



Dress Rehearsal Review Pilot Assessment: RUCIO

Riccardo Di Maria

CERN

December 9th, 2020 - 2nd ESCAPE WP2/DIOS Workshop, CERN



ESCAPE DataLake Components

From CERN Team

Rucio

XCache

From CERN-IT

FTS

OpenStack

perfSONAR

CRIC

Grafana

OracleDB

k8s

From ESCAPE Partners

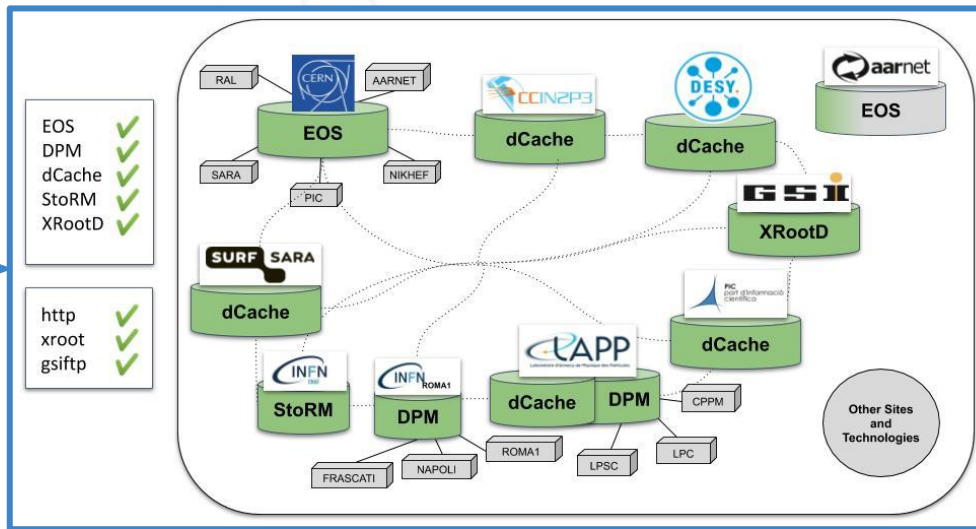
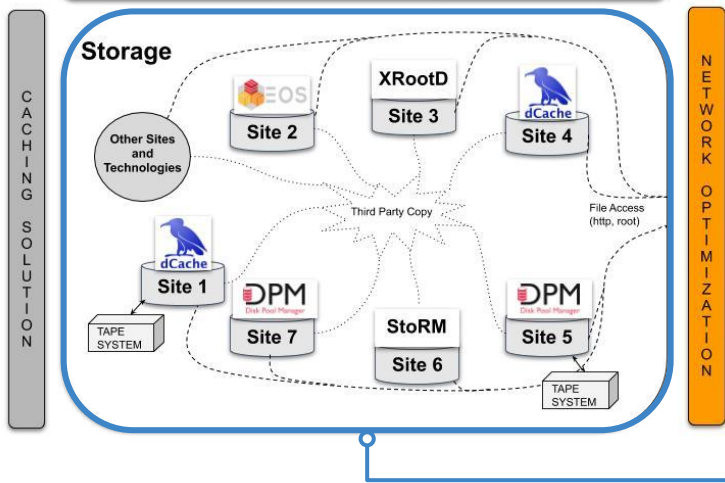
IAM



Goal: exercise covering **experiment data workflow** needs on a single day (data injection, replication, and access).

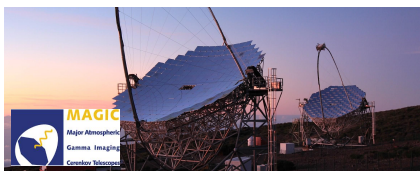
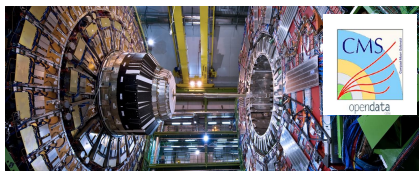
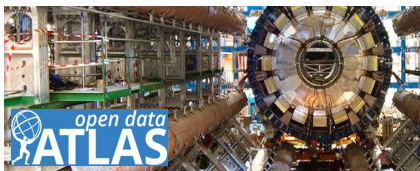
Three fold goal: perspective from **scientists**, perspective from **sites**, and the assessment of the **ESCAPE DataLake tools and services** under **pseudo-production conditions**: RUCIO, FTS, CRIC, IAM, perfSONAR, monitoring, QoS, clients, etc.

DataLake Objectives: stable infrastructure: 10 sites, 5 storages technologies, 3 protocols; monitoring: automated tests; 1M files - to demonstrate stable and sizeable data movement; 3 QoS: CRIC as reference point.



- Hiding complexity and providing transparent access to data.
- Heterogeneous federated storage and operations model.
- Some centers joining even if not funded by ESCAPE.

Further info: https://wiki.escape2020.de/index.php/WP2_-_DIOS#Datalake_Status



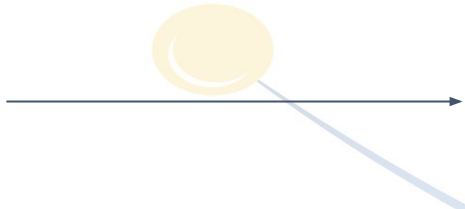
RSE	Quota	WM
ALPAMED-DPM	100 TB	10 TB
CNAF-STORM	10 TB	1 TB
DESY-DCACHE	40 TB	4 TB
EULAKE-1	300 TB	30 TB
GSI-ROOT	1 TB	10 GB
IN2P3-CC-DCACHE	60 TB	1 TB
INFN-NA-DPM	68 TB	5 TB
INFN-NA-DPM-FED	46 TB	5 TB
INFN-ROMA1	2 TB	200 GB
LAPP-DCACHE	10 TB	1 TB
LAPP-WEBDAV	100 GB	90 GB
PIC-DCACHE	28 TB	27.99 TB
PIC-INJECT	28 TB	27.99 TB
SARA-DCACHE	98 TB	140 GB

ESCAPE DataLake

- Total Quota:
891 TB
- Watermark:
113.44 TB
- 10+ RSEs
- 9 sciences
- 50+ accounts



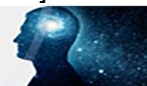
Kubernetes Cluster @ CERN

- OpenStack VMs:
 - 1 master: 4 CPU, 8 GB RAM
 - 10 nodes: 8 CPU, 16 GB RAM
 - K8s cluster:
 - filebeat (per node) and logstash for cluster monitoring
 - rucio-client with root account and admin privileges for DataLake managing
 - escape-crons pod
 - OracleDB:
 - quota raised from 15 GB to 50 GB
-
- IAM-Rucio sync
 - IAM-Gridmap (EOS) sync
 - CRIC-Rucio sync
 - noise production (100MB file upload per RSE + add rule per RSE)
 - Gfal SAM and FTS tests
- 



Kubernetes Cluster @ CERN

- Rucio (HELM-charts-based):
 - UI (escape-rucio.cern.ch)
 - Auth Server
 - Main Server (2)
 - Daemons:
 - **Abacus Account** [updating account (counter) usages]
 - **Abacus Collection Replica** [updating collection replica]
 - **Abacus RSE** [updating RSE (counter) usages]
 - **Conveyor Submitter** (3 x 4 threads) [managing non-tape file transfers - preparing and submitting jobs]
 - **Conveyor Poller** (3 x 4 threads) [checking status of submitted transfers]
 - **Conveyor Finisher** (2 threads) [updating Rucio internal state for finished transfers]
 - **Hermes** [delivering messages via STOMP to a message broker]
 - **Judge Injector** (2) [asynchronously injecting replication rules]
 - **Judge Evaluator** (3 x 3 threads) [executing and reevaluating replication rules]
 - **Judge Repairer** (2 x 5 threads) [repairing stuck replication rules]
 - **Judge Cleaner** (2 x 5 threads) [cleaning expired replication rules]
 - **Reaper2** (2 x 4 threads) [deleting replicas]
 - **Transmogrifier** [creating replication rules for DIDs matching a subscription]
 - **Undertaker** [managing (deleting) expired DIDs]



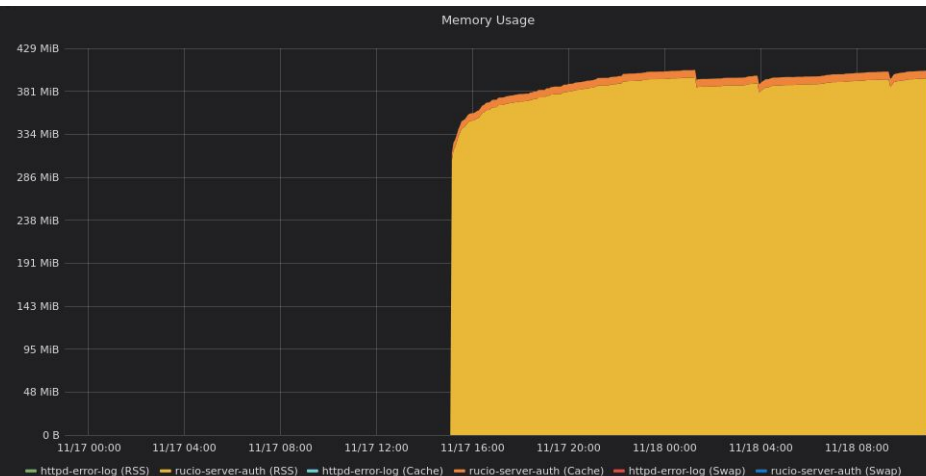
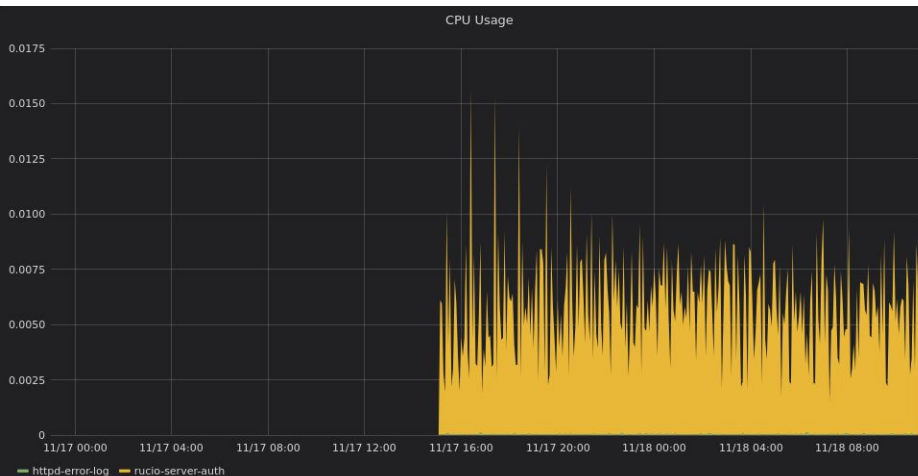
Auth - before setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server-auth	-	-	0.02 (0.12) 550 MiB (1.25 GiB)	Errors and restart due to no limits set.



