HammerCloud ATLAS - Job shaping

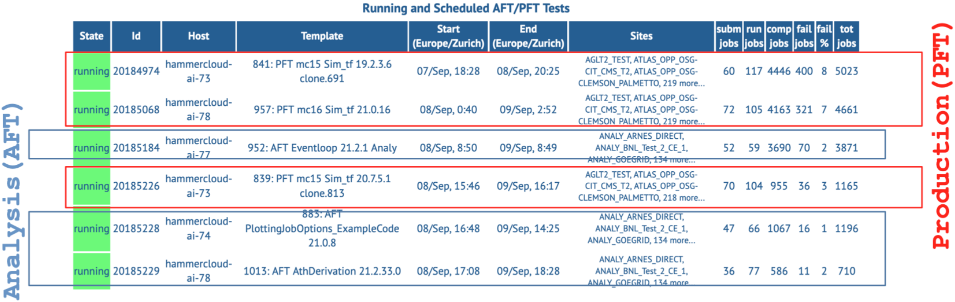
David Hohn presenting work by Michael Böhler

Jan. 25th 2021

Albert-Ludwigs-Universität Freiburg

Reminder: HammerCloud and auto-exclusion

- framework/service to test, commission, benchmark ATLAS Distributed Computing (ADC) resources with realistic workloads
- automated site exclusion and recovery based on production- and analysis-type tests (3 templates each)
    - recent templates are documented here
- interfaces mainly with the ATLAS workflow management system (WFMS) „PanDA“
- the current tests can be monitored here



Running and Scheduled AFT/PFT Tests

State	Id	Host	Template	Start (Europe/Zurich)	End (Europe/Zurich)	Sites	subm jobs	run jobs	comp jobs	fail jobs	fail %	tot jobs
running	20184974	hammercloud-ai-73	841: PFT mc15 Sim_tf 19.2.3.6 clone.691	07/Sep, 18:28	08/Sep, 20:25	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 219 more...	60	117	4446	400	8	5023
running	20185068	hammercloud-ai-78	957: PFT mc16 Sim_tf 21.0.16	08/Sep, 0:40	09/Sep, 2:52	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 219 more...	72	105	4163	321	7	4661
running	20185184	hammercloud-ai-77	952: AFT Eventloop 21.2.1 Analy	08/Sep, 8:50	09/Sep, 8:49	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 134 more...	52	59	3690	70	2	3871
running	20185226	hammercloud-ai-73	839: PFT mc15 Sim_tf 20.7.5.1 clone.813	08/Sep, 15:46	09/Sep, 16:17	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 218 more...	70	104	955	36	3	1165
running	20185228	hammercloud-ai-74	883: AFT PlottingJobOptions_ExampleCode 21.0.8	08/Sep, 16:48	09/Sep, 14:25	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 134 more...	47	66	1067	16	1	1196
running	20185229	hammercloud-ai-78	1013: AFT AthDerivation 21.2.33.0	08/Sep, 17:08	09/Sep, 18:28	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 134 more...	36	77	586	11	2	710

http://hammercloud.cern.ch/hc/app/atlas/

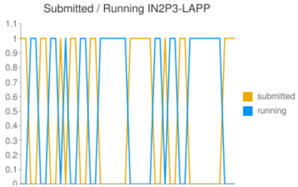
Description – Job Shaping Qualification Task

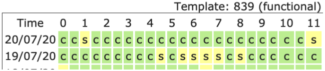
- dynamically send **more** and specialised **jobs** to failing site
  - collect more information about the problem/root cause
  - **speed up exclusion/recovery decisions**
- develop new templates focused on testing specific components of PQ functionality
- develop new views to interpret the data
- expose error source (currently HC team and sites are informed by email)
- identify main sources of problems over time, failure patterns
- provide interpreted data to speed up problem solving

ATLASSCQT-28

Job Shaping – 1. Step

- Analyse last 3h of HC test jobs
  - *golden templates* used for site exclusion
  - currently 1 job per template per site
- If there is no progress
  - Job stuck in states
    - submitted
    - running
    - No new job submitted
  - > 2h increase number of parallel running jobs
- When jobs succeed → decrease number of parallel running jobs



