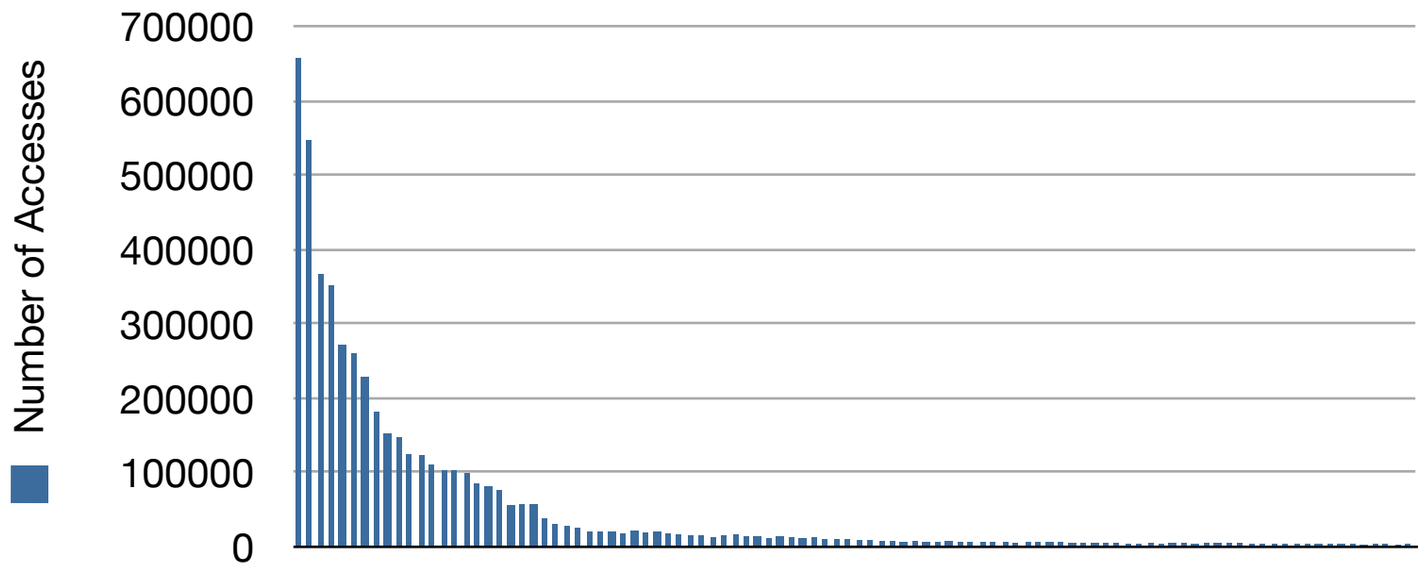
ATLAS Demonstrator Update: PanDA Dynamic Data Placement

Kaushik De, Tadashi Maeno, Torre Wenaus, Alexei Klimentov, Rodney Walker, Graeme Stewart



Reminder of the problem

- For user chaotic analysis it was (and is) hard to predict what data is going to be used
  - Pre-placing data during the early LHC running worked
    - But clogged networks and disks
    - Very poor 'hit rate' on data
- Try to do something more responsive
  - But without impacting on users

