DDM operations

Tomáš Kouba
ADC Tier-1/Tier-2/Tier-3 Jamboree
2012/12/11

Outline

- Recent issues with full T1 datadisk
- Space tokens overview
  - changes and plans
- Automatic management procedures
  - cleaning, blacklisting
- Tape usage situation
- Lost files
- T2D status

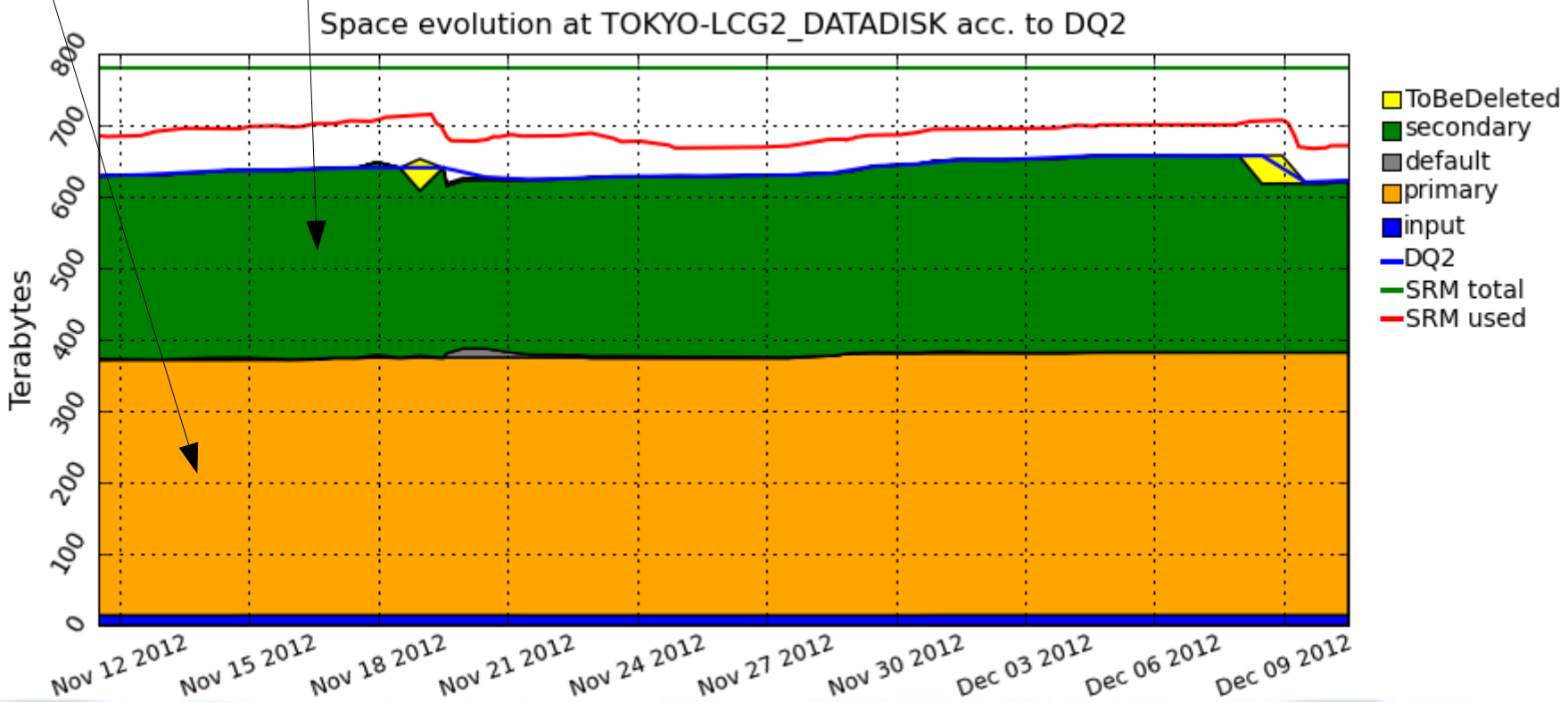
Recent issues with full T1 DATADISK

- During recent months DDM operations faces problems with full T1 DATADISKs
- Intermediate production output stays on T1 DATADISK, without a proper procedure for cleaning
  - Group production
    - a procedure and auto-cleaning have been set up and running
  - Official production
    - manual interventions to be automatized
  - Reprocessing
    - need to be done in an organized way
- Not a site issue but sites may be affected
- Thanks to T1s that provide more space than pledged and/or deployed space for 2013 already

Primary/secondary

Primary – data to be kept at the site

Secondary – extra data, can be deleted



Space tokens - overview

- ATLASDATADISK – ATLAS central data according to policies + on demand
- ATLASHOTDISK - files used by many jobs, replicated by the site within same SE (if needed)
- ATLASGROUPDISK - data managed by groups
- ATLASPRODDISK - buffer for central production
- ATLASSCRATCHDISK - temporary user data
- ATLASLOCALGROUPDISK – local user data
- ATLASDATATAPE – custodial RAW and derived data from T0 and reprocessing
- ATLASMCTAPE – custodial copy of MC data

DDM endpoints

- Entities used by DDM services
  - Often reported by shifters in tickets
- Defined in TiersOfAtlas (aka ToA)
  - Set of python-like variables
  - http://atlas.web.cern.ch/Atlas/GROUPS/DATABASE/project/ddm/releases/TiersOfATLASCache.py
  - since September 2012 ToA is generated from AGIS and AGIS Web UI is used for editing the data
- Usually <SITENAME>_<SPACETOKEN>
  - e.g. BNL-OSG2_DATADISK, FZK-LCG2_PRODDISK
  - Space token used without “ATLAS” prefix
- ToA defines what SE, space token and path is used for the endpoint:
  - IN2P3-CC_HOTDISK:
    `token:ATLASHOTDISK:srm://ccsrm.in2p3.fr:8443/srm/managerv2?SFN=/pnfs/in2p3.fr/data/atlas/atlashotdisk/`

