CMS and Alice at CC-IN2P3

Artem Trunov
CMS and Alice support

Topics

- Numbers, policies (overview)
- Tools
- Storage model

CMS numbers – 2008

(Credit: Dave Nebold, Bristol)

- Data reduction chain
  - RAW – 1.5 MB
    - RECO - 0.25 MB
      - AOD - 0.05 MB
- Events
  - 92 days of running
  - ~300Hz trigger rate
    - 1.2 B events - 450 MB/s
- Size
  - FEVT (RAW+RECO) = 1.2 PB
  - ReReco = 300TB x 3 = .9 PB
  - AOD = 60 x 4 = 240 TB
  - Sim to real ratio is 1:1 (+30% of size overhead)

Alice Numbers

(credit: Yves Schutz)

- Running conditions of a standard year:
  - pp: 10^7 seconds leading to 10^9 events/year
  - AA: 2×10^6 seconds of AA leading to 2×10^8 events/year
- A similar amount of MC data will be produced yearly
  - pp: 10^9 events/year
  - AA: 2×10^7 background events/year reused 10 times with different signals
- The recording rate is 100Hz in pp and AA
- The corresponding data rate is 0.11GB/s for pp and 1.38GB/s for AA
- Startup scenario

	2007	2008	2009	2010
Pp	**7%**	**40%**	**60%**	**100%**
AA	**0%**	**10%**	**50%**	**100%**

New 2008	CERN				External			Total
	Tier0	CAF	Tier1	Total	Tier1s	Tier2s	Total	
CPU(MSI2K)	1.80	0.52	1.44	1.94	6.92	7.76	14.68	16.61
DISK(PB)	0.024	0.064	0.97	1.06	1.84	0.95	2.79	3.85
MS (PB)	0.79	-	0.83	1.62	3.28	-	3.28	4.90

CMS Policies

- RAW on disk 100%
  - Copy at CERN (on tape) and each T1
- RECO on disk 100%
  - Second pass RECO on disk and tape at production T1
- AOD on disk
  - “Current” (most recent version + 1 older) 100%
  - On tape at production T1 site
  - Replicated on disk at all other T1s
- SIM data – 10% on disk
- Totals:
  - ~6PB of disk, ~10 PB on tape

CMS and Alice site roles

- Tier0
  - Initial reconstruction
  - Archive RAW + REC from first reconstruction
  - Analysis, detector studies, etc
- Tier1
  - Archive a fraction of RAW (2nd copy)
  - Subsequent reconstruction
  - “Skimming” (off AOD)
  - Archiving Sim data produced at T2s
  - Serve AOD data to other T1 and T2s
  - Analysis
- Tier2
  - Simulation Production
  - Analysis

