Grid Computing: European Data Grid Project

Ben Segal
Les Robertson
Fabrizio Gagliardi
Federico Carminati

Presented By Arie van Praag

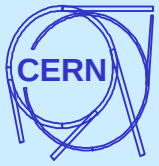
CERN

Information Technology Division

October, 2000

b.segal@cern.ch, a.van.praag@cern.ch

Originally IEEE Oct. 2000

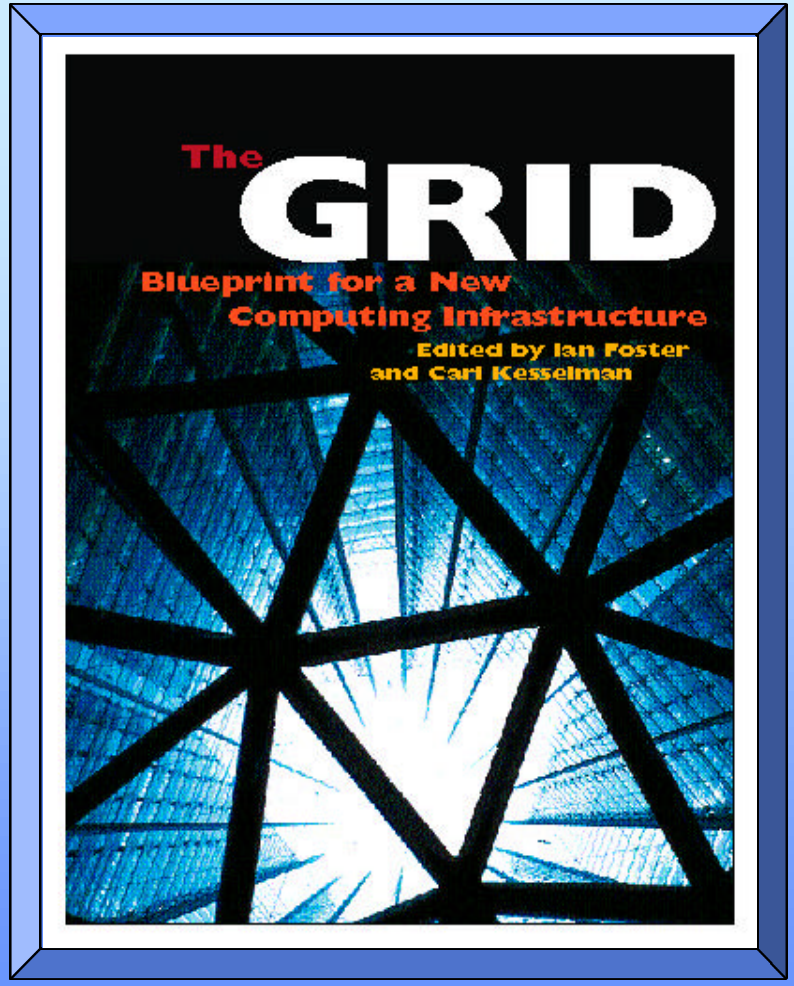


The GRID metaphor

- Analogy with the electrical power GRID
- Unlimited ubiquitous distributed computing
- Transparent access to multi peta byte distributed data bases
- Easy to plug in
- Hidden complexity of the infrastructure

