Update on ALICE Computing

F.Carminati

November 08, 2006

ALICE computing model

- For pp similar to the other experiments
  - Quasi-online data distribution and first reconstruction at T0
  - Further reconstructions at T1's
- For AA different model
  - Calibration, alignment, pilot reconstructions and partial data export during data taking
  - Data distribution and first reconstruction at T0 in the four months after AA run (shutdown)
  - Further reconstructions at T1's
- T0: First pass reconstruction, storage of RAW, calibration data and first-pass ESD's
- T1: Subsequent reconstructions and scheduled analysis, storage of a collective copy of RAW and one copy of data to be safely kept, disk replicas of ESD's and AOD's
- T2: Simulation and end-user analysis, disk replicas of ESD's and AOD's

Resource situation

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)	3.6	5.8	11.4	12.9	18.9	20.0	22.9	23.5
	Balance %	-28%	-38%	-42%	-52%	-45%	-59%	-36%	-60%
Disk	Requirement (PB)	0.9	0.77	3.4	1.6	6.5	4.0	9.5	5.3
	Balance %	29%	-0.5%	-21%	-7%	-32%	-45%	-33%	-42%
MS	Requirement (PB)	1.7	-	6.4	-	12.2	-	19.2	-
	Balance %	-15%	-	-46%	-	-46%	-	-48%	-

- We are trying to discuss with FAs and to find new resources
  - But we will 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
  - We will not record all data
  - We will do less data analysis
  - Impact on physics reach and timeliness of results

ALICE computing model evolution

- The computing model has not changed
  - Some aspects have been better defined
- The resources have been re-profiled to take into account the new accelerator schedule
- The storage strategy is clear, however it is being deployed/tested only now
- The analysis model is being tested, but wait for surprises here...

T1-T2 relations

- Current “tentative” megatable assignments

GridKa FZK	1 FZU AS Prague 1 RDIG 1 GSI 1 Muenster 4 Total
INFN CNAF	1 INFN Tier2 Federation 1 Total
UK Tier1	1 UK Tier2 Federations 0 Birmingham 1 Total
NL Tier1	0 SARA 0 Total
PDSF	1 US Tier2 Federation 0 Brazil T2 Federation 0 UNAM Mexico 1 Total
NDGF	0 0 0 Total

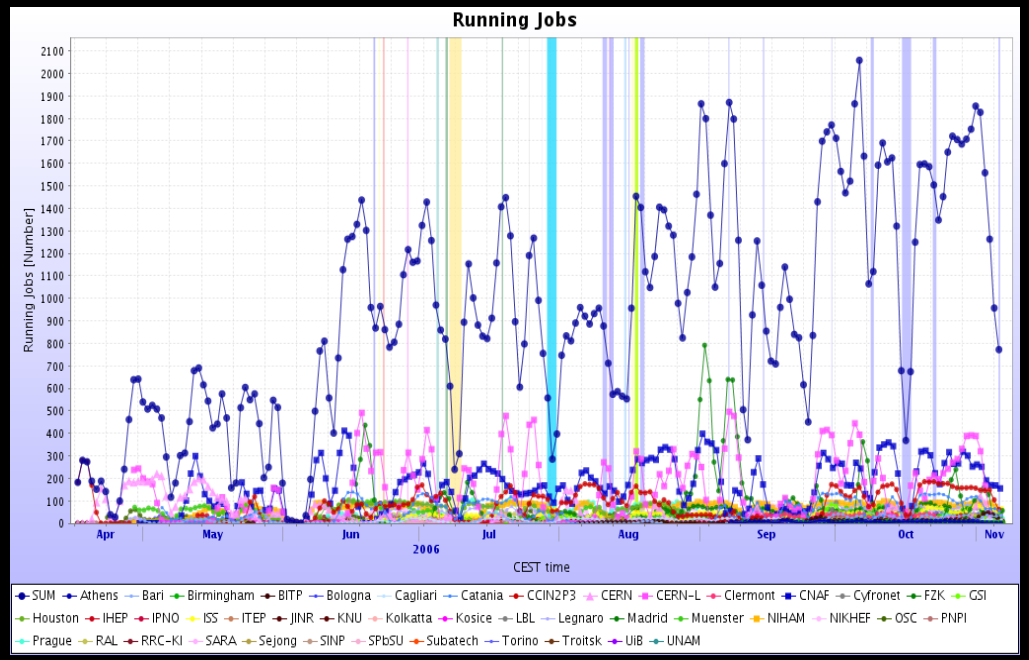
CCIN2P3	French Tier-2 Federation 1 Paris 1 Clermont-Ferrand 1 Nantes 1 Lyon 1 Sejong (Korea) 0 Kisti (Korea) 1 Madrid (Spain) 6 Total
CERN (CAF)	1 Cape Town 1 VECC/SINP Kolkata 1 Romanian Tier-2 Federation 1 RMKI (Hungary) 0 Athenes 1 Slovakia Federation 1 Ukraine Tier2 Federation 1 Polish Tier-2 Federation 0 Hiroshima 1 Wuhan 8 Total

