

Comprehensive Grid and Job Monitoring with Fifemon

...

Kevin Retzke

OSG All-Hands Meeting, March 2016



Landscape



Fermilab

Why Do We Need Monitoring?

Grid admins want to know:

- Overall health of the batch system
- Worker node status and availability
- Efficiency in matching jobs to resources
- Identify and fix problems quickly (before users and stakeholders notice... and open tickets)

Users want to know:

- State of their jobs
- Availability of resources
- WHY ISN'T MY JOB RUNNING?!

Stakeholders want to know:

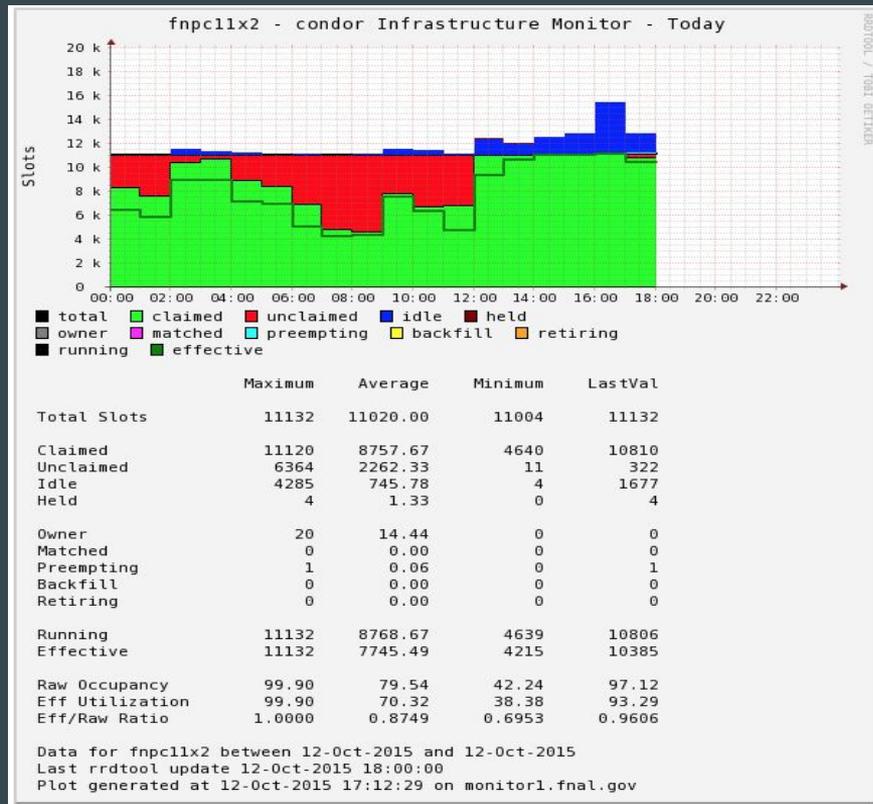
- Each group is getting the resources it needs
- Resources are being used effectively

Fermigrid Monitor (ca. 2004)

Monitoring for local HTCondor grid (GPGrid).

- Aggregate metrics for grid and VO's.
- No offsite information, no user information.
- Difficult to alter or expand.

OK for grid admins, good for stakeholders, bad for users.

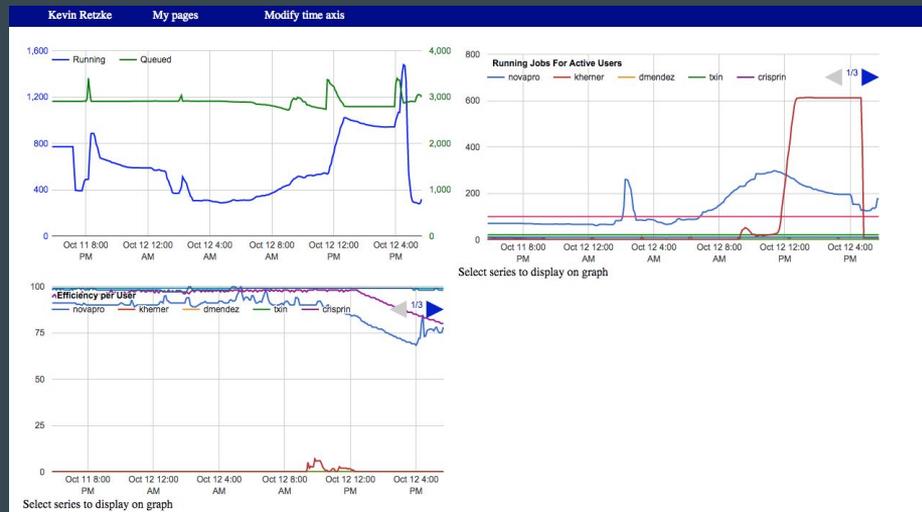


Fifemon v1 (ca. 2014)

Growing usage of offsite resources through OSG; needed new monitoring.

- Aggregate metrics for users and VOs.
- No grid-level information.
- Cumbersome to maintain and expand.

OK for grid admins, bad for stakeholders,
OK for users.

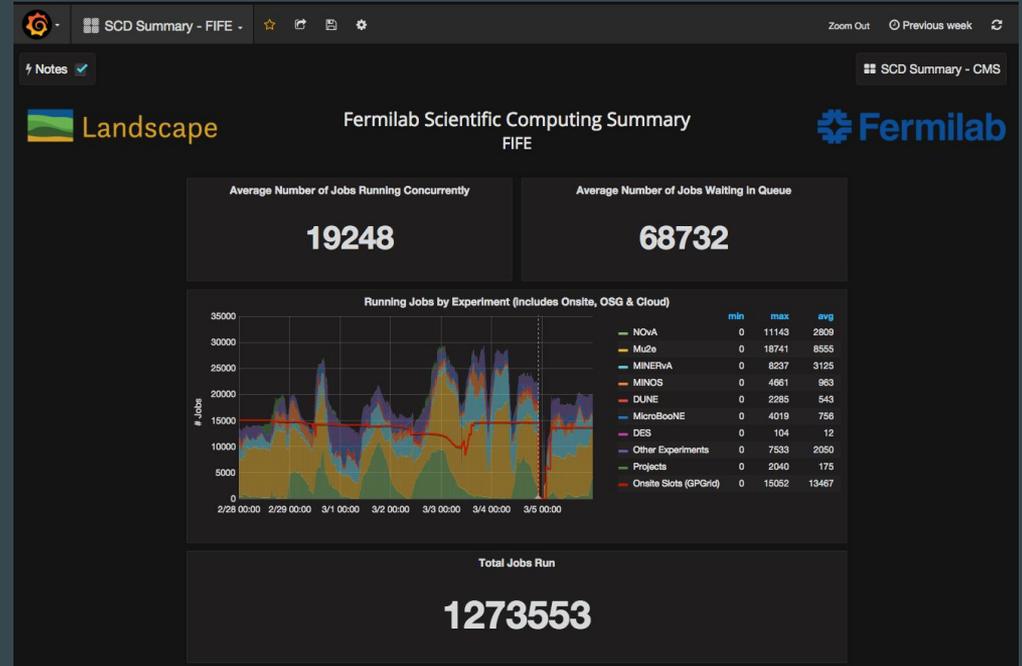


Fifemon v2+ (ca. 2015)

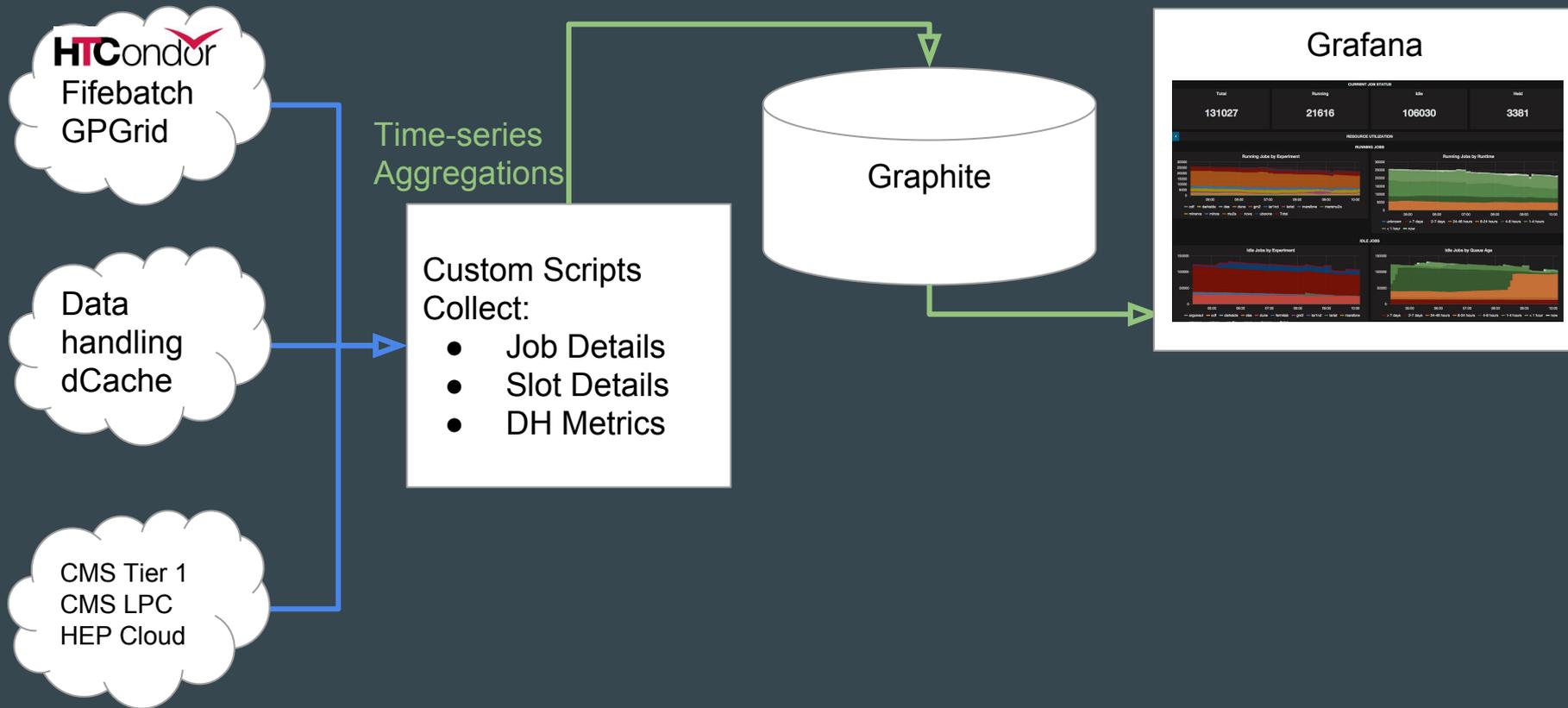
Dedicated effort (~½ FTE) to developing comprehensive monitoring.

