PD2P : PanDA Dynamic Data Placement for ATLAS

T. Maeno[1], K. De[2], S. Panitkin[1], for the ATLAS Collaboration

[1] Brookhaven National Laboratory, USA

[2] University of Texas at Arlington, USA

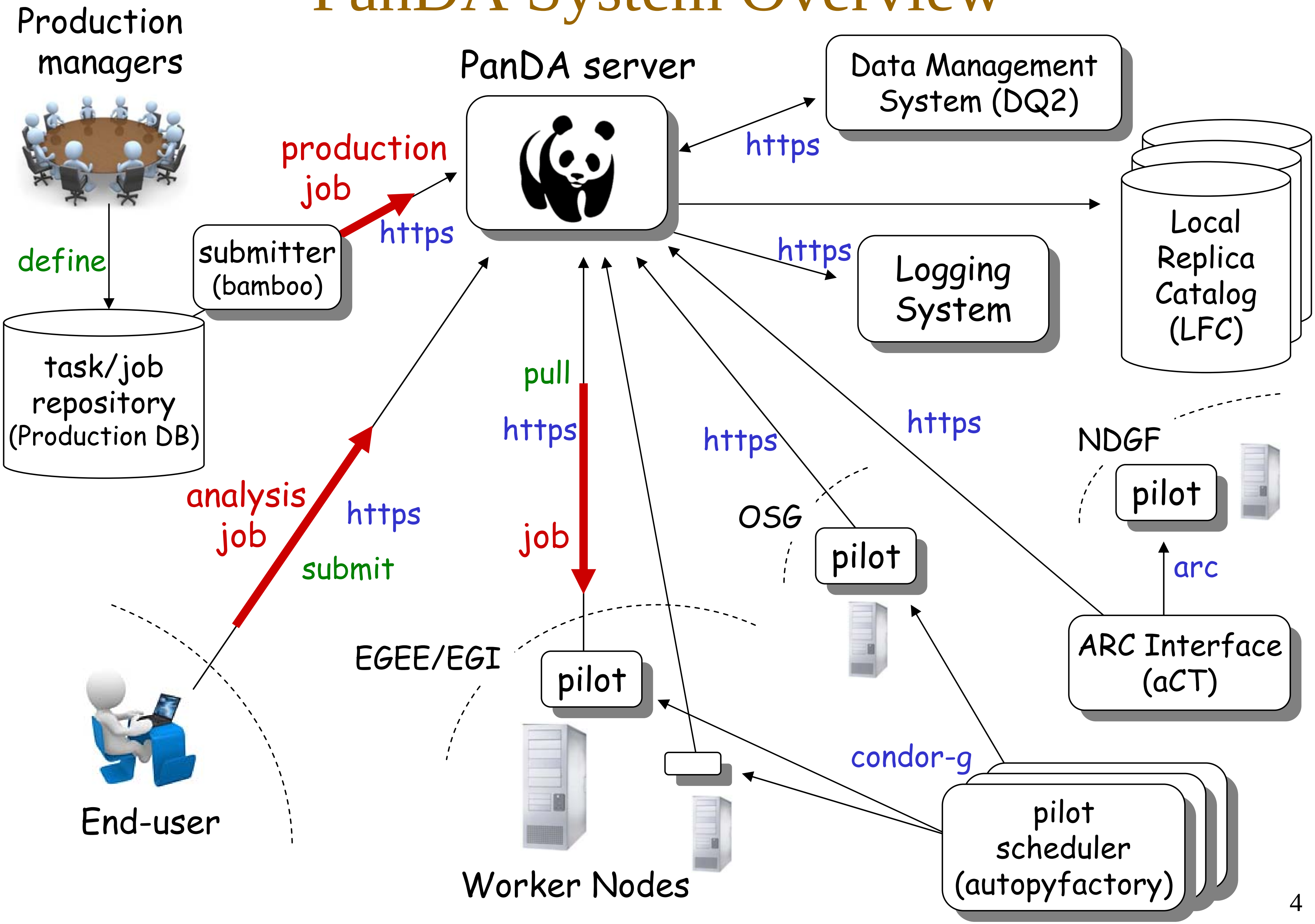
Outline

- Introduction
- Issues and Goals
- Implementation of PD2P
- Results
- Future Plans
- Conclusions

