A Light for Science

European Synchrotron Radiation Facility

The XRAY Grid

ESRFUP-WP11

Outline

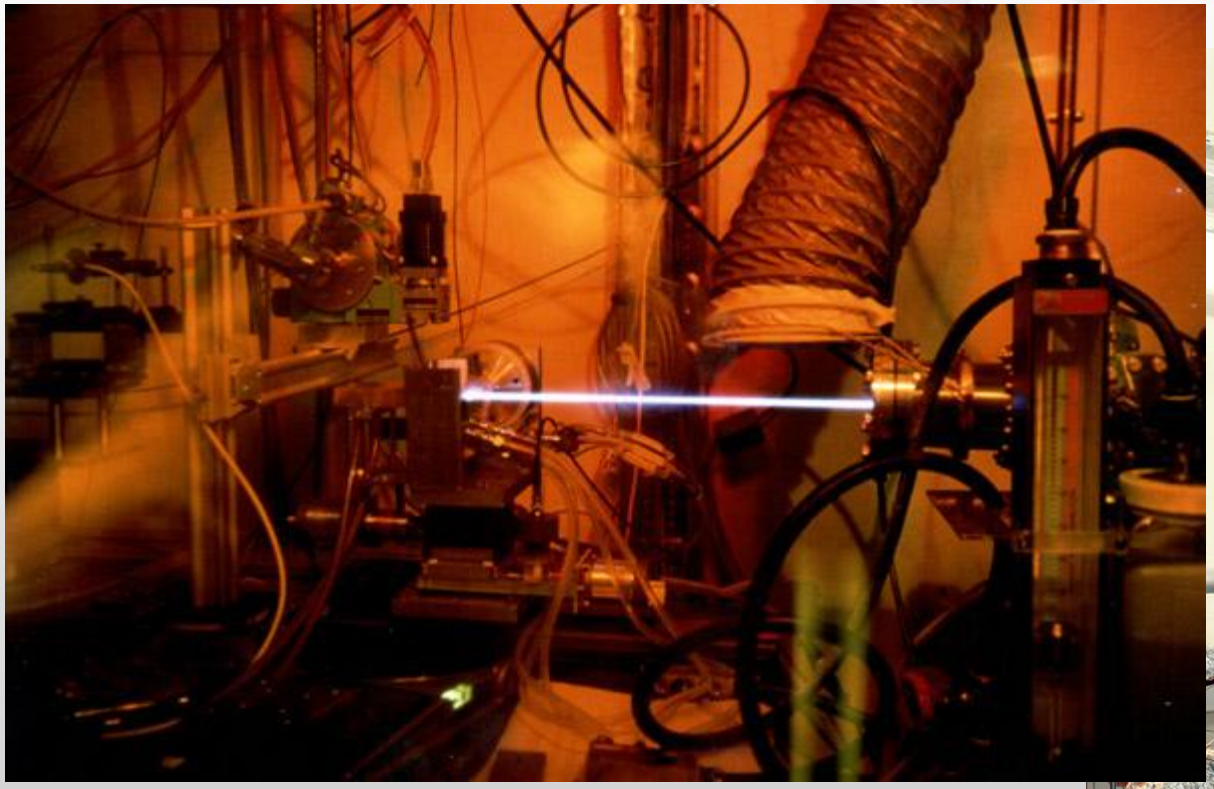
- The European Synchrotron Radiation Facility
- The grid evaluation project (WP11 of ESRFUP)
- Setting up the XRAY infrastructure
- Test Cases
- Community
- Conclusion