- Leverage open-source monitoring technology
- Focus on incorporating new data sources and new dashboards
- Rapid development and iteration of tailored views for each target audience.

Good for grid admins, stakeholders, and users alike!



Fifemon v2



Fifemon v2 Components

Data collection:

- Generic HTCondor probe; adding a new pool is a matter of configuration
- Several other centrally-run probes querying other specific resources
- Data handling services directly reporting to Graphite

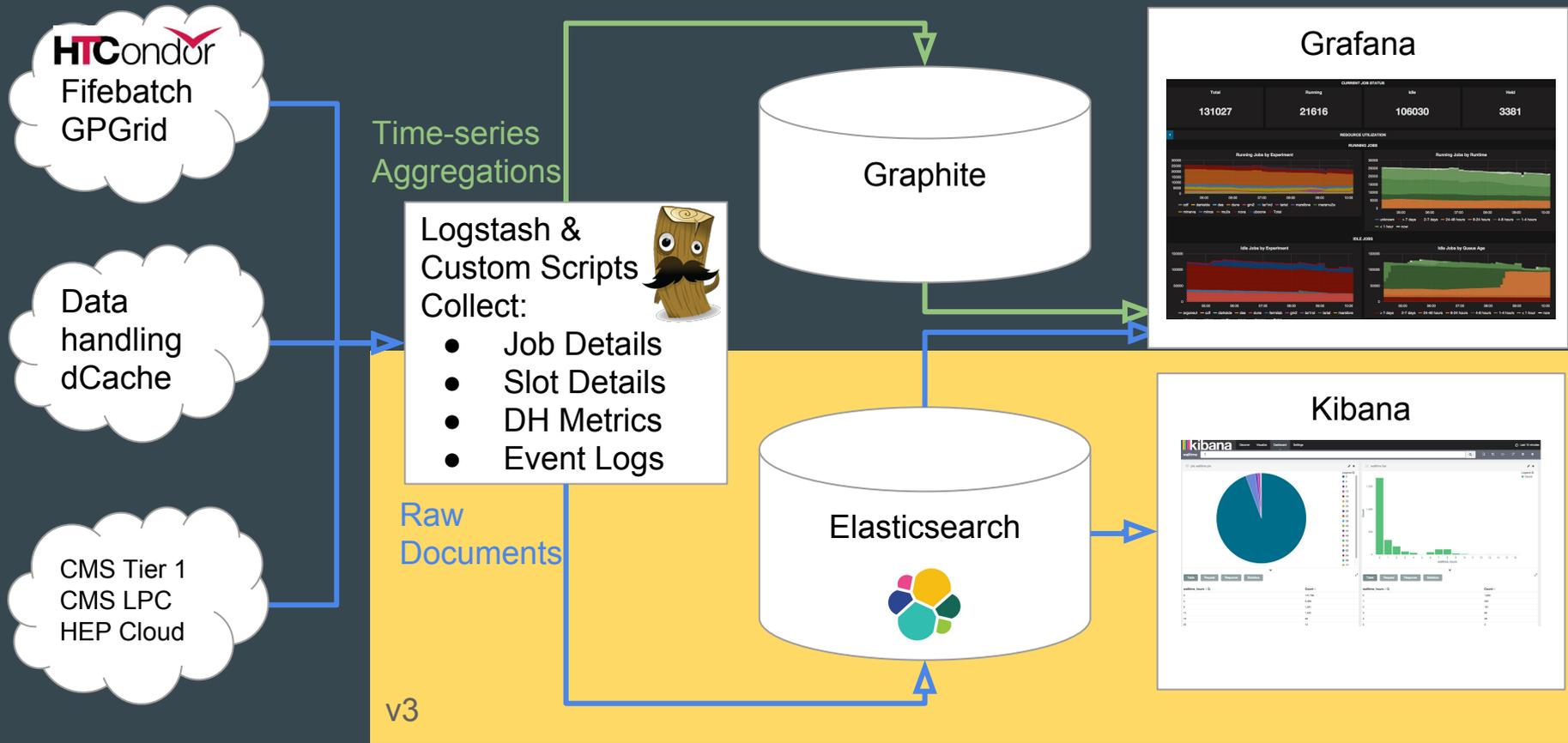
Graphite:

- Time-series database, stores data in files similar to RRD, but adds caching and powerful manipulation library.

Grafana:

- Time-series visualization dashboard platform.
- Supports numerous data sources (Graphite, InfluxDB, Elasticsearch, etc).
- Several auth methods (LDAP, OAuth, proxy).
- Rich user interface for graphing metrics and building dashboards.

Fifemon v3



Fifemon v3 Components

Data collection:

- Logstash to collect and manipulate event data (i.e. logs).
- Current focus is on HTCondor EventLog.

Elasticsearch:

- “NoSQL” document database, powered by Apache Lucene.
- Store full details on jobs, batch slots, and logs.
- Data adds up quickly (Fifebatch: 4-5 MM documents, 7-8 GB per day) and keeping history becomes prohibitively expensive

Grafana:

- Enhancing dashboards with current/recent status information from Elasticsearch.
- Adding custom tables and views with basic JavaScript and HTML, still using Grafana UI.

Kibana:

- Restricted access, mainly for Grid Admin analysis and troubleshooting

Case Studies

“There’s a dashboard for that...”

Case Study: Grid Admin

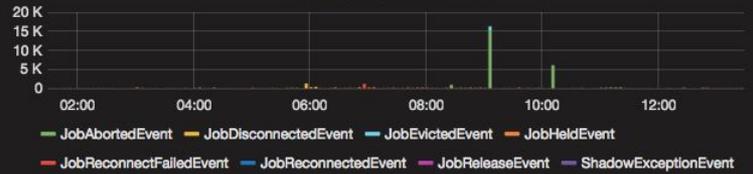
“Is the batch system healthy?”



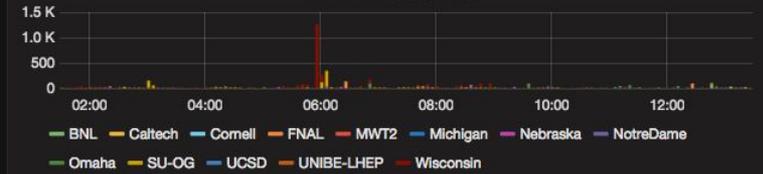
Cinnamon



Abnormal Condor Events



Disconnects by Site



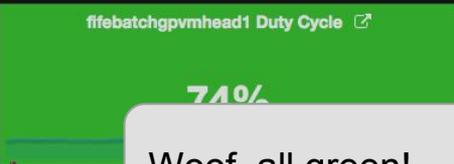
Disconnected Jobs



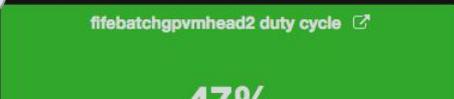
SCHEDD



NEGOTIATOR



COLLECTORS



Woof, all green!





