User Report: CMS

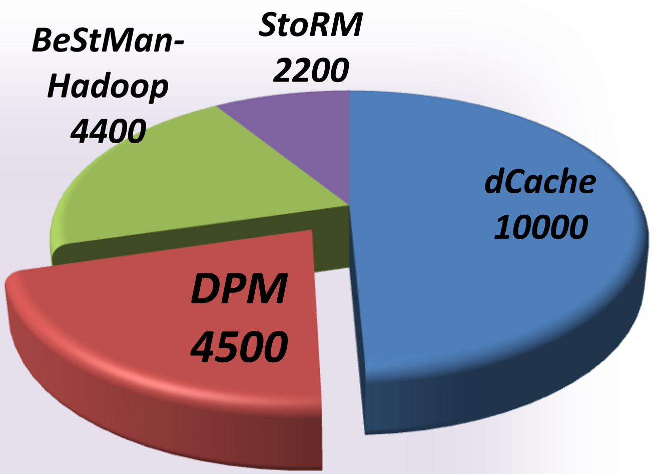
A. Sartirana (LLR, E.Polytechnique, Paris)

On behalf of CMS Computing

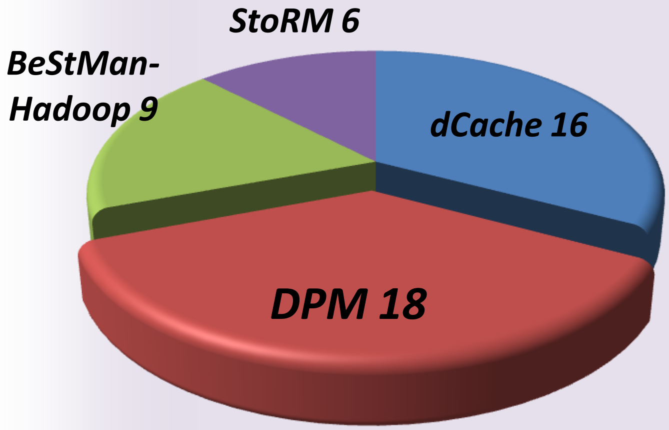
DPM Community resources

Important fraction of CMS Resources

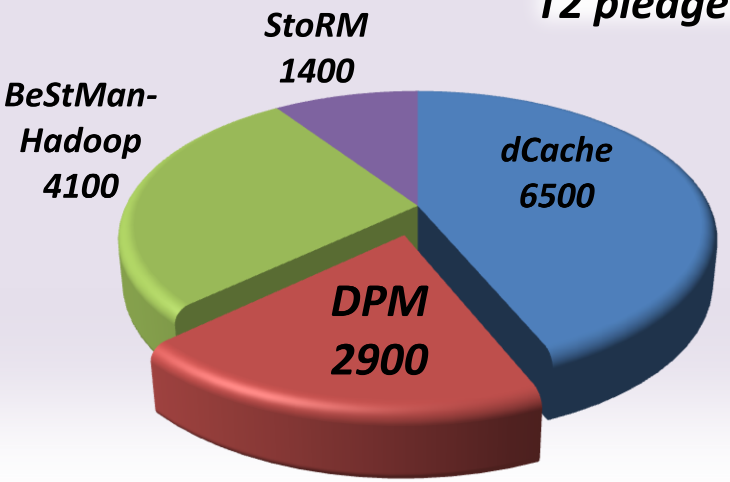
- 37% of CMS T2 Sites;
- 21% of CMS pledges for Q4 2012;
- 19% of currently used PhEDEx space.



T2 pledged TBs (Q4 2012)



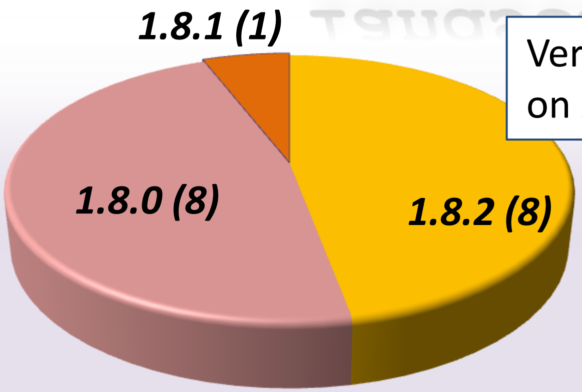
T2 sites/storage tech



T2 used TBs (PhEDEx data)

