Sample Applications

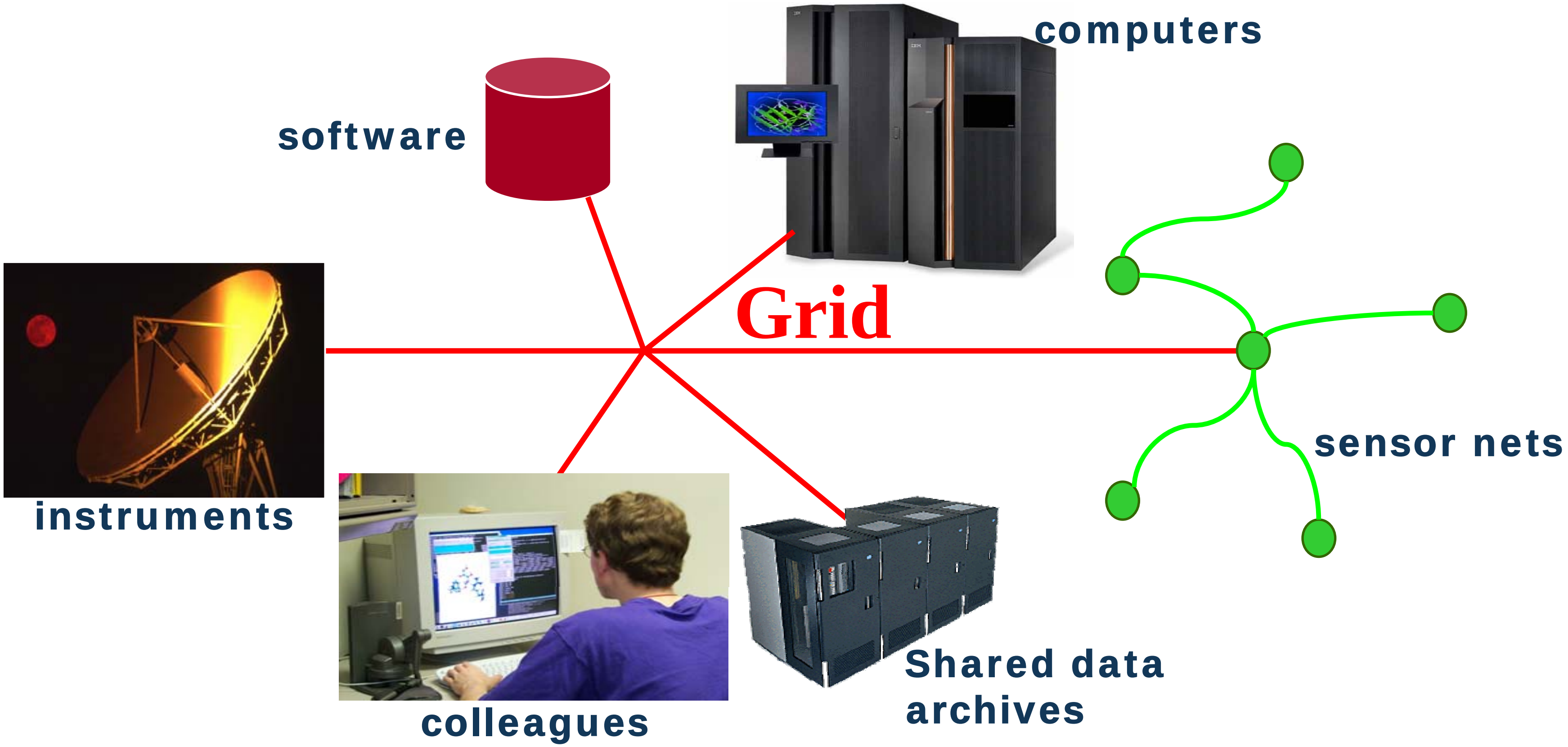
Dr. Richard Sinnott

http://csperkins.org/teaching/2004-2005/gc5/

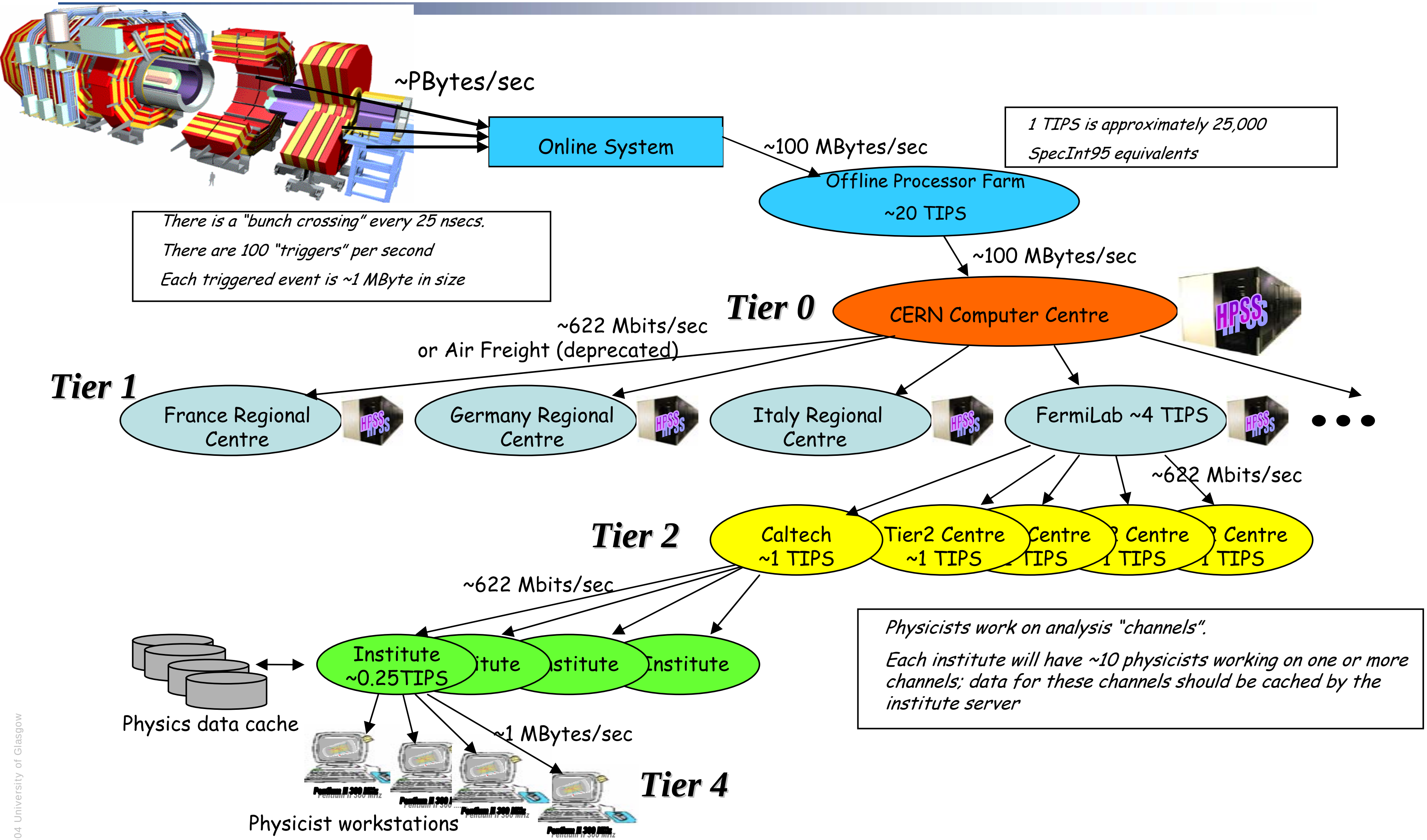
UNIVERSITY
of
GLASGOW

Foundation for e-Science

- e-Science 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



~PBytes/sec

Online System

~100 MBytes/sec

1 TIPS is approximately 25,000

SpecInt95 equivalents

Offline Processor Farm

~20 TIPS

There is a "bunch crossing" every 25 nsecs.

There are 100 "triggers" per second

