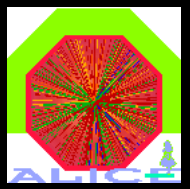
ALICE Use Cases for LCG-SC4

LCG-SC4 Workshop
Mumbai, 10-12th,
February, 2006

Piergiorgio Cerello (cerello@to.infn.it)
INFN

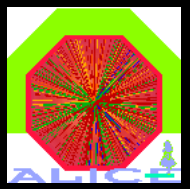
Overview



- ALICE Data Challenge 2005 (-2006)
  - Status
  - Lessons learnt
- ALICE Data Challenge(s) 2006
  - Final verification of computing model and GRID services readiness
    - Analysis Use cases
    - Plans

Part I
ALICE Data Challenge 2005

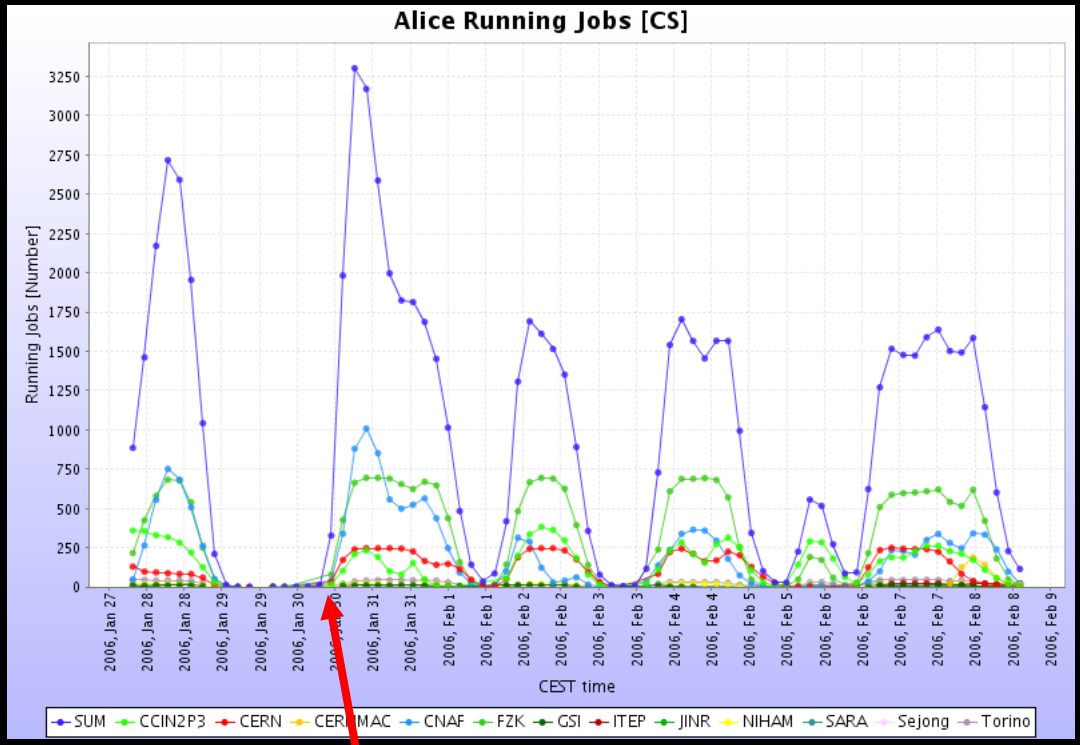
ALICE Data Challenge 2005 (-2006)



- Status
  - preparation started in the summer of 2005
  - Expected starting date: Sep, 1st, 2005
  - Ongoing
    - History of PDC'05 - ALICE MonALISA repository http://alimonitor.cern.ch:8889
    - Hard to define a “starting date”
    - Several issues identified and solved on the way
    - Continuous process of tuning
    - Steady improvement in performance

ALICE Data Challenge 2005 (-2006)

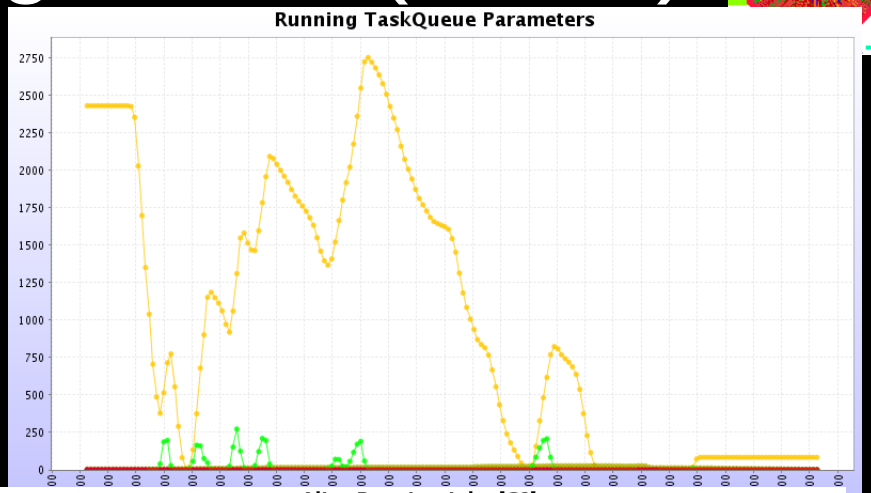
- Jobs (last month)

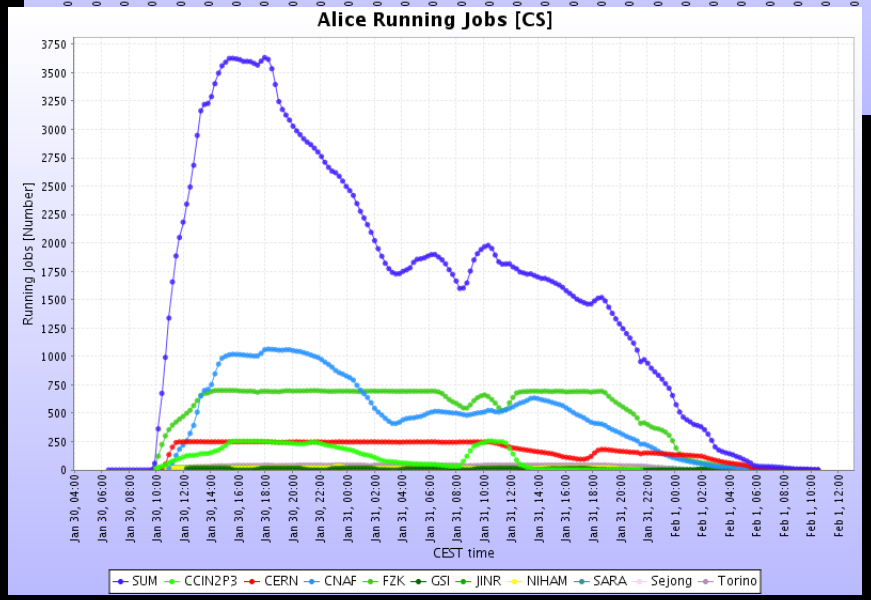


- Focus on one “run”

ALICE Data Challenge 2005 (-2006)

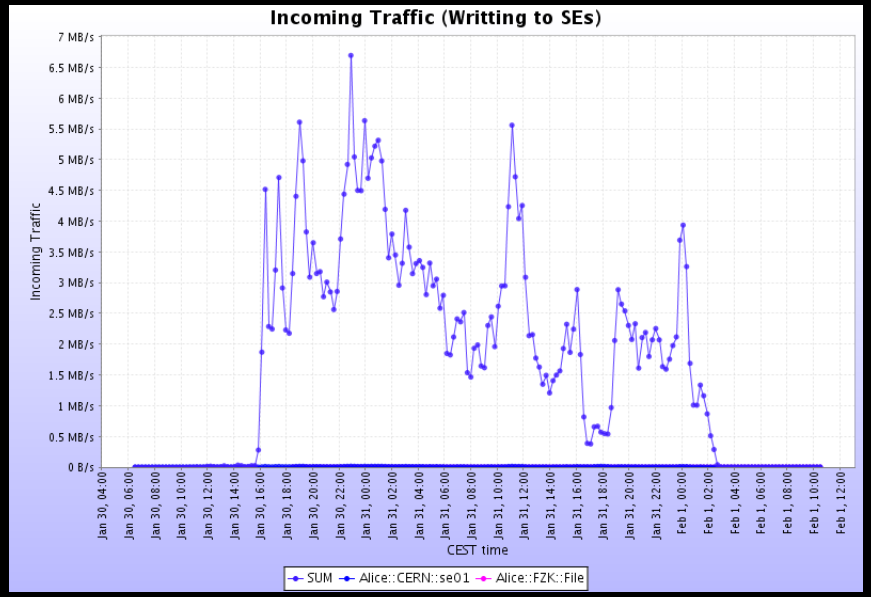
- Jobs
  - In the Task Queue
  - In Execution