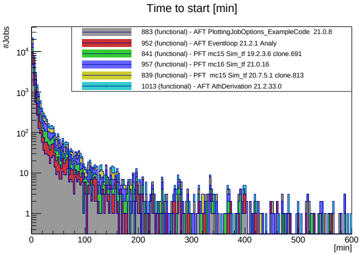


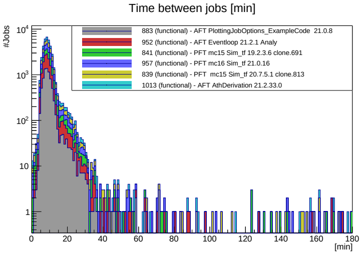
Template: 839 (functional)

Time	0		1		2		3		4		5		6		7		8		9		10		11
20/07/20	c	c	s	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c	s
19/07/20	c	c	c	c	c	c	c	c	c	s	c	s	s	s	s	c	s	c	c	c	c	c	c

- Currently parameters:
  - Run script every 30 mins
  - max_jobs (parallel) 1 → 5
  - Time threshold = 2h

Some distributions





- time [min] between job was submitted by HC and job started
- sum over all jobs in all templates on all sites

- time [min] between job finished and next job submitted
- sum over all jobs in all templates on all sites

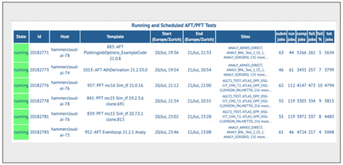
Technical Details

Job Shaping

Each func template (tpl)
- ATF (e.g 883,1013,952)
- PFT (e.g. 957,841,839)

Get sites per tpl

> consider sites with
HC_param = „AutoExclusion"



Running and Scheduled AFT/PFT Tests

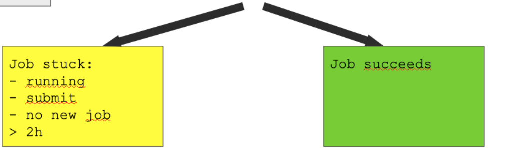
State	Id	Host	Template	Start (Europe/Zurich)	End (Europe/Zurich)	Sites	subm jobs	run jobs	comp jobs	fail jobs	fail %	tot jobs
running	20182771	hammercloud-ai-78	883: AFT PlottingJobOptions_ExampleCode 21.0.8	20/Jul, 19:36	21/Jul, 21:33	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 131 more...	63	44	5266	261	5	5634
running	20182773	hammercloud-ai-74	1013: AFT AthDerivation 21.2.33.0	20/Jul, 19:54	21/Jul, 20:54	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 131 more...	46	61	3435	257	7	3799
running	20182776	hammercloud-ai-76	957: PFT mc16 Sim_tf 21.0.16	20/Jul, 21:12	21/Jul, 22:00	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 216 more...	62	112	4147	473	10	4794
running	20182778	hammercloud-ai-78	841: PFT mc15 Sim_tf 19.2.3.6 clone.691	20/Jul, 21:54	21/Jul, 20:55	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 216 more...	55	119	3305	334	9	3813
running	20182781	hammercloud-ai-76	839: PFT mc15 Sim_tf 20.7.5.1 clone.813	20/Jul, 23:02	21/Jul, 23:28	AGLT2_TEST, ATLAS_OPP_OSG-CIT_CMS_T2, ATLAS_OPP_OSG-CLEMSON_PALMETTO, 216 more...	55	119	3972	337	8	4483
running	20182783	hammercloud-ai-73	952: AFT Eventloop 21.2.1 Analy	20/Jul, 23:46	21/Jul, 23:08	ANALY_ARNES_DIRECT, ANALY_BNL_Test_2_CE_1, ANALY_GOEGRID, 131 more...	61	46	4724	217	4	5048

Get active test

> retrieve max_jobs per site

Get Job Results

> check latest job (within last 3h)



