ALICE plans and requirements now to first collisions

January 22, 2007

Federico Carminati

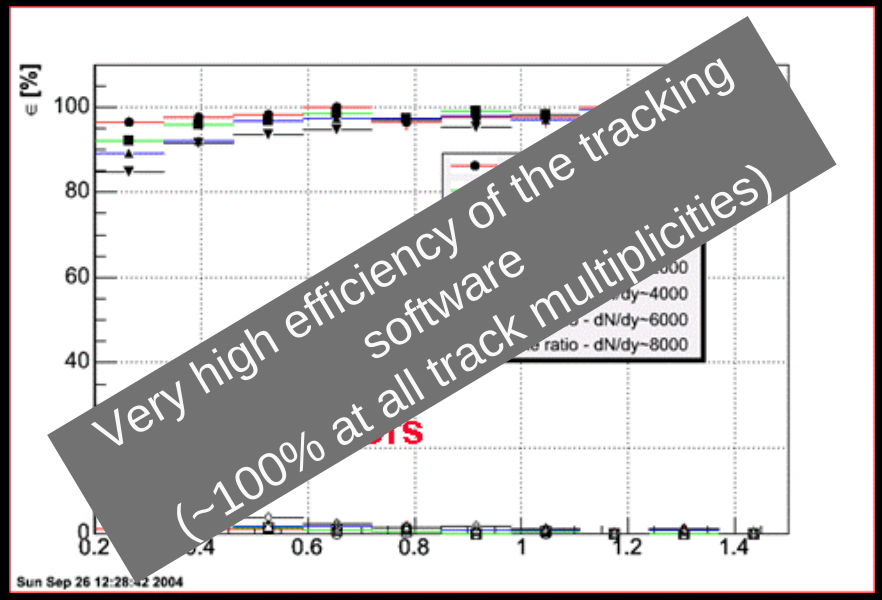
Summary

- Present status
  - AliRoot
  - Distributed computing
- Experience with Data Challenge
- A recall of ALICE computing model
- Planning from here to data taking

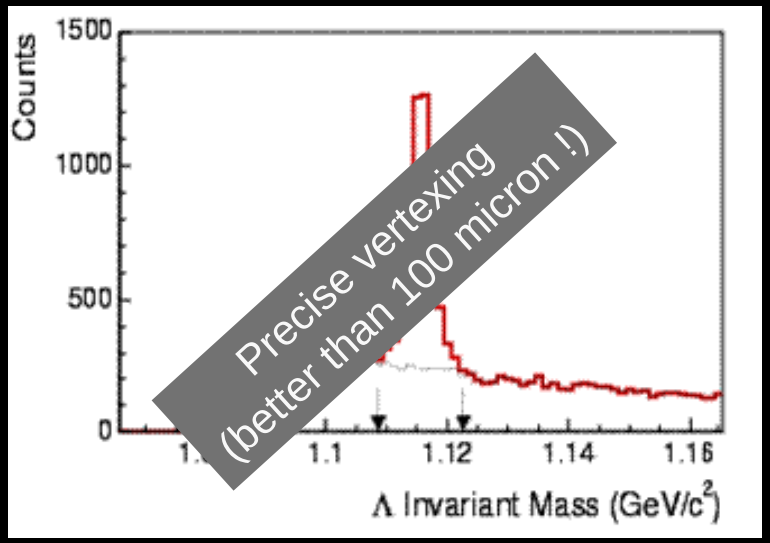
AliRoot status

- Used since 1998 for the detector & computing TDRs and the PPR
  - Integrated with DAQ, DCS and HLT
  - Linux (SLC3 & SLC4! IA32/64, Opteron), Solaris and MacOS
- Simulation
  - FLUKA interface validated, not yet in production
  - G4 interface validation started
  - Services and structures "as built"
- Reconstruction
  - Efficiency and PID at TDR values or better for PbPb ($dn/dy_{ch} \leq 8000$) and pp
  - Almost all parameters taken from Offline Condition DB
  - QA in place for most detectors
  - Measured magnetic field map in preparation
- Code still growing very fast
  - Memory & Speed continuously improving
  - Coding conventions and effc++ actively enforced