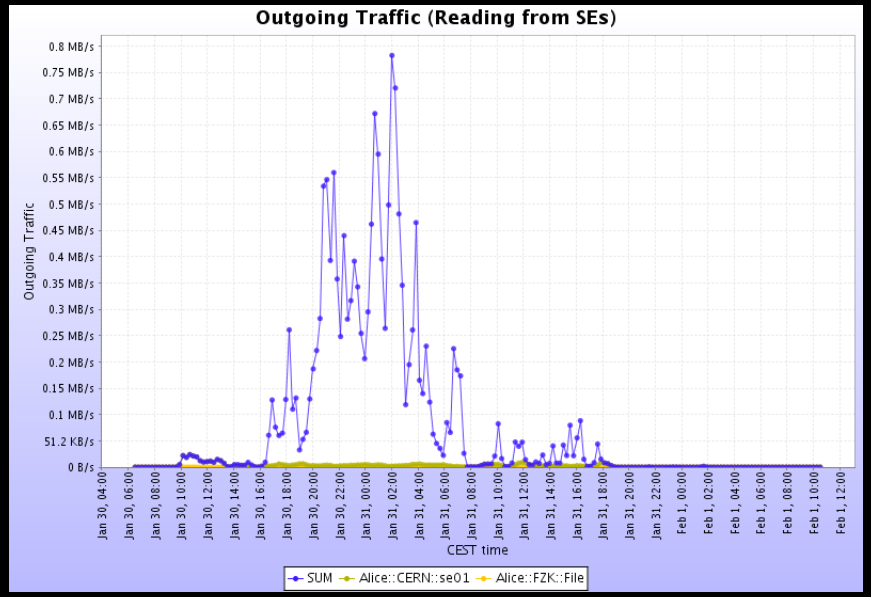




ALICE Data Challenge 2005 (-2006)

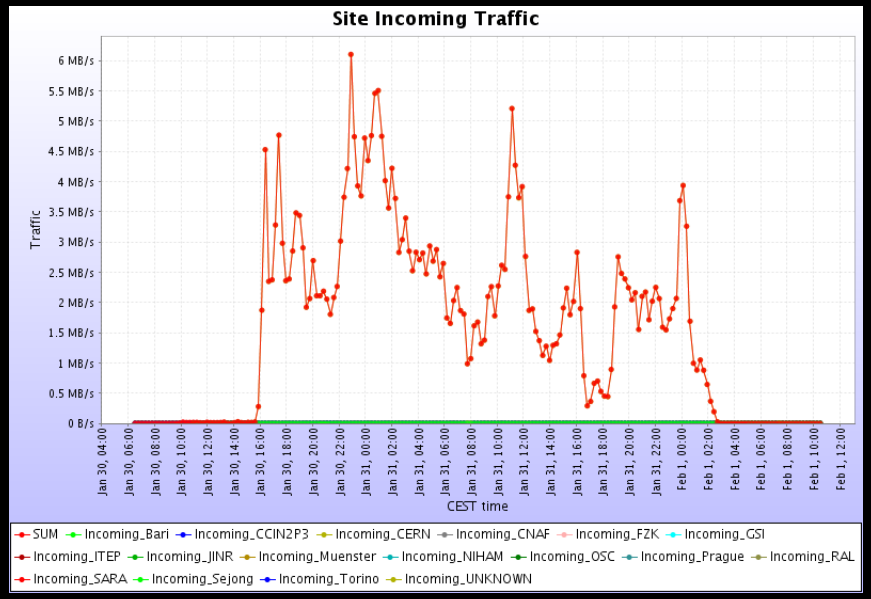
- Storage
  - Writing up to 7 MB/s
  - Reading up to 0.8 MB/s (testing mode)

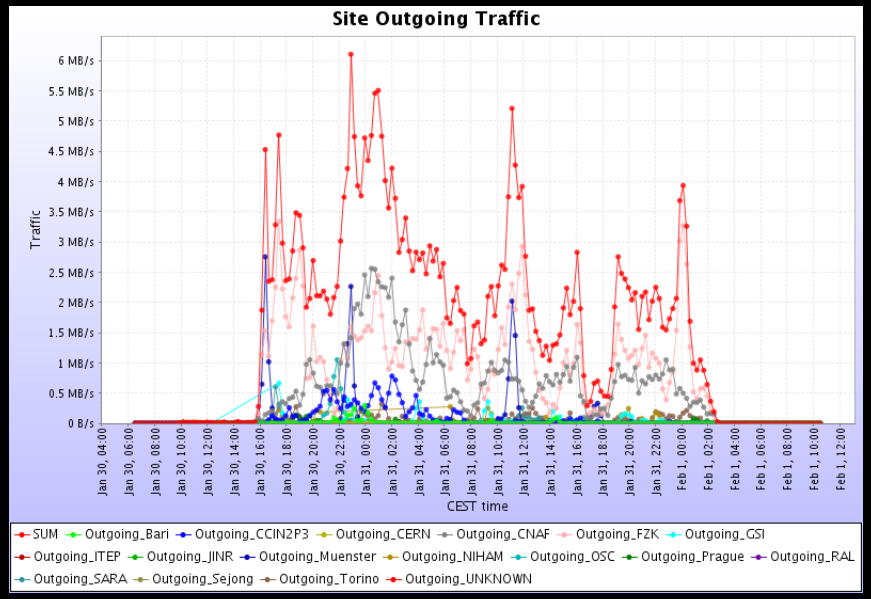




ALICE Data Challenge 2005 (-2006)

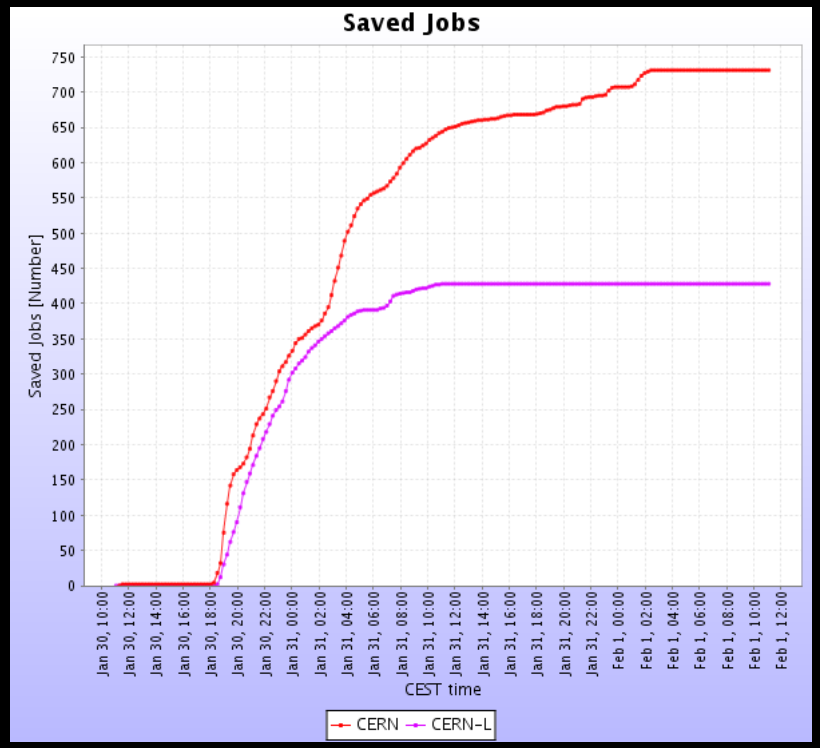
- Network
  - Incoming to CERN * Outgoing (other sites)

