LHC Computing Grid Project - LCG

The LHC Computing Grid
Looking towards 2007

2nd LCG Workshop

Les Robertson – LCG Project Leader
CERN – European Organization for Nuclear Research
Geneva, Switzerland

les.robertson@cern.ch

pdate: 23/03/2004

Where are we?
Open issues
Where we need to go

- Applications
- Fabric & Networking
- Grid Deployment
- Middleware & ARDA
- Summary & Conclusions

- Applications
- Fabric & Networking
- Grid Deployment
- Middleware & ARDA
- Summary & Conclusions

Applications Area - Achievements

- POOL persistency for event data delivered and integrated in three experiments
  - successful production usage with millions of events
  - expected to be production event store in 2004 DCs for three experiments
- Robust LCG dictionary
- Comprehensive software development infrastructure meeting AA needs and spreading well beyond AA: experiments, other LCG areas, EGEE, other projects (CLHEP, probably Geant4)
- Important steps in simulation physics validation
  - first round of Geant4 em and hadronic physics validation completed ("as good as or better than Geant3")
  - simulation physics requirements of the four experiments documented
  - good collaboration on validation work with both Geant4 and FLUKA

- Generator library GENSER developed, populated, and is being evaluated/adopted by experiments
- Strong CERN Geant4 program squarely focused on LHC priorities
  - successfully deployed in production in CMS and pre-production in ATLAS
- Successful, and deepening, collaboration with ROOT
  - data store technology in POOL
  - analysis environment used either directly or via interfaces (pyROOT, pyLCGDict, AIDA ROOT)

Highlights for the next year

- Common conditions DB
- Common math library development
- Closer relations with ROOT –
  - Aiming for convergence with ROOT on mathlib and dictionary
  - ROOT will use LCG AA software components, as well as vice versa
- Physicist-level event collections: collaboration AA/ROOT/ARDA
- POOL and Geant4 in Data Challenge production in CMS, ATLAS and LHCb
- Experiment adoption and validation will continue to be the measure of success

Longer term

- Current development program should be completed in 12-18 months

  Is there more common work to be done?

- Thereafter emphasis will be on maintenance, and supporting the scale and complexity needed for LHC data taking

- We need to understand this in more detail this year, establish scope, objectives, resources needed for Phase 2 and beyond

  .. and identify where these will come from

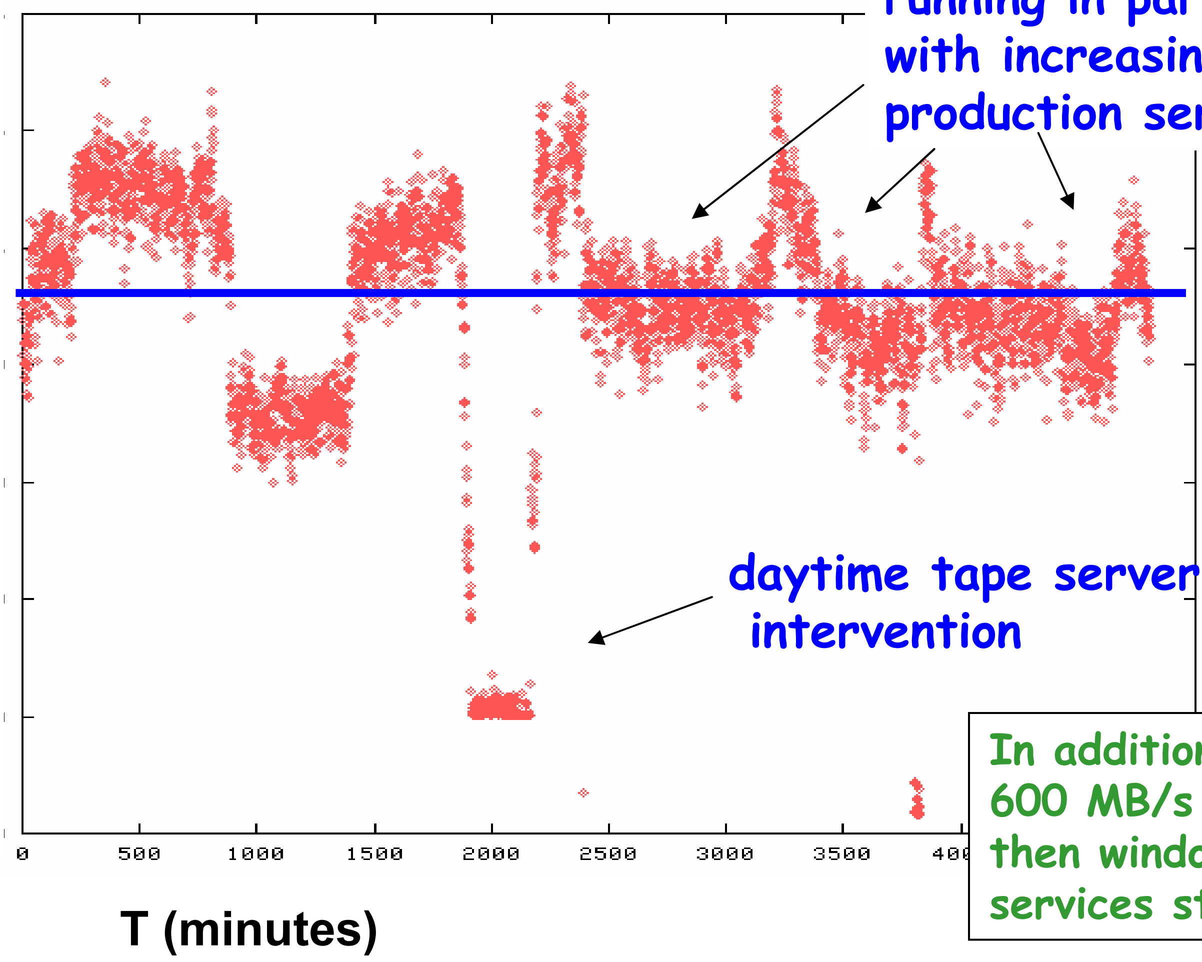
- Applications
- Fabric & Networking
- Grid Deployment
- Middleware & ARDA
- Summary & Conclusions

Fabric – Preparations for Phase 2

- **High performance data distribution**
  - **Data exchange between mass storage systems over a Wide Area Network**
  - **FNAL – CMS – CERN project starting now**

- **High performance data recording**
  - **ALICE – Mass Storage Data Challenges at CERN**
    - □ **2002 – target 200 MB/sec sustained – achieved 280 MB/s**
    - □ **2003 – target 300 MB/sec – achieved 280 MB/sec**
    - □ **2004 → 450 MB/s, 2005 → 700 MB/s -- --**
      -- -- CDR in 2008 → 1.2 GB/s
  - **File system → network → tape storage → 1 GB/s in April 2003**

1 Gbyte/s Computing Data Challenge → Observed rates



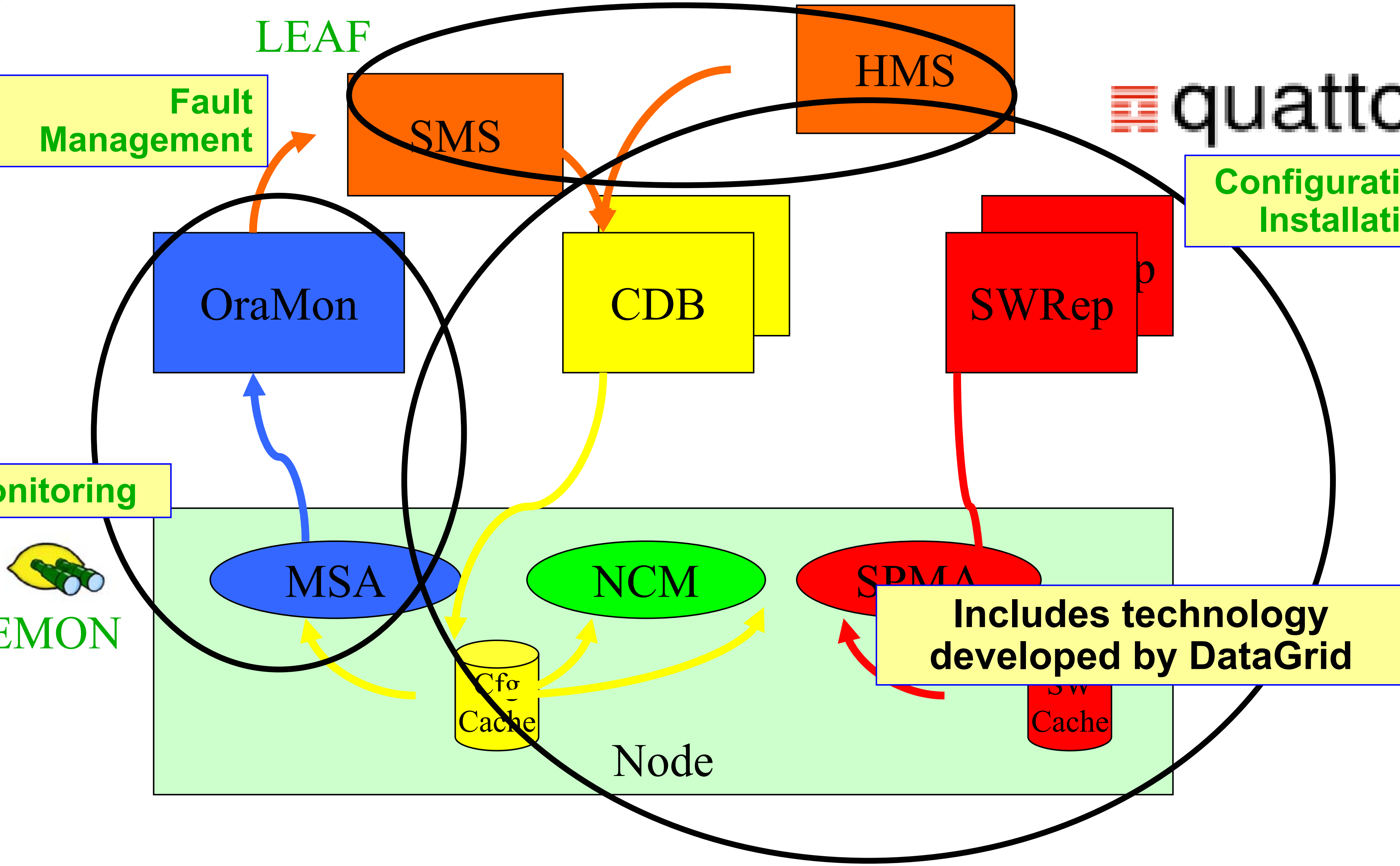
running in parallel with increasing production service

