ALICE Status experience with SC4

F.Carminati

September 25, 2006

Summary

- Resource situation
- Status of DC4
- Major issues
- Outlook
- Conclusion

Resource situation

- For pp (similar to others)
  - Quasi-online data distribution and first reco @ T0
  - Further reconstructions @ T1's
- For AA
  - Calibration, alignment, pilot recos and partial data export during data taking
  - Data distribution and first reco @ T0 in four months after AA run (shutdown)
  - Further reconstructions at T1s
- Scheduled analysis at T1s, end user analysis at T2s
- T0: storage of RAW, calibration data and first-pass ESD's
- T1: storage of a portion of RAW and one copy of data to be safely kept

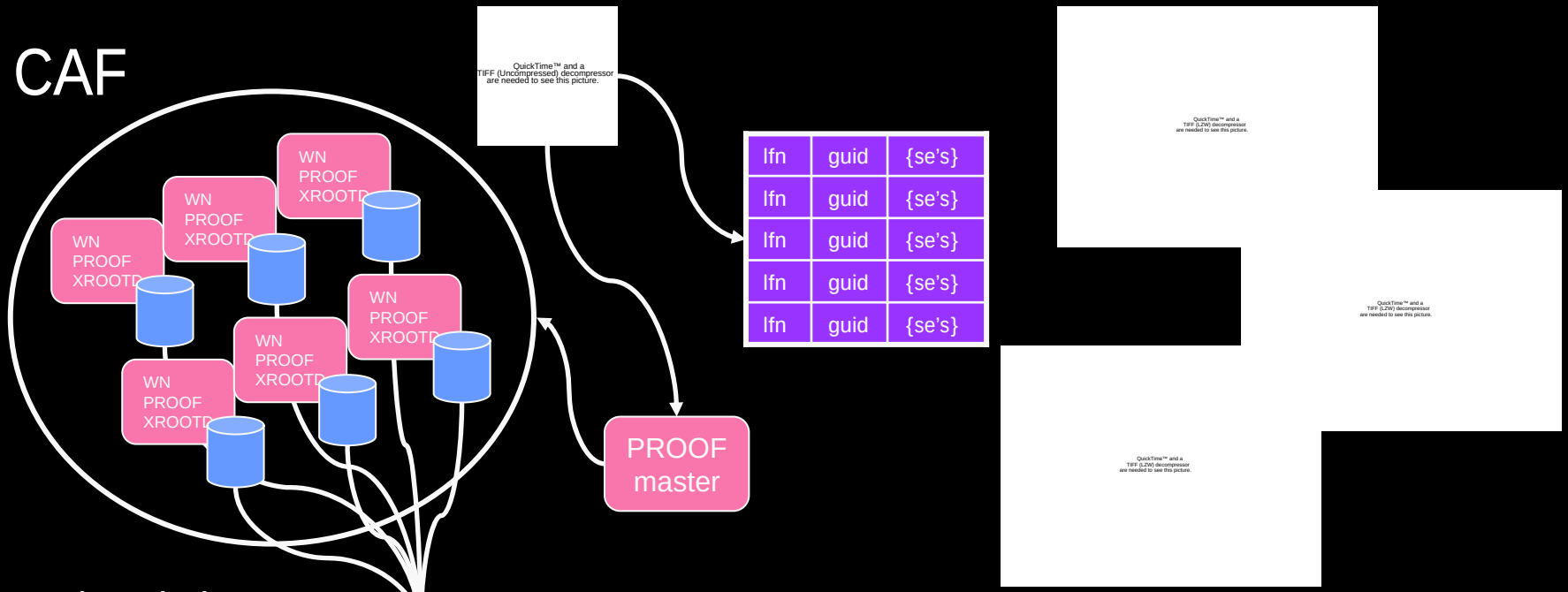
Pledged by external sites versus required (new LHC schedule) MoU only

		2007		2008		2009		2010	
		T1	T2	T1	T2	T1	T2	T1	T2
CPU	Requirement (MSI2K)	3.0	4.2	10.2	9.6	18.4	16.0	22.9	19.0
	Missing %	-14%	-14%	-35%	-36%	-44%	-53%	-36%	-50%
Disk	Requirement (PB)	1.0	0.8	4.2	1.6	7.9	4.0	9.8	5.3
	Missing %	16%	-5%	-35%	-7%	-44%	-45%	-34%	-42%
MS	Requirement (PB)	2.0	-	7.0	-	14.0	-	20.9	-
	Missing %	-28%	-	-51%	-	-53%	-	-53%	-

Resource situation

- We are trying to discuss with FAs and to find new resources
  - But we will certainly not cover the deficit
- We are reassessing the needs
  - But this tends to push them up rather than down
- The deficit is so large that it hardly makes sense to develop an alternative within the pledged resources
  - At the moment the loss in scientific output would be too high
- If we could reduce the gap (10%-20%), then it would make sense to develop a set of alternative scenarios
- If we cannot, then the investment by the Fas to build ALICE will be only partly exploited

