

# Wide Area Network Data Access Requirements in Analysis

Doug Benjamin  
Duke University



# Setting the stage



- Since the beginning of the year, how much analysis was done?
- Why Jan 1 to May 1?
  - has analysis ramp up ahead of major Winter conferences (Moriond series)
  - Post conference activity perhaps indicative of steady state activity well into LS1 period
- What data do analyzers use?



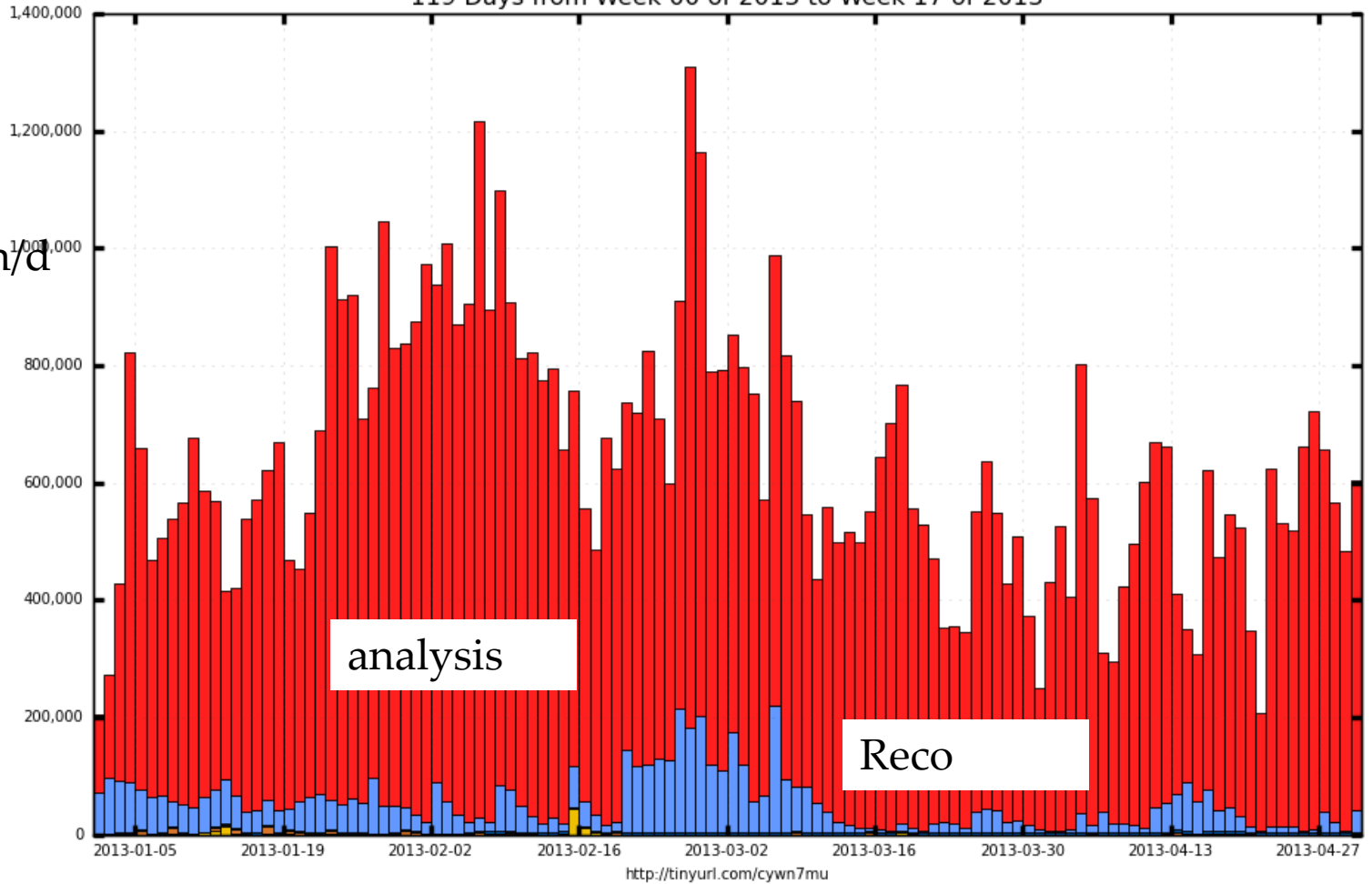
# Completed Jobs (1-Jan to 1 May 2013)



1 million/day

dashboard

Completed jobs  
119 Days from Week 00 of 2013 to Week 17 of 2013

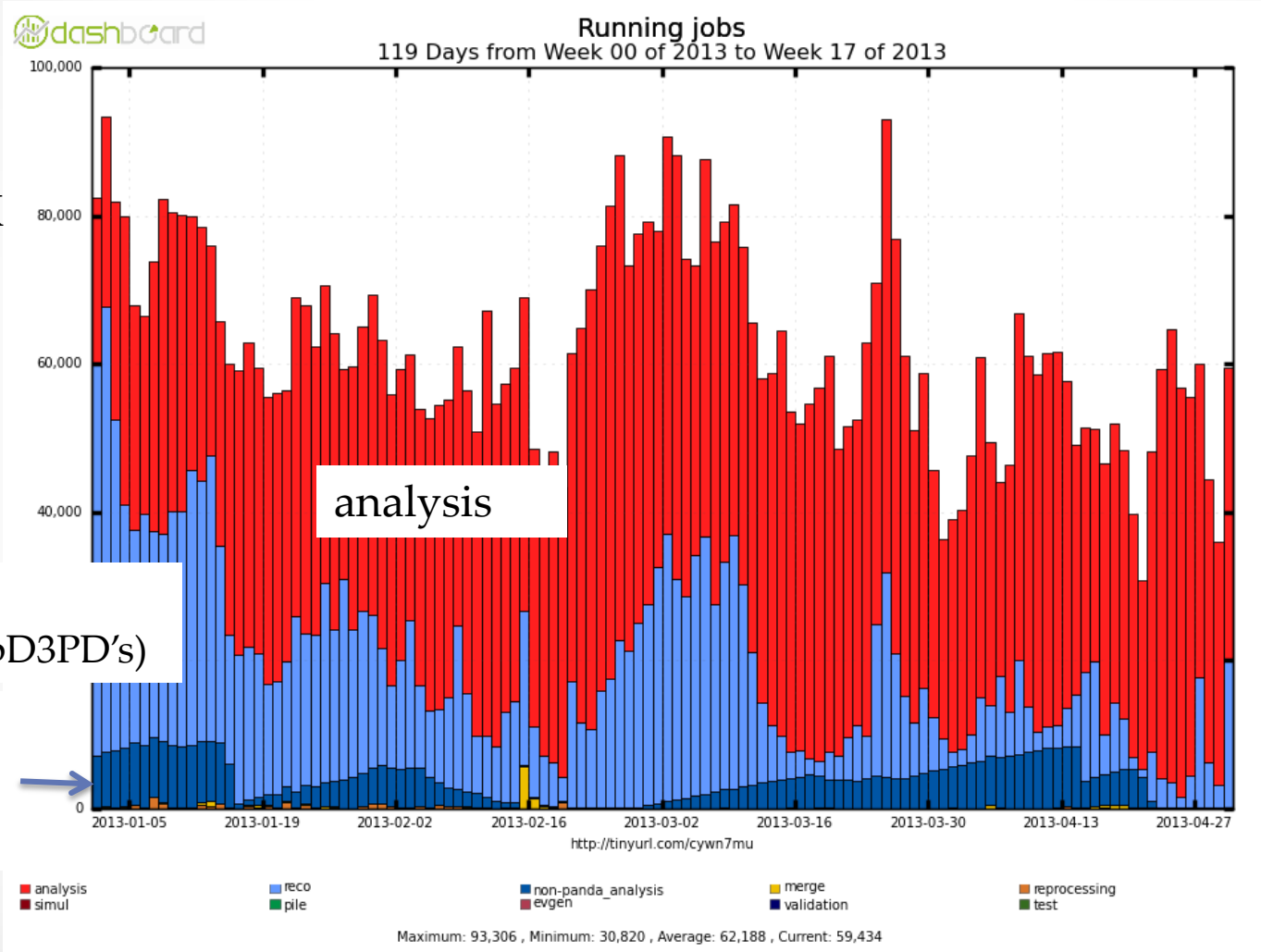


- analysis
- reco
- non-panda\_analysis
- reprocessing
- merge
- pile
- simul
- validation
- evgen
- test

