DDM solutions for disk space resource optimization

Fernando H. Barreiro Megino
(CERN-IT Experiment Support)
on behalf of ATLAS DDM

WLCG Collaboration Workshop

8 July 2010

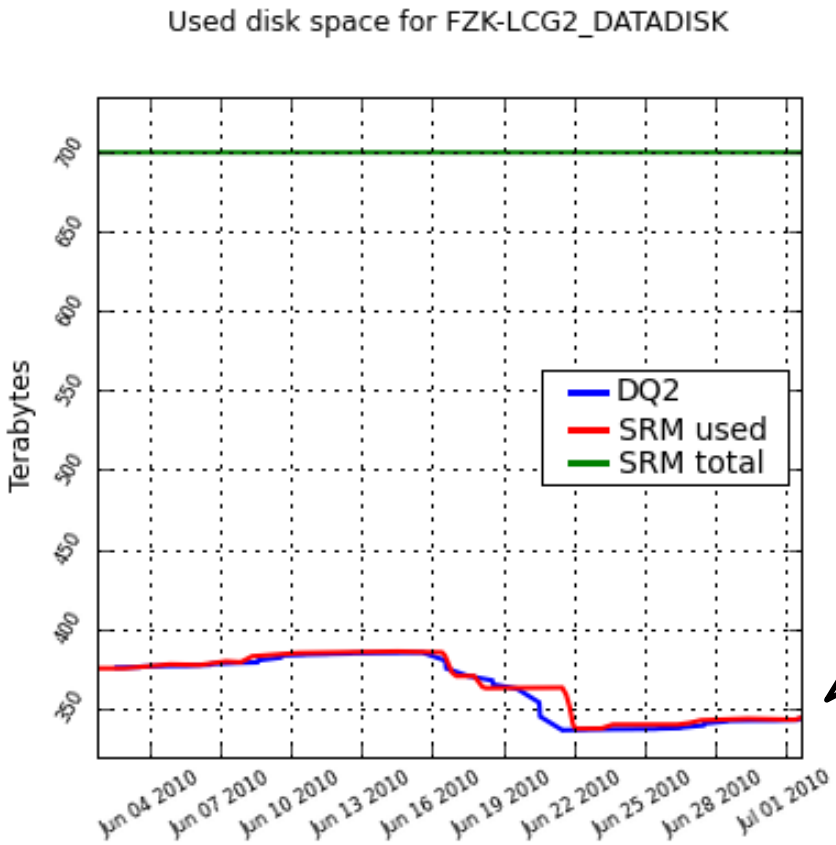
- Optimize space resources
  - Minimize unused space in sites
  - Keep sites full with **interesting** data
  - Improve control and monitoring
- Proposed solutions
  - DDM Accounting
  - Data ranking
  - Automatic replica reduction
  - Automatic DDM site exclusion

DDM Accounting

http://bourricot.cern.ch/dq2/accounting/global_view/30/

- Disk space and usage information at sites collected by different agents and presented in a web frontend
  - Volumes of space, number of datasets and files registered at sites according to DDM (Vincent Garonne)
    - possibility to break down volumes by metadata
  - Total and used disk space according to SRM

- Original main purpose: Follow up consistency between data volumes known by DDM and SRM

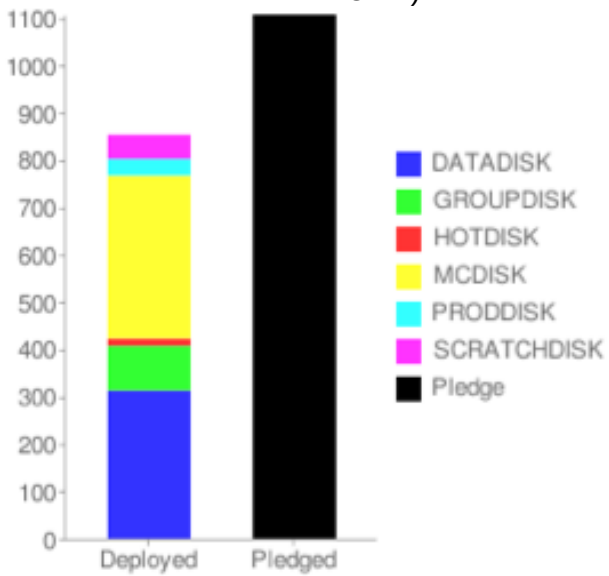


Good example of consistency: FZK running regular checks (Cedric Serfon)

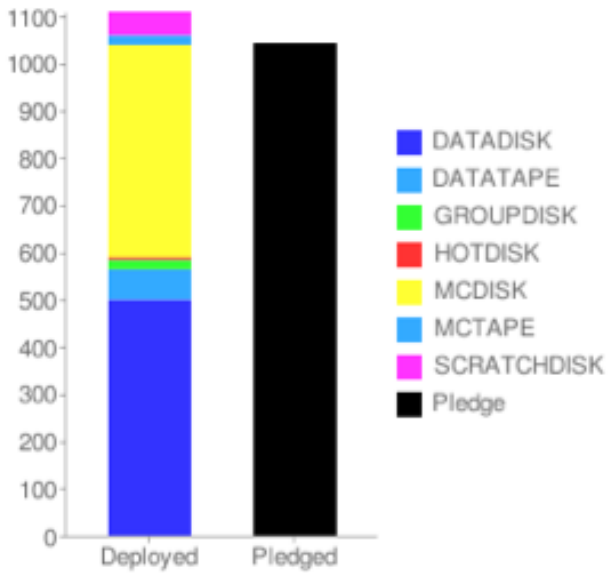
http://bourricot.cern.ch/dq2/accounting/site_view/FZK-LCG2_DATADISK/30/

