Patricia Méndez Lorenzo (CERN)

THE ALICE GRID MODEL

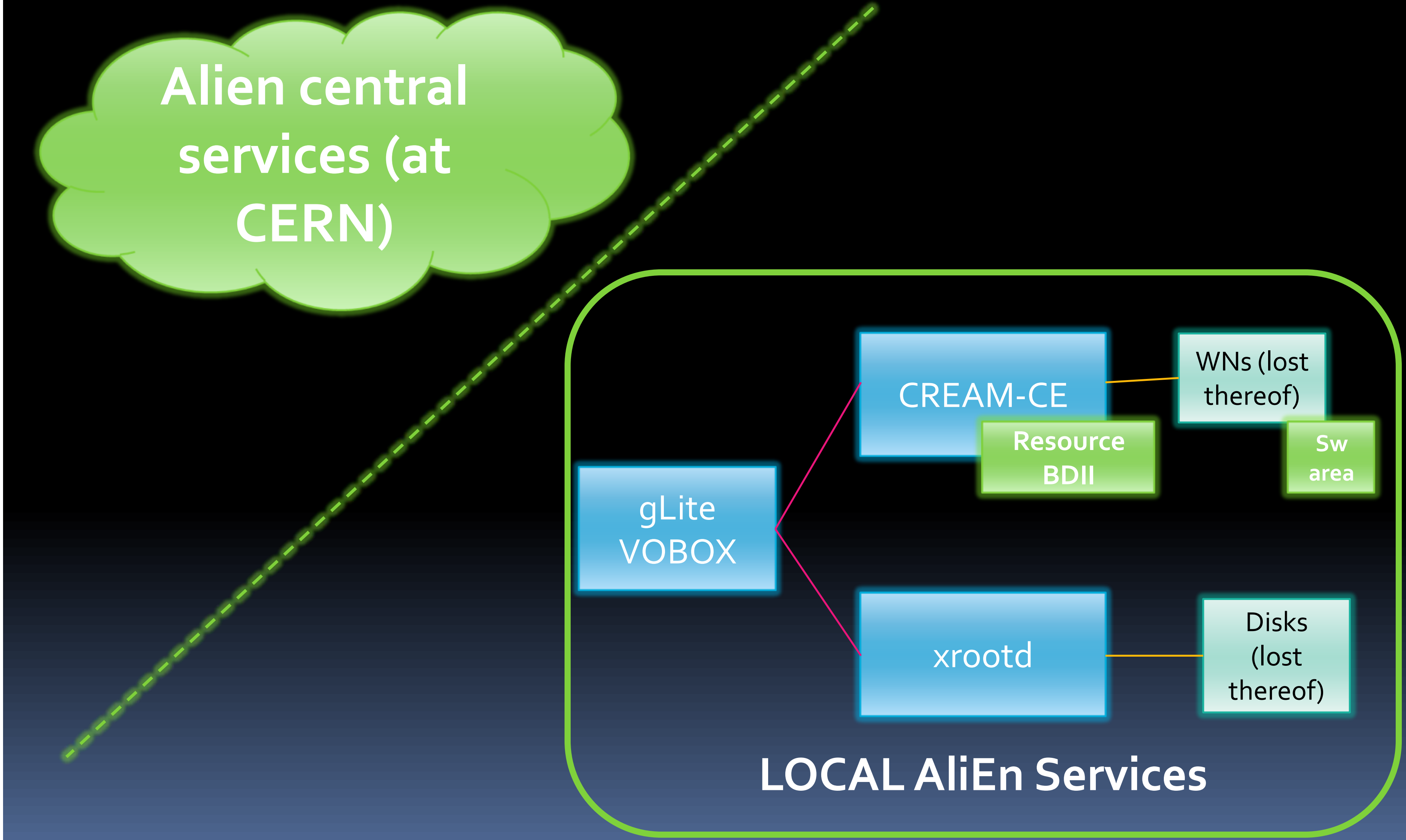
Outlook

- Your ALICE site
  - ALICE Services
  - Generic Services
- Interoperability
- User Analysis
- Monitoring
- Operations

The ROC_LA

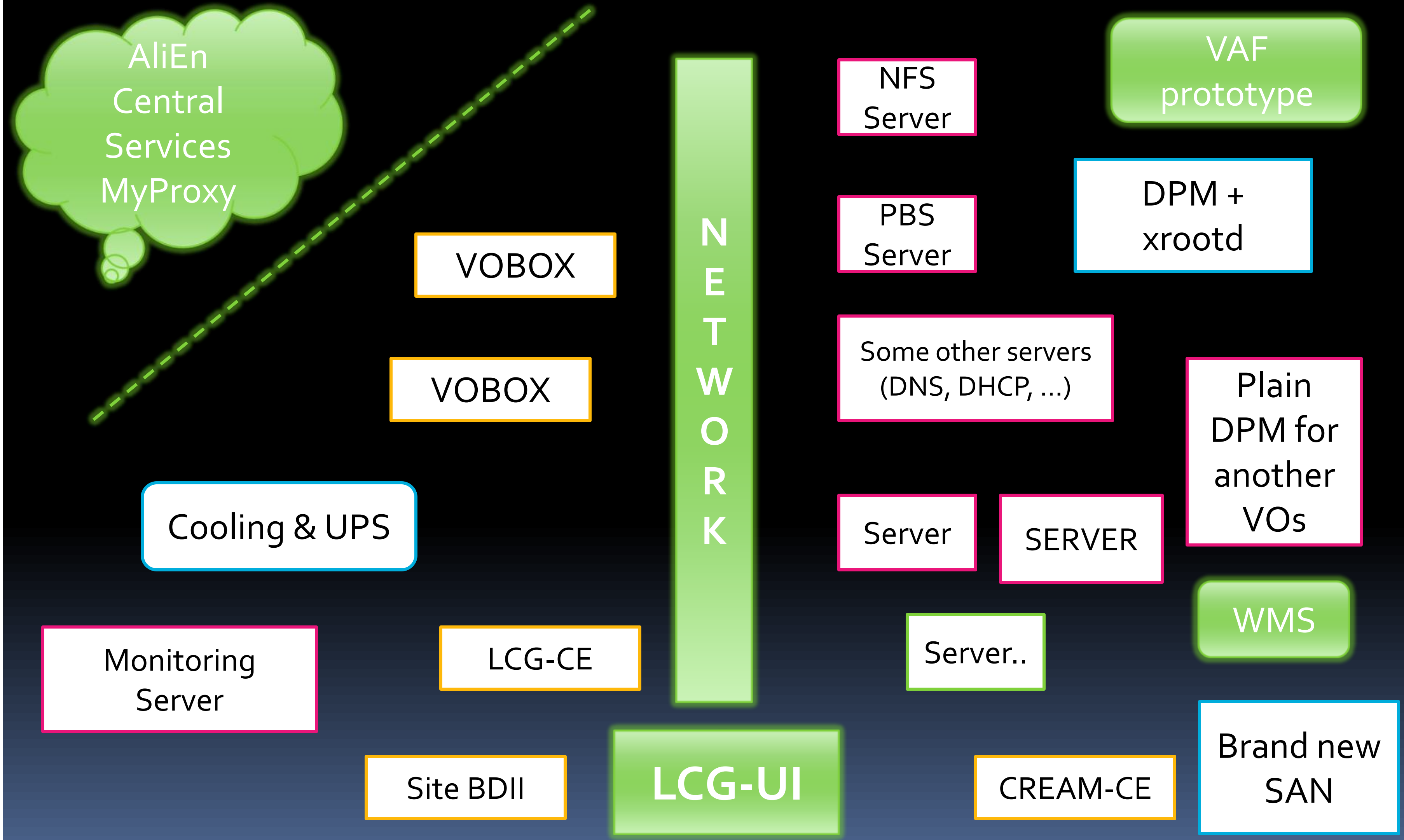
- ALICE clearly supports the creation of a ROC infrastructure and is willing to provide expertise
- The most important aspect to consider: THE NETWORK
- ALICE has established a faithfully communication channel with the Network experts at CERN who are helping the USA and ASIA sites to improve the bandwidth infrastructure with CERN

