Science and Technology Facilities Council

ATLAS

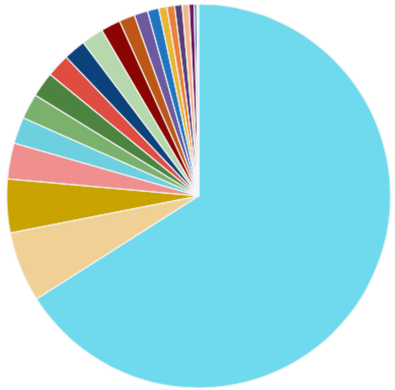
J. Walder

19/4/2020

General Picture

- Use Cern MONIT to compare to the Tier-1s: (20th April - 28th)
  - Timeout errors are higher fractionally at RAL than elsewhere

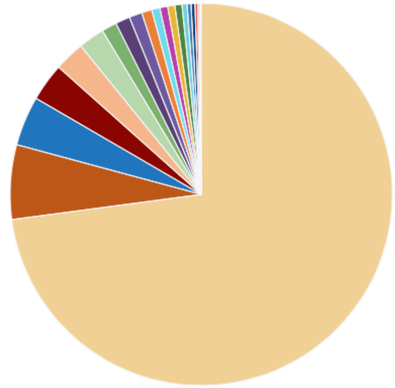
Panda Failure Categories - Pie chart



All

	current	percentage
TaskBuffer Error: Timeout	587 K	66%
TaskBuffer Error 300	52.7 K	6%
Pilot/DDM Error: get error	40.1 K	5%
Transformation Error: not installed in CE	26.8 K	3%
DDM Error	19.8 K	2%
Athena/Exec Error: Athena	18.9 K	2%
Pilot Error	18.8 K	2%
Execution Error 65	17.4 K	2%
Pilot/PanDA Error: Put error	16.6 K	2%
TaskBuffer Error	16.6 K	2%
Pilot/PanDA Error: Get error	14.63 K	2%
Pilot/PanDA Error	11.79 K	1%
Job Dispatcher Error: lost heartbeat	10.20 K	1%
Execution Error 68	8.47 K	1%
Athena/Exec Error: Proot	6.08 K	1%
Execution Error	5.80 K	1%
Pilot/Runtime Error	5.60 K	1%
Transformation Error	5.07 K	1%