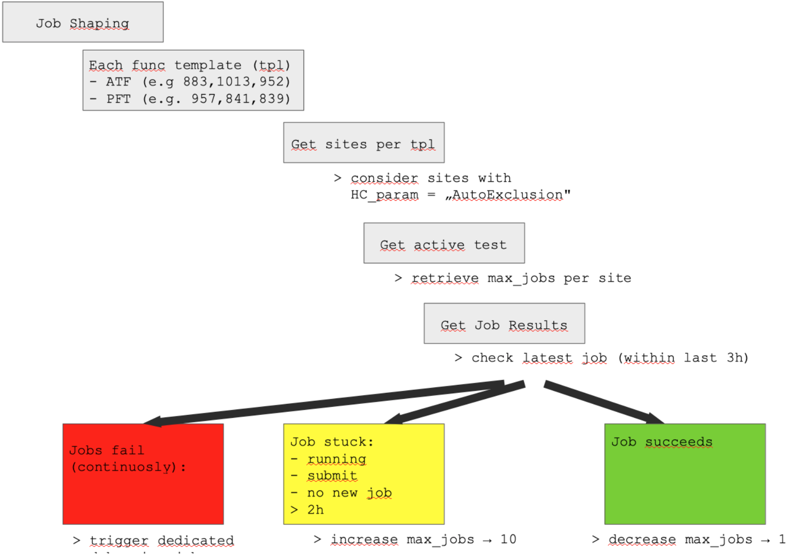
Job stuck:
- running
- submit
- no new job
> 2h

Job succeeds

> increase max_jobs → 10

> decrease max_jobs → 1

Technical Details - extendable



Job Shaping

Each func template (tpl)
- ATF (e.g 883,1013,952)
- PFT (e.g. 957,841,839)

Get sites per tpl

> consider sites with HC_param = „AutoExclusion"

Get active test

> retrieve max_jobs per site

Get Job Results

> check latest job (within last 3h)

Jobs fail (continuosly):

> trigger dedicated

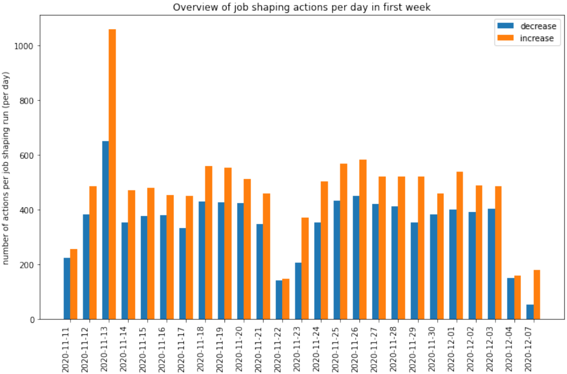
Job stuck:
- running
- submit
- no new job
> 2h

> increase max_jobs → 10

Job succeeds

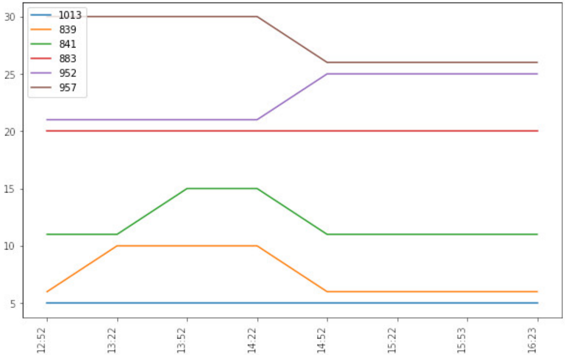
> decrease max_jobs → 1

First plots from test phase



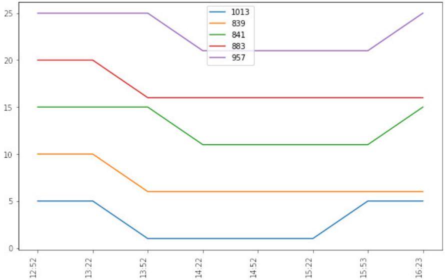
- days with less entries == job shaping was not running the full day
- more increase than decrease: most likely subset of sites only increased, but no successfull job triggered decrease again

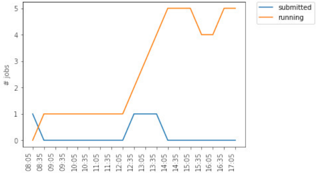
Prototype of job shaping plots



- y-axis show settings for max parallel jobs (values are 1 or 5) [+ 5,10,.. for visualization]
- e.g. tpl 839 was increased to 5 at 13:22 and reduced to 1 at 14:52
- jobs of tpl 883 and 1013 work running smoothly

