Grid Challenges and Experience

Heinz Stockinger

Outreach & Education Manager

EU DataGrid project

CERN (European Organization for Nuclear Research)

Grid Technology Workshop,

Islamabad, Pakistan, 20 October 2003

Outline

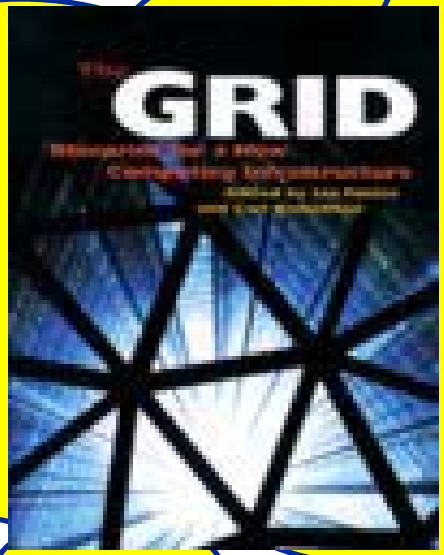
1. What is a Grid?
2. General Grid Technologies
3. Grid Projects – Focus on EU DataGrid Project
4. Experience

Introduction

- Grids are currently under development and show promising results
- Used for solving computing and/or data intensive applications
- Basic Grid concept will be explained in that talk
- What are the technologies that are used to build Grids?
  - Several projects available – we use one major example
- Details on early results and experience

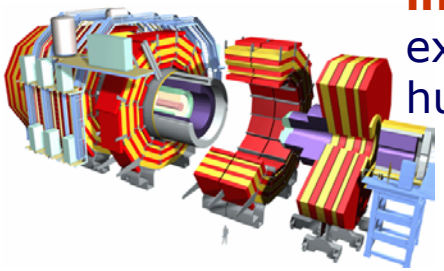
The Grid Vision (1)

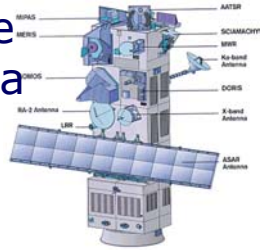
Researchers perform their activities regardless of *geographical location*, interact with colleagues, share and access data



Grid Middleware provides part of the software infrastructure

Scientific instruments and experiments provide huge amount of data





The Grid Vision (2)

- A Grid is:
  - Special form of distributed computing
  - Computing and storage resources are distributed over several locations (sites)
  - Sites are typically connected via wide-area network links
  - Site normally has a local-area network which itself has distributed computing and data storage resources
- Check list given by Ian Foster:

… coordinate resources that are not subject to centralized control …

… using standard, open, general-purpose protocols and interfaces …

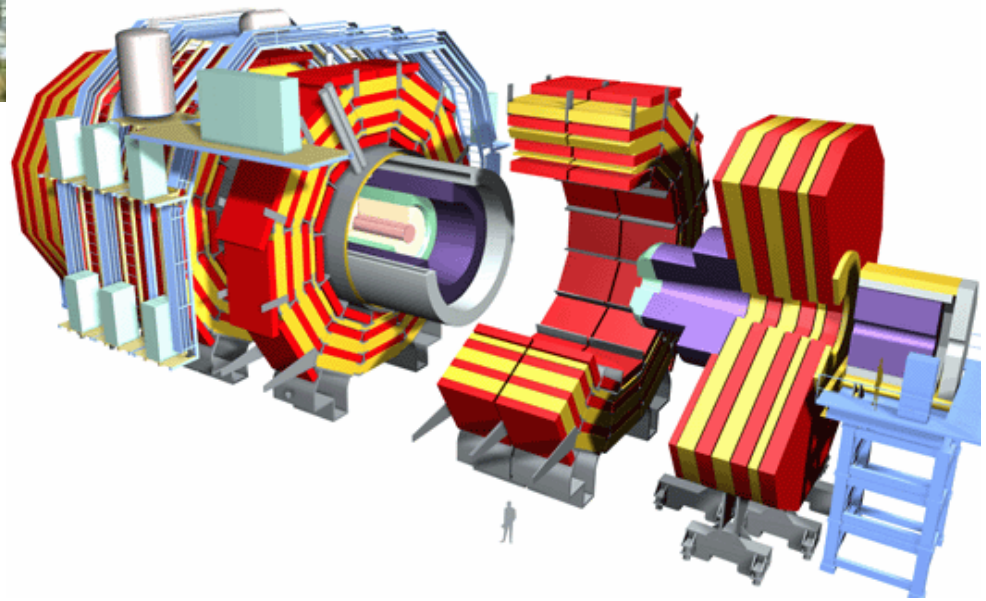
… deliver non-trivial qualities of service …