- Space deployed by federations
- Distribution of resources amongst spacetokens
- Space pledged per federation

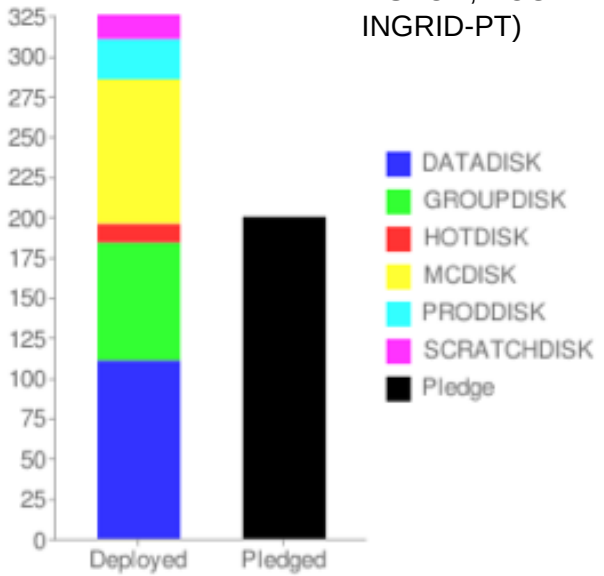
ES-ATLAS-T2 (IFAE, IFIC-LCG2, UAM)



ES-PIC (PIC)



PT-LIP-LCG-Tier2 (LIP-COIMBRA, LIP-LISBON, NCG-INGRID-PT)



Some federations don't deploy their pledge at once

Other federations deploy more than they have pledged

http://bourricot.cern.ch/dq2/accounting/federation_reports/SPAINSITES/

Management resource overviews

CERN

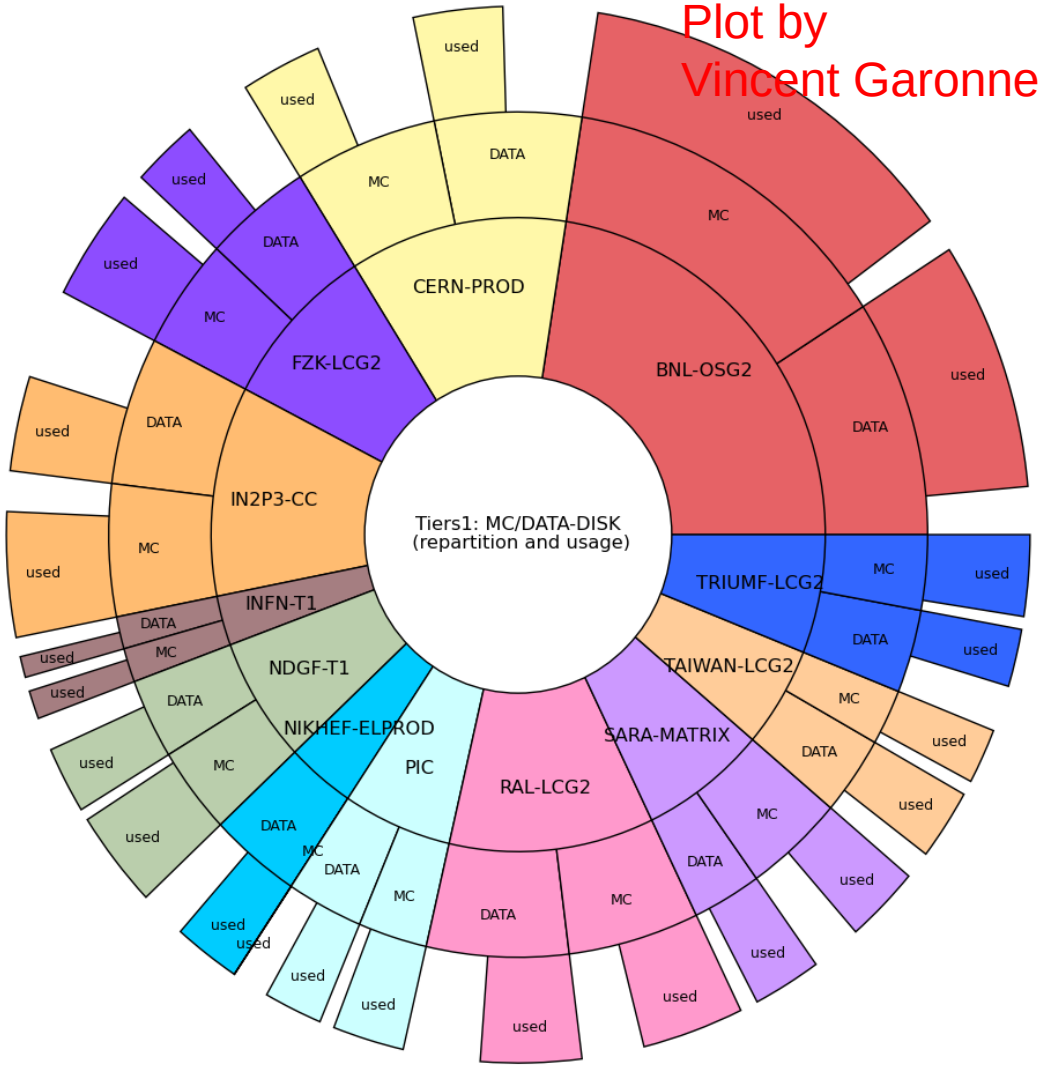
SPACETOKEN	FREE(TB)	USED(TB)	TOTAL(TB)
DATADISK	477	480	957
DATATAPE	68	581	650
GROUPDISK	84	137	221
LOCALGROUPDISK	123	85	208
MCDISK	513	421	934
MCTAPE	13	95	108
SCRATCHDISK	59	54	113
SPECIALDISK	19	2	21
TOTAL	1356	1856	3212

