The South African National Compute Grid : What it means for SA-CERN

03/18/08

Bruce Becker, UCT-CERN Research Centre

UCT CERN

Outline

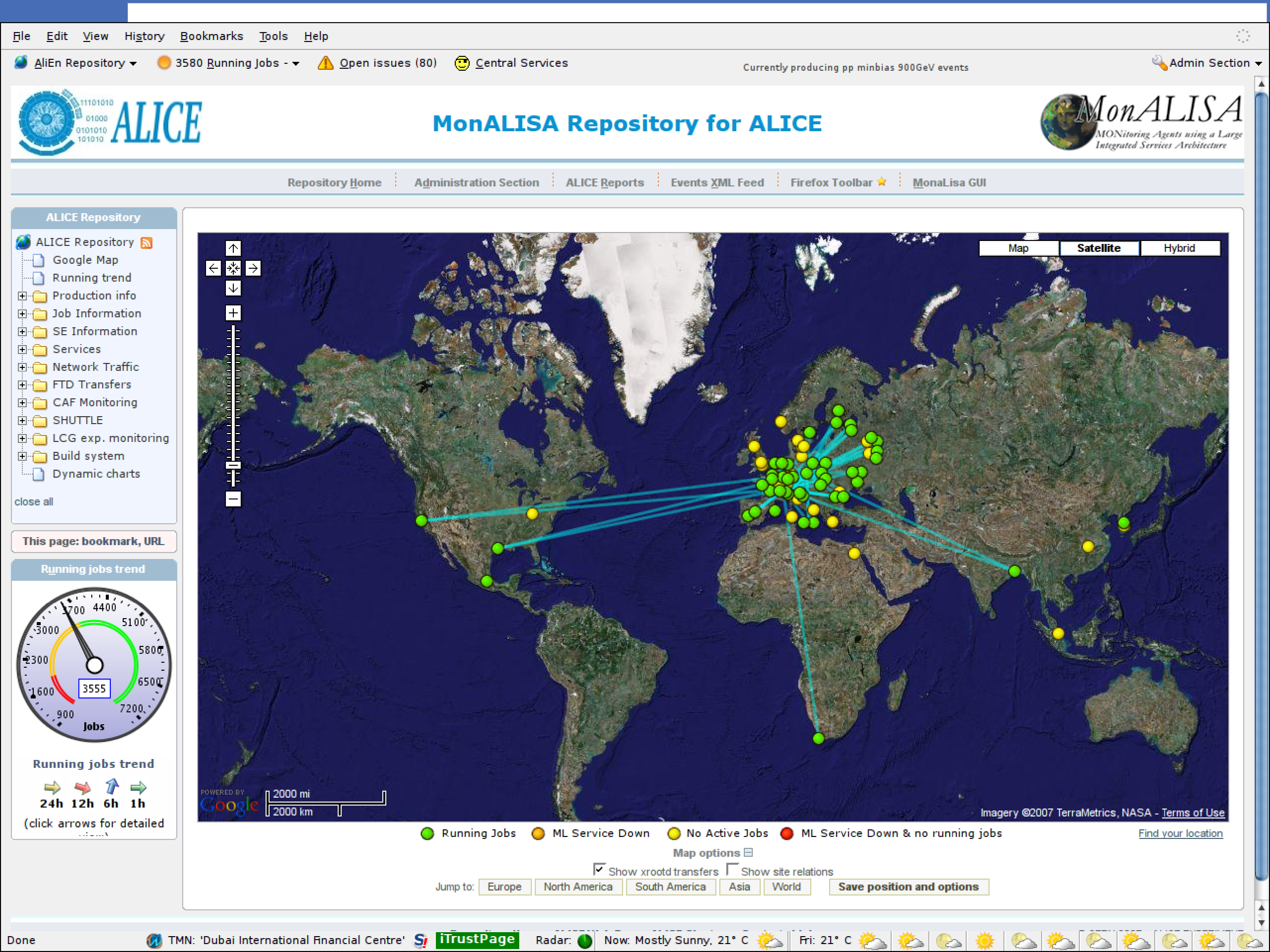
- Introduction to SA National Compute Grid and related projects
- ALICE experiment, CERN, and UCT-CERN Research Centre
- Existing sites
- How to make use of the grid in SA

The state of play in South Africa : December 2007

- Decision to build the CHPC made serious HPC a priority; several disciplines relying more and more on HPC, domains several different applications
- But
  - still concentrated in one place
  - almost exclusively for "flagship" projects.
  - Not "open" for smaller groups, which still relied on setting up departmental and group-level clusters, on various campuses
- Missing :
  - No security infrastructure
  - No proper collaboration model in SA (but several ad-hoc collaborations) and internationally.
  - No enabler for smaller groups
  - No model for fast-turnaround computing
  - No model for industrial/commercial/academic collaboration in computing
- All of which and more can be provided by a well-funded and widely supported national grid initiative

What was happening in the meantime

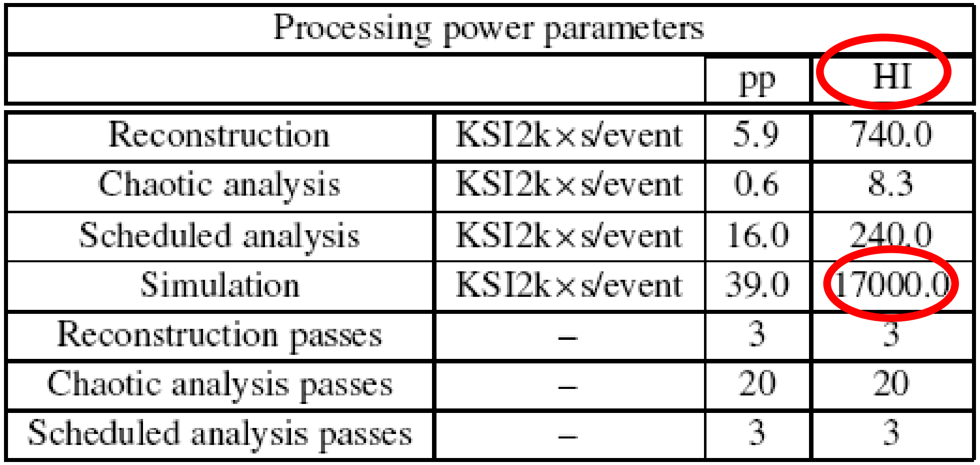
- “Gigantic” transnational infrastructure may have started at CERN, but it's now grown far beyond that.
  - LHC experiments still take up a very large part of computing requirements provided by grid in EU/US/Japan
  - Data challenges prove that the infrastructure can work, in the hierarchical model of the experiments
  - LHC experiment Commissioning currently under way... first beam towards the end of the year, a very exciting time in physics, entirely impossible to do without grid computing.
- However, EU projects in several fields of science now rely on grid computing - Considered a national imperative in most EU countries.
- Several regional efforts (e.g. Nordugrid, CyberSar in Sardinia, and Consorzio COMETA, PI2S2 in Sicily...)
  - Not to mention ***EUIndiaGrid, EUChina, EELA, EUMedGrid... (all managed by INFN)***
- Middleware development
  - Ease of use
  - Standardised – user base covers most of the world
  - Extended functionality – supports mpich jobs natively, handles distributed licences of proprietary software, etc etc...