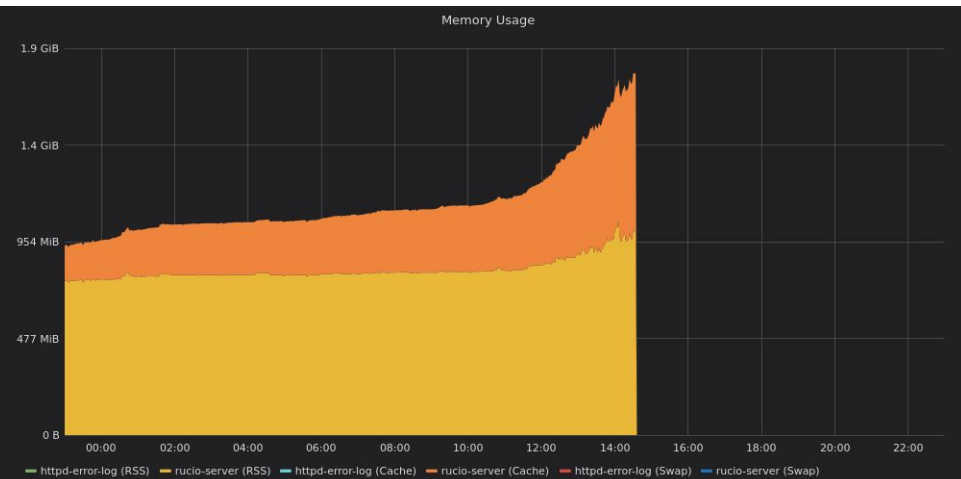
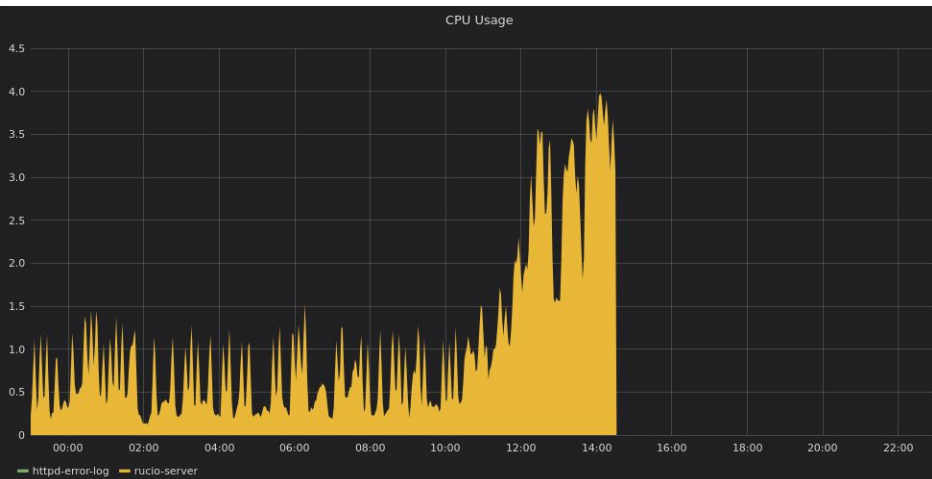
Auth - after setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server-auth	4 2500 MiB	4 2500 MiB	0.02 500 MiB	OK



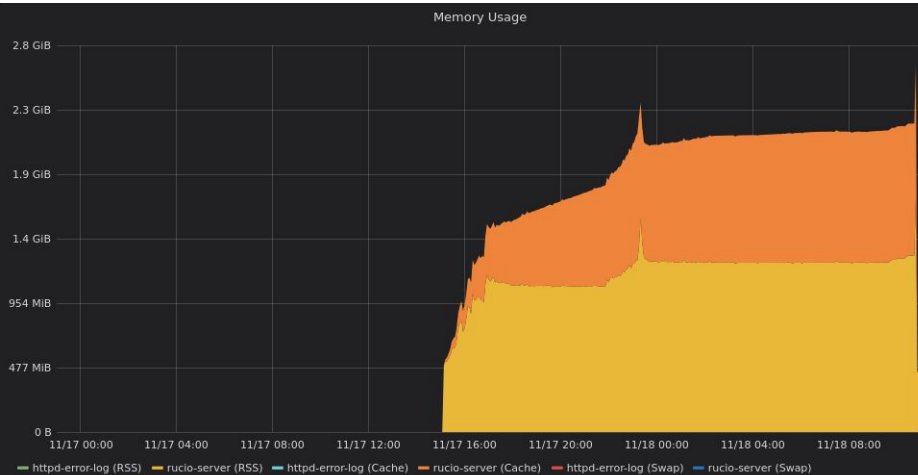
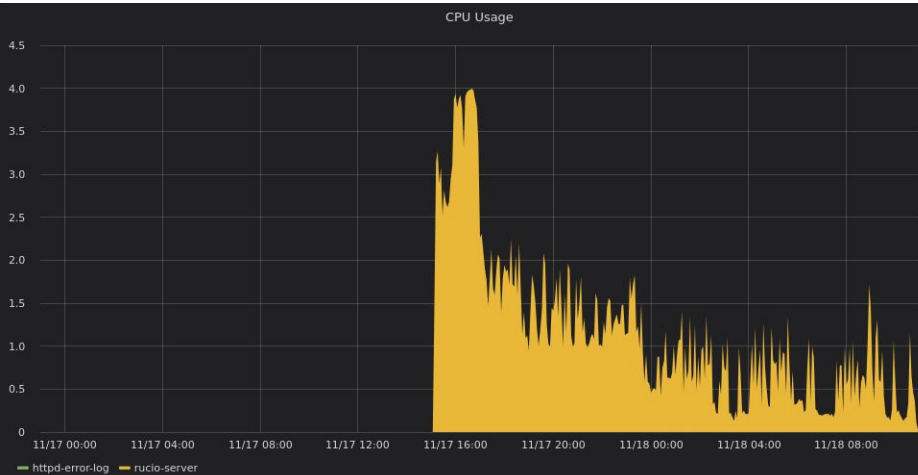
Server - before setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server (2)	-	-	2-4 1.25 GiB (2 GiB)	Manual restart due to no limits set.



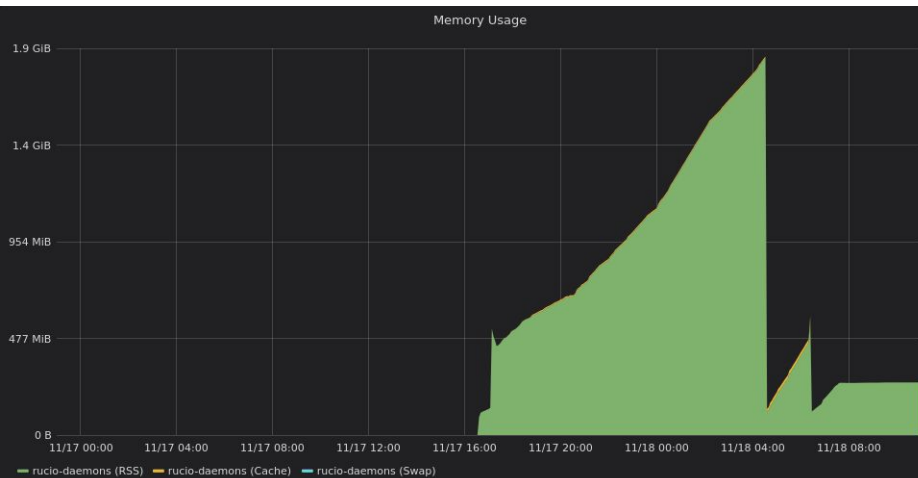
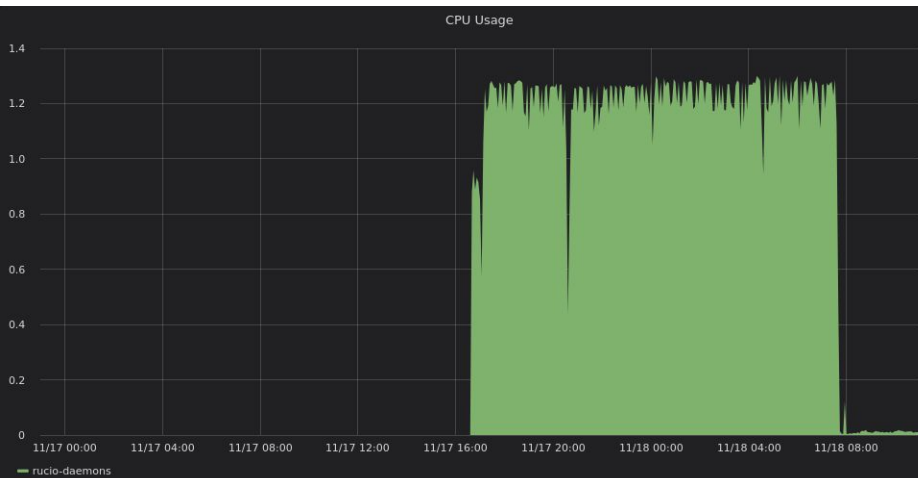
Server - after setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server (2)	4 2500 MiB	4 2500 MiB	2-4 1-2 GiB (>2.5 GiB cache)	2 restarts due to memory limit.



Daemon: Judge Evaluator

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-evaluator (3 x 3 threads)	2 2000 MiB	2 2000 MiB	1.3 1-2 GiB	Problematic 2, 9, 2 restarts. Tried different configurations (2 x 5 threads). Not understood.