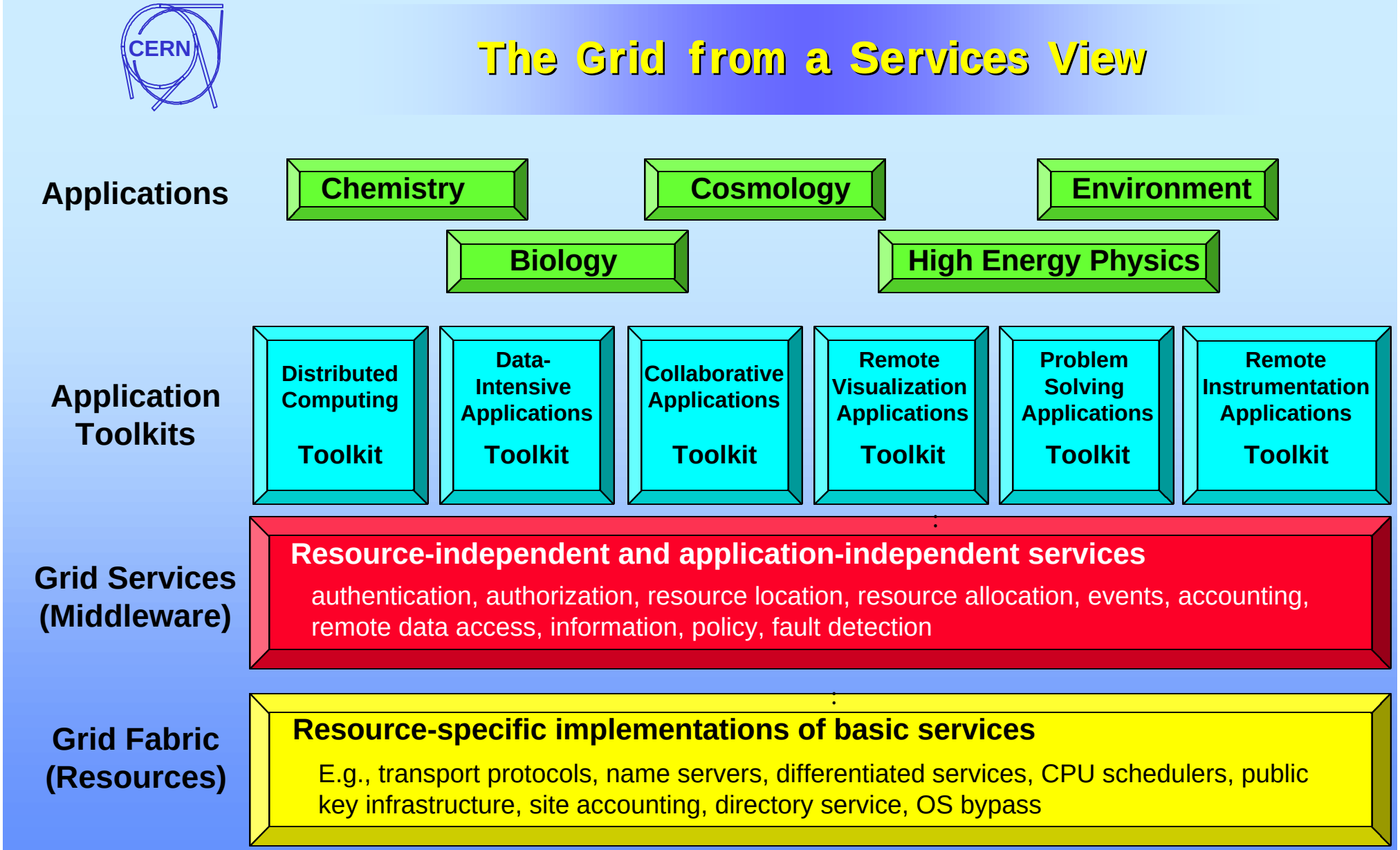
Five Emerging Models of Networked Computing From *The Grid*

- **Distributed Computing**
  - || synchronous processing
- **High-Throughput Computing**
  - || asynchronous processing
- **On-Demand Computing**
  - || dynamic resources
- **Data-Intensive Computing**
  - || databases
- **Collaborative Computing**
  - || scientists