Job Type - Group Analysis, Group production, User analysis



# User Analysis jobs 1-Jan – 1-May 2013



S

Job Type - Group Analysis, Group production, User analysis



# Data popularity



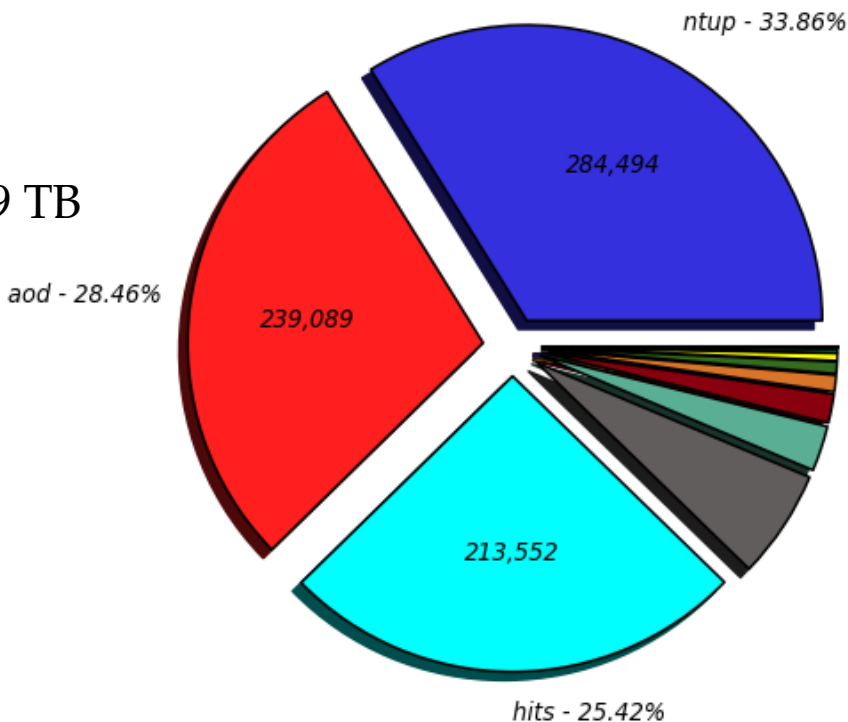
(used datasets 2012-09-22 to 2013-03-21)



Relative share of number of TB in used datasets (2012-09-22 - 2013-03-21) (Sum: 840,241)

AOD  
239,089 TB

NTUP  
284,494 TB



AOD's used  
to make  
NTUP's

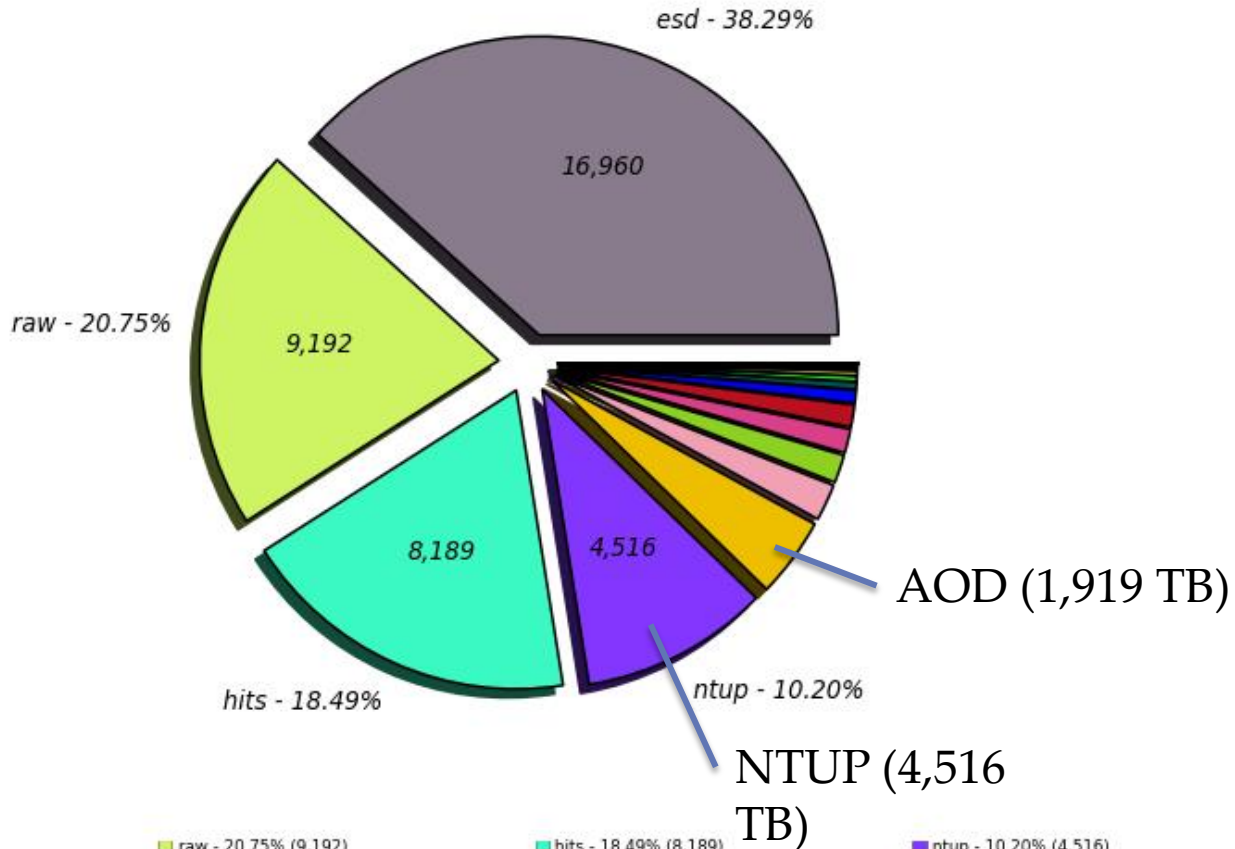
- ntup - 33.86% (284,494)
- raw - 2.48% (20,820)
- evnt - 0.40% (3,387)
- d2aod - 0.02% (178.00)
- txt - 0.00% (1.00)
- pool - 0.00% (0.00)
- mcprod - 0.00% (0.00)
- muonalignment - 0.00% (0.00)
- aod - 28.46% (239,090)
- desd - 1.58% (13,272)
- rdo - 0.21% (1,774)
- log - 0.01% (57.00)
- other - 0.00% (0.00)
- atlasphysics - 0.00% (0.00)
- v3 - 0.00% (0.00)
- tagprod - 0.00% (0.00)
- hits - 25.42% (213,553)
- daod - 0.90% (7,544)
- hist - 0.05% (399.00)
- d2aodm - 0.00% (30.00)
- atlasproduction - 0.00% (0.00)
- atlasprod1 - 0.00% (0.00)
- muoncalibration - 0.00% (0.00)
- esd - 5.95% (49,963)
- desdm - 0.65% (5,466)
- draw - 0.03% (211.00)
- tag - 0.00% (2.00)
- gen - 0.00% (0.00)
- trigmc - 0.00% (0.00)
- evgentp4mc11 - 0.00% (0.00)



# Data unpopularity (Unused datasets)



Relative share of number of TB in unused datasets (2012-09-22 - 2013-03-21) (Sum: 44,306)



- esd - 38.29% (16,960)
- aod - 4.33% (1,919)
- log - 1.19% (528.00)
- daod - 0.26% (116.00)
- d2aod - 0.03% (14.00)
- d2esdm - 0.01% (5.00)
- pt20 - 0.00% (0.00)
- o1 - 0.00% (0.00)
- v15 - 0.00% (0.00)
- v15000000 - 0.00% (0.00)

- raw - 20.75% (9,192)
- rdo - 1.98% (876.00)
- cbnt - 0.68% (302.00)
- evnt - 0.12% (52.00)
- tag - 0.02% (8.00)
- d2esd - 0.00% (2.00)
- atlasproduction - 0.00% (0.00)
- mbts - 0.00% (0.00)
- 15 - 0.00% (0.00)
- fit1iet - 0.00% (0.00)

- hits - 18.49% (8,189)
- desd - 1.52% (675.00)
- hist - 0.39% (175.00)
- dpd - 0.09% (40.00)
- gen - 0.02% (7.00)
- rpcwbeam - 0.00% (2.00)
- pool - 0.00% (0.00)
- v140500 - 0.00% (0.00)
- l1caloem - 0.00% (0.00)
- v2 - 0.00% (0.00)

- ntup - 10.20% (4,516)
- desdm - 1.29% (569.00)
- d2aodm - 0.29% (129.00)
- draw - 0.05% (23.00)
- other - 0.01% (5.00)
- v1451 - 0.00% (0.00)
- txt - 0.00% (0.00)
- tcwbeam - 0.00% (0.00)
- mcut - 0.00% (0.00)
- ... plus 86 more



# Interim summary



- 1-Jan to 1-May 2013 – peak of over 1M analysis jobs per day
- Prior to Moriond ~800 K jobs/day
- Post Moriond ~ 400 – 600 K jobs/day
- Jobs are typically short (most < 4 hours)
- In 6 month period End – Sep 2012 to end March 2013, NTUP's most popular data type 284 PB used (~ 67 PB/month)
- In same period - 4.5 PB NTUP's unused (~ 1.5%)