How I see the ALICE sites



The rest of the talk will concentrate on how I see your site

How you see your site



AliEn Central Services MyProxy

VAF prototype

NFS Server

PBS Server

DPM + xrootd

VOBOX

VOBOX

NETWORK

Some other servers (DNS, DHCP, …)

Plain DPM for another VOs

Cooling & UPS

Server

SERVER

WMS

Monitoring Server

LCG-CE

Server..

Brand new SAN

Site BDII

LCG-UI

CREAM-CE

The ALICE Computing model

- From the detector …

AliEn

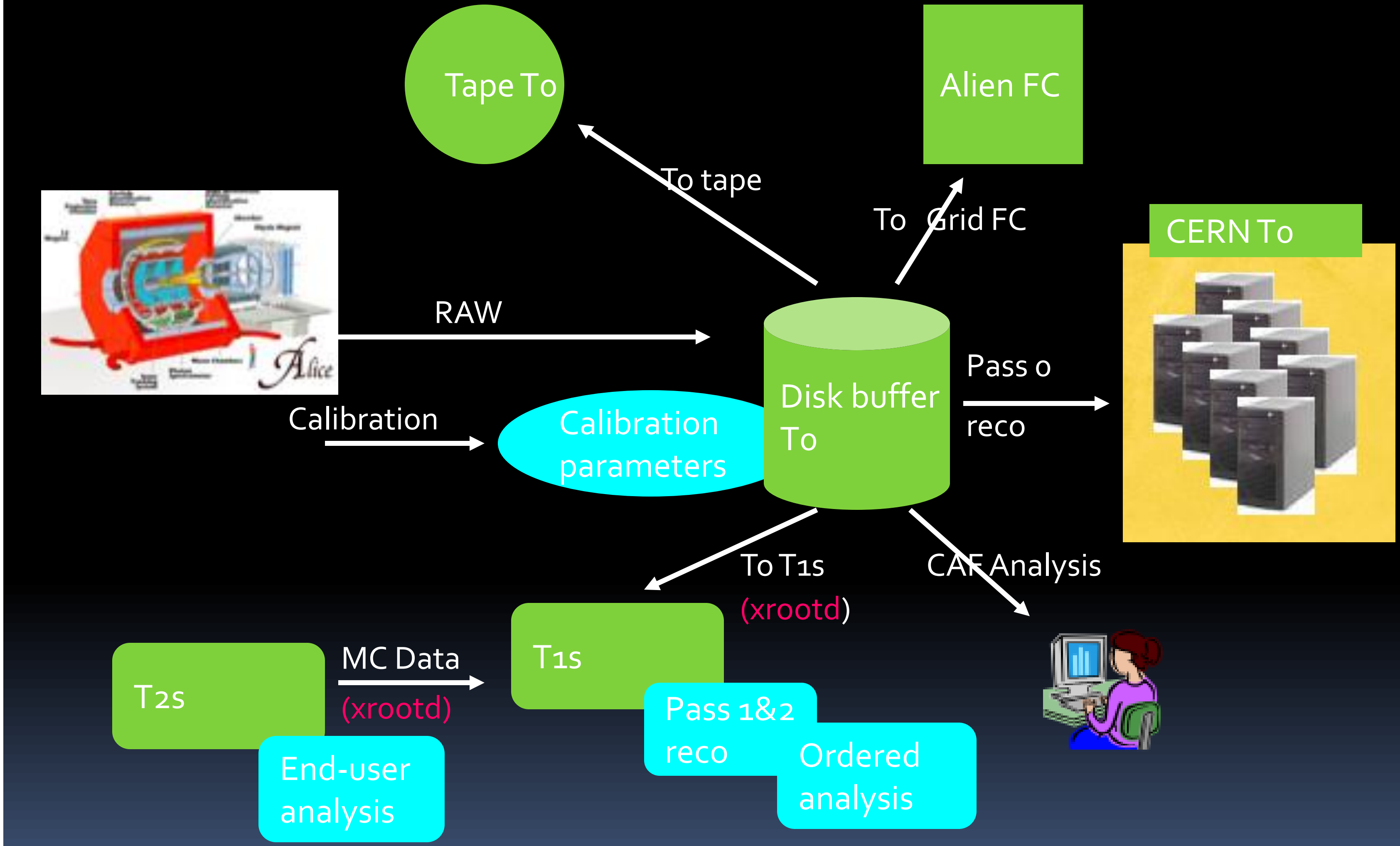
- Reconstruction
- Simulation
- Storing data and distribution
- Analysis

- … to the physics paper

ALICE submission in 6 points

- Job agent infrastructure
  - Submitted through the Grid infrastructure available at the site
- Real jobs held in a **central queue** handling priorities and quotas
- Agents submitted to provide a standard environment (job wrapper) across different systems
- Real jobs **pulled** by the sites
- Automatic operations
- Extensive monitoring

Computing Model – (pp case)



Tape T0

Alien FC

To tape

To Grid FC

CERN T0

RAW

Disk buffer T0

Pass 0 reco

Calibration

Calibration parameters

To T1s (xrootd)