"Typical" Grids

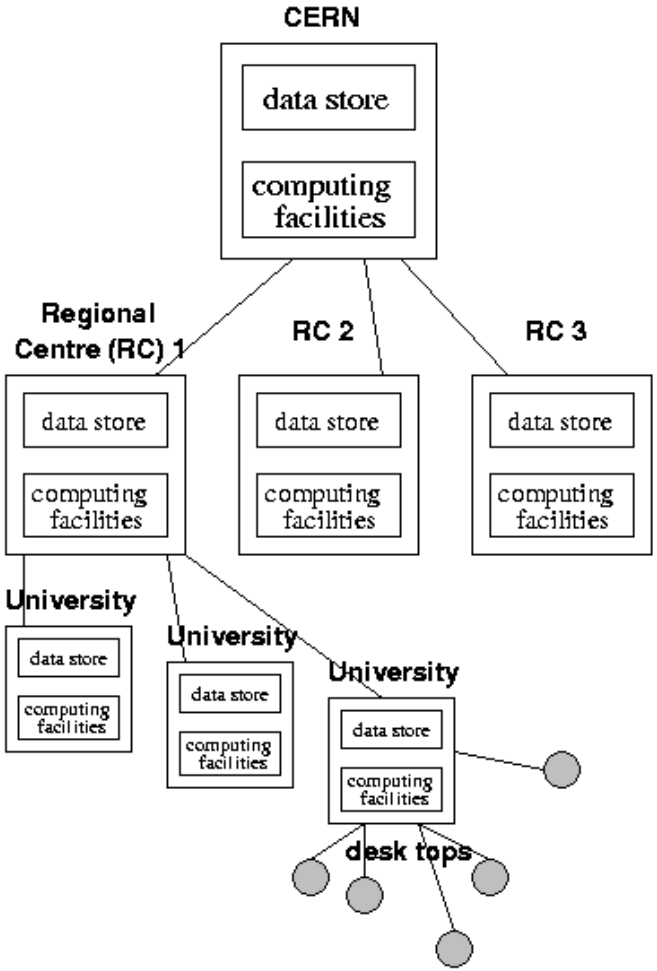
- The Grid vision can be applied best to applications that have the following features:
  - Distributed user community
  - Lots of computing power is required (Computational Grid)
  - Lots of storage capacity is required (Data Grid)
  - Distributed storage locations etc.
- Grids can be applied in academia and industrial environments
- Currently, mainly in computing intensive sciences:
  - High Energy Physics, Earth Observation, Biology, Biomedicine
  - Engineering, Multimedia
- Example Grid:
  - CERN + High Energy Physics application

