National e-Science Centre

Grids, Security and the Life Sciences

Dr Richard Sinnott
Technical Director National e-Science Centre
|||
Deputy Director Technical Bioinformatics Research Centre
University of Glasgow

ros@dcs.gla.ac.uk

21st July 2005

National e-Science Centre

Me?

I promise not to gloat too much about football… ;o)

Overview

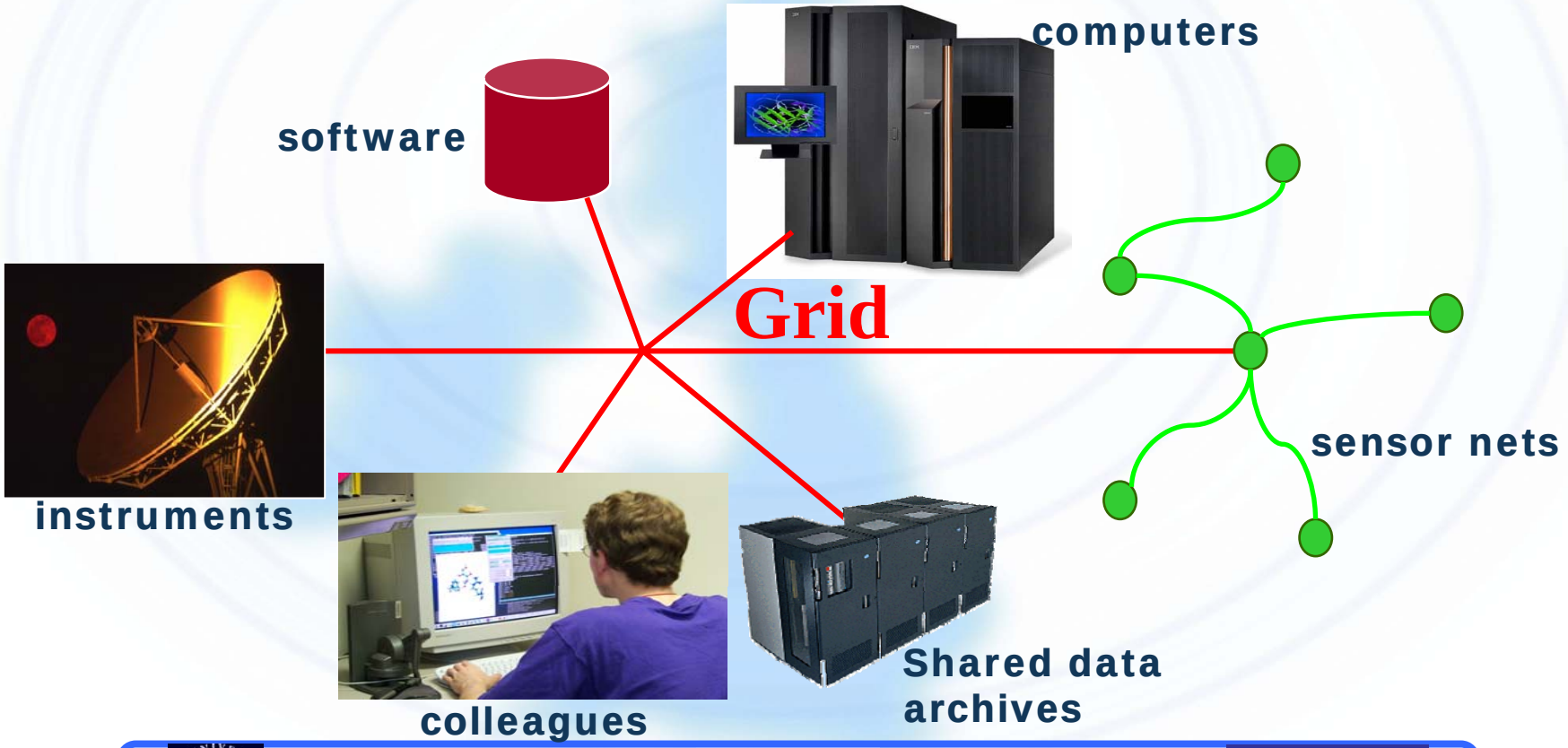
National e-Science Centre

- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

Grids? E-Science? E-Research?

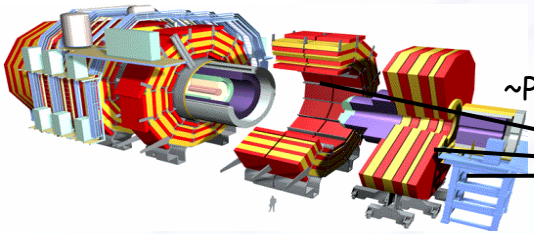
National e-Science Centre

- methodologies transforming science, engineering, medicine and business
  - driven by exponential growth in data, compute demands
    - enabling a whole-system approach



Data Grids for High Energy Physics

National e-Science Centre



~PBytes/sec

Online System

~100 MBytes/sec

1 TIPS is approximately 25,000 SpecInt95 equivalents

Offline Processor Farm
~20 TIPS

~100 MBytes/sec

There is a "bunch crossing" every 25 nsecs.
There are 100 "triggers" per second
Each triggered event is ~1 MByte in size

Tier 0

CERN Computer Centre

~622 Mbits/sec

Tier 1

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

~622 Mbits/sec

Tier 2

Caltech
~1 TIPS

Tier2 Centre
~1 TIPS

~622 Mbits/sec

Institute
~0.25TIPS

Physics data cache

~1 MBytes/sec

Tier 4

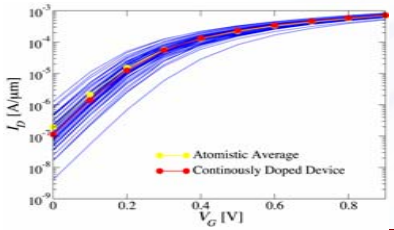
Physicist workstations

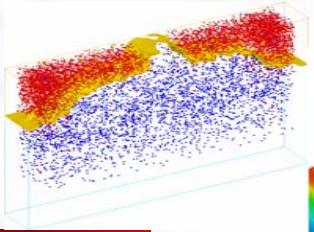
Physicists work on analysis "channels".
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server

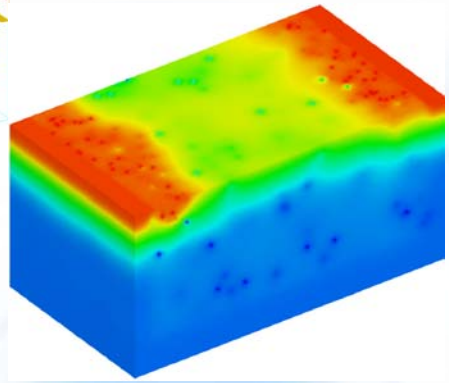
GridToday (4.4.2005)
IBM Shatters Data Management Challenge at CERN
Using IBM TotalStorage SAN File System storage virtualization software, the internal tests shattered performance records during a data challenge test by CERN by reading and writing data to disk at rates in excess of 1GB/second for a total I/O of over 1 petabyte (1 million gigabytes) in a 13-day period

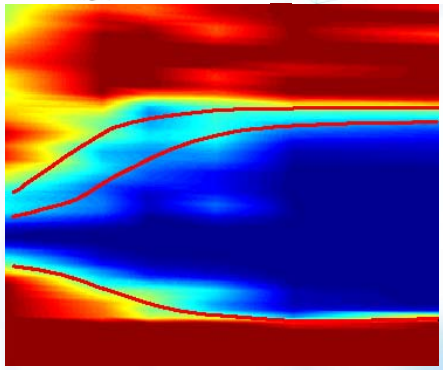
Next Generation Transistor Design

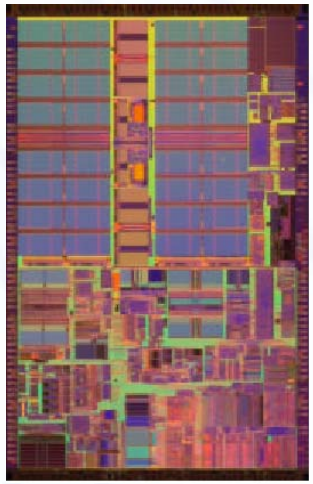
National e-Science Centre







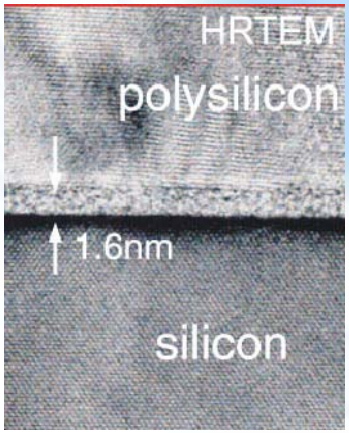


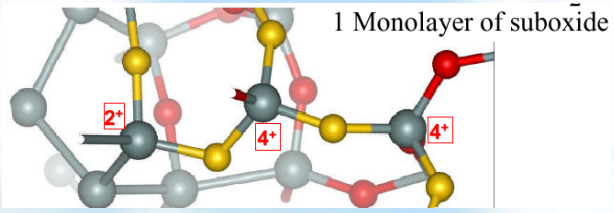


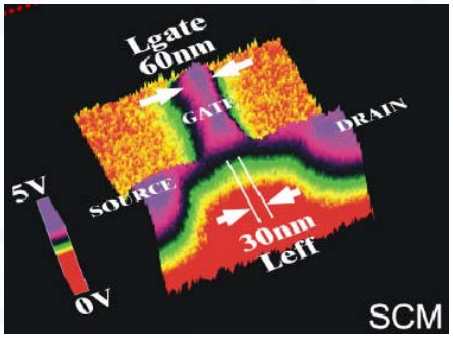
3D

+

Statistical

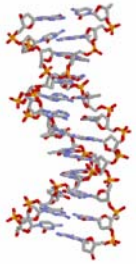


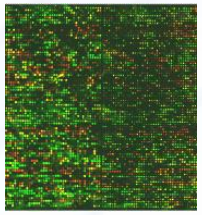


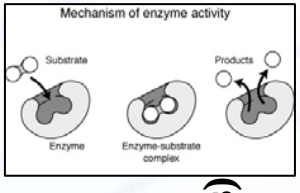


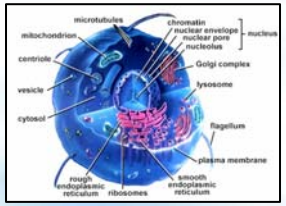
Systems Biology?