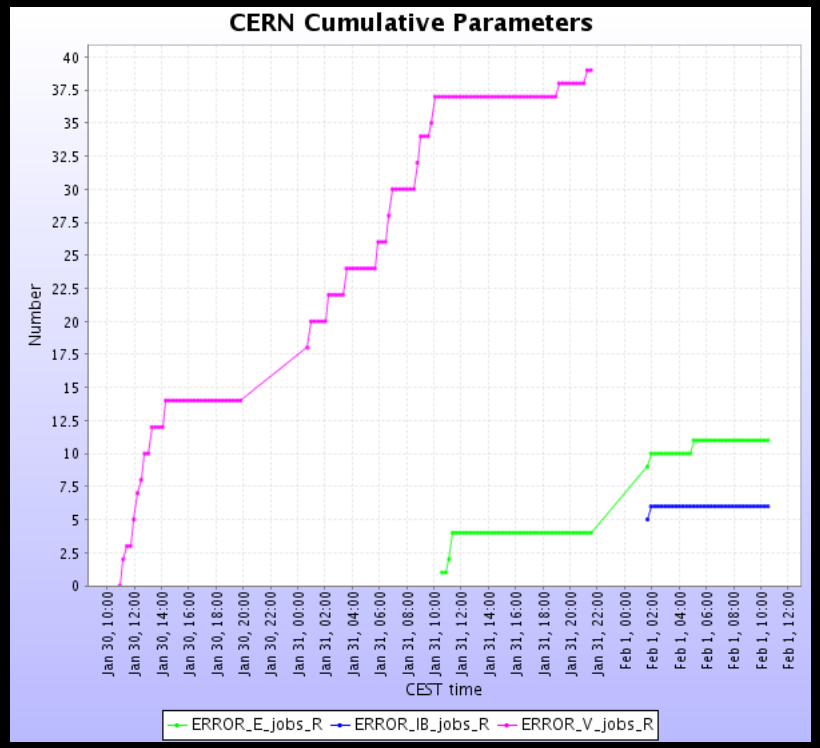




ALICE Data Challenge 2005 (-2006)

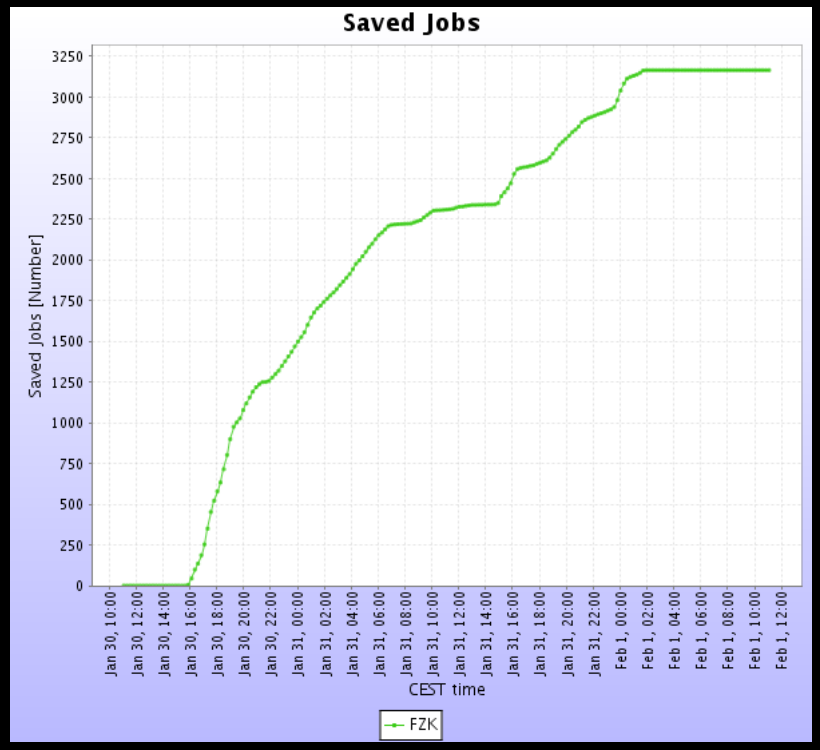
- Sites view: CERN

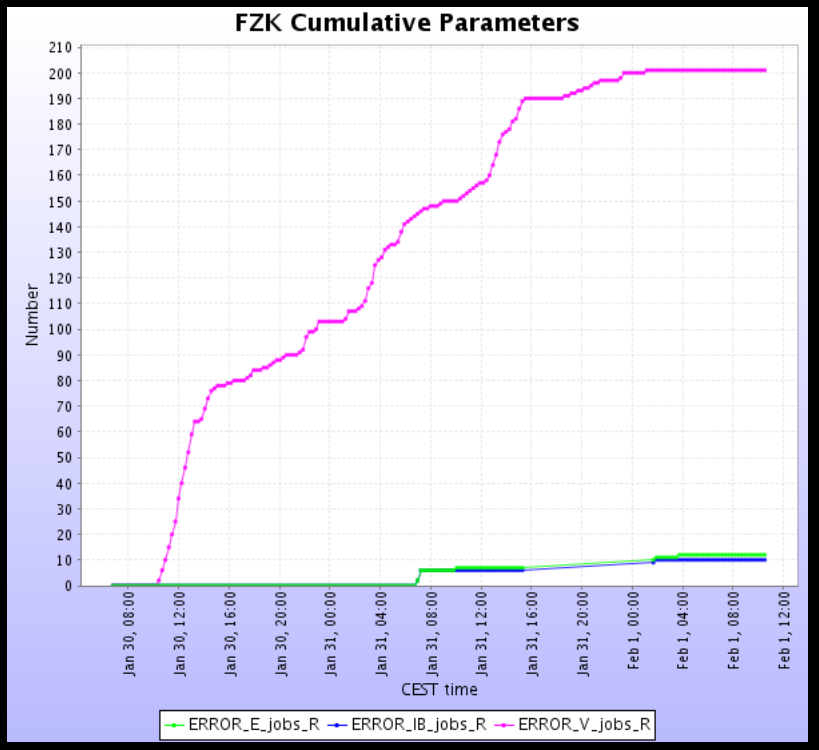




ALICE Data Challenge 2005 (-2006)