File
Edit
View
History
Bookmarks
Tools
Help
AliEn Repository
3580 Running Jobs -
Open issues (80)
Central Services
Currently producing pp minbias 900GeV events
Admin Section
ALICE
MonALISA Repository for ALICE
MonALISA
MONitoring Agents using a Large Integrated Services Architecture
Repository Home
Administration Section
ALICE Reports
Events XML Feed
Firefox Toolbar
MonaLisa GUI
ALICE Repository
ALICE Repository
Google Map
Running trend
Production info
Job Information
SE Information
Services
Network Traffic
FTD Transfers
CAF Monitoring
SHUTTLE
LCG exp. monitoring
Build system
Dynamic charts
close all
This page: bookmark, URL
Running jobs trend
3700
4400
5100
3000
5800
2300
6500
1600
3555
7200
900
Jobs
Running jobs trend
24h
12h
6h
1h
(click arrows for detailed
Map
Satellite
Hybrid
POWERED BY Google
2000 mi
2000 km
Imagery ©2007 TerraMetrics, NASA - Terms of Use
● Running Jobs
● ML Service Down
● No Active Jobs
● ML Service Down & no running jobs
Find your location
Map options
☑ Show xrootd transfers
□ Show site relations
Jump to:
Europe
North America
South America
Asia
World
Save position and options
Done
TMN: 'Dubai International Financial Centre'
iTrustPage
Radar:
Now: Mostly Sunny, 21° C
Fri: 21° C

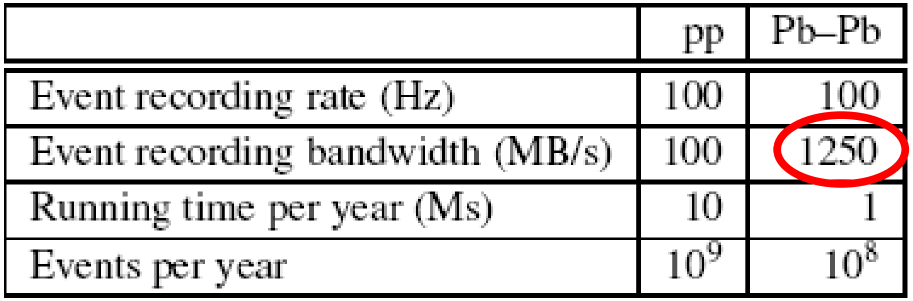
The ALICE data : Processing requirements and software framework

- The software environment for ALICE provides everything from the basic frameworks to specific user code for analysis of Data
  - all ROOT functionality
  - Simulation classes
  - Reconstruction classes
  - Raw data readers
  - Analysis frameworks
  - Monte-Carlo services
- http://aliceinfo.cern.ch/Offline



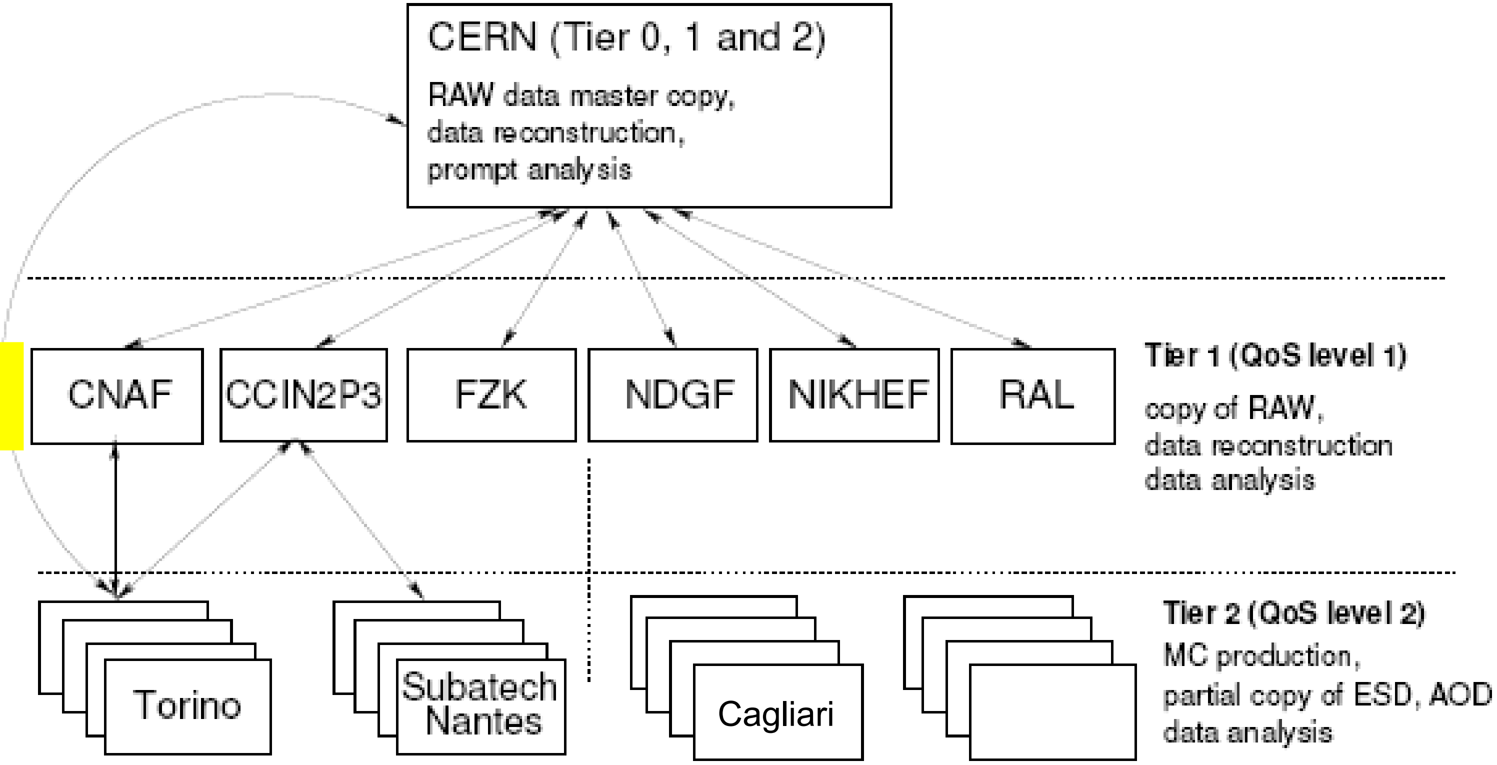
Processing power parameters

		pp	HI
Reconstruction	KSI2k×s/event	5.9	740.0
Chaotic analysis	KSI2k×s/event	0.6	8.3
Scheduled analysis	KSI2k×s/event	16.0	240.0
Simulation	KSI2k×s/event	39.0	17000.0
Reconstruction passes	–	3	3
Chaotic analysis passes	–	20	20
Scheduled analysis passes	–	3	3



	pp	Pb–Pb
Event recording rate (Hz)	100	100
Event recording bandwidth (MB/s)	100	1250
Running time per year (Ms)	10	1
Events per year	10^9	10^8