cluster: fifebatch

☰ Fifebatch

CURRENT JOB STATUS

Total

116018

Running

21974

Idle

91392

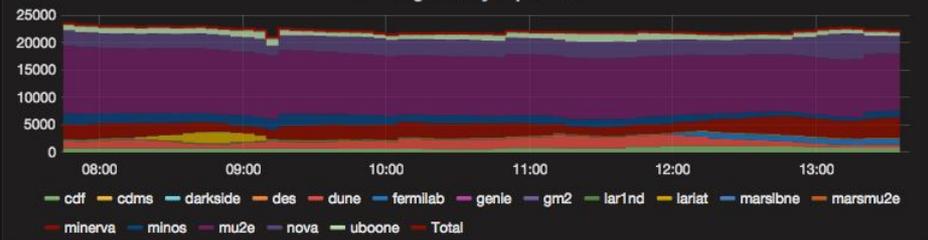
Held

2652

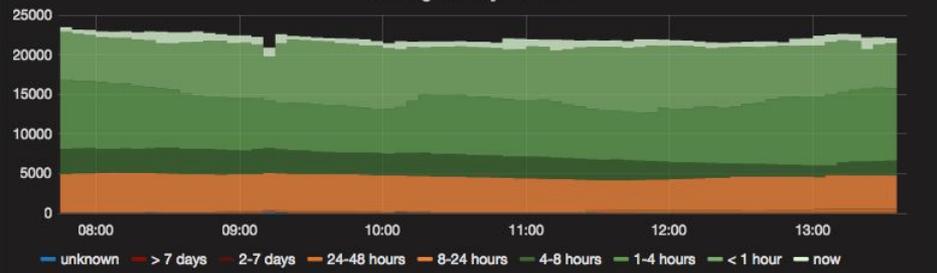
RESOURCE UTILIZATION

RUNNING JOBS

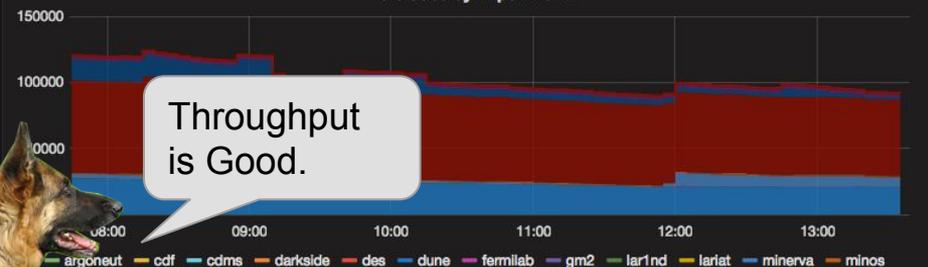
Running Jobs by Experiment



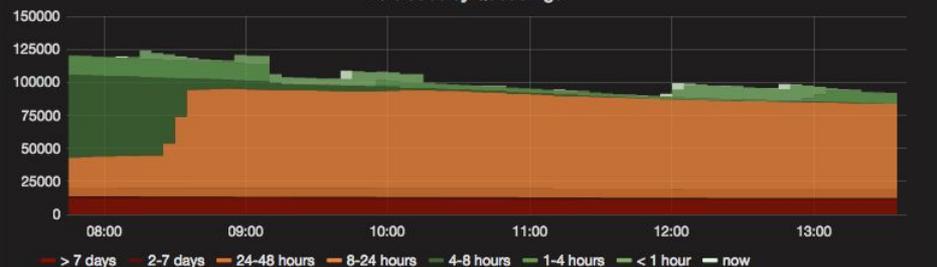
Running Jobs by Runtime



Idle Jobs by Experiment



Idle Jobs by Queue Age





Grid: gpgrid

FIFE Onsite Summary

GPGGrid

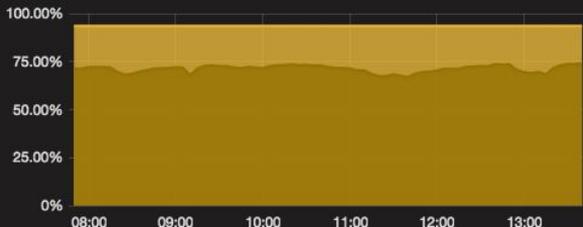
GPGGrid Group

Why Are There Unused Slots on GPGGrid?

PAGE HELP

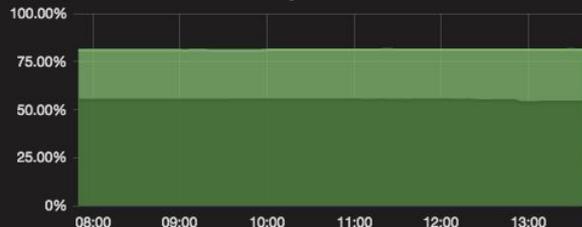
RELATIVE UTILIZATION

CPU Utilization



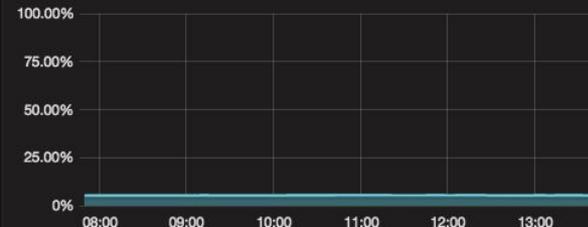
	min	max	avg	current
CPU Claimed	93.488%	93.822%	93.549%	93.580%
CPU Utilized	66.928%	74.206%	71.271%	74.206%

Memory Utilization



	min	max	avg	current
Memory Claimed	80.853%	81.224%	81.063%	81.180%
Memory Utilized	53.703%	55.235%	54.930%	54.102%

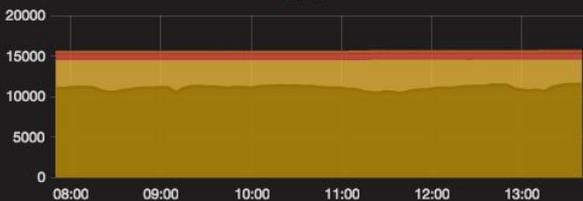
Disk Utilization



	min	max	avg	current
Disk Claimed	5.301%	5.438%	5.386%	5.415%
Disk Utilized	3.938%	4.053%	3.990%	3.954%

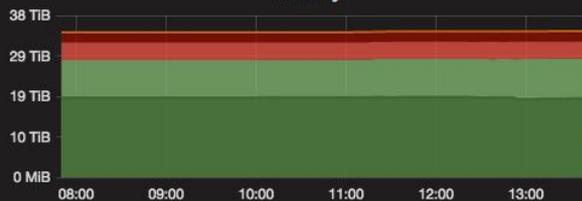
ABSOLUTE UTILIZATION

CPU



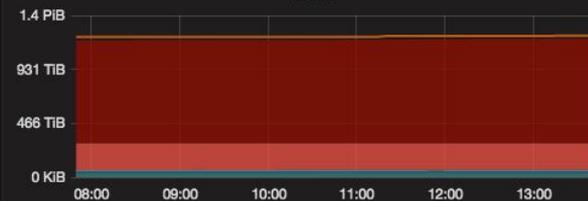
	min	max	avg	current
Claimed	1008	14521	14606	14606
Unclaimed	0	1001	1002	1002
Unusable	0	0	0	0
Total	1008	15523	15608	15608
Effective	1002	11063	11582	11582

Memory



	min	max	avg	current
Claimed	27.6 TIB	28.0 TIB	27.8 TIB	27.9 TIB
Unclaimed	4.1 TIB	4.2 TIB	4.1 TIB	4.1 TIB
Unusable	2.3 TIB	2.4 TIB	2.3 TIB	2.4 TIB
Total	34.1 TIB	34.4 TIB	34.2 TIB	34.4 TIB
Effective	18.4 TIB	18.9 TIB	18.8 TIB	18.6 TIB

Disk



	min	max	avg	current
Claimed	63 TIB	65 TIB	64 TIB	65 TIB
Unclaimed	230 TIB	232 TIB	231 TIB	230 TIB
Unusable	887 TIB	901 TIB	892 TIB	900 TIB
Total	1.184 PIB	1.193 PIB	1.187 PIB	1.193 PIB
Effective	47 TIB	48 TIB	47 TIB	47 TIB

Grid utilization is OK.



GROUP UTILIZATION



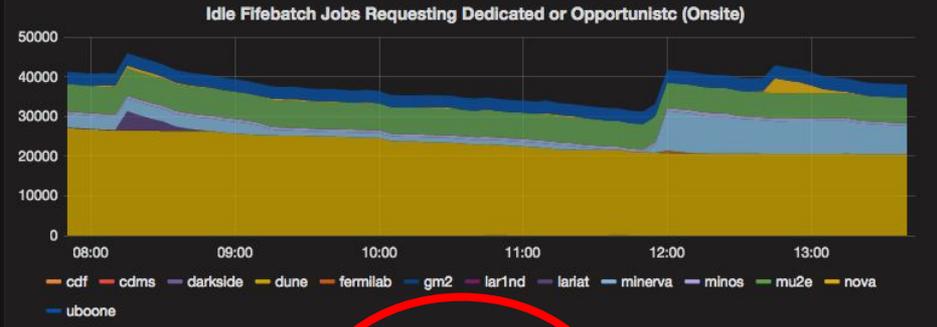
PAGE HELP

Are there FIFE jobs requesting onsite resources?

If jobs are requesting only OFFSITE, they will not run on GPGGrid, unless they come back through the OSG opportunistic gatekeeper.

See also:

- FIFE Onsite Summary
- Fifebatch - Onsite

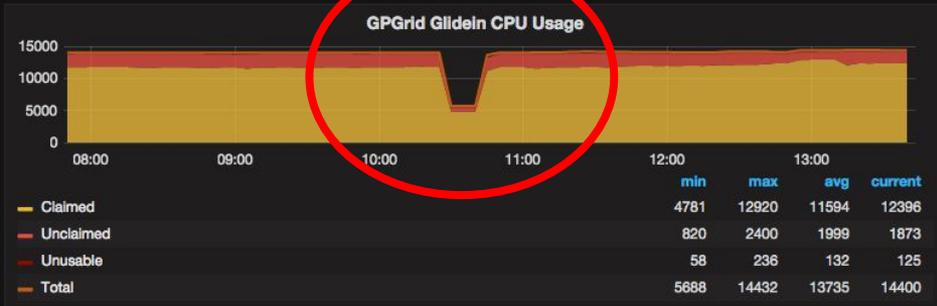


Are the remaining resources in the Glideins "unusable"?

If there are lots of multicore or high-memory (>2 GB) jobs running there will be unusable resources left in the glideins.

See also:

- Grid Utilization (GPGGrid CE)
- Fifebatch Slots (GPGGrid)



Are the remaining resources in the Glideins "unusable"?

"Unusable" above... jobs requesting resources within the limits of... that's remaining

See also:

- Fifebatch Slots (GPGGrid)
- Fifebatch Slots Unclaimed (GPGGrid)

Let's check anyways... what happened here?