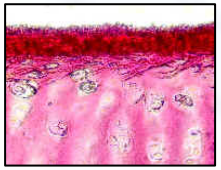
National e-Science Centre

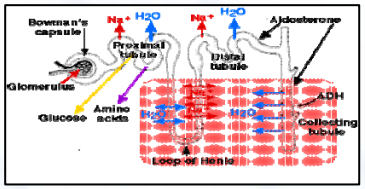












Nucleotide sequences

Nucleotide structures

Gene expressions

Protein Structures

Protein functions

Protein-protein interaction (pathways)

Cell

Cell signalling

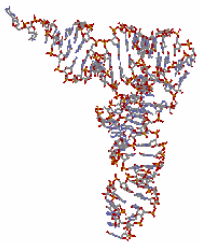
Tissues

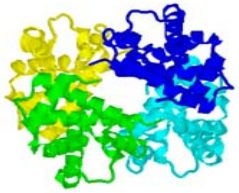
Organs

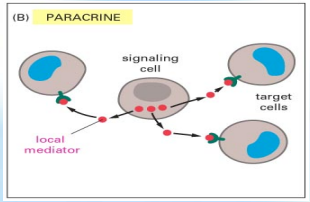
Physiology

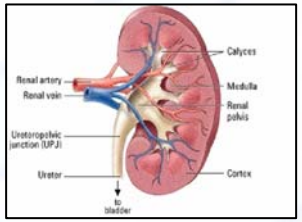
Organisms

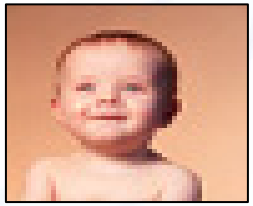
Populations

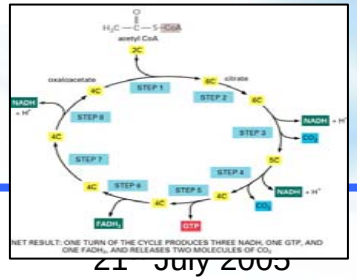












+ links to plant/crops, environmental, health, … information sources

Overview

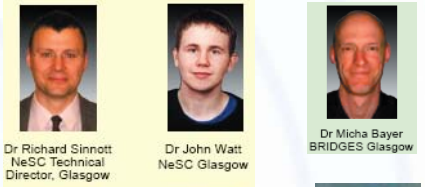
National e-Science Centre

- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

Glasgow e-Science Hub

National e-Science Centre

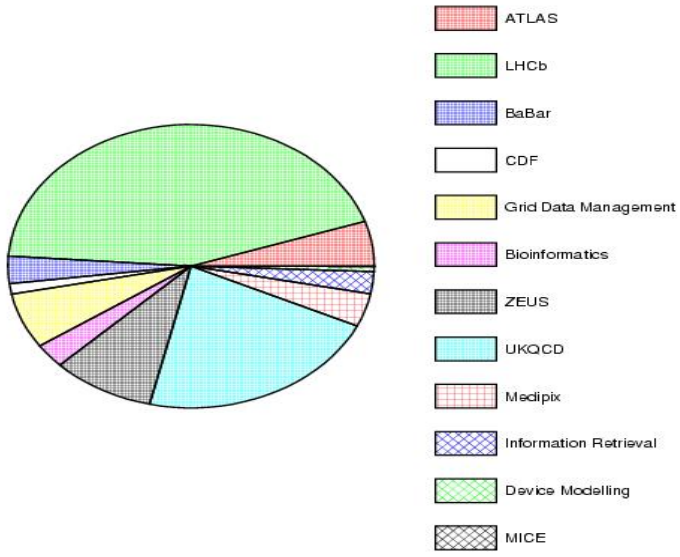
- E-Science Hub
  - Externally
    - Glasgow end of NeSC
      - Involved in UK wide activities
        - » **Involved in numerous life science/security related projects (more later)**
      - Public visibility of NeSC
        - » responsible for NeSC web site (www.nesc.ac.uk)
  - Internally
    - Focal point for e-Science research/activities at Glasgow
    - Work closely with foundation departments
      - » Department of Computing Science
      - » Department of Physics & Astronomy
    - Also working with other groups including
      - » Bioinformatics Research Centre
      - » Biostatistics
      - » Electronics and Electrical Engineering
      - » Clinicians, Hospitals, across Scotland, …



Glasgow e-Science Infrastructure

National e-Science Centre

- Consolidating resources
  - Story started with building around ScotGrid
    - Providing shared Grid resource for wide variety of scientists inside/outside Glasgow
      - HEP, CS, BRC, EEE, …
        - » Target shares established
        - » Non-contributing groups encouraged



- ScotGrid [Disk ~15TB
  CPU ~ 330 1GHz]
- Over 2 million CPU hours completed (May 2005)
- Over 230,000 jobs completed
  - Includes time out for major rebuilds
- Typically running at ~90% usage

Grid Security

National e-Science Centre

- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- **Grid Security**
  - **Grid Requirements, Concepts, Technologies, …**
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

Grid Security

National e-Science Centre

- Why is Grid security so important?
- The Challenge of Grid Security
  - Multifarious sets of applications running on
  - …a myriad of evolving end systems which are
  - …accessing/using a plethora of potentially changing data resources
  - …across multiple heterogeneous distributed institutions
  - …by remote and changing collections of end users
- Technical challenges
  - Technologies to help make Grids secure
    - Public Key Infrastructures
  - Security always depends on the weakest link
- Social challenges
  - Educating users in security issues
- Manageability
  - Systems must be easily configurable, changeable when security threats arise / have arisen
- Usability
  - Systems must be usable by non-computer scientists
- Scalability
  - Must allow for a multitude of different classes of user

Why is Grid security so important?

National e-Science Centre

- If it is not secure
  - Large communities will not engage
    - medical community, industry, financial community …
  - Legal and ethical issues possible to be violated with all sorts of consequences
    - e.g. data protection act violations and fines incurred
  - Expensive (impossible?) to repeat some experiments
    - Huge machines running large simulations for several years
  - Trust (more later) is easily lost and hard to re-establish
  - Grid resources are a dream for hackers
    - Huge file storage for keeping their "dodgy data"
    - Perfect environment for launching attacks like distributed denial of service
      - Not just access to one machine
        - » whole interconnected networks of ultra performant machines which can be used for cracking passwords, codes, launching distributed denial of service attacks, …

The Challenge of Grid Security

National e-Science Centre

- Grids allow (or should allow!) dynamic establishment of virtual organisations (VOs)
  - These can be arbitrarily complex
    - Grids (VOs) might include highly secure supercomputing facilities through to single user PCs/laptops
    - Need security technologies that scales to meet wide variety of applications
      - from highly secure medical information data sets through to particle physics/public genome data sets
      - Using services for processing of patient data through to “needle in haystack” searching of physics experiments or protein sequence similarity of genomic data
  - Should try to develop generic Grid security solutions
    - Avoid all application areas re-inventing their own (incompatible/inoperable) solutions

The Challenge of Grid Security …ctd

National e-Science Centre

- Grids allow scenarios that stretch inter-organisational security
  - Imagine two distributed virtual organisations agreeing to share resources, e.g. compute/data resources to accomplish some task with sharing done across internet
    - Could have policies that restrict access to and usage of resources based on pre-identified users, resources
    - But what if new resources added, new users added, old users go,…?
    - What if organisations decide to change policies governing access to and usage of resources?
    - What if want to transfer large data sets between different organisations – how to ensure that data is not cached somewhere it might be compromised?
    - …

Prelude to Grid Security

National e-Science Centre

- What do we mean by security anyway?
  - Secure from whom?
    - From sys-admin? From rogue employee? …
  - Secure against what?
    - Security is never black and white but is a grey landscape where the context determines the accuracy of how secure a system is
      - e.g. secure as given by a set of security requirements
  - Secure for how long?
    - "I recommend overwriting a deleted file seven times: the first time with all ones, the second time with all zeros, and five times with a cryptographically secure pseudo-random sequence. Recent developments at the National Institute of Standards and Technology with electron-tunnelling microscopes suggest even that might not be enough. Honestly, if your data is sufficiently valuable, assume that it is impossible to erase data completely off magnetic media. Burn or shred the media; it's cheaper to buy media new than to lose your secrets…."

      » -*Applied Cryptography 1996*, page 229

Prelude to Grid Security …ctd

National e-Science Centre

- Note that security technology ≠ secure system
  - Ultra secure system using 2048+ bit encryption technology, packet filtering firewalls, …
    - …. on laptop in unlocked room
    - … on PC with password on “post-it” on screen/desk
    - …
  - Famous quote to muse over:
    - “…if you think that technology can solve your security problems then you don’t know enough about the technology, and worse you don’t know what your problems are…”
      - » Bruce Schneier, Secrets and Lies in a Digital Networked World

Technical Challenges of Grid Security

National e-Science Centre

- Key terms that are typically associated with security
  - Authentication
  - Authorisation
  - Audit/accounting
  - Confidentiality
  - Privacy
  - Integrity
  - Fabric management
  - Trust

All are important for Grids but some applications may have more emphasis on certain concepts than others

Security Concepts::Authentication

National e-Science Centre

- Authentication is the establishment and propagation of a user's identity in the system
  - e.g. so site X can check that user Y is attempting to gain access to resources
    - Note does not check what user is allowed to do, only that we know (and can check!) who they are
      - Masquerading always a danger (and realistic possibility)
      - Need for user guidance on security
        - Password selection
        - Treatment of certificates
        - …
- Typically achieved using Public Key Infrastructures
  - More later…

Security Concepts::Authorisation

National e-Science Centre

- *Authorisation*
  - concerned with controlling access to services based on policy
    - Can this user invoke this service making use of this data?
    - Complementary to authentication
      - Know it is this user, now can we restrict/enforce what they can/cannot do
  - Many different contenders for authorisation infrastructures
    - PERMIS
    - CAS
    - VOMS
    - AKENTI
    - VOM
      - We have lots of experience with PERMIS from University of Kent
        - » (most advanced authorisation infrastructure???)

Security Concepts::Auditing

National e-Science Centre

- *Auditing*
  - the analysis of records of account (e.g. security event logs) to investigate security events, procedures or the records themselves
    - Includes logging, intrusion detection and auditing of security in managed computer facilities
      - well established in theory and practice
        - » Grid computing adds the complication that some of the information required by a local audit system may be distributed elsewhere, or may be obscured by layers of indirection
        - » e.g. Grid service making use of federated data resource where data kept and managed remotely
    - Need tools to support the generation of diagnostic trails
      - Do we need to log all information?
      - How long do we keep it for?
      - …

Security Concepts::Confidentiality

National e-Science Centre

- *Confidentiality*
  - is concerned with ensuring that information is not made available to unauthorised individuals, services or processes
    - It is usually supported by access control within systems, and encryption between systems
      - Confidentiality is generally well understood, but the Grid introduces the new problem of transferring or signalling the intended protection policy when data staged between systems

Security Concepts::Privacy

National e-Science Centre

- Privacy
  - particularly significant for projects processing personal information, or subject to ethical restrictions
    - e.g. projects dealing with medical, health data
  - Privacy requirements relate to the use of data, in the context of consent established by the data owner
    - Privacy is therefore distinct from confidentiality, although it may be supported by confidentiality mechanisms.
    - Grid technology needs a transferable understanding of suitable policies addressing privacy requirements/constraints
      - Should allow to express how such policies can be
        - » defined,
        - » applied,
        - » implemented,
        - » enforced, …

Security Concepts::Integrity

National e-Science Centre

- *Integrity*
  - Ensuring that data is not modified since it was created, typically of relevance when data is sent over public network
    - Technical solutions exist to maintain the integrity of data in transit
      - checksums, PKI support, …
    - Grid also raises more general questions
      - e.g. provenance
        - » maintaining the integrity of chains or groups of related data

Security Concepts::Fabric Management

National e-Science Centre

- *Fabric Management*
  - consists of the distributed computing, network resources and associated connections that support Grid applications
    - impacts Grid security in two ways:
      - an insecure fabric may undermine the security of the Grid
        - » Are all sites fully patched (middleware/OS)?
      - Can we limit damage of virus infected machine across Grid?
        - » Identify it, quarantine it, anti-virus update/patch, re-instate into VO, …
      - fabric security measures may impede grid operations
        - » e.g. firewalls may be configured to block essential Grid traffic
  - (I was up to 3am Wednesday morning writing a bid to solve this problem!) ;o)
    - EARWIGS – e-resEARch frameWork for Integrated Grid Security
      - Possibly the best acronym ever?

