Extending PD2P

Kaushik De

Univ. of Texas at Arlington

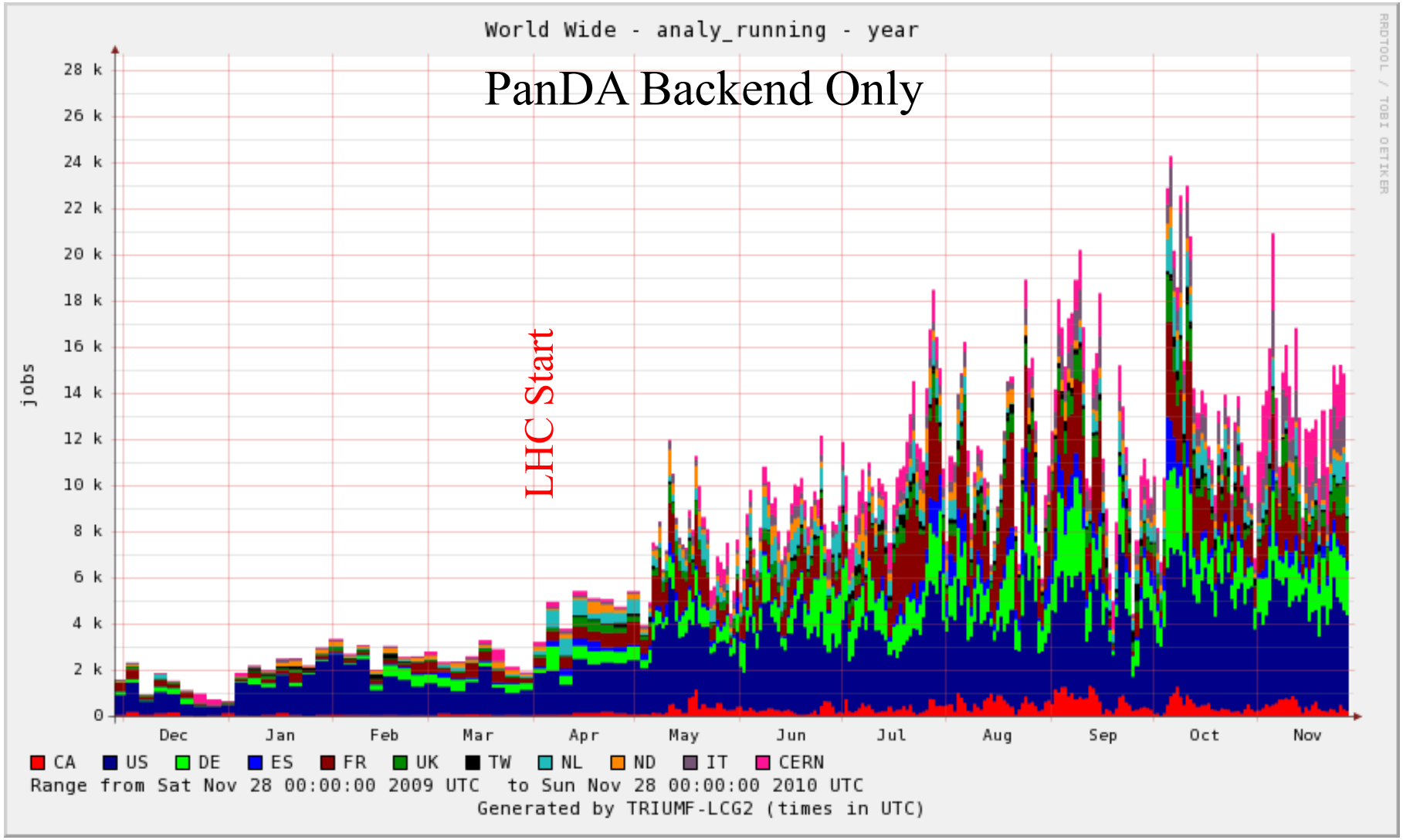
S&C Week, CERN

Dec 3, 2010

Introduction

- PD2P was introduced ~5 months ago
- PD2P adiabatically but dramatically changed data distribution model for analysis users
  - Instead of pushing data to Tier 2 sites, we now mostly pull them as needed for user analysis (data AOD's continue to be pushed)
  - Jobs still go to data – the pull is done asynchronously for future jobs
    - First user jobs go to Tier 1 sites since they still get pushed data
  - So far, working very well!
- In this talk, I will present
  - PD2P status (see parallel session talks for more details)
  - Plans for extending PD2P to Tier 1 sites
  - Simultaneously, reduce pushed AOD's to Tier 2 sites