Panda Failure Categories - Pie chart

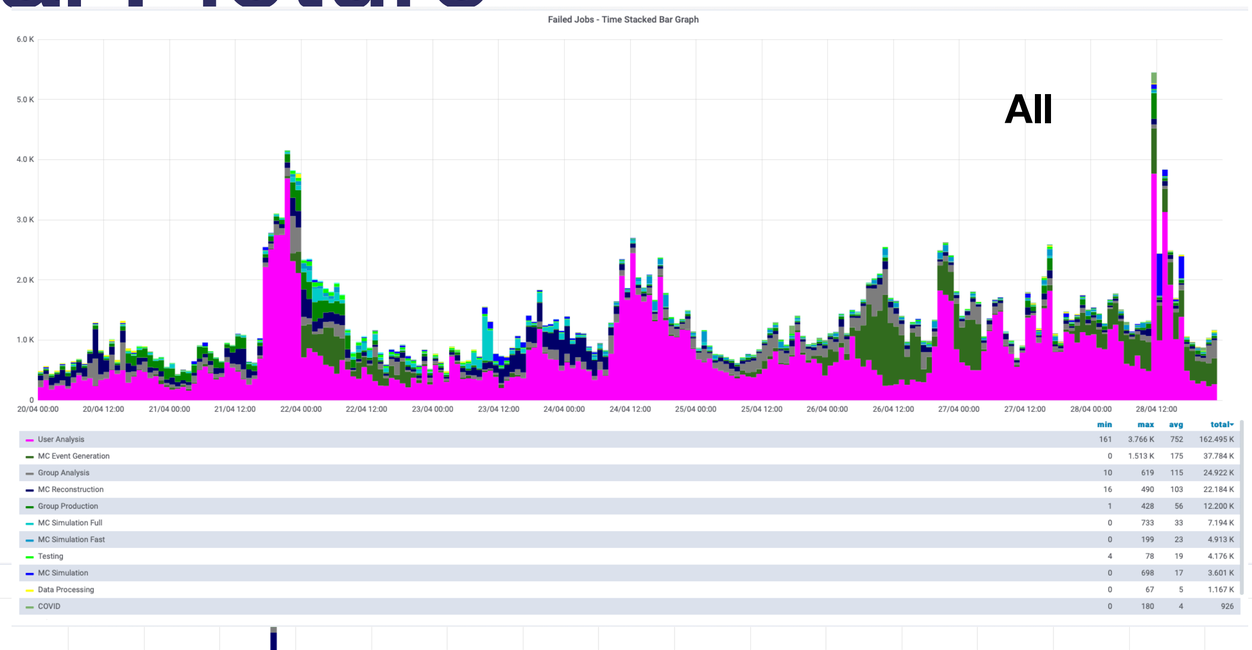


RAL

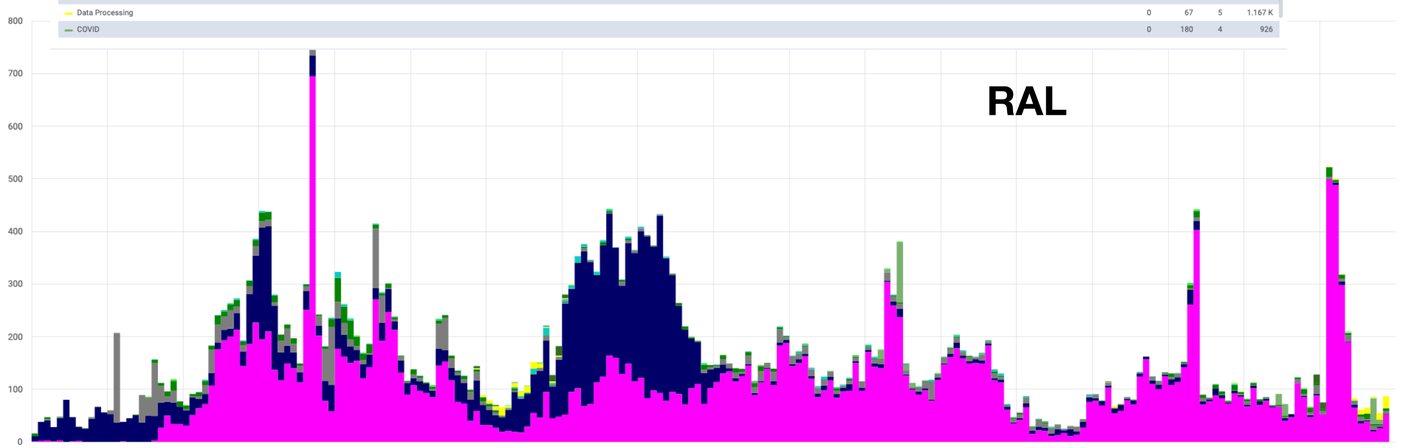
	current	percentage
TaskBuffer Error: Timeout	104.6 K	73%
Pilot/PanDA Error: Get error	9.03 K	6%
Execution Error 68	6.01 K	4%
Pilot/PanDA Error: Put error	4.60 K	3%
Transformation Error: not installed in CE	3.68 K	3%
TaskBuffer Error 300	3.16 K	2%
Athena/Exec Error: Athena	1.838 K	1%
TaskBuffer Error	1.763 K	1%
Pilot Error	1.568 K	1%
Execution Error	1.250 K	1%
Transformation Error	989	1%
Job Dispatcher Error: lost heartbeat	948	1%
Athena/Exec Error: Proot	872	1%
Pilot/DDM Error: get error	859	1%
DDM Error	611	0%
Pilot/PanDA Error	477	0%
Pilot/PanDA Error: killed by panda server	472	0%
Execution Error 65	327	0%

General Picture

- Total avg rate of 1.3k failures:
  - RAL contributes avg 167 failures to the rate
- User analysis jobs ~ 1/2 of failures
- Squid/ATLAS issue included in this period



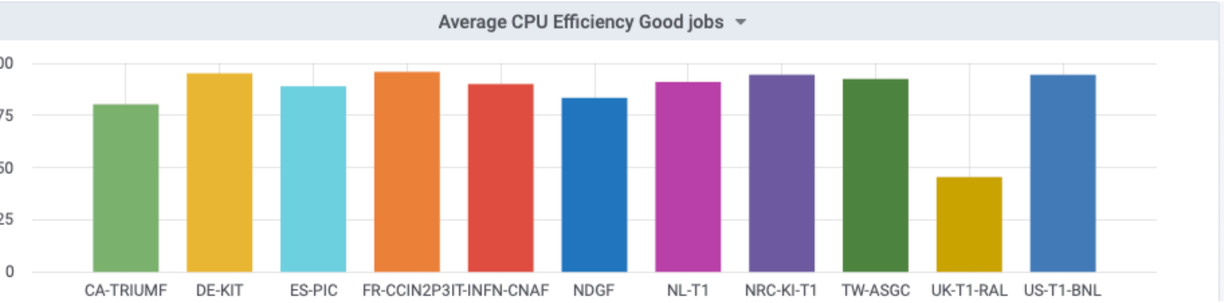
	min	max	avg	total
User Analysis	161	3.766 K	752	162.495 K
MC Event Generation	0	1.513 K	175	37.784 K
Group Analysis	10	619	115	24.922 K
MC Reconstruction	16	490	103	22.184 K
Group Production	1	428	56	12.200 K
MC Simulation Full	0	733	33	7.194 K
MC Simulation Fast	0	199	23	4.913 K
Testing	4	78	19	4.176 K
MC Simulation	0	698	17	3.601 K
Data Processing	0	67	5	1.167 K
COVID	0	180	4	926

