Enabling Grids for E-sciencE

Grid Computing: definitions, first use case and requirements

Dr. Rüdiger Berlich,
Forschungszentrum Karlsruhe / Germany
Mumbai, 11.02.06

Some slides contributed by EGEE team, FZK members

www.eu-egee.org

Information Society

Definition, Oracle

From an Oracle talk (September 2005)

"What is Grid Computing ?
In basic terms, grids are clusters of interconnected servers, enabling shared computing resources utilization"

(originally from "Defining Grid Computing", Giga Research, August 2002)

Focus: Clustering ?

Definition, Sun

"[A grid] is a collection of computing resources that perform tasks. It appears to users as a large system, providing a single point of access to powerful distributed resources. Users treat the grid as a single computational resource. Resource management software, such as Sun Grid Engine, accepts jobs submitted by users and schedules them for execution on appropriate systems in the grid based on resource management policies. Users can literally submit thousands of jobs at a time without being concerned where they run." -- Grid Engine 5.3 Administration and User's Guide, p 23

found on http://sysadmin.cs.caltech.edu/docs/help/software/gridengine_help

Focus: distributed computing, but Sun Grid Engine comparable to PBS, LSF ?

Definition, IBM

Grid computing enables the virtualization of distributed computing and data resources such as processing, network bandwidth and storage capacity to create a single system image, granting users and applications seamless access to vast IT capabilities. Just as an Internet user views a unified instance of content via the Web, a grid user essentially sees a single, large virtual computer.

taken from http://www-1.ibm.com/grid/about_grid/what_is.shtml

Focus: Virtualisation, distributed computing, comparison with the Internet and the Web

Definition, HP

"Grid is a software environment based on open standards and protocols that make it possible to share disparate, loosely coupled IT resources across organizations and geographies. IT resources are freed from their physical boundaries and offered as services. They can potentially include almost any IT component -- computer cycles, storage spaces, databases, applications, files, sensors or scientific instruments."

From http://h71028.www7.hp.com/enterprise/cache/125371-0-0-225-121.html

Focus: Wide area, Open Standards, Webservices ...

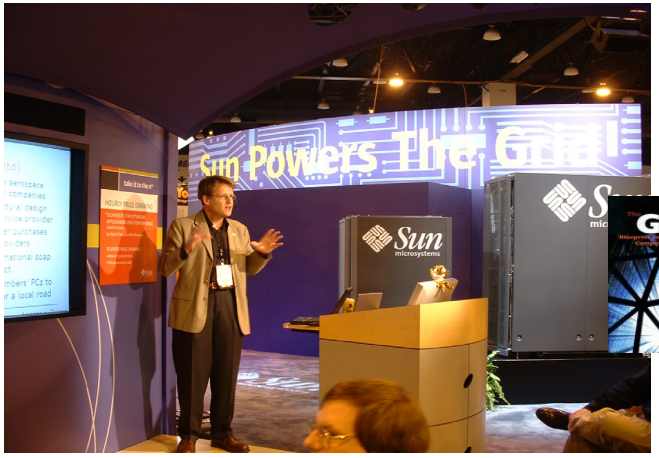
Meaning, Apple

Xgrid

- Recycling of "spare" compute power
- similar to Condor ??
- automated configuration
- plug & play

http://www.apple.com/server/macosx/features/xgrid.html

Definition, Ian Foster



Definition, Ian Foster

"Grid computing is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations" (I.Foster)

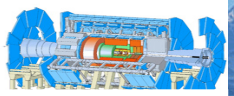
A *Virtual Organisation* is:

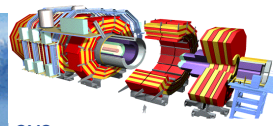
- People from different institutions working to solve a common goal
- Sharing distributed processing and data resources
- Technically not too different from Unix rights management (**access control**)

Focus: Wide area, collaboration, virtual organisations

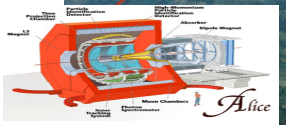
The Large Hadron Collider

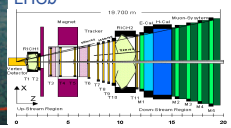
Enabling Grids for E-sciencE


Atlas


CMS


LHC


Alice

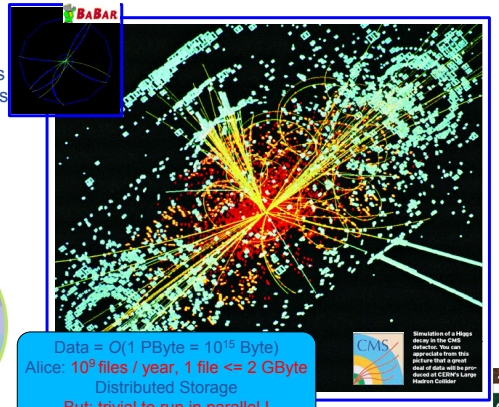

LHCb

LHC / CMS

Enabling Grids for E-sciencE

In LHC:
Expect data rates of
10 - 40 Petabytes
for all experiments
per year.

But: trivial to run in parallel ...

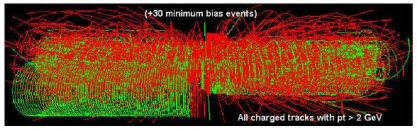


Data = *O*(1 PByte = 10^{15} Byte)
Alice: 10^9 files / year, 1 file <= 2 GByte
Distributed Storage
But: trivial to run in parallel !