Huge Rise in Analysis Activity

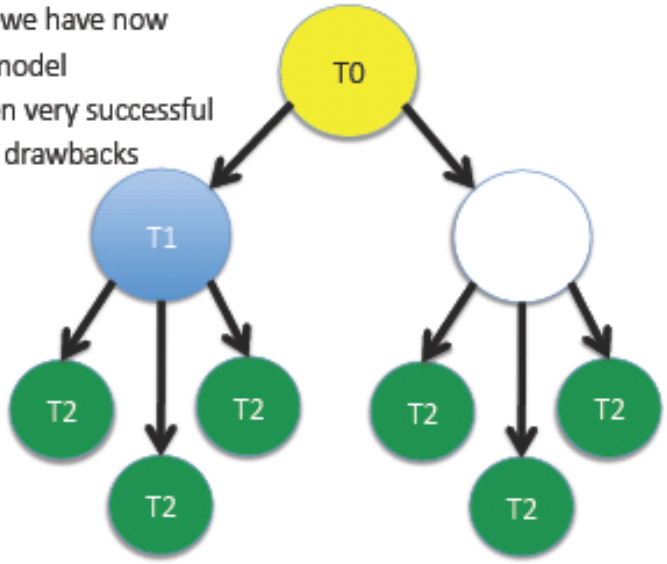


Data Distribution is Very Important

- Most user analysis jobs run at Tier 2 sites
  - Computing model: push data out to Tier 2 sites
  - Jobs are sent to data
  - Difficult since many data formats and many sites
  - We adjusted frequently the number of copies and data types in April & May
  - But Tier 2 sites were filling up too rapidly, and user pattern was unpredictable
  - Most datasets copied to Tier 2's were never used

Data placement model
The"Monarch Model"

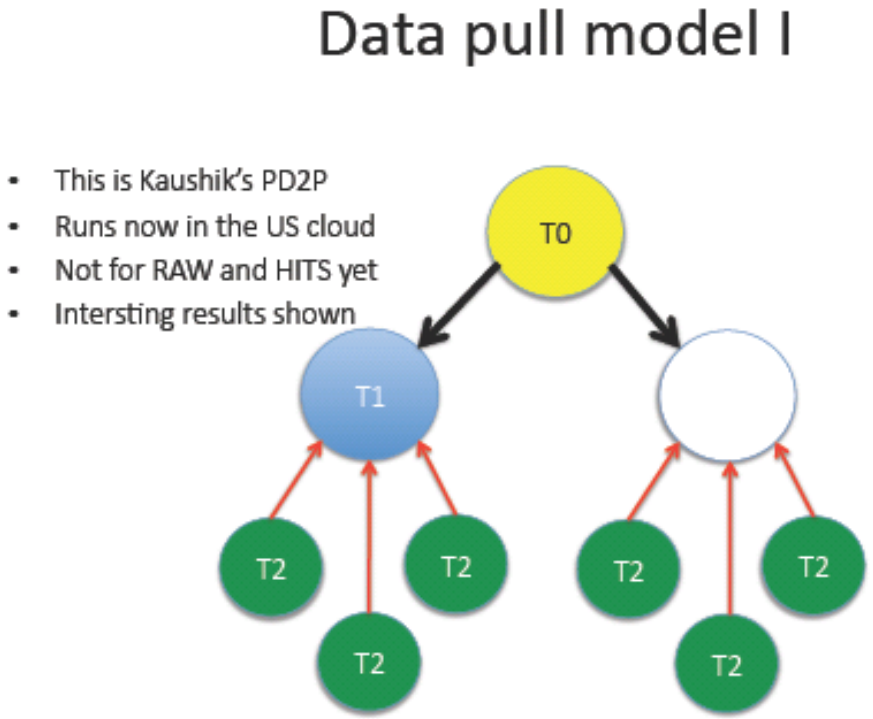
- This is what we have now
- It is a push model
- And has been very successful
- But has also drawbacks