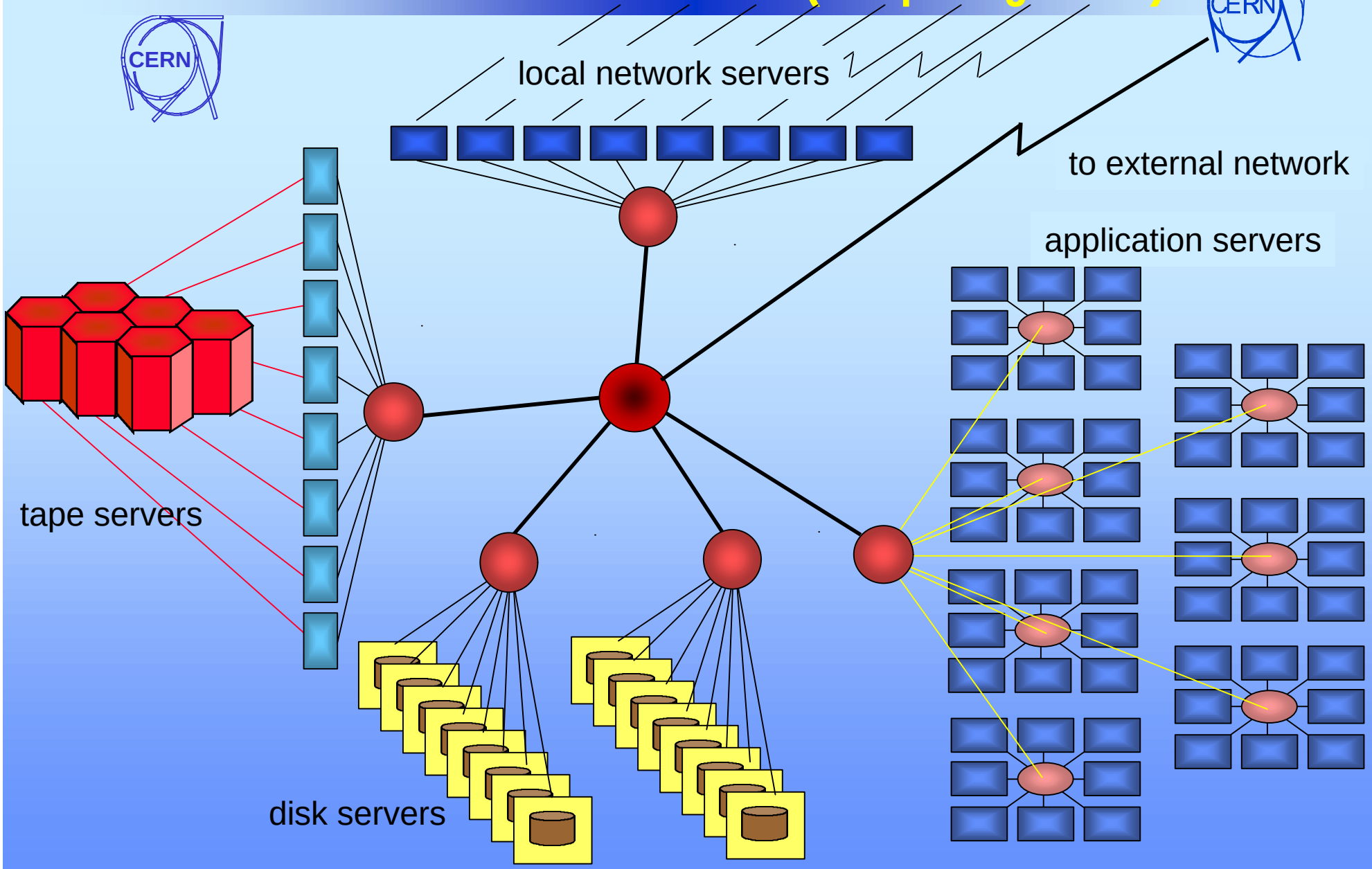


Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999, http://www.mkp.com/grids

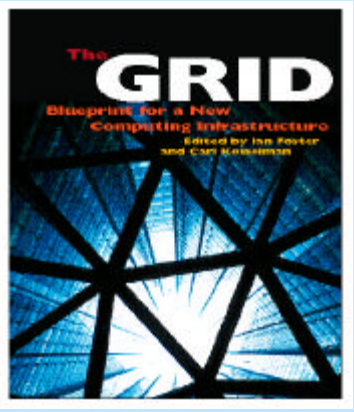
The Grid from a Services View



Generic model of a Fabric (computing farm)