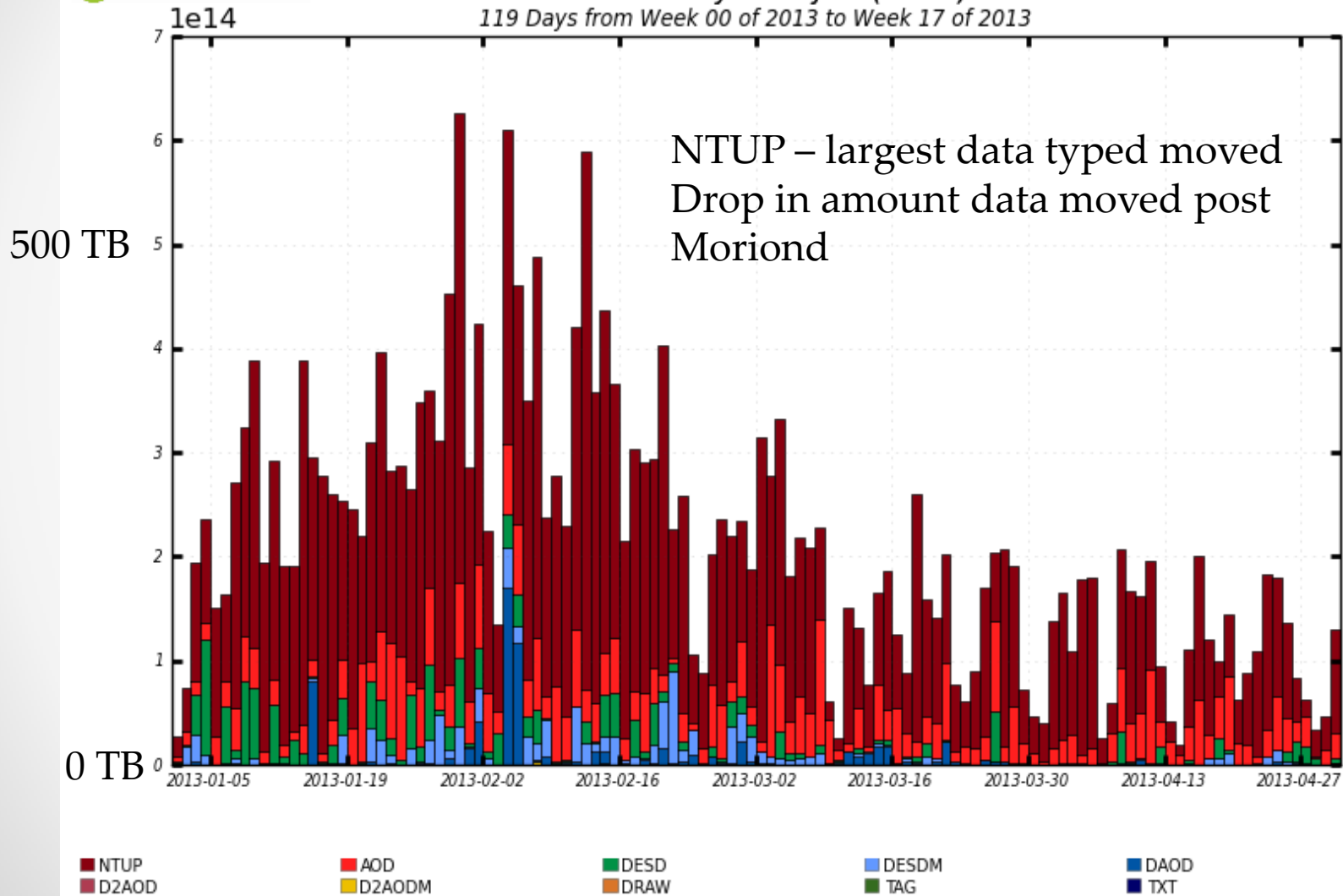


# Analysis triggered data motion (PD2P)



dashboard

Number of Physical Bytes (in TBs)  
119 Days from Week 00 of 2013 to Week 17 of 2013



Maximum: 625,444,452,253,292 , Minimum: 18,567,501,102,876 , Average: 213,946,526,963,683 , Current: 130,511,950,942,669



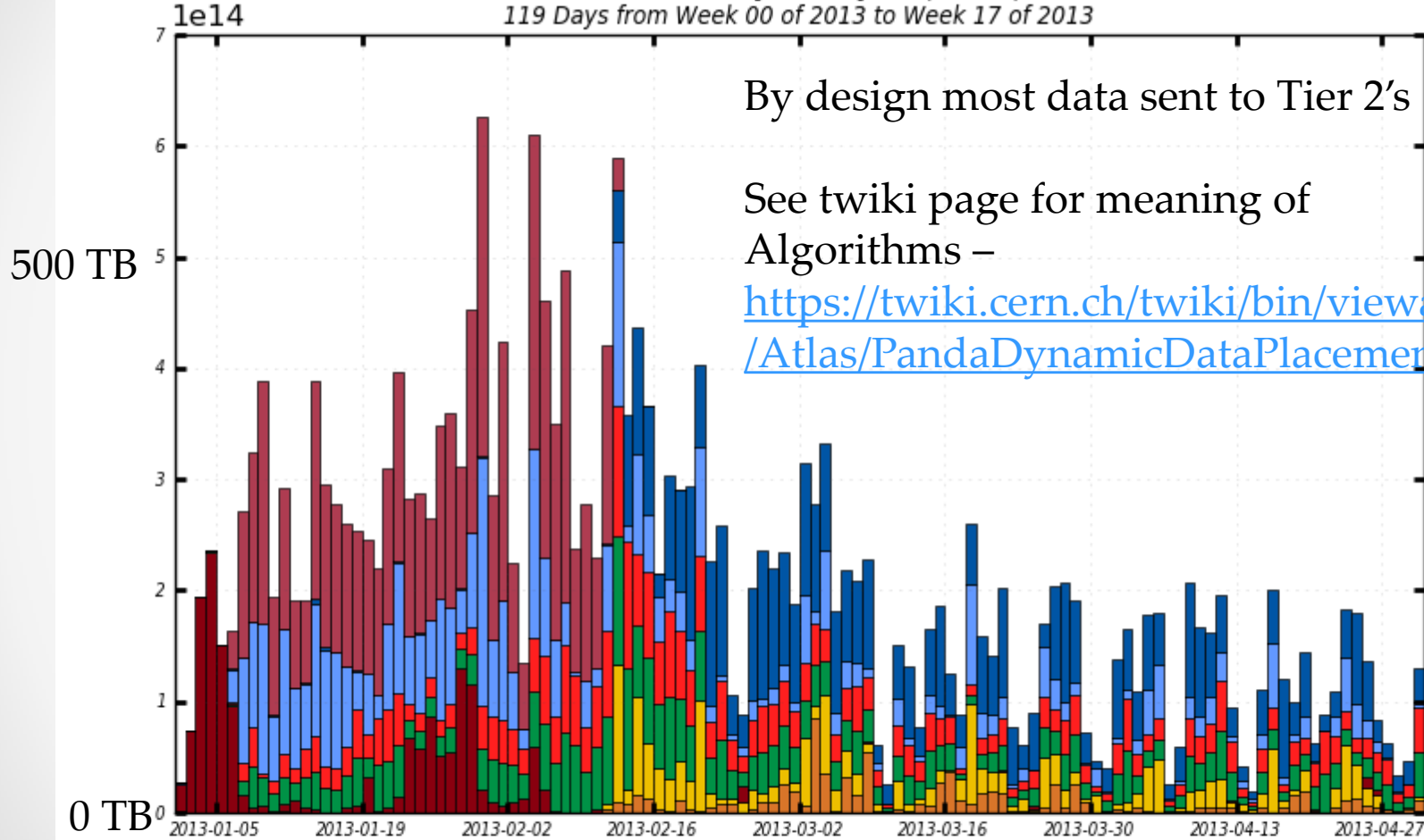


# PD2P by Algorithm



dashboard

Number of Physical Bytes (in TBs)  
119 Days from Week 00 of 2013 to Week 17 of 2013



By design most data sent to Tier 2's

See twiki page for meaning of Algorithms –

<https://twiki.cern.ch/twiki/bin/viewauth/Atlas/PandaDynamicDataPlacement>

■ SELECTEDT2	■ SELECTEDT2_NOREP	■ SELECTEDT2_T2MOU	■ SELECTEDT1	■ SELECTEDT2_T1MOU
■ unknown	■ SELECTEDT2_JOB	■ SELECTEDT2_WAIT		

Maximum: 625,444,452,253,292 , Minimum: 18,567,501,102,876 , Average: 213,946,526,963,683 , Current: 130,511,950,942,669