Slots with remaining resources exceeding JobSub defaults

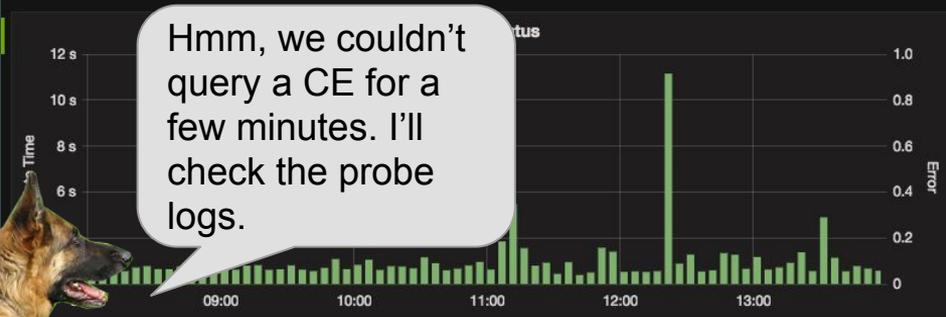
100



probe: gpce01_status + gpce02_status

Update Time

Metric	Min	Max	Avg	Current
awsmonitor	-	-	-	-
cmssrv14_status	1.61 s	8.98 s	2.05 s	1.84 s
cmssrv274_status	0.32 s	1.02 s	0.39 s	0.38 s
cmssrv39_status	0.86 s	2.34 s	1.40 s	1.37 s
condor_pool_jobs	-	-	-	-
fifebatch-pp_status	1.18 s	11.11 s	1.71 s	1.26 s
fifebatch2_status	3.90 min	5.89 min	4.77 min	3.93 min
fifebatch_status	4.07 min	5.72 min	4.73 min	4.28 min
fnpccm1_status	-	-	-	-
gpce01_status	2.38 s	11.15 s	3.01 s	2.57 s
gpce02_status	3.24 s	9.05 s	3.79 s	3.34 s
gpcollector01_status	1.99 s	2.04 min	25.28 s	2.42 s
gpgrid	-	-	-	-



Hmm, we couldn't query a CE for a few minutes. I'll check the probe logs.



Case Study: Stakeholder

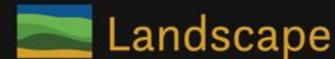
“Is my experiment getting the resources it needs and using them effectively?”



Hazel



FIFE Batch Monitoring



QUICK LINKS

Help

About Fifemon

FIFE Summary

CMS Summary

Experiments

NOvA

MINOS

MINERvA

MINOS

DUNE

MicroBooNE

DES

Other

For Users

User Batch Details

Why Isn't My Job Running?

Grid Status

FIFE Onsite Summary

Fifebatch

GPGrid (CE)

GPGrid (Condor)

DASHBOARDS

Main Dashboards

About Fifemon



Experiment Overview



Fifebatch



GPGrid



Grid Utilization



Help



Jobs Exceeding Resource Request



SCD Summary - CMS



Starred dashboards

Fifebatch Health



Fifebatch Slots



Job Cluster Summary

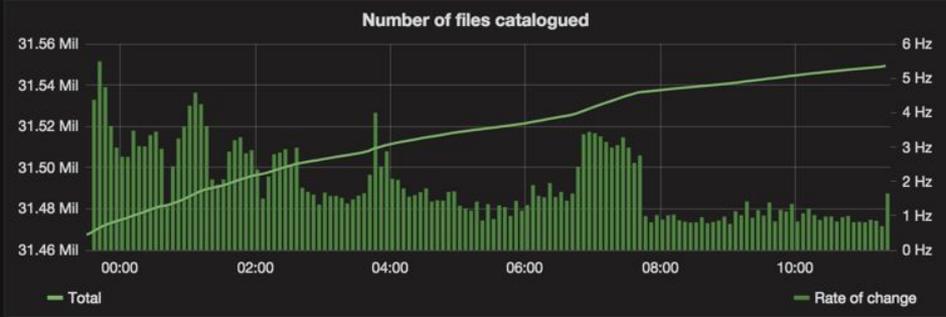
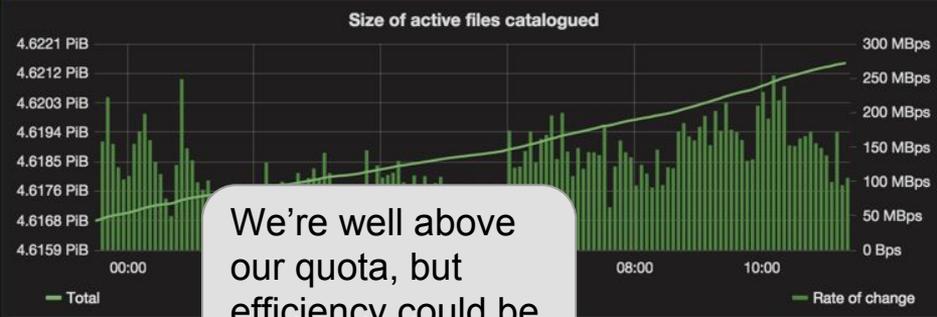
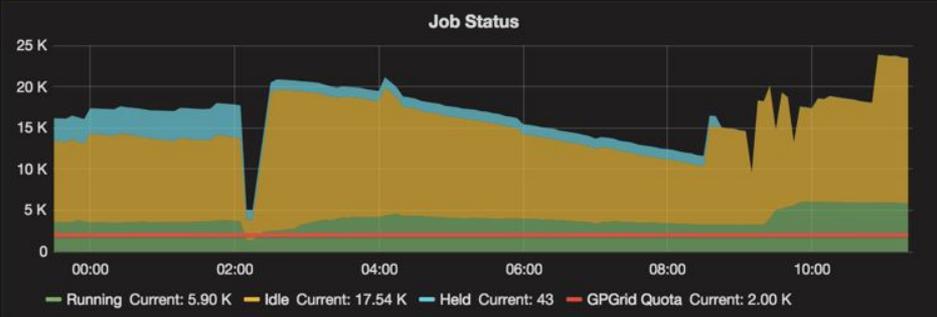


Probe Status

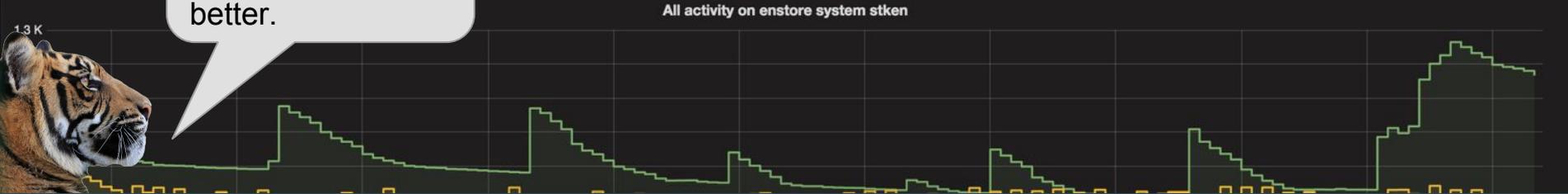




nova



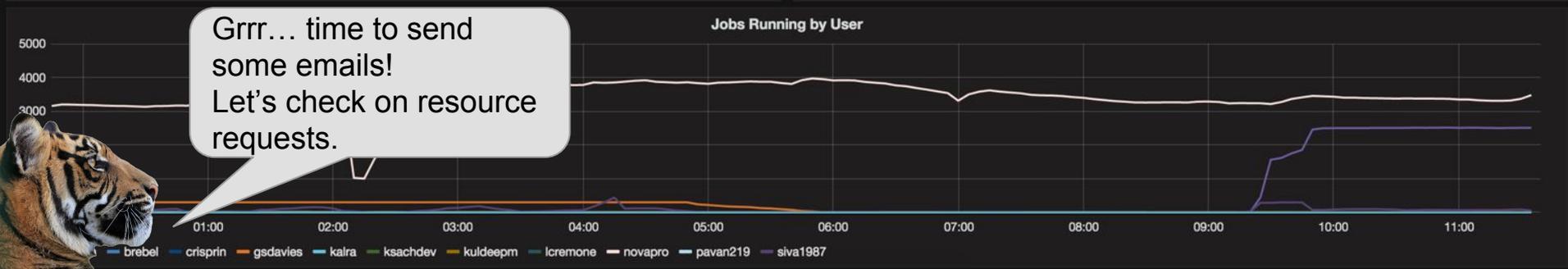
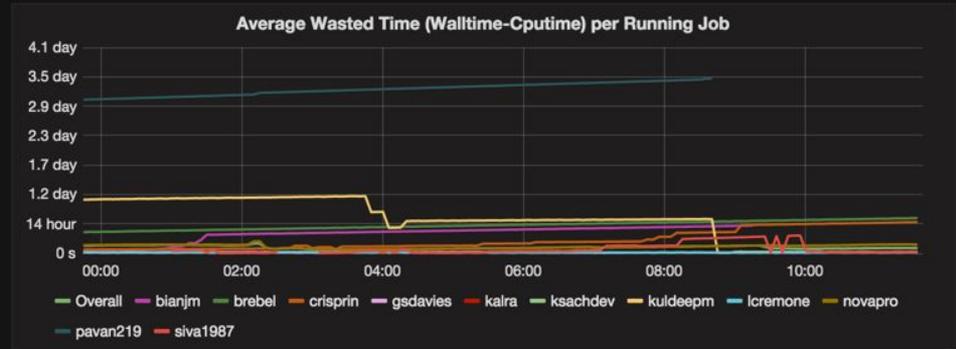
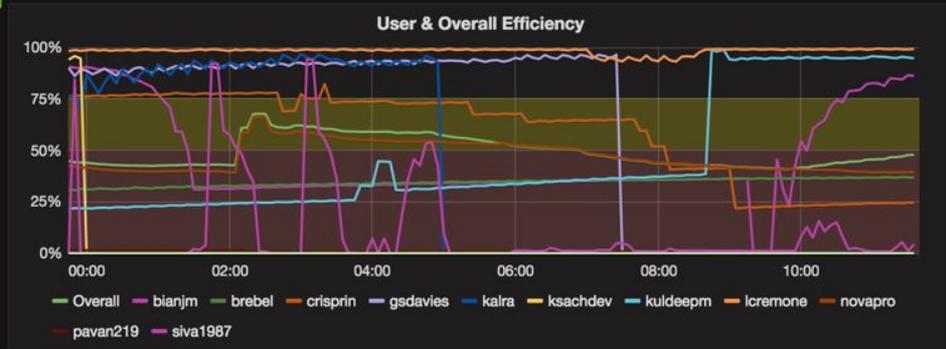
We're well above our quota, but efficiency could be better.