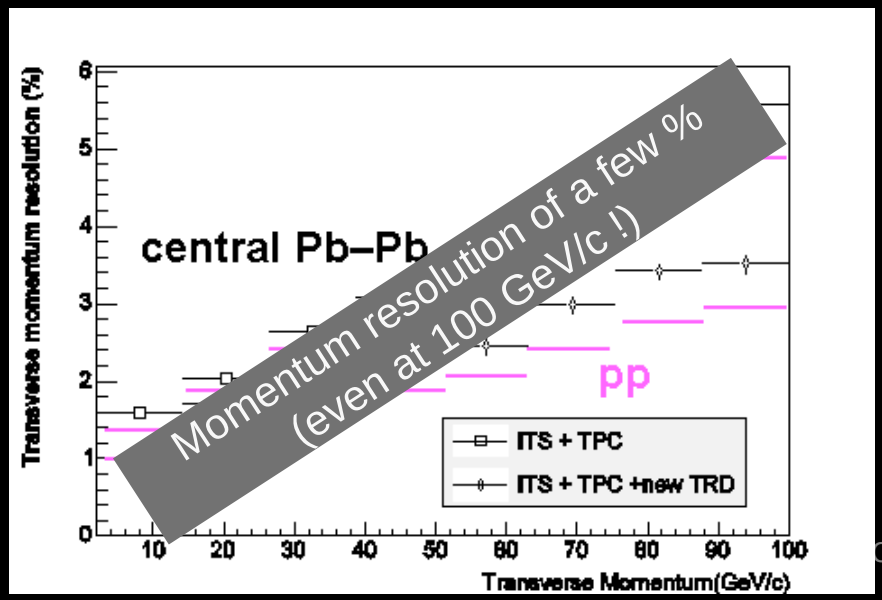
Reconstruction



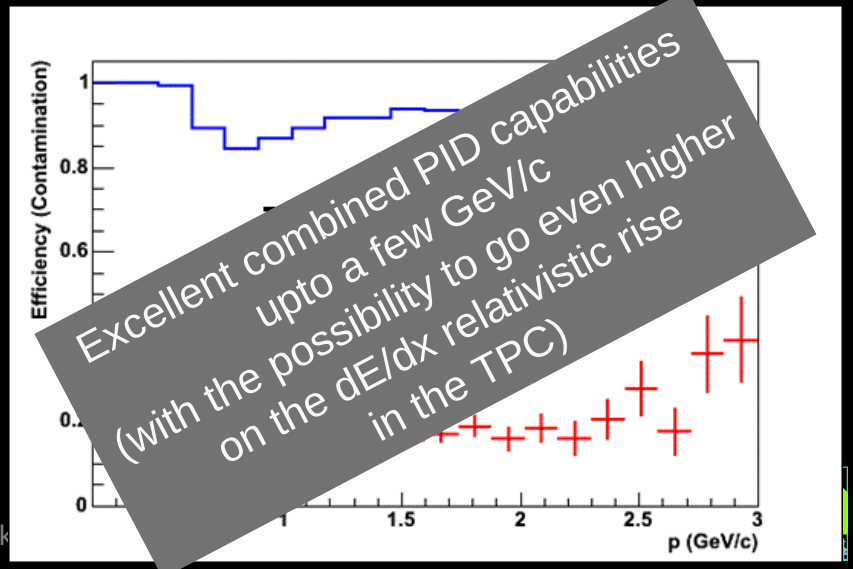
Very high efficiency of the tracking software
(~100% at all track multiplicities)



Precise vertexing
(better than 100 micron !)



Momentum resolution of a few %
(even at 100 GeV/c !)

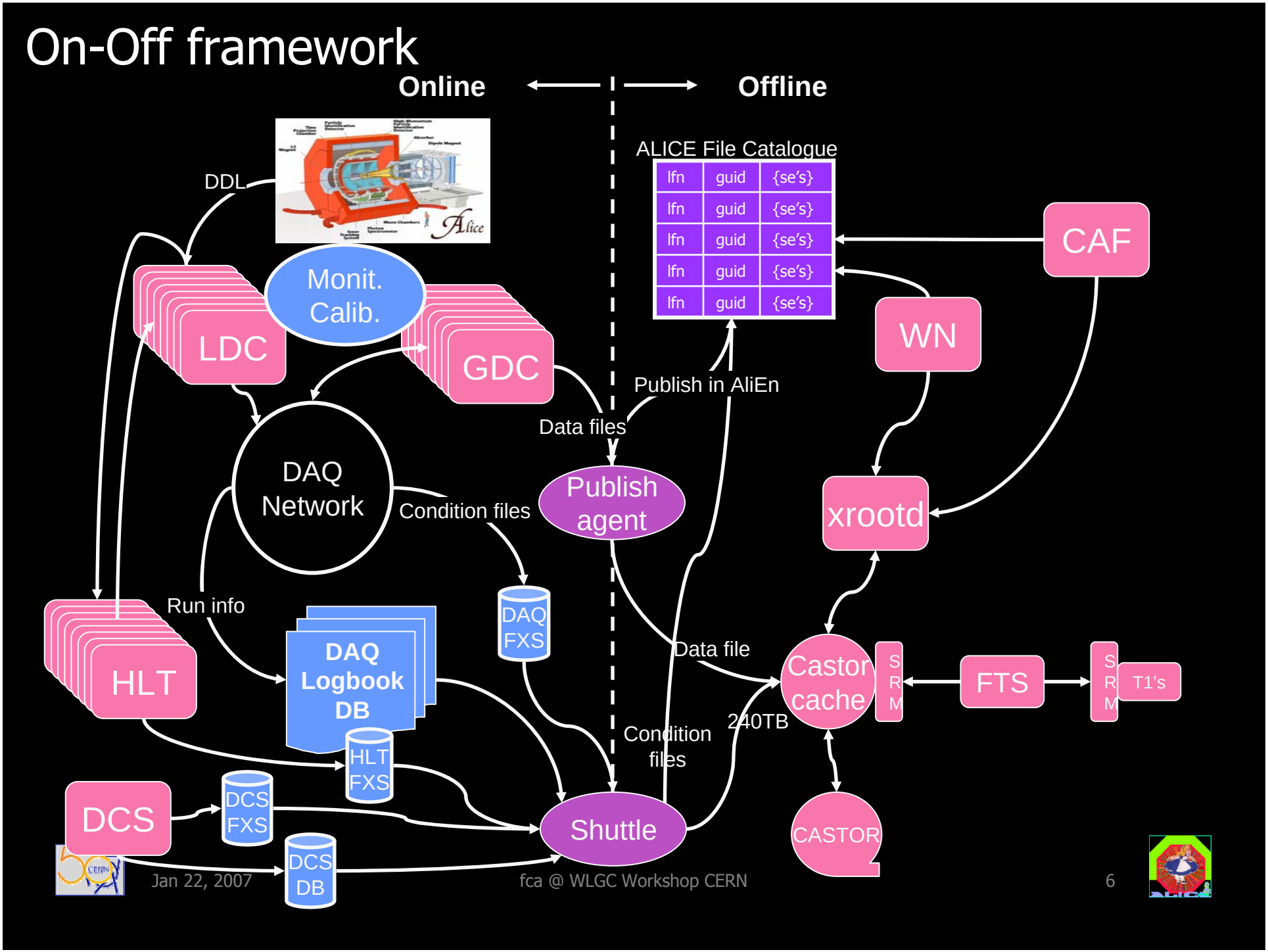


Excellent combined PID capabilities
upto a few GeV/c
(with the possibility to go even higher
on the dE/dx relativistic rise
in the TPC)