T1-T2 relations

- We have very few T1s
- NDGF is still in an “undefined” state
- NIKHEF(SARA) and RAL are providing very little storage
- The bulk of the load is shared by 4 T1s: CERN, FZK, CCIN2P3 and CNAF
  - This drives up the requirements for MS and disk space for these centres
- Two factors can possibly alleviate this
  - Three out of four centres in US have “custodial storage capabilities”
  - Some of the T2s can have custodial storage capabilities (KISTI, Spain-EELA)

PDC'06

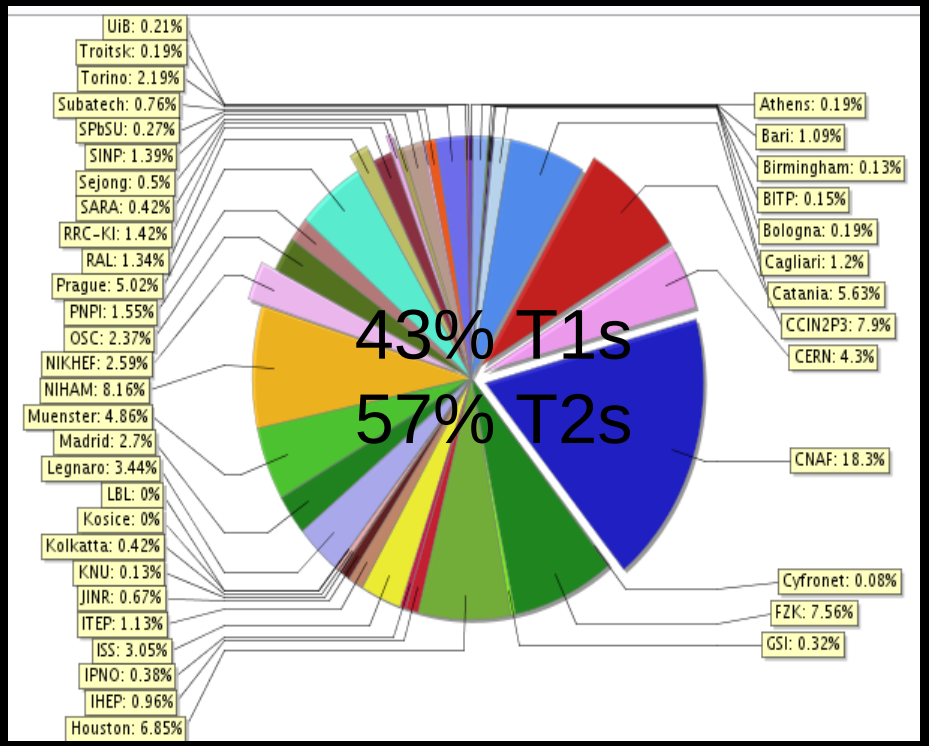
- The longest running Data Challenge in ALICE
  - Continuously running since 15 April ***(7 months!)***
  - 46 sites - 6 Tier 1s, 40 T2s, 40 in production, 6 setting up
  - We could only use 50% of pledged resources



- 588K jobs total
  - 463K production
  - 43K DAQ
  - 82K user
- 3.1M hours total
- 320TB, 15MFiles

PDC'06 - statistics (2)