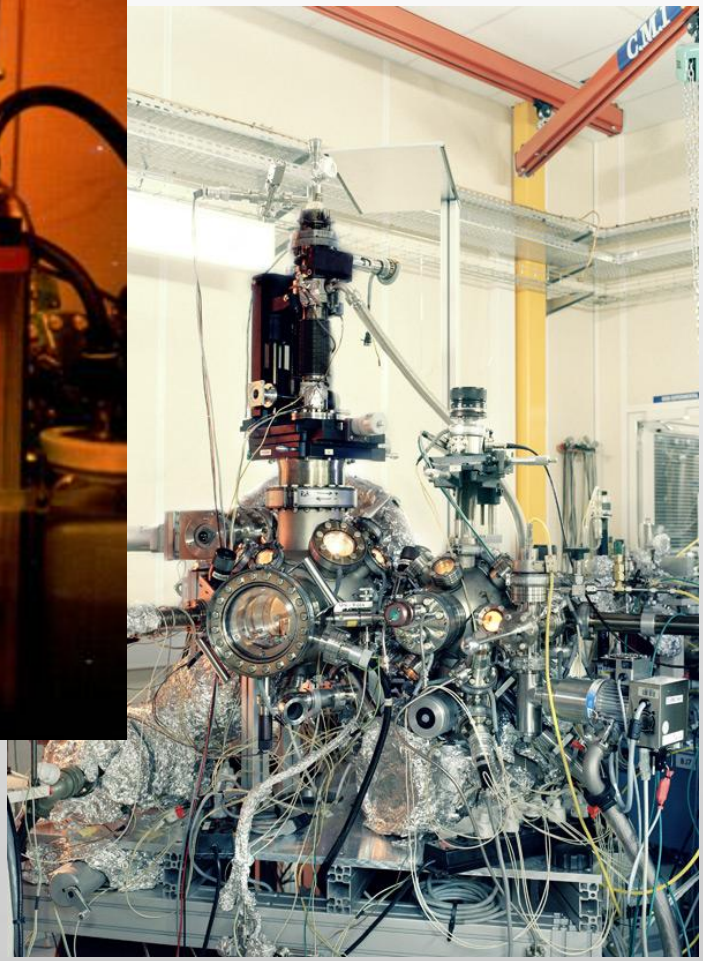
European Synchrotron Radiation Facility

www.esrf.eu

The beamlines

Wide range of Experiments and setups





Solid State Physics, Material Science, Chemistry, Life Science, Paleontology, …

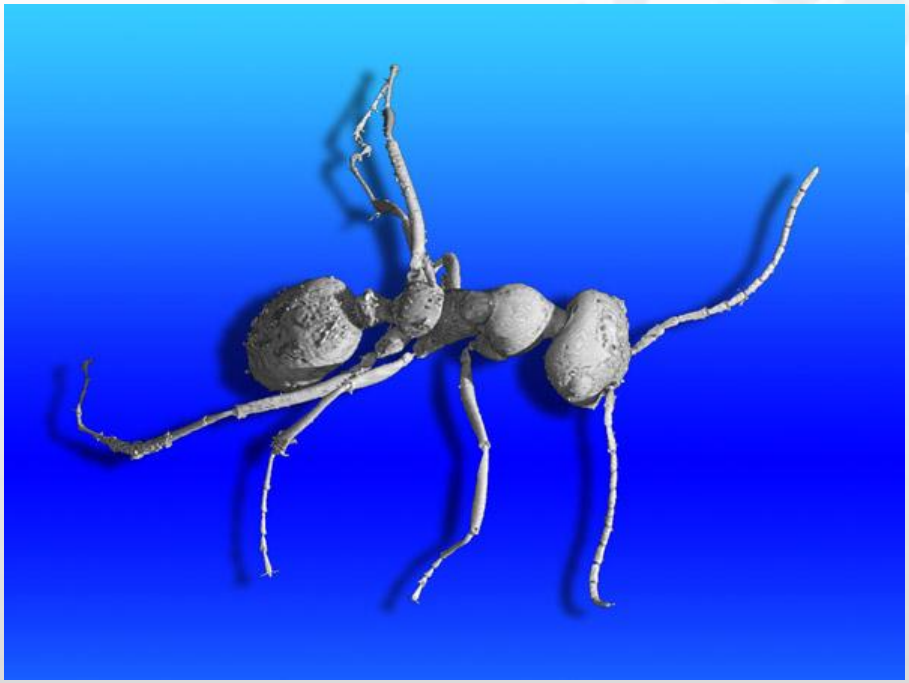
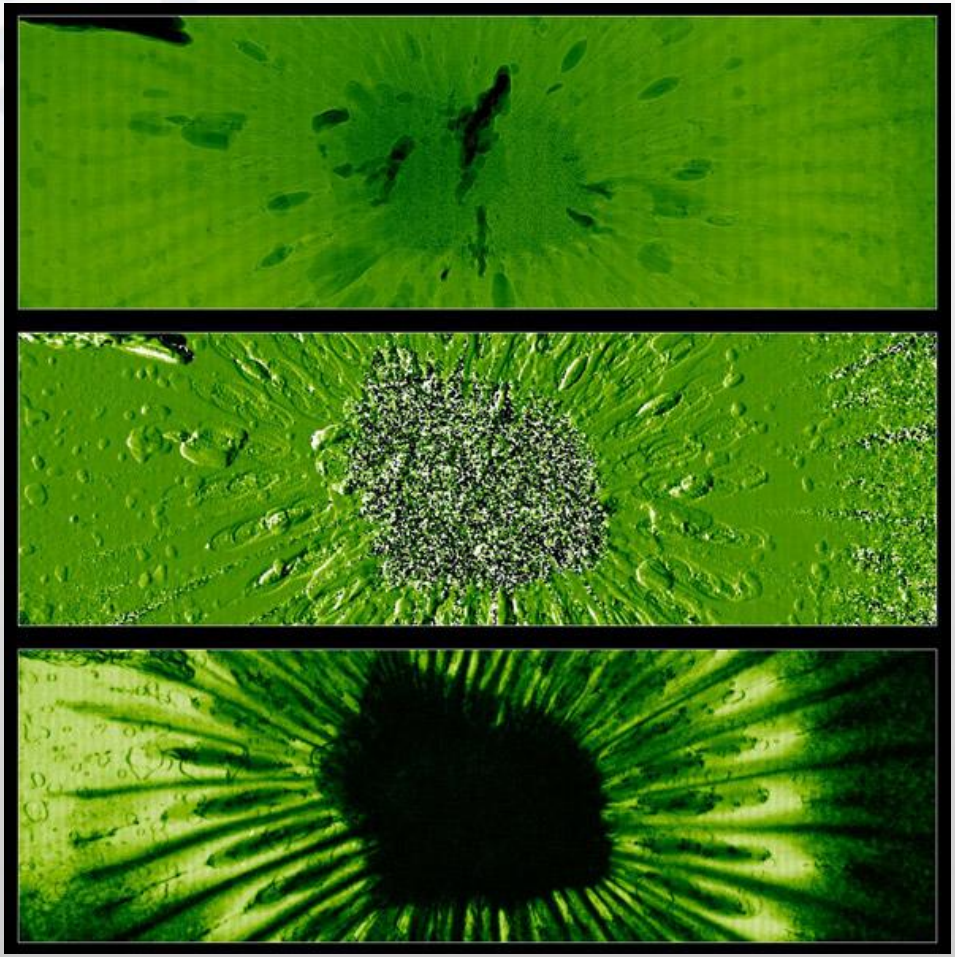


Image 3D d'une fossile de 100 million d'annees
optenue par microtomographie avec rayon X

Plus sur -> paleo.esrf.eu

Radiographies en interférométrie à réseaux d'une tranche de kiwi
énergie des photons rayons X : 27 keV