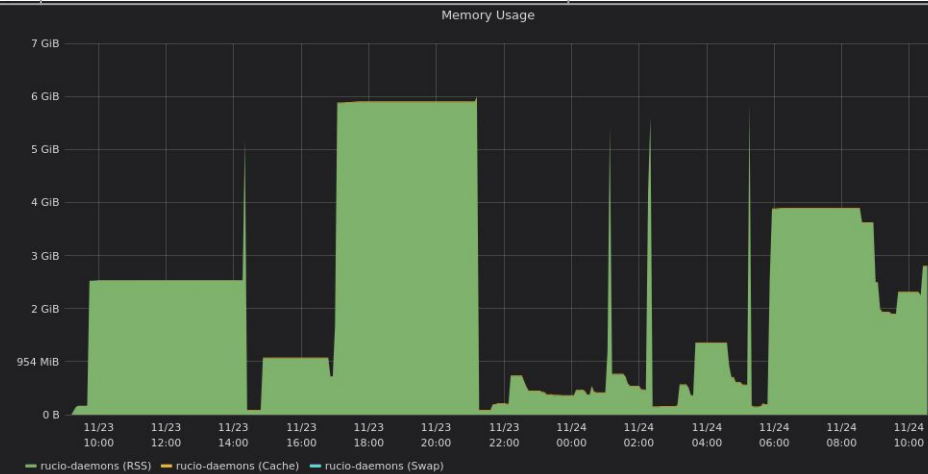
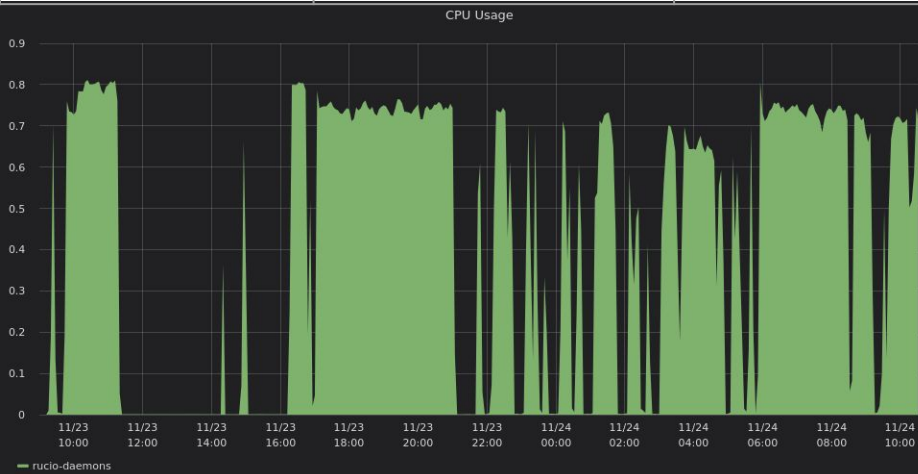


- Rucio → (#replicas) [CPU (*limits)|Memory]:
 - UI → (1) [0.1|500 (*800) MiB];
 - Auth Server → (2) [0.2 (*1)|0.5 (*1) GiB];
 - Main Server → (2) [2 (*4)|2 (*4) GiB];
 - Daemons:
 - Abacus Account → (1) [0.1|150 MiB];
 - Abacus Collection Replica → (1) [0.4|200 MiB];
 - Abacus RSE → (1) [0.1|150 MiB];
 - Conveyor Submitter → (3 x 4 threads) [0.8|400 MiB];
 - Conveyor Poller → (3 x 4 threads) [0.5|250 MiB];
 - Conveyor Finisher → (1 x 2 threads) [1(*1.5)|250 (*500) MiB];
- Hermes → (1) [0.1|200 MiB];
- Judge Injector → (2) [0.1 (*0.8)|200 (*400) MiB];
- Judge Evaluator → (3 x 3 threads) [2|3 GiB];
- Judge Repairer → (2 x 5 threads) [1|0.8 (*6) GiB];
- Judge Cleaner → (2 x 5 threads) [1|400 MiB];
- Reaper2 → (2 x 4 threads) [0.4|400 (*800) MiB];
- Transmogriifier → (1) [0.1|200 MiB];
- Undertaker → (1) [1|400 MiB].
- Total → (29) [21.3|21.60 GiB].
→ *28.8|38.75 GiB
- OpenStack VMs → (6 nodes) [8|16 GiB].



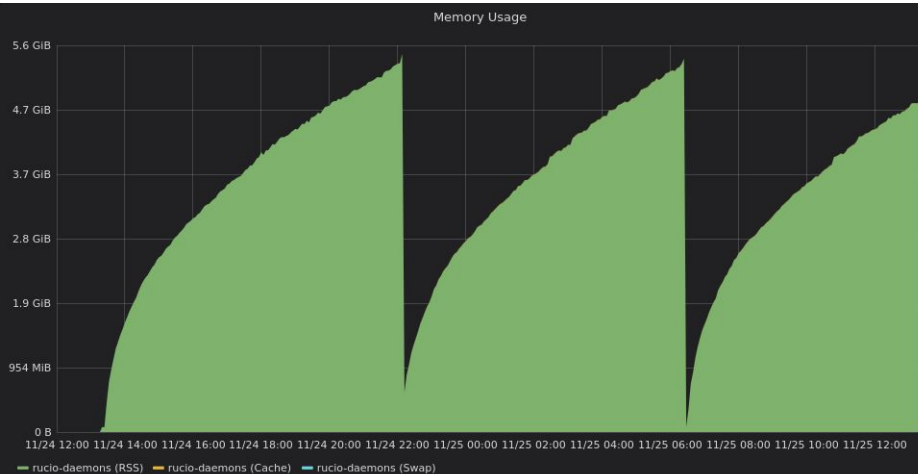
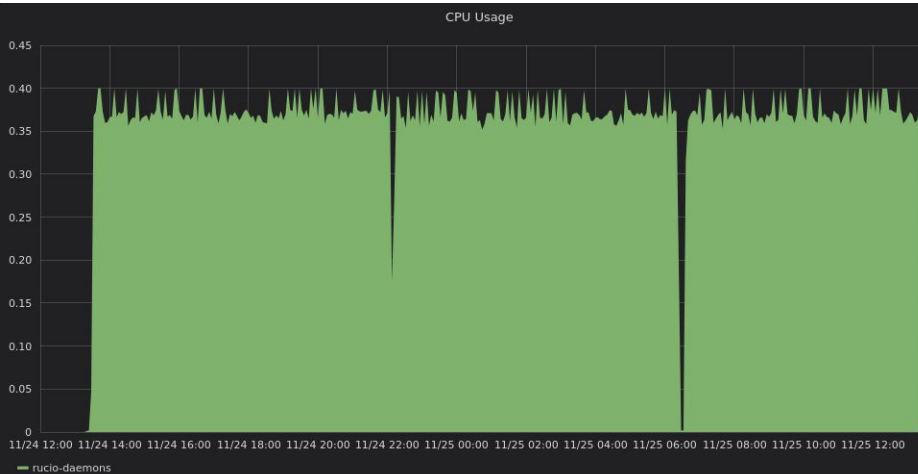
Daemon: Abacus Collection Replica - FDR Takeaway

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	Post-FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
abacus-collection-replica	1 4 GiB	1 6 GiB	0.8 4.5 (>6) GiB	Restarts due to OOMKilled. Cannot keep up. Millions of rows in UPDATED_COL_REP. Limits changed again to: (*8) GiB

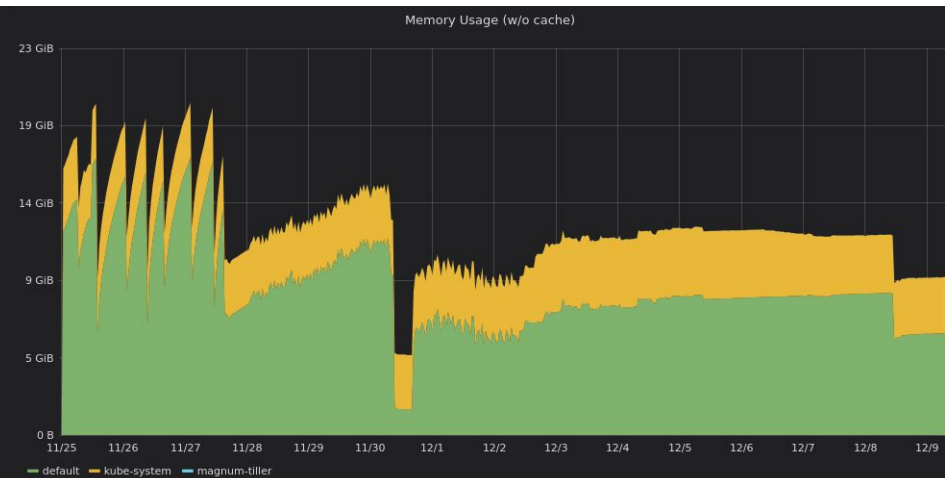
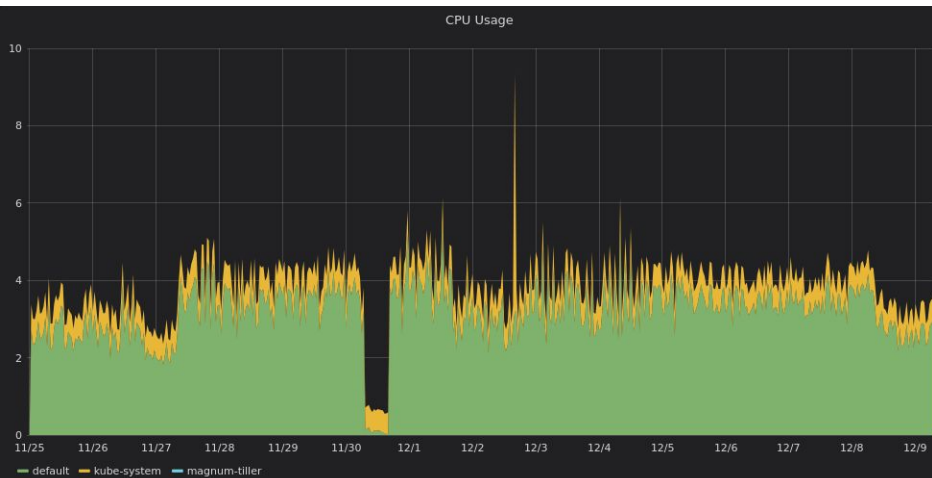
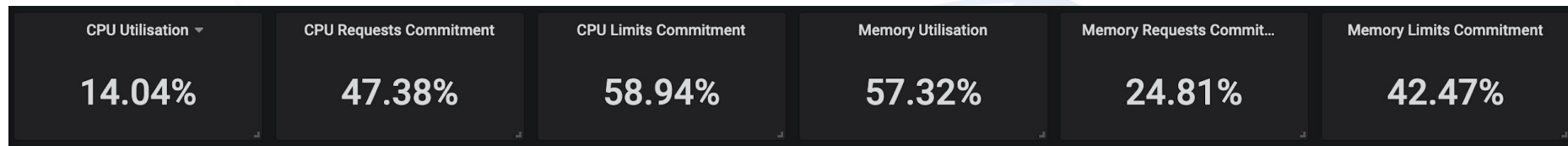


Daemon: Judge Injector - FDR Takeaway

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	Post-FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-injector (2)	0.2 400 MiB	0.4 5.5 GiB	0.4 >6 GiB	Problematic Limits changed again to: 0.4(*1) 1(*10) GiB



K8s Cluster Utilisation



Full Dress Rehearsal Takeaway (1/2)

- Infrastructure should be resource-aware for the project sustainability (**minimal env.**).
 - ATLAS and CMS already run Rucio with a great amount of resources at disposal.
 - Full Dress Rehearsal proved:
 - OpenStack VMs → (6 nodes) [8 | 16 GiB]
 - ESCAPE Rucio needs **21.3 (*28.8) CPUs** and **21.6 (*38.75) GiB** for 29 k8s-pods
- FDR highlighted few problems - infrastructure should be improved:
 - **Rucio(daemons)** - synergy is key for Rucio exportability (**Rucio team very engaged**). Exploring new Rucio phase space → hidden from ATLAS due to established model.
 - Main and Auth servers limits: **SOLVED** on-the-fly.
 - Abacus Collection Replica Daemon: **SOLVED** post FDR (@Martin - thanks!).