920 MB/s average over a period of 3 days with an 8 hour period of 1.1GB/s and peaks of 1.2GB/s

daytime tape server intervention

In addition :
600 MB/s into CASTOR for 12 hours , then window of opportunity closed services started

Fabric Automation at CERN



LEAF

Fault Management

SMS

HMS

quatto

Configurati
Installati

OraMon

CDB

SWRep

onitoring

EMON

MSA

NCM

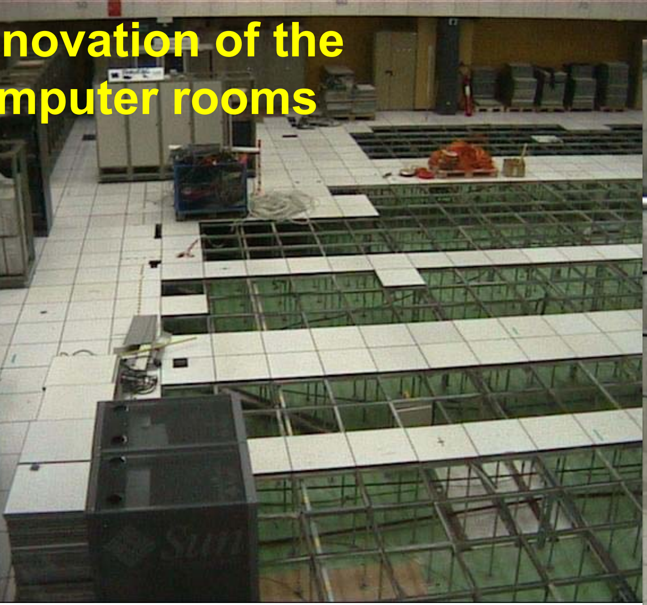
SPMA

Cfg Cache

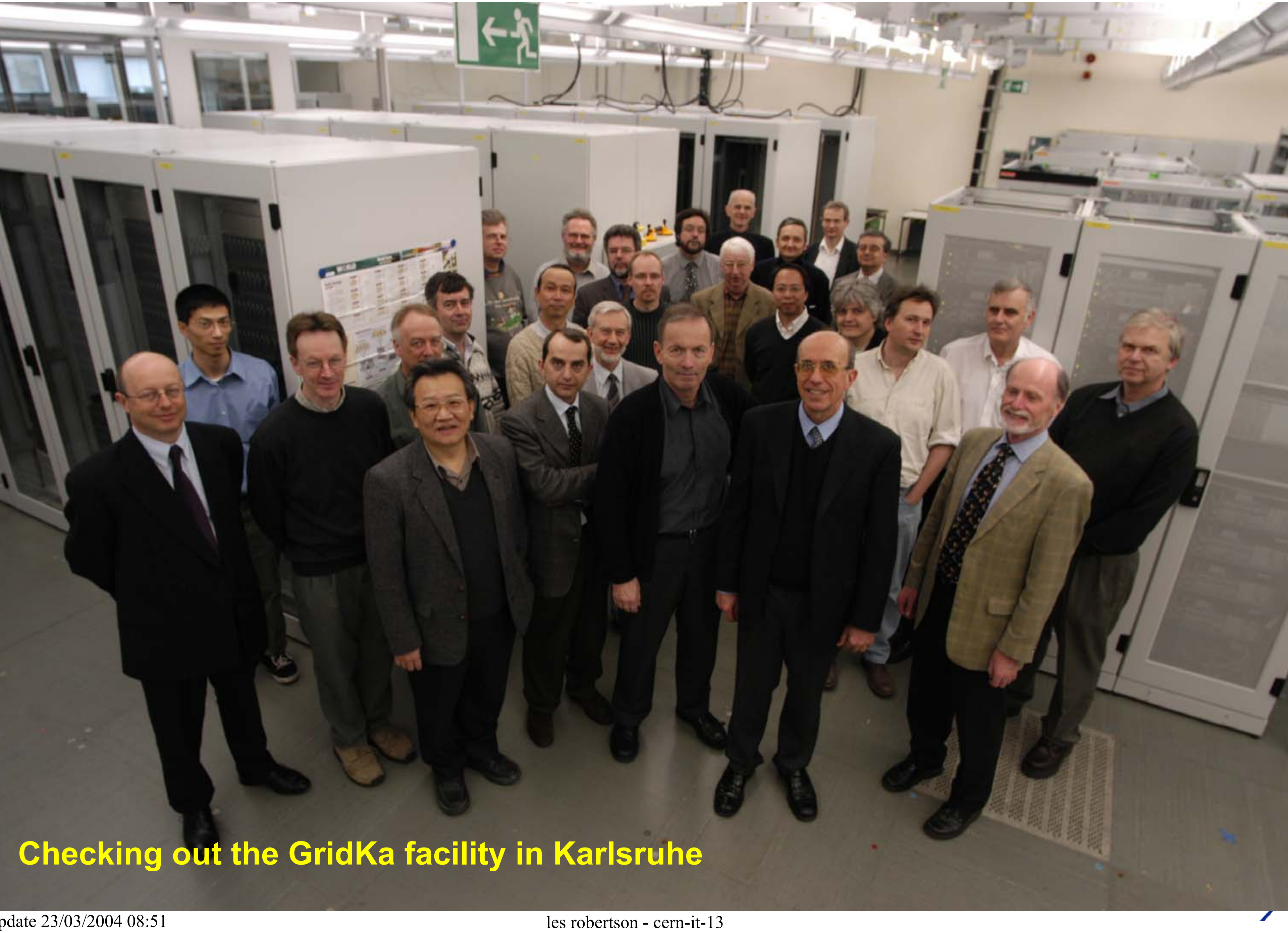
SW Cache

Node

Includes technology developed by DataGrid



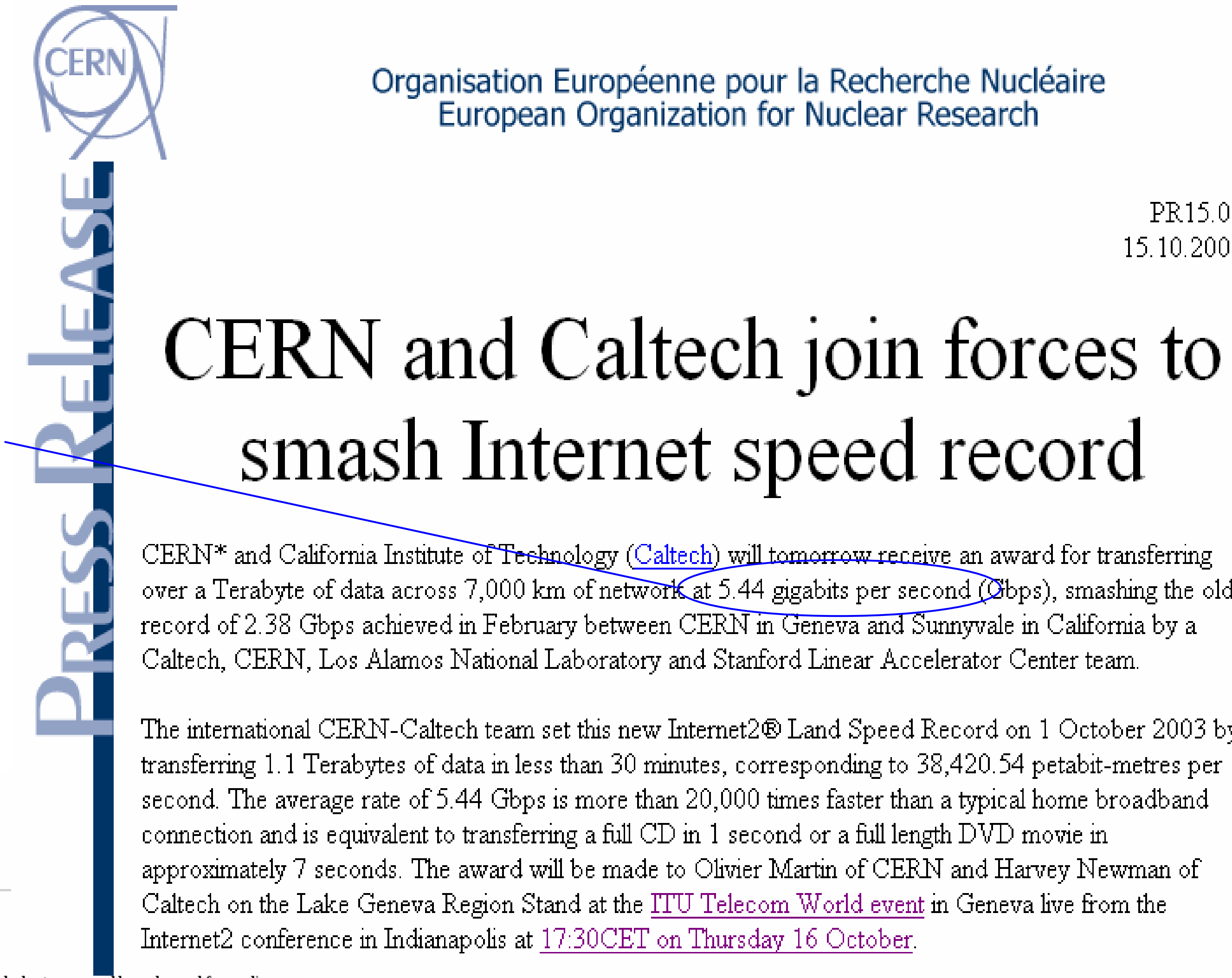
Preparing the Tier 0+1 computer centre



Checking out the GridKa facility in Karlsruhe

WAN connectivity

5.44 Gbps
1.1 TB in 30 mins



Organisation Européenne pour la Recherche Nucléaire
European Organization for Nuclear Research

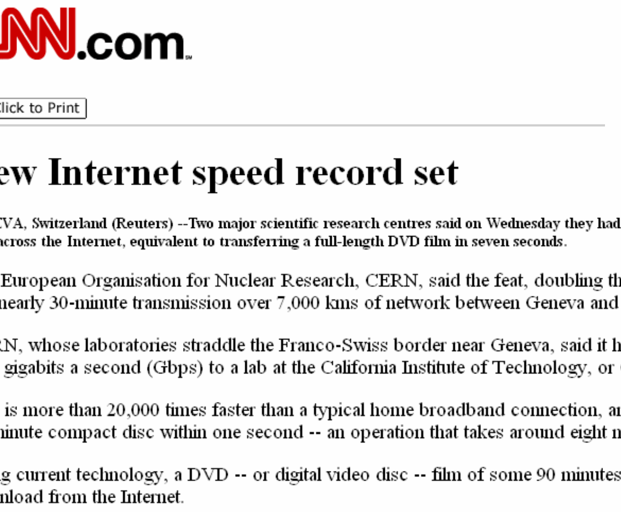
PRESS RELEASE

PR15.0
15.10.200

CERN and Caltech join forces to smash Internet speed record

CERN* and California Institute of Technology (Caltech) will tomorrow receive an award for transferring over a Terabyte of data across 7,000 km of network at 5.44 gigabits per second (Gbps), smashing the old record of 2.38 Gbps achieved in February between CERN in Geneva and Sunnyvale in California by a Caltech, CERN, Los Alamos National Laboratory and Stanford Linear Accelerator Center team.

The international CERN-Caltech team set this new Internet2® Land Speed Record on 1 October 2003 by transferring 1.1 Terabytes of data in less than 30 minutes, corresponding to 38,420.54 petabit-metres per second. The average rate of 5.44 Gbps is more than 20,000 times faster than a typical home broadband connection and is equivalent to transferring a full CD in 1 second or a full length DVD movie in approximately 7 seconds. The award will be made to Olivier Martin of CERN and Harvey Newman of Caltech on the Lake Geneva Region Stand at the ITU Telecom World event in Geneva live from the Internet2 conference in Indianapolis at 17:30CET on Thursday 16 October.



NN.com

lick to Print

ew Internet speed record set

VA, Switzerland (Reuters) --Two major scientific research centres said on Wednesday they had set a new world speed record for sending across the Internet, equivalent to transferring a full-length DVD film in seven seconds.

European Organisation for Nuclear Research, CERN, said the feat, doubling the previous top speed, was achieved nearly 30-minute transmission over 7,000 kms of network between Geneva and a partner body in California.

N, whose laboratories straddle the Franco-Swiss border near Geneva, said it had sent 1.1 Terabytes of data at gigabits a second (Gbps) to a lab at the California Institute of Technology, or Caltech, on October 1.