From Kors – last SW week

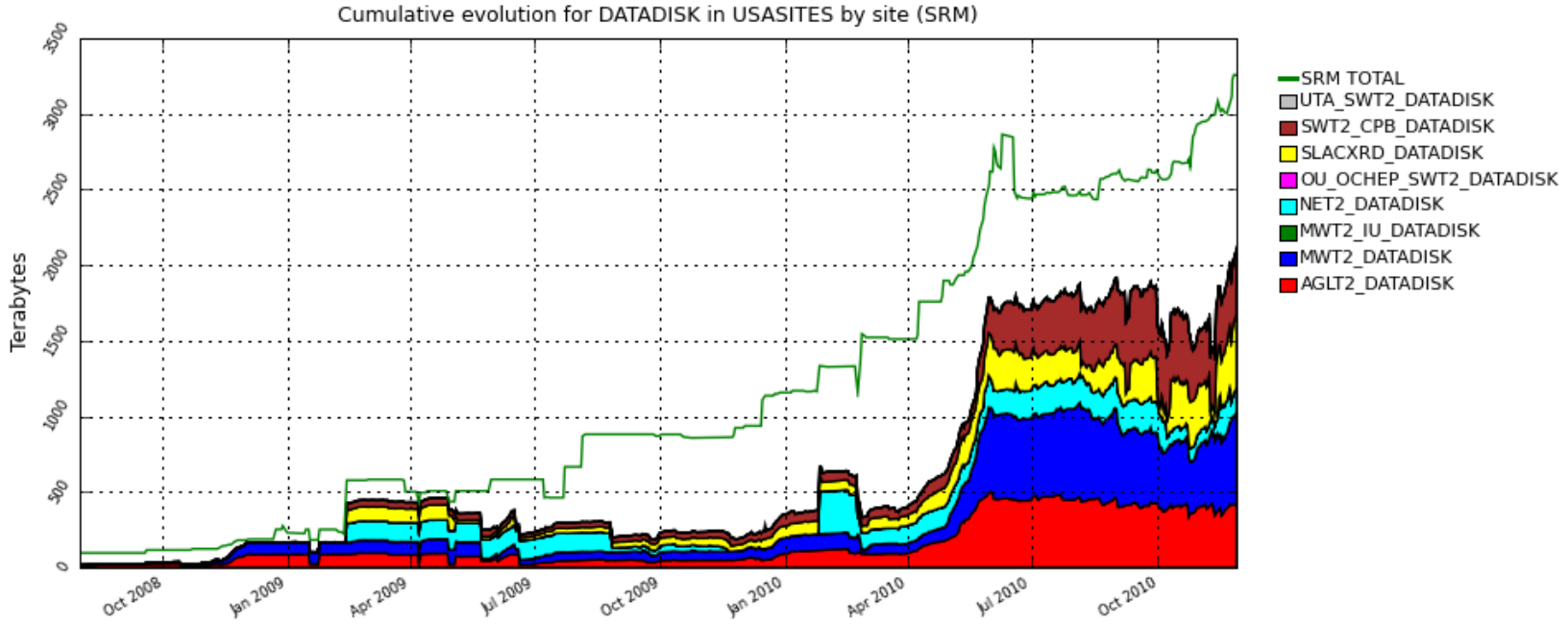
We Changed Data Distribution Model

- Reduce pushed data copies to Tier 2's
  - Only send small fraction of AOD's automatically
  - Pull all other data types, when needed by users
  - PanDA decides when and where to push data
  - Note: for production we have always pulled data
- But users were insulated from this change
  - No delays in running jobs
  - No change in user workflow