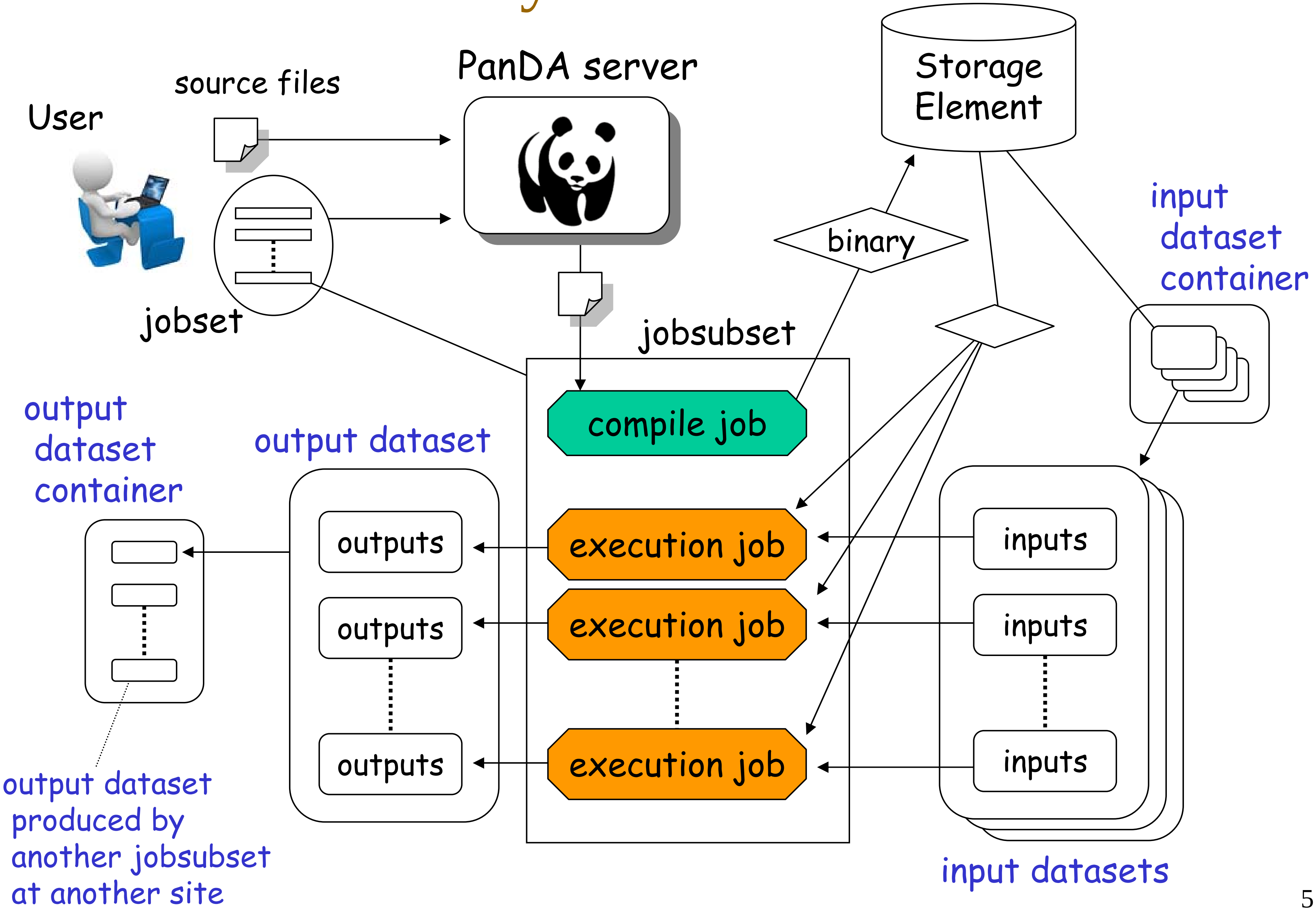
Introduction

- PanDA = Production and Distributed Analysis System
  - Designed to meet ATLAS production and analysis requirements for a data-driven workload management system capable of operating at LHC data processing scale
  - All Monte-Carlo simulation and data reprocessing jobs in addition to user and group analysis jobs
  - Working for 6 years including 2 years of LHC data-taking
  - 5 million jobs in total per week
  - More than 1400 users submitted analysis jobs in 2011
- PD2P = PanDA Dynamic Data Placement
  - An intelligent subsystem of PanDA to automatically replicate data
  - Developed to cope with difficulties of data placement for ATLAS

PanDA System Overview



Analysis Workflow



Analysis Workflow (cntd)

- Job set → job subsets → jobs
- One job subset per site
  - Jobs in each job subset are routed to one site
- Matchmaking per site
  - Scratch disk size on WN
  - Memory size on WN
  - Software availability
  - Downtime
  - Job statistics
  - Availability of input data
    - Jobs read or transfer input files from local storage elements at sites

Issues

- ~100 sites are available to users
  - Analysis jobs can run at Tier-0, Tier-1, and Tier-2 sites
  - Non-uniform distribution of CPU and storage resources
- ATLAS is producing huge amounts of data
  - Limited storage capacities at sites → No sites can hold all data → Non-uniform data distribution
- Analysis jobs go to data
  - Tight correlation between data and job distribution
- Policy-based dataset distribution
  - Two copies of data pushed to Tier-2 sites following the ATLAS computing model
  - ATLAS had used it and found some drawbacks
    - Tier-2 sites had been filling up too rapidly
    - Most data copied to Tier-2 sites were never used
    - Job distribution was unbalanced
- Unpredictable and evolving usage pattern of data
  - Each user has their own analysis requirements which are being changed rapidly for various analysis use cases
- Data distribution must be dynamically optimized
  - to fit the non-uniform CPU/storage resource distribution
  - to meet rapidly evolving user requirements

Goals

- Machinery for request based data replication → PD2P