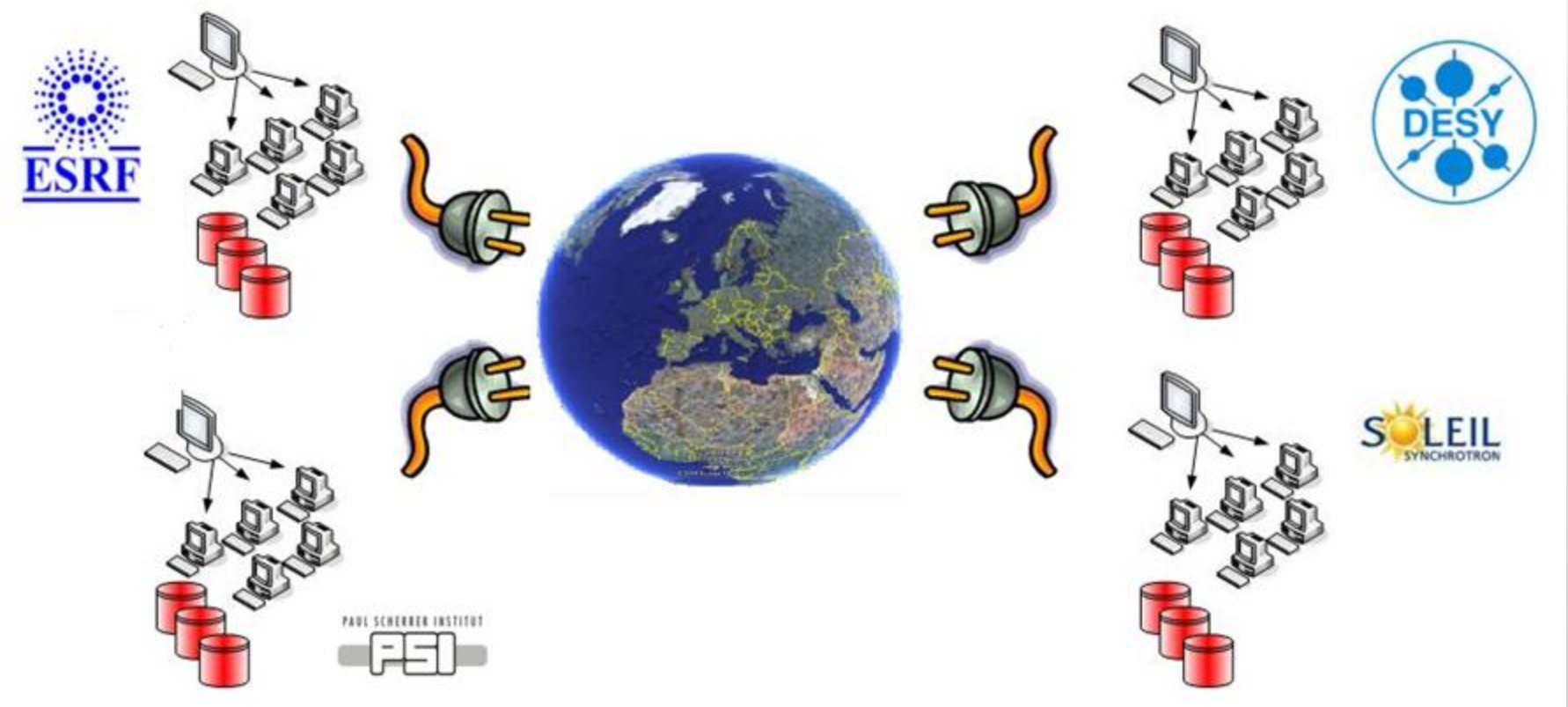
The deliverables of ESRFUP-WP11

- Deliverable: Collaboration Agreement between 3(-4) partner labs

for the creation of a Synchrotron Radiation Virtual Organization (Jan '09)

- Purchase compute and storage system, install and test gLite to form Grid Sites at the ESRF and 2(-3) other partner sites
  - Deliverable: ESRF Grid Site operational (Jan '09)
  - Deliverable: Partner Grid Sites operational (Jun '09)
- Organize a workshop with possible future partners (Dec '08)
- Gridify one resource intensive applications on the test bed, write wrapper software, make added value analysis
  - Deliverable: Test case software operational (Nov '09)
- Final report on operational experiences with the international test bed installation including future orientations for photon science grid activities (Jan '10)

XRAY Grid Testbed of ESRFUP-WP11



Creation of a Synchrotron Radiation Virtual Organization

- The **XRAY VO** has been created
  - .. and registered with EGEE https://cic.gridops.org/index.php?section=vo&vo=xray.vo.eu-egee.org
- **V**irtual **O**rganization **M**anagement **S**ervice was set up
  - Enrolment URL:
    - https://grid-voms.esrf.eu:8443/voms/xray.vo.eu-egee.org
- Registered users are granted access to the resources of a VO according to their group membership and assigned role

Hardware specs

- Storage Components
  - Sun Fire X4500 Server (Thumper)
  - 2 dual core AMD Opteron, 16 GB
  - OpenSolaris based OS (SunOS 5.10)
  - 24 TB internal storage

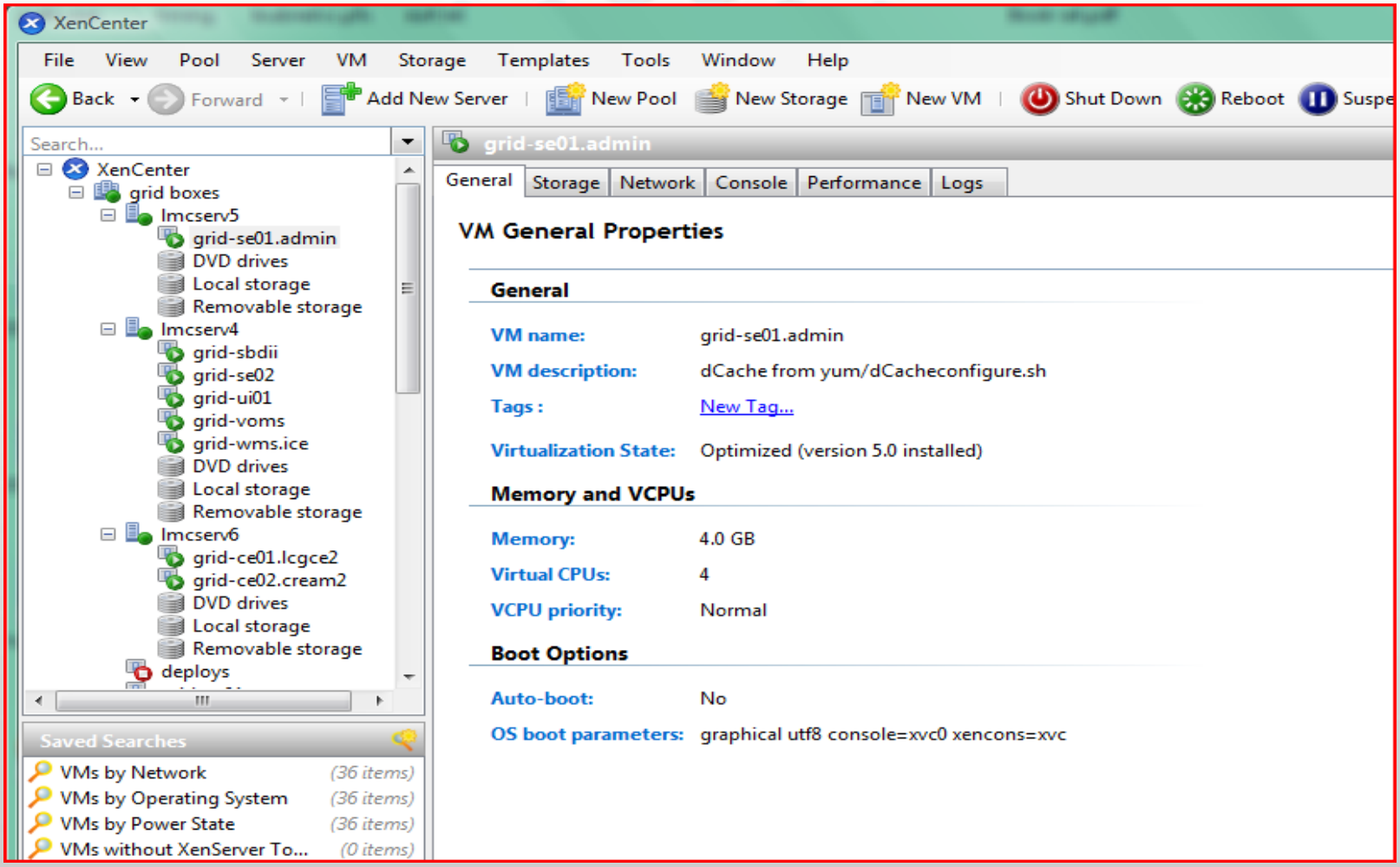
- Compute Nodes
  - Sun Fire X2200 M2 Server
  - 2 quad core AMD Opteron
  - 16 GB main memory, 250 GB HDD
  - Scientific Linux 4.7

