Computing and Data Analysis for Future HEP Experiments

ICHEP 2002

Presented by

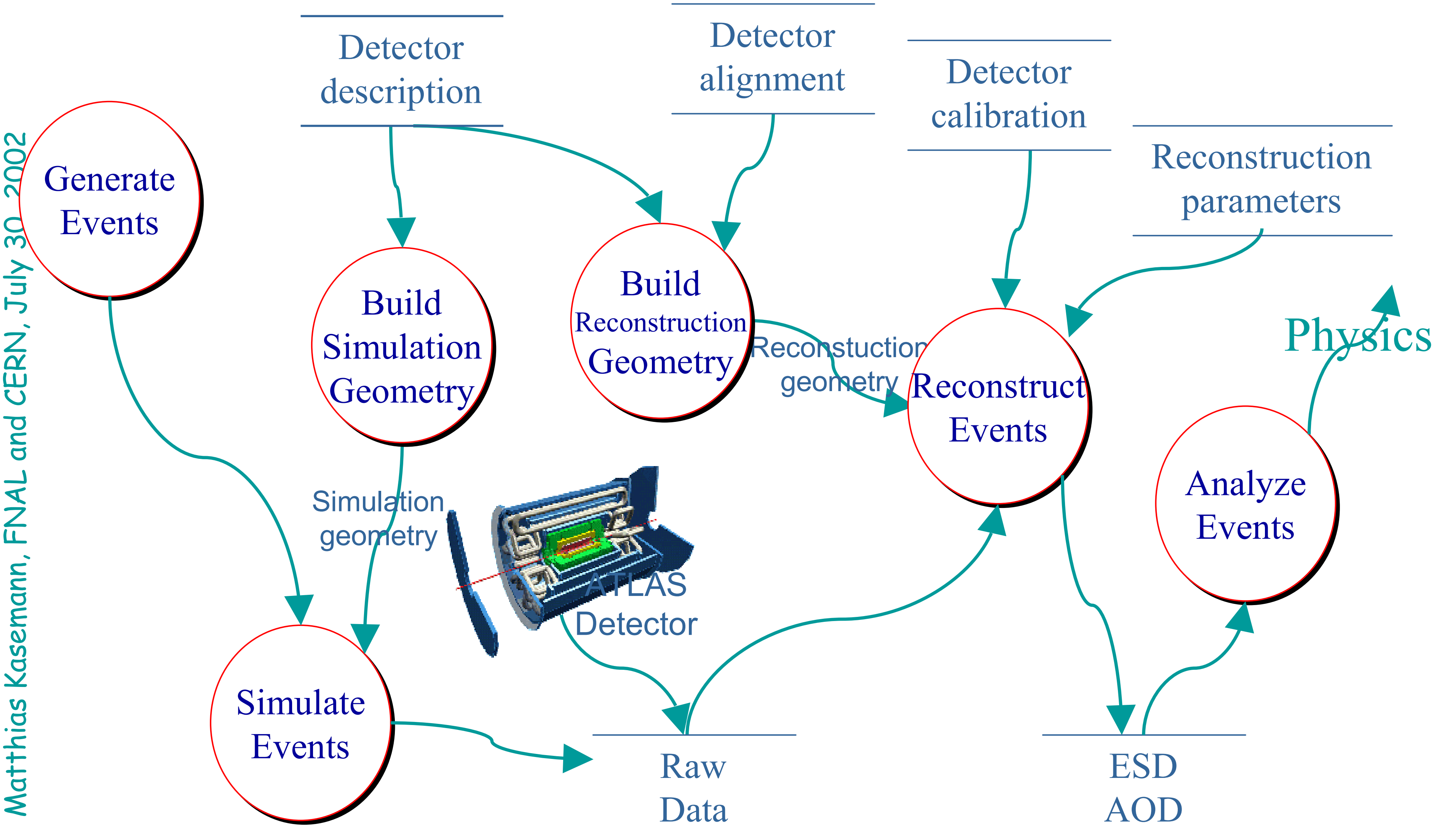
Matthias Kasemann
FNAL and CERN

31st INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS AMSTERDAM

Outline

- Computing for current HEP experiments, lessons learned
  - BaBar and Belle computing
  - CDF and D0 computing
- Technology developments
  - Computing and networking
  - What about Grid computing?
- Computing for LHC experiments
  - Challenges and requirements
  - Next steps and Organization of work