ALICE (and WLCG) Data Grid Tier structure



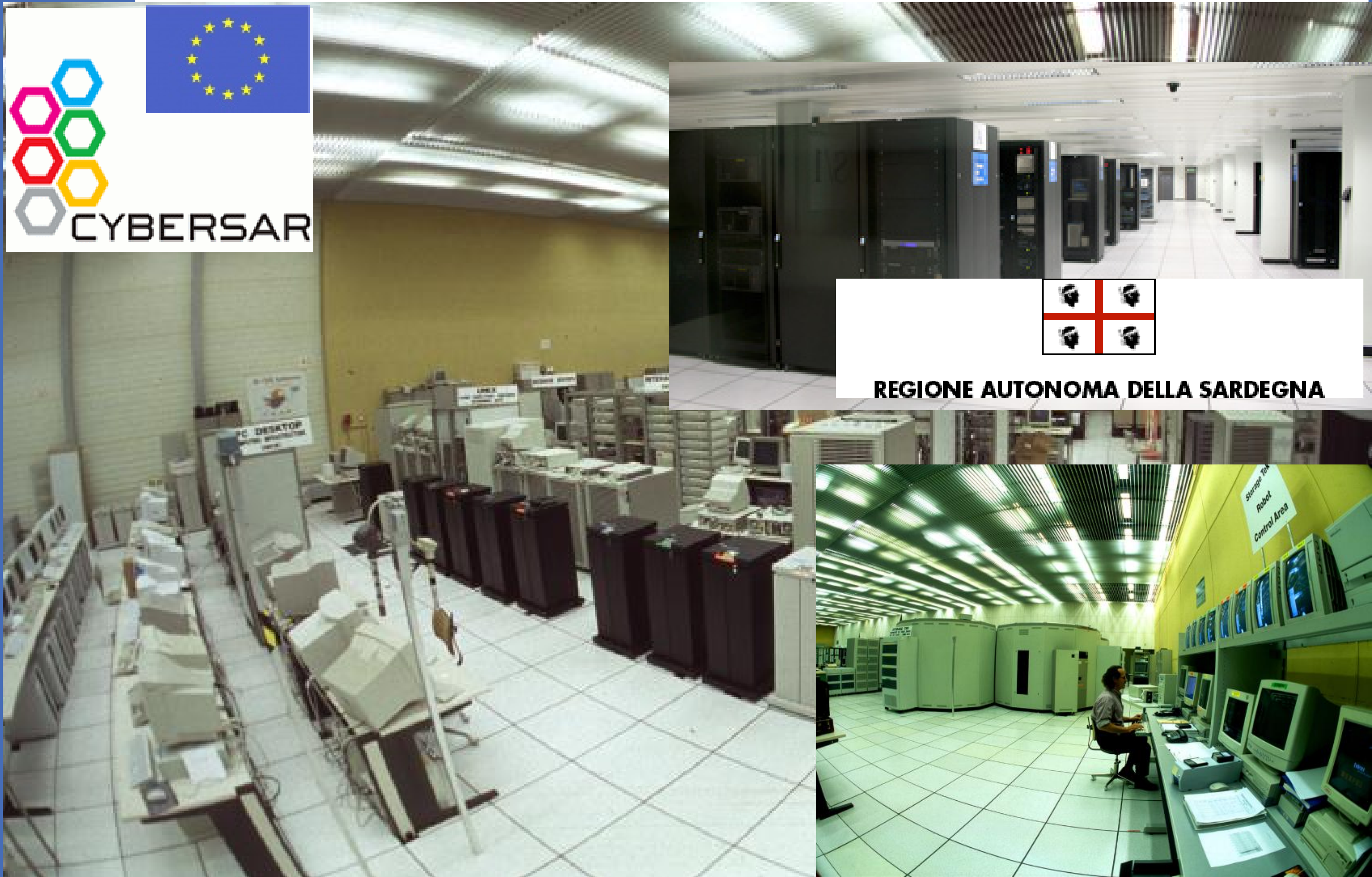
Tier 1 reference for AliSA

SAGrid is not strictly hierarchical – much more modular

SO WHAT ? How can that possibly help SA researchers ? If you're not on ALICE or ATLAS, you can't use it....

But I would like people in SA to have access to this infrastructure and massive computing power ....

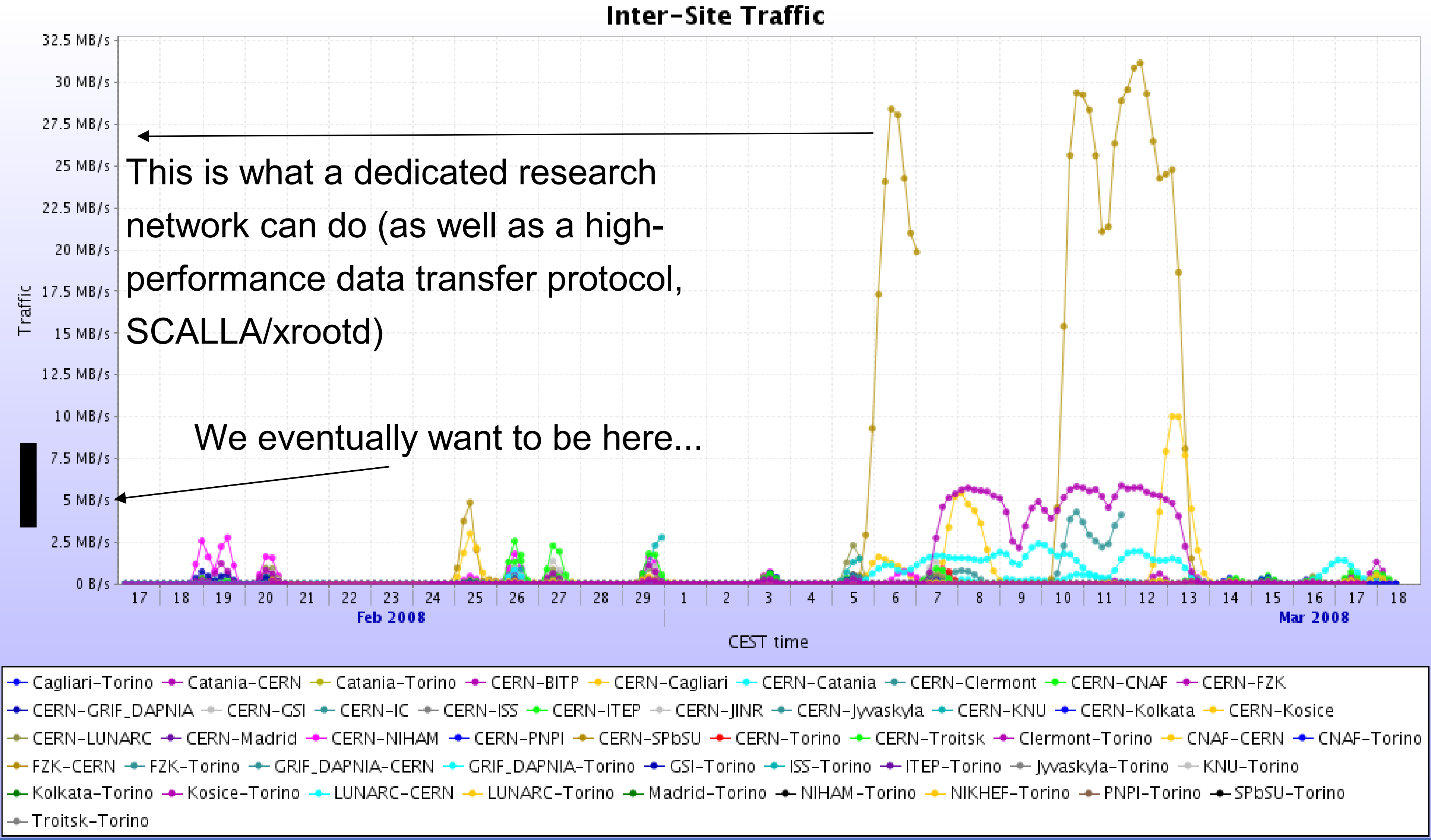
Using the grid to access the computing facilities around the world, especially where we already have existing collaborations



CYBERSAR

REGIONE AUTONOMA DELLA SARDEGNA

And access to data and networks



This is what a dedicated research network can do (as well as a high-performance data transfer protocol, SCALLA/xrootd)

We eventually want to be here...