Security Concepts::Trust

National e-Science Centre

- *Trust*
  - characteristic allowing one entity to assume that a second entity will behave exactly as the first entity expects
  - Important distinction between 'trust management' systems which implement authorisation, and the wider requirements of trust
    - e.g. health applications require the agreement between users and resources providers of restrictions that cannot be implemented by access control
      - e.g. restrictions on the export of software, or a guarantee that personal data is deleted after use
    - therefore a need to understand and represent policy agreements between groups of users and resource providers
      - such policies may exist inside or outside the system, and are typically not supported by technical mechanisms

Grid Security v. Basics

National e-Science Centre

- We want all of the above, but…
  - Little consensus on most concepts
  - Best practice to copy
- Main area of agreement and adoption by Grid community is idea of Public Key Infrastructure (PKI)

Introduction to PKI

National e-Science Centre

- In the beginning of the internet security was not prime concern
  - No longer the case
    - Ever growing dependencies on security over internet
      - Banking, finances, shopping, …
  - Question is how do we implement it?
    - Collection of approaches, standards, solutions, …
  - Public Key Infrastructures (PKI) offer one possibility
    - Most obvious advantage of PKIs to Grid community is single sign-on

Public Key Infrastructures (PKI)

National e-Science Centre

- Public Key Infrastructure (PKI) responsible for deciding policy/managing, enforcing certificate validity checks
- Central component of PKI is *Certificate Authority* (CA)
  - CA has numerous responsibilities
    - Issuing certificates
      - Often need to delegate to local Registration Authority
        - » Prove who you are, e.g. with passport
    - Revoking certificates
      - Certificate Revocation List (CRL) for expired/compromised certificates
    - Storing, archiving
      - Keeping track of existing certificates, various other information, …
  - CA often (but not always) is trusted organisation to you/your organisation
    - UK e-Science has CA in Rutherford Appleton Labs
      - Strict, policies and procedures for getting Grid certificates

PKI Trust Model

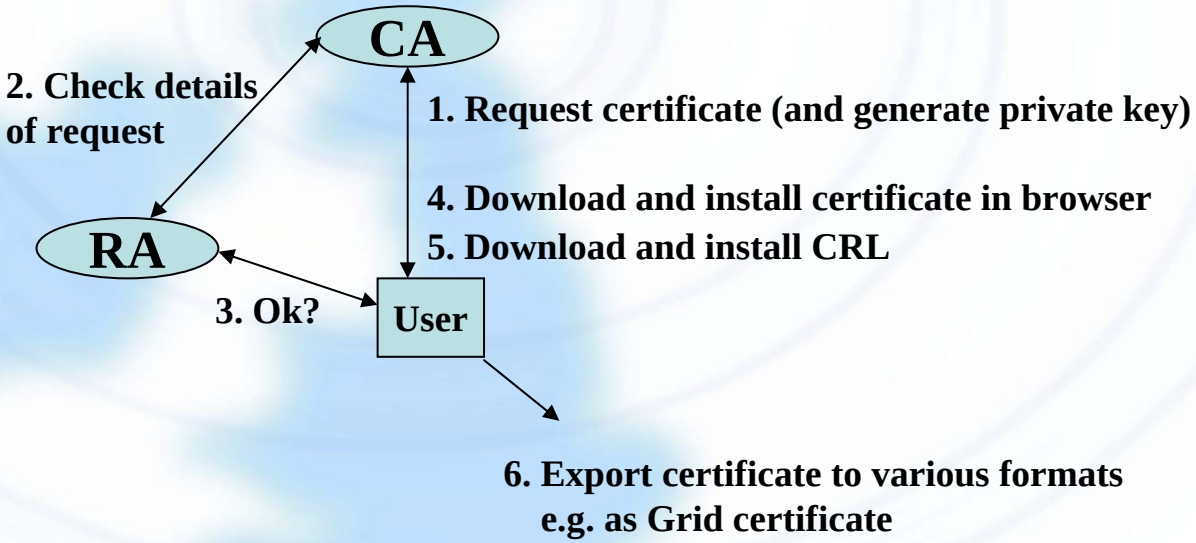
National e-Science Centre

- CA issues certificates
  - Could be to users, resources, other CAs, …
    - CA certificates can describe/limit trust relationship
- Issuing certificate is indication of trust
  - CA trusts it is really you who is applying for and going to use this certificate
  - You (and others using this CA) trust that certificates are managed correctly
- How to decide if CA is trustworthy?
  - Different choices
    - User decides to trust CA
    - CAs decide if they trust one another
      - *Certification paths* used to track trust relationships
- Different architectural choices for PKI impact upon certification paths and validity checking

UK e-Science PKI

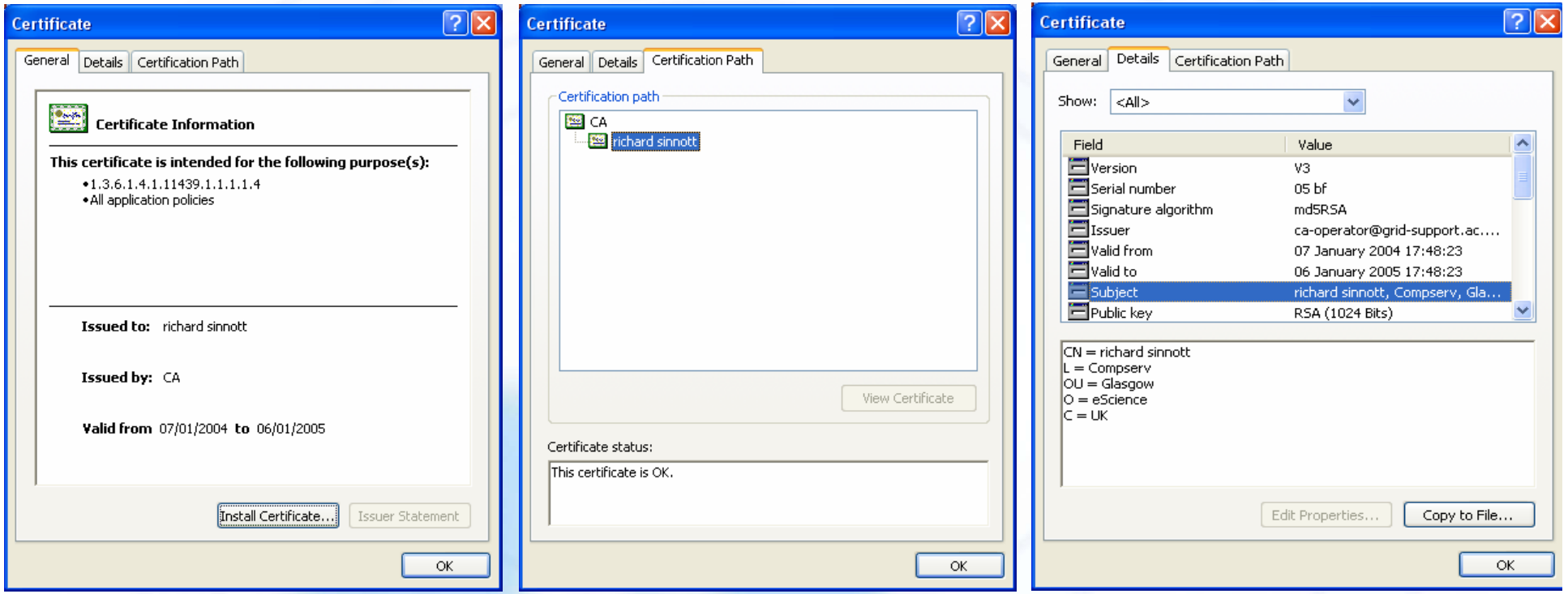
National e-Science Centre

- Based on statically defined centralised CA with direct single hierarchy to users
- Typical scenario for getting Grid certificate



Example X.509 Certificate

National e-Science Centre



So what has this to do with Grids?

National e-Science Centre

- PKIs allow for single sign on!
- For example, assume Grid used Globus infrastructure with GSI
  - (Assume you have heard about this already?)
- Basic idea is all sites in VO have locally administrated Grid map file
  - Globus uses gridmap file consisting of Distinguished Name : local account
    - "/C=UK/O=eScience/OU=Glasgow/L=Compserv/CN=richard sinnott" ros
    - …
  - VO sites can check that invoker has appropriate credentials
    - e.g. cert is issued by UK e-Science CA
  - Provided everyone trusts UK e-Science CA then I can get access to any site where my cert is recognised
    - Hence single password for my cert is used to get me access to all sites
  - Often manual process for grid mapfile although technologies like VOMS can be used to dynamically update numerous grid mapfiles

PKI Issues

National e-Science Centre

- So what is wrong with PKI
  - Only authentication support (not authorisation)
    - Not able to restrict user actions
    - Collections of users identified and statically defined trust relationships
    - But what if want to dynamically establish a VO where different users have different roles, different responsibilities and resources themselves are changing…?
      - PKIs in themselves do not support this possibility
  - And need all other security aspects (auditing, privacy, …)

What we need is…

National e-Science Centre

- Technologies for establishment of arbitrary VOs
  - need rules/contracts (policies)
    - Who can do what, on what, in what context, …
- Policies can be direct assertions/obligations/prohibitions on specific resources/users
  - Policies can be local to VO members/resources
    - e.g. user X from site A can have access to P% resource B on site C
      - (site C responsible for local policy – autonomy!!!)
  - Policies can be on remote resources
    - users from site A can access / download data Y from site B provided they do not make it available outside of site A
      - …site B trusts site A to ensure this is the case
        - » and possibly to ensure that the security is comparable with site B
        - » … trust!!!

Authorization Technologies for VO

National e-Science Centre

- Various technologies for authorization including
  - PERMIS
    - PrivilEge and Role Management Infrastructure Standards Validation
      - http://www.permis.org
  - Community Authorisation Service
    - http://www.globus.org/security/CAS/
  - AKENTI
    - http://www-itg.lbl.giv/security/akenti
  - CARDEA
    - http://www.nas.nasa.gov/Research/Reports/Techreports/2003/nas-03-020-abstract.html
  - VOMS
    - http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html
  - All of them predominantly work at the local policy level