CAF Analysis

T2s

MC Data (xrootd)

T1s

Pass 1&2 reco

Ordered analysis

End-user analysis

- PRINCIPLE OF OPERATION: THE VOBOX

The ALICE VOBOX

ALICE REQUIRES THE GENERIC GLITE3.2 VOBOX

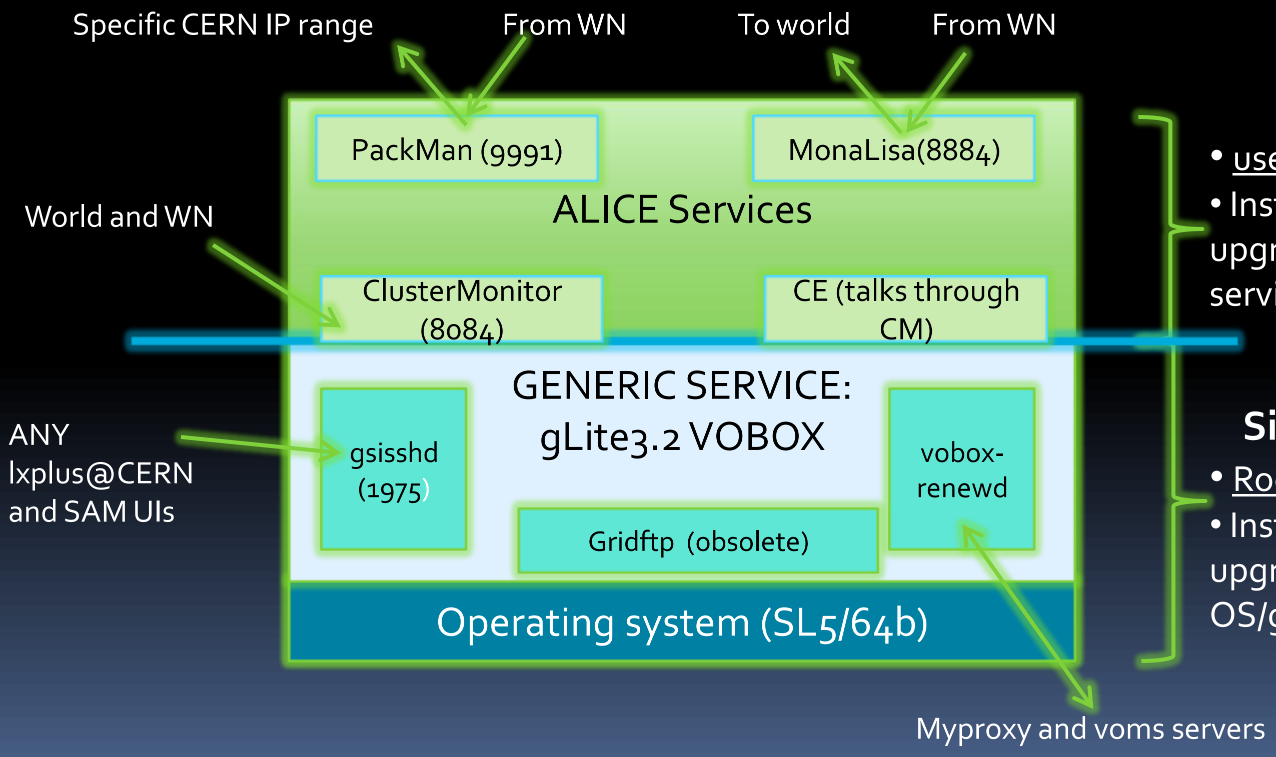
- Primary service required at every site
  - Service scope:
    - Run (local) specific ALICE services
    - File system access to the experiment software area
  - Mandatory requirement to enter the production
    - Independently of the middleware stack available
    - **Same concept valid for gLite, OSG, ARC and generic AliEn sites**
- Same setup requirements for all sites
  - Same behavior for T1 and T2 sites in terms of production
  - Differences between T1 and T2 sites a QoS matter only
    - Applicable to ALL generic services provided at the sites
- Service related problems should be managed by the site administrators
- Support ensured at CERN for any of the middleware stacks

The gLite3.2 VOBOX

- The gLite-VOBOX is a gLite-UI with 2 added values:
  1. Automatic renewal of the user proxy
     - Needed to ensure a good behavior of the experiment services
  2. Direct access to the experiment software area
     - Access restricted to the voms role: lcgadmin
     - User mapped to SINGLE local account
       - Unique service not accepting pool accounts
  - Access to the machine based on a valid voms-aware user proxy

The gLite-VOBOX layers

- 70% of the ALICE VOBOXES are running the gLite middleware



ALICE responsibility

- user privileges (gsissh)
- Installation, maintenance, upgrades and follow up for ALICE services

Site admin responsibility

- Root privileges
- Installation, maintenance, upgrades and follow up for OS/generic service

- THE ALICE WORKLOAD MANAGEMENT SYSTEM

ALICE Workload Management System

- The ALICE tendency in the last two years is to minimize the number of services used to submit job agents
  - Already ensured with ARC, OSG and AliEn native sites
  - gLite sites can also follow this approach
    - Through the CREAM-CE system
    - ALICE is asking for the deployment of this service at all sites since 2008
    - ALICE has deprecated the gLite3.X-WMS with the deployment of AliEn v.2.18
      - The WMS submission module still available in AliEn v.2.18
      - ALL ALICE sites have to provide a (at least) CREAM

The CREAM-CE

- 1st experiment to put CREAM-CE in production and provide developers with important feedback
  - 2008 approach: Testing
    - Testing phase (performance and stability) at FZK
  - 2009 approach: AliEn implementation and distribution
    - System available at T0, all T1 (but NIKHEF) and several T2 sites
      - Dual submission (LGC-CE and CREAM-CE) at all sites providing CREAM
      - Second VOBOX was required at the sites providing CREAM to ensure the duality approach asked by the WLCG-GDB
  - 2010 approach: Deprecation of the WMS
    - At this moment ALL ALICE sites provide at least one CREAM-CE
- ALICE is activelly involved in the operation of the service at all sites together with the site admins and the CREAM-CE developers