HEP analysis chain: common to most experiments



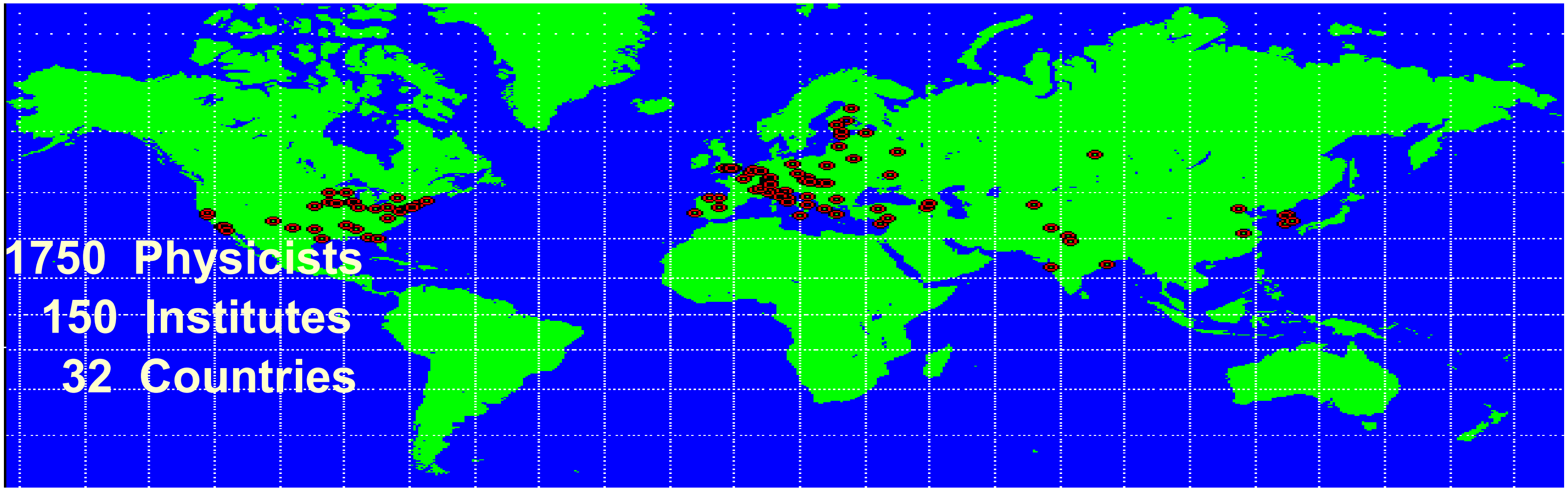
From HEPAP - Long Range Planning Report(2001)

- Information technology (IT) has become an **integral part** of high-energy physics research,
- We are facing a major challenge in filtering, storing and analyzing the data.
- Have to **invest significant resources** in IT research and development, and adapt cutting edge technologies to our purposes, often in partnership with industry.
- We have **profited enormously from the IT advances** of the past two decades.
  - we have benefited from the advances in data handling, retrieval and processing.
  - At the same time, our enormous data volumes, distributed environments and use of networking have **pushed IT** in directions with broad future applications.

Particle Physics Computing Challenges

- Geographical dispersion: of people and resources
- Complexity: the detector and the data
- Scale: Petabytes per year of data per experiment

Example: CMS Experiment



Major challenges associated with:

Communication and collaboration at a distance

Distributed computing resources

Remote software development and physics analysis

BaBar computing: numbers and strategies

- BaBar computing challenges:
  - Choice and setup of (scaling) software and computing model
  - Keep-up with resource requirements (cost)
- Data stored in OODBMS (Objectivity):
  - over 654.1 TB stored, (Mon Jul 15 2002) rate: ~1 TB/day (at FNAL for Run2: ~300 TB (now) rate: ~1 TB/day)
- Changed from **central computing model to distributed model**:
  - major driving force was lack of funding
  - promise: improve physics analysis output (subjective metric...)
- For now: can live with OC12 (622Mbps) between centers

BaBar computing: truly distributed

- Distributed Computing and Analysis:
  - TierA sites in SLAC and LYON,
    - both have full set of analysis data in objectivity
    - SLAC: site for first reconstruction
  - TierA site at RAL for ROOT based analysis data distribution
  - TierA site in Padova ready for data reprocessing (initially)
  - MC production distributed over 15 sides (incl. Lyon), stable
  - people have free and transparent access to Lyon and SLAC
- Data copies at TierA sites <u>improve access performance</u>
- Questions to assess now:
  - manpower is a serious issue to solve and maintain problems and maintain two analysis branches!
  - Re-evaluate (streamline?) the data formats used for analysis.
  - how many copies of the data do they need on disk for performance?
  - how to best use the 4 TierA sites?