Each triggered event is ~1 MByte in size

~100 MBytes/sec

Tier 0

CERN Computer Centre

~622 Mbits/sec

or Air Freight (deprecated)

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

Centre ~1 TIPS

2 Centre ~1 TIPS

2 Centre ~1 TIPS

~622 Mbits/sec

Institute ~0.25TIPS

Institute

Institute

Institute

Physics data cache

Physicists work on analysis "channels".

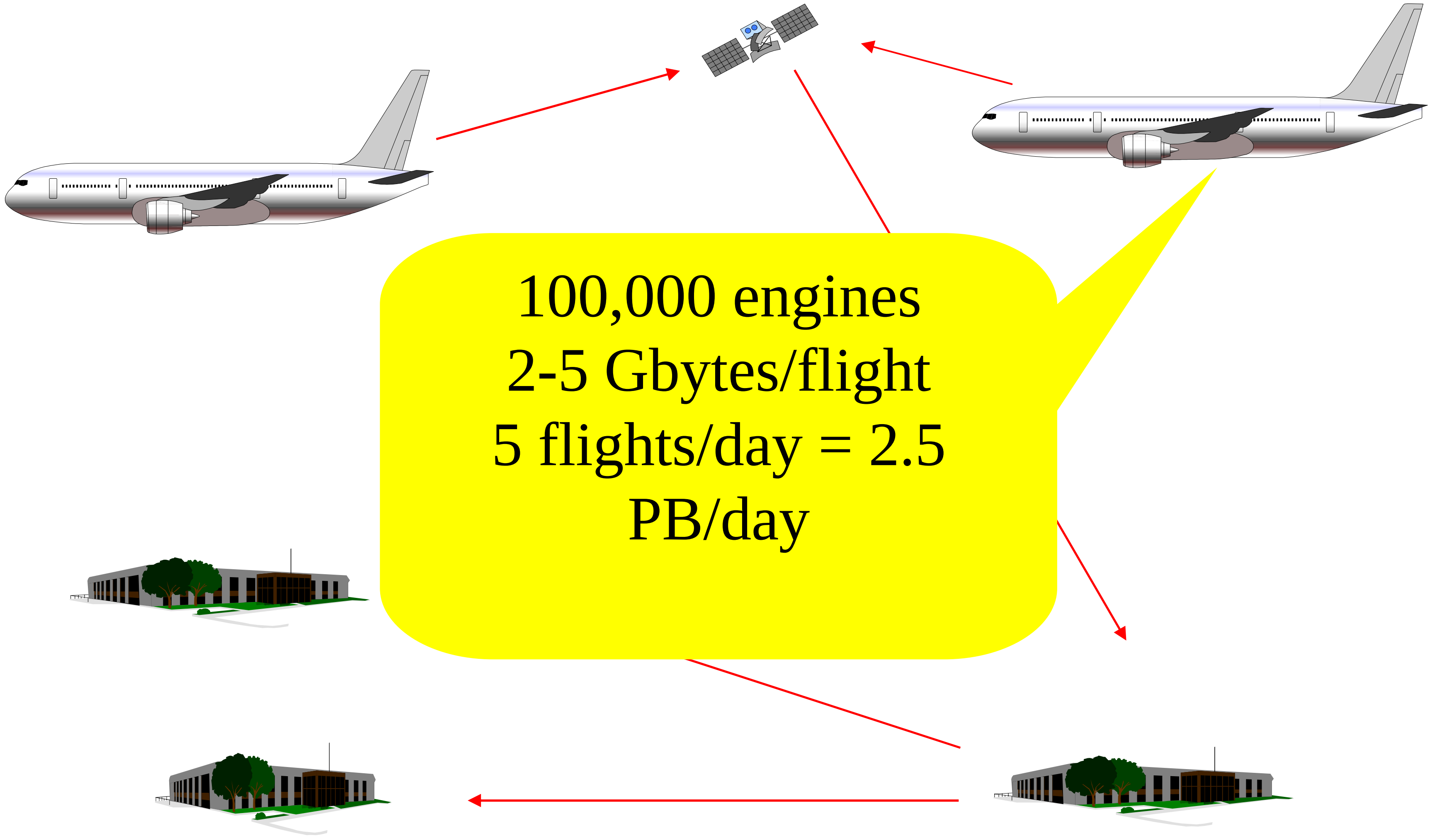
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server