Hardware Specs

- Middleware servers
  - Sun Fire X4150
  - 2 x quad core Intel Xeon E5440
  - 16GB main memory, 2x73 GB HDD
  - Citrix XenServer 5.0 Enterprise Edition
  - Virtual Machines on Scientific Linux 4.7 (i386 or x86_64)

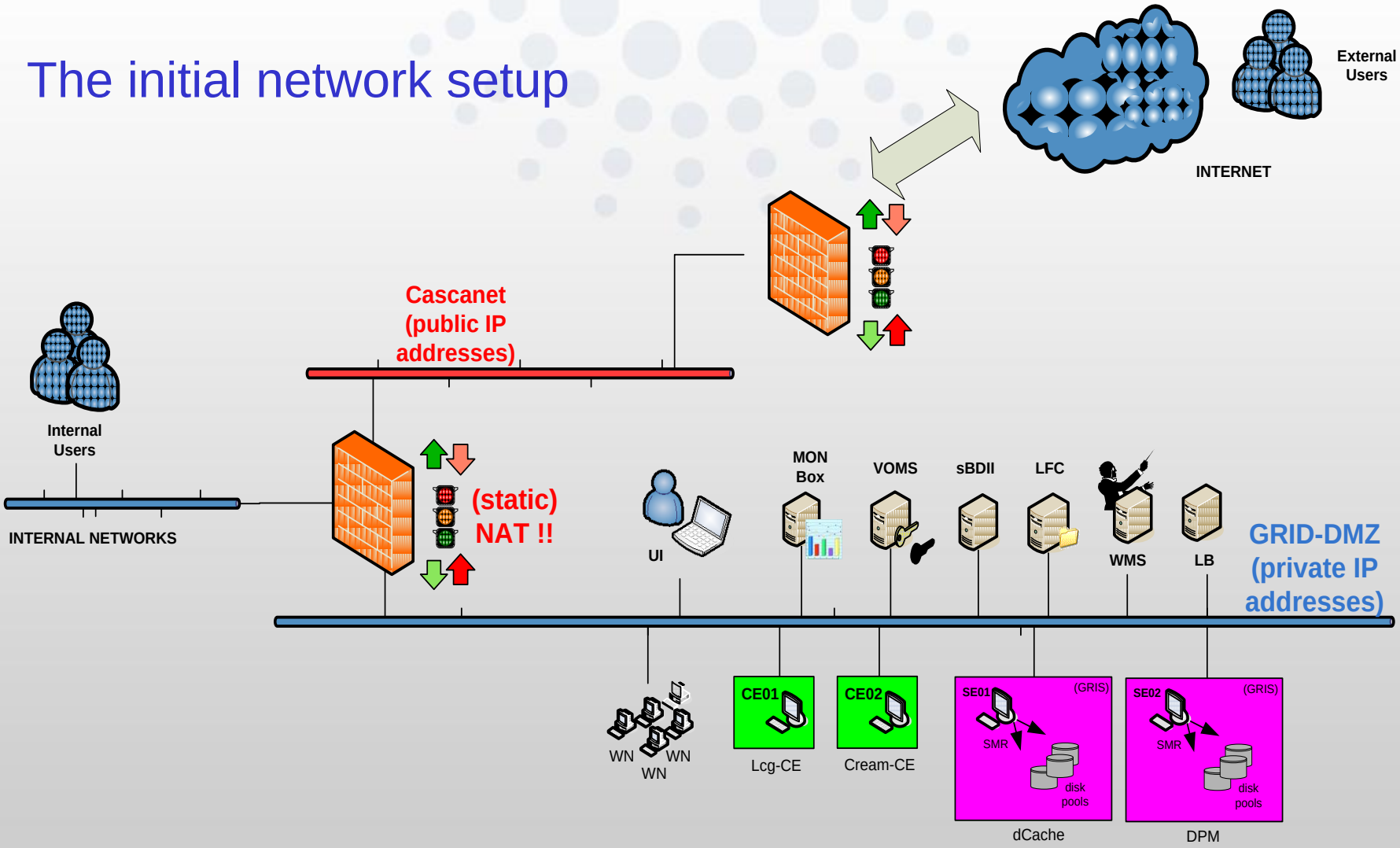
- Rack
  - Sun Rack 900-38
  - Network Switch, Extreme x450e-48p