is more than 20,000 times faster than a typical home broadband connection, and is also equivalent to transferring a inute compact disc within one second -- an operation that takes around eight minutes on standard broadband.

g current technology, a DVD -- or digital video disc -- film of some 90 minutes length takes some 15 minutes to nload from the Internet.

Network Requirements for 2007→

- The rapid improvements over the past ten years in wide area network bandwidth and costs are key enablers of data intensive grids

- Current estimates for the effective bandwidth required in 2007 at Tier-1s is 10 Gbps, with 40 Gbps at the Tier-0, rising to perhaps 100 Gbps by the end of the decade.

- These estimates are of course highly dependent on the experiment computing models that are being developed now.

- And conversely the costs and performance of wide area networking will enable or constrain the evolution of the LHC Grid and the computing models.

- This extends out through Tier-2 → Tier-3 .....
  where technology may not be the only issue

1st International Grid Networking Workshop – GNEW2004

co-organized by CERN/DataTAG, DANTE, ESnet, Internet2, TERENA

15-16 March 2004, CERN, Geneva

DataTAG is a project funded by the European Commission
under contract IST-2001-32459

The Network is not Infinite or Free

- **Whatever is provided, we need to deal with many basic impediments to high performance end-end.**
  - Eliminate firewall performance issues.
  - Use of optimised stacks for WAN transfers.
  - End-End performance issues – applications, internal busses, disks, campus networks ….
- **Network community is starting to propose hybrid networks for high-performance needs**
  - General purpose packet switched network for most uses.
  - Circuit switched infrastructure or statically provisioned “community network” for specialised used.
- **We need a realistic approach to set up a true production high performance network infrastructure in 2006**
  - It must NOT depend on hope and promises but a realisation that **good ideas take a long time to be implemented** as a true production system.
  - We need to **start** setting this up **now** – with a pragmatic approach

Main Points – Fabrics and Networks

- Automation, cost containment in the common facility at CERN is on schedule

- Performance – Data recording, reconstruction, distribution of data
  - The basic technology – performance, costs - looks to be on track
    - – but we need to put it all together
    - – need more computing challenges that test real scenarios and include the full hierarchy Tier-0..→..Tier-3
    - – including LAN and WAN
- Wide Area Networking
  - We need to complete **this year** the planning for the **service** for Tier-1s and Tier-0