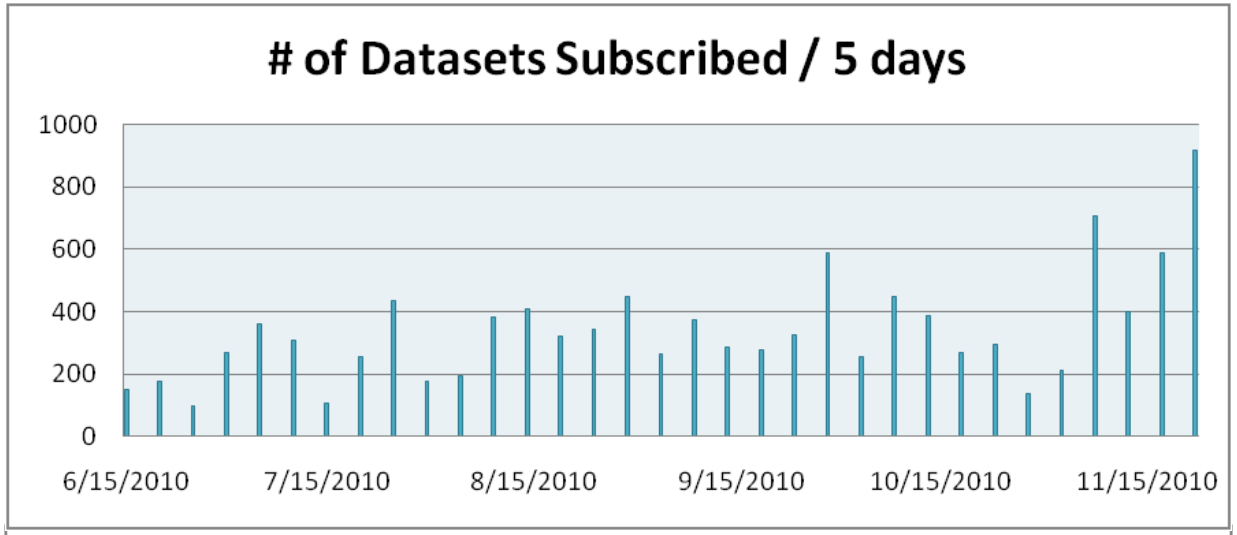
From Kors – last SW week

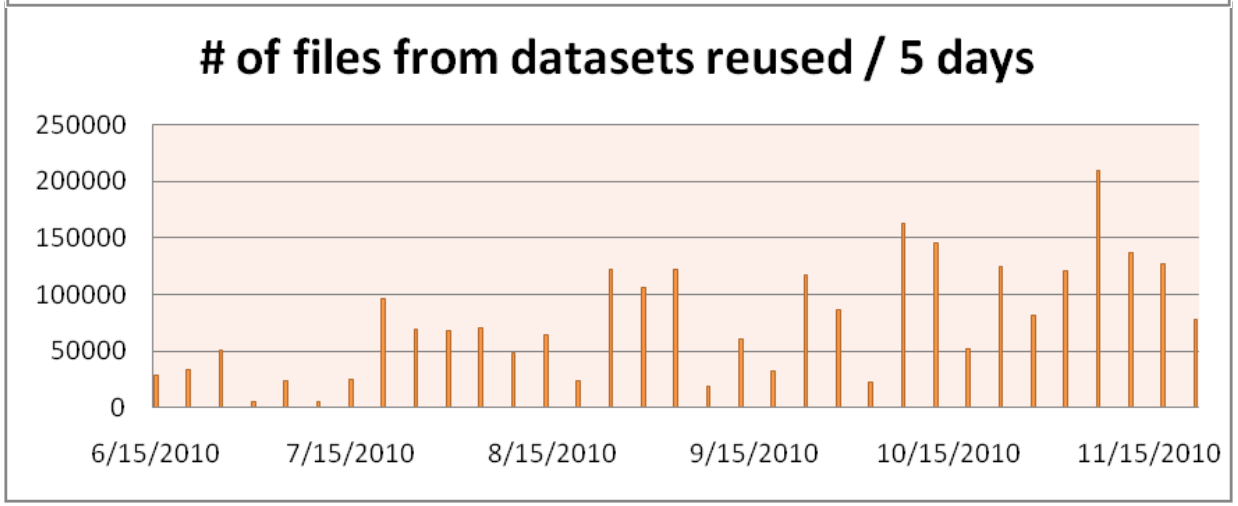
Data Flow to Tier 2's



- Example above is from US Tier 2 sites
  - Exponential rise in April and May, after LHC start
  - We changed data distribution model end of June – PD2P
  - Much slower rise since July, even as luminosity grows rapidly
- PD2P is now running in all clouds ATLAS-wide

