Virtual Laboratory for e-Science

Jelmer Barhorst Pieter de Boer
Wouter Borremans Bart Dorlandt
Aziz Ait Massaoud

24 februari 2005

vl·e virtual laboratory for e·science

Introduction

What's up doc? (1)

Information has become the fuel of our knowledge society and our ability to digest this information, to understand and to share it will determine scientific, economic and social progress.

What's up doc? (2)

Exceptional increase in available resources
(CPU, storage, bandwidth)

⇓

digital revolution

⇓

new research paradigm:
(digitally) enhanced science or e-Science

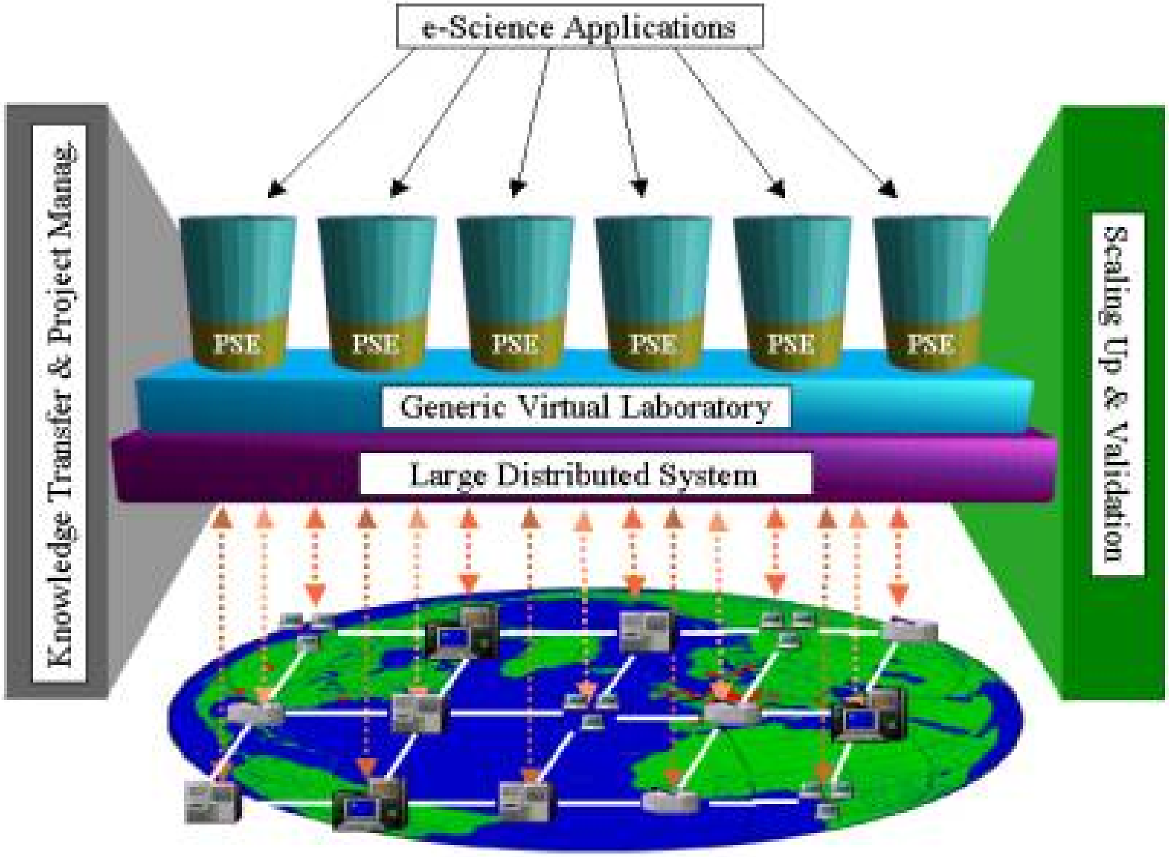
Project aim

Bridge the gap between the technology push of the high performance networking and the Grid and the application pull of a wide range of scientific experimental applications.

Mission

To boost e-Science by creating an e-Science environment and carrying out research on methodologies.