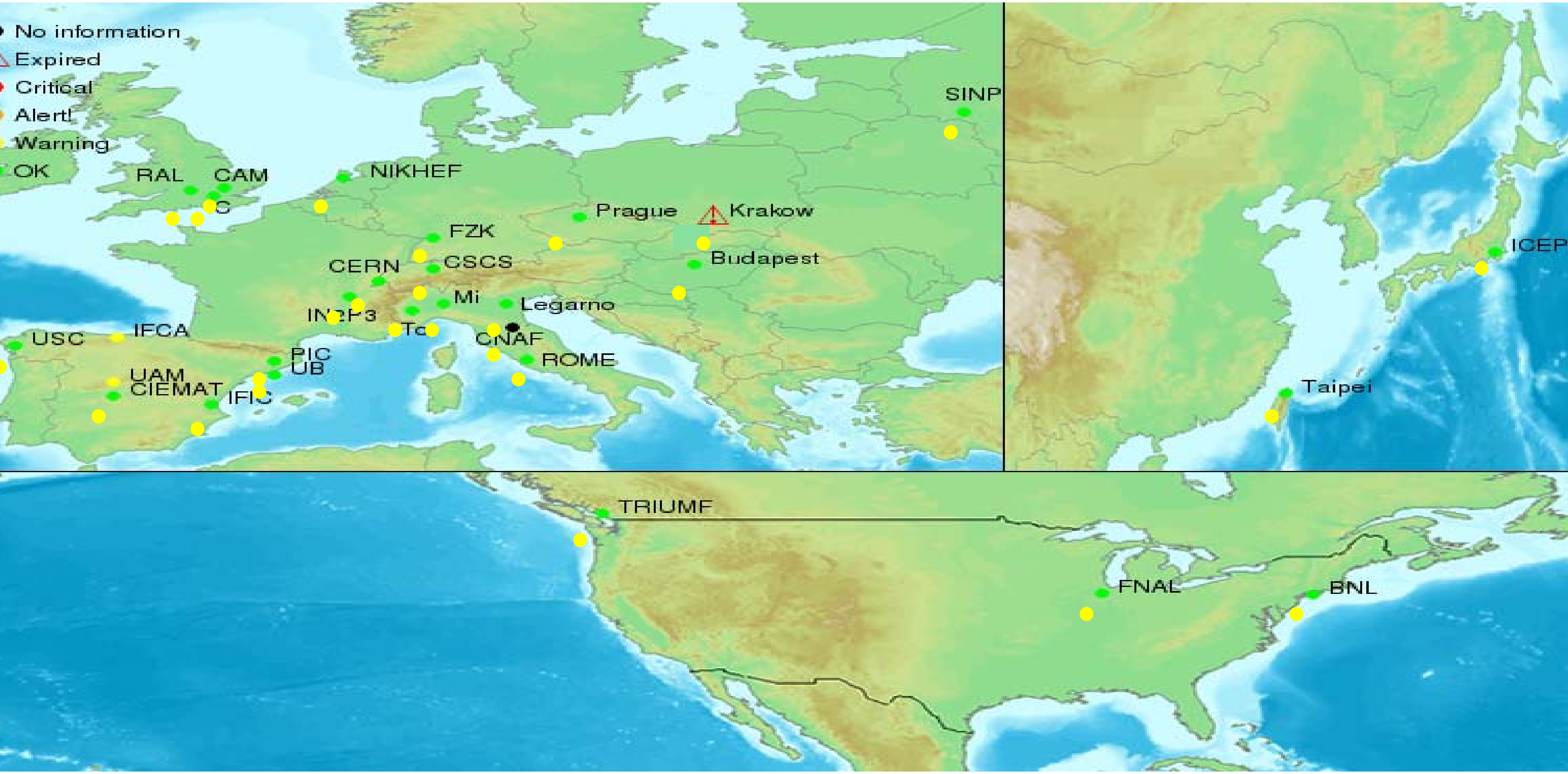
- By this time next year the technology choices and the costs must be clear

- Applications
- Fabric & Networking
- Grid Deployment
- Middleware & ARDA
- Summary & Conclusions

LHC Grid Deployment

LCG-1 –

- Service opened in September, ~30 sites at year-end
- Significant use by CMS-Italy in last days of 2003 for production
- Intermittent use by small groups
- Missed opportunity for preparing for data challenges?



LCG

LCG for the Data Challenges

- **Migrating to an upgraded version of the grid software (LCG-2)**
  - Target is the 2004 data challenges
  - Over 1,800 processors available now at core sites - migration of remaining LCG-1 sites has started

  - Data challenges have started this month – ALICE (PDC3), CMS (DC04)
  - LHCb and ATLAS – start in May

  - NIKHEF to coordinate VO support for D0
  - Hewlett Packard to provide “Tier 2-like” services for LCG, initially in Puerto Rico.

Grid Operations Centre at RAL

CCLRC

User Support Centre at FZK

GridKa

Planning for a second operations & support centre in Taipei

ASCC ACADEMIA Sinica Computing Centre

中央研究院計算中心

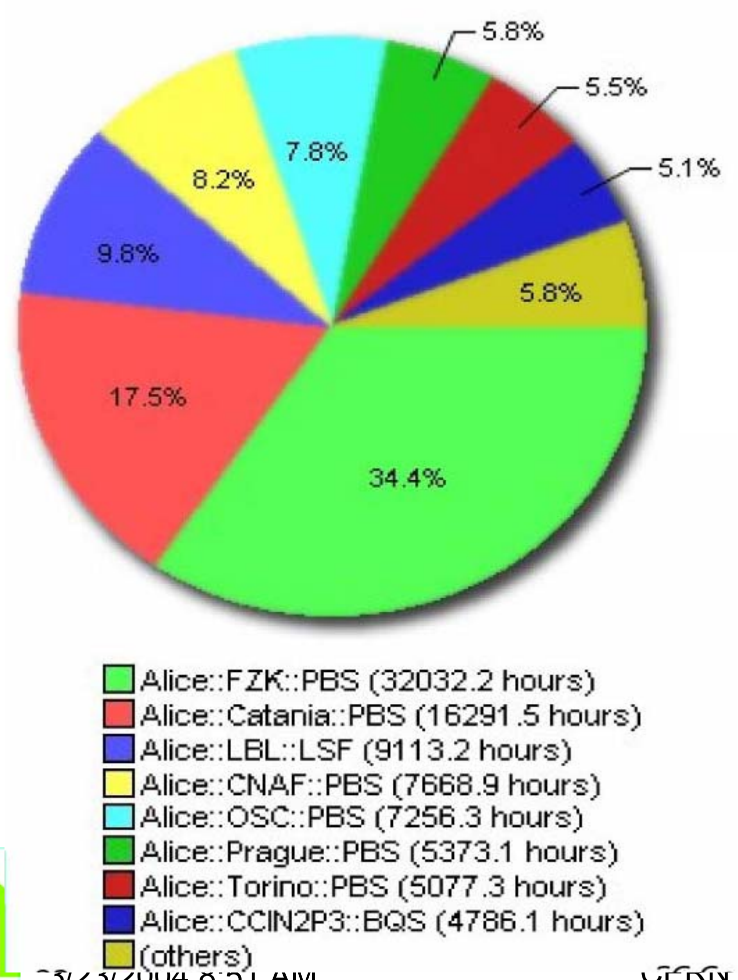
LCG-2 Support Agreements

VDT (US tools)	VDT team at Wisconsin (NSF funding)
DataGrid resource broker	INFN/CERN
DataGrid replica management	CERN
DataGrid relational information system (RGMA)	RAL
GridIce monitoring tools	INFN
GLUE schema	INFN
VOMS	INFN
dCache storage manager	DESY
CASTOR storage manager	CNAF/PIC/CERN
Security & VO policies and procedures	Agreed by all of the GDB countries

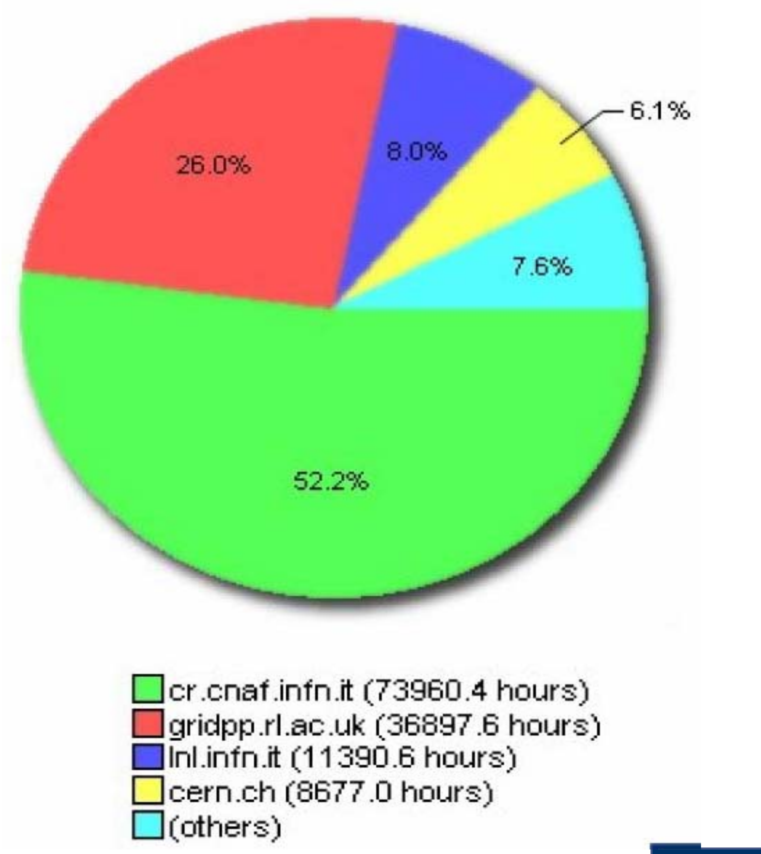
Sun, 14 Mar 2004: PDC Record - 1010 Jobs - 2.0 THz*s