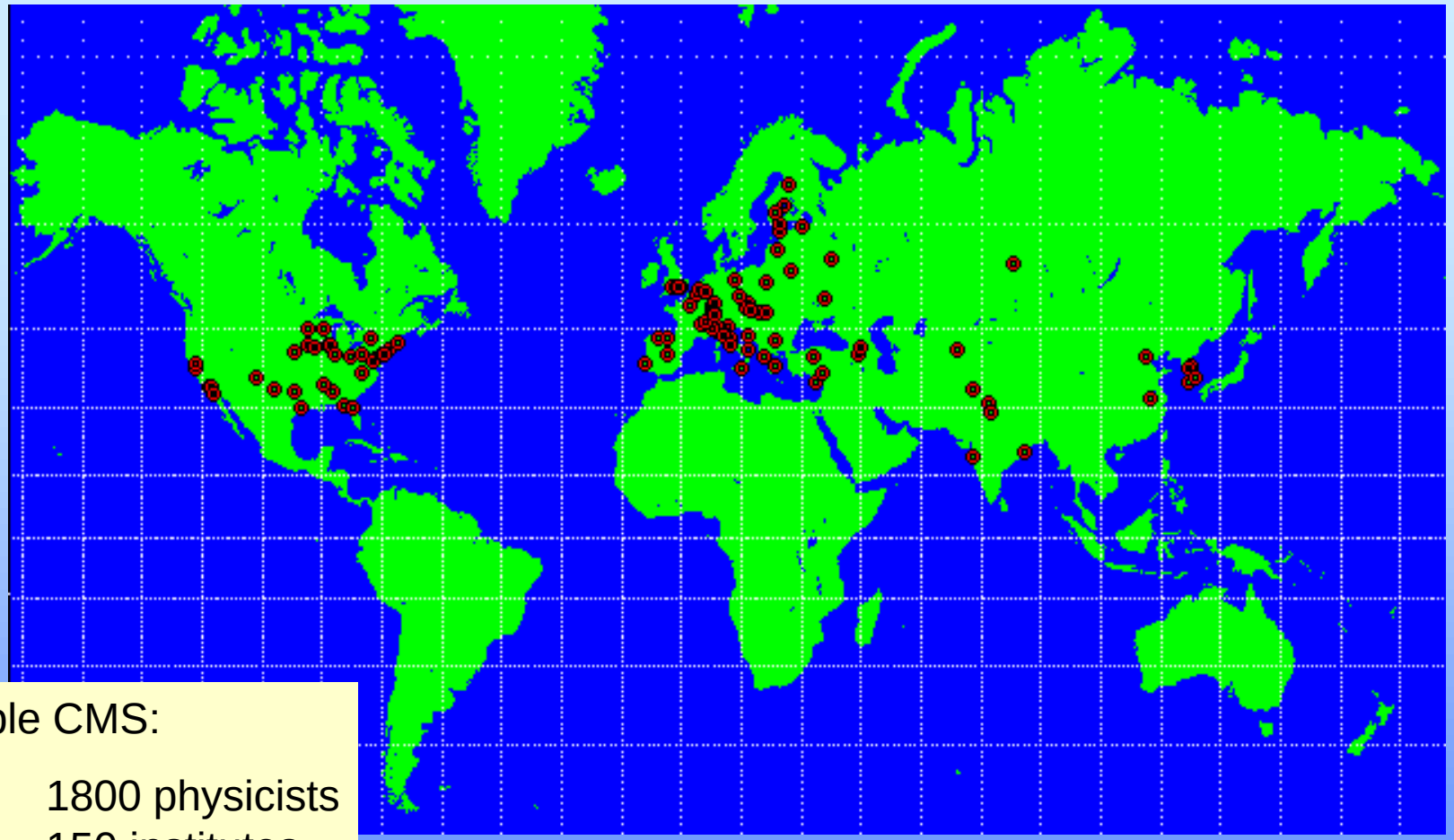
Are Grids a solution?



- Change of orientation of US Meta-computing activity:
  - From inter-connected super-computers towards a more general concept of a computational Grid (The Grid – Ian Foster, Carl Kesselman)

- Has initiated a flurry of activity in HEP:
  - US – Particle Physics Data Grid (PPDG)
  - GriPhyN – data grid proposal submitted to NSF
  - Grid technology evaluation project in INFN
  - UK proposal for funding for a *prototype grid*
  - NASA Information Processing Grid
  - European Data Grid Project