  - Data replications are triggered by user requests to run jobs
- Many replicas for popular data
  - The number of data replicas is defined based on usage
- No delay for users
  - Users should not experience delay due to data movement
- Transparent changes to users
  - Any changes should not affect on-going critical analysis activities

Implementation of PD2P

General Policies

- Two algorithms
  - One for Tier-1 and the other for Tier-2
    - Tier-1 sites are used as data repository while Tier-2 sites are used more for execution of analysis jobs
- PD2P considers only official datasets
  - Users can submit jobs with private data but those data are not replicated by PD2P
- Replication policies for data types are defined by the ATLAS computing model
- Copies are made at sites where enough disk space is available
- PD2P is triggered only when users submit jobs to sites that are dedicated to analysis activities
  - E.g., production or site testing jobs are ignored
- Only online sites and data replicas at online sites are used. Also each Tier-2 site can be configured as to whether or not PD2P is used
  - E.g., once HammerCloud or SiteStatusBoard blacklists sites based on site status and/or downtime, PD2P doesn't use those sites or replicas at those sites

Tier-1 Algorithm (1/2)

- Primary copies of ATLAS data are placed at Tier-1 sites based on the Memorandum of Understanding (MoU) share
  - MoU share specifies the contributions expected from the corresponding region
- PD2P makes secondary copies at Tier-1 sites when
  - PD2P didn't make a replica of the data during the past week to a Tier-1 site
  - The number of data replicas at Tier-1 sites is less than int(log10(*Nused*))
    - *Nused* = how many times the data was used per job set
  - *Nused* is 10 to the power of X, where X is an integer larger than 0
    - i.e., *Nused* = 10,100,1000,…

Tier-1 Algorithm (2/2)

- One Tier-1 site is selected based on MoU share
- Replication request is sent to ATLAS Distributed Data Management (DDM) system
- When a copy is made at a Tier-1 site, another copy is made at a Tier-2 site at the same time
  - One Tier-2 site is selected based on MoU share
  - To have popular data not only at Tier1 sites but also at Tier2 sites

