virtual laboratory for e-science

BiG Grid
the dutch e-science grid

Grid Computing and Site Infrastructures

David Groep, NIKHEF

UvA SNE 2009

vl·e virtual laboratory for e·science

BiG Grid the dutch e-science grid

NIKHEF pdp

eGee Enabling Grids for E-sciencE

Graphics: Real Time Monitor, Gidon Moont, Imperial College London, see http://gridportal.hep.ph.ic.ac.uk/rtm/

The case for grid: applications
Security models for authN and authZ: PKI, federations and VOs
The Grid: the protocols, the information system and its anomalies
At the site: cluster architectures and networks
Scaling up the infrastructure: power, systems management
Monitoring: things that break, keeping an eye on things
Operational Security: 'interesting' users, policies, and the SSCs
Putting it together for growth: layering, growth, and grid models
Sustainable infrastructure: standards and EGI
What Next?

GRID COMPUTING AND INFRASTRUCTURES

news.com.au

Broadband network soon to be o

By Ryan Emery
April 07, 2008 03:44am

BY the time Australia upgrades its broadban
could be obsolete - thanks to a high-speed i
in Geneva.

The new network, called "the grid", is more than 1
than a typical broadband connection.

It is a system of fibre-optic cables and modern ro
movies and entire music catalogues can be dow
hours.

The grid, devised by scientis
Nuclear Research, and ho
of data from their Large H
also transmit holographic
telephony for the price of

physics professor David

De Telegraaf DigiTaal

SNELNIEUWS

Maandag 21 april

binnen- en Buitenland RSS

10:44 Zoon aangezien voor kalkoen

10:32 Opcenten sinds 2000 verdubbeld

10:30 Stelling: NAVO moet meer...

10:09 Zondags al malen

HOME > NIEUWS > DIGITAAL

ma 07 apr 2008, 12:29

Internet binnenkort 10.000 keer sneller

door onze redactie

AMSTERDAM - Het internet zoals wij dat kennen kan binnenkort wel eens sterk verouderd zijn. De wetenschappers die aan de wieg stonden van het huidige internet zijn namelijk bezig met een variant die tot 10.000 keer sneller zal zijn dan het snelste huidige breedbandnetwerk.

De Large Hadron Collider, de deeltjesvernieller van het Europese onderzoeksbureau CERN.

CERN," zegt professor David Britton, n de universiteit van Glasgow in de

en in Zwitserland dat de Large we deeltjesversneller jaarlijks zoveel veel als op 56 miljoen cd'tjes zou t daardoor het hele internet

Telegraph.co.uk

Home News Sport Business Travel Jobs Motoring Telegraph TV SEARCH

apse as video demand soars

within two years under the pressure of booming demand rned.

ce world wide web

Webwereld

ALTIJD HET LAATSTE ICT-NIEUWS

Nieuws
Column
Video
Dossier
Blog
Beveiliging ...

Tip ons Archief Whitepapers

Nieuws

Nederland grote hulp bij grid-project

Dinsdag 26 april 2005, 15:54 - Acht computercentra, waaronder het Nederlandse Sara, zijn met elkaar verbonden om binnen tien dagen 500 terabyte aan data uit te wisselen.

Door Edwin Feldmann

Bij het zogeheten LHC Computing Grid-project zijn diverse Nederlandse instellingen betrokken waaronder het Nederlandse Sara en het Nikhef. De centra gaan de Large Hadron Collider (LHC) testen.

Doel van het project is om voldoende reken-, opslag- en netwerkfaciliteiten te verschaffen om wetenschappelijke experimenten te laten slagen.

De verbindingen zullen binnen tien dagen ononderbroken gegevens uitwisselen met een gemiddelde snelheid van 600 MBps. In totaal zal er aan het einde ongeveer 500 terabyte (512.000 gigabyte) aan data zijn verstuurd. "Wanneer er gebruik zou zijn gemaakt van een eenvoudige 512 Kbps-verbinding zou hiervoor 250 jaar nodig zijn", aldus de organisatie.

Onderzoekers staan te dringen om plaatsje op Nederlands wetenschappelijk grid

BIG GRID officieel gelanceerd

Op het BIG GRID-lanceringsevenement leken de aanwezigen elkaar de loef af te willen steken met de vele petabytes (1000 TB) die ze genereren met hun onderzoek. Een ding was duidelijk: een onderzoeksgrid voor opslag en verwerking van al die data is hard nodig. Er wordt aan gewerkt. Twee jaar geleden werd er door de te-

Een snel netwerk is de basis voor BIG GRID. Met het Nederlandse SURFnet is dat er al. Daar hangt al de nodige apparatuur aan, zoals de nieuwe SARA-supercomputer, die al op gridachtige wijze wordt gebruikt en gedeeltelijk uit de pot van BIG GRID is betaald. Die infrastructuur en apparatuur worden in de komende jaren aangevuld tot een grootschalig grid voor wetenschappelijk gebruik. Daarbij zijn ook industriële partners welkom, zoals

- Name "Grid" chosen by analogy with electric power grid (Foster and Kesselman 1997)
- Vision: plug-in computer for processing power just like plugging in toaster for electricity.

The idea has been around for decades

'*distributed computing*',
'*metacomputing*'

- *and will be around: 'Web 2.0', 'Virtualisation', 'Cloud Computing'*

Grids in Science

The Grid is 'more of everything' as science struggles to deal with ever increasing complexity

more than one computer

more than one place on earth

more than one science!

more than ...

Why would we need it?

Enhanced Science needs more and more computations and

Collected data in science and industry grows exponentially

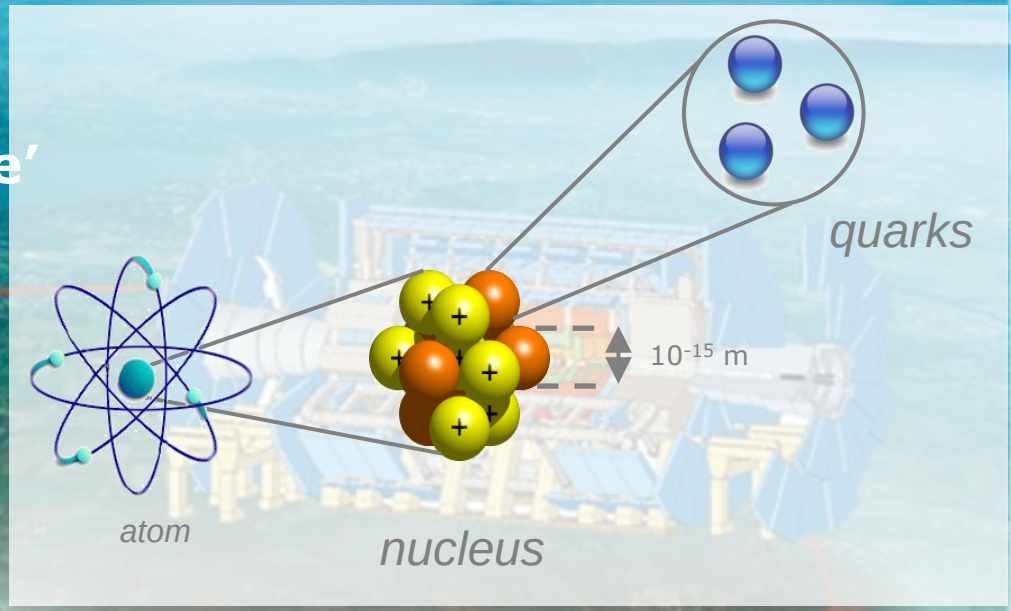
The Bible	**5**	**MByte**
X-ray image	5	MByte/image
Functional MRI	1	GByte/day
Bio-informatics databases	500	GByte each
Refereed journal papers	1	TByte/yr
Satellite world imagery	5	TByte/yr
US LoC contents	20	TByte
Internet Archive 1996-2002	100	TByte
Particle Physics 2005	1	PByte/yr
Particle Physics Today: LHC	**20**	**PByte/yr**

1 Petabyte = 1 000 000 000 Megabyte

LHC Computing

Large Hadron Collider

- 'the worlds largest microscope'
- 'looking at the fundamental forces of nature'
- 27 km circumference
- Located at CERN, Geneva, CH

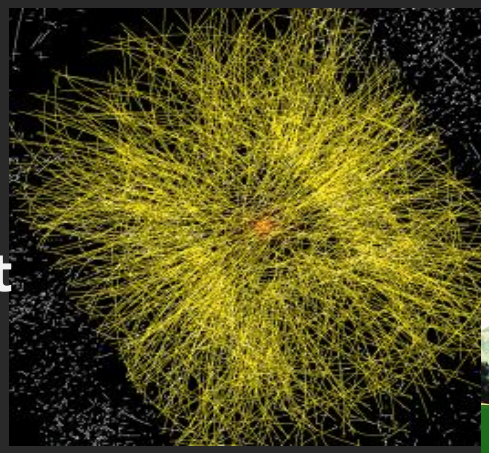


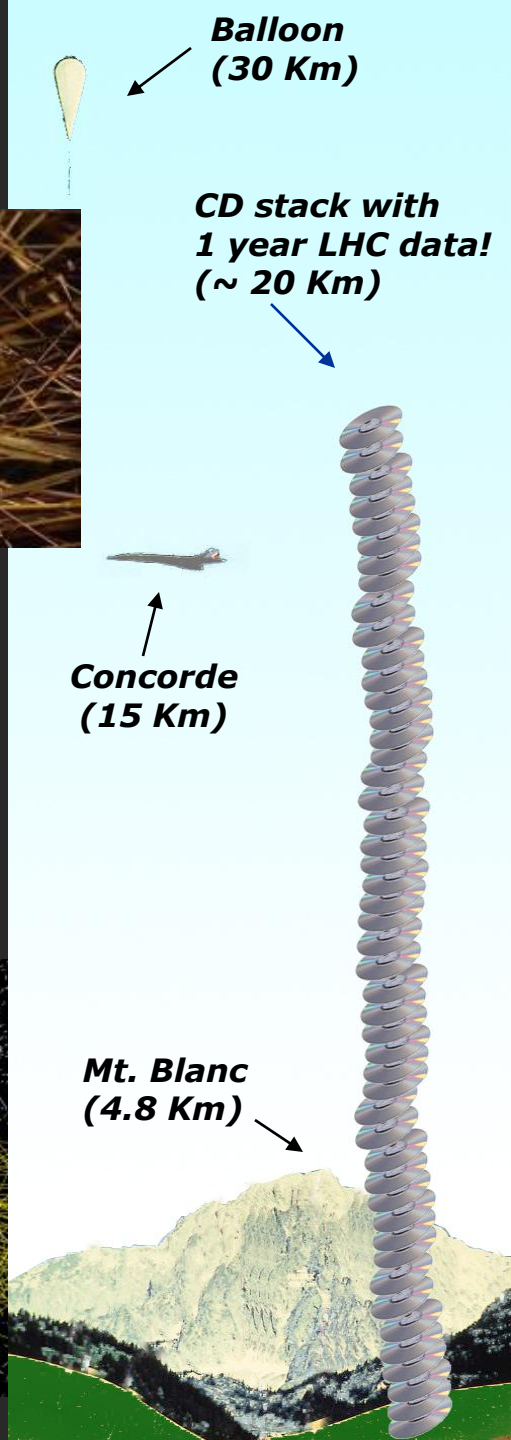
~ 20 PByte of data per year, ~ 60 000 modern PC style computers

- Signal/Background 10^{-9}

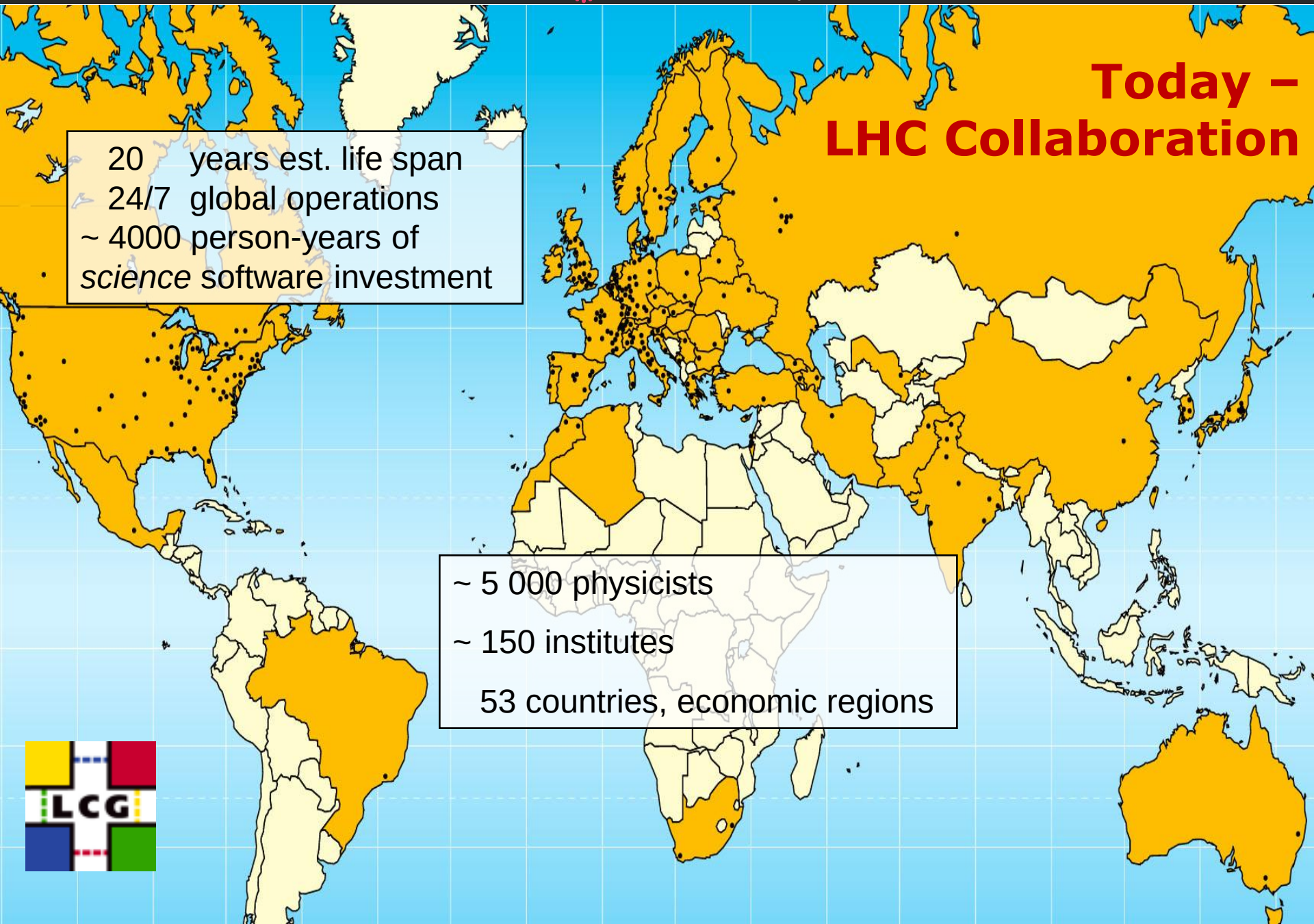
- Data volume
  - (high rate) **X**
    (large number of channels) **X**
    (4 experiments)
  - **➔ 20 PetaBytes of new data each year**

- Compute power
  - (event complexity) **X**
    (number of events) **X**
    (thousands of users)
  - **➔ 60'000 of (today's) fastest CPUs**





Today – LHC Collaboration



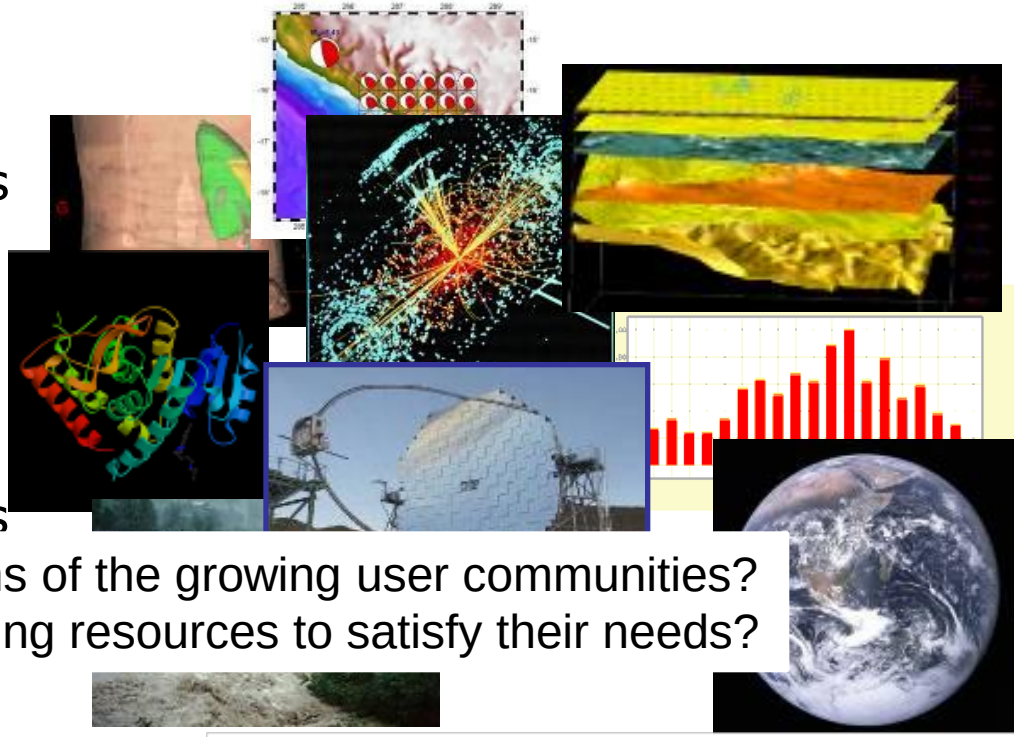
20 years est. life span
24/7 global operations
~ 4000 person-years of *science* software investment

~ 5 000 physicists
~ 150 institutes
53 countries, economic regions

LCG

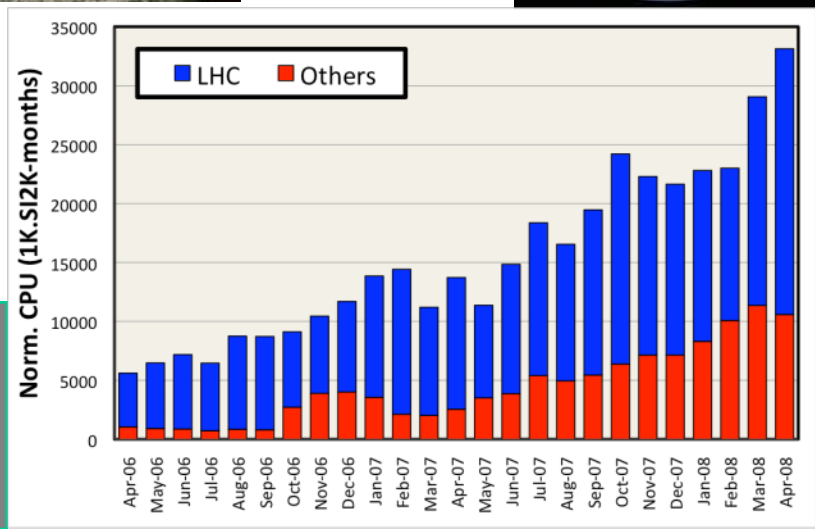
EGEE Achievements - Applications

- >270 VOs from several scientific domains
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics



How do we match the expectations of the growing user communities?
Will we have enough computing resources to satisfy their needs?

  - High Energy Physics
  - Life Sciences
- Further applications under evaluation

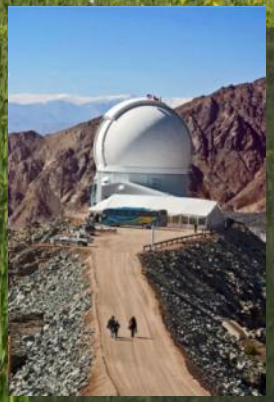


Applications have moved from testing to routine and daily usage
~80-95% efficiency

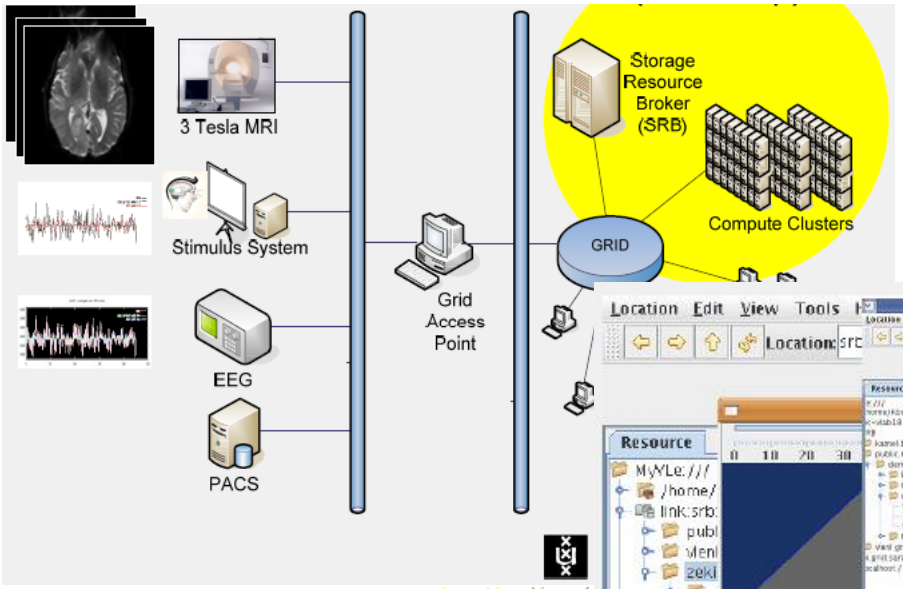
Astronomy & Astrophysics

LOFAR large distributed radio telescop

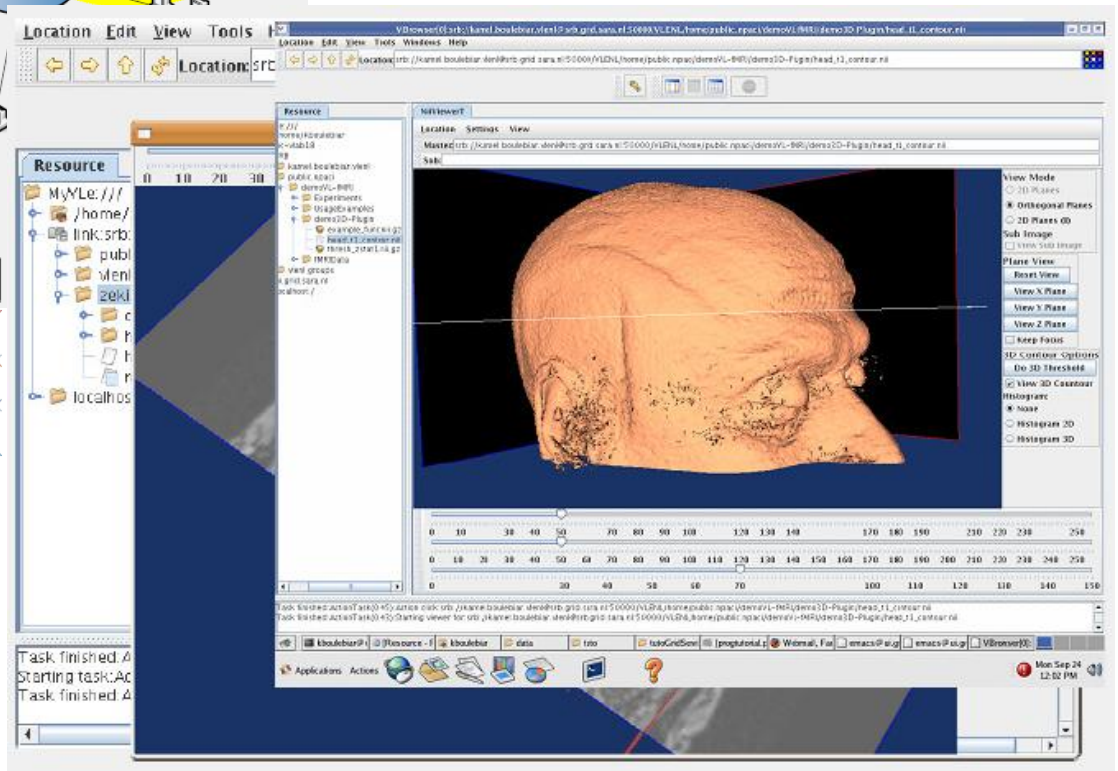
AUGER & ARGO Cosmic Ray Observato



Functional MRI analysis



Storage of fMRI research data for sharing between groups and processing of image alignments



vl·e

Research work by:
Silvia Olabarriaga (AMC, UvA)
Tristan Glatard (IvI,UvA)
Abdullah Ozsoy (IvI,UvA)

In silico drug discovery

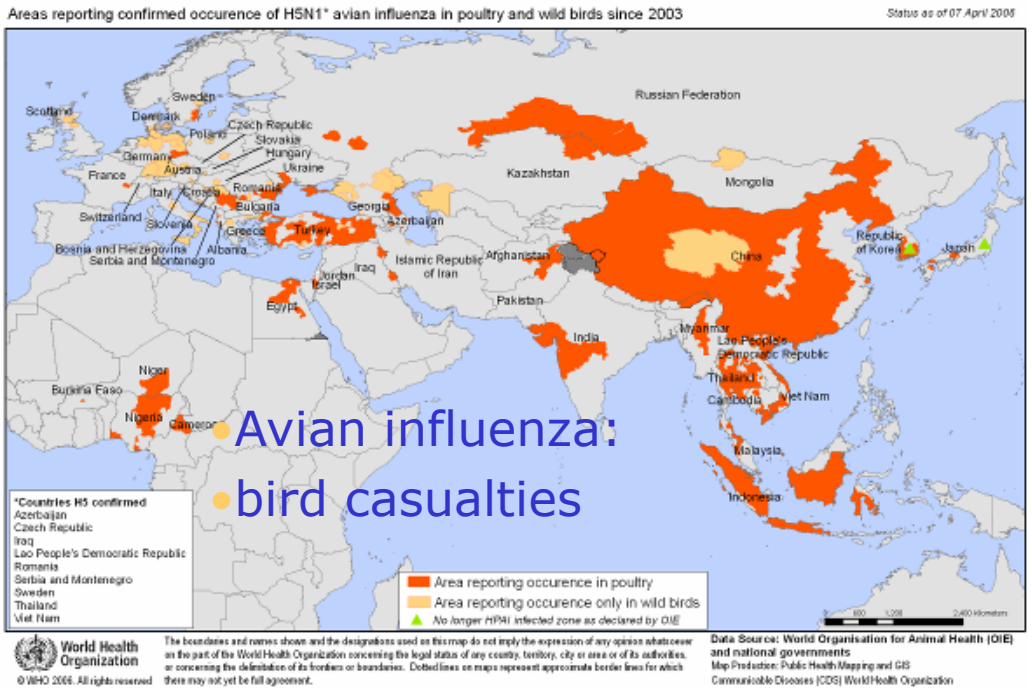
- Diseases such as HIV/AIDS, SRAS, Bird Flu, Malaria etc. are a threat to public health due to world wide exchanges and circulation of persons

- Grids open new perspectives to *in silico* drug discovery
  - Reduced cost and adding an accelerating factor in the search for new drugs

International collaboration is required for:

- Early detection
- Epidemiological watch
- Prevention
- Search for new drugs

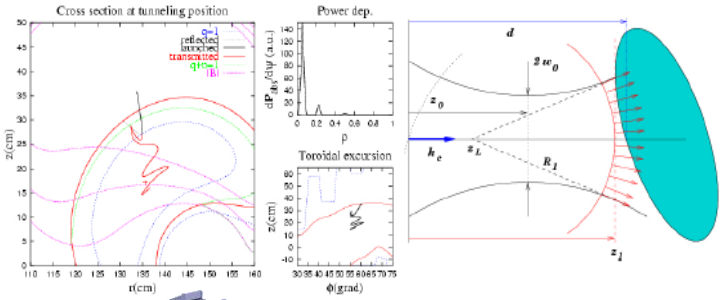
WISDOM
Initiative for grid-enabled drug discovery against neglected and emergent diseases



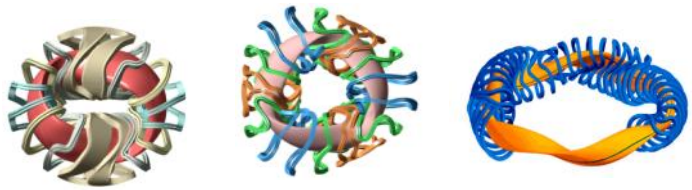
Fusion

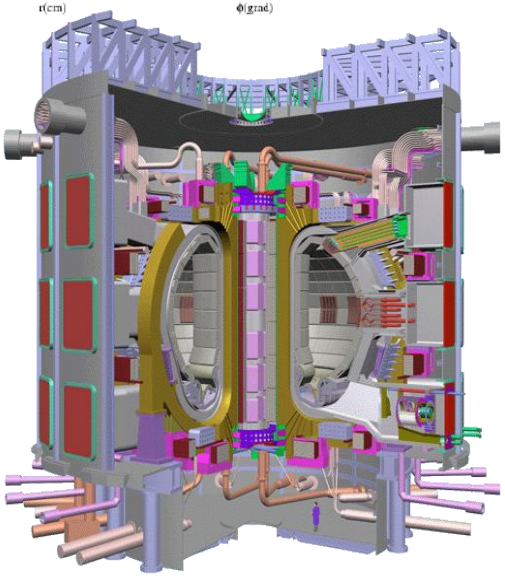
Commercial exploitation of fusion energy still needs to solve several outstanding problems requiring exceptional computing facilities including supercomputers and cluster-based grids

- Ion Kinetic Transport
- Massive Ray Tracing
- Stellarator Optimization



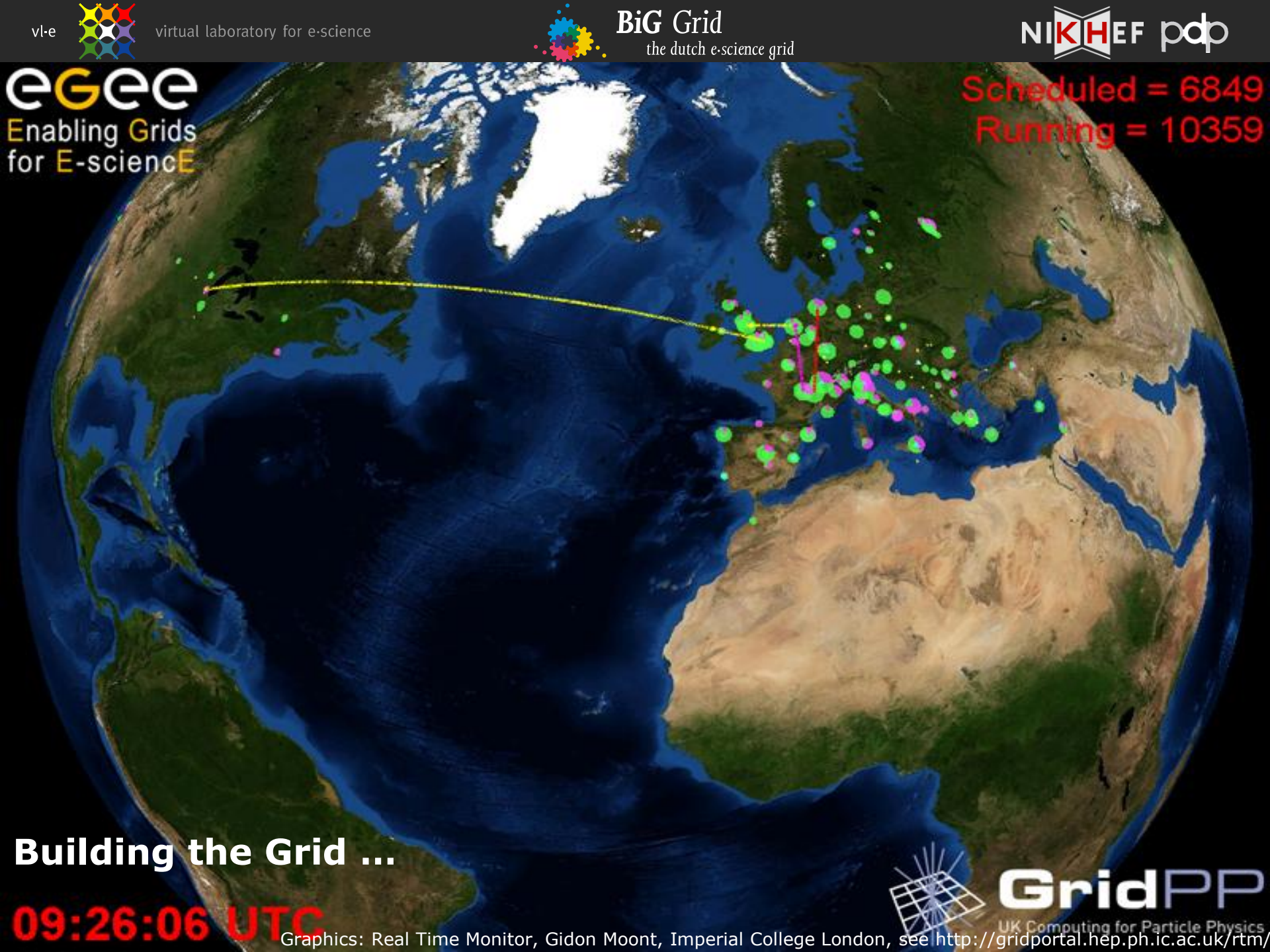
Interworking course-grained clusters and MPP systems across both the EGEE and DEISA grids





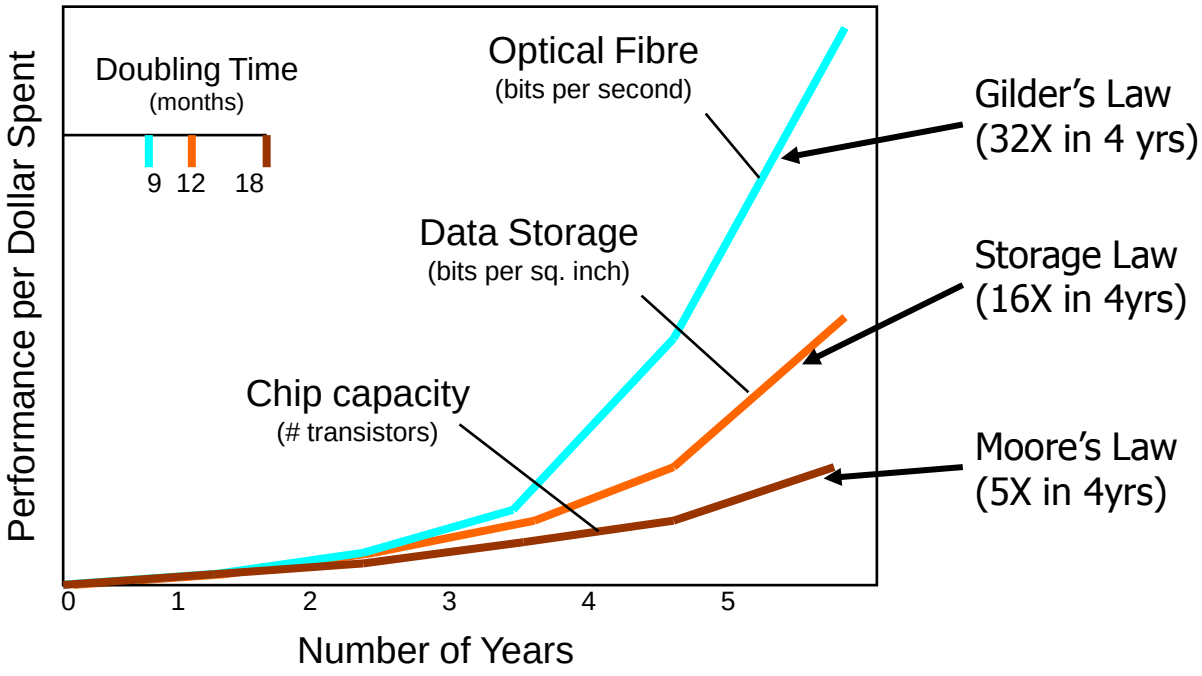
Enterprise

- Transaction processing
- Finance (what-if analyses)
- Pharma (in-silico drug design)
- Aerospace (fluid dynamics)


vl·e
virtual laboratory for e·science
BiG Grid
the dutch e-science grid
NIKHEF
pdp
eGee
Enabling Grids
for E-sciencE
Scheduled = 6849
Running = 10359
Building the Grid ...
09:26:06 UTC
GridPP
UK Computing for Particle Physics
Graphics: Real Time Monitor, Gidon Moont, Imperial College London, see http://gridportal.hep.ph.ic.ac.uk/rtm/

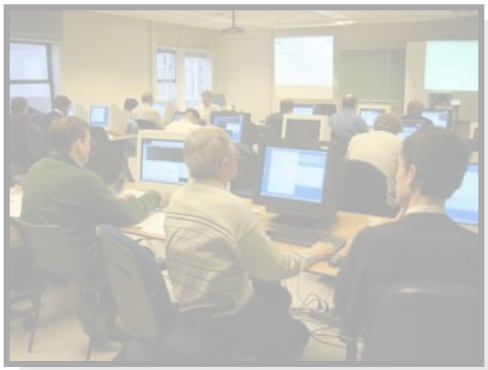
Why Grid computing – today?