Data Challenge statistics

PU speed x time non LCG sites (93,001 GHz*s)



CPU speed x time LCG sites (141,760 GHz*s



Data Challenge DC04 Underway (Mar 1 - Apr 30)

- 70M MC events (20M with G4) produced in pre-challeng
  - Classic production Centers and LCG and US/GRID3 heavily used

- Challenge:
  - (Not a "CPU" challenge, but a full-chain demonstration)
  - Reaching a sustained 25Hz reconstruction rate in the Tier-0 farm (25% of the target conditions for LHC startup)
    - (This, however, is a lot of CPU, ~500)
  - Use of CMS and LCG software to record the DST, catalog the data and Meta-Data
  - Distribution of the reconstructed data to six Tier-1 centers using available GRID and other tools
  - Close to real-time reprocessing of that data at some of the Tier-1 centers
  - Production of new data-sets at the T1 with their subsequent distribution to Tier-2 centers for analysis purposes
  - Monitoring and archiving of performance criteria of the ensembl of activities for debugging and post-mortem analysis

LCG-2 components in DC04

RLS (Replica Location Service)

- Many clients:
  - RLS Publishing Agent: converting the XML catalogue of Tier-0 job into the RLS
  - Configuration Agent: querying the RLS metadata to assign files to Tier-1
  - Export buffer Agents: inserting/deleting the PFN for the location in the EB
  - Tier-1 Agents: inserting PFN for the destination location, in some cases dumping the RLS into local MySQL POOL catalogue
- Scalability problems…
  - Understand bottlenecks, e.g. use C++ API instead of the (java) command line
  - Reduce the load on RLS
  - Mirror the RLS
    - Mirror at CNAF is ready since last week. But for not yet in use

Data transfer between LCG-2 Storage Elements

- Export Buffer at Tier-0 with disk based SE
  - Production system delivered by IT end last week with ~ 1 TB of disk
    - Before were using a system provided by EIS team
    - added CPU and 2TB disk space today
  - Serving transfers to CASTOR SEs at PIC and CNAF via the Replica Manager
    - Also replicating files from CNAF to Legnaro for muon streams

Services and SW installation

- Dedicated information indexes at CERN supported by LCG
  - CMS may add its own resources and remove problematic sites
- Dedicated Resource Broker at CERN supported by LCG
- Virtual Organization tools are the official LCG-2 ones
- Dedicated GridICE monitoring server at CNAF:
  - monitor resources registered in the CMS-LCG information index
  - active on all service machines (CE, SE, RB, etc…)
  - WN monitoring on at CNAF/PIC/Legnaro
- CMS Software installation:
  - With new LCG-2 tools CMS software manager can:
    - install the software in an LCG site (with a shared area between CE and WNs)
    - advertise in the Information system what has been installed
  - Working on two kind of CMS software distribution:
    - DAR (for production activities)
    - CMSI-based tool to install RPM's (for analysis activities)

Grid2003 Demonstrator

id2003 Project follow-on of US Atlas and US CMS Grid testbeds

- ➔ Demonstration for SC2003 and U.S. funding agencies: performance demonstrator for functional multi-VO Grid
- ➔ Collaboration of US LHC and Grid projects, labs and universities Including both U.S. Tier-1 and all U.S. Tier-2 centers

id2003 approach

- ➔ experiment projects/VOs (US CMS, US Atlas and others) bring their grid-ified applications into multi-VO Grid3 environment
- ➔ Grid2003 team works with sites to provide basic Grid services:
  - processing and data transfer, software packaging/deployment, monitoring, information providers, VO/authentication management, basic policies
  - simple/non-intrusive installation based on VDT and EDG middleware
  - iVDGL iGOC cross-VO operations support, including trouble tickets

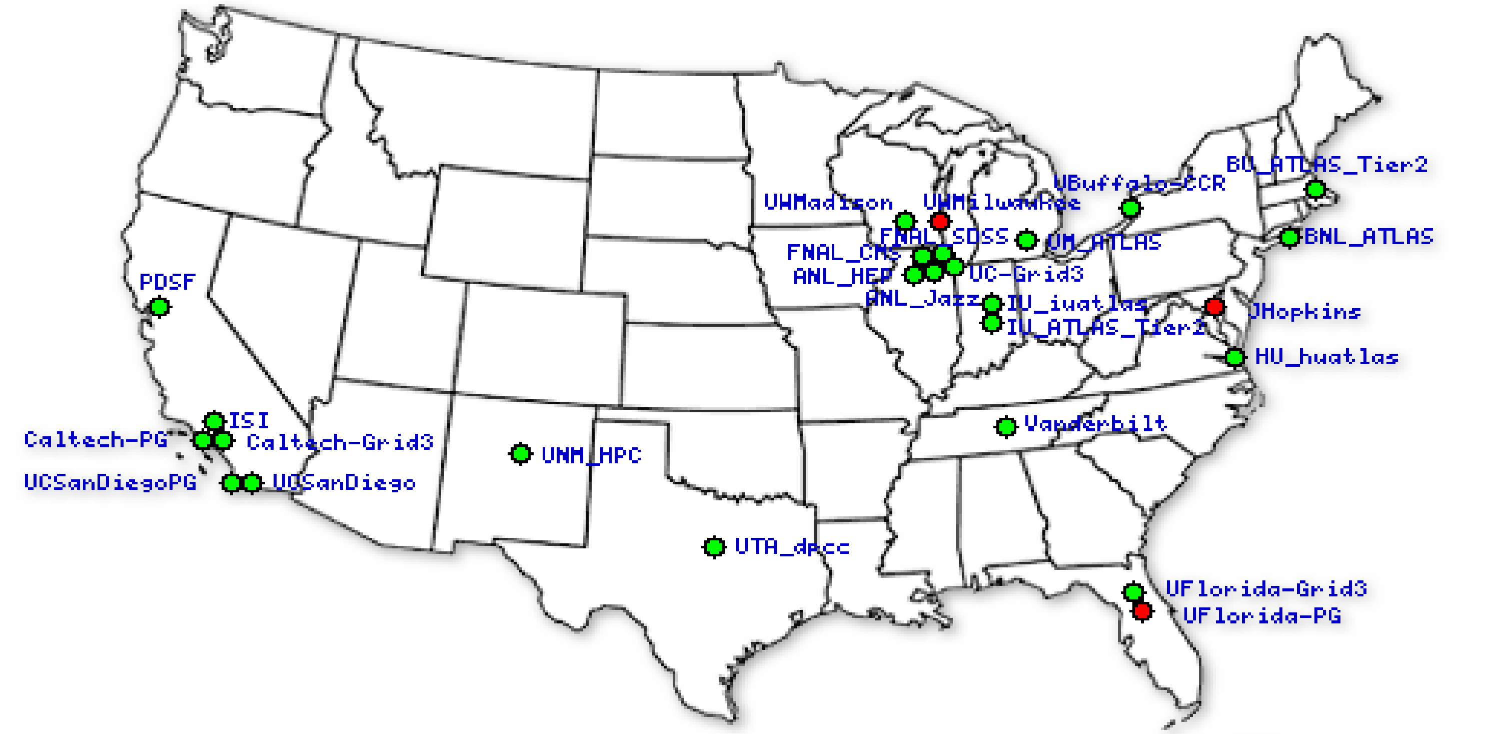
sites, 2800 CPUs, running fairly stable since SC2003 (Nov 2003)

- ➔ e.g., 13M CMS full detector simulation events produced on Grid3 -- and counting
- ➔ represents about 100 processor years of computing

Toward the US Open Science Grid

ilding partnerships on US Grid infrastructure for LHC and other sciences

- ➔ LHC application driving this effort, Grid3 is a great initial step
- ➔ Federate US resources with the LCG, the EGEE and other national and international Grids



S LHC experiment projects, regional centers, universities and Grid projects

rmulated a roadmap towards the “Open Science Grid”