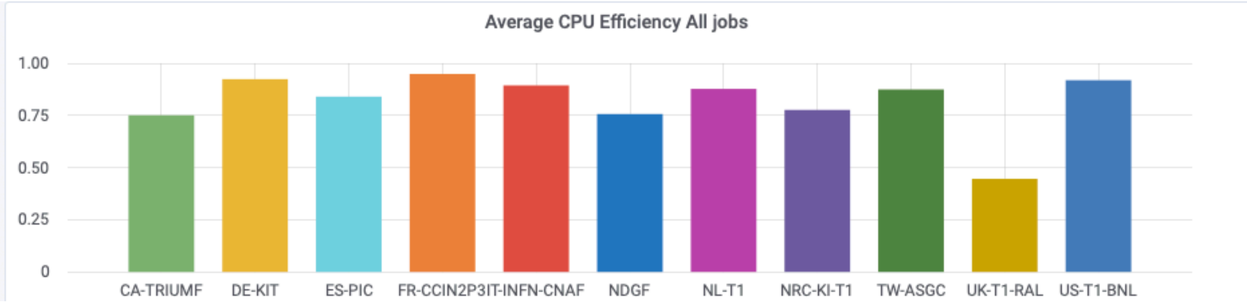


	min	max	avg	total
User Analysis	0	695	104	22.463 K
MC Reconstruction	0	337	44	9.486 K
Group Analysis	0	170	10	2.233 K
Group Production	0	46	4	952
COVID	0	115	1	259
MC Simulation Full	0	14	1	246
Data Processing	0	23	1	185
MC Event Generation	0	18	0	90
Testing	0	3	0	52
Others	0	1	0	7
MC Simulation Fast	0	2	0	3

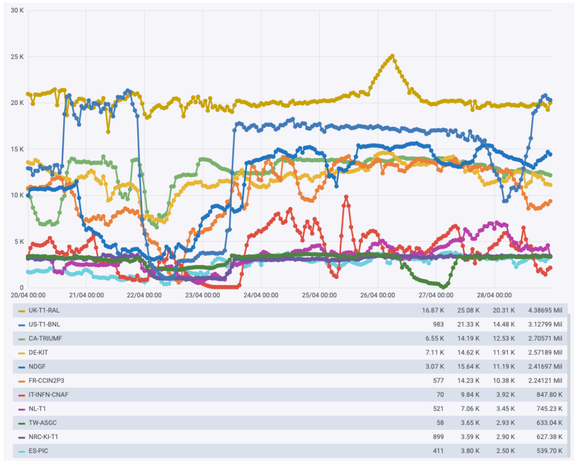
CPU Efficiency

- RAL < 50% cpu efficiency:
  - Other sites closer to 75%





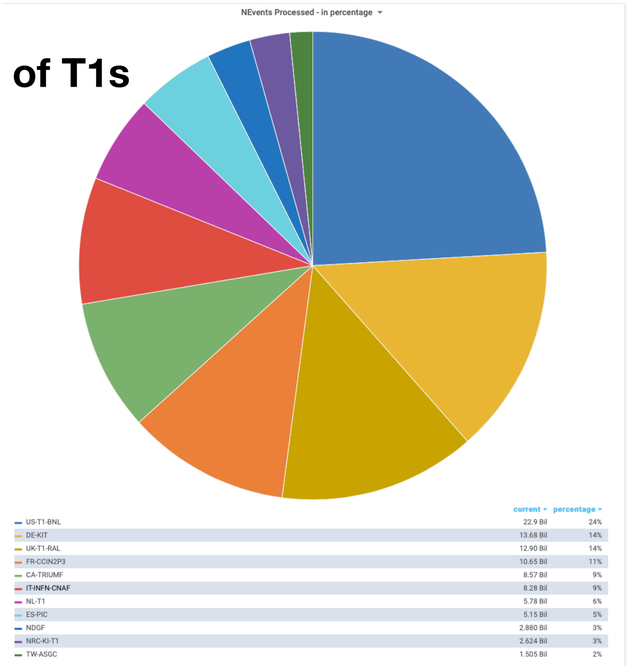
- RAL contributes typically most slots



By HS06, rank slightly lower

	min	max	avg	total
CA-TRIUMF	104.3 K	252.5 K	217.6 K	47.0001 Mil
UK-T1-RAL	168.5 K	250.7 K	203.0 K	43.8443 Mil
US-T1-BNL	11.9 K	262.4 K	177.3 K	38.3066 Mil
NDGF	41.9 K	215.4 K	155.1 K	33.5077 Mil
DE-KIT	89.1 K	184.7 K	149.8 K	32.3532 Mil
FR-CCIN2P3	6.2 K	153.2 K	111.6 K	24.1066 Mil
NL-T1	7.2 K	108.2 K	49.3 K	10.6588 Mil
NRC-KI-T1	14.1 K	55.9 K	45.3 K	9.7928 Mil
IT-INFN-CNAF	329	107.3 K	42.6 K	9.2103 Mil
TW-ASGC	759	50.6 K	40.4 K	8.7203 Mil
ES-PIC	5.6 K	52.7 K	34.7 K	7.4985 Mil