BaBar: Summary

- Major **review of computing model every 1.5-2 years** to adapt to experience and changing technology
- Based on experience:
  Computing model is expected to scale to 500 fb-1
  - Both, data rate and volume double every year.
  - Certain Computing loads **scale with rate** , e. g.
    - Prompt reconstruction and Monte Carlo generation.
  - Certain Computing loads **scale with total data volume** , e. g.
    - Data storage and Analysis load.
- There is a formal agreement ("Master MOU") that establishes the framework for Tier A's
- Computing costs go down factor of two every 1.5-2 years, data volume grows faster!

→Simple scaling requires ~ 40% more funds per year.
  - Dominated by disk costs.

BaBar computing: alternatives

- Reduced data sample.
  - Tighter trigger, a la hadron machines, unpopular to unacceptable.
  - Tighter event filter during reconstruction.
- Smaller disk- resident fraction.
  - Larger fraction for more important streams.
  - Physics optimization.
- Greater reliance on staging.
  - Technological improvements such as tapes, drives, robots, mass-storage systems
  - Better data management, e. g. data clustering.
  - Smarter usage, i. e. don't touch a variable unless really necessary.
- Multiple centers !!!

Data storage + Software

- Raw data 1GB/pb-1 (100TB for 100 fb-1)
- Generic MC: MDST: ~10TB/year
- Object Oriented (C++)
  - gcc3 (compiles with SunCC)
- No commercial software
  - QQ, (EvtGen), GEANT3, CERNLIB, CLHEP, Postgres DB
- Legacy FORTRAN code
  - GSIM/GEANT3/ and old calibration/reconstruction code
- I/O: experiment built serial I/O package+zlib
  - The only data format for all stages
    (from DAQ to final user analysis skim files)

Proven: successful computing and analysis model.

Scale of CDF & D0 Computing

- Scope of current computing: 2 experiments @
  ~ 12-15 MB/sec each raw data rate
  ~ 12 MB/sec into reconstruction farms
  ~ 4 - 16 MB/sec out of reconstruction farms
  ~ 150 MB/sec each – total offline capacity for data movement
  - Raw data ~ 150 TB /yr / experiment
  - Total datasets up to 500 TB /yr /experiment
  - Central disk storage now ~ > 150 TB (growing!)

- Computing hardware and infrastructure **cost**:
  - Initial investment ~ $15M / experiment
  - Operating and upgrades: ~$3M / yr / experiment
- Essential:
  - About 35 people / experiment for software and computing