Idea of a Prototype of job shaping plots





- combine info of settings from jobShaping with evolution of running jobs
- this example is for tpl 957(upper line left plot)
- keep in mind, that there are always delays from set parallel from 5 → 1, until additional jobs finished - here they did not finish and again job shaping increased back to 5

Conclusion

- Current state:
  - Modules for triggering automatic increase of max running jobs in test phase
- Next steps:
  - put job shaping in production
    - add dedicated model class to HC to write job shaping events into db
    - run job shaping completely automatic
  - add dedicated view to hc web page for monitoring job shaping
  - extend functionality to submit debug jobs to generate more/dedicated info of failing sites
- draft of description of this project ready in the overleaf document

Back-up

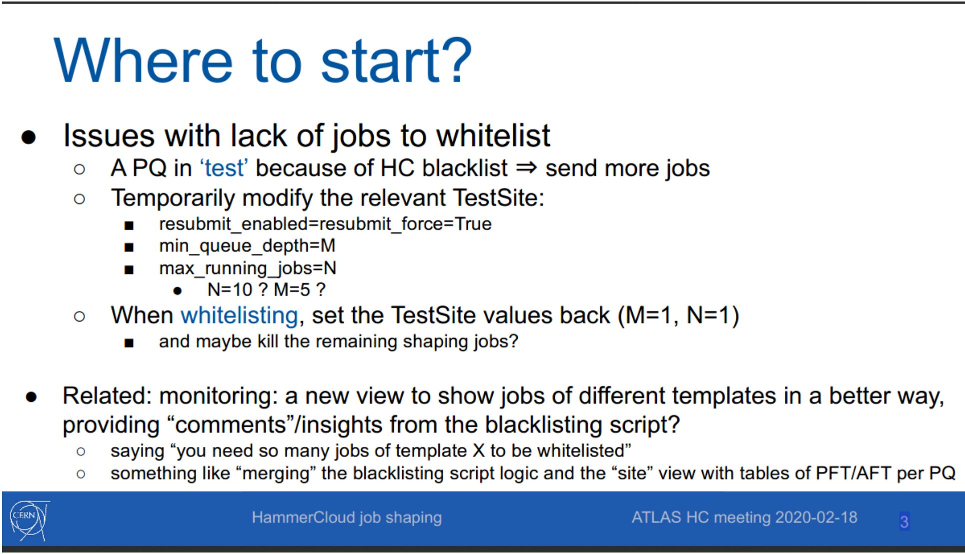
Reminder: site exclusion policy

- Status is computed every 30 min
- Panda queues in state: online or test are tested
- Time window of resent results: past 3 hours (AFT) / past 4 hours (PFT)
- Panda Queue is excluded if jobs fail requirements
  - last 3 jobs from one template
  - or last 2 jobs from one template and last job from another tpl
  - or the last job from each of the 3 templates have failed.
- Exclusion: PanDA queue status: online → test
- Re-included: last **2 jobs from each test** are **successful**

- Unified queues the following policy is applied:
  - Above policies fail for the PFTs <u>or</u> AFTs → exclude
  - Above policies pass for both PFTs <u>and</u> AFTs → re-include

https://twiki.cern.ch/twiki/bin/viewauth/IT/HammerCloudTutorialATLASsiteAdmins

Proposal



Where to start?

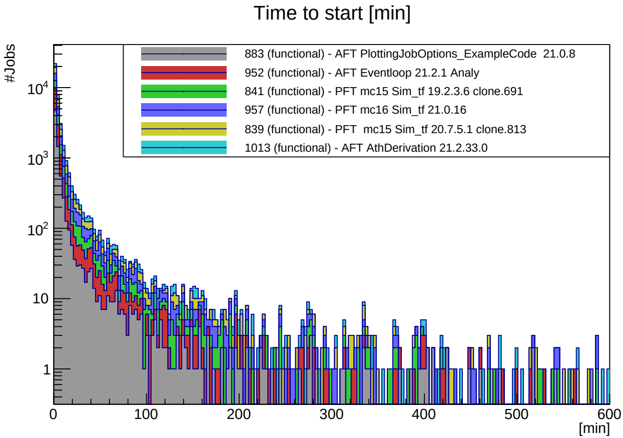
- Issues with lack of jobs to whitelist
  - A PQ in 'test' because of HC blacklist ⇒ send more jobs
  - Temporarily modify the relevant TestSite:
    - resubmit_enabled=resubmit_force=True
    - min_queue_depth=M
    - max_running_jobs=N
      - N=10 ? M=5 ?
  - When whitelisting, set the TestSite values back (M=1, N=1)
    - and maybe kill the remaining shaping jobs?