14% of Events processed of T1s



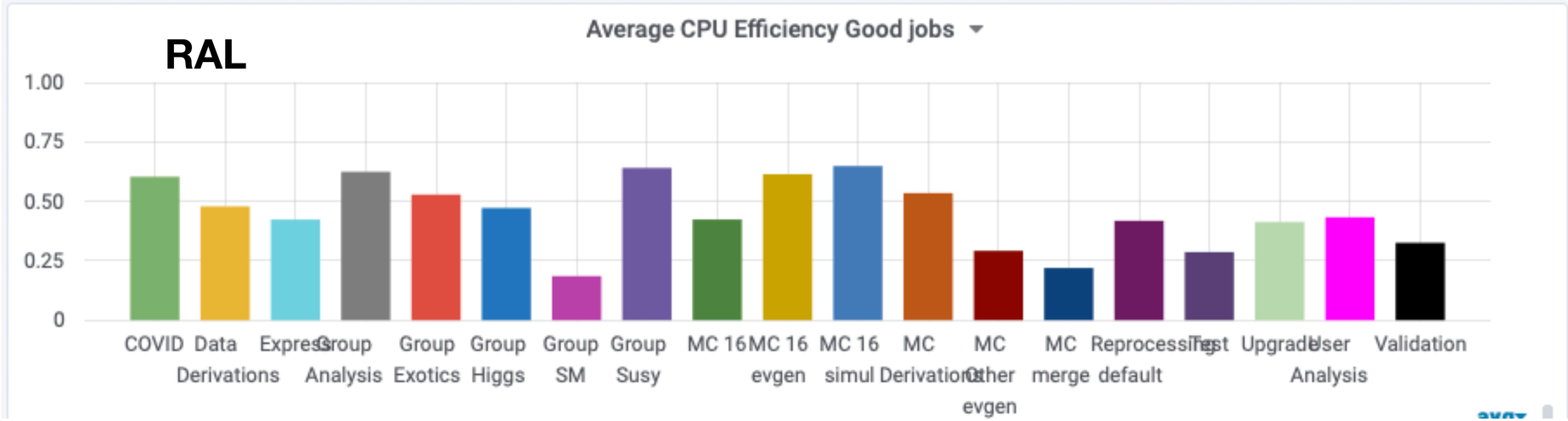
CPU Efficiency by Job Type

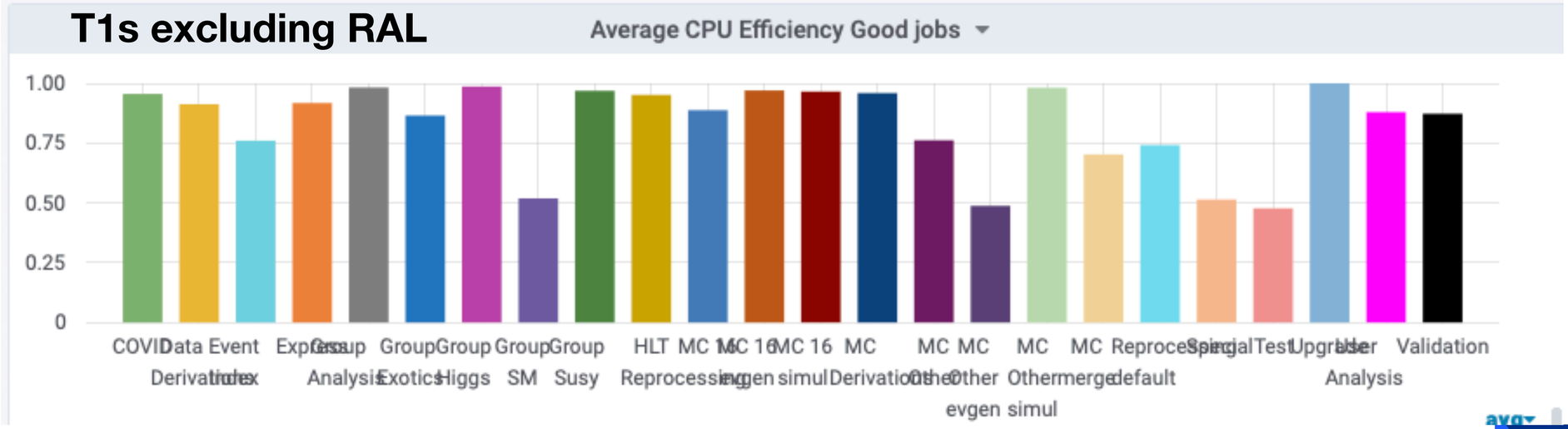
- RAL Efficiency comparison with other T1s

- Interesting points in COVID and MC evgen for example

  Not IO heavy jobs (MC merging is …)

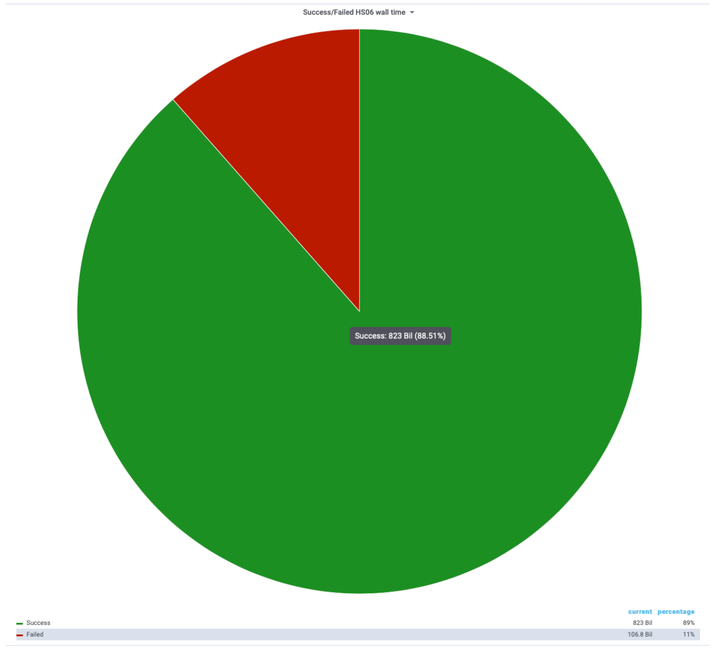
- Suggests that there is some other bottleneck?