CMS Transfers

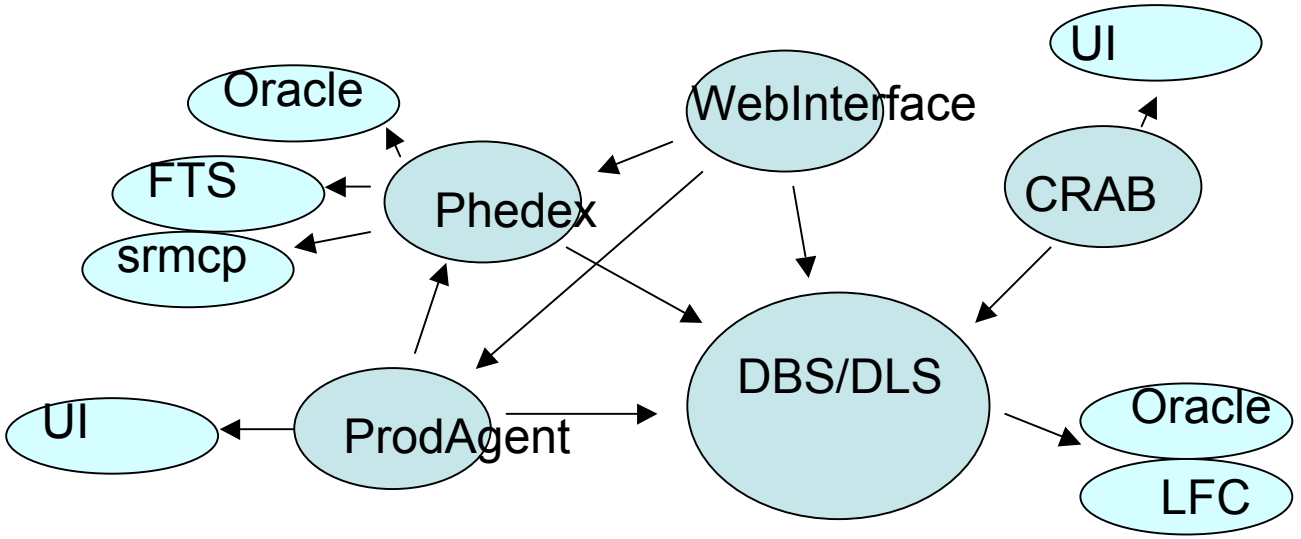
- T0
  - RAW+RECO to T1s at ~250MB/s
  - Some services to “orphan” T2s
- T1 (*number for a typical T1 like IN2P3*)
  - RAW+RECO from CERN – 25 MB/s
  - AOD exchange - ~100MB/s in and out (14 days)
  - SIM data from T2s – 15MB/s
  - All data to T2s - ~100MBs (30 days)
- *In CMS any T2 can transfer from Any T1 – not a rigid Tier model*

Alice Policies and Transfers

- T0 → T1
  - Distributed copy of raw data
  - Distributed copy of ESD from first pass reconstruction
- T1 ↔ T1
  - Distributed copy of ESD from MC and 2nd+3rd pass reconstruction
  - Distributed copy of AOD from scheduled analysis
- T2 → T1
  - MC data (raw and ESD) produced exclusively in T2s
  - AOD produced by user driven analysis
- T1 → T2
  - ESD and AOD from scheduled analysis
- *Ideally, ALICE want to implement a cloud model, where site have no specific roles*

CMS Tools

- Bookkeeping tools
  - Description of data datasets, provenance, production information, data locatoin
  - *DBS/DLS (Data Bookkeeping Service/Data Location Service)*
- Production tools
  - Manage scheduled or personal production of MC, reconstruction, skimming
  - *ProdAgent*
- Transfer tools
  - All scheduled intersite transfers
  - *Phedex*
- User Tools
  - For end user to submit analysis jobs.
  - *CRAB*



Bookkeeping Tools

- Centrally set up database
- Has different scopes (“local”, “global”), to differentiate official data from personal/group data.
- Use Oracle server at CERN.
- Data location service uses either Oracle or LFC.
- *Very complex system*

Production Agent

- Managing all production jobs at all sites
- Instances of it are used by production operators
  - 1 for OSG + 3 EGEE sites
- Merging production job output to create a reasonable size files.
- Registering new data in Phedex for transfer to final destination
  - MC data produced at T2 to T1.
- All prod info goes to DBS

Phedex

- The only tool that is required to run at all sites
  - Work in progress to remove this requirement
- Set of site-customizable agents that perform various transfer related tasks.
- Uses Oracle at CERN to keep it's state information
- Can use FTS to perform transfer, or srmcp
  - Or another mean, like direct gridftp, but CMS requires SRM at sites.
- Uses 'pull' model of transfers, i.e. transfers are initiated at the destination site.
- One of the oldest and stable SW component of CMS
  - Secret of success: development is carried out by the CERN+site people who are/were involved in daily operations
- Uses someone's personal proxy certificate

CRAB

- A user tools to submit analysis jobs
- Users specifies a data set to be analyzed, his application and config.
- CRAB:
  - locates the data with DLS
  - splits jobs
  - submits the to the Grid
  - tracks jobs
  - collects results
- Can be configured to upload job output to some Storage Element.

