Status and Prospects of The LHC Experiments Computing

computing models, computing commissioning and its practical problems

CHEP, Prague Kors Bos, NIKHEF&CERN March 23, 2009

This Talk

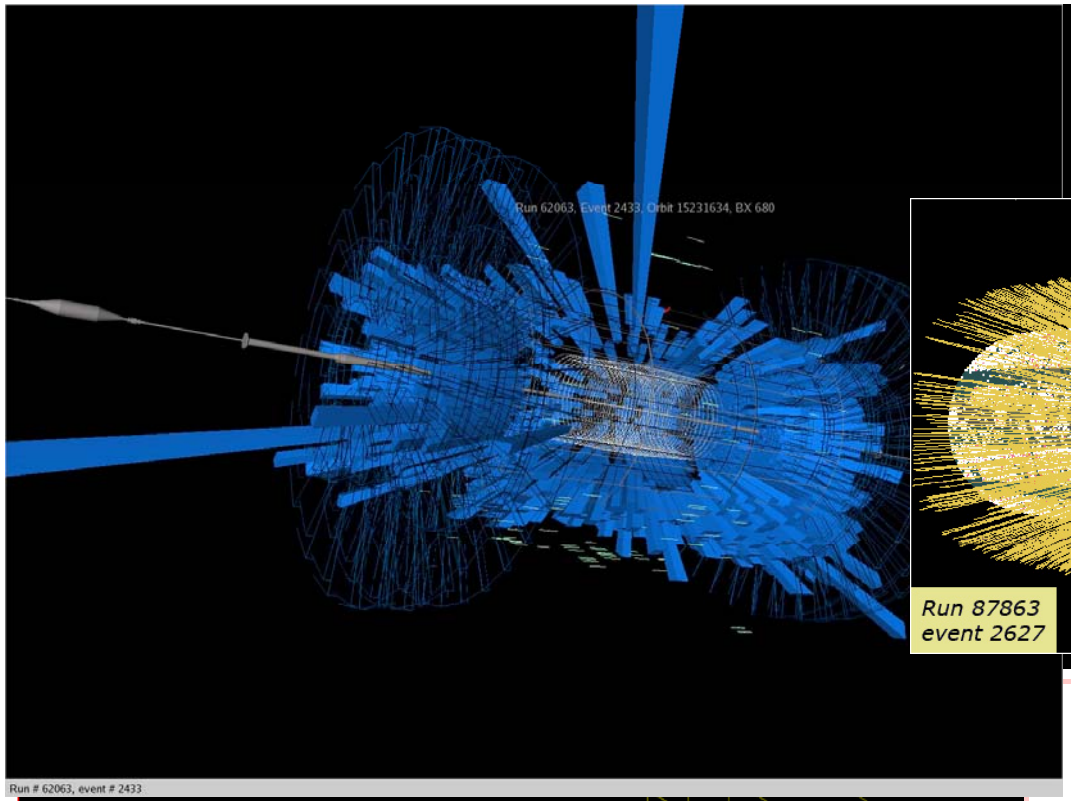
- **Disclaimer 1**: The title that Milos gave me cannot be done in 20 minutes and maybe not even in 20 hours. A good fraction of this whole conference is about this. So this merely will be an introduction

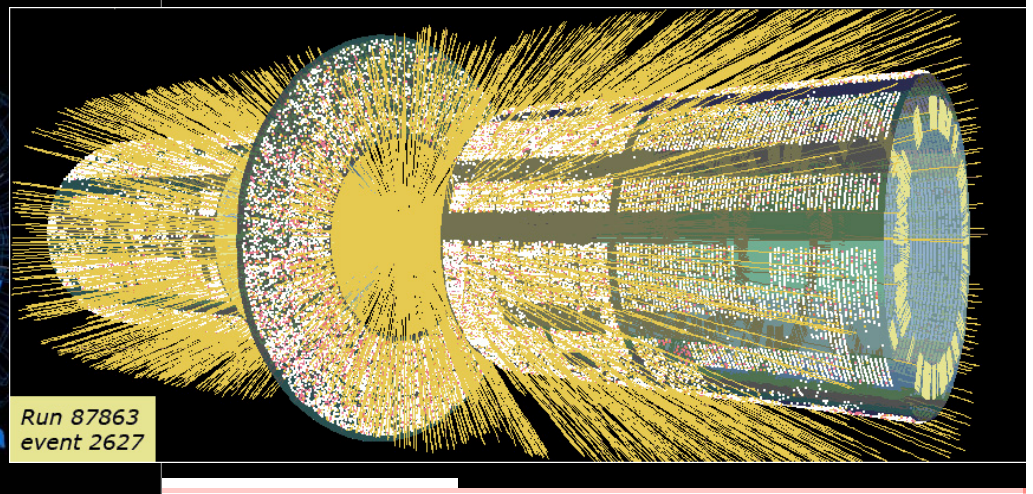
- **Disclaimer 2**: I try to talk about all 4 LHC experiments but I am obviously biased towards one …

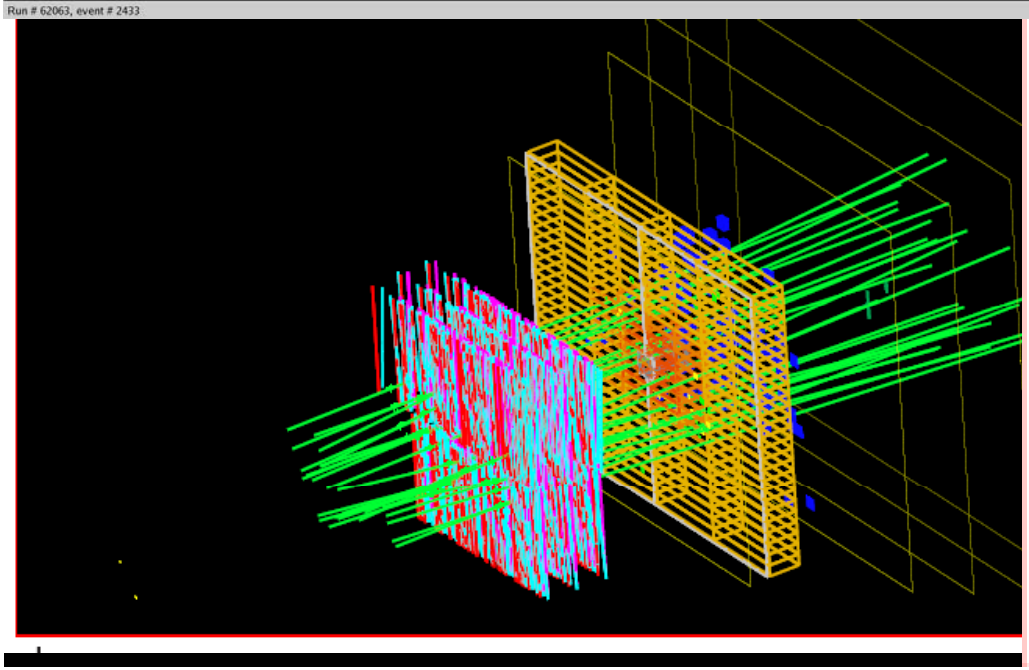
- **Disclaimer 3**: I may get things not completely right when talking about other VO's than my own and I apologize beforehand and refer to all specialized talks at this conference