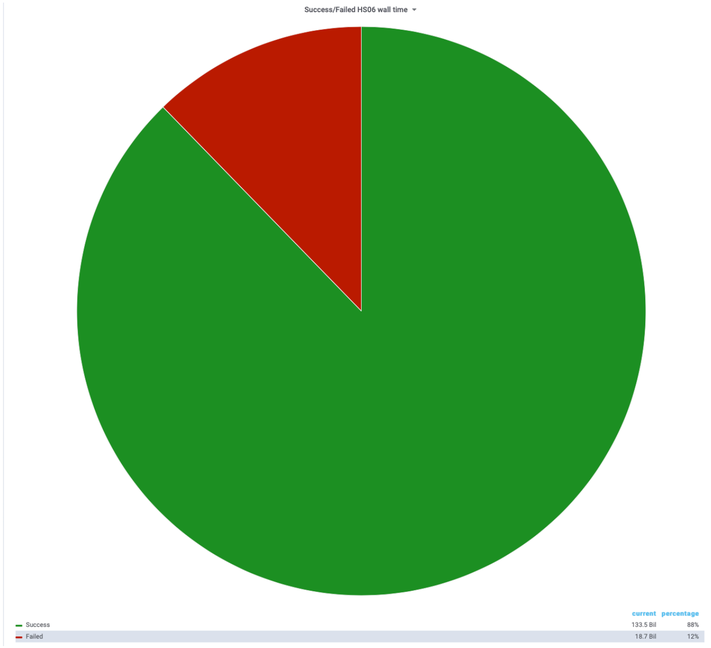




Lost Walltime

- Overall lost wall time [HS06s] in failed jobs between T1s and RAL not so different:
  - RAL ~ 12%
  - T1s (excluding RAL) ~11%





Overview

- Startimes from 20th April - 28th: 17k entries

Time	First	Last
creationtime	2020-04-13 17:31:59	2020-04-28 16:02:36
starttime	2020-04-20 00:01:13	2020-04-28 16:14:44
modificationtime	2020-04-20 00:18:32	2020-04-28 16:21:43
endtime	2020-04-20 00:13:35	2020-04-28 16:17:01

	Fraction
Succeeded	0.87
Failed	0.13

hasSSD	succeeded
False	0.85
True	0.90

-

No SSD	109970
hasSSD	60478

- Number of jobs, and success fraction by gShare (Job type)

- General improvement in Succesful jobs in SSDs across all types (not separated by failure mode),

gshare	(Count, noSSD)	(Count, SSD)	(Success Fraction, noSSD)	(Success Fraction, SSD)
COVID	51	23	0.96	0.96
Data Derivations	2428	2442	0.90	0.98
Express	1905	856	0.76	0.71
Group Analysis	1882	1119	0.89	0.97
Group Exotics	2633	1054	0.91	0.96
Group Higgs	1258	456	0.83	0.90
Group SM	61	29	0.89	1.00
Group Susy	1643	805	0.92	0.93
MC 16	13523	14517	0.72	0.77
MC 16 evgen	49	3	0.39	0.67
MC 16 simul	629	765	0.98	0.97
MC Derivations	5730	5098	0.94	0.98
MC Other evgen	46	2	0.20	1.00
MC merge	4284	1883	0.89	0.99
Reprocessing default	240	34	0.83	0.97
Test	2016	215	0.99	1.00
Upgrade	17	8	0.94	1.00
User Analysis	71562	31154	0.86	0.93
Validation	13	15	0.92	1.00

Error Types

- Counts (and by fraction) of failures by error code (piloterrorcode) :
  - “0” is mainly (not not all) succesful

	Count All	Count no SSD	Count SSD	Fraction of (All)[%]	Fraction of(no SSD)[%]	Fraction of (SSD)[%]
0	159746	100424	59322.00	93.72	91.32	98.09
1151	5130	5062	68.00	3.01	4.60	0.11
1152	2844	2724	120.00	1.67	2.48	0.20
1165	797	437	360.00	0.47	0.40	0.60
1324	406	399	7.00	0.24	0.36	0.01
1168	375	256	119.00	0.22	0.23	0.20
1203	338	120	218.00	0.20	0.11	0.36
1150	326	256	70.00	0.19	0.23	0.12
1171	108	67	41.00	0.06	0.06	0.07
1099	91	60	31.00	0.05	0.05	0.05
1326	69	32	37.00	0.04	0.03	0.06
1	66	25	41.00	0.04	0.02	0.07
1137	50	29	21.00	0.03	0.03	0.03
1353	25	16	9.00	0.01	0.01	0.01
1344	23	23	NaN	0.01	0.02	NaN
1355	18	15	3.00	0.01	0.01	0.00
1212	11	9	2.00	0.01	0.01	0.00
1352	11	7	4.00	0.01	0.01	0.01
1187	10	6	4.00	0.01	0.01	0.01
1181	2	1	1.00	0.00	0.00	0.00
1322	1	1	NaN	0.00	0.00	NaN
1236	1	1	NaN	0.00	0.00	NaN