Role Based Access Controls

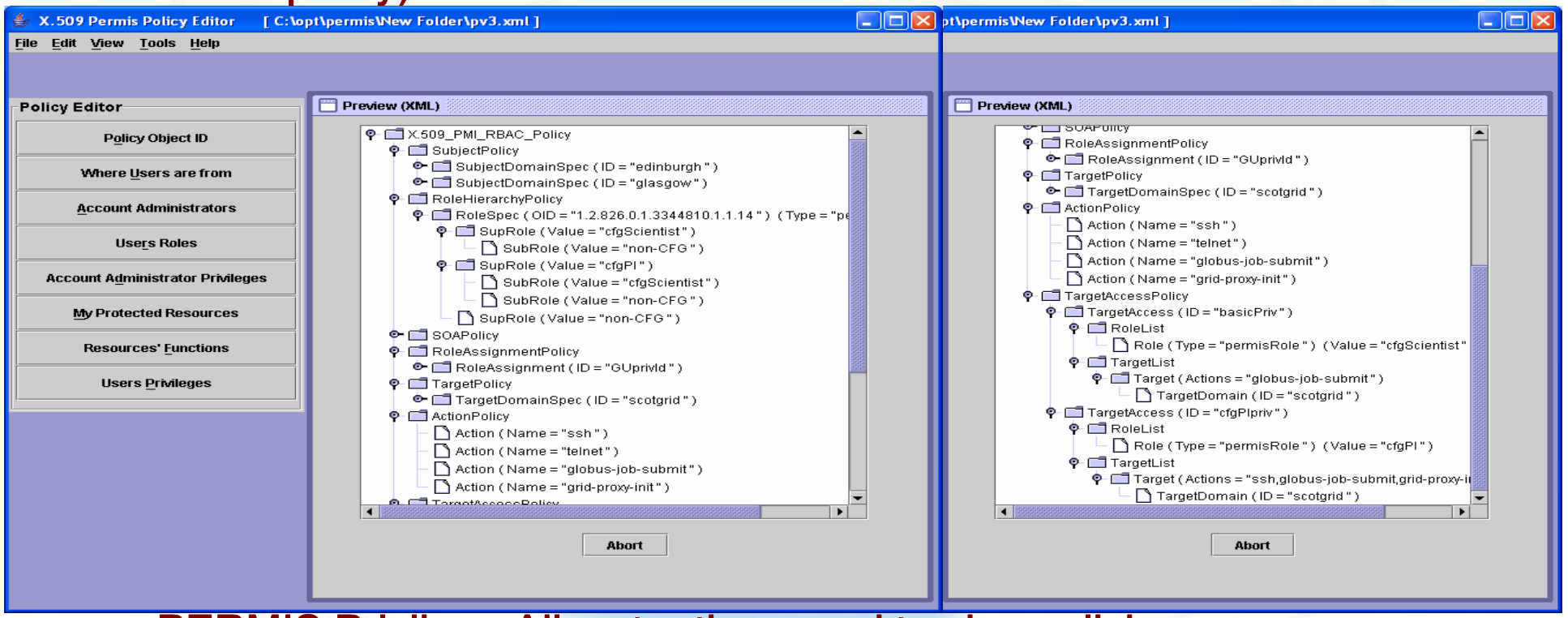
National e-Science Centre

- Need to be able to express and enforce policies
  - Common approach is role based authorisation infrastructures
    - PERMIS, CAS, …
- Basic idea is to define:
  - *roles* applicable to specific VO
    - roles often hierarchical
      - Role X ≥ Role Y ≥ Role Z
      - Manager can do everything (and more) than an employee can do who can do everything (and more) than a trainee can do
  - *actions* allowed/not allowed for VO members
  - *resources* comprising VO infrastructure (computers, data resources etc)
- A policy then consists of sets of these rules
  - *{ Role x Action x Target }*
    - Can user with VO role X invoke service Y on resource Z?
  - Policy itself can be represented in many ways,
    - e.g. XML document, SAML, XACML, …

PERMIS Based Authorisation

National e-Science Centre

- PERMIS Policies created with PERMIS PolicyEditor (output is XML based policy)



- PERMIS Privilege Allocator then used to sign policies
  - Associates roles with specific users
    - Policies stored as attribute certificates in LDAP server

Grid APIs for Generic Authorisation

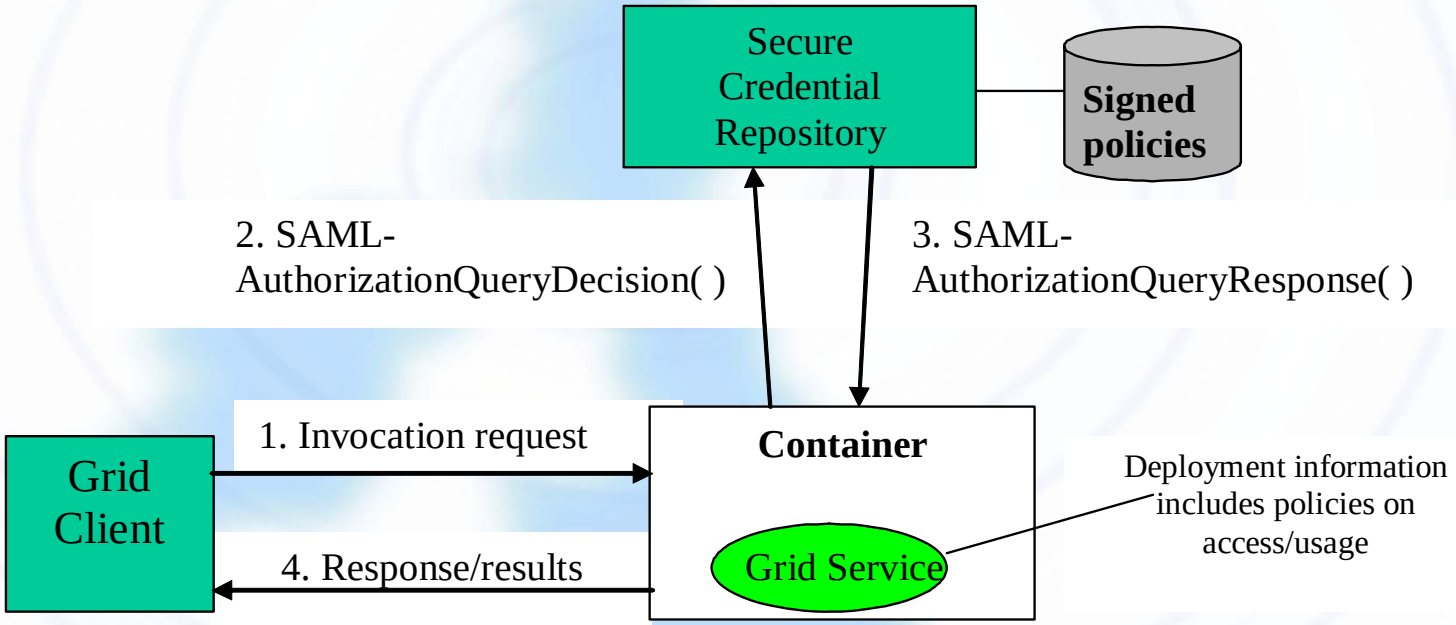
National e-Science Centre

- We don't want to have to re-engineer services to make them secure!
- GGF have defined generic API for Grid service authorization
  - SAML AuthZ specification defines a number of elements for making assertions and queries regarding authentication, authorization decisions
  - Includes message exchange between a policy enforcement point (PEP) and a policy decision point (PDP)
    - consisting of *AuthorizationDecisionQuery* flowing from the PEP to the PDP, with an assertion returned containing some number of *AuthorizationDecisionStatements*
    - *AuthorizationDecisionQuery* itself consists of
      - A *Subject* element containing a *NameIdentifier* specifying the initiator identity
      - A *Resource* element specifying the resource to which the request to be authorized is being made.
      - One or more *Action* elements specifying the actions being requested on the resources
    - Result is a *SimpleAuthorizationDecisionStatement* (granted/denied Boolean) and an *ExtendedAuthorizationDecisionQuery* that allows the PEP to specify whether the simple or full authorization decision is to be returned

Grid APIs for Generic Authorisation …ctd

National e-Science Centre

- SAML AuthZ specification provides generic PEP approach for ALL Grid services
  - … or at least all GT3.3+ based services



- PDP application specific
  - Default behaviour is if not explicitly granted by policy, then rejected

Life Sciences & Grids

National e-Science Centre

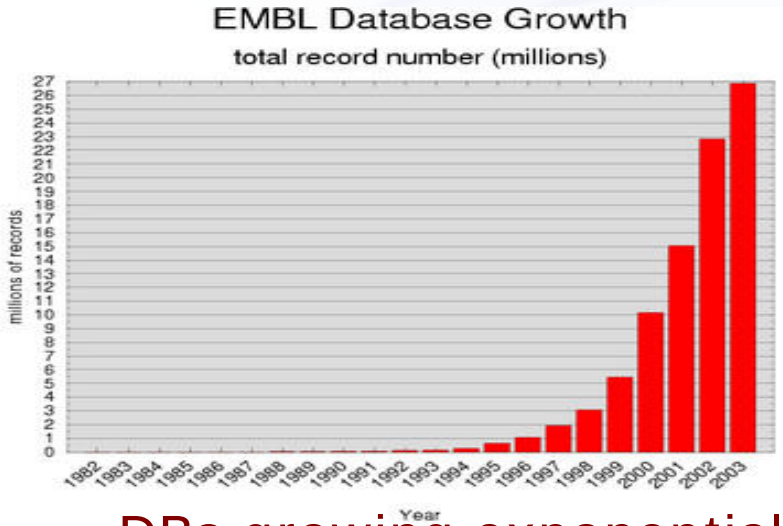
- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

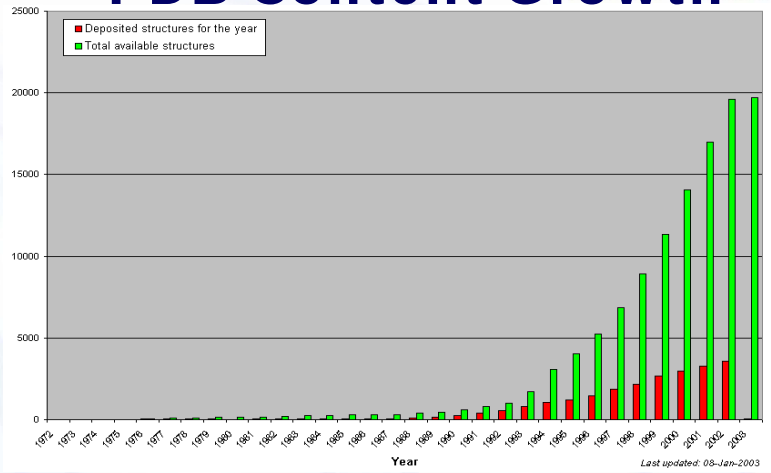
Grids & Life Sciences

National e-Science Centre

- Extensive Research Community
  - >1000 per research university
- Extensive Applications
  - Many people care about them
    - Health, Food, Environment, …
- Interacts with many disciplines
  - Physics, Chemistry, Maths/Statistics, Nano-engineering, …
- Huge and expanding number of databases relevant to bioinformatics community
  - Heterogeneity, Interdependence, Complexity, Change, Dirty…
- Linking using in co-ordinated, secure manner full of open issues to be addressed
- Compute demands growing as more *in-silico* research undertaken

Database Growth





- DBs growing exponentially!!!
  - Biobliographic (MedLine, …)
  - Amino Acid Seq (SWISS-PROT, …)
  - 3D Molecular Structure (PDB, …)
  - Nucleotide Seq (GenBank, EMBL, …)
  - Biochemical Pathways (KEGG, WIT…)
  - Molecular Classifications (SCOP, CATH,…)
  - Motif Libraries (PROSITE, Blocks, …)

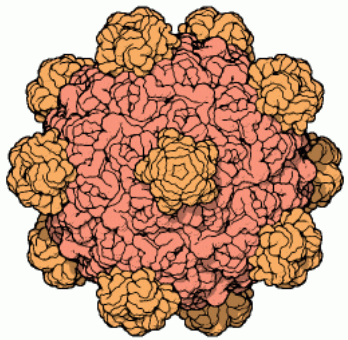
Distributed and Heterogeneous data

National e-Science Centre

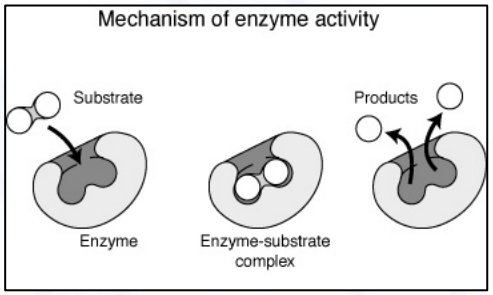
Sequence

```
LPSYVDWRSA GAVVDIKSQG
ECGGCWAFSA IATVEGINKI
TSGSLISLSE QELIDCGRTQ
NTRGCDGGYI TDGFQFIIND
GGINTEENYP YTAQDGDCDV
```