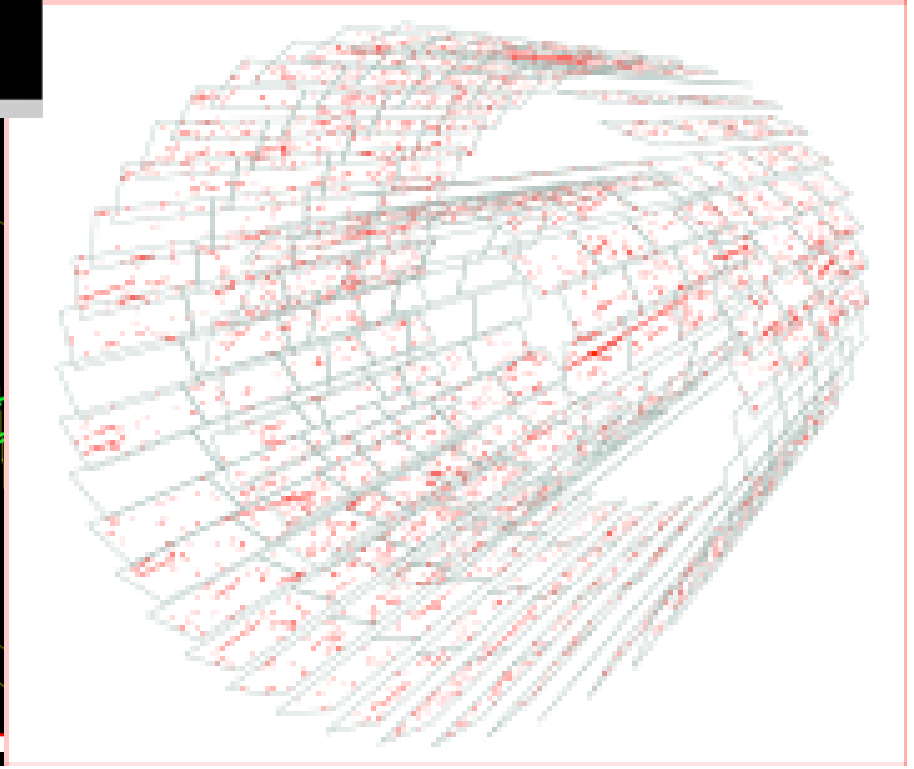
- **Disclaimer 4**: I can not guarantee that I will explain all acronyms, but I will try

First events





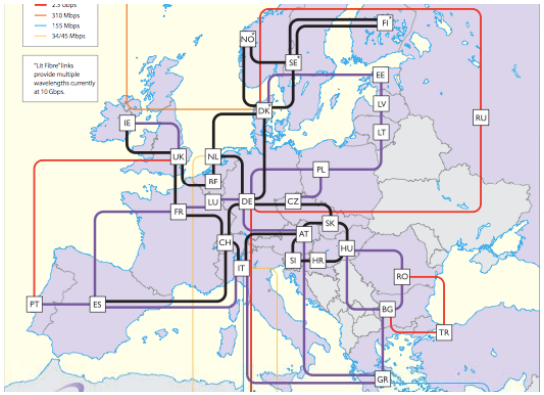




Status and Prospects of The LHC Experiments Computing

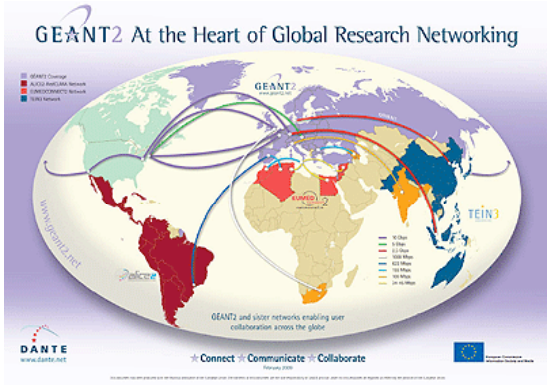
computing models, computing commissioning and its practical problems

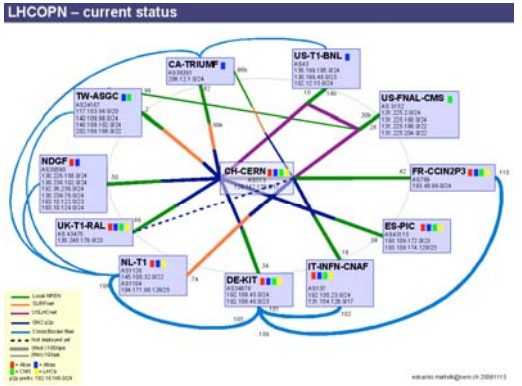
CHEP, Prague Kors Bos, NIKHEF&CERN March 23, 2009



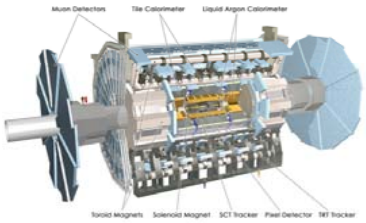
Ubiquitous Wide Area Network Bandwidth

- First Computing TDR's assumed not enough network bandwidth
- The Monarch project proposed multi Tier model with this in mind
- Today network bandwidth is our least problem
- But we still have the Tier model in the LHC experiments
- Not in all parts of the world ideal network yet (last mile)
- LHCOPN provides excellent backbone for Tier-0 and Tier-1's
- Each LHC experiment has adopted differently





ATLAS Workflows



Prompt Reconstruction

Calibration & Alignment
Express Stream Analysis

Tier-0

CAF

CASTO
R

650 MB/sec

50-500 MB/sec

RAW Re-processing
HITS Reconstruction

Tier-1

Tier-1

Tier-1

50-500 MB/sec

Simulation
Analysis

Tier-2

Tier-2

Tier-2

CMS Workflows

Prompt Reconstruction

TIER-0

CASTOR

CAF

Calibration
Express-Strea
Analysis

600MB/s

50-500MB/s

TIER-1

TIER-1

TIER-1

Re-Reco
Skims

50-500MB/s

~20MB/s

Simulation
Analysis

TIER-2

TIER-2

TIER-2

TIER-2

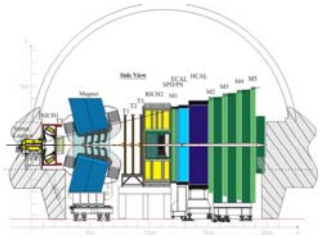
Similarities & Differences
CMS vs ATLAS

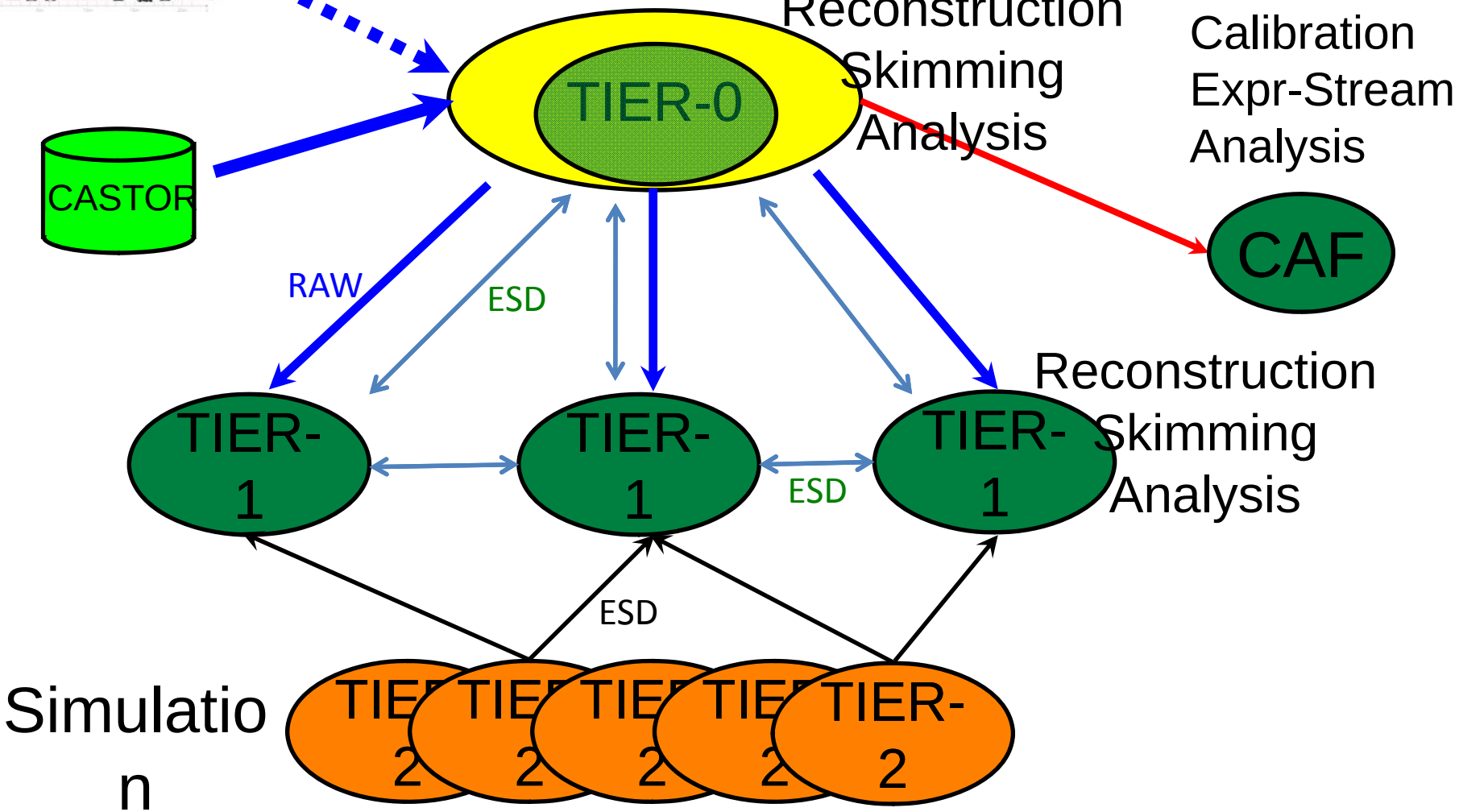
- Tier-0 and CAF very much the same functionality
- Rates are quite similar
- Functionality of Tier-1's much the same: re-reconstruction
- Functionality of Tier-2's much the same: Simulation and analysis