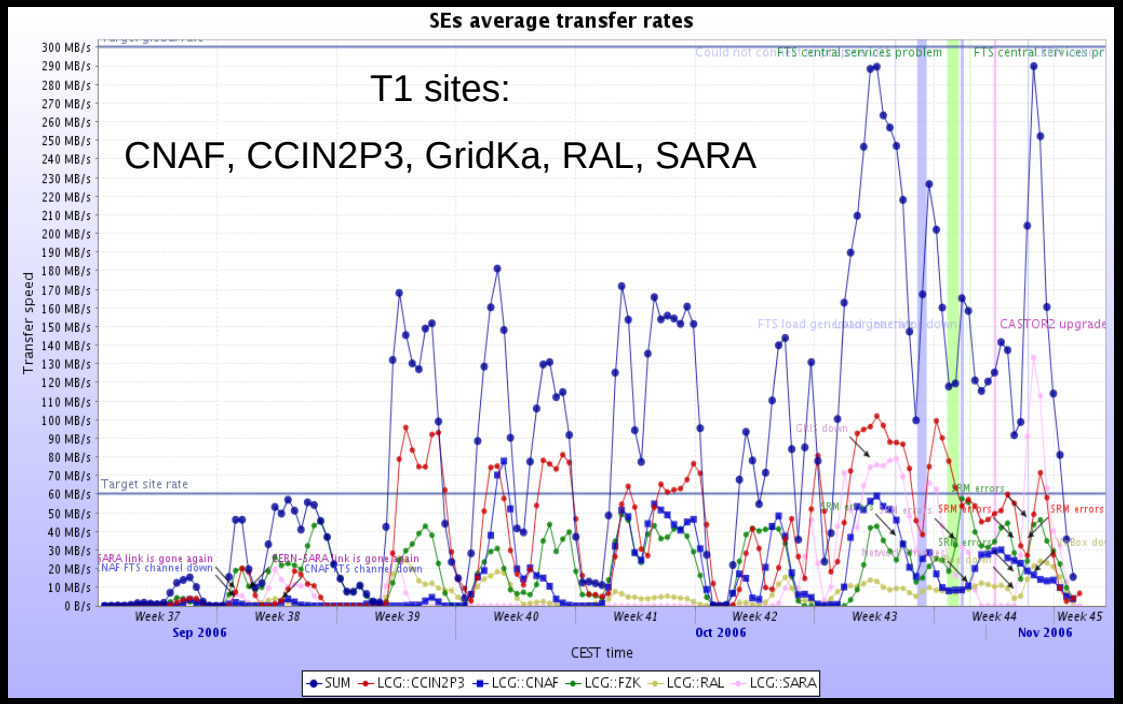
- Event statistics:
  - Full statistics available here
  - Total 12.5 M events
- Conditions:
  - p+p minimum bias
  - di- and single-μ events
  - Jet kinematics
  - Vertex displacement
  - Low CM collision energy
  - ***All available for user analysis*** (some of these already being analysed)
  - But only at CERN because of problems with storage deployment



GRID software

- AliEn
  - Single point of entry to ALICE GRID
  - 4 releases during PDC-6
  - Stability of central services now better than 90%
    - GRID catalogue, job submission and tracking, user authentication, storage management, monitoring… near production quality
  - Next big issue to tackle is storage and its reliability
- Workload management
  - LCG Resource Broker/CE: extensively tested, no problems
  - AliEn - ARC interface (NorduGrid): running at Bergen, to be expanded to other NDGF sites as they become operational
  - AliEn - OSG interface: work on it will start soon
- Transfer tools (File Transfer Service FTS) - file replication
  - Continuous test since 10 September of stability and throughput for T0->T1 transfers (RAW data replication)
  - Good and steady progress - will reach the design goals of the exercise soon

FTS transfers - history



- Instabilities achieving the expected rate (300MB/s) in November-December
- Not clear why it is working so well now
- We are still observing dips speed from time to time

- The whole exercise has used 3 out of 5 sites for most of the time
  - SARA has been rather instable (VOBOX down, SE down..)
  - At RAL the speed is very low. We are in "negotiations" with the site manager to have access to CASTOR2