~1 MBytes/sec

Tier 4

Physicist workstations

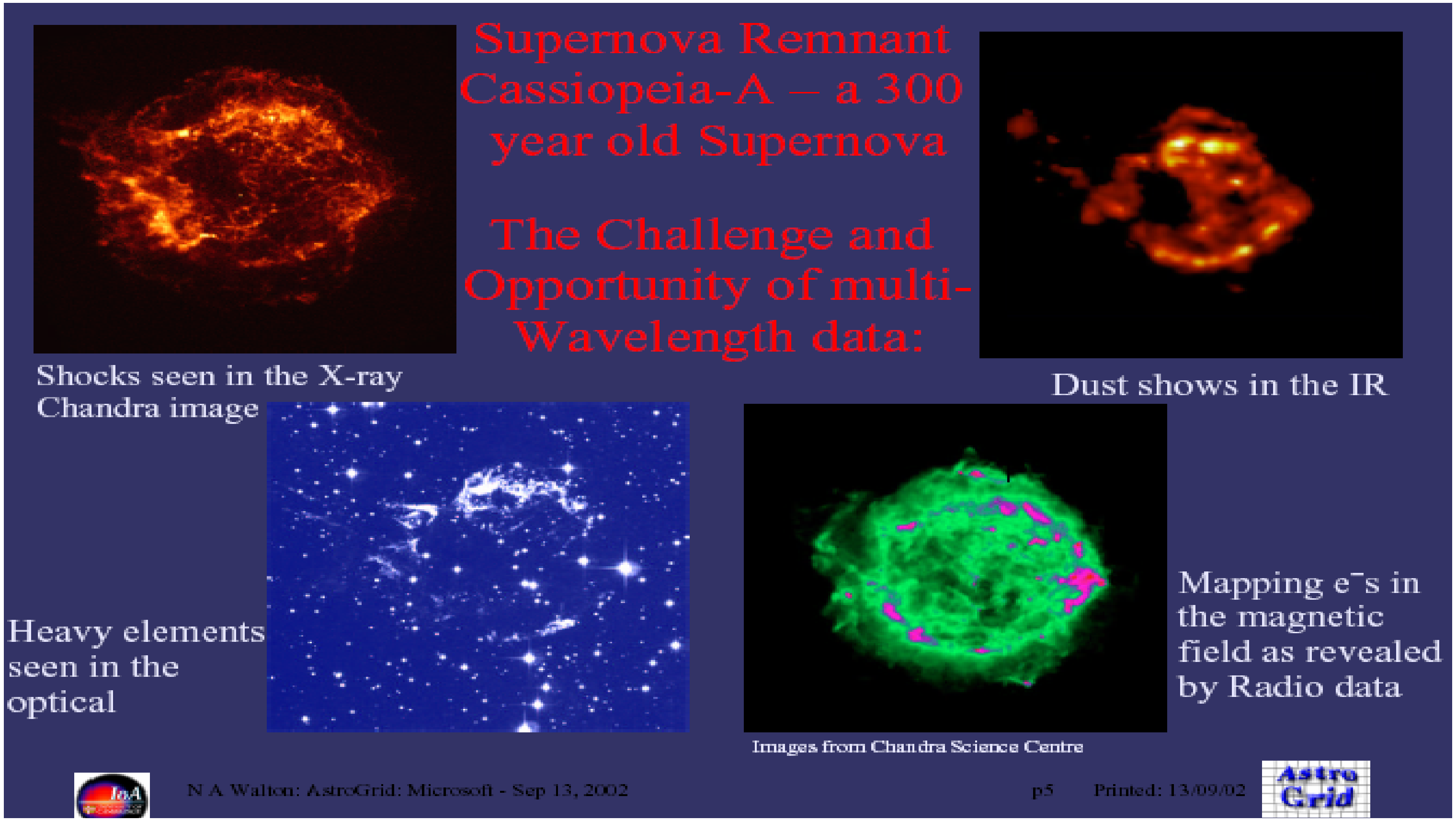
Global in-flight engine diagnostics



100,000 engines
2-5 Gbytes/flight
5 flights/day = 2.5 PB/day

Distributed Aircraft Maintenance Environment: Universities of Leeds, Oxford, Sheffield &York

Virtual Observatories



Supernova Remnant Cassiopeia-A – a 300 year old Supernova

The Challenge and Opportunity of multi-Wavelength data:

Shocks seen in the X-ray Chandra image

Dust shows in the IR

Heavy elements seen in the optical

Mapping e^-s in the magnetic field as revealed by Radio data

Images from Chandra Science Centre

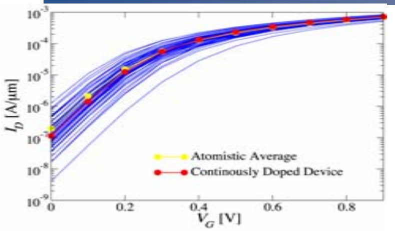
N A Walton: AstroGrid: Microsoft - Sep 13, 2002

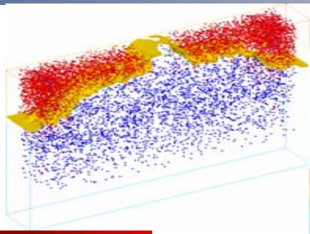
p5

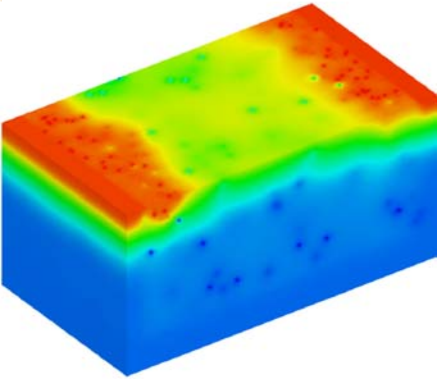
Printed: 13/09/02

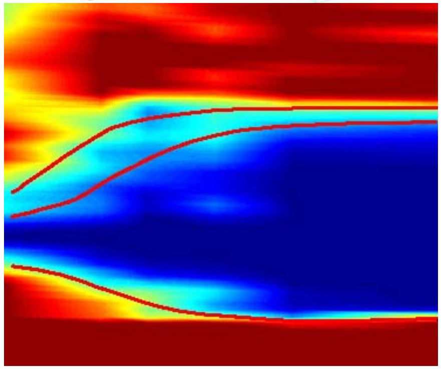
- Huge data sets
  - AstroGrid over 15TB data first week online
- Huge computations
  - Cross referencing data
  - Remove all junk from data sets
    - satellites, aeroplanes…