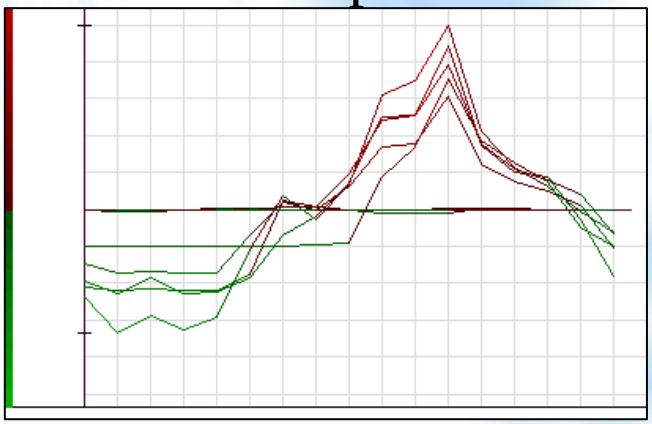
Structure



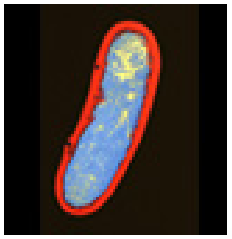
Function



Gene expression



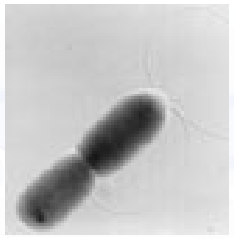
Morphology



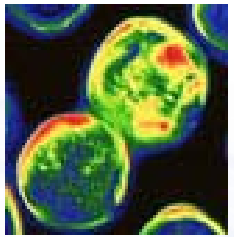
Yersinia pestis

Arabidopsis thaliana

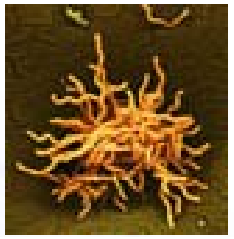
Buchnerasp. APS



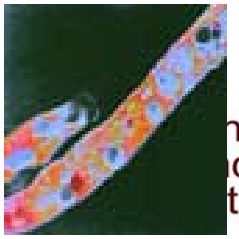
Aquifex aeolicus



Archaeoglobus fulgidus

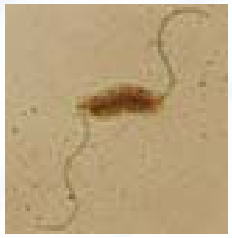


Borrelia burgorferi

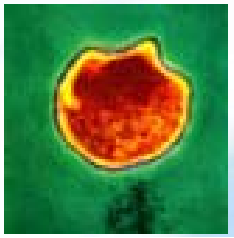


Mycobacterium tuberculosis

Caenorhabitis elegans

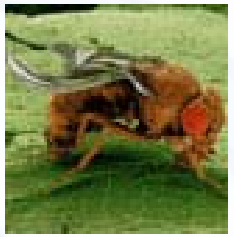


Campylobacter jejuni



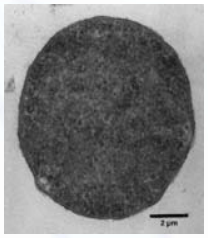
Chlamydia pneumoniae

Vibrio cholerae

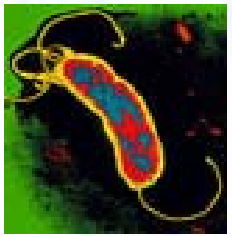


Drosophila melanogaster

Escherichia coli

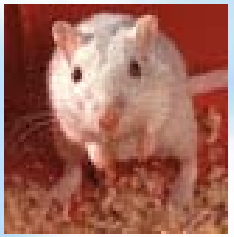


Thermoplasma acidophilum



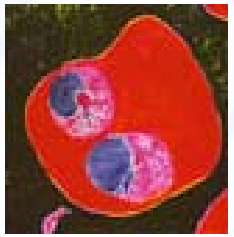
Helicobacter pylori

Mycobacterium leprae

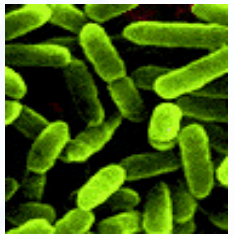


mouse

Neisseria meningitidis Z2491



Plasmodium falciparum

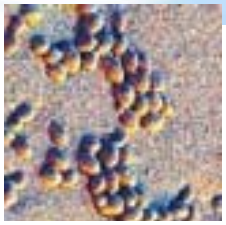


Pseudomonas aeruginosa

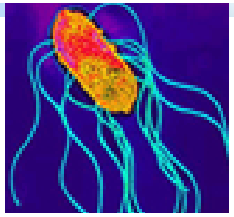
Ureaplasma urealyticum

rat

Rickettsia prowazekii

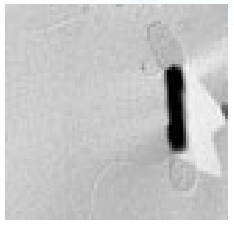


Saccharomyces cerevisiae

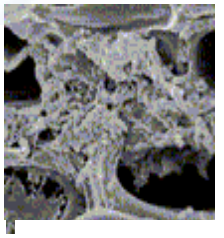


Salmonella enterica

Bacillus subtilis



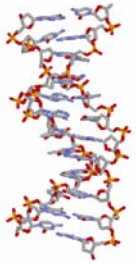
Thermotoga maritima

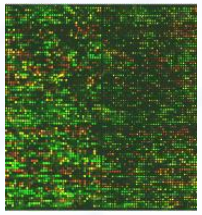


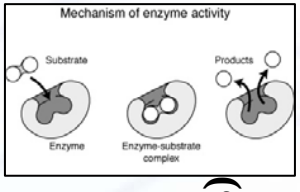
Xylella fastidiosa

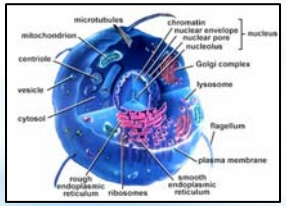
Systems Biology

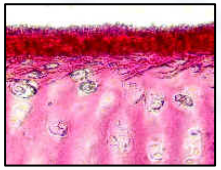
National e-Science Centre

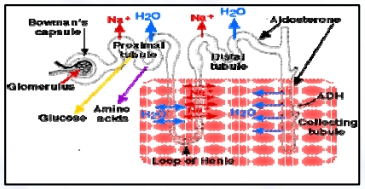












Nucleotide sequences

Nucleotide structures

Gene expressions

Protein Structures

Protein functions

Protein-protein interaction (pathways)

Cell

Cell signalling

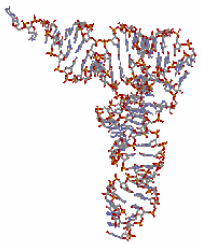
Tissues

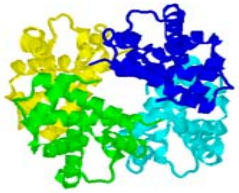
Organs

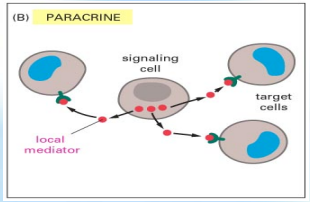
Physiology

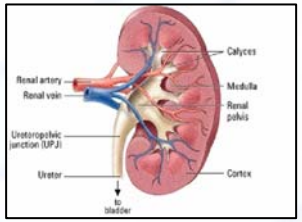
Organisms

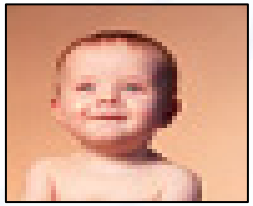
Populations

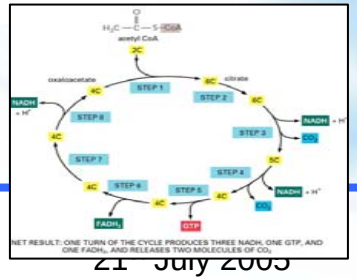












+ links to plant/crops, environmental, health, … information sources

e-Science dti

UNIVERSITY of GLASGOW

Is Grid the Answer?

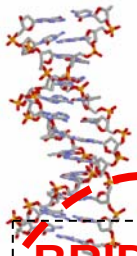
National e-Science Centre

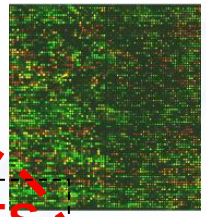
- Some key problems to be addressed
  - Tools that *simplify* access to and usage of data
    - Internet hopping is not ideal!
  - Tools that *simplify* access to and usage of large scale HPC facilities
    - qsub [-a date_time] [-A account_string] [-c interval] [-C directive_prefix] [-e path] [-h] [-l] [-j join] [-k keep] [-l resource_list] [-m mail_options] [-M user_list] [-N name] [-o path] [-p priority] [-q destination] [-r c] [-S path_list] [-u user_list] [-v variable_list] [-V] [-W additional_attributes] [-z] [script]
  - Tools designed to *aid understanding* of complex data sets and relationships between them
    - e.g. through visualisation
  - Make it all easy to use!
    - Scientists should not have to be Linux script experts,
    - …nor set up/configure complex Grid software or follow complex procedures for getting, using Grid certificates,
    - …nor have detailed understanding of low level data schemas for all data sites,
    - … etc etc

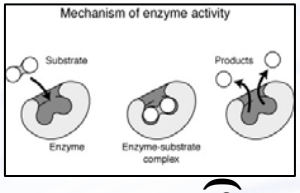
Overview of BRIDGES

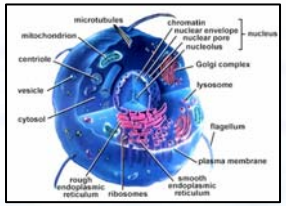
National e-Science Centre

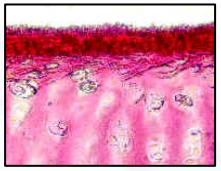
- Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES)
  - NeSC (Edinburgh and Glasgow) and IBM
  - Started October 2003
- Supporting project for CFG project
  - Generating data on hypertension
  - Rat, Mouse, Human genome databases
- Variety of tools used
  - BLAST, BLAT, Gene Prediction, visualisation, …
- Variety of data sources and formats
  - Microarray data, genome DBs, project partner research data, …
- Aim is integrated infrastructure supporting
  - Data federation
  - Security

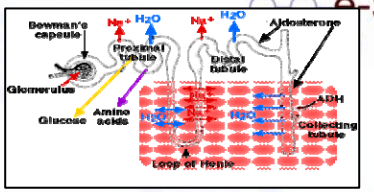












National e-Science Centre

BRIDGES

Nucleotide sequences

Nucleotide structures

Gene expressions

Protein Structures

Protein functions

Protein-protein interaction (pathways)