Virtualization using Citrix XenServer 5.0 EE

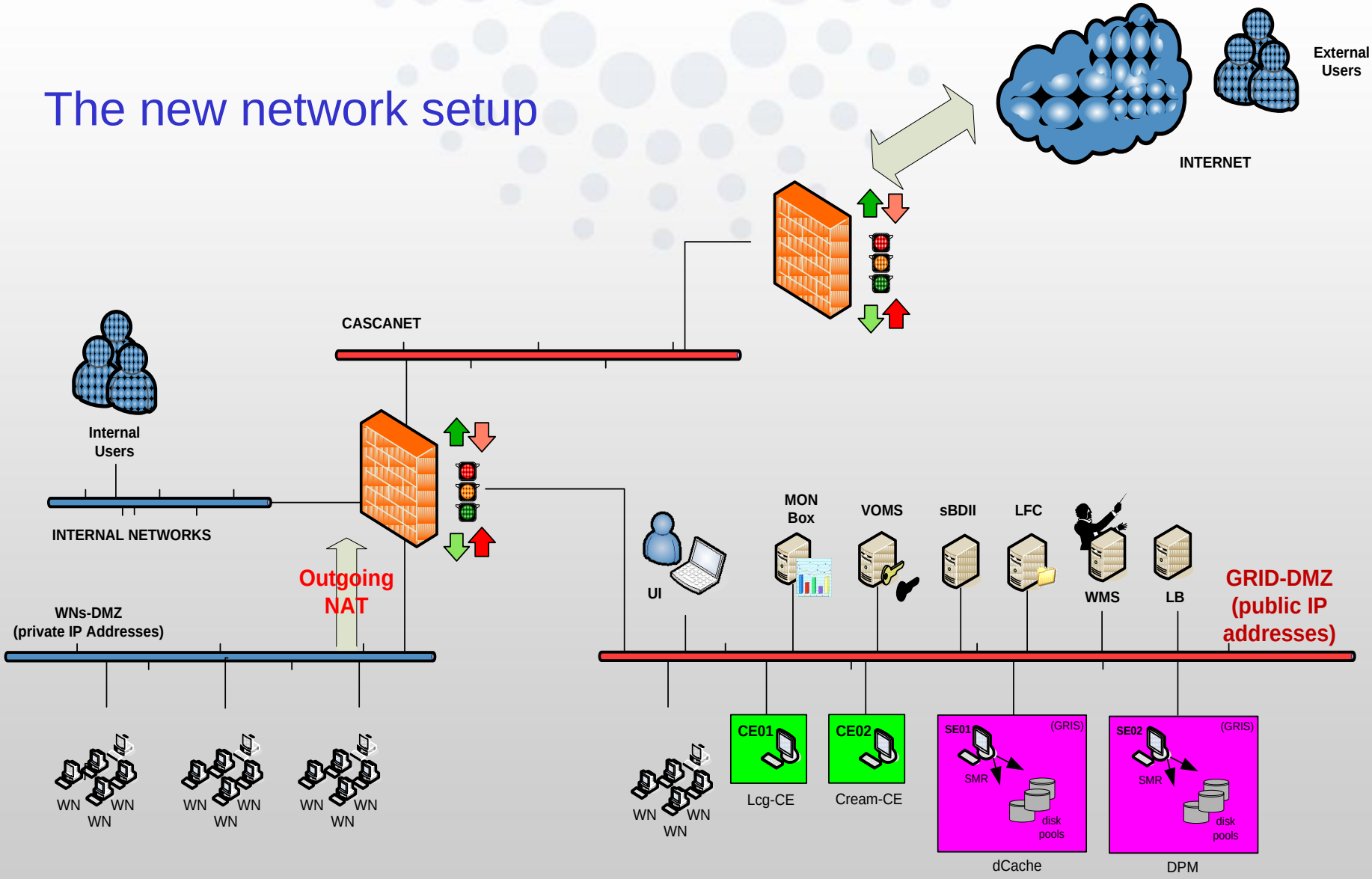


XRAY Grid

The initial network setup



The new network setup



XRAY setup @ ESRF

- **14 Worker Nodes** with altogether **80 Cores**
- **12 TB of disk** space (RAID-Z2)
- **2 Computing Elements** in Test Bed
  - **Lcg-CE** MPI enabled
  - **Cream CE**
- **2 Storage Elements**
  - **dCache**
  - **DPM**
- **1 Site BDII**
- **2 User interfaces** 32bit (internal + external)
- **1 VOMS**
- **1 WMS**
- **(1 local LFC, 1 Myproxy server)**

Hardware set up @ partner sites

XRAY setup @ DESY

- DESY enabled XRAY on their existing grid site
- Shipped one thumper (~**20TB**) for their dCache pool
- DESY has added **WMS, LFC** and **AMGA** as core services to the XRAY VO

XRAY setup @ PSI

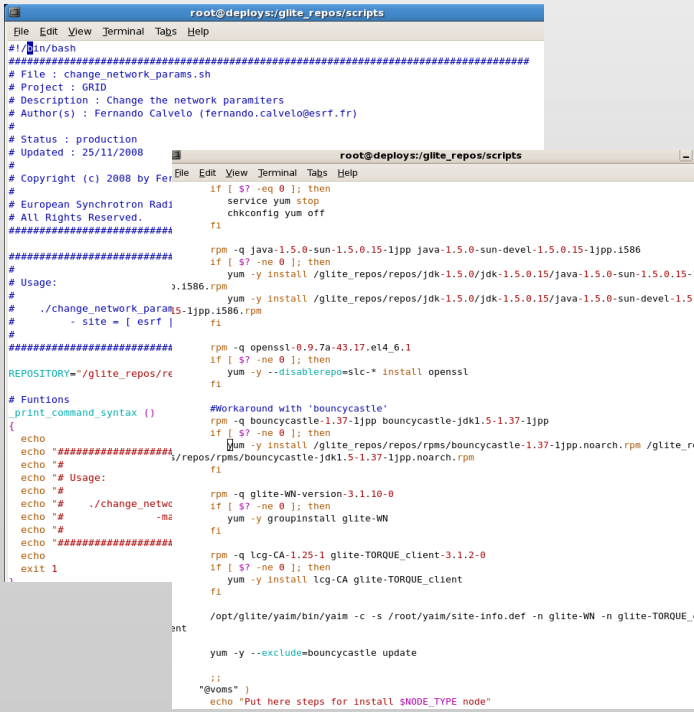
- 6 Worker Nodes with **48 Cores**
- Up to **20TB** of disk space
- 1 Computing Element: **lcg-CE**
- 1 Storage Element: **dCache** (1.9.0-8)
- 1 User interface, Site-BDII and Monbox
- Operational

