Identifying problematic WNs using Pilots

Chris Walker, Alastair Dewhurst

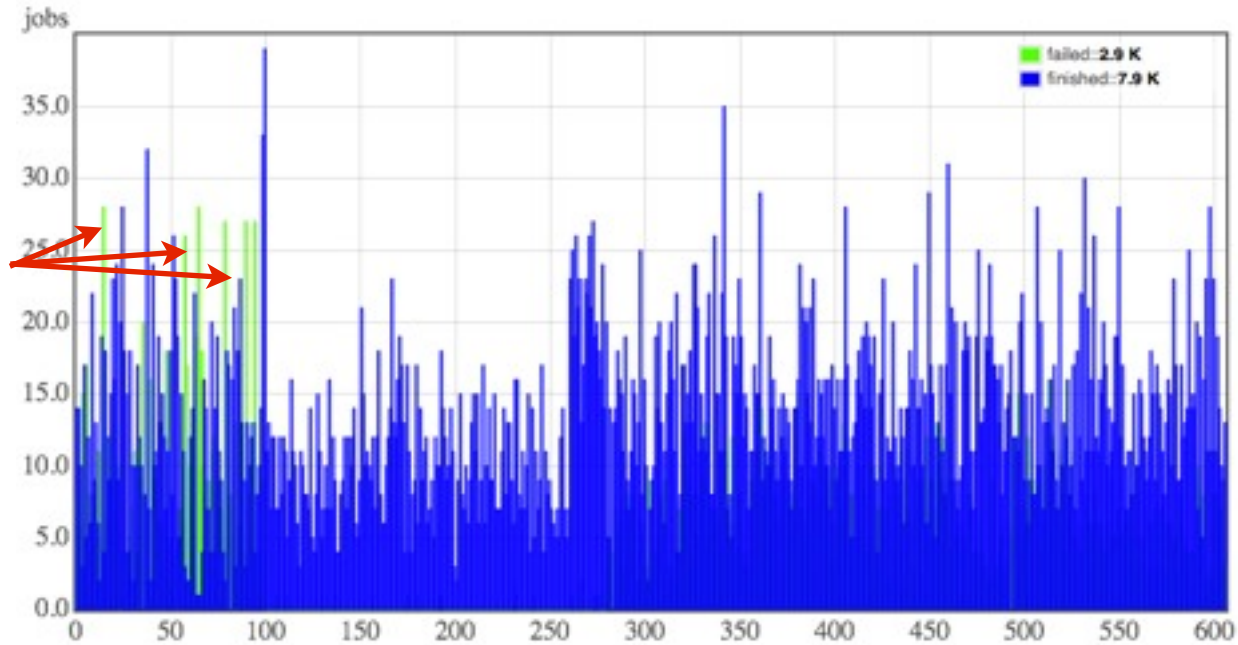
Motivation

- WNs are frequently abused by VO Jobs.
- No surprise that they occasionally get into a 'broken' state.
- One or two broken WN may not result in a large error rate.
  - Broken machines may persist for a long time
  - Errors message are likely to be obscure and waste a lot of time debugging. (e.g. CVMFS inode issue)
- Some but not all types of problems are known in advance and test/checks can be pre-written.

Current method

- Shifter/User notices job failures.
  - If smart/lucky will identify problem is restricted to just a machine or two.
- GGUS ticket to site (possibly with helpful information!).
- ATLAS have some of the better VO monitoring but shifters frequently don't have time.

Broken nodes, bad luck or is the site testing some new feature?



Idea

- The pilot jobs already send error messages back to their respective VOs, why not also inform the sites by leaving a message on the WN?
- A site would create a directory where pilot jobs could write error messages.
  - If it didn't exist pilot would just skip step
  - tmp watch to clean up old messages
- When a pilot job exits with an error it leaves a message in this directory.
- Spoken to Pilot developers of main VOs - No opposition to suggestion and would be trivial to implement.

Error message

- What form should the error message take?
  - Something easily parseable: XML?
- What fields do we need?
  - Time of start and end of job.
  - VO name.
  - VO job ID (to get more details if necessary).
  - VO Job type (e.g. Test, User Analysis, High memory)
  - Error message (jobDispatcher: lost heartbeat : 2014-05-29 01:18:43)
- Any particular naming convention for each file?

Site checks

- Sites will still need basic checks for sudden job deaths (eg. power cut).
- What site checks could be done?
  - If more than X error messages are left in Y minutes.
  - Problems from multiple VOs?
- Any useful metrics?
  - Number of failed jobs and lost CPU time easy to measure
  - Daily summary of frequent error messages
- Any volunteers to write some checks?

Machine/Job Features

- Spoken to Stefan Roiser, chair of the WLCG Machine / Job Feature taskforce, who thought it a useful feature to have:
  - The only issue I have is that with the communication back from the VO we initially designed to have this within the "pilot/payload area" of the VO, i.e. as soon as the payload goes away (e.g. dies for some reason) this area would be cleaned up and not available anymore to the resource provider, but this may be a minor issue we'd need to change.
- He said he would bring it up at the next meeting, although this is yet to be scheduled.
- https://twiki.cern.ch/twiki/bin/view/LCG/MachineJobFeatures

Conclusions

- Technically not a problem to implement.
- WLCG task force reasonably happy to include it, so a standard between VOs should be possible to reach
- The more sites involved the better it will be.
  - Do UK site admins want this?
- Discuss?