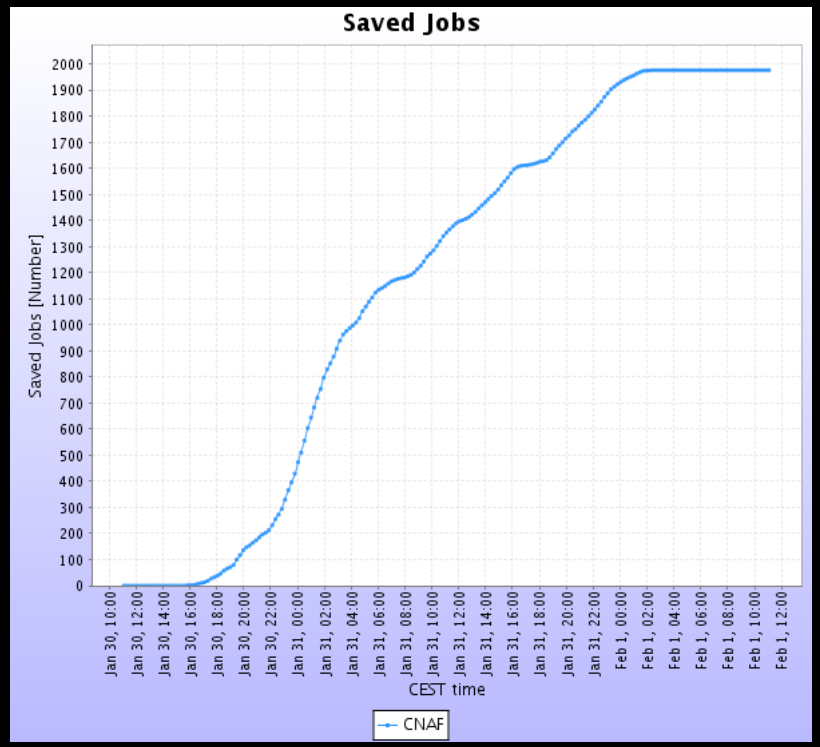
- Sites view: FZK

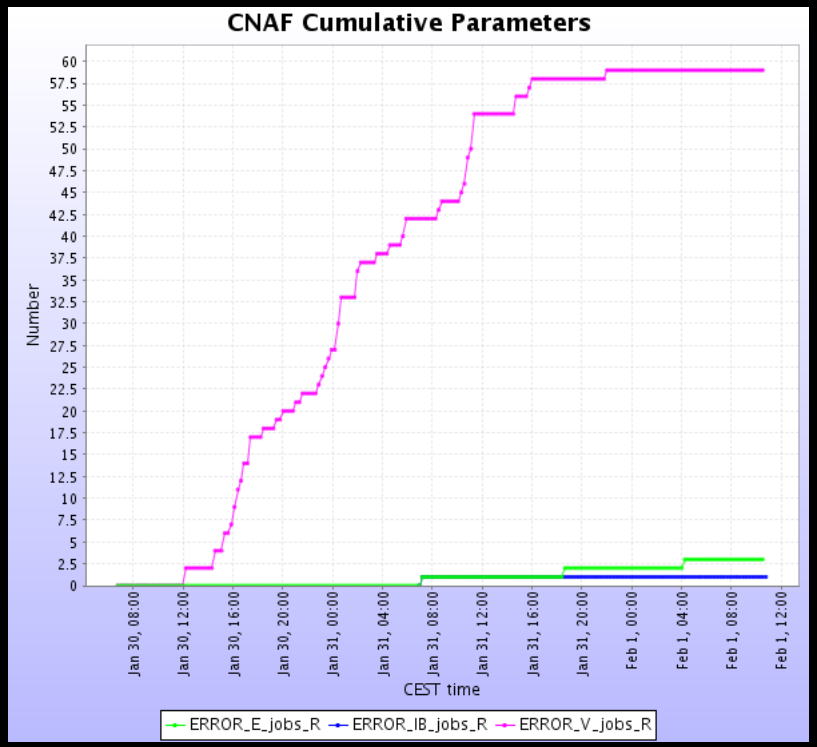




ALICE Data Challenge 2005 (-2006)

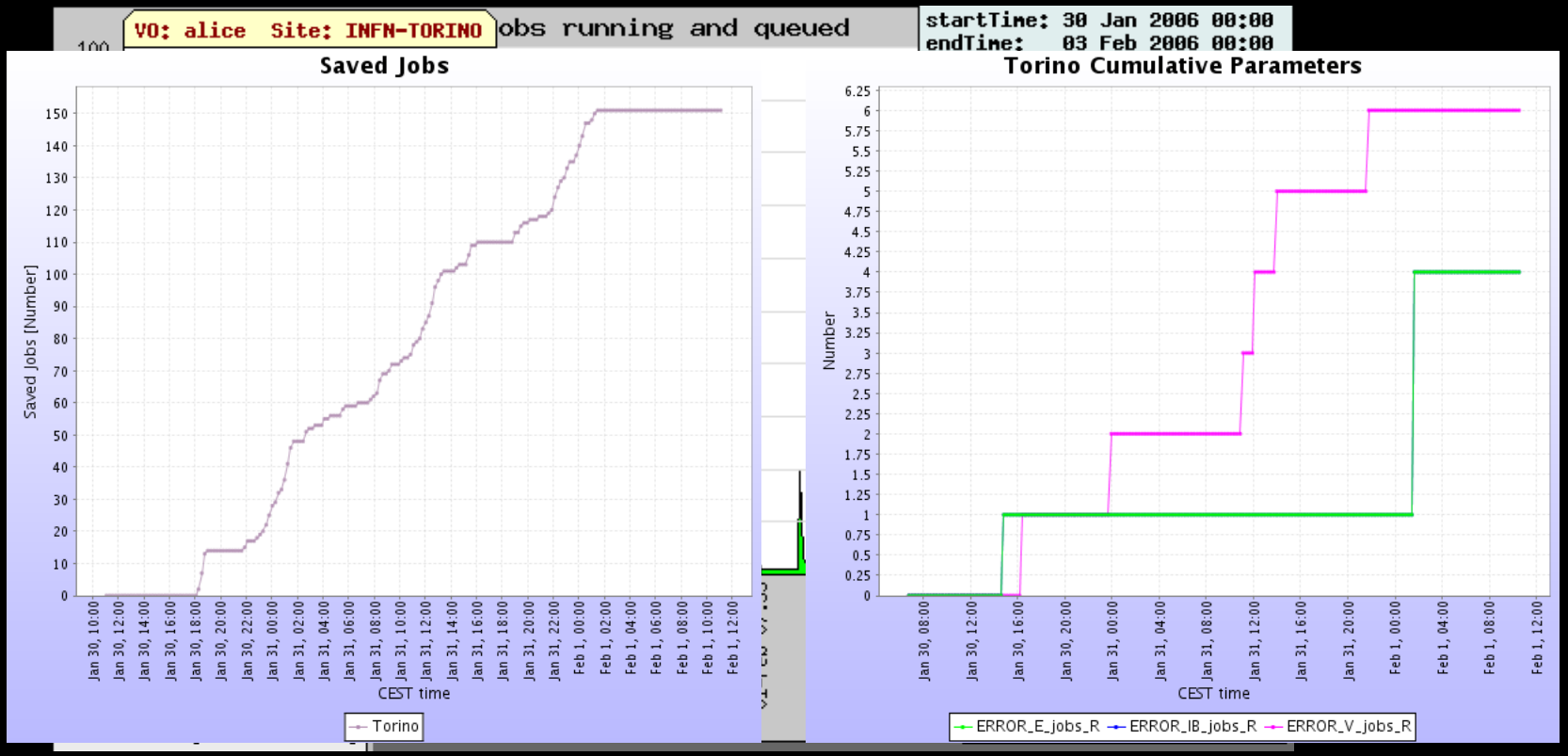
- Sites view: CNAF





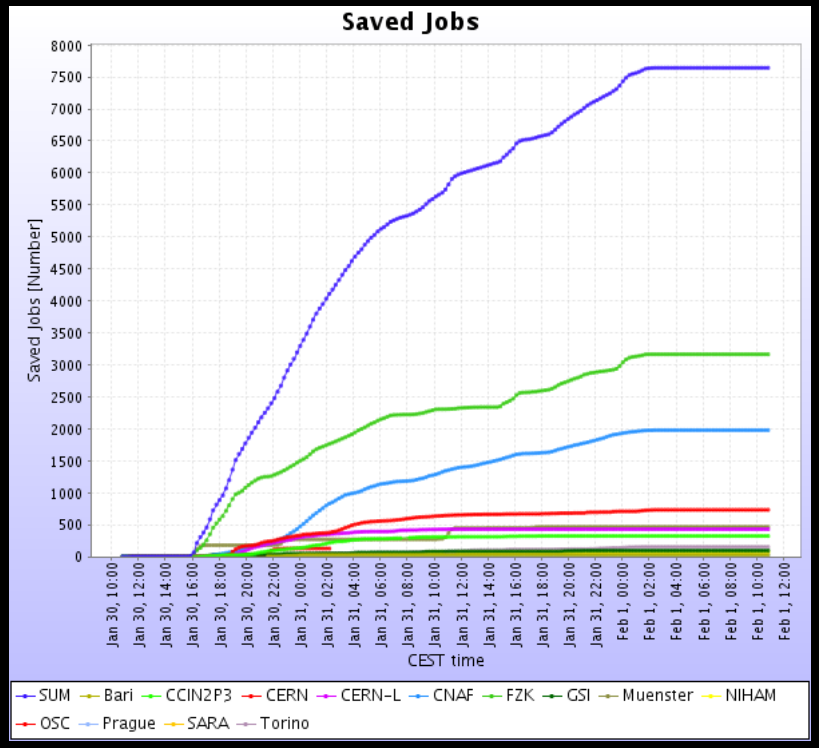
ALICE Data Challenge 2005 (-2006)