CERN – European Organization for Nuclear Research

- Over a **Petabyte of data** per year
- Several thousand users



World-wide distributed Regional Centres

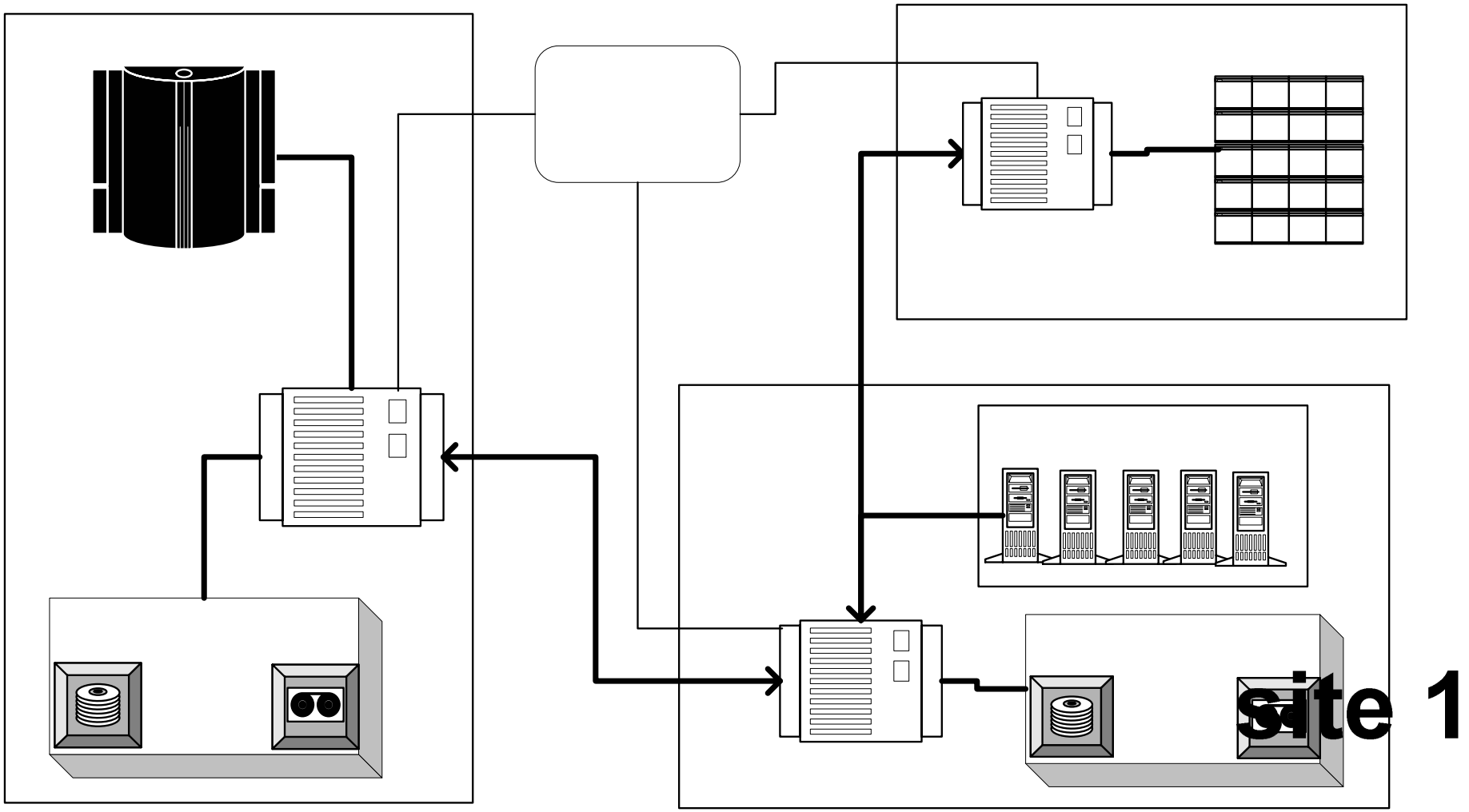


- Part of the distributed computing model
- Complement the functionality of the CERN Centre
- Enable physics analysis all over the world

- **Computational** vs. **Data Grid**
  - we have both components
- European DataGrid project

  http://www.eu-datagrid.org

Data Grid Storage Model



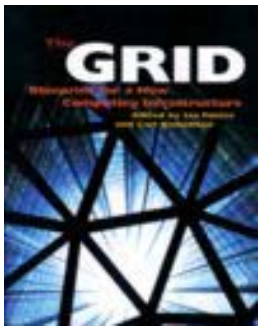
Brief History of Grid Technology (1)

Grid computing has it's roots in classical parallel and distributed computing domain:

- Mainly designed for expensive and specialised parallel computing hardware
- Initially, special purpose interconnects and programming languages were used
- Common standards and programming models were required (standards like MPI and PVM were created)
- Standardisation also paved the way for the more general cluster computing approach
- High performance versus high throughput computing
- Originally, the parallel and distributed community dealt with CPU intensive applications
- Later, applications became more data intensive and several parallel I/O techniques were developed (http://www.cs.dartmouth.edu/pario)

Brief History of Grid Technology (2)

- The nature of distributed resources has a stronger impact on the Grid
- First emergence of Grid computing ideas in many of the early meta computing projects
- HTTP (Hyper Text Transfer Protocol)
  - made possible world-wide information sharing
  - The Web that exploded in the early 1990ies can be considered as one of the direct predecessors of the Grid.
- Building on several of the Internet protocols and ideas from parallel and distributed computing, the first Grid ideas gained world-wide interest around 1998/99
  - HTTP mainly allows for information sharing
  - Grid allows for all kinds of resource sharing
    - Computing and data (information) resources
- Many Grid projects have been created since that time
  - Grid projects all over the world



Note that world-wide distributed applications exist already much longer but the term "Grid" was created around 1998 by Ian Foster and Carl Kesselman

Outline

1. What is a Grid?
2. General Grid Technologies
3. Grid Projects – Focus on EU DataGrid Project
4. Experience

The Grid Today

- Many aspects of the Grid vision have already been or are being realised:
  - Still many steps to go through in order to make the Grid popular to a "conventional" user since
  - Currently considerable expertise is still required in order to make efficient use of Grid technology.
- There is no "single" Grid
  - Several projects, middleware software toolkits etc.
- Several different technologies are used
  - Pointed out in that talk
- Grids need to work together
  - Need for standardization: Global Grid Forum