Code	Error
1151	File transfer timed out during stage-out:
1152	transfer timed out during stage-in
1165	Local output file is missing
1324	Service not available at the moment
1168	Total file size too large

CPU efficiency

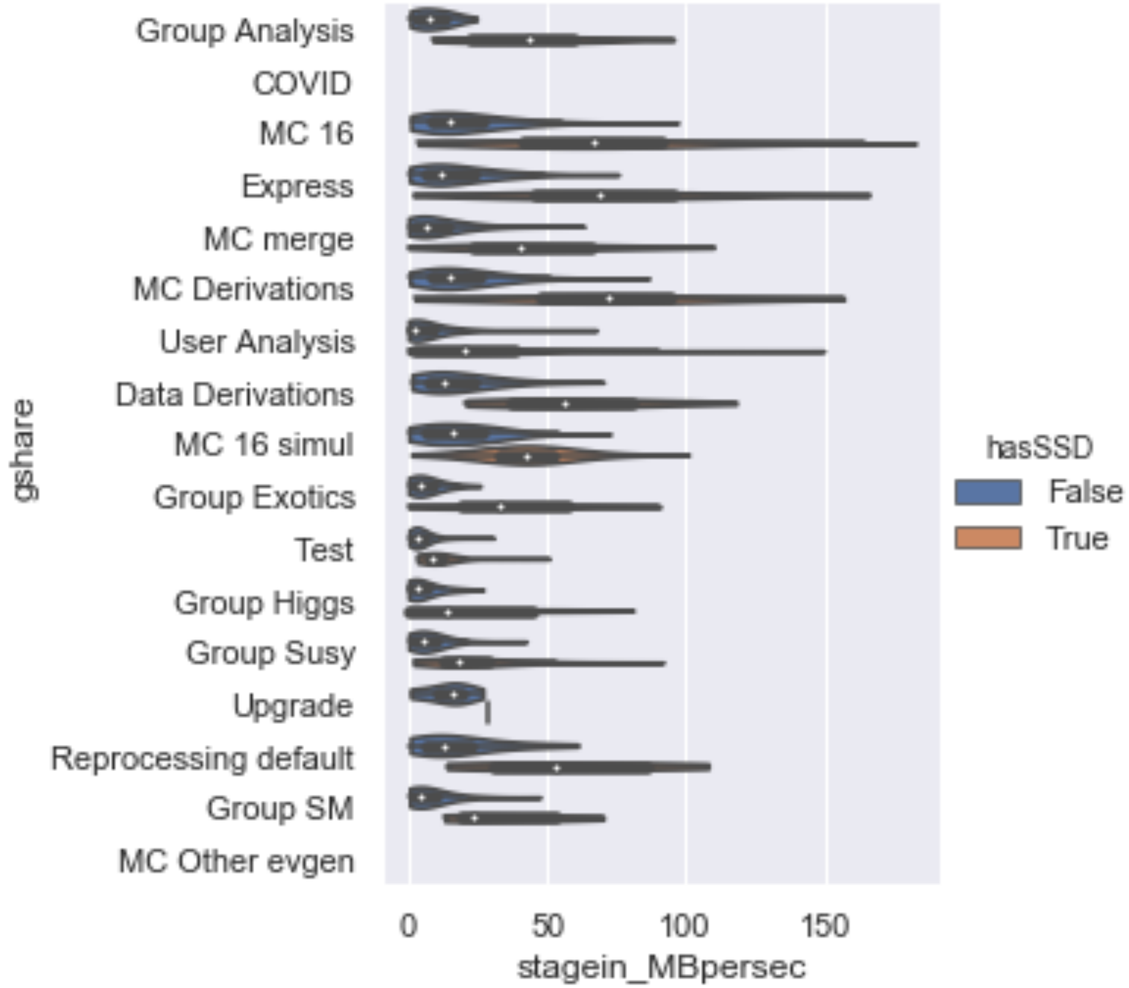
- CPU efficiency for successful jobs:
  - Broadly mirrors the Monit pages
  - Stats small in some entries (uncertainties not included)
  - Shows interesting features; e.g. for COVID SSD, no improvement (but small stats)

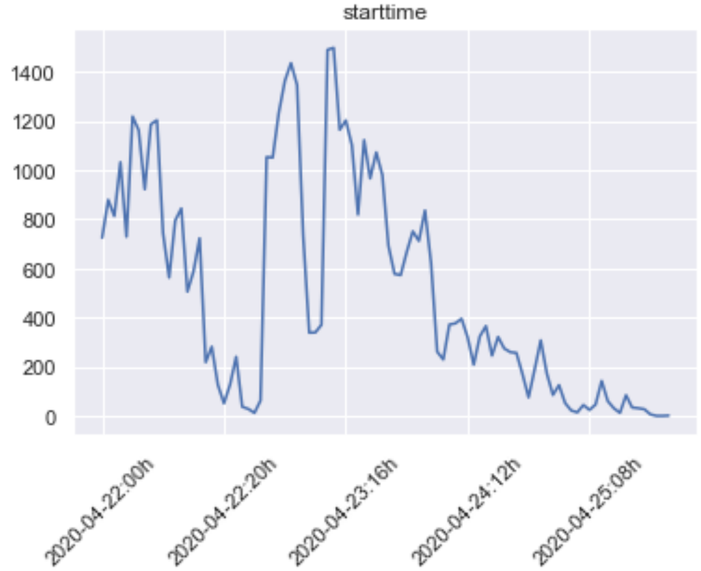
gshare	COVID	Data Derivations	Express	Group Analysis	Group Exotics	Group Higgs	Group SM	Group Susy	MC 16	MC 16 evgen	MC 16 simul	MC Derivations	MC Other evg
Total	0.57	0.44	0.41	0.68	0.55	0.53	0.14	0.68	0.44	0.63	0.63	0.53	0.
SSD	0.55	0.56	0.59	0.60	0.56	0.59	0.40	0.58	0.56	0.80	0.62	0.59	0.
No SSD	0.58	0.37	0.37	0.72	0.55	0.51	0.10	0.73	0.37	0.62	0.63	0.48	0.