AliRoot status

- Analysis
  - Physics WGs moving to new framework based on TSelector
  - Active discussion on the redesign of AODs
- Condition framework
  - Offline framework ready
  - Online frameworks on track
  - Alignment and online calibration algorithms being developed
- Data
  - Size of data (RAW, ESD, AOD) under final evaluation
  - Framework exercised with raw data from test beams
  - Raw data de/en–code 90% validated
- Metadata
  - First version defined and implemented
- Documentation
  - Good doc for AliRoot, grid doc needs to be consolidated
  - Tutorials held monthly

On-Off framework

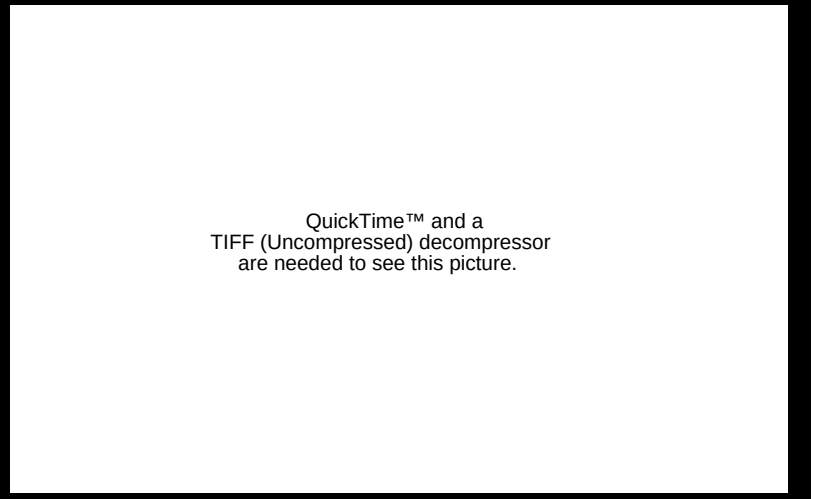


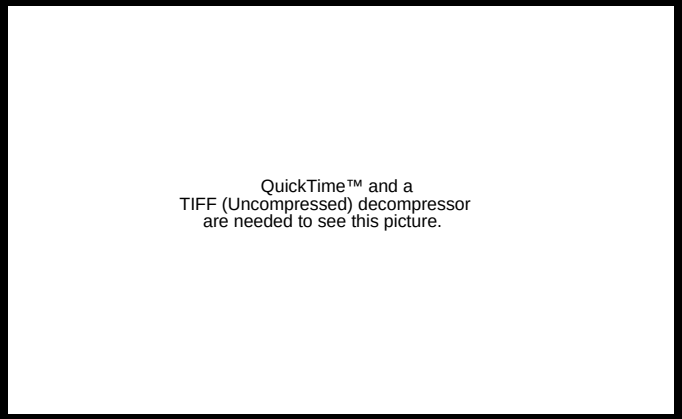
On-Off framework

- Shuttle Framework
  - Core: Done
  - DAQ File Exchange Server & Logbook connection validated
  - HLT File Exchange Server implemented
  - DCS File Exchange Server under development (critical!)
- Full test setup
  - Nightly build & shuttle run with publishing of output
  - TRD, TOF, PHOS, EMCAL, PMD & T0 preprocessors to be validated
  - Many preprocessors still missing in CVS
  - DA framework missing for DCS and HLT
- Detector algorithms nearly all still missing
- xrootd-CASTOR2 interface under testing

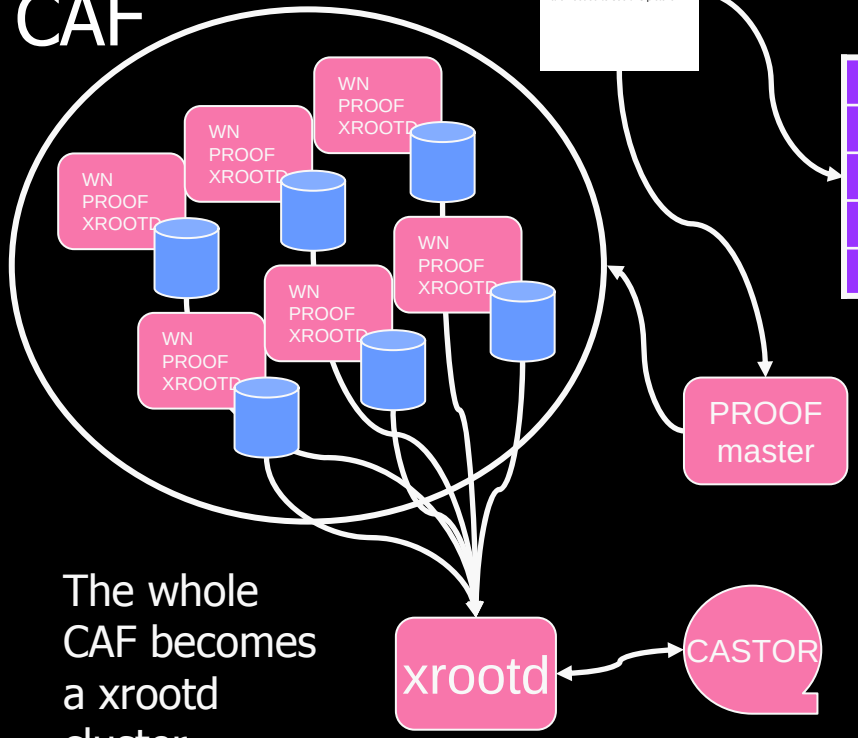
Visualisation

- Framework developed in collaboration with the ROOT team
- Current version provides most of the design functionality
- Another iteration with 3D graphics / user-interaction support in ROOT needed to allow building of fully-fledged multi-view end-user application