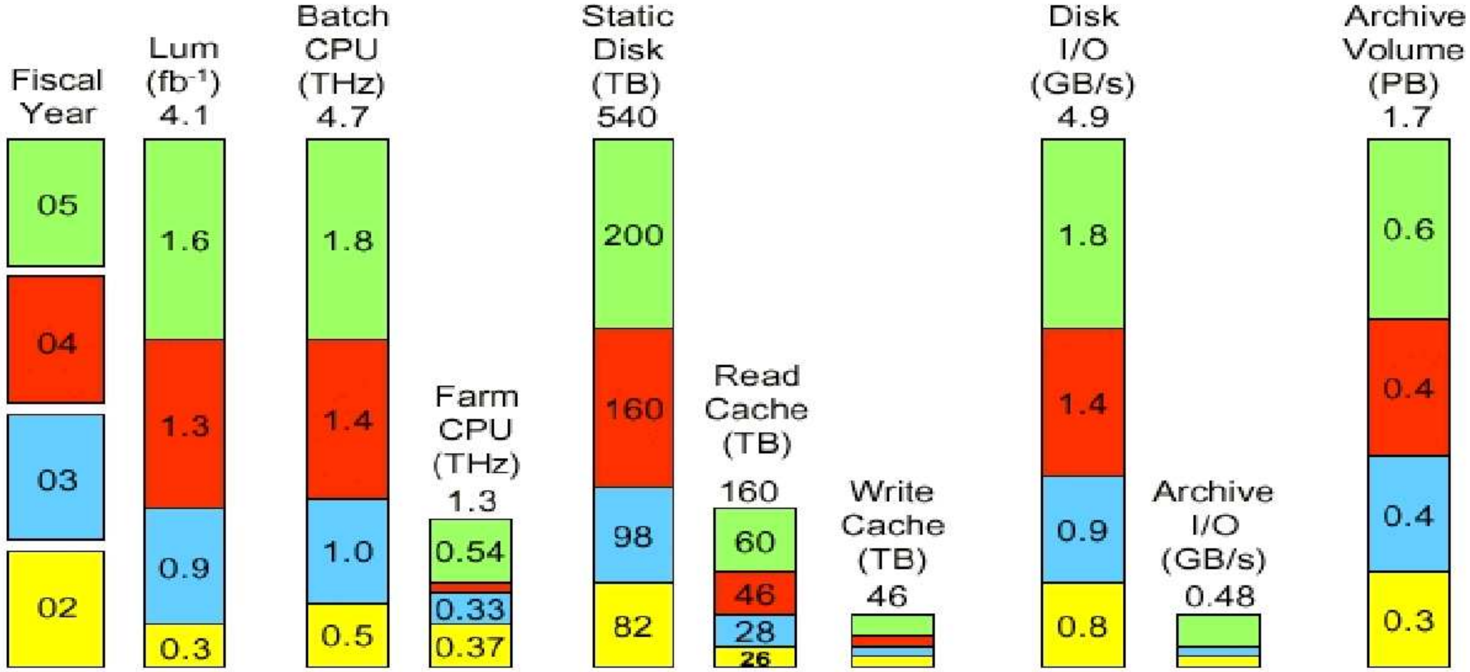
CDF & D0 Software Infrastructure

- Databases - based on ORACLE
- Common code – C++ class library (ZOOM) & CLHEP
- Compilers and Debuggers - common choice of CD-centrally supported products
- User analysis framework - based on ROOT
  - Both experiments use ROOT (freeware with a support arm in the Fermilab CD) as their tool for end-user analysis (making ntuples and histograms)
    [CDF also uses ROOT I / O as its persistent data format.]
- Simulation code - based on common set of physics generators and on the **GEANT3 detector simulation** (although each experiment has its own fast parameterized simulation programs)

Almost all the infrastructure choices are common to the two experiments!

This has been a successful effort to maximize the support benefits from the fixed central Computing Division resources available.

CDF
Computing Requirements



Requirements set by goal:

200 simultaneous users to analyze secondary data set (10^7 evts) in a day

Need ~700 TB of disk and ~5 THz of CPU by end of FY'05:

→ **need lots of disk→ need cheap disk → IDE Raid**

→ **need lots of CPU→ commodity CPU → dual Intel/AMD**

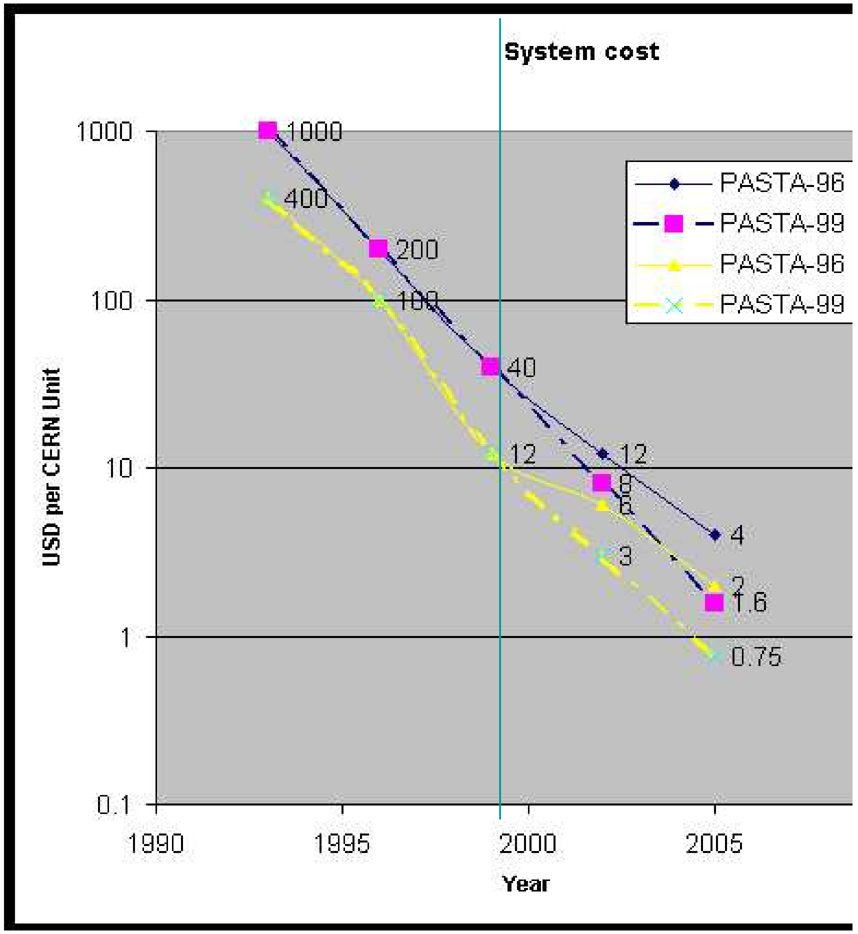
CDF, D0 & Grid Computing

- CDF and D0 developed data **handling systems** and analysis models over the last 3 years, they are successfully deployed and in use
  (see results at this conference)
- At FNAL **the SAM distributed data handling system** was developed for the D0 experiment,
  - it is heavily used for D0
  - CDF starts to deploy it now
- FNAL-CD, CDF and D0 are participating in US- and European based Grid projects
  - **Standard Grid tools will replace functionality** as they become available

Kansas
Michigan St.
Michigan
Oklahoma
Indiana
Fermilab
Lancaster
Imperial
NIKHEF
SAM data transfer
Monte Carlo files
Lyon
Prague
Wuppertal
Munich
Arizona
Texas
Boston
Worldwide Data Grid status in March 2002
5 remote Monte Carlo generation sites + more coming
15 SAM stations for remote analysis + more coming
Flags show the 18 member nations of the collaboration

Technology: HEP computing in 2002

- Bulk computing done on **<u>large clusters of Linux</u>** computers
  - Share of other UNIX's decreasing steadily
    - Used for special servers, data bases, web servers etc
    - Danger: maturity of software development!!
  - HEP computing is trivially parallel (events)
- For planning: extrapolating cost and capacity using Moore's law:
  - doubling every ~1.5 years, expect 10GHz by 2005-07
  - Large uncertainty in any cost extrapolations