Other CMS tools

- Software installation service
  - In CMS software is installed in VO area via grid jobs.
- Dashboard
  - Collecting monitoring information (mostly jobs) and presenting through a web interface.
- ASAP
  - Automatic submission of analysis jobs using web interface.

CMS site services

- “CMS does not require a VO Box”
  - But it needs a machine at site to run Phedex agents
- However VO Box is a very useful concept that CMS can take advantage of.
  - Gsissh access
    - Passwordless! Uses your grid proxy
  - Proxy-renewal service
    - Automatically retrieves your personal proxy, required to make transfer, from myproxy server.
- Other services can be run on the VO Box, like ProgAgent.
- At CC-IN2P3 we have migrated CMS services to a standard VO Box, and also made a setup shared with Alice services, so one can interchange a VO Box in case of failure.

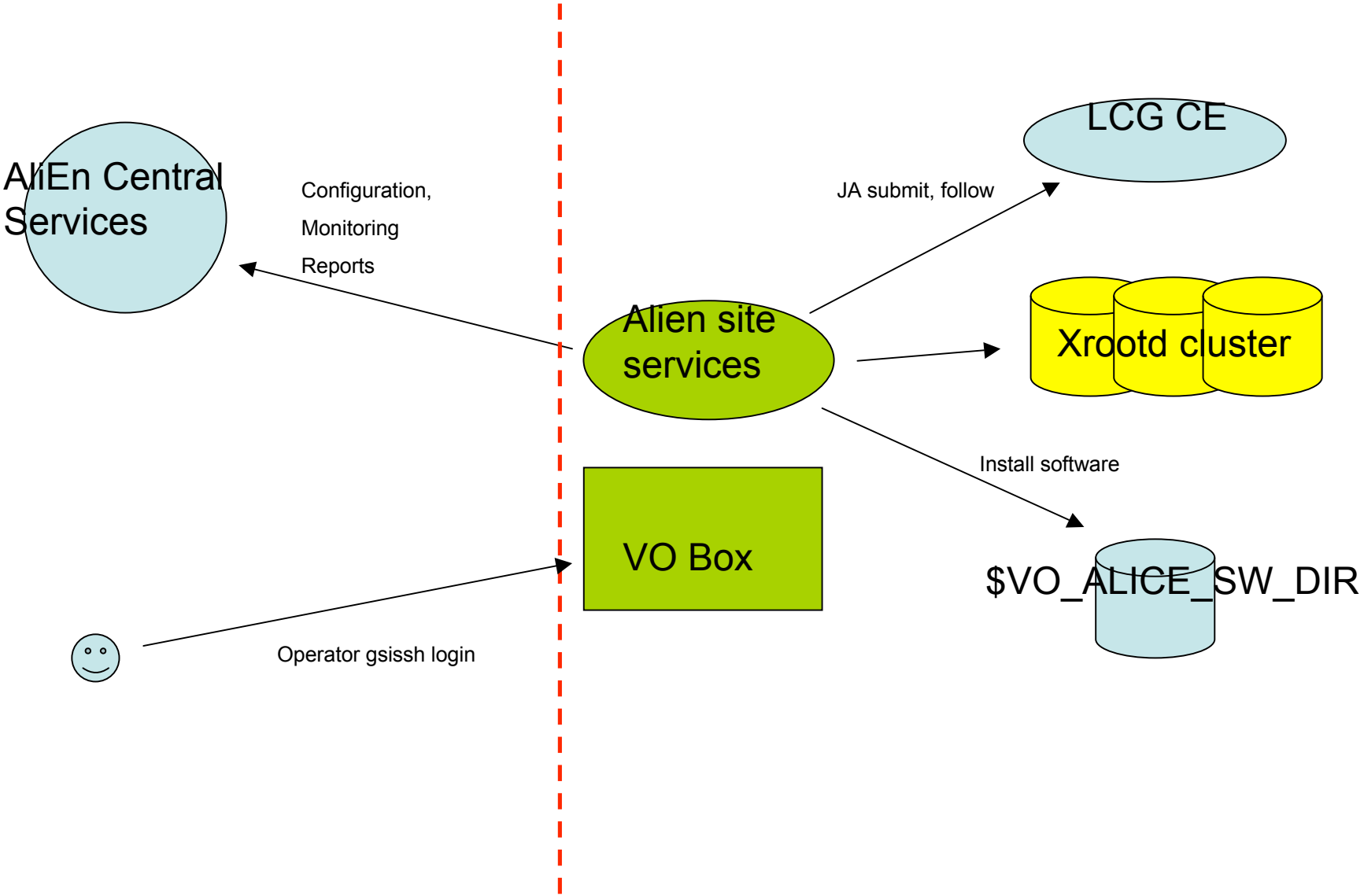
CMS also requires a site support person

- Some needs to install Phedex and look after transfers.
- It may seem to much, but this works for CMS
  - Big collaboration, complex computing model, and some one needs to bring this in order
- These people also contribute back to the experiment