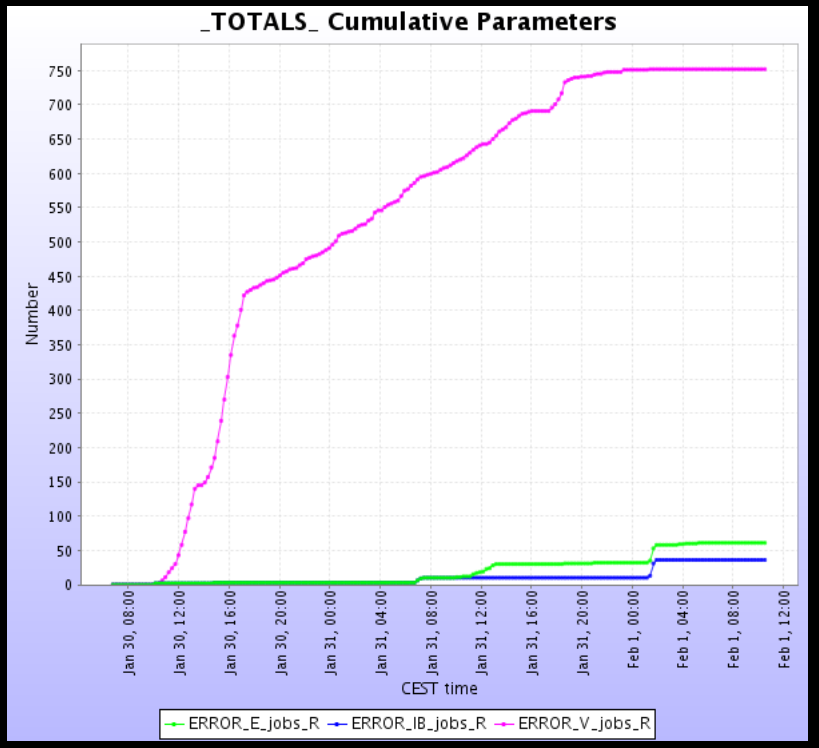
- Sites view: Torino



ALICE Data Challenge 2005 (-2006)