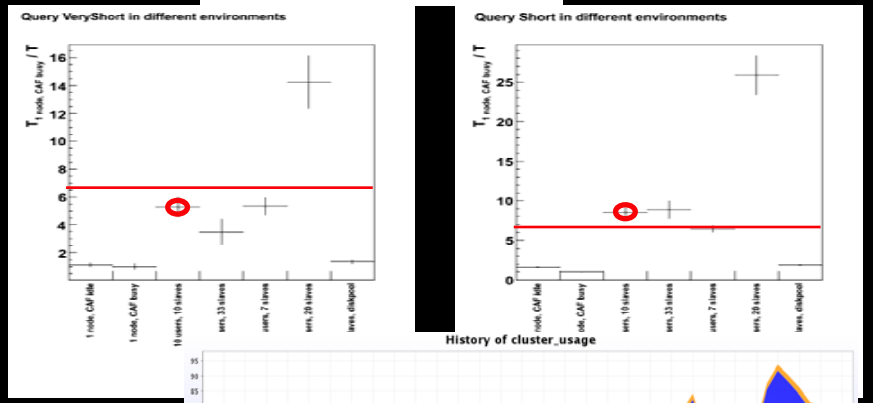
CAF

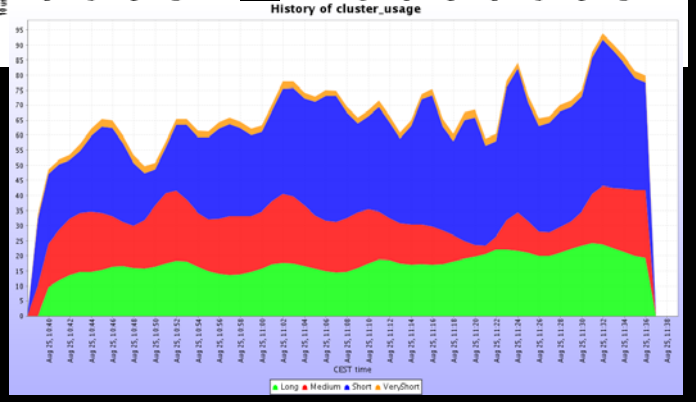


lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

The whole CAF becomes a xrootd cluster

- For offline monitoring and early physics
- Working with the PROOF team on optimisation
- S/W versions handling now in PROOF
- Quotas / load balancing plans ready
- Data access at the moment "by hand"
- Disk space management to be implemented





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

Distributed Computing

- AliEn
  - Single point of entry to ALICE GRID, developed and operated with very limited manpower
  - Stability of central services now better than 90%
  - 7 releases, 199 bugs, problems, requests (195 resolved)
  - 5 tutorials, 130 participants, >200 registered users, 55 active
  - ARC interface (NorduGrid): running at Bergen, to be expanded to other NDGF sites as they become operational
  - OSG interface: we hope work will start soon

