e-Science, Archives and the Grid

Tony Hey
Director of UK e-Science
Core Programme
Tony.Hey@epsrc.ac.uk
Outline

• Technology Trends

• What is e-Science?

• What is ‘the’ Grid?

• Present Grid Projects
Increase in MIPS per Chip

MIPS/Chip

Year


MIPS - Millions of instructions per second

*Pentium, 286, 386 and 486 are registered trademarks of Intel Corp.
e-Science

‘e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.’

‘e-Science will change the dynamic of the way science is undertaken.’

John Taylor
Director General of Research Councils
Office of Science and Technology
The vision for the Information Power Grid is to promote a revolution in how NASA addresses large-scale science and engineering problems by providing persistent infrastructure for

- “highly capable” computing and data management services that, on-demand, will locate and co-schedule the multi-Center resources needed to address large-scale and/or widely distributed problems
- the ancillary services that are needed to support the workflow management frameworks that coordinate the processes of distributed science and engineering problems

NASA’s IPG
Multi-disciplinary Simulations (1)

Wing Models
- Lift Capabilities
- Drag Capabilities

Airframe Models
- Braking performance
- Steering capabilities
- Traction
- Dampening capabilities

Crew Capabilities
- Accuracy
- Perception
- Stamina
- Reaction times

Stabilizer Models
- Deflection capabilities
- Responsiveness

Human Models
- SOPs

Engine Models
- Thrust performance
- Reverse Thrust performance
- Responsiveness
- Fuel Consumption

Landing Gear Models

Whole system simulations are produced by coupling all of the sub-system simulations
Multi-disciplinary Simulations (2)

National Air Space Simulation Environment

44,000 Wing Runs
48,000 Human Crew Runs

66,000 Stabilizer Runs

22,000 Commercial US Flights a day

50,000 Engine Runs

22,000 Airframe Impact Runs

132,000 Landing/Take-off Gear Runs

Simulation Drivers

(Flying pulled together under the NASA AvSP Aviation ExtraNet (AEN))

• FAA Ops Data
• Weather Data
• Airline Schedule Data
• Digital Flight Data
• Radar Tracks
• Terrain Data
• Surface Data

Many aircraft, flight paths, airport operations, and the environment are combined to get a virtual national airspace
e-Science Examples

- Bioinformatics/Functional genomics
- Collaborative Engineering
- Medical/Healthcare informatics
- Earth Observation Systems
- TeleMicroscopy
- Virtual Observatories
- Robotic Telescopes
- Particle Physics at the LHC
What is the Grid?

• Computing cycles, Data Storage, Bandwidth and Facilities viewed as commodities as in Electric Power Grid

• Need software and hardware infrastructure to support ‘Grid’ model of ‘Information Utilities’ on demand

• Grid offers uniform access to more than just html pages and information
The GRID Vision

Computing resources

Data

Knowledge

Complex problem

GRID

Solution

Instruments

People
The Challenge of the Grid

- The Grid is an emergent infrastructure capable of delivering dependable, pervasive and uniform access to a set of globally distributed, dynamic and heterogeneous resources.
- Problems of scalability, interoperability, fault tolerance, resource management and security.
- A useful abstraction of the Grid architecture is in terms of a three layered model going from data and computation to information and knowledge.
Data, Information and Knowledge

- **Data**
  Uninterpreted bits and bytes

- **Information**
  Data equipped with meaning

- **Knowledge**
  Information applied to achieve a goal, solve a problem or enact a decision
Three Layer GRID Abstraction

Knowledge Grid

Information Grid

Computation/ Data Grid

Data to Knowledge

Control
US Grid Projects

- NASA Information Power Grid
- DOE Science Grid
- NSF National Virtual Observatory
- NSF GriPhyN
- DOE Particle Physics Data Grid
- NSF Distributed Terascale Facility
- DOE ASCI Grid
- DOE Earth Systems Grid
- DARPA CoABS Grid
- NEESGrid
- DOH BIRN
- NSF iVDGL
EU Grid Projects

- DataGrid (CERN, ..)
- EuroGrid (Unicore)
- DataTag (TTT...)
- Astrophysical Virtual Observatory
- GRIP (Globus/Unicore)
- GRIA (Industrial applications)
- GridLab (Cactus Toolkit)
- CrossGrid (Infrastructure Components)
- EGSO (Solar Physics)
National Grid Projects

- UK e-Science Grid
- Japan – Grid Data Farm, ITBL
- Netherlands – VLAM, PolderGrid
- Germany – UNICORE, Grid proposal
- France – Grid funding approved
- Italy – INFN Grid
- Eire – Grid proposals
- Switzerland - Grid proposal
- Hungary – DemoGrid, Grid proposal
- ApGrid
- ......
UK e-Science Initiative (1)

