.opensciencegrid.org	systemd: elasticsearch.service	100.00%
gracc.opensciencegrid.org	systemd: elasticsearch_exporter.service	100.00%
gracc.opensciencegrid.org	systemd: graccarchive@ps-itb.service	100.00%
gracc.opensciencegrid.org	systemd: graccarchive@ps-prod.service	100.00%
gracc.opensciencegrid.org	systemd: graccarchive@raw.service	100.00%
gracc.opensciencegrid.org	systemd: graccarchive@transfers.service	100.00%
gracc.opensciencegrid.org	systemd: graccbackup@ps-itb.service	100.00%
gracc.opensciencegrid.org	systemd: graccbackup@ps-prod.service	100.00%
gracc.opensciencegrid.org	systemd: graccbackup@raw.service	100.00%
gracc.opensciencegrid.org	systemd: graccbackup@transfers.service	100.00%
gracc.opensciencegrid.org	systemd: graccurator@cvmfs-logs.service	100.00%
gracc.opensciencegrid.org	systemd: graccurator@glidein-logs.service	100.00%
gracc.opensciencegrid.org	systemd: graccurator@htcondor-xfer.service	100.00%
gracc.opensciencegrid.org	systemd: graccsummerperiodic.service	100.00%
gracc.opensciencegrid.org	systemd: graccsummerperiodicyearly.service	100.00%
gracc.opensciencegrid.org	systemd: grafana-server.service	100.00%
gracc.opensciencegrid.org	systemd: kibana.service	100.00%
gracc.opensciencegrid.org	systemd: node_exporter.service	100.00%
gracc.opensciencegrid.org	systemd: prometheus.service	100.00%
Summary		100.00%

GRACC

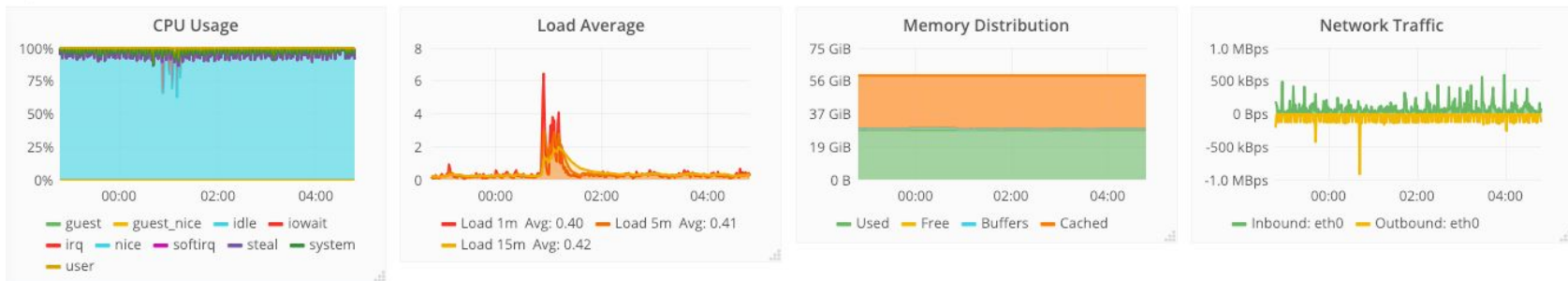
Monitoring dashboards

- status of ES health
- status of nodes

Cluster



gracc-data1.anvil.hcc.unl.edu



Transfer and Cache Accounting

In addition to jobs, we use GRACC for transfer and cache accounting

StashCache Working Set Size by Host and Directory

Cache Hostname ↕	logical_dirname.keyword: Descending ↕	Working Set ▾	Total Read ↕
red-gridftp1.unl.edu	/user/ligo	2.581TB	552.518GB
red-gridftp4.unl.edu	/user/ligo	2.484TB	531.206GB
red-gridftp5.unl.edu	/user/ligo	2.444TB	502.28GB
198.17.101.66	/user/ligo	2.069TB	4.784TB
osg-gftp.pace.gatech.edu	/user/ligo	662.879GB	3.234TB
osg.chic.nrp.internet2.edu	/pnfs/fnal.gov/usr/minerva	346.096GB	6.644TB
osg.kans.nrp.internet2.edu	/pnfs/fnal.gov/usr/minerva	331.676GB	1.839TB
red-gridftp7.unl.edu	/user/ligo	166.961GB	231.755GB
145.146.100.30	/user/ligo	164.316GB	266.428GB
osg.newy32aoa.nrp.internet2.edu	/user/ligo	146.106GB	173.994GB

TCP Transfer Statistics

- Finding network issues between submit hosts and worker nodes
- Using Filebeats for uploading XferLogs from HTCondor

Average Retransmissions by source and destination

Source	Destination	Unique IPs	Average Retransmissions	Average Rordering	Sum of bytes
login.duke.ci-connect.net	cinvestav.mx	1	11698	3	1.818GB
login.duke.ci-connect.net	syr.edu	3	1992	66	5.627GB
login.uscms.org	ac.uk	3	1446.333	220	108.992MB
login.uscms.org	gridka.de	1	515.5	151.5	450.093MB
login02.osgconnect.net	syr.edu	3	317	26.5	1.343GB
login.duke.ci-connect.net	unl.edu	1	234	127	5.066GB
login.uscms.org	org.br	1	138.167	225.917	495.509MB
login.uscms.org	infn.it	1	89.8	218.8	172.332MB
login.uscms.org	jlnr-t1.ru	2	87	31	141.175MB
login.duke.ci-connect.net	iu.edu	1	85	3	239.492MB

Wishlist



- Interested in roll-ups for summarization. Not sure about enriching the records
- Some life-cycle management with Curator, could be expanded

Concerns



- ES can be slow, but it's probably our hosting platform
- We are scared of **drive-by** attacks
- We have done disaster recovery exercises, takes >48 hours to restore the platform and data from snapshots.
 - Likely days from tape...
- We inherit projects from others, and we are scared of ingesters
 - Writing a **good** ingester from message bus to ES is hard, so many error conditions