Interoperation of US Grids with the LCG

US Atlas and US CMS working on interoperability of LCG and US Grid

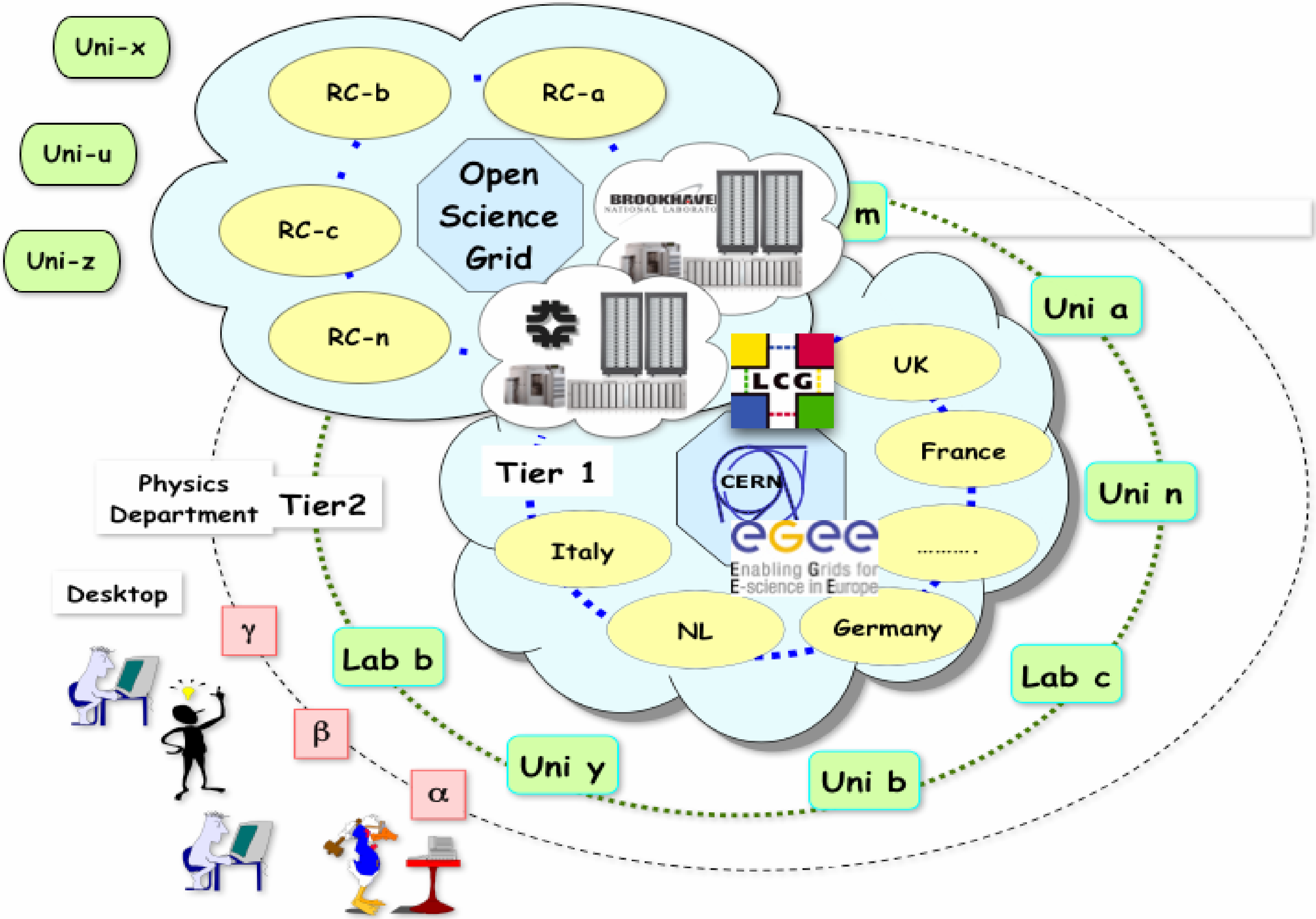
- ➔ First steps already achieved
  - On storage service, middleware, VO management and application level
- ➔ Atlas DC2 application running across LCG, NorduGrid, US Grid3
- ➔ CMS DC04 data transfers and management of dataset replicas between storage services on LCG and US Grid3 sites

Next step: US Tier-1 centers to federate US resources with LCG service

- ➔ Realistic near term goals:
  - Fermilab Grid installation available to LCG resource broker through existing LCG-2 installation at Fermilab Tier-1
  - Reconciling LCG and US Grid VO management (VOMS)
- ➔ Next steps this year
  - Managed storage across Grids
  - Include access to US Tier-2 centers and other US Grid sites from LCG

Emerging ARDA approach to middleware and end-to-end systems will help in facilitating this

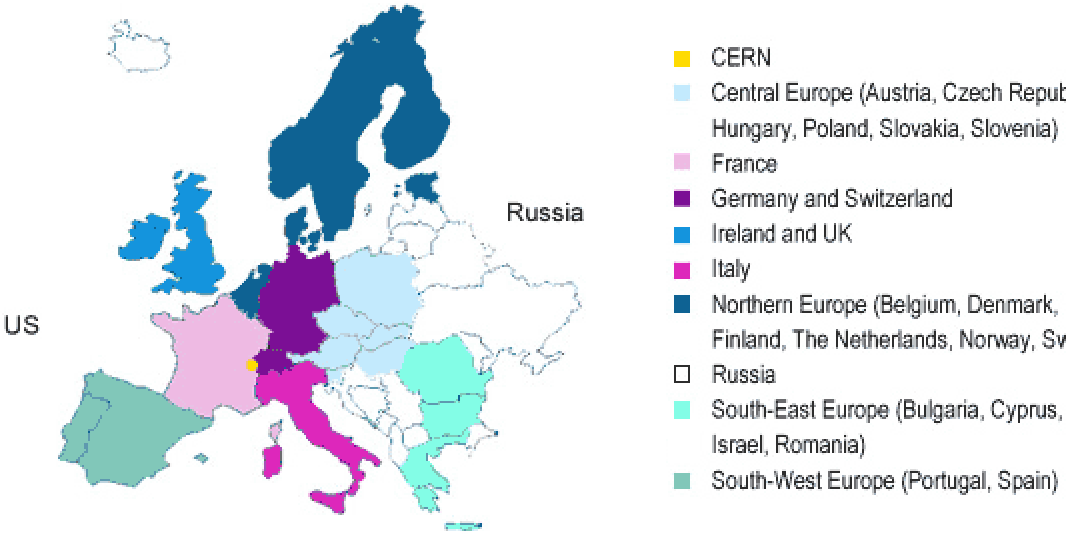
Federating Worldwide Resources for the LHC



Relation to EGEE Project
Enabling Grids for E-Science in Europe

- EU funding for EGEE starts in April
  - 70 partners in Europe, Russia, Middle East, US
  - Major overlap with LCG sites
- The EGEE grid will grow out of LCG
  - Shared infrastructure and management
  - Starts with same grid middleware – LCG-2

egee
Enabling Grids for
E-science in Europe


US
Russia
CERN
Central Europe (Austria, Czech Repub
Hungary, Poland, Slovakia, Slovenia)
France
Germany and Switzerland
Ireland and UK
Italy
Northern Europe (Belgium, Denmark,
Finland, The Netherlands, Norway, Sv
Russia
South-East Europe (Bulgaria, Cyprus,
Israel, Romania)
South-West Europe (Portugal, Spain)

Grid Deployment Coordination and Management

- Grid Deployment Board
  - National members (regional centre managers)
  - Experiment members
- Policies, agreements, decisions and standards
- Definition and schedule of LCG releases
- Coordinates and plans grid resources for physics and computing data challenges
- Security group

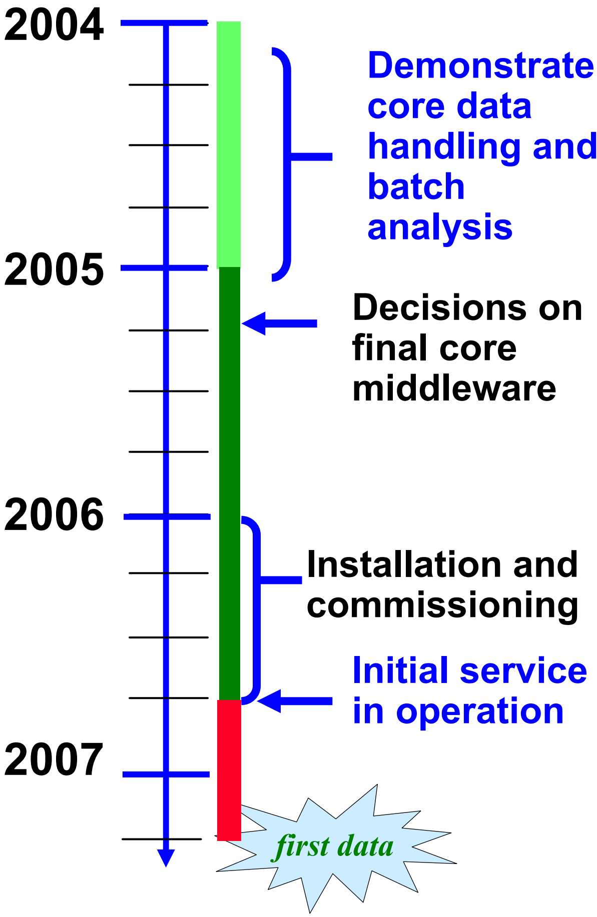
- How to extend or adapt this to include EGEE, OSG, other contributors (e.g. HP), and other VOs?
  - Discussions going on with EGEE
  - OSG-LCG - GDB group set up to define the issues
    Meeting next month at BNL to discuss inter-operation