- £120M 3 Year Programme to create the next generation IT infrastructure to support e-Science and Business
- SR2000 – Funded UK e-Science Grid and Grid Support Centre, e-Science Application research projects and industrial collaboration
- SR2002 – Bidding for additional funding to extend scope of e-Science programme
- Essential that UK plays a leading role in Global Grid development with the USA, EU and Asia
UK e-Science Initiative (2)

- £120M Programme over 3 years
- £75M is for Grid Applications in all areas of science and engineering
- £10M for Supercomputer upgrade
- £35M for development of ‘industrial strength’ Grid middleware

- Require £20M additional ‘matching’ funds from industry
Centres will be Access Grid Nodes

- Access Grid will enable informal and formal group to group collaboration

- It enables:
  - Distributed lectures and seminars
  - Virtual meetings
  - Complex distributed grid demos

- Improves user experience ("sense of presence") - natural interactions (natural audio, big display)
Manchester Access Grid Node
Generic Grid Middleware

- All e-Science Centres will donate resources to form a UK ‘national’ Grid
- All Centres will run same Grid Software
  - Starting point will be based on Globus, Storage Resource Broker and Condor
- Work with Global Grid Forum and major computing companies to move Grid software on towards realizing VO vision
Globus Grid Middleware

- **Single Sign-On**
  - Proxy credentials, GRAM
- **Mapping to local security mechanisms**
  - Kerberos, Unix, GSI
- **Delegation**
  - Restricted proxies
- **Community authorization and policy**
  - Group membership, trust
- **File-based**
  - GridFTP gives high performance FTP integrated with GSI
Storage Resource Broker (1)

- Open Source software developed by Reagan Moore and the DICE group at the San Diego Supercomputer Center
- SRB approach separates organization of distributed digital objects into a collection from their physical storage location
  - Metadata catalog to manage attributes about digital objects
  - Data handling system to manage interaction with remote storage systems
• SRB allows access through federated servers
  - file systems, databases, archival systems
• Collection-based data handling system
• Extensible collection attributes
• Extensible support for access to any type of storage system
• SRB only interim solution – need well-defined Grid middleware interface to Databases
IBM Grid Press Release

Irving Wladawsky-Berger
(Lead for IBM Corporate on Grid)

• ‘Grid computing is a set of research management services that sit on top of the OS to link different systems together’
• ‘We will work with the Globus community to build this layer of software to help share resources’
• ‘All of our systems will be enabled to work with the grid, and all of our middleware will integrate with the software’
EPSRC e-Science Projects (1)

- **Comb-e-Chem**: Structure-Property Mapping
  - Southampton, Bristol

- **DAME**: Distributed Aircraft Maintenance Environment
  - York, Oxford, Sheffield, Leeds

- **Reality Grid**: A Tool for Investigating Condensed Matter and Materials
  - QMW, Manchester, Edinburgh, IC, Loughborough, Oxford
EPSRC e-Science Projects (2)

- **My Grid**: Personalised Extensible Environments for Data Intensive *in silico* Experiments in Biology
  - Manchester, EBI, Southampton, Nottingham, Newcastle, Sheffield
- **GEODISE**: Grid Enabled Optimisation and Design Search for Engineering
  - Southampton, Oxford, Manchester
- **Discovery Net**: High Throughput Sensing Applications
  - Imperial College
Comb-e-Chem: Structure-Property Mapping

- Goal is to integrate structure and property data sources within knowledge environment to find new chemical compounds with desirable properties
- Accumulate, integrate and model extensive range of primary data from combinatorial methods
- Support for provenance and automation including multimedia and metadata
- Southampton, Bristol, Cambridge Crystallographic Data Centre
- Roche Discovery, Pfizer, IBM
MyGrid: An e-Science Workbench

- Goal is to develop ‘workbench’ to support:
  - Experimental process of data accumulation
  - Use of community information
  - Scientific collaboration
- Provide facilities for resource selection, data management and process enactment
- Bioinformatics applications
  - Functional genomics, database annotation
- Manchester, EBI, Newcastle, Nottingham, Sheffield, Southampton
- GSK, AstraZeneca, Merck, IBM, Sun, ...
PPARC e-Science Projects

- **GridPP**
  - links to EU DataGrid, CERN LHC Computing Project, U.S. GriPhyN and PPGrid Projects

- **AstroGrid**
  - links to EU AVO and US NVO projects

- **VISTA**
  - under consideration
Support for e-Science Projects

- ‘Grid Starter Kit’ available from July 2001
- Set up Grid Support Centre
  - International dimensions: EU DataGrid, and US iVDGL projects
- Grid Network Team will identify bottlenecks and elucidate Testbed requirements
- Training Courses and Research Seminars
  - Coordinated by National e-Science Centre
The Grid and Virtual Organisations

[Foster and Kesselman – ‘Take 2’]

The Grid is a software infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources

- includes computational systems and data storage resources and specialized facilities
- enabler for transient ‘virtual organisations’

➢ Must also address access to digital archives