SA National grid : Partners
(not a closed list, but not a free-for-all)

- A national grid initiative doesn't happen in isolation, collaboration is required both between national institutes and international partners
- Our collaborators
  - **South African Universities and National Labs**
  - Istituto Nazionale di Fisica Nucleare, Sezione di Cagliari and Catania
  - The EGEE training infrastructure **GILDA** (**G**rid **I**NFN **L**aboratory for **D**issemination **A**ctivities). Also supported by EU FP7 Project 'EPIKH'
  - Centre National de Recherche Scientifique (CNRS)
  - UNESCO/HPLabs anti brain-drain project with African countries
  - Staff exchange and training funding from EU FP7 project “EPIKH” - Exchange Programme to advance e-Infrastructure Know-How. (8 Meuros for 4 years)
- (Current) Sponsors
  - UCT-CERN Research Centre Consortium project, via CHPC
  - Institutes hosting phase-1 sites
  - Microsoft South Africa
  - Sun Microsystems EMEA
  - HP South Africa

SA National Grid : Who and where

- We currently have a small team of operators and a support team in Catania.
- Functional sites provide most of the manpower
- INFN, Sezione di Catania and CNRS (Orsay, Lyon) provide the training
  - 4 training events by the end of the year (2 in SA !)
  - Bilateral agreements between SA/IT and SA/FR
  - Multilateral FP7 Exchange Programme EPIKH : staff transfer, and several **week-long grid schools in South Africa**
- Inside SA-CERN
  - Central services admin : Gareth de Vaux
  - TLABS site admin : Sean Murray
  - UJ and Wits site admin : Norman Ives and Sergio Ballestrero
  - SA Grid CA has a Registration Authority at tLABS : Sean Murray
    - **GET A GRID CERTIFICATE FROM HIM !**
  - National Coordinator (me)
    - Tell me what you want or what you're not sure about.
- SA-CERN is placed in a very advantageous position to use the grid...
- ***Please do so !***

Moan moan moan, “What's the point, we don't have enough bandwidth” : Erm. Yes, we do. And if we don't we will negotiate it.

- SANREN is currently being deployed, it will take a while, but it will be there
- Experience in EU, elsewhere, has shown that when a second operator enters the market, a third soon follows and prices drop like a stone.
- Even if that doesn't happen, the CSIR is very aggressive on the deployment of SANREN
- International bandwidth is as always an issue... but new agreements are coming into place between EU/US/Indian NRENs and SANREN
- With a tightly-coordinated national grid infrastructure, we have a very strong bargaining power,
  - Major universities, national research labs
  - Bilateral agreements with IT, FR
  - International development programme backed by HP/UNESCO
  - Several development projects with MS, Sun, IBM, NICE, INFN, Amazon, Cosmolab, COMETA, ... always more on the horizon
  - EU FP7 “capacities” proposal December : O(1M) Euros
- A very different scenario than going to your “IT guy” and saying “please can we have some more bandwdith” !

SA National Grid : Situation Report

In April, I said : “What we could see within 18-36 months” :

Two major regional sites, 3-4 smaller regional sites, connected together in a non-centralised grid. Mirror Information Indices north and south...

Full integration into WLCG and EGEE

Collaboration with African sites :

Senegal

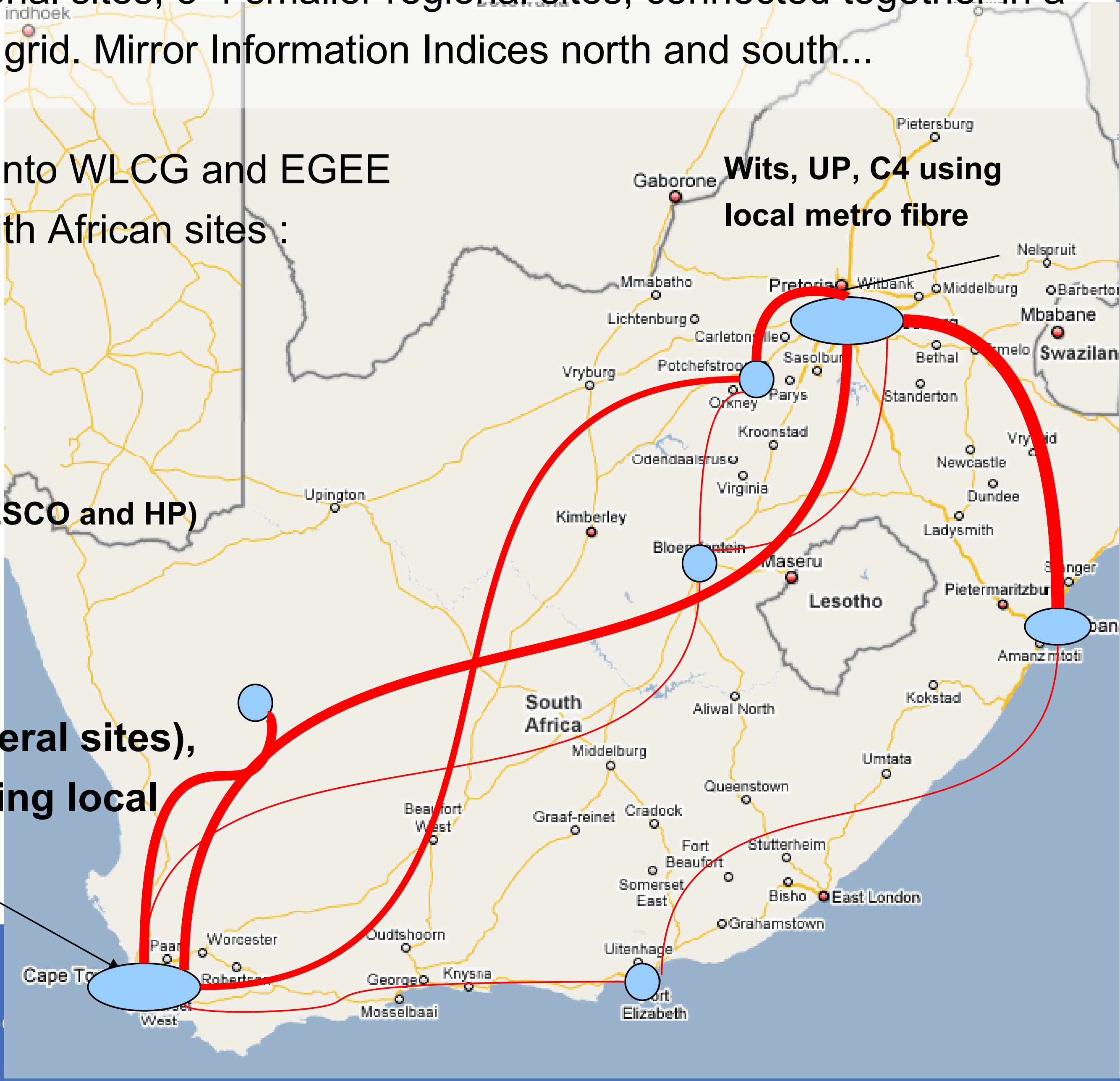
Nigeria

Zimbabwe

Ghana

Egypt

(together with UNESCO and HP)