- New applications need larger amounts of data or computation
- Larger, and growing, distributed user community
- Network grows faster than compute power/storage



Graphic: "The Triumph of Light", Scientific American, January 2001

What is Grid?



Cycle scavenging

- harvest idle compute power
- improve RoI on desktops

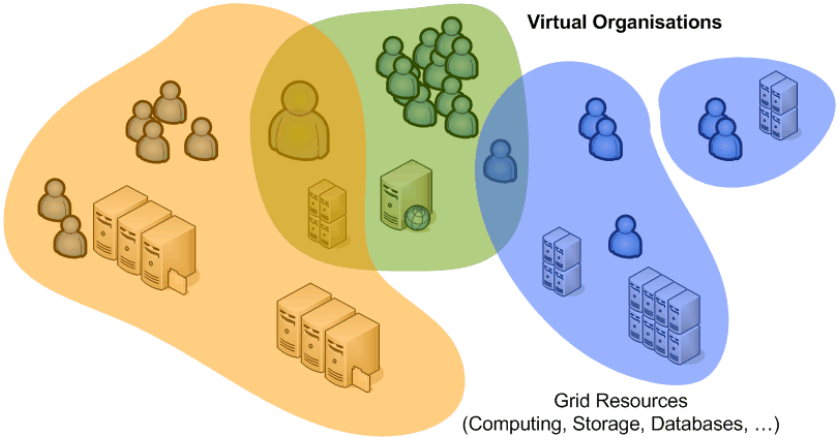
Cluster computing and storage

- What-if scenarios
- Physics event analysis
- Improve Data Centre Utilization

Cross-domain resource sharing

- more than one organisation
- more than one application
- more than one …

- open protocols
- collective service



Community Building
- authentication
- authorization
- virtual organizations

Scheduling and clustering
- resource management
- prioritization and fair-share

Hardware Infrastructures
- compute clusters
- disk and tape storage
- database services

Operational Security Policy
- distributed incident response
- policies

Managing Complexity
- systems management
- scaling
- multi-national infrastructures

Grid Structures

Definition of inter-organizational grids

Virtual Organizations

COMMUNITY BUILDING

Three essential ingredients for Grid

'inter-organizational resource sharing'

A grid combines resources that

- Are not managed by a single organization
- Use a common, open protocol … that is general purpose
- Provide additional qualities of service, *i.e.*, are usable as a collective and transparent resource

GRID today

DAILY NEWS AND INFORMATION FOR THE GLOBAL GRID COMMUNITY / JULY 22, 2002: VOL. 1 NO. 6

WHAT IS THE GRID? A THREE POINT CHECKLIST
By Ian Foster Argonne National Lab & University of Chicago

The recent explosion of commercial and scientific interest in the Grid makes it timely to revisit the question: What is the Grid, anyway? I propose here a three-point checklist for determining whether a system is a Grid. I also discuss the critical role that standards must play in defining the Grid.

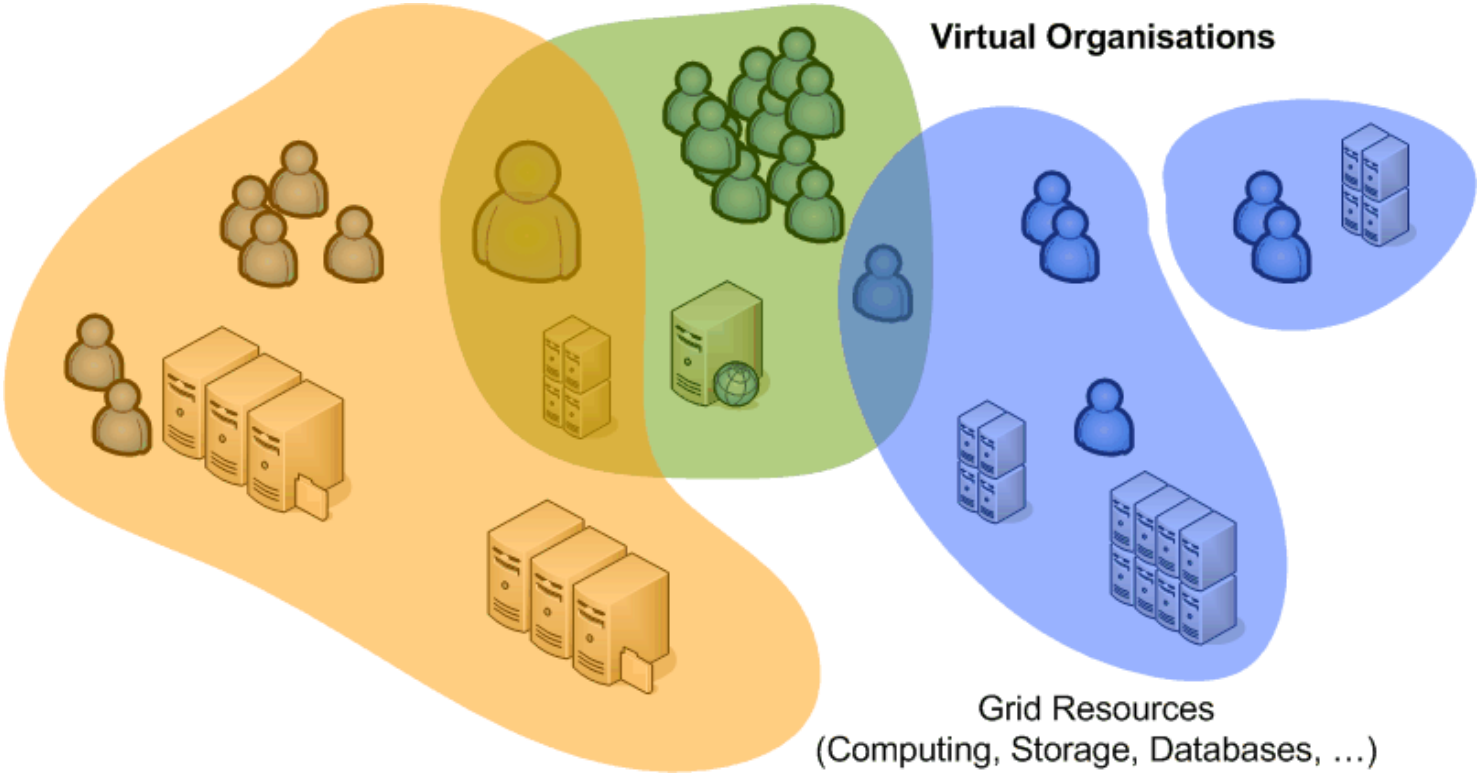
The Need for a Clear Definition Grids have moved from the obscurely academic to the highly popular. We read about Compute Grids, Data Grids, Science Grids, Access Grids, Knowledge Grids, Bio Grids, Sensor Grids, Cluster Grids, Campus Grids, Tera Grids, and Commodity Grids. The skeptic can be forgiven for wondering if

Source: Ian Foster in *Grid Today*, July 22, 2002; Vol. 1 No. 6, see http://www-fp.mcs.anl.gov/~foster/Articles/WhatIstheGrid.pdf

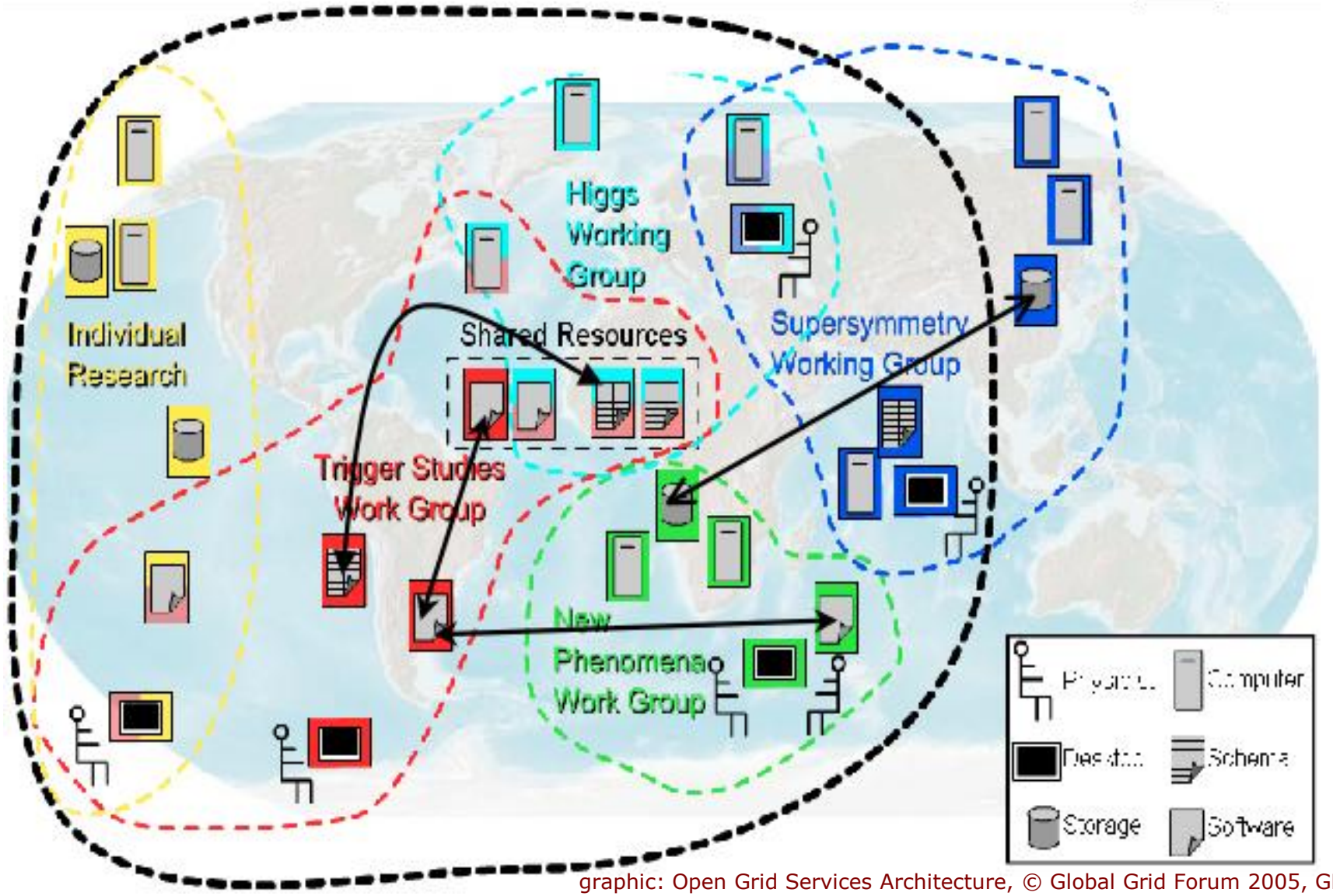
Virtual Organisations

The communities that make up the grid:

- **not under single hierarchical control**,
- (temporarily) **joining forces** to solve a particular problem at hand,
- bringing to the collaboration a subset of their resources,
- sharing those **at their discretion** and each **under their own conditions**.



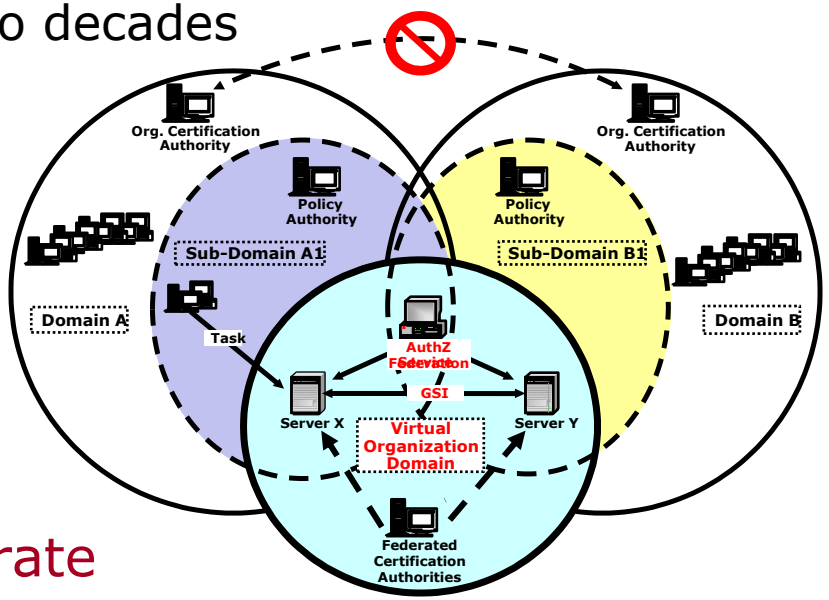
Although nothing is ever quite that neat …



graphic: Open Grid Services Architecture, © Global Grid Forum 2005, GFD.30

Federation in Grid Security

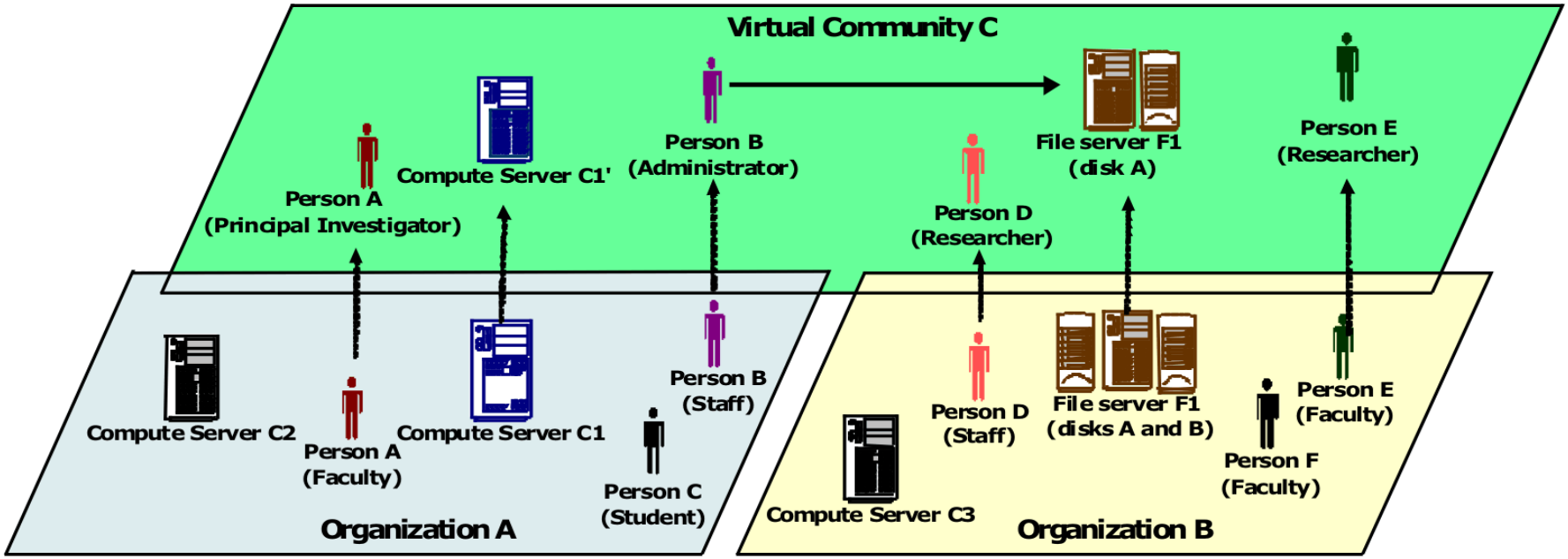
- There is no *a priori* trust relationship between members or member organisations!
  - VO lifetime can vary from hours to decades
  - VO not necessarily persistent (both long- and short-lived)
  - people and resources are members of many VOs



- but who to trust and how to federate
  - at the organisation level?
    eduroam™, inCommon, SWITCHaai, UK Access Mngt Federation
  - at the user and VO level?
    user AuthN and VO-centric AuthZ authorities 'orthogonal' to the org structure

Organizing people

'Identity is not enough'



'virtual' organization roles are independent of home organization roles and **authority for the VO roles rests with the VO**

Graphic: Open Grid Forum OGSA Working Group, GFD.30

Authentication vs. *Authorization*

For user-centric delegation and VO-based grids

- **Single** Authentication token ("passport")
  - issued by a party trusted by all,
  - recognised by many resource providers, users, and VOs
  - satisfy traceability and persistency requirement
  - in itself does not grant any access, but provides a unique binding between an identifier and the subject

- Per-VO (per 'UHO') Authorisations ("visa") attributes
  - granted to a person/service via a virtual organisation
  - based on the 'passport' name
  - embedded in the single-sign-on token (proxy)
  - acknowledged by the resource owners
  - providers can obtain lists of authorised users per VO, but can still ban individual users

Federated PKI: the IGTF policy bridge

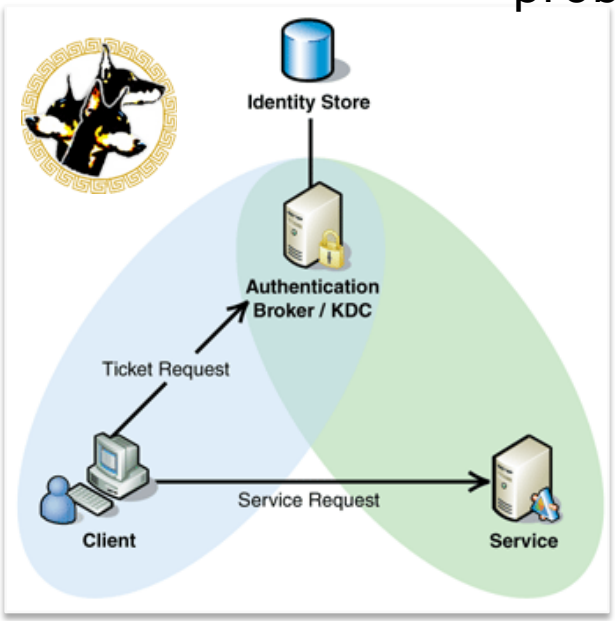
Home-organization based federations

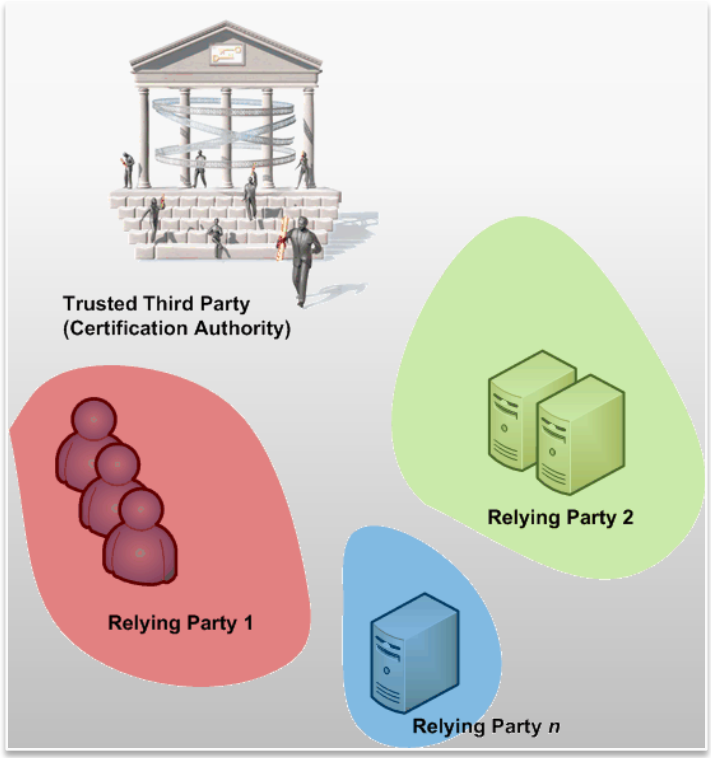
User centric identity

IDENTITY AND AUTHENTICATION

Security Trust Mechanisms

Making the order of the problem manageable …





Intra-organizational security vs. global grids

Direct (username-password) authN

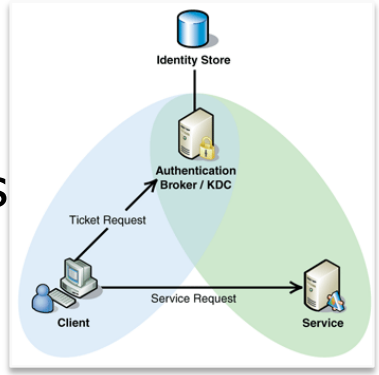
- Dedicated to each site where you want access
- Usually strongly linked to authorization
  - different accounts for different roles
- In a multi-organizational problem is

$$\mathcal{O}(n_{sites}) * \mathcal{O}(n_{users})$$

Federation technologies (see later) help in some respects

Kerberos

- Common trust domain around a KDC
- Based on service tickets, derived from a TGT
  - Encrypted with the service key from the target service
  - Whether you talk to the 'right' server is implicit in it's ability to decode your service ticket
- Cross-domain trust by recognizing KDC tickets
  - interesting in presence of symmetric crypto
  - but usually, alignment mismatch between organizations is the limiting factor
  - For multi-domain gets to be $\mathcal{O}(n^2)$ for *n* sites

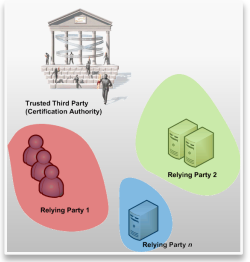


PKI

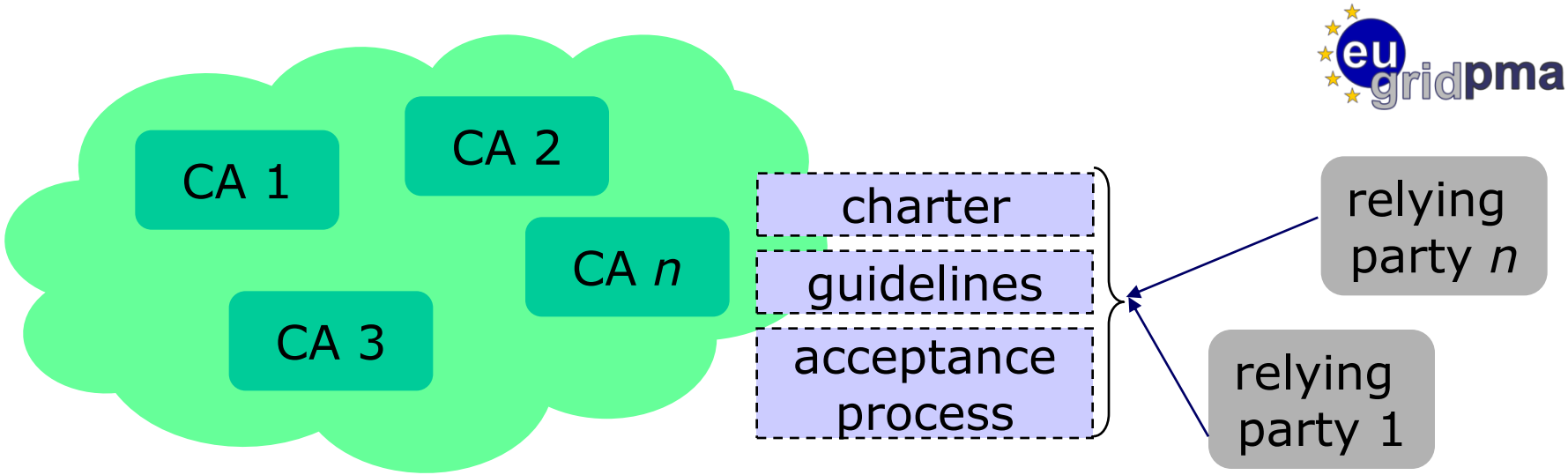
- Relying parties (sites and users) all recognise a trusted third party (CA)
- Problem is now $\mathcal{O}(n_{CA})$

  and n_{CA} is hopefully << n_{sites}

- But there will be more than one CA as well …



Federated PKI for authentication

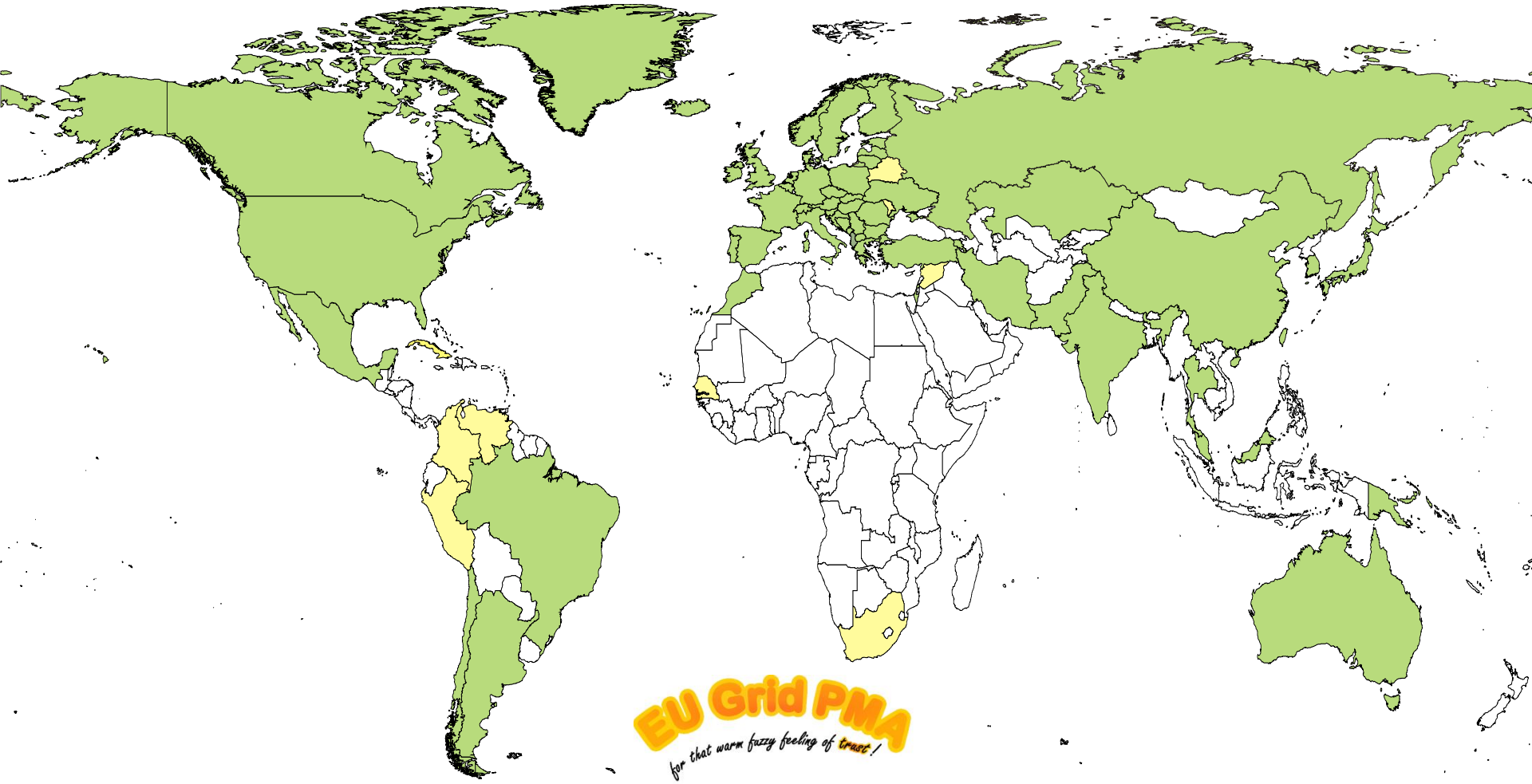


- A Federation of many independent CAs (a 'policy bridge')
  - common minimum requirements
  - trust domain as required by users and **relying parties**
  - well-defined and peer-reviewed acceptance process
- User has a single identity
  - from a local CA close by
  - works across VOs, with single sign-on via impersonation 'proxies' (RFC3820)
  - certificate itself also usable outside the grid