PD2P Statistics

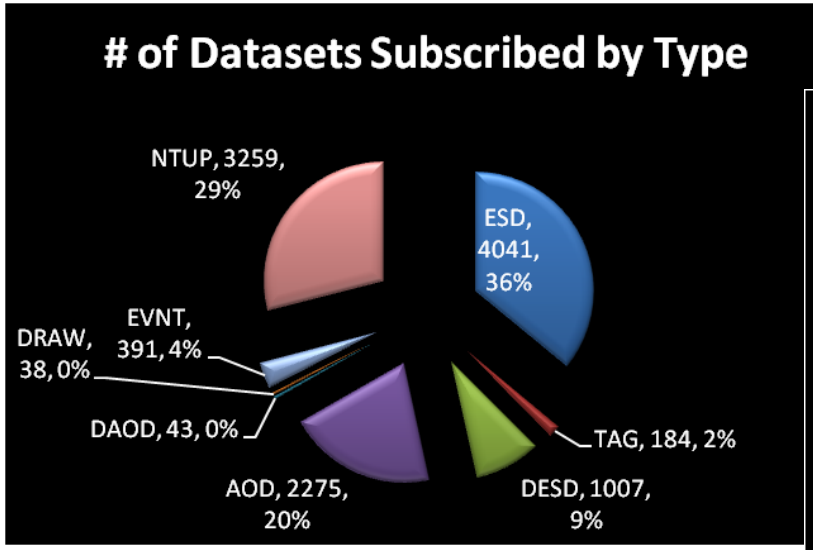


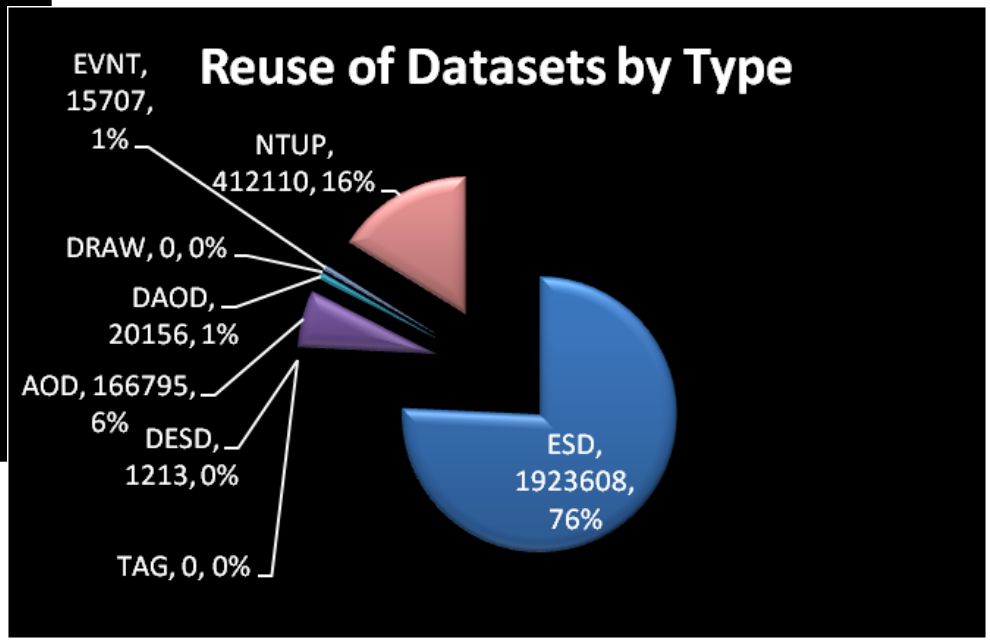


Tens of thousands of datasets are never subscribed and not in these plots

Patterns of Data Usage

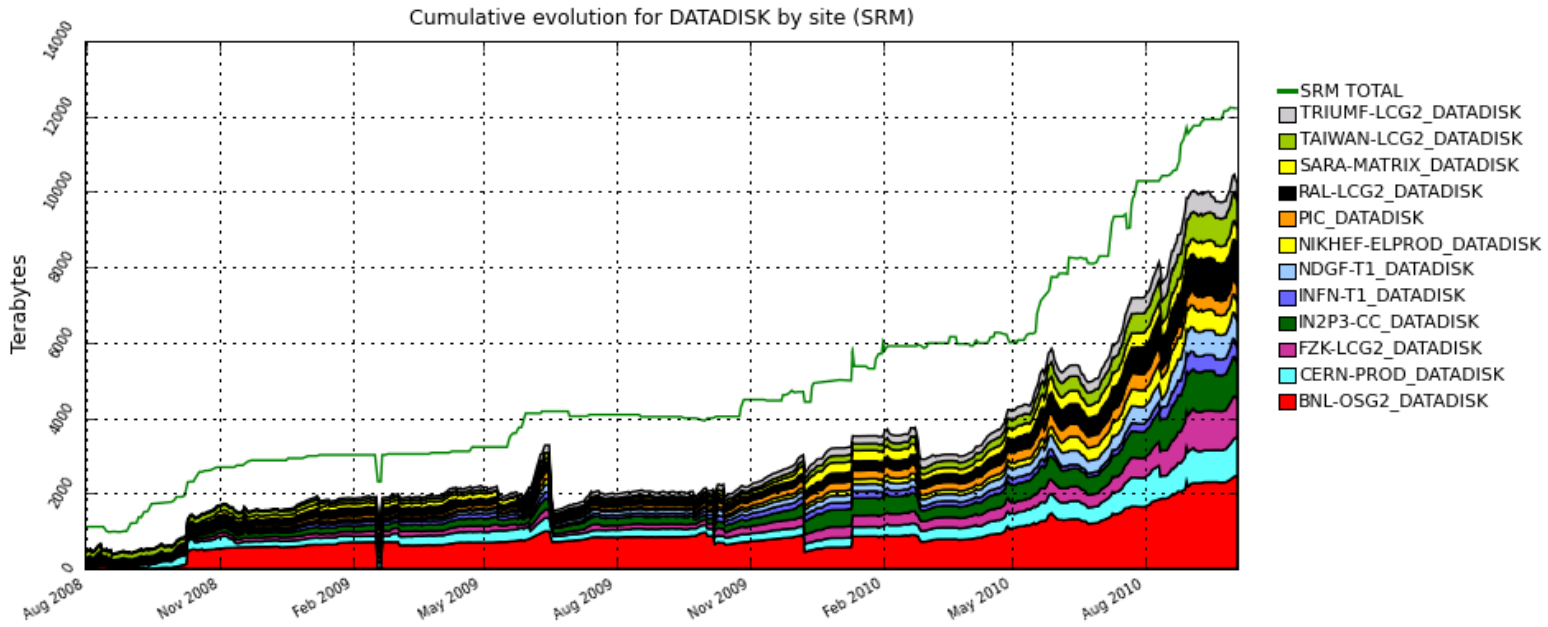
- Interesting patterns observed by data type
- During past ~5 months:
  - All types of data subscribed: ESD, NTUP, AOD, DESD are popular
  - But highest reuse (counting files): ESD (and data AOD's which are automatically distributed, hence not in these charts)