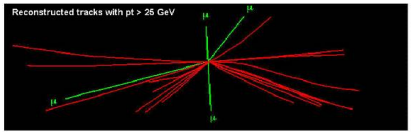
The needle and the haystacks

Enabling Grids for E-sciencE

Starting from this event



Looking for this "signature"

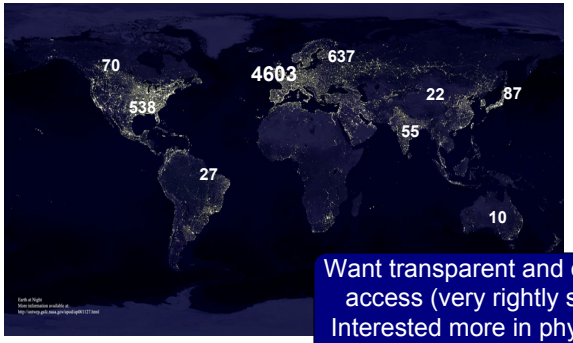


→ **Selectivity: 1 in 10^{13}**
(Like looking for a needle in 20 million haystacks)

LHC

Enabling Grids for E-sciencE

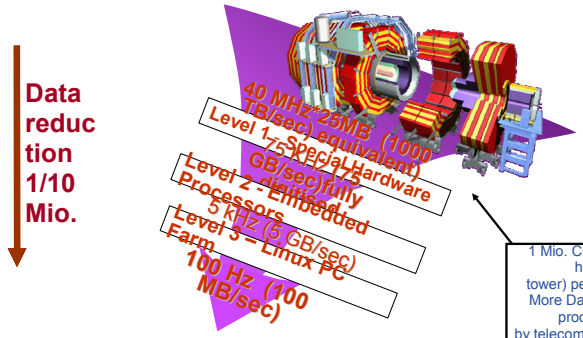
Over 6000 LHC Scientists world wide



Want transparent and quick access (very rightly so). Interested more in physics results, than computing revolutions

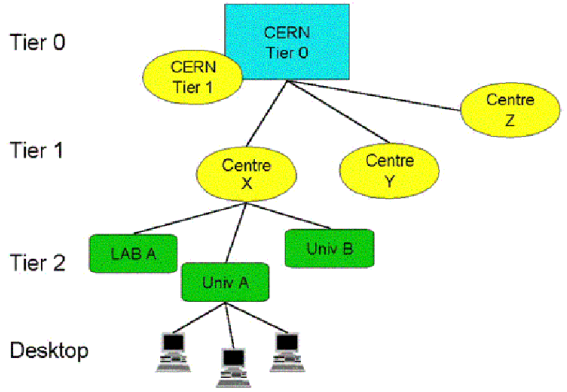
Europe: 267 Institutes, 4603 Users
Other: 208 Institutes, 1632 Users

LHC / CMS



several PB of data per year and experiment
... and 6000 physicist that want to access it !

Tier Centres

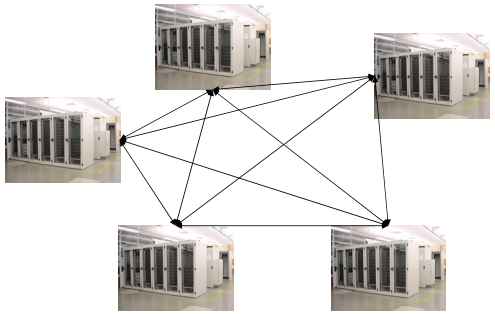


The GridKa Cluster: A Tier-1 centre in LCG

ca. 1500 CPUs
(approx. 4500 in 2007)

Scientific Linux

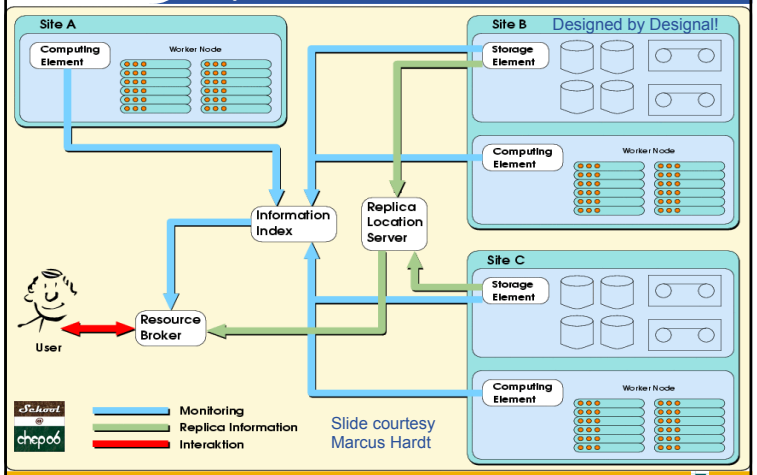
Practical definition in PP



Grid Computing ==
Clustering Clusters; Building a global batch submission system ...

The Grid as a "Cluster of Clusters"

Enabling Grids for E-sciencE



Slide courtesy
Marcus Hardt

Requirements of production quality Grid

Enabling Grids for E-sciencE

Need:

- transparent access to data
  - replication, virtualisation, global filesystems, ...
- secure storage, authentication and authorisation
  - access control (Unix ...), PKI infrastructure, CA, agreed policies, VO
- accounting (computing costs money)
  - not really solved
- training, support
  - GGUS, EGEE
- fast networks (low latency, high bandwidth)
  - Geant, DFN, ...

Need:
(a) software layer "middleware"
(b) fast networks
(c) common policies
(d) services