HammerCloud Blacklisting Study

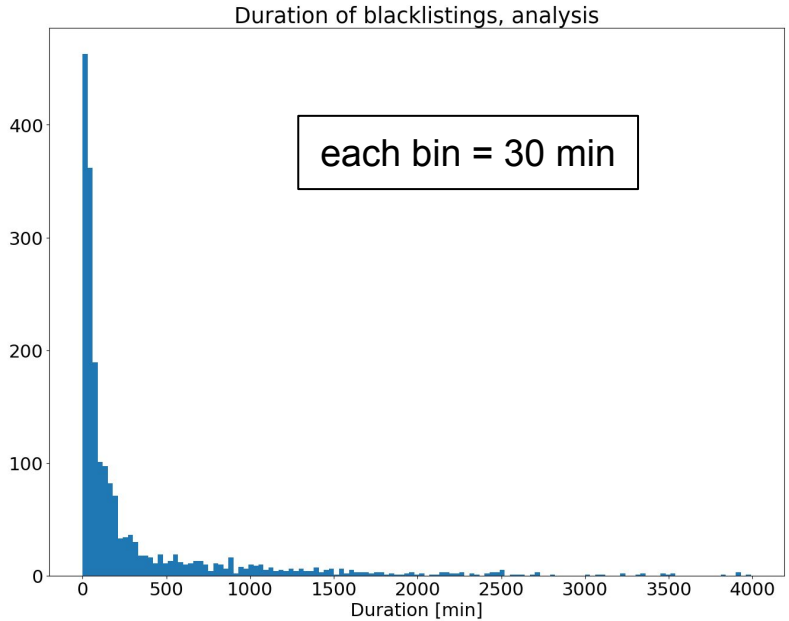
Thomas Maier

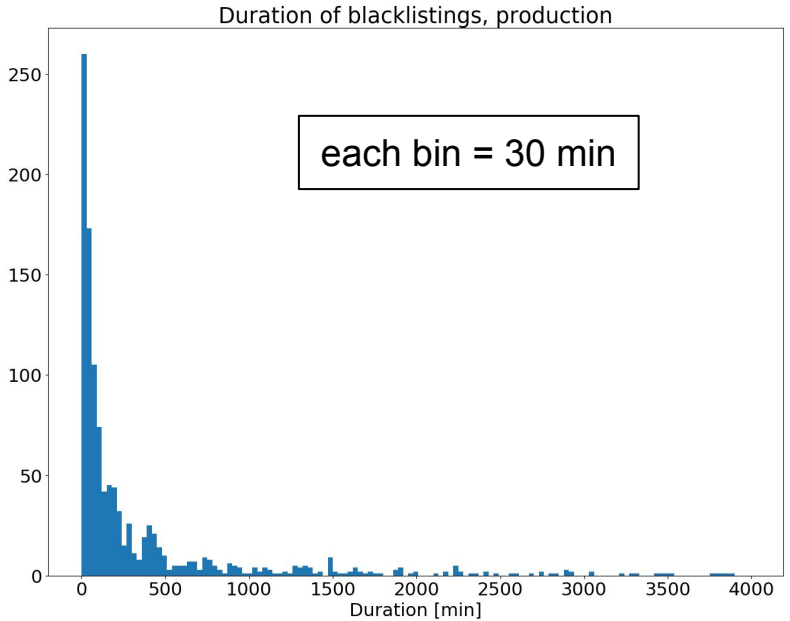
Ludwig-Maximilians-Universität München

Introduction

- Blacklisting log data period: 13.03.2018 - 10.10.2018
  - Information on blacklisting and whitelisting actions
    - Count “test periods” and their duration by scanning through the log for a given queue and selecting blacklisting actions with a following whitelisting action
    - Evaluate “blacklisting chains” by counting consecutive test periods with a blacklisting action following the previous whitelisting action within a defined time gap
  - Ignoring “slave” blacklistings, i.e. when PQs are set to TEST due to the decision on the master queue
- jobs_archive data
  - Information on the example job panda IDs given in the blacklisting reason string
- Previous study in 2017: link
- Git repo: link