DDM endpoints for phys groups

- For groups the name looks like <SITENAME>_<GROUPNAME>
  - e.g. `PRAGUELCG2_PHYS-HI`
- Group endpoints use ATLASGROUPDISK space token:
  - `token:ATLASGROUPDISK:srm://golias100.farm.particle.cz:8446/srm/managerv2?SFN=/dpm/farm.particle.cz/home/atlas/atlasgroupdisk/phys-hi/`
  - ToA defines quotas for every group, sum of endpoints quotas == size of ATLASGROUPDISK
- Every once and then sites were asked to resize ATLASGROUPDISK
- Free space cannot be used for other ADC activities
  - Not even temporarily
  - Groups do not replicate/delete popular data
  - Sites complain about unused space

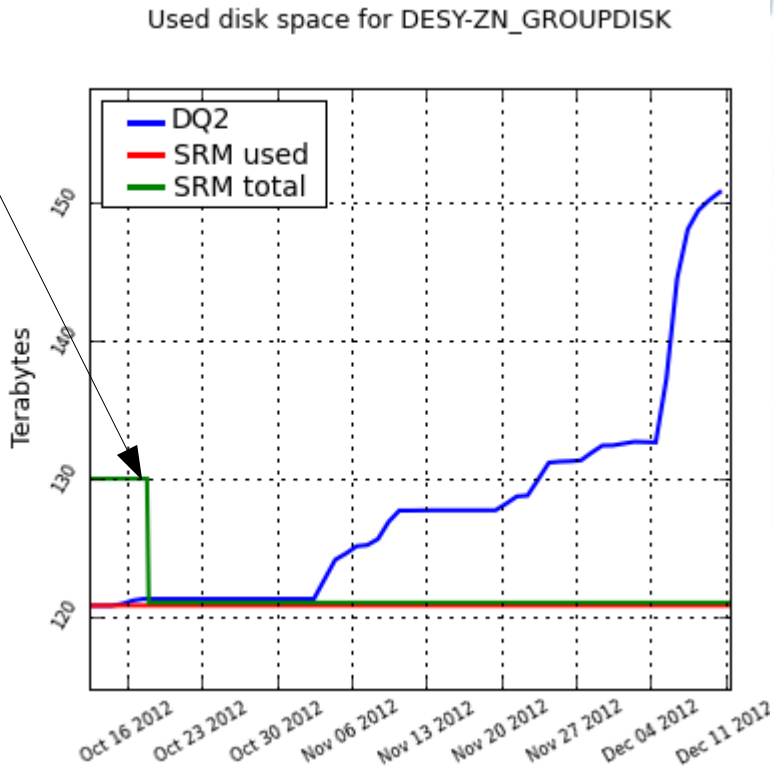
ATLASDATADISK for group endpoints

- We would like to use the underused group space for secondary replicas of datasets
- Can we change:
  - `token:ATLASGROUPDISK:srm://golias100.farm.particle.cz:8446/srm/managerv2?SFN=/dpm/farm.particle.cz/home/atlas/atlasgroupdisk/phys-hi/`
- To:
  - `token:ATLASDATADISK:srm://golias100.farm.particle.cz:8446/srm/managerv2?SFN=/dpm/farm.particle.cz/home/atlas/atlasgroupdisk/phys-hi/`
- ?
- We can if the space token is not bound to the namespace (depends on SE implementation of space tokens)

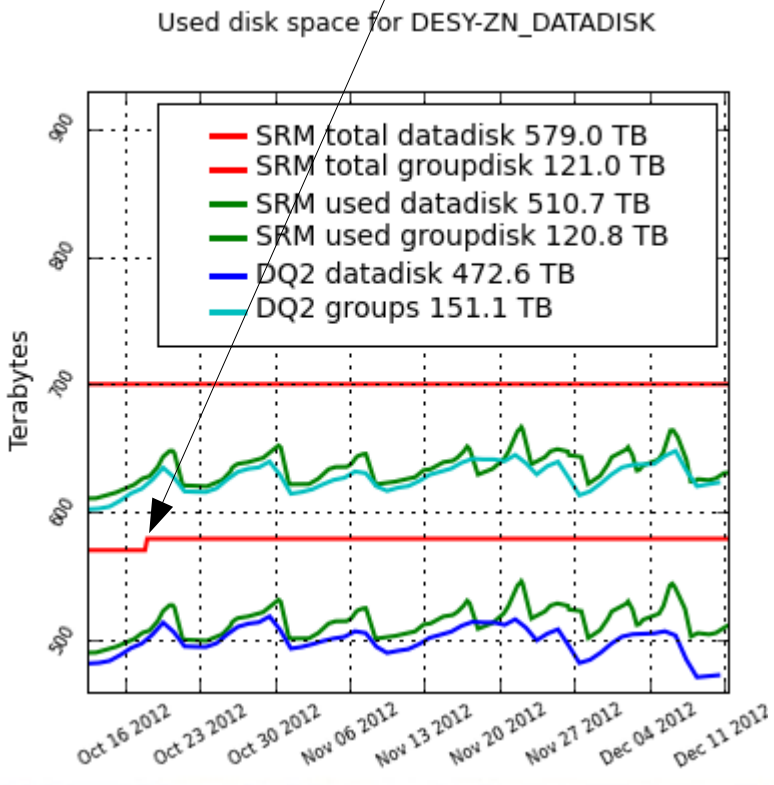
GROUPDISK => DATADISK concerns

- Will the space promised to group be available?
  - Yes, but DATADISK must have enough secondaries (so we can delete them when group actually starts using the space)
- Is this supported by all SE types?
  - No. StoRM and EOS bind space token with namespace (path)
    - So we cannot have DATADISK for `.../atlasgroupdisk/...` path
- Is this step needed and useful for all sites?
  - No. US sites tweak DATADISK/GROUPDISK size on regular basis (automaticly, sometimes manually, but without DDM operations request)
- Do we (or site) need to migrate data?
  - No. After some time site will be asked to move reallocate free space from ATLASGROUPDISK to ATLASDATADISK.
- How will this affect monitoring/accounting?
  - Bourricot is being adopted
  - The new DDM accounting offers grouping by space token or DDM endpoint (among other parameters)
- More details in:
  - SW&C workshop: https://indico.cern.ch/getFile.py/access?contribId=1&sessionId=5&resId=3&materialId=slides&confId=169697
  - ADC weekly: https://indico.cern.ch/getFile.py/access?contribId=10&resId=0&materialId=slides&confId=210753

