Analysis of empty ATLAS pilot jobs

P A Love[1], M Alef[2], S Dal Pra[3], A Di Girolamo[4], A Forti[5], J Templon[6], E Vamvakopoulos[7]

1. Lancaster University 2. Karlsruhe Institute of Technology 3. INFN-CNAF 4. CERN 5. University of Manchester 6. NIKHEF 7. Centre de Calcul IN2P3

Summary

In this analysis we quantify the wallclock time used by short empty pilot jobs on a number of WLCG compute resources. Pilot factory logs and site batch logs are used to provide independent accounts of the usage. Results show a wide variation of wallclock time used by short jobs depending on the site and queue, and changing with time. The mean fraction of wallclock time used by short jobs over a single month can range from 0.1% to 0.9% depending on the site. The variation in wallclock usage may be explained by different workloads for each resource with a greater fraction when the workload is low. Aside from the wall time used by empty pilots, we also looked at how many pilots were empty as a fraction of all pilots sent. This fraction ranged from 2-40% and the large number was due to pilot factory settings and consequently reduced after tuning these settings.

Introduction

- ATLAS pilot model has been used for many years and provides a late-binding model which allows the ATLAS workflow management system (PanDA) to retain control of job execution priority.
- AutoPyFactory (APF) has a flexible system to control the rate of pilot submission based on Activated workload and other input.
- Tuning of this configuration is required to match the number pilot submissions with the amount of work available in PanDA.
- ATLAS prefers to supply more pilots to the site to maximise the job throughput at the expense of submitting pilots where no job payload is available. These are so-called empty pilots.
- Developments in APF should reduce the number of empty pilots whilst maintaining job throughput.
- In this work we quantify the number of empty pilots and summarise the amount of wallclock time used by empty pilots.

Identifying empty pilots

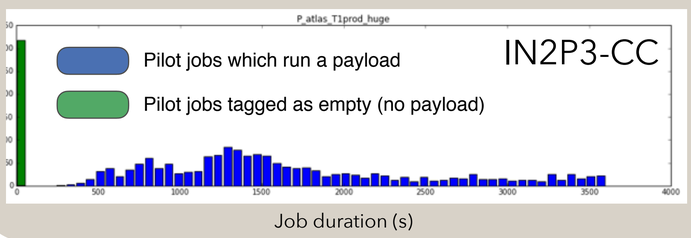
- The ATLAS pilot does not persistently store information about whether it receives a job payload or not so we need to combine logs from various sources.
- The following four methods of identifying (and tagging) empty pilots are used:

1. Join APF job records with PanDA JobsArchive records
2. Join site batch records with PanDA JobsArchive records
3. Filter batch records using a CPU time and Wallclock time thresholds
4. Filter APF job records using a Wallclock time threshold

The first method is capable of tagging jobs for all sites without information from the site itself.

The second method requires collaboration from the site in order to obtain batch records and also collaboration from ATLAS to provide the JobsArchive records

The third method may be used by the site without collaboration from ATLAS.



This plot shows the clean identification of empty pilots (green) by combining site batch records with PanDA JobArchive records (method 2). This result validates the use of CPUtime threshold for other sites (method 3).

Sites' summary

- The follow sites have analysed batch system records to determine the fraction of short pilots and the sum of wallclock time used by these jobs. This sample is for single core jobs only.
- All sites tagged short jobs using a threshold of (cputime<60 & wallclock <60) (method 3)
- There are wide variations between sites. A more detailed study is needed to understand the results presented here but a high number of short jobs are seen with low workloads.
- Wallclock time used is that measured by the batch system (and accounted to ATLAS).
- On the whole, wallclock time used by empty pilots <0.1%.
- Sites have found that the real impact on the resource utilisation is much higher because of gaps after each job until the batch system starts the next one.