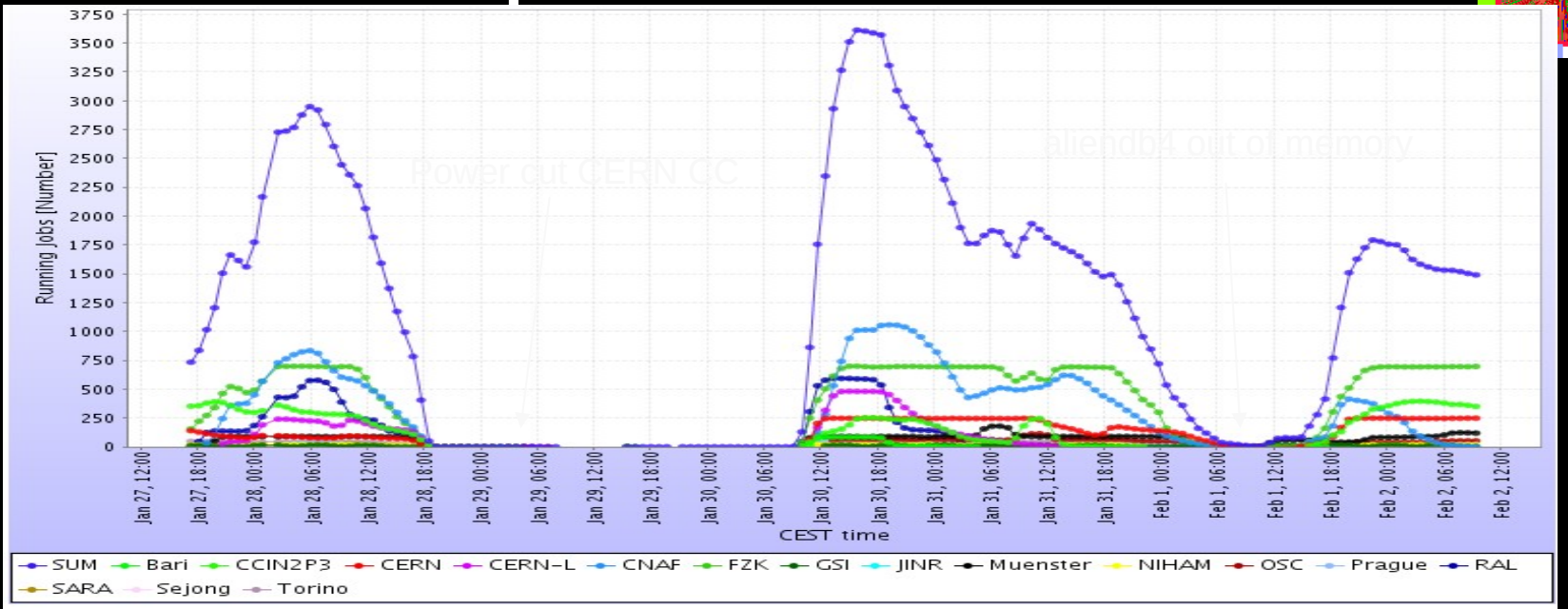
- Summary





About 10% errors, half of them from AliRoot

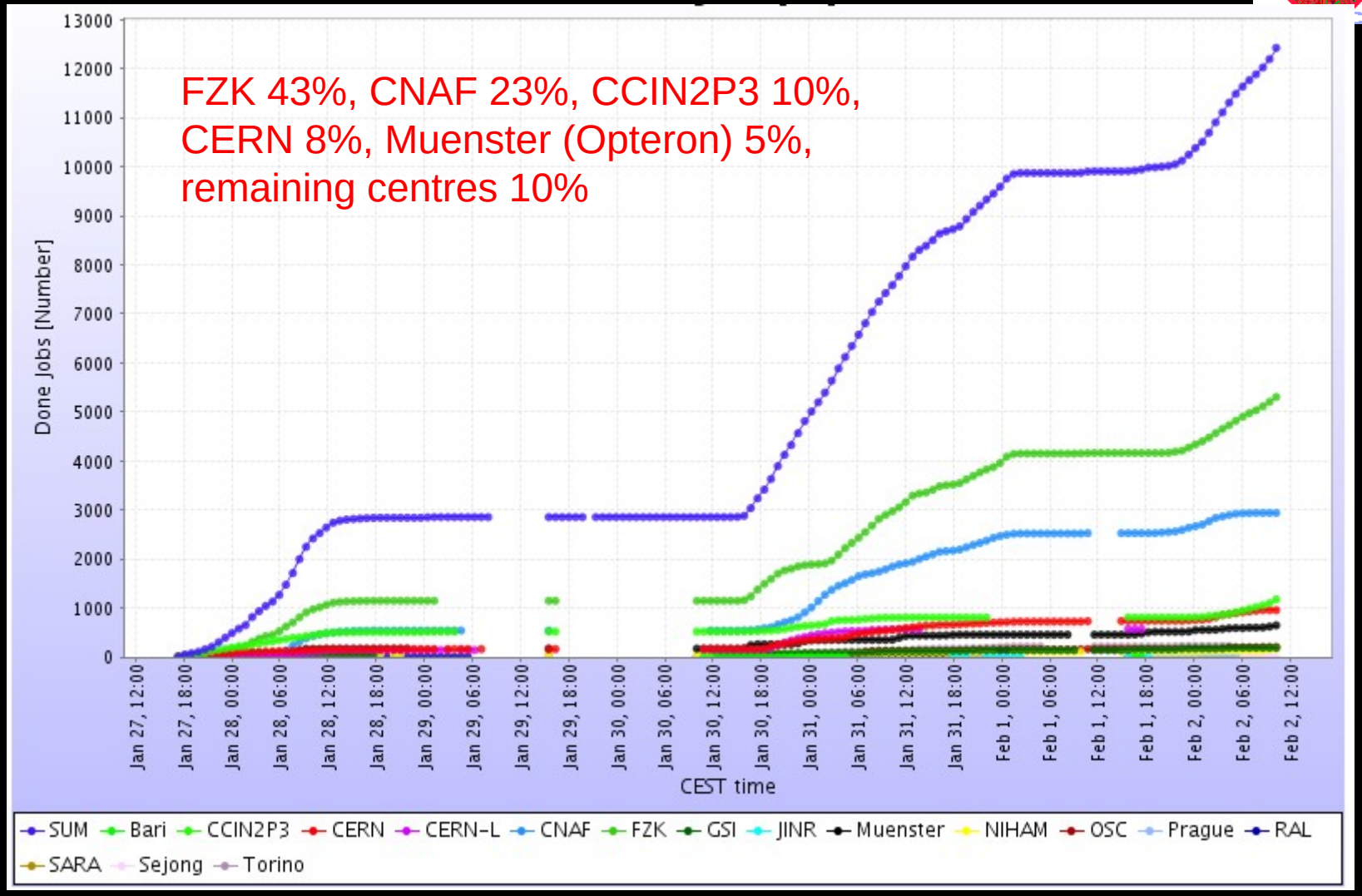
One week production status



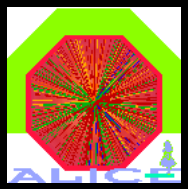
Alice Running Jobs [CS]				
Farm	Last value	Min	Avg	Max
SUM	1490	0	1183	3651
Bari	0	0	34.01	84
CCIN2P3	349	0	176.1	399
CERN	248	0	137.5	248
CERN-L	0	0	146.1	483
CNAF	4	0	363.7	1072
FZK	696	0	442.9	700
GSI	0	0	8.524	18
JINR	0	0	0.277	7
Muenster	119	0	76.63	180
NIHAM	20	0	14.95	34
OSC	52	0	45.17	120
Prague	0.425	0	11.5	68
RAL	0	0	161.4	595
SARA	0	0	10.49	30
Sejong	2	0	1.675	2
Torino	0	0	29.33	47

15 centres participating

1 week job completion

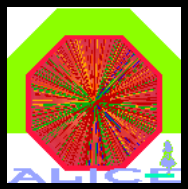


ALICE Data Challenge 2005(-2006) 1 week operation



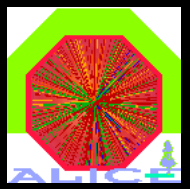
- Task queue:
  - Up to 3750 concurrently running jobs (1200 average), same amount of jobs waiting in the TQ
  - 12500 completed jobs:
    - 100 MSi2K CPU/hours
    - 2 TB output (CASTOR2@CERN) - 160 MB/job
- Central AliEn services (v.2-6):
  - Stable - no interventions
  - Very responsive: both catalogue and job management
  - Approximately 5 sec/job from submission to WAITING state in the TQ (with the current hardware – 17K jobs/day).
- Site agents/services (v.2-6):
  - Stable - no interventions

ALICE Data Challenge 2005(-2006) 1 week operation



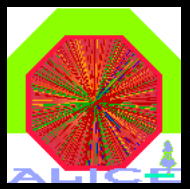
- Storage:
  - CASTOR2 – stable
    - Setting-up of combined xrootd-CASTOR disk pool at CERN
    - Hardware setup already exists at GSI and Lyon – need installaton/operation procedures
    - Site storage and scratch SEs - stable

ALICE Data Challenge 2005(-2006)



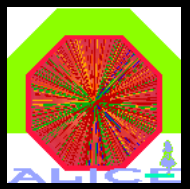
- Site tuning:
  - Added RAL – local queue issue is now resolved
  - SARA has no computing capacity (20-30 jobs max) – still to explore the possibility to submit from SARA VO-box to NIKHEF queues
  - Bari – submitting to 'short' queue
  - Catania – almost there
  - Russian T2s – Mikalai is working
  - US – at the moment Houston and OSC (Itanium queue being added)
  - ***Potentially 10 more T2s to be included in the production***

ALICE Data Challenge 2005 (-2006)



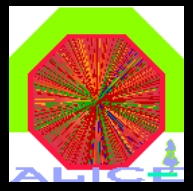
- Lessons learnt
  - Remarkable improvement in amount of available resources and stability wrt 2004
    - Stability of services was very good
  - ALICE VO production operations can be managed by a small team
  - Good and efficient coordination of operation with computing centres and LCG deployment team through ALICE-LCG task force group

ALICE Data Challenge 2005 (-2006)

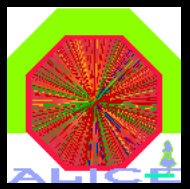


- Lessons learnt
  - Satisfactory integration with baseline LCG services through the VO-box setup
  - Demonstrated capability of central services and site components to manage computing resources on a level similar to the one expected in 2006
  - We are confident that the “production mode” for simulation and reconstruction will work

Part II
ALICE Data Challenge 2006

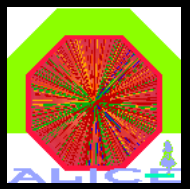


ALICE Data Challenges 2006



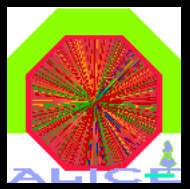
- Last chance to show that things are working together (i.e. to test our computing model)
- whatever does not work here is likely not to work when real data are there
  - So we better plan it well and do it well

ALICE Data Challenges 2006



- Three main objectives
  - Computing Data Challenge
    - Final version of rootifier / recorder
    - Online data monitoring
  - Physics data challenge
    - Simulation of signal events: 10^6 Pb-Pb, 10^{7-8} p-p
    - Final version reconstruction
    - Data analysis
  - PROOF data challenge
    - Preparation of the fast reconstruction / analysis framework

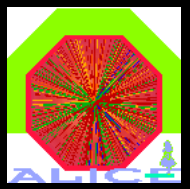
SC4 – Service Validation



- Service
  - Identification of key TierX ←→ TierY transfers
    - “dteam validation”
  - Validation by experiment productions
    - Service improvement

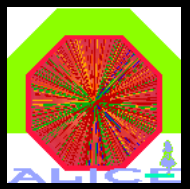
- Experiments: functionality and scalability
  - Full demonstration of experiment production
    - Full chain – data taking through to analysis!
  - Expansion of services from production to general users
  - Ramp-up in number of sites; service level
  - Ramp-up in compute / storage / throughput capacity
  - Accompanied by agreed monitoring of actual and historical service level delivery

ALICE SC4 Use Cases



- Not covered so far in Service Challenges:
  - T0 recording to tape (and then out)
  - Reprocessing at T1s
- Calibrations & distribution of calibration data
  - HEPCAL II Use Cases
  - Individual (mini-) productions (if / as allowed)
- Additional services to be included
  - Full VOMS integration
  - PROOF, xrootd, ... (analysis services in general...)