Tier-2 Algorithm (1/2)

- Executed independently of the Tier-1 algorithm
- PD2P makes two additional copies at Tier-2 sites when
  - There is no replica within Tier-2 sites, or not enough replicas are available while many jobs are waiting in the queue
  - The number of the data replicas at Tier-2 sites is less than 5
  - No more than two copies are concurrently being replicated

Tier-2 Algorithm (2/2)

- One Tier-2 site is selected based on MoU share
- The other Tier-2 site is selected by using
    - *$List_d$*: a list of Tier-1 and Tier-2 sites where the input dataset is already available
    - *$List_c$* : a list of Tier-2 sites that have fast FTS channels to one of the sites in *$List_d$*
    - *W* : the weight per site which is calculated using the number of active WNs at the site, site reliability, job statistics at the site, and the total number of replicas made by PD2P at the site for the last 24 hours
  - The weight *W* is calculated for each site in *$List_c$* and the Tier-2 with the largest *W* is used
- Two copies
  - To have one copy quickly available at a reliable site while distributing another copy, even if slowly, by following MoU share

Rebrokerage

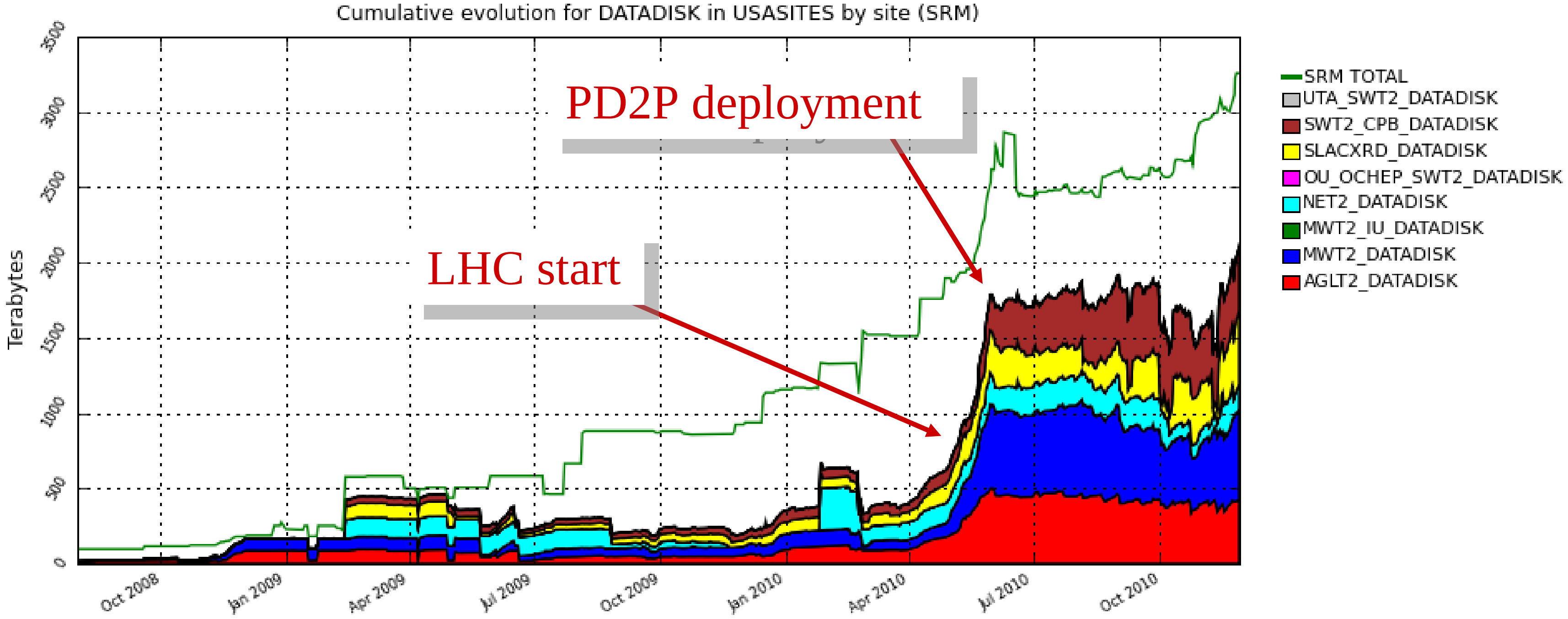
- PD2P relies on future reuse of data for its effectiveness
  - The data copy triggered by the initial job is not used unless subsequent jobs reuse it
  - The initial job remains at the original site although a new copy was replicated at free sites by PD2P
- The rebrokerage mechanism
  - Periodically reassigns jobs to other sites if they are waiting in the queue for a while
  - To increase reuse of PD2P replicas