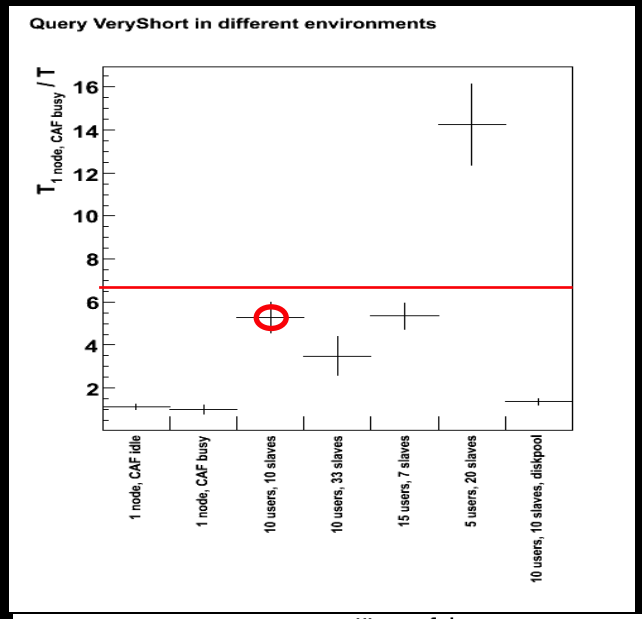
CAF

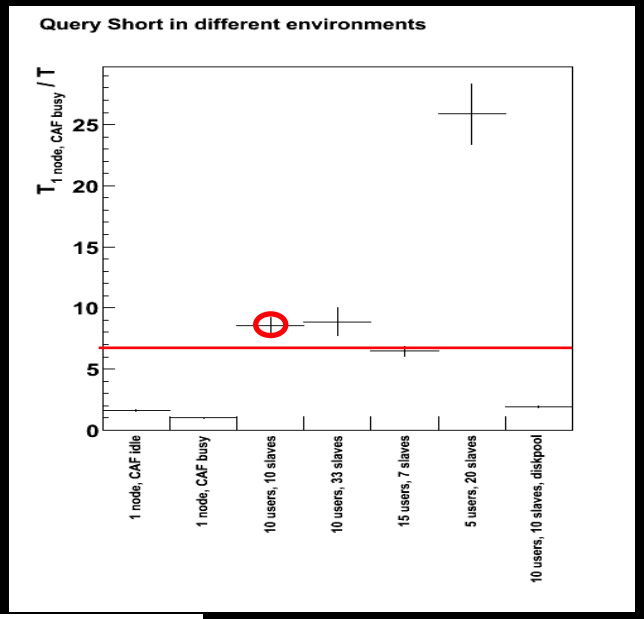


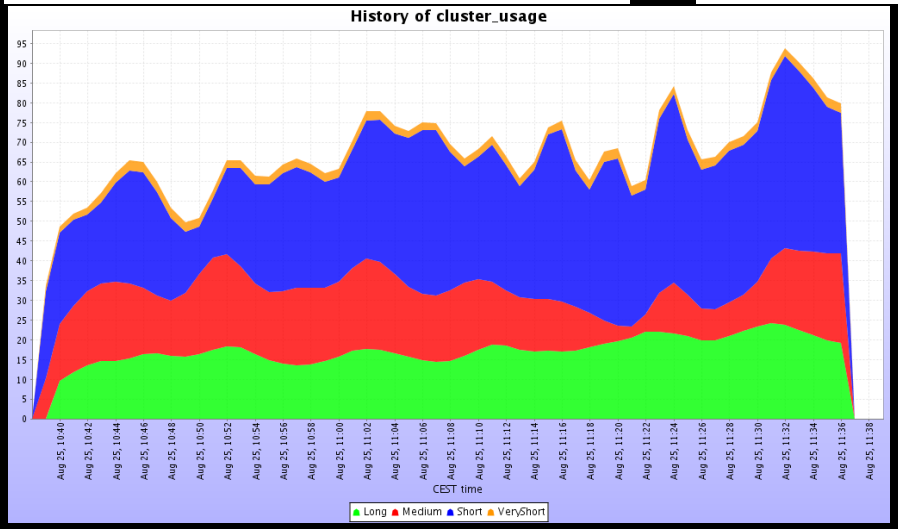
The whole CAF becomes a xrootd cluster

- CAF capacity approx 1.6MSI2k
- Reco
  - pp 1MB@40kSI2k•s: 40ev/s@40MB/s
  - HI 12.5MB@3600kSI2k•s: 0.5ev/s@6.5MB/s
- Analysis
  - pp 50kB@0.2kSI2k•s: 80kev/s@4GB/s
  - HI 5MB@2kSI2k•s: 800ev/s@4GB/s
- Calibration
  - Anything between the two above