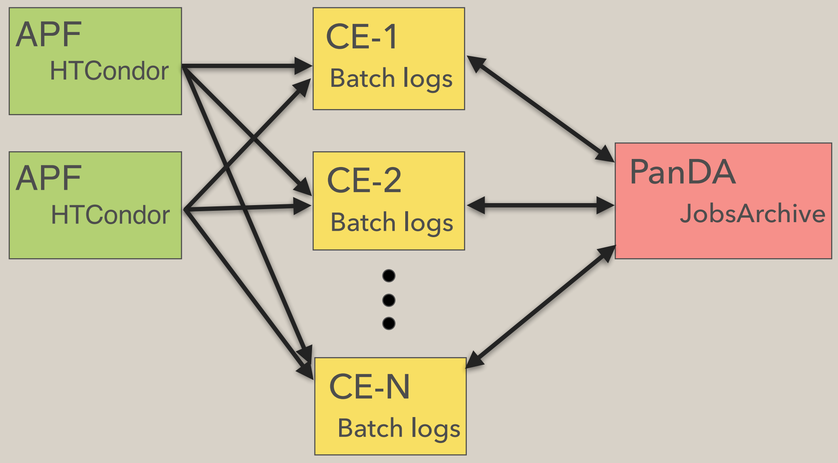
Site (August 2016 daily data)	Fraction of wallclock for short jobs (mean ± stddev)	Fraction of short jobs (mean ± stddev)
CC-IN2P3	(0.08 ± 0.11)%	(25 ± 14)%
FZK-LCG2	(0.22 ± 0.69)%	(40 ± 24)%
INFN-T1	(0.01 ± 0.01)%	(2 ± 3)%
MANC-HEP	(0.14 ± 0.13)%	(28 ± 17)%
NIKHEF-ELPROD	(0.88 ± 1.81)%	(41 ± 30)%

Pilot submission in ATLAS

- Most ATLAS pilots are submitted by AutoPyFactory (APF) where 12 factories service over 450 individual resources
- Each resource is handled by at least 2 factories to provide redundancy
- Globally the site resources have a diverse number of job slots ranging from ~100 to ~10k
- APF provides a rich 'sched' plugin system to moderate the number of pilots submitted to each of these sites.
- The plugins form a chain of logic taking input from the previous plugin and moderating the number of pilots submitted based on APF configuration

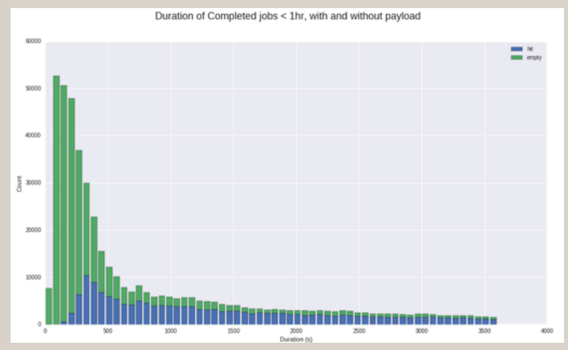
Ready	Checks the number of jobs ready to be run in the Workload Management Service (WMS), the number of previously submitted pilot still in idle state, and calculates the difference.
Scale	Multiplies by a factor the decision made by the previous plugin in the chain.
MaxPerCycle	Limit the maximum number of pilots to be submitted each cycle.
MinPerCycle	Limit the minimum number of pilots to be submitted each cycle.
StatusTest	Set number of pilots to submit when the WMS queue is in internal status test.
StatusOffline	Set number of pilots to submit when the WMS queue is in internal status offline.
MaxPending	Limit the number of pilots pending in the resource queue.

- The following sched plugins are currently used:



Wallclock time distributions

- In this panel APF job records are tagged as having a payload if a corresponding GTAG is found in the PanDA JobsArchive.
- This sample consists of jobs from a single day, and include records from 264 unique PanDA Queues.
- Empty pilots have the expected short wallclock time.
- Wallclock time is that measured by the APF HTCondor (EnteredCurrentStatus-QDate)



Pilot jobs which run a payload

Pilot jobs tagged as empty (no payload)

- These plots show the single-core ATLAS production queues for four different WLCG sites.
- Job records are from a single day (15 May 2016) in order to show the diversity of behaviour.
- In absolute terms NIKHEF and INFN-T1 have many more records tagged as empty pilots.
- The variation in empty job distributions between sites is expected due to the difference in available slots at each site and also the difference in Activated workload for each ATLAS queue.