Deletion

- DDM takes care of data deletion
  - Once access to a data replica has decreased to zero the replica gets deleted
- The Detail of the DDM deletion service is described in Garonne V., *The ATLAS Distributed Data Management project: Past and Future*, CHEP2012

Results

Cumulative Evolution of Data Flow to US Tier-2 Sites

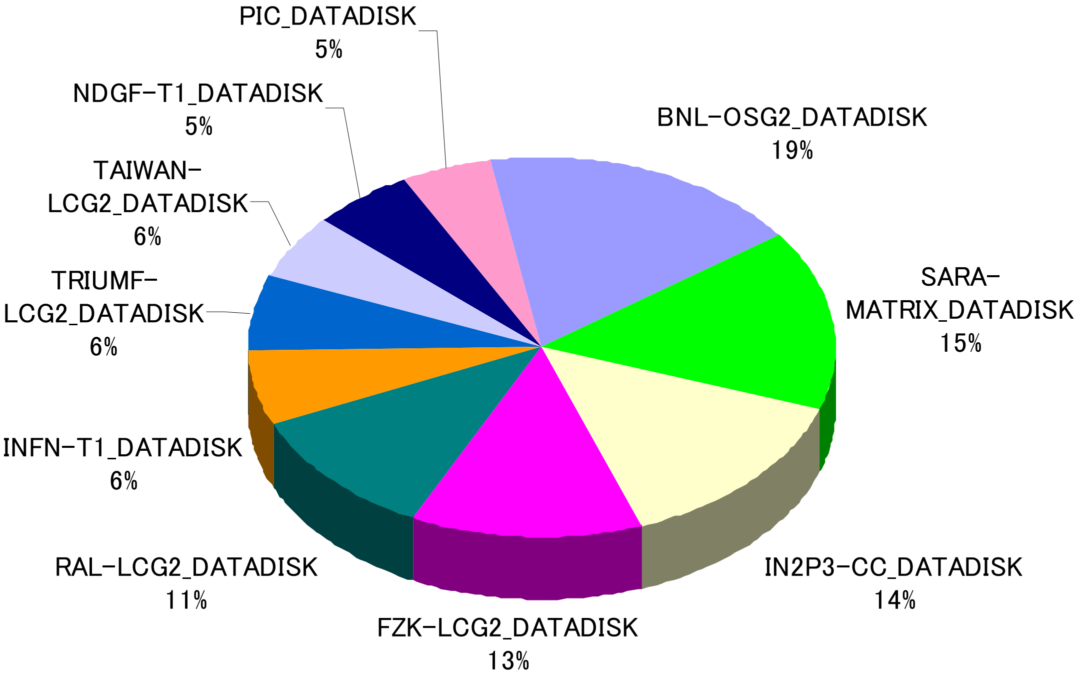


PD2P deployment

LHC start