CURRENT



USER HISTORY



Grrr... time to send some emails!
Let's check on resource requests.





User Jobs

User	I	R	C	X	H	Max Memory/Request	Max Disk/Request	Max Time/Request
anorman	0	0	0	0	9	0.78	0.00	0.00
arrieta1	100	0	0	0	3	0.00	0.00	0.00
bianjm	825	2506	0	0	0	0.37	0.00	0.73
boyd	50	0	0	0	0	0.00	0.16	0.00
brebel	0	1	0	0	0	0.00	0.00	3.27
crisprin	0	3	0	0	0	0.01	0.00	8.55
dmendez	0	0	0	0	6	1.00	0.01	0.00
khemer	4	0	0	0	0	0.00	0.00	0.00
kretzke	1	0	0	0	0	0.00	0.00	0.00
kuldeepm	0	10	0	0	0	0.34	0.00	6.07
lcremone	0	2	0	0	0	0.29	0.13	5.54
novapro	22154	3464	0	0	14	1.05	1.01	12.04
pavan219	0	0	0	0	11	0.95	0.11	0.00
siva1987	0	0	0	0	0	0.66	0.00	3.39

Disk and Memory requests look good, lots of users exceeding request time though.

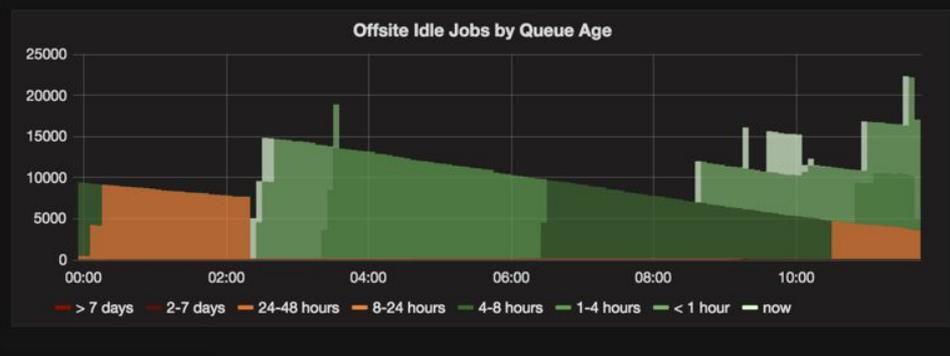
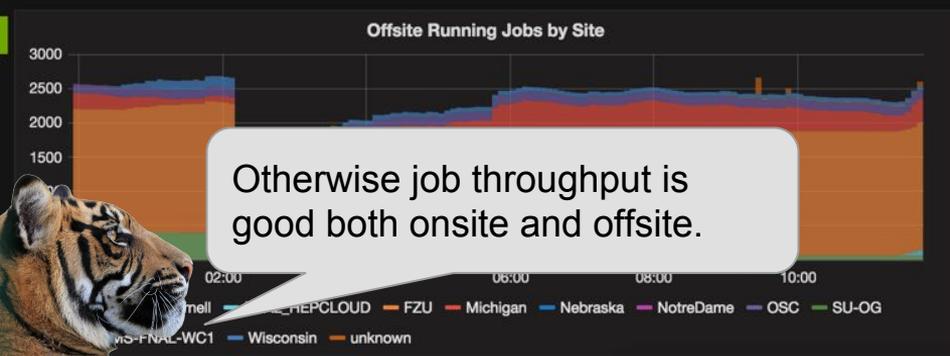
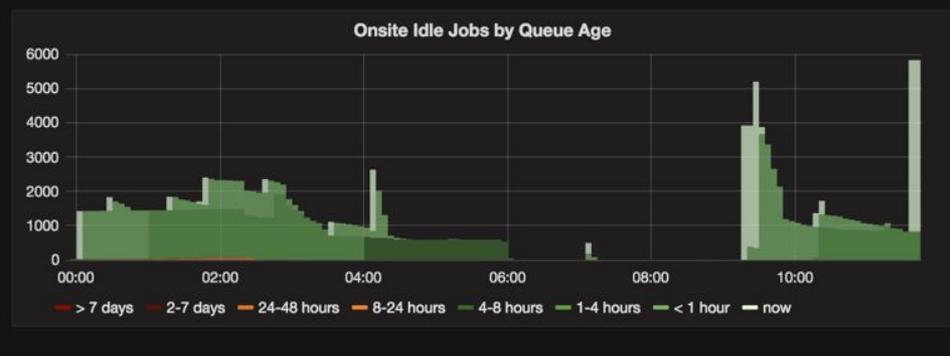
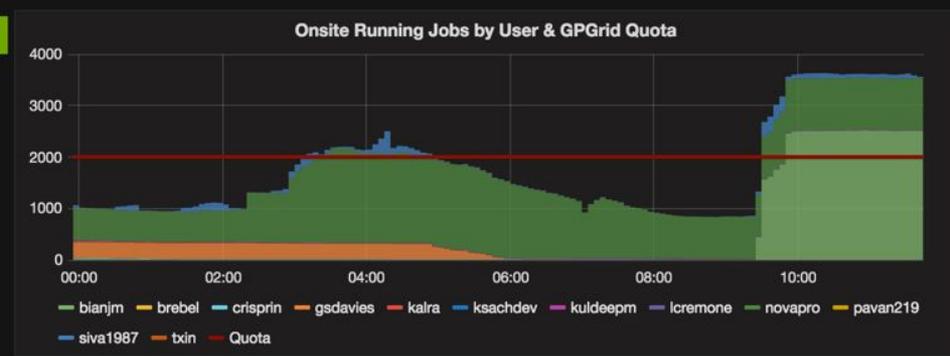
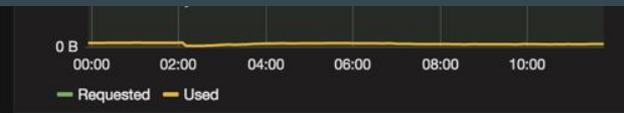


Memory Usage



Disk Usage





Otherwise job throughput is good both onsite and offsite.



Case Study: User 1

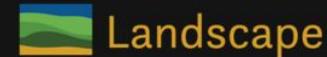
“What’s the status of my jobs?”



Cocoa



FIFE Batch Monitoring



QUICK LINKS

Help

About Fifemon

FIFE Summary

CMS Summary

Experiments

NOvA

Mu2e

MINERvA

MINOS

DUNE

MicroBooNE

DES

Other

For Users

User Batch Details

Why Isn't My Job Running?

Grid Status

FIFE Onsite Summary

Fifebatch

GPGGrid (CE)

GPGGrid (Condor)

DASHBOARDS

Main Dashboards

- About Fifemon ☆
- Experiment Overview ☆
- Fifebatch ☆
- GPGGrid ☆
- Grid Utilization ☆
- Help ☆
- Jobs Exceeding Resource Request ☆
- SCD Summary - CMS ☆
- SCD Summary - FIFE ☆

Starred dashboards

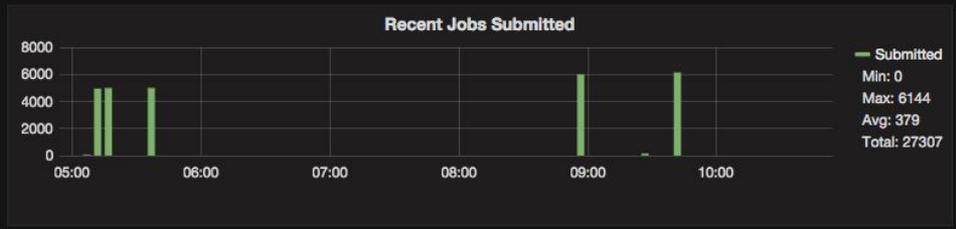
- Fifebatch Health ★
- Fifebatch Slots ★
- Job Cluster Summary ★
- Probe Status ★