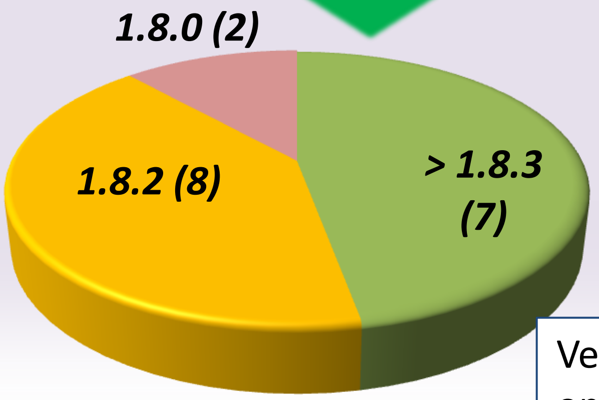
DPM Community landscape

Site	v.
T2_AT_Vienna	1.8.4
T2_FR_GRIF_IRFU	1.8.3.1
T2_FR_GRIF_LLR	1.8.3.1
T2_FR_IPHC	1.8.3.1
T2_GR_Ioannina	1.8.2
T2_HU_Budapest	1.8.0
T2_IN_TIFR	1.8.0
T2_PK_NCP	1.8.2
T2_PL_Warsaw	1.8.3.1
T2_RU_INR	1.8.2
T2_RU_PNPI	1.8.2
T2_RU_RRC_KI	1.8.3
T2_RU_SINP	1.8.2
T2_TR_METU	1.8.0
T2_TH_CUNSTDA	1.8.4
T2_TW_Taiwan	1.8.2
T2_UA_KIPT	1.8.2
T2_UK_London_Brunel	1.8.5

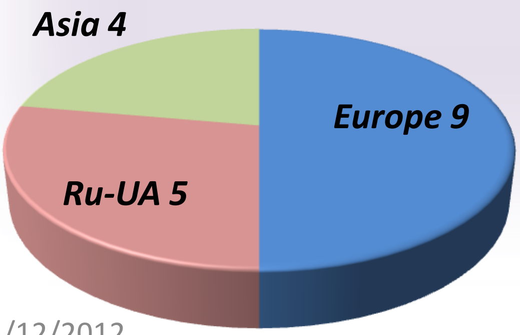


Versions deployed on ***Feb 2012***

Pushed by EMI upgrade



Versions deployed on ***Nov 2012***



DPM Community *feedback*

- ***DPM has proven*** to be ***stable, performing and easy to administrate***
  - issues related to the storage system itself are very rare;
  - ***good perfs*** (e.g. <job eff.>) within CMS sites standards [*];
- important new features appeared or are about to, e.g.
  - new draining tool;
  - new xrootd plugin: more efficient and federation aware;
- still some ***open issues***, for example
  - mgmt of ACL is painful(…and non recursive);
  - buggy/painful pools/groups mapping mgmt;
  - checksum calculation reset file ctime;
- ***very responsive Dev. Team and community.***