TIER1s

SPACETOKEN	FREE(TB)	USED(TB)	TOTAL(TB)
DATADISK	2533	4804	7337
DATATAPE	371	79	450
GROUPDISK	504	195	699
HOTDISK	165	17	181
LOCALGROUPDISK	164	81	245
MCDISK	1439	5865	7304
MCTAPE	243	91	333
SCRATCHDISK	406	301	708
USERDISK	169	534	703
TOTAL	5995	11966	17961

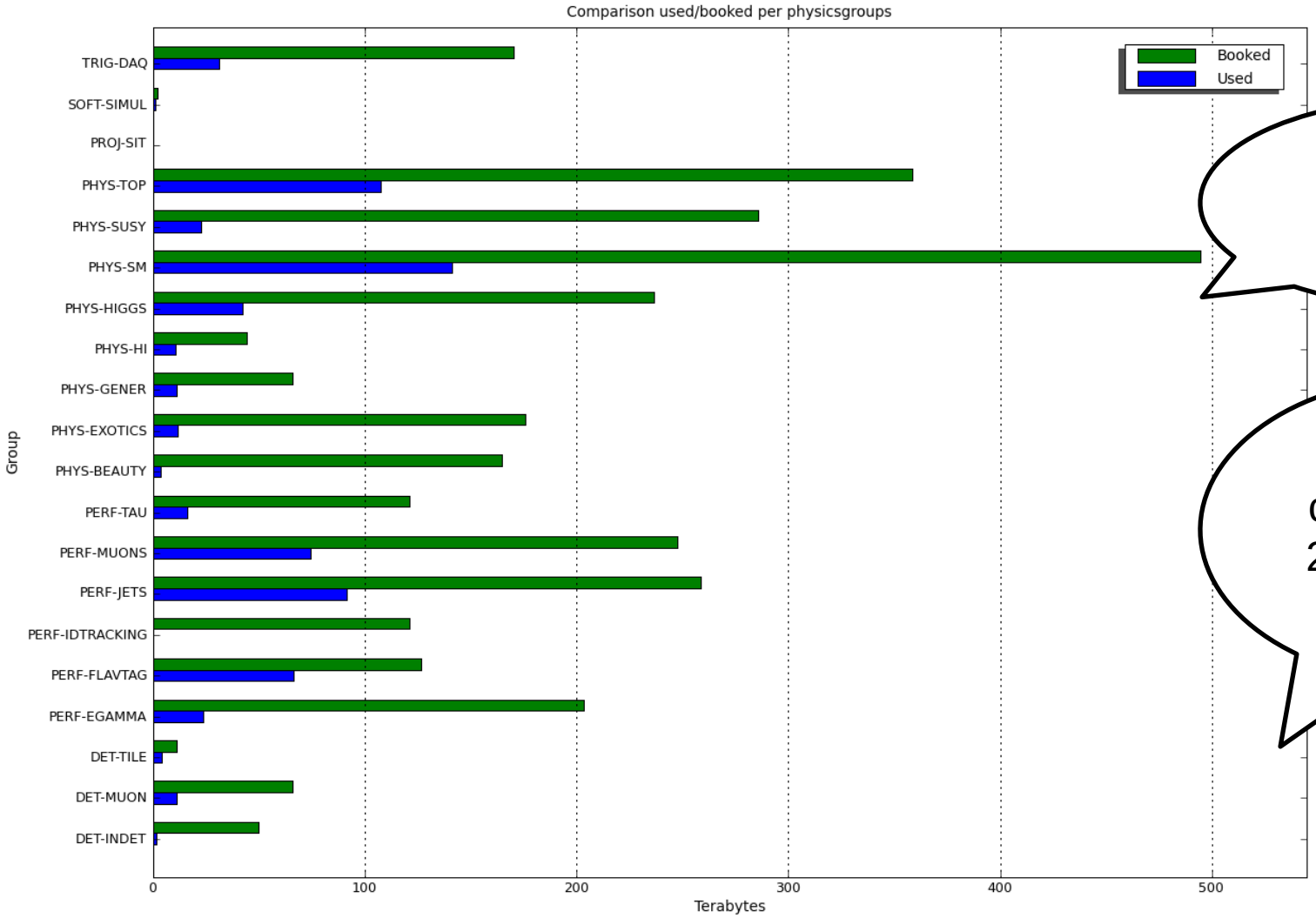
http://bourricot.cern.ch/dq2/accounting/atlas_stats/