Grid Standardisation

- Grid development is based on establishing protocols and building services and software development kits.
- Community that seeks for standard protocols (like the Internet community)
  - "Too many projects – too many implementations"
  - Implementations need to "speak" to each other
- Early standardisation process was very much influenced by the Globus project (e.g. GRAM protocol for resource management)
- Global Grid Forum (GGF, http://www.ggf.org)
  - Standardisation of Grid services, protocols and interfaces
  - GGF mission is to focus on the promotion and development of Grid technologies and applications via the development and documentation of "best practices,"
  - Many working groups in several different areas:
    - Applications and Programming Environments, Architecture, Data, Information Systems, Peer-to-Peer, Scheduling and Resource Management, Security

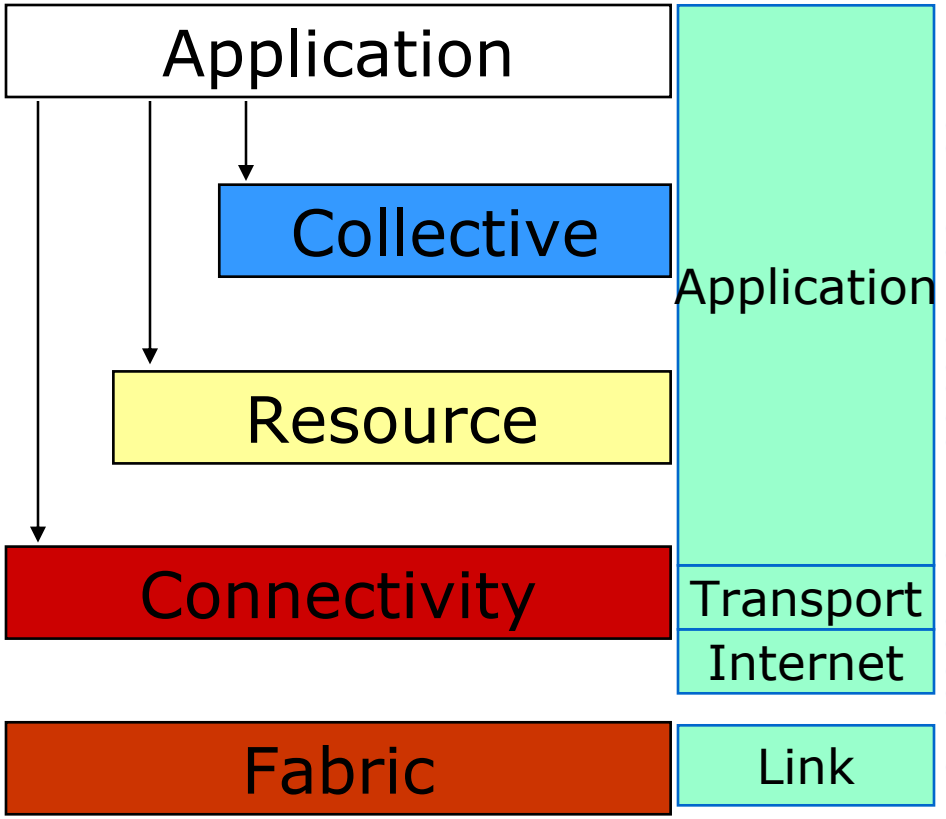
Layered Grid Architecture

"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services

"Sharing single resources": negotiating access, controlling use

"Talking to things": communication (Internet protocols) & security

"Controlling things locally": Access to, & control of, resources



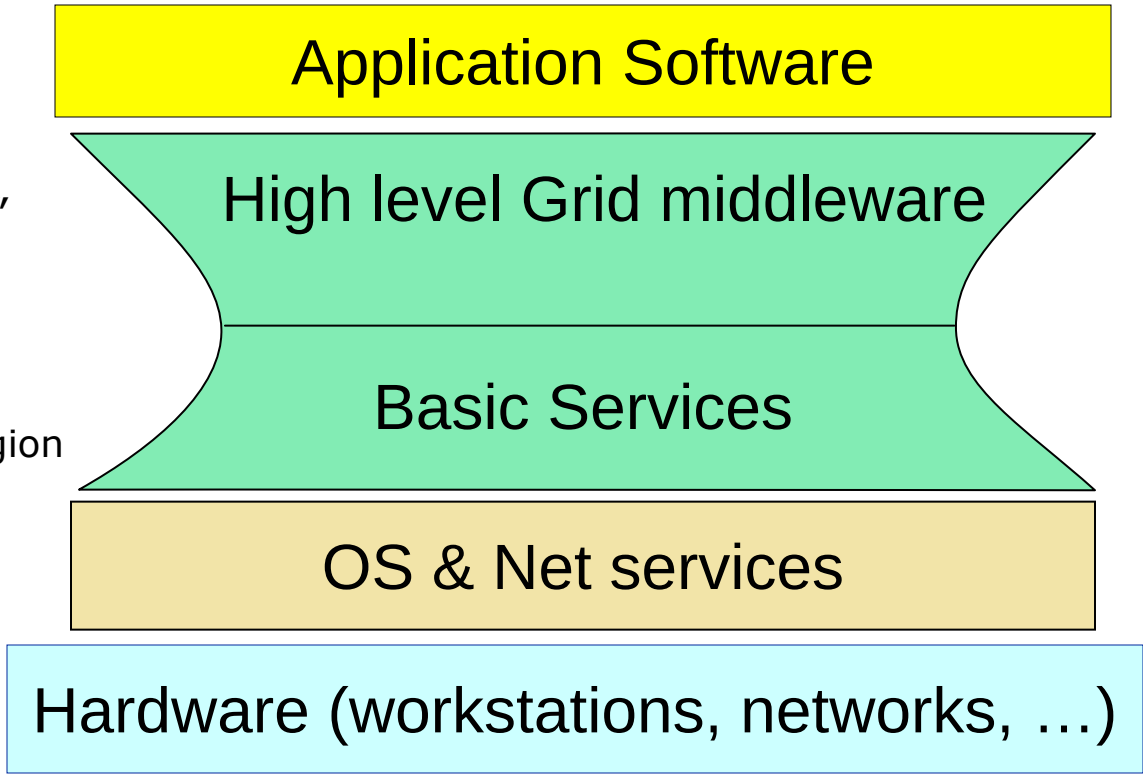
slide by Ian Foster

Main Services of a Grid Architecture

- Service providers
  - Publish the availability of their services via information systems
  - Such services may *come-and-go or change* dynamically
  - e.g. a testbed site that offers *x* CPUs and *y* GB of storage
- Service brokers
  - Register and categorize published services and provide search capabilities
  - e.g. 1) **Resource Broker** selects the best site for a "job"
    2) **Catalogues** of data held at each testbed site
- Service requesters
  - **Single sign-on**: log into the grid once
  - Use brokering services to find a needed service and employ it
  - e.g. physicists submit a simulation job that needs 12 CPUs for 6 hours and 15 GB which gets scheduled, via the Resource Broker, on the CERN testbed site