[*] slide Backup::Performances

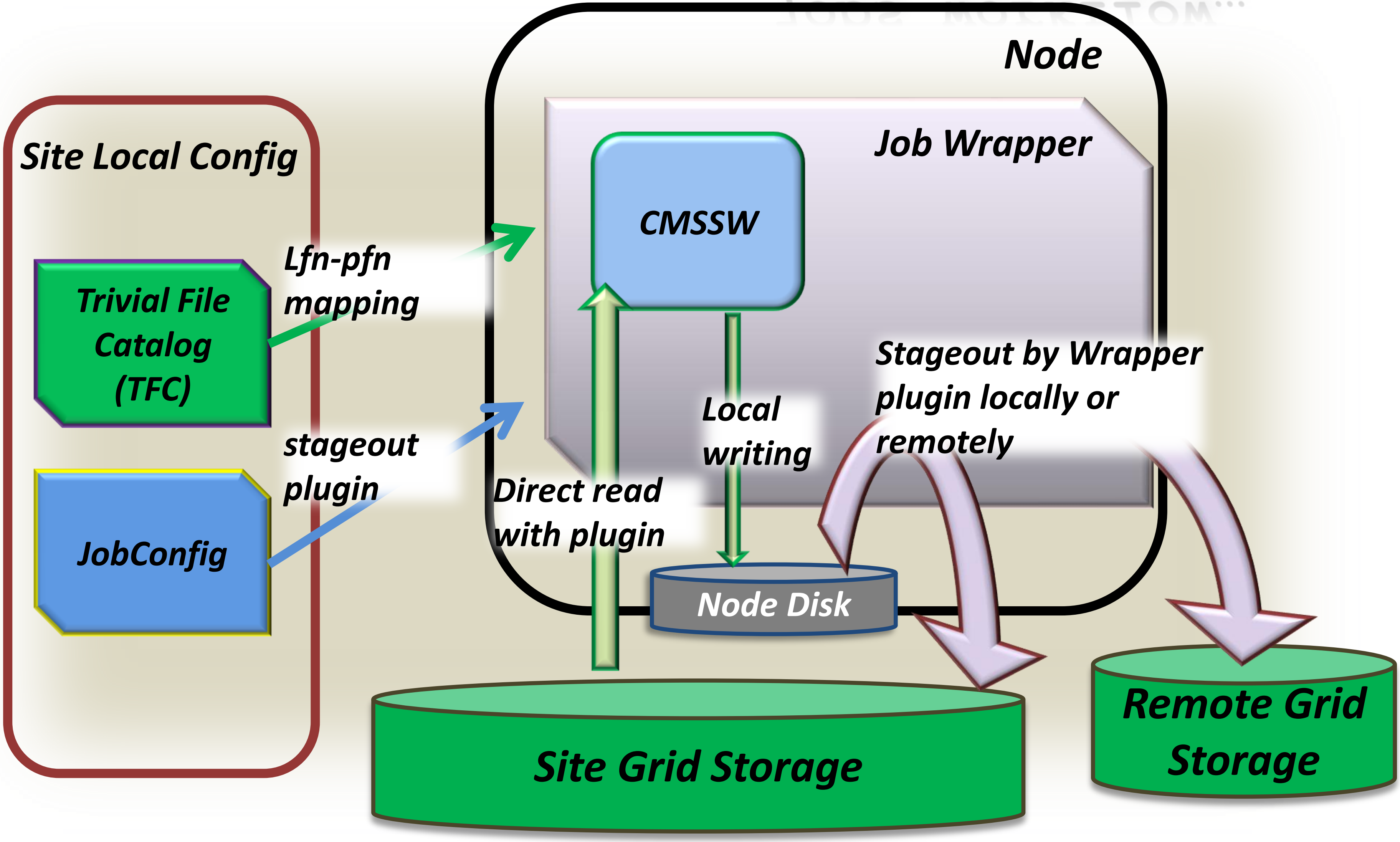
DM Evolving intro

- ***CMS Computing System*** is designed to meet the needs for ***storage and processing of CMS data***
  - original computing model (2005) [*]: ***static, hierarchical*** and ***local***
    - transfers flow hierarchically from T0 to T1 to T2;
    - ***jobs access data locally at sites;***
  - thanks to good ***reliability and performance of networks***: data distribution evolved (2008) into a ***"full mesh"***;
  - today evolving into a "***less data-driven***" ***model*** with the deployment of an ***xrootd federation***
    - allows ***jobs to access data remotely***;
- in the next 2 slides: ***sketch*** of the ***Job Workflow***
  - ***basic info*** and ***evolution*** with the xrootd fed;
  - ***more info*** can be found in the ***backup slides [**]***.

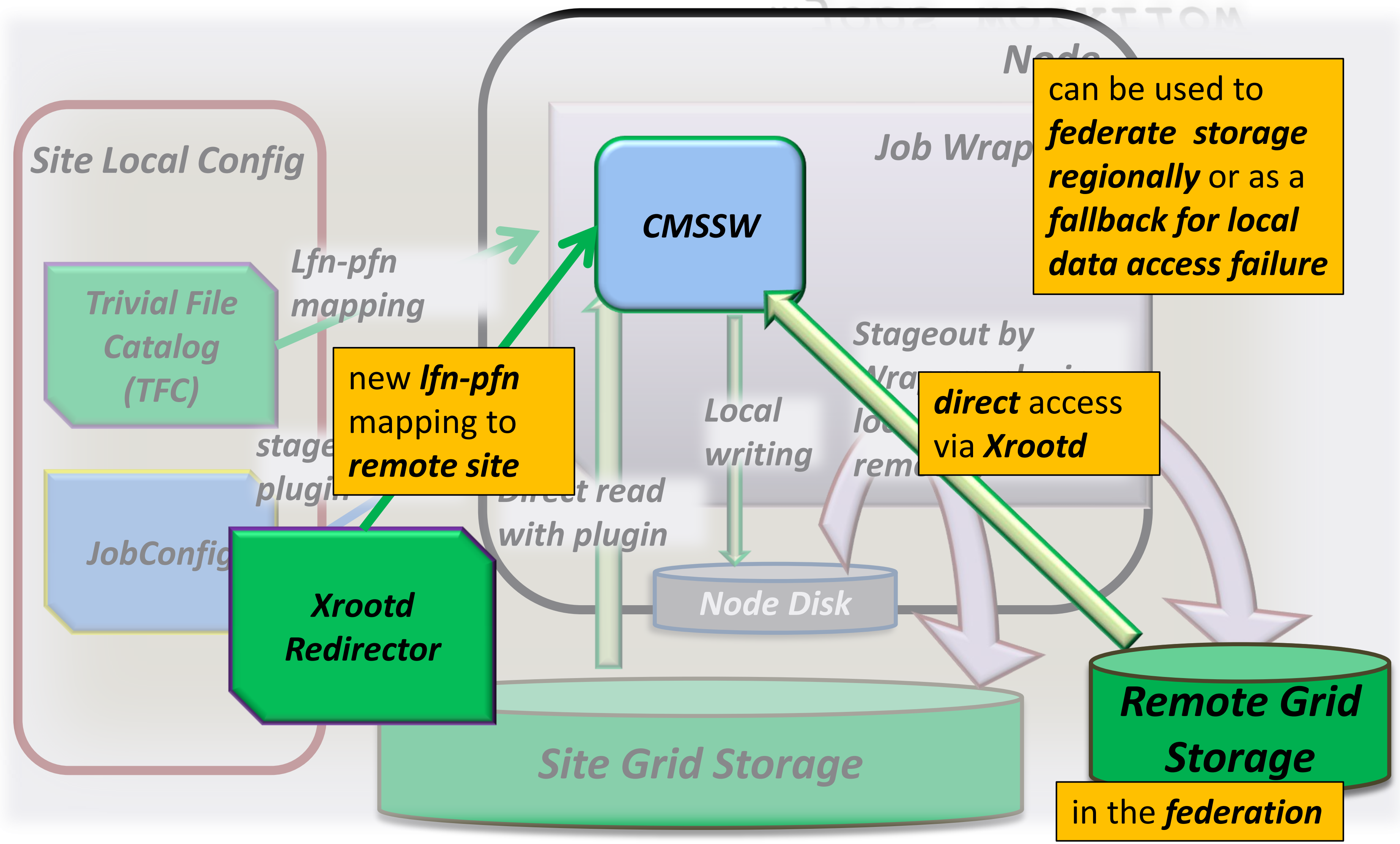
[*] CMS C-TDR released (CERN-LHCC-2005-023)