International Grid Trust Federation and EUGridPMA, see http://www.gridpma.org/

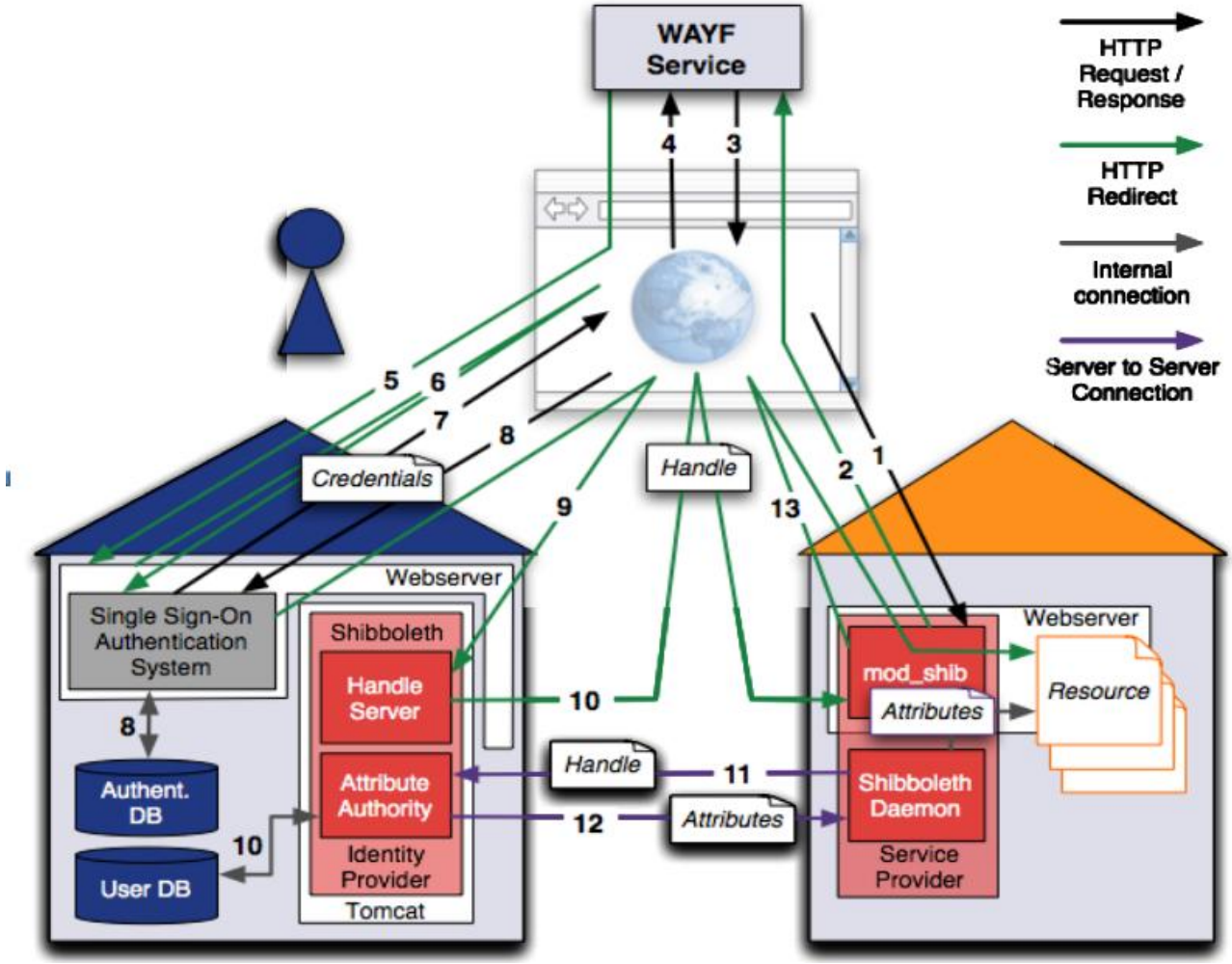
IGTF
International Grid Trust Federation
AP | EU | TAG

Federation of 3 Regional "PMAs", that define common guidelines and accredit credential-issuing authorities



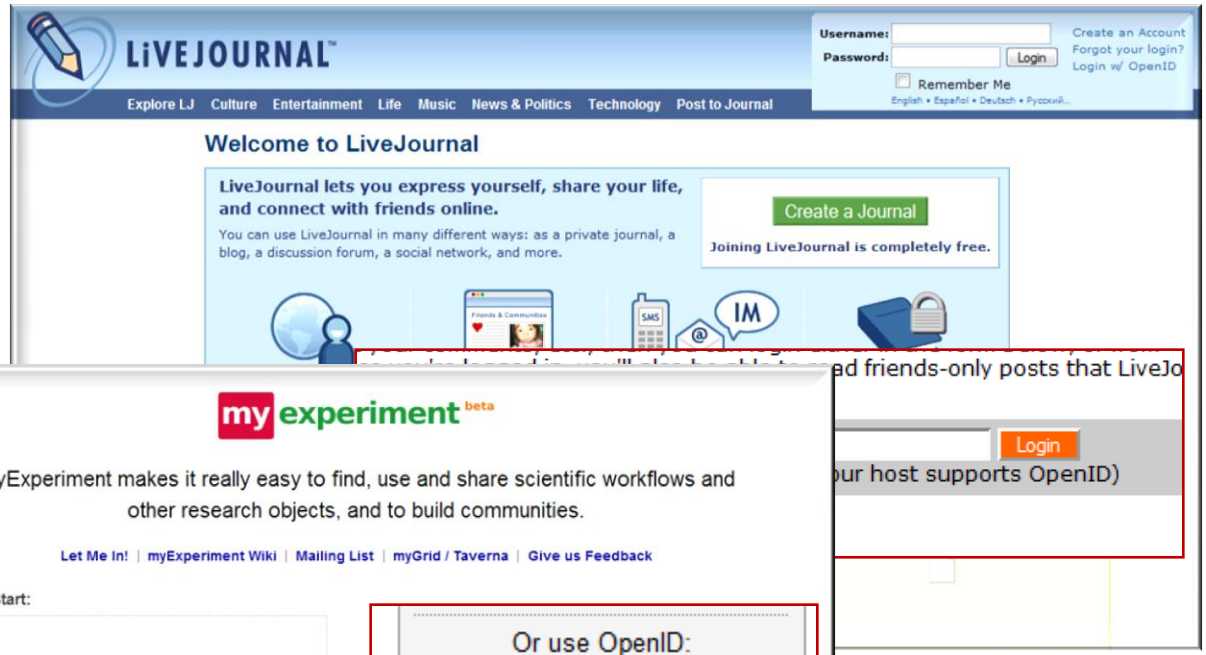
Federations rising

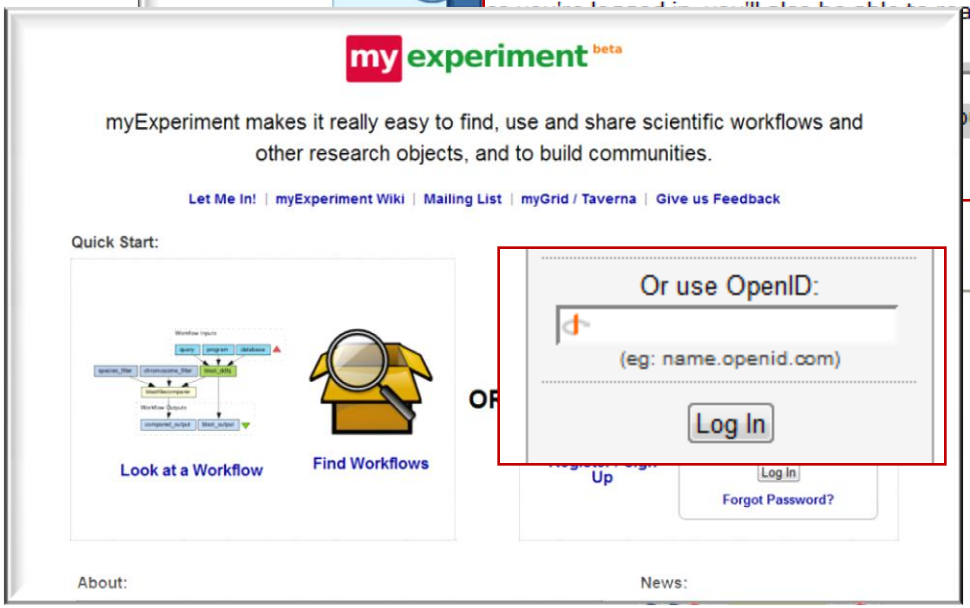
hiding PKI from users



Archetype: shibboleth (from Internet2 Middleware Services)

Federation techniques getting popular





OpenID

A-SELECT

based around web services security protocols and SAML assertions

Federated Authentication

- Users authenticate to their home organization
- There they have a set of attributes
  - With a release policy
  - Home organisation authoritative for them
- Service Providers make access decision based on the attributes related to an abstract handle
  - User's name (eduPersonPrincipalName) is also an attribute
- Home org cannot make assertions on VOmembership
  - We need to move to a multi-authority world
  - But is very good for identity: translatable to a specific PKI, where certificates derive from the federated identity (SLCS/MICS CAs)

User Centric Identity?

- CardSpace, project Higgins,
- …

see, e.g., Burton Group's blog
http://identityblog.burtongroup.com/

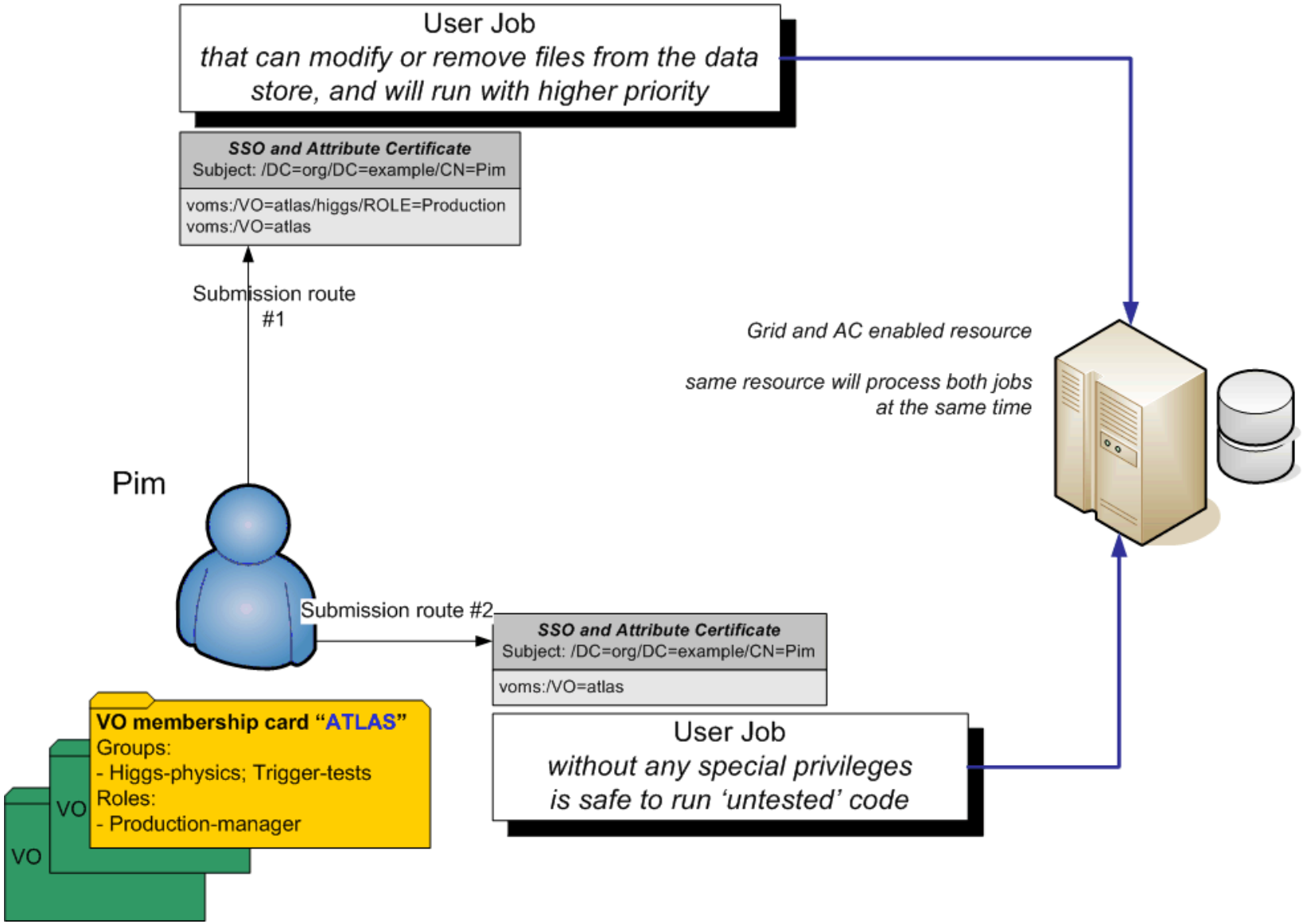
- Based on Web Services and 'SAML' assertions
  - Self-assertions
  - Assertions 'filled in' by trusted third parties, such as Visa, MC, etc.
- Required assurance depends on the target system

- *Interop testing just starting, see, e.g.*
  http://identityblog.burtongroup.com/bgidps/2007/08/recapping-the-c.html
- *See Kim Cameron's Identity blog*

Roles in communities and your home organisation
Delegation

ORGANISING COMMUNITIES

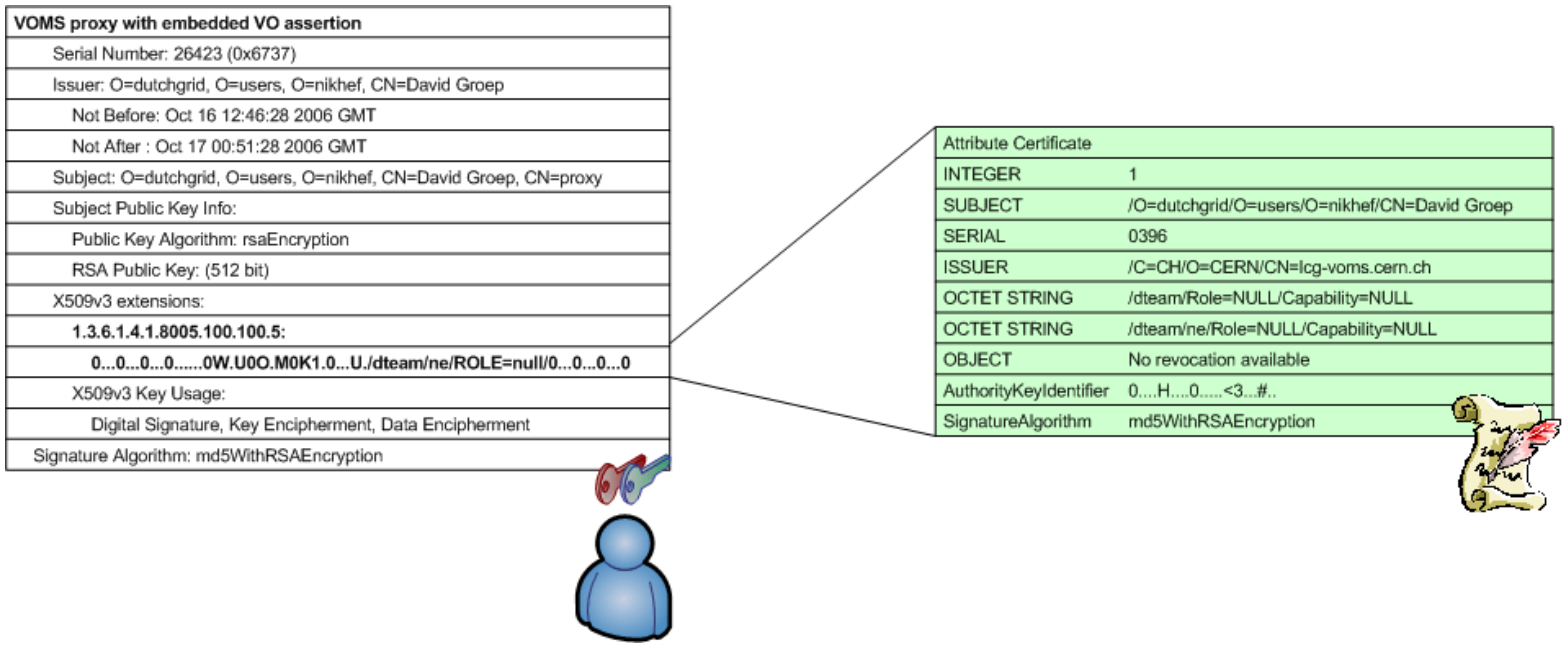
Role-based access control



VOMS: Assertions in X.509 AC or SAML

Virtual Organisation Management System (VOMS)

- push-model for signed VO membership tokens
  - using the traditional X.509 'proxy' certificate for shipping
  - TLS/X.509 is just so a convenient carrier of data ...



Organisations Working Together as an Infrastructure

Accessing resources

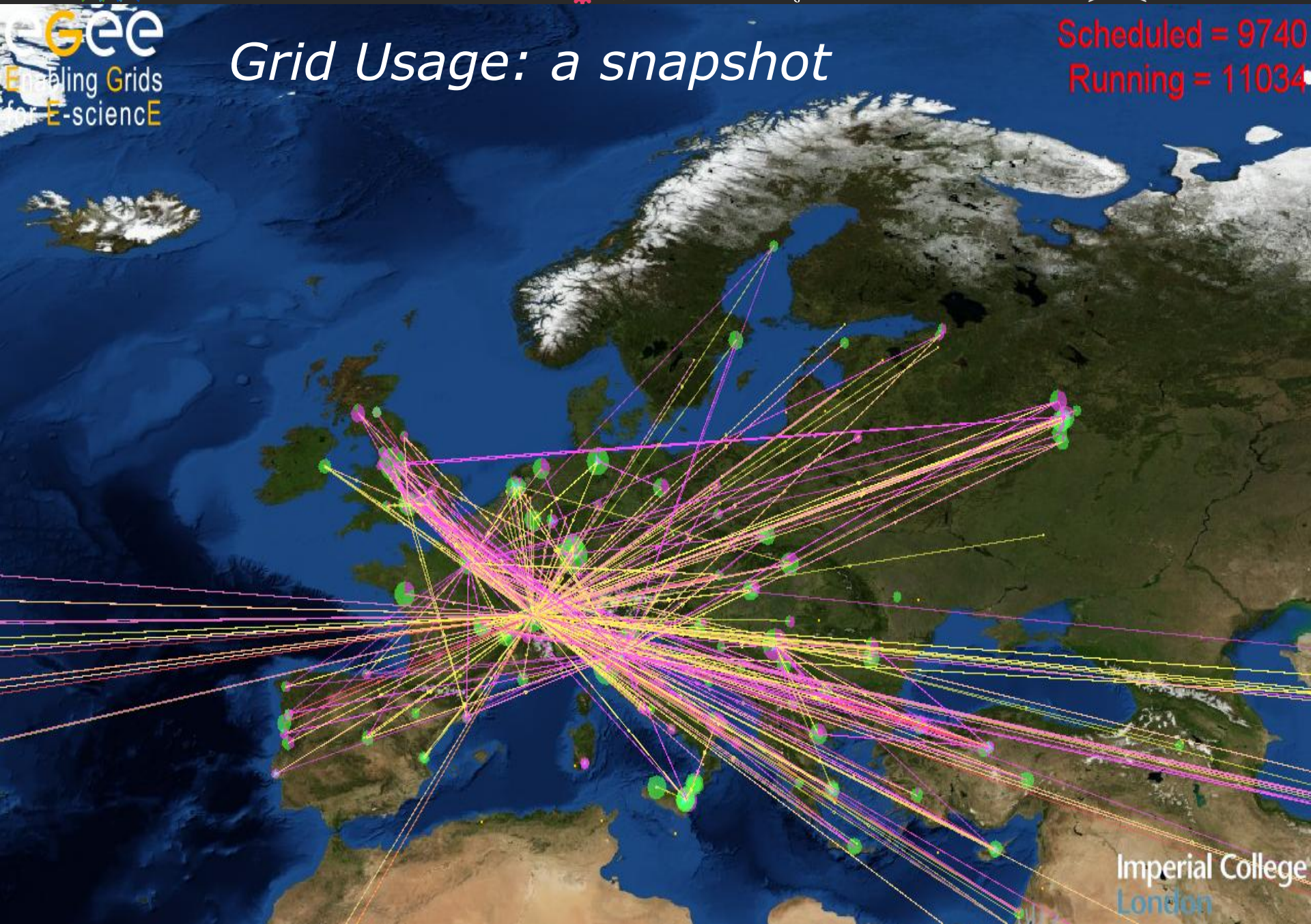
Resource Brokering

Data access

Resource Information Systems

USING THE GRID

Grid Usage: a snapshot



vl·e virtual laboratory for e·science

BiG Grid the dutch e-science grid

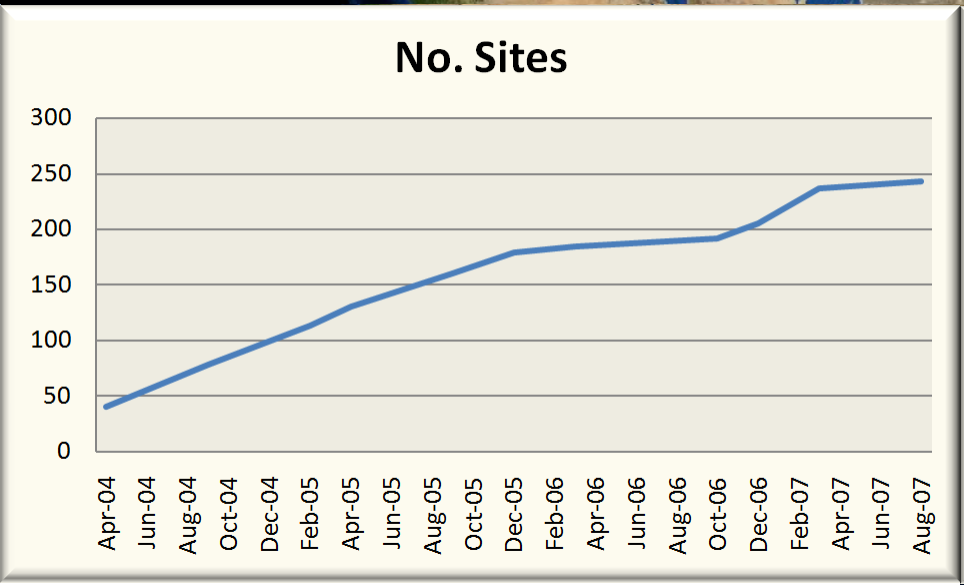
NIKHEF pdp

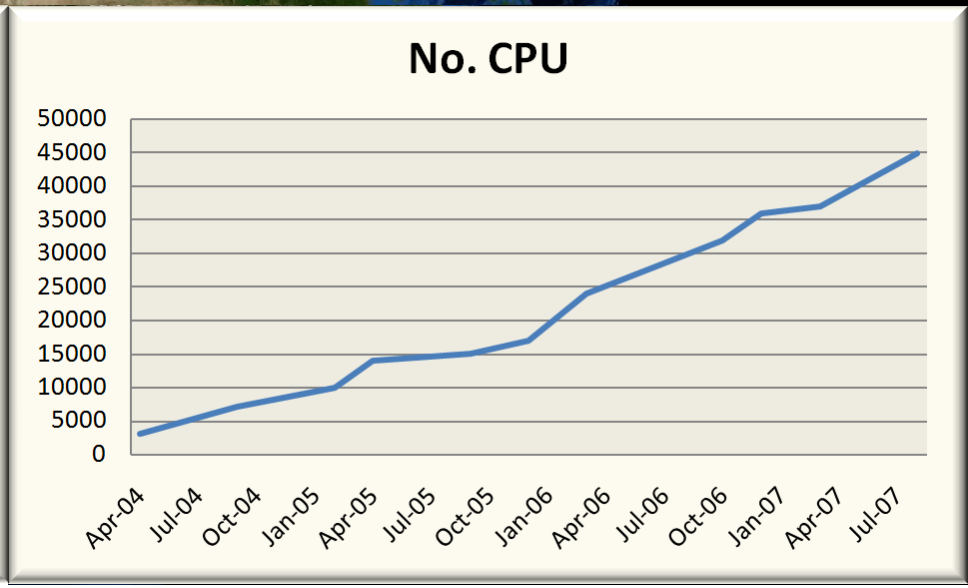
eGee Enabling Grids for E-sciencE

EGEE: ~250 sites, >45000 CPU

24% of the resources are contributed by groups external to the project

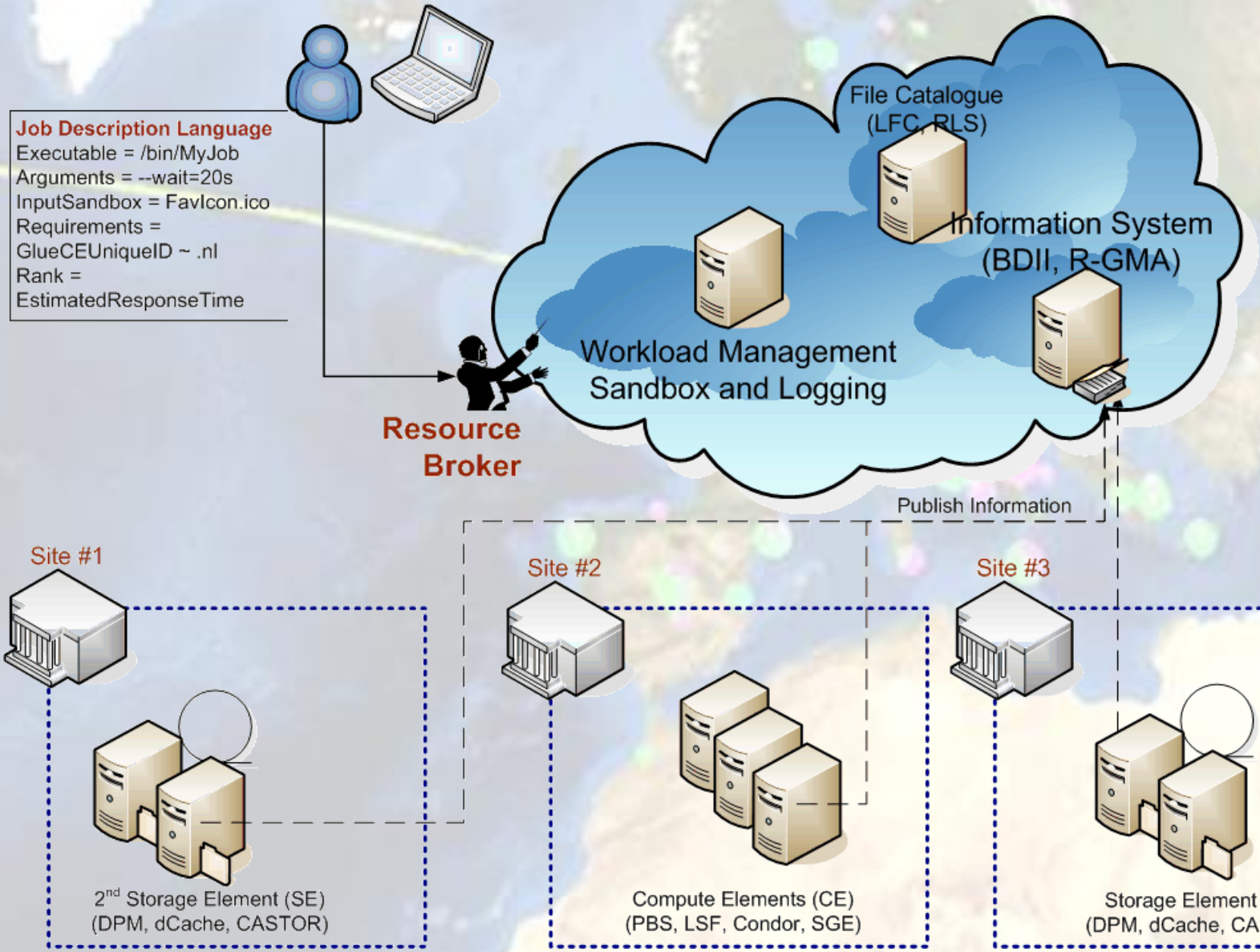
~>20k simultaneous jobs





Graphs: Ian Bird, EGEE SA1, EGEE07 Conference October 2007

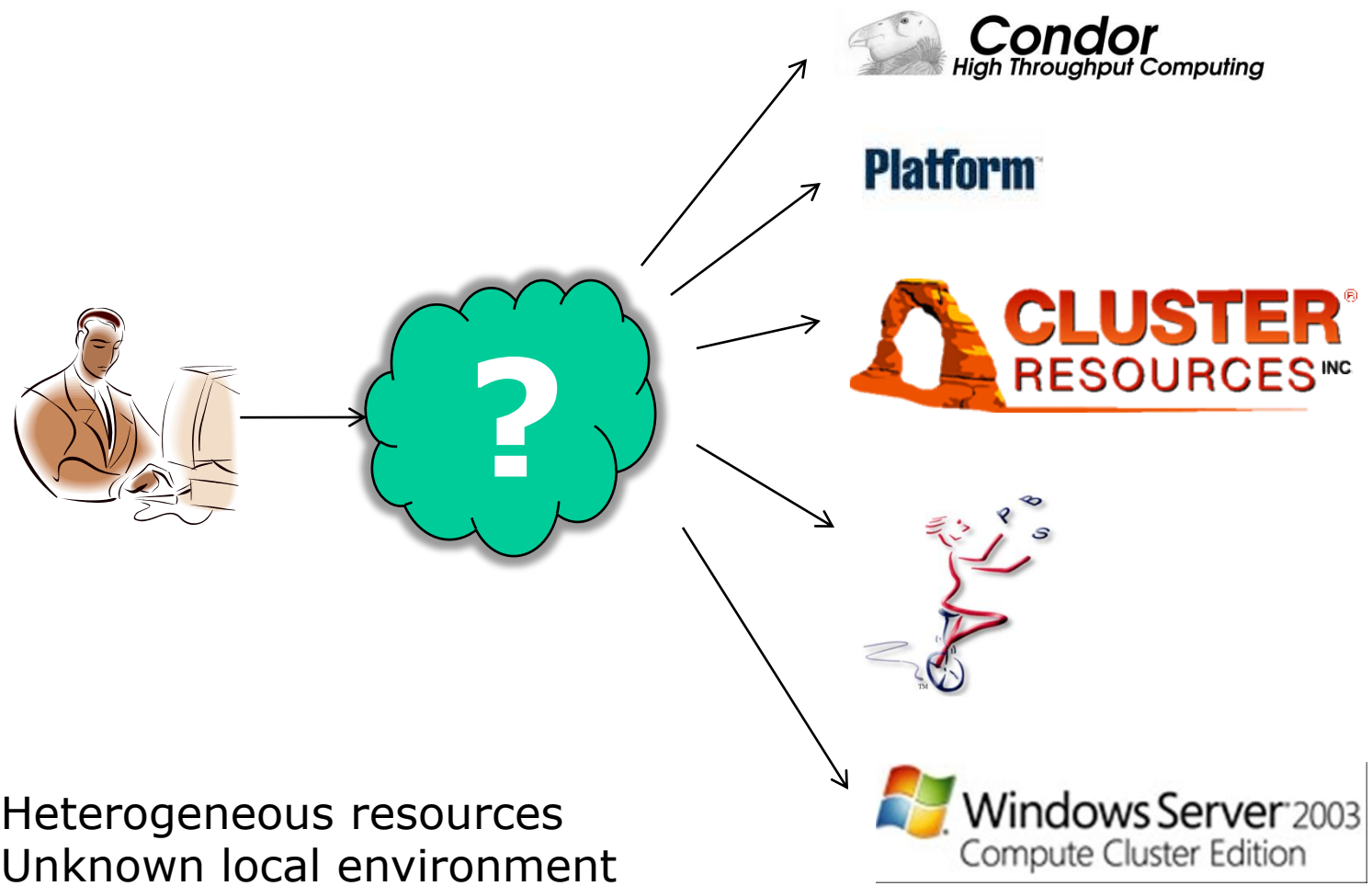
Typical grid topology for computational jobs



Services in a Grid

- Computing Element "front-end *service* for (set of) computers"
  - Cluster computing: *typically Linux with IP interconnect* with a 'head node' that batches or forwards requests to the cluster
  - Capability computing: *typically shared-memory supercomputers*
- Storage "front-end *service* for disk or tape"
  - Both disk and tape based
    - Varying retention time, QoS, uniformity of performance
  - Expressing ACLs in grid terms in challenging: *mapping of grid authorization to e.g. POSIX ACLs*
- File Catalogues … naming (data) objects in the Grid
  - for the really courageous people: represent computing, storage and data all as 'named objects' in a single 'grid name space'
- Information System … finding out resources on the Grid
  - Directory-based for static information
  - Monitoring and bookkeeping for real-time information
- Resource Broker …
  - Matching user job requirements to offers in the information system
  - WMS allows disconnected operation of the user interface