Plot by Vincent Garonne



http://bourricot.cern.ch/dq2/accounting/t1_reports/

Physics groups reports



Usage very different between groups

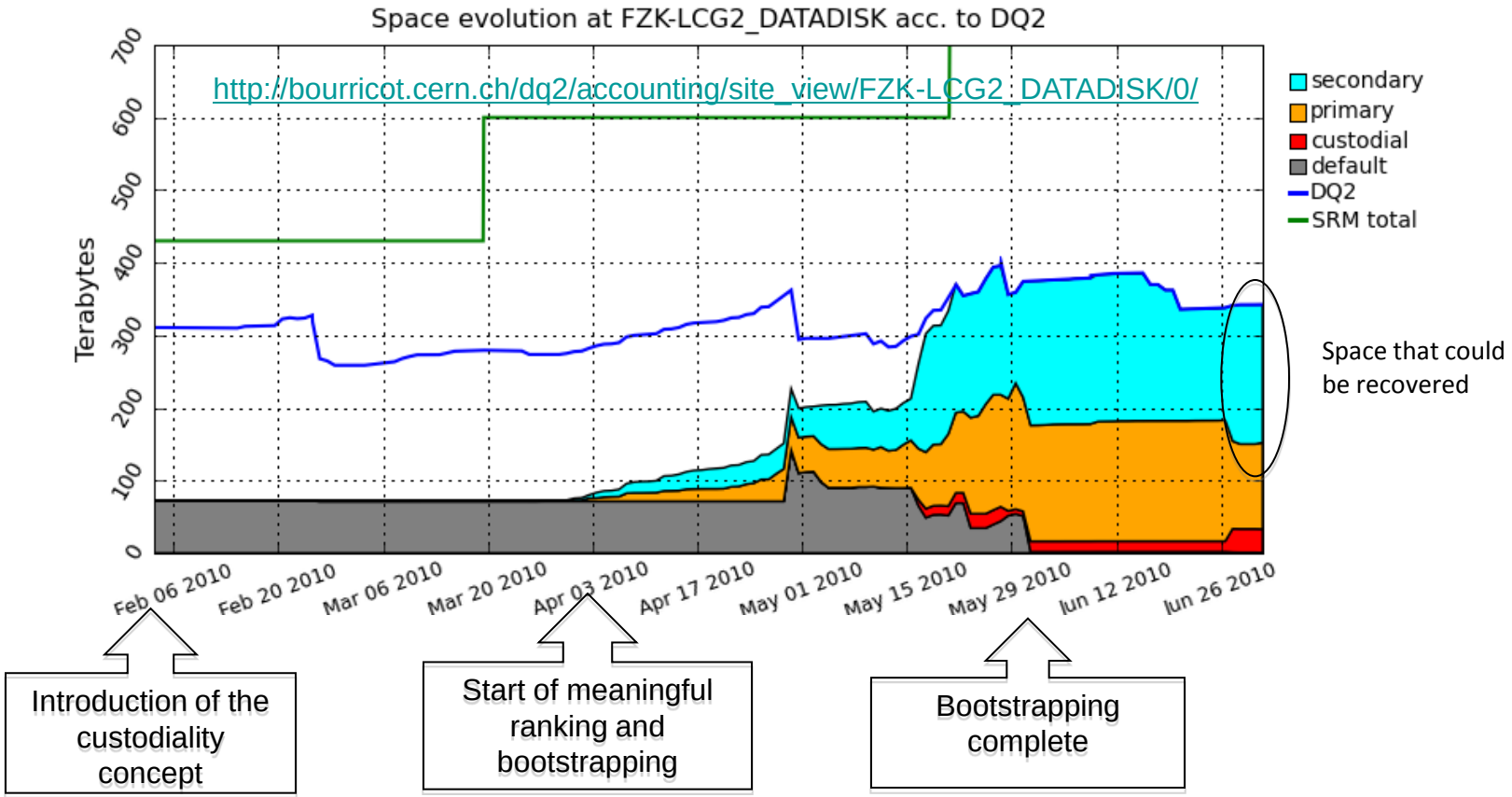
Group quota converging to 25TiB booked per site