What is required to "build a Grid"?

- Software
  - Application software …
  - "Higher level middleware"
    - E.g. provided by EDG
  - "Low level" Middleware (toolkit)
    - E.g. Globus, Unicore, Legion etc.
  - Operation System (Unix, Linux)
- Hardware
  - Standards PCs or Unix workstations
  - Wide-area networks
- Focus on middleware and what technologies are used



Grid Technologies? Where to start?

- Big amount of existing projects and Grid tools
  - Need to focus on a small amount of representative technologies
- Globus Toolkit ™ is regarded to be the de-facto standard
  - Builds Grid middleware services
  - Deployed in several Grid projects and testbeds

- European DataGrid projects builds higher level Grid middleware
  - Based on Globus
  - Interactions and collaborations with several major Grid projects

- Here: Main focus on Data Grids and data management

Outline

1. What is a Grid?
2. General Grid Technologies
3. Grid Projects – Focus on EU DataGrid Project
4. Experience

Major existing Grid projects (1/2)

- **Europe-based projects**:
  - **European DataGrid (EDG) : 2001-2003** www.edg.org
  - **LHC Computing GRID (LCG): 2002-2008 -….** cern.ch/lcg
  - **CrossGrid : 2002-2005** www.crossgrid.org
  - **DataTAG : 2002-2003** www.datatag.org
  - **GridLab : 2002-2004** www.gridlab.org
  - **EGEE : 2004-2007 ?** www.cern.ch/egee

European National Projects:

- **INFNGRID, UK-GridPP, NorduGrid(Nordic test bed for wide area computing)…**

Major existing Grid projects (2/2)

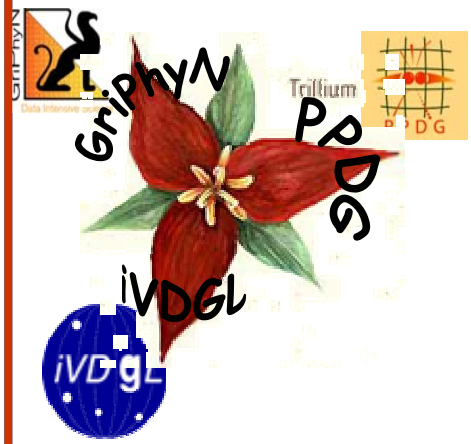
- **US projects:**
  - GriPhyN HEP www.griphyn.org
  - PPDG HEP www.ppdg.net
  - iVDGL (joint GriPhyN, PPDG) www.ivdgl.or
  - TERAGRID (NSF) www.teragrid.org
    - IBM, Intel Qwest ,Myricom, Sun Microsystems, Oracle.
  - National Middleware Initiative (NSF NMI) www.nsf-middleware.org
  - ESG www.earthsystemgrid.org
  - NEESgrid virtual lab earthquake engineering www.neesgrid.org
  - BIRN biomedical informatics research network birn.ncrr.nih.gov/birn/

