

# Considerations on CPU and Memory limits in the WN

Alessandra Forti

ATLAS Jamboree

CERN

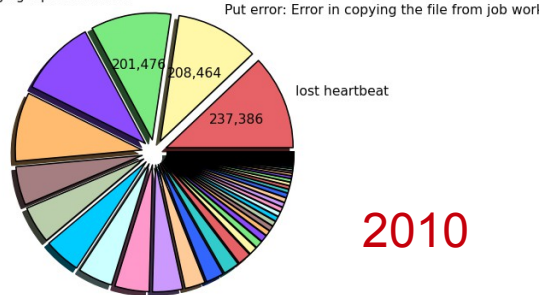
10th December 2012

# This talk

- Lost heartbeat: follow up of talks given in July
  - Review of CPU hours and walltime (Alessandra)
  - CPU and Wall time follow-up (Rod)
- Memory: how to limit memory leaks without killing all the jobs
- A proposal for a new dashboard

# Lost Heartbeat: Quantified

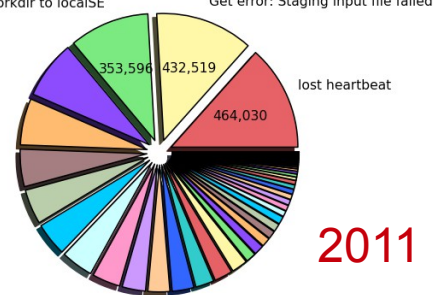
Panda Failures by ExitCode (Pie Graph) (Sum: 1,969,091)  
Get error: Staging input file failed



2010

- lost heartbeat (237,386)
- Get error: Staging input file failed (201,476)
- ATH\_FAILURE - Athena non-zero exit (173,162)
- unknown (93,552)
- job killed from the batch system: SIGTERM (87,371)
- Trf installation dir does not exist and could not be installed (75,409)
- Get error: input file staging timed out (44,495)
- trf is not installed in the CE (32,943)
- General pilot error, consult batch log (18,465)
- This job was killed by panda server (13,142)
- Put error: Error in copying the file from job workdir to localSE (208,464)
- Undocumented Execution Error Code (192,130)
- DQ2 server error (96,823)
- Expired three days after submission (92,232)
- Transfer timeout (2weeks) (86,152)
- Get error: Failed to get LFC replicas (51,029)
- Athena release is not installed in the CE, or trf failed due to "Unknown Problem" (43,478)
- Aborted by Extif (20,139)
- Get error: No such file or directory (16,669)
- misc 68 more

Panda Failures by ExitCode (Pie Graph) (Sum: 3,464,002)  
error in copying the file from job workdir to localSE



2011

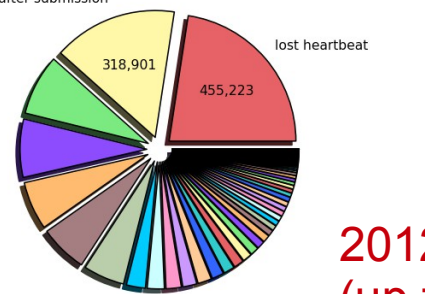
- lost heartbeat (464,030)
- Put error: Error in copying the file from job workdir to localSE (432,519)
- Undocumented Execution Error Code (255,936)
- ATH\_FAILURE - Athena non-zero exit (205,552)
- Get error: Failed to get LFC replicas (153,681)
- Get error: No such file or directory (103,628)
- Adder could not add files to the output datasets (96,348)
- Trf installation dir does not exist and could not be installed (79,460)
- Transfer timeout (2weeks) (43,748)
- General pilot error, consult batch log (38,953)
- misc 76 more
- Get error: Staging input file failed (432,519)
- Undocumented Execution Error Code (255,936)
- Get error: Replica not found (165,178)
- Expired three days after submission (146,895)
- Aborted by Extif (124,239)
- job killed from the batch system: SIGTERM (96,670)
- DQ2 server error (81,073)
- Athena release is not installed in the CE, or trf failed due to "Unknown Problem" (43,478)
- trf is not installed in the CE (43,478)
- misc 76 more

- All Atlas T2 errors per year.

Lost heartbeat is cumulatively the most dominant error in atlas production at T2s.

It is important also in analysis but in the monitoring it's squashed by more common Athena problems.

Panda Failures by ExitCode (Pie Graph) (Sum: 2,015,203)  
Expired three days after submission



2012  
(up to July)

- lost heartbeat (455,223)
- Unknown Transform error (155,929)
- ATH\_FAILURE - Athena non-zero exit (124,013)
- Undocumented Execution Error Code (105,963)
- Get error: Failed to get LFC replicas (143,659)
- Exception caught by runjob (36,014)
- job killed from the batch system: SIGTERM (30,233)
- Trf installation dir does not exist and could not be installed (25,327)
- Adder could not add files to the output datasets (21,259)
- Put error: LFC registration failed (18,388)
- misc 97 more
- Expired three days after submission (318,901)
- Put error: Error in copying the file from job workdir to localSE (149,546)
- Get error: Staging input file failed (122,687)
- Aborted by Extif (48,755)
- Get error: Replica not found (39,244)
- Unspecified error, consult log file (35,181)
- DQ2 server error (27,757)
- Bad replica entry returned by lfc\_getreplicas(): SFN not set in LFC for this guid (2)
- General pilot error, consult batch log (19,653)
- misc 97 more

# Lost Heartbeat: what is it?

- Lost heartbeat happens when the jobs loose connection with the panda server.