New data for group endpoints are written to ATLASDATADISK so site could reduce the ATLASGROUPDISK



The space was assigned to ATLASDATADISK



Bottom red line is SRM total for DATADISK

Top red line shows sum space of DATADISK and GROUPDISK (did not change after the free space had been reassigned)

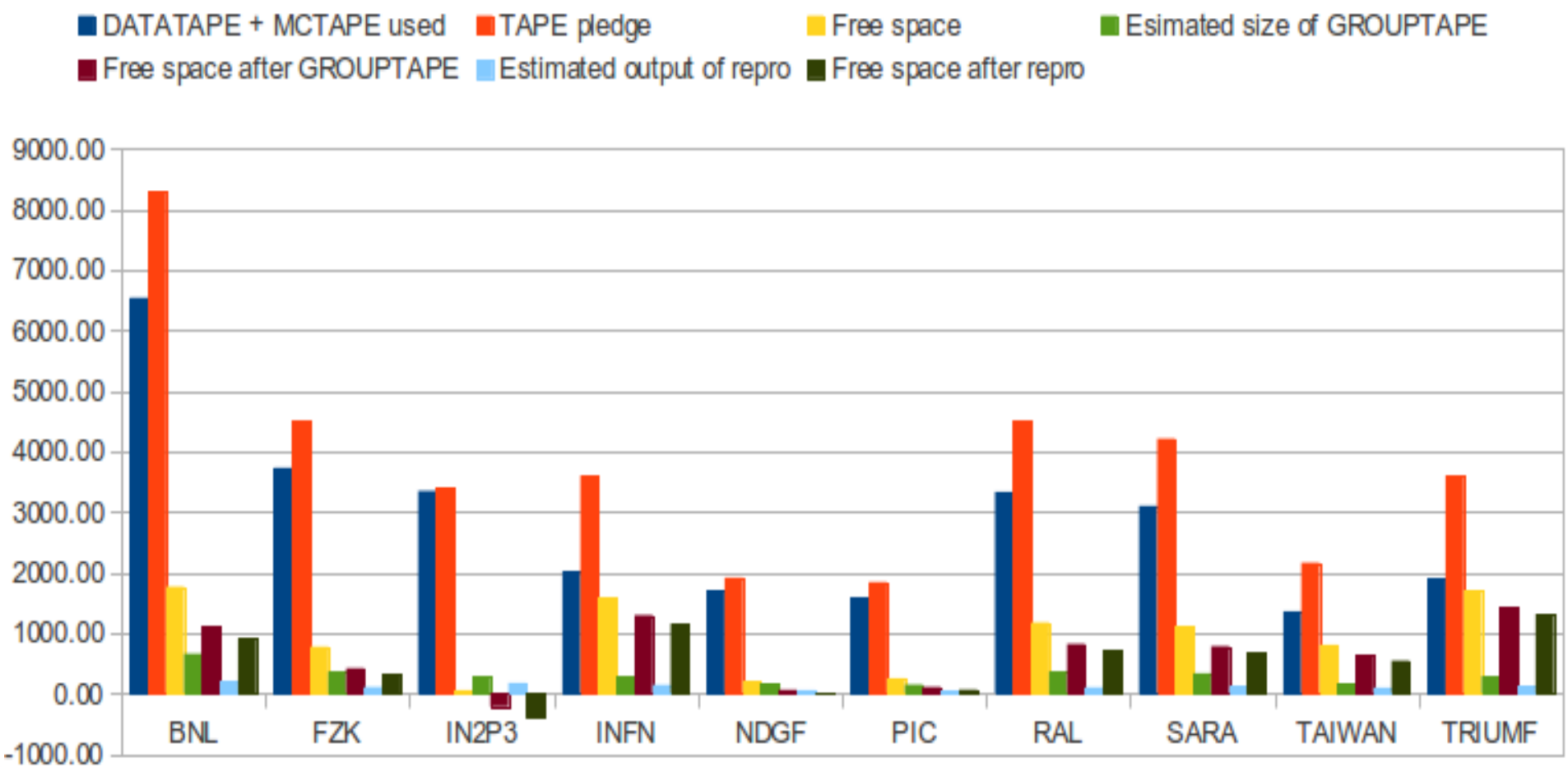
Dark green lines for SRM used space

Blue lines for DQ2 accounting (again DATADISK and DATADISK+GROUPDISK)