[**]Backup::{Definitions|PhEDEx}

DM Evolving jobs workflow…



DM Evolving …jobs workflow



Node

Job Wrap

Site Local Config

Trivial File Catalog (TFC)

Lfn-pfn mapping

stage plugi

JobConfig

CMSSW

Local writing

Stageout by Wra lo rem

Direct read with plugin

Node Disk

Xrootd Redirector

Site Grid Storage

Remote Grid Storage

can be used to ***federate storage regionally*** or as a ***fallback for local data access failure***

new ***lfn-pfn*** mapping to ***remote site***

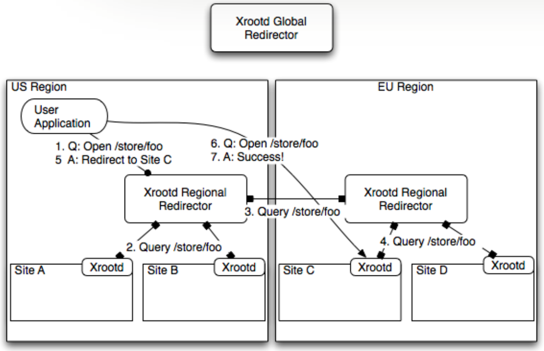
direct access via ***Xrootd***

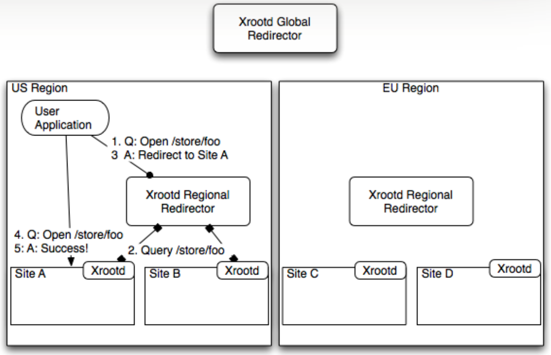
in the ***federation***

Xrootd Fed *AAA project*

- USCMS (OSG) project [*] to ***develop and test*** tools for an ***xrootd federation within*** the ***CMS*** data management
    - ❖ ***prototype*** architecture was developed and tested [**]
    - ❖ ***16 sites*** took part to the prototype:
        - ❖ ***no DPM sites***. glite plugin version was not compliant(?);
        - ❖ after a successful test phase ***CMS is pushing to adopt this*** as an evolution of its data management.

Site	storage
T1_US_FNAL	dCache
T2_CH_CERN	Xrootd/EOS
T2_DE_DESY	dCache
T2_IT_Bari	StoRM
T2_IT_Legnaro	dCache
T2_IT_Pisa	StoRM
T2_UK_London_IC	dCache
T2_US_Caltech	bestman
T2_US_Florida	bestman
T2_US_MIT	bestman
T2_US_Nebraska	bestman
T2_US_Purdue	bestman
T2_US_UCSD	bestman
T2_US_Vanderbilt	bestman
T2_US_Winsconsin	bestman
T3_US_FNALPC	dCache