Area’s of research



Applications

Goal

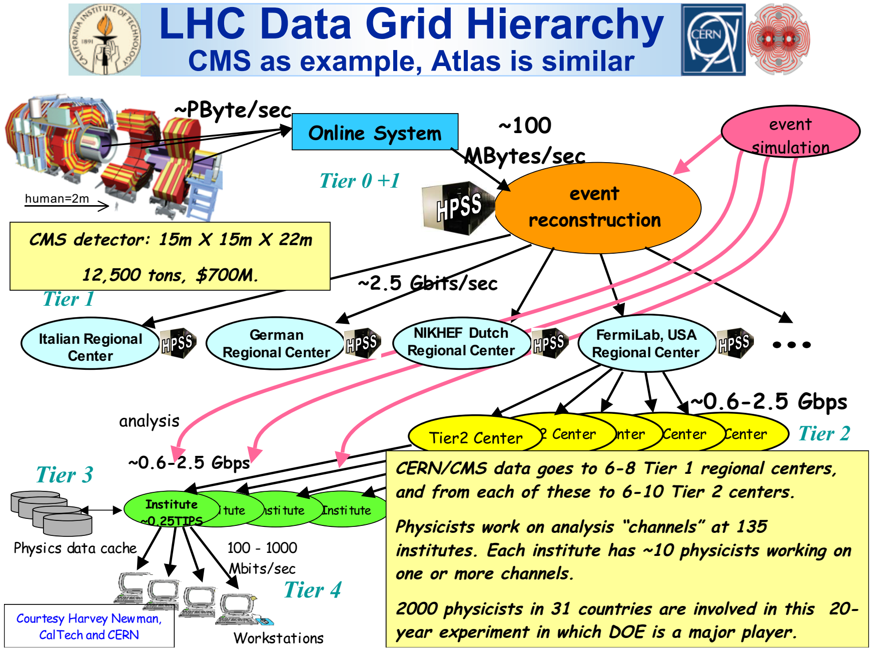
Create several research prototypes of advanced e-Science application specific Problem Solving Environments (PSEs)

Subprogammes

- Data-Intensive Science
- Food Informatics
- Medical Diagnosis & Imaging
- Biodiversity
- Bioinformatics ASP
- The Dutch Telescience Laboratory

Data-Intensive Science

LOFAR



LHC Data Grid Hierarchy

CMS as example, Atlas is similar

CERN/CMS data goes to 6-8 Tier 1 regional centers, and from each of these to 6-10 Tier 2 centers.

Physicists work on analysis "channels" at 135 institutes. Each institute has ~10 physicists working on one or more channels.

2000 physicists in 31 countries are involved in this 20-year experiment in which DOE is a major player.

Courtesy Harvey Newman, CalTech and CERN

CERN

UvA, KNMI, NIKHEF

vl·e virtual laboratory for e·science

Food informatics

- Aims at the design and development of a problem-solving environment for Dutch food research institutes
- Develop efficient information management systems methods for using high-speed network technologies
- Enhance the competitive position

Food informatics

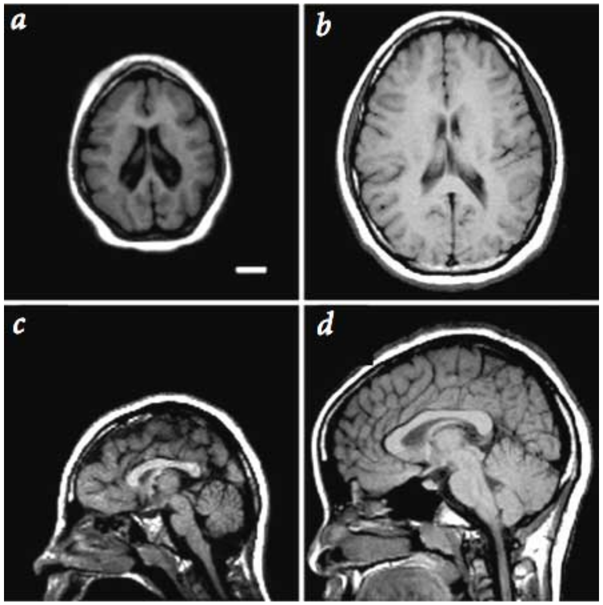
Parties involved:

- Unilever
- FCDF (Friesland Coberco Diary Foods)
- etc.

Recent innovations:

- Complete automation of bread bakery industry

Medical Diagnosis & Imaging



UvA, Philips Research & Medical Systems, AMC, VUmc, TU Delft

Biodiversity

Goals:

- Develop and implement a generic data model for VOFF organisations
- Develop and implement the virtual spatial database EcoGrid
- Develop generic methodologies and tools for scale conversion of data, and data mining of ecological data

- PSE for integrated analysis of observations and model results

Biodiversity

EcoGrid

- Collecting scientifical data about populations
- Generate models
- Predictions

Bioinformatics ASP

Tools for analysing, modelling and integrating experiment data of ‘omics’ known in the domains of life sciences (i.e. genomics ⇒ study of genetic structure of organisms)

Integrated Bioinformatics Unit, UvA & Structure and Functional Organization of the Cell Nucleus, Swammerdam Institute for Life Sciences, UvA

The Dutch Telescience Laboratory

- Better known as DUTELLA (?)
- Flexible problem solving environment for scientific experimentation and collaboration
- Share raw data
- Processing tools
  - Automatically process data / charts

The Dutch Telescience Laboratory

The DUTELLA project comprises three subprojects:

- Biomarker discovery with high resolution LC-FTICRMS
- Molecular imaging
- Combining various types of data about a sample

Generic Virtual Lab Methodology

Generic Virtual Lab Methodology

Research on fundamental knowledge of generic virtual laboratory methodologies for e-Science.

