e-Research Infrastructures for e-Science

Axel Berg
SARA national HPC & e-science support center
RAMIRI, June 15, 2011
Science Park Amsterdam
a world of science in a city of inspiration

- Faculty of Science of the “University of Amsterdam”
- National Institute for Nuclear Physics and High Energy Physics (NIKHEF)
- Institute for Atomic and Molecular Physics (AMOLF)
- National Research Institute for Mathematics and Computer Science (CWI)
- SARA national HPC & e-science support center
- + 80 innovative companies
SARA’s Mission: Support Science & Innovation

SARA Foundation is an independent (hybrid) organization with ~140 fte’s in 2 locations (Amsterdam and Almere)

The mission of SARA is 2-fold:
1. Supporting research in the Netherlands by providing high-end not-for-profit ICT services to research communities [SARA for Science & Innovation]

2. Offering commercial high-end commodity ICT services based on the expertise built in the high-end activities [Vancis for a VANCed Ict Services]
SARA National Supercomputing Center

> SARA supports research in the Netherlands by providing high-end computing-, networking-, storage-, visualization- and e-Science support services and expertise

> These services are guaranteed by:

- deployment of integrated HPC services and infrastructure
- provision of multidisciplinary expertise and support in ICT-technology and applications
- conducting necessary innovation, engineering and development to support and sustain those services
- participation in National and International e-Science and Grid project’s
- connection to and integration into international e-Infrastructures, collaboration
**e-Infrastructures SARA**

**Huygens National Super**
IBM Power 6, 3328 cores, 15.25 TB of memory, 700 TB of disk space, 60 TFlop/s

**LISA National Compute Cluster**
Dell cluster 4480 cores, 12 TB of memory, 20 TFlop/s

**Grid Resources**
2376 Cores, 3408 TB of disk, 2000 TB tape
12 BioInfo Sites Life Science Grid
High Energy Physics, Astronomy, Bio Info

**Visualization**
High Resolution Tiled Panel Display
Remote Visualization

**Innovative Infrastructures**
Cloud
GPU
Hadoop
iRODs

**Network**
SURFnet 6
Netherlight
Connectivity is key

The European Optical Exchange Point @SARA
Advances in Science

Molecular Sciences

Weather & climate prediction

Astronomy

Earth Sciences

Health care

slide from Thom Dunning, NCSA
Advances in Society
Stability of the Atlantic Meridional Overturning Circulation (SAMOC)

- Huygens P6: 750,000 core hours
- Ocean model resolution: 10 km (0.1 degree)
- 1 year model simulation ~ 1 day on about 1200 cores on natl. supercomputer

- Utrecht University, Netherlands
  - Prof. dr. ir. H.A. Dijkstra

- Climate System Modeling Group, Los Alamos Natl. Lab., USA
  - M. Maltrud, M. Hecht, P. Jones, W. Weijer

- Simulation data produced: > 100 TB
Science towards Exascale

- Scientific experiments and data-intensive computing generate today exabytes of data
- Driven by e.g. detector and HW developments
  - Increased resolution, automation & robotization
    - Medical imaging (fMRI): ~1 GByte per measurement
    - Satellite world imagery: ~5 TByte/year
    - Climate modelling: 100 TB per simulation
    - Astronomy LOFAR: >2,5 PByte per year
    - LHC physics: 10-30 PByte per year
- Entering new area of science at Petascale/Exascale level: more than just evolutionary approach will be needed
High Throughput Data Analysis for the Large Hadron Collider (CERN)
SARA & NIKHEF receive, store and analyze 10% of all LHC data
LIFEWATCH: e-science & technology infrastructure for biodiversity data and observatories
Problem Solving Environments: FlySafe: development of bird avoidance models

Slides courtesy of prof. W. Bouten, UvA
HPC is recognized as an important infrastructure (USA, Europe, Japan, India, China)
HPC supports innovation
Out-Compute to Out-Compete™ *
Research communities and research infrastructures

> **ESFRI**: European Strategy Forum on Research Infrastructures

> Examples of research communities/research infrastructures:

> **CLARIN**: Common Language Resources and Technology Infrastructure

> **ENES**: European Network for Earth System Modeling

> **ELIXIR**: European Life Sciences Infrastructure For Biological Information

> **LIFEWATCH**: e-science and technology infrastructure for biodiversity data and observatories

> **wLCG**: High Energy Physics

> **LOFAR/SKA**: astronomy

> **VPH**: Virtual Physiological Human
Dutch national and European e-infrastructures

<table>
<thead>
<tr>
<th>National e-infrastructure</th>
<th>National</th>
<th>International</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid &amp; Cloud Computing</td>
<td>BiG Grid <a href="http://www.biggrid.nl">www.biggrid.nl</a></td>
<td><a href="http://www.egi.eu">www.egi.eu</a></td>
</tr>
<tr>
<td>Network</td>
<td>SURFnet <a href="http://www.surfnet.nl">www.surfnet.nl</a></td>
<td><a href="http://www.geant.net">www.geant.net</a></td>
</tr>
<tr>
<td>Data services</td>
<td>BiG Grid <a href="http://www.biggrid.nl">www.biggrid.nl</a></td>
<td>‘www.eudat.eu’</td>
</tr>
</tbody>
</table>
ESFRI project requirements for Pan-European e-infrastructure resources and facilities
EUdat: Collaborative Data Infrastructure

Data Generators

Users

Data Curation

Community Support Services

Common Data Services

Trust

User functionalities, data capture & transfer, virtual research environments

Data discovery & navigation, workflow generation, annotation, interpretability

Persistent storage, identification, authenticity, workflow execution, mining
Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.

Open Access and Permanent Access

> Open access to scientific information and data

> Significant difference between ‘Open Access to research results’ and ‘Permanent Access to research data’

> Permanent Access can be addressed at a technical, operational and funding level

> Open Access is policy-based and primarily the issue of the data owners

> Service providers can facilitate services that enable policy-based access
Data explosion in science (experimental sciences & data-intensive computing); data is generated centrally and/or distributed

Scientific communities are getting organized and are global, and so are e-infrastructures (PRACE, EGI, GEANT, Eudat)

Research requires a tightly integrated e-infrastructure service that contains all elements (compute, storage, network, visualization, support)

Next PRACE Research Infrastructure as an example