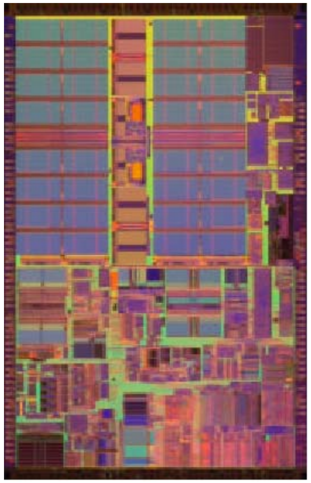
Next Generation Transistor Design







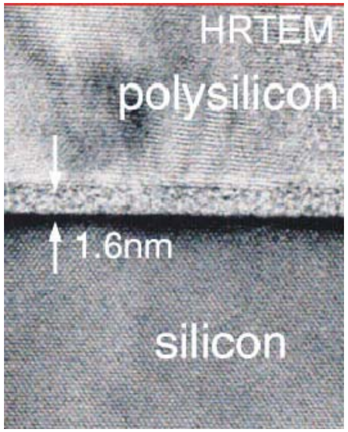


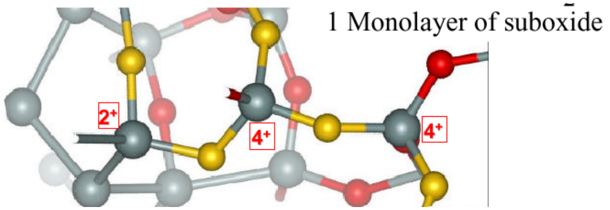


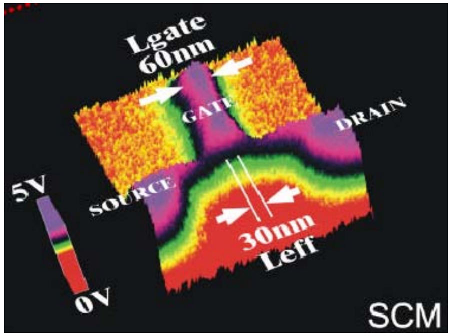
3D

+

Statistical

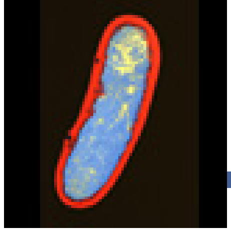




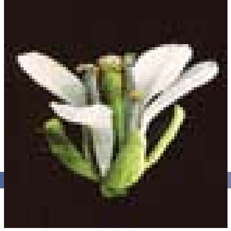


Life Sciences

- Extensive Research Community
  - >1000 per research university
- Extensive Applications
  - Many people care about them
    - Health, Food, Environment
- Interacts with virtually every discipline
  - Physics, Chemistry, Maths/Stats, Nano-engineering, …
- 450+ databases relevant to bioinformatics (and growing!)
  - Heterogeneity, Interdependence, Complexity, Change, …
- Wonderful Scientific Questions
  - How does a cell work?
  - How does a brain work?
  - How does an organism develop?
  - What happens to the biosphere when the earth warms up?
  - Why do people who eat less tend to live longer?
  - …