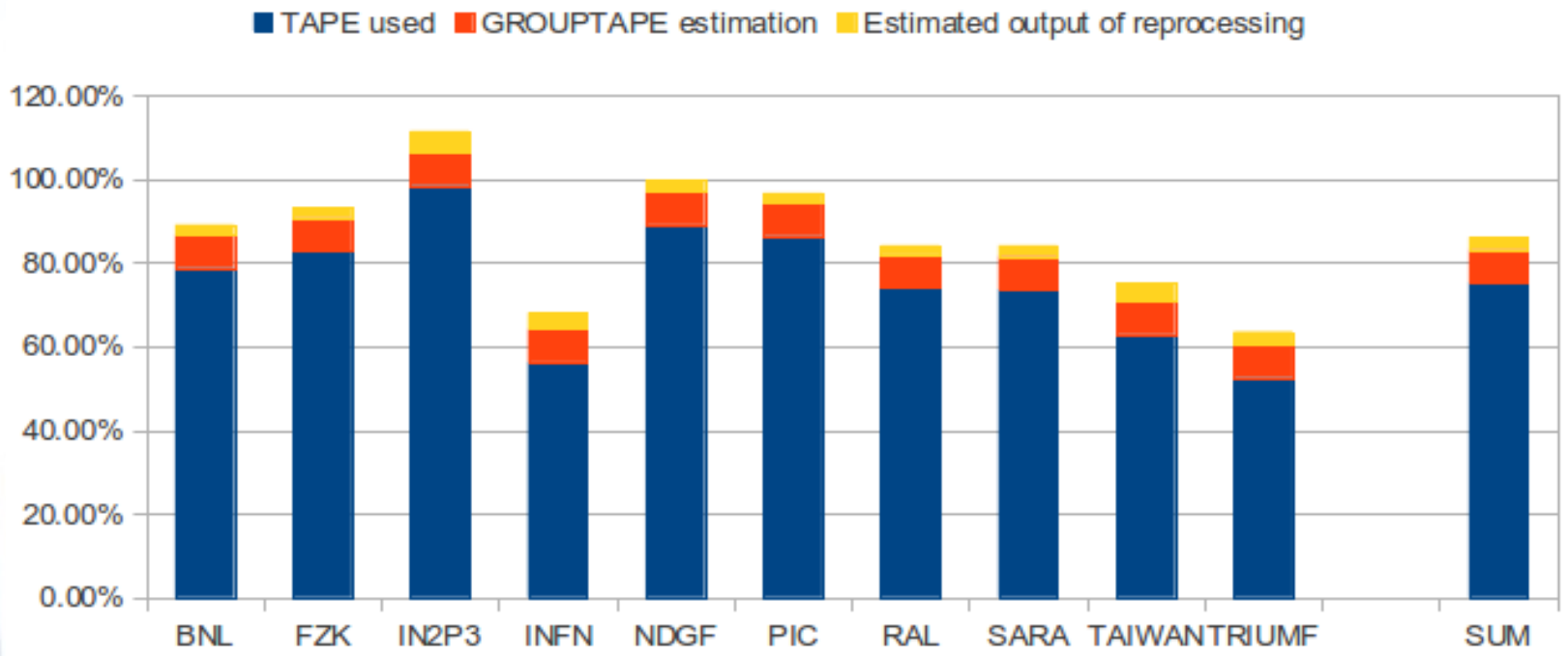
Tape space for groups

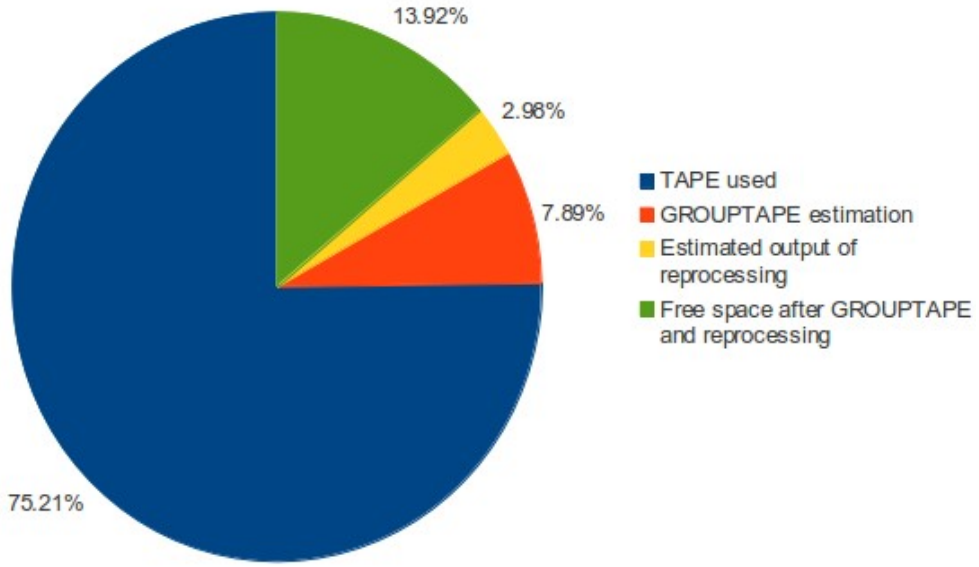
- ATLAS now provides tape space for groups
  - order of PBs
  - uses ATLASMCTAPE space token
    - with one exception ATLASGROUPTAPE at INFN-T1
  - uses `.../atlasgrouptape/<group name>` subdirectory
- e.g. SARA-MATRIX_GROUPTAPE_PHYS-TOP
  - `token:ATLASMCTAPE:srm://srm.grid.sara.nl:8443/srm/managerv2?SFN=/pnfs/grid.sara.nl/data/atlas/atlasgrouptape/phys-top/`
- The guide for groups how to use the space:
  - https://twiki.cern.ch/twiki/bin/viewauth/Atlas/GroupsOnGrid#Archiving_group_datasets_into_Ta

General tape situation



- These numbers are taken from DQ2 accounting.
- The tape compression is not take into account.
  - i.e. the situation reported by tape system can be different


TAPE used
GROUPTAPE estimation
Estimated output of reprocessing
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
BNL
FZK
IN2P3
INFN
NDGF
PIC
RAL
SARA
TAIWAN
TRIUMF
SUM


13.92%
2.98%
7.89%
75.21%
TAPE used
GROUPTAPE estimation
Estimated output of reprocessing
Free space after GROUPTAPE and reprocessing

Tape buffers

- ATLAS asked sites to have separate disk buffers for MCTAPE, DATATAPE
  - so the T0 export is not slowed by MC
- We (DDM ops) would like to hear from the sites about their experiences and observation about the two buffers
- There is no T0 export during LS1
  - what should we do during LS1? merge the two buffers?
  - what should we do after LS1? have them separate again?