- ALICE DATA MANAGEMENT SYSTEM AND RAW DATA TRANSFERS

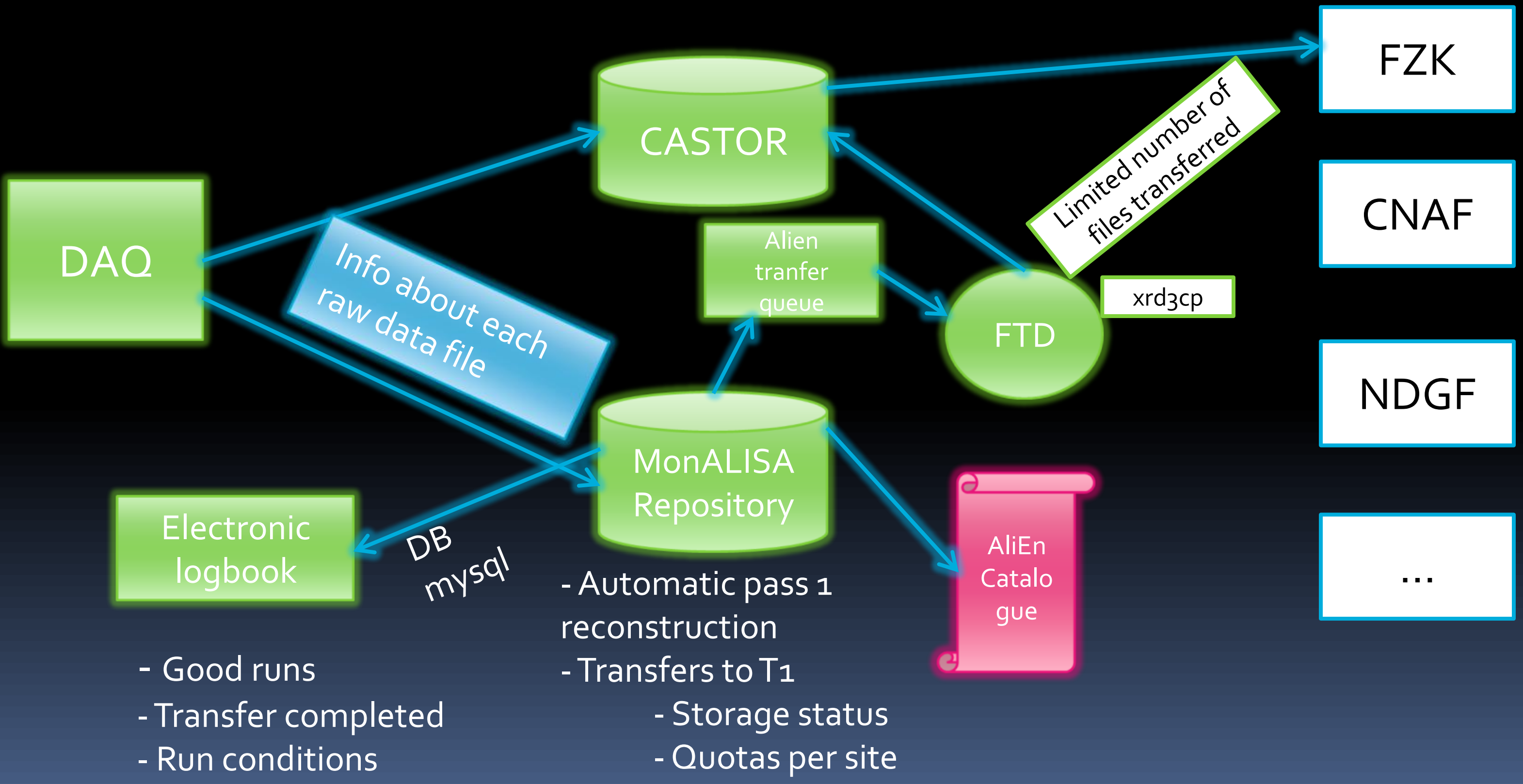
Data Management system

- Data access and file transfer: via xrootd
  - Data access
    - T0 and T1 sites: Interfaced with all supported MSS
    - T2 sites: xrootd pure setup (ALICE advice)
  - File transfer
    - T0-T1: raw data transfer for custodial and reconstruction purposes
      - No specific assignments between data and sites
    - T1-T2: reconstructed data and analysis

    (NOTE: T1-T2 assignment not predefined in the ALICE computing model)
- Default behavior: Jobs are sent where data are physically stored
  - External files can be locally copied or remotely access based on the user decision
  - Outputs of the jobs locally stored and replicated to (in default) two external sites
    - Choice based on ML status report of the SE and physical proximity

Raw data transfers

- To-T1 transfers performed via xrootd
  - 3rd party copy (xrd3cp) enabled
    - Co-existing with xrdcp



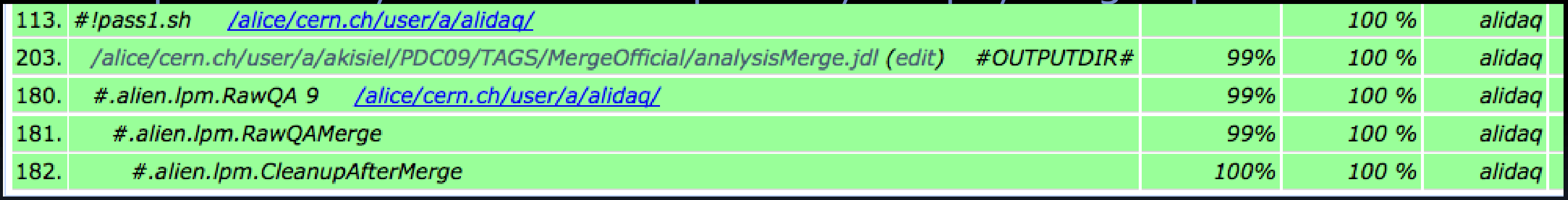
Site connection and interoperability

- Sites connection
  - Only to the data management level
  - In terms of job submission and services sites are independent with each other
    - ALICE computing model foresees same service infrastructure at all sites
  - ALICE assumes the T1-T2 connection established by the ROC infrastructure for site control and issue escalation
- Interoperability between sites
  - Defined to the AliEn level
    - Definition of plug-ins for each middleware stack
    - Same AliEn setup for all sites

- USER LAYER: ANALYSIS

User Analysis (I)

- Analysis trains
  - Grouping many analysis tasks in a common data set
    - Allows for better CPU/Wall and reduces load on the storage servers
    - Pass 1 reconstruction is automatically followed by specific Pilot trains
      - Assess the quality of the run (detector by detector, global tracking and vertex position and stability)
    - Specific analysis activities required by the physics groups



113.	#!pass1.sh /alice/cern.ch/user/a/alidaq/		100 %	alidaq
203.	/alice/cern.ch/user/a/akisiel/PDC09/TAGS/MergeOfficial/analysisMerge.jdl (edit) #OUTPUTDIR#	99%	100 %	alidaq
180.	#.alien.lpm.RawQA 9 /alice/cern.ch/user/a/alidaq/	99%	100 %	alidaq
181.	#.alien.lpm.RawQAMerge	99%	100 %	alidaq
182.	#.alien.lpm.CleanupAfterMerge	100%	100 %	alidaq