But you are not there yet …

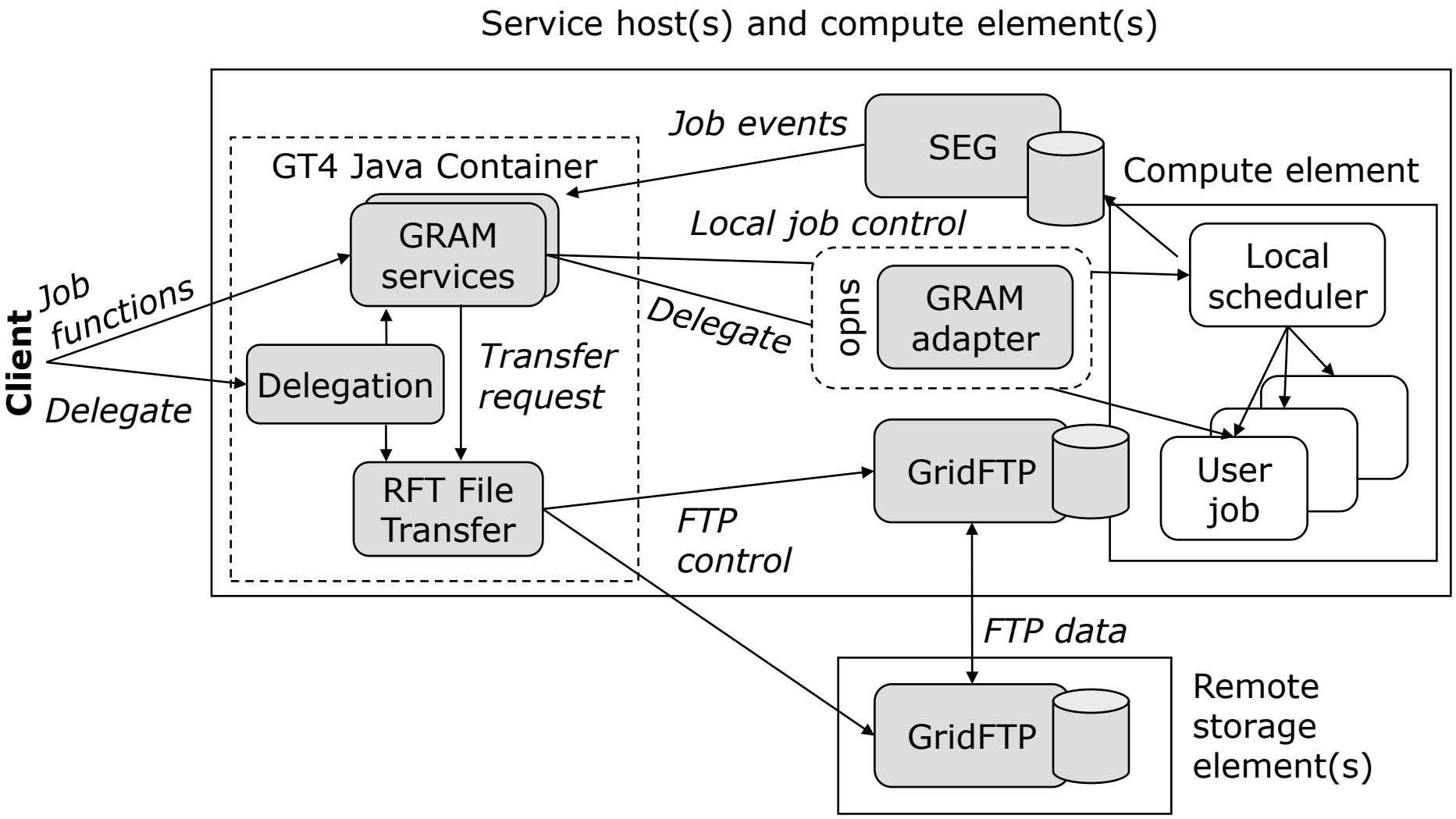


- Heterogeneous resources
- Unknown local environment
- Unknown access policies

A seemingly simple task

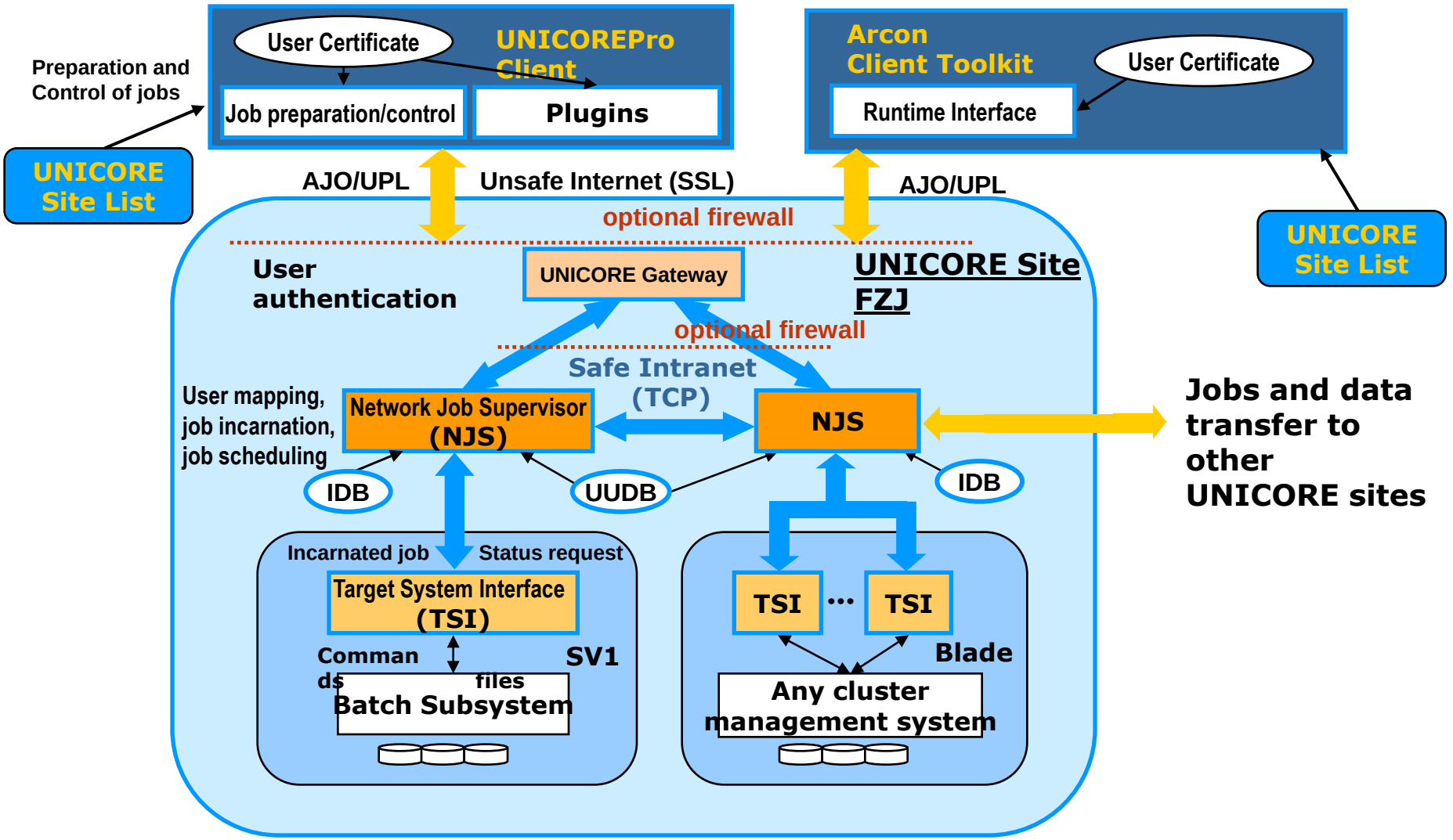
- The CE seems conceptually simple
  - submit a job
  - wait for it to run
  - retrieve the results
  - or kill it prematurely …
  - *but:* there are a bazillion ways to implement it
    - with implicit or explicit data staging
    - hide the entire site structure and use forwarding nodes
    - or even allow automatic forwarding to another site
    - policy and prioritization
- the user does not want to know the difference
  - and an automatic resource broker needs a backend for every type
- back-end is usually just a simple old batch system

Example: GT4 WS GRAM Architecture



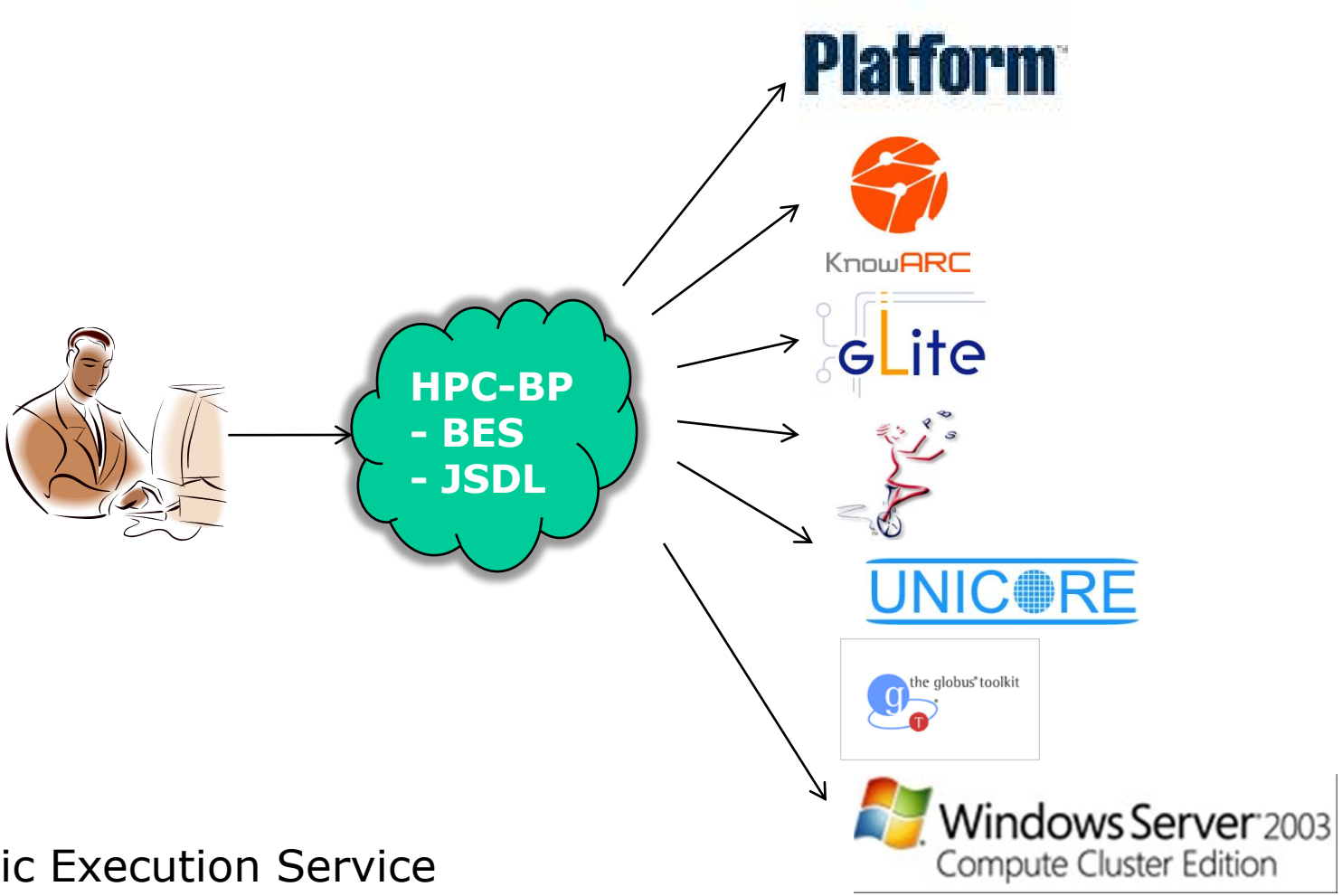
Unicore CE Architecture

UNICORE



Graphic from: Dave Snelling, Fujitsu Labs Europe, "Unicore Technology", Grid School July 2003

Interoperability – but only basics at first



Various grid middleware interface solutions working towards common standards

Basic Execution Service
- Job submission works, but
- security model, file staging, etc. still need to be resolved

Computing: user expectations?

- Different user scenarios are possible and valid
  - paratrooper mode: come in, take all your equipment (files, executable &c) with you, do your thing and go away
    - you're supposed to clean up, but the system will likely do that for you if you forget. In all cases, garbage left behind is likely to be removed
  - two-stage 'prepare' and 'run'
    - extra services to pre-install environment and later request it
    - see later on such Community Software Area services
  - don't think but just do it
    - blindly assume the grid is like your local system
    - expect all software to be there
    - expect your results to be retained indefinitely
    - … realism of this scenario is unclear for 'production' grids
      - it does not scale to larger numbers of users
      - but large user communities hold 'power' over the resource providers (or the customers run away)

Using these systems

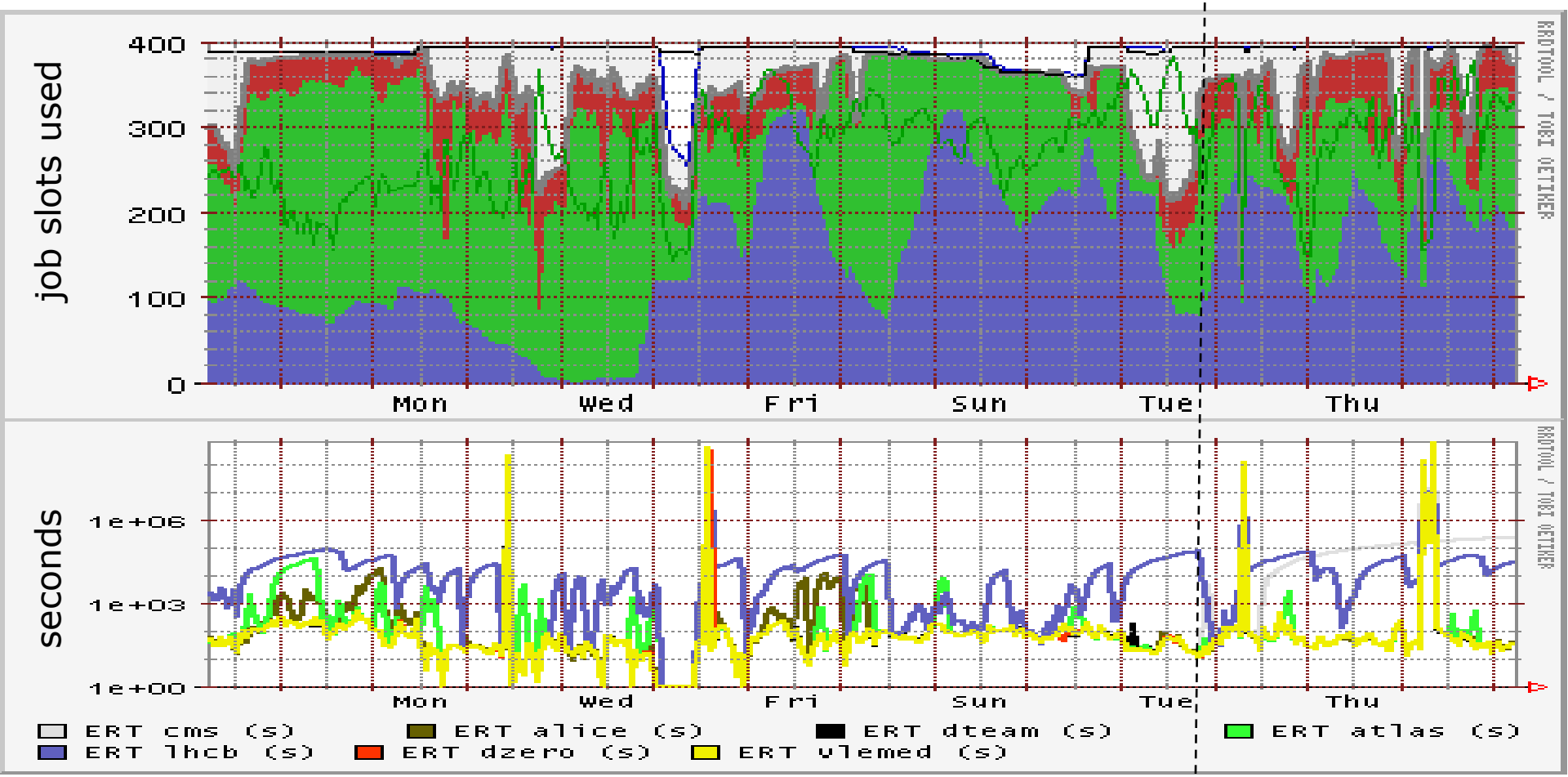
- As both clusters and capability systems are both 'expensive' (i.e. not on your desktop), they are resources that need to be *scheduled*

- interface for scheduled access is a *batch queue*
  - job submit, cancel, status, suspend
  - sometimes: checkpoint-restart in OS, e.g. on SGI IRIX
  - allocate #processors
    (and amount of memory, these may be linked!)
    as part of the job request

- systems usually also have smaller interactive partition
  - more a 'user interface', not intended for running production jobs …

Fair shares and estimated response time

local 'fair shares', used to satisfy overall SLA requirements, need to be translated to an 'estimated response time' for the grid VOs and groups – an unsolved problem

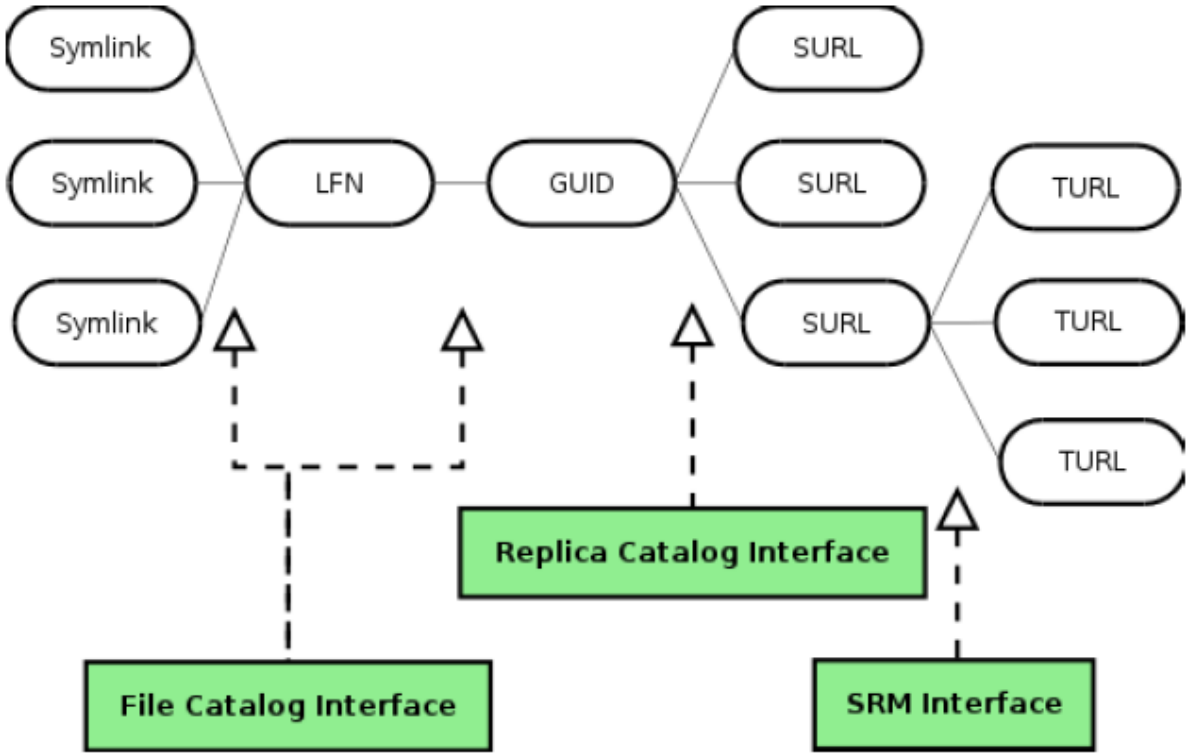


GlueCEStateEstimatedResponseTime(*VO*, *t*)

Storage

- More complex than computing
  - (but will not talk about it much here …)
- Also here: different types and back ends
  - Simple disks: file system, GPFS, Lustre, …
  - MSS: DMF, HPSS, dCache/Enstore, CASTOR, …
- Separate functions of storage
  1. Presentation: file system view, logical naming
  2. Storage resource management:
     relocation, pinning, routing -- SRM
  3. Transfer protocols:
     GridFTP, Byte-IO, gsidcap, gsirfio
  4. Storage:
     file system, tape libraries
  - Today, the grid interfaces expose all of these levels …
    and, e.g., NFSv4 tried to combine all of that …

Storage layering and interfaces



graphic: Peter Kunszt, EGEE DJRA1.4 gLite Architecture

How to you see the Grid?

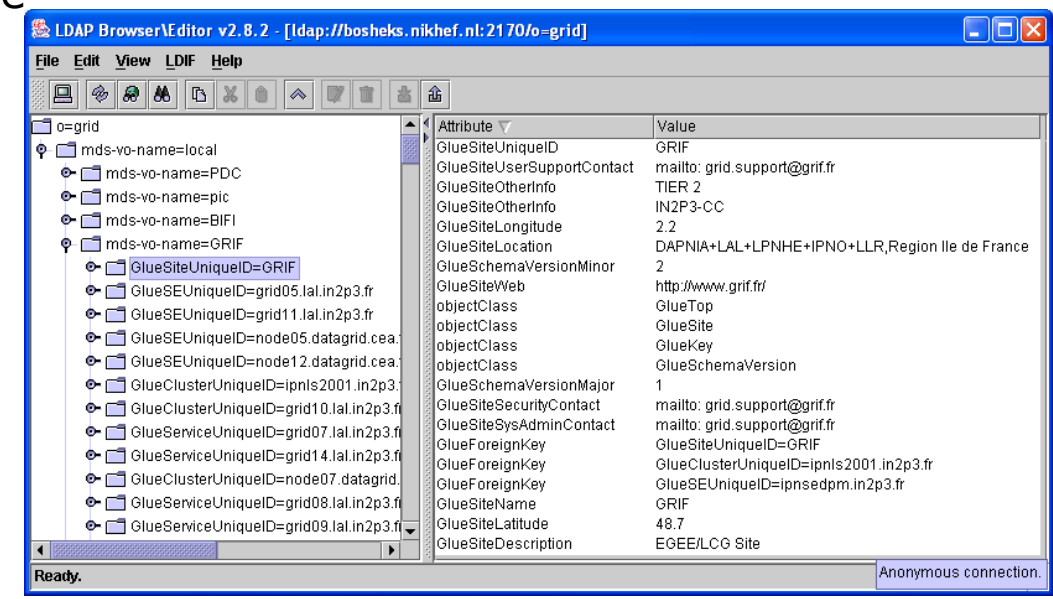
Broker matches the user's request with the site

- 'information supermarket' matchmaking (using Condor Matchmaking)
- uses the information published by the site

Grid Information system

'*the only information a user ever gets about a site*'

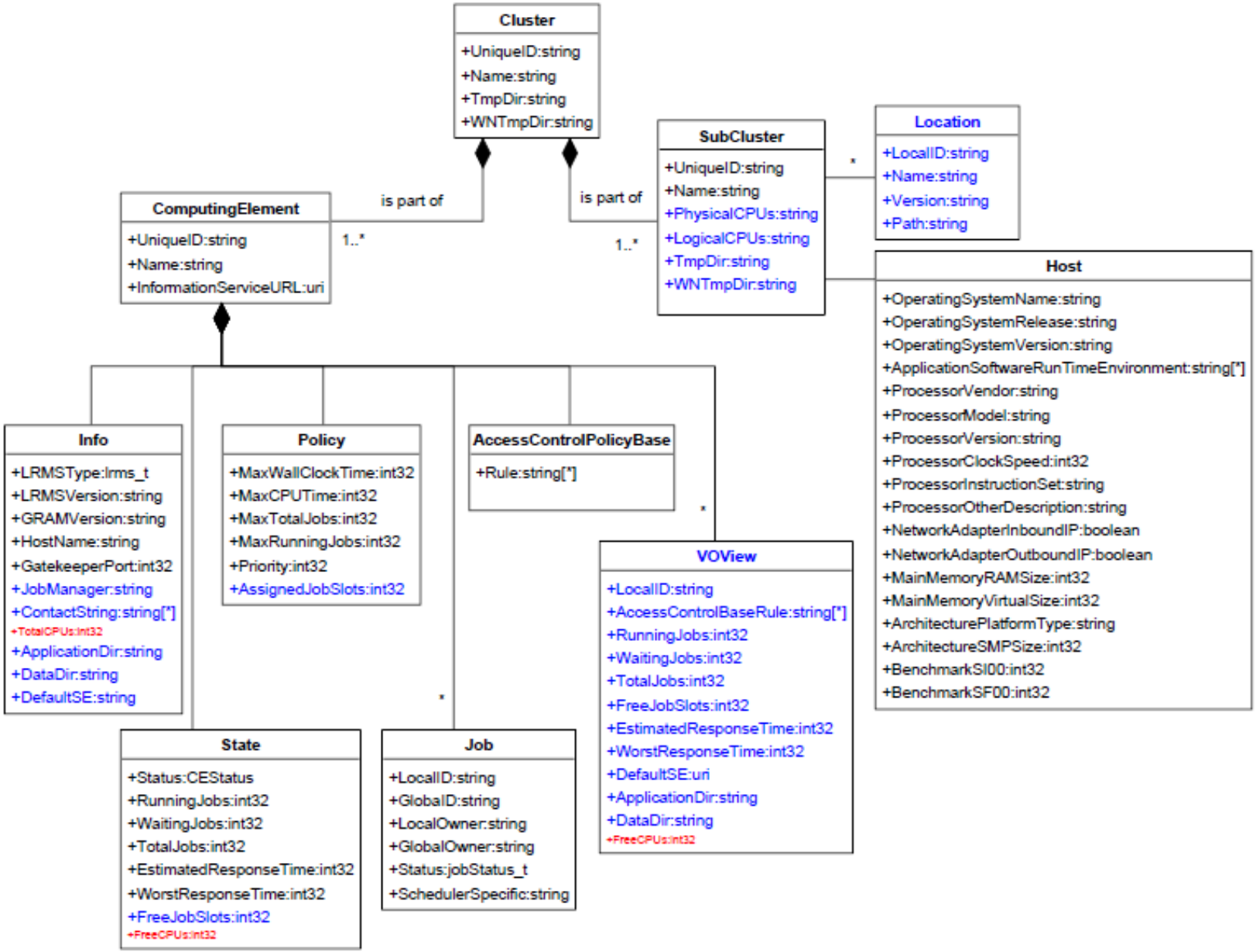
- So: should be 'reliable', consistent* and complete*
- Standard schema (GLUE) to describe sites, queues, storage *(complex schema semantics)*
- Usually presented as an LDAP directory



LDAP Browser Jarek Gawor: www.mcs.anl.gov/~gawor/ldap

Information system and brokering issues

- Without the information system, the user is 'blind' on the grid
- Size of information system scales with #sites and #details
  - already 12 MByte of LDIF
  - matching a job takes ~15 sec
  - Static and dynamic information is mixed ← this is ReallyBad™
- Scheduling policies are infinitely complex
  - no static schema can likely express this information
  - but negotiation processes take time at each request
    WS-Agreement is not really popular, at least not yet …
- Much information (still) needs to be set-up manually ☹
  … anything human will go wrong
- Broker tries to make optimal decision based on this information
  … but a `reasonable' decision would have been better



From: the GLUE Information Model version 1.2, see document for details

Glue Attributes Set by the Site

- Cluster info

  GlueSubClusterUniqueID=gridgate.cs.tcd.ie

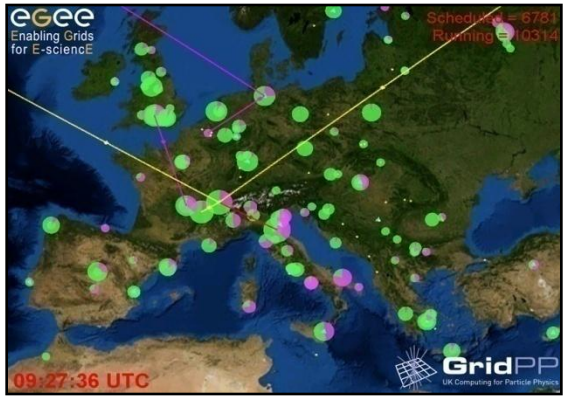
  HostApplicationSoftwareRunTimeEnvironment: VO-atlas-release-10.0.4
  HostBenchmarkSI00: 1300
  GlueHostNetworkAdapterInboundIP: FALSE
  GlueHostNetworkAdapterOutboundIP: TRUE
  GlueHostOperatingSystemName: RHEL
  GlueHostOperatingSystemVersion: 3
  …

- Scheduler status information per VO

  GlueCEStateEstimatedResponseTime: 519
  GlueCEStateRunningJobs: 175
  GlueCEStateTotalJobs: 248

- Storage has similar info
  (paths, max number of files, quota, retention, …)

Working at scale



Grid is an error amplifier …

'passive' controls are needed
to push work away
from failing resources

Failure-ping-pong – or *creeper and reaper* revisited

Resource information systems are the
backbone of any real-life grid

Grid is much like the 'Wild West'

- almost unlimited possibilities – but as a community plan for scaling issues, and a novel environment
- users and providers *need to interact* and articulate needs

Example: GlueServiceAccessControlRule

```
For your viewing pleasure: GlueServiceAccessControlRule
261 distinct values seen for GlueServiceAccessControlRule

(one of) least frequently occuring value(s):
1 instance(s) of GlueServiceAccessControlRule:
   /C=BE/O=BEGRID/OU=VUB/OU=IIHE/CN=Stijn De Weirdt

(one of) most frequently occuring value(s):
   310 instance(s) of GlueServiceAccessControlRule: dteam

(one of) shortest value(s) seen:
   GlueServiceAccessControlRule: d0

(one of) longest value(s) seen:
   GlueServiceAccessControlRule: anaconda-ks.cfg configure-
   firewall install.log install.log.syslog j2sdk-1_4_2_08-
   linux-i586.rpm lcg-yaim-latest.rpm myproxy-addons myproxy-
   addons.051021 site-info.def site-info.def.050922 site-
   info.def.050928 site-info.def.051021 yumit-client-2.0.2-
   1.noarch.rpm
```

Example: GlueHostOperatingSystemRelease

```
Today's attribute: GlueHostOperatingSystemRelease
      1  GlueHostOperatingSystemRelease: 3.02
      1  GlueHostOperatingSystemRelease: 3.03
      1  GlueHostOperatingSystemRelease: 3.2
      1  GlueHostOperatingSystemRelease: 3.5
      1  GlueHostOperatingSystemRelease: 303
      1  GlueHostOperatingSystemRelease: 304
      1  GlueHostOperatingSystemRelease: 3_0_4
      1  GlueHostOperatingSystemRelease: SL
      1  GlueHostOperatingSystemRelease: Sarge
      1  GlueHostOperatingSystemRelease: sl3
      2  GlueHostOperatingSystemRelease: 3.0
      2  GlueHostOperatingSystemRelease: 305
      4  GlueHostOperatingSystemRelease: 3.05
      4  GlueHostOperatingSystemRelease: SLC3
      5  GlueHostOperatingSystemRelease: 3.04
      5  GlueHostOperatingSystemRelease: SL3
     18  GlueHostOperatingSystemRelease: 3.0.3
     19  GlueHostOperatingSystemRelease: 7.3
     24  GlueHostOperatingSystemRelease: 3
     37  GlueHostOperatingSystemRelease: 3.0.5
     47  GlueHostOperatingSystemRelease: 3.0.4
```

The Most Popular Site Location

Compute Clusters

The Impact of Scale

Data versus compute

GRID SITE INFRASTRUCTURE

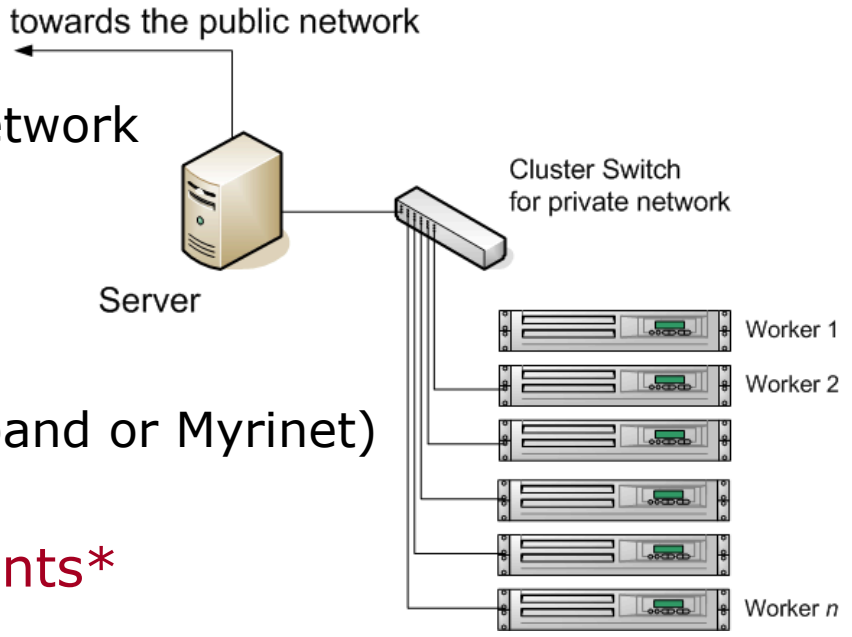
High Performance or High Throughput?