HOTDISK removal

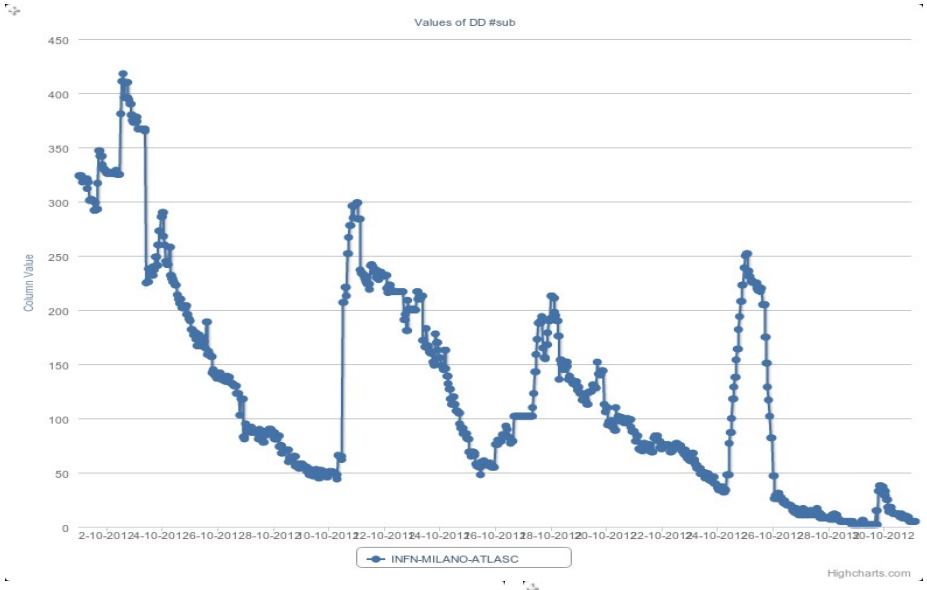
- Sites with working CVMFS do not need HOTDISK anymore
  - Pilot already uses files from CVMS if available
  - PanDA is able to broker jobs to sites if they have CVMFS installed regardless of HOTDISK presence
  - We need to keep HOTDISK at T1s for file distribution to T2s where reconstruction might run
  - We will not remove HOTDISKs until the end of reprocessing
  - Still some issues being discussed
    - https://twiki.cern.ch/twiki/bin/viewauth/Atlas/ADCOperations#HOTDISK
- Cleaning and removing HOTDISK is possible (and being done at ZA-UJ)
- Sites might be asked to remove this space token and add the space to DATADISK

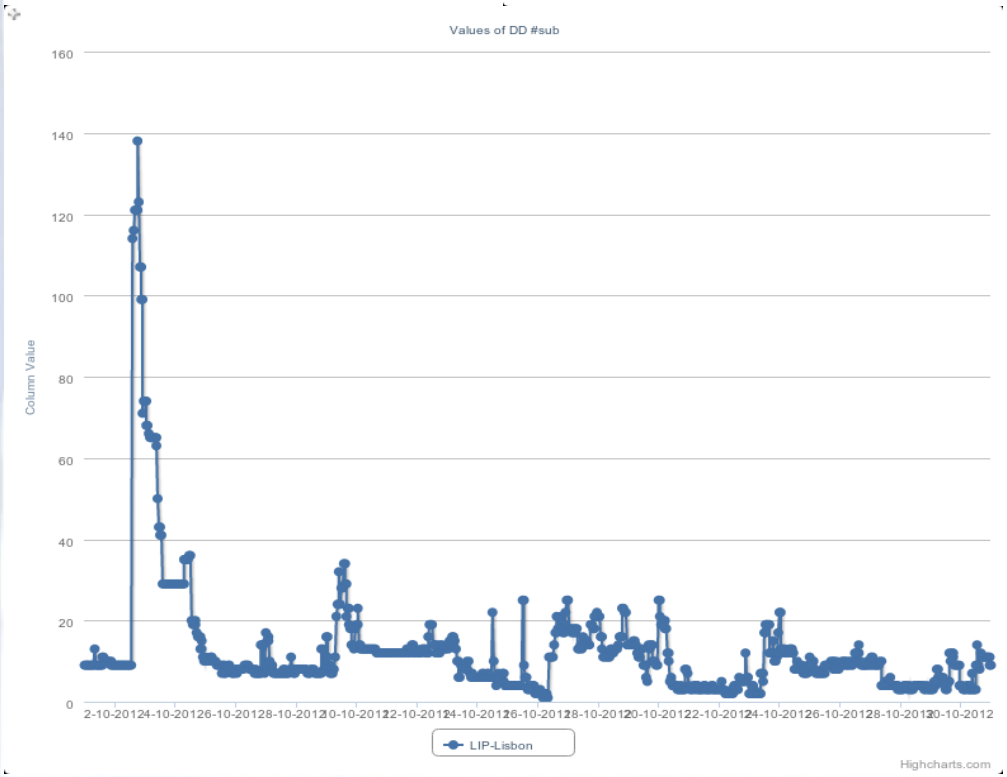
PRODDISK data in DATADISK?