- Reconstruction is started as soon as data is registered in CASTOR
- Analysis trains automatically triggered at the end of the reconstruction activities
- At the end of the analysis, merge and cleanup procedures are executed

User Analysis (II)

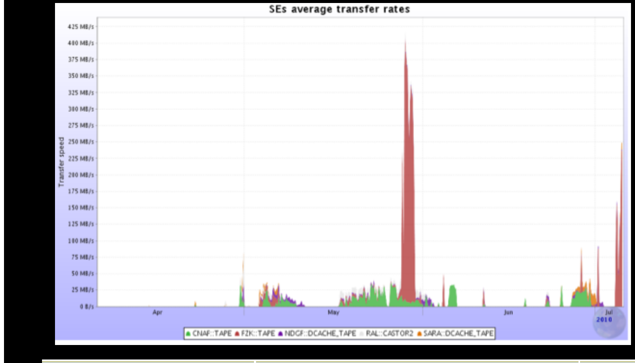
- Chaotic analysis
  - User analysis on the Grid
  - Internal ALICE prioritization within the common task queue works well
    - Production user is demoted in favor of users in the queue
    - Generic pilot jobs assure fast execution of user jobs at the sites

User Analysis (III)

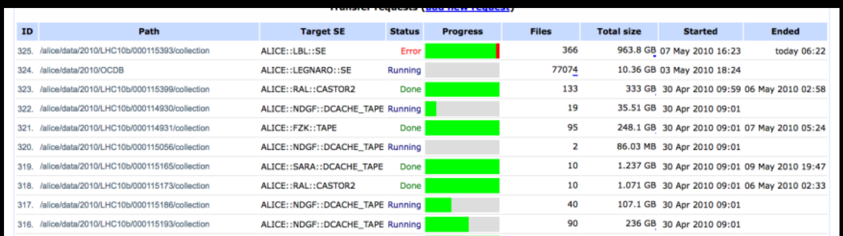
- Fast and interactive analysis in PROOF clusters
  - Processing of reconstructed data, calibration and alignment
    - Set of pre-stage files already available at the WNs of the PROOF clusters
    - Possibility to pre-stage any kind of file required by the users and registered in the AliEn file catalogue
    - Limitations:
      - Available space in WNs
      - Memory consumption affecting the proof master
  - Very popular system for the users
  - CAF@CERN
    - For the moment unique stable system
    - 1 proof master and 26 nodes distributed in two sets (3 and 4 disk nodes)
  - GSIAF@GSI, SAF@Subatech, LAF@CCIN2P3
  - Status and development weekly followed during the TF&AF meeting

- MONITORING

Examples of MonALISA monitoring information

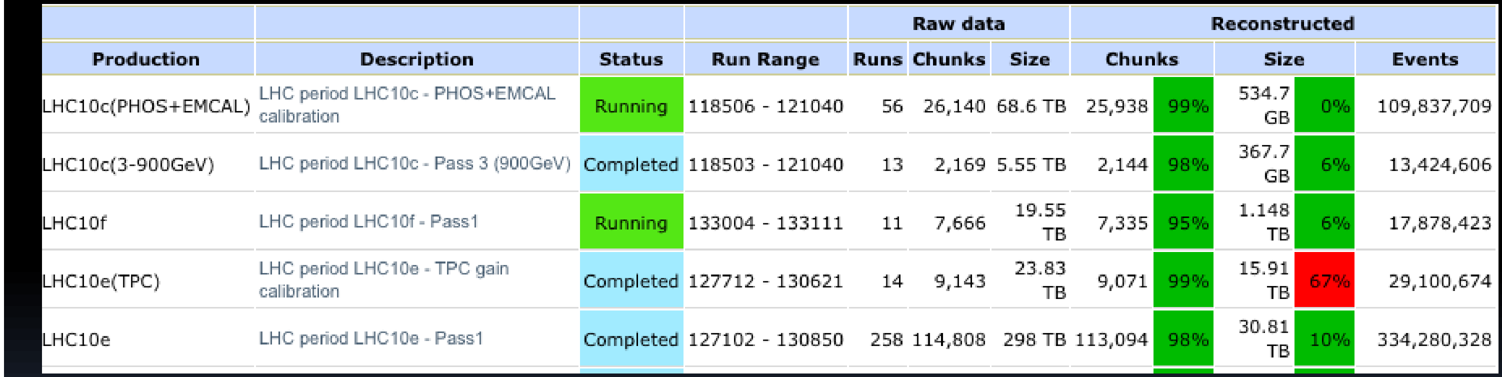


Raw data transfers



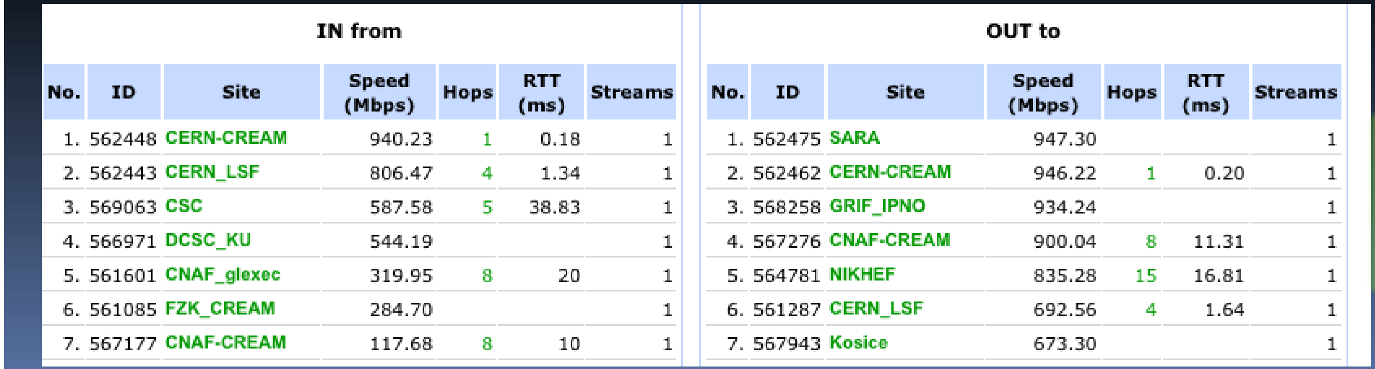
Transfer requests (add new request)