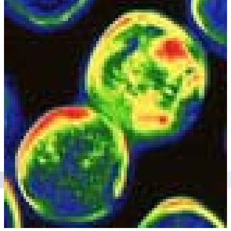
Yersinia
pestis



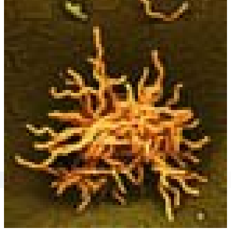
Arabidopsis
thaliana

Buchnerasp.
APS

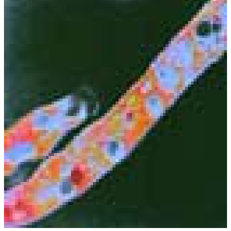
Aquifex
aeolicus



Archaeoglobus
fulgidus

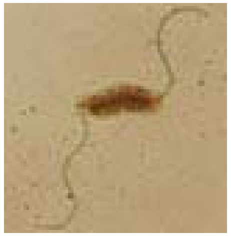


Borrelia
burgorferi

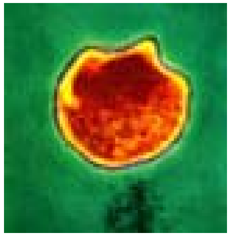


Mycobacterium
tuberculosis

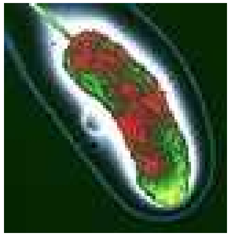
Caenorhabitis
elegans



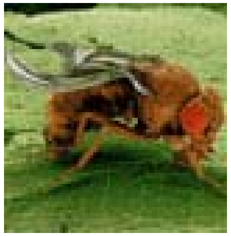
Campylobacter
jejuni



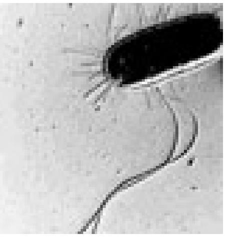
Chlamydia
pneumoniae



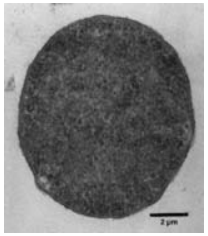
Vibrio
cholerae



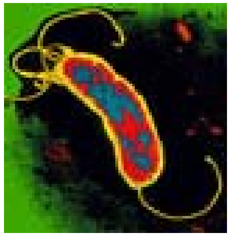
Drosophila
melanogaster



Escherichia
coli

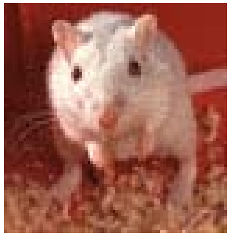


Thermoplasma
acidophilum



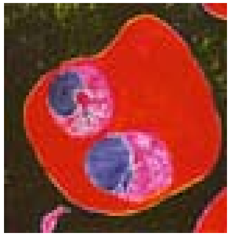
Helicobacter
pylori

Mycobacterium
leprae

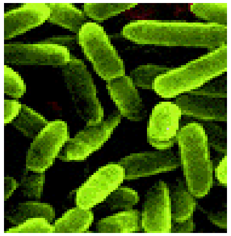


mouse

Neisseria
meningitidis
Z2491



Plasmodium
falciparum

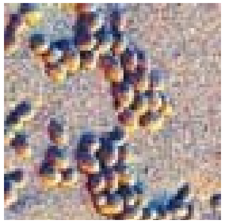


Pseudomonas
aeruginosa

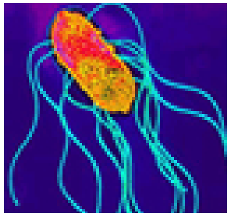
Ureaplasma
urealyticum

rat

Rickettsia
prowazekii

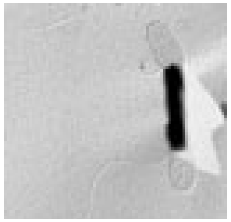


Saccharomyces
cerevisiae

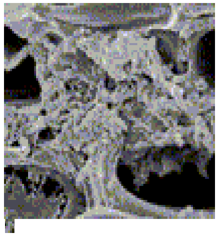


Salmonella
enterica

Bacillus
subtilis



Thermotoga
maritima



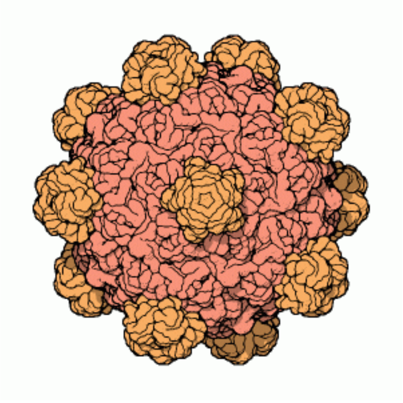
Xylella
fastidiosa

Distributed and Heterogeneous data

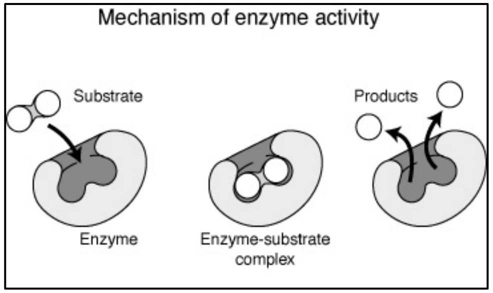
Sequence

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

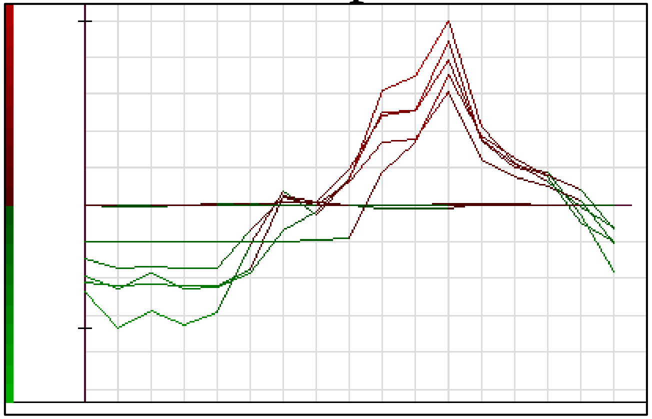
Structure



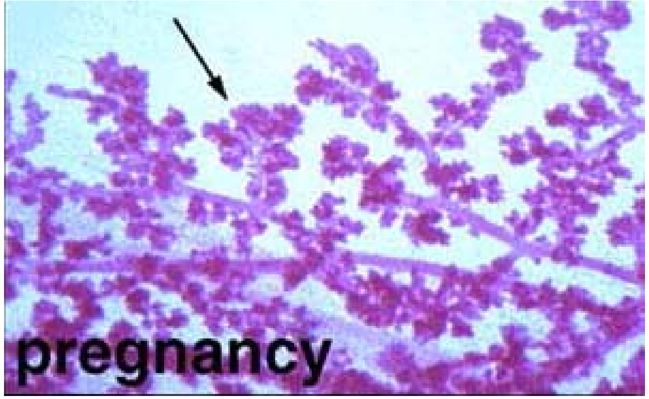
Function



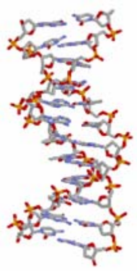
Gene expression

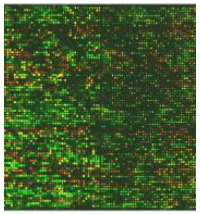


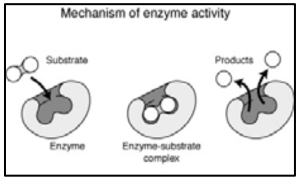
Morphology

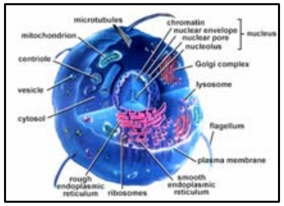


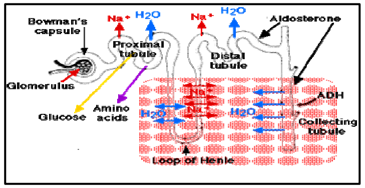
Data Sets associated with Systems-Biology



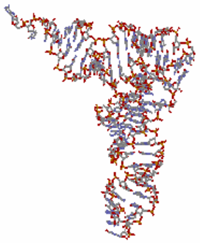


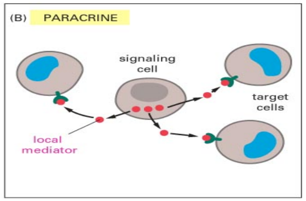


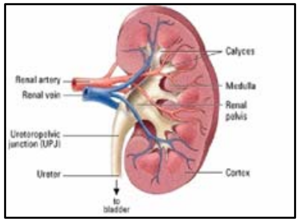


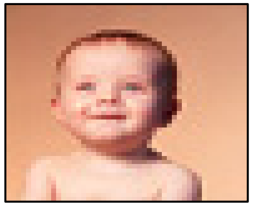


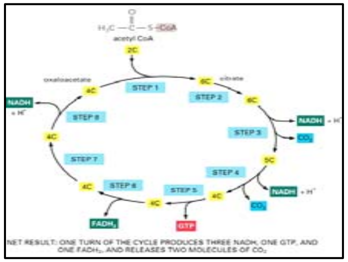
- Nucleotide sequences
- Nucleotide structures
- Gene expressions
- Protein Structures
- Protein functions
- Protein-protein interaction (pathways)
- Cell
- Cell signalling
- Tissues
- Organs
- Physiology
- Organisms
- Populations