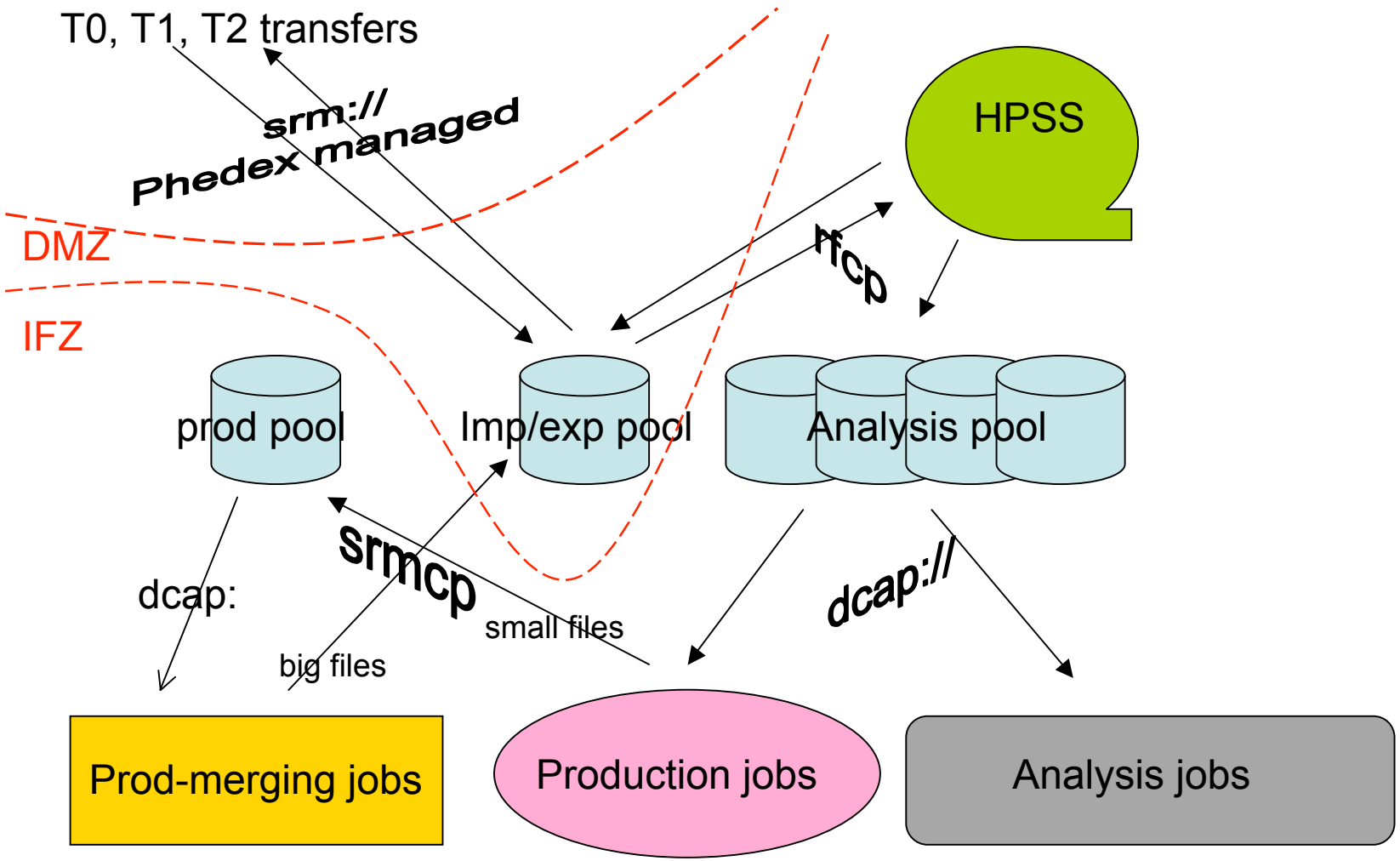
Alice services - AliEn

- “Alice Environment’ - independent implementation of GRID middleware.
- Central services:
  - Job Queue
  - File catalog
  - Authorization
- Site services include:
  - Computing element
    - With LCG/gLite support
  - Storage element – xrootd protocol
  - PackMan - SW installation
  - File Transfer Deamon
    - with FTS suport
  - ClusterMonitor
  - MonalLisa agent
- Services run at each site on the standard VO Box, optimized for low maintenance and unattended operations.
- Job Agent model
  - Once on a WN, they contact a database and pick up real tasks
- Also a fully featured user login shell