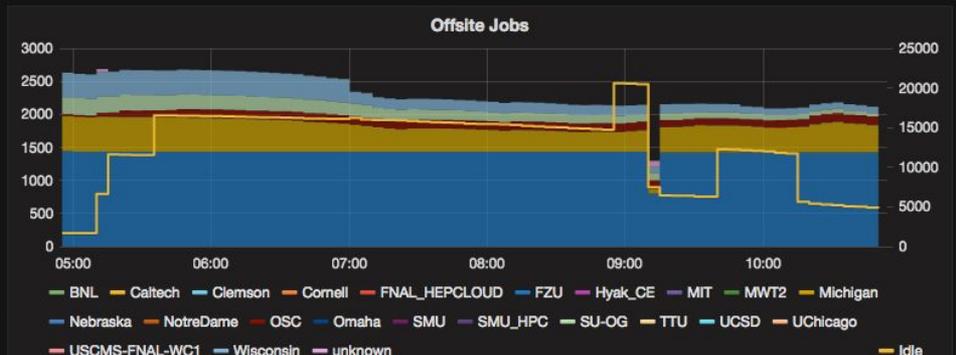
cluster: fifebatch user: cocoa

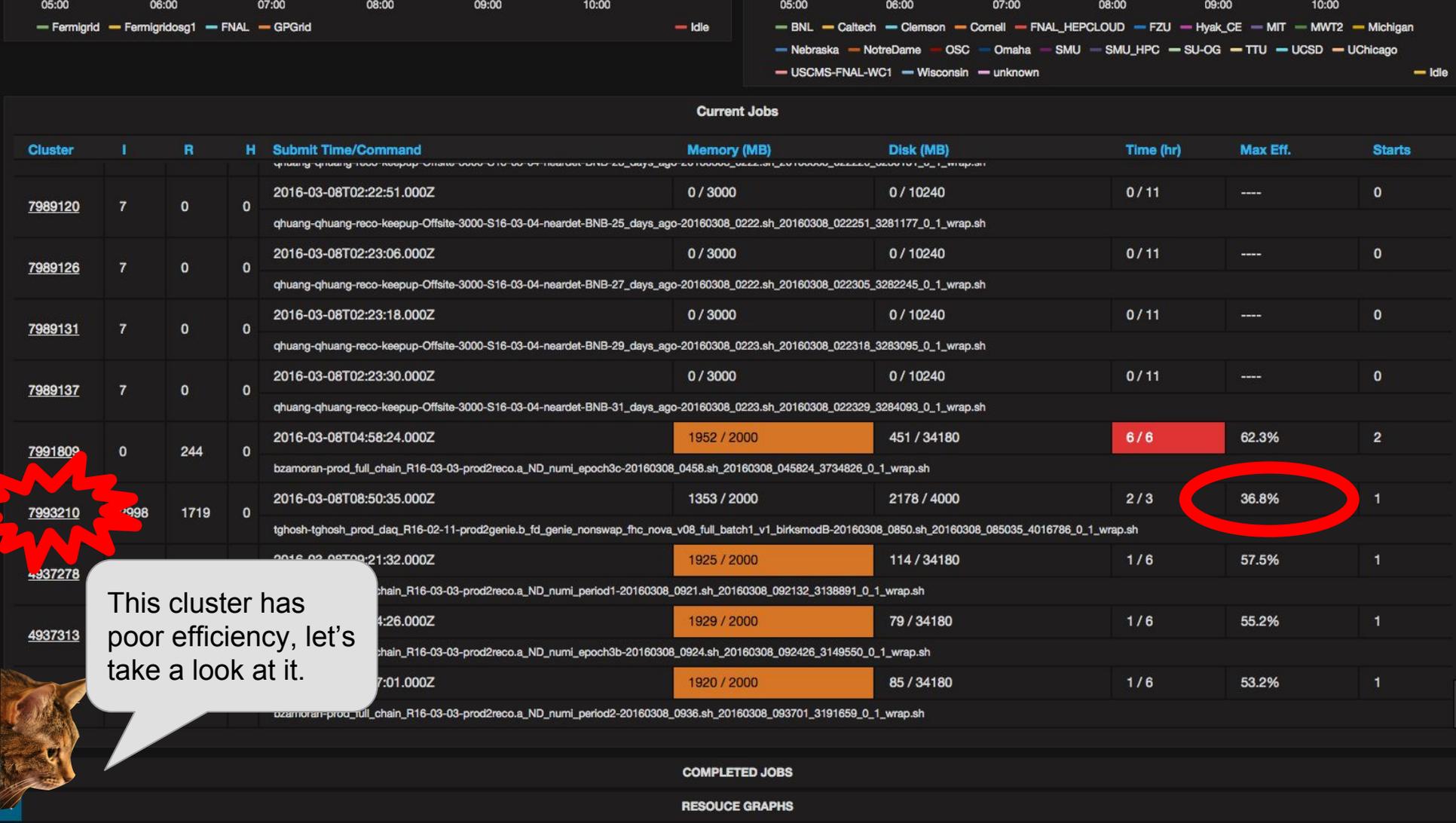
User Efficiency Details Why Are My Jobs Held? Why Isn't My Job Running?

HELD JOBS



Yay my Jobs are starting, but my efficiency is dropping!





- Fermigrid
- Fermigridosg1
- FNAL
- GPGGrid
- Idle
- BNL
- Caltech
- Clemson
- Cornell
- FNAL_HEPCLOUD
- FZU
- Hyak_CE
- MIT
- MWT2
- Michigan
- Nebraska
- NotreDame
- OSC
- Omaha
- SMU
- SMU_HPC
- SU-OG
- TTU
- UCSD
- UChicago
- USCMS-FNAL-WC1
- Wisconsin
- unknown
- Idle

Current Jobs

Cluster	I	R	H	Submit Time/Command	Memory (MB)	Disk (MB)	Time (hr)	Max Eff.	Starts
7989120	7	0	0	2016-03-08T02:22:51.000Z qhuang-qhuang-reco-keepup-Offsite-3000-S16-03-04-needet-BNB-25_days_ago-20160308_0222.sh_20160308_022251_3281177_0_1_wrap.sh	0 / 3000	0 / 10240	0 / 11	----	0
7989126	7	0	0	2016-03-08T02:23:06.000Z qhuang-qhuang-reco-keepup-Offsite-3000-S16-03-04-needet-BNB-27_days_ago-20160308_0222.sh_20160308_022305_3282245_0_1_wrap.sh	0 / 3000	0 / 10240	0 / 11	----	0
7989131	7	0	0	2016-03-08T02:23:18.000Z qhuang-qhuang-reco-keepup-Offsite-3000-S16-03-04-needet-BNB-29_days_ago-20160308_0223.sh_20160308_022318_3283095_0_1_wrap.sh	0 / 3000	0 / 10240	0 / 11	----	0
7989137	7	0	0	2016-03-08T02:23:30.000Z qhuang-qhuang-reco-keepup-Offsite-3000-S16-03-04-needet-BNB-31_days_ago-20160308_0223.sh_20160308_022329_3284093_0_1_wrap.sh	0 / 3000	0 / 10240	0 / 11	----	0
7991809	0	244	0	2016-03-08T04:58:24.000Z bzamorán-prod_full_chain_R16-03-03-prod2reco.a_ND_numi_epoch3c-20160308_0458.sh_20160308_045824_3734826_0_1_wrap.sh	1952 / 2000	451 / 34180	6 / 6	62.3%	2
7993210	998	1719	0	2016-03-08T08:50:35.000Z tghosh-tghosh_prod_daq_R16-02-11-prod2genie.b_fd_genie_nonswap_fhc_nova_v08_full_batch1_v1_birksmodB-20160308_0850.sh_20160308_085035_4016786_0_1_wrap.sh	1353 / 2000	2178 / 4000	2 / 3	36.8%	1
4937278				2016-03-08T09:21:32.000Z chain_R16-03-03-prod2reco.a_ND_numi_period1-20160308_0921.sh_20160308_092132_3138891_0_1_wrap.sh	1925 / 2000	114 / 34180	1 / 6	57.5%	1
4937313				2016-03-08T09:26:00.000Z chain_R16-03-03-prod2reco.a_ND_numi_epoch3b-20160308_0924.sh_20160308_092426_3149550_0_1_wrap.sh	1929 / 2000	79 / 34180	1 / 6	55.2%	1
				2016-03-08T09:27:01.000Z bzamorán-prod_full_chain_R16-03-03-prod2reco.a_ND_numi_period2-20160308_0936.sh_20160308_093701_3191659_0_1_wrap.sh	1920 / 2000	85 / 34180	1 / 6	53.2%	1



This cluster has poor efficiency, let's take a look at it.



COMPLETED JOBS

RESOURCE GRAPHS



cluster: 7714932

PAGE HELP

JOB INFORMATION

Job ID:	7714932.0@fifebatch2.fnal.gov	Resources Requested	
Submit Date:	2016-02-26T18:09:46	CPU:	1
Experiment:	mu2e	Memory:	3994 MB
User:	mu2epro (mu2epro/cron/mu2egpvm01.fnal.gov@FNAL.GOV)	Disk:	9216 MB
Usage Model:	OFFSITE	Runtime:	9 hr
Sites Requested:	BNL,Caltech,FERMIGRID,FNAL,MIT,Michigan,Nebraska,Omaha,SU-OG,Wisconsin,UCSD,NotreDame,MWT2		

[View sandbox files](#)

[View available slots](#)

PROCESS STATUS

Total Processes	Idle Processes	Running Processes	Held Processes
9175	6065	2898	4
	Failed Processes (nonzero exit code)	Disconnected Processes	
	26	408	



A few failed processes, and a bunch are disconnected.

RESOURCES USED

Max Memory Usage	Max Disk Usage	Max Walltime
-------------------------	-----------------------	---------------------

Completed Processes (exit code 0)

1011

Failed Processes (nonzero exit code)

26

Disconnected Processes

408

RESOURCES USED

Max Memory Usage

1.934 GiB

Max Disk Usage

7.91 GiB

Max Walltime

11.11 hour

Memory Usage

Min	Max	Average
10.02 MIB	1.93 GiB	1.31 GiB

Disk Usage

Min	Max	Average
1.75 GiB	7.91 GiB	5.34 GiB

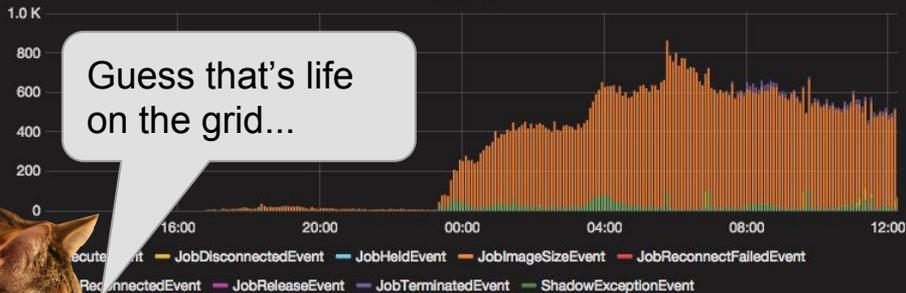
Walltime

Min	Max	Average
33.70 min	11.11 hour	5.36 hour

PROCESS LIST

CONDOR EVENTS

All Events



Abnormal Events



STATS BY SITE

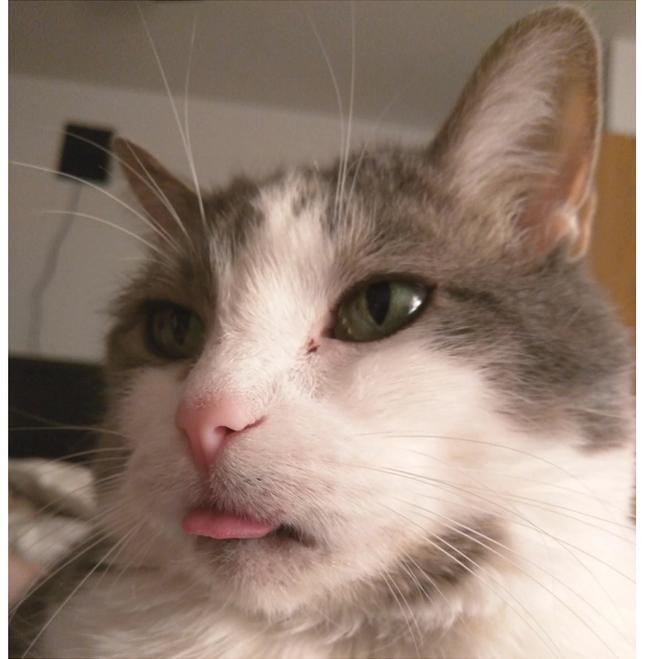
JOBSUB



Guess that's life on the grid...