ID	Path	Target SE	Status	Progress	Files	Total size	Started	Ended
325.	/alice/data/2010/LHC10b/000115393/collection	ALICE::LBL::SE	Error		366	963.8 GB	07 May 2010 16:23	today 06:22
324.	/alice/data/2010/OCDB	ALICE::LEGNARO::SE	Running		77074	10.36 GB	03 May 2010 18:24	
323.	/alice/data/2010/LHC10b/000115399/collection	ALICE::RAL::CASTOR2	Done		133	333 GB	30 Apr 2010 09:59	06 May 2010 02:58
322.	/alice/data/2010/LHC10b/000114930/collection	ALICE::NDGF::DCACHE_TAPE	Running		19	35.51 GB	30 Apr 2010 09:01	
321.	/alice/data/2010/LHC10b/000114931/collection	ALICE::FZK::TAPE	Done		95	248.1 GB	30 Apr 2010 09:01	07 May 2010 05:24
320.	/alice/data/2010/LHC10b/000115056/collection	ALICE::NDGF::DCACHE_TAPE	Running		2	86.03 MB	30 Apr 2010 09:01	
319.	/alice/data/2010/LHC10b/000115165/collection	ALICE::SARA::DCACHE_TAPE	Done		10	1.237 GB	30 Apr 2010 09:01	09 May 2010 19:47
318.	/alice/data/2010/LHC10b/000115173/collection	ALICE::RAL::CASTOR2	Done		10	1.071 GB	30 Apr 2010 09:01	06 May 2010 02:33
317.	/alice/data/2010/LHC10b/000115186/collection	ALICE::NDGF::DCACHE_TAPE	Running		40	107.1 GB	30 Apr 2010 09:01	
316.	/alice/data/2010/LHC10b/000115193/collection	ALICE::NDGF::DCACHE_TAPE	Running		90	236 GB	30 Apr 2010 09:01	



				Raw data			Reconstructed				
Production	Description	Status	Run Range	Runs	Chunks	Size	Chunks		Size		Events
LHC10c(PHOS+EMCAL)	LHC period LHC10c - PHOS+EMCAL calibration	Running	118506 - 121040	56	26,140	68.6 TB	25,938	99%	534.7 GB	0%	109,837,709
LHC10c(3-900GeV)	LHC period LHC10c - Pass 3 (900GeV)	Completed	118503 - 121040	13	2,169	5.55 TB	2,144	98%	367.7 GB	6%	13,424,606
LHC10f	LHC period LHC10f - Pass1	Running	133004 - 133111	11	7,666	19.55 TB	7,335	95%	1.148 TB	6%	17,878,423
LHC10e(TPC)	LHC period LHC10e - TPC gain calibration	Completed	127712 - 130621	14	9,143	23.83 TB	9,071	99%	15.91 TB	67%	29,100,674
LHC10e	LHC period LHC10e - Pass1	Completed	127102 - 130850	258	114,808	298 TB	113,094	98%	30.81 TB	10%	334,280,328

Production status



IN from

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	562448	CERN-CREAM	940.23	1	0.18	1
2.	562443	CERN_LSF	806.47	4	1.34	1
3.	569063	CSC	587.58	5	38.83	1
4.	566971	DCSC_KU	544.19			1
5.	561601	CNAF_glexec	319.95	8	20	1
6.	561085	FZK_CREAM	284.70			1
7.	567177	CNAF-CREAM	117.68	8	10	1

OUT to

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
1.	562475	SARA	947.30			1
2.	562462	CERN-CREAM	946.22	1	0.20	1
3.	568258	GRIF_IPNO	934.24			1
4.	567276	CNAF-CREAM	900.04	8	11.31	1
5.	564781	NIKHEF	835.28	15	16.81	1
6.	561287	CERN_LSF	692.56	4	1.64	1
7.	567943	Kosice	673.30			1

Bandwidth per site

MonALISA

- ALICE Grid computing infrastructure is monitored through MonALISA
  - Services status (central and site VO-boxes)
  - Jobs (status, resources consumptions)
  - SEs (status, occupancy)
  - RAW data transfers (status, rates)
  - WLCG services (through SAM/Nagios)
  - Production status and control
  - Network
  - Access to administrative tasks
  - …
- More than a monitoring infrastructure
  - Alarms, services restarts and actions

SAM/NAGIOS

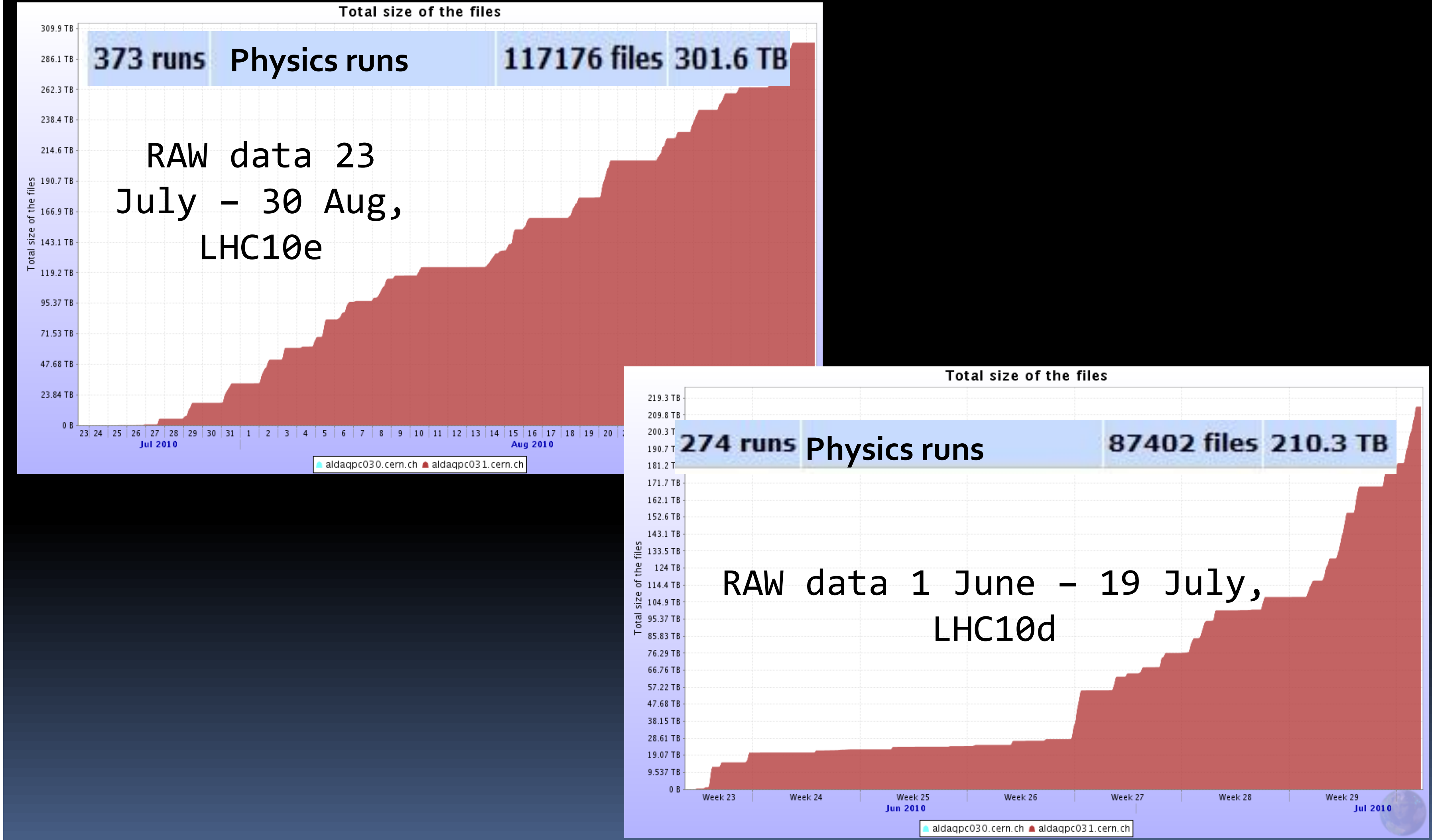
- WLCG generic services monitoring
  - Critical for site availability/reliability reports
  - The results are published in MonALISA (in addition to the standard portal)
- The monitoring includes:
  - Monitoring of the gLite3.2 VOBOXES
    - Specific ALICE test suite created for SAM and now migrated to Nagios
  - Monitoring of CEs
    - Standard "ops" test suite
  - Monitoring fully migrated to Nagios (production status)