Full Dress Rehearsal Takeaway (2/2)

- Judge (Injector, Evaluator, Repairer) Daemons 1M-file rule: **SOLVED** post FDR (@RucioTeam - thanks!). Using/testing a new algorithm (also for ATLAS).
- k8s/CERN-GitOps for R&D work and stable production environment: **ON-GOING** e.g. syncs and tests are temporarily down.
- DB (devdb19u) problematic: **SOLVED** post FDR - moved to PROD+DEV.
- Sites involved and responsive.
 - GSI-ROOT RSE space issue (contributing with quota 1 TB) **SOLVED** centrally on-the-fly.
- Sciences and experiments aligned.
 - Contributing with more and more realistic use cases and workflows ([logbook](#)).
 - LSST batch issue immediately **SOLVED** with a workaround.



On-Going and To-Do-ASAP

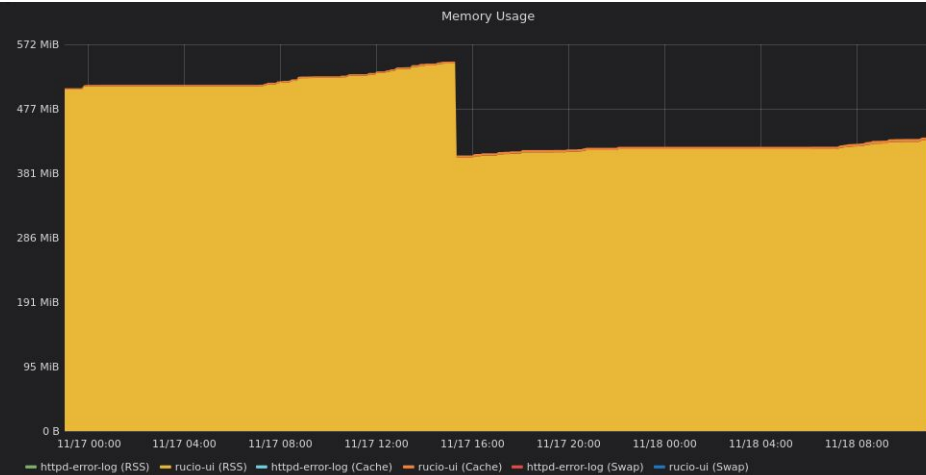
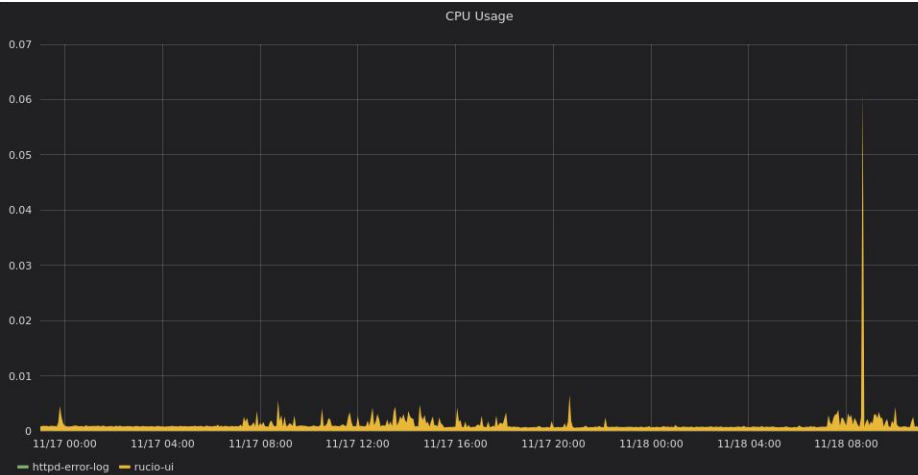
- **Upgrade infrastructure as a primary goal → k8s + CERN-GitOps**
- Improve tests (incl. HammerCloud) and monitoring
- **Collaboration for a second/test rucio instance (DB at CERN for the time being)**
- **Enable or develop more features, e.g. Rucio multi-VO, tokens, etc.**
- Investigate data corruption
- HammerCloud ready to run realistic research infrastructure workloads
- Real data distribution and analysis for non-HEP RI (LOFAR, CTA, LSST, MAGIC)
- Ability to plug heterogeneous clouds (commercial)
- Changing QoS within a site and across sites



Backup

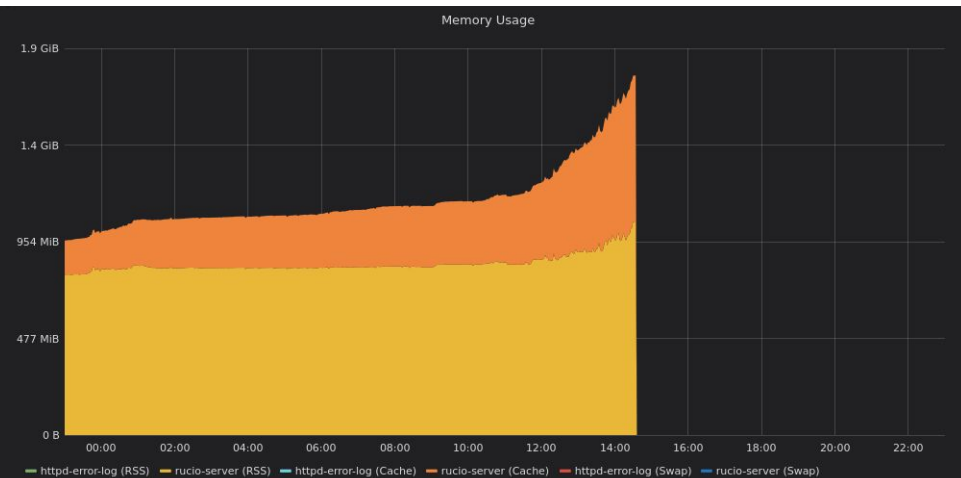
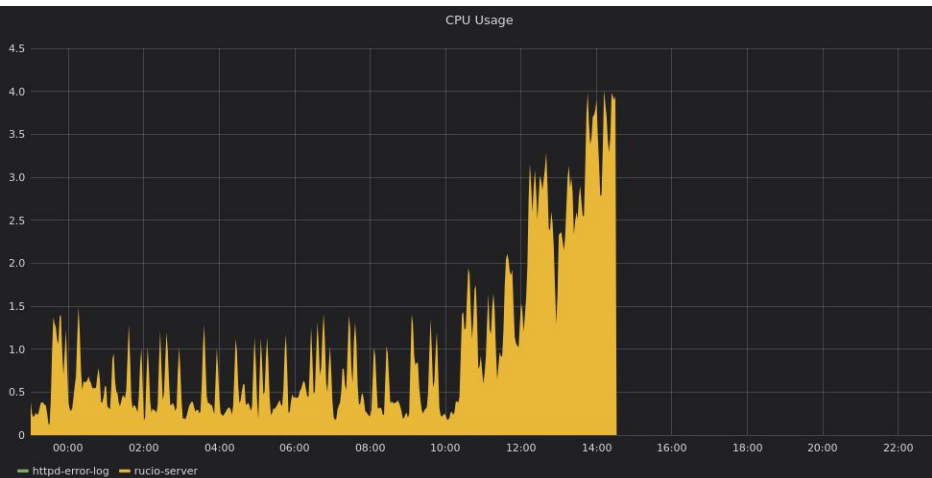


Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
ui	-	-	0.004 (0.06) 550 MiB	OK



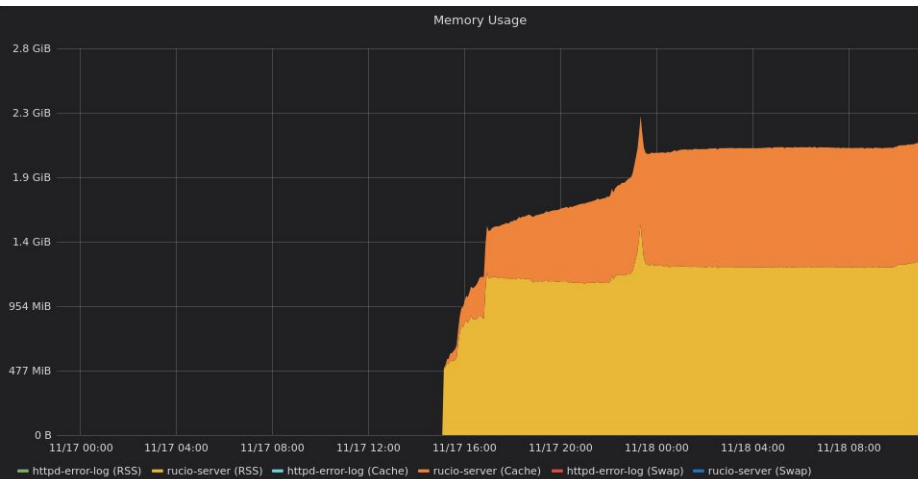
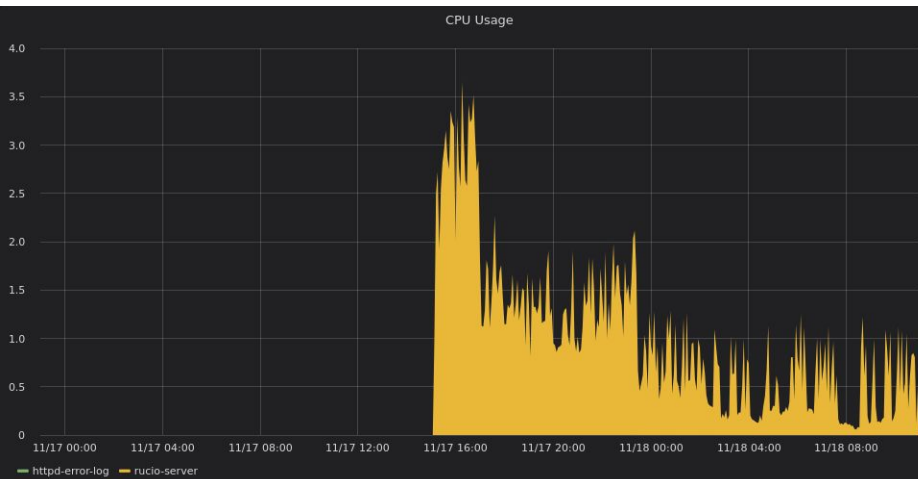
Server - before setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server (2)	-	-	2-4 1.25 GiB (2 GiB)	Manual restart due to no limits set.