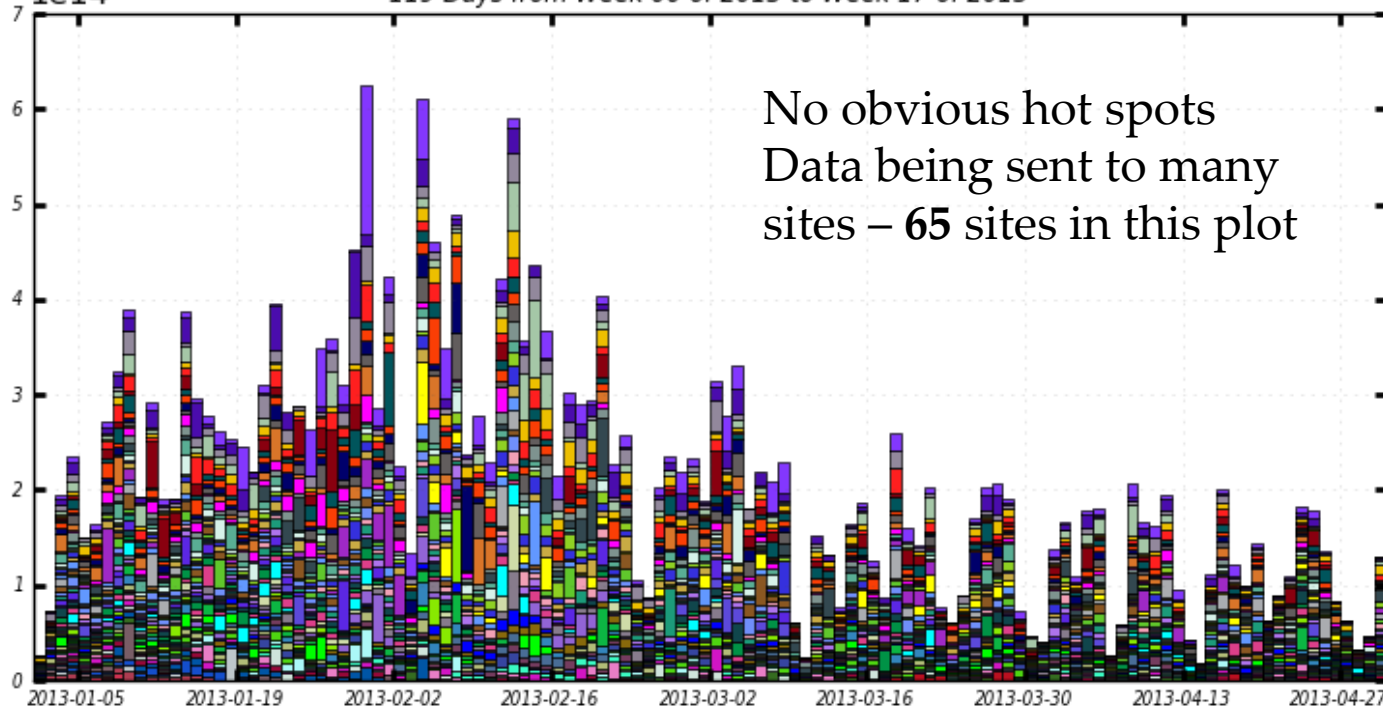


# PD2P sends data everywhere



1e14

Number of Physical Bytes (in TBs)  
119 Days from Week 00 of 2013 to Week 17 of 2013



No obvious hot spots  
Data being sent to many  
sites – 65 sites in this plot

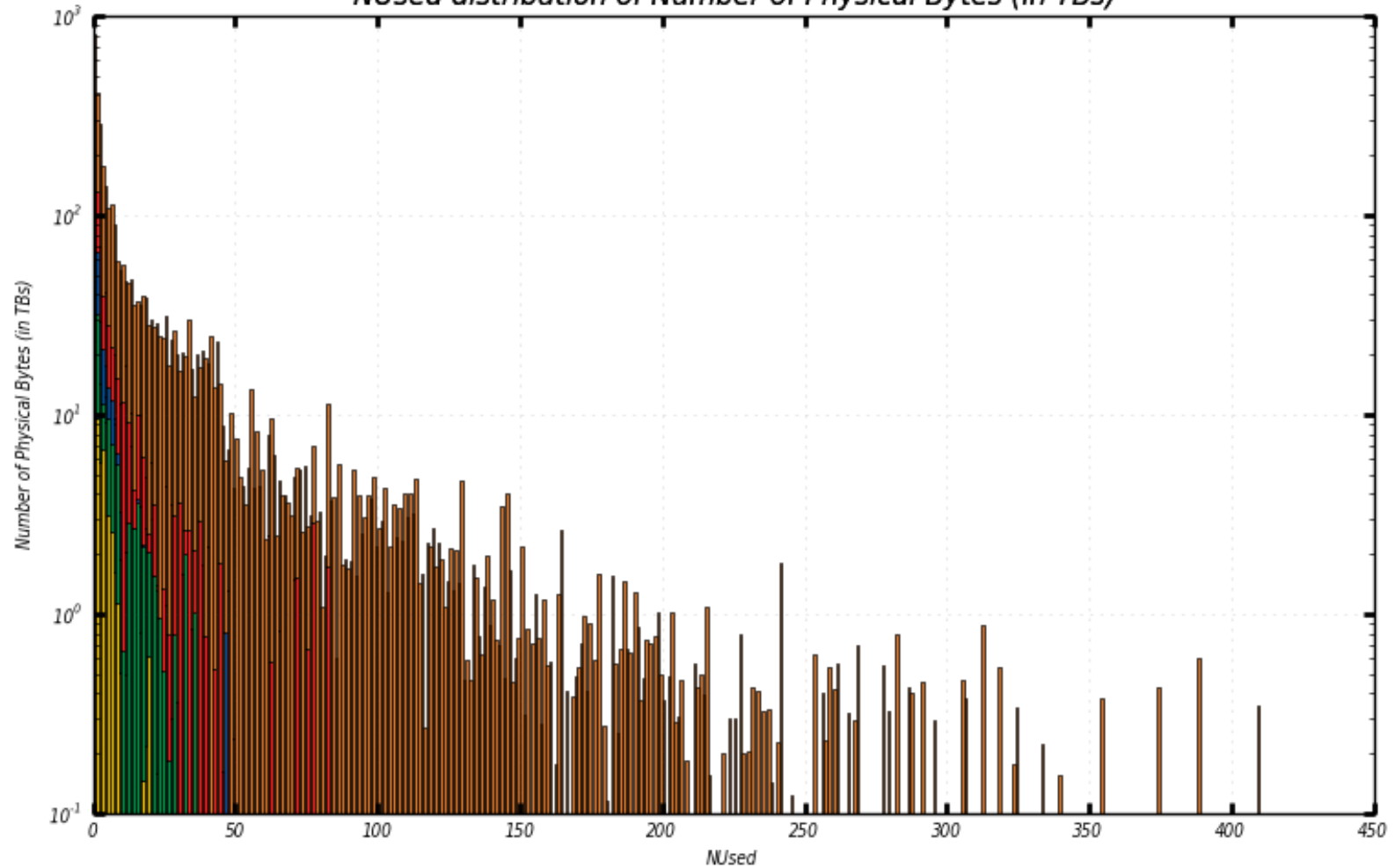
- |                                  |                                 |                                    |
|----------------------------------|---------------------------------|------------------------------------|
| ■ slacxrd_datadisk               | ■ uki-It2-qmul_datadisk         | ■ swt2_cpb_datadisk                |
| ■ wuppertalprod_datadisk         | ■ bnl-osg2_datadisk             | ■ aglt2_datadisk                   |
| ■ cscs-lcg2_datadisk             | ■ nwt2_datadisk                 | ■ tokyo-lcg2_datadisk              |
| ■ desy-hh_datadisk               | ■ desy-zh_datadisk              | ■ ca-victoria-westgrid-t2_datadisk |
| ■ grif-irfu_datadisk             | ■ sfu-lcg2_datadisk             | ■ uki-It2-rhul_datadisk            |
| ■ infn-napoli-atlas_datadisk     | ■ ru-protvino-ihep_datadisk     | ■ in2p3-cc_datadisk                |
| ■ in2p3-lpc_datadisk             | ■ ific-lcg2_datadisk            | ■ uki-scotgrid-glasgow_datadisk    |
| ■ goegrid_datadisk               | ■ ca-scinet-t2_datadisk         | ■ lrz-lmu_datadisk                 |
| ■ uki-northgrid-liv-hep_datadisk | ■ uki-southgrid-ox-hep_datadisk | ■ uki-scotgrid-ecdf_datadisk       |
| ■ triumph-lcg2_datadisk          | ■ infn-roma1_datadisk           | ... plus 36 more                   |



# Dataset reuse



NUSED distribution of Number of Physical Bytes (in TBs)