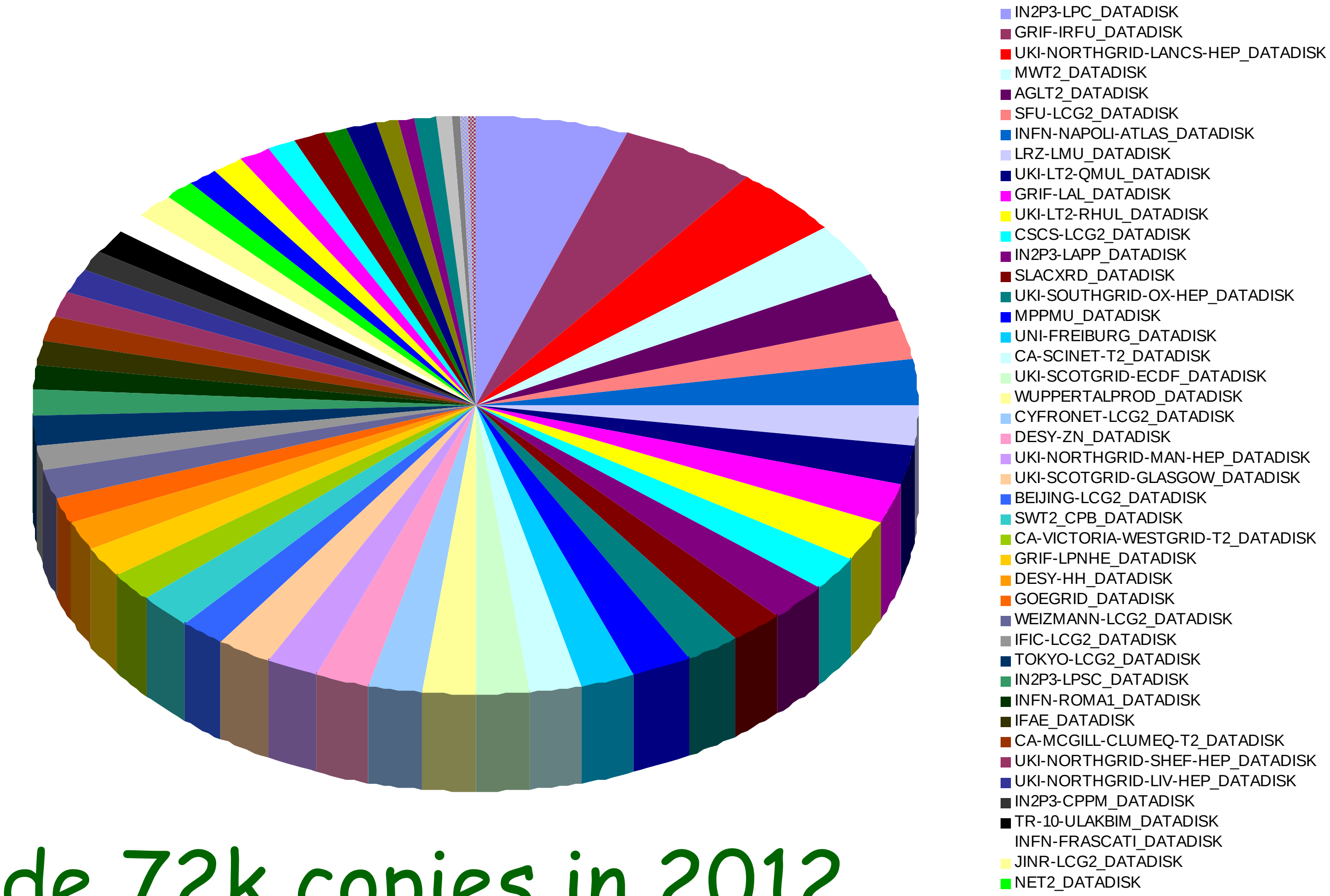
- A large improvement in terms of disk usage efficiency and manageability over the policy-based data distribution model
- Data placement policy has since evolved to a hybrid combination of PD2P based automated replication and limited policy-based distribution of data known to be most popular
  - To rapidly engage Tier-2 sites fully in the analysis of new data