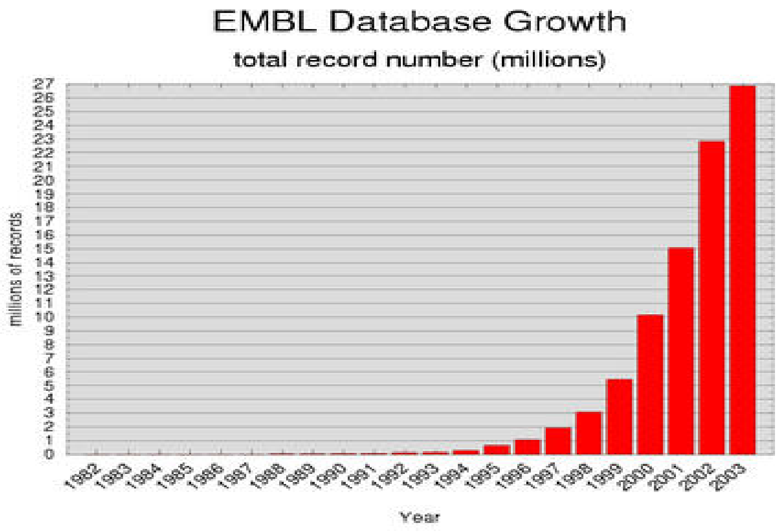




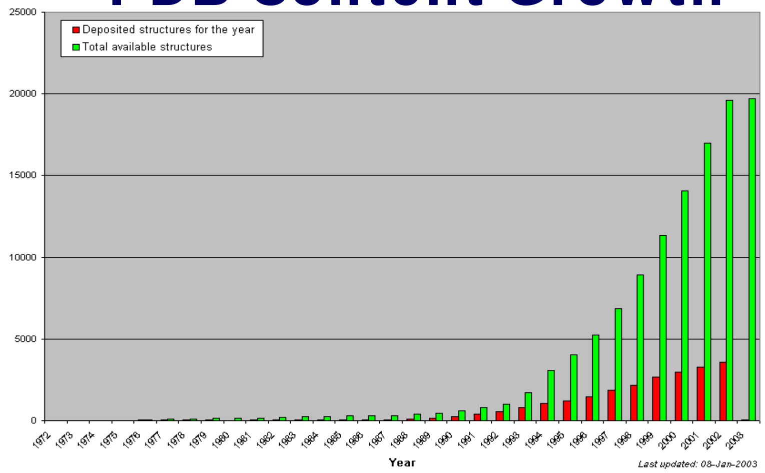
NET RESULT: ONE TURN OF THE CYCLE PRODUCES THREE NADH, ONE GTP, AND ONE FADH₂, AND RELEASES TWO MOLECULES OF CO₂

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

Database Growth



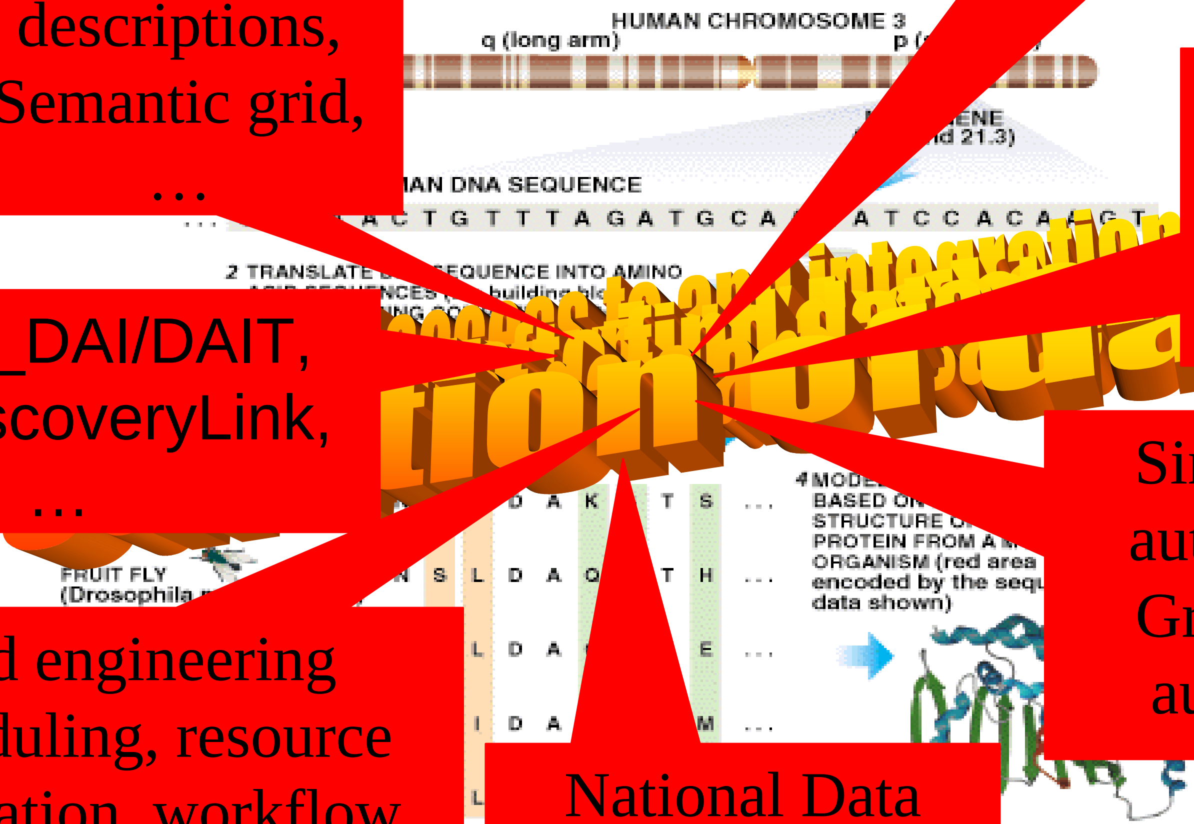
PDB Content Growth



- 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, ...)

Bioinformatics Grid Needs

Workflow / Virtual Organisatio



BioInf community, Database schemas, …

WSDL descriptions, Semantic grid, …

UDDI repositories, BioInf portals, …

OGSA_DAI/DAIT, IBM DiscoveryLink, …

Single sign on authentication, Granularity of authorisation

Grid engineering (scheduling, resource reservation, workflow enactment, …)

National Data curation centre

Goble myGrid presentation

Is Grid the Answer?

- 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] [-I] [-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

Access to and Usage of Data

- Grid technology should allow to
  - hide heterogeneity,
  - deal with location transparency,
  - address security concerns,
  - …

- Data Access and Integration Specification (DAIS) being defined by GGF
  - OGSA-DAI and DAIT projects key role in shaping these standards
- Other commercial solutions (IBM Information Integrator, …)
  - More later!

Access to and Usage of HPC facilities

- Consider whole genome-genome (2*3*10^9 bp) comparisons between two species
  - Current strategy essentially chops up one genome and fires searches for those fragments in the other then re-assembles results
    - messy approximate matching - re-assembly difficult
    - important correlations can be lost
      - to make this tractable so called junk DNA ignored
      - chopping may introduce artefacts or hide phenomena

➢**Better to put both full genomes in memory and perform a useful complete comparison**
➢**Only possible with very high-end machines** (available via grids)

- Should not have to be script writer/Linux sys-admin to use these facilities

Cognitive aspects of Data

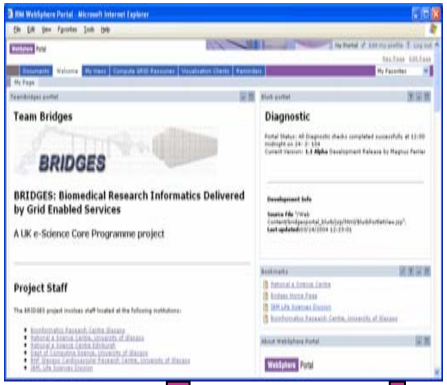
- Life science data can be “ugly”
  - Raw data sets messy
  - Requires significant effort to understand
  - Schemas/data models evolving
  - …