NTUP  
D2AOD

AOD  
DRAW

DESD  
TXT

DESDM

DAOD

Maximum: 814.96 , Minimum: 0.00 , Average: 13.92

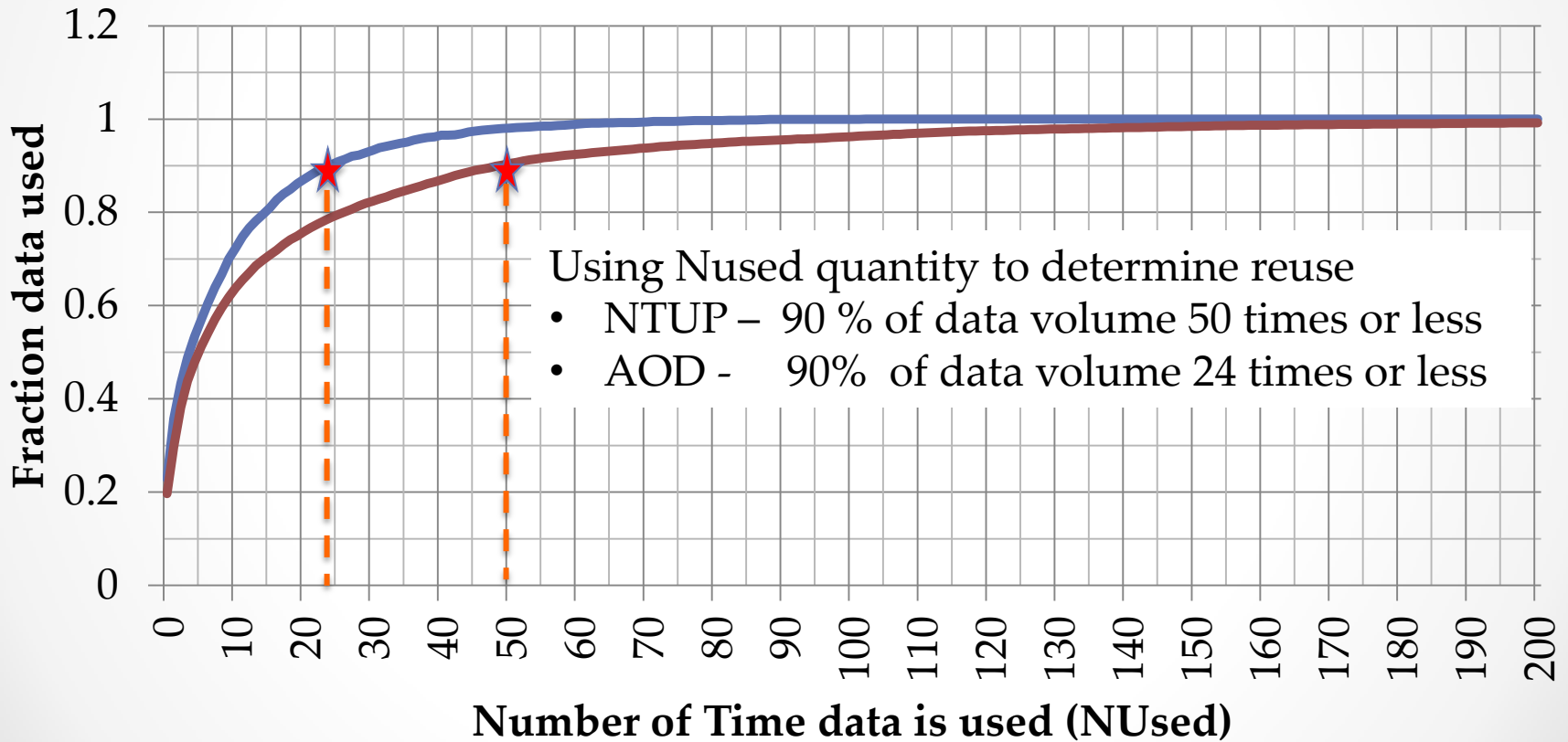


# Fraction of Data volume reused



Fraction Data used vs PD2P NUsed

— AOD — NTUP





# PD2P interim summary



- Jan-Apr 2013 – PD2P moved at peak ~ 600 TB/day
- NTUP most popular data type to move
  - Not a surprise given the popularity plots
- Most data moved to Tier 2 sites
- Data moved to a wide variety of sites (> 67 sites)
- ... something about dataset reuse....

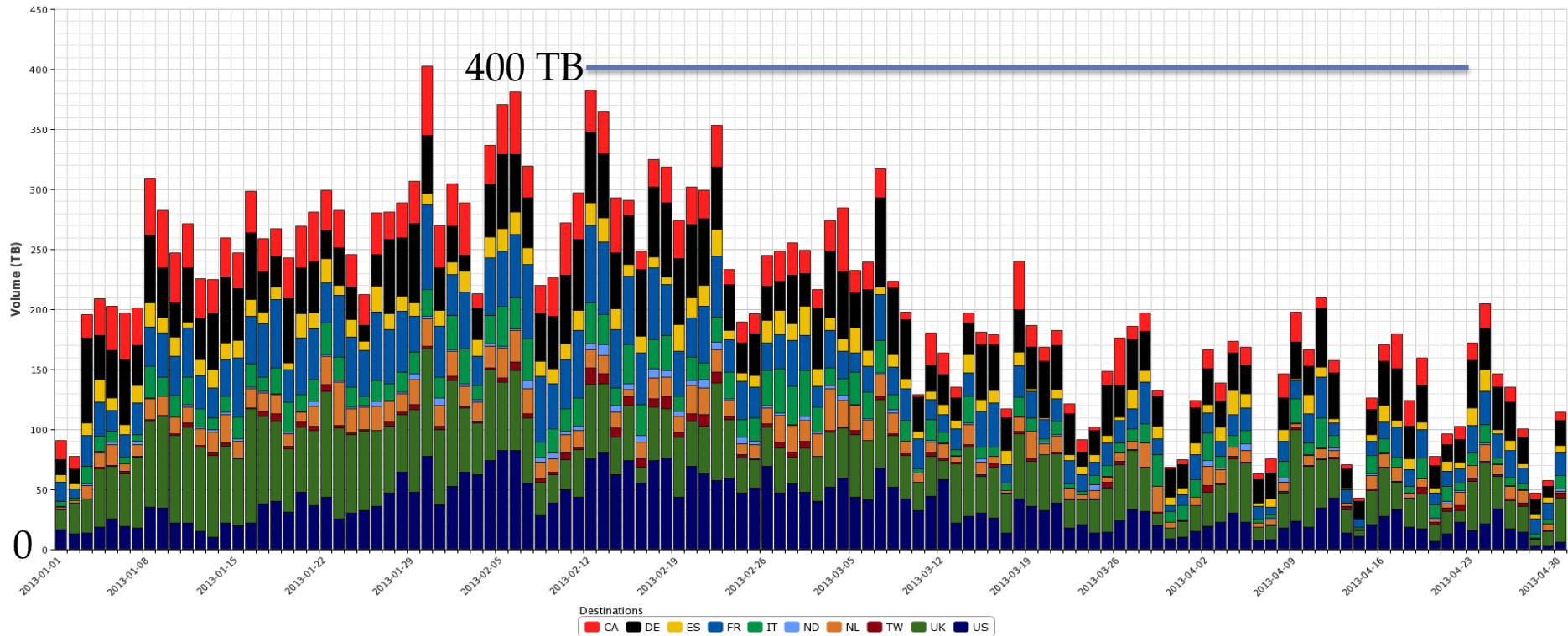


# Planned data replication



dashboard

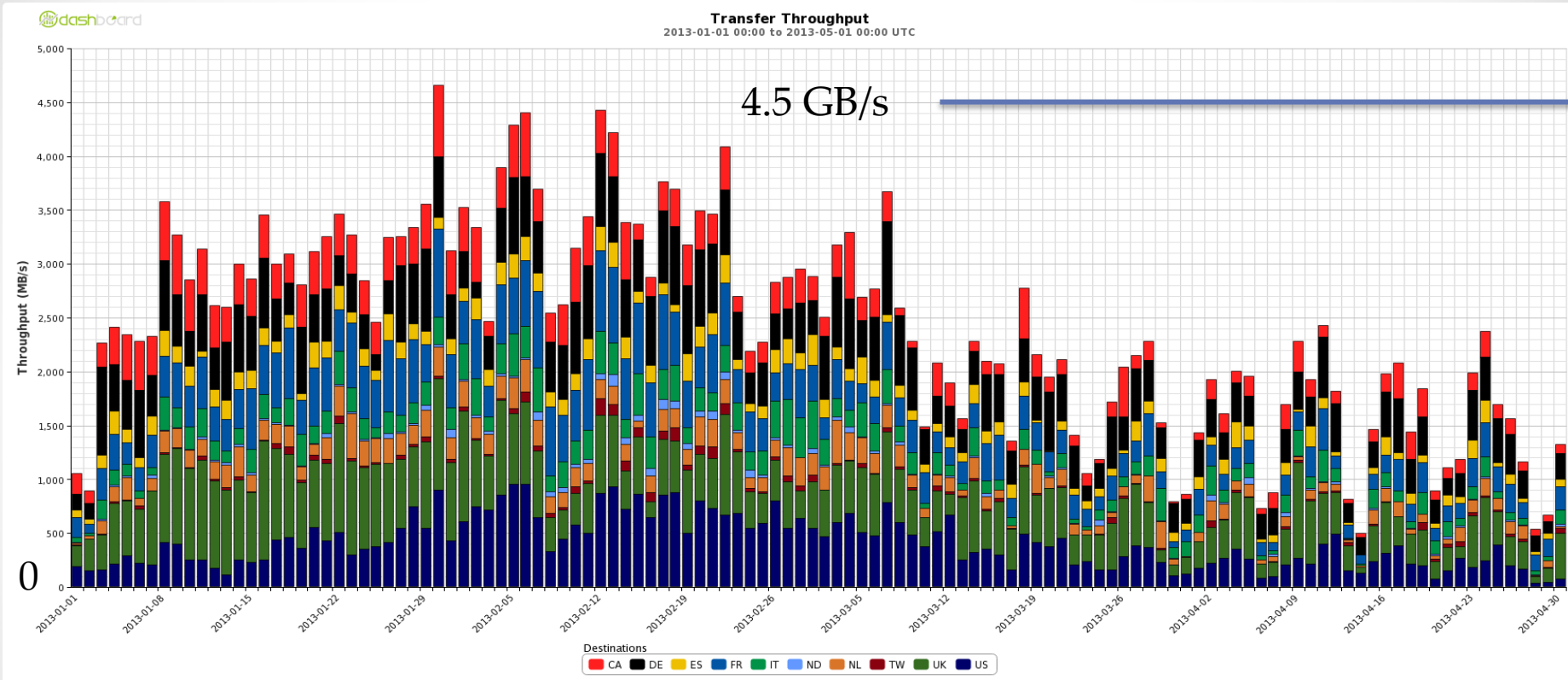
Transfer Volume  
2013-01-01 00:00 to 2013-05-01 00:00 UTC



DDM – Daily Data Brokering data Volume (1-Jan to 1-May 2013)