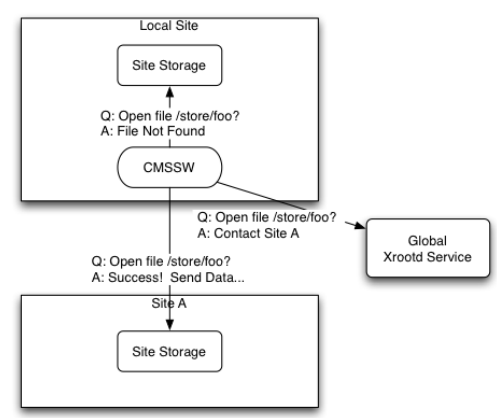


[*] https://twiki.grid.iu.edu/bin/view/Management/AnyDataAnyTimeAnyWhere
http://osg-docdb.opensciencegrid.org/0010/001025/001/AnyDataAnyTimeAnyWhere.pdf

[**] https://twiki.cern.ch/twiki/bin/view/Main/CmsXrootdArchitecture

Xrootd Fed deploying

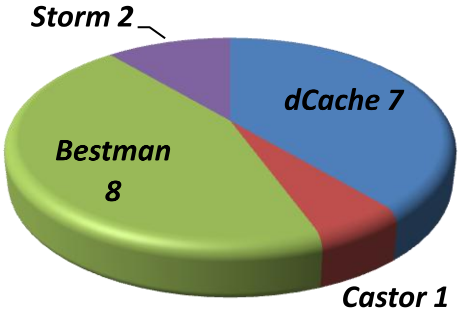
- ***CMS pushes*** all the sites to be at least ***"passively" in the fed*** within ***Christmas*** by enabling ***xrootd fallback***
  - only ***client side, integrated in CMSSW,*** no need to have a xrootd server
    - just a 2 lines change in the job config and tfc [*];



- ***next*** step is ***enlarging the federation***
  - members need a ***configured xrootd*** [**];
  - ***currently*** there are ***no DPM sites*** in the fed (to my knowledge).

New sites in the fed.

- ***T1_UK_RAL***
- ***T2_IT_Rome***
- ***T2_UK_SGrid_RALPP***



[*] https://twiki.cern.ch/twiki/bin/view/Main/ConfiguringFallback

[**]https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Xroot/ManualSetup#CMSfederation

Xrootd Fed deploying

- CMS pushes all the sites to be at least "passively" in

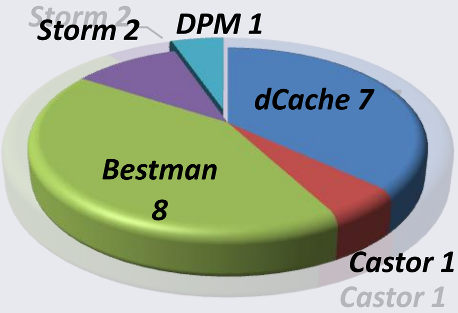
First Feedback

- ***setup*** and check of **basic functionalities** are ***ok;***
- few things to add/correct in the wiki;
- ***no*** real ***feedback on perfs*** yet.
- LLR =~ ***CMS only*** storage. What about ***multiple feds?***

- membe… ***…figured xrootd*** [**];
- ***curre…*** …here are ~~***no DPM sites***~~ in the fed (to my knowledge).

Entered the fed on 29/11

T2_FR_GRIF_LLR
T1_UK_RAL
T1_IT_Rome
T2_UK_SGrid_RALPP



[*] https://twiki.cern.ch/twiki/bin/view/Main/ConfiguringFallback
[**]https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/Xroot/ManualSetup#CMSfederation