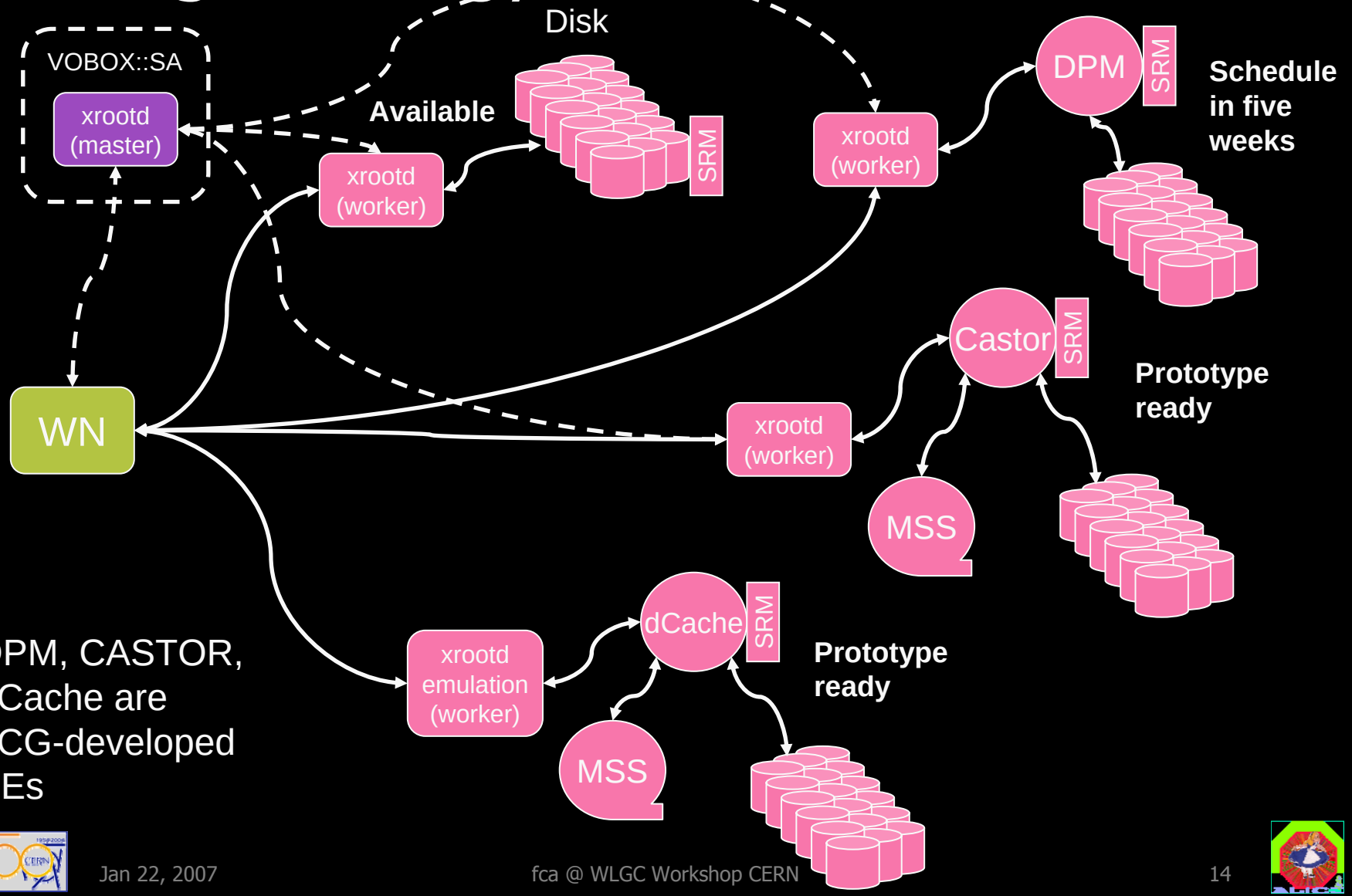
Distributed Computing

- We are using most services of LCG
- ALICE specific services as required by our computing model
  - Complementary to LCG
  - Installed centrally at CERN and locally on a single node (VO-Box)
- LCG-WMS under control
  - We are successfully testing the gLite-WMS
- Full monitoring information obtained via MonALISA
- File replication tools (File Transfer Service FTS)
  - Since Sep 06 continuous T0->T1 RAW data replication test
  - Progress toward design goals (sustained 300 MB/s)
  - Still many stability issues

Distributed Computing

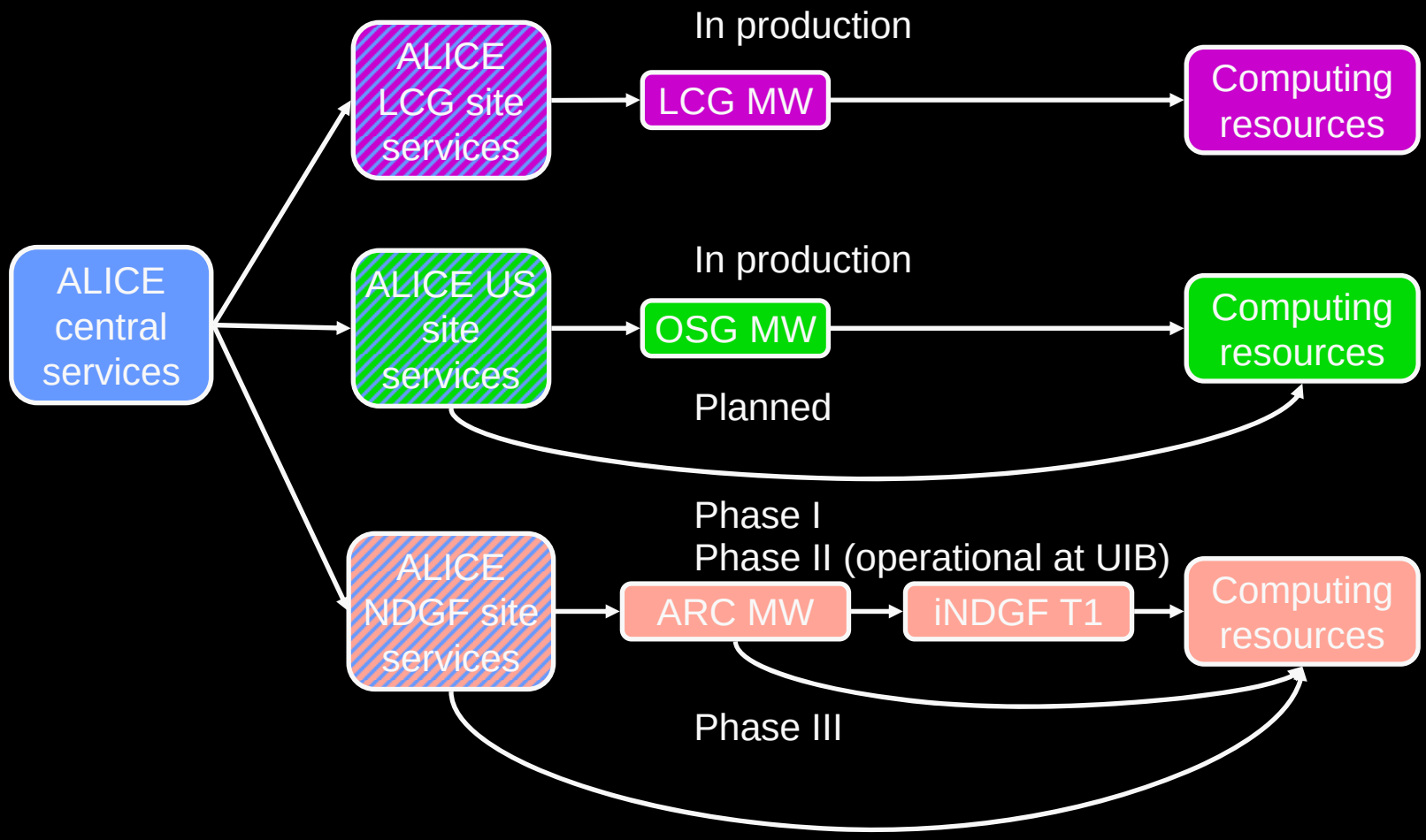
- Data Management tools still not under control
  - The xrootd choice is excellent but requires developments
  - Prototypes of dCache (advanced), DPM and CASTOR2 with xrootd support under test
  - Not particularly depending on SRM functionality – it has to be there and stable
  - No configuration / deployment of SEs yet apart from CERN
  - New file catalogue to enter production soon
  - Data access for analysis not yet tested in realistic situations
  - CASTOR2 still evolving
- Global management of resources in prototyping stage
  - Will enter production soon and take a very long time to be tuned
  - Emergencies will be handled by hand (at a high adrenaline cost!)