CAF performance







- Still several issues to be solved but the progress is steady
- Strong support from the ROOT/PROOF team

Brief summary of PDC'06 goals

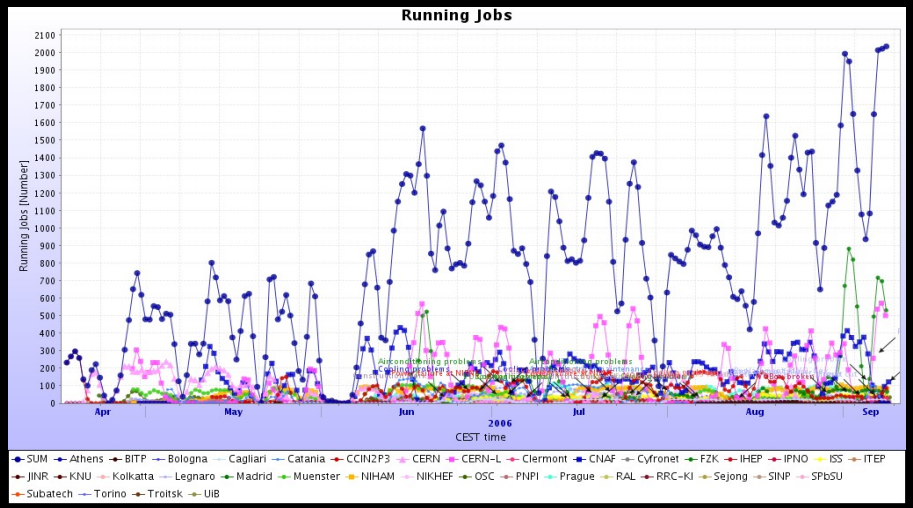
- Simulation for detector and software performance studies
- Verification of the ALICE distributed computing model
  - Integration and validation of GRID components
    - LCG services (Resource Broker, UI, L-File Catalogue, FTS...) & VO-Box
    - AliEn central services (File Catalogue, job submission and control, task queue, monitoring)
  - Distributed calibration and alignment framework
  - Full data chain – RAW data from DAQ, registration in the AliEn FC, first pass reconstruction at T0, replication at T1
  - Computing resources – verification of scalability and stability of the on-site services and building of expert support
  - End-user analysis on the GRID

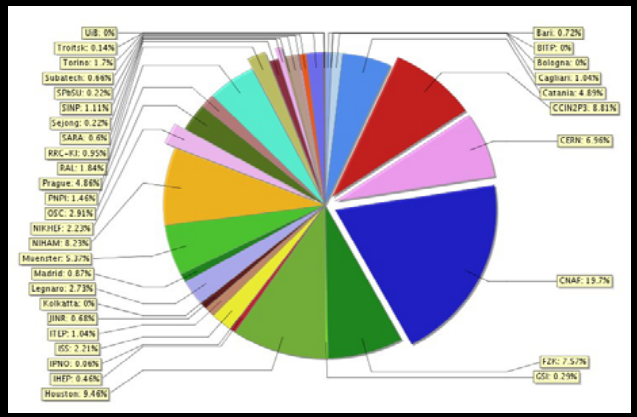
History of PDC'06 (2)

- Gradual inclusion of sites in the ALICE Grid
  - 6 T1s: CCIN2P4, CERN, CNAF, GridKA, NIKHEF, RAL
  - 30 T2s
  - Spain just joined, Japan (Hiroshima) may join soon
  - Successful meeting with with NGDF on 19/9/06
- Available CPU = 2000 (expected ~4000)
  - Competing for resources with the other LHC experiments
  - Computing centres are waiting for the last moment to buy hardware – will get more for the same price
  - Expect additional resources from Nordic Countries, US and Korea
- 0.5 PB registered in CASTOR2
  - 300kFiles (1.6GB/file) combined in archives for optimal load on MSS
- 6M p+p events in 60 runs (ongoing)
  - ESDs, simulated RAW and ESD tag files
- 50k Pb+Pb event in 100 runs
  - ESDs and ESD tag files