http://bourricot.cern.ch/dq2/accounting/group_reports2/

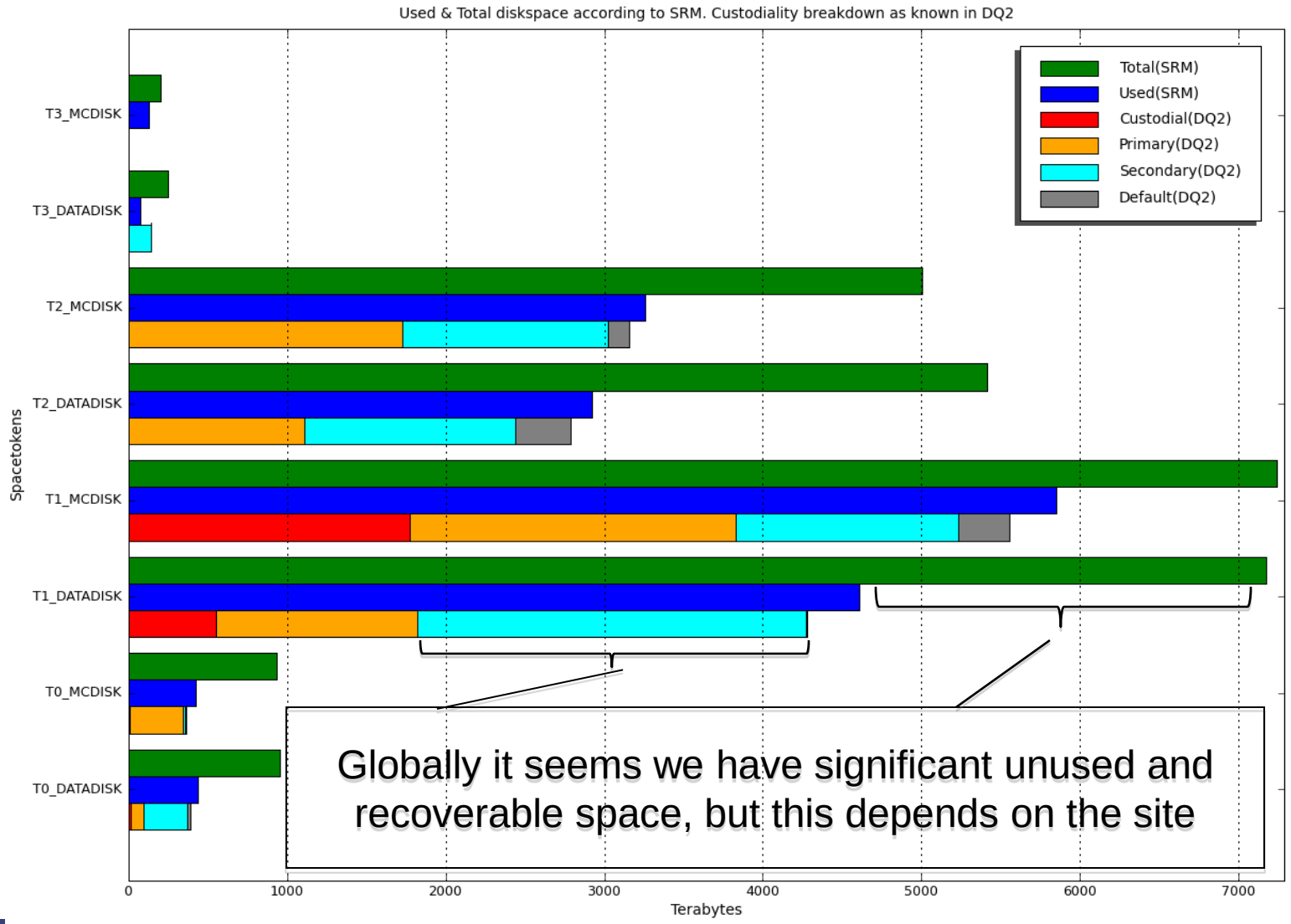
Data ranking

Monitoring of data ranking evolution

- Distinguish Computing Model replicas from additional replicas
  - Custodial: Master replica (Must be kept)
  - Primary: Following Computing Model (CM)
  - Secondary: In excess of CM, may be deleted
  - Default: Not ranked yet