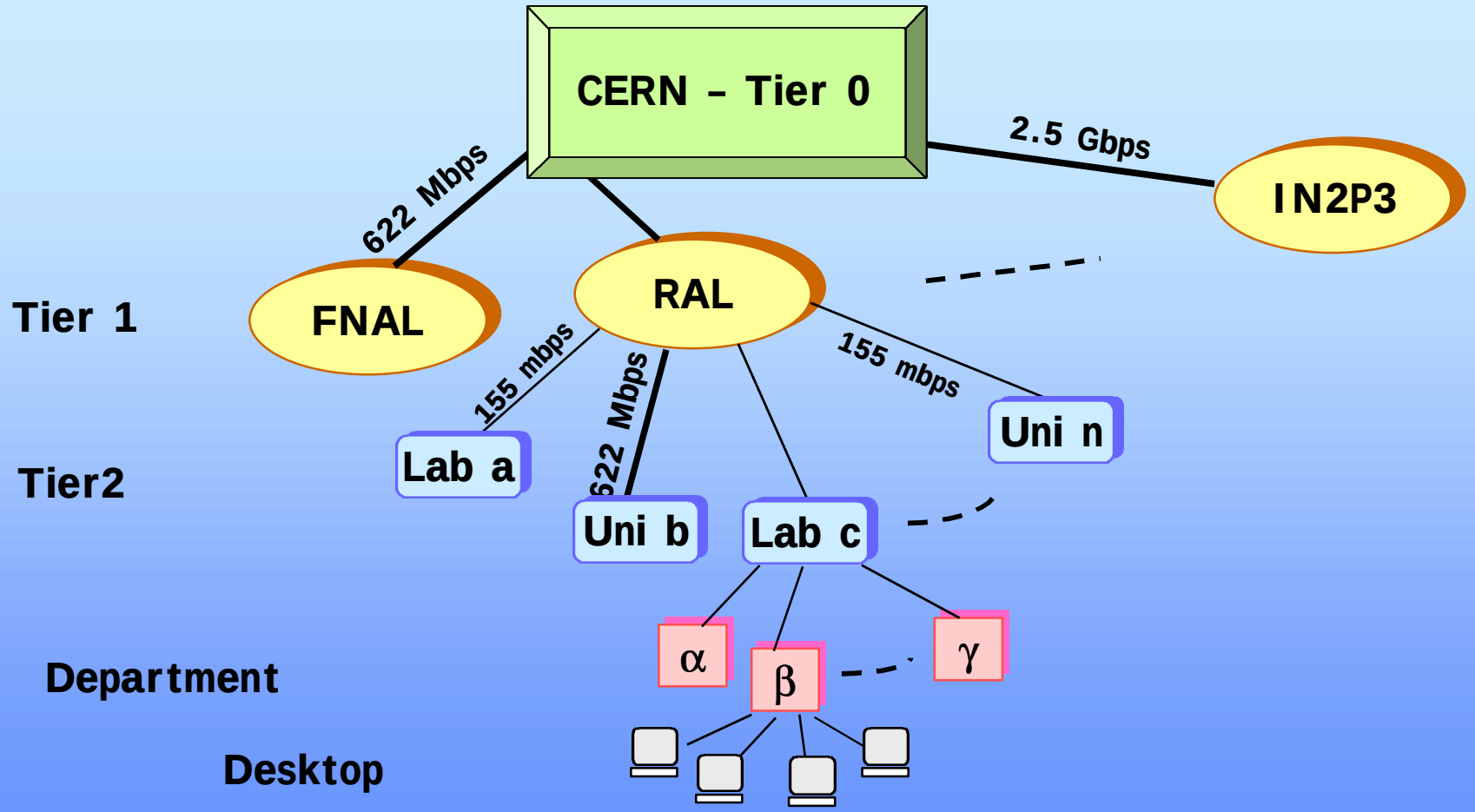
HEP World Wide Collaboration
⇒ distributed computing & storage capacity

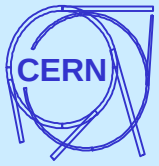


Example CMS:

1800 physicists
150 institutes
32 countries

Regional Centres - a Multi-Tier Model





European Data Grid R&D requirements

Local Fabric

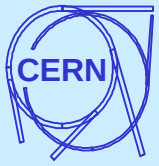
- **Management of giant computing fabrics**
  - auto-installation, configuration management, resilience, self-healing
- **Mass storage management**
  - multi-PetaByte data storage, "real-time" data recording requirement, active tape layer – 1,000s of users