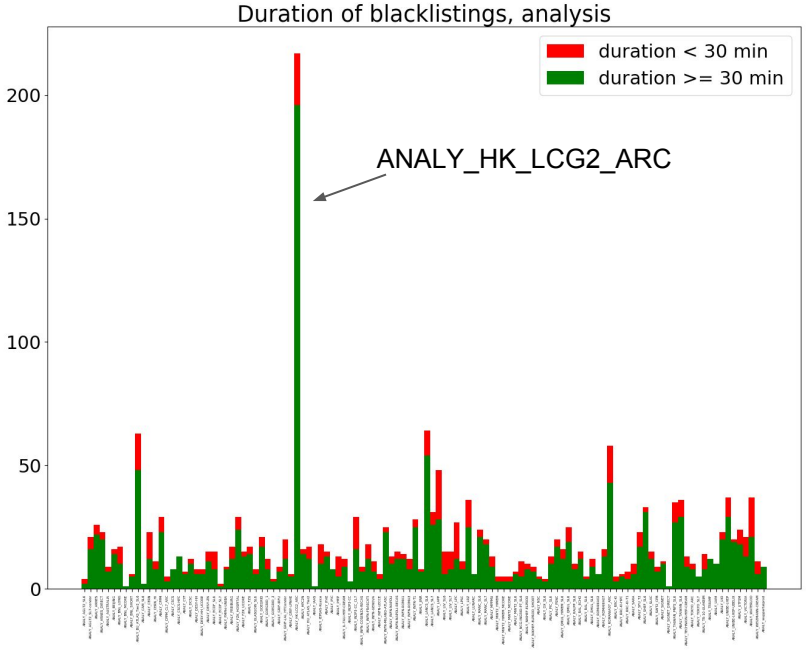
Duration of blacklistings (“test periods”)



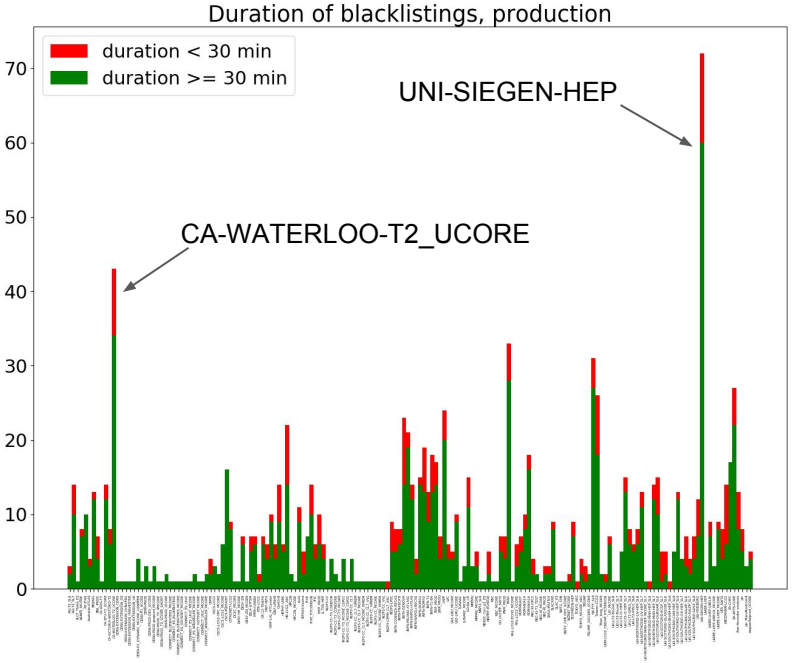


- Most test periods only last for up to a couple of hours
- Long periods possibly due to untracked/manual whitelistings
  - However, there are some PQs that are just constantly blacklisted, up to 90d!