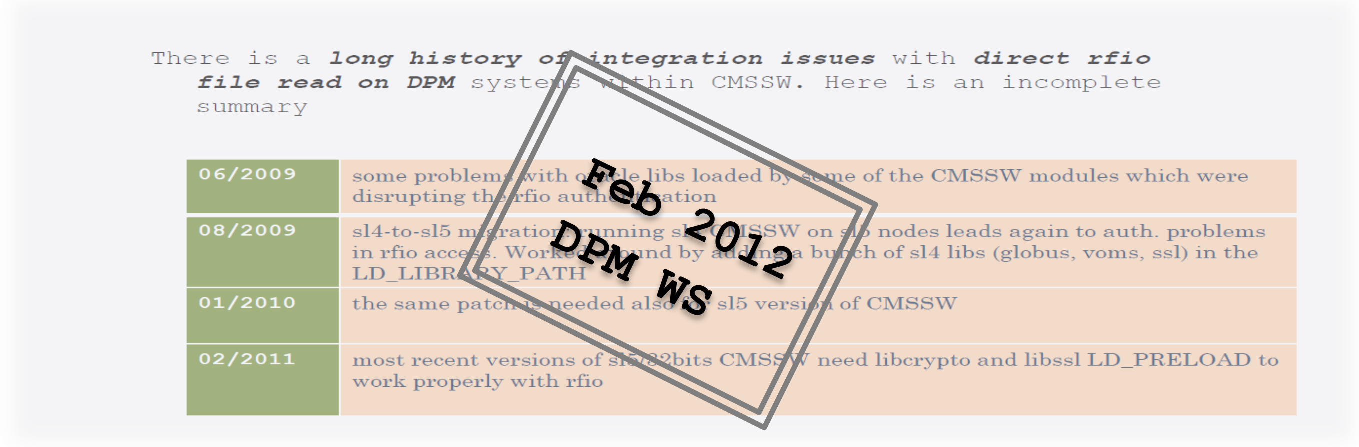
Integration setup

- All DPM/CMS sites ***use rfio plugin*** for direct ***file reading within CMSSW***
  - ❖ the evolution into the ***xrood federation*** may ***encourage*** sites to pass to *xrootd local access*
    - ❖ note that the local access and fallback/federation conf are ***decoupled;***
- ***stageout*** is performed by means of ***lcg-utils/srmcp/rfcp*** plugins
  - ❖ rfcp writes VOLATILE files so sites should be careful;
- ***PhEDEx*** implements a ***dpm namespace plugin*** for file validation/deletion (which uses dpns commands) and for datasets verification
  - ❖ ***standard and more performing*** interface for all PhEDEx agents (will substitute bash local scripts);
  - ❖ ***checksum verification*** of transfers relies now on ***FTS*** and is well integrated with DPM.

not really changed since 02/2012

Integration issues

And the "CMSSW vs DPM" blues goes on…



There is a ***long history of integration issues*** with ***direct rfio file read on DPM*** systems within CMSSW. Here is an incomplete summary

Feb 2012 DPM WS

06/2009	some problems with oracle libs loaded by some of the CMSSW modules which were disrupting the rfio authentication
08/2009	sl4-to-sl5 migration: running sl4 CMSSW on sl5 nodes leads again to auth. problems in rfio access. Worked around by adding a bunch of sl4 libs (globus, voms, ssl) in the LD_LIBRARY_PATH
01/2010	the same patch is needed also for sl5 version of CMSSW
02/2011	most recent versions of sl5/32bits CMSSW need libcrypto and libssl LD_PRELOAD to work properly with rfio

…

04/2012	some versions of CMSSW (5_3_X and older) need a LD_PRELOAD of liblcgdm.so to run on EMI-1 and EMI-2/sl5
07/2012	need for a LD_PRELOAD of libssl.so.10 in to run fine with EMI-2/sl6

Great effort from both CMS and DPM to debug and document workarounds[*]:

- CMS should ***reduce the shipped library*** to the essential (already in new rel.);
- ***still*** the effort relies on volunteer sites and admins;

[*] https://twiki.cern.ch/twiki/bin/view/CMSPublic/CompOpsT2DPMInstructions

Summary

- There is a ***wide community*** of **CMS T2/T3** sites deploying ***DPM storage***
  - ***~30% of T2/T3 sites*** corresponding to ***~20% of T2 resources***;
  - sites are ***well integrated*** in CMS Computing System and give ***important contribution*** with good performances;

- CMS data management is ***evolving*** with the deployment of a ***xrootd federation***
  - with the ***new xrootd plugin*** DPM sites ***should be ready*** to enter such federation;
  - ***first feedback*** from T2_FR_GRIF_LLR is ***good;***

- ***DPM/CMSSW integration problems*** are still not over
  - bad times with ((glite+EMI1+EMI2) x sl5/sl6) but so far ***all problems*** that appeared have been ***fixed (with workarounds)***;
  - the direction in which CMSSW is moving seems to be a good one for avoiding new problems in the future.

Backup definitions…

- ***CMSSW*:** core software framework (simulation, reconstruction analysis)
  - input files access based on plugins: posix, rfio (DPM/Castor), dcap (dCache), xroot, http, etc.;

- ***Trivial File Catalog (TFC)*:** site-local configuration xml file with regexp rules for lfn-pfn mapping
  - used by CMSSW, job submission tools, PhEDEx;
  - defines CMSSW input file access plugin (by the pfn protocol)