Case Study: User 2

“Why isn’t my job running yet?!”



Peanut



FIFE Batch Monitoring



QUICK LINKS

- [Help](#)
- [About Fifemon](#)
- [FIFE Summary](#)
- [CMS Summary](#)

Experiments

- [NOvA](#)
- [Mu2e](#)
- [MINERvA](#)
- [MINOS](#)
- [DUNE](#)
- [MicroBooNE](#)
- [DES](#)
- [Other](#)

FIFE Users

- [User Batch Details](#)
- [Why Isn't My Job Running?](#)

Grid Status

- [FIFE Onsite Summary](#)
- [Fifebatch](#)
- [GPGrid \(CE\)](#)
- [GPGrid \(Condor\)](#)

DASHBOARDS

Main Dashboards

- [About Fifemon](#) ☆
- [Experiment Overview](#) ☆
- [Fifebatch](#) ☆
- [GPGrid](#) ☆
- [Grid Utilization](#) ☆
- [Help](#) ☆
- [Jobs Exceeding Resource Request](#) ☆
- [SCD Summary - CMS](#) ☆
- [SCD Summary - FIFE](#) ☆

Starred dashboards

- [Fifebatch Health](#) ★
- [Fifebatch Slots](#) ★
- [Job Cluster Summary](#) ★
- [Probe Status](#) ★



Username: peanut -

☰ Fifebatch

☰ Troubleshooting Guides

General Tips

How long ago did you submit your jobs?

It can take several hours (up to a day) for jobs to start, the grid is generally running at full capacity. Remember, the batch system is for large-scale computing. If you need immediate results on a small scale you should be using the interactive nodes provided for your experiment.

What resources did you request?

If you're not sure, you can see these listed on your [User Batch Details](#) page or [Job Cluster Summary](#) in the table below (select your username from the dropdown above). If you didn't request any your job got the defaults. Your job will only start in a slot that has at least your requested resources available; the more you request, the fewer slots that will be available.

What is your usage model?

- DEDICATED or OPPORTUNISTIC: your job will run on GPGrid, and how long your job takes to start is dependent on:
 - how many other jobs are vying for slots on GPGrid (take a look at the [FIFE Onsite Summary](#) dashboard).
 - how much your experiment is using; your experiment has a quota on GPGrid (visible on the [Experiment Overview](#) page), usage over this number is purely opportunistic
 - what resources are available in the remaining slots on GPGrid (see the [Fifebatch Slots](#) dashboard)
- OFFSITE: your job will run on the OSG, where availability is opportunistic and highly variable (with some exceptions, e.g. FZU for NOvA).
 - Did you request any specific sites? Some sites have restrictions on resources, runtime, or experiments (see the [FIFE wiki](#) for details)
 - take a look at the [Fifebatch Slots](#) dashboard to see where we are currently getting slots

IDLE JOBS

Idle Jobs

jobid	Submit Date	Group	CPUs	Memory	Disk	Runtime	Usage Model	Sites
4924655.0@fifebatch1.fnal.gov	2016-03-08 00:17:42	nova	1	4.88 GiB	10.00 GiB	3.00 hour	OFFSITE	Wisconsin
7936096.0@fifebatch2.fnal.gov	2016-03-07 00:17:26	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	UCSD
4924412.0@fifebatch1.fnal.gov	2016-03-08 00:17:42	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	Caltech
4924398.0@fifebatch1.fnal.gov	2016-03-08 00:17:42	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	BNL
4924398.0@fifebatch1.fnal.gov	2016-03-08 00:17:03	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	Nebraska
4924398.0@fifebatch1.fnal.gov	2016-03-08 00:17:03	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	UCSD
4924398.0@fifebatch1.fnal.gov	2016-03-06 00:22:21	nova	1	3.42 GiB	10.00 GiB	3.00 hour	OFFSITE	UCSD

Let's look at the job details





cluster: 4924655 ▾

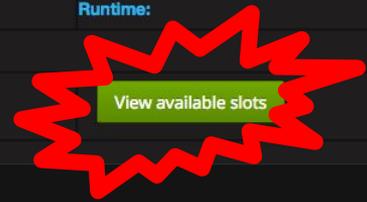


PAGE HELP

JOB INFORMATION

Job ID:	4924655.0@fifebatch1.fnal.gov	Resources Requested	
Submit Date:	2016-03-08T00:17:42	CPU:	1
Experiment:	nova	Memory:	5000 MB
User:	novapro (UNKNOWN)	Disk:	10240 MB
Usage Model:	OFFSITE	Runtime:	3 hr
Sites Requested:	Wisconsin		

View sandbox files



View available slots

PROCESS STATUS

Total Processes	Idle Processes	Running Processes	Held Processes
1	1	0	0

Completed Processes (exit code 0)

Failed Processes (nonzero exit code)

Disconnected Processes

Are there any slots available?

N/A

N/A

RESOURCES USED

Max Memory Usage

Max Disk Usage

Max Walltime





cluster: fifebatch - Sites: Wisconsin - Min. Cores: 1 - Min. Memory (MB): 5000 - Min. Time (hr): 3.00 - Available to: nova -

☰ Fifebatch

Unclaimed CPUs

0

Claimed CPUs

0

Unclaimed by Site

Claimed by Site

There are no Glideins running at Wisconsin with 5GB memory!





cluster: fifebatch Sites: All Min. Cores: 1 Min. Memory (MB): 5000 Min. Time (hr): 3.00 Available to: nova

☰ Fifebatch

Unclaimed CPUs

25

Claimed CPUs

134

Unclaimed by Site

Site	# Cpus	# Glideins
GPGrid	25	5

Claimed by Site

Site	# Cpus	# Glideins
GPGrid	134	134

Hey, there are some Glideins on GPGrid that could run a job needing 5GB memory! Maybe I should submit there instead.



Comprehensive Batch Monitoring with Fifemon Increases Grid Utilization and Job Throughput

(and makes everyone's life easier)

Next Generation Accounting



Architecting a Replacement for Gratia

Motivation

- Gratia is showing its age - written in 2004 in Java 3
- Changes/incompatibilities in underlying libraries (Hibernate ORM) and database (MySQL) have made housekeeping cleanup (deleting old records) non-performant.
- Rigid SQL schema (controlled through Hibernate mappings) makes tracking new record types and metrics difficult.

Considerable effort would be required to maintain and update Gratia to serve the needs of the OSG for the next ten years.

Introducing



A flexible accounting and monitoring system based on open-source technology.

Compatible with existing Gratia infrastructure:

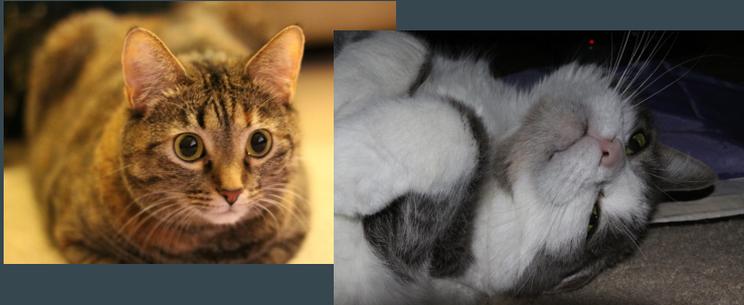
- NO changes to probes required
- Historical data easily migratable

Etymology:

- Grid Accounting
- Gratia-Compatible Collector
- Grok: “to understand” (Heinlein, *Stranger in a Strange Land*)

