(BiG) Grid: from research to e-Science

Jan Just Keijser, Nikhef
janjust@nikhef.nl

August 30th, 2010
What is the Grid?
“If the customer calls it a grid, it **is** a grid”
• 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’

The Grid vision is to realise this on a global scale
Why would we need a grid?

Enhanced Science needs more and more computations and
Collected data in science and industry grows exponentially

<table>
<thead>
<tr>
<th>Reference: Bible</th>
<th>5 MByte</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray image</td>
<td>5 MByte/image</td>
</tr>
<tr>
<td>Functional MRI</td>
<td>1 GByte/day</td>
</tr>
<tr>
<td>Bio-informatics databases</td>
<td>500 GByte each</td>
</tr>
<tr>
<td>Refereed journal papers</td>
<td>1 TByte/jr</td>
</tr>
<tr>
<td>Satellite world imagery</td>
<td>5 TByte/jr</td>
</tr>
<tr>
<td>US Library of Congress</td>
<td>20 TByte</td>
</tr>
<tr>
<td>Internet Archive 1996-2002</td>
<td>100 TByte</td>
</tr>
<tr>
<td>Particle Physics 2005</td>
<td>1 PByte/jr</td>
</tr>
<tr>
<td><strong>Particle Physics Today: LHC</strong></td>
<td><strong>20 PByte/jr</strong></td>
</tr>
</tbody>
</table>

1 Petabyte = 1 000 000 000 000 Megabyte
From research to e-Science

• The use of computers and ICT in experiments is still increasing
• Some experiments would no longer be possible without computers
• Because ICT is becoming an integral part of the experiment, the experiment itself changes:

  Research becomes e-Science
CERN, Where the web was born ...

- Previous generation of HEP experiments (LEP) involved hundreds of scientists, thousands of engineers, and people working remotely.

- Users at CERN, founded 1954 as Europe’s first international organisation, needed worldwide information sharing.

- This need to share information inspired Tim Berners-Lee to create the ‘World Wide Web’ in 1990.
LHC Computing

Large Hadron Collider

- ‘the worlds largest microscope’
- ‘looking at the fundamental forces of nature’
- 27 km circumference
- CERN, Genève

~ 20 PByte of data per year, ~ 60 000 modern PC style computers
Beyond the Web: Grid for LHC and Science

Work regardless of geographical location, interact with colleagues, share and access data

The GRID: networked data processing centres and “middleware” software as the “glue” of resources (computers, disks, mass storage).

Scientific instruments, libraries and experiments provide huge amounts of data

based on: Federico.Carminati@cern.ch
How does the Grid work?

• It relies on advanced software, called middleware.

• Middleware automatically locates data the scientist needs, and the computing power to analyse it.

• Middleware balances the load on different resources. It also handles security, accounting, monitoring and much more.

*Slide courtesy: GridCafe.org and Bob Jones, EGEE (www.eu-egee.org)*
Grids in e-Science

Grid means 'more than one'

More than one location

More than one computer

More than one Field of science

More than ...
Cross-domain and global e-Science grids

The communities that make up the grid are:

- **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**.
Grid Infrastructure

To bring this about and sustain it requires a persistent infrastructure based on standards

**Hardware infrastructure**
- clusters, supercomputers, databases, mass storage, visualisation, networks

**Trust and AAA infrastructure**
- authentication, authorization, accounting, billing and settlement

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

**Application infrastructure**
- user support, and ICT experts ... with domain knowledge
Nikhef (NDPF)

- 2500 processor cores
- 2000 TByte disk
- 160 Gbps network

SARA (GINA+LISA)

- 4800 processor cores
- 1800 TByte disk
- 2000 TByte tape
- 80 Gbps network

RUG-CIT (Grid)

- 120 processor cores
- 8800 GByte disk
- 10 Gbps network

Philips Research Ehv

- 1600 processor cores
- 100 TByte disk
- 1 Gbps network
What is BiG Grid?

- Collaborative effort of the NBIC, NCF and Nikhef
- Aims to set up a grid infrastructure for scientific research.
- This research infrastructure contains compute clusters, data storage, combined with specific middleware and software to enable research which needs more than just raw computing power or data storage.
- We aim to assist scientists from all backgrounds in exploring and using the opportunities offered by the Dutch e-science grid.

www.biggrid.nl
Opportunities: scaling up

Grid especially means scaling up:
• Distributed computing on many, different computers,
• Distributed storage of data,
• Large amounts of data (Giga-, Tera-, Petabytes)
• Large number of files (millions)

This allows scientists to do research on a scale they could never achieve alone
Virtual Laboratory for e-Science

Data integration for genomics, proteomics, etc. analysis
Timo Breit et al.
Swammerdam Institute of Life Sciences

Avian Alert and FlySafe
Willem Bouten et al.
UvA Institute for Biodiversity
Ecosystem Dynamics, IBED

Medical Imaging/fMRI
Silvia Olabarriaga et al.
AMC and UvA IvI

Bram Koster et al.
LUMC
Microscopic Imaging group

Molecular Cell Biology and 3D Electron Microscopy
Image sources: BiG Grid Consortium Partners

**BiG Grid**

**SCIAMACHY**
Wim Som de Cerff et al.
**KNMI**

**LOFAR:**
LOw Frequency ARray radio telescope

**MPI Nijmegen:**
Psycholinguistics