Matthias Kasemann, FNAL and CERN, July 30, 2002

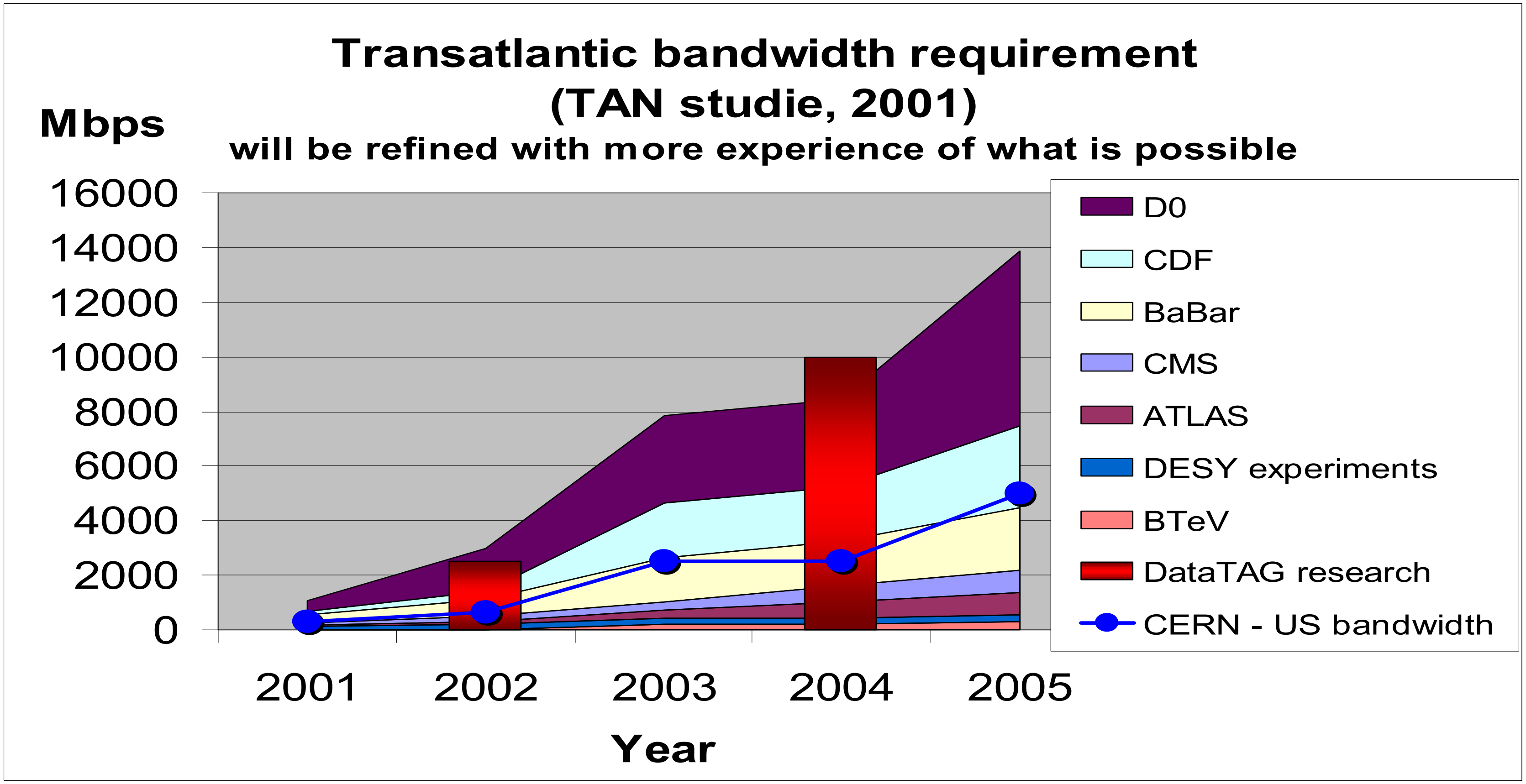
Technology: Data Storage

- **Disk file servers**: Linux, other Unix's
  - SCSI, IDE and fiber channel, RAID and Storage Area Network configurations
  - Low cost IDE disk server: 2 TB for $10.000 expected to drop by x10 until 2006

- **Mass storage**: magnetic tapes used nearly 100%
  - Media cost: $1/GB (2002), expected to drop by x10 until 2006
  - Tape slots in robots (incl. drives): ≥$300/TB

- Tape/disk comparison (at FNAL in 2002)
  - Tape including robotics: ≥$1.300/TB
  - Disk: $5000/TB