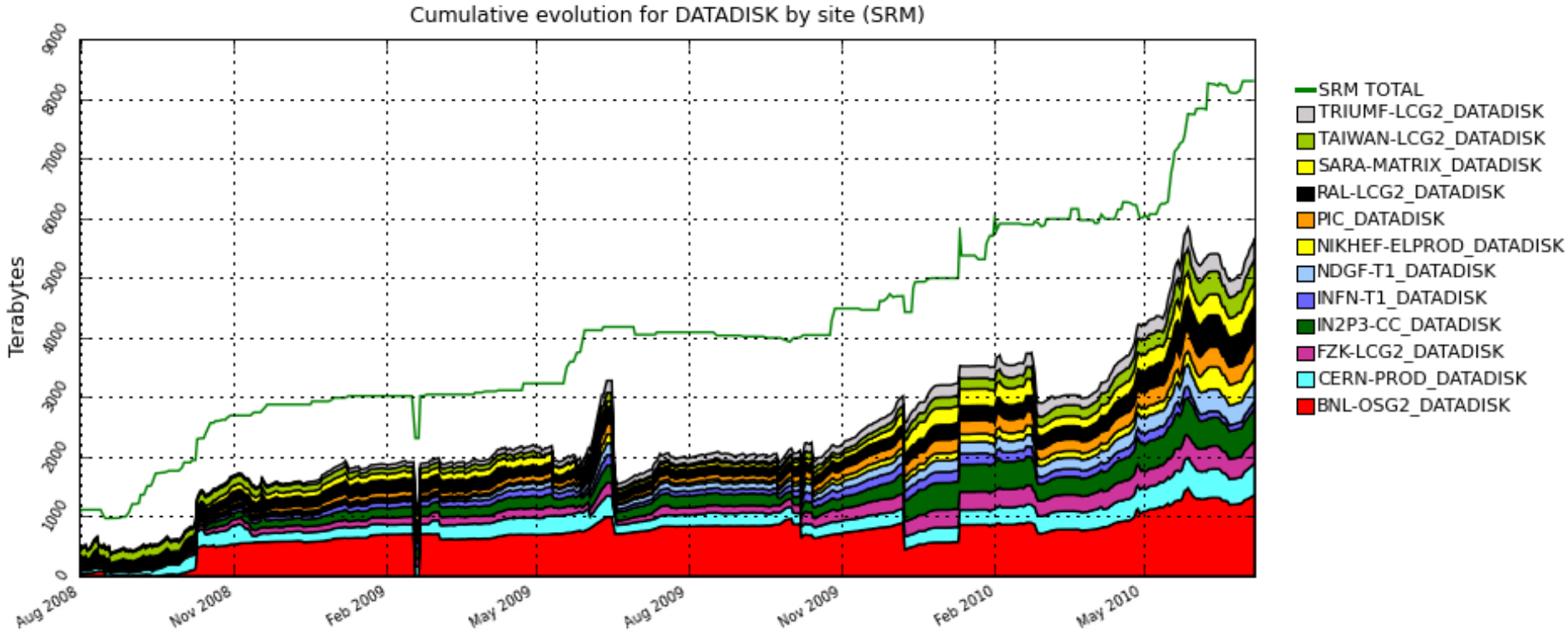


Data ranking



Automatic replica reduction

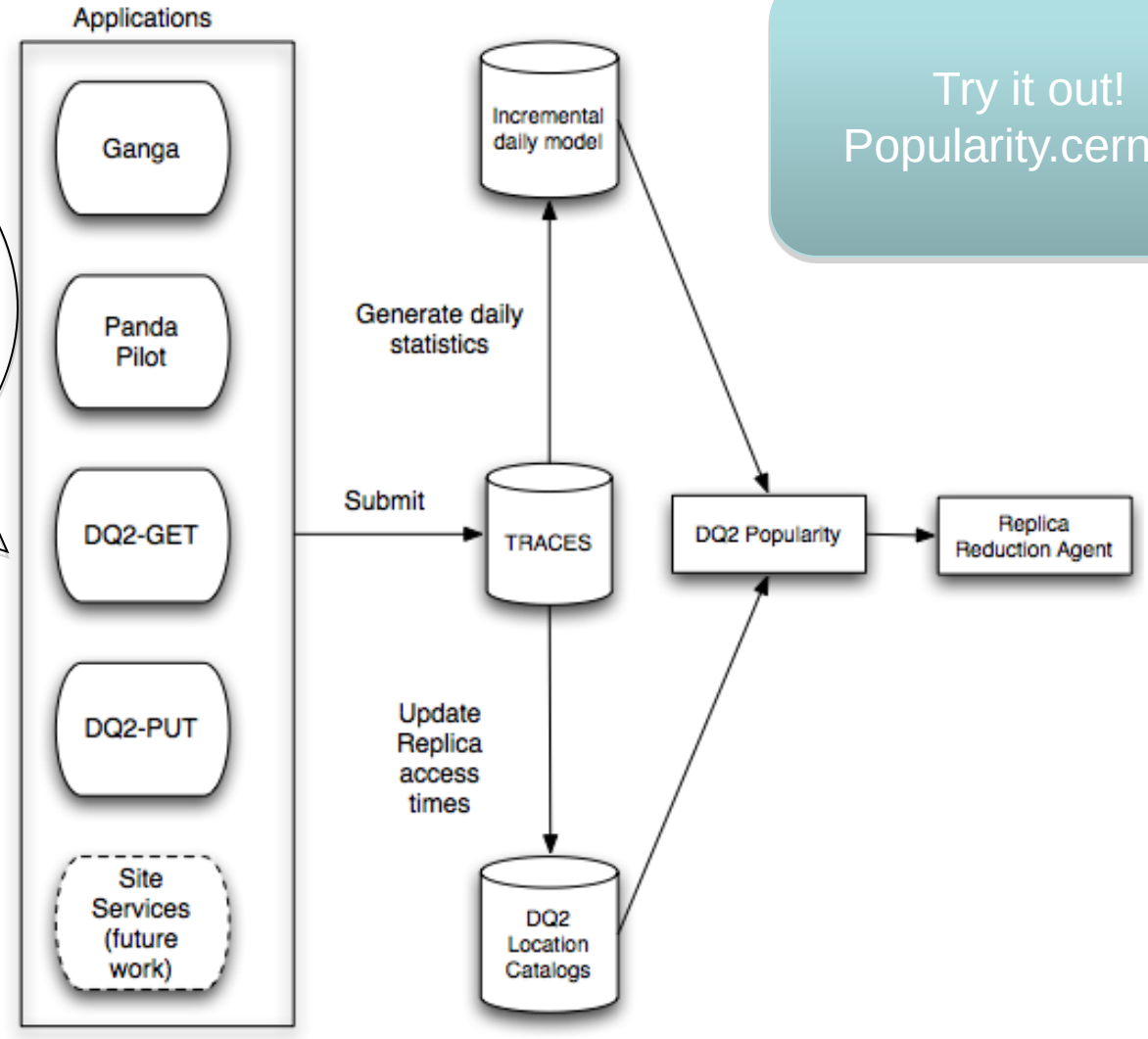
Estimated pledge for T1 DATADISK until **June 2011: 13PB**



Reminder on DDM Popularity

Angelos Molfetas (CERN)

Only actions by official tools can be considered



Try it out!
Popularity.cern.ch

Automatic replica reduction in theory

Andrii Tykhonov (Jozef Stefan Institute, Ljubljana, Slovenia)

- Agent that looks once a day for full MCDISK and DATADISK endpoints
  - Full endpoint: free space<min(10%, 25.0 TB)
- For each full site it will try to select **secondary** replicas to delete
  - Successfully cleaned endpoint: free space>min(20%, 50.0 TB)
- Selection of replicas for deletion based on DDM popularity
  - Older than 15 days
  - From oldest to newest
  - Not accessed recently
- Datasets will be moved into the deletion queue after a grace period of 24 hours.