- **Asia-based projects**:
  - ApGRID www.apgrid.org
  - TWGRID www.twgrid.org
  - Many Grid projects in : Korea, Japan, China, Australia

Major US & European Grid Projects, many with strong HEP participation

the globus project™
www.globus.org

The Virtual Data Toolkit (VDT)



Many national, regional Grid projects - GridPP(UK), INFN-grid(I), NorduGrid, Dutch Grid, …

The DataGrid Toolkit

US projects

European projects

Outline

1. What is a Grid?
2. General Grid Technologies
3. Grid Projects – Focus on EU DataGrid Project
4. Selected Areas + Technologies
   - Security – Information Systems – Data Management
   - Web Service - OGSA

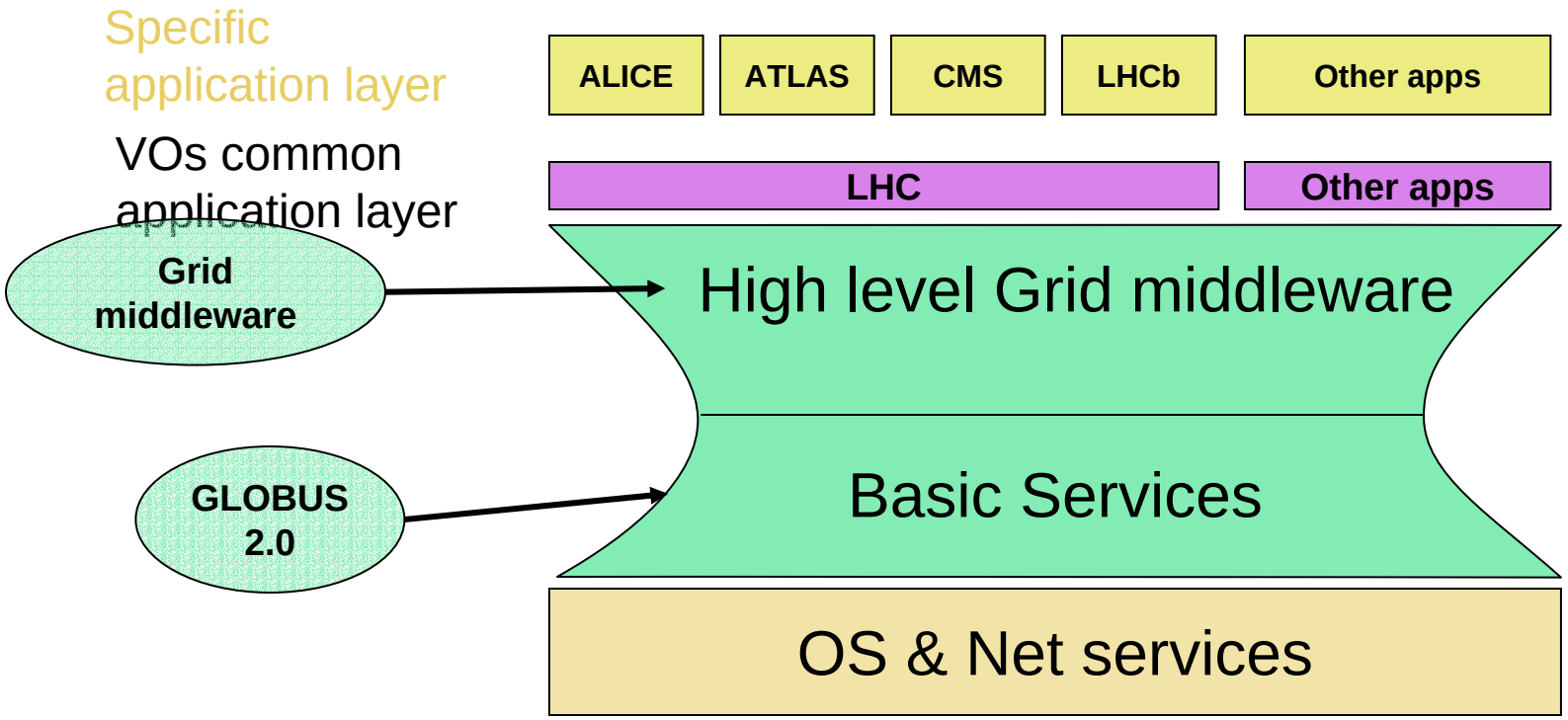
The European Data Grid Project (EDG)