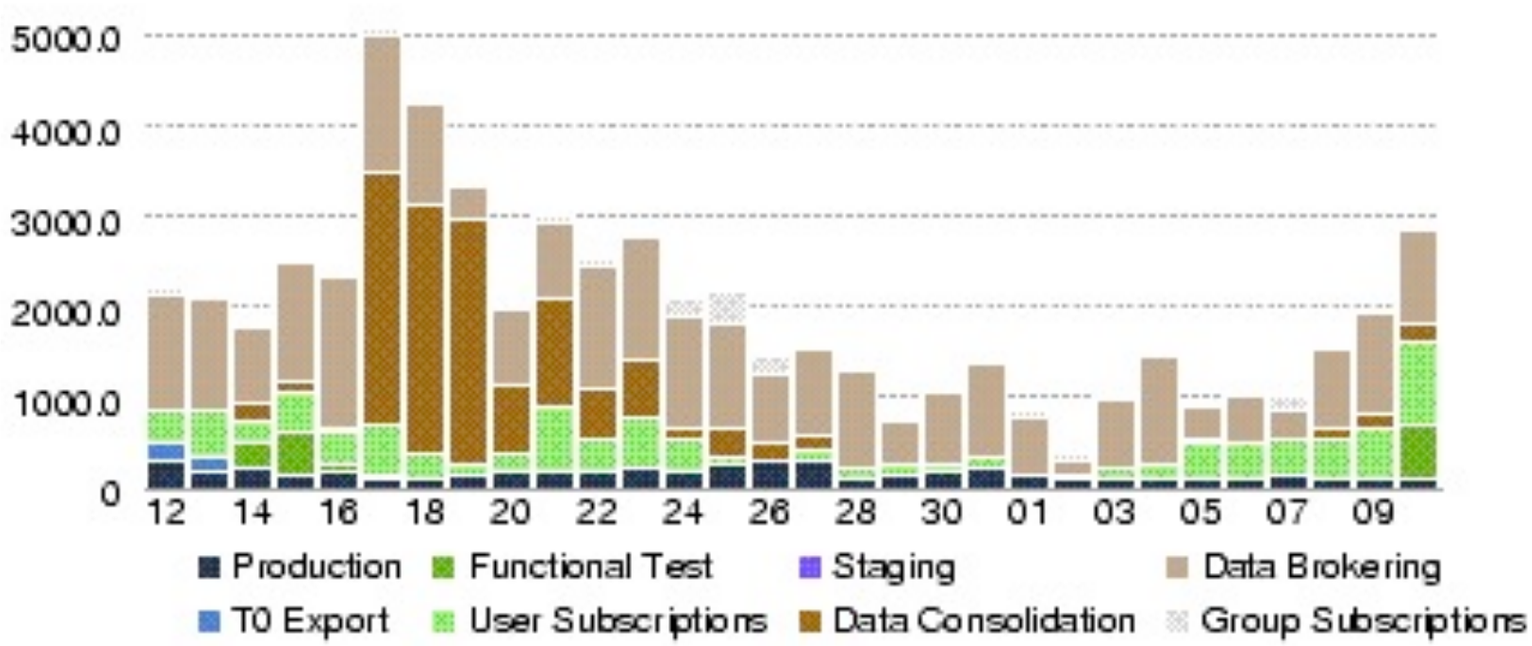
Basic PD2P Model

- Continue automatic replication to T1s (these are the repositories)
- Reduce, or eliminate, pre-placement of data at T2s
- Trigger a replication of data from T1 to T2 as soon as a user submits jobs on a dataset
  - But run jobs in the T1 initially → no slow down for the user
  - Make additional replicas if needed → based on jobs in queue now
  - Select T2 by normal brokering rules
  - Clean up done by Victor (DDM popularity service)

Recent Improvements

- Extended to almost all ATLAS clouds
  - Including CERN for the LST demonstrator
- Re-brokering of analysis jobs is now in place
  - Jobs can hop from T1 to T2 once the PD2P triggered replica arrives
- Space check at target site implemented
  - Site is excluded if there is not enough free space
  - Although clean-up was done by DDM, sometimes this was too slow
- Stop empty sites pulling in too many subscriptions
  - Making many subscriptions: compete with one another and stop rapid completion
  - Negative weight for recent subscriptions