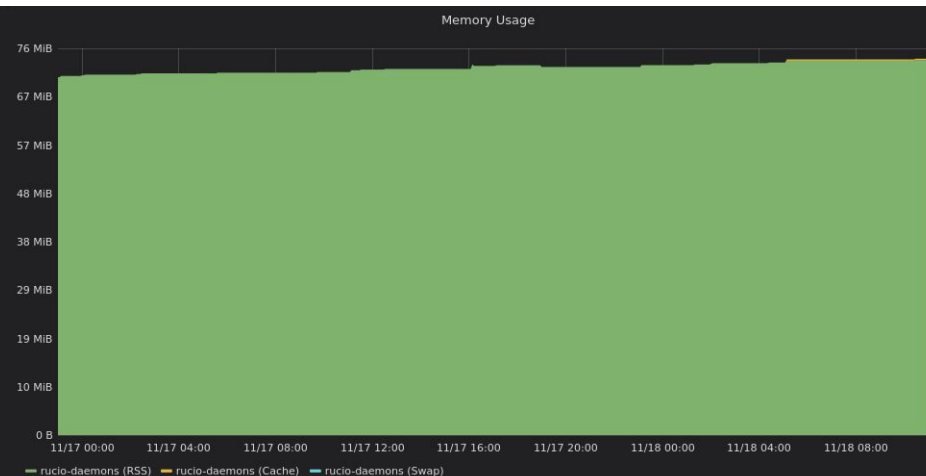
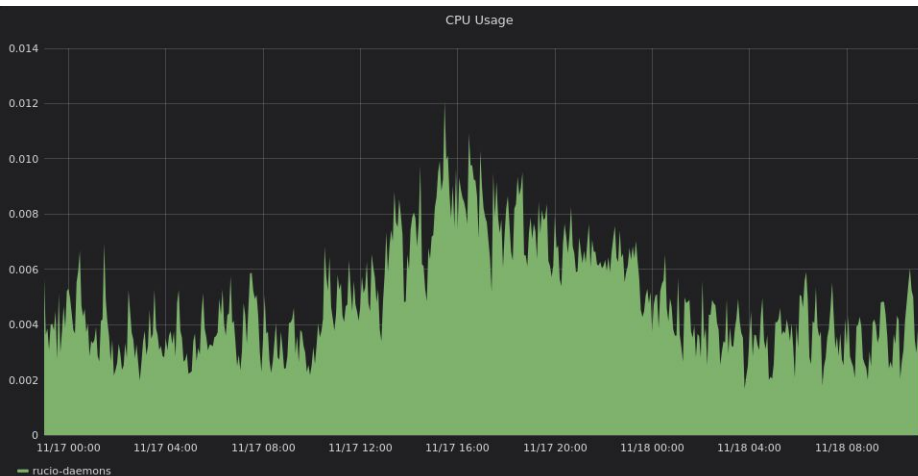
Server - after setting resources requests/limits

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
server (2)	4 2500 MiB	4 2500 MiB	2-4 1-2 GiB (>2.5 GiB cache)	2 restarts due to memory limit.



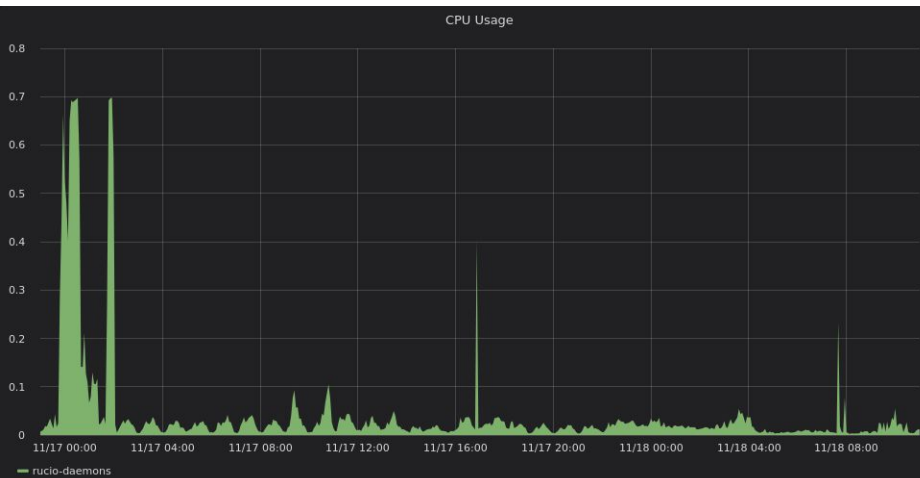
Daemon: Abacus Account

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
abacus-account	0.70 200 MiB	0.70 200 MiB	0.01 80 MiB	OK



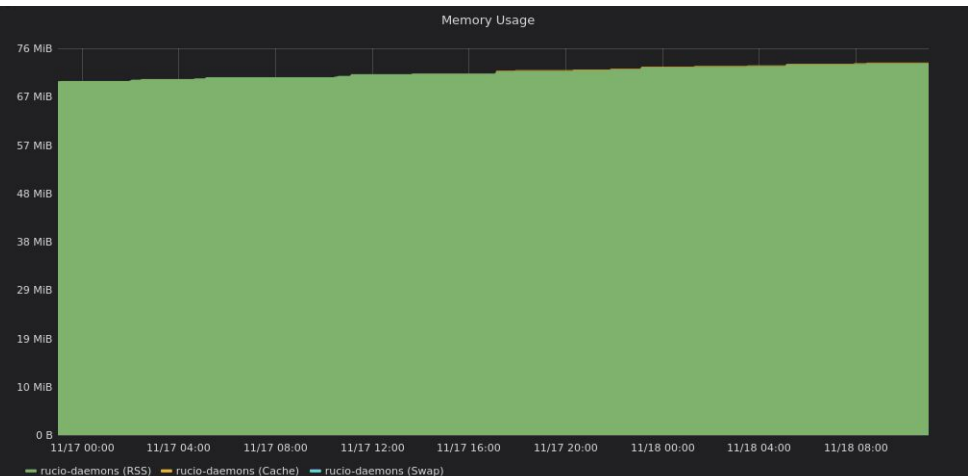
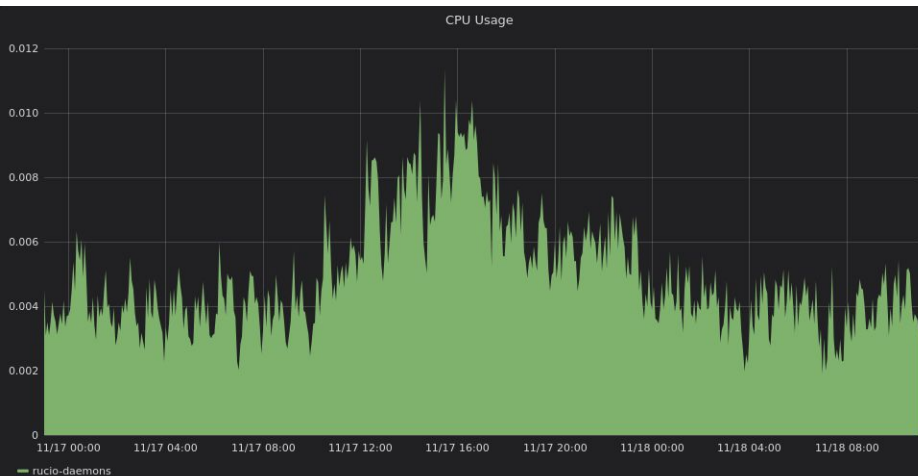
Daemon: Abacus Collection Replica

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
abacus-collection-replica	0.70 7000 MiB	0.70 7000 MiB	0.01-0.2 (0.7) 3 GiB	OK 1 restart due to OOMKilled but not shown in plot. Not understood.



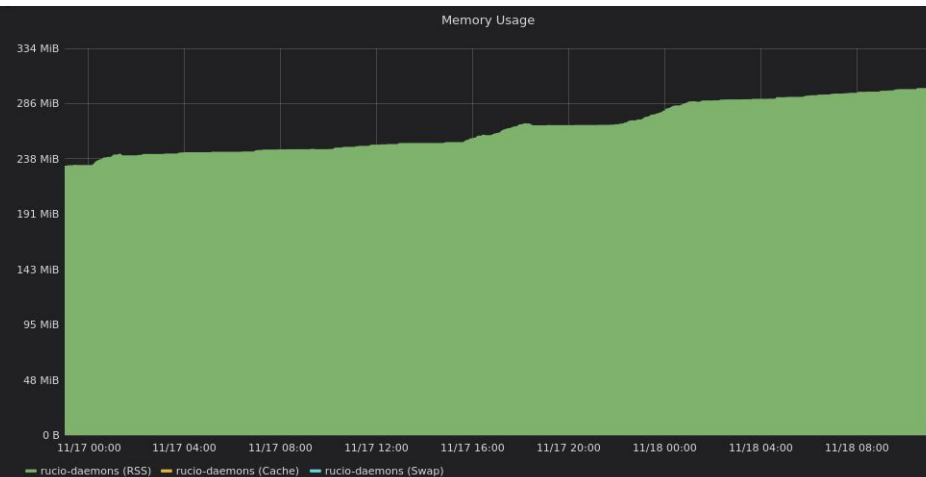
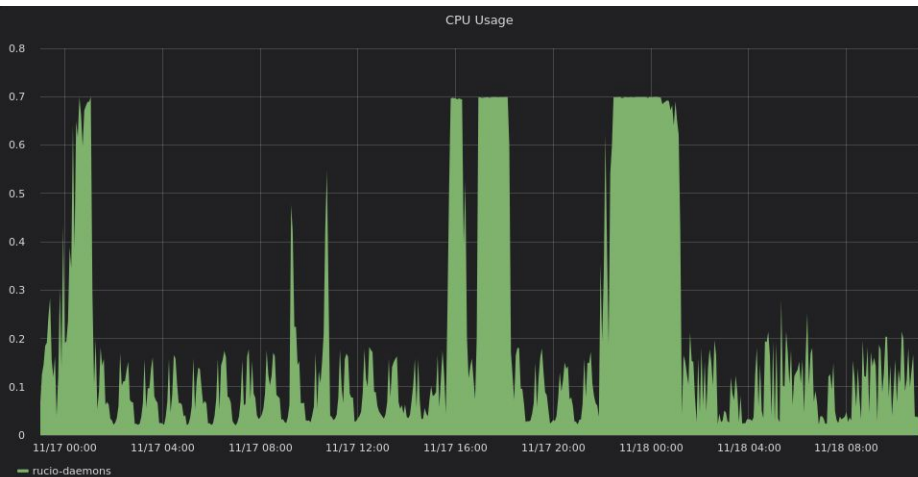
Daemon: Abacus RSE