Great... English please?

Subprogammes

So, what does it actually mean?

- Interactive Problem Solving Environments
- Adaptive Information Disclosure
- UI and VR-based Visualization
- Collaborative Information Management
- Virtual Laboratory and System Integration

Interactive Problem Solving Environments

- What is an iPSE?
- What use does it have?

Adaptive Information Disclosure

- What is it?
- Some keywords: semantic models, agent technology, formal concept analysis, datamining, text mining, grid mining, grammar induction, information extraction, question answering

UI and VR-based Visualization

- Visualization using Grids
- Visualization of Grids

Collaborative Information Management

- Design of collaboration tools
- Using the grid as data back-end
- Automagic database schema creation

Virtual Laboratory and System Integration

- Provide for an open forum for research
- Collaborate on software creation and distribution
- Provide feed back to the community

Conclusion

Fascinating stuff!

Large Scale Distributed Systems

Large-Scale Distributed computing

One of the essential components of the total e-Science technology chain

- a Large-Scale Distributed computing development area,consisting of high performance networking and grid parts

Focus

The focus of P3 is fundamental research in the area of large-scale distributed computing systems, based on, high-performance networking and grid technology.

Why needed?

ICT developments

- Processing power doubles every 18 month
- Memory size doubles every 12 month
- Network speed doubles every 9 month
- Something has to be done to harness this development

The programme (1)

In this subprogramme a Java-centric grid programming environment will be build

- Called Ibis
- High-performance applications and a scheduling infrastructure for co-allocation which are easy to use, highly portable ("run anywhere") and robust.

The programme (2)

Such environments are heterogeneous, their resources have different performance and failures (or withdrawal of resources) are more likely to occur than in (small-scale) multi-computers.

Thus, building a programming environment and a co-allocating scheduler with the above properties for such highly dynamic environments is a challenging task

Corresponding key questions

Will it be possible to create a secure and reliable distributed hardware/software infrastructure base that can be used to provide access to grid computing, storage and visualisation resources, instrumentation and information?

Will it be possible to scale the developed methodology and resulting software in the VL-e project to real-life applications?

Properties

In a large scale distributed system:

- Data sources are typically in high numbers, autonomous (under strict local control) and very heterogeneous in size and complexity.
- Data consistency and the performance of data access are crucial.

Research

To address these general problems, we have pursued two complementary research actions.

- Data replication in cluster systems
- Distributed data processing

Data replication in cluster systems

Clusters of PC servers provide a cost-effective alternative to tightly-coupled multiprocessors.

- To improve performance, applications and data can be replicated at different nodes so that users can be served by any of the nodes depending on the current load.

- Successfully used by Web search engines (e.g., Google?).

Challenge

To obtain high-performance and high-availability, databases (and DBMSs) are replicated at several nodes, so they can be accessed in parallel through applications.

Then the main problem is to assure the consistency of autonomous replicated databases.