Cell

Cell signalling

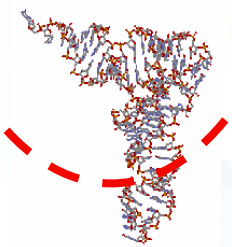
Tissues

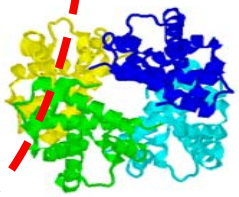
Organs

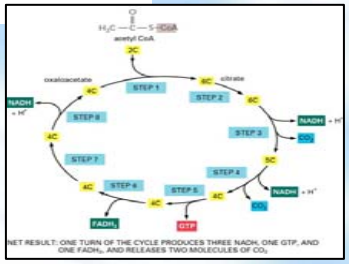
Physiology

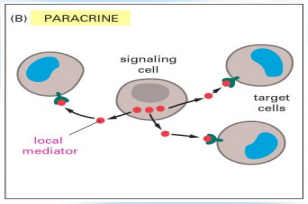
Organisms

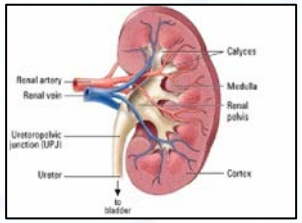
Populations

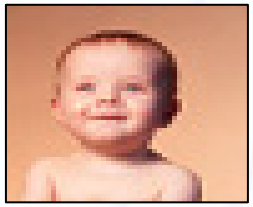








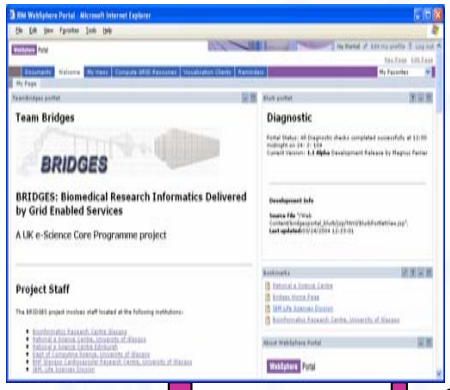


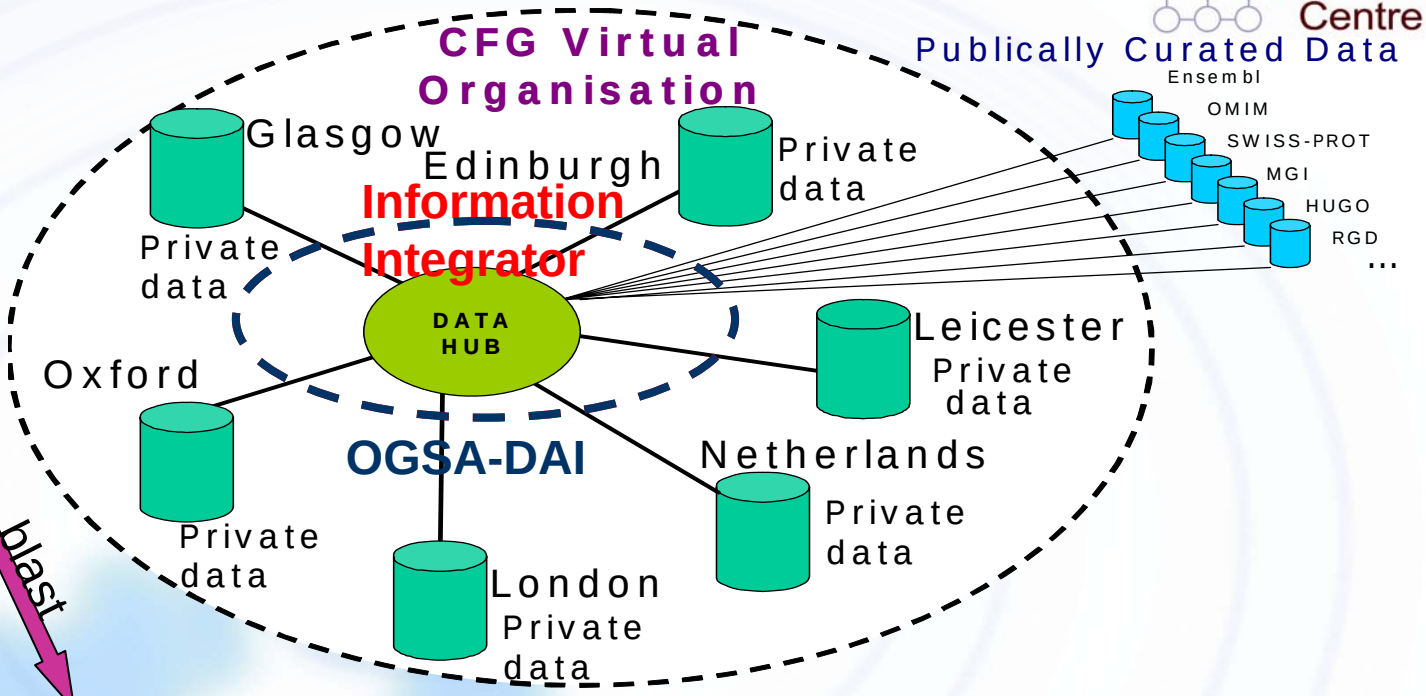


BRIDGES Project

National e-Science Centre

VO Authorisation

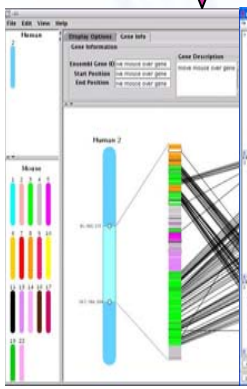


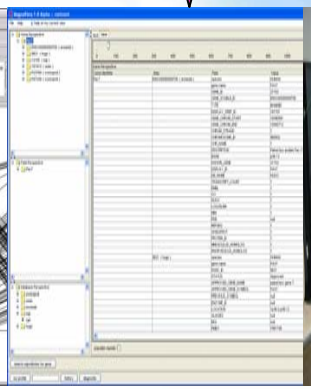


Synteny Service

Magna Vista Service

blast

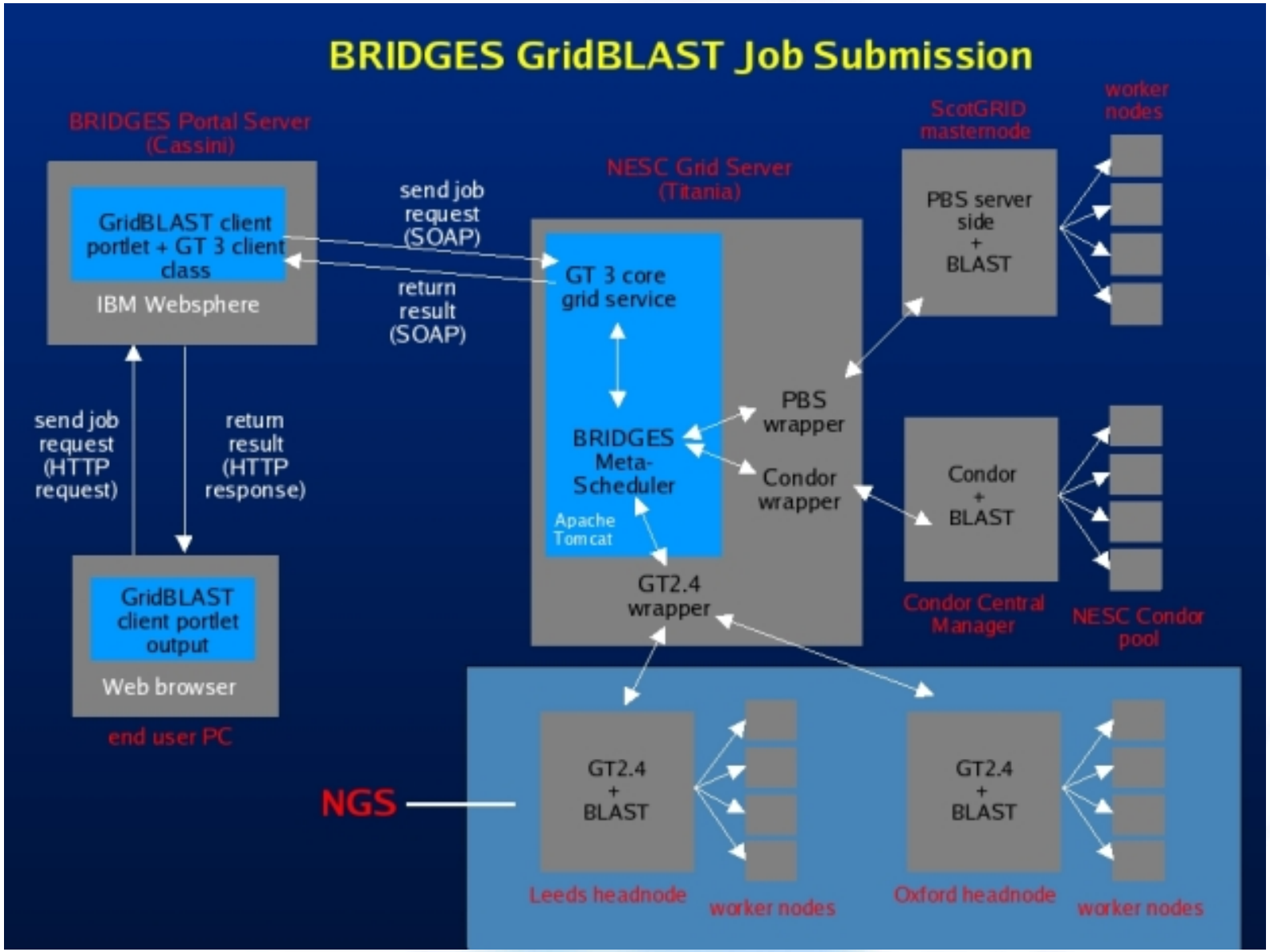




+

+

+



National e-Science Centre

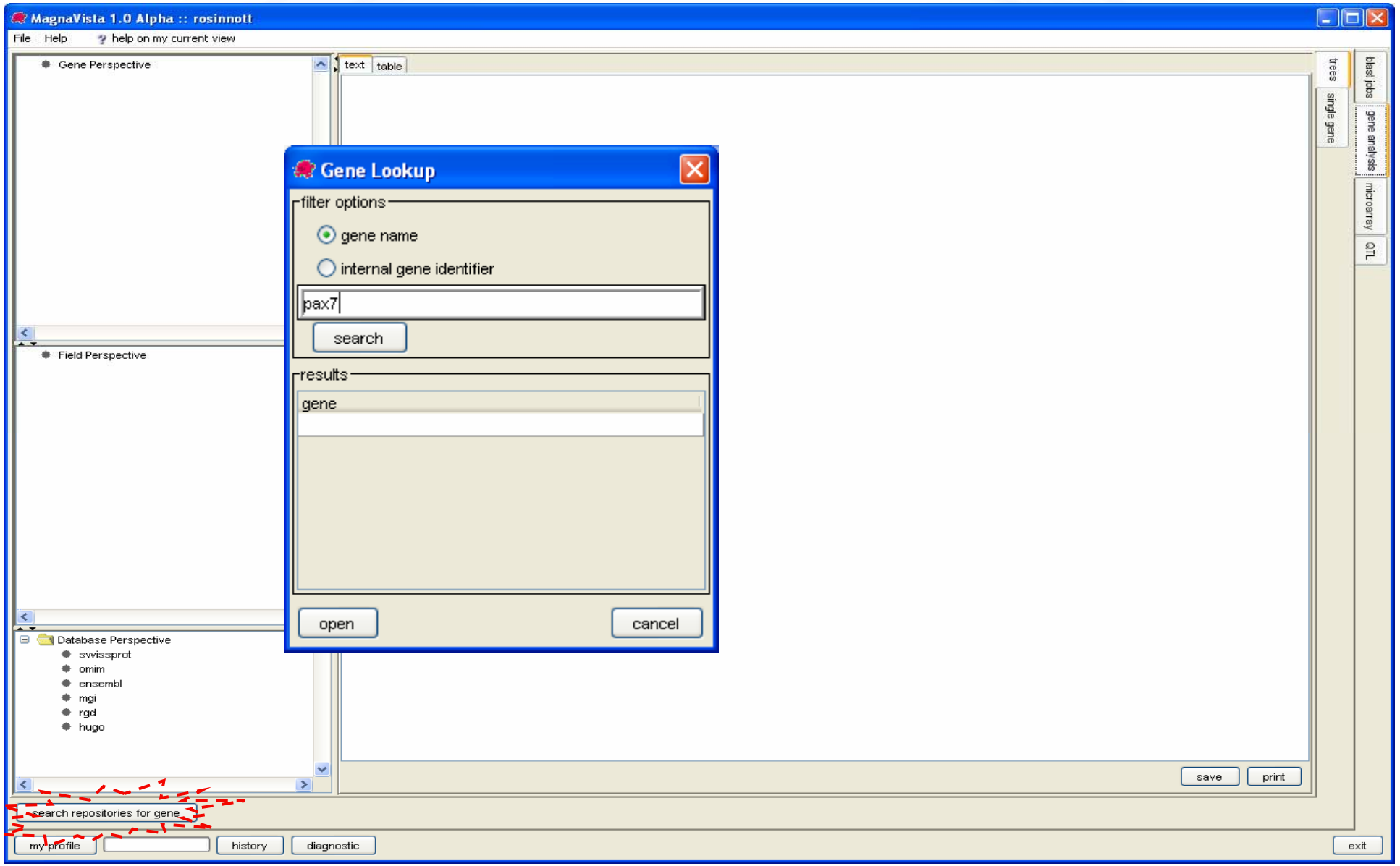
Demonstrations

National e-Science Centre

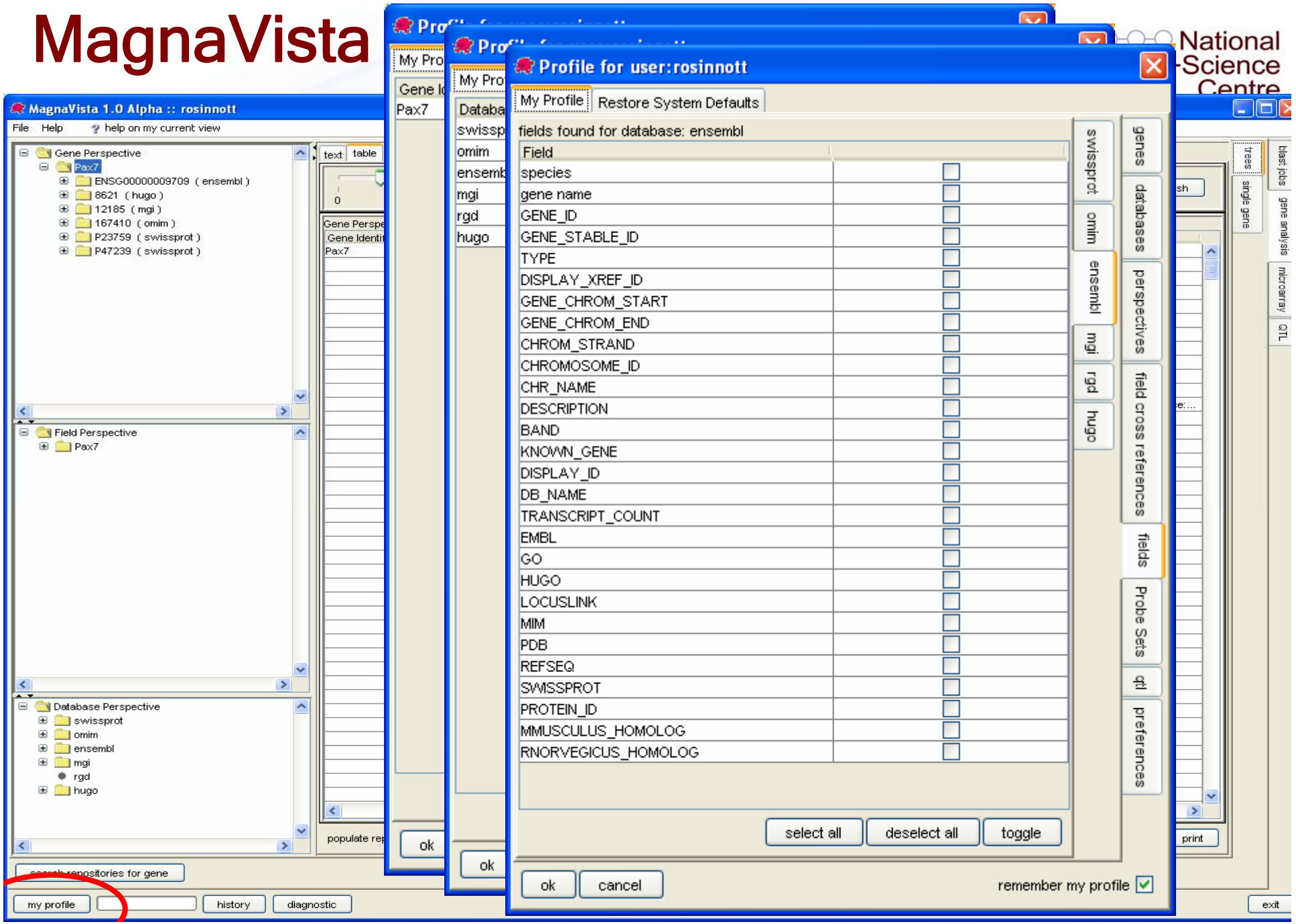
- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

MagnaVista

National e-Science Centre



MagnaVista


MagnaVista 1.0 Alpha :: rosinnott
File Help help on my current view
Gene Perspective
Pax7
ENSG00000009709 (ensembl)
8621 (hugo)
12185 (mgi)
167410 (omim)
P23759 (swissprot)
P47239 (swissprot)
Field Perspective
Pax7
Database Perspective
swissprot
omim
ensembl
mgi
rgd
hugo
text table
search repositories for gene
my profile
history
diagnostic
Profile for user:rosinnott
My Profile
Restore System Defaults
fields found for database: ensembl
Field
species
gene name
GENE_ID
GENE_STABLE_ID
TYPE
DISPLAY_XREF_ID
GENE_CHROM_START
GENE_CHROM_END
CHROM_STRAND
CHROMOSOME_ID
CHR_NAME
DESCRIPTION
BAND
KNOWN_GENE
DISPLAY_ID
DB_NAME
TRANSCRIPT_COUNT
EMBL
GO
HUGO
LOCUSLINK
MIM
PDB
REFSEQ
SWISSPROT
PROTEIN_ID
MMUSCULUS_HOMOLOG
RNORVEGICUS_HOMOLOG
select all
deselect all
toggle
ok
cancel
remember my profile
swissprot
omim
ensembl
mgi
rgd
hugo
genes
databases
perspectives
field cross references
fields
Probe Sets
qtl
preferences
National Science Centre
trees
single gene
blast jobs
gene analysis
microarray
QTL
print
exit

Other NeSC Projects

National e-Science Centre

- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future

Scottish Bioinformatics Research Network

National e-Science Centre

- Four year proposal expected to start imminently
  - Funded (£2.4M) by Scottish Enterprise, Scottish Higher Education Funding Council, Scottish Executive Environment and Rural Affairs Department
    - Involves Glasgow, Dundee, Edinburgh, Scottish Bioinformatics Forum