Causes can be many: WN crashes, CE crashes, power cuts, network problems, panda crashes, WMSglidein problems....

But the most common is: jobs exceed batch system limits on resources. In particular CPU or Wall time.

- It's not because sites don't give enough resources
- VOid card: 3120 mins (52h) on a 1 kSI2k machine
  - ~2.2 kSI2k should require only 23.6h
  - Mcr with 48h queue still failed
  - Other sites have even longer queues even up to 96h

# Superficial dig on tasks

- There are some tasks that have bugs and they simply loop until the batch system kills them.

<https://savannah.cern.ch/support/?129431>

- Some other tasks seem just to require a very long time and are not brokered correctly.

# Wasted resources

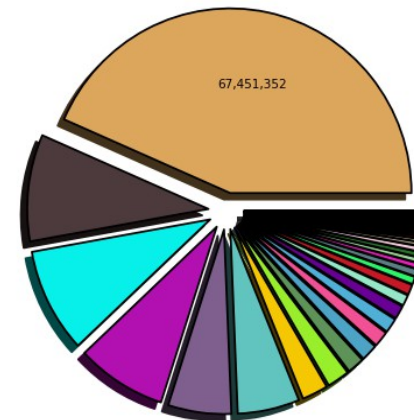
- Problem is not the number of jobs is the amount of CPU/Wallclock hours wasted

8,535\*48=409,680 hours or 17,070 days

- Prototype dashboard now reports CPU time/Wall time per error.



WallClock Consumption of Panda Failed jobs by ExitCode (Sum: 155,188,321)



- Just as indication

UK last week

Lost heartbeat 43%

lost heartbeat (67,451,352)  
 Reached batch system time limit (14,223,004)  
 Athena core dump or timeout, or conddb DB connect exception (8,741,812)  
 ATH\_FAILURE - Athena non-zero exit (3,596,983)  
 Athena crash - consult log file (2,336,261)  
 Unspecified error, consult log file (2,124,413)  
 Put error: Local output file missing (3,820,085)  
 Not documented. Exitcode: 15\_ Reached batch system time limit (1,379,726)  
 Not documented. Exitcode: 9\_ Put error: File copy timed out (969,090)  
 Not documented. Exitcode: 0\_ Reached batch system time limit (600,015)  
 Put error: Error in copying the file from job workdir to localSE (14,513,800)  
 Not documented. Exitcode: 143\_ Reached batch system time limit (12,470,818)  
 Not documented. Exitcode: 8 (8,566,893)  
 Get error: Input file staging timed out (3,093,298)  
 Unknown Transform error (2,276,563)  
 Job killed from the batch system: SIGTERM (1,985,445)  
 Not documented. Exitcode: 8\_ Athena ran out of memory (1,478,402)  
 Athena crash - consult log file\_ Reached batch system time limit (1,012,498)  
 Athena crash - consult log file\_n/a (730,712)  
 n/a 60 mins

# Two problems

- Why these jobs do not terminated with a different error?

## Pilot problem

- Why jobs that require long time aren't scheduled on longer queues?

## Brokering problem

# Pilot Problem

- Most batch systems send a SIGTERM before a SIGKILL. Pilot doesn't seem to catch this consistently

Sometimes it does and we get “Job killed by signal 15: Signal handler has set job result to FAILED, ec = 1201” or “Reached maximum batch system time limit “

Sometimes it doesn't and we get “lost heartbeat”

- Catching the signal consistently would help

Report a correct error code rather than lost heartbeat

- And identify which jobs are using more than they should

Terminate the job cleanly

- This requires to configure a time delay between at SIGTERM and SIGKILL and the pilot using it.

- Fixing this doesn't avoid the waste of resources from jobs that really take long time.



# Brokering Problem

- To avoid wasting resources we need to better broker the jobs.

- A new algorithm since last August

Use scout jobs measures to predict how much time they require and use maxtime parameter in schedconfig to broker the jobs.

This helped reducing dramatically the number of “lost heartbeat” due to exhausted CPU time/Wall time.

- Didn't work at all sites

RAL was complaining still few weeks ago

- Looping jobs might be the culprit though

# maxtime

- There were setting maxtime automatically for all queues

Schedconfig/bdii inconsistencies

Obsolete queues

It looks fixed now

- The algorithm is an approximation it doesn't guarantee all jobs will be schedule correctly

From this week a higher cpucount is assigned to jobs that fail

- Automatically set at 96h if jobs fail with “Reached maximum batch system time limit”.