- Related: monitoring: a new view to show jobs of different templates in a better way, providing "comments"/insights from the blacklisting script?
  - saying "you need so many jobs of template X to be whitelisted"
  - something like "merging" the blacklisting script logic and the "site" view with tables of PFT/AFT per PQ

HammerCloud job shaping   ATLAS HC meeting 2020-02-18   3

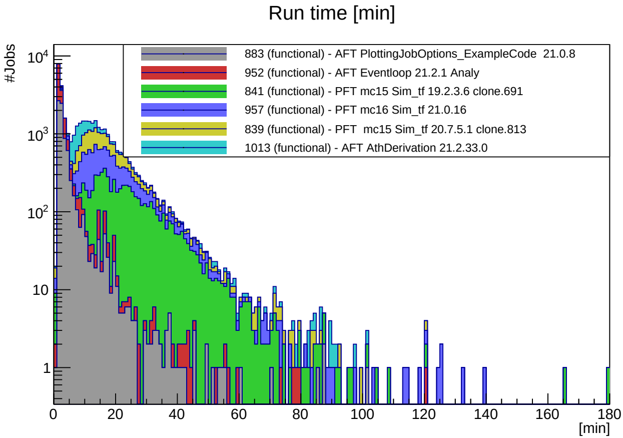
https://indico.cern.ch/event/889846/contributions/3754427/attachments/1989369/3316077/20200218_HCjobshape_jschovan.pdf

Time to start – submit time



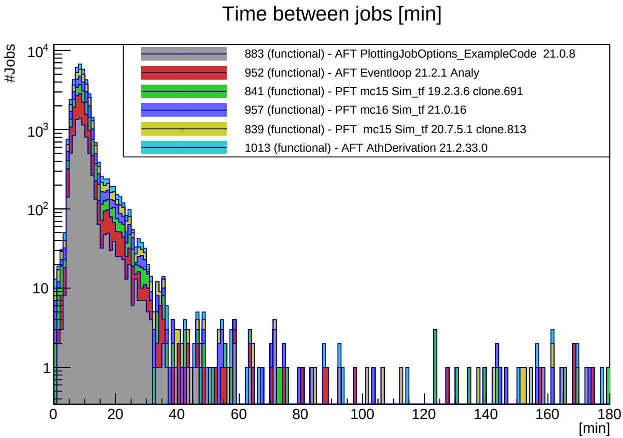
- time [min] between job was submitted by HC and job started
- sum over all jobs in all templates on all sites

Typical runtime of HC jobs – by template



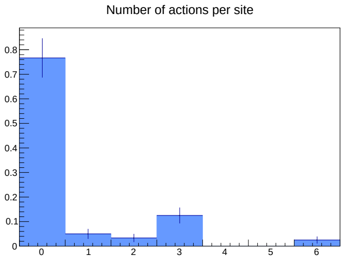
- time [min] of completed jobs (end – start) time
- sum over all jobs in all templates on all sites

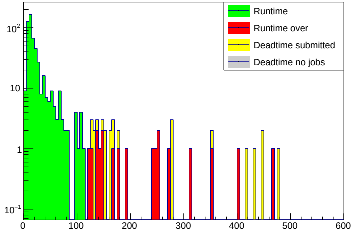
Time between two HC jobs - dead time



- time [min] between job finished and next job submitted
- sum over all jobs in all templates on all sites

Some distributions – job shaping (dry-run)





- Left: number of actions for a given site (either 3 or 6 functional tests are running)
  - bin 0 == all fine (bin 6 = no HC jobs are running)
- Right: time/dead time of analysed jobs