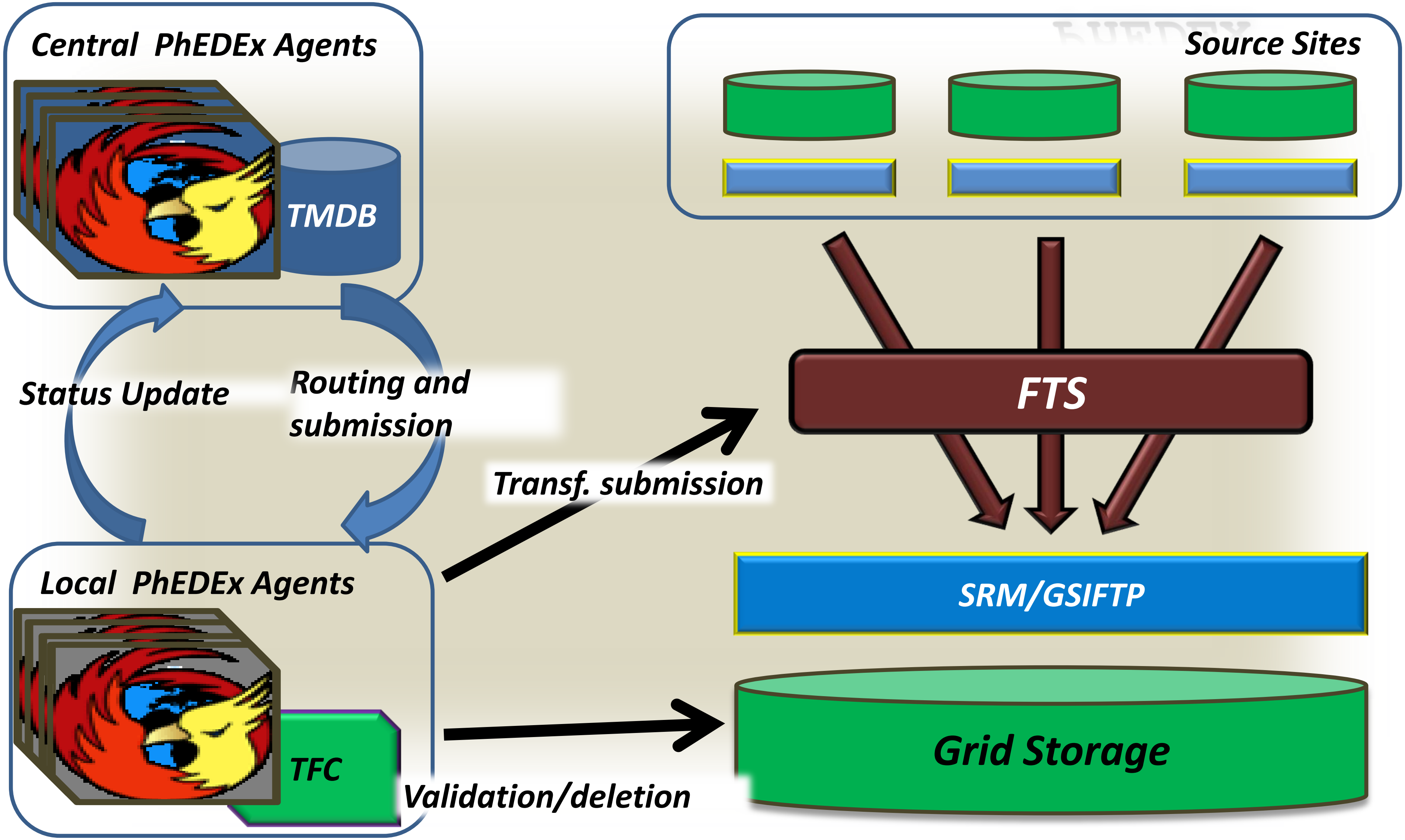
- ***Job Config*:** site-local configuration xml file with the information for CMS application
  - location of the TFC;
  - defines plugin to use for for output stageout;

Backup …definitions

- ***PhEDEx***: data transfer and placement system. Routes requested data from all possible sources
  - central agents and a DB (TMDB) at CERN: information about replicas and routes;
  - actual transfers performed by FTS;
  - local site agents: interaction with storage for transfer validation, file deletions and consistency checks

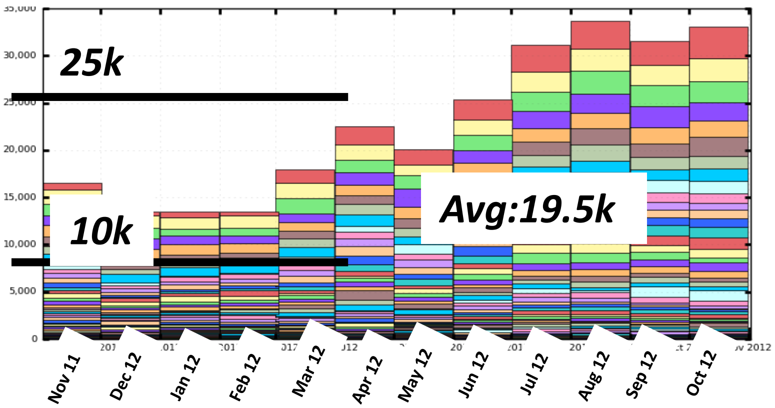
- ***Job submission tools*** (CRAB, ProdAgent…): implement the CMS data-driven grid model (jobs run where data stored).
  - manage transparently the interaction with Grid MW, with the CMS data bookkeeping tools and with monitoring;
  - use plugin method for stageout files onto different storage technologies;
  - plugins: srmv2 (dcache srm client), lcg-srmv2 (lcg utils), rfio (rfcp,…), posix, etc.