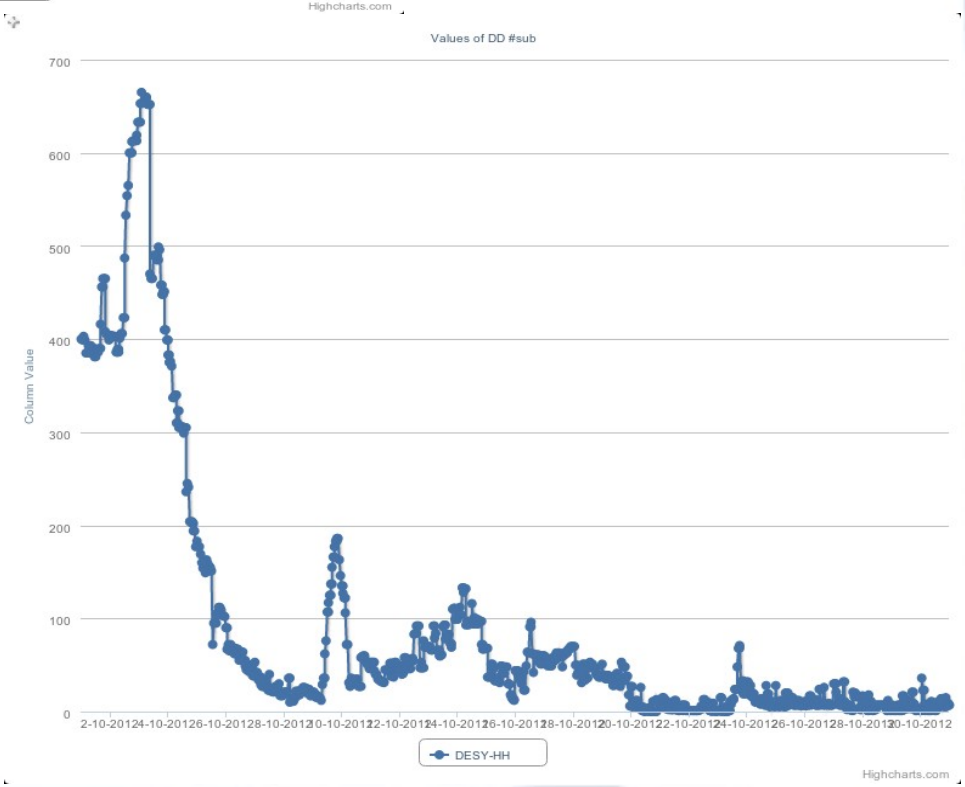
- These space tokens are only accessible to `/atlas/Role=production` VOMS role
- They were set separate in the beginning because PRODDISK was anticipated to have problem with dark data
  - This does not happen much recently
    - LFC registration is done centrally by PanDA
- We are investigating the possibility to use DATADISK for temporary data for production
  - Deeper understanding of dark data is needed
  - Being tested at T1s

T2D status

- Some sites qualified as T2D but do not deliver
  - We consider to remove them from the list
  - Clouds were contacted so they can explain the reasons or plans to improve the throughput
  - Details in Guido Negri's slides at ADC weekly
    - https://indico.cern.ch/getFile.py/access?contribId=8&resId=1&materialId=slides&confId=218989
- Why this is important: T2D sites are treated differently
  - Only T2D site can become *alpha* from *bravo* in ABCD system
  - *Alpha* gets twice more data from T1-T2 replication and from PD2P
  - New group space is created only at T2D sites
  - DDM does not consider multi-hop transfer for T2Ds


Values of DD #sub
Column Value
450
400
350
300
250
200
150
100
50
0
2-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012
INFN-MILANO-ATLASC
Highcharts.com

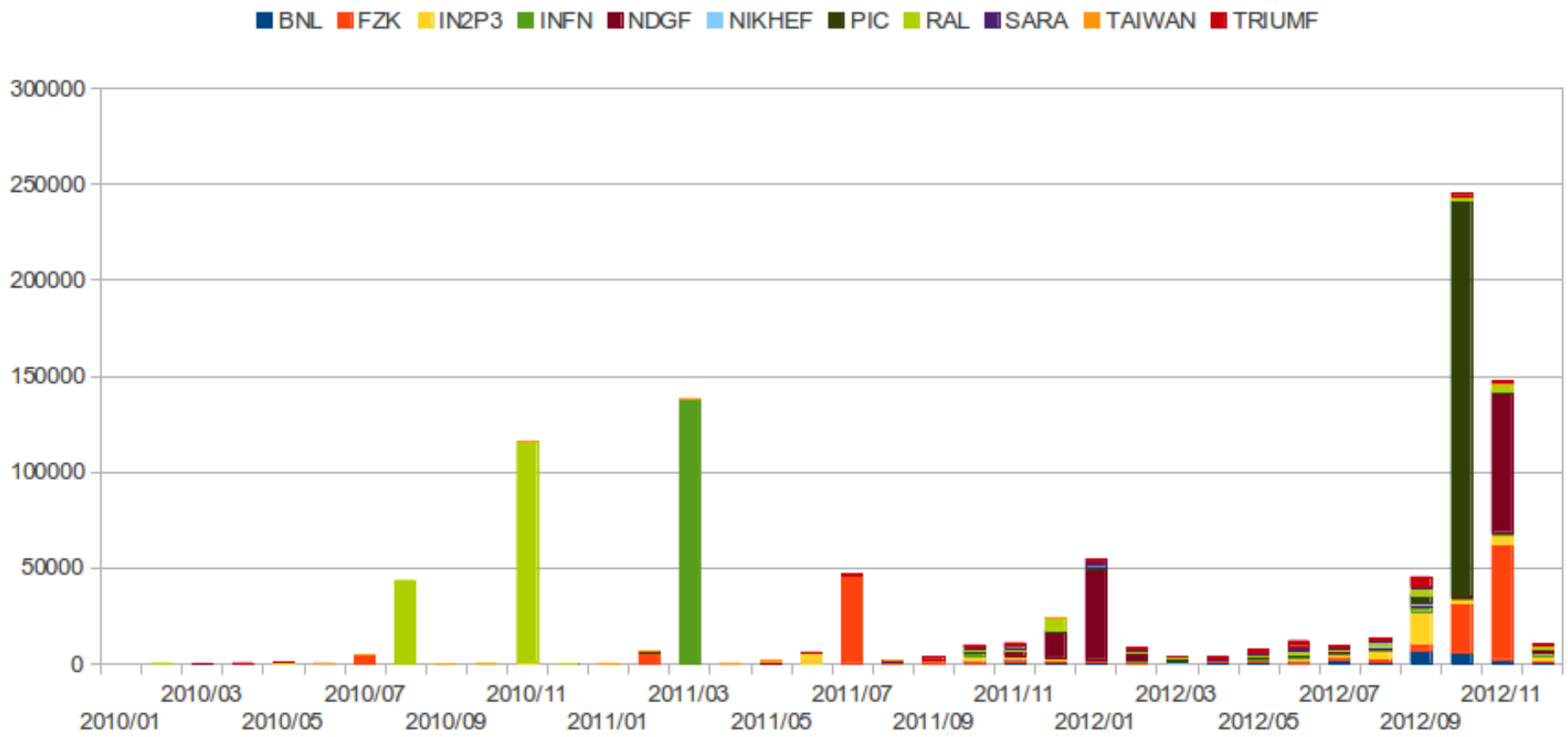

Values of DD #sub
Column Value
160
140
120
100
80
60
40
20
0
2-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012
LIP-Lisbon
Highcharts.com