Evolution of the cleaning strategy from previous presentations

Automatic replica reduction in practice

http://lxvm0338.cern.ch/victor/diskspacemonitor_deletion.html

This webpage contains overflowed* sites

Cloud	Tier	Site name	Total space	'ToBeDeleted' space	Free space (before)		Cleaned space	Free space (after)		"Secondary" space: Young/Popular	Site info	Cleaned**
FRANCESITES	Tier2	GRIF-LAL_MCDISK	104.45 TB	0.00 TB	5.02 TB	5%	5.78 TB	10.79 TB	11%	7.05 / 20.90 TB	info.txt	False
		GRIF-SACLAY_MCDISK	98.96 TB	0.00 TB	8.27 TB	9%	4.88 TB	13.15 TB	14%	3.34 / 14.84 TB	info.txt	False
FZKSITES	Tier2	CYFRONET-LCG2_MCDISK	37.38 TB	0.00 TB	2.86 TB	8%	0.06 TB	2.91 TB	8%	1.03 / 18.52 TB	info.txt	False
		DESY-HH_MCDISK	122.79 TB	0.00 TB	11.79 TB	10%	7.13 TB	18.92 TB	16%	1.41 / 15.88 TB	info.txt	False
NLSITES	Tier2	RRC-KI_MCDISK	80.26 TB	0.00 TB	7.53 TB	10%	9.03 TB	16.56 TB	21%		info.txt	True
USASITES	Tier2	SLACXRD_MCDISK	230.85 TB	0.00 TB	11.28 TB	5%	3.13 TB	14.41 TB	7%	26.04 / 71.06 TB	info.txt	False

Last time stamp: 2010/07/01 12:52:32

Actual deletion started this week. Until now our ops team was manually deleting the published lists

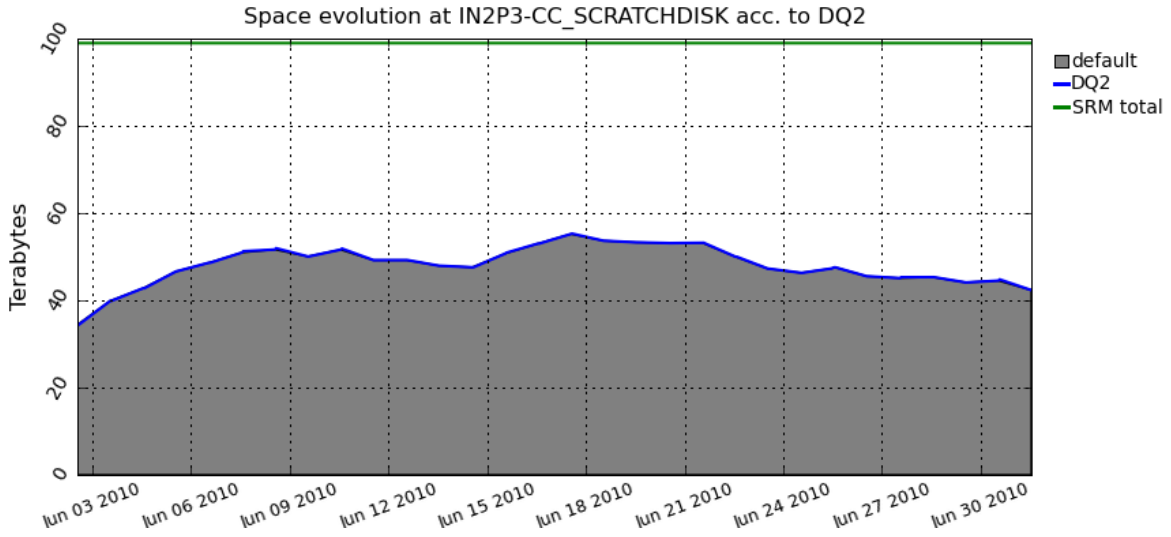
Some sites can not be completely cleaned: The data is too fresh and/or too popular.

Somes notes on automatic replica reduction

- Early stage of experience with the system
  - **Reduce secondary replicas more aggressively**
  - Follow closely un-cleaned DATADISKs and MCDISKs
  - Detect exceptions and provide fine tuning by sites
- Sites who deploy disk space at once will be the ones to fully profit of this model
- Replica reduction is allied with the Dynamic Data Placement (D2PD) model

Future work on automatic replica reduction

- Last week's decision: expand automatic replica reduction to SCRATCHDISK
  - SCRATCHDISK stores analysis output for users
  - At the moment: Fixed policy - datasets older than 30 days are automatically deleted
  - In the future: More flexible policy
    - Keep between 20 and 10% free in spacetoken
    - Once spacetoken has less than 10% free, remove data from oldest to newest until 20% free
    - Data ranking or popularity do not play a role for this spacetoken