Distributed data processing

A new dynamic technique

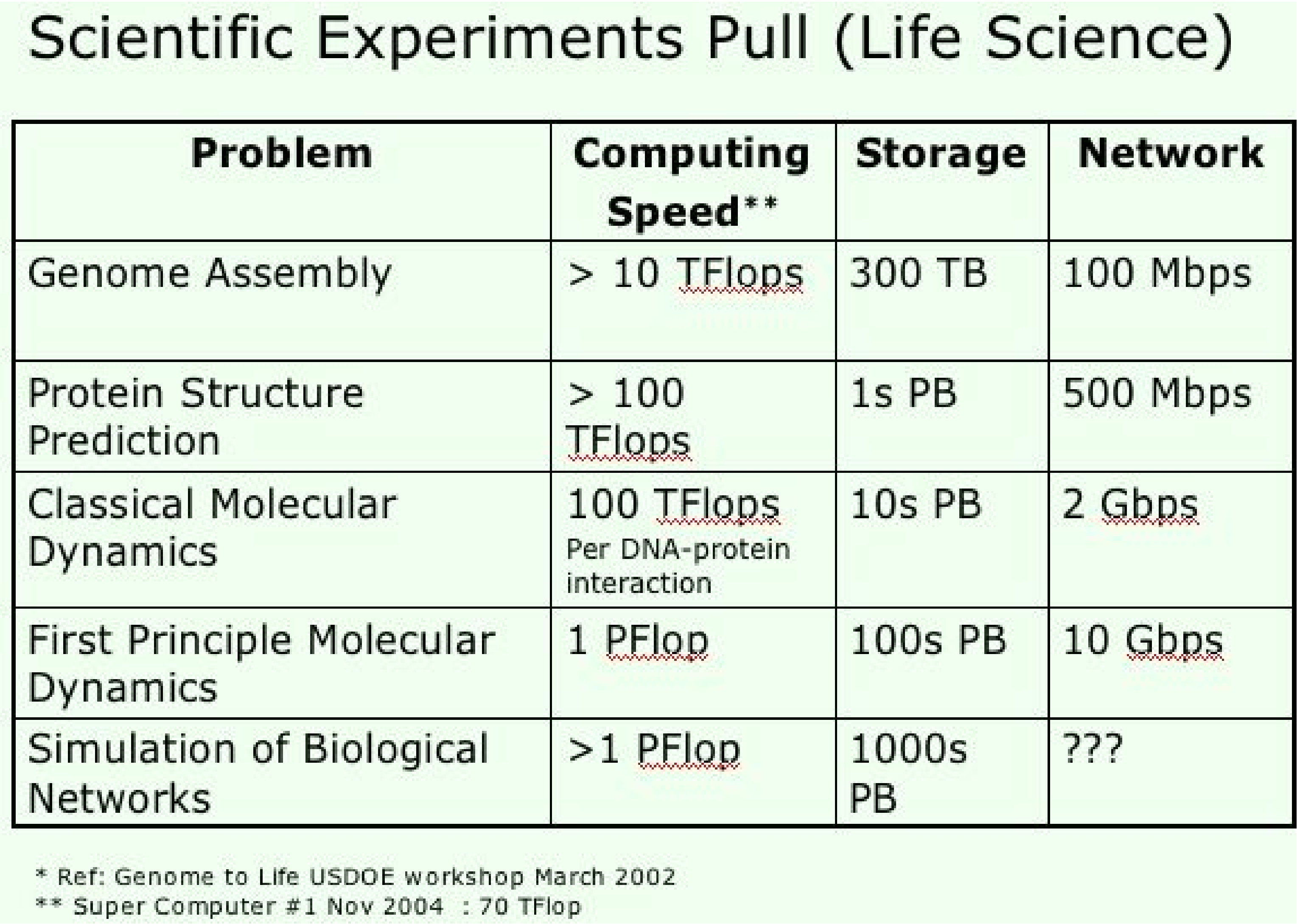
- Optimistic database replication is used with freshness control
- Algorithms are used to evaluate data freshness

Validation

Scaling & Validation

The core of the VL-E scientific methodology is building e-Science problem solving environments (PSEs) and use real-life applications in proof-of-concept environments to validate the research results

Why



Scientific Experiments Pull (Life Science)

Problem	Computing Speed**	Storage	Network
Genome Assembly	> 10 TFlops	300 TB	100 Mbps
Protein Structure Prediction	> 100 TFlops	1s PB	500 Mbps
Classical Molecular Dynamics	100 TFlops Per DNA-protein interaction	10s PB	2 Gbps
First Principle Molecular Dynamics	1 PFlop	100s PB	10 Gbps
Simulation of Biological Networks	>1 PFlop	1000s PB	???

* Ref: Genome to Life USDOE workshop March 2002

** Super Computer #1 Nov 2004 : 70 TFlop

Where does it start?

- Grid software
- VL-e software

Vl-e aims

- creating integrated environments for validating methodology and software
- assembling infrastructures to enable validation and scaling to real-life applications
- creating real-life proof-of-concepts in the diverse application areas to validate environment consistency and reliability

Testing & Expanding. . .

- Storage (NCF)
- Networking (Gigaport, SURFnet)
  - More Complex
  - Security
- Applications
  - Performance
  - Acceptance

Research Levels

- Local
- National
- International

Implementation Plan

- Creation of comprehensive and consistent environments, such that applications and PSEs can be scaled to a real-life size and the methodology can be validated.
- Creation of environments at multiple scales for validating proof-of-concepts.
- Validation - enabling applications to use the environment, and study usage patterns to

extract common requirements on the generic components.

- Investigate and define the parameters that have to be determined to enable a stable, available and reliable environment.

- Besides the abovementioned tasks, knowledge migration is a natural extension of the work to be accomplished in this program, and this program will also give a significant contribution to the knowledge migration center.

Software & Training

- Consistent
- Open Source
- Easy

Conclusion

- 8 FTE
- still one position to be filled @NIKHEF
- Wait and see. . .