Preparing for 2007

- 2003 – has demonstrated event production
- In 2004 we must show that we can also handle the data – even if the computing model is very simple

  -- This is a key goal of the 2004 Data Challenges

- Target for end of this year –
  - Basic model demonstrated using current grid middleware
  - All Tier-1s and ~25% of Tier-2s operating a reliable service
  - Validate security model, understand storage model
  - Clear idea of the performance, scaling, and management issues


2004
2005
2006
2007
Demonstrate core data handling and batch analysis
Decisions on final core middleware
Installation and commissioning
Initial service in operation
first data

Main Points – Grid Deployment

- This year the data challenges must show that we can handle data

- Still need to work on the collaboration between regional centres
  - shared planning and priorities
  - the experiments must see a single service
  - effective operation – feeling the pulse of the data challenges -- the GOC has a key role here

- Merging with EGEE will be a challenge
- And we must also understand what *federating* with OSG means

- Applications
- Fabric & Networking
- Grid Deployment
- **Middleware & ARDA**
- Summary & Conclusions

New Middleware Development

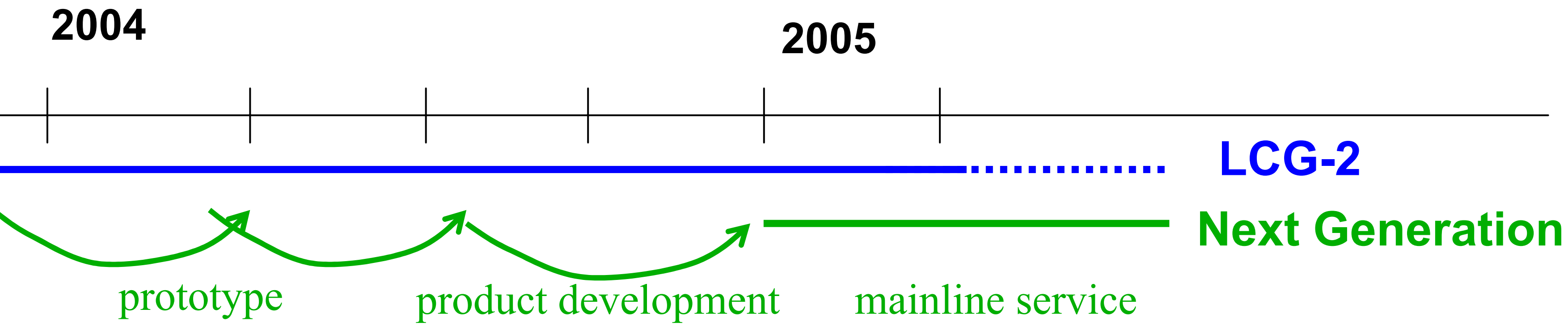
- Exploiting and integrating experience, expertise and technology from DataGrid (EU), the Virtual Data Toolkit (US), AliEn (ALICE), NorduGrid
- Joint EGEE-VDT design team
- Focus on HEP requirements + bio-medical
- Strongly coupled to ARDA - a new LHC distributed analysis project

- We need to see an early prototype soon,
  involving HEP applications and users

  and a "usable system" with a year - stability, performance as important as functionality

- By this time next year we will have to start making decisions about the middleware to be used in 2007

LCG-2 and Next Generation Middleware



2004

2005

LCG-2

Next Generation

prototype

product development

mainline service

- LCG-2 will be the main service for the 2004 data challenges
- This will provide essential experience on operating and managing a global grid service – and will be supported and **developed**
- Target is to establish a base (fallback) solution for early LHC years

- LCG-2 will be maintained until the new generation has proven itself

Expectation and Reality

- The past two years has taught us that grid computing is much harder than we thought

- We knew that -
  - the basic technology was immature, there was limited practical experience
  - developing software is easier than delivering it as part of a production service
  - distributed systems are difficult to design and to test
  - independent computing centres have to learn how to collaborate

- But we underestimated the costs that come with the liberal funding available for *GRIDs*

  → the size and complexity of the grid community

  → the constraints and commitments of non-HEP funding

  → the many different agendas - national, regional, personal

  → the *HYPE* → the exaggerated expectations

Aiming at the right Goal

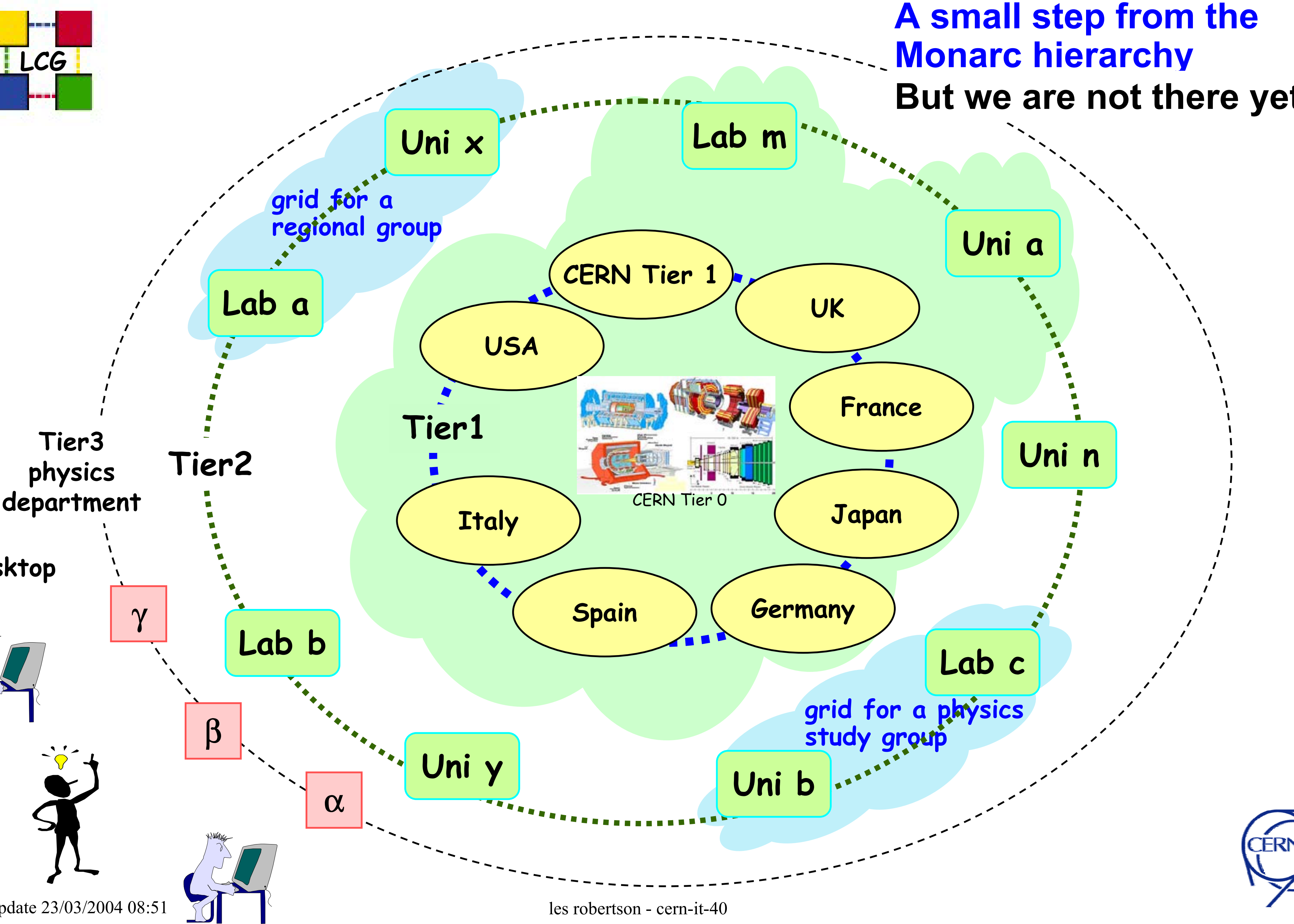
- Our goal is straightforward - to set up a computing environment for LHC

- The grid is only a means to that end

- We have to set our priorities by the practical needs of the experiments
  - Focus on data challenges
  - Evolve in stages a workable computing model
  - With the experience of this year's data challenges we must set realistic goals for 2007

That is what the middleware must address

The *Cloud* from 2001
A small step from the Monarc hierarchy
But we are not there yet



ARDA – A Realisation of Distributed Analysis

ARDA working group recommendations

EGEE-VDT Middlew

- New service decomposition
  - Strong influence of Alien system
- Role of experience, existing technology…
  - Web service framework