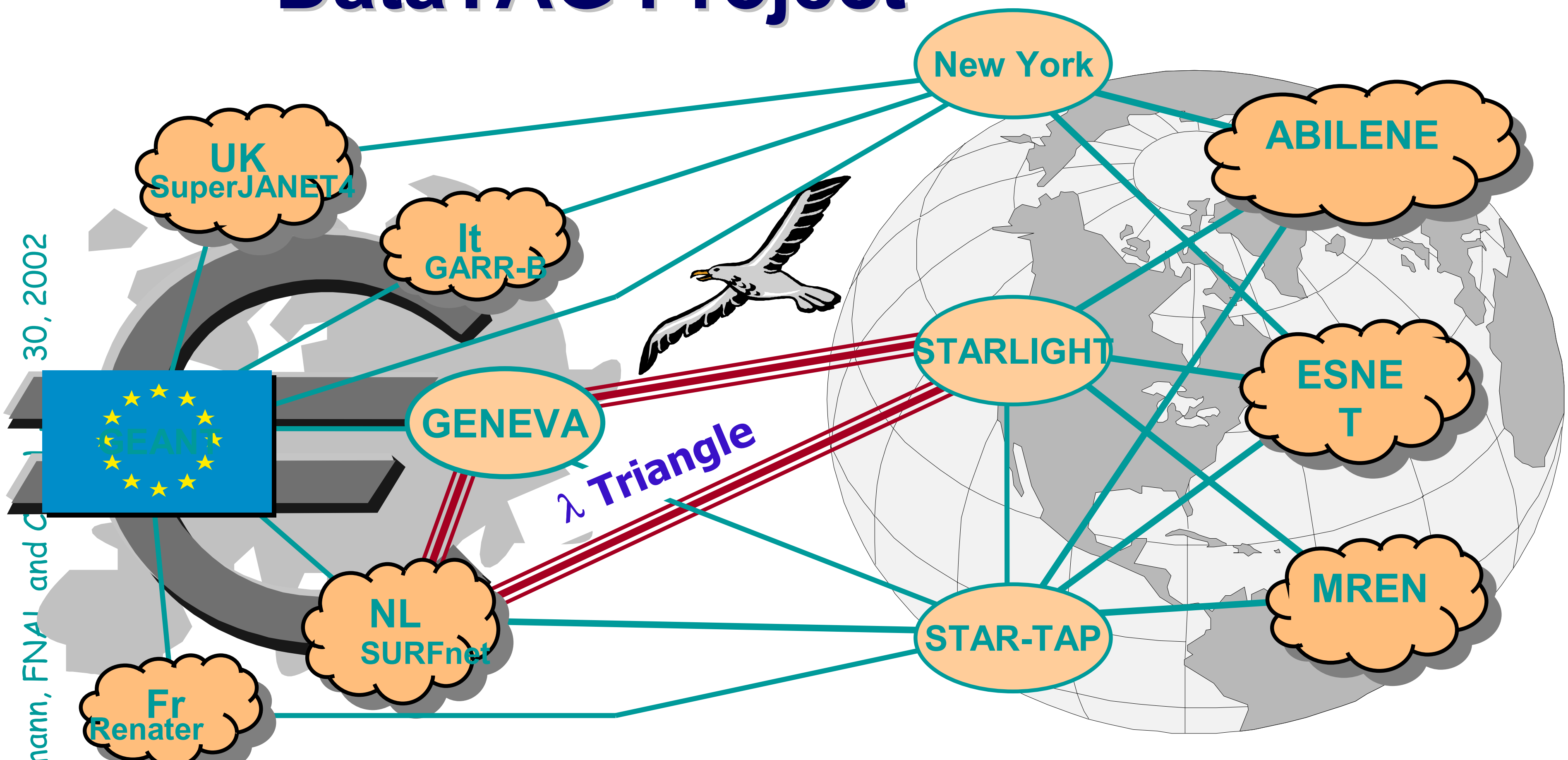
Networking needs: e.g. CERN-US



Installed bandwidth, Maximum Link occupancy of 50% assumed
See: ***http://gate.hep.anl.gov/lprice/TAN***

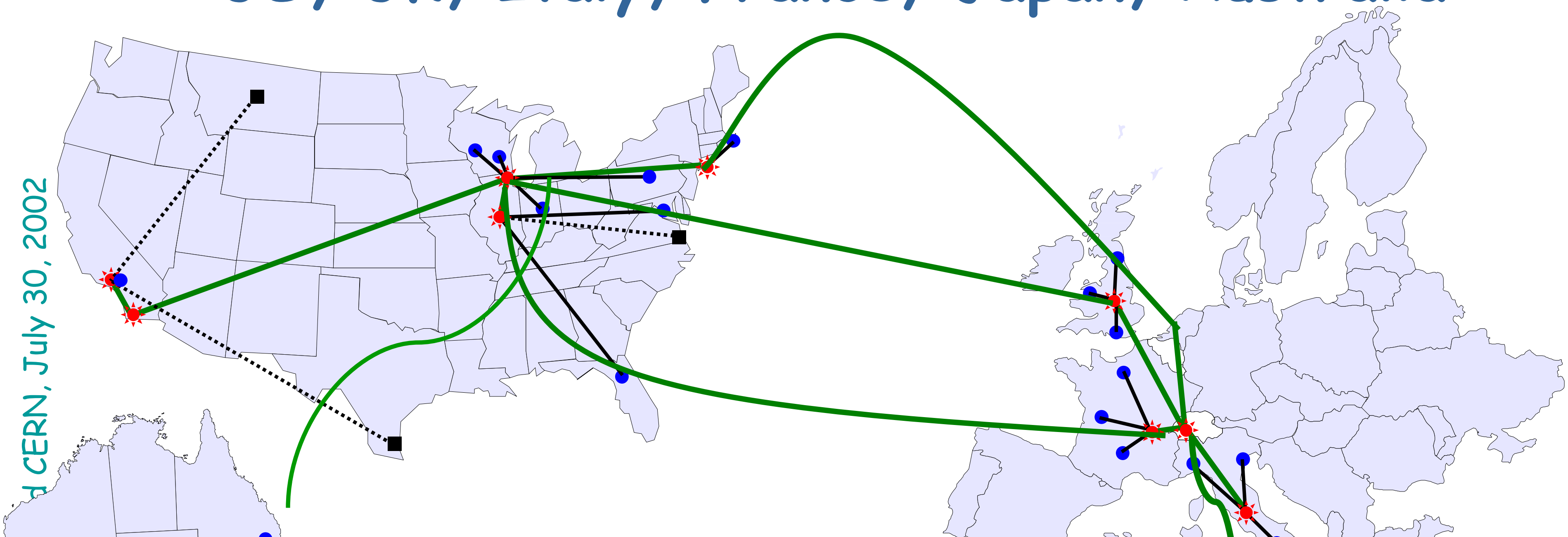
Project: DataTAG 2.5 Gbps Research Link in Summer 2002;
10 Gbps Research Link in ~2003 or Early 2004