- CMS: analysis jobs in Tier-2's can get data from any Tier-1
- ATLAS: analysis jobs in Tier-2's can get data only from Tier-1 within the same cloud

- CMS: analysis coordinated per Tier-2
- ATLAS: coordinated per physics group and/or cloud

LHCb Workflows





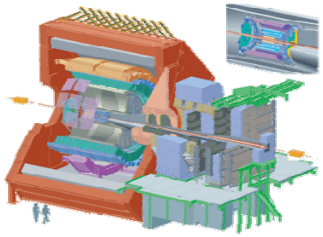
Similarities & Differences
CMS & ATLAS vs LHCb

- CAF very much the same functionality
- Rates are much higher but data volume much smaller
- Different functionality of Tier-1: reconstruction, skimming and analysis
- The Tier-0 acts as another Tier-1: reconstruction, skimming and analysis

- The Tier-2's do only simulation (+digitization +reconstruction) production
- Output from simulation (DST) can be uploaded to any Tier-1
- No cloud concept

- RAW and RDST (output from reconstruction) go to tape in Tier-0/1
- DST (output from skimming) goes to all Tier-0/1's on disk

ALICE Workflows



Prompt Reconstruction

Calibration & Alignment
Express Stream Analysis

Tier-0

CAF

CASTOR

Storage hypervisor – xrootd global redirector

RAW Re-processing
Simulation, analysis
(if free resources)

T1 AF

Tier-1

Tier-1

Tier-1

Simulation
Analysis

Tier-2

Tier-2

Tier-2

Tier-2

Tier-2

T2 AF

Similarities & Differences
CMS vs ATLAS vs ALICE

- Tier-0 and CAF very much the same functionality
- Functionality of Tier-1's much the same: re-reconstruction
  - If resources available, T1s can do MC and analysis (ALICE job queue prioritization)
- Functionality of Tier-2's much the same: Simulation and analysis

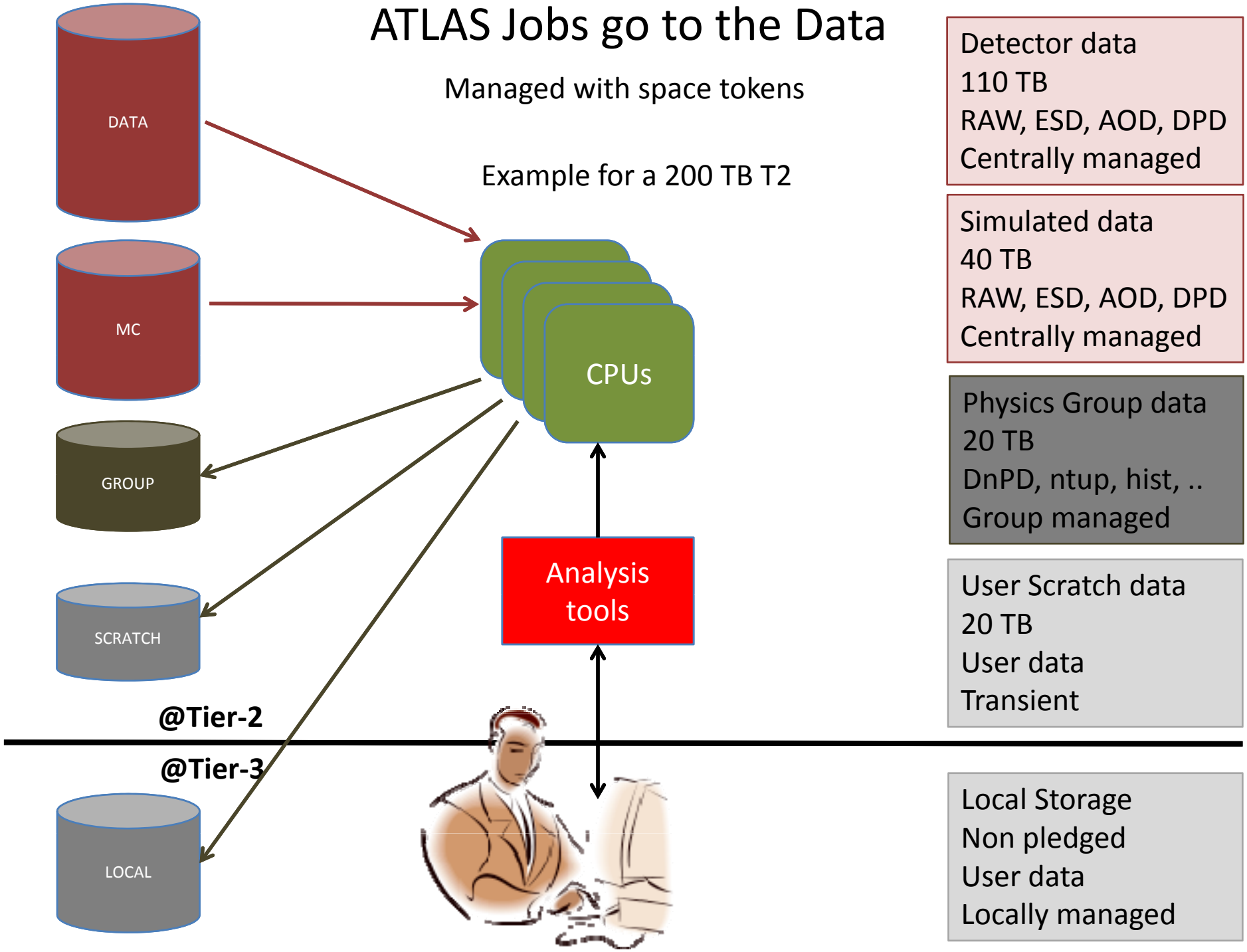
- ALICE: analysis jobs are allowed to 'pull' data from any storage in case of local data not found (Grid catalogue-SE discrepancy)
  - Through xrootd global redirector (SE collaboration on Grid scale)
  - Network is ubiquitous, limited 'ad hoc' data transfers do not pose a problem
  - Allow the job to complete and fix the discrepancy afterwards

- ESDs/AODs can be stored at any T1/T2 depending on the resources availability, there is no 'targeted, per data or physics type' data placement