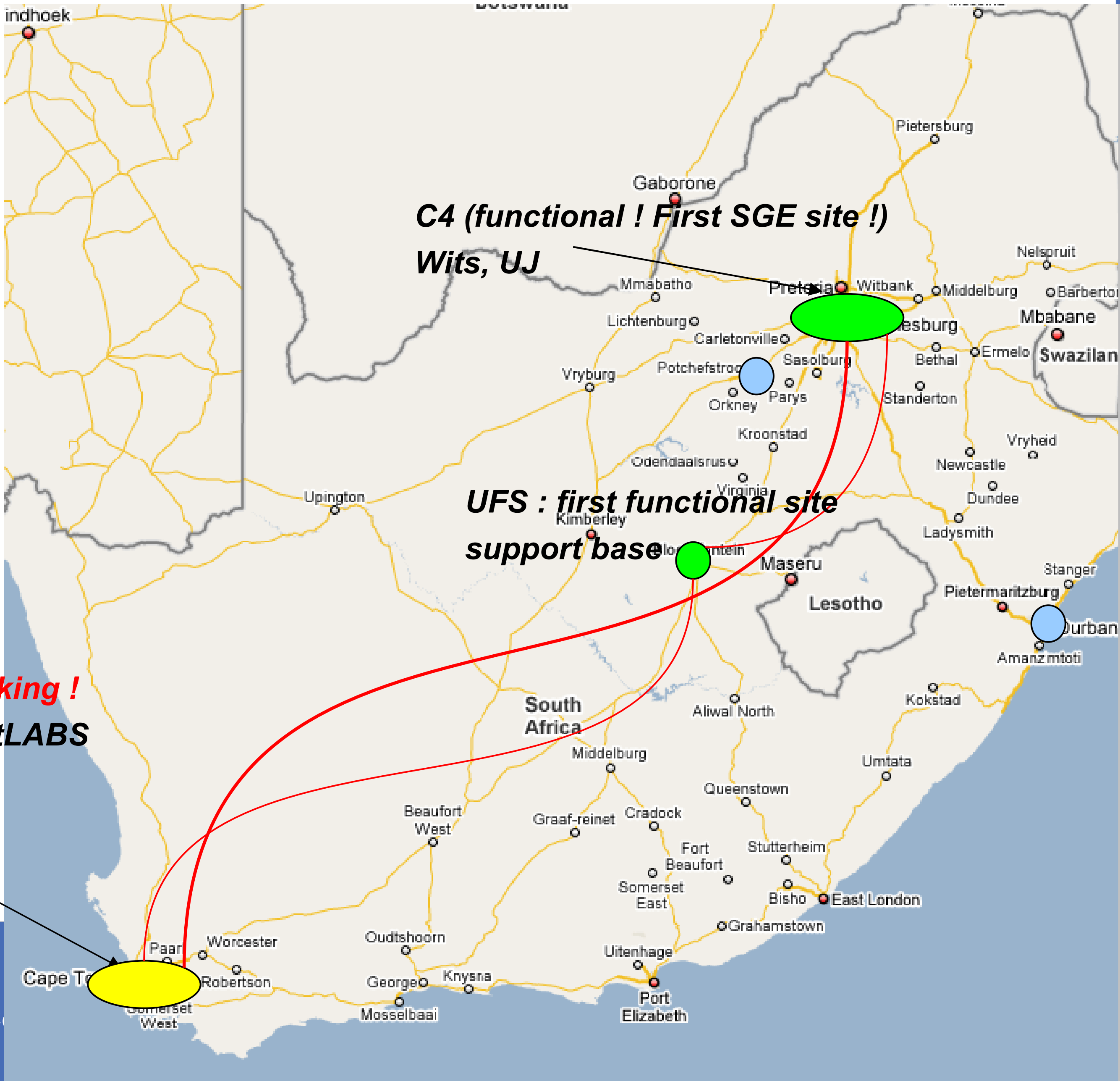
Wits, UP, C4 using local metro fibre

CHPC, UCT (several sites), iThemba LABS using local metro fibre

In November, we have :

Standing by :
US
UKZN
NWU
UP
SAAO

All the users and their applications



C4 (functional ! First SGE site !)

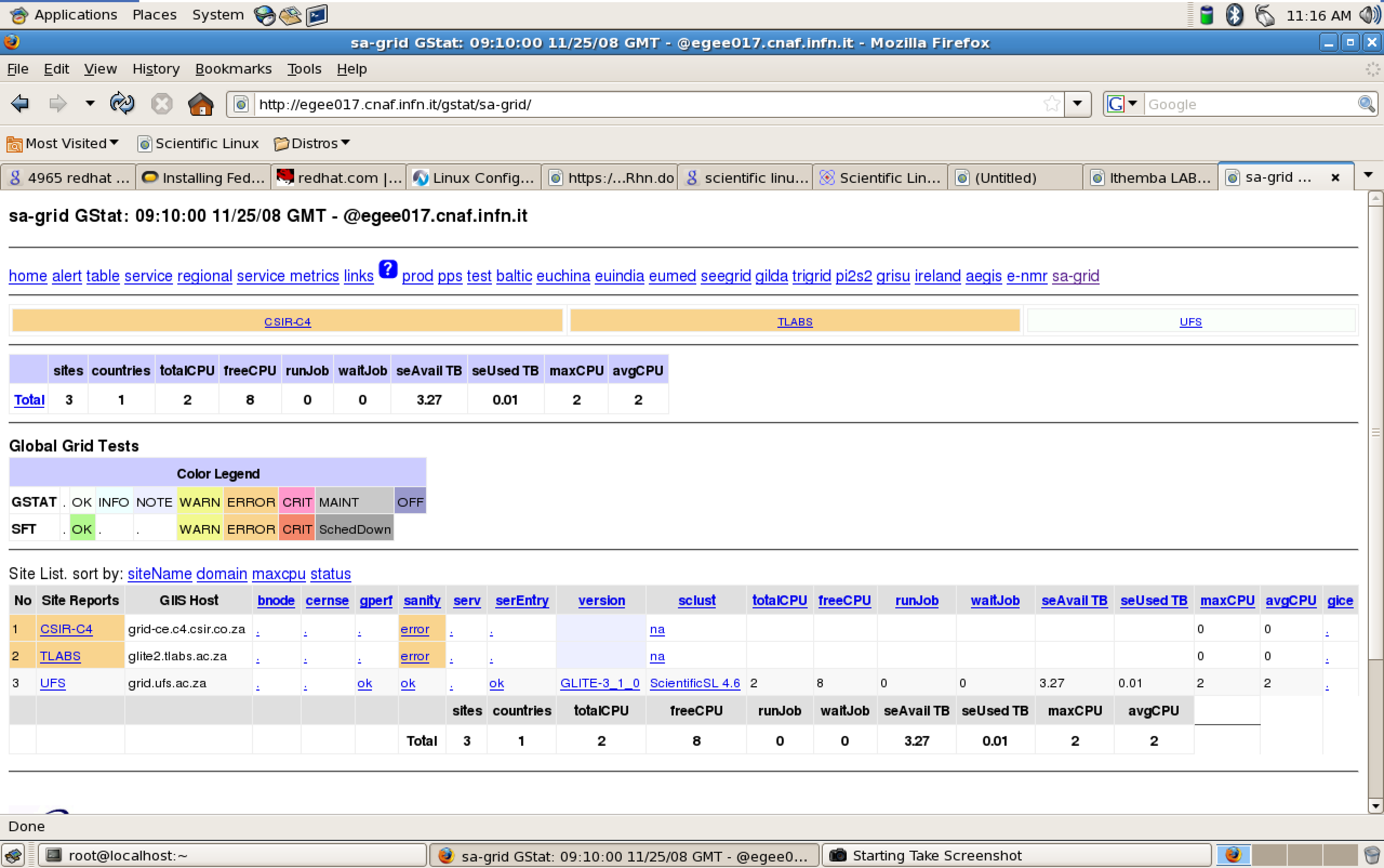
Wits, UJ

UFS : first functional site support base

Central Services working !

Compute element at tLABS

Breaking News : We are on the Global Operations Centre monitor


Applications Places System
11:16 AM
sa-grid GStat: 09:10:00 11/25/08 GMT - @egee017.cnaf.infn.it - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://egee017.cnaf.infn.it/gstat/sa-grid/
Google
Most Visited Scientific Linux Distros
4965 redhat ... Installing Fed... redhat.com |... Linux Config... https:/...Rhn.do scientific linu... Scientific Lin... (Untitled) Ithemba LAB... sa-grid ...
Done
root@localhost:~
sa-grid GStat: 09:10:00 11/25/08 GMT - @egee0...
Starting Take Screenshot

sa-grid GStat: 09:10:00 11/25/08 GMT - @egee017.cnaf.infn.it

home alert table service regional service metrics links ? prod pps test baltic euchina euindia eumed seegrid gilda trigrid pi2s2 grisu ireland aegis e-nmr sa-grid

CSIR-C4	TLABS	UFS

	sites	countries	totalCPU	freeCPU	runJob	waitJob	seAvail TB	seUsed TB	maxCPU	avgCPU
Total	3	1	2	8	0	0	3.27	0.01	2	2

Global Grid Tests

Color Legend									
GSTAT	.	OK	INFO	NOTE	WARN	ERROR	CRIT	MAINT	OFF
SFT	.	OK	.	.	WARN	ERROR	CRIT	SchedDown	

Site List. sort by: siteName domain maxcpu status

No	Site Reports	GIIS Host	bnode	cernse	gperf	sanity	serv	serEntry	version	sclust	totalCPU	freeCPU	runJob	waitJob	seAvail TB	seUsed TB	maxCPU	avgCPU	gice
1	CSIR-C4	grid-ce.c4.csir.co.za	.	.	.	error	.	.		na							0	0	.
2	TLABS	glite2.tlabs.ac.za	.	.	.	error	.	.		na							0	0	.
3	UFS	grid.ufs.ac.za	.	.	ok	ok	.	ok	GLITE-3_1_0	ScientificSL 4.6	2	8	0	0	3.27	0.01	2	2	.
							sites	countries	totalCPU	freeCPU	runJob	waitJob	seAvail TB	seUsed TB	maxCPU	avgCPU			
						Total	3	1	2	8	0	0	3.27	0.01	2	2			

We are as ready as you are... so, be ready !