Values of DD #sub
Column Value
700
600
500
400
300
200
100
0
2-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012 22-10-2012 24-10-2012 26-10-2012 28-10-2012 20-10-2012
DESY-HH
Highcharts.com

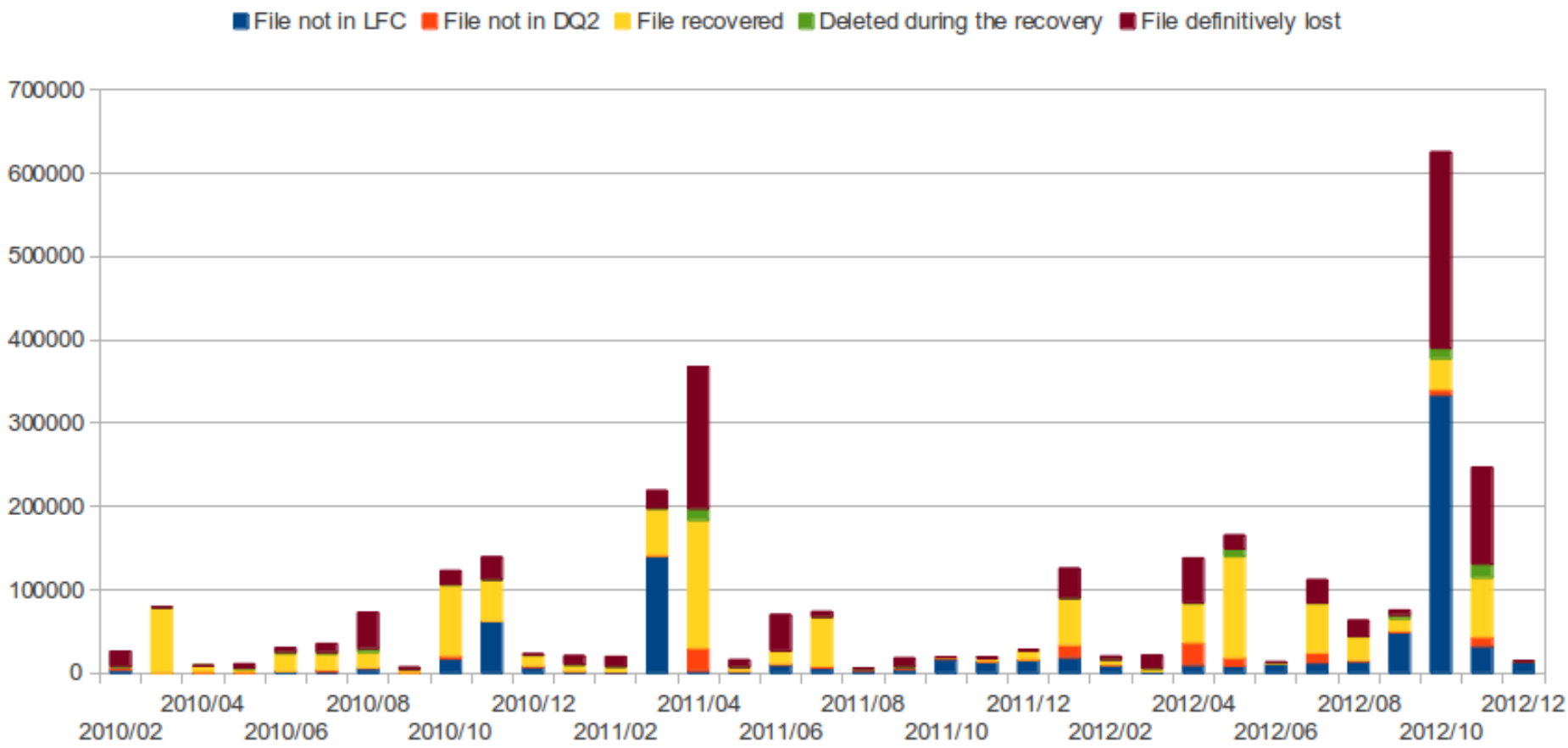
Lost files

- Loosing file is a common problem
- Recovery (aka consistency) service deals with lost files
  - http://bourricot.cern.ch/dq2/consistency/
- Big recent losses:
  - FZK problems with dCache pools
  - dCache pool failure at NDGF
  - accidental deletion of 09 files at PIC
- Most serious is loosing files from tapes
  - Recovery service accounts loosing files from tape disk pool as loosing files from tape