Wide-area

- **Workload management**
  - no central status
  - local access policies
- **Data management**
  - caching, replication, synchronisation
  - object database model
- **Application monitoring**

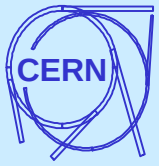
Note: Build on existing components such as Globus Middleware

(done by Foster's (Argonne) and Kesselman's (University of Southern California) groups)



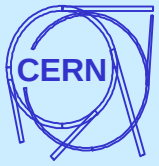
Preliminary programme of work

WP 1	Grid Workload Management	C. Vistoli/INFN
WP 2	Grid Data Management	B. Segal/CERN
WP 3	Grid Monitoring services	R. Middleton/PPARC
WP 4	Fabric Management	T. Smith/CERN
WP 5	Mass Storage Management	J. Gordon/PPARC
WP 6	Integration Testbed	F. Etienne/CNRS
WP 7	Network Services	C. Michau/CNRS
WP 8	HEP Applications	F. Carminati/CERN
WP 9	EO Science Applications	L. Fusco/ESA
WP 10	Biology Applications	C. Michau/CNRS
WP 11	Dissemination	G. Mascari/CNRS
WP 12	Project Management	F. Gagliardi/CERN



WP 1 GRID Workload Management

- **Goal:**
  define and implement a suitable architecture for distributed scheduling and resource management in a GRID environment.
- **Issues:**
  - **Distributed scheduling (data and/or code migration) of unscheduled/scheduled jobs**
  - **Uniform interface to various local resource managers**
  - **Priorities, policies on resource (CPU, Data, Network) usage**
  - **Optimal co-allocation of data, CPU and network for specific "grid/network-aware" jobs**

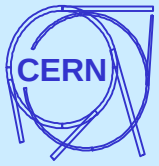


WP 2 GRID Data Management

- **Goal:**

  to specify, develop, integrate and test tools and middle-ware infrastructure to coherently manage and share petabyte-scale information volumes in high-throughput production-quality grid environments.

- **Issues:**
  - Universal Name Space
  - WAN Data Access / Authentication / Authorization
  - Efficient Data Transfer / Caching & Replication
  - Interfacing to Mass Storage Systems
- **Also:**
  - Meta Data Publishing & Management



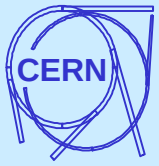
WP 3 GRID Monitoring Services

- **Goal:**

  to specify, develop, integrate and test tools and infrastructure to enable end-user and administrator access to status and error information in a Grid environment.

- **Goal:**

  to permit both job performance optimisation as well as allowing for problem tracing, crucial to facilitating high performance Grid computing.



WP 4 Fabric Management

- **Goal:**

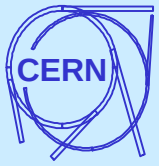
  to facilitate high performance grid computing through effective local site management.

- **Goal:**

  to permit job performance optimisation and problem tracing.

- **Goal:**

  using experience of the partners in managing clusters of several hundreds of nodes, this work package will deliver a computing fabric comprised of all the necessary tools to manage a centre providing grid services on clusters of thousands of nodes



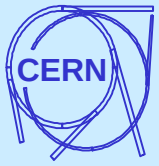
WP 5 Mass Storage Management

- **Goal:**

  Recognising the use of different existing Mass Storage Management Systems by the HEP community, provide extra functionality through common user and data export/import interfaces to all different *existing* local mass storage systems used by the project partners.

- **Goal:**

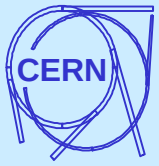
  Ease integration of local mass storage system with the GRID data management system by using these interfaces and through relevant information publication.



WP 6 Integration test bed

Goals:

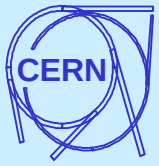
- plan, organise, and enable test beds for the end-to-end application experiments, which will demonstrate the effectiveness of the Data Grid in production quality operation over high performance networks.
- integrate successive releases of the software components from each of the development work packages.
- demonstrate by the end of the project test beds operating as production facilities for real end-to-end applications over large trans-European and potentially global high performance networks.



