Introduction:

- There are some job failures (order of 1/1000) that have been observed in big workflows at USDF.
- The failures are not too harmful because they usually succeed at retries, then the whole workflow is successful.
- To understand those failures we did some scaling test at FrDF, using the DC2 step1 pipeline and up to 3000 cores.
- We also had a successful run at UKDF with 100 cores. (The UKDF has just increased to 3000 cores)
- The tests didn’t show severe scaling issues for both the FrDF site and PanDA.
- Similar failures as seen at USDF did appear both in the connectivity to the GCS and CERN, but at rather low rate (single digit out of 57K)
- From the tests, the error “timeout in holding” was identified to be the DB lock in multithread. Wen has already reported it to the PanDA core team. **Page 2**
- Clustering was tested for this pipeline. The 5 pipetasks was grouped into 1 task, and this helps reduce the processing time from 4 hrs to 2 hrs at FrDF. **Page 3**
- The failures caused by log rotation were observed in the run at UKDF too. **Page 3**
- The majority of USDF failures are not due to NAT, but caused by the log rotation. The log rotation problem was identified by Wen earlier, but it’s hard to distinguish it from the error caused by the update of job status getting out of order. They have the same error message. A new test with a big workflow at USDF over the weekend confirms this reasoning by avoiding the log rotation time. (My apology, I should have understood and tested it earlier). **Page 4**

Conclusion:

At the current scale, there should be no concerns with the resources. The failure rate is rather low after the squid proxy has been improved. The majority of the remaining failures are caused by log rotation, DB lock and jobs hanging on the nodes if using the new dev version of pilot. These have already been reported to the PanDA core team.

Tests at FrDF with more cores.

Run **3916**, DC2 step1, **57K** jobs. **FrDF, 1000 cores**, w_2023_07

Error code	Nerrors	description	Cause
jobdispatcher:103 (sup:2)	22	timeout in holding : last heartbeat	**FrDF tests help understand this error**: database lock in multithread
pilot 1137	6	Failed to stage-out file: exception caught in gs client	No NAT or squid, so might be related to the GCS server
transformation:1	5	ssl.SSLZeroReturnError: TLS/SSL connection has been closed	butler file transfer failure

https://panda-doma.cern.ch/tasks/?days=100&reqid=3916

Run **3929**,   DC2 step1, **57K** jobs. **FrDF, 3000 cores**, w_2023_07

Error code	Nerrors	description	Cause
pilot 1137	4	Failed to stage-out file: exception caught in gs client	No NAT or squid, so might be related to the GCS server

https://panda-doma.cern.ch/tasks/?days=100&reqid=3929

Conclusion: no major issues from the FrDF site or PanDA with 3000 cores. The gs failures were observed, but at a rather low rate. The test helps us understand the error “timeout in holding”.

Tests on clustering:

Run **3933**, DC2 step1 with clustering, **11K** jobs. **FrDF, 3000 cores**, w_2023_07

Error code	Nerrors	description	Cause
taskbuffer:300	1	The worker was finished while the job was running	update job status out of order: network issue (**No log rotation caused failures**)

https://panda-doma.cern.ch/tasks/?days=100&reqid=3933

Run **3927**, DC2 step1 with clustering, **11K** jobs. **FrDF, 3000 cores**, w_2023_15

Error code	Nerrors	description	Cause
transformation:137	3	Transform received signal SIGKILL	Merge job failed for 3 attempts. Tim found errors in butler data transfer.

https://panda-doma.cern.ch/tasks/?days=100&reqid=3927

Conclusion: There are no major issues at FrDF site or PanDA. w_2023_15 has some butler issue that caused the merge job to fail (**has been fixed by Fabio yesterday**). The clustering groups 5 pipetasks into 1, and reduces the pipeline processing time from 4 hours to 2 hours.

DC2 at UKDF:

Run **3915**, DC2 step1, **57K** jobs. **UKDF, 100 cores**, w_2023_07

Error code	Nerrors	description	Cause
taskbuffer:300	19	The worker was finished while the job was running	**log rotation**
pilot 1137	3	Failed to stage-out file: exception caught in gs client	No NAT or squid, so might be related to the GCS server
pilot 1144	1	This job was killed by panda server	update job status out of order: network issue (concurrent runs 3922-3925 at USDF)

https://panda-doma.cern.ch/tasks/?days=100&reqid=3915

To understand better the errors in bigger runs. Take RC2 step4 for example.

Run **3853**, RC2 step4, **260K** jobs. **2023-04-10** to **2023-04-11, USDF**

Error code	Nerrors	description	Cause
taskbuffer:300	420	The worker was finished while the job was running	1. **log rotation** (major one) 2. update job status out of order: network issue
jobdispatcher:100	197	lost heartbeat	appear in new pilot version, some jobs hang on the nodes
jobdispatcher:103 (sup:2)	11	timeout in holding : last heartbeat	FrDF tests help understand this error: database lock in multithread
jobdispatcher:102	26	Sent job didn't receive reply from pilot within 30 min	jobs failed to update “starting” status: network issue
pilot 1305	66	Failed to execute payload:invalid value encountered in sqrt	payload failure
transformation:139	15	Transform received signal SIGSEGV	payload failure