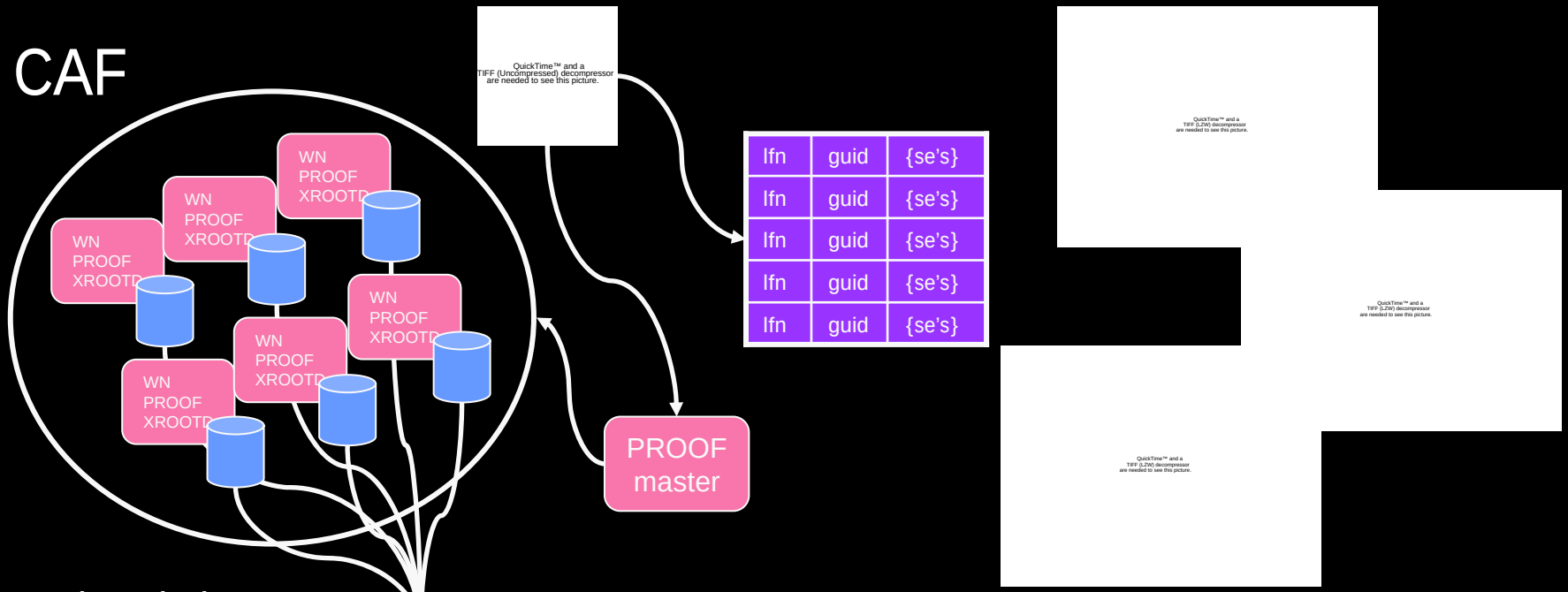
PDC'06 support

- Grid operation
  - Out ultimate goal is to automatise as much as possible the GRID operations - small team of experts take care of everything
    - Regional experts (1 per country/region) are responsible for the site operations (VO-boxes) and interactions with the local system administrators
    - Total of 15 people are responsible for the daily operations and support of the ALICE GRID (with the help of site admins)
      - New sites installation (95% of all) - Patricia Mendez Lorenzo (CERN/ARDA)
      - France - Artem Trunov (CCIN2P3), Jean-Michrl Barbet (Subatech)
      - Spain - Patricia Mendez Lorenzo
      - Italy - Stefano Bagnasco (INFN), Marisa Lusivetto (INFN)
      - Germany - Kilian Schwarz (GSI), Jan Fiete Grosse Oetringhaus (Muenster)
      - Russia, Greece - Mikalai Kutouski (JINR)
      - Nordic Sites - Csaba Anderlik (NDGF)
      - Romania - Claudiu Shiaua (NIHAM)
      - India - Tapas Samanta (VECC)
      - South Korea - Chang Choi (Sejong)
      - USA - Latchezar Betev (CERN)
      - Czech Republic - Dagmar Adamova (Prague)
      - Everything else (still looking for regional experts) Patricia Mendez Lorenzo
  - Still, this is quite a strain on very few people – expecting that with the more mature software, the load will go down
  - Operational experience is documented (still incomplete) in various HowTo's (alien.cern.ch)

Preliminary plan for PDC'07

- General purpose - continue and expand the tasks performed in PDC'06, increase the complexity of the exercise
- Begins early 2007, continuous until beginning of data taking
- Tasks
  - Continuation of user data analysis on the GRID
  - Tests and deployment of SE with integrated xrootd (CASTOR2, dCache, DPM)
  - Production of MC data for physics and detector performance studies - new request from ALICE PWGs
  - Testing and validation of new releases of application software: AliRoot, ROOT, Geant3, Fluka, conditions data infrastructure
  - Testing and deployment of new AliEn releases
  - Testing and integration of gLite RB/CE, further test of FTS stability and transfer throughput
  - GRID experts training, user training
  - Gradual introduction of new computing centres in the ALICE GRID, exercising the resources in the already installed sites