Main points

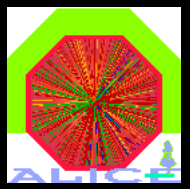


- Data flow

- Realistic system stress test
- Network stress test

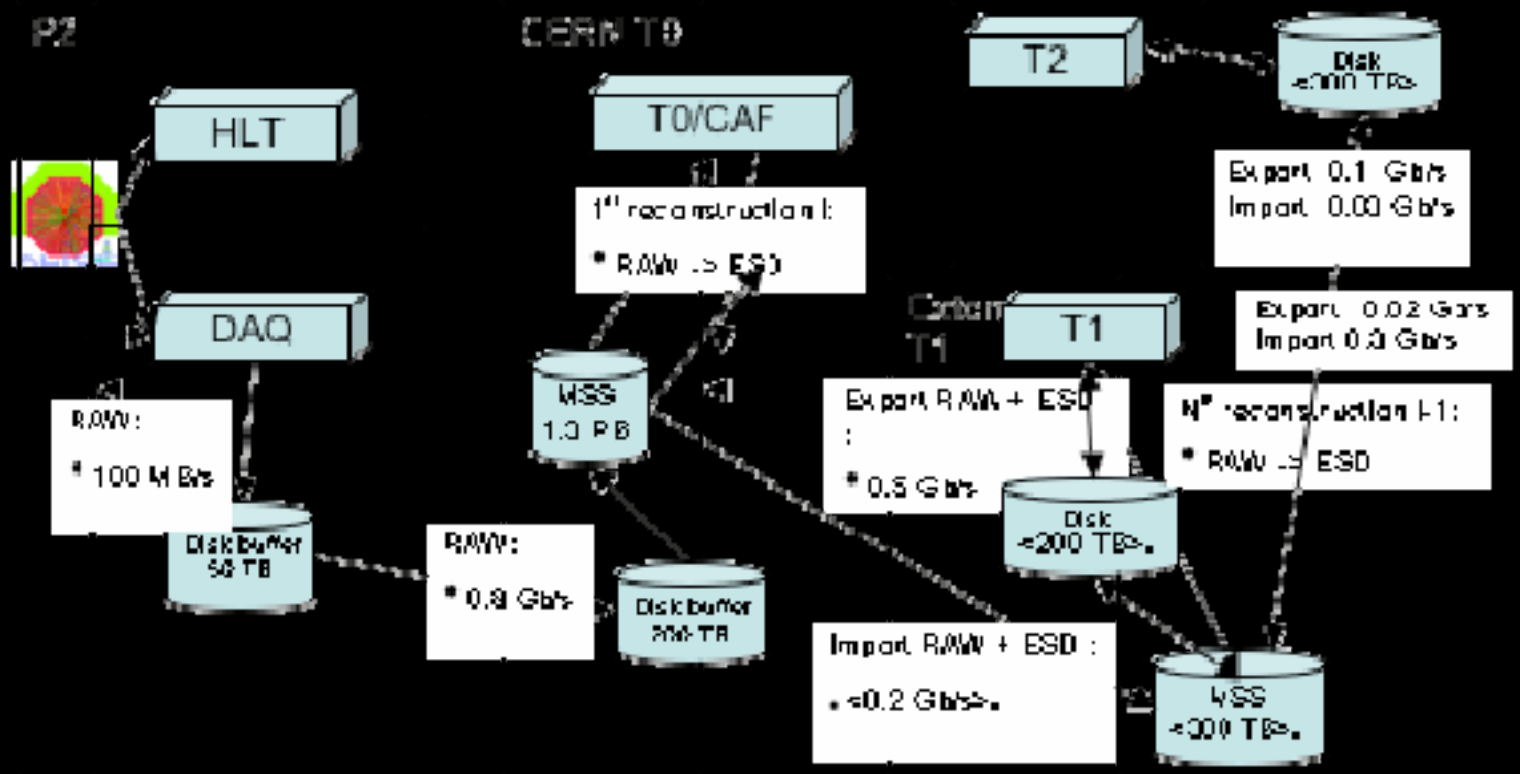
- SC4 Schedule
- Analysis activity

Data Flow

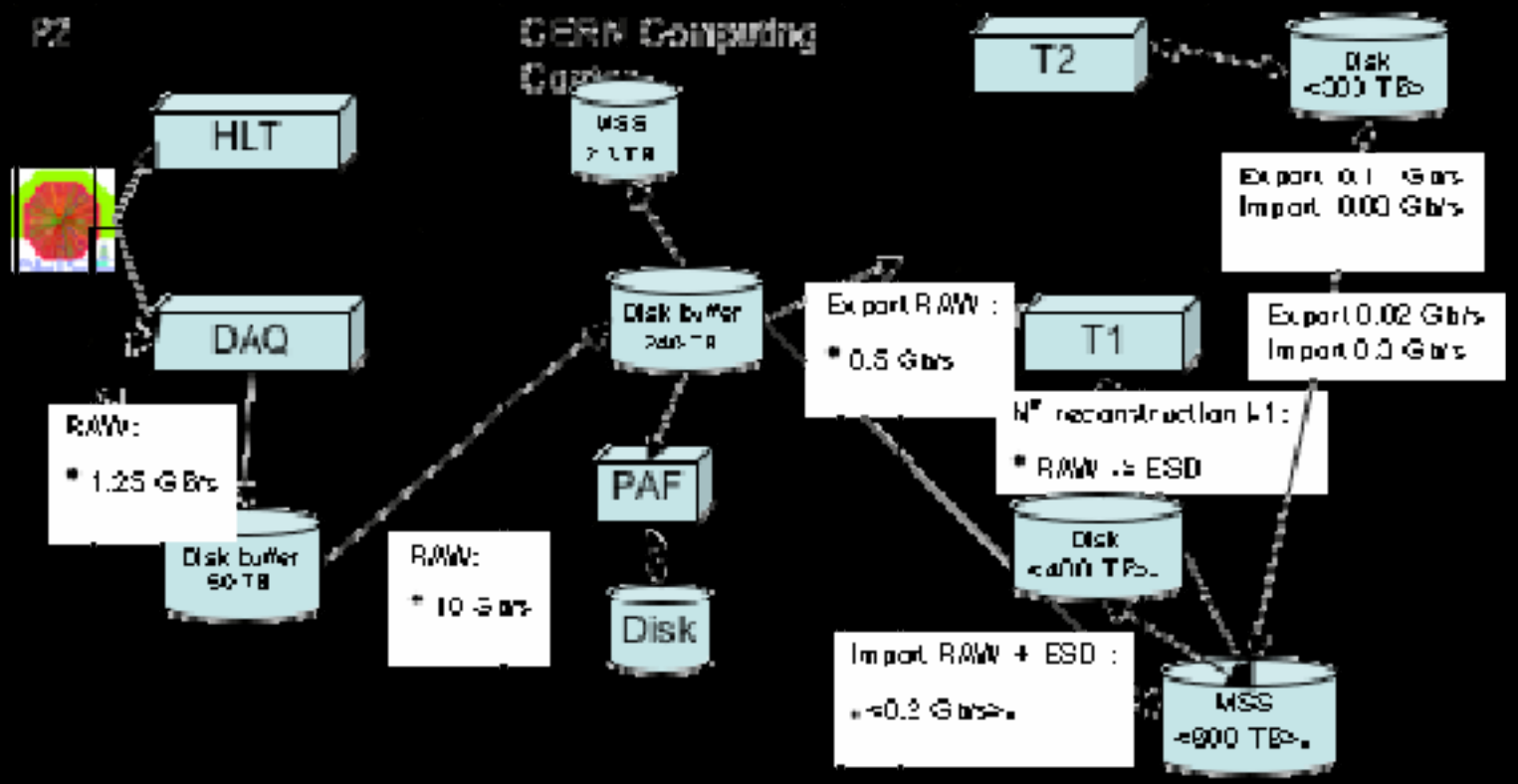


- Not very fancy… always the same
- Distributed Simulation Production
  - Here we stress-test the system with the number of jobs in parallel
- Data back to CERN
- First reconstruction at CERN
  - RAW/ESD Scheduled “push-out” – here we do the network test
- Distributed reconstruction
  - Here we stress test the I/O subsystem
- Distributed (batch) analysis
  - “And here comes the proof of the pudding” - FCA