Automatic DDM site exclusion

Automatic DDM site exclusion

http://bourricot.cern.ch/blacklisted_production.html

Read — Write

CLOUD	SITE	ENDPOINT	STATUS							
			f		r		u		w	
FRANCESITES	IN2P3-LPC	SCRATCHDISK					DISKSPACE	Free:0.3TB Total:3.3TB	DISKSPACE	Free:0.3TB Total:3.3TB
	RO-02-NIPNE	DATADISK, HOTDISK, MCDISK, PRODDISK, SCRATCHDISK, SOFT-TEST						problems with power line which produce cooling problems.looking for a solution. [Classif. UNSCHEDULED] [Sev. OUTAGE]	manual	https://savannah.cern.ch /support/index.php?113614
NLSITES	RU-MOSCOW-MEPHI-LCG2	DATADISK, LOCALGROUPDISK, MCDISK, PRODDISK, SCRATCHDISK, SOFT-TEST						https://savannah.cern.ch /support/index.php?112518	manual	https://savannah.cern.ch /support/index.php?112518
	TR-10-ULAKBIM	DATADISK, HOTDISK, MCDISK, PRODDISK, SCRATCHDISK, SOFT-TEST	AGIS	ULAKBIM will be unavailable, due to maintenance of power infrastructure of ULAKBIM [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	TR-10-ULAKBIM will be unavailable, due to maintenance of power infrastructure of ULAKBIM [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	TR-10-ULAKBIM will be unavailable, due to maintenance of power infrastructure of ULAKBIM [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	TR-10-ULAKBIM will be unavailable, due to maintenance of power infrastructure of ULAKBIM [Classif. SCHEDULED] [Sev. OUTAGE]
TAIWANSITES	TW-FTT	SCRATCHDISK					DISKSPACE	Free:0.5TB Total:16.1TB	DISKSPACE	Free:0.5TB Total:16.1TB
UKSITES	UKI-LT2-IC-HEP	DATADISK, HOTDISK, MCDISK, PRODDISK	AGIS	Moving dCache outside firewall. [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	Moving dCache outside firewall. [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	Moving dCache outside firewall. [Classif. SCHEDULED] [Sev. OUTAGE]	AGIS	Moving dCache outside firewall. [Classif. SCHEDULED] [Sev. OUTAGE]
	UKI-LT2-QMUL	DATADISK, HOTDISK, LOCALGROUPDISK, MCDISK, PERF-JETS, PHYS-TOP, PRODDISK, SCRATCHDISK	AGIS	Emergency downtime to alter the Voltage optimisation equipment installed 2 weeks ago. This requires all the machines to be shut down. [Classif. UNSCHEDULED] [Sev. OUTAGE]	AGIS	Emergency downtime to alter the Voltage optimisation equipment installed 2 weeks ago. This requires all the machines to be shut down. [Classif. UNSCHEDULED] [Sev. OUTAGE]	AGIS	Emergency downtime to alter the Voltage optimisation equipment installed 2 weeks ago. This requires all the machines to be shut down. [Classif. UNSCHEDULED] [Sev. OUTAGE]	AGIS	Emergency downtime to alter the Voltage optimisation equipment installed 2 weeks ago. This requires all the machines to be shut down. [Classif. UNSCHEDULED] [Sev. OUTAGE]
USASITES	DUKE	LOCALGROUPDISK					DISKSPACE	Free:0.1TB Total:10.2TB	DISKSPACE	Free:0.1TB Total:10.2TB
	MWT2_UC	GROUPDISK, LOCALGROUPDISK, PERF-JETS, PERF-TAU, PHYS-HIGGS					DISKSPACE	Free:0.3TB Total:3.7TB	DISKSPACE	Free:0.3TB Total:3.7TB
	SLACXRD	PERF-FLAVTAG, PERF-IDTRACKING, PERF-JETS, PHYS-BEAUTY, PHYS-SM							manual	Offline due to lack of space

(red)	OFF: Site is EXCLUDED from DDM for the particular permission
(green)	ON: Site is INCLUDED in DDM for the particular permission. Automatic collectors will NOT be taken into account.
(white)	AUTO: Allows the automatic collectors to decide whether a site should be included or not.

Published GOCDB and OIM downtimes are taken into account automatically

Full spacetokens will be excluded as destination

All the information is propagated to the **Site Status Board**

- Effort from ATLAS Distributed Computing in optimizing space resources provided by sites
- Need to become more aggressive in secondary data deletion
- Reduce need of manual operations
- Monitoring and deletion tools are in place and will be improved and tuned with experience

Acknowledgements

- Simone Campana
- Alessandro Di Girolamo
- Stephane Jézéquel
- I. Ueda

Backup slides

2010 and 2011 Data Distribution

Following the Comp. Model and the AMFY recommendations

RAW

- 1 copy to tape at CERN
- 1 copy to tape distributed over the T1s
- ~~1 copy to disk distributed over the T1s~~ a fraction to disk distributed over the T1s
- ~~1 copy on disk at BNL, CCIN2P3 and SARA~~

ESD

- 1 copy to tape at CERN
- ~~1 copy to disk at CERN~~
- 2 copies to disk distributed over the T1s
- ~~1 copy distributed over the T2s of each cloud~~
- ~~Some extra copies to disk in specific T2s~~ a fraction to disk in the T2s

AOD ~~& DESD~~

- ~~Like the ESD above~~
- 2 copies to disk distributed over the T1s
- 10 copies to disk distributed over all T2s (nota bene: not per cloud)

DESD

- 10 copies to disk distributed over all T2s of all clouds (nota bene: not per cloud)

Others

- Many copies of TAG, TAG_COMM, NTUP, HIST, …

All data can always be requested

Slide from Kors Bos

The Solution

- Dynamic caching model for data at Tier 2's
- Panda already uses cache for production – but need something different for user analysis to preserve quick job start
- Try a dynamic data placement model = PD2P
  - Continue automatic distribution to Tier 1's as decided by ACM/CREM2
  - Reduce automatic data subscriptions to Tier 2's
  - Panda will subscribe a dataset to Tier 2, if no other copies are available (except at a Tier 1), as soon as any user needs the dataset
    - User jobs will still go to Tier 1 while data is being transferred – no delay
  - Panda will subscribe replicas to additional Tier 2's, if needed, based on user demand for datasets
  - Cleanup will be done by central DDM popularity based cleaning service
  - Try this as demo in U.S., starting with DATADISK and MCDISK
  - Exclude RAW, RDO and HITS datasets from PD2P
  - Restrict transfers within cloud for now – next step intra-cloud transfers

Slide from Kaushik De