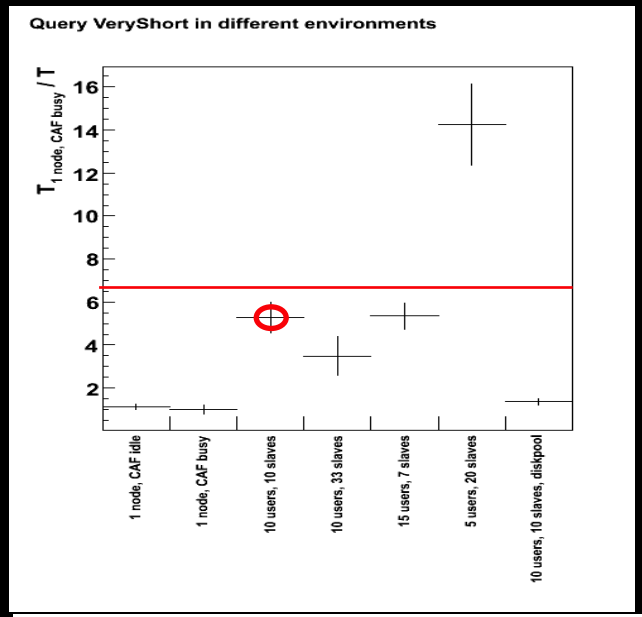
CAF

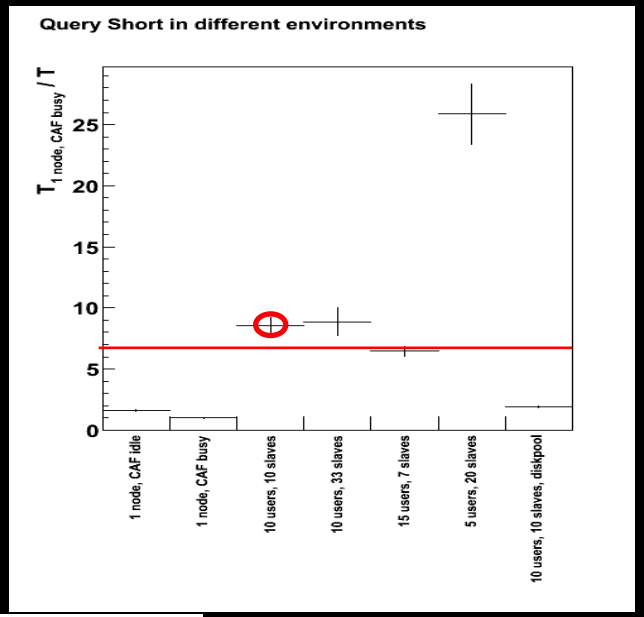


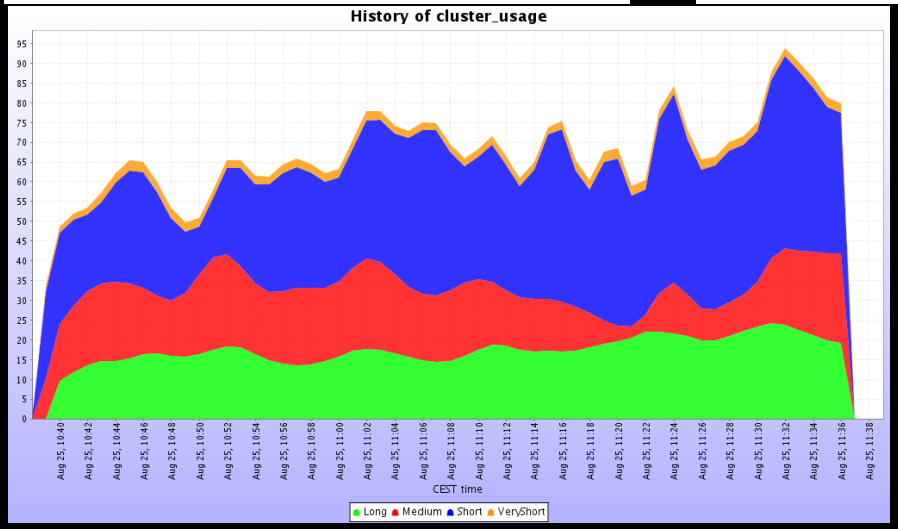
The whole CAF becomes a xrootd cluster

- 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





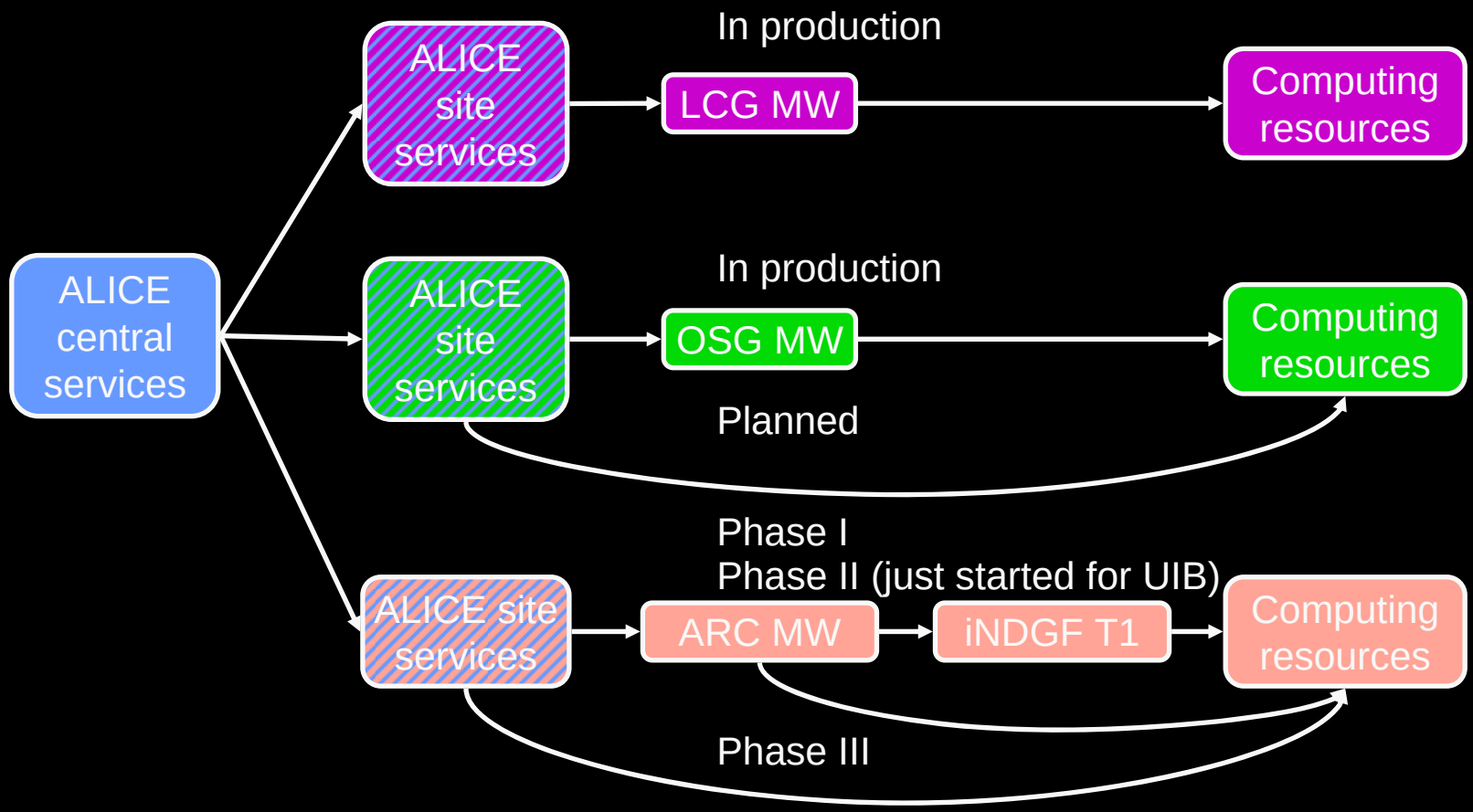


- Still several issues to be solved but the progress is steady
- Strong support from the ROOT/PROOF team

MW status

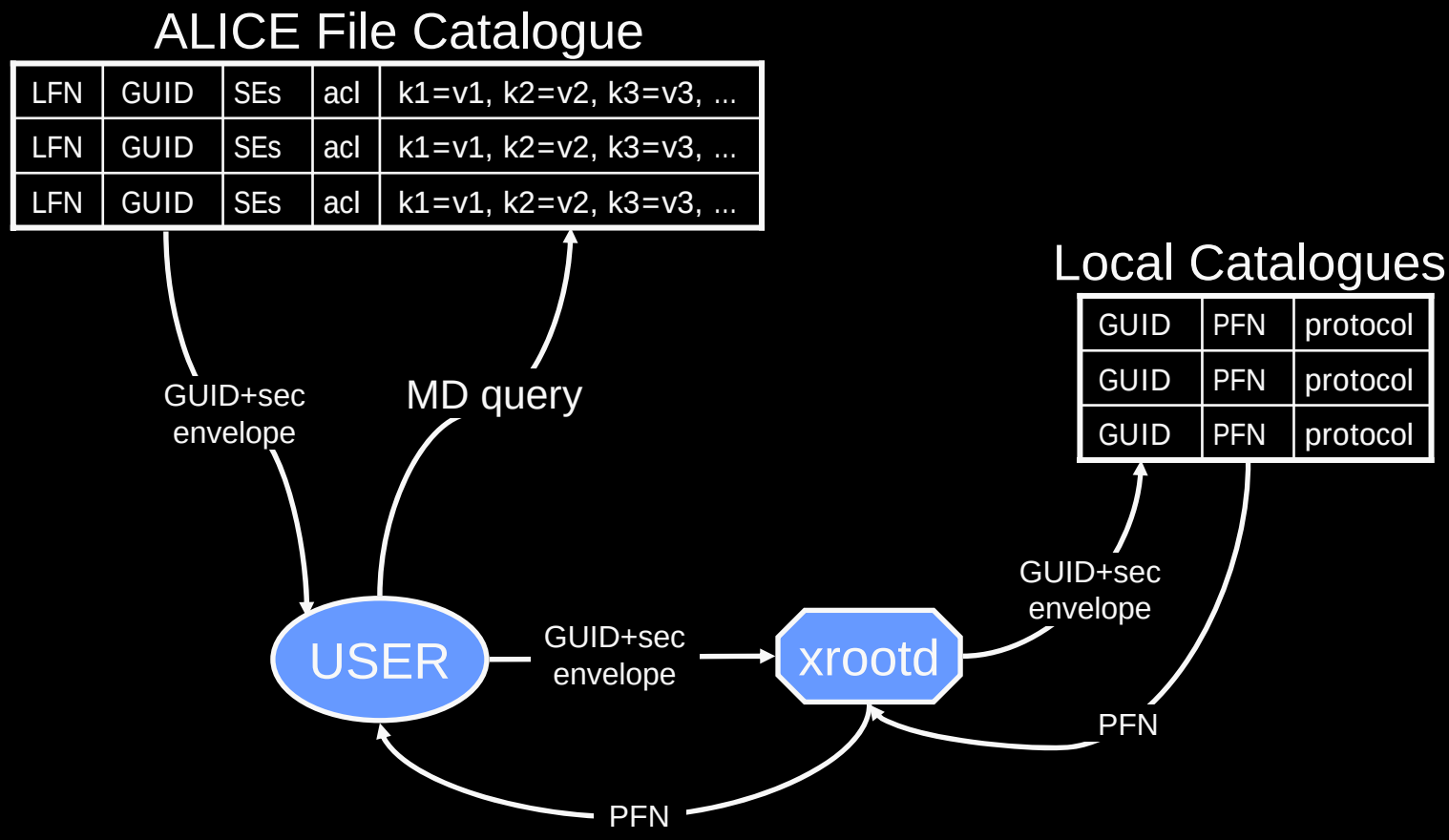
- We are using most services of LCG, complemented by ALICE specific ones that are required by our computing model
  - We have a single production system and very limited manpower to make this working
- ALICE specific services are installed centrally at CERN and on a single node in each computing centre (VO-Box)
- The design is evolving on the basis of the feedback
- Current workload management is under control
  - We started testing the gLite-CE
- Storage is still developing
  - The decision to use xrootd is excellent technically but requires developments
  - We are testing the prototypes of dCache, DPM and CASTOR2 with xrootd support
  - Not particularly depending on SRM functionality – it has to be there and stable