- Tools needed to
  - Simplify understanding
  - Improve analysis
  - Navigate through potentially huge data sets
    - e.g. to find genes of interest in chromosomes of different species

Overview of BRIDGES

- Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES)
  - NeSC (Edinburgh and Glasgow) and IBM
  - 2 year project (£330k) funded by DTI 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

Bridges Project



IBM WebSphere Portal - Microsoft Internet Explorer

File Edit View Favorites Tools Help
WebSphere Portal
My Portal | Edit my profile | Log out
New Page Edit Page
Documents | Welcome | My Views | Compute GRID Resources | Visualization Clients | Reminders
My Favorites
My Page

Team Bridges

BRIDGES

BRIDGES: Biomedical Research Informatics Delivered by Grid Enabled Services

A UK e-Science Core Programme project

Project Staff

The BRIDGES project involves staff located at the following institutions:

- Bioinformatics Research Centre Glasgow
- National e-Science Centre, University of Glasgow
- National e-Science Centre Edinburgh
- Dept of Computing Science, University of Glasgow
- BHF Glasgow Cardiovascular Research Centre, University of Glasgow
- IBM Life Sciences Division

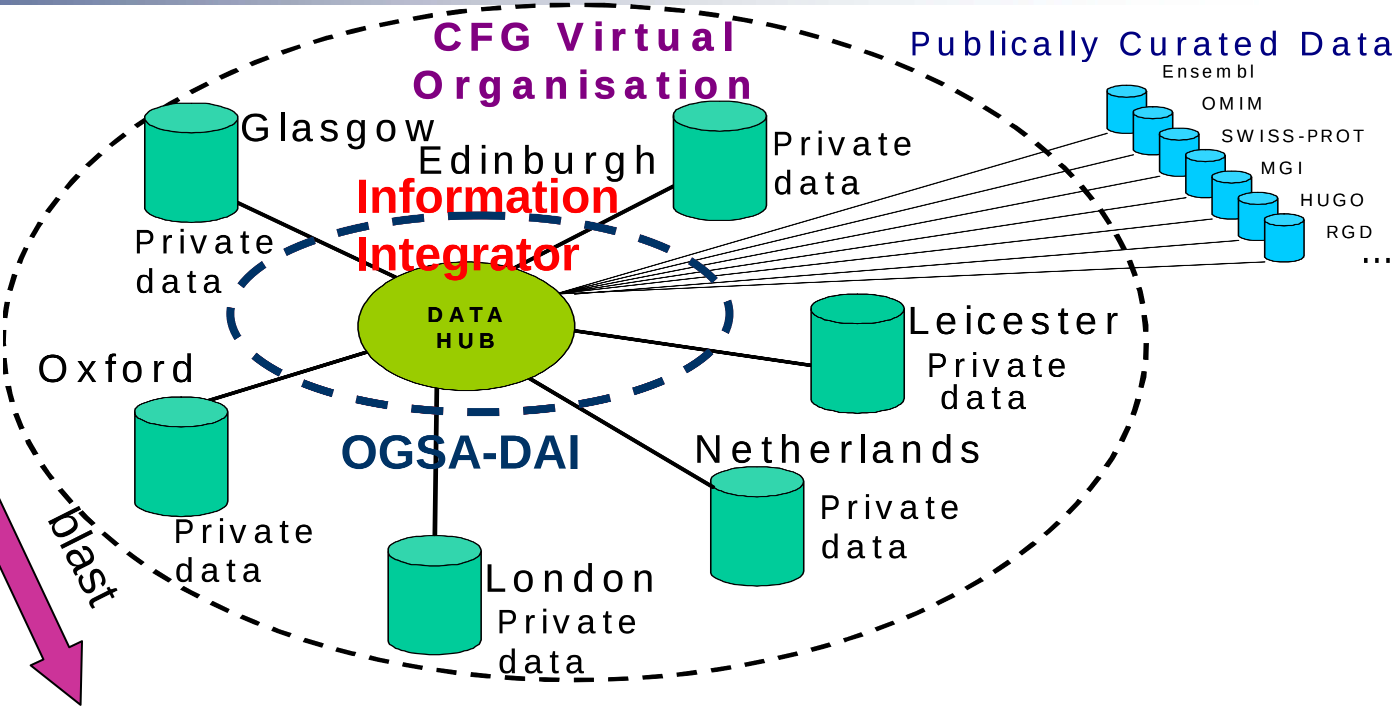
Diagnostic

Portal Status: All Diagnostic checks completed successfully at 12:00 midnight on 24/2/104
Current Version: **1.1 Alpha** Development Release by Magnus Ferrier

Development Info

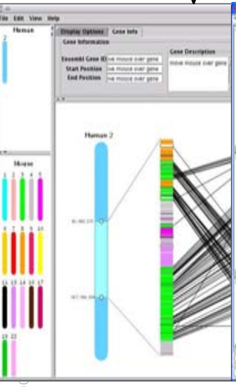
Source File "/Web Content/bridgesportal_blub/jsp/html/blubPortletView.jsp".
Last updated 03/24/2004 12:23:01

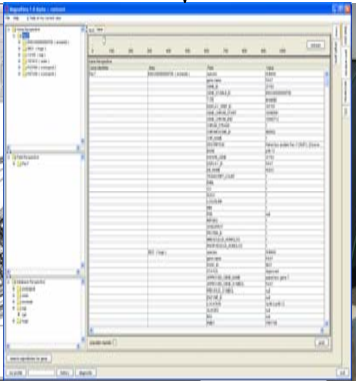
Bookmarks
National e-Science Centre
Bridges Home Page
IBM Life Sciences Division
Bioinformatics Research Centre, University of Glasgow
About WebSphere Portal
WebSphere Portal



Synteny Service

Magna Vista Service





+

Where we are today!