Duration of blacklistings per PQ

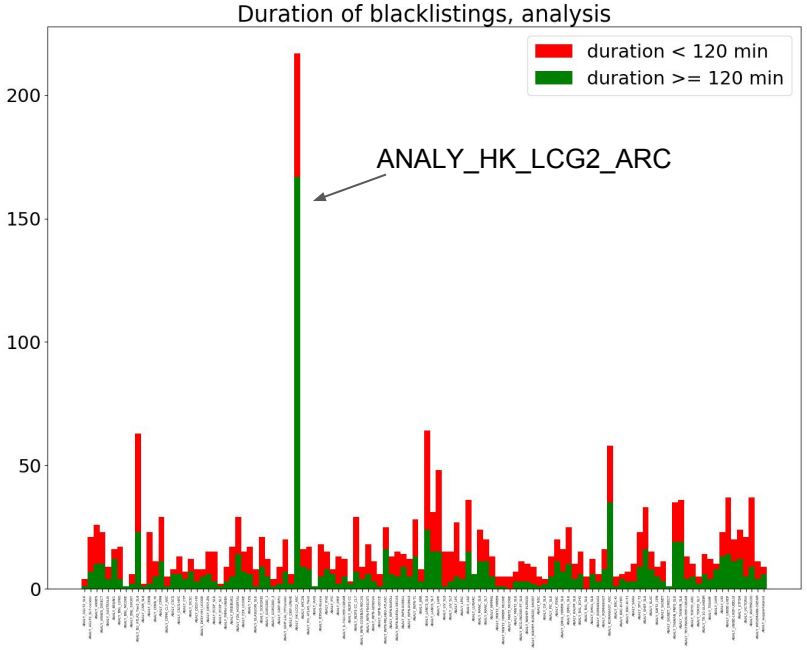


PDF

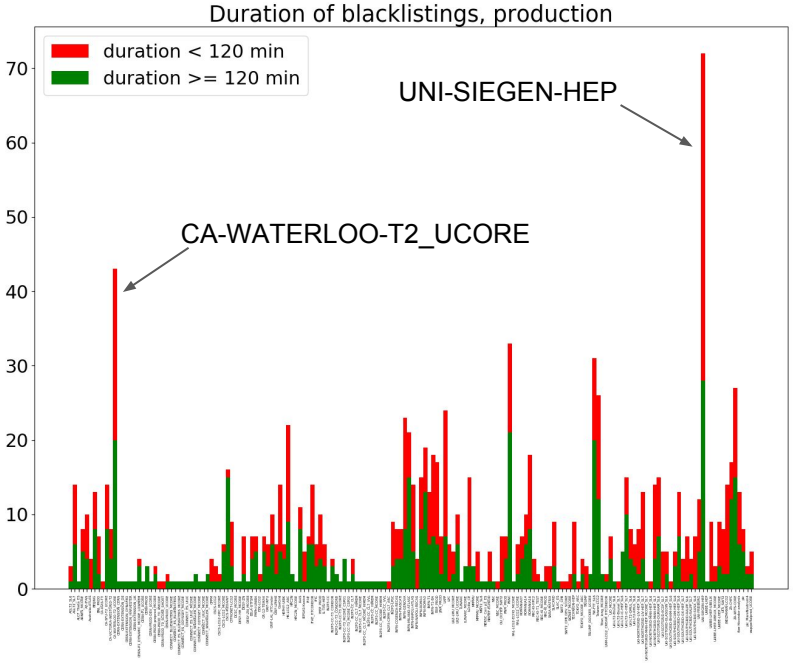


PDF

Duration of blacklistings per PQ

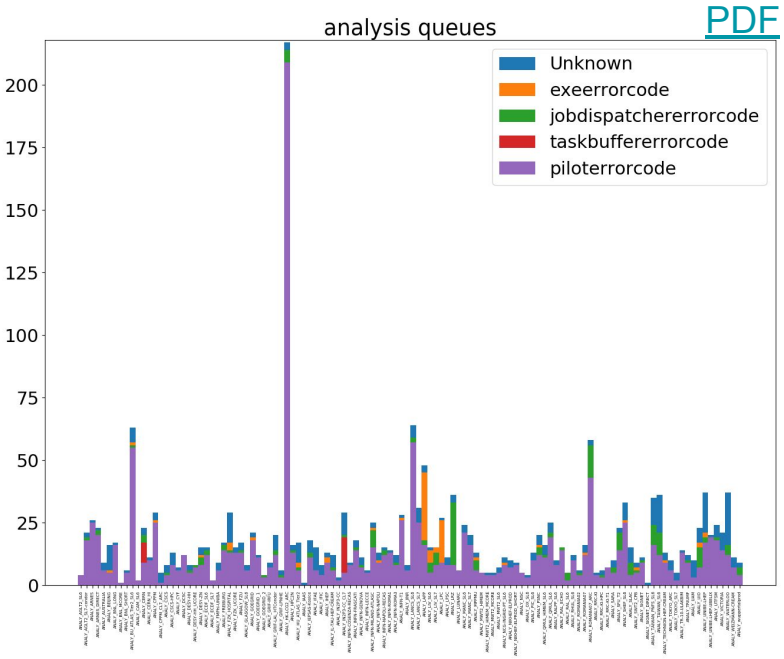


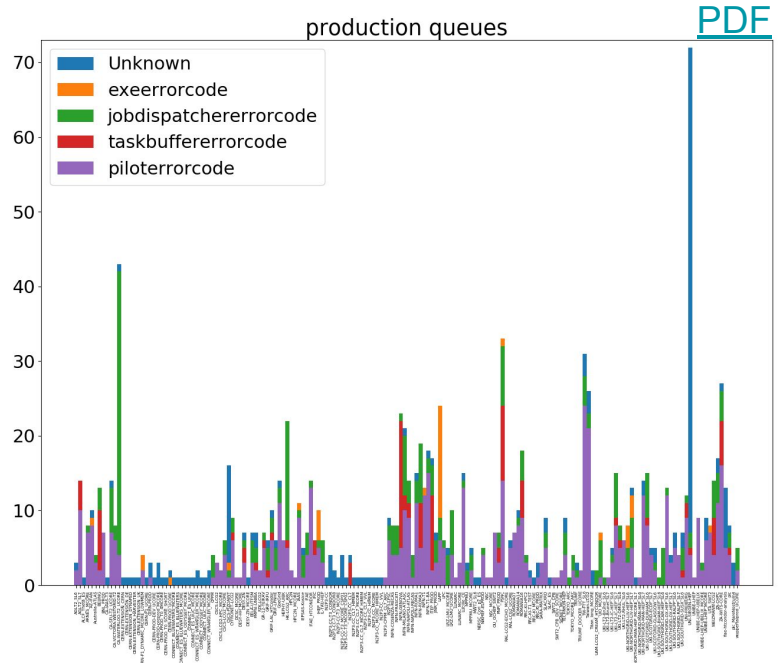
PDF



PDF