- Mostly, SSD shows improvement,
  - But doesn’t get to the near 100% CPU efficiencies seen in other T1s

ATLAS / Vande plots

- Smaller subset of data
  - Correlate ATLAS (panda) information with Vande (influxDB).
  -



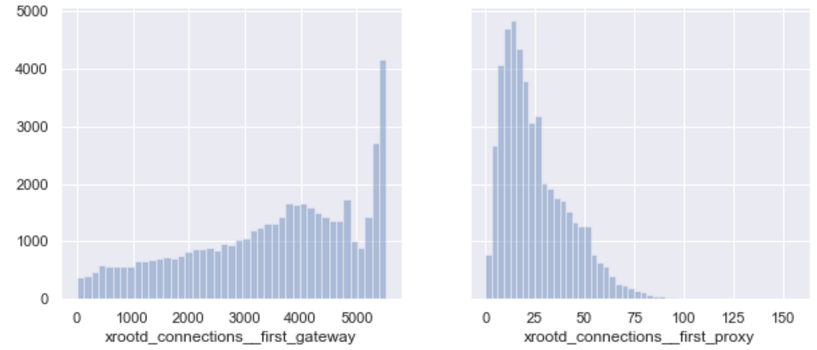


Worker node proxy connections

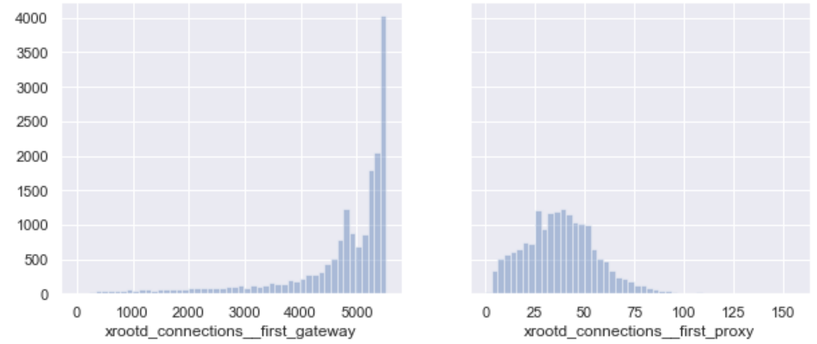
- Not sure of exact meaning:
  - Measurement xrootd_connections gateway and proxy (initial value taken)

- are these cumulative or instantaneous values?