- SA National Grid runs regular user training sessions
  - All material kept on the agenda server : http://indico.sagrid.ac.za
  - Next event is at Durban : http://indico.sagrid.ac.za/conferenceDisplay.py?confId=6 **SIGN UP !**
- Our grid will handle any application you can already run, guaranteed.
  - Don't sit back and waste time because you think you have not enough access to resources, talk to me.
- We are running dedicated application development and integration tutorials
  - No excuse not to use the grid
  - If you have a good idea but don't know how to integrate or develop the application for it, we can get someone to help you probably
  - Come talk to me

The South African National Compute Grid for SA-CERN

- A large part of SA-CERN activities revolve around computing :
  - Production computing (ALICE)
  - Analysis (ALICE, ATLAS)
  - Data acquisition and management, various code development, etc
- For production computing, there are two ways to integrate into the experiment-specific grid
  - Use the lightweight middleware of the experiment directly (AliEn)
  - Arrange a higher-level agreement between WLCG and a separate infrastructure (which clearly should already exist).
- But what about “utility” or “individual” computing ?
  - Individual members or smaller groups have access to computing power as they need it (no need to waste money on hardware
- In order to be a functional collaboration, we need to be talking the same language, in terms of computing
  - Similar methodology between site admins will allow them to help each other out
  - Transparency between sites means work can be portable (visitors, etc)
  - Student mobility is increased, there is no stupid barrier to entry (one framework to learn)

How to use this stuff...

- If you missed a training session, here is a quick intro to how the grid works, to demystify if for you...

SA Nat'l Grid is based on production services from gLite, by EGEE
gLite : General Services

The basic Glite middleware consist of:

- **Authentication** and **Authorization System**
- **Information System**
- **Workload Management System**
- **Data Management System**

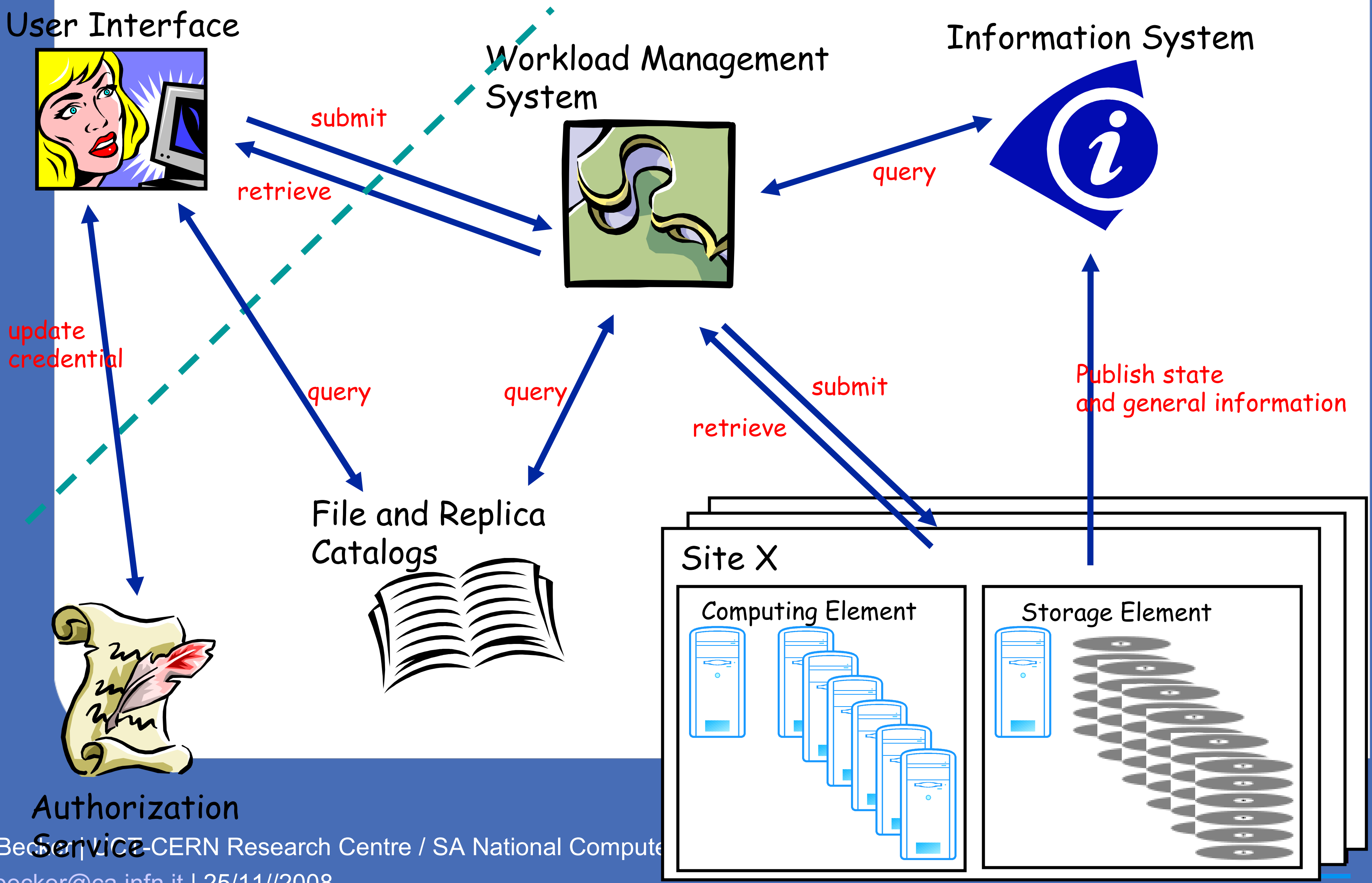
some sites run :

- **Various Monitoring Services**

and a lot of sites with:

- **Computing Element** Grid interface to local computing cluster
- **Worker Nodes** the local computing cluster
- **Storage Element** the local storage solution

Grid Topology and Services



General JDL

```
Type = "Job";
JobType = "Normal";
Executable = "/bin/sh";
StdOutput = "std.out";
StdError = "std.err";
InputSandbox = {“script.sh",“program.exe"};
OutputSandbox = {"std.out","std.err"} ;
Arguments = “script.sh";
Requirements = other.GlueCEUniqueId == "ce-
  cybr.ca.infn.it:2119/jobmanager-lcglsf-infinite"
```

Do what you would usually do in a script, just tell the job where to land (or leave that up to the WMS)

“GLUE Schema” is the way we describe site configurations (an official EGEE metadata structure, but can be extended)

User Applications

- How does the grid know what your application is or where it is ?
  - Usually applications are installed on the compute element
  - Widely-used applications, used throughout a VO, will be installed in a standard place.
  - Standard configuration is published in the site Information Index
  - Non-standard configuration can be passed with the job
  - The Application can also be passed with the job
    - You can even send an entire virtual machine with your application pre-installed with the job (in the InputSandbox).
- You need to know how to use your application, when it is installed, tell the site admin the configuration you or your team would like, and this will be inserted into the Information Index
- It is not so hard... actually very very easy... please TRY IT !