Storage strategy



DPM, CASTOR, dCache are LCG-developed SEs

ALICE GRID model



ALICE
Data Challenge VII

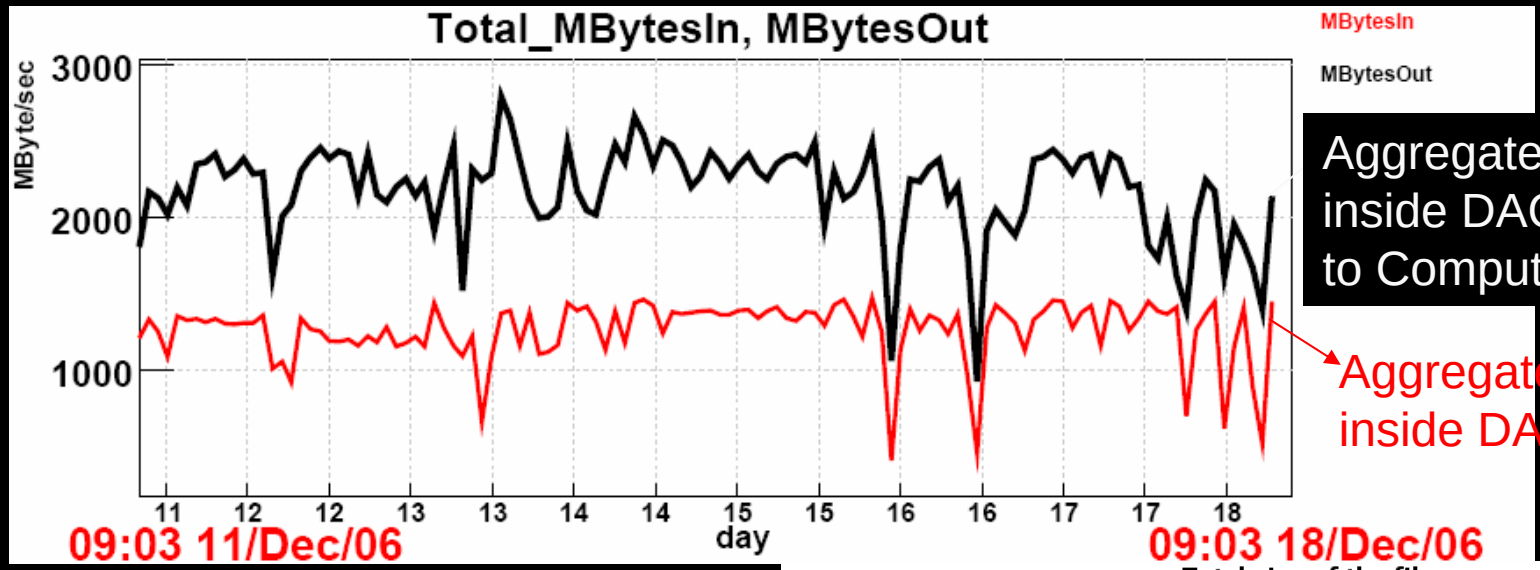
Data Generation:
45 optical link interfaces

ALICE Data Acquisition
at LHC Point 2

Data Storage in the GRID Tier0
in CERN Computing Centre
on the Meyrin site

Realistic test of the whole data flow
from the ALICE experiment
to the IT Computing Centre