Key question: max. granularity of decomposition:

- Have you got one big problem or a bunch of little ones?
  - To what extent can the "problem" be decomposed into sort-of-independent parts ('grains') that can all be processed in parallel?
- Granularity
  - fine-grained parallelism – the independent bits are small, need to exchange information, synchronize often
  - coarse-grained – the problem can be decomposed into large chunks that can be processed independently
- Practical limits on the degree of parallelism –
  - how many grains can be processed in parallel?
  - degree of parallelism v. grain size
  - grain size limited by the efficiency of the system at synchronising grains
  - IO throughput versus computation

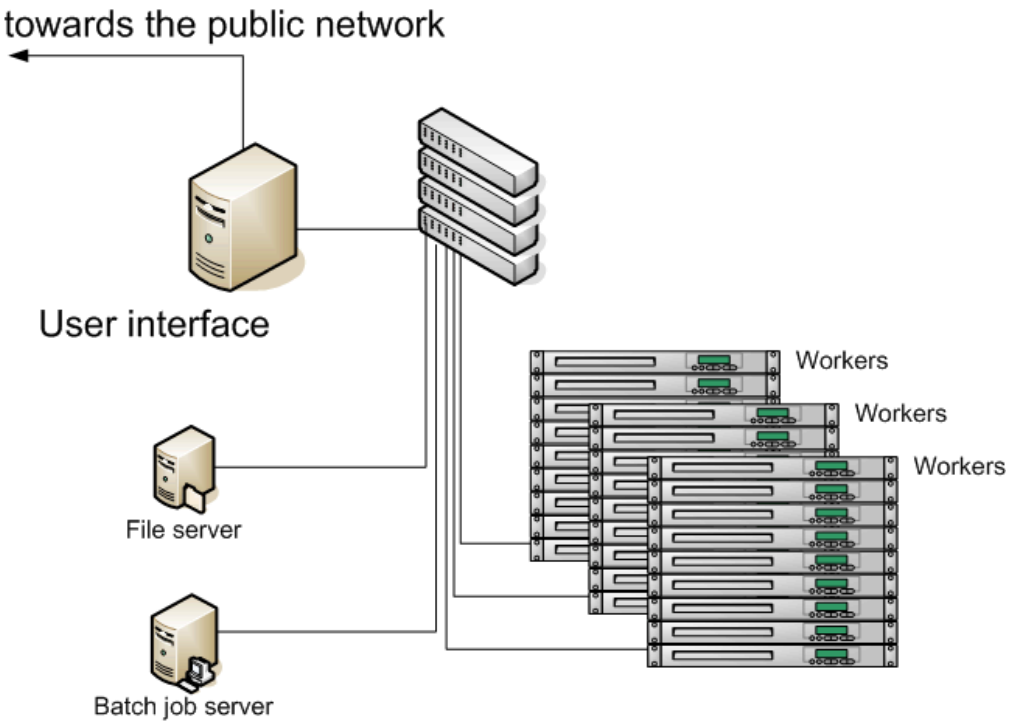
Cluster architectures: Beowulf

- 'Beowulf' virtual supercomputers
  - entire cluster managed by the server
  - users interact only with the server to start and manage jobs
  - geared towards CPU intensive application with few data
- classic network architecture
  - server connected to the public network
  - all WNs on a cluster-local LAN
  - usually using private IP space
  - no communication from the WNs to the outside world
  - optional fast interconnect (Infiniband or Myrinet)
- installs can even use diskless clients*
  - PXE boot refers to NFS root fs
  - all IO is done remotely on the server



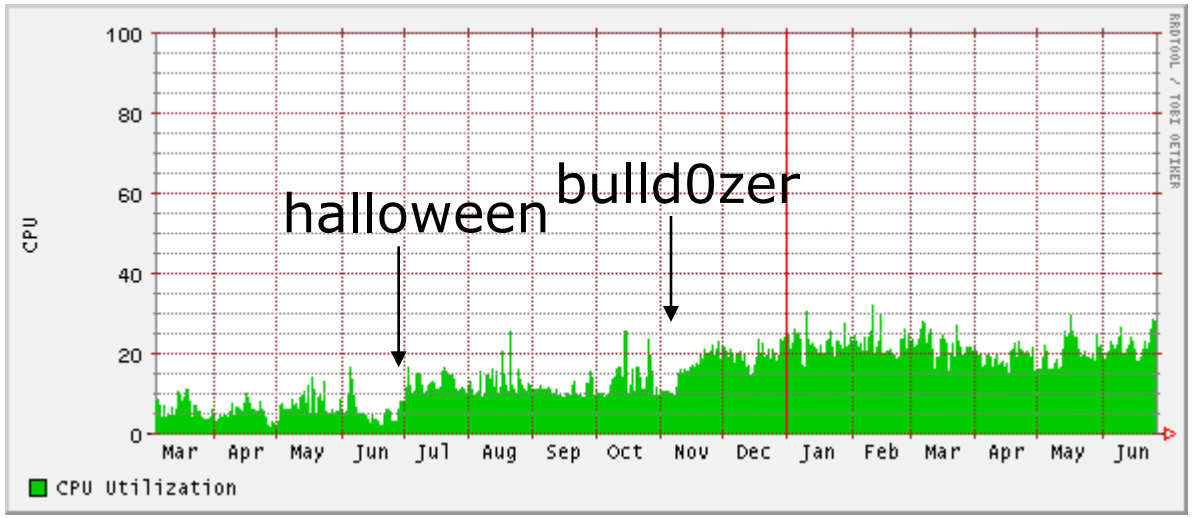
Growing your cluster

- Larger clusters accommodated by more switches, but
  - file I/O (headnode load) becomes bottleneck
    - system booting (PXE, NFS roots)
    - home directories
    - cluster job management
  - function separation (boot server, IO server) within the cluster helps only little
  - Local-IO support is better
  - Not for 'global' use



But NAT does not help

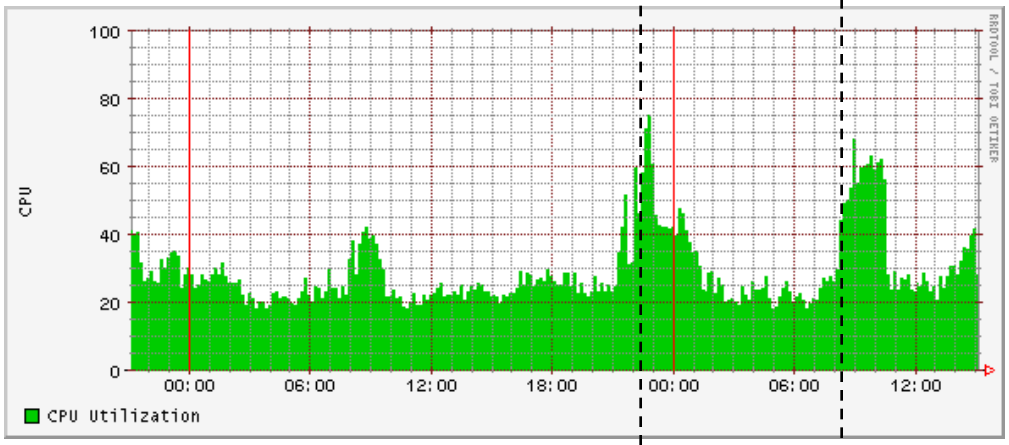
- The NAT kludge leads to several problems
  - with FTP-like protocols for data-transfer
  - with the load on the NAT box
- *and is certainly not the solution for protecting the WNs from attacks from the public internet, as commonly perceived*
  - *can do that easily with 'permit tcp established' followed by 'deny any any'*

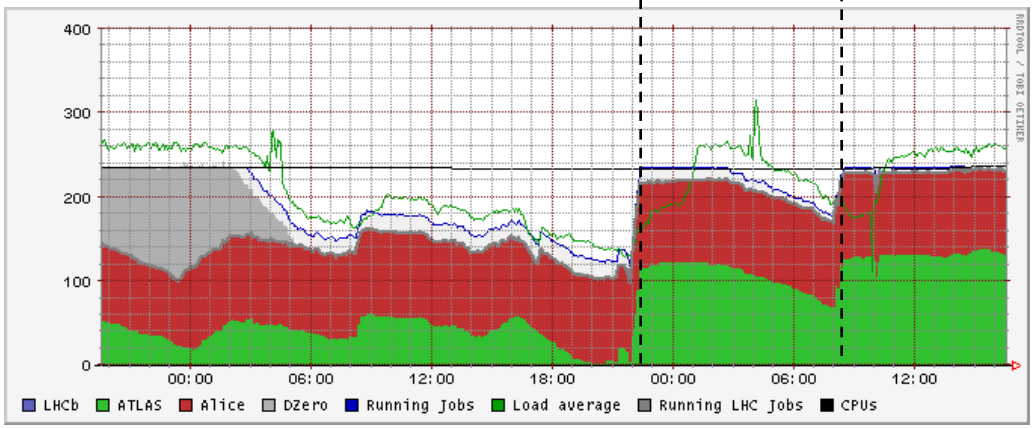


CPU load average 'deel.nikhef.nl'
(Foundry BigIron 15k with 2x BMGR8 Mngt-IV module)

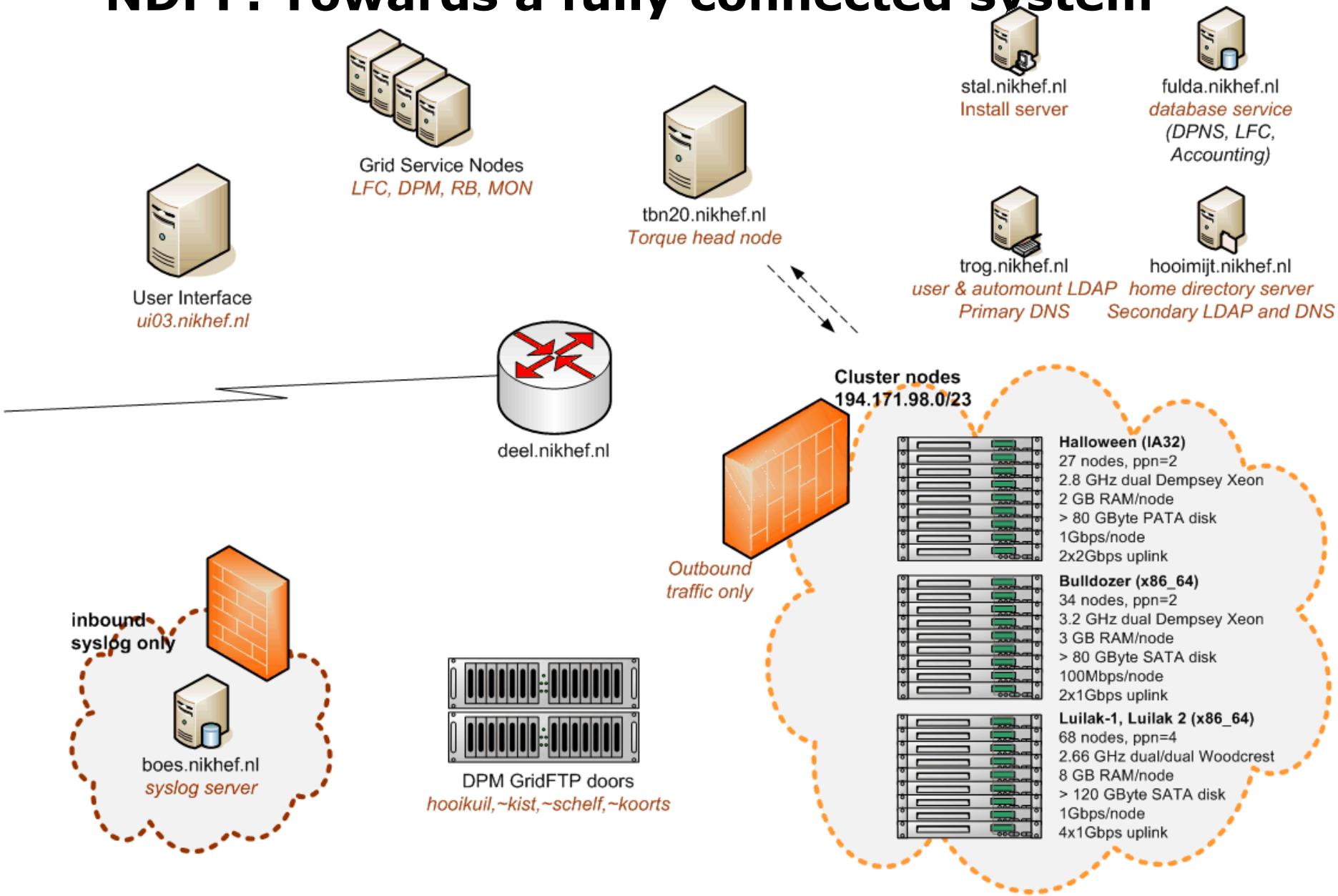
Data intensive jobs

ATLAS HEP jobs retrieving input data sets





NDPF: Towards a fully connected system



Extreme Data Intensive computing

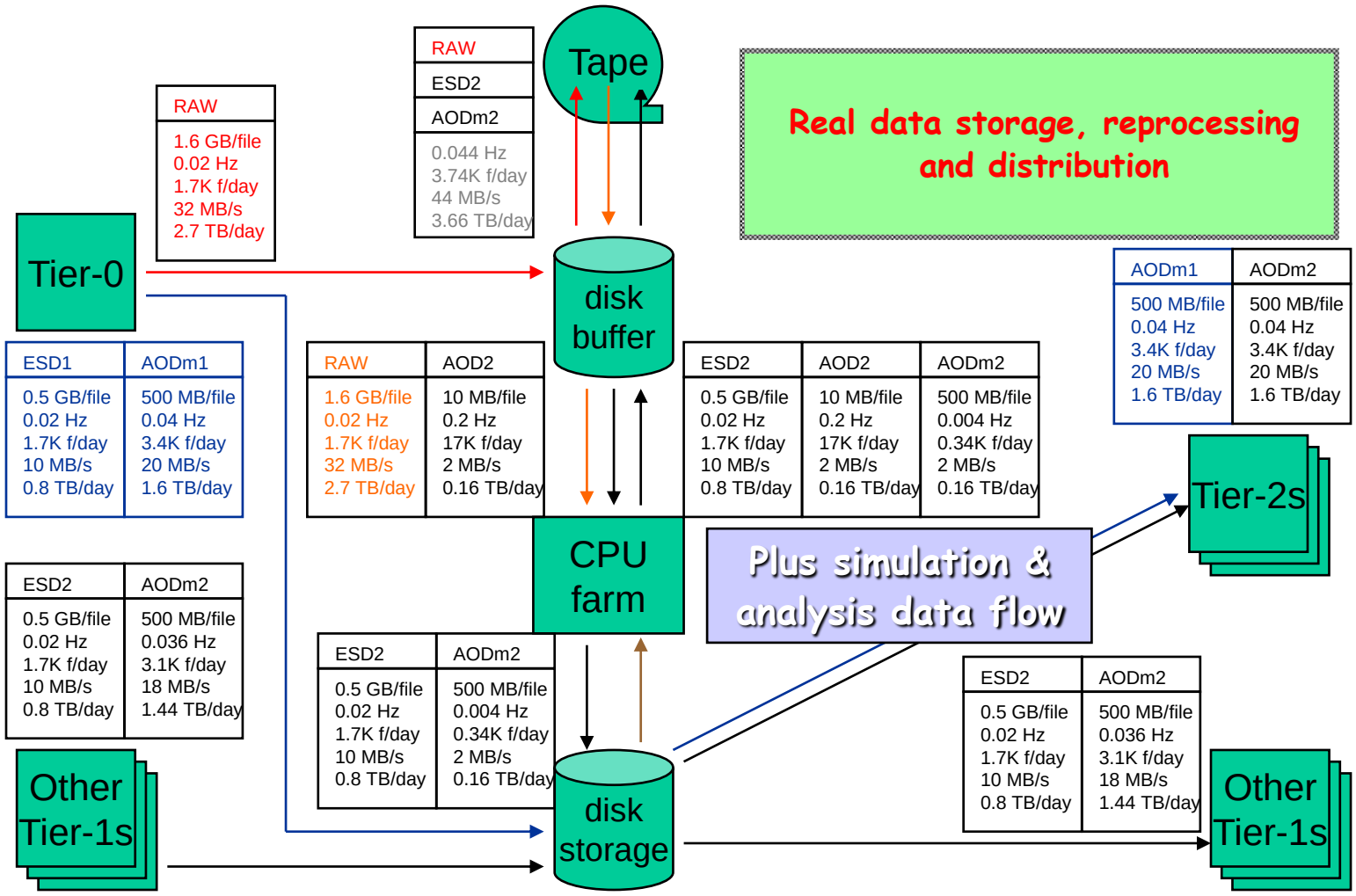
- Most schemes work for mild data-intensive work (1–10 Gbyte/hr in-job, ~100–1000 instructions/byte)
- For extreme operations (FFT, noise cancellation, etc) with 0.1–1 instructions/byte, you need different things
  - Data partitioning across worker nodes with large disks
  - Multi-tiered storage (RAM, 1st level SSD, 2nd level SAS/SATA)
  - If the application cannot be decomposed: cluster file systems (Lustre, GFS, GPFS, …)
  - Extremely fast interconnect can help as well (block device access over QDR infiniband)
- If you just need lots of compute but limited data, PS3 clusters are also nice ☺

The LHC OPN

OPN Routing Creativity

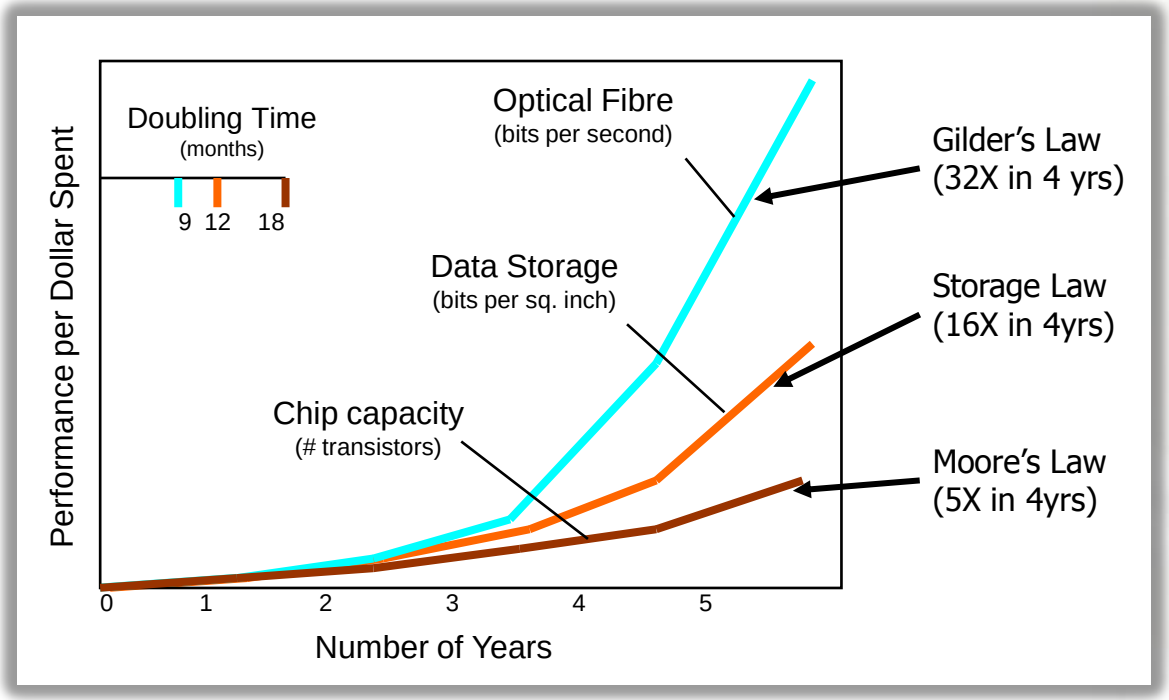
NETWORK

Remembering the Atlas Tier-1 data flows



ATLAS data flows (draft). Source: Kors Bos, NIKHEF

There's always a network close to you



NL Light

SURF NET

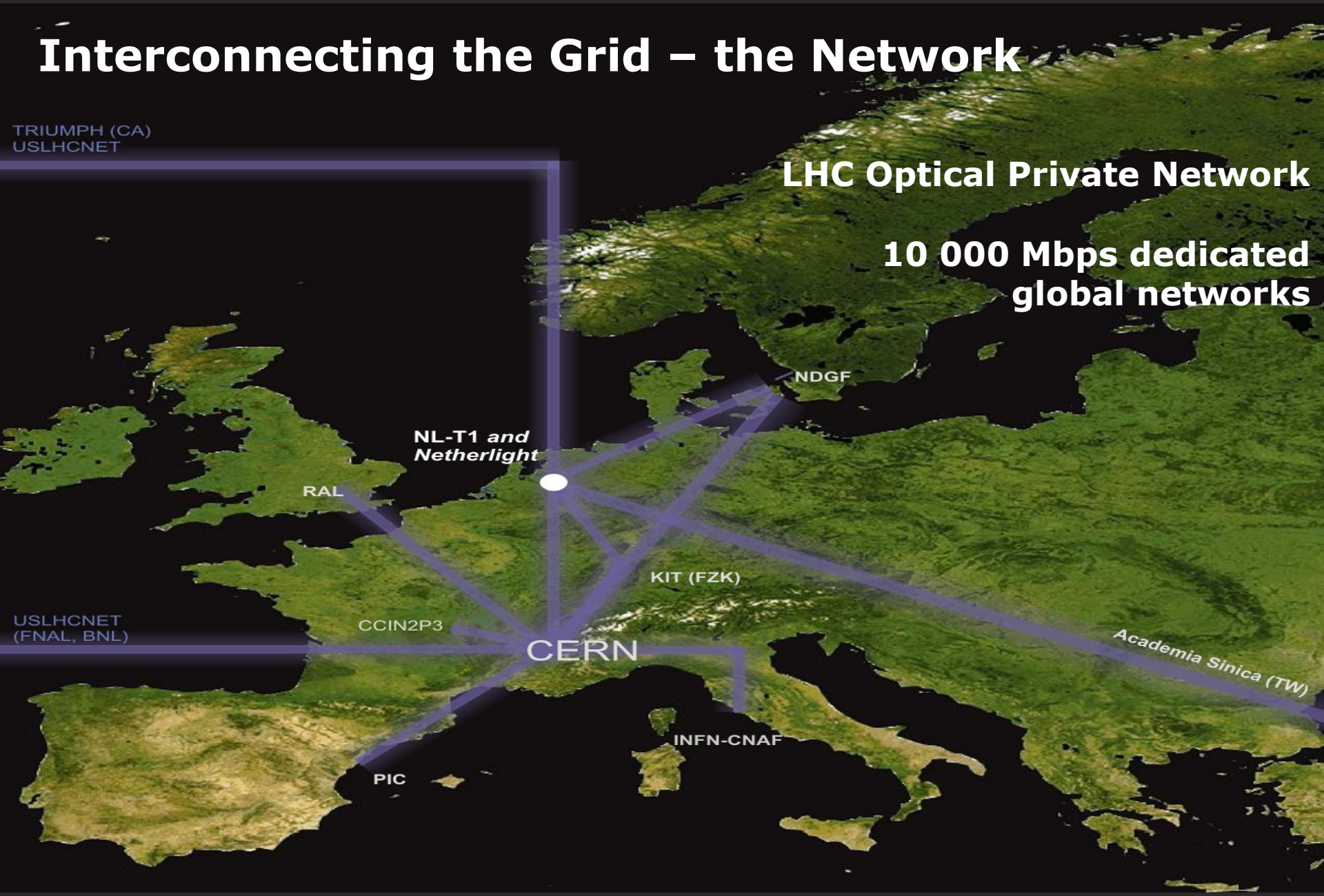
SURFnet pioneered 'lambda' and hybrid networks in the world

- and likely contributed to the creation of a market for 'dark fibre' in the Netherlands

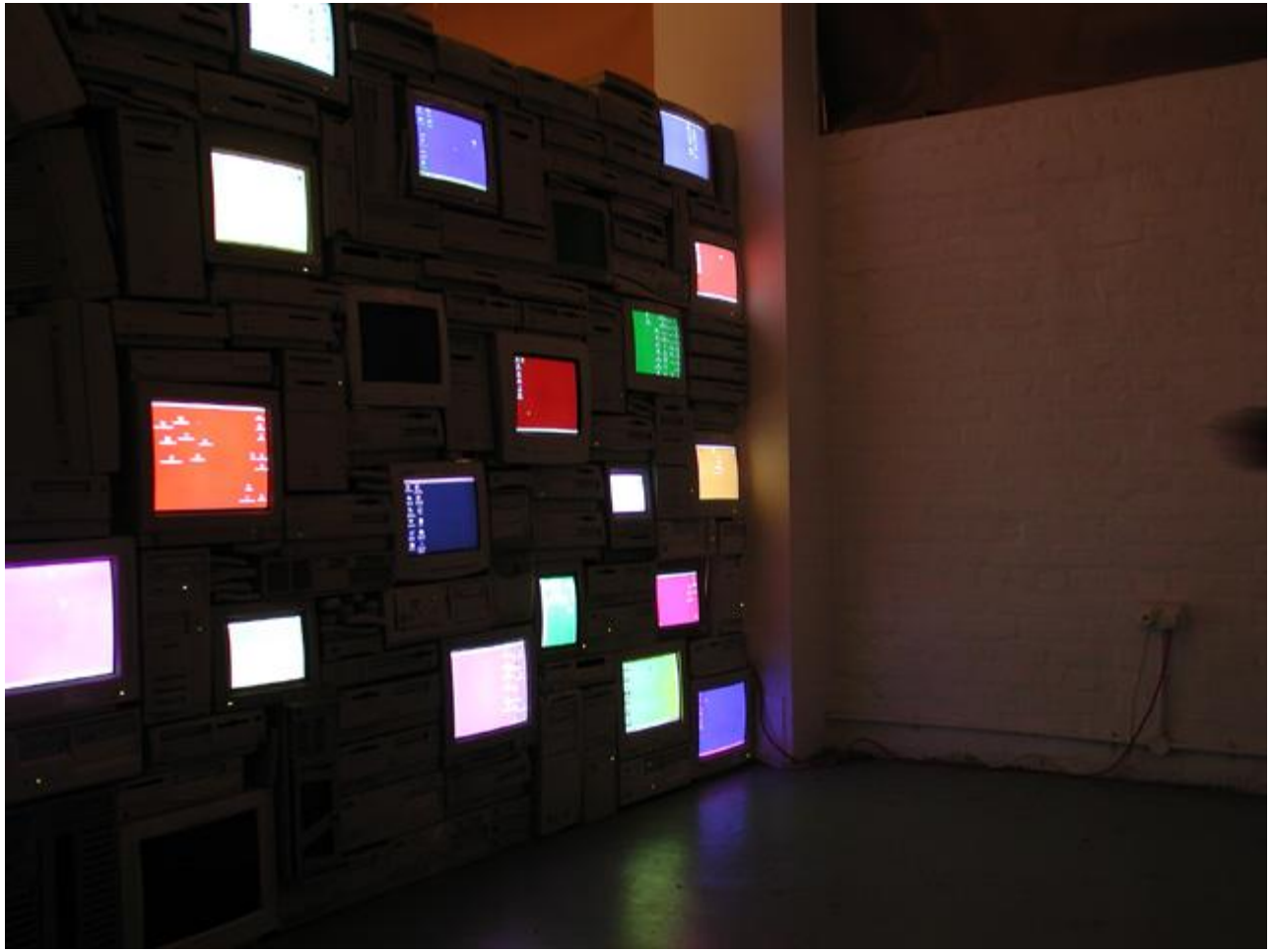
There's always fibre within 2 miles from you – where ever you are!