Distribution of PD2P Tier-1 Copies



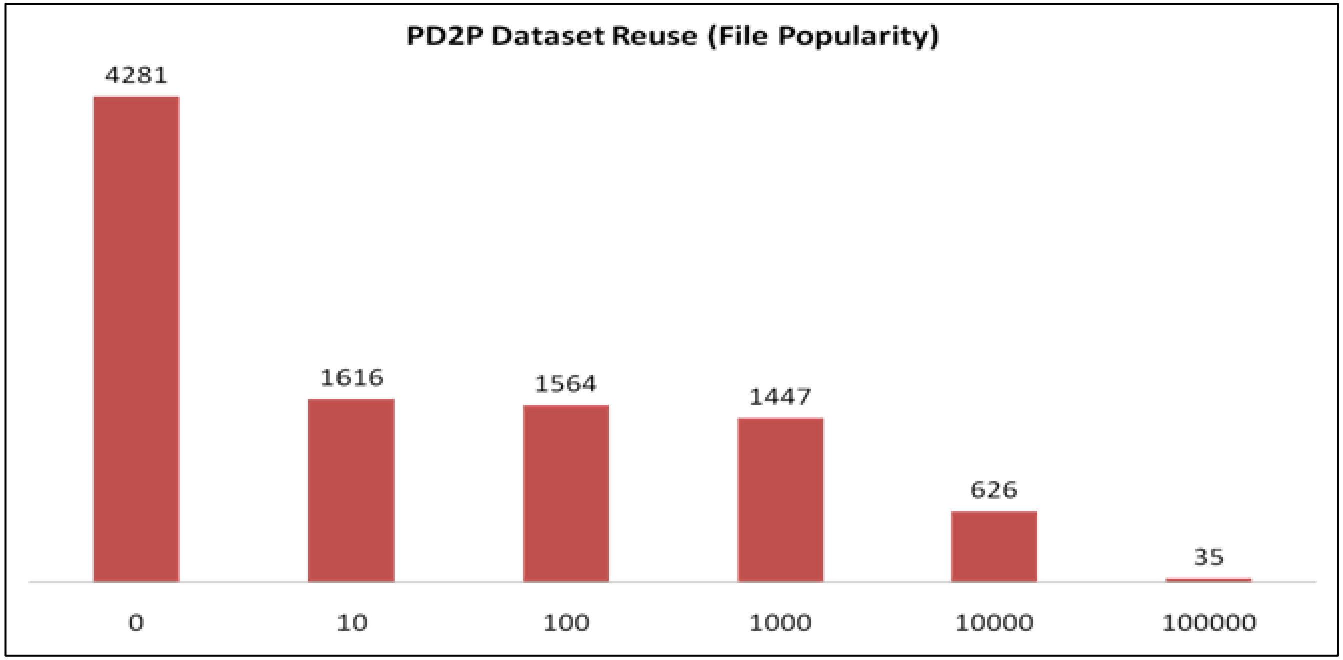
- PD2P made 9k copies in 2012
- Roughly proportional to MoU share, as expected