# Data Brokering transfer Rate



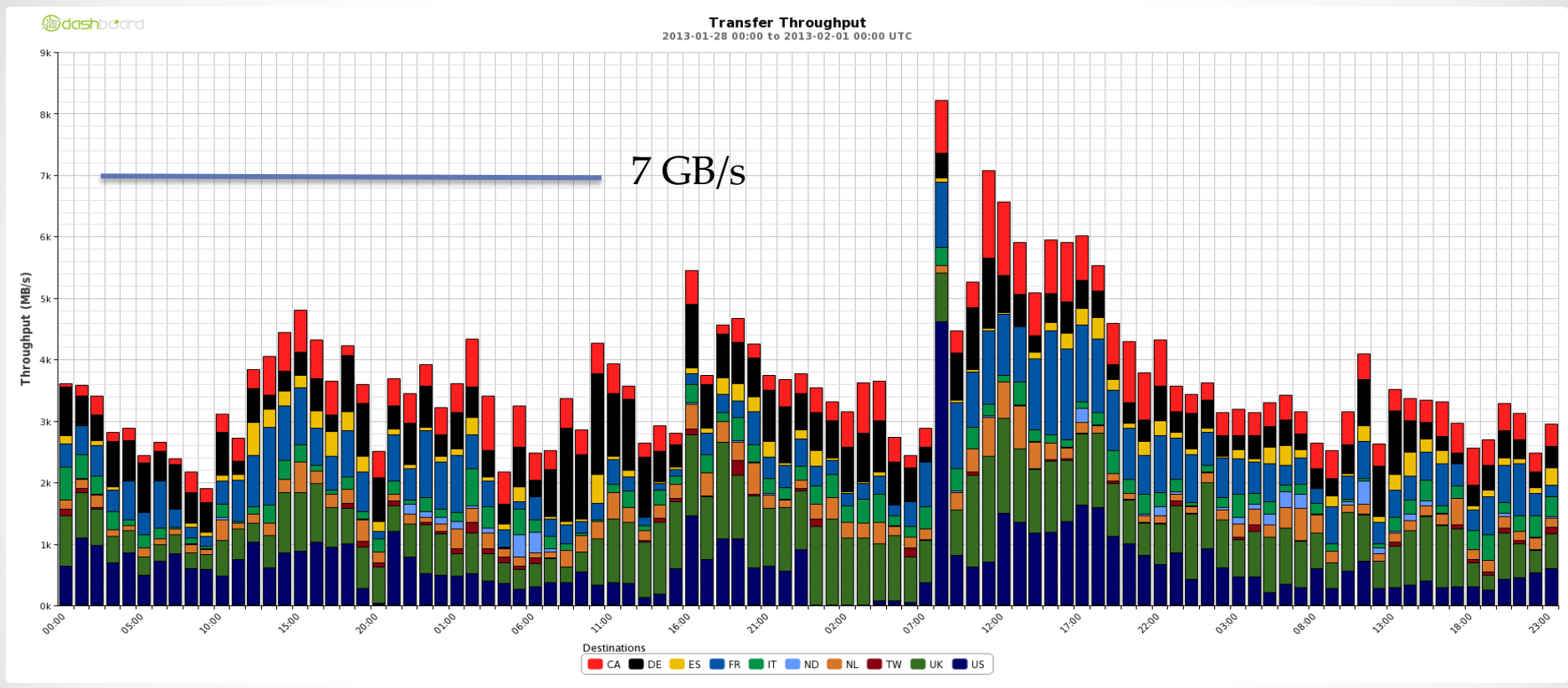
Transfer Throughput (averaged over a day) 1-Jan to 1-May 2013