History of PDC'06

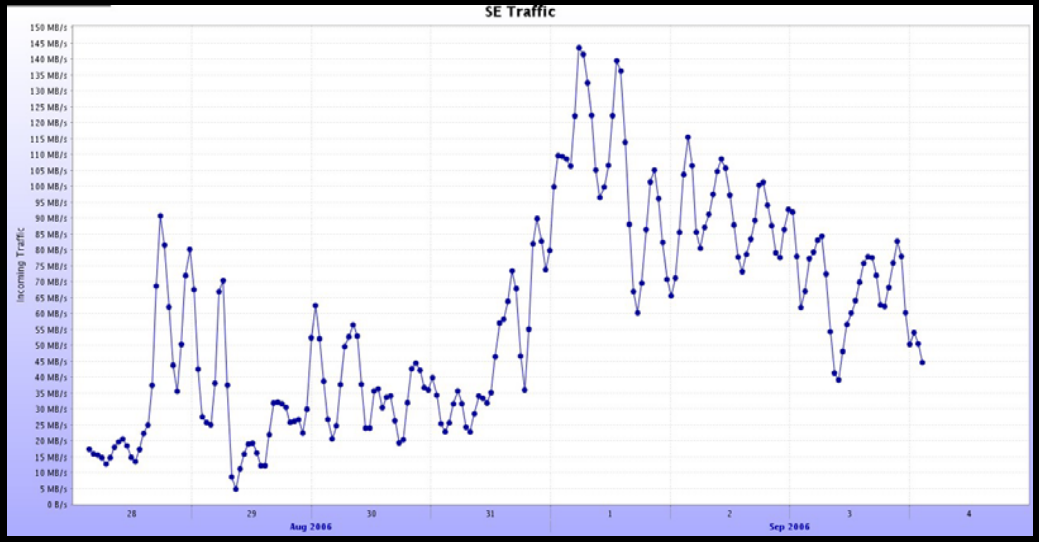
- Resources come 50% from T1s, 50% from T2s – as pledged
  - The role of the T2 is very high!
  - Equivalent to 500 CPUs running continuously





Data movement

- Produced data is sent to CERN
  - Up to 150 MB/sec data rate (limited by the available CPUs) – ½ of the rate during PbPb data export
- Registration of RAW
  - Dummy registration in AliEn and CASTOR2 – continuous since July 15
  - Registration and automatic reconstruction of TPC test data