XRAY setup @ Soleil

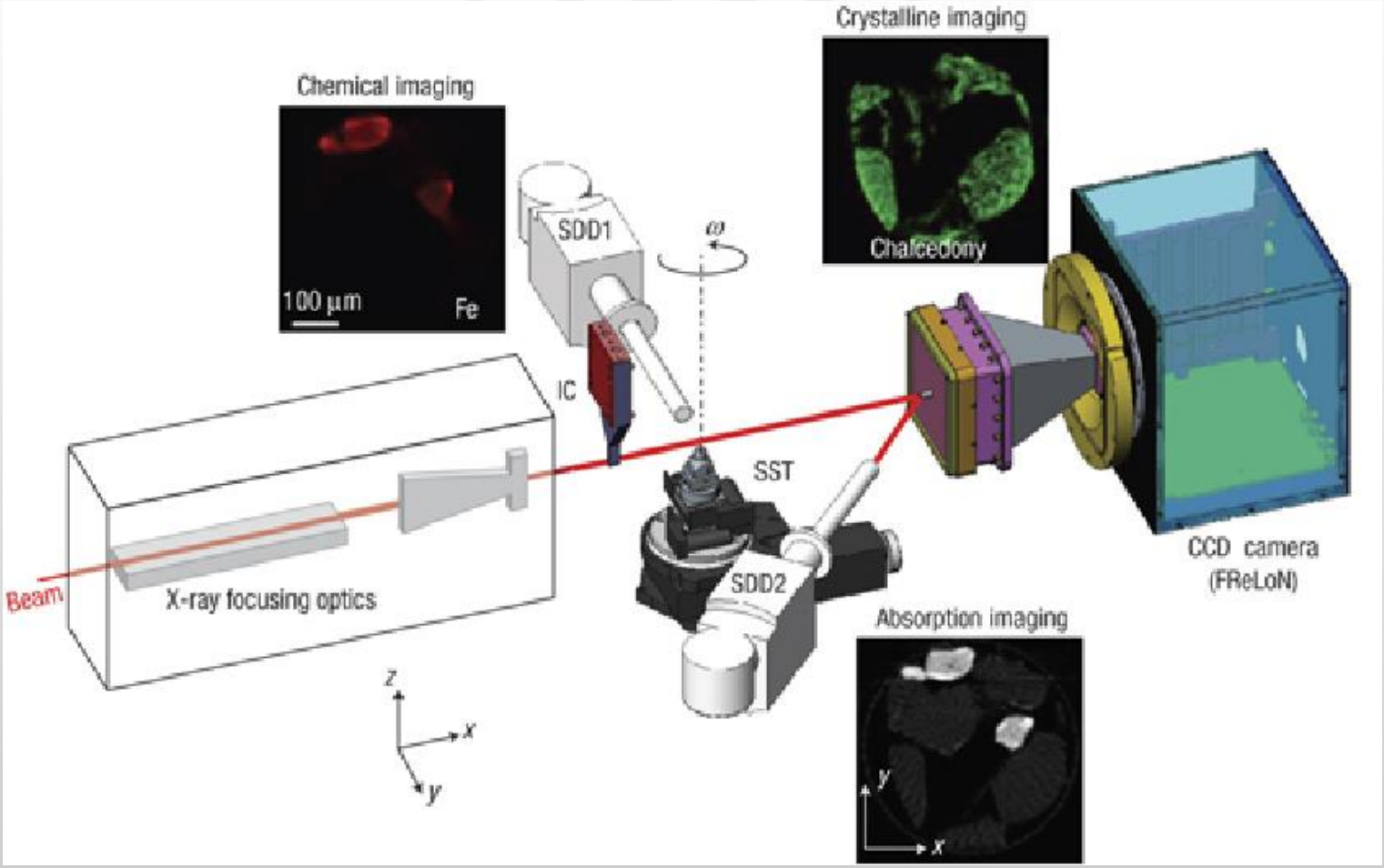
- 6 Worker Nodes with **48 Cores**
- Up to **20TB** of disk space
- 1 Computing Element: lcg-CE
- 1 Storage Element: **DPM**
- 1 User interface, Site-BDII, and Monbox
- Cluster delivered
- Pending final configuration

Handling the site administration

- Created a couple of deployment and configuration scripts
  - Similar to the current practice at the ESRF
- Facilitates the deployment and update of services
- Enables us to keep the configuration homogeneous
- Monitoring status with Ganglia and Nagios



A typical setup

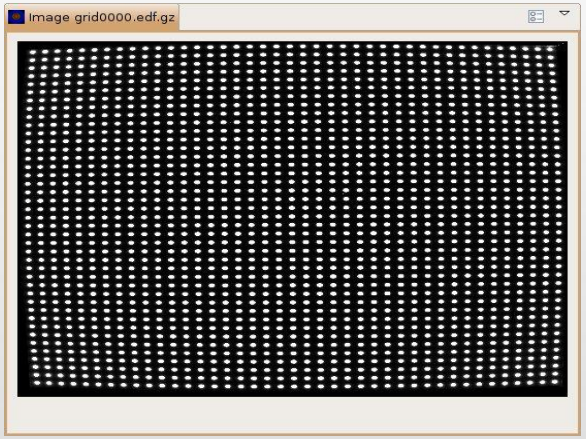


Requirements of synchrotron users

- Lots of small jobs -- 100000's running for minutes
- Lots of I/O → usually 100's of image files per job
- Experiments generate lots of data → Tera Bytes per day
- Large number of small user groups from institutions distributed all over Europe
- Many diverse and changing experiments 1000's per year

Test program - spd

- ***Spd is a program to correct 2D images for*** :
  - Spatial distortion : 2d spline curve
  - Flood field : image division
  - Background field : image subtraction



  - One image takes about 17 seconds
  - Additional images take a fraction of a second

  - Simple but typical of many programs used on 2d images :
    - Low on CPU, High in I/O

  - Typical data set is 180 images x 8 i.e. 1.44 GB
  - Typical experiment measures HUNDREDS of data sets

Issues with Image Processing on the grid

- Large overhead of Grid Submission times
  - LFC too slow for many files
  - WMS far too slow for rapid submission of large job numbers
    - Parametric jobs quite helpful
  - All components esp. batch system needs tuning
    - Cream with better response times, lcg-CE has also improved
- Accessing the data
  - Heavy I/O is considerable bottleneck
    - Copying to and from WNs
    - Will be impossible for data sets > 500GB
  - Direct mounts proved unworkable with dCache
    - Improvements promised, however
  - Posix access via GFAL: huge effort in modifying the code seems impractical
- Storage commands perceived as unusable
  - Interested users need heavy assistance, maintenance a problem

Test Case: Image reconstruction

- Image Reconstruction (Tomography)
  - Large datasets: currently 10s of Giga Bytes – Tera Bytes
  - Needs lots of memory: 8GB
  - CPU time can vary from 1 hour to weeks depending on the sophistication of the employed method
  - Efficient data access absolutely necessary