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
abacus-rse	0.70 200 MiB	0.70 200 MiB	0.01 80 MiB	OK



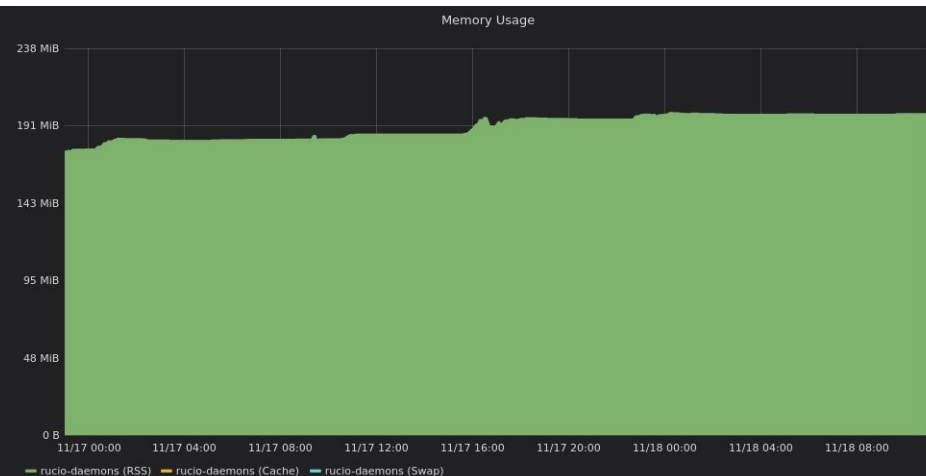
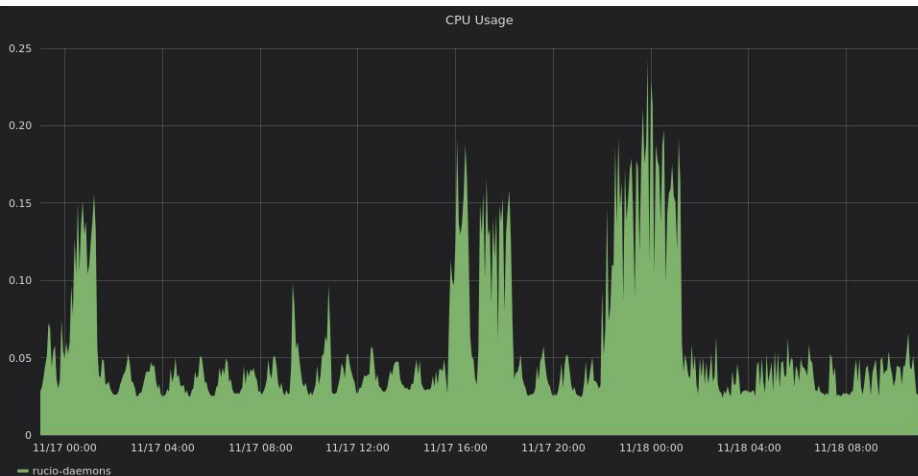
Daemon: Conveyor Submitter

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
conveyor-submitter (3 x 4 threads)	0.70 800 MiB	0.70 800 MiB	0.2-0.7 300 MiB	OK



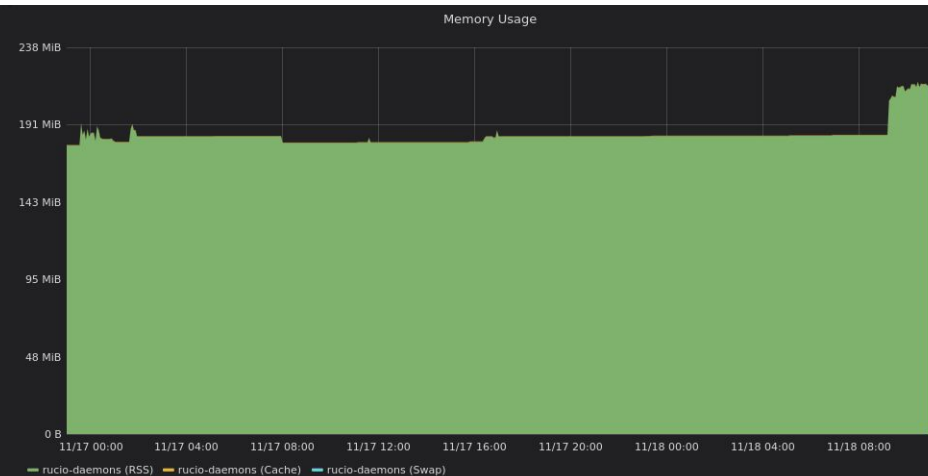
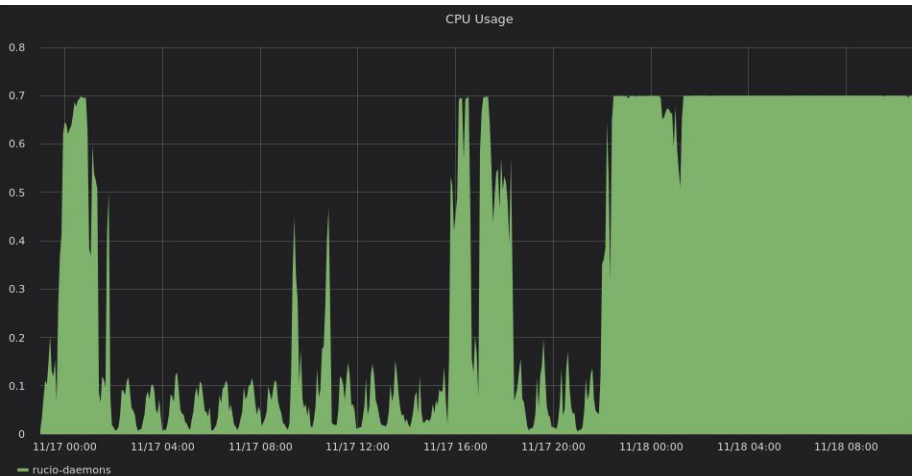
Daemon: Conveyor Poller

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
conveyor-poller (3 x 4 threads)	0.70 800 MiB	0.70 800 MiB	0.05-0.25 (0.4) 200 MiB	OK



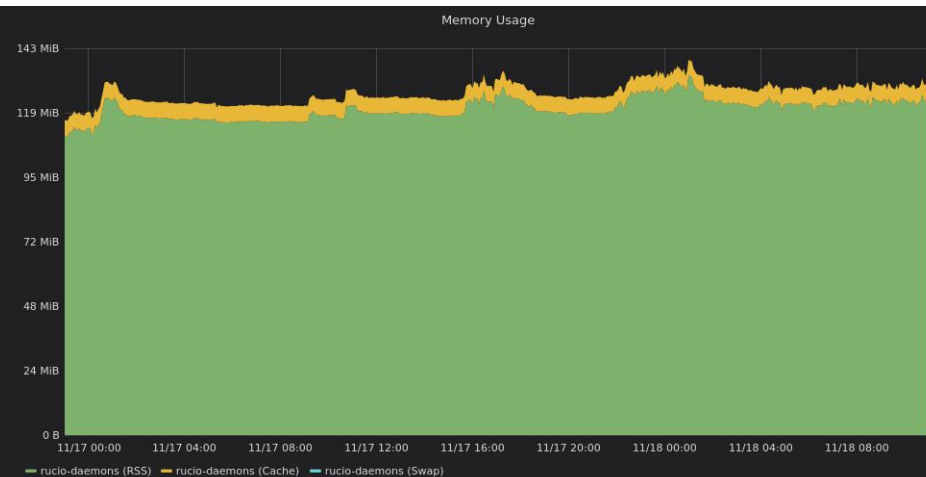
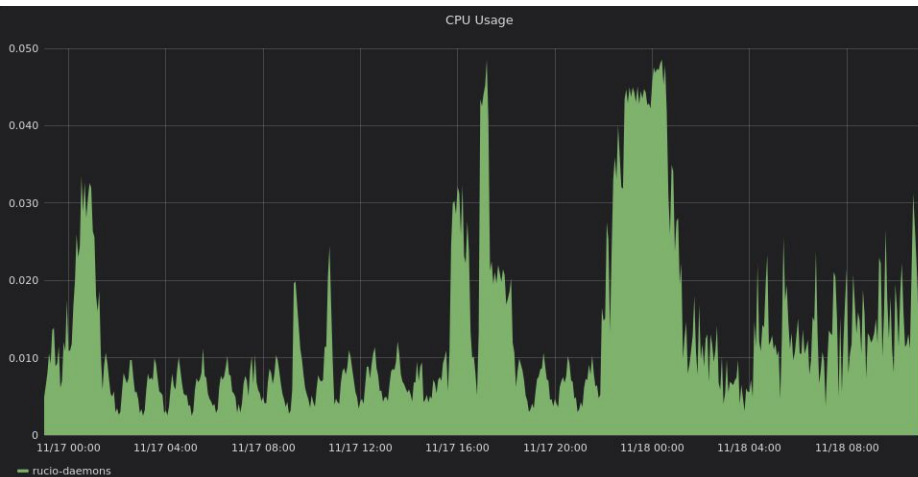
Daemon: Conveyor Finisher

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
conveyor-finisher (2 threads)	0.70 400 MiB	0.70 400 MiB	0.1-0.7 (0.7 maintained) 200 MiB	OK Probably worth to increase CPU.