PanDA error codes per PQ, taken from example jobs

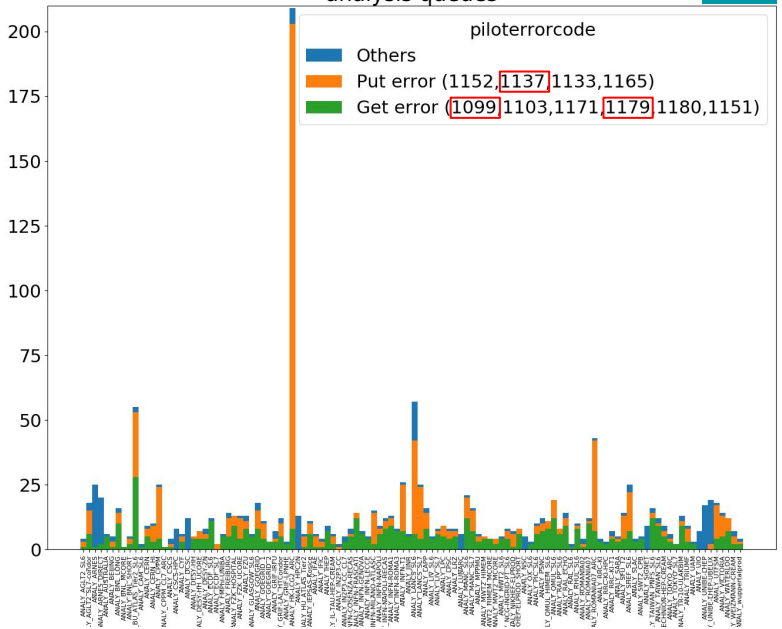




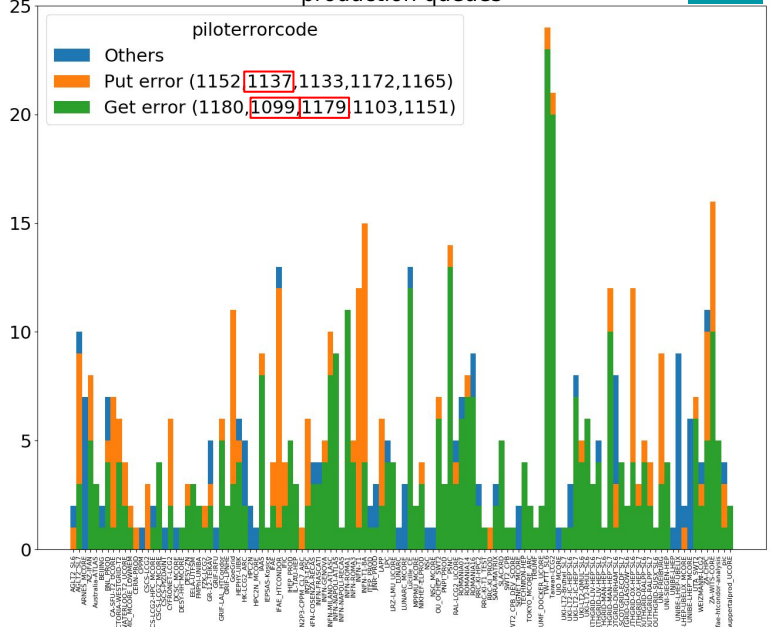
- Priority for marking a given test period with a particular error code
  - piloterror > taskbuffer > jobdispatcher > exeerror