Lost files per T1



Lost files per final state



Conclusion

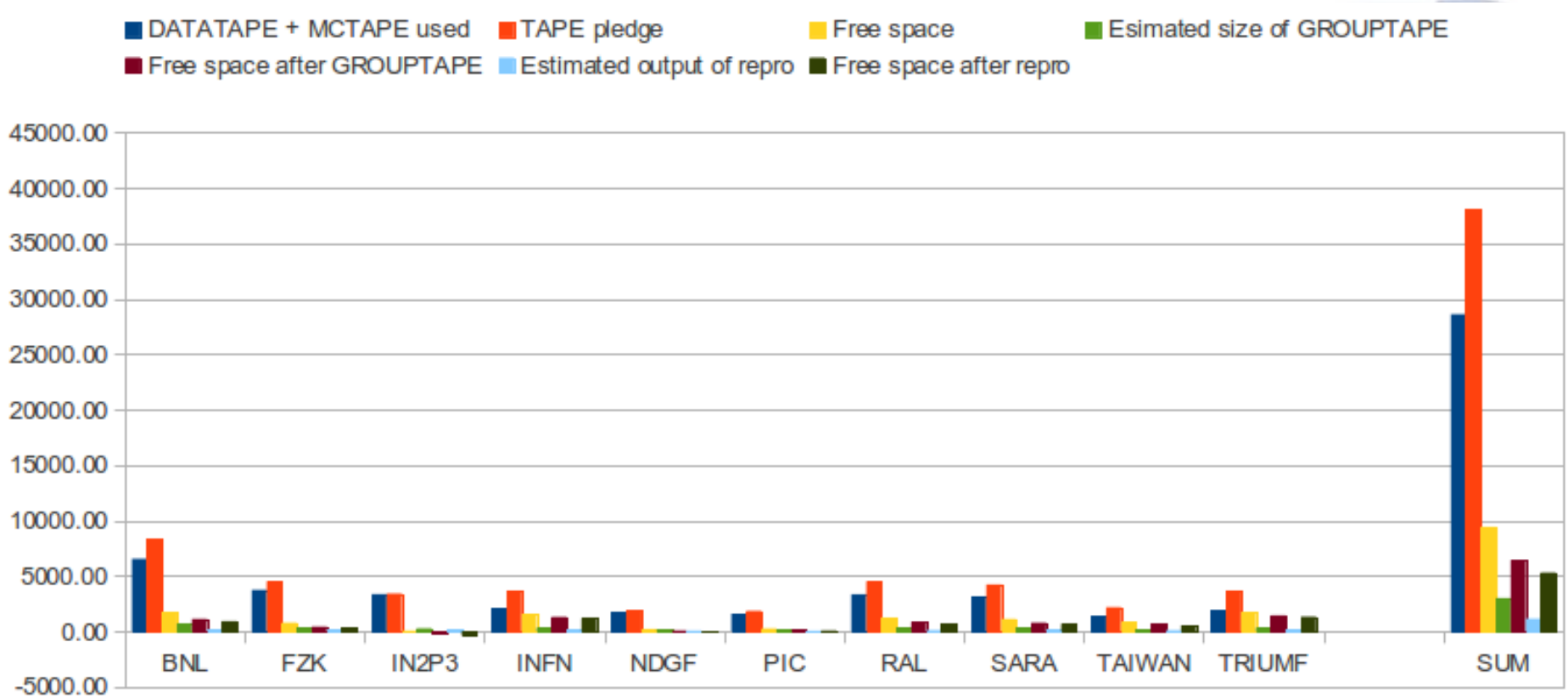
- Nothing important changes before LS1
- Full DATADISK sped up GROUPDISK merge
  - Sites might be asked to reassign free space
- Groups will be soon provided with tape space
  - Action is up to DDM operations
- If site wants to merge the tape buffers, please discuss with DDM operations
- HOTDISK is going to be drained on CVMFS-enabled sites
  - Sites will be asked to reassign the space to DATADISK
- DDM operations is going to regularly check and ask T2D sites about the throughput
  - Sites should not be surprised

Backup slides

Automatic deletion and blacklisting

- DDM endpoints are automatically cleaned by cleaning agent (aka Victor)
  - Works on the following endpoints (free space % that triggers cleaning – target free space in %)
    - PRODDISK (25% - 30%), SCRATCHDISK (50% - 55%), DATADISK (10% - 15%)
  - The default is sometimes seen too strict for T2s (the free space should be minimal), but also too loose for T1s (deletion should start sooner)
  - The thresholds can be set per endpoint if the site feels the defaults are not optimal for them
- An endpoint is automatically blacklisted
  - Based on upcoming downtime, failed transfers, lack of free space
  - For details about the first two see the talk “Operations Automation” (following this one)
- Blacklisting based on free space: http://bourricot.cern.ch/blacklisted_production.html
  - The thresholds are now system wide
  - The plan is to have threshold per site
    - similar to cleaning

Conditions for DISKSPACE exclusions	
DATADISK and MCDISK	Exclude when less than 1.0TB AND 10% free. Reinclude when more than 1.5TB OR 15% free.
DATATAPE and MCTAPE	Exclude when less than 1.0TB. Reinclude when more than 3TB.
EOSDATADISK	Exclude when less than 5.0TB AND 10% free. Reinclude when more than 7.5TB OR 15% free.
HOTDISK	Exclude when less than 0.05TB free. Reinclude when more than 0.1TB free.
T1 SCRATCHDISK	Exclude when less than 10.0TB free. Reinclude when more than 11TB free.
Other spacetokens	Exclude when less than 0.5TB AND 10% free. Reinclude when more than 1.0TB OR 15% free.



T1	DATATAPE + MCTAPE used	TAPE pledge	Free space	Percent share of total TAPE pledge	Esimated size of GROUPTAPE	Free space after GROUPTAPE	Estimated output of repro	Free space after repro
BNL	6535.49	8300.00	1764.51	21.83%	654.97	1109.54	200.13	909.41
FZK	3727.35	4500.00	772.65	11.84%	355.10	417.55	104.99	312.56
IN2P3	3346.62	3400.00	53.38	8.94%	268.30	-214.92	172.80	-387.72
INFN	2022.41	3600.00	1577.59	9.47%	284.08	1293.50	140.35	1153.15
NDGF	1701.11	1911.00	209.89	5.03%	150.80	59.09	53.91	5.17
PIC	1585.26	1836.00	250.74	4.83%	144.88	105.86	44.11	61.75
RAL	3329.09	4500.00	1170.92	11.84%	355.10	815.81	100.08	715.73
SARA	3098.17	4210.00	1111.83	11.07%	332.22	779.61	109.88	669.74
TAIWAN	1356.74	2160.00	803.26	5.68%	170.45	632.81	94.41	538.41
TRIUMF	1890.23	3600.00	1709.77	9.47%	284.08	1425.69	111.71	1313.98
SUM	28592.455	38017	9424.545	100.00%	3000	6424.55	1132.36	5292.18