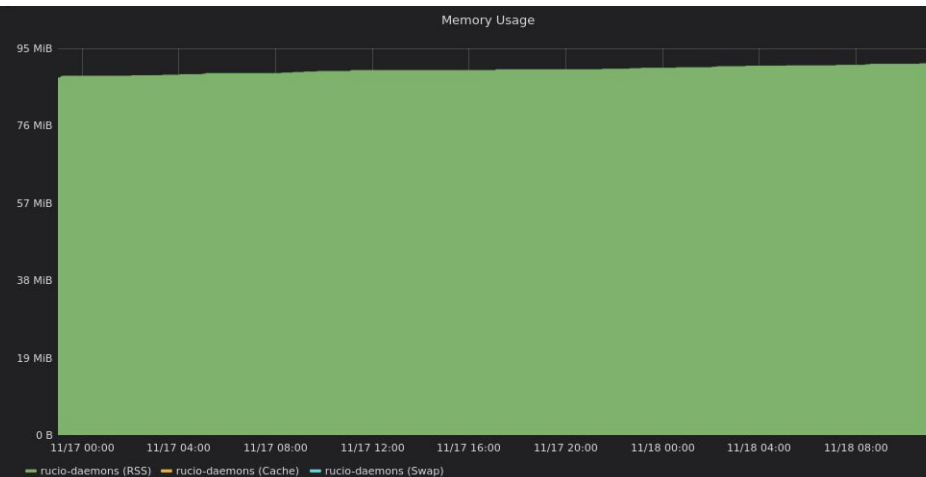
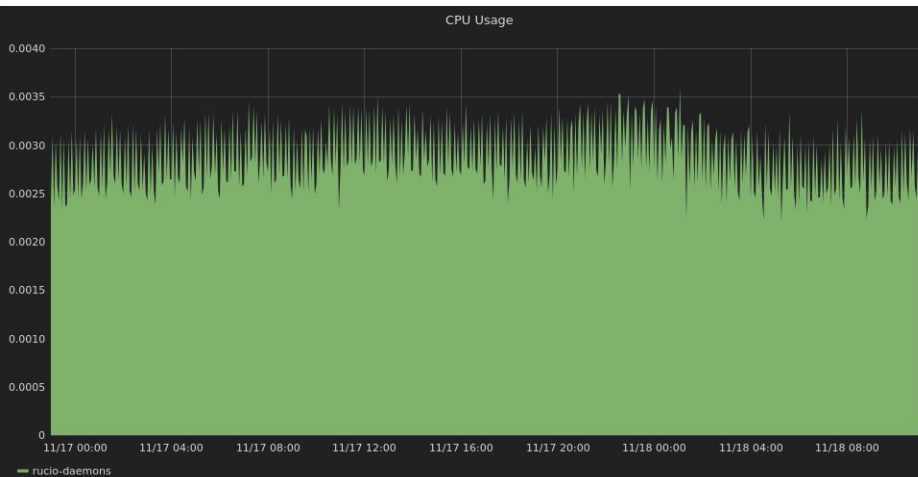
Daemon: Hermes

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
hermes	0.70 700 MiB	0.70 700 MiB	0.05 140 MiB	OK



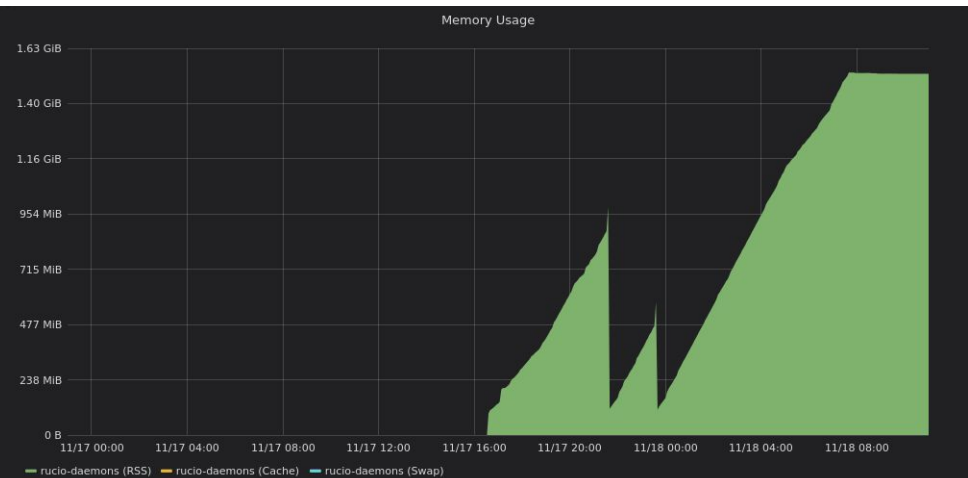
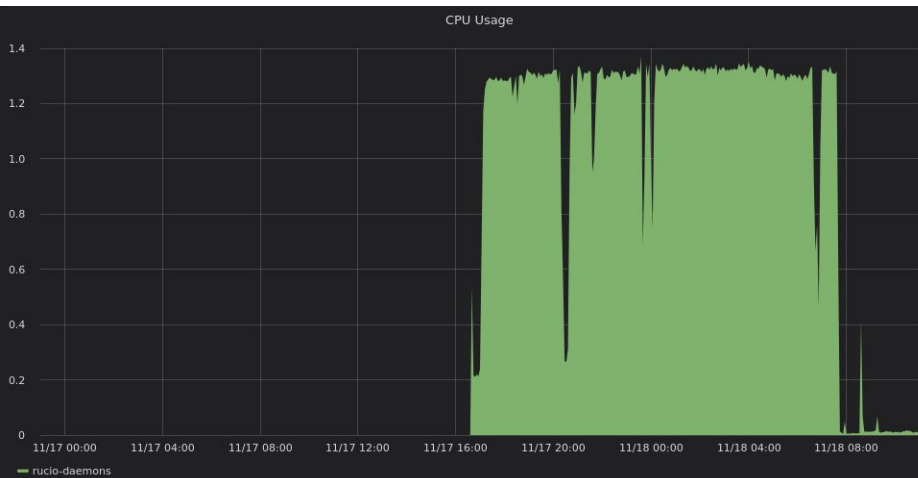
Daemon: Judge Injector

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-injector (2)	0.70 8000 MiB	0.70 8000 MiB	0.003 100 MiB	OK Limits are so high because was problematic in the past.



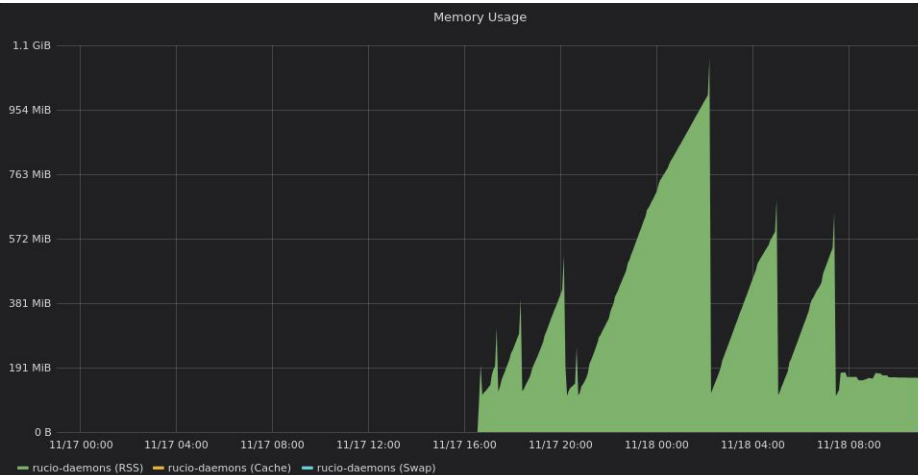
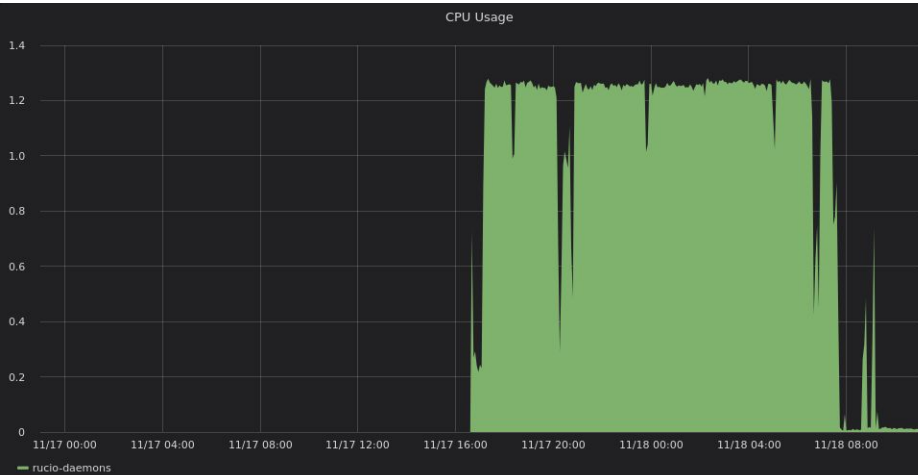
Daemon: Judge Evaluator

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-evaluator (3 x 3 threads)	2 2000 MiB	2 2000 MiB	1.3 1-2 GiB	Problematic 2, 9, 2 restarts. Tried different configurations (2 x 5 threads). Not understood.



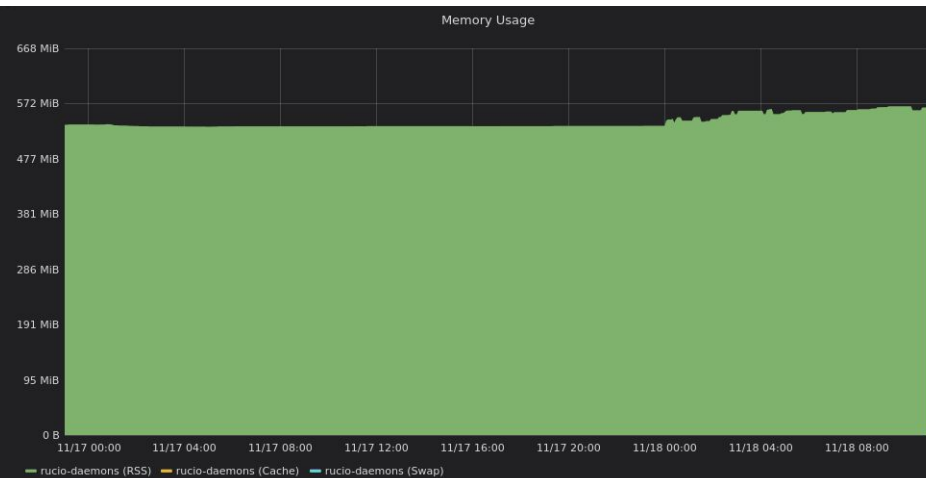
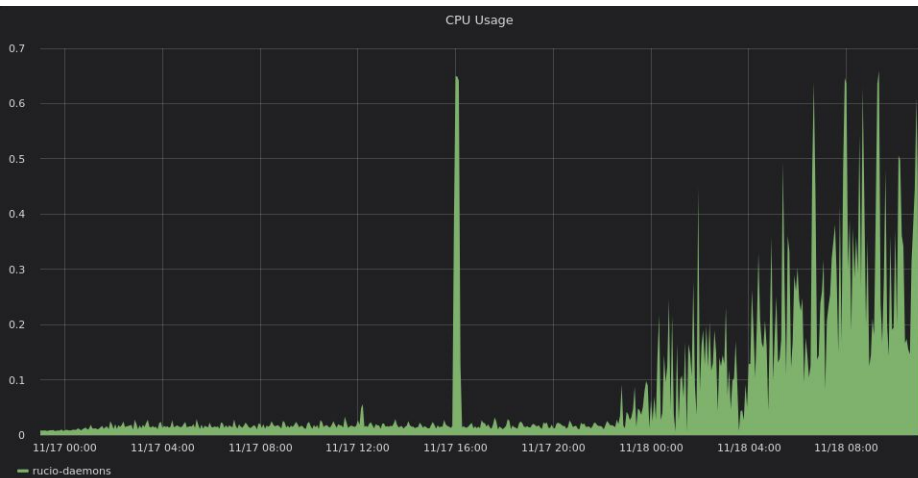
Daemon: Judge Evaluator

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-evaluator (3 x 3 threads)	2 2000 MiB	2 2000 MiB	1.3 1-2 GiB	Problematic 2, 9, 2 restarts. Tried different configurations (2 x 5 threads). Not understood.