Test Case: Simulations

- Monte Carlo
  - PENELOPE: calculates radiation dose deposition
  - Compute intensive, low data i/o
  - Allows optimization between number of jobs and execution time
  - Runs on our infrastructure

- Ab initio calculations
  - FDNMES: calculates xray absorption spectra
  - Small to modest in data i/o
  - Runs in serial or MPI mode
  - Currently in testing

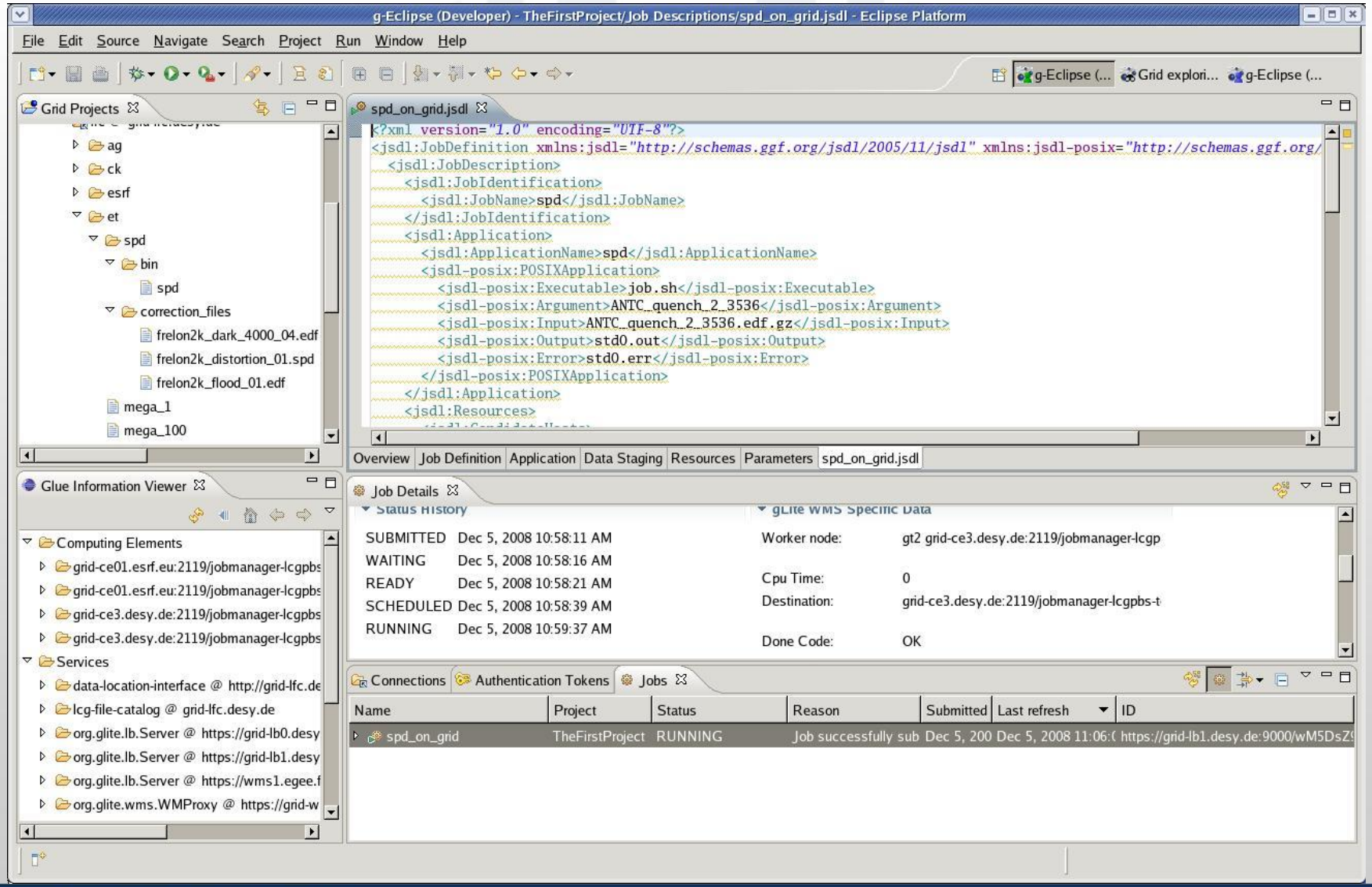
Other possible candidates: GASBOR: protein structure reconstruction

Data Transfers

- Transfer tests performed between
  - ESRF – PSI (CH), DESY (D), APS (US)
- Initially very disappointing transfer rates (100kB-1MB)
- Needs a lot of tuning (tcp_buffer sizes,…)
- Needs a lot of interaction with network people
  - A strict site (security) policy might slow things down considerably
- Once tuned, we get to reasonable rates between partner sites (30MB/s),
  - …but not necessarily to user's laptops (without tuning)
- Overhead with small files (MB)
  - tar datasets helps, but do not make them too large (TB)

→Again no ease of use (yet) (what about Mona Lisa FDT?)