Pilot error codes per PQ, taken from example jobs

PDF



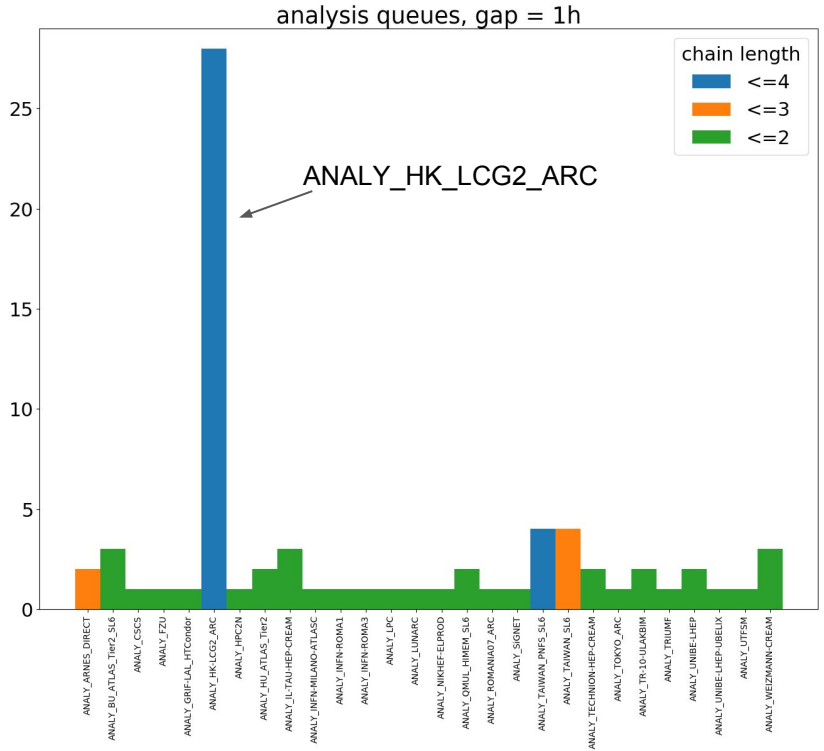
PDF



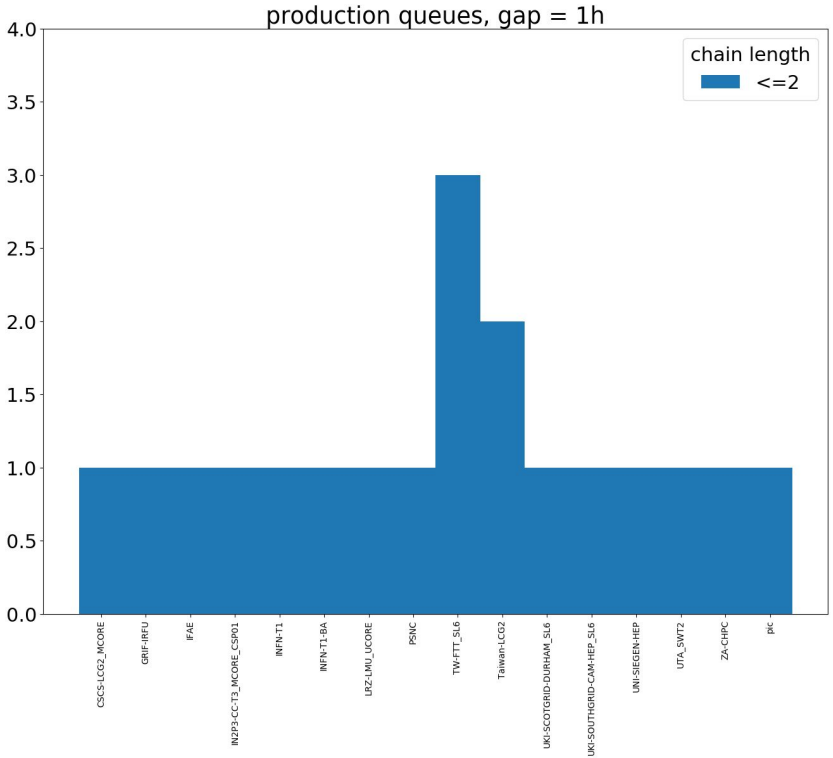
- For the majority of blacklisting incidents the problem was either getting the input to the worker node or missing output / uploading the output