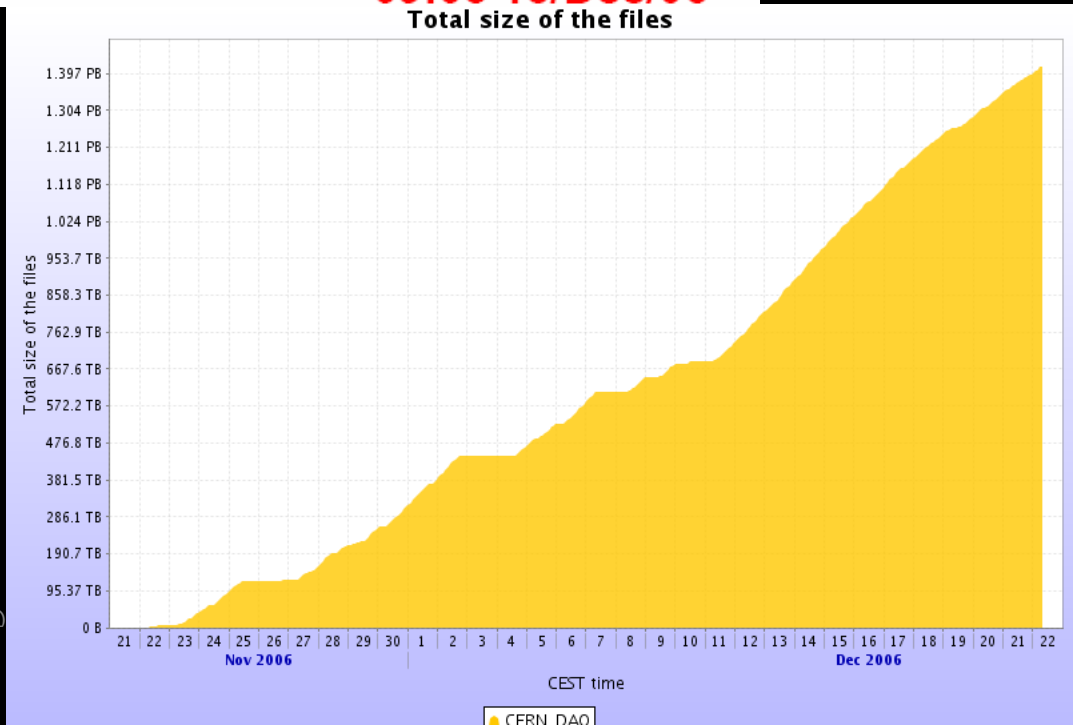
ALICE Data Challenge performances



Aggregate throughput inside DAQ + migration to Computing Center

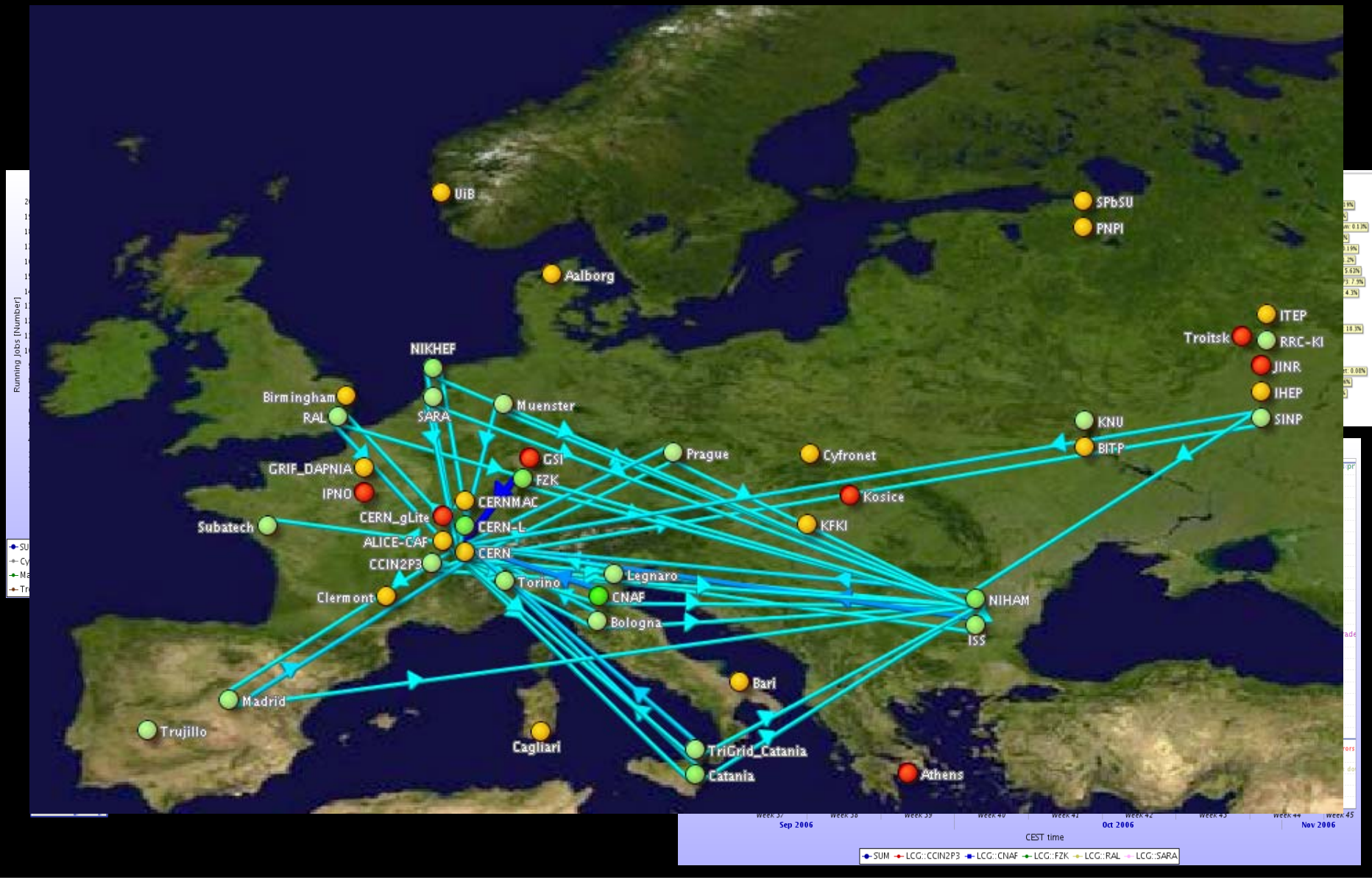
Aggregate throughput inside DAQ

ALICE DAQ,
ROOT formatting,
IT Mass Storage
Global Performances:
- 4 days at 1020 Mbytes/s
- 1.4 PB of data



Grid data challenge - PDC'06

- The longest running Data Challenge in ALICE



PDC 06 experience

- FTS still not production quality
  - Instabilities and dips in FTS, storage, VO-boxes, ALICE services to sustain the goal rate (300MB/s) in November-December
  - We have now a good understanding how to use FTS and its limitations
  - FTS is fully integrated in the ALICE high-level services
  - Expert support was excellent
  - The whole exercise could only use 3 out of 5 sites for most of the time
    - Instabilities at SARA (VOBOX down, SE down..)
    - At RAL the speed has improved only after we had access to CASTOR2
- Only 50% percent of pledged resources used during the PDC06
  - Undeployed resources
  - Competition from other experiments and instabilities in SW/HW
  - Lack of competence / manpower to support ALICE at some centres
  - Everything works when we have motivated competent people on site
- New centres should be brought in the picture asap
  - It takes a long time to start working together effectively
- Still a lot of manual operations needed

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
    - 15 people for the daily operations and support of the ALICE GRID (plus help of sysadmins)
      - 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
  - 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)

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 for GFAL
- Implementation of glexec
  - First on the testbed and then on the LCG nodes
- Overall stability of the services!

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

Computing model / resources

year	Time for physics (s)	
	pp	PbPb
2007	7×10^5	0
2008	4×10^6	2×10^5
2009	6×10^6	1×10^6
2010	1×10^7	2×10^6

- Missing computing resources are a threat to ALICE physics goals
- 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 makes no sense to develop an alternative within the pledged resources
  - The loss in scientific output would be too high
- If we could reduce the gap (10%-20%), 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

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)	2.2	3.2	6.9	7.8	19.9	15.8	37.5	28.5
	Missing %	10%	10%	-7%	-22%	-47%	-49%	-61%	-65%
Disk	Requirement (PB)	1.0	0.68	2.9	1.6	7.3	3.8	28.9	9.6
	Missing %	2%	11%	-13%	-9%	-42%	-41%	-79%	-69%
MS	Requirement (PB)	1.0	-	5.0	-	16.4	-	37.9	-
	Missing %	36%	-	-31%	-	-57%	-	-73%	-

T1-T2 relations

- We have few T1s
  - CERN, CNAF, CCIN2P3, FZK, RAL, NIKHEF, NGDF, US
  - NDGF is still in an "undefined" state
  - RAL & NIKHEF provide little resources to ALICE
  - A request to NIKHEF to "support" some T2s is still unanswered
- The bulk of the load is shared by 5 T1s: CERN, FZK, CCIN2P3, CNAF and US (for Mexico and Brazil)
  - This drives up the requirements for MS and disk for these centres
- Two factors can possibly alleviate this
  - Three out of four centres in US have "custodial storage capabilities"
  - Some T2s have custodial storage capabilities (KISTI, Spain-EELA)

T1-T2 relations

- Current "tentative" megatable assignments

To be reviewed!