- Information Integrator DB repository established and populated
  - … with public data sets (OMIM, HUGO, RGD, SWISS-PROT)
  - … linked to relevant resources (ENSEMBL- rat, human, mouse, MGI)
- GT3 based Grid services developed (BLAST) using own meta-scheduler
  - General usage of ScotGrid and local Condor pool
- Portal developed using IBM WebSphere
- Genome visualisation browsers
  - SyntenyVista – for viewing synteny between local/remote data sets
  - MagnaVista – for exploring genetic information across multiple (remote) resources
- Gaining experience with security technologies
  - Setting up policies with Grid security authorisation software etc
- <u>Rolled-out Alpha version of system to CFG group July ‘04</u>

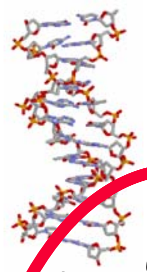
Demo

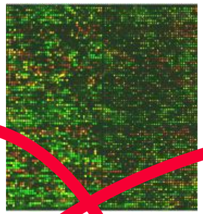
www.nesc.ac.uk/hub/projects/bridges

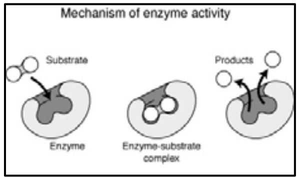
Lessons learned

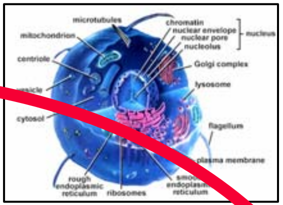
- Public data resources openness
  - Often cannot query directly
  - Often not easy/possible to find schemas
  - Joint Data Standards Study investigating this
    - Started on 1st June and involves
      - Digital Archiving Consultancy
      - Bioinformatics Research Centre (Glasgow)
      - NeSC (Edinburgh and Glasgow)
    - Look at technical, political, social, ethical etc issues involved in accessing and using public life science resources
      - Will liase with NDCC
      - Interview relevant scientists, data curators/providers
    - 8 month project with final report in January
      - Funded by MRC, BBSRC, Wellcome Trust, JISC, NERC, DTI
- GT3 not without pain! (… understatement!!!!)
  - Hopefully GT4 will be better?

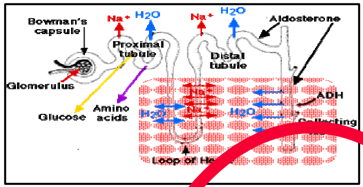
Complexity of Biological Data











Nucl equences

Nucl uctures

Gene ssions

P uctures

P unctions

Protein-protein i tion (pathways)

Cell s ng

Tissues

Organs

Physiology

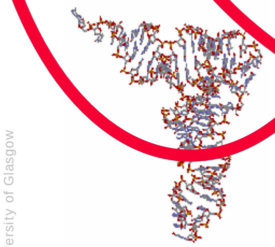
sms

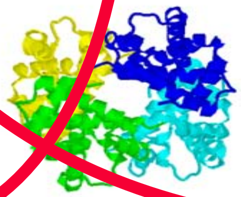
Po ns

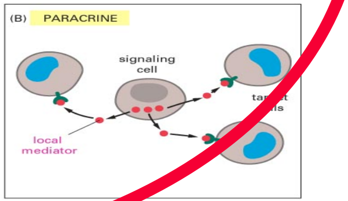
BRIDGES

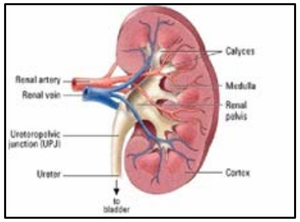
SBRN

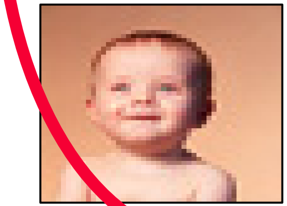
VOTES

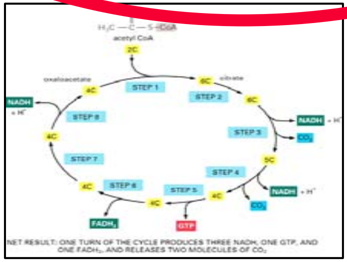








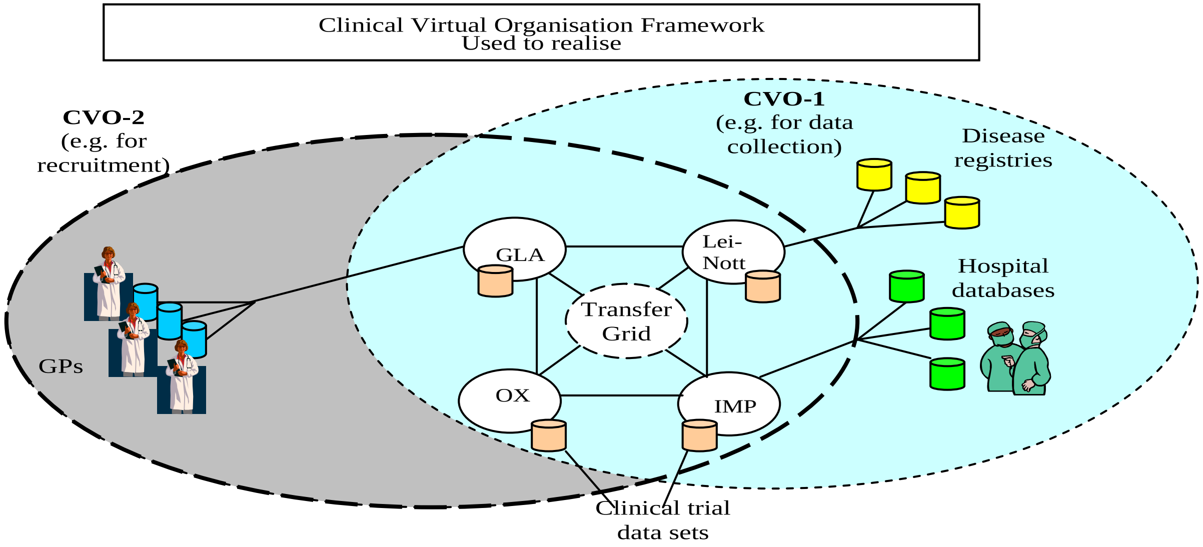




NET RESULT: ONE TURN OF THE CYCLE PRODUCES THREE NADH, ONE GTP, AND ONE FADH2, AND RELEASES TWO MOLECULES OF CO2

VOTES

- Virtual Organisations for Trials and Epidemiological Studies
  - 3 year (£2.8M) MRC funded project expected to start imminently
  - Plans to develop *framework for producing Grid infrastructures* 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, Imperial



Scottish Bioinformatics Research Network

- Four year proposal (£2.5M) expected to start imminently
  - Funded 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

Conclusions

- Numerous application domains exploring e-Science/Grid technologies
- Consolidation of know-how/technologies essential
  - EGEE
  - OMII
  - UK e-Science task forces (ETF, STF, ATF, …)
  - NDCC
  - NeSC
- Do we know how best to build Grids?
  - Different domains coming up with own ways of building Grids
    - OGSA needed asap
  - Clear that various domains have issues which must be resolved before Grid can make significant **and long lasting** impact