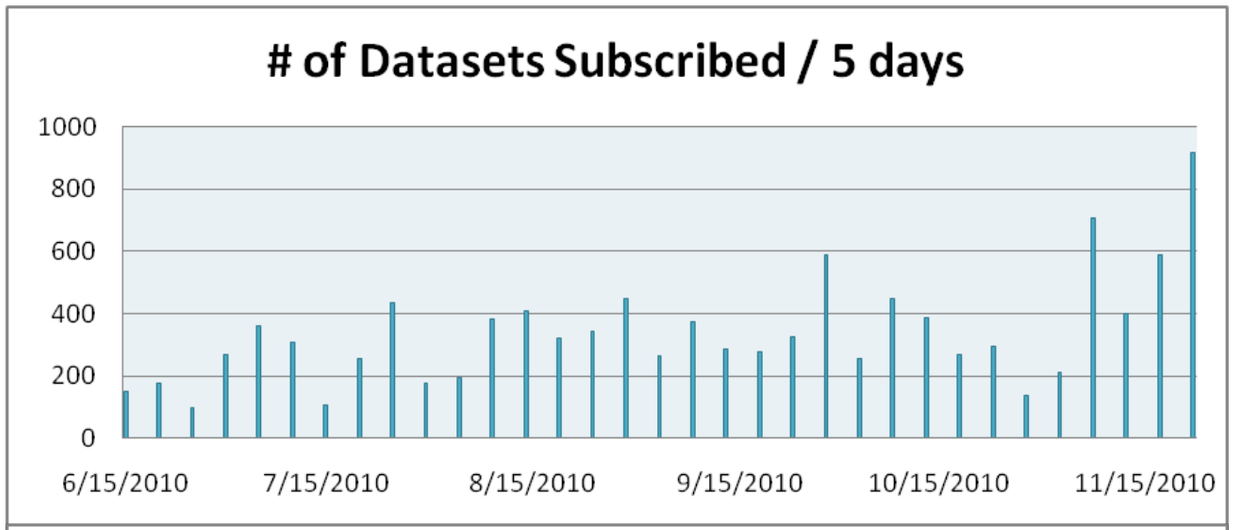
PD2P Activity

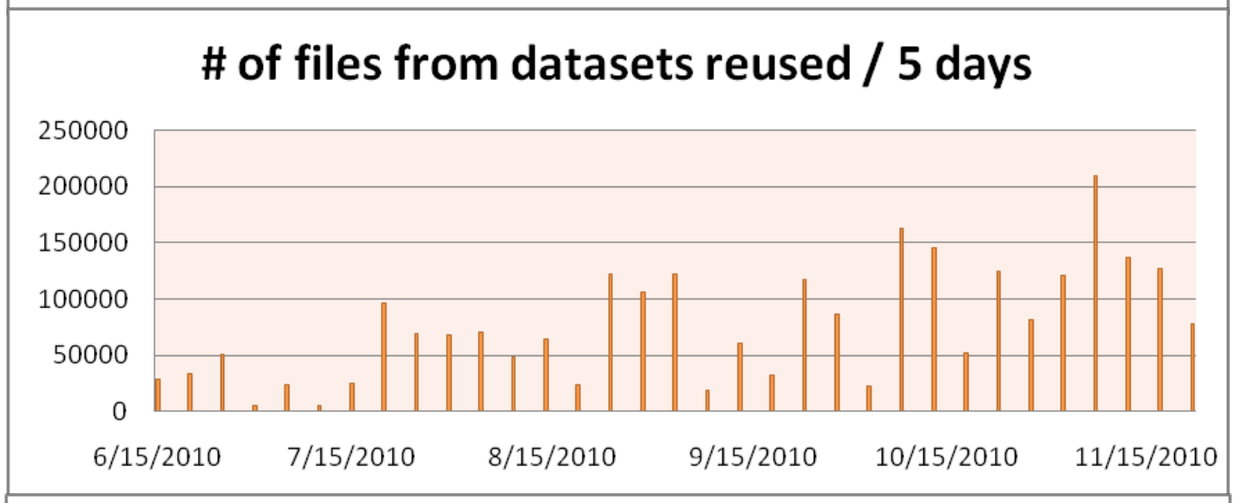


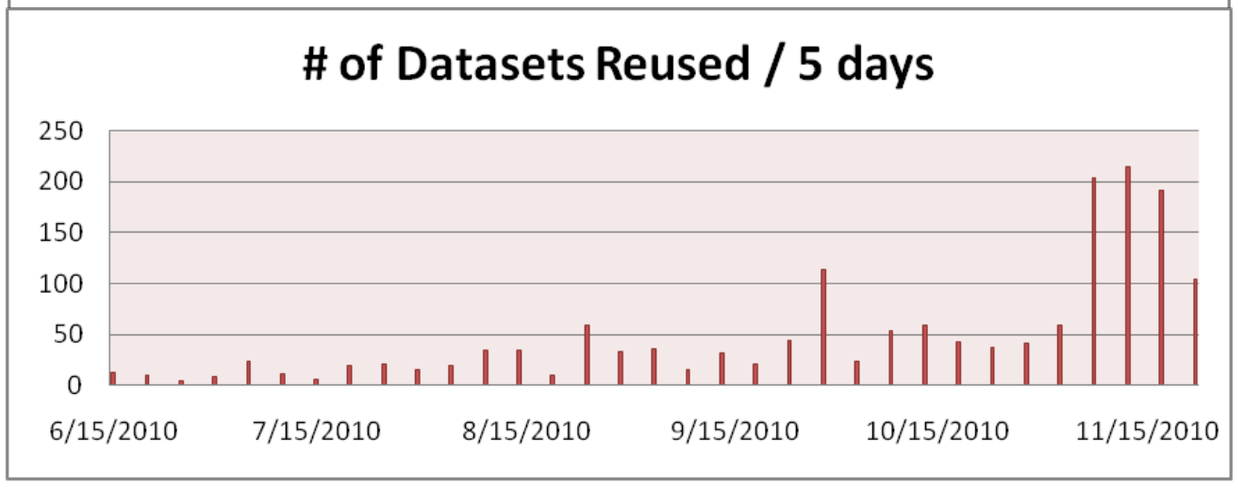
- PD2P now responsible for significant data movement on the grid