Distribution of PD2P Tier-2 Copies



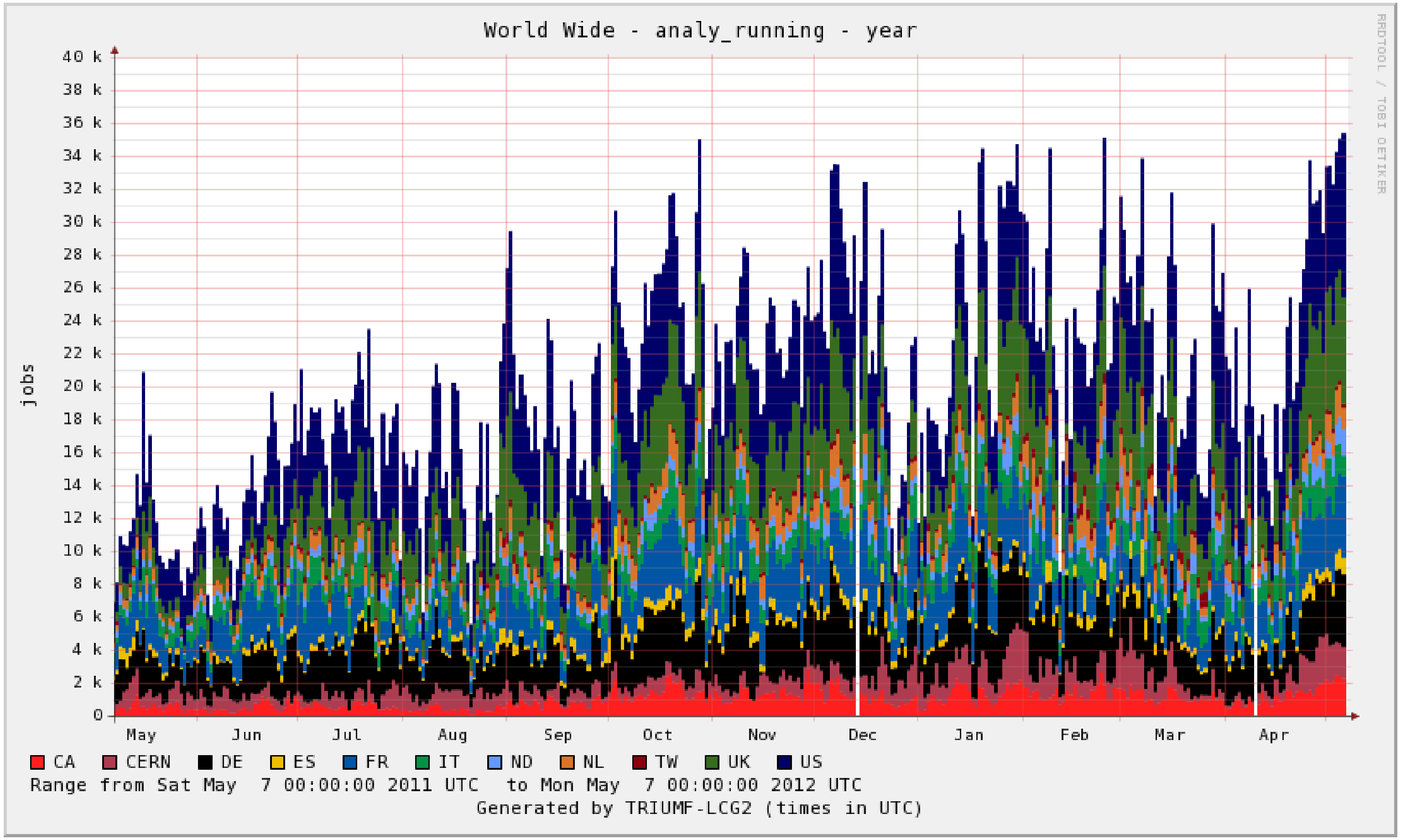
- Made 72k copies in 2012
- Well balanced
  - Broad usage of Tier-2 resources

Frequency of Reuse for PD2P Data



- 45% of data were never reused while others were well reused
- Users' interest is largely unpredictable and volatile
  - The request based model of PD2P suits user behaviour

Analysis Jobs Concurrently Running ’11 –’12



- Steady increase of analysis activities

Future Plans and Conclusions

Future Plans

- PD2P performs course-grained caching
  - Replicates a set of data files (dataset)
  - PD2P uses Tier-1 sites as primary data repositories while using Tier-2 sites as temporary storage for cache data
  - With PD2P, caching is at the dataset level
- ATLAS intends to extend caching scheme with more fine-grained approaches
  - The file level caching with the federated xrootd system
  - Below the file level, with advantage of a ROOT-based caching mechanism such as TTreeCache
  - Improvement of PanDA's brokerage to have cache-awareness

Conclusions

- The PD2P system has been developed to cope with the challenges of data placement to effectively serve the ATLAS analysis community amid limited storage and processing resources
- PD2P shows a large improvement in terms of disk usage efficiency, while distributing ATLAS data (and thereby analysis processing) at Tier-1 and Tier-2 sites broadly
- ATLAS's caching scheme will be extended from PD2P's dataset level to more fine-grained levels in the future, based on storage and I/O developments presently underway