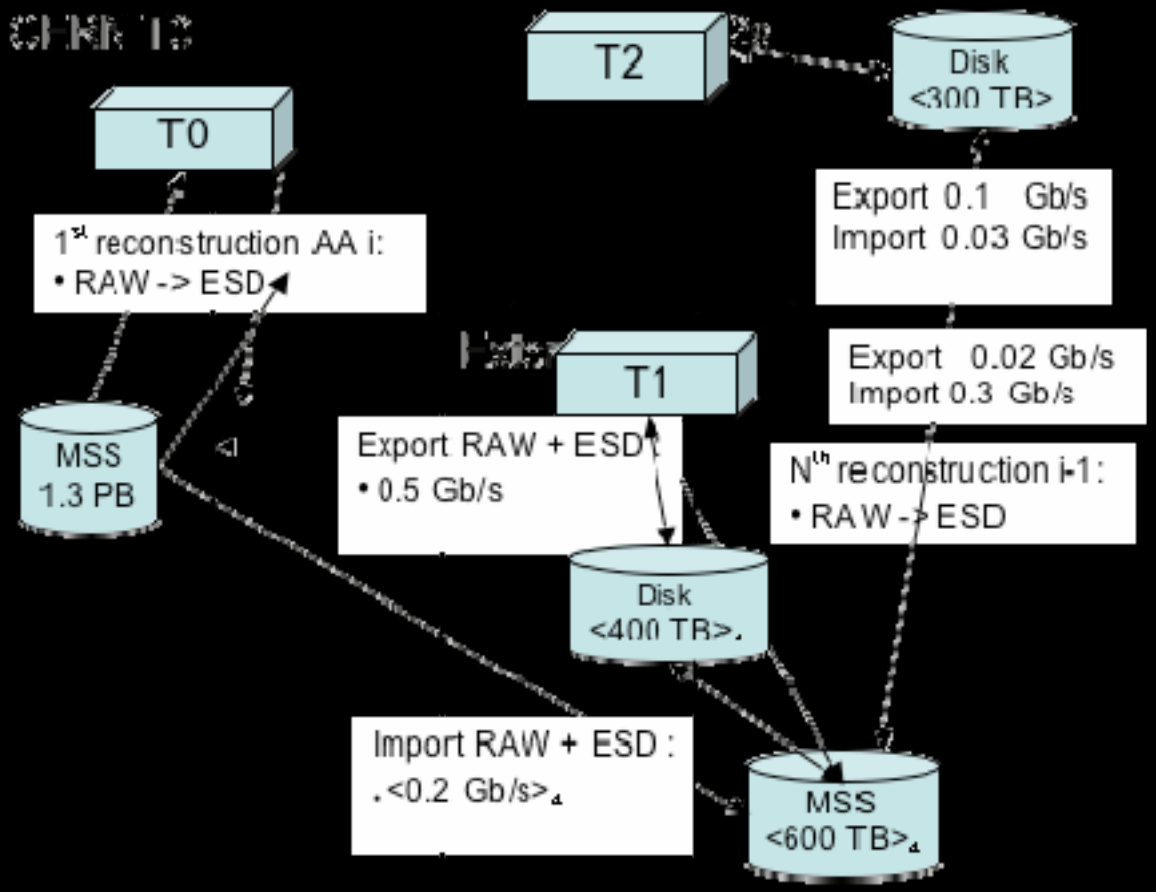
Proton - Proton Run



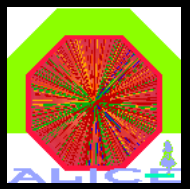
Heavy Ion Run



Shutdown Periods

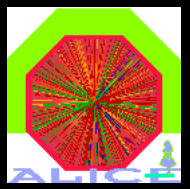


Realistic system stress test



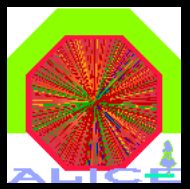
- Number of jobs (up to 3750 until now)
  - MC production
- Number of jobs reading/storing files
  - Reconstruction
- Number of files accessed by a job, job splitting, many users
  - Analysis
- Network transfer
  - Push out of files before distributed reconstruction

System Parameters / Services



- Simulation / Reconstruction
  - Number of jobs (UI, RB, CE - up to 3750 until now)
- Network
  - Push out of files before distributed reconstruction (FTS)
- Analysis
  - Number of users & roles --> VOMS
  - Number of queries: Metadata --> AliEn File Catalogue
  - Batch mode
    - Job splitting (AliEn)
  - Interactive mode
    - Parallel Distributed Analysis (PROOF)
  - Number of files accessed by a job --> xrootd, ...

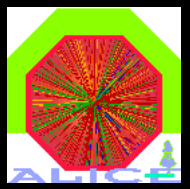
Network stress test



- T0⇔T1 easy to do, it comes with production
- T1⇔T(x>0) has to be scheduled or folded with analysis
- We plan to use FTS, via jobs in the AliEn TQ or directly
- Expected flow for the first test:

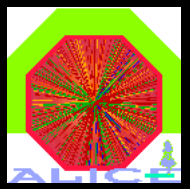
T1 centre	Number of CPUs	Number of reconstruction jobs per day	Incoming data rate [MB/s]	Data volume in local storage [GB/day]
CNAF	396	4000	12	90
CCIN2P3	220	2240	7	50
GridKa	363	3400	11	82
Total	979	9380	30	222

Analysis



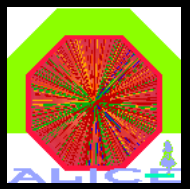
- We have to analyse data while we take them
  - Fast feedback to production
    - Calibration / Alignment
    - Data filtering and search for signals
- We need an operational proof@caf (March 2006) at CERN
  - 1MSI2k
  - 500 today's bi-processors
  - 170pp/s or 1.2PbPb/s

SC3 -> SC4 Schedule



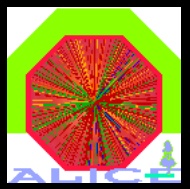
- February 2006
  - Rerun of SC3 disk – disk transfers (max 150MB/s X 7 days)
  - Transfers with FTD, either triggered via AliEn jobs or scheduled
  - T0 -> T1 (CCIN2P3, CNAF, Grid.Ka, RAL)
- March 2006
  - T0-T1 “loop-back” tests at 2 x nominal rate (CERN)
  - Run bulk production @ T1,T2 (simulation+reconstruction jobs) and send data back to CERN
  - (We get ready with proof@caf)
- April 2006
  - T0-T1 disk-disk (nominal rates) disk-tape (50-75MB/s)
  - First Push out (T0 -> T1) of simulated data, reconstruction at T1
  - (First tests with proof@caf)
- July 2006
  - T0-T1 disk-tape (nominal rates)
  - T1-T1, T1-T2, T2-T1 and other rates TBD according to CTDRs
  - Second chance to push out the data
  - Reconstruction at CERN and remote centres
- September 2006
  - Scheduled analysis challenge
  - Unscheduled challenge (target T2’s?)

SC4 Rates - Scheduled Analysis



- Users
  - Order of 10 at the beginning of SC4
- Input
  - 1.2M Pb-Pb events, 100M p-p events, ESD stored at T1s
- Job rate
  - Can be tuned, according to the availability of resources
- Queries to MetaData Catalogue
  - Time/Query to be evaluated (does not involve LCG services)
- Job splitting
  - Can be done by AliEn according to the query result (destination set for each job)
  - CPU availability is an issue (sub-jobs should not wait too much for delayed executions)
  - Result merging can be done by a separate job
- Network
  - Not an issue

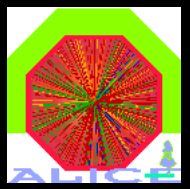
SC4 Rates - Scheduled Analysis



- Some (preliminary) numbers
  - Based on 20 minutes jobs

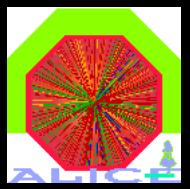
Centre	ESD volume in local storage [TB]	Number of available CPUs	Number of analysis jobs/day	Number of passes over the ESD sample/day	Data I/O per CPU/day [GB]	Aggregated I/O from local storge [GB/s]
CERN	26	1000	72,000	16	430	4.8
CCIN2P3	10.5	220	16,000	9	430	2.7

SC4 Rates - Unscheduled Analysis



- To be defined

Summary



- Simulation / Reconstruction is likely to work well
- Scheduled Analysis is not much different (but it requires few more services)
  - It includes calibration/alignment
  - If required, it can be run on input data concentrated in one Tx (x=0,1)
- Unscheduled analysis will introduce more complexity, but it will come later

Resources

Site	*CPU (MKSI2K)*	*Disk (TB)*	*Tape (TB)*	*BW to CERN/T1 (Gb/s)*
CERN	693 (107%)	231 (109%)	170 (105%)	10
CCIN2P3	220 (100%)	105 (100%)	95 (100%)	10
CNAF*	396 (100%)	187 (100%)	187 (100%)	10
INFN T2*	363 (100%)	59 (100%)	106 (100%)	10
GridKa	363 (87%)	59 (68%)	106 (62%)	10
RAL	20 (83%)	2 (20%)	2 (20%)	10
UK T2 *	121 (100%)	16 (100%)	0	
NDGF*	164 (100%)	58 (100%)	101 (100%)	5
USA	423 (235%)	21 (54%)	25 (100%)	
FZU Prague	60 (100%)	14 (100%)	0	1
RDIG	240 (48%)	10 (6%)	0	1
French T2	130 (146%)	28 (184%)	0	0.6
GSI	100 (100%)	30 (100%)	0	1
U. Muenster	132 (100%)	10 (100%)	0	1
Polish T2*	198 (100%)	7.1 (100%)	0	0.6
Slovakia	25 (100%)	5 (100%)	0	0.6
Total	4142 (132%)	913 (105%)	689 (100%)	