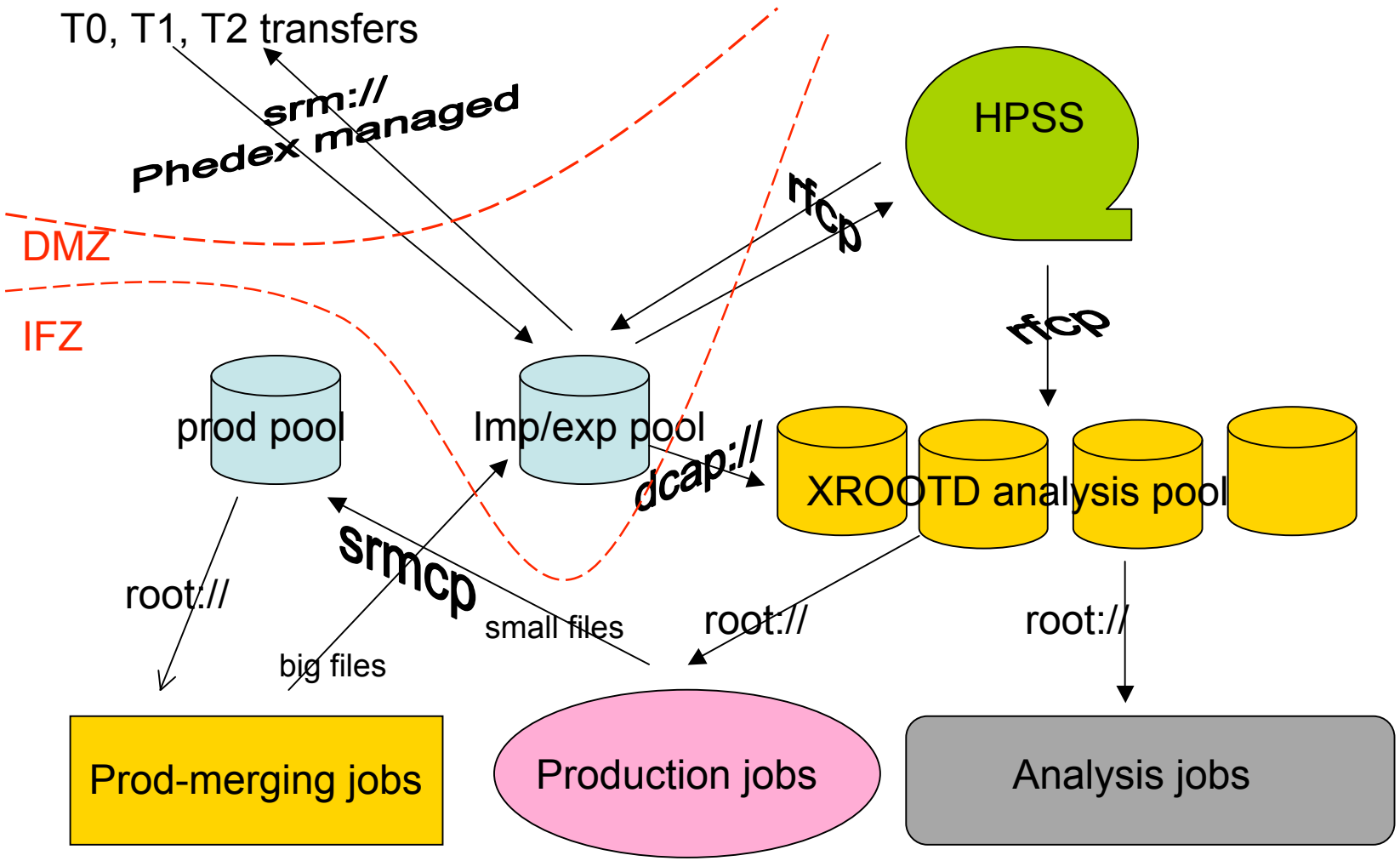
Alice Services at a site



CMS storage at CC-IN2P3



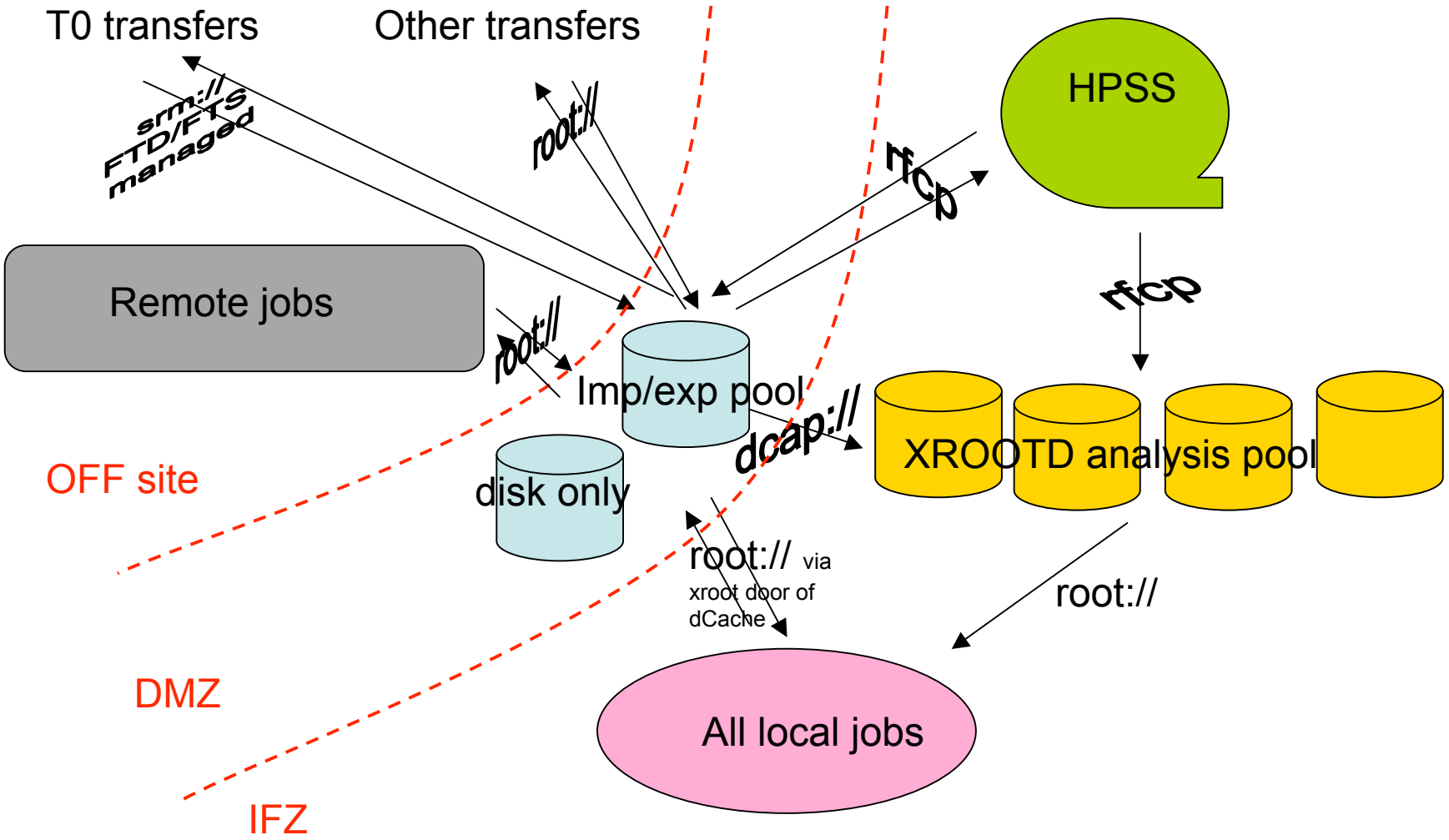
CMS storage at CC-IN2P3



Experiments with xrootd

- We see the following use of xrootd: large **read-only** disk cache in front of tape storage
  - No need of security overhead
  - Very low maintenance effort
  - Transfer (srm) and fancy data management functions provided by dCache