Summary

- SA National Grid is working to a very fast timeline... about 2-3x faster than we had anticipated
- Mostly thanks to help and support from GILDA and INFN
- It is a national infrastructure, supported in a federation by almost all University IT departments to a greater or lesser extent
  - Eventually by DST, EU
- SA CERN is providing a large amount of the manpower and support for operations and development (UJ, iThemba, UCT-CERN)
- It's time for users to USE...
  - Go to your RA (Sean Murray) with your ID book and get a grid certificate.
  - Come to the training session (or download the previous ones)
  - Contact me with your application
  - Or fill out the questionnaire : http://grid.ct.infn.it/infn/questionario/index.php?sid=62599&lang=eng

Extra slides

gLite : Description of some general services – Security and Information Systems

- The **Authentication** and **Authorization System**:
  - Contains the list of all the people authorized to use gLite
    - divided by VO
    - downloaded to all machines running Grid services
    - map the gLite users to the local users of the machine
- The **Information System**:
  - provides information about gLite resources and their statuses.
  - Information published by the individual resources and copied into central databases.
  - Used by:
    - WMS/RB: match resources against job requirements and to rank them
    - DMS: storage resources and file catalog
    - monitoring systems

gLite : Description of some general services – WMS, DMS

- The **Workload Management System**:
  - manages jobs submitted by users
    - matches the job requirements to the available resources
    - schedules the job for execution on an appropriate computing cluster
    - tracks the job status
    - allows the user to retrieve the job output when ready
- The **Data Management System**:
  - Allows users to
    - move files in and out of the Grid
    - replicate files among different locations
    - locate files.
  - This is achieved by:
    - transferring data via a number of protocols (GridFTP is the most commonly used)
    - interacting with a central file catalog

gLite site services

- **The CE runs a gatekeeper**
  - Accepts jobs from Condor-G
  - Creates a **job manager** ***(JM)*** per job
    - Generic interface to the batch system
  - The JM *only* submits or cancel a job
  - The **grid monitor** queries the status of the jobs
    - ***One instance per CE per user***

- **The local batch system**
  - Last element of the chain
  - Often a server runs on the CE node

gLite site services – the last compute layer

- **The WNs are the host executing the job**

- A set of WNs managed by a CE by **local batch system** constitutes a compute cluster

- A cluster MUST be homogeneous
  - Similar hardware
  - Same OS, configuration …

- The gLite WN does NOT run any service
  - Requires a minimal amount of Grid middleware

- The WN runs a job wrapper
  - Wrapper around the user executable
  - Transports the input/output sandbox from/to the RB

Authentication and Authorization System

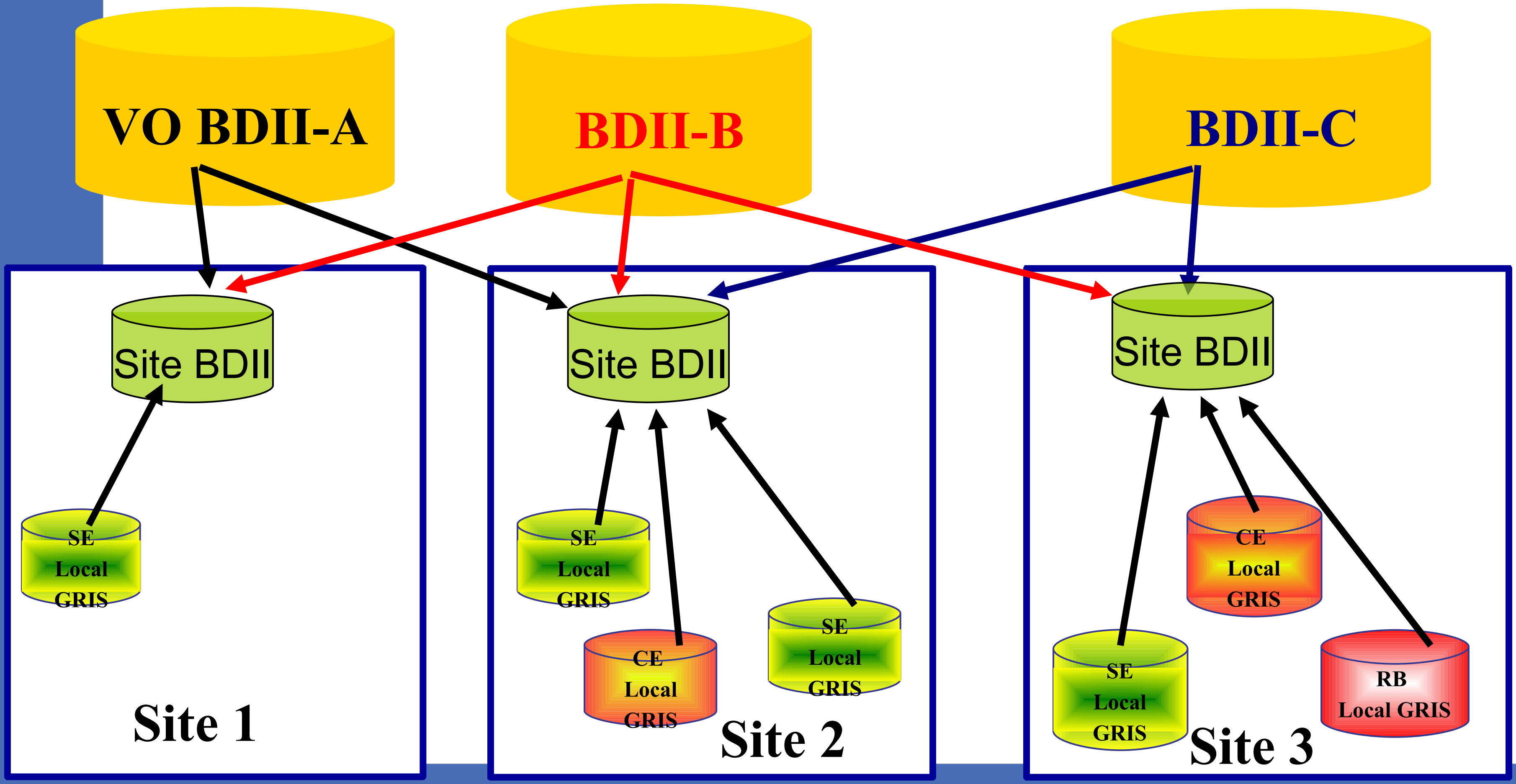
- User authentication based on central databases
  - one database per VO
  - database contains the certificate subjects of all gLite users.

- Databases accessed by RBs (a part of WMS), CEs and SEs
  - locally build a list of authorized users (/etc/grid-security/grid-mapfile)
  - The list maps user certificate subjects to local “pool” accounts

```
. . .
"/C=IT/O=INFN/OU=Personal Certificate/L=COSMOLAB/CN=Francesca Mocci" .cybersar
"/C=IT/O=INFN/OU=Personal Certificate/L=COSMOLAB/CN=Giuseppe Saba" .cybersar
"/C=IT/O=INFN/OU=Personal Certificate/L=COSMOLAB/CN=Ignazio Pillai" .cybersar
. . .
```

  - Users with a ‘.’ in are illegal in UNIX.
    - Signal to the globus libraries to allocate a pool account, eg cybersar005

Information System Hierarchy



The Data Management system

- The DMS relies on two kind of services:
  - The File Catalog (FC)
  - The Storage element (SE)
- The **File Catalog**
  - Needed to maintain mappings between
    - Global Unique Identifiers of files (GUID)
    - Logical file names (LFN)
    - Phisical File Names i.e. Locations (PFN)
  - Is an application server frontend to a database backend
    - Currently ORACLE, but support also for MySQL
- The **Storage Element**
  - Provide uniform access to storage space
  - SE can manage several kinds of back-ends...