gEclipse


g-Eclipse (Developer) - TheFirstProject/Job Descriptions/spd_on_grid.jsdl - Eclipse Platform
File Edit Source Navigate Search Project Run Window Help
Grid Projects
ag
ck
esrf
et
spd
bin
spd
correction_files
frelon2k_dark_4000_04.edf
frelon2k_distortion_01.spd
frelon2k_flood_01.edf
mega_1
mega_100
spd_on_grid.jsdl
<?xml version="1.0" encoding="UTF-8"?>
<jsdl:JobDefinition xmlns:jsdl="http://schemas.ggf.org/jsdl/2005/11/jsdl" xmlns:jsdl-posix="http://schemas.ggf.org/
<jsdl:JobDescription>
<jsdl:JobIdentification>
<jsdl:JobName>spd</jsdl:JobName>
</jsdl:JobIdentification>
<jsdl:Application>
<jsdl:ApplicationName>spd</jsdl:ApplicationName>
<jsdl-posix:POSIXApplication>
<jsdl-posix:Executable>job.sh</jsdl-posix:Executable>
<jsdl-posix:Argument>ANTC_quench_2_3536</jsdl-posix:Argument>
<jsdl-posix:Input>ANTC_quench_2_3536.edf.gz</jsdl-posix:Input>
<jsdl-posix:Output>std0.out</jsdl-posix:Output>
<jsdl-posix:Error>std0.err</jsdl-posix:Error>
</jsdl-posix:POSIXApplication>
</jsdl:Application>
<jsdl:Resources>
Overview Job Definition Application Data Staging Resources Parameters spd_on_grid.jsdl
Glue Information Viewer
Computing Elements
grid-ce01.esrf.eu:2119/jobmanager-lcgpbs
grid-ce01.esrf.eu:2119/jobmanager-lcgpbs
grid-ce3.desy.de:2119/jobmanager-lcgpbs
grid-ce3.desy.de:2119/jobmanager-lcgpbs
Services
data-location-interface @ http://grid-lfc.de
lcg-file-catalog @ grid-lfc.desy.de
org.glite.lb.Server @ https://grid-lb0.desy
org.glite.lb.Server @ https://grid-lb1.desy
org.glite.lb.Server @ https://wms1.egee.f
org.glite.wms.WMProxy @ https://grid-w
Job Details
Status History
SUBMITTED Dec 5, 2008 10:58:11 AM
WAITING Dec 5, 2008 10:58:16 AM
READY Dec 5, 2008 10:58:21 AM
SCHEDULED Dec 5, 2008 10:58:39 AM
RUNNING Dec 5, 2008 10:59:37 AM
gLite WMS Specific Data
Worker node: gt2 grid-ce3.desy.de:2119/jobmanager-lcgp
Cpu Time: 0
Destination: grid-ce3.desy.de:2119/jobmanager-lcgpbs-t
Done Code: OK
Connections Authentication Tokens Jobs
Name Project Status Reason Submitted Last refresh ID
spd_on_grid TheFirstProject RUNNING Job successfully sub Dec 5, 200 Dec 5, 2008 11:06:(https://grid-lb1.desy.de:9000/wM5DsZ

Community building…

- Very many and diverse user groups
- strong competition between groups of same field

→VO concept does is not perceived as a good fit

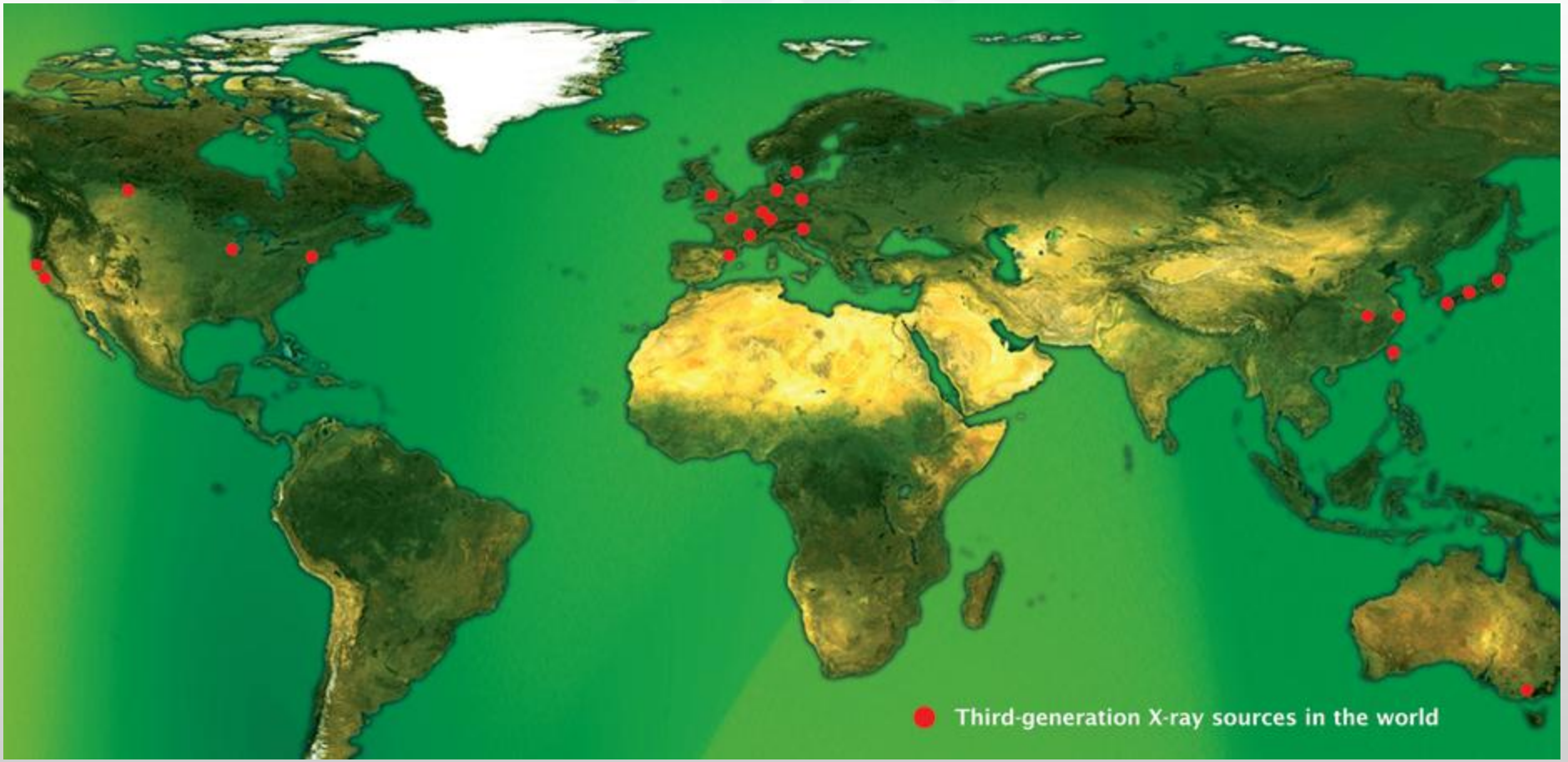
- Secure remote access via long term x509 certificates
- Perceived as too complex by the users
- Not so easily scalable to many 1000s of users from 100s of different institutions
- Transition from existing authentication methods disruptive

→Single sign on via Shibboleth (and related Short Lived Credential Service) now seems more acceptable

Conclusion

- The synchrotron community is facing a serious data problem
  - Individual users do not yet see it as their problem, however
- EGEE Grid is not an obvious solution to this problem
  - In addition a lot of good will gets lost due to its user friendliness
  - For certain simulations with not so intensive I/O suitable
- A better return on investments could come from optimized local clusters with high performance file and batch systems
  - Also more from many-core, GPUs, FPGAs
- A combination with grid technology of the type National Analysis Facility at DESY an interesting option

Tomorrow's XRAY Grid!?



Skepticism rules !