Detailed overview of pilot error codes

Blacklisting chains (gap = 1h) per PQ

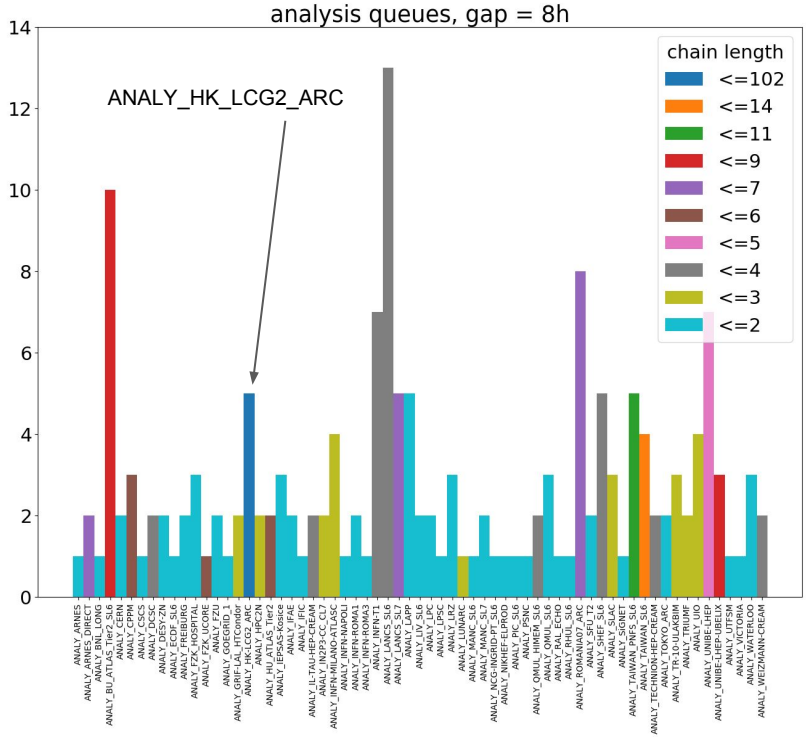


PDF

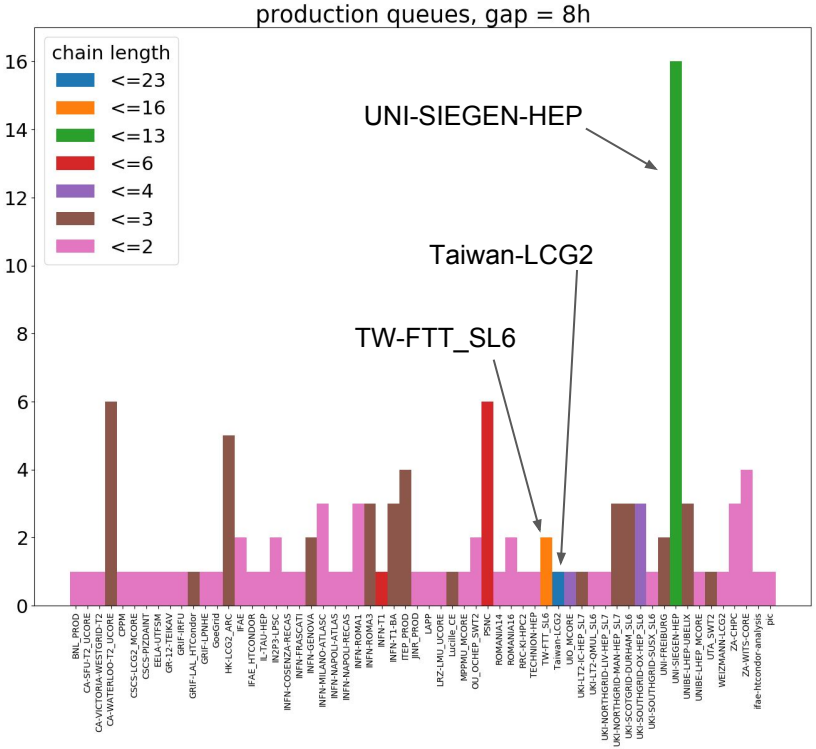


PDF

Blacklisting chains (gap = 8h) per PQ



PDF



PDF

Summary

- Most blacklisting test periods only last a couple of hours
- For the majority of blacklistings, failing retrieval/upload of the input/output is the error reason
- From these studies and past blacklisting reports in general, I don’t see a good way to prevent “unnecessary” blacklisting in a dynamic way → the information HammerCloud receives from PanDA is not really sufficient to evaluate if an issue is transient or not
- However, acting on chained blacklisting behaviour could be something reasonable to include in HammerCloud
    - Keep a PQ blacklisted if a blacklisting chain of length X occured (with gap = Y hours) and notify the responsible people that the PQ will stay in TEST until it’s investigated
    - AFAIK we already do something similar with the EventService blacklisting

The line must be drawn here! This far, no further!

Jean-Luc Picard

piloterrorcodes

	total/analysis/production	
1137:	727/591/136	Put error: Error in copying the file from job workdir to localSE
1179:	421/285/136	Get error: Failed to get LFC replicas
1099:	335/173/162	Get error: Staging input file failed
1008:	121/80/41	General pilot error, consult batch log