GridKa FZK	1 FZU AS Prague 1 RDIG 1 GSI 1 Muenster 4 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
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	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
NDGF	0 0 0 Total		

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…

Plans in 2007

- AliRoot and SHUTTLE framework ready end Q1’07
- CAF and Grid MW will continue evolving, with an important checkpoint end Q3’07
- Continuous “data challenge” mode, aiming at using all resources requested
  - Physics studies
  - System stability and functionality
  - Stable and well defined support for all services and sites
  - Integration of new centres
  - Exercising of operational procedures
  - Improvement of MW
- We are planning a T0 combined test in Q2’07

Plans for AliRoot

- To be provided by end Q1'07
  - AODs
  - Final analysis framework
  - Alignment algorithms
  - Raw data format
  - Visualisation framework
  - Validation of survey to alignment procedures
- Since April'07 release a stricter policy on backward compatibility will be applied
- Algorithms for pp physics frozen in September'07

Plans for SHUTTLE

- SHUTTLE framework to be completed by 1Q'07
  - Data flow defined for all detectors
  - FXS for HLT, DAQ & DCS
  - Data access for HLT, DAQ & DCS to the FXS
  - Detectors "plugins" working
  - Data registration with MD into ALICE FC
  - Synchronization with DAQ, ECS & HLT

Resources requested for 2007

Nothing else than the MegaTable

	CERN				External			Total
	Tier0	CAF	Tier1	Total	Tier1s	Tier2s	Total	
CPU (MSI2k)	0.053	0.026	0.90	0.90	3.63	5.79	9.41	10.3
DISK (PB)	0.014	0.051	0.26	0.32	0.86	0.77	1.63	1.95
MSS (PB)	0.066	-	0.63	0.70	1.66	-	1.66	2.36

- Bandwidth
  - CERN → T2 30 MB/s
  - T2 → CERN 50 MB/s
  - CERN → T1s 80MB/s
  - T1↔T1 & T2↔T2 change from centre to centre
- Rates (7x105s)
  - Raw 7.7 x 104 GB = 40k files (2GB/file) = 5k files/day
  - ESD 4.2 x 103 GB = 2k files = 0.25k files/day
- In the PDC06 we have done
  - Write: CASTOR2 - 10k/day, total all SEs: 20k/day
  - Read: total 32k/day (mostly conditions data).

Preliminary plan for PDC'07

- General purpose - continue and expand the tasks performed in PDC'06, increase the complexity of the exercise
- Started… continuous until beginning of data taking
- Tasks
  - 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, Geant4, 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
  - Scheduled and end-user analysis of produced data

Combined T0 test 2007

- First combined test for ALICE
- Standard MDC
  - Simulated events from LDC to GCS's + rootification & registration in AliEn with RAW MD
- Condition framework
  - DCS + HLT + DAQ + ECS access, run of available DA's, creation of condition files via SHUTTLE, registration in AliEn with MD
- Quasi-online reconstruction
  - Automatic job trigger at T0
- Data distribution
  - Automatic FTS transfer trigger
- CAF offline monitoring
- Data analysis at T1's with replication of ESD/AOD from T0

Courtesy of J.Shiers

WLCG Commissioning Schedule

Inter-Site Rates – Revised Megatable

Centre	T0->T1	T1->T2	T2->T1	T1<->T1
	Predictable – Data Taking	Bursty – User Needs	Predictable – Simulation	Scheduled Reprocessing
IN2P3, Lyon	220	286.2	85.5	496.0
GridKA, Germany	220	384.9	84.1	395.6
CNAF, Italy	190	321.3	58.4	583.8
FNAL, USA	110	415.0	52.6	417.0
BNL, USA	300	137.7	24.8	358.0
RAL, UK	120	108.3	36.0	479.4
NIKHEF, NL	160	34.1	6.1	310.4
ASGC, Taipei	120	126.5	19.3	281.2
PIC, Spain	100	167.1	23.3	294.5
Nordic Data Grid Facility	60	-	-	62.4
TRIUMF, Canada	60	-	-	59.0

2006

2007

2008

- Continued testing of computing models, basic services
- Testing DAQ → Tier-0 (? ?) & integrating into DAQ → Tier-0 → Tier-1 data flow
- Combined T0 test
- Building up end-user analysis support
- Finalisation of CAF & Grid
- Exercising the computing systems, ramping up job rates, data management performance, ….
- The real thing

- SC4 – becomes initial service when reliability and performance goals met
- Introduce residual services Full FTS services; 3D; gLite 3.x; SRM v2.2; VOMS roles; SL(C)4
- Initial service commissioning – increase performance, reliability, capacity to target levels, experience in monitoring, 24 x 7 operation, ….
- 01jul07 - service commissioned - full 2007 capacity, performance
- first collisions in the LHC. Full FTS services demonstrated at 2008 data rates for all required Tx-Ty channels, over extended periods, including recovery (T0-T1).

Conclusions

- Development and deployment of our distributed computing infrastructure is proceeding
  - We can say today that we have an "almost" working system (AliEn+other MW), progress is steady and objective in sight
  - 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 loss of key people would be unrecoverable
  - The SFT & EGEE/ARDA contribution are instrumental
- The resource situation is such that we cannot attempt yet a rescaling
  - We hope to reach the situation where such an exercise can be done meaningfully
- At the moment we are on track, however there are a very large number of elements that have to converge in a very short time
  - Contingency is very scarce


"That's
all
folks!"