ATLAS Jobs go to the Data

Managed with space tokens

Example for a 200 TB T2



Detector data
110 TB
RAW, ESD, AOD, DPD
Centrally managed

Simulated data
40 TB
RAW, ESD, AOD, DPD
Centrally managed

Physics Group data
20 TB
DnPD, ntup, hist, ..
Group managed

User Scratch data
20 TB
User data
Transient

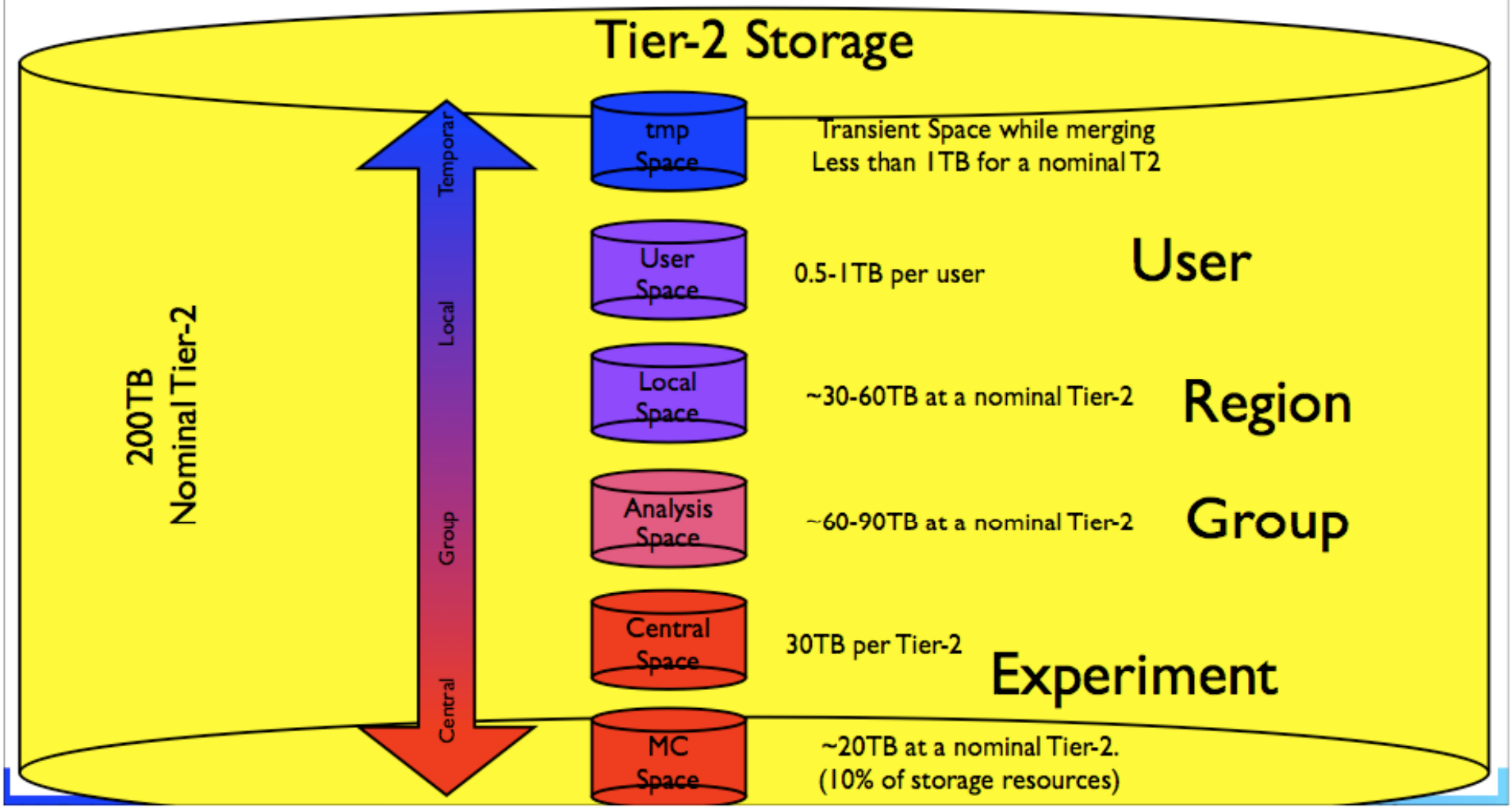
Local Storage
Non pledged
User data
Locally managed

In CMS Jobs go to Data

How is the Storage managed? (without space tokens)

Storage at Tier-2 centers is broken into 6 pieces

- Transient and unmanaged to more persistent and centrally managed



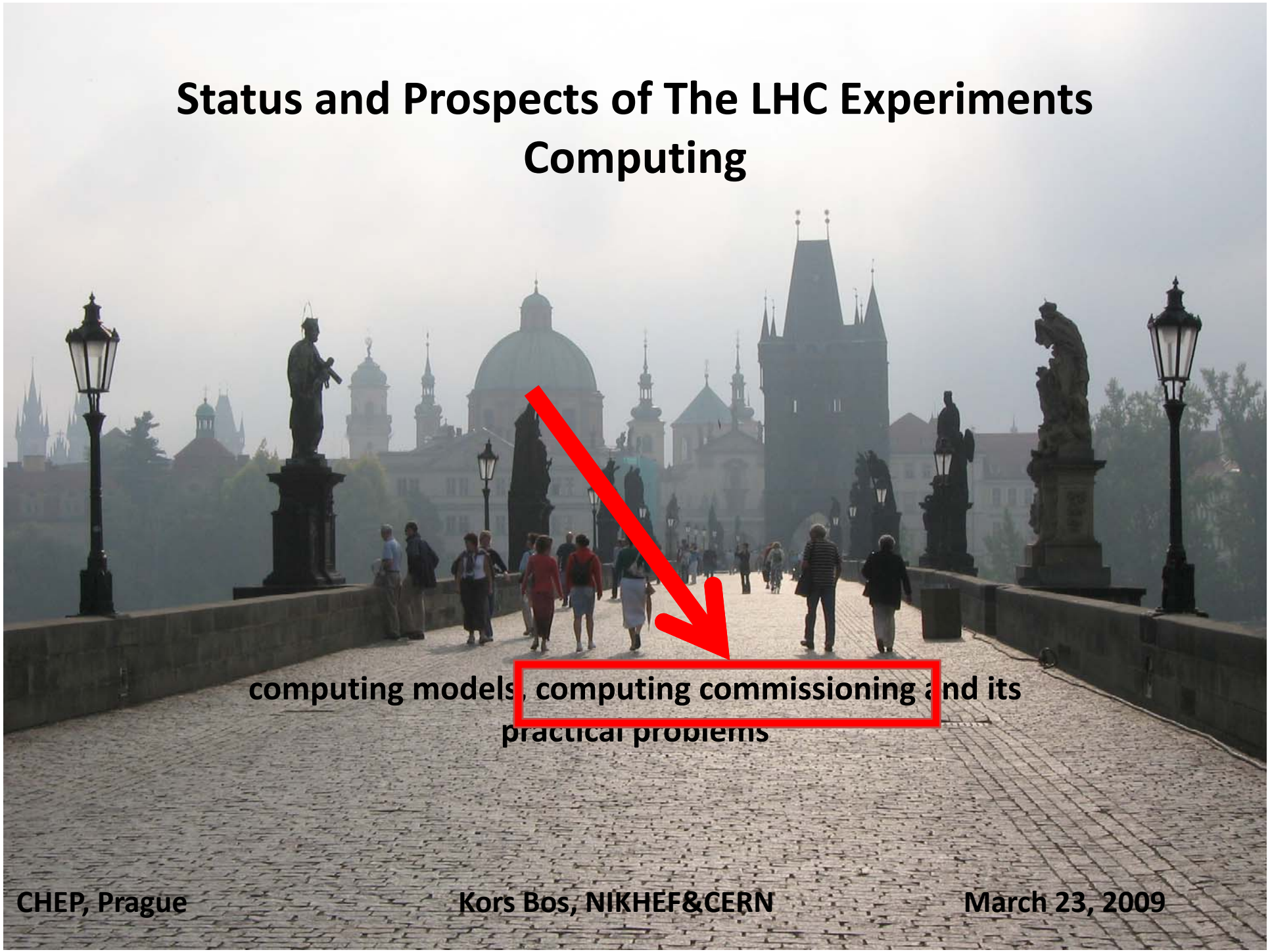
- Analysis is done in the place (Tier-0 and Tier-1's) where the already data is

- *LHCb uses 6 space tokens*

- Jobs go to the data
  But…
- Data can also go to the jobs depending on where the free resources are