- To build on the emerging Grid technology to develop a sustainable computing model for effective share of computing resources and data
- Start : Jan 1, 2001 End : Dec 31, 2003
- Specific project objectives:
  - Middleware for fabric & Grid management (mostly funded by the EU)
  - Large scale testbed (mostly funded by the partners)
  - Production quality demonstrations (partially funded by the EU)
- To collaborate with and complement other European and US projects
- Contribute to Open Standards and international bodies:
  - Co-founder of Global Grid Forum and host of GGF1 and GGF3
  - Industry and Research Forum for dissemination of project results

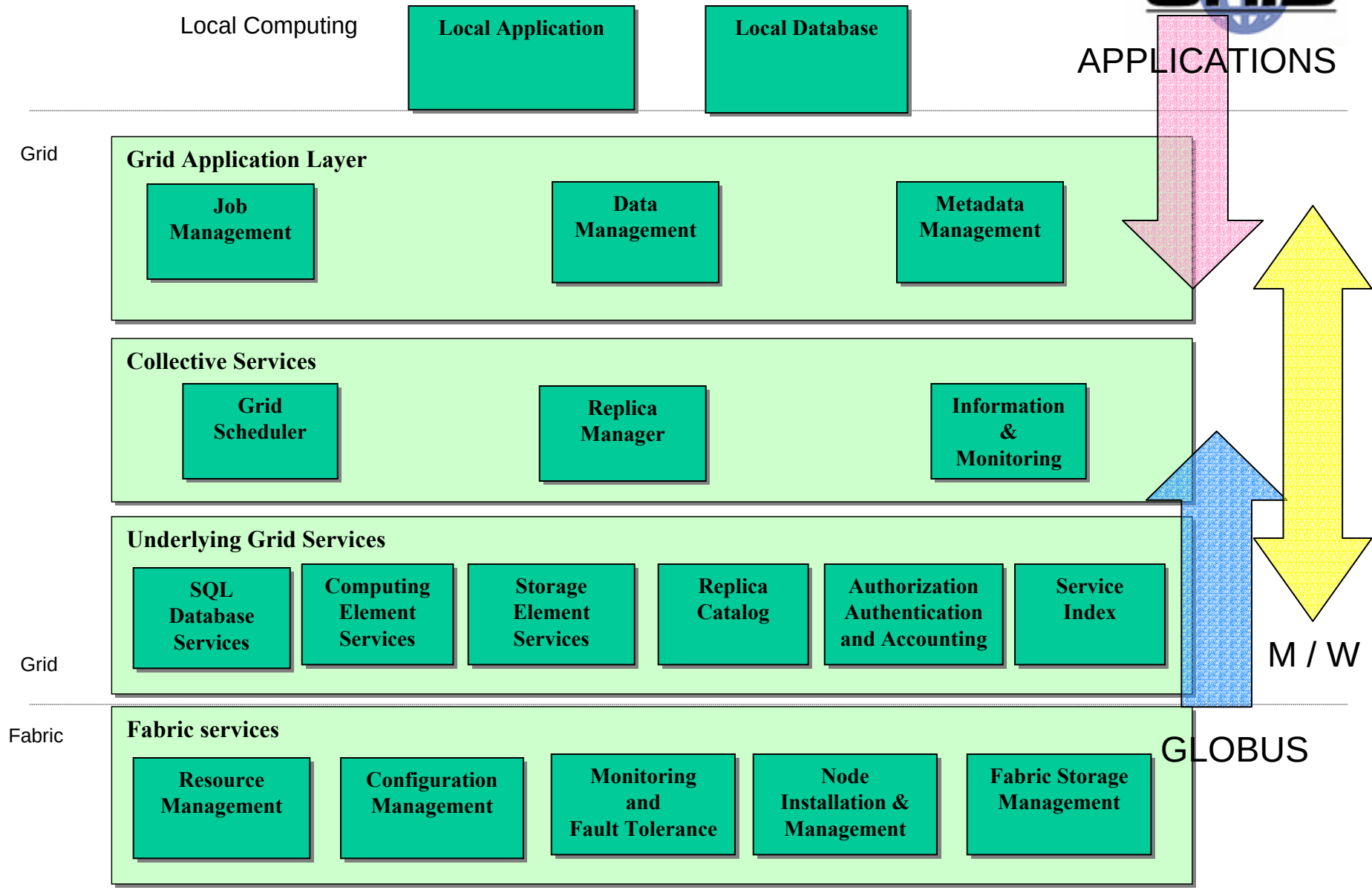
http://www.eu-datagrid.org or http://www.edg.org