(it's just that last mile to your home that's missing – and the business model of your telecom provider…)

Interconnecting the Grid – the Network



Firewall



"*Firewall*" by Sandy Smith,
www.computersforart.org

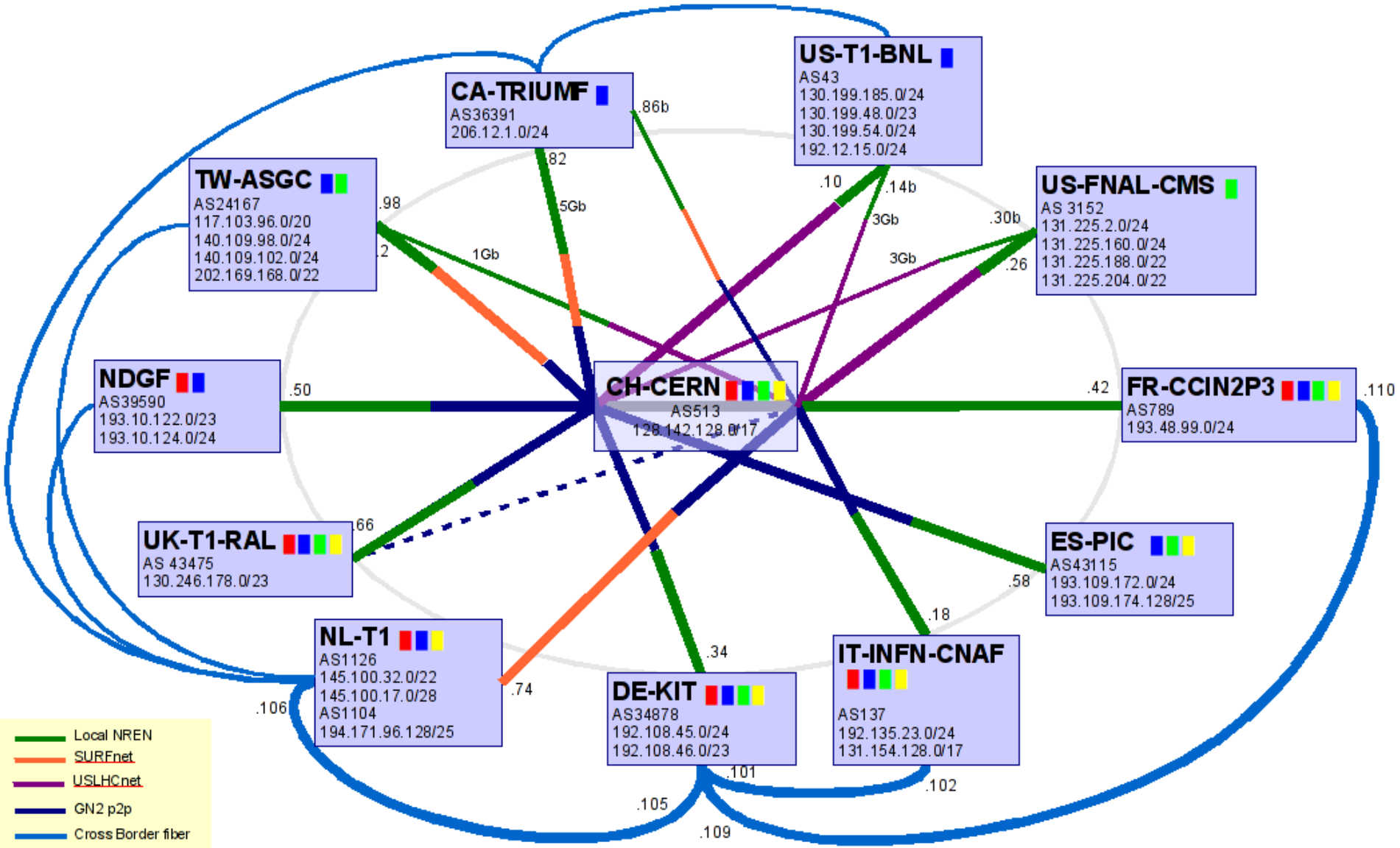
Streams and Firewalls

- Data transfer target:
  300 MByte/s out of CERN to **each** of the ~10 T1s
  - 24 GBit/s aggregate bandwidth
  - you cannot traverse firewalls at that speed
  - For those of you who still believe in firewalls

- OPN – an Optical Private Network for the LHC
  - internal routing only (BGP)
  - all participants sign up to a common policy
  - exclusively for data transfers
  - no direct connections to 'The Internet'
  - allow un-firewalled connection

"*Firewall*" by Sandy Smith, www.computersforart.org

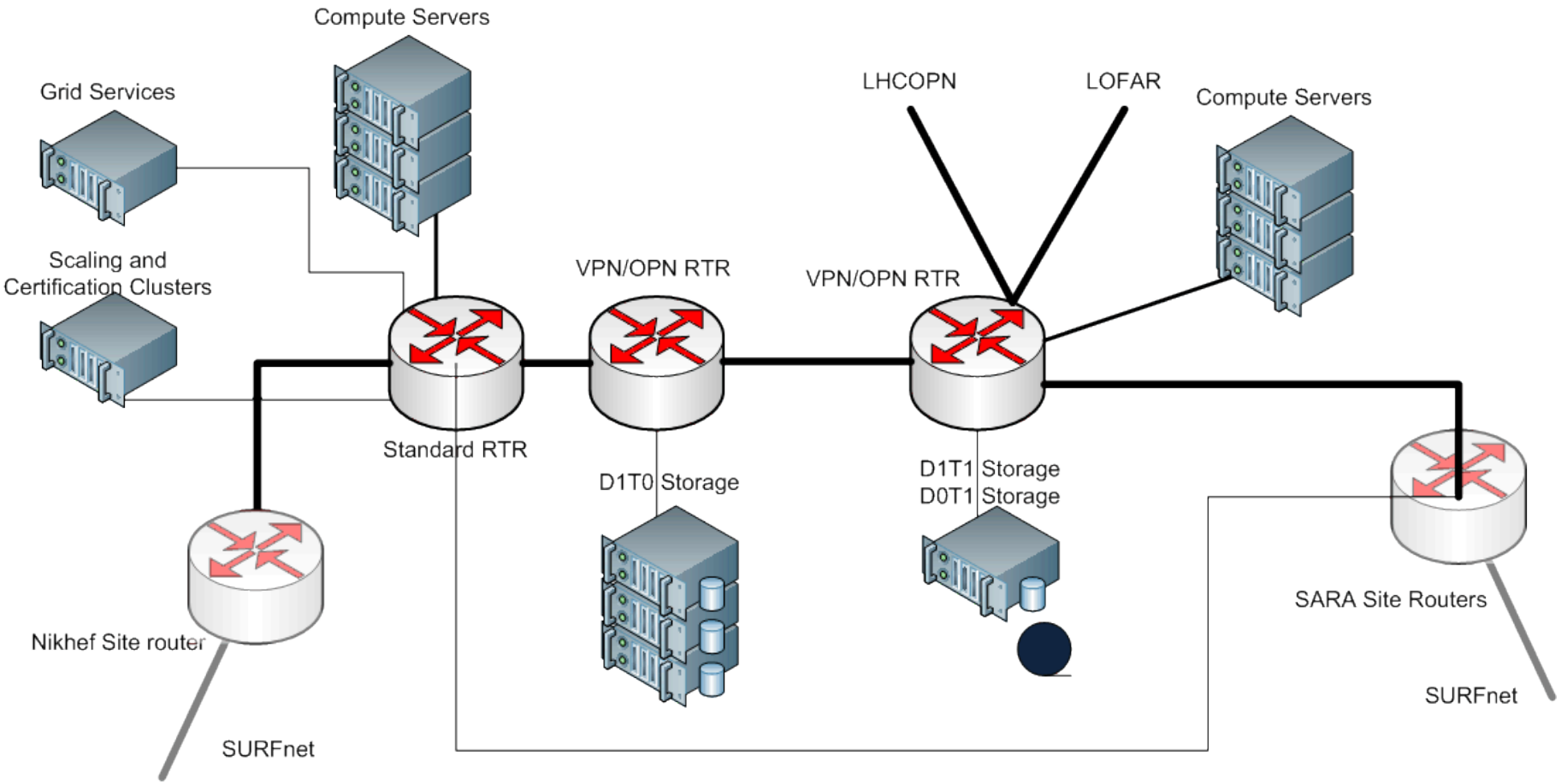
LHCOPN – current status



edoardo.martelli@cern.ch 20090130

Policy impact of the OPN

- Since only storage systems (not WNs) may use the OPN, router needs to distinguish between the two classes
  - If you have a single core router in your grid cluster where you want to terminate the OPN, you are almost forced to use source-based routing
  - but then you loose the features of BGP for fail-over &c
  - since a single router has a single routing policy, you need a *second* router to get the policy right …
  - With two independent OPNs, you need 3 routers
  - With three independent OPNs, you need 4 routers
  - ...
  - you actually need virtual routers in your box ☺



From: BiGGrid Network Plan 2008 BIGGRID-TC-2008-015

Managing many heterogeneous systems

OS level tricks

Procuring your systems: Help! I'm a publicly (co)funded shop ...

SCALING UP: SYSTEMS MANAGEMENT

Think BIG

Examples: CERN Computer Centre

- not only systems management
- but also asset mngt and facilities
- *and you are not even allowed to look inside Google's data centres!*

A undue warm and fuzzy feeling

- More nodes means more power
  - TCO over 3 years at Nikhef/Sara determined by
  - 50% investment, 10% floor space, 40% power (*approximate figures)*
  - But installing power is far more time consuming than buying computers or disk
- But in tender processes, vendors find 'power' the most difficult thing to measure
  - Under what load conditions?
  - What is 'maximum load' – or how to put a system is 'realistic' (~70%) utilization?
  - What is measured: kVA or kW?
  - What is for you the most critical factor? …

Installation

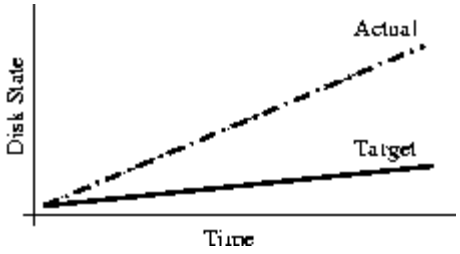
Installing and managing a large cluster requires a system that

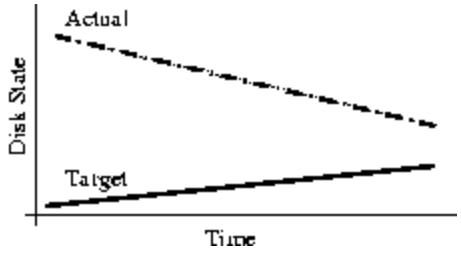
- Scales to $\mathcal{O}$(10 000) nodes, with
  1. a wide variety in configuration ('service nodes')
  2. and also many instances of identical systems ('worker nodes')
- Is *predictable* and *consistent*
- Can rapidly recovery from node failures
  by commissioning a new box (i.e. in minutes)
- Preferably ties in with monitoring and recovery ('self-healing')

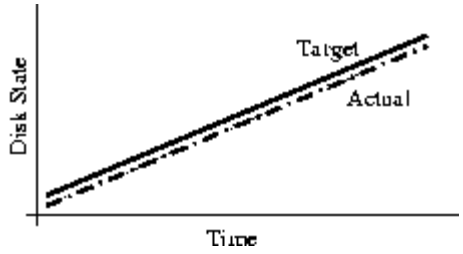
Popular systems include

- Quattor
- xCAT, NPACI Rocks
- SystemImager & cfEngine
- LCFGng

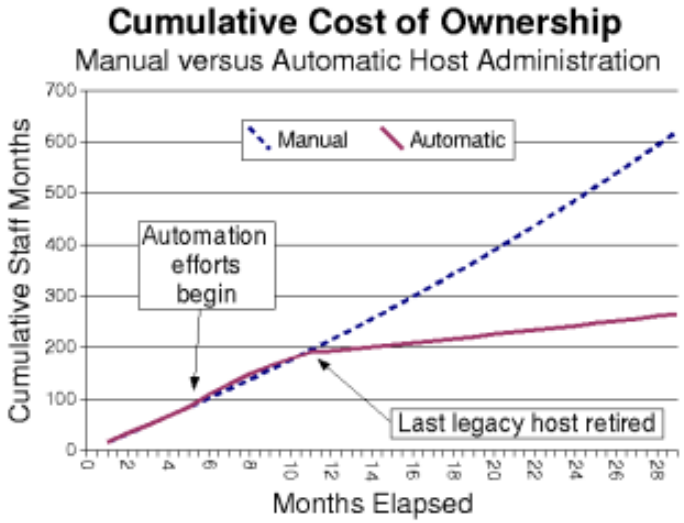
Divergent, Convergent, and Congruent Systems







- Different characteristics
  - Incremental: cfengine, LCFGng
  - Deterministic by re-install: xCAT, Rocks
  - Ordered transactional: Quattor
- Can a self-modifying system reach consistent (or even stable) state without repeatable deterministic ordering of changes on a host?



See also
http://www.infrastructures.org/papers/turing/turing.html
(figures are from paper referenced)

Managing complexity: Quattor

- Designed to manage O(10 000) systems that
  - may all have different configurations
  - but whose configuration is known at any time
- Obeys the deterministic requirement
  - different parts of a configuration have explicit dependencies
- Can configure almost any kind of system today
  - ~250 different configuration drivers, from Apache to Xen
  - 'Components' ensure that the system will be pushed into the desired state, whatever its current state, autonomously
  - New components are being written weekly (e.g. Nagios configuration, new grid services, …)
- Fully Open Source: see quattor.sf.net
- Used by industry, finance and academia

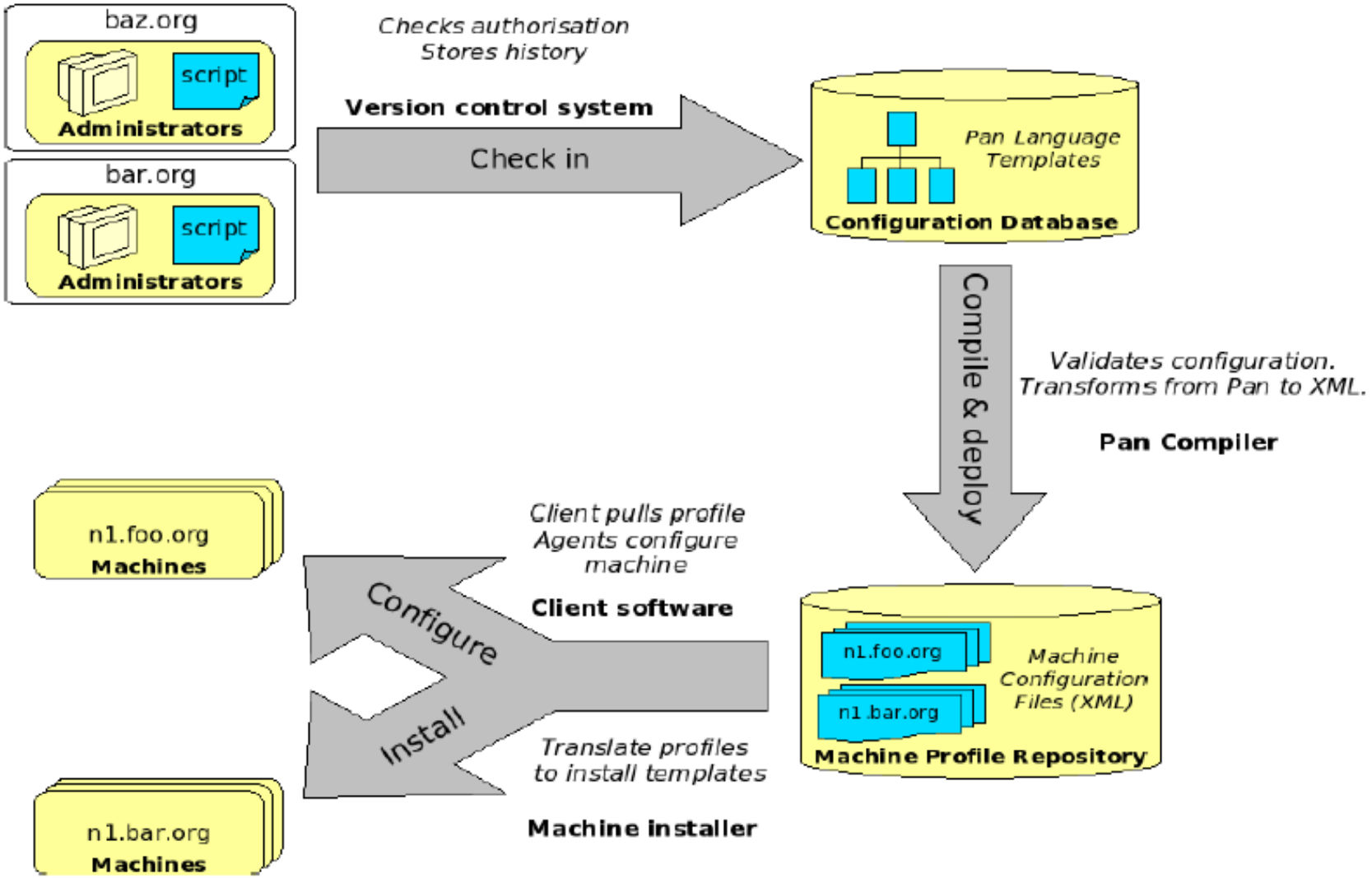
- Quattor was developed to meet the above requirements
  - Aimed to improve on its ancestor LCFG
  - Uses a high-level *declarative* configuration language – *Pan*
    - Hierarchical schema
    - Modularization for data reuse and customization
    - Pre-deployment checks through *validation*
  - Allows different service deployment strategies
  - Provides a full "configuration distribution"
    - Out-of-the box solutions for *gLite* grid services

Table 1: Quattor deployments

	Distributed			Single-site			
Metric	**BEGrid**	**Grid-Ireland**	**GRIF**	**CERN**	**CNAF**	**Nikhef**	**UAM**
Managed machines	260	417	575	8000	800	301	553
Administrators	8	11	25	100	10	4	3
Physical sites	6	18	6	1	1	1	1

Devolved management workflow...

quattor



...Devolved management workflow...

quattor

- Configuration management system
  - Subsystem deployment can be
    - → *Centralized* for strict operation control on the server
      - Sort of broker-based
    - → *Distributed* for more operational flexibility
      - Easier autonomous handling of configuration parts
  - Authentication via X.509/Kerberos5/encrypted passwords
  - Authorization via access control lists (ACLs)
- Automatic installation of managed nodes (all operations can be done remotely)
  - Retrieves information from machine *profiles*
  - Configures DHCP and PXE
  - Generates Kickstart files

...Devolved management workflow

quattor

- Node configuration management
  - Nodes are notified of changes and download fresh profiles
  - Autonomous agents (“components”) triggered by changes in specific parts of the configuration schema
    - Can also deploy manually (automatic dispatching disabled)
    - Pre/post runtime dependencies ensure correct service configuration
  - Idempotent (repeated actions have the same effect)
- Software management
  - Separation of *repository* and *configuration*
    - Different repositories accessed via HTTP
    - Package lists in Pan templates
  - Modes
    - *Strict* -- install only listed packages, remove manual installations
    - *Flexible* -- allow multiple versions, respect manual installation
  - Rollbacks can be easily performed

What works on one machine …

Monitoring is all that matters

THINGS THAT BREAK

User and system directories and maps

Large number of alternatives exists (nsswitch.conf/pam.d)

- files-based (/etc/passwd, /etc/auto.home, …)
- YP/NIS, NIS+
- Database (MySQL/Oracle)
- LDAP

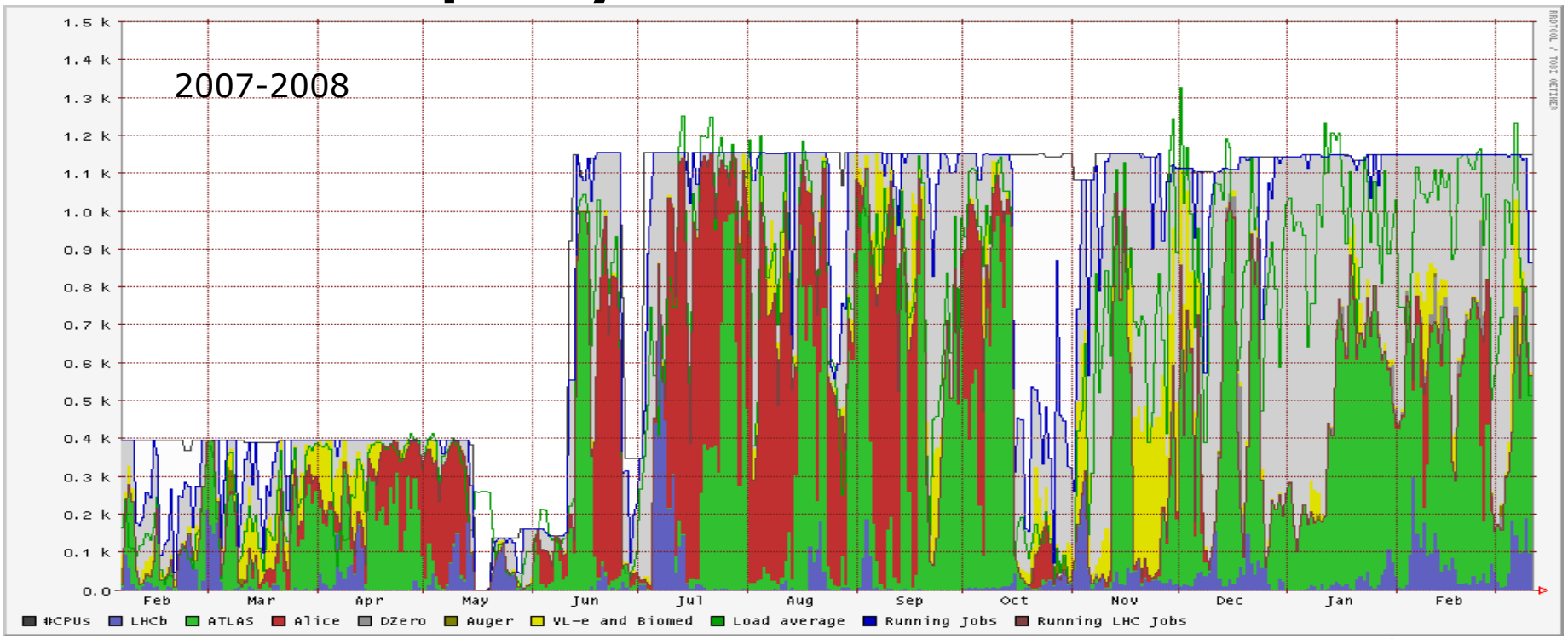
We went with LDAP:

- information is in a central location (like NIS)
- scales by adding slave servers (like NIS)
- is secure by LDAP over TLS (unlike NIS)
- can be managed by external programs (also unlike NIS)
  (we can even do real-time grid credential mapping to and from uid's)

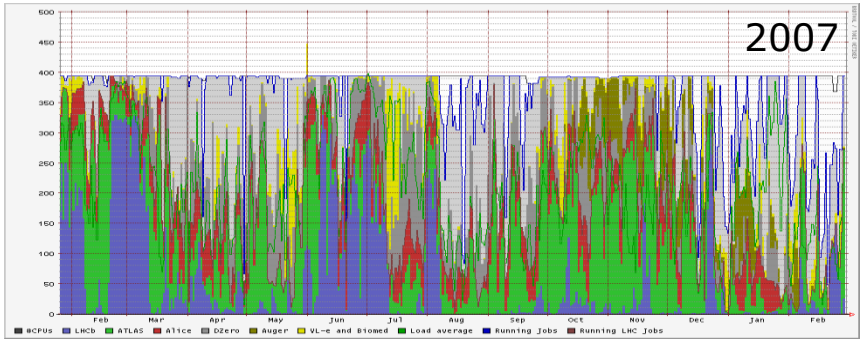
But you will need to run *nscd*, or a large number of slave servers

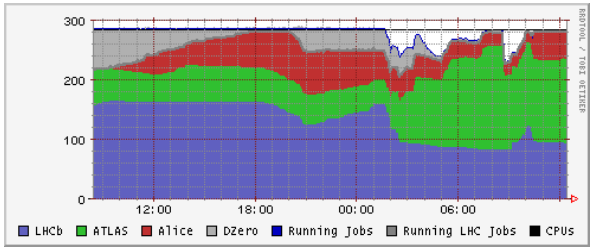
- with *nscd*, a single server can easily handle ~200 nodes/500 cores
- in rare cases, (statically linked) programs run into trouble

NDPF Occupancy

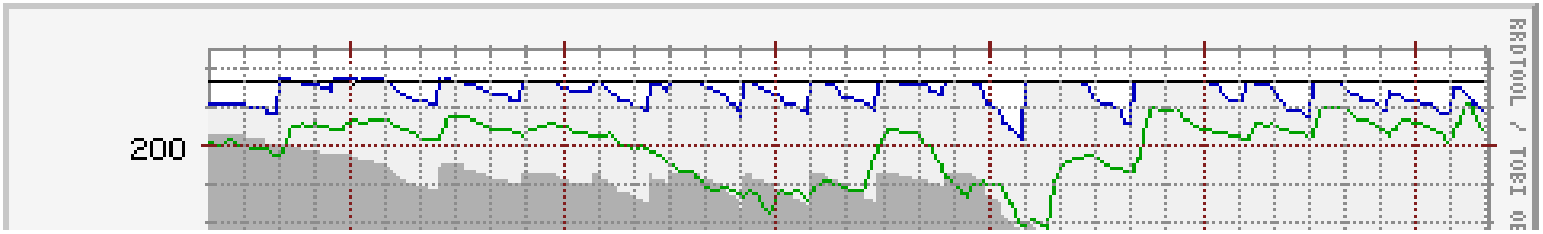


Average occupancy 2007, 2008 > 90%

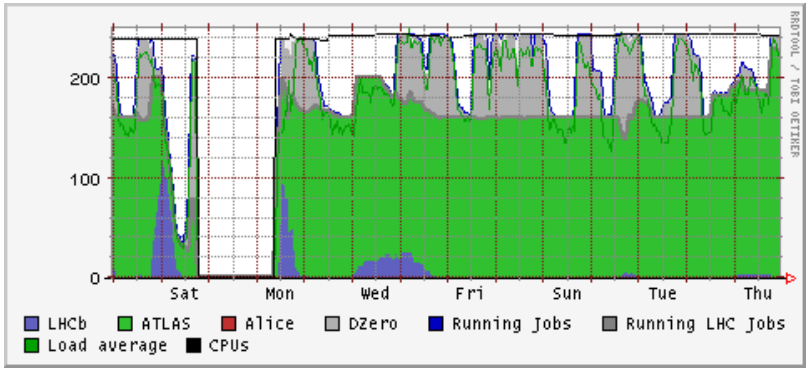




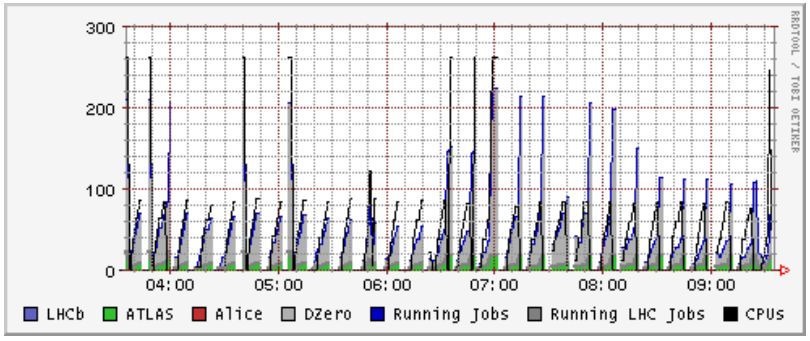
each colour represents a grid VO, black line is #CPUs available



An unresponsive node causes the scheduler MAUI to wait for 15 minutes, then give up and start scheduling again, hitting the rotten node, and …

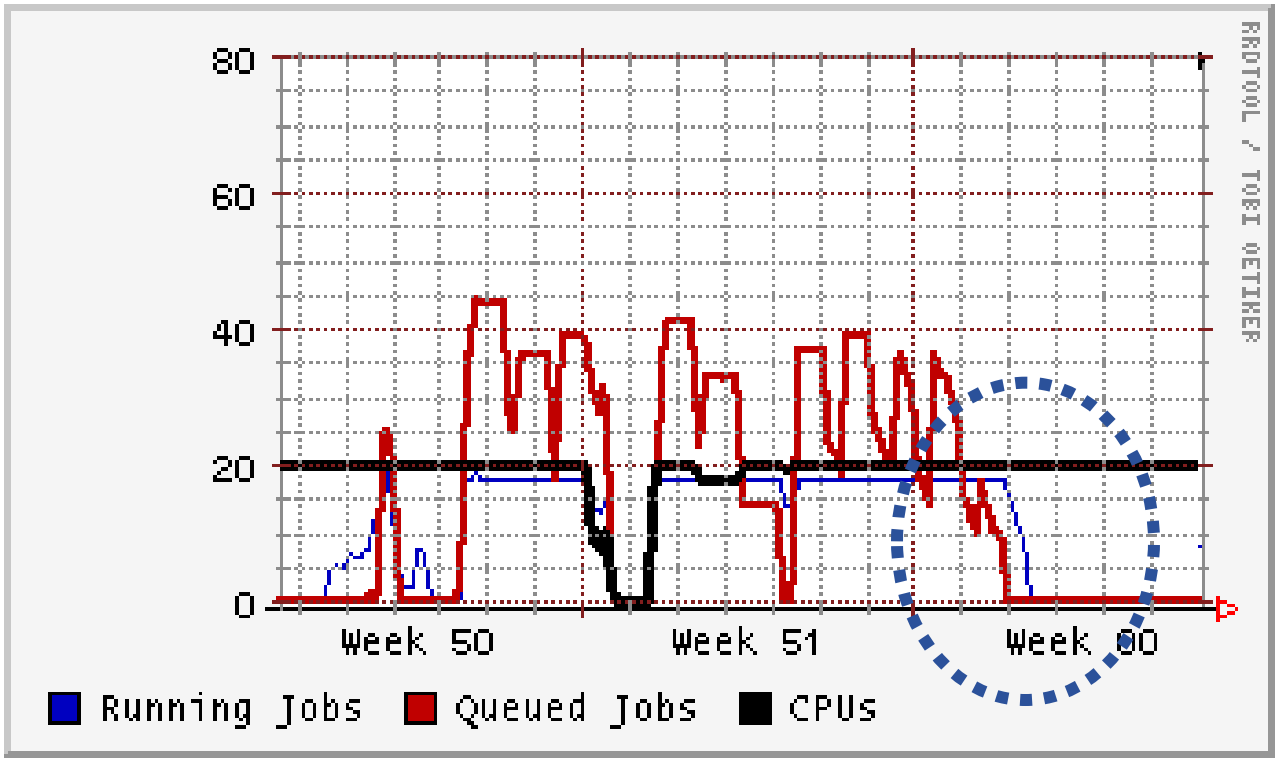


Auditing Indicent: a disk with less than 15% free makes the syscall-audit system panic, new processes cannot write audit entries, which is fatal, so they wait, and wait, and …
a head node has most activity & fails first!



PBS Server trying desparately to contact a dead node who's CPU has turned into Norit … and unable to serve any more requests.

Black Holes



A mis-configured worker node accepting jobs that all die within seconds.
Not for long, the entire job population will be sucked into this black hole…

Clusters: what did we see?

- the Grid (and your cluster) are error amplifiers
  - "black holes" may eat your jobs piecemeal
  - dangerous "default" values can spoil the day ("GlueERT: 0")
- Monitor! (and allow for (some) failures, and design for rapid recovery)

- Users don't have a clue about your system beforehand
  (that's the downside of those 'autonomous organizations')
- If you want users to have clue, you push publish your clues correctly
  (the information system is all they can see)
- Grid middleware may effectively do a DoS on your system
  - doing *qstat* for every job every minute, to feed the logging & bookkeeping …

- And finally: all investments in documentation and tidiness pay off,
  … or your colleague will not find that #$%^$*! machine
  in the middle of night…

Logging and Auditing

Auditing and logging

- syslog (also for grid gatekeeper, gsiftp, credential mapping)
- process accounting (psacct)

For the paranoid – use tools included for CAPP/EAL3+: LAuS

- system call auditing
- highly detailed:
  useful both for debugging and incident response
- default auditing is critical: system will halt on audit errors ☺
  - and once in a while you hit a kernel bug that cannot be reproduced, as we did in RHEL3 ☹

If your worker nodes are on private IP space

- need to preserve a log of the NAT box as well

Grid and Cluster Logging

Grid statistics and accounting

- *rrdtool* views from the batch system load per VO
  - combine *qstat* and *pbsnodes* output via script, cron and RRD
- *cricket* network traffic grapher
- *ganglia* monitoring
- *Nagios* probe-based alarms and (*grid)* monitoring
- extract *pbs accounting data* in dedicated database
  - grid users have a 'generic' uid from a dynamic pool – need to link this in the database to the grid DN and VO
- from accounting db, upload (*anonymized)* records
  - grid accounting system for VOs and funding agencies
  - accounting db also useful to charge costs to projects locally
  - but remember to consider DPA restrictions
    - *define data usage explicitly*
    - *make users agree, and make sure your click-through actually holds up*
    - *don't expose if you don't need to*

Nagios display

Nagios

General
- Home
- Documentation

Monitoring
- Tactical Overview
- Service Detail
- Host Detail
- Hostgroup Overview
- Hostgroup Summary
- Hostgroup Grid
- Servicegroup Overview
- Servicegroup Summary
- Servicegroup Grid
- Status Map
- 3-D Status Map
- Service Problems
- Host Problems
- Network Outages

Show Host:

- Comments
- Downtime
- Process Info
- Performance Info
- Scheduling Queue

Reporting
- Trends
- Availability
- Alert Histogram
- Alert History
- Alert Summary
- Notifications
- Event Log

Configuration
- View Config

Current Network Status
Last Updated: Wed Jun 6 11:53:57 CEST 2007
Updated every 90 seconds
Nagios® - www.nagios.org
Logged in as *nagiosadmin*
- Notifications are disabled