- *Alice doesn't use space tokens at all*

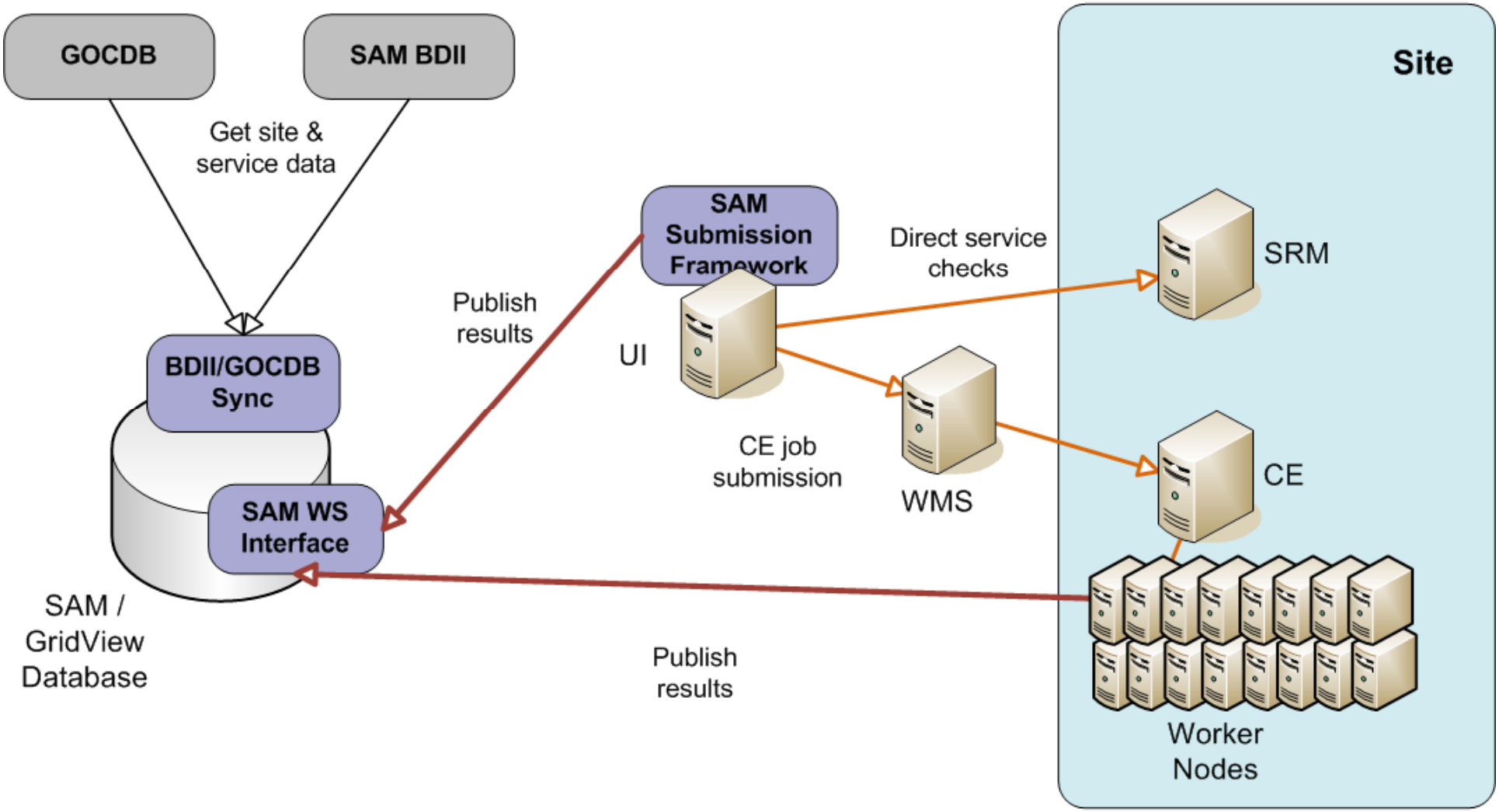
Status and Prospects of The LHC Experiments Computing



computing models, computing commissioning and its practical problems

CHEP, Prague Kors Bos, NIKHEF&CERN March 23, 2009

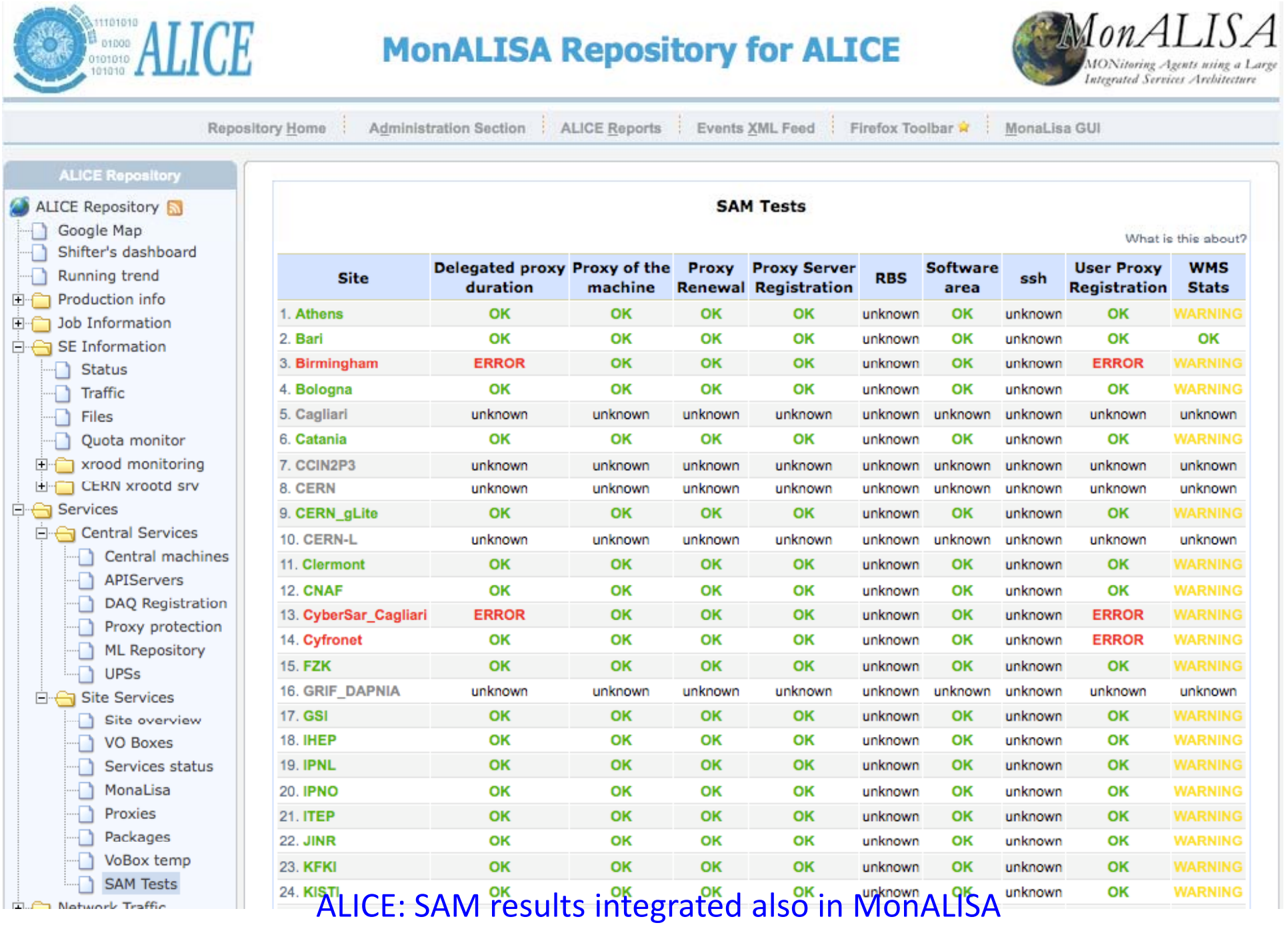
Service Availability Monitoring (SAM)



How SAM works

Sitename	Service Type	Service Name	js	swdir	PR	PSR	PM	SA	UPR
		ce129.cern.ch	ok	ok					
	VOBOX	voalice03.cern.ch			ok	ok	ok	ok	ok
		voalice06.cern.ch			error	ok	ok	ok	ok
FZK-LCG2	CE	ce-1-fzk.gridka.de	ok	ok					
		ce-2-fzk.gridka.de	ok	ok					
		ce-3-fzk.gridka.de	ok	ok					
		ce-4-fzk.gridka.de	ok	ok					
		ce-5-fzk.gridka.de	ok	ok					
	VOBOX	alice-fzk.gridka.de			ok	ok	ok	ok	ok
INFN-T1	CE	ce04-lcg.cr.cnaf.infn.it	ok	ok					
		ce05-lcg.cr.cnaf.infn.it	ok	ok					
		ce06-lcg.cr.cnaf.infn.it	ok	ok					
NIKHEF-ELPROD	CE	gazon.nikhef.nl	ok	ok					
		trekker.nikhef.nl	ok	ok					
	VOBOX	erf.nikhef.nl			ok	ok	ok	ok	ok
RAL-LCG2	CE	lcgce02.gridpp.rl.ac.uk	ok	ok					
	VOBOX	lcgvo0597.gridpp.rl.ac.uk			ok	ok	ok	ok	ok

ALICE latest results (VOBOXes and CE)