Experiment Dashboard

- Currently used for the monitoring of the raw data transfers
  - Complementary information (in addition to MonALISA rates)
  - In some cases the error messages are cryptic – to be improved
- It can also be used to monitor the status of the ALICE real jobs (not the JobAgents)
  - Jobs need to be instrumented with message infrastructure to talk to Dashboard

- STATUS OF THE OPERATIONS

Status of the operations

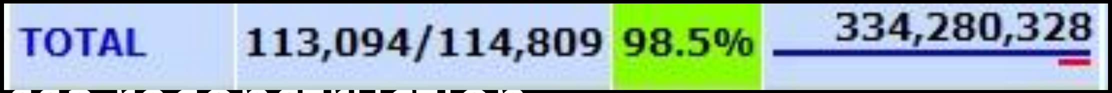


RAW data 23 July – 30 Aug, LHC10e

RAW data 1 June – 19 July, LHC10d

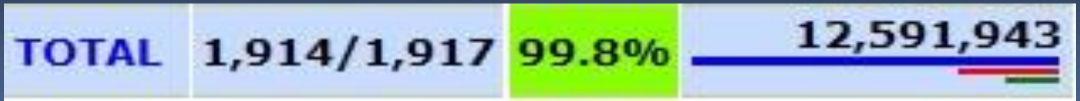
Pass1 and Pass2 Processes

- Pass 1 reconstruction
  - Quasi-online, follows the registration of RAW in CASTOR@CERN
  - The data is reconstructed fully ~24 hours after data taking
  - Typical reconstruction efficiency



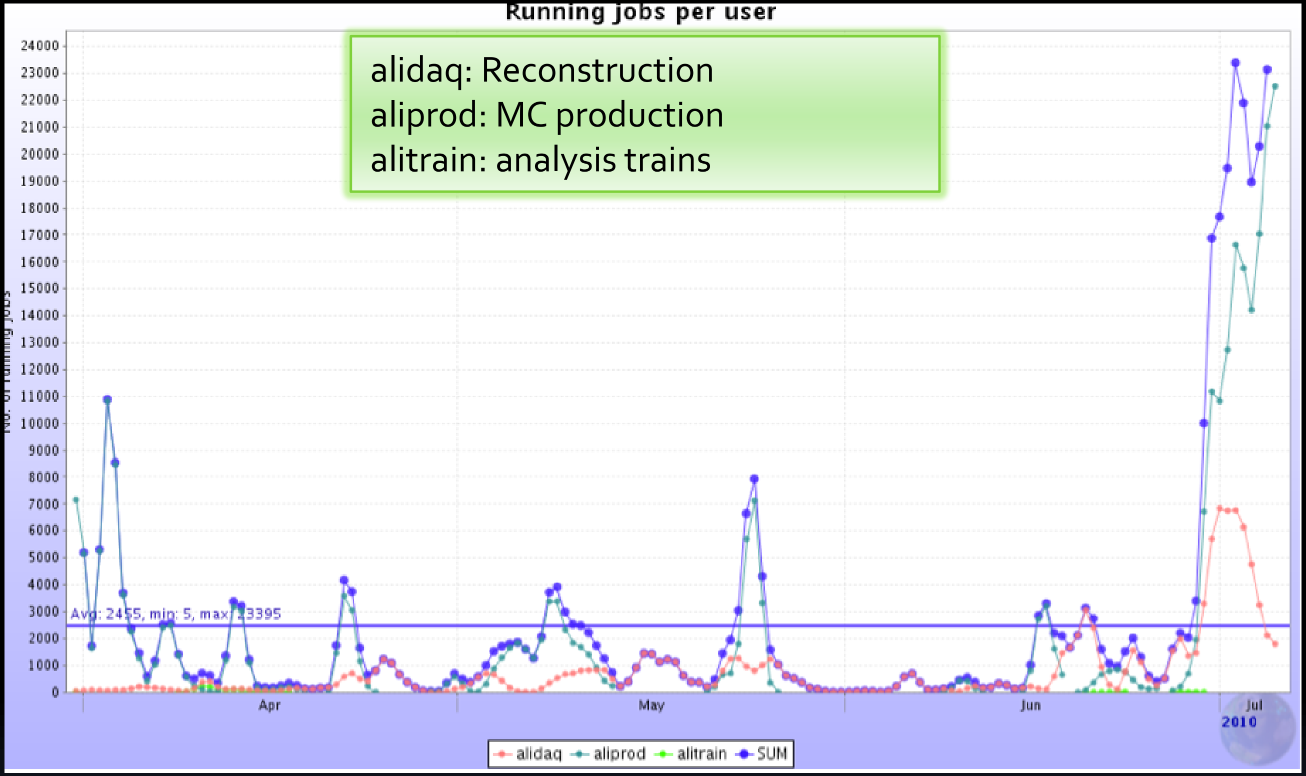
TOTAL	113,094/114,809	98.5%	334,280,328

- Pass2 reconstruction
  - ~1 month after data taking @T1s
  - Updated software, updated conditions
    - Improved detectors calibration from Pass1 reconstruction ESDs (calibration trains)
  - Typical reconstruction efficiency