View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
25	0	0	1

All Problems	All Types
0	26

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
158	0	137	24	0

All Problems	All Types
161	319

Service Status Details For All Hosts

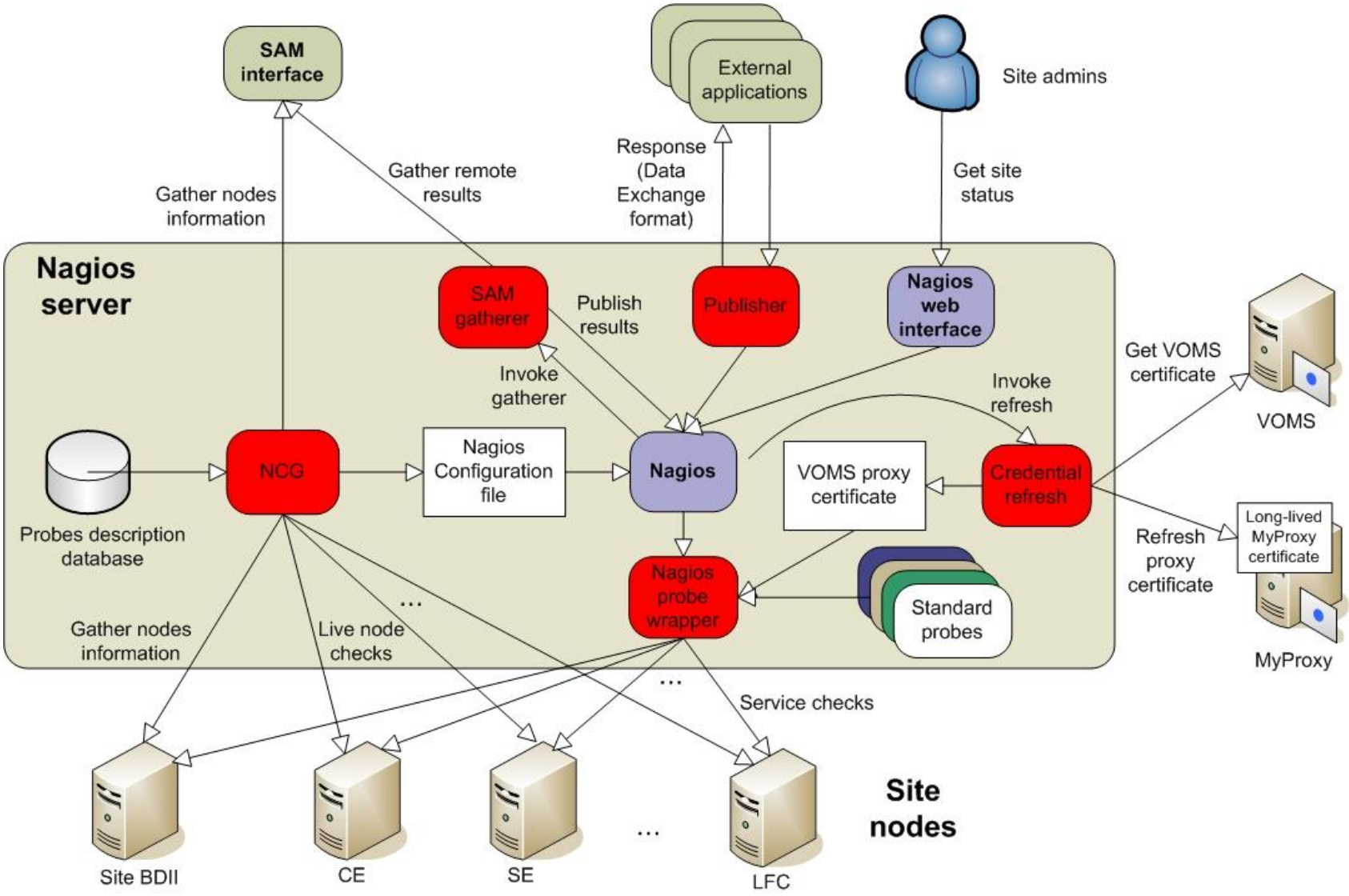
Host	Service	Status	Last Check	Duration	Attempt	Status Information
castorgrid.cern.ch	GridFTP-Ping	OK	06-06-2007 11:53:33	11d 20h 16m 10s	1/4	FTP OK - 0.039 second response time on port 2811 [220 castorgrid04.cern.ch CASTOR GridFTP Server 1.12 GSS/ Globus/GSI wu-2.6.2(cern-2) (gcc32dbg, 1069715860-42
	GridFTP-Transfer	OK	06-06-2007 11:16:03	0d 0h 37m 54s	1/4	Upload to remote computer succeeded. Download from re computer succeeded. File successfully removed from rem computer. Received file is valid.
	SE-host-cert-valid-OPS-remote	OK	06-06-2007 11:38:08	0d 18h 29m 35s	1/1	SAM status: ok
	SE-lcg-cp-Atlas-remote	OK	06-06-2007 11:03:53	0d 18h 49m 54s	1/1	SAM status: ok
	SE-lcg-cp-CMS-remote	OK	06-06-2007 09:59:00	0d 1h 54m 57s	1/1	SAM status: ok
	SE-lcg-cp-DTeam-remote	OK	06-06-2007 11:47:54	0d 18h 21m 11s	1/1	SAM status: ok
	SE-lcg-cp-OPS-remote	OK	06-06-2007 11:00:03	0d 19h 2m 36s	1/1	SAM status: ok
	SE-lcg-cr-Atlas-remote	OK	06-06-2007 11:03:50	0d 18h 49m 59s	1/1	SAM status: ok
	SE-lcg-cr-CMS-remote	OK	06-06-2007 09:58:48	0d 1h 55m 9s	1/1	SAM status: ok
	SE-lcg-cr-DTeam-remote	OK	06-06-2007 11:47:51	0d 18h 21m 14s	1/1	SAM status: ok
	SE-lcg-cr-OPS-remote	OK	06-06-2007 11:00:00	0d 19h 2m 39s	1/1	SAM status: ok
	SE-lcg-del-Atlas-remote	OK	06-06-2007 11:03:56	0d 18h 49m 51s	1/1	SAM status: ok
	SE-lcg-del-CMS-remote	OK	06-06-2007 09:59:05	0d 1h 54m 52s	1/1	SAM status: ok

Host	Service	Status	Duration	Information
NDPF				
! tbn19.nikhef.nl	free disk space /var	critical	57s	CHECK_NRPE: Socket timeout after 10 seconds.

www.nikhef.nl 1 host down 2 service warnings 1 critical service

Graph: Ian Bird

Nagios monitoring in a grid environment



Graph: Ian Bird

Breaking the egg shell approach

Towards policy harmonization

Open Issues: personal data in the grid

SECURITY IN A DISTRIBUTED WORLD

Hardening your cluster

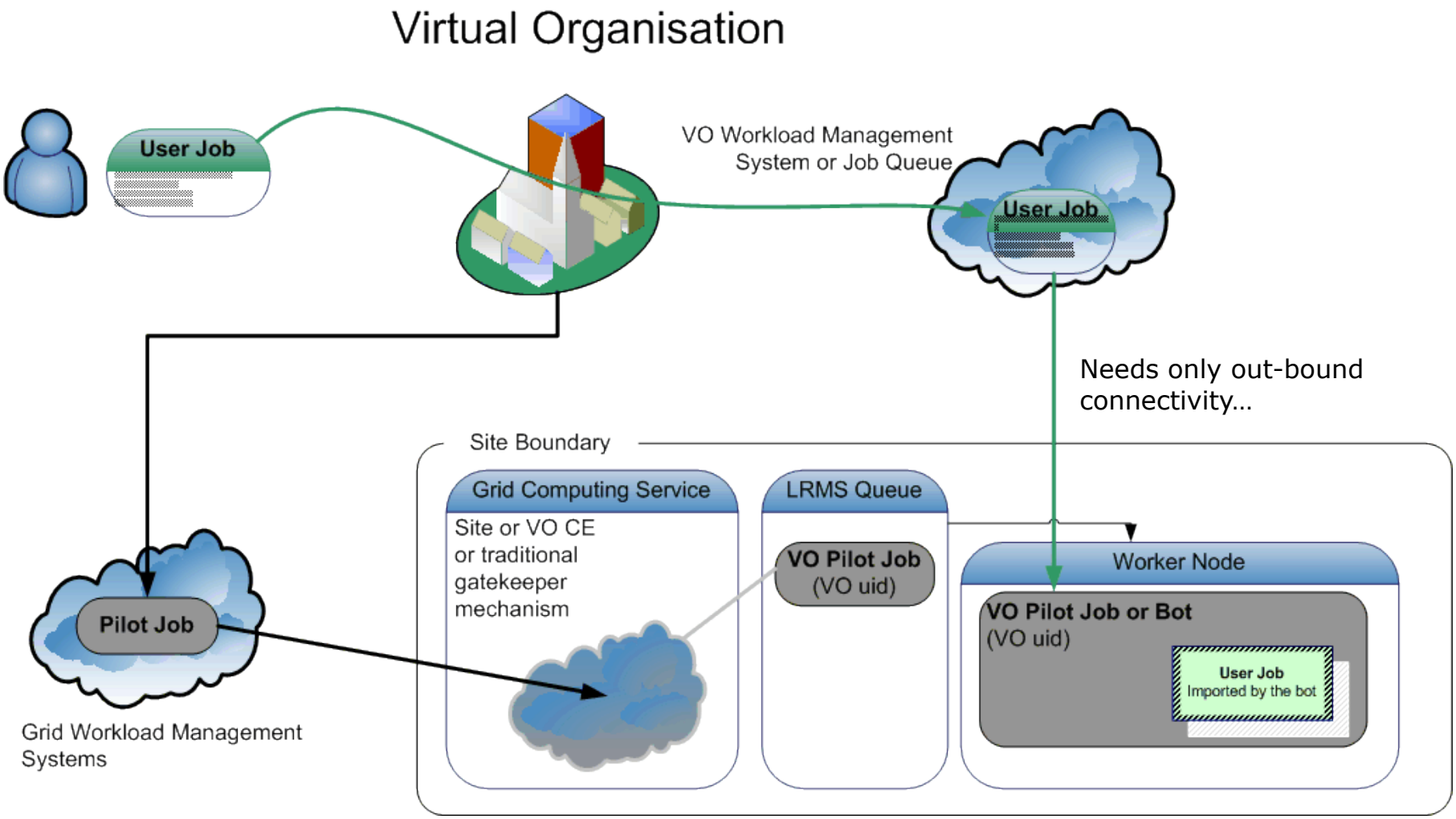
A firewall feels quite secure … initially …

- with grid resource sharing, the 'eggshell' approach breaks
  - local users are no longer local:
    - local exploits will be used
    - malicious users will try to 'escape' from the worker nodes
    - anyway, O(10k) systems in one go is quite attractive ;-)
      and has real market value
  - if you support an 'user interface' system for some remote users to get onto the grid, they *will* the same password as everywhere else
  - the most common attack on distributed clusters today is still the ssh trojans and password sniffers

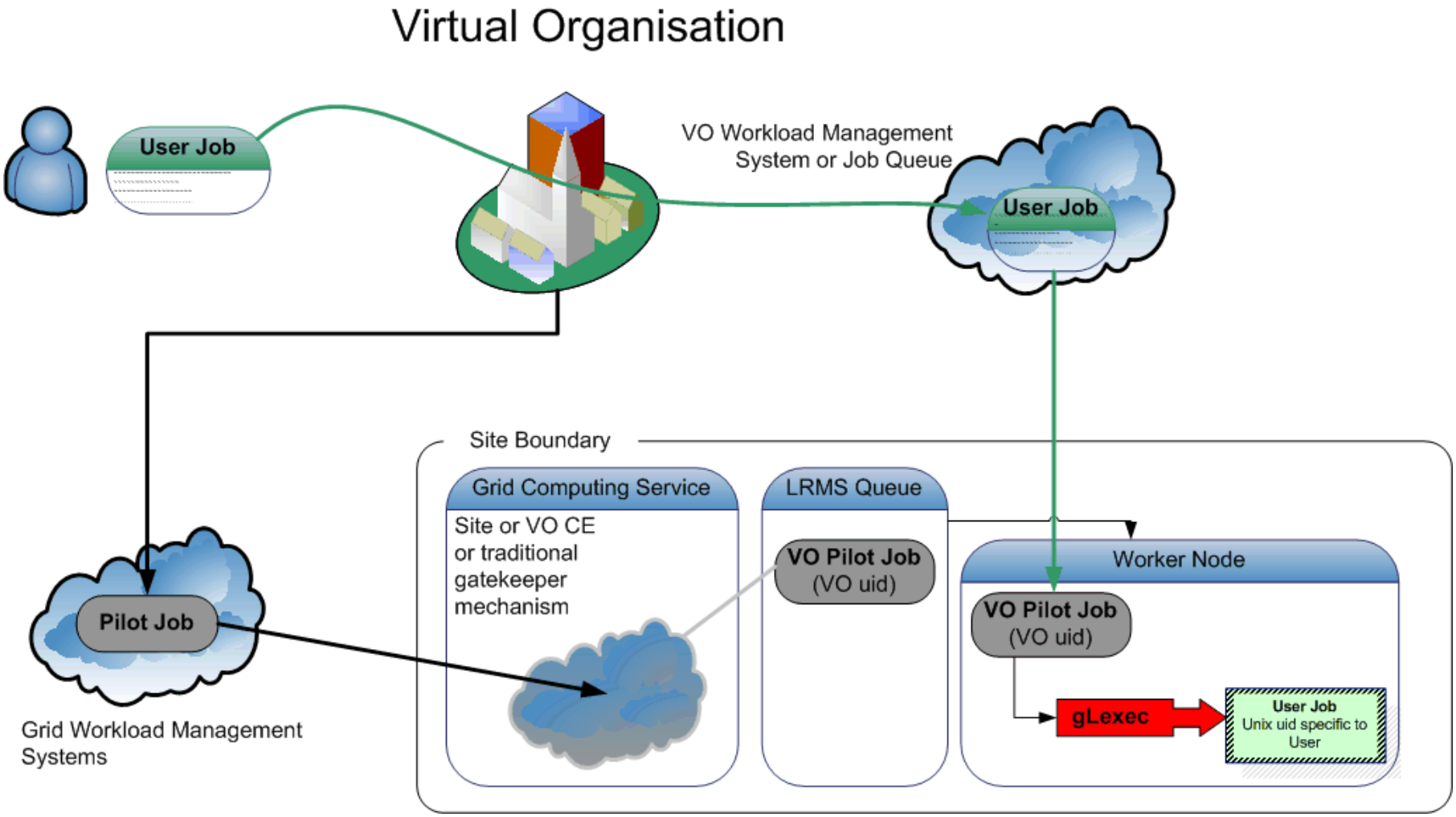
- you need global coordination,
  or you will be re-compromised from other 'partner' sites

read ***http://www.nsc.liu.se/~nixon/stakkato.pdf***
for some real-life experience

What A VO Community Can Do To You



Working with VO to respect policy, isolation

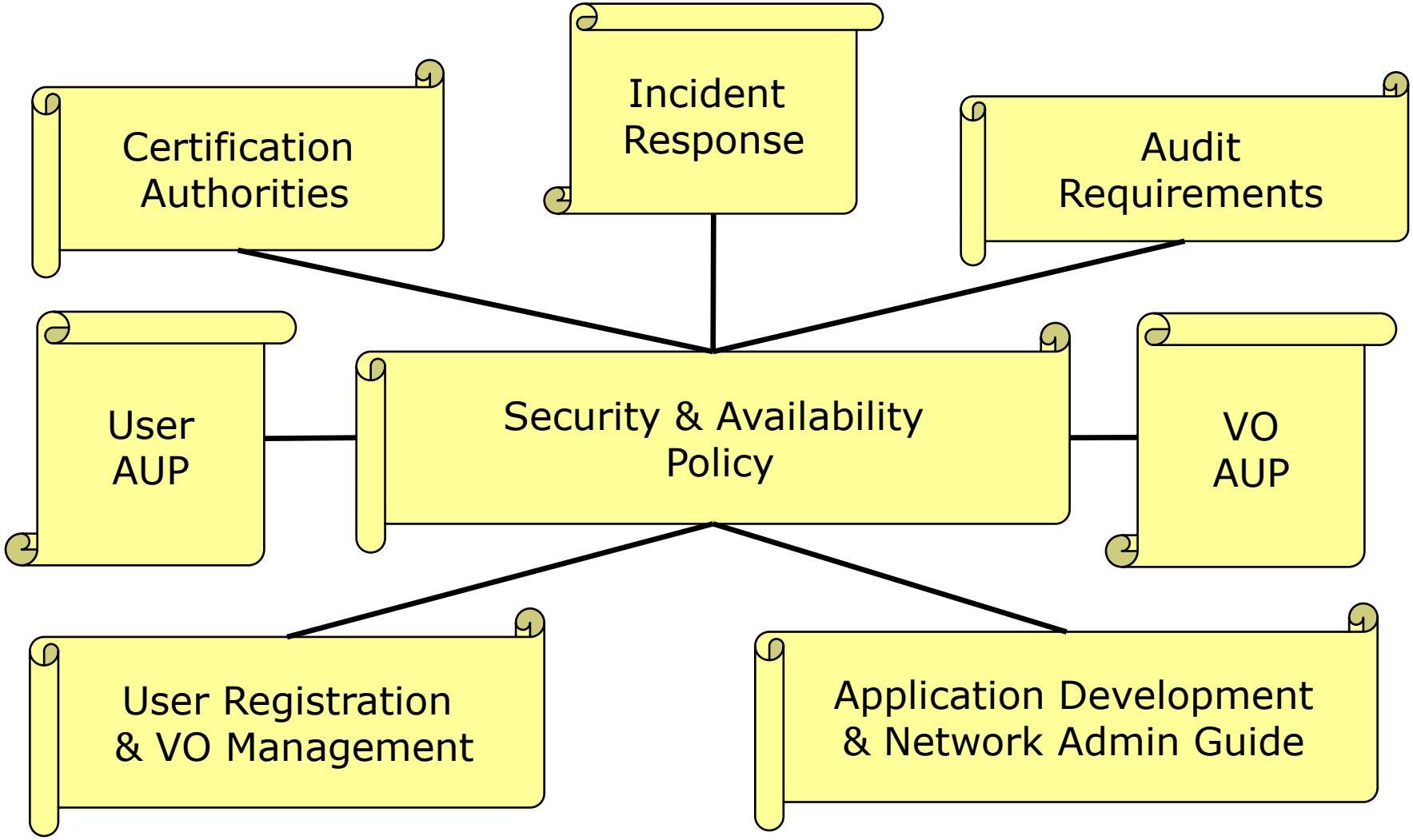


- At least prevent stealing VO pilot job credentials
- Allow cooperative policy compliance

Compliance

- Users will usually try to circumvent policy
- Enforcement (technical but mainly managerial) needed
  - Do some ethical hacking against these systems
    … *and escalate exploits found up to the first management level that reacts, until finally you go must go public …*
  - May even be fun (for some weird definitions of fun)
  - Code reviews and audits
    e.g. *http://pages.cs.wisc.edu/~kupsch/vuln_assessment/*
  - Ensure training for programmers

EGEE/LCG Security Policies



picture: Ian Neilson

http://cern.ch/proj-lcg-security/documents.html

strike balance between security and usability …

Example: Grid User AUP

By registering with the Virtual Organization as a GRID user
you shall be deemed to accept these conditions of use:

1. *You shall only use the GRID to perform work, or transmit or store data consistent with the stated goals and policies of the VO of which you are a member and in compliance with these conditions of use.*
2. *You shall not use the GRID for any unlawful purpose and not (attempt to) breach or circumvent any GRID administrative or security controls. You shall respect copyright and confidentiality agreements and protect your GRID credentials (e.g. private keys, passwords), sensitive data and files.*
3. *You shall immediately report any known or suspected security breach or misuse of the GRID or GRID credentials to the incident reporting locations specified by the VO and to the relevant credential issuing authorities.*
4. *Use of the GRID is at your own risk. There is no guarantee that the GRID will be available at any time or that it will suit any purpose.*
5. *Logged information, including information provided by you for registration purposes, shall be used for administrative, operational, accounting, monitoring and security purposes only. This information may be disclosed to other organizations anywhere in the world for these purposes. Although efforts are made to maintain confidentiality, no guarantees are given.*
6. *The Resource Providers, the VOs and the GRID operators are entitled to regulate and terminate access for administrative, operational and security purposes and you shall immediately comply with their instructions.*
7. *You are liable for the consequences of any violation by you of these conditions of use.*

What's in a Policy

- What do lawyers typically look for
  - Consistency of Terminology
  - Describe in exact and limitative terms

- How binding is it?
  - The signer must be explicitly aware of his or her action
  - Use default-deny
  - On web forms: at least use a pop-up box
  - *But this has only been marginally tested in court*

- What about the subjects
  - Keep it simple and short
  - 'Separate the policy from the actions' –
    but, indeed, then they'll never read the policy
  - Short lists work best (also for agreeing on policy)

Balancing incident response to privacy

What personal data are you allowed to keep?

- There are a couple of exemption clauses, for
  - Computer, communications and access control
  - but limited to max 6 months

  … otherwise, you actually ought to register your administration or accounting database
- Write down what you keep, why, and for how long
- Keep as little data as possible
- Limit logs to traffic analysis, not content
- But
  - keep enough to trace people in case of incidents
  - and to support your peers in dealing with incidents

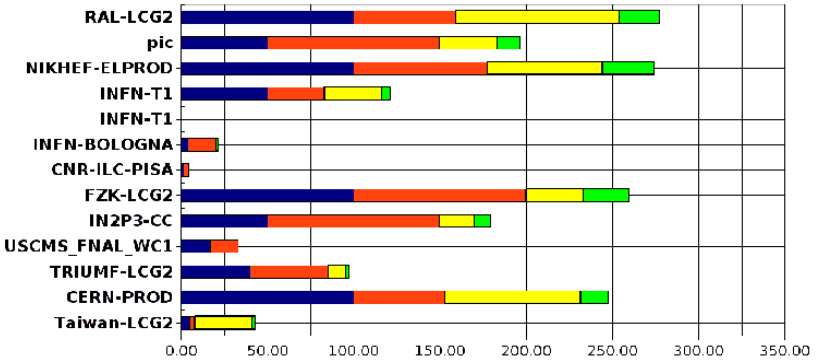
See e.g.

- http://www.cbpweb.nl/documenten/av_21_Goed_werken_in_netwerken.stm
- http://www.cbpweb.nl/HvB_website_1.0/i1.htm

Exercising incident response: the SSCs

- In a distributed multi-domain system, periodic exercises of the procedure have proved *very* useful
  - Reminds sites about the required steps
  - Finds holes and outdated contact information
  - Assess site compliance and capability
  - Experience: response to real incidents highly correlates with the test incidents!

- Evaluation made public after tests
  - Helps to get managerial backing for those poor site admins



https://twiki.cern.ch/twiki/bin/view/LCG/LCGSecurityChallenge
http://osct.web.cern.ch/osct/ssc.html

Distributed Systems Architecture
It is all about scaling
Grid-level Monitoring

PUTTING TOGETHER AN INFRASTRUCTURE

Grid Infrastructure

Realizing ubiquitous computing requires a *persistent infrastructure*, based on standards

Organisation

resource providers, user communities and virtual organisation

Operational Services

execution services, workflow, resource information systems, database access, storage management, meta-data

Support and Engineering

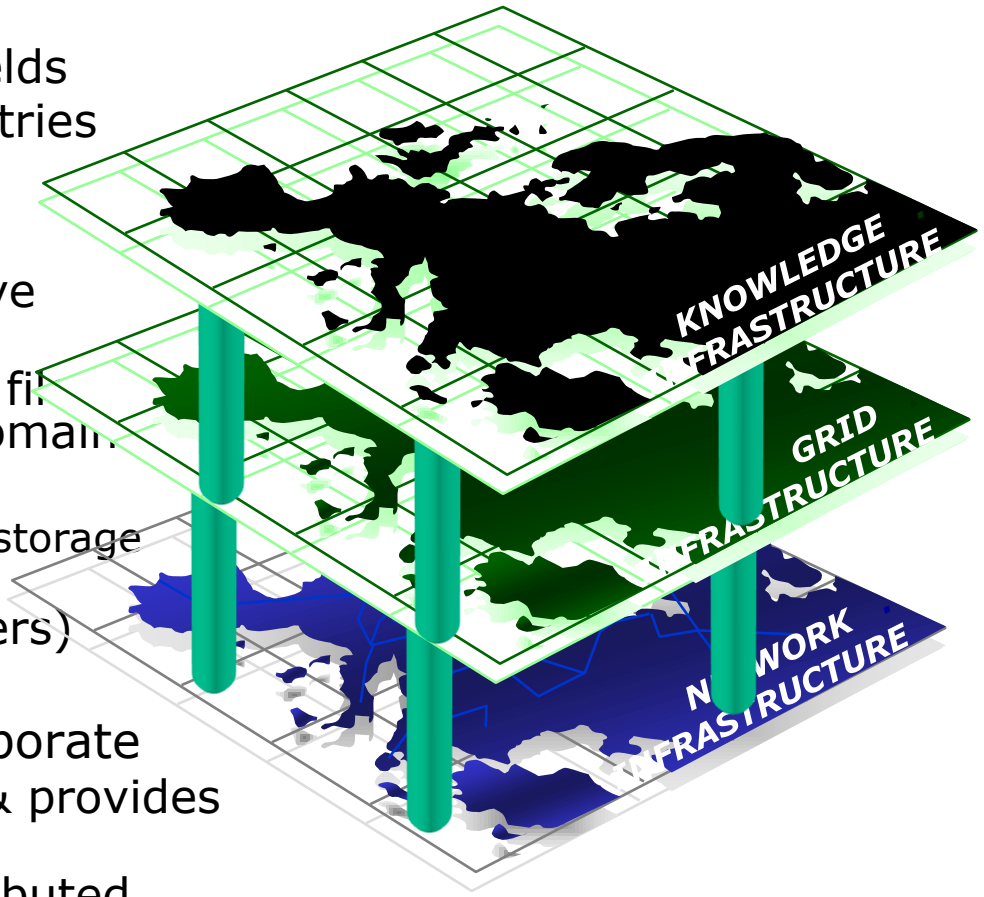
user support and ICT experts … with domain knowledge

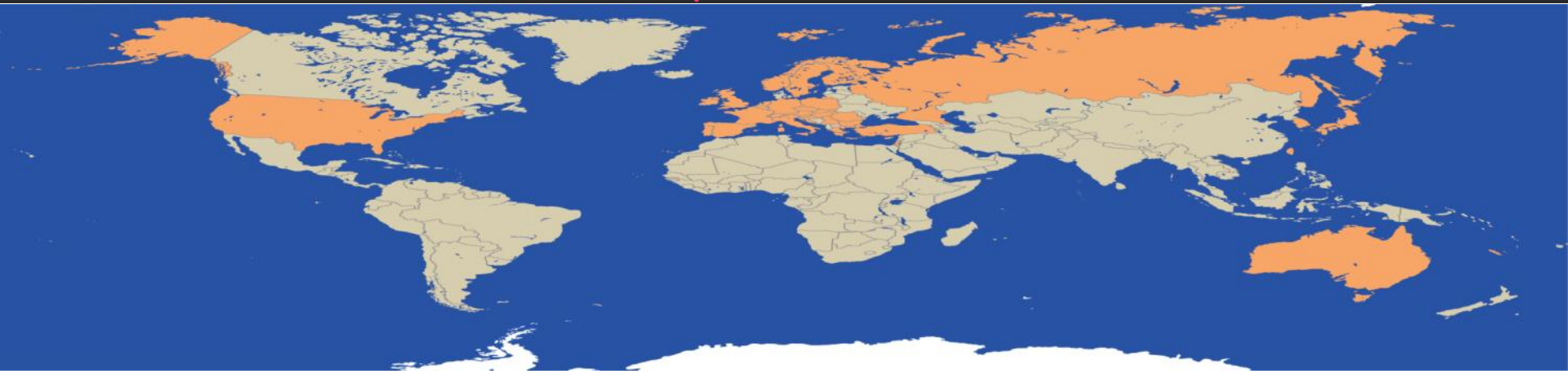
How e-Infrastructures help e-Science

- e-Infrastructures provide easier access for
  - Small research groups
  - Scientists from many different fields
  - Remote and still developing countries

- … to new technologies
  - Produce, store and search massive amounts of data
  - Transparent access to millions of fi
    across different administrative domai
  - Low cost access to resources
    - Mobilise large amounts of CPU & storage on short notice (PC clusters)
  - High-end facilities (supercomputers)

- And help to find new ways to collaborate
  - Eases distributed collaborations & provides new ways of community building
  - Develops applications using distributed complex workflows
  - Gives easier access to higher education

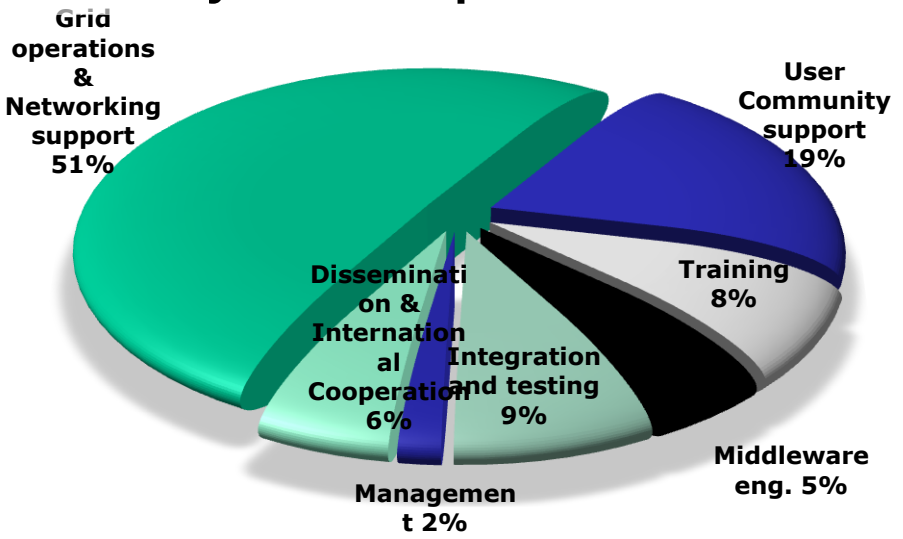




Flagship Grid infrastructure project co-funded by the European Commission

Main Objectives

- Expand/optimise existing EGEE infrastructure, include more resources and user communities
- Prepare migration from a project-based model to a sustainable federated infrastructure based on National Grid Initiatives

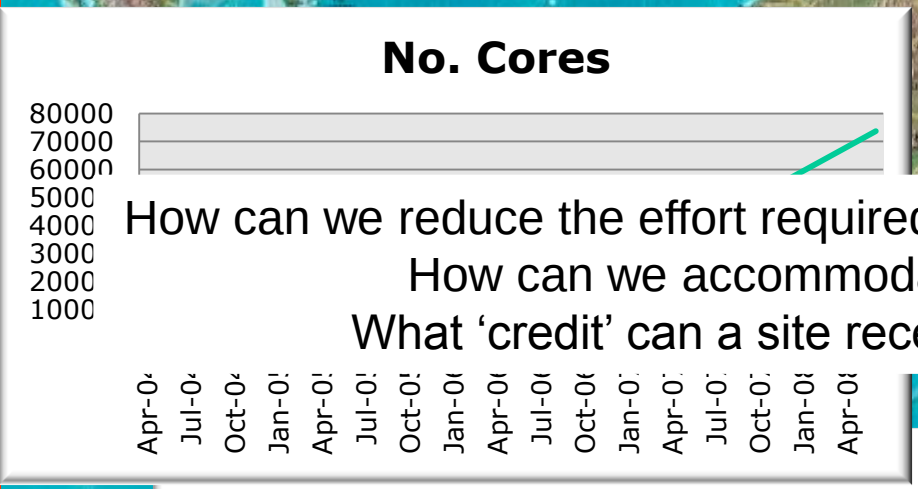


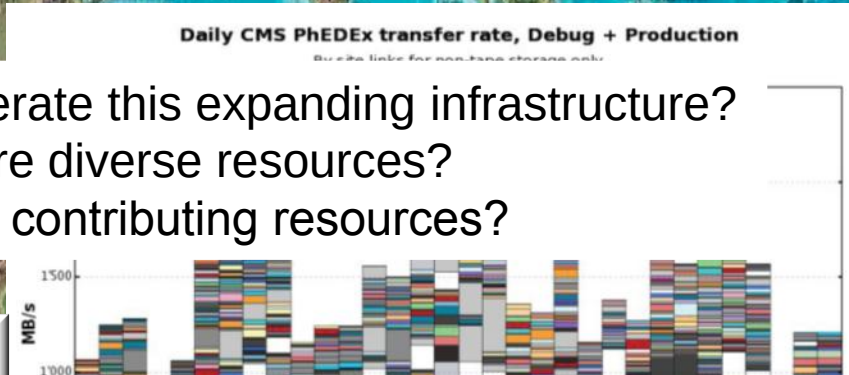
Duration: 2 years
Consortium: ~140 organisations across 33 countries
EC co-funding: 32Million €

EGEE Production Grid Infrastructure

Steady growth over the lifetime of the project

Improved reliability

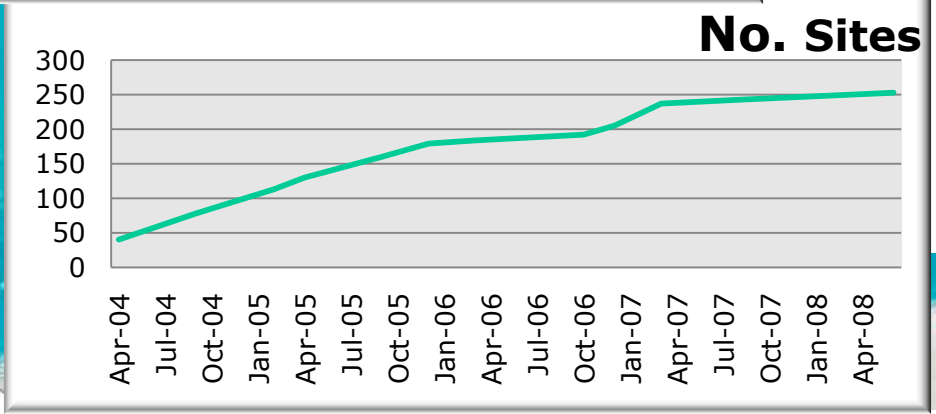


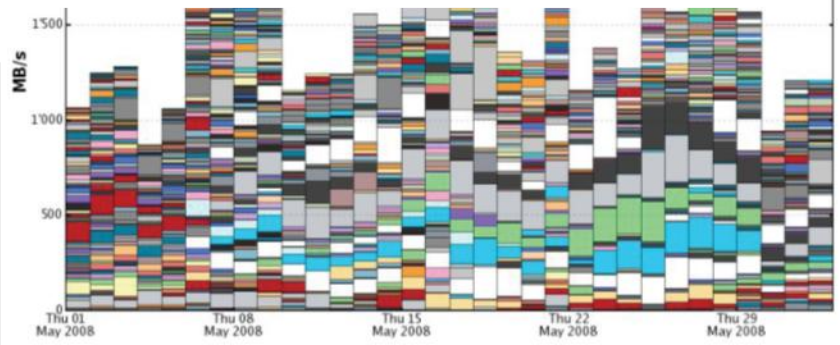


How can we reduce the effort required to operate this expanding infrastructure?