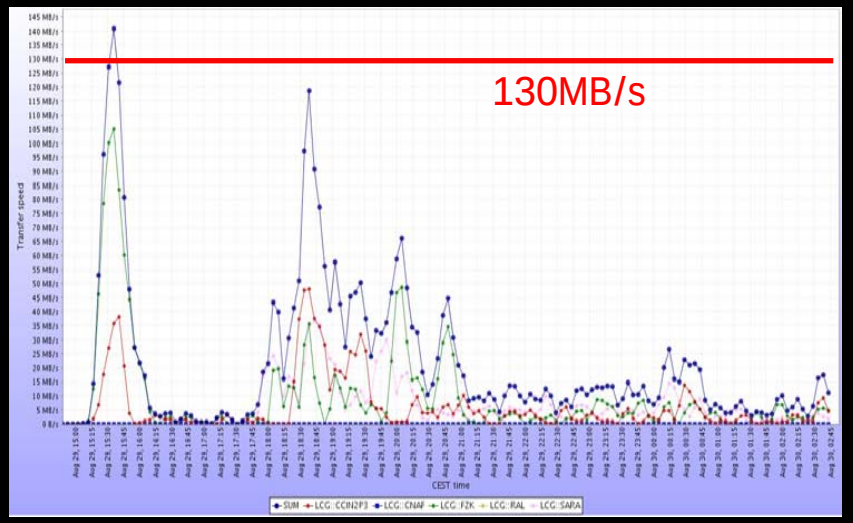


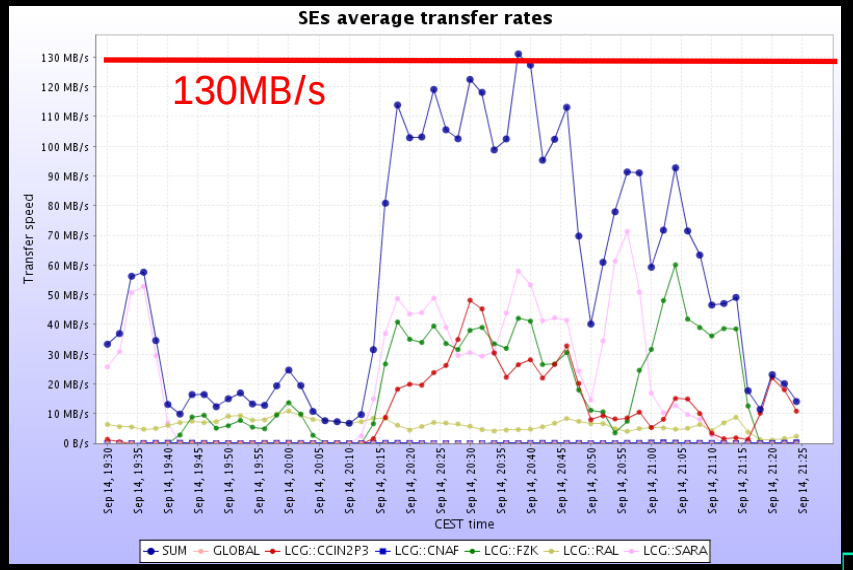
The FTS Service for ALICE

- Lower level tool for data replication between sites
- Used as plug-in in the AliEn FTD that runs in all VO-BOXES (T1 & T2)
  - Interfaced to LFC and SRM SE
- The main goal is the testing of the FTS stability and the integration with FTD
- Phases:
  - T0-T1: Migration of raw data and 1st pass reconstructed data
    - 7 days @ 300MB/s distributed according the MSS resources pledged
    - 5 T1 sites: CCIN2P3, SARA, CNAF, RAL, GridKa
  - T1-T2-T1: Transfers of ESD and AOD (T1-T2) and MC data for custodial storage (T2-T1)
    - 2 days @ rates specified: “Plans of the ALICE FTS transfers 2006”, 06/06
  - T1-T1: Replication of ESD and AOD
    - 2 days

Status of the transfers

- Problems with CASTOR2 as source (CERN) and sink (CNAF)
- FTS server at CERN hanging
- Problems with the access to the catalogue in all sites
- Instabilities in the VOBOXEs (CERN and SARA)
- From the ALICE side
  - Increased the size of the transferred files
  - Increase the number of simultaneous transfers per site