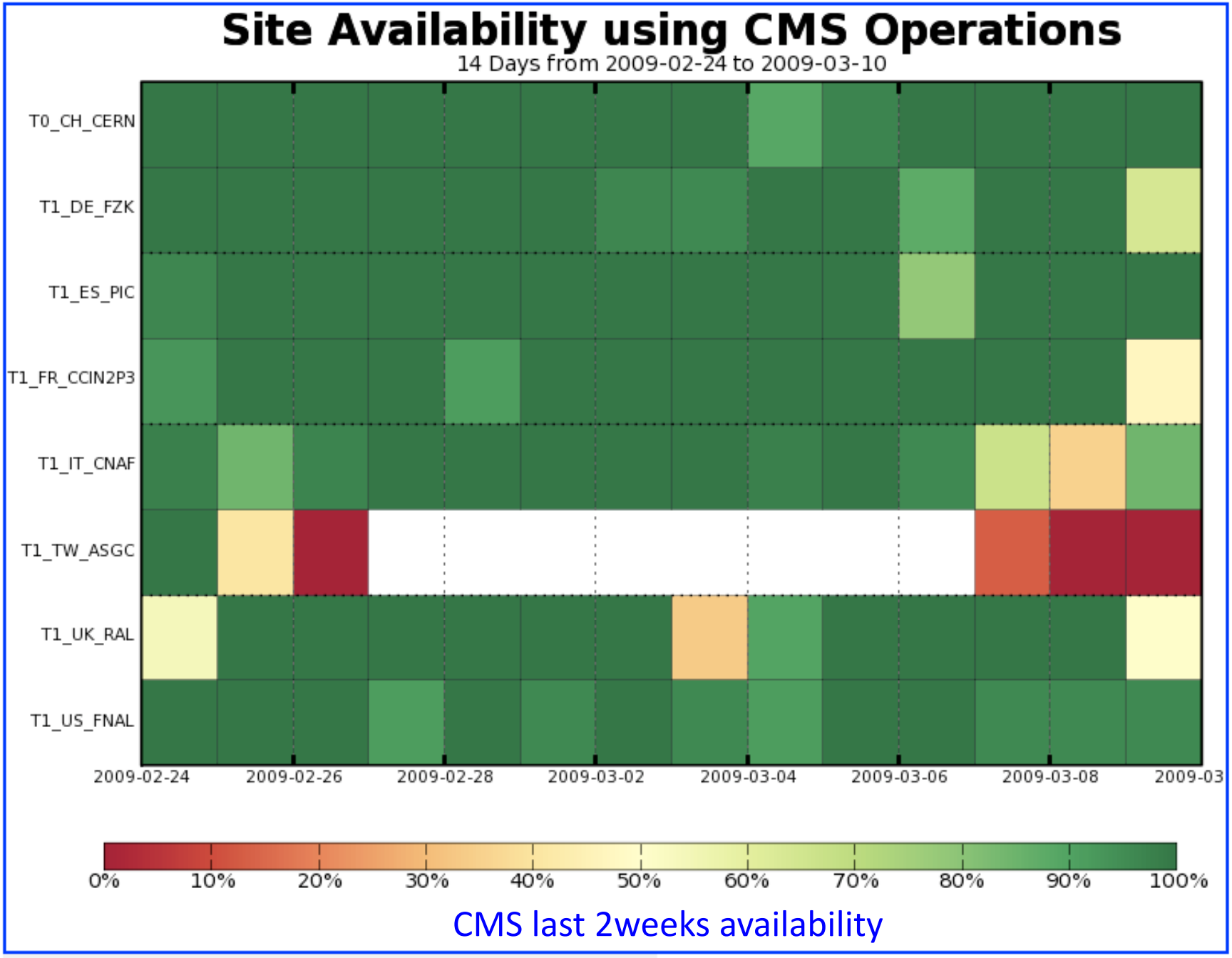


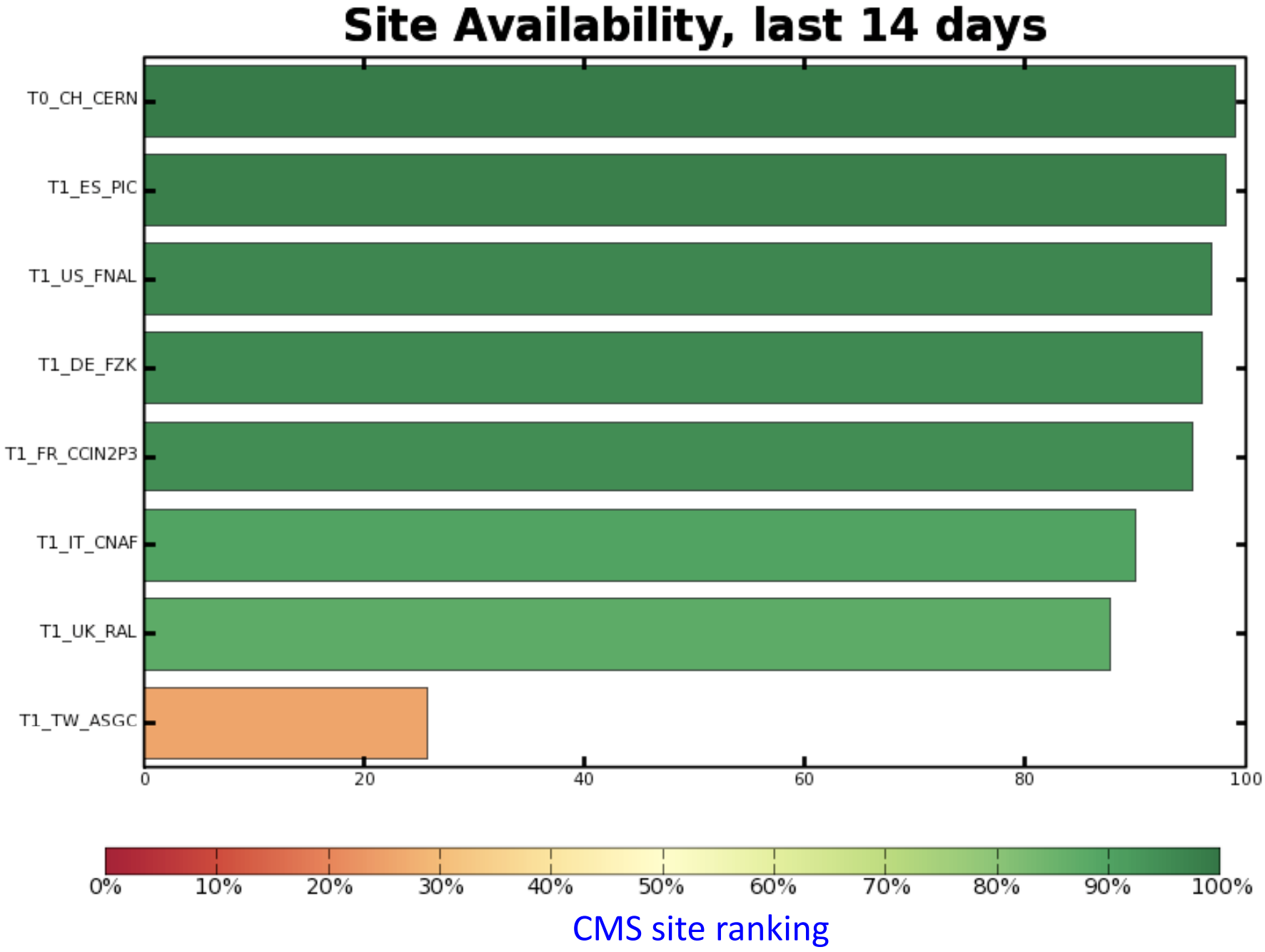
ALICE: SAM results integrated also in MonALISA