WP 7 Networking Services

Goals:

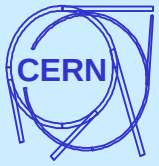
- review the network service requirements of Data Grid and make detailed plans in collaboration with the European and national actors involved.
- establish and manage the Data Grid Virtual Private Network.
- monitor the traffic and performance of the network, and develop models and provide tools and data for the planning of future networks, especially concentrating on the requirements of grids handling significant volumes of data.
- Deal with the distributed security aspects of Data Grid.



WP 8 HEP Applications

- **Goal:**

  Develop and/or adapt High Energy Physics applications (Simulation, Data Analysis, etc.) for the geographically distributed community using the functionality provided by the Data Grid, i.e. transparent access to distributed data and high performance computing facilities.



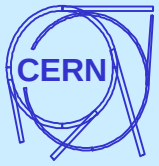
WP 9 Earth Observation Science Applications

- **Goal:**

  to define and develop EO specific components to integrate the GRID platform and bring GRID-aware applications to the earth science community.

- **Goal:**

  provide a good environment to exploit Earth Observation Science (EO) applications that require large computational power and access large data files distributed over geographical archive.



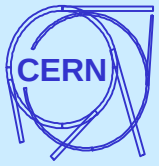
WP 10 Biology Science Applications

Goals:

- Production, analysis and data mining of data produced within projects of sequencing of genomes or in projects with high throughput for the determination of three-dimensional macromolecular structures

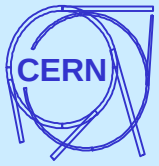
- Production, storage, comparison and retrieval of measures of the genetic expression levels obtained through systems of gene profiling based on micro-arrays, or through techniques that involve the massive production of non-textual data as still images or as Video.

- Retrieval and in-depth analysis of the biological literature (commercial and public) with the aim of the development of a search engine for relations between biological entities



WP 11 Information Dissemination and Exploitation

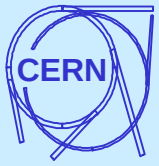
- Goal: to create the critical mass of interest necessary for the deployment, on the target scale, of the results of the project. This allows the development of the skills, experience and software tools necessary to the growth of the world-wide Data Grid
- Goal: promotion of the Data Grid middleware in industry projects and software tools
- Goal: coordination of the dissemination activities undertaken by the project partners in the European countries
- Goal: Industry & Research Grid Forum initiated as the main exchange place of information dissemination and potential exploitation of the Data Grid results



WP 12 Project Management

Goals:

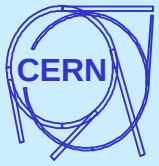
- Overall management and administration of the project
- Coordination of technical activity within the project
- Conflict and resource allocation resolution
- External relations



Participants

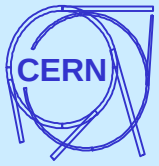
Main partners:	**CERN, INFN(I), CNRS(F), PPARC(UK), NIKHEF(NL), ESA-Earth Observation**
Other sciences:	**Metrology (KNMI,NL), Biology, Medicine**
Industrial participation:	**CSSI/F, DataMat/I, IBM/UK**
Associated partners:	**Czech Republic, Finland, Germany, Hungary, Spain, Sweden (mostly computer scientists)**

Formal collaboration with USA being established



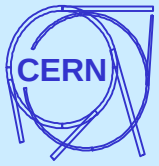
Status

- Prototype work already started at CERN and in most of collaborating institutes (Globus initial installation and tests)
- Proposal to the EU submitted on May 8th, second draft submitted in September, now under final review by EU.



EU Data Grid Main Issues

- Project is by EU standards very large in funding and participants
- Management and coordination will be a challenge
- Coordination between National and Data Grid programmes
- Coordination with US Grid activity
- Coordination of HEP and other sciences objectives
- Very high expectations already raised, could bring disappointments
- EU Project not yet finally approved



Conclusions for HEP

- The Grid is a useful metaphor to describe an appropriate computing model for LHC and future HEP computing
- Middleware, APIs and interface are general enough to accommodate many different models for science, industry and commerce
- Still important R&D to be done
- Major funding agencies prepared to fund large test beds in USA, EU and Japan
- Excellent opportunity for HEP computing
- We need to deliver up to the expectations, therefore adequate resources needed ASAP, not obvious since IT skilled staff scarce in HEP institutes and difficult to hire in the present IT labour market situation