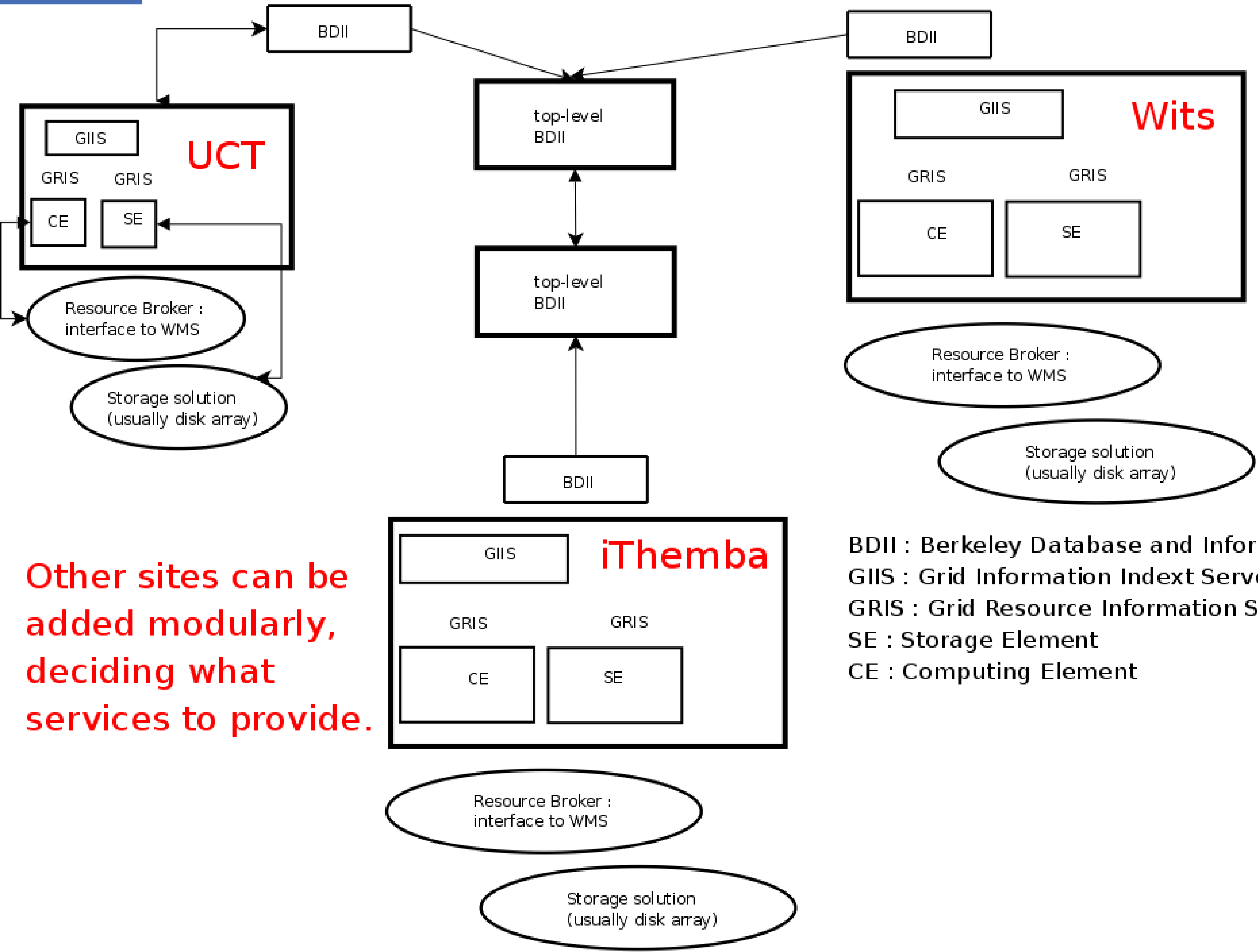
Job Preparation – Job Description Language

- Information to be specified when a job has to be submitted:
  - Job characteristics
  - Job requirements and preferences on the computing resources
    - Also including software dependencies
  - Job data requirements

- Information specified using a Job Description Language (JDL)
  - Based upon Condor’s *CLASSified ADvertisement language (ClassAd)*
    - Fully extensible language
    - A ClassAd
      - Constructed with the classad construction operator
      - It is a sequence of attributes separated by semi-colon (;).

- So, the JDL allows definition of a set of attribute, the WMS takes into account when making its scheduling decision

ZANuclGrid : services


BDII
BDII
top-level BDII
top-level BDII
BDII
GIIS
UCT
GRIS
GRIS
CE
SE
Resource Broker : interface to WMS
Storage solution (usually disk array)
GIIS
Wits
GRIS
GRIS
CE
SE
Resource Broker : interface to WMS
Storage solution (usually disk array)
GIIS
iThemba
GRIS
GRIS
CE
SE
Resource Broker : interface to WMS
Storage solution (usually disk array)

Other sites can be added modularly, deciding what services to provide.

BDII : Berkeley Database and Information Index
GIIS : Grid Information Indext Server
GRIS : Grid Resource Information Service
SE : Storage Element
CE : Computing Element

This is entirely extensible... and open !

- This is being done for now for physics departments, but very soon, everyone will be welcome
  - It's not just CERN-related, or ALICE related
  - It's not just physics
  - It can handle almost any kind of applications
- Especially the CHPC
- We see this as providing
  - a national infrastructure for researchers
  - The right kind of interface for groups nationwide to use the CHPC
- We clearly need SANReN, but the days of being stuck without a computing resource ***will be over.***
- This brings incredible opportunities... but also some challenges

A national computing grid... the up side

- What does a national computing grid mean to South African researchers ?
  - vastly improved possibilities for collaboration within the country
  - opportunities to learn and use a tool which has transformed research in all areas in the rest of the world
  - coherence with HPC developments in the country (“no lab left behind”)
  - Full and efficient utilisation of SANReN
  - Access to the grid for smaller labs who may not have HPC facilities or are just starting out
  - Learn the technology and adapt it to our local environment
- The “down side” is we have to work together, co-ordinate activities, do some more work,push for funds, etc... (not really a downside)

But we cannot do this by ourselves !

Support from EGEE and INFN and others

- I have been working to raise the level of support from CERN, EGEE, INFN for Cape Town (this is actually moving faster than I can handle !)
- It takes just a little bit more to do it for the rest of the country
- EGEE has an interest in spreading use of gLite in the world, and Africa is a big “market” for them
- INFN has already committed to send gLite trainers to South Africa this week
- INFN, CT are applying for EU funds (~ 8M Euro) for trainers, forsee a site-manager and user-level training courses later this year
- Some support from HP Labs in Geneva
- Almost all major vendors in SA are interested
  - Breakpoint (Sun), BCX (IBM), HP...

End...
... or is it the beginning ?

- The aim of this presentation is to inform – we do not ask direct participation yet from IT departments
- However, researchers are moving forward with implementations right now...
- We would like IT to play a significant role in enabling a national computing grid
- This essential tool for collaboration and massively distributed computing will be built anyway – but will be painful without co-operation and support
- If you think your department would like to be part of the protoype or partake in planning, please email bruce.becker@ca.infn.it

Essential JDL

- **At least one has to specify the following attributes:**
  - the name of the executable
  - the files where to write the standard output and standard error of the job
  - the arguments to the executable, if needed
  - the files that must be transferred from UI to WN and viceversa

```
[
Executable = “ls -al”;
StdError = “stderr.log”;
StdOutput = “stdout.log”;
OutputSandbox = {“stderr.log”, “stdout.log”};
]
```

Job Submission

```
(glite/edg)-job-submit [-r <res_id>] [-c
  <config file>] [-vo <VO>] [-o <output
  file>]  <job.jdl>
```

-r the job is submitted directly to the computing element identified by <*res_id*>

-c the configuration file <*config file*> is pointed by the UI instead of the standard configuration file

-vo the Virtual Organisation (if user is not happy with the one specified in the UI configuration file)

-o the generated edg_jobId is written in the <*output file*>