Sitename	Service Type	Service Name	LHCb CE-lhcb-job-Boole	LHCb CE-lhcb-job-Brunel	LHCb CE-lhcb-job-DaVinci	LHCb CE-lhcb-job-Gauss	LHCb CE-lhcb-os	LHCb CE-lhcb-queues	csh	js	swdir	DiracTestRAW	DiracTestUSER	FileAccessV2
LCG.GRIDKA.de	CE	ce-1-fzk.gridka.de	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce-2-fzk.gridka.de	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce-3-fzk.gridka.de	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce-4-fzk.gridka.de	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce-5-fzk.gridka.de	ok	ok	ok	ok	ok	ok	ok	ok	ok			
	SRMv2	gridka-dCache.fzk.de										ok	ok	ok
LCG.NIKHEF.nl	CE	ce.gina.sara.nl	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		celisa.grid.sara.nl								error				
		gazon.nikhef.nl	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		trekker.nikhef.nl	ok	ok	ok	ok	ok	ok	ok	ok	ok			
	SRMv2	srm.grid.sara.nl										ok	ok	ok
LCG.PIC.es	CE	ce05.pic.es	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce06.pic.es	ok	ok	ok	ok	ok	ok	ok	ok	ok			
		ce07.pic.es	ok	ok	ok	ok	ok	ok	ok	ok	ok			
	SRMv2	srmlhcb.pic.es										ok	ok	ok
LCG.RAL.uk	CE	lcgce04.gridpp.rl.ac.uk	ok	ok	ok	ok	ok	info	ok	ok	ok			
		lcgce05.gridpp.rl.ac.uk	ok	ok	ok	ok	ok	info	ok	ok	ok			
	SRMv2	srm-lhcb.gridpp.rl.ac.uk										ok	ok	ok

LHCb latest results

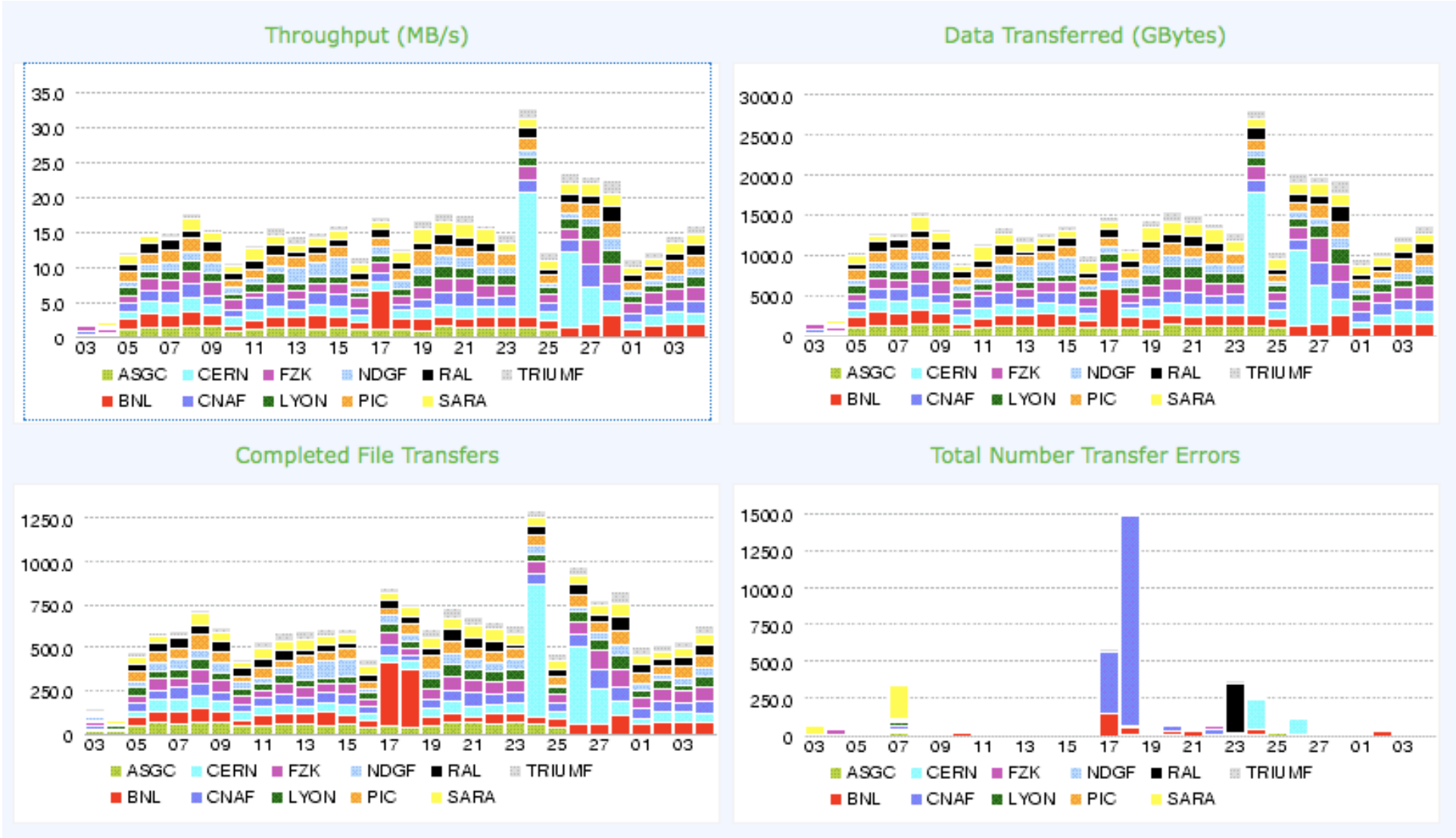
A snapshot similar to the CMS and LHCb one could be retrieved also to ALTAS


Site Availability using CMS Operations
14 Days from 2009-02-24 to 2009-03-10
T0_CH_CERN
T1_DE_FZK
T1_ES_PIC
T1_FR_CCIN2P3
T1_IT_CNAF
T1_TW_ASGC
T1_UK_RAL
T1_US_FNAL
2009-02-24
2009-02-26
2009-02-28
2009-03-02
2009-03-04
2009-03-06
2009-03-08
2009-03
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
CMS last 2weeks availability