# Memory errors

- Another subset of “Lost heartbeat”
- Requested 2GB pmem 4GB vmem

Most sites at Atlas request do not set limits in batch system

- Multicore machines have enough memory to allow jobs exceed the requirement without problem

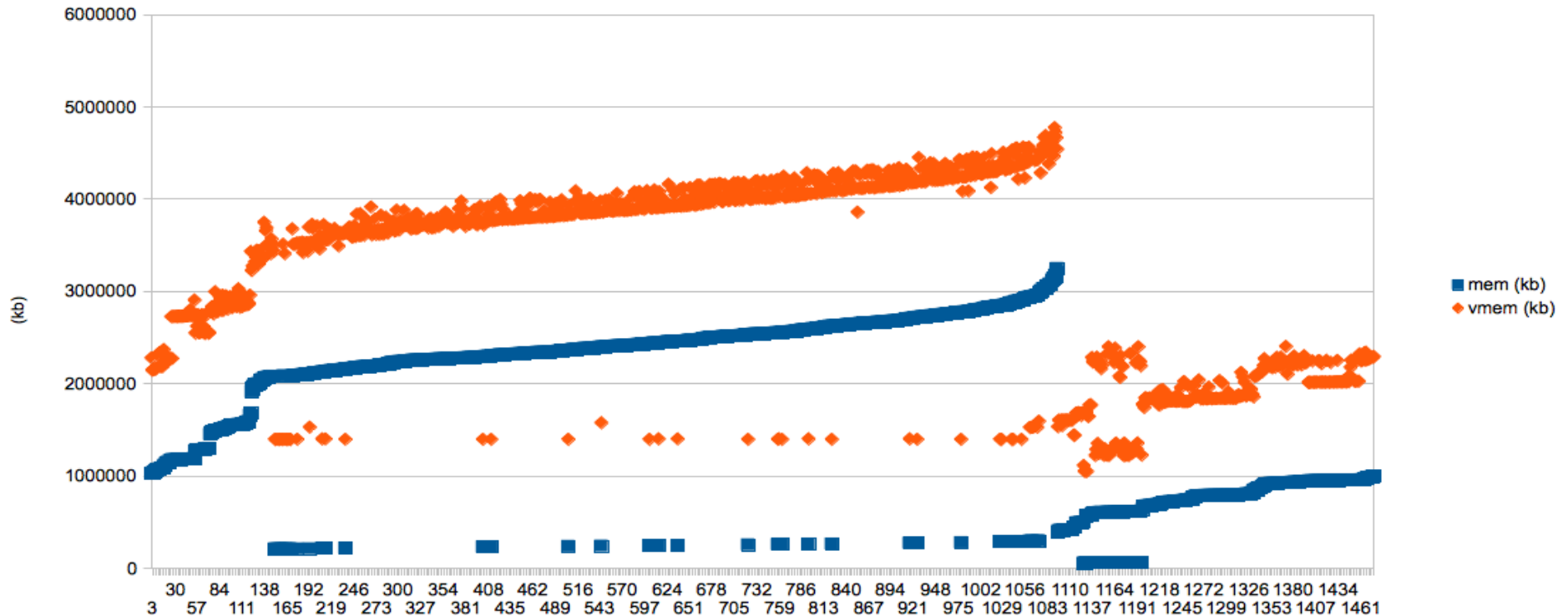
Still the physical memory is not inexhaustible

No way to contain memory leaks

WNs crashing without much reason

- Other (VOs) jobs affected
- Wasted resources

# Problem



- Want to limit memory leaks without affecting most jobs

# A way to limit

- Batch System

Recipes are batch system dependent

To limit in the batch system without killing most jobs you need to over allocate resources.

Torque/Maui

<http://tinyurl.com/d692cro>

Simplest recipes affects all the users

- Often not as demanding but...

- SchedConfig/AGIS queue config

1 memory parameter to set

Batch system independent

Uses ulimit to set vmem per process

Leaves a trace in the log files

- Setting VMEM limit to 3000000kB

No worry about pilot/batch system communication

Atlas self contained

Haven't had a crashing node since.

# New dashboard proposal

- In the past 1.5 year I've looked at a number of failure modes

Cvmfs files corruption, cvmfs timeout, lost heartbeat, replica not found, memory problems, missing shared libraries.....

- Used the historical dashboard, panda monitor matching batch system logs
- Historical dashboard helped a lot giving aggregate information

## But it has some limitations

- No task information
- No links to the jobs
- No possibility to select errors
- Difficult to understand if errors spread in multiple sites

# New dashboard proposal

- DDM2 can do all this for transfers
- Proposal to have a “DDM2” style dashboard for jobs

Task vs sites matrix

Success and errors rates

- Errors grouped by type

Links back to job errors

Capability to plot single or multiple errors

- Per task
- Per site

# Conclusions

- Lost heartbeat has been tamed by better brokering production jobs (requires maxtime).

Analysis and group jobs might still cause problems

- There is neither “user” nor group scout scheme

Pilot catching SIGTERM **consistently** would help identify the jobs that still are not brokered correctly.

- Memory errors can be put under control by limiting the memory parameter in schedconfig on all the queues.
- A DDM2 style dashboard for job errors would drastically cut down the debugging effort.