EDG Globus-based middleware architecture

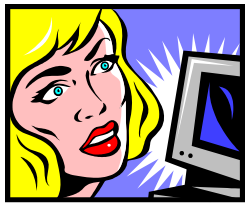
- Current EDG architectural functional blocks:
  - **Basic Services** (authentication, authorization, Replica Catalog , secure file transfer, Info Providers) rely on Globus 2.0
  - **Higher level EDG middleware.**(developed within EDG)
  - **Applications** (HEP,BIO,EO)



EDG middleware Grid architecture

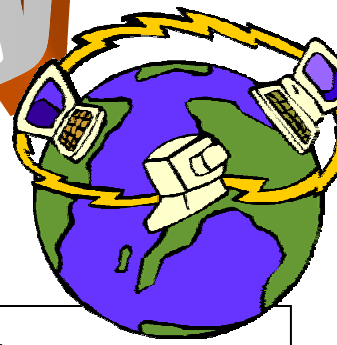


Interaction of Services

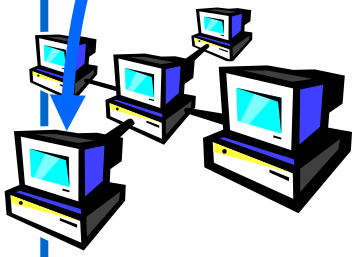


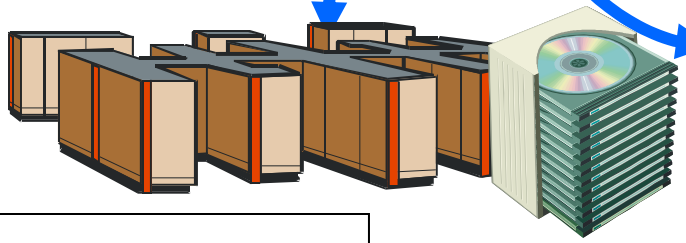
Workload Management Services

Data Management Services

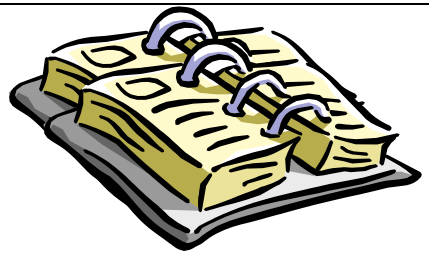


Networking





Information Service



Fabric Management

Current Software Status

- EDG currently provides a set of middleware services
  - Job & Data Management
  - Grid & Network monitoring
  - Security, Authentication & Authorization tools
  - Fabric Management
- EDG release 2.0 currently deployed to the EDG-Testbeds
  - GNU/Linux RedHat 7.3 on Intel PCs
  - Most of release 2.0 is in LCG-1 (except R-GMA and SE)
  - ~15 sites in application testbed actively used by application groups
    - Core sites CERN(CH), RAL(UK), NIKHEF(NL), CNAF(I), CC-Lyon(F)
  - EDG sw (release 1.4) also deployed at total of ~40 sites via CrossGrid, DataTAG and national Grid projects
- Many applications ported to EDG testbeds and actively being used
- Final Release 2.0 with several new technologies is currently finalised and deployed

Outline

1. What is a Grid?
2. General Grid Technologies
3. Grid Projects – Focus on EU DataGrid Project
4. Experience

Experience (1)

- CERN has been involved in Grid activities since early 2000
- Originally, very much **pioneer effort**:
  - Hardly any Globus knowledge in Europe at this time
    - Globus was still rather new and relatively immature
    - Early co-operation and collaboration with the Globus Alliance and the Codor team
- INFN provided first binary Globus distribution
  - Was used in early pilot projects and also influenced Globus' software distribution effort
- Early Grid project to test and evaluate Globus
  - GDMP (Grid Data Mirroring Package): first Grid tool that was used in production in the CMS experiment

Experience (2)

- Early projects helped to find bottlenecks, design and scalability problems in Grid middleware
  - Lots of issues have been sorted out in the meantime
  - Lots of input also from the EU DataGrid user community
- CERN and HEP community have adopted the Grid computing model for the physics data challenges
- Many new collaborations and projects have been established
  - EU DataGrid, PPDG and GriPhyN were among the very first ones
  - Currently, about 20 Grid projects funded by the European Union
- Grid Today
  - many aspects of the Grid vision are being realised
  - still many steps make the Grid popular to a "conventional" user
  - Grid is not yet "finalised"

Conclusion

- Grids Technologies become more and more popular for CPU or data intensive applications
- Several projects and technologies are available and standardization effort tries to make them interoperable
- Further information
  - EU DataGrid Project: www.eu-datagrid.org
  - LCG Project: www.cern.ch/lcg
  - Global Grid Forum: www.ggf.org
  - European Grid projects: www.gridstart.org
  - Globus Alliance: www.globus.org

Thanks to the EU for the support of this work