1246:	94/93/1	"User tarball (source unknown) cannot be downloaded from PanDA server"
1103:	76/35/41	Get error: No such file or directory
1244:	54/54/0	"No release candidates found"
1180:	30/15/15	Get error: Globus system error
1098:	20/13/7	No space left on local disk
1144:	19/0/19	This job was killed by panda server
1151:	16/7/9	Get error: Input file staging timed out
1152:	15/14/1	Put error: File copy timed out
1194:	14/13/1	"STAGEOUT FAILED: File verification failed ..."
1221:	12/11/1	File already exist
1110:	6/2/4	Failed during setup
1165:	5/4/1	Put error: Local output file missing
1171:	5/5/0	Get error: adler32 mismatch on input file
1133:	5/3/2	Put error: Fetching default storage URL failed
1201:	3/1/2	Job killed by signal: SIGTERM
1223:	1/0/1	TRF failed due to bad_alloc
1112:	1/1/0	Exception caught by pilot
1172:	1/0/1	Put error: adler32 mismatch on output file
1111:	1/0/1	Exception caught by runJob
1176:	1/1/0	Pilot has no child processes (job wrapper has either crashed or did not send final status
1212:	1/1/0	Payload ran out of memory

jobdispatchererrorcodes

	total/analysis/production	
100:	300/157/143	lost heartbeat
102:	172/31/141	"Sent job didn't receive reply from pilot within 30 min"
101:	4/3/1	job recovery failed for three days

taskbuffererrorcodes

300: "The worker was cancelled while the job was starting :..."