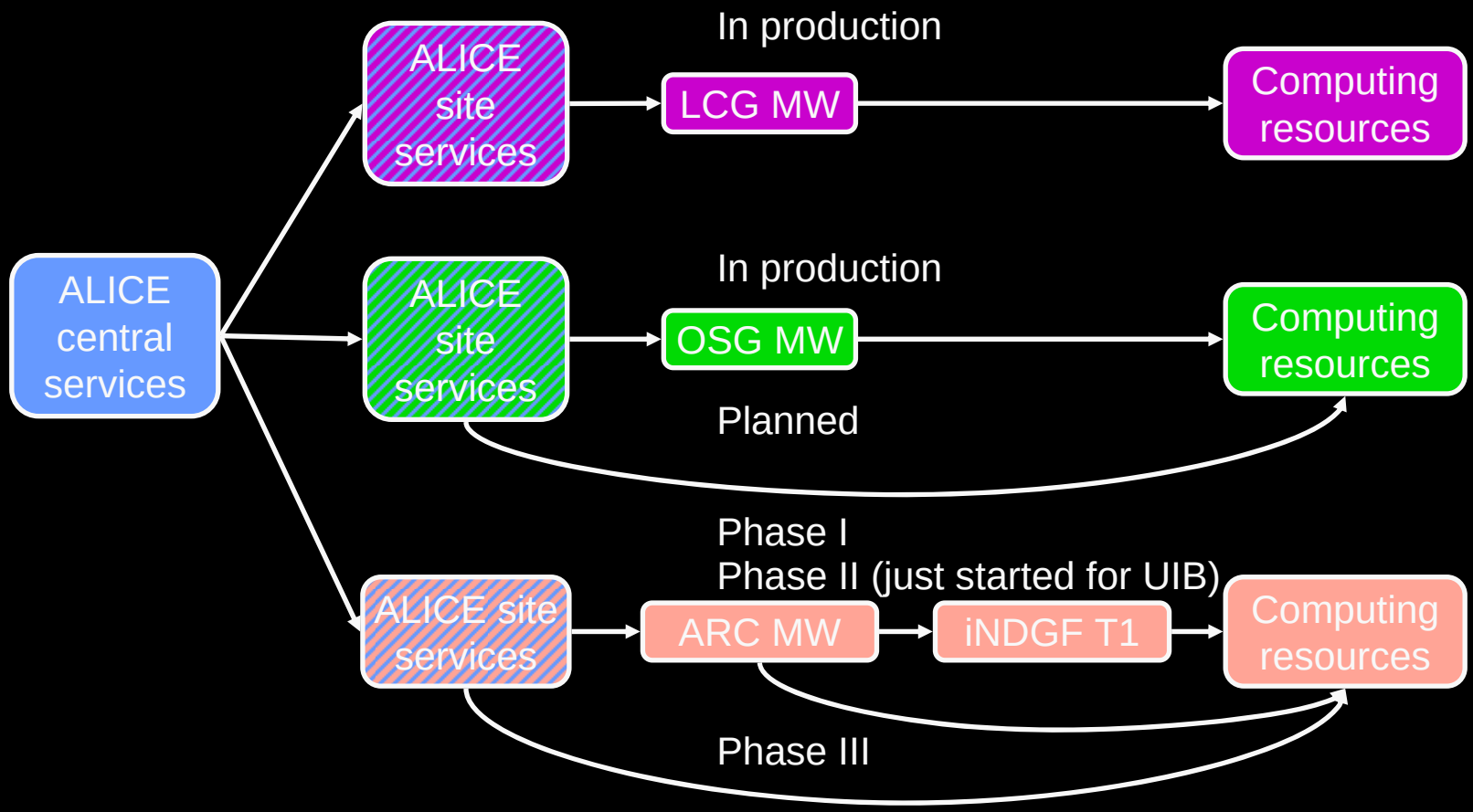




MW status

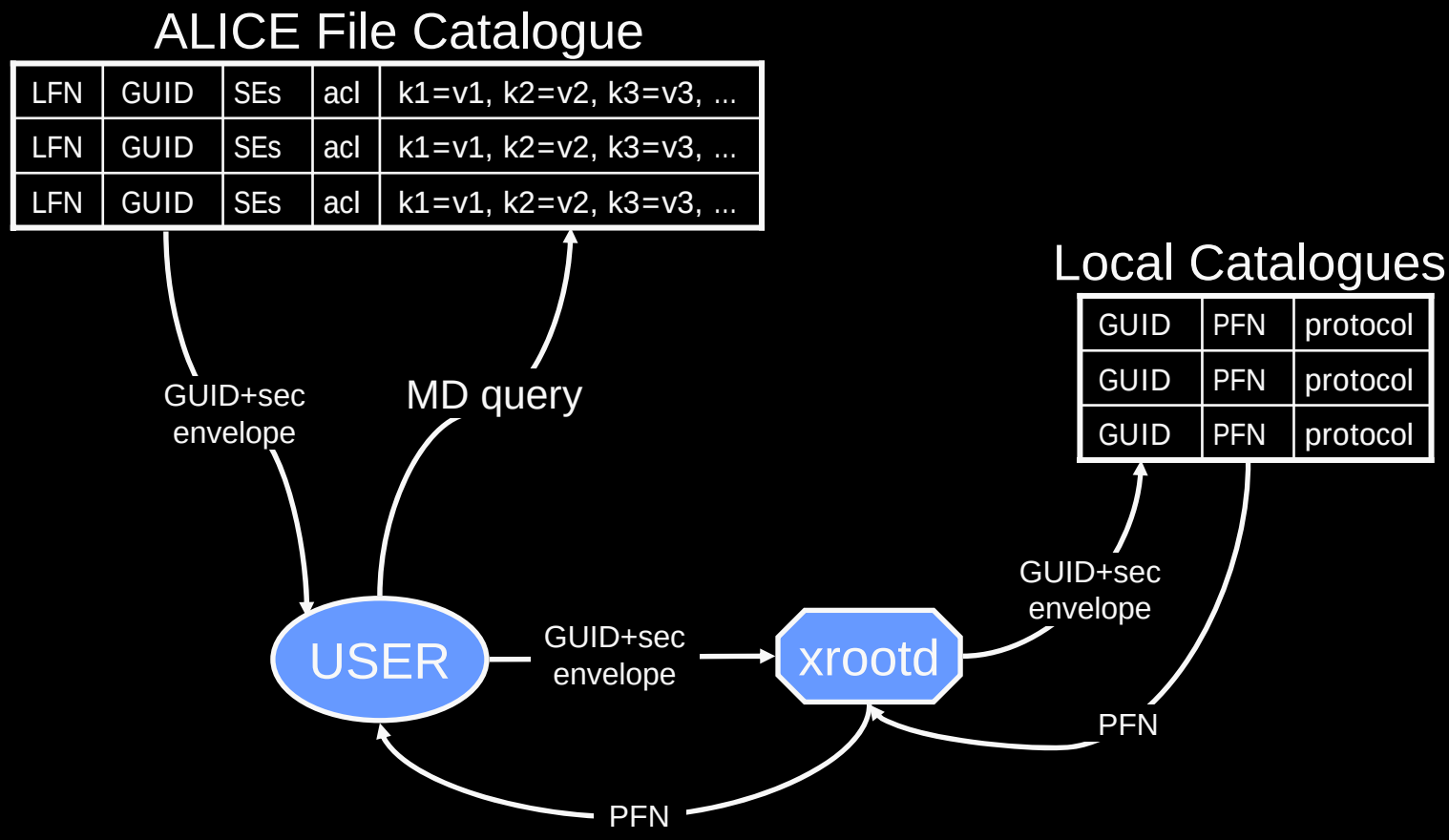
- We are using most services of LCG, complemented by ALICE specific ones that are required by our computing model
- ALICE specific services are installed centrally at CERN and on a single node in each computing centre (VO-Box)
- The design is evolving on the basis of the feedback
- Workload management seems to be under control
- Storage is still developing
  - The decision to use xrootd is excellent technically but requires developments, which however are ongoing
  - Not particularly depending on SRM functionality – it has to be there and stable