Reuse Improving

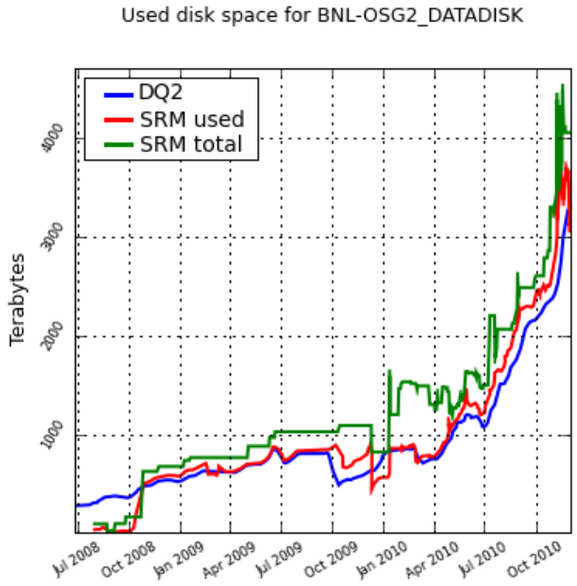
- Helped by re-brokering

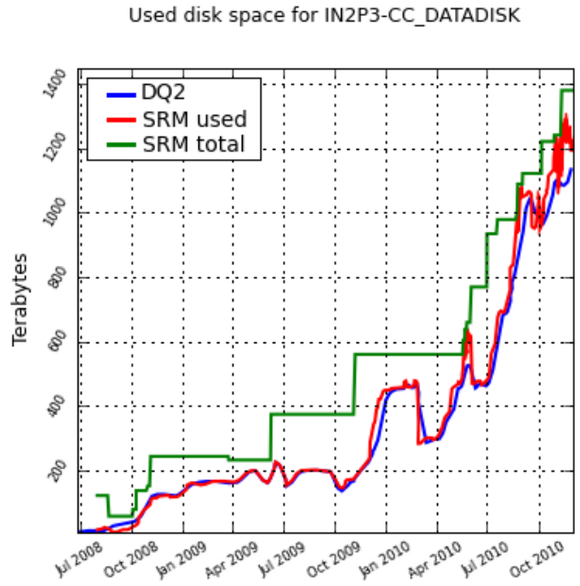


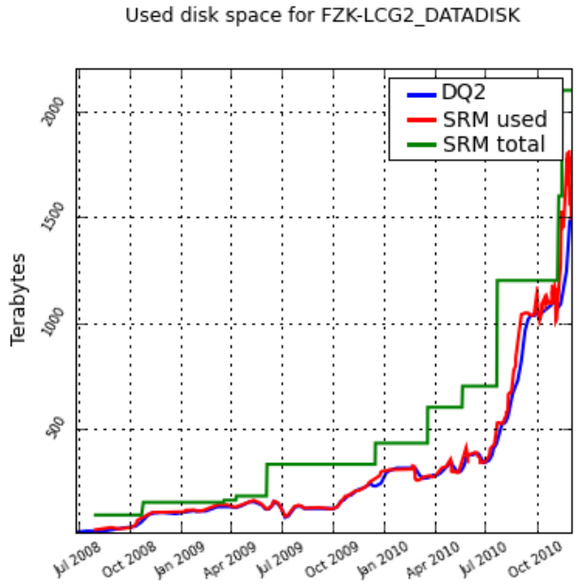




T1 Disk Crisis







- T1 Disk now running very, very close to full
- 2011 run still to come

PD2P at Tier-1s

- Reduce the number of primary datasets held at T1s
- Secondary replicas are made by PanDA – usage based
  - Initial copies are made at Tier-2's (using current PD2P algorithm)
  - Check the number of waiting analysis jobs for any dataset needed by user
    - If too many waiting jobs (based on some lo-threshold), and no copies already made by PD2P, start replication to Tier-1 and Tier-2
    - Use MoU share to decide which Tier-1 gets this extra copy, and use brokerage to decide which Tier-2 gets copy
    - If still too many waiting jobs (that is, more than some hi-threshold), make another copy (could be at Tier-1 or Tier-2)
- Minimally, 2 copies of all data are available ATLAS-wide, more copies are only made for hotly used data

Further Improvements

- T1 PD2P
- Look in detail at what data turns out to be hot
  - System can cope with current levels of replication, but we'd like to continue to improve reuse
- Improve the algorithm for selecting sites
  - Better feedback between DDM and PanDA (but without overcomplicating)
- Always keen to link up with other projects, e.g., LST