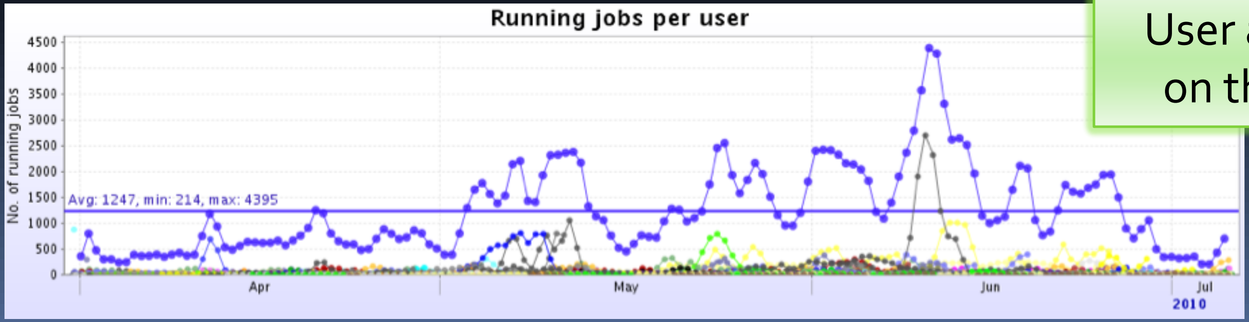


TOTAL	1,914/1,917	99.8%	12,591,943

Job profiles per users

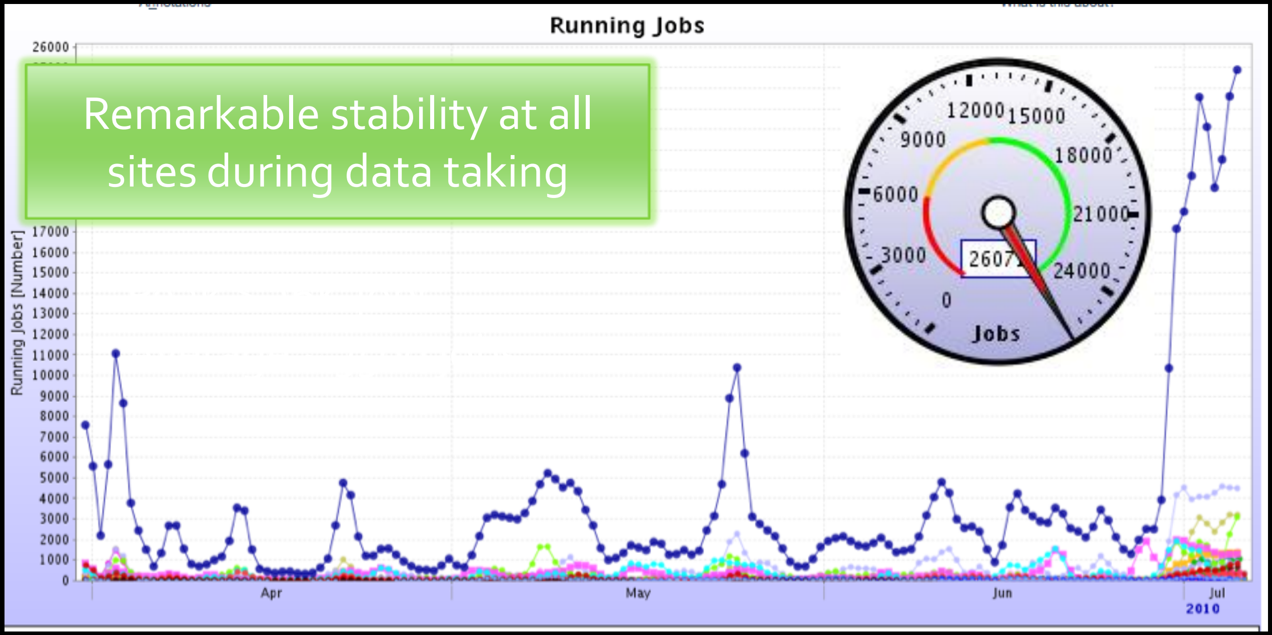


alidaq: Reconstruction
aliprod: MC production
alitrain: analysis trains

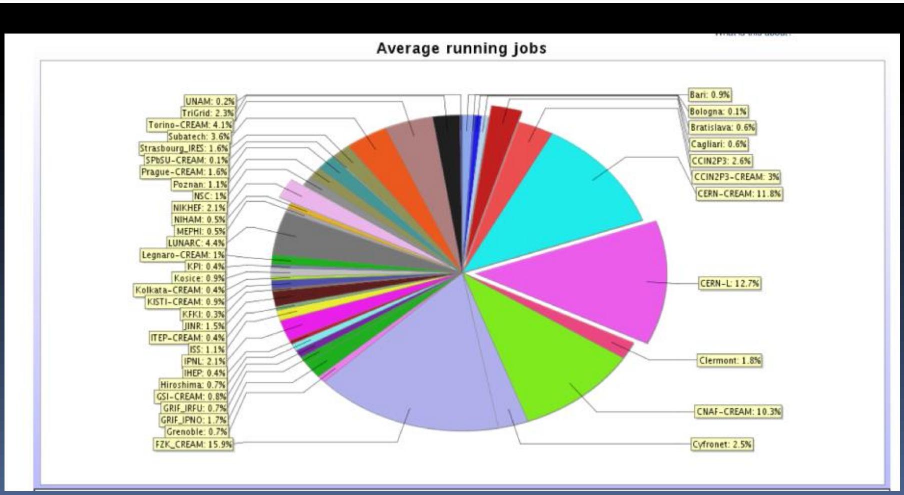


User analysis on the Grid

Job profiles per site



Remarkable stability at all sites during data taking



More than 50% of the work in ALICE is done by T2

Not really clear what is the difference between T1 and T2 apart from data custodial and better network

Summary

- Two words…
  - AliEn & MonALISA
- Three mayor services…
  - VOBOX, CREAM, xrootd
- Your challenge
  - Bandwidth
- Support and expertise ensured by the ALICE core team
- ALICE always welcome new T2 sites in production
- Current Status
  - Grid operation is now fairly routine
  - This has not been easy. This has been a long time effort