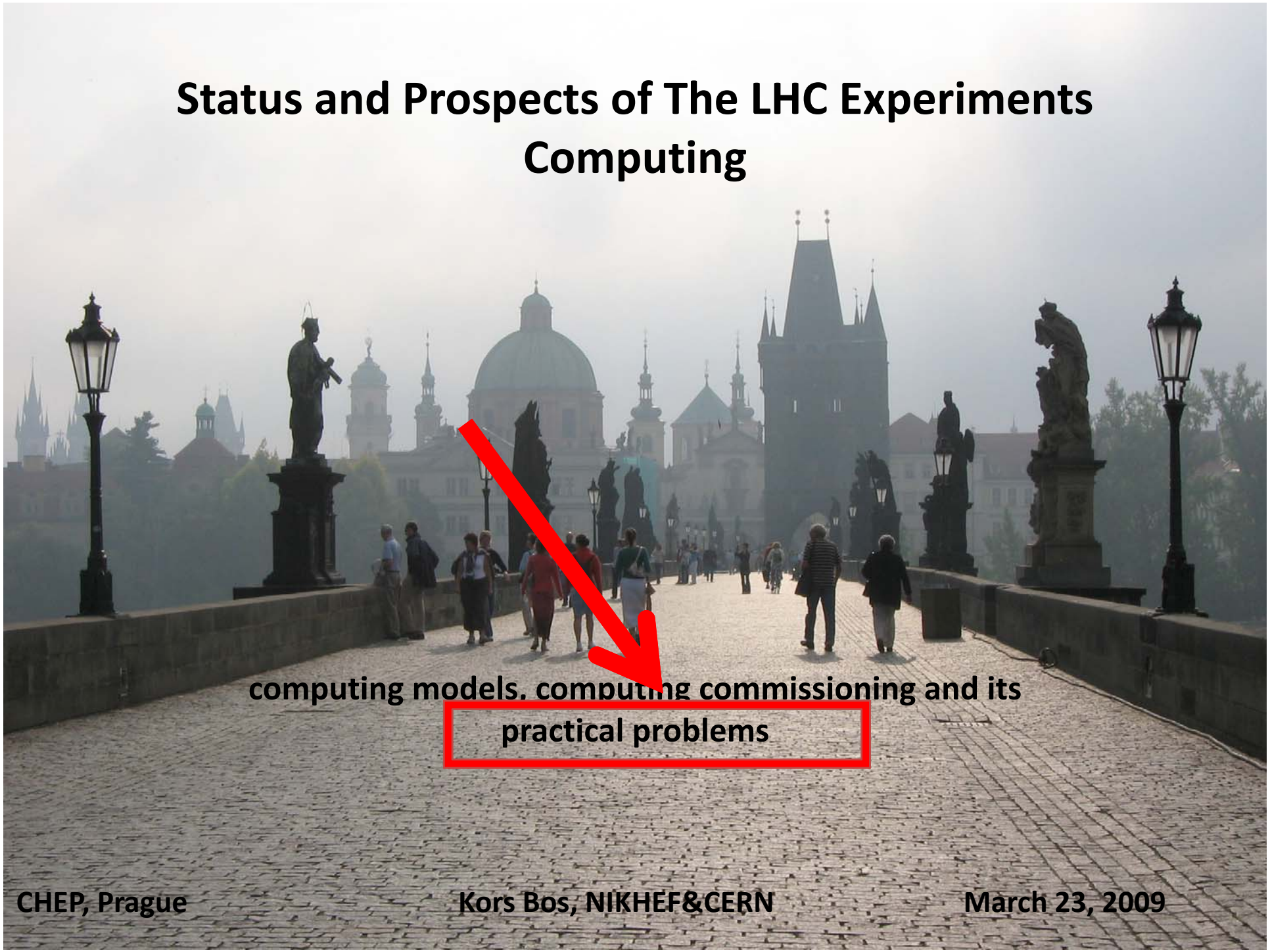


CMS site ranking

Functional tests in ATLAS



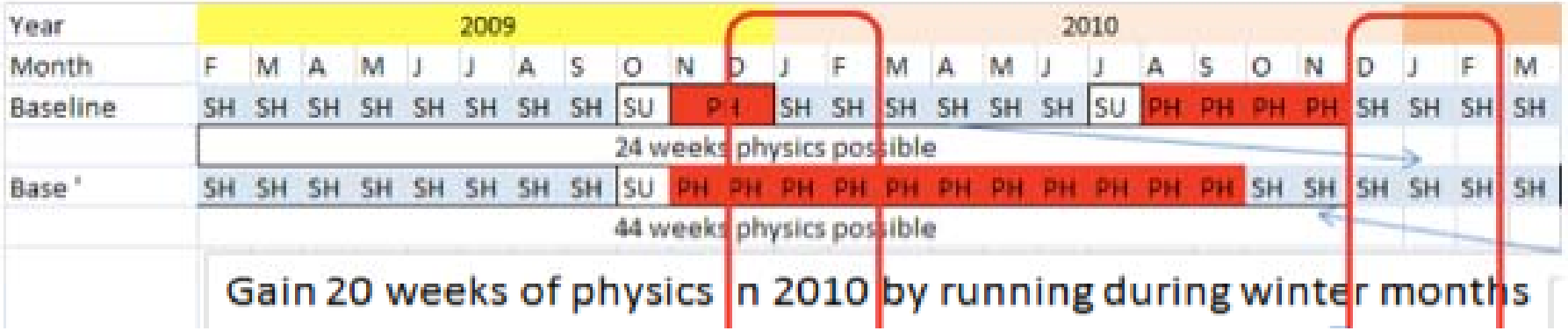
Status and Prospects of The LHC Experiments Computing



computing models, computing commissioning and its practical problems

CHEP, Prague Kors Bos, NIKHEF&CERN March 23, 2009

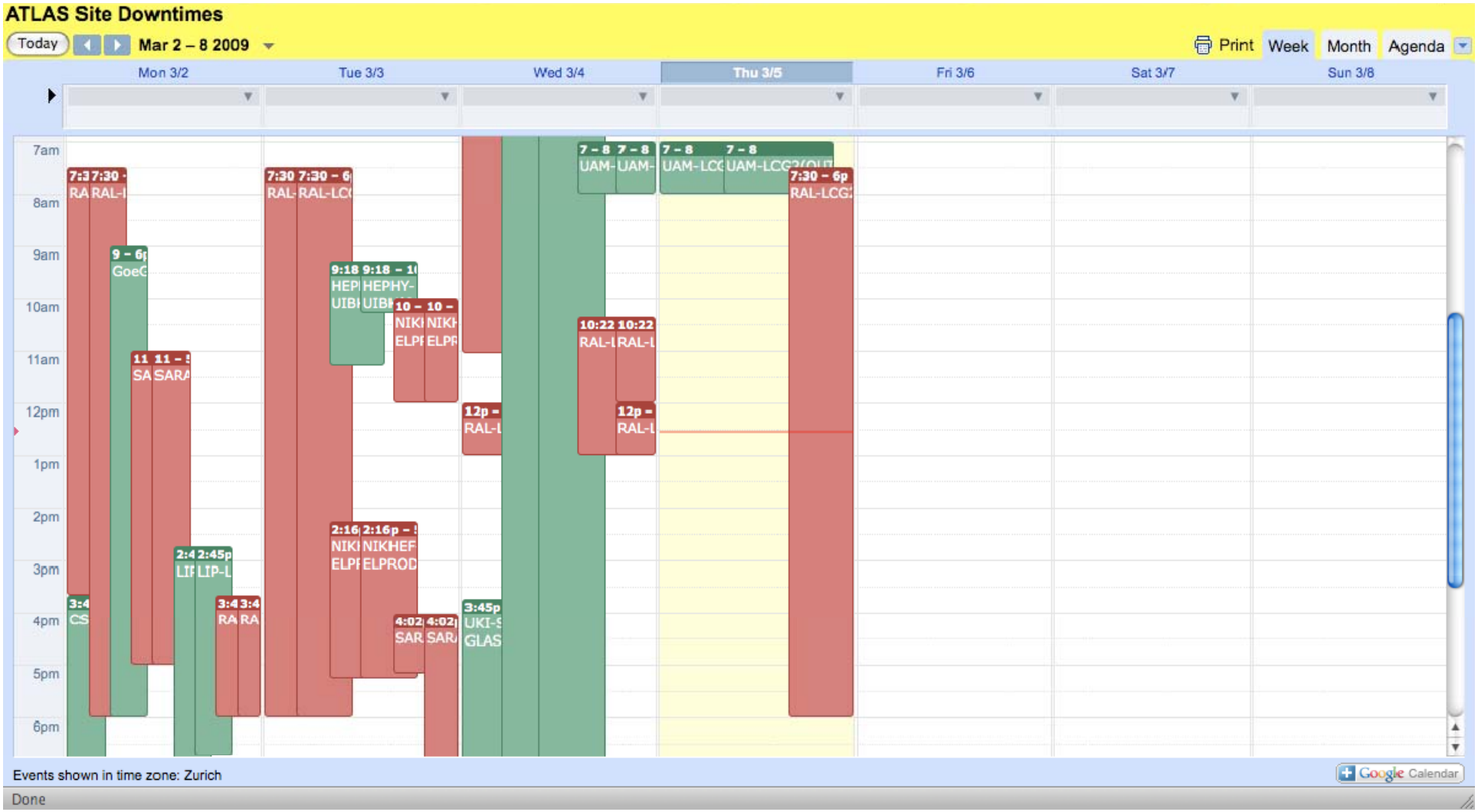
Practical Problem 1: Big Step at once



- A run of ~1 year without interruption
  - Without having had a chance to test in a short period
  - Without having ran all services of all 4 VO's at the same time
- Do we have the bandwidth everywhere ?
- Do we have the people to run all shifts ?
- Have sites appreciated what it means ?
  - Only very short (max 1 day) scheduled downtimes
  - ..

Scheduled down times of the sites

we better be prepared that not all sites are always up ..



Practical Problem 2 : Tapes

but calculable

- ATLAS writes RAW data and G4 HITS to tape and ESD from re-processing
  - ATLAS read RAW back from tape for re-processing and HITS for (re-)reconstruction
- CMS writes RAW data to tape
  - CMS reads RAW data back fro re-processing
- LHCb writes RAW data to tape and RDST from reconstruction
  - LHCb reads RAW data back for re-processing
- Alice writes RAW data to tape as well as ESD and AOD
  - And reads RAW back for re-processing

- All these processes have been tested individually
  - *But not all together !*
- A Tier-1 supporting all 4 experiments needs to worry about
  - Tape families
  - Number of tape drives
  - Bandwidth to/from tape
  - Buffer sizes

- Probably one of the biggest unknown for the next run
  - Very hard to plan & test beforehand

Practical Problem 3 : Users

2009-2010 Run

the calculable and non-calculable

- Data acquisition will work and also the data distribution
- Calibration and alignment will work and also the reconstruction in the sites
- Monte Carlo Simulation production will work

- Tape writing will work … scales with the hardware available
- Tape reading may be trickier … hard to do it all efficiently

- CPU's will work … but there will never be enough
- Bandwidth to the data may become an issue

- Users will be the big unknown … and yet it is the most important
- Only this will validate or falsify the computing models

We will know better in Taipei !