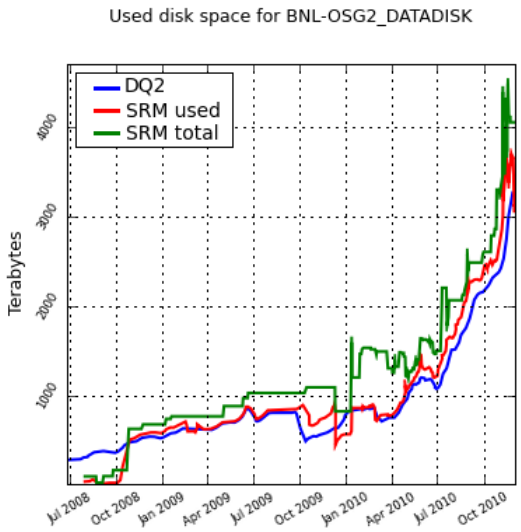


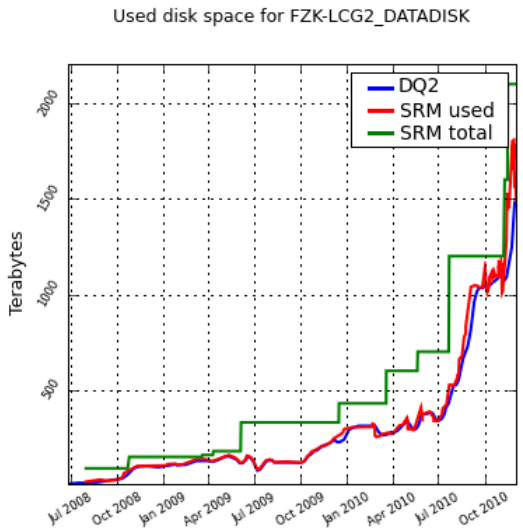
Extending PD2P

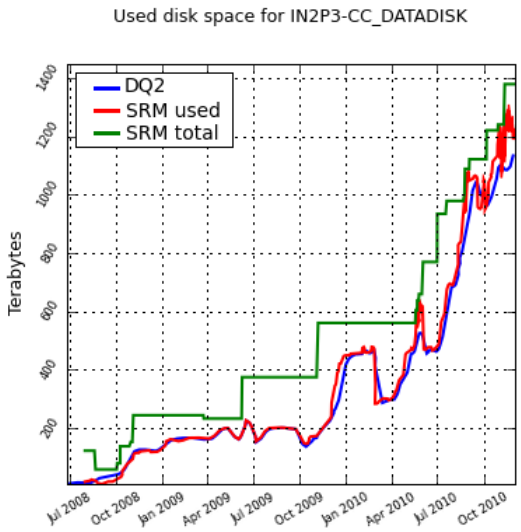
- PD2P has worked very well for Tier 2 sites
  - We can reduce pushed data even more
- But now we have disk crises at Tier 1 sites
  - Last month, storage at many Tier 1 sites got full
  - Reprocessed data, continuing LHC data, more MC…