ALICE GRiD model

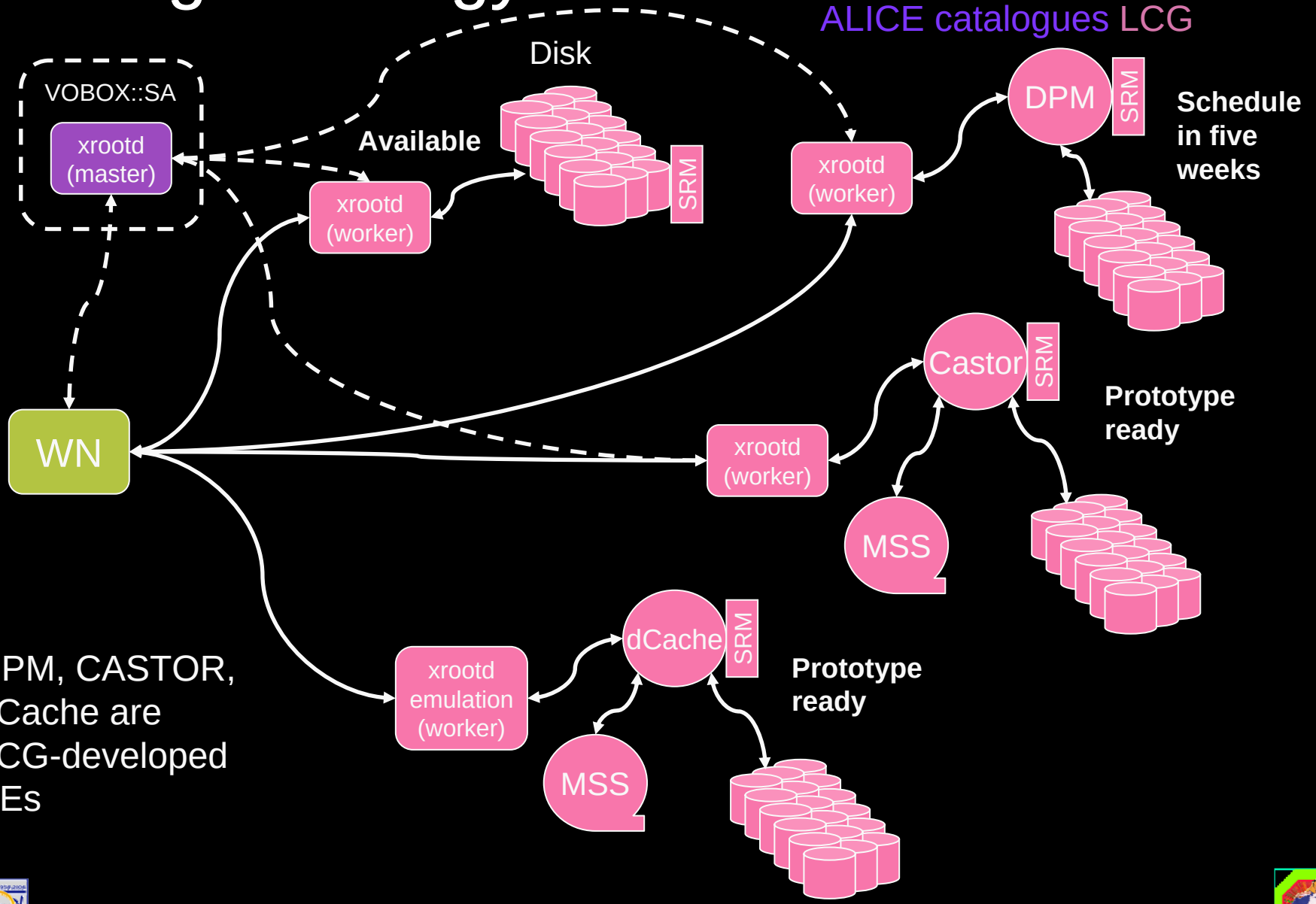


ALICE Files

/year/acc period/run/…



Storage strategy



Computing strategy

- Jobs are assigned where data is located
  - We use VOMS groups and roles moderately
- WMS efficiency not an issue thanks to JAs
- Resources are shared
  - No “localization” of groups
  - Equal Group/Site Contribution and Consumption will be regulated by accounting system
  - Prioritisation of jobs in the central ALICE queue
- Data access only through the GRID
  - No backdoor access to data
  - No “private” processing on shared resources

