



Enabling Grids for E-scienceE

FTS failure handling

Gavin McCance

Service Challenge technical meeting

21 June 2006

www.eu-egee.org
www.glite.org



- **What to do if a channel starts to fail**
 - It starts dropping files (after retries)
 - It “drains” pretty fast
 - Default for a file is 3 retries with 10 minutes between tries
- **We sometimes observe intermittent problems at 1am**
 - The SRM goes bad
 - Drops 10,000 files after multiple retries
 - 35 minutes later, it all works again, apparently without human intervention
 - Do it get you out your bed or not?

- Make use of the FTS **Hold** state
- Jobs that have failed after X retries go to **Hold** instead of **Failed**
- Twice daily procedure: check your channel for **Hold** jobs, understand and fix the problems, set **Hold** jobs to **Pending** for another retry
- VO dependent
 - Not all VOs want this: they prefer fail-fast
 - The retry policy and “Hold” policy is configurable per VO
- This has always been part of FTS
 - But we don’t run the procedure just now

- **On repeated failures**
 - Some criteria to detect this
 - 100 files failed consecutively, say
 - and/or 70% failure over last 1000 files
- **Halt the channel and send an alarm**
 - GOOD: you don't drain many jobs
 - BAD: if it's 1 am, you get someone out their bed -- since no transfers will run until someone restarts the channel (after fixing the problem), and you can't afford to lose 8 hours transfers

- **On repeated failure**
- **Stuttering halt**
 - Halt the channel for X minutes
 - and send notification
 - Then retry
 - If it's still bad, send alarm
 - BAD: you drop more jobs on each retry if the problem is still there
 - GOOD: the pause may be sufficient to allow the storage to recover from intermittent problems / overload

- Using the **Hold** state to limit the failures from “job drains”
 - Jobs can always be rescued in the morning
 - This is a daily operations procedure
- Use **stuttering halt** to prevent 8 hours downtime for what was a 45 minute intermittent problem
 - Send notification for first couple of halts
 - Send alarm (get-out-of-bed) for repeated failures

- **FTS server at the tier-0 has an alarm for you**
 - How does your local monitoring know about this?
 - Web-page that your local monitoring system polls?
 - Either binary – “it’s good / it’s bad”
 - or... X failures in last hour – you decide what’s worth getting out of bed for
- **Once you’re debugging...**
 - Exposing the information the FTS has about what’s going wrong
 - It does have a lot of information: we don’t expose it well enough
 - performance figures, aggregate failure classes, log files of individual transfers, all the error messages we get back
 - Work in progress to make this more accessible
 - Critical for operations: this is why the FTS is there...