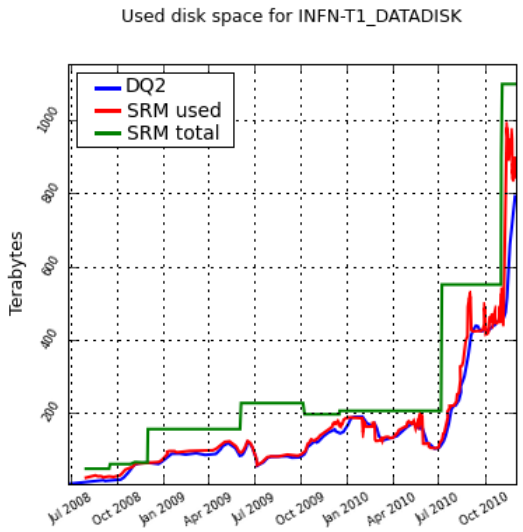


Some Tier 1 sites – Close to Full

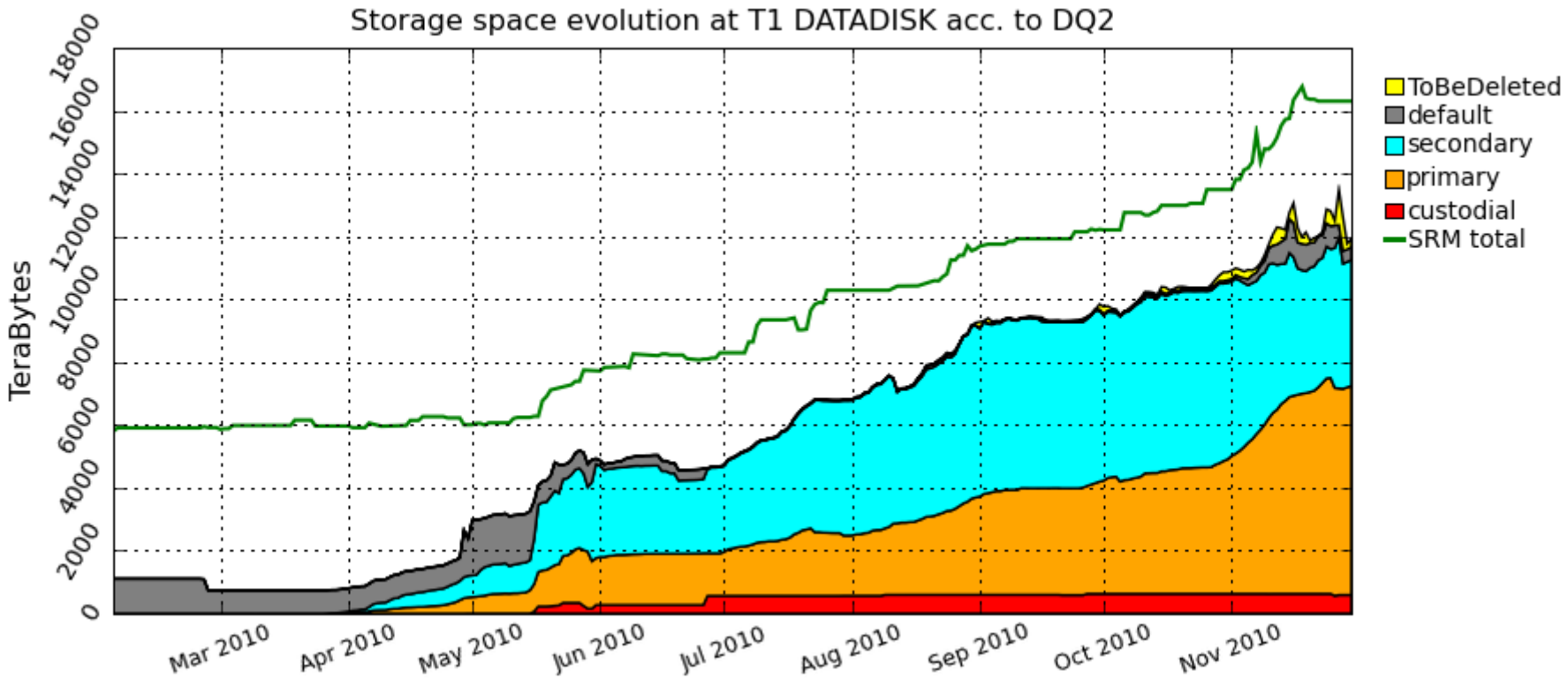








What Can be Reduced?



A Possible Solution – PD2P for Tier 1's

- Step 1: no change to custodial data placement
  - But only TAPE copies should be labeled custodial
- Step 2: minimize the number of primary datasets
  - Primary datasets are still pushed automatically
    - Typically, primary data only goes to Tier 1 sites
    - Retain the flexibility to place primary data at Tier 2 sites
  - Always first copy should be primary, if it is not custodial
  - If data is kept only on disk, a second primary copy is made
- Step 3: secondary dataset replication is based on usage
  - Secondary replicas are made by PD2P based on analysis usage
  - Can be deleted by automatic deletion agent if storage is full
  - Still allow manually requested secondary datasets through DaTri

Policy for Secondary Replicas at Tier 1

- Secondary replicas are made by PanDA – usage based
  - Initial copies are made at Tier 2's (using current PD2P algorithm)
  - But always 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
- We plan to try this in January
  - Introduce gradually – users should not experience any slowdown