Backup


Central PhEDEx Agents
TMDB
Source Sites
Status Update
Routing and submission
FTS
Transf. submission
SRM/GSIFTP
Local PhEDEx Agents
TFC
Grid Storage
Validation/deletion

Backup performances

T2 sites months/month CPU time in 2012

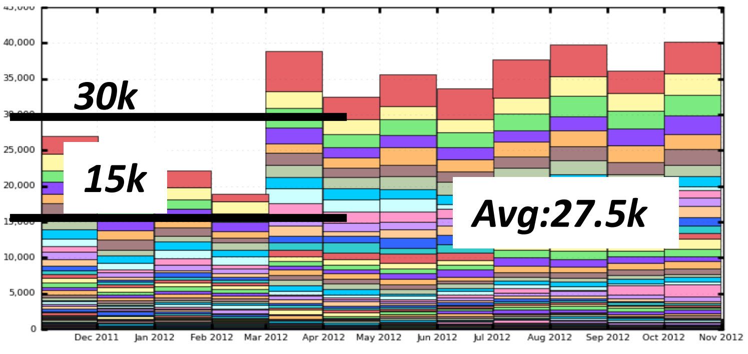


25k

10k

Avg:19.5k

T2 sites months/month WC time in 2012



30k

15k

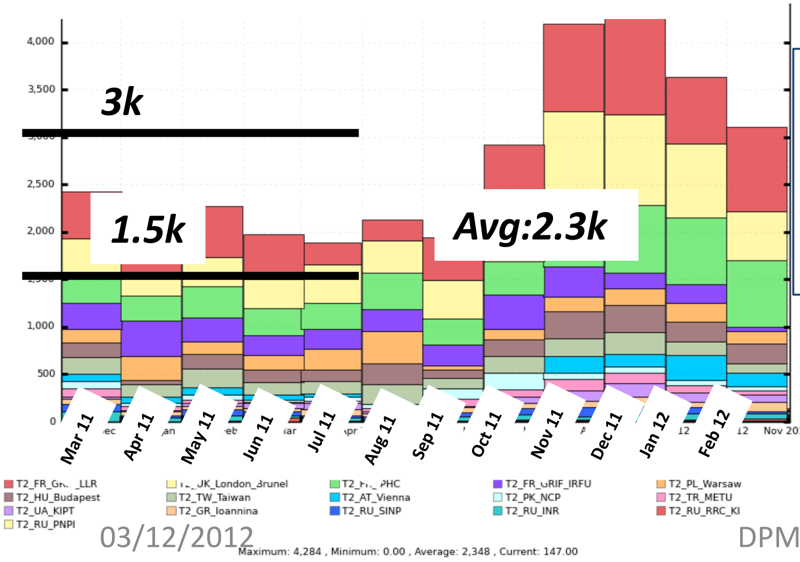
Avg:27.5k

DPM sites contribution ~ 12%

CMS sites eff. ~ 71%

DPM sites eff. ~ 76% [*]

DPM sites months/month CPU time in 2012



Maximum: 4,284 , Minimum: 0.00 , Average: 2,348 , Current: 147.00

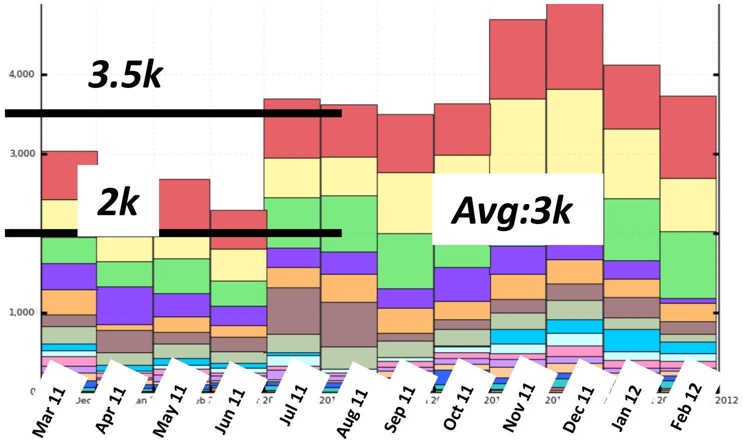
3k

1.5k

Avg:2.3k

[*] *Efficiency depends on many variables* local to sites and to jobs. Here I just want to show that *DPM substantially performs in line with any other storage system.*

DPM sites months/month WC time in 2012



3.5k

2k

Avg:3k