# Data Brokerage transfer rate



## 28-Jan – 31-Jan (48 hrs)



Transfer Throughput (one hour bins) 28-Jan to 1-Feb 2013



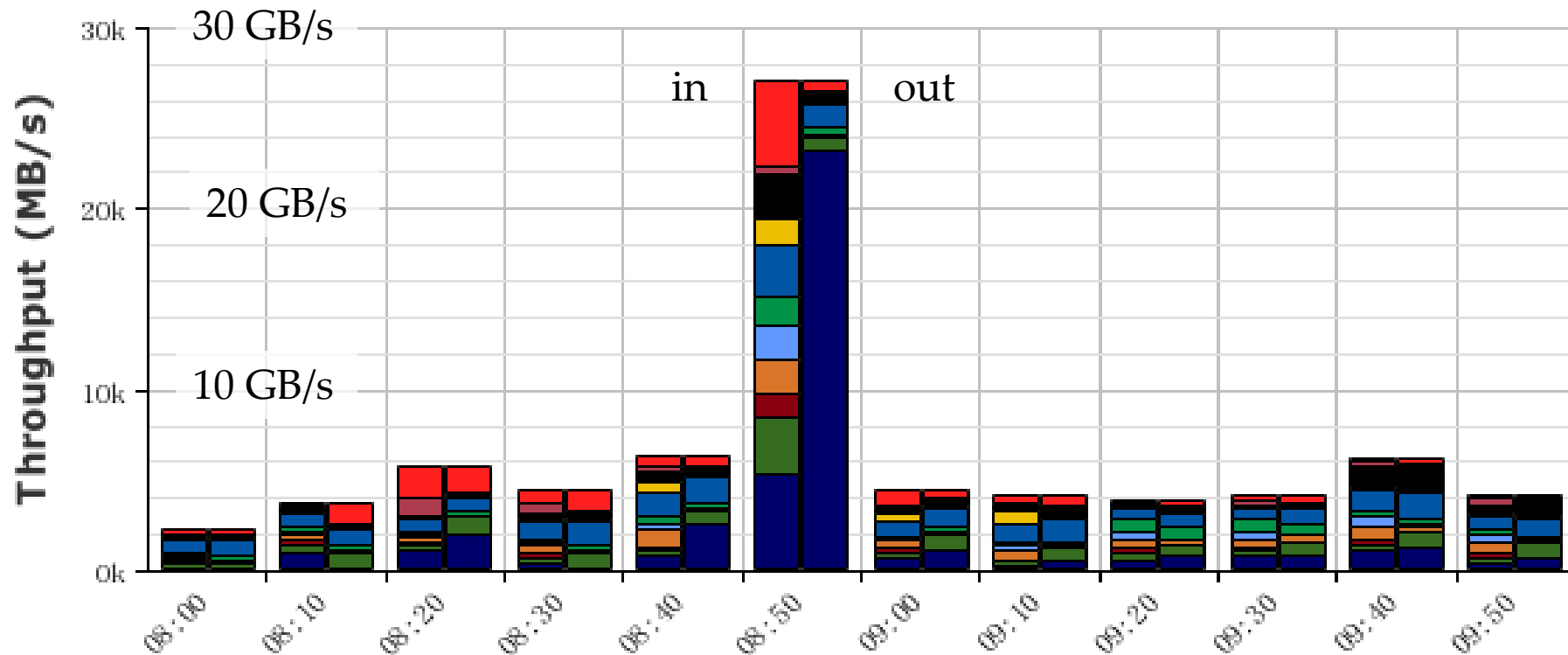


# Transfer rate zoom in further



## Transfer Throughput In/Out

2013-01-30 08:00 to 2013-01-30 10:00 UTC



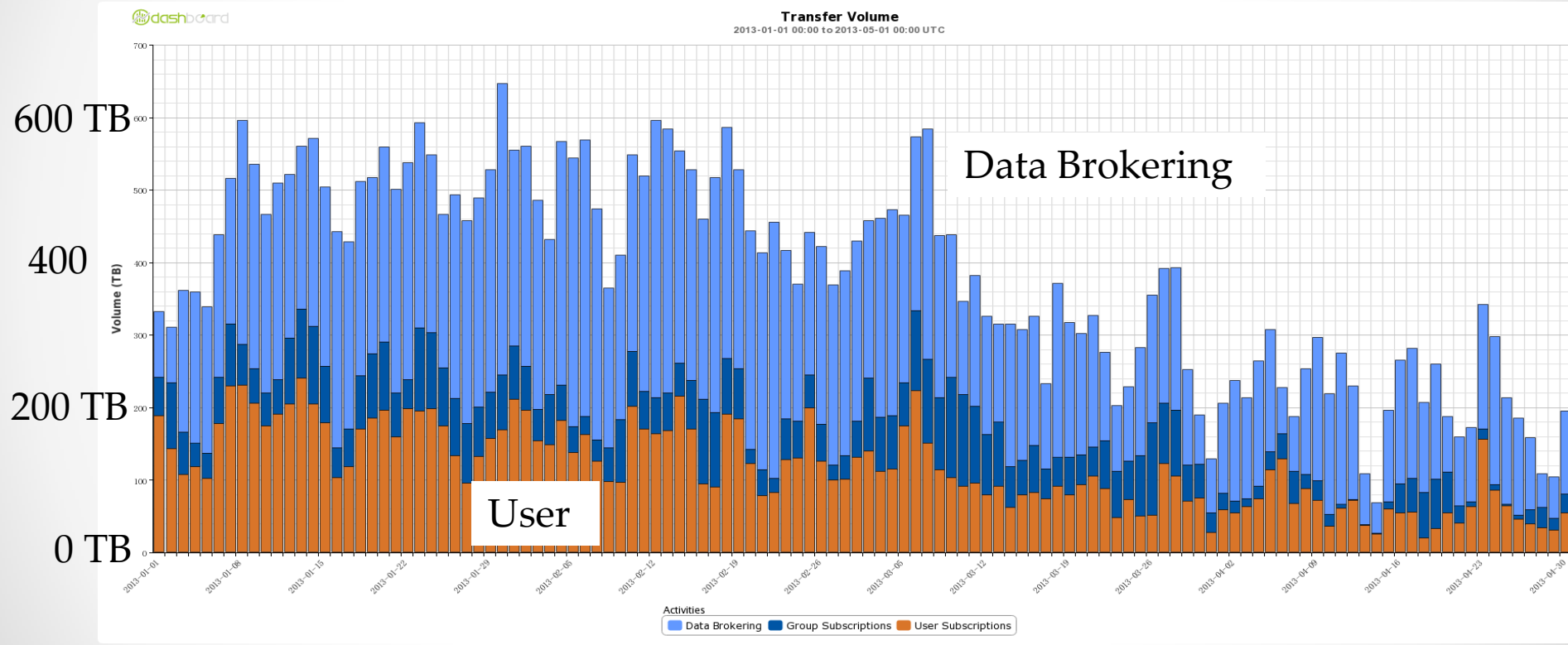
Sources / Destinations





# Planned transfers

## Data Brokering, Group Subscriptions, User subscriptions





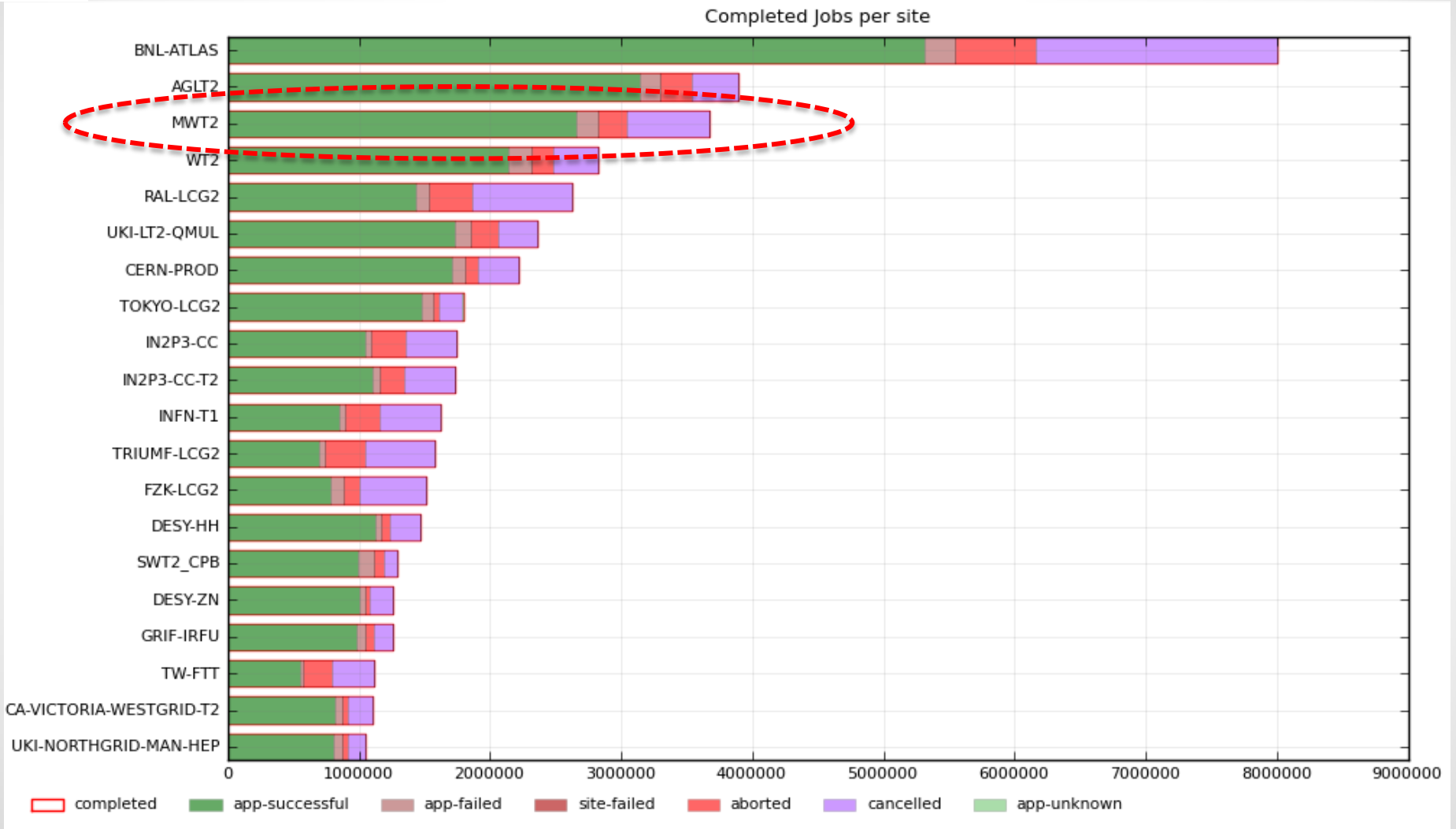
# Planned Data Brokerage summary



- Peak 400 TB/day and ~ 200 TB/day post Moriond
- Transfer rates
  - Peak 4.5 GB/s (day resolution)
  - During busiest few days 28-Jan through 31-Jan-13 (48 hrs)  
7 GB/s (one hour resolution)
  - Jan 30, 2013 (08:00 – 10:00) peak time, in Peak 10 minute period  
30 GB/s - reading mostly from US Cloud to rest of ATLAS  
~ 5 GB/s other 10 minute periods
- User data subscriptions  
150-200 TB/day pre Moriond  
50-100 TB/day post Moriond
- All networking plans should include accounting for User Output



# Analysis Sites



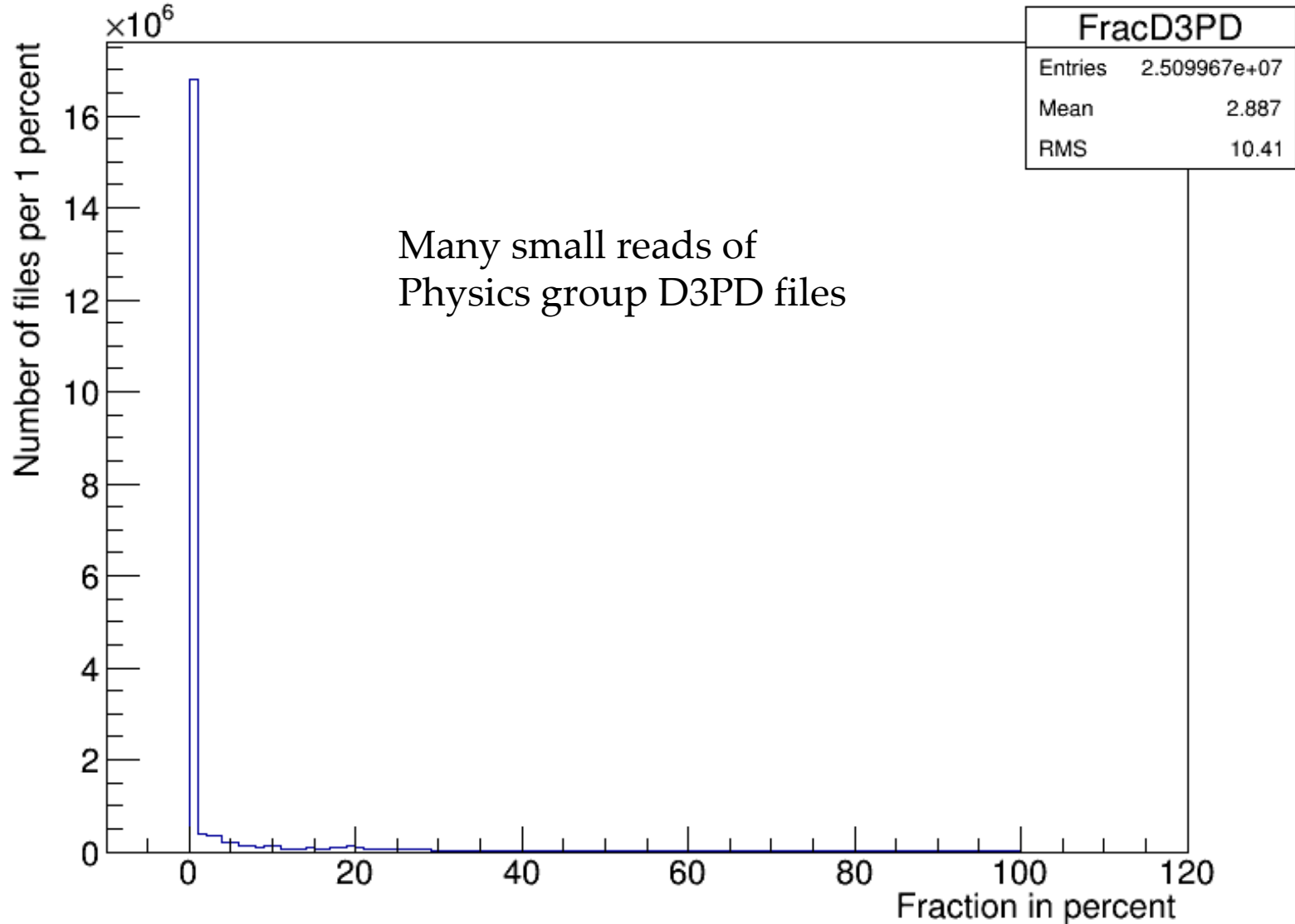
Jan 1 to April 30 2013 - MWT2 – direct read site