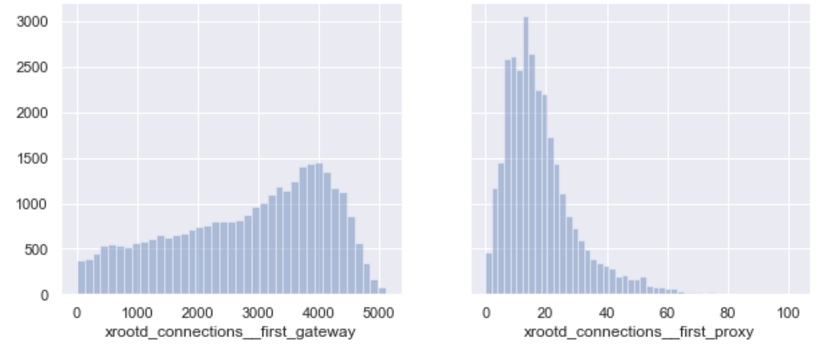
All



SSD



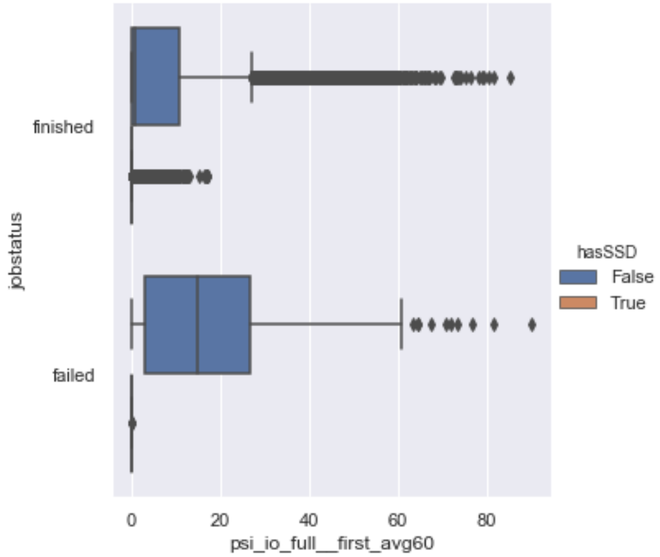
! SSD

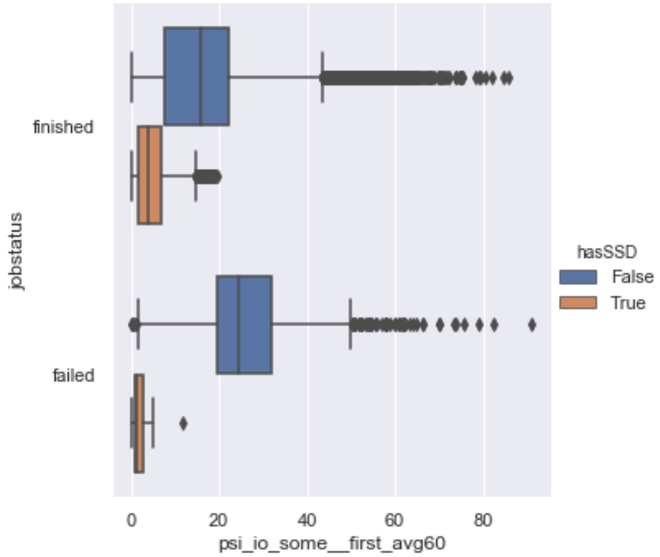


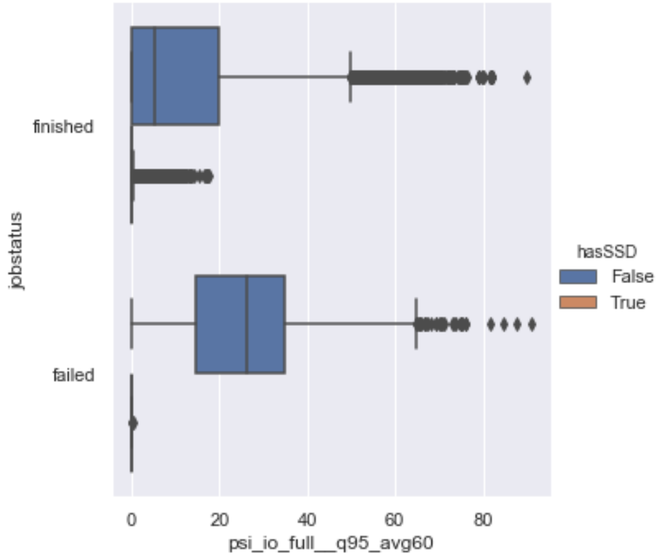
Pressure Stall Information: IO

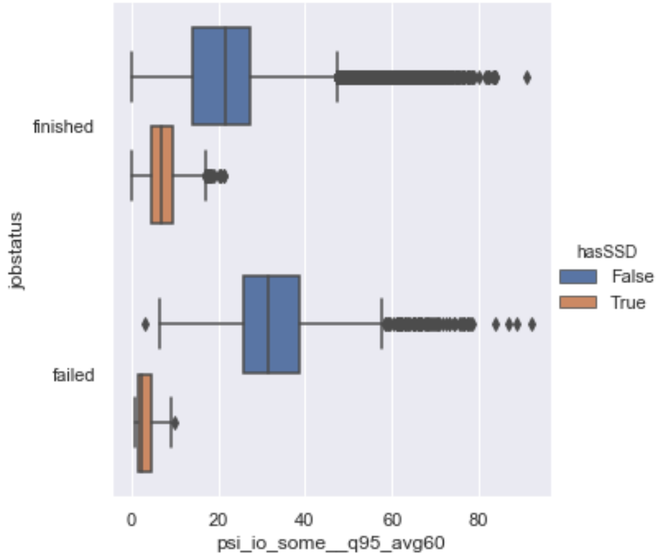
- Clear improvement in PSI

Definition: [measurement]__[aggregation]_[field]









- => mitigation rather than solution ?

General Summary

- File transfer problems are highest failure point for RAL (and Tier-1s in general)
  - Stage-out (and perhaps stage-in timeouts) source unknown (rucio / gfal problems) ? Need more logging information:
    - Some suggestions provided:

Also, do you have any logs from the RUCIO mover? gfal2 python bindings use the standard Python logging facility (afair), and xrootd logs can be enabled using XRD_LOGLEVEL envar (by default the will go to stderr, but can be redirected to file using XRD_LOGFILE). This could be extremely helpful!

Is RUCIO sharing gfal2 context between connections?
There are few parameters on XRootD client side that could have an impact on RUCIO mover performance:

- could you try increasing the thread pool size (by default it is using only 3 threads), e.g. XRD_WORKERTHREADS=16 (or more)
- could you please try setting XRD_STREAMERRORWINDOW=0
- in RUCIO framework could you try forcing xrootd framework to use multiple connections by adding a virtual user to the URL (otherwise all requests are multiplexed over a single physical connection, which most of the time is what you want, but in some corner case scenarios could cause a glitch)

- However overall CPU efficiency seems to be around ~50%, even with SSDs.
  - Before ATLAS extended it's wall time limit, transfer speed meant many jobs killed by LRMS
  - SSDs (in current configuration) show significant overall improvements;
    - User analysis jobs need some further investigation

- Collect more stats …
- Look again at PSI correlations, etc.

- What other metrics to look at:
  - CVMFS 'efficiency' ?
  -