ALICE GRiD model

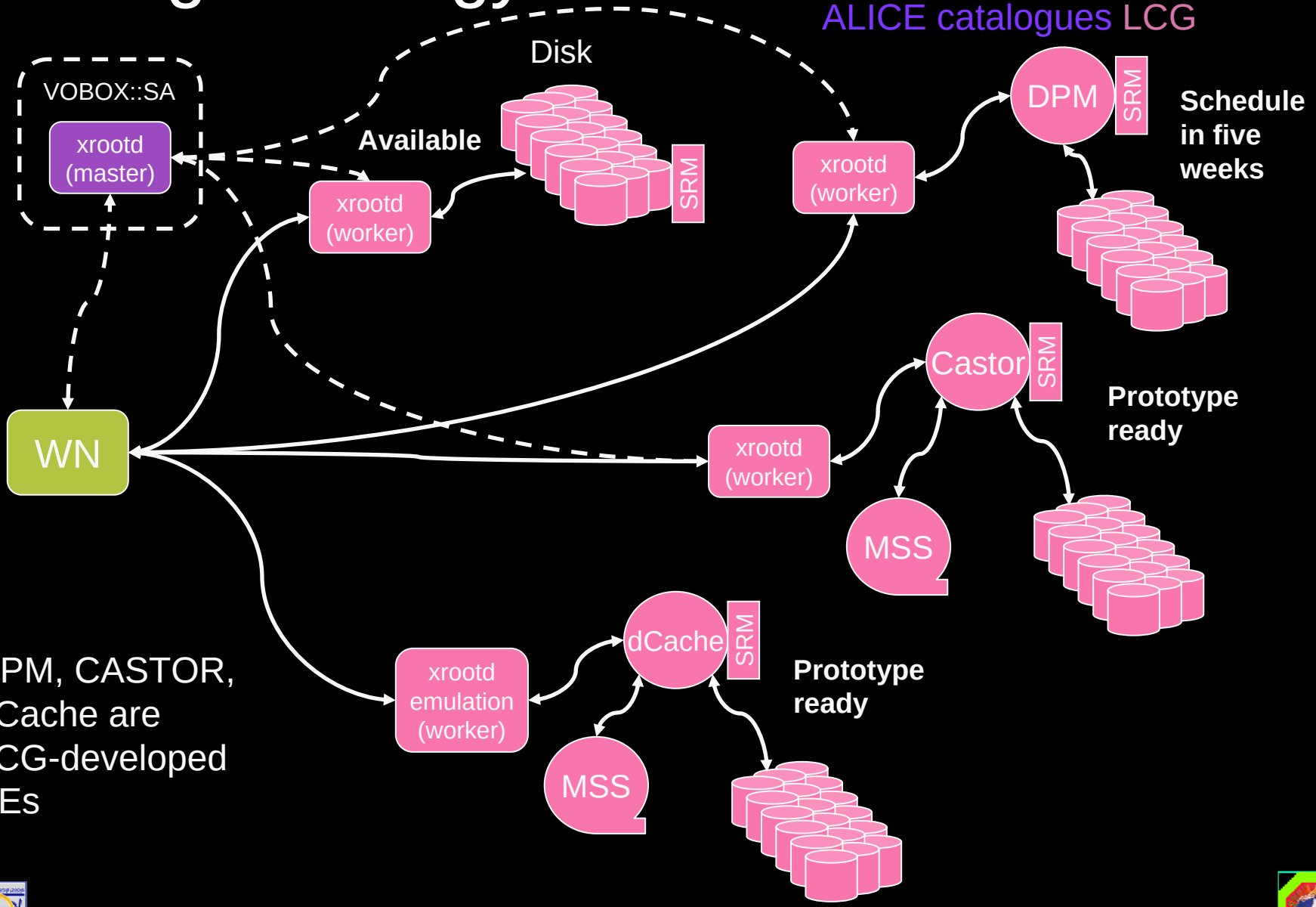


ALICE Files

/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

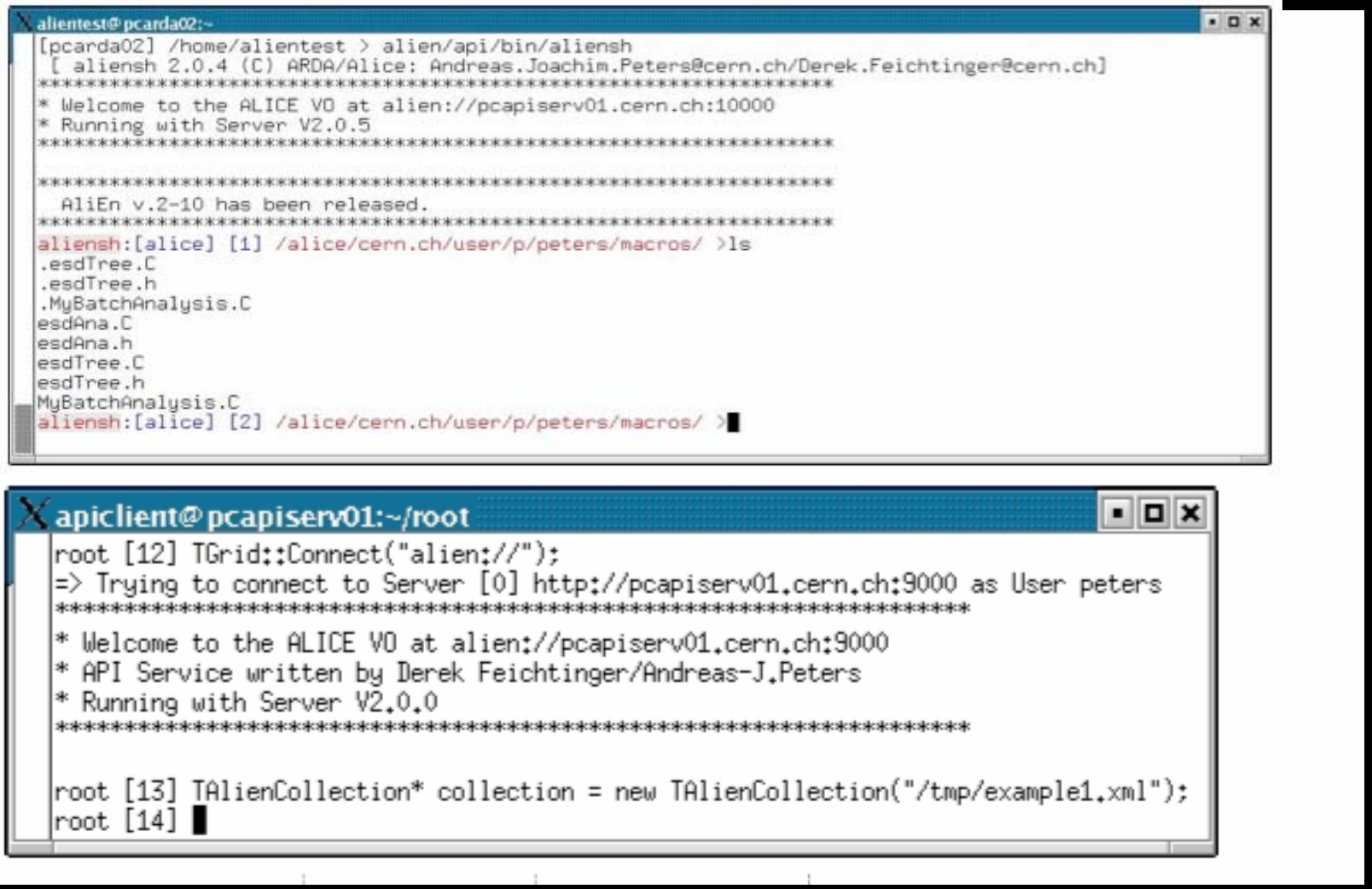
Analysis model

- 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 to LCG

- Improved FTS and underlying storage stability
  - Continue central (CERN) and site experts pro-active follow up on problems
- 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
- 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!"