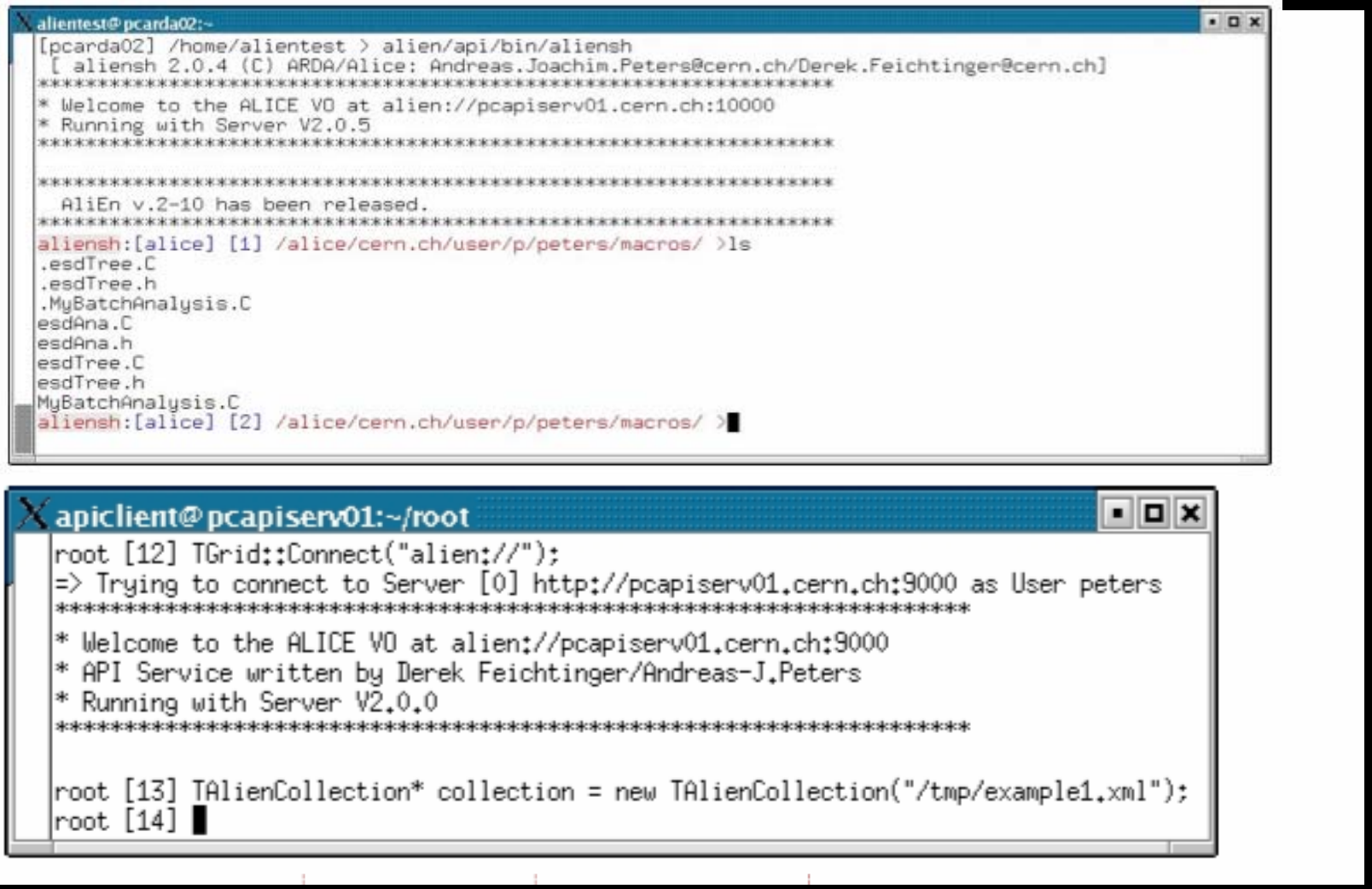
Analysis model

- Two types
  main difference: data access patterns, storage, code change frequencies
  - Scheduled
    - Analyses all data of a given type
    - Centralised – like data filtering for “Sub-Analysis”
    - Output typically ESD/AOD (+ control histograms)
  - Chaotic
    - Focused on single physics tasks
    - Based on filtered data
    - Many iterations on “random” subsamples of data
    - Output typically histogram files + event lists

Tier 1

Tier 1/2

ROOT / AliEn UI



Main requirements from LCG

- Major FTS robustification and optimisation
- xrootd interfaces to DPM and CASTOR2
- Inclusion of xrootd in the standard storage element would really help
  - And probably “cost” very little
  - We have no need of GFAL
- Implementation of glexec
  - First on the testbed and then on the LCG nodes
- Major improvement in the overall stability of the system

Conclusions

- Development and deployment of our distributed computing infrastructure is proceeding
  - We cannot honestly say that we have today a working system (AliEn+other MW) but progress is steady
  - Some developments from LCG are on the critical path and we depend on them – these should be pursued vigorously
    - FTS, xrootd->(DPM, CASTOR2), glexec
- The manpower situation has improved, but any perturbation (reduction or loss of key people) would be unrecoverable
  - The EGEE/ARDA contribution is instrumental
- The resource situation is so bad that we cannot even attempt yet a rescaling
  - We strongly hope to reach soon the situation where such an exercise can be done meaningfully


"That's
all
folks!"