Alice storage at CC-IN2P3



Transfers are successful

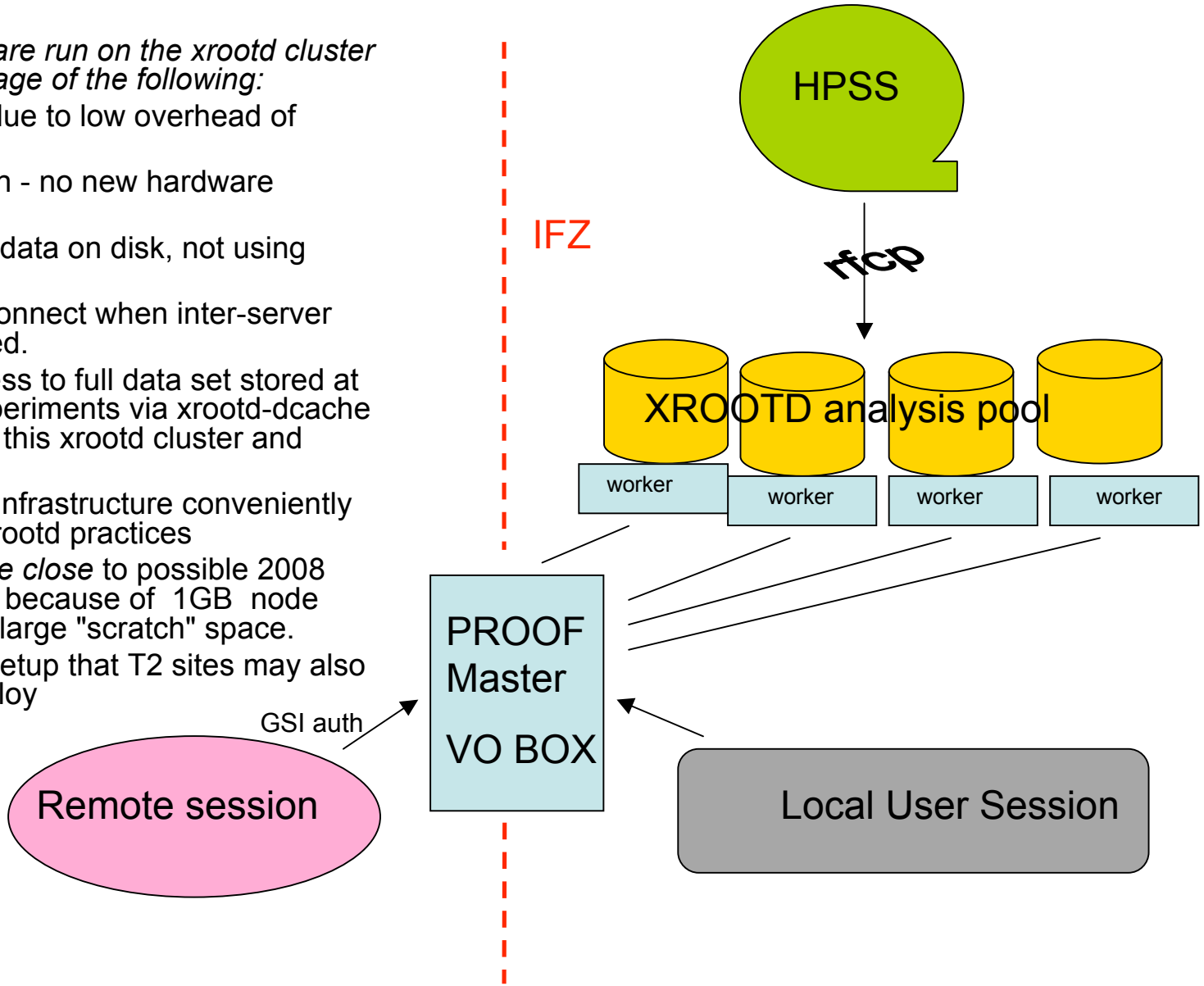
- We are able to transfer at >250 MB/s
- Stability of the site's storage and transfers are recognized by experiments and LCG.
  - Credit goes to Storage group, in particular to Lionel and Jonathan.

Another tests - PROOF

- PROOF is a part of ROOT and gives a mechanism to parallelize analysis of ROOT trees and ntuples.
  - Requires high speed access to data, not often possible in current setups
    - WN and server bandwidth limitation
      - WN: 1GB per rack -> 1.5MB/s per cpu core.
      - Servers: usually have 1Gb/s, but disk subsystem may perform better.
- Solution endorsed by PROOF team:
  - Large dedicated clusters with locally pre-loaded data
    - Not preferable at the moment.

PROOF at CC-IN2P3

- *PROOF agents are run on the xrootd cluster and take advantage of the following:*
- free cpu cycles due to low overhead of xrootd
- zero cost solution - no new hardware involved
- Direct access to data on disk, not using bandwidth
- 1GB node interconnect when inter-server access is required.
- transparent access to full data set stored at our T1 for all experiments via xrootd-dcache link deployed on this xrootd cluster and dynamic staging
- management of infrastructure conveniently fit into existing xrootd practices
- this setup is *more close* to possible 2008 PROOF solution because of 1GB node connection and large "scratch" space.
- this is a kind of setup that T2 sites may also considers to deploy



The End!

CMS and ALICE will be ok!