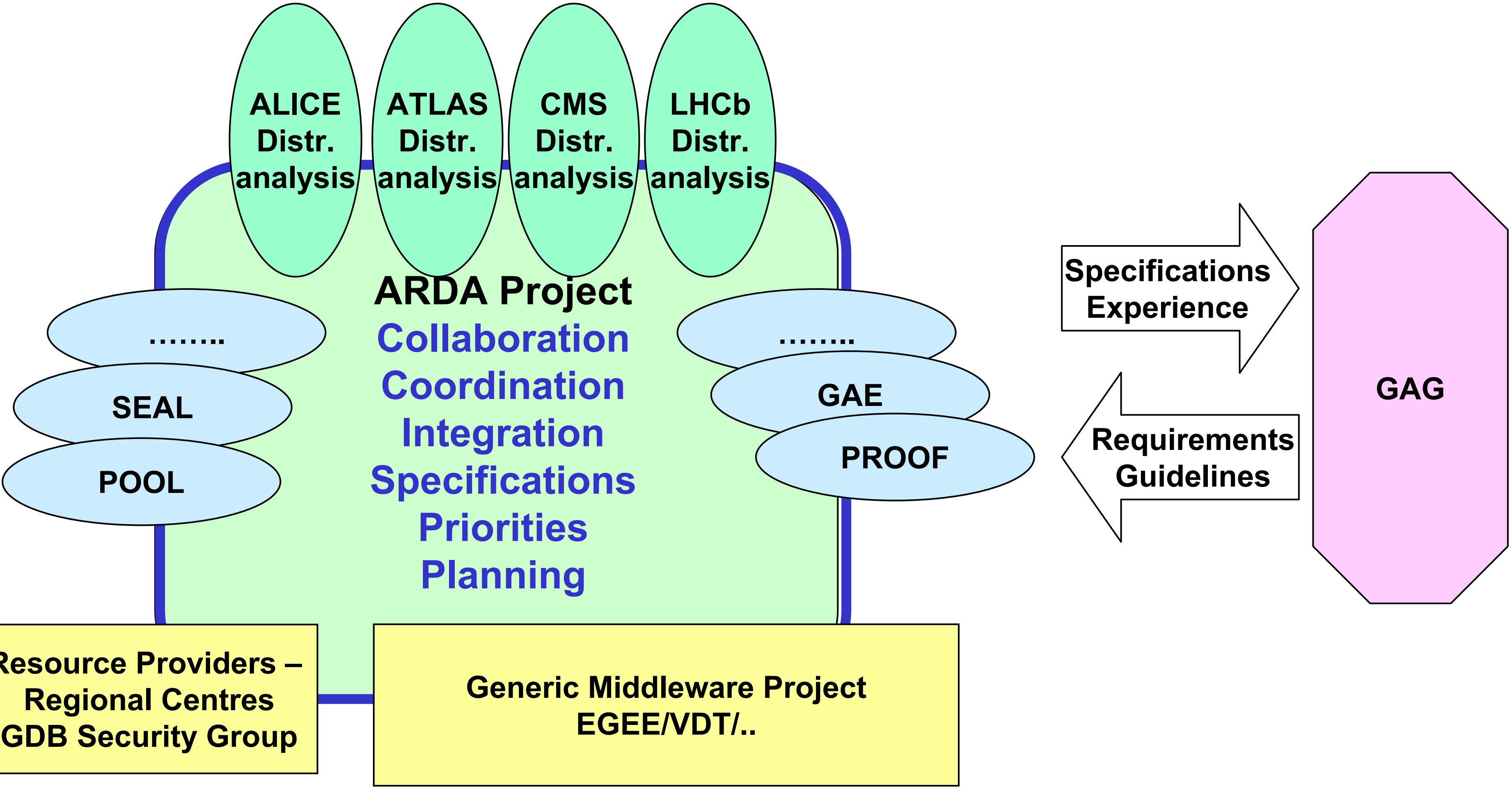
ARDA project

- Interfacing to existing middleware to enable their use in the experiment frameworks
- Early deployment of (a series of) prototypes to ensure functionality and coherence

ARDA - End-to-end prototypes

- Provide a fast feedback to the EGEE MW development team
  - Avoid uncoordinated evolution
  - Coherence between users expectations and final product
- Guarantee the experiments are ready to benefit from the new MW as soon it becomes available
  - Expose the experiments (and the community in charge of the deployment) to the current evolution of the whole system, to be prepared to use it in the best and quickest way
- Move forward towards new-generation real systems
  - Prototypes should be exercised with realistic workload and conditions (experiments absolutely required for that!)
    - No academic exercise or synthetic demonstrations

The ARDA Project



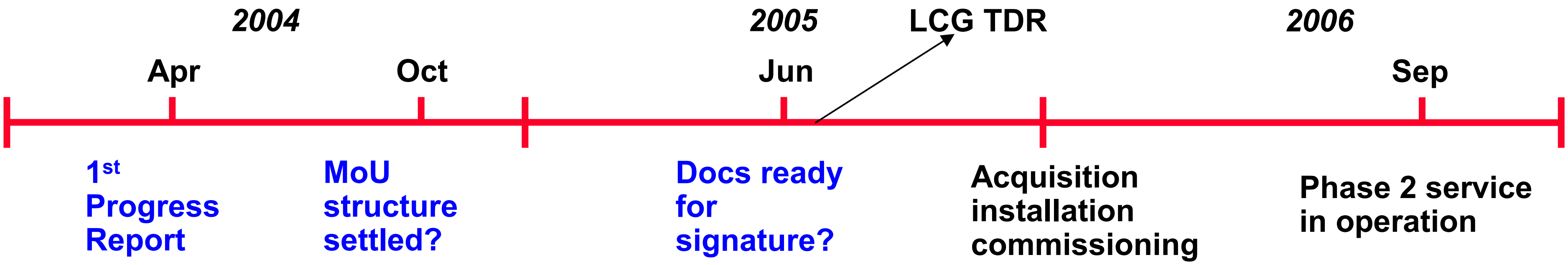
Main Points – Middleware and ARDA

- **As we start to plan the second generation of middleware**
  - Concentrate on prototyping, rapid development cycle, and integration with applications
- **ARDA –**
  - Specifically targeted at the “new middleware”
  - End-to-end distributed analysis – from prototypes → services
- **The complexity generated by large projects, orthogonal funding will be a major challenge for the new middleware**
- **Until the new middleware has proved itself – solid support must be maintained for the current `tools**

- Applications
- Fabric & Networking
- Grid Deployment
- Middleware & ARDA
- Summary & Conclusions

Assembling Funding for the Phase 2 Grid

- **Memorandum of Understanding for Phase 2 and beyond**
  - **Task Force established – some of the funding agencies + all experiments**
  - **Covering host lab, Tier 0 and Tier 1s, maybe also Tier 2s**
  - **A re-assessment of the requirements for Tier 0, Tier 1, Tier 2 being prepared – four experiments together**
  - **First report to the Computing Resource Review Board in April**



2004

Apr

1st Progress Report

Oct

MoU structure settled?

2005

Jun

Docs ready for signature?

LCG TDR

Acquisition installation commissioning

2006

Sep

Phase 2 service in operation

- **The Phase 2 services for Tier-0 and Tier-1 must be in operation by September 2006**
- **Acquisition process for the scale of computing required is very long in some centres -- and is starting now at CERN**

→ **Tier-0, Tier-1 centres will have to do their planning before the MoU is signed**

Summary of where we are

- LCG has been established as a collaboration –
  - experiments, developers and regional centres
  - working organisation in place – at many different levels
  - Computing Resource Review Board now drafting the MoU
  - reviewed by the LHCC - “5th experiment”
- Scope of Phase 1 of the project defined – and products and services are being delivered
- Major LCG applications now in use by experiments
- Grid - agreements reached on security, registration, accounting, operation, middleware
- Improving coupling with grid projects – but more to be done
- Demonstrated that grid technology is good for simulation
- Now starting to tackle data movement and distributed analysis
- Good progress on basic technologies – farms, farm management, disk and tape storage, mass storage management, LANs, WANs
- Improved understanding of the costs of all this to feed into the experiments’ computing models and the LCG TDR

Where we need to go

- **Experience this year** must decide the **basic computing model** for 2007-8
  - we need to know the scale, performance, resources needed
  - and we have to ensure commitments from regional centres, grid operations, networks
- We have to decide on the longer term need for **common applications support** – because we must also look for commitments to provide these resources - this **will not all be done at CERN**
- Essential that the **next round of middleware** is developed in **close collaboration with experiments**
- In the meantime we must **maintain VDT/LCG-2** **as a solid backup**
- **2007 is not so far away –**
  - development must now give way to **delivery, integration, services**
  - **end-to-end data challenges** are essential to **verify realistic scenarios** and see where we need to improve

Halfway through Phase 1 of the project we now see practical results

thanks to your hard work and solid support for the collaboration