  - Aim to provide bioinformatics infrastructure for Scottish health, agriculture and industry
    - Infrastructure support at Dundee, Edinburgh and Glasgow to support first-rate research in bioinformatics at each academic institute
    - Infrastructure support at three institutes, to support inter-institutional sharing of compute and data resources through application of Grid computing
    - Outreach and training activities mediated by the Scottish Bioinformatics Forum

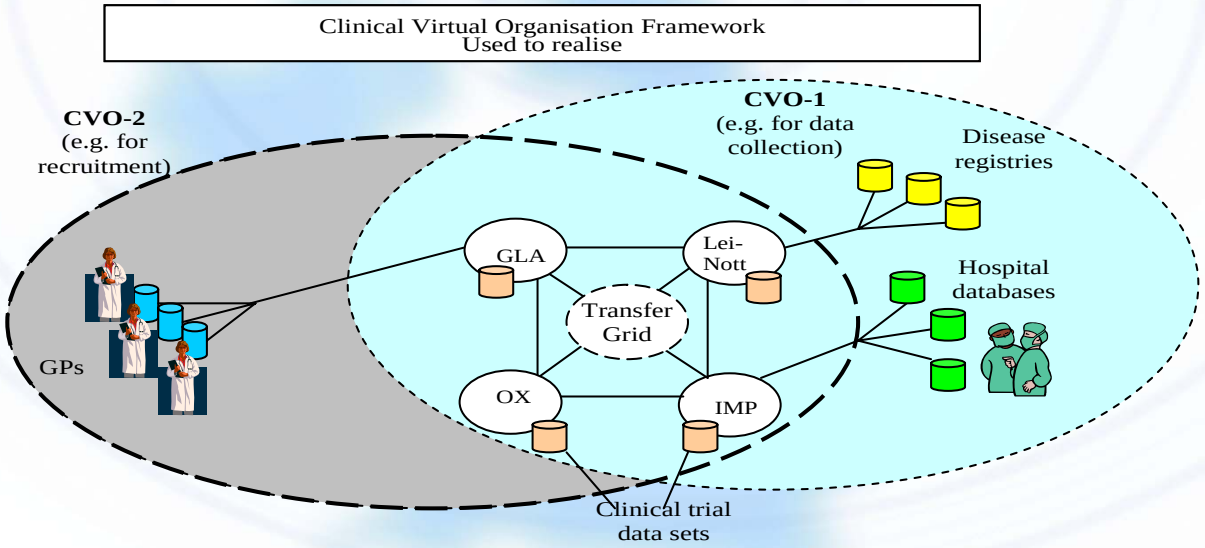
Grid Enabled Microarray Expression Profile Search

National e-Science Centre

- 1 year project expected to start 1st September
  - Funded (£61k) by BBSRC
    - Involves Glasgow, Cornell University, US, Riken Institute, Japan
  - Aim to provide tools for discovery, comparison and analysis of microarray data sets
    - How does my data compare to others?
    - How do these experiments compare?
    - Can we improve the way we establish how genes in different species are linked?
  - Requires data access, integration and move towards data mining
  - Built upon fine grained security
    - Microarrays expensive and contain potentially important (valuable) data sets

VOTES

- Virtual Organisations for Trials and Epidemiological Studies
  - 3 year MRC (£2.8M) funded project just started
  - Plans to develop Grid infrastructure to address key components of clinical trial/observational study
    - Recruitment of potentially eligible participants
    - Data collection during the study
    - Study administration and coordination
      - Involves Glasgow, Oxford, Leicester, Nottingham, Manchester



Generation Scotland
Scottish Family Health Study

National e-Science Centre

- Five (2+3) year proposal (£4.4M) just started
  - Funded by Health Department and Department for Enterprise and Lifelong Learning
    - Involves Glasgow, Dundee, Edinburgh, Aberdeen

      - focus of genetics as applied to healthcare

      - first two years emphasis on providing a platform for research into the genetic basis of common complex diseases in Scotland
        - » Mental health, cardiovascular, …
        - » Plan to establish 15,000 family-based intensively-phenotyped cohort recruited from the East and West of Scotland

      - basis for neutralising heritable (genetic) risk factors in disease surveillance, treatment optimisation, avoidance of adverse drug events and prediction of response to therapy, health care planning and drug discovery, …