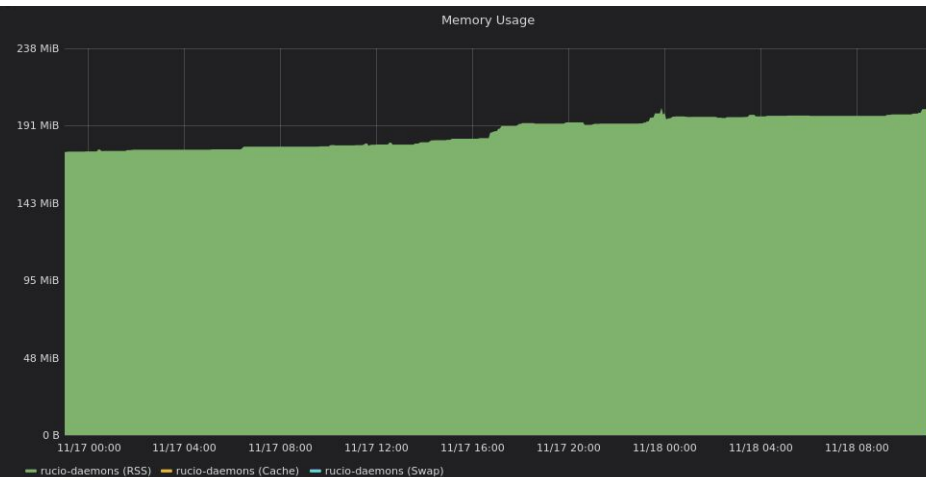
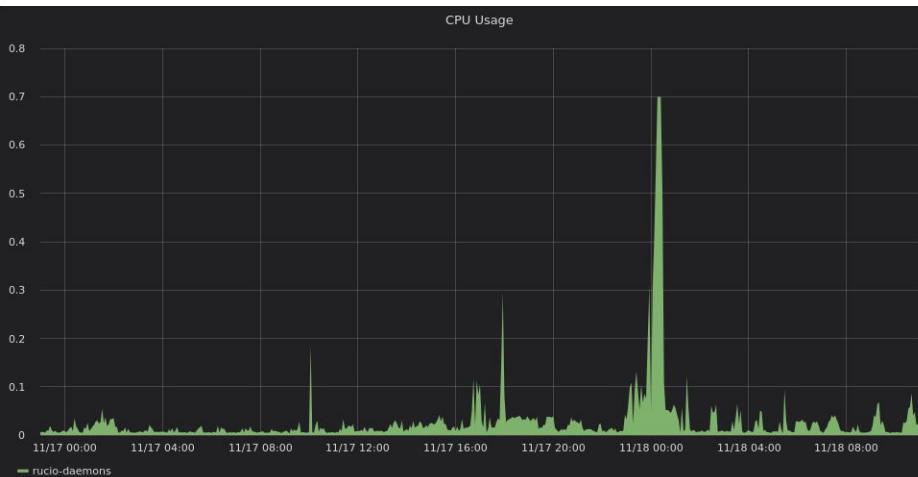
Daemon: Judge Repairer

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-repairer (2 x 5 threads)	0.70 1000 MiB	0.70 1000 MiB	0.2 (0.7) 600 MiB	OK



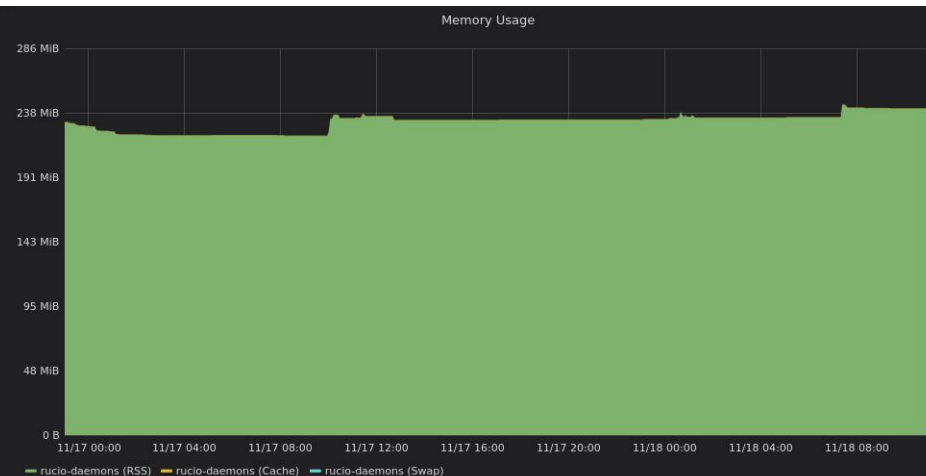
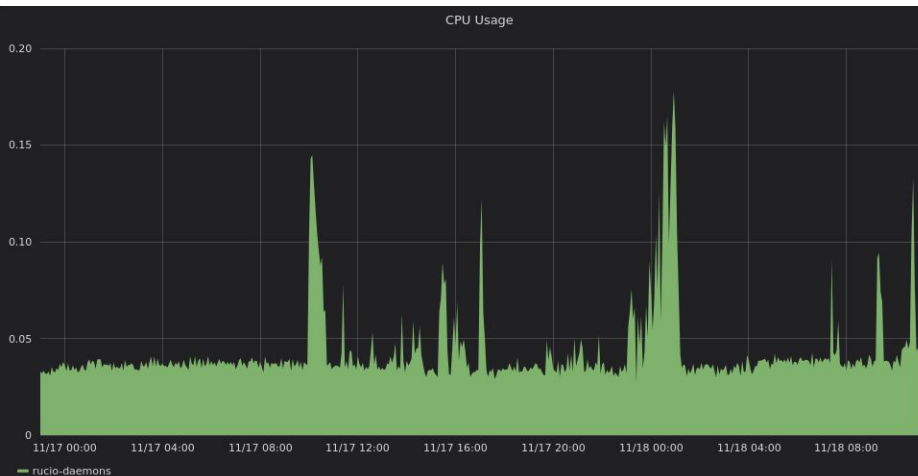
Daemon: Judge Cleaner

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
judge-cleaner (2 x 5 threads)	0.70 500 MiB	0.70 500 MiB	0.1 (0.7) 200 MiB	OK



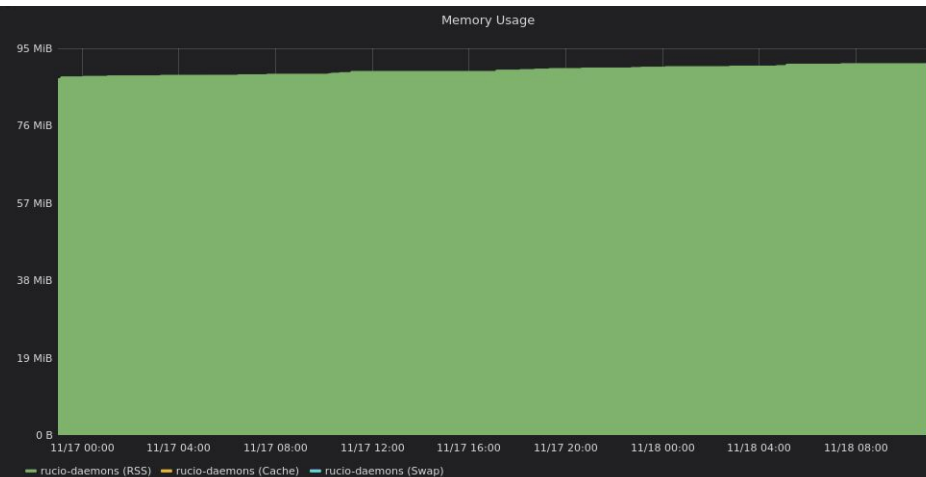
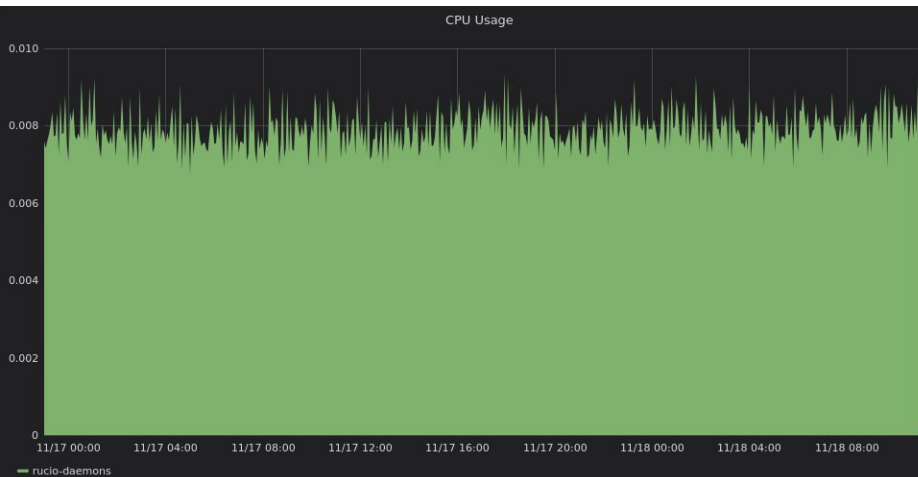
Daemon: Reaper2

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
reaper2 (2 x 4 threads)	0.70 400 MiB	0.70 400 MiB	0.05 (0.2) 250 MiB	OK



Daemon: Transmogrifier

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
transmogrifier	0.70 200 MiB	0.70 200 MiB	0.01 100 MiB	OK



Daemon: Undertaker

Service	Resources Requests [CPU Memory]	Resources Limits [CPU Memory]	FDR Usage of Resources [CPU (peak) Memory (peak)]	Restarts/Comments
undertaker	0.70 400 MiB	0.70 400 MiB	0.03 (0.7) 100 MiB (200 MiB)	OK Memory limit raised at 400 MiB due to OOMKilled at 200 MiB.