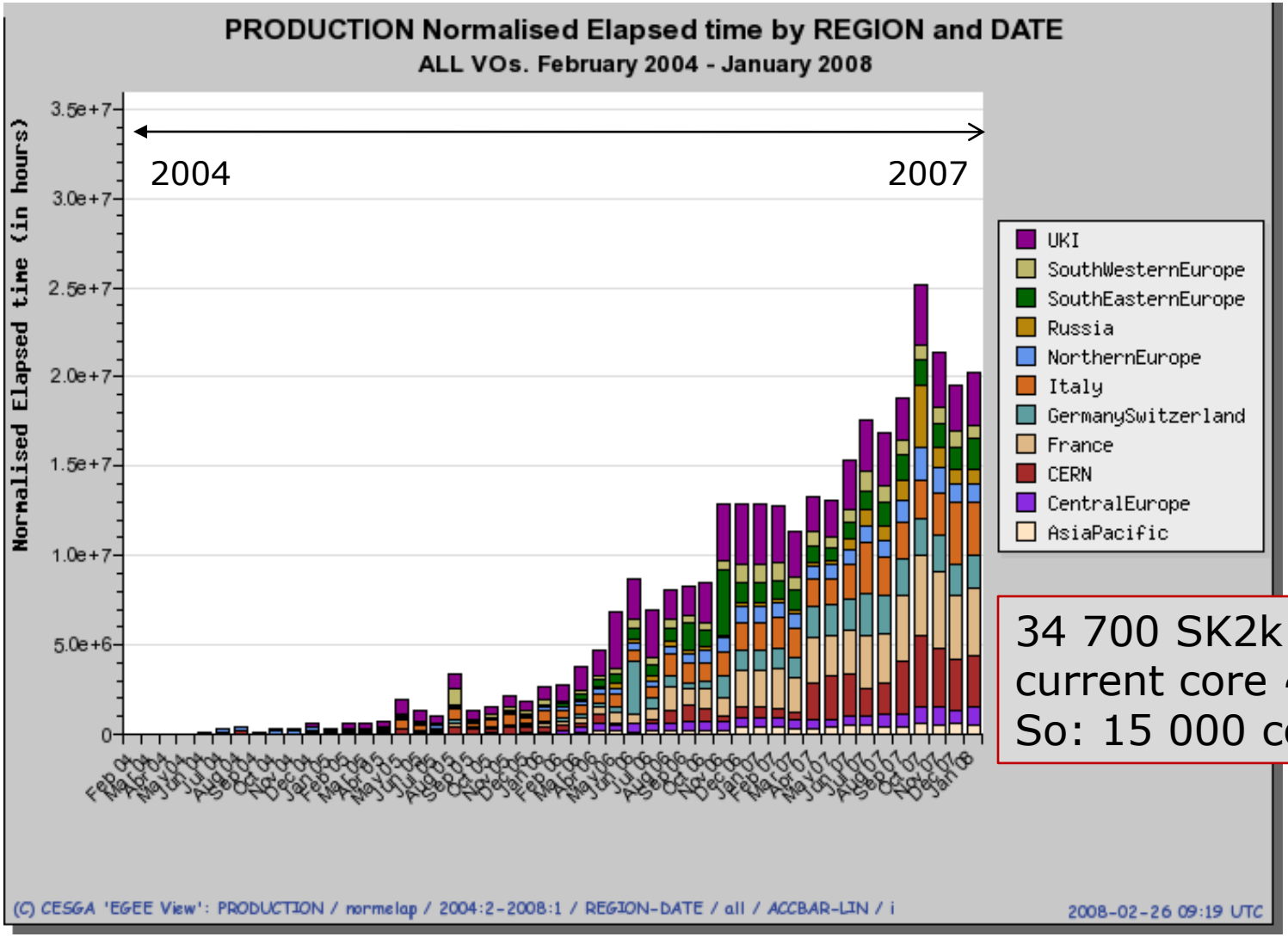
How can we accommodate more diverse resources?

What 'credit' can a site receive for contributing resources?





25 million SI2000 CPU hours reported/month

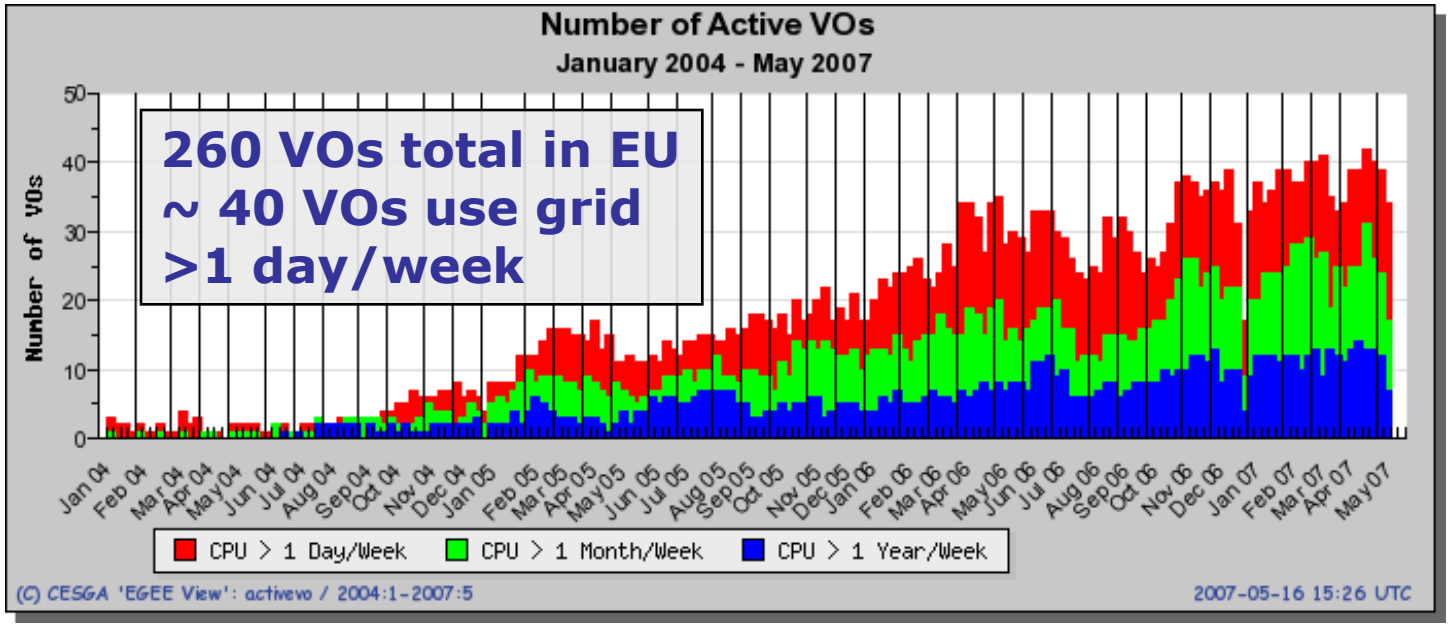


eGee Enabling Grids for E-sciencE

34 700 SK2k hr/hr
current core ~ 2 SI2k
So: 15 000 cores avg.

data: EGEE monitoring, RAL and CESGA, http://goc.grid-support.ac.uk/gridsite/accounting/

Grid Infrastructures Works!



Enabling Grids for E-sciencE

Number of **active** Vos in EU since 2004

over 40 VOs hosted in NL

A reliable Grid Infrastructure needs operational support:

- availability monitoring
- reporting and follow-up
- user support

data: EGEE monitoring, RAL and CESGA, http://goc.grid-support.ac.uk/gridsite/accounting/

Latest SAM results, Site Status, for 'OPS' VO, 27 Sep 2007 13:39 GMT.
Size of site rectangles is number of CPUs from BDII.
Certified Production sites, grouped by regions.

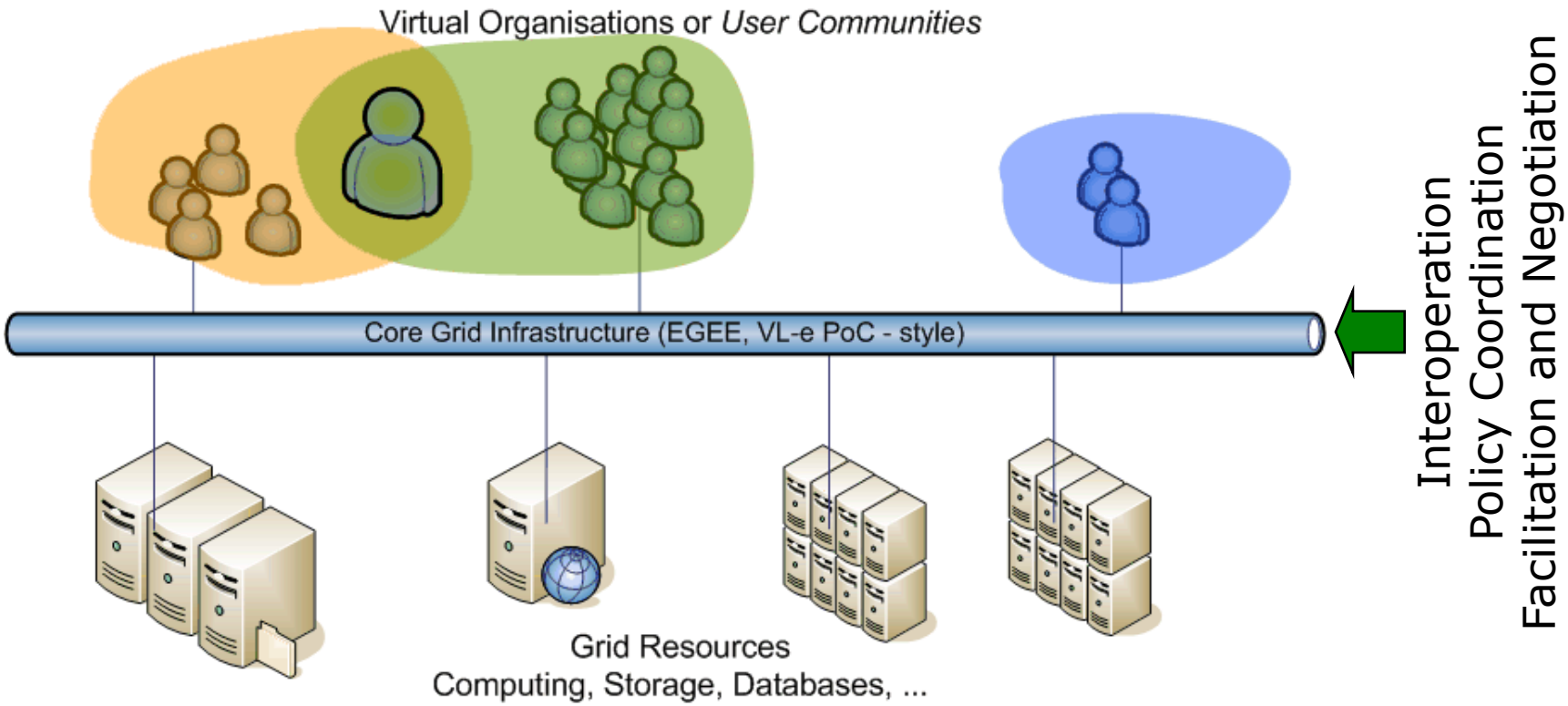
Graphs: Ian Bird, EGEE SA1, EGEE07 Conference October 2007

vl·e virtual laboratory for e·science

BiG Grid the dutch e·science grid

NIKHEF pdp

Building Grid Infrastructures

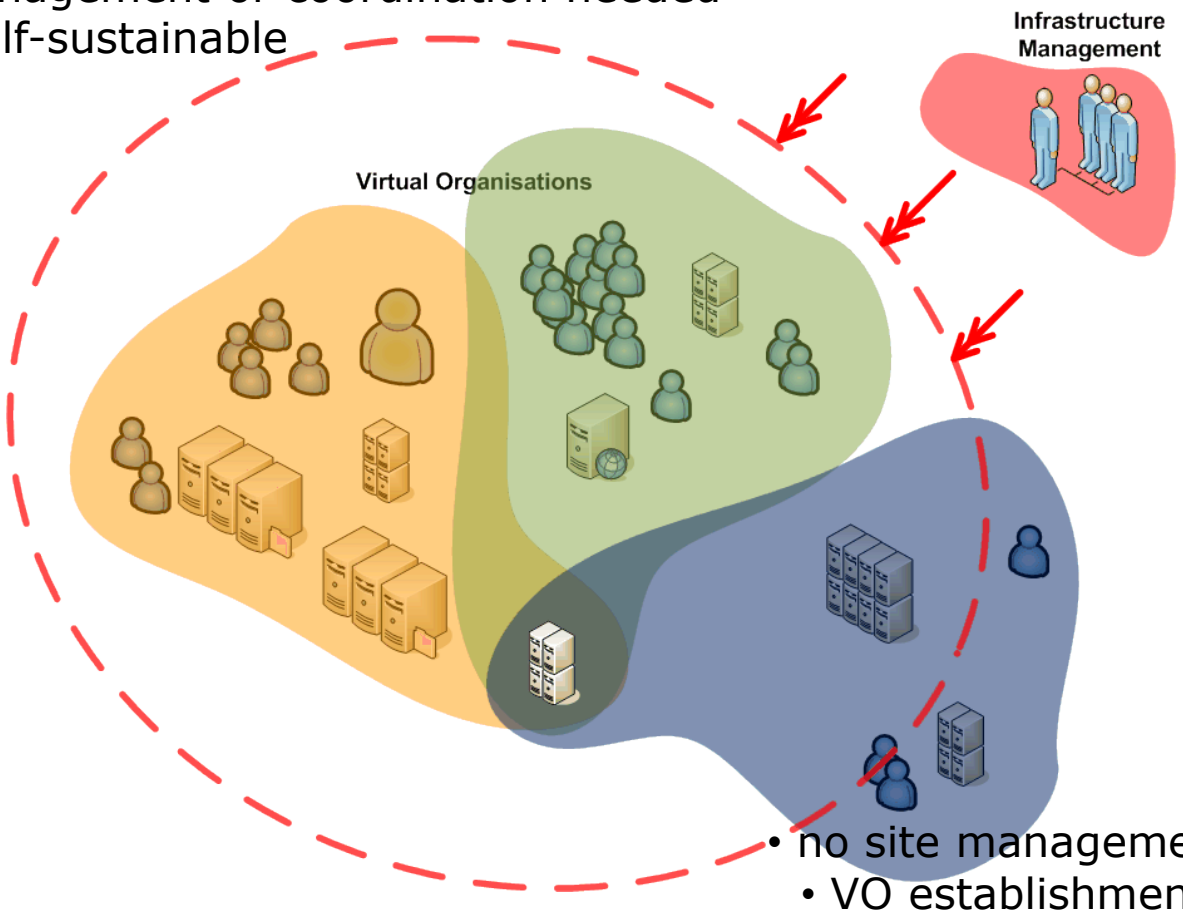


- Interop: common syntax and semantics for grid operations
- Policy Coordination: User and VO AUPs, operations, trust
- Facilitating negotiation: VO meta-data, SLAs, op. environment

VO-centric infrastructure ('OSG style')

What happens if you do not coordinate infrastructure from the beginning …

Advantages
- no site management or coordination needed
- VOs are self-sustainable



Disadvantages
- no site management or coordination
- VO establishment is more complex
- infrastructure itself is transient and harder to sustain

Managing Complexity and Standards

Towards a sustained infrastructure organisation

SUSTAINING THE INFRASTRUCTURE

Interoperation and standards

Coordination of Infrastructures is a 'must'

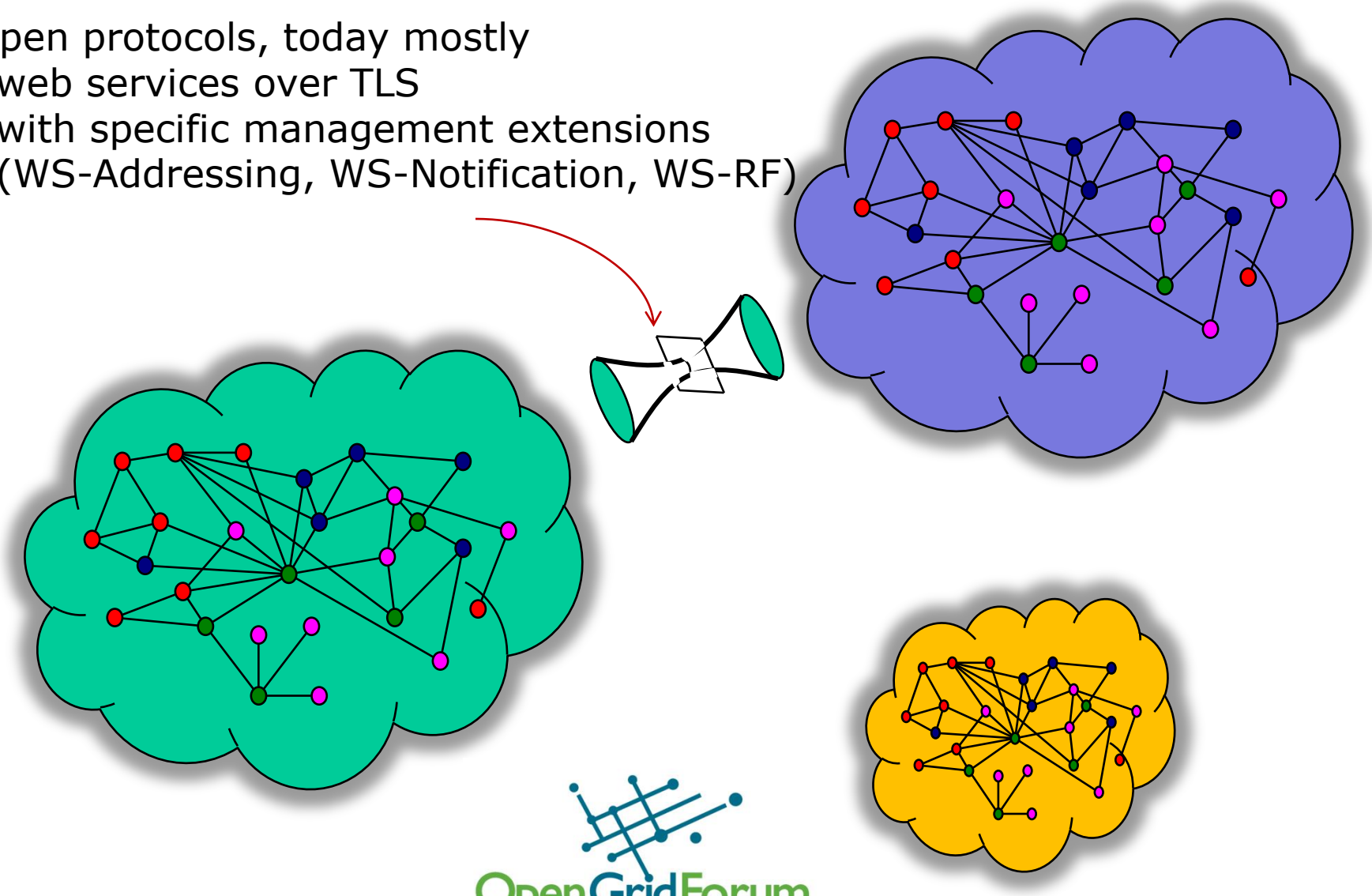
- stability and consistency vary widely
- self-healing and verification are largely absent
- Global issues require coordinated response (e.g. Incidents, brokering of access, etc.)

Two parallel tracks

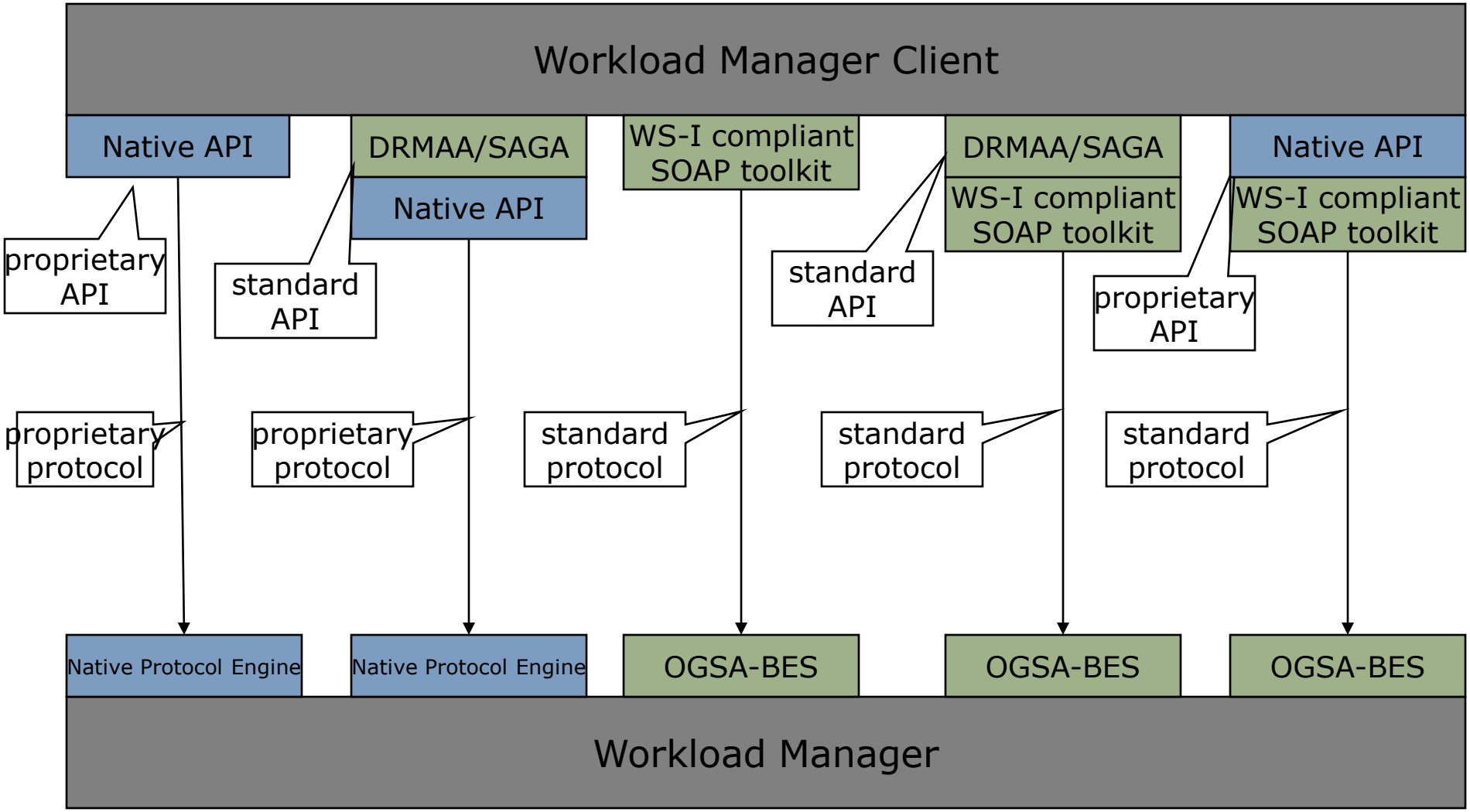
- Middleware: global standards
- Europe now moving towards this persistent infrastructure with EGI

Interoperation – between the clouds?

Open protocols, today mostly
- web services over TLS
- with specific management extensions (WS-Addressing, WS-Notification, WS-RF)



Introducing standards



Graphic: Open Grid Forum 2007

Standards

- Standards, such as those by IETF, OASIS, OGF, &c aid interoperability and reduce vendor lock-in

- as you go higher up the stack, you get less synergy
  - Transport: IP/TCP, HTTP, TLS/SSL, &c well agreed
  - Web services: SOAP and WS-Security used to be the solution for all … but 'Web 2.0' shows alternatives tailored to specific applications gaining popularity
  - Grid standards:
    low-level job submission (BES, JSDL), management (DRMAA), basic security (OGSA-BSP Core, SC) there
  - higher-level services still need significant work …

see also http://www.ogf.org/

Why not standardize?

- A technology might be "too new"
  - 'you stifle innovation with standardization, which focuses on commonality'
- A technology might be very niched
  - De-facto standards will emerge in this case and in perhaps not so niche areas like KML in Google maps
- Standards take too long;
  - get your product out today and grab market – then your API is the de facto standard
- Organizations with a strong proprietary product might try and succeed derailing standards that would enable competition

European Grid Initiative

Goal:

- Long-term sustainability of grid infrastructures in Europe

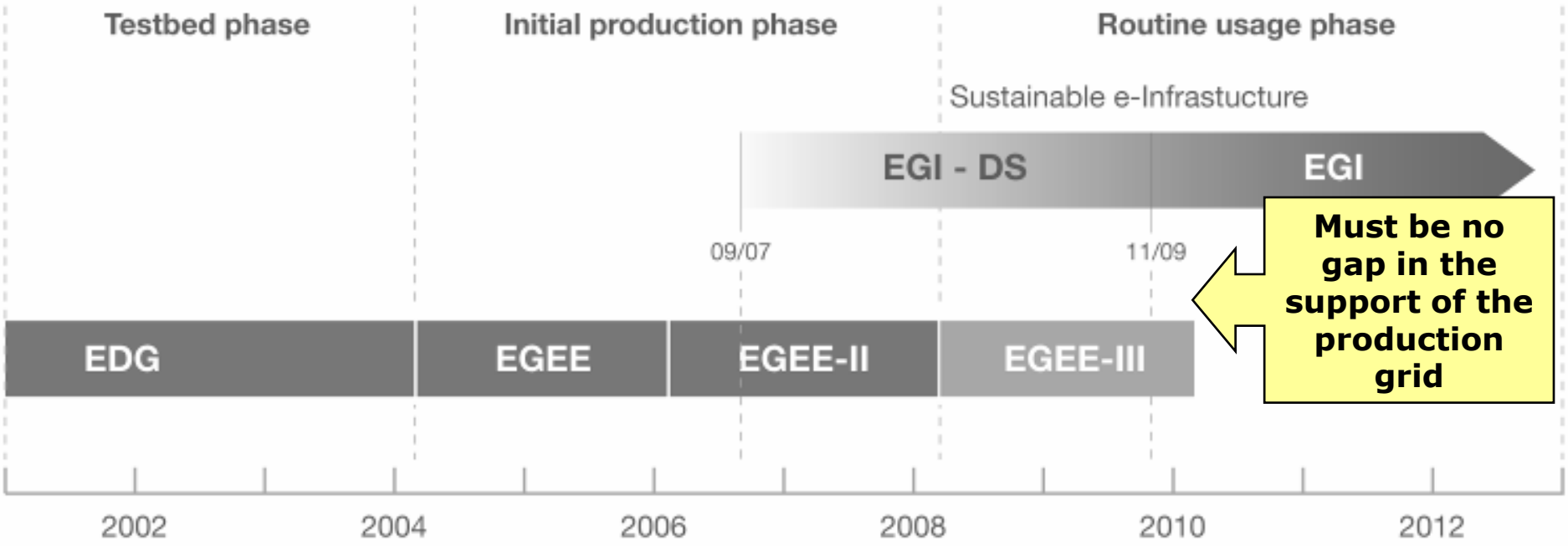
Approach:

- Establishment of a new federated model bringing together NGIs to build the EGI Organisation

EGI Organisation:

- Coordination and operation of a common multi-national, multi-disciplinary Grid infrastructure
  - To enable and support international Grid-based collaboration
  - To provide support and added value to NGIs
  - To liaise with corresponding infrastructures outside Europe

European Grid Initiative timeline



European Grid Initiative

- EGI Design Study proposal approved by the European Commission (started 1st September'07)
- Supported by 35+ National Grid Initiatives (NGIs)
  http://web.eu-egi.eu/partners/ngi/
- 2 year project to prepare the setup and operation of a new organizational model for a sustainable pan-European grid infrastructure
- Draft EGI Blueprint produced:
  Blueprint Proposal http://www.eu-egi.eu/blueprint.pdf
  Functions Description http://www.eu-egi.eu/functions.pdf

http://www.eu-egi.org

Amsterdam to host EGI!

European Grid Initiative

»Towards a sustainable production grid infrastructure«

Amsterdam to host EGI.org

Amsterdam has been chosen to host EGI.org, the coordinating organization responsible for managing the European Grid Initiative (EGI).

Amsterdam was selected as the host city at the last EGI policy board meeting in Catania on Monday 2 March 2009, ahead of seven other European cities that also expressed their interest in hosting the EGI Organization.

"The choice of the location of the EGI.org headquarters is a further and decisive step towards the implementation of a sustainable European grid infrastructure", said **Gaspar Barreira**, Chairman of the EGI Policy Board. "From now on we will be all mobilised for the real establishment of a new international research infrastructure in Europe, where a large number of countries will put together and operate the world's largest grid computing facility."

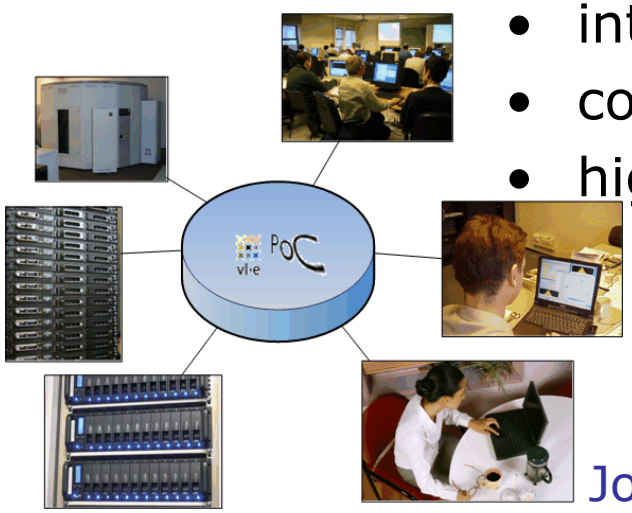
"We are very honoured that the European grid community has chosen Amsterdam to host EGI.org", said **Patrick Aerts**, Director of the National Compute Facility (NCF), the Netherlands, after the announcement of the decision. "We thank all EGI-Policy Board members, and especially our runners-up for the trust that they have placed in the Netherlands. This is of course a very positive result for the Science Park Amsterdam, The City of Amsterdam and NWO/NCF, which together represent the Netherlands Grid Initiative (NGI), but it is above all a shared EGI achievement."

But I Just Want it to Work!

vl·e PoC

In the end, the infrastructure will be user driven

In NL: a common infrastructure for e-Science is provided by **BiG Grid** and the *VL-e Proof-of-Concept*



- interoperable interfaces to resources
- common software environment
- higher-level 'virtual lab' services

Central Facilities:
SARA, NIKHEF, RUG-CIT, Philips

Join yourself: user-interfaces,
distributed clusters, storage

BiG Grid
the dutch e·science grid

http://poc.vl-e.nl/distribution/

Does it work

How can we make it better

GOING FROM HERE

Going from here

Many nice things to do:

- In many cases, a single OS is a nice feature for users, since they know what they get
  - but users will need SLES, Debian, Gentoo, … or specific libraries
  - Guaranteed execution environment for users
  - … but sites don't want to change OS
- Virtualisation (Xen, VMware) to hide user OS from system OS?
- Enabling applications: the integration of software and grid!
- Auditing and user tracing in this highly dynamic system
  can we know for sure who is running what where? Or whether a user is DDoS-ing the White House right now?
  - Out of 221 sites, we know for certain there is a compromise!

More things to do …

- Data access: access data efficiently over the wide area
  - The file system abstractions seems to have broken down
  - But the storage container object (like Amazon's S3 objects) is counter-intuitive for users)

- Can we do something useful with the large disks in all worker nodes?
  (our 1200 cores share ~ 48 TByte of unused space!)

- There are new grid software releases every month, and the configuration comes from different sources …
  how can we combine and validate all these configurations fast and easy?

- Apply for a job in engineering,development @Nikhef☺

A Bright Future!

Imagine that you could plug your computer into the wall and have direct access to huge computing resources immediately, just as you plug in a lamp to get instant light. …

Far from being science-fiction, this is the idea the [Grid] is about to make into reality.

The EU DataGrid project brochure, 2001

vl·e

http://www.vl-e.nl/

BiG Grid

the dutch e-science grid

NIKHEF

pdp