JDSS Project

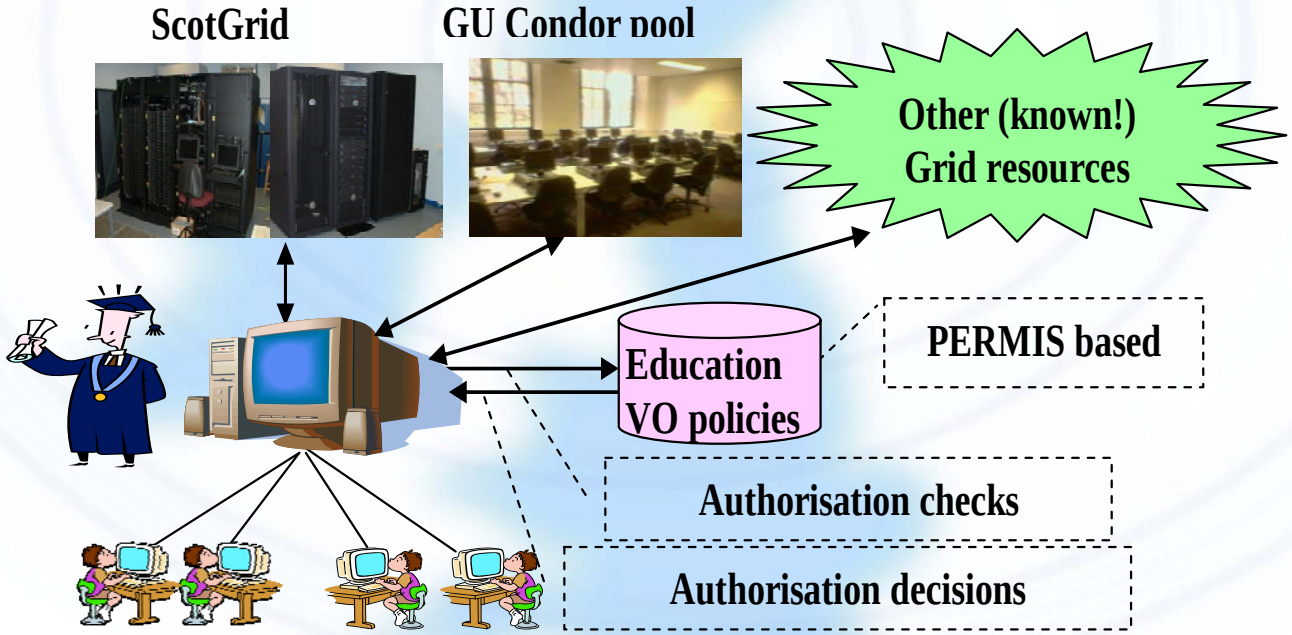
National e-Science Centre

- Public data resources openness
  - Often cannot query directly
  - Often not easy/possible to find schemas
  - Joint Data Standards Study investigating this
    - Started on 1st June and involves
      - Digital Archiving Consultancy
      - Bioinformatics Research Centre (Glasgow)
      - NeSC (Edinburgh and Glasgow)
        - » Funded by MRC, BBSRC, Wellcome Trust, JISC, NERC, DTI
    - Look at technical, political, social, ethical etc issues involved in accessing and using public life science resources
      - Interview relevant scientists, data curators/providers
    - 8 month project with final report due imminently

DyVOSE Project

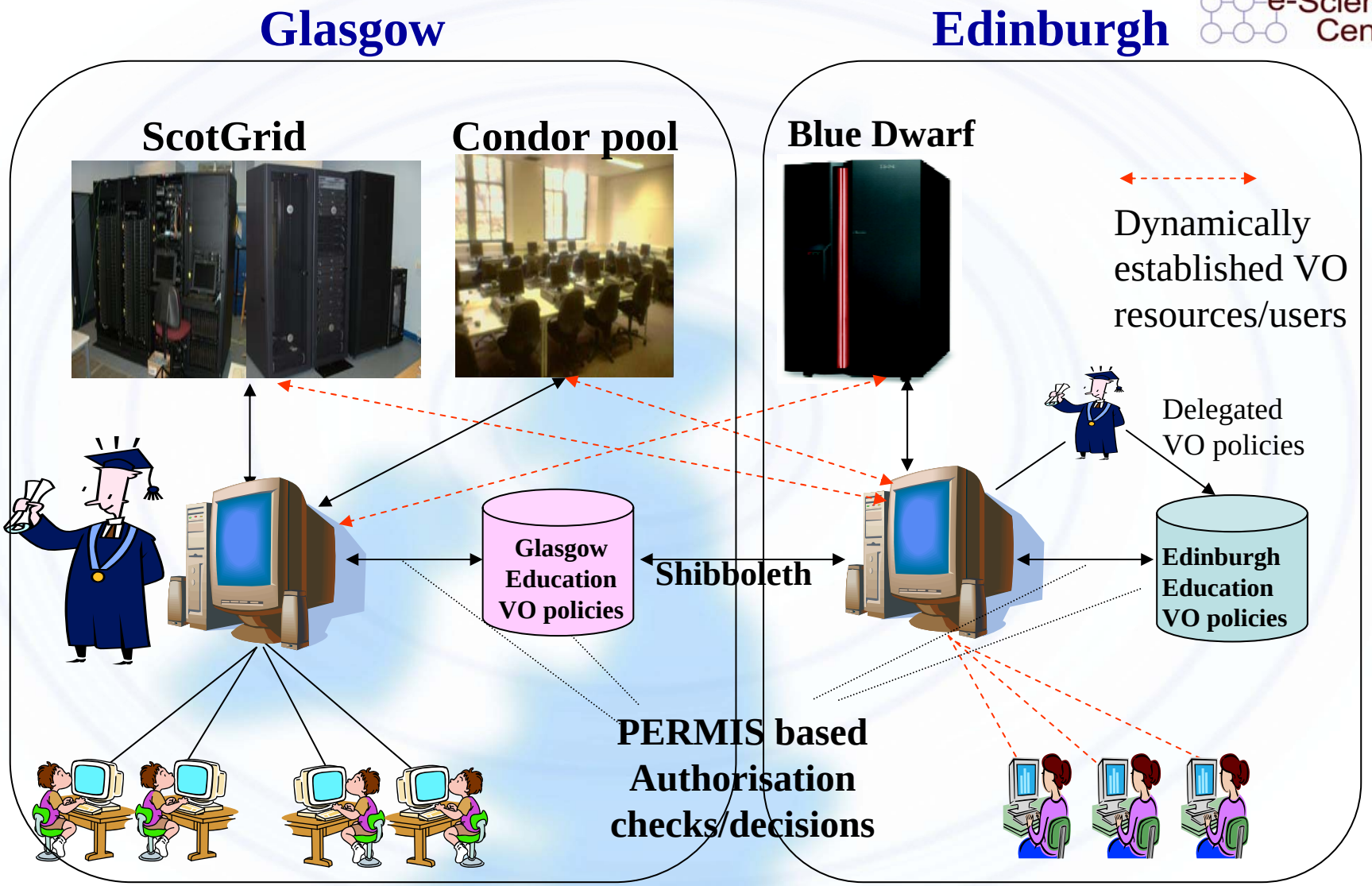
National e-Science Centre

- Dynamic Virtual Organisations for e-Science Education (DyVOSE) project
  - Two year project started 1st May 2004 funded by JISC
  - Exploring advanced authorisation infrastructures for security
    - … in Grid Computing Module as part of advanced MSc at Glasgow
      - Provide insight into rolling Grid out to the masses!



DyVOSE Phase 2/3

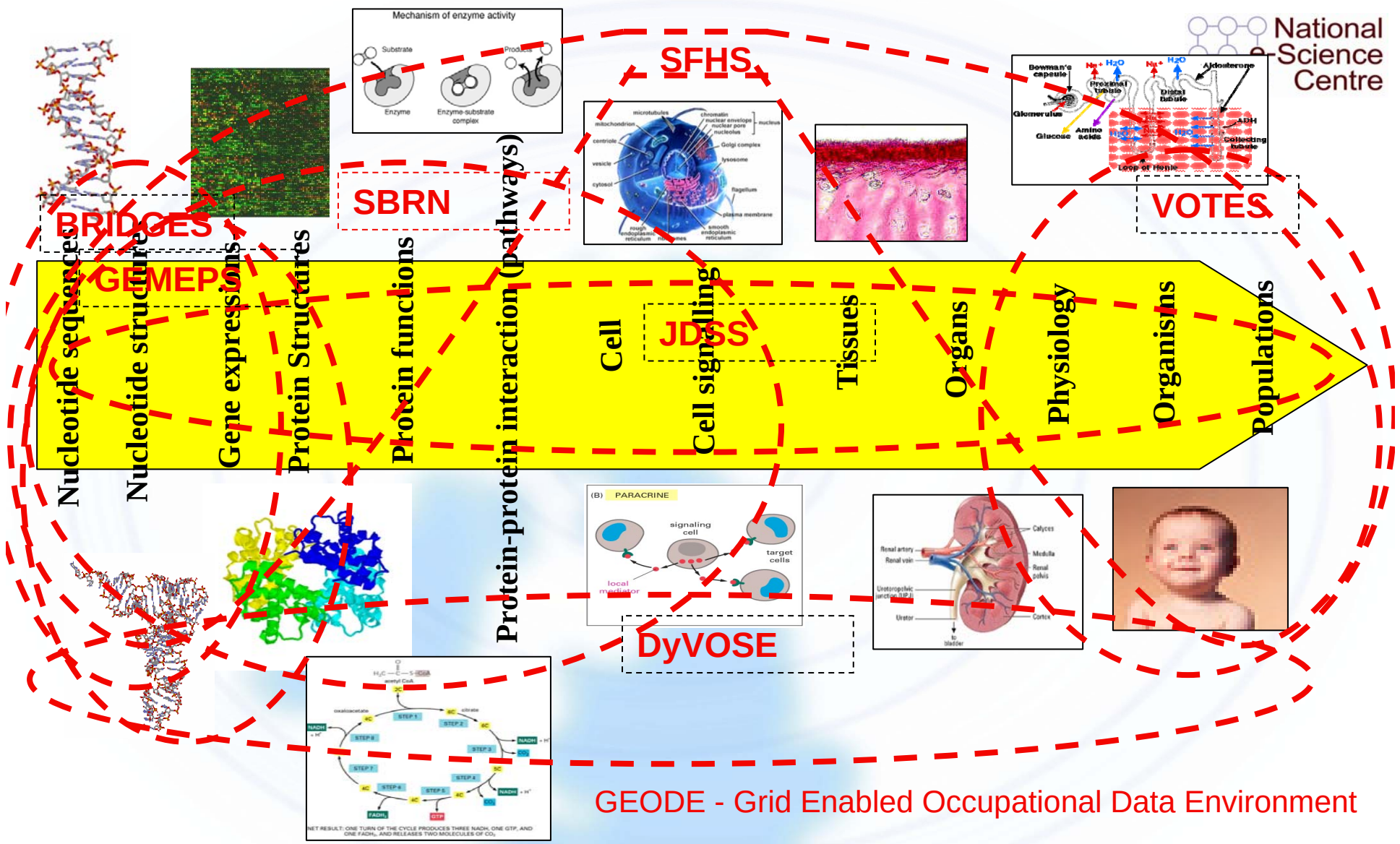
National e-Science Centre



Outlook

National e-Science Centre

- Grids and Grid Research
  - Classic “big-science”
- NeSC
  - NeSC at Glasgow
- Grid Security
  - Concepts, Grid Requirements, Technologies, …
- Break (10 mins?)
- Life Sciences and Grids
- Demonstrations
- Related NeSC projects
- Outlook for the future


National e-Science Centre
SFHS
SBRN
BRIDGES
GEMEP
VOTES
JDSS
DyVOSE
Nucleotide sequences
Nucleotide structure
Gene expressions
Protein Structures
Protein functions
Protein-protein interaction (pathways)
Cell
Cell signalling
Tissues
Organs
Physiology
Organisms
Populations
Mechanism of enzyme activity
Substrate
Enzyme
Enzyme-substrate complex
Product
(B) PARACRINE
signaling cell
target cells
local mediator
GEODE - Grid Enabled Occupational Data Environment
… many other bids submitted for parts of this picture

Systems Biology = the gNeSC-gNiche?

National e-Science Centre

- Once we have (securely) connected all relevant data sets and simplified access to and usage of HPC resources, wrapped your favourite bioinformatics applications as Grid services, linked them to clinical data sets…
  - what questions would you like to ask?
    - How does a cell work?
    - Why do people who eat less tend to live longer?
    - How many people across Scotland had a heart attack in the last 5 years took drug X, and of those that did where genes A or B influenced by this drug?
    - Who has performed an experiment similar to mine and where their results similar?
    - …