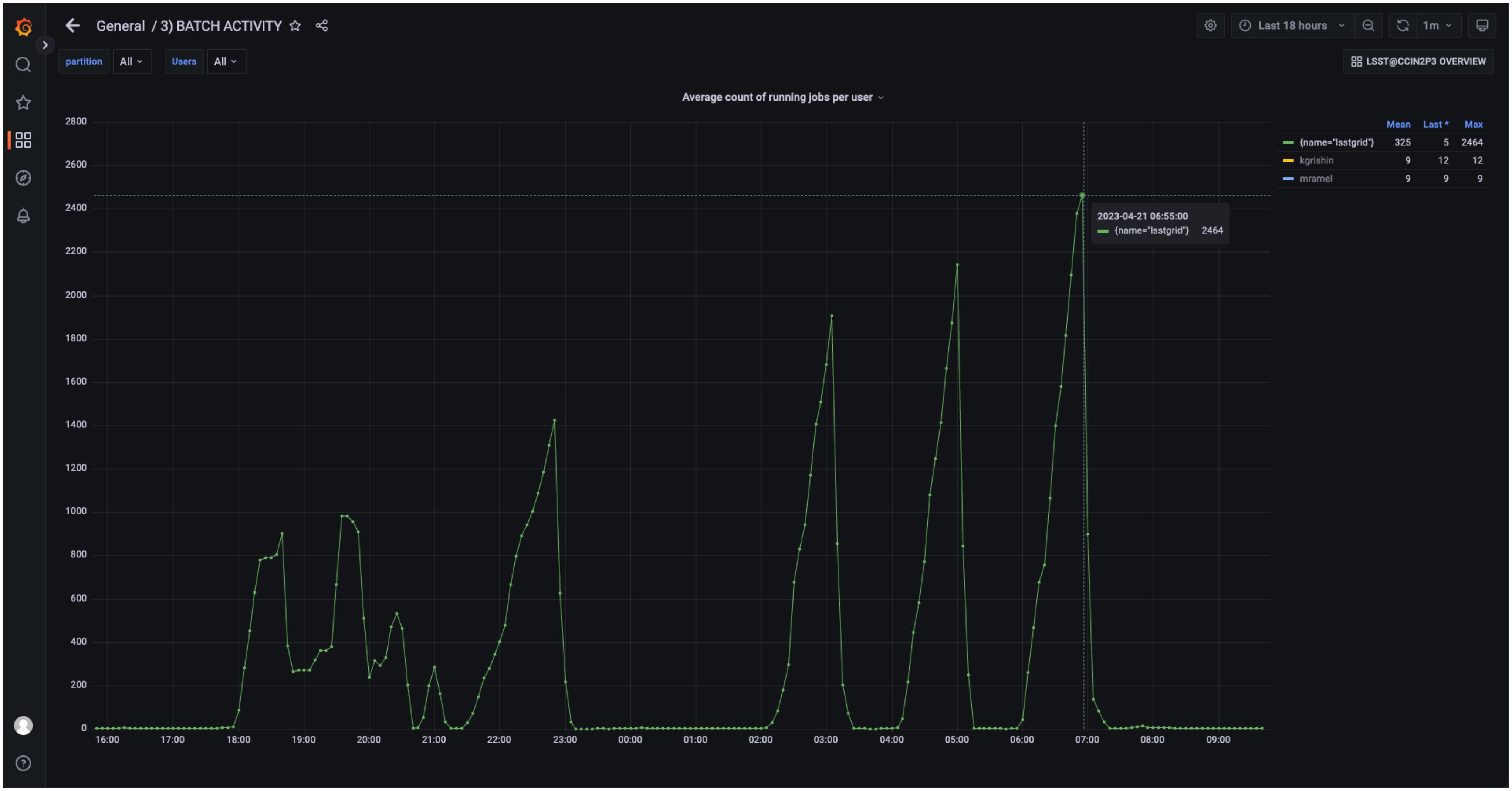
https://panda-doma.cern.ch/tasks/?days=100&reqid=3853

Run **3944**, RC2 step4, **260K** jobs. **2023-04-22, USDF** (to avoid log rotation time)

Error code	Nerrors	description	Cause
taskbuffer:300	**1**	The worker was finished while the job was running	update job status out of order: network issue (**No log rotation caused failures**)
jobdispatcher:100	42	lost heartbeat	appear in new pilot version, some jobs hang on the nodes
jobdispatcher:103 (sup:2)	3	timeout in holding : last heartbeat	FrDF tests help understand this error: database lock in multithread
pilot 1137	7	Failed to stage-out file: exception caught in gs client	Maybe related to rubin-panda-iam-dev using squid too, as Yee pointed out. **Didn’t show in run 3853**
pilot 1305 (transformation:1)	53	Failed to execute payload:Unable to calculate psf matching kernel	payload failure

https://panda-doma.cern.ch/tasks/?days=100&reqid=3944

Monitoring at FrDF



General / 3) BATCH ACTIVITY

partition All

Users All

Last 18 hours

1m

LSST@CCIN2P3 OVERVIEW

Average count of running jobs per user

	Mean	Last *	Max
{name="lsstgrid"}	325	5	2464
kgrishin	9	12	12
mramel	9	9	9

2023-04-21 06:55:00
{name="lsstgrid"} 2464