# How much of D3PD do users read?



Fraction of file read in User analysis job w/ group Ntuple

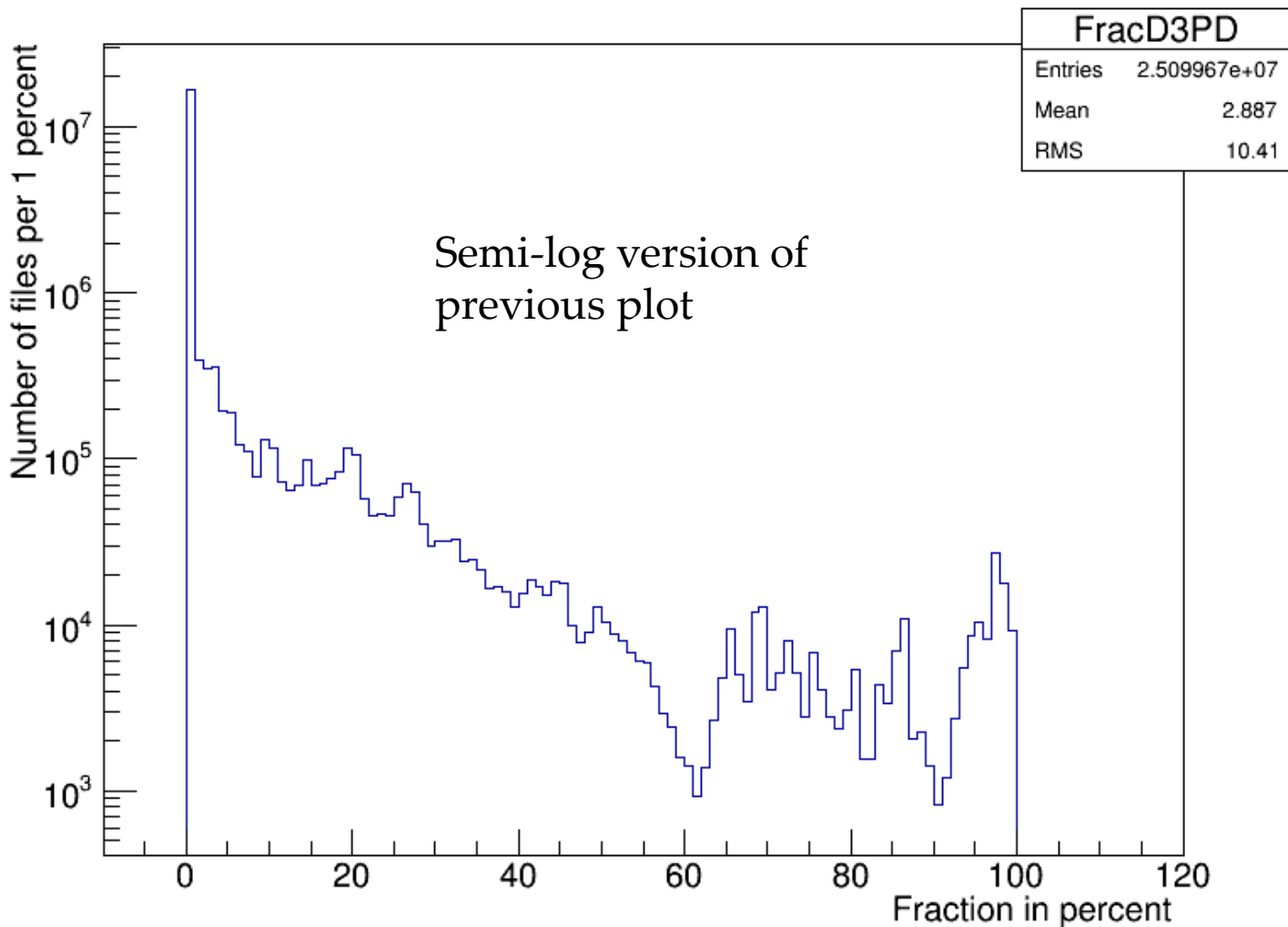




# How much of D3PD do users read?



Fraction of file read in User analysis job w/ group Ntuple

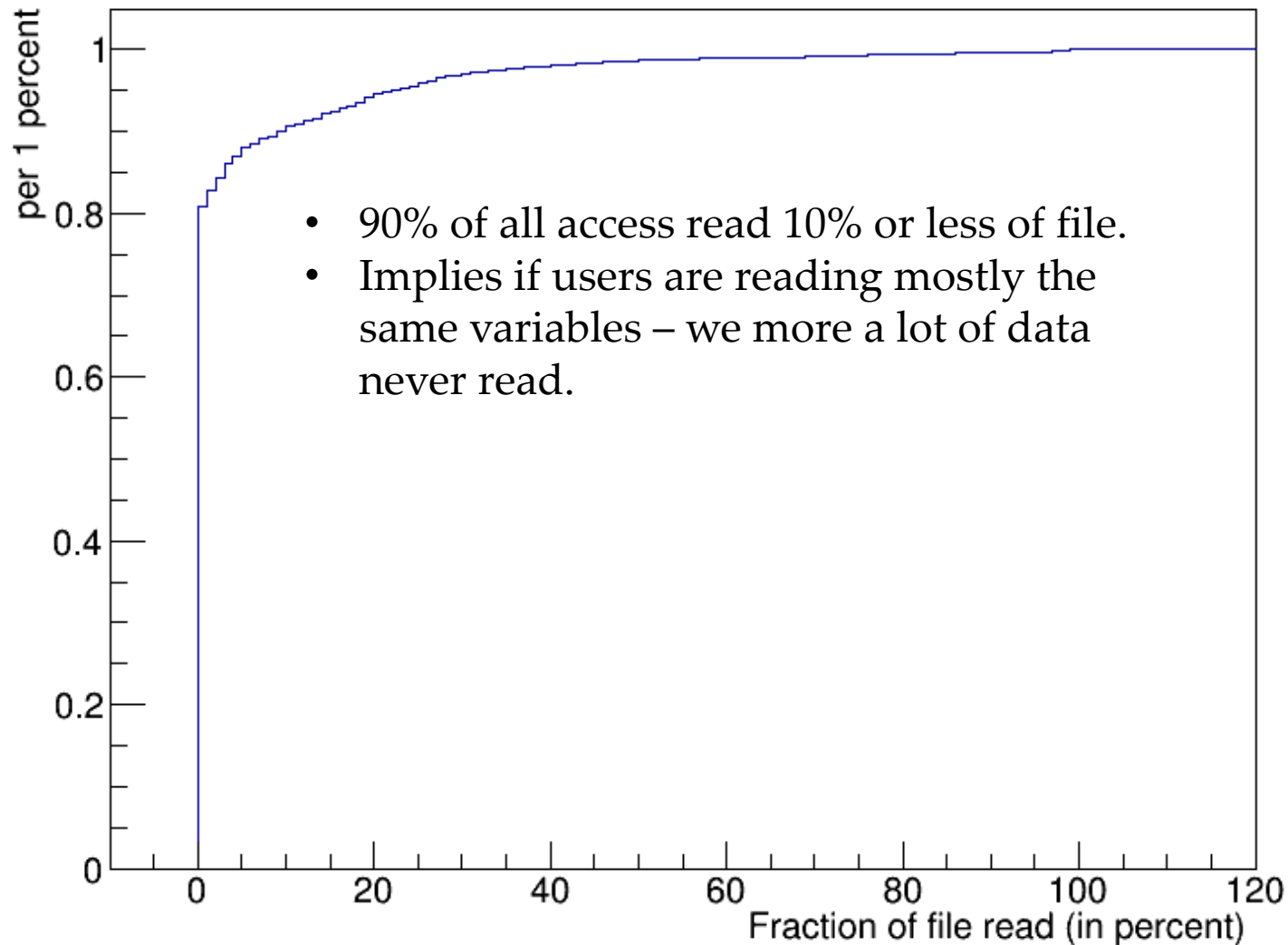




# How much of D3PD do users read?



Integral in User analysis job w/ group Ntuple





# Interim summary user access of D3PD's



- Current D3PD's from the Physics groups typically very large
- They contain more Branches than the users typically use
- New analysis model is working to merge AOD and D3PD's
  - Implication – input files will contain much more information that the users actually read
  - Efficient Skim/Slim centralized service will be crucial
- Need some mechanism for capturing what variables the users are really using and then provide them files with mostly those variables.





# Open questions needing further study



- How much PD2D data is read only once or twice after it has been replicated via PD2P?
- How long does a file stay popular (ie read frequently)?
  - We want to replicate the popular files and not the other ones?
- Can we reduce the amount of data in the Analysis files that is rarely read?
- Should we have caches for data files at the Tier 1 and Tier 2 sites ?
  - Do for analysis data what was done for database data and software releases (frontier/squid and cvmfs/squid).
  - What will it take to have partial event caches?
- Can we estimate the network growth needs when jobs and input data are in separate locations (WAN access from jobs)



# Conclusions



- During busy times – 1 Million user analysis and D3PD production jobs per day, user analysis jobs are short duration.
- PD2P moves at peak ~ 600 TB/day and planned replication of similar files ~ 400 TB/day, User subscriptions ~ 200 TB/day
- 90% of NTUP datasets (by volume) are read 50 times or less. (according to PD2P values)
- Users still read a small fraction of the centrally produced D3PD's (NTUP's). Implies much of the data moved is not read
- What network issues will be seen w/ factor of 3 increase in Trigger Rate in 2015-2016?