Gracc Architecture

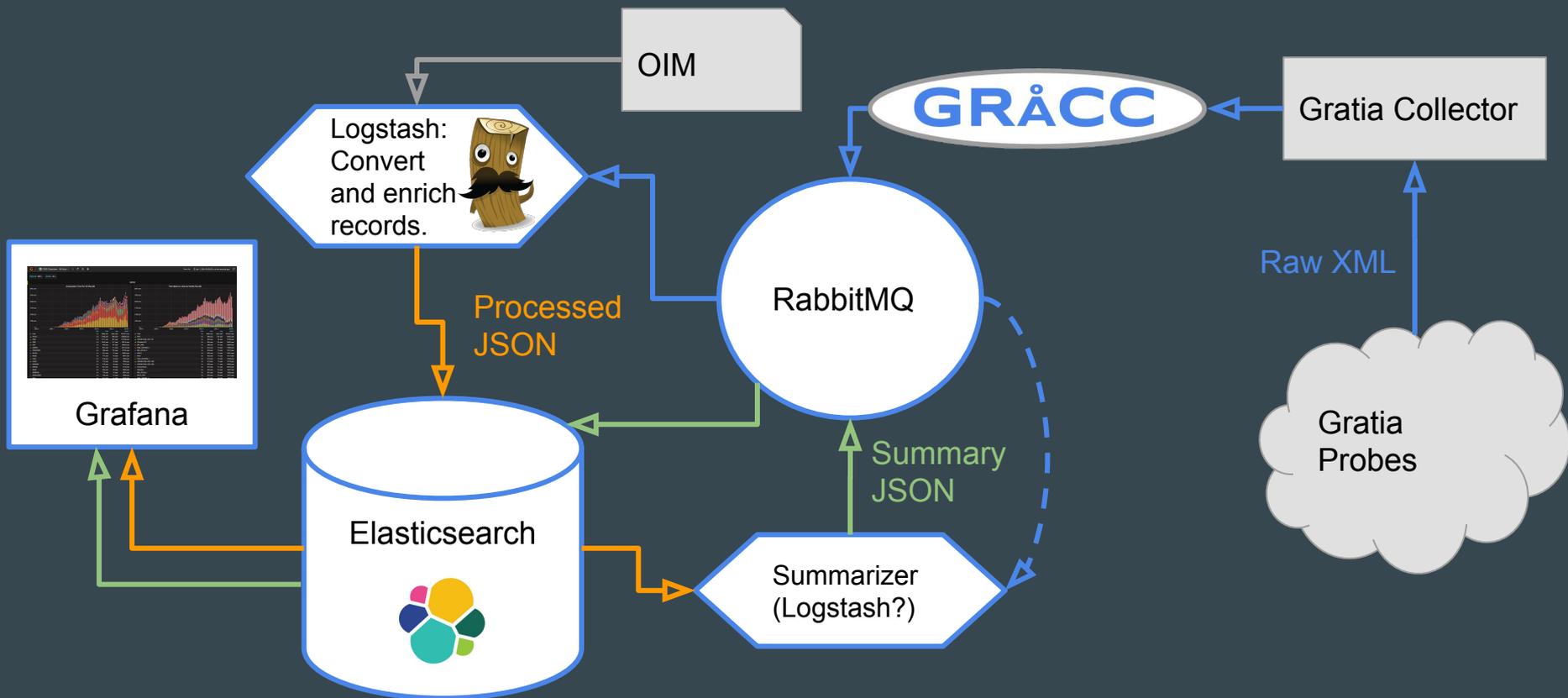
- Swappable, independent components that communicate through a data exchange
- Gratia was a monolithic 800-lb gorilla, Gracc will be composed of several 10-lb kitties (they're cuter...)



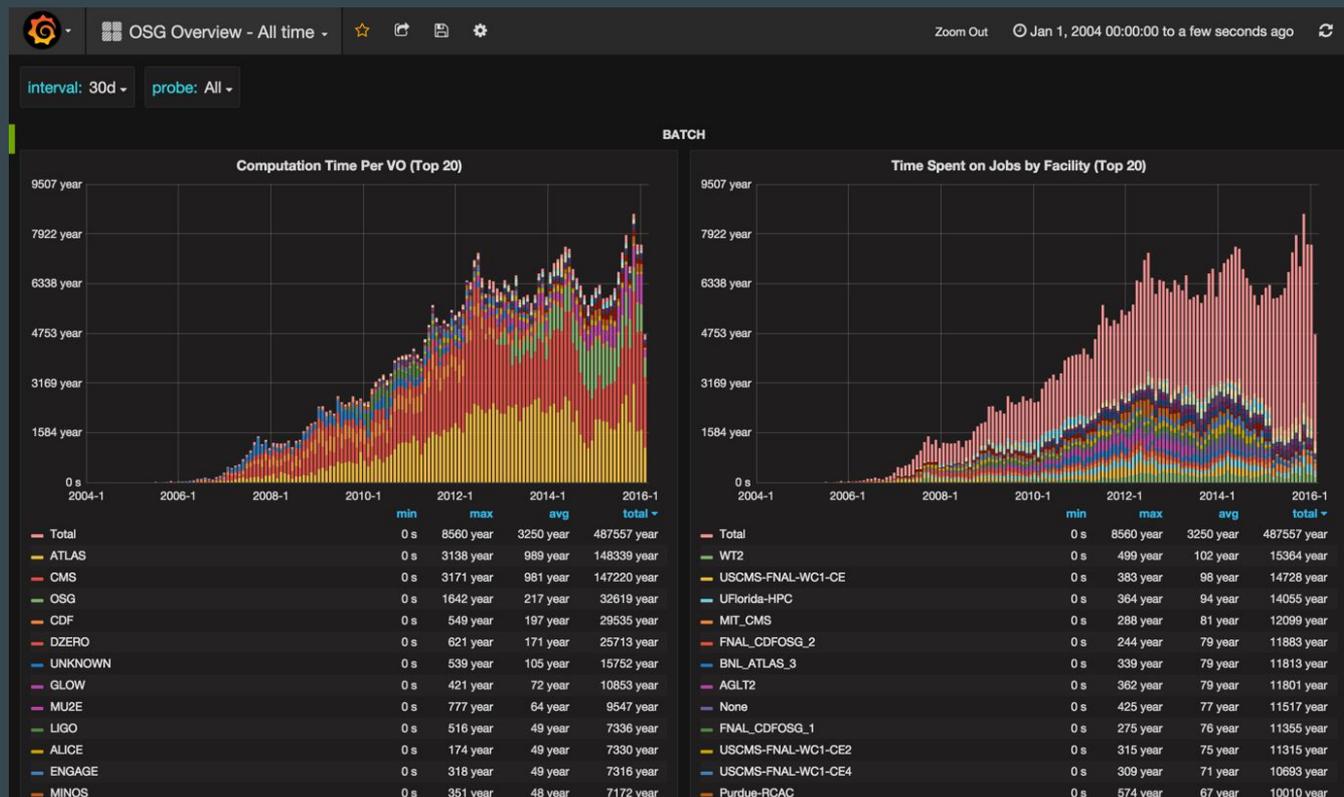
Prototype Components:

- Elasticsearch - data storage
- Grafana - user interface
- Logstash - data handling
- RabbitMQ - data exchange

Gracc Architecture

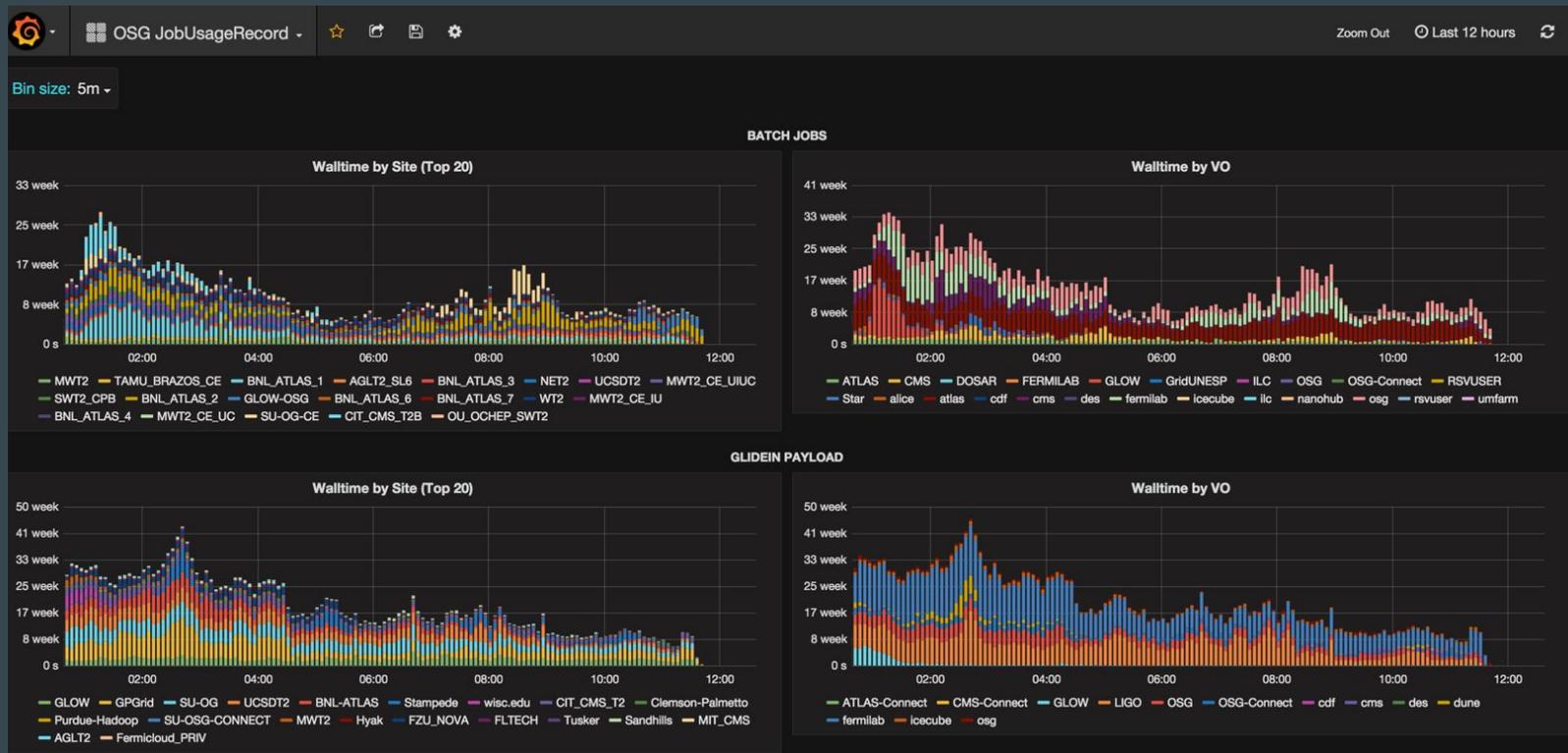


Prototype - Summary Data



<https://hcc-anvil-175-6.unl.edu/dashboard/db/osg-overview-all-time>

Prototype - JobUsageRecord



<https://hcc-anvil-175-6.unl.edu/dashboard/db/osg-jobusagerecord>

Prototype: Self-Monitoring



<https://hcc-anvil-175-6.unl.edu/dashboard/db/gracc-monitor>

**Gracc will provide a
flexible and extensible
platform for OSG
monitoring and
accounting.**