DataTAG Project



- EU-Solicited Project. CERN, PPARC (UK), Amsterdam (NL), and INFN (IT); and US (DOE/NSF: UIC, NWU and Caltech) partners
- Main Aims:
  - Ensure maximum interoperability between US and EU Grid Projects
  - Transatlantic Testbed for advanced network research
- 2.5 Gbps wavelength-based US-CERN Link 6/2002 (10 Gbps ~2003 or 2004)

GriPhyN iVDGL Map Circa 2002-2003
US, UK, Italy, France, Japan, Australia



International Virtual-Data Grid Laboratory
- Conduct Data Grid tests "at scale"
- Develop Common Grid infrastructure
- National, international scale Data Grid tests, leading to managed ops (iGOC)

Components
- Tier1, Selected Tier2 and Tier3 Sites
- Distributed Terascale Facility (DTF)
- 0.6 - 10 Gbps networks

Planned New Partners
- Brazil T1
- Russia T1
- *Pakistan T2*
- China T2
- …

d CERN, July 30, 2002

HEMP Major Links:
Bandwidth Roadmap (Scenario) in Gbps

Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005	10	2-4 X 10	λ Switch; λ Provisioning
2007	2-4 X 10	~10 X 10; 40 Gbps	1st Gen. λ Grids
2009	~10 X 10 or 1-2 X 40	~5 X 40 or ~20-50 X 10	40 Gbps λ Switching
2011	~5 X 40 or ~20 X 10	~25 X 40 or ~100 X 10	2nd Gen λ Grids Terabit Networks
2013	~Terabit	~MultiTerabit	~Fill One Fiber

From: ICFA SCIC, H. Newman, Feb, 2002

HENP Networks: Outlook and High Performance Issues

- Higher speeds are soon going to reach limits of existing protocols
  - TCP/IP 25 years old; built for 64 kbps
  - Ethernet 20 years old; built for 10 Mbps
- We need to understand how to use and deploy new network technologies in the 1 to 10 Gbps range
  - Optimize throughput: large windows; perhaps many streams;
  - Will then need new concepts of fair sharing and managed use of networks
  - New [sometimes expensive] hardware; and new protocols
    - GigE and soon 10 GigE on some WAN paths
    - MPLS/GMPLS for network policy; "QoS"
  - Alternatives to TCP ?? (e.g. UDP/RTP + FEC)
  - DWDM and management of Lambdas at 2.5 then 10 Gbps

From: ICFA SCIC, H. Newman, Feb, 2002

The Grid vision of computing

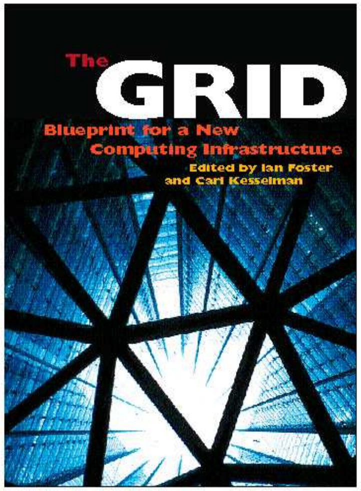
- Flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resource
  - From "The Anatomy of the Grid: Enabling Scalable Virtual Organizations"
- Enable communities ("virtual organizations") to share geographically distributed resources as they pursue common goals -- *assuming the absence* of…
  - central location,
  - central control,
  - omniscience,
  - existing trust relationships.

the globus project
www.globus.org

Merrill Lynch

Globus Grid Computing—the Next Internet
by John Roy/Steve Milunovich

The Internet was first a network and is now a communications platform. The next evolutionary step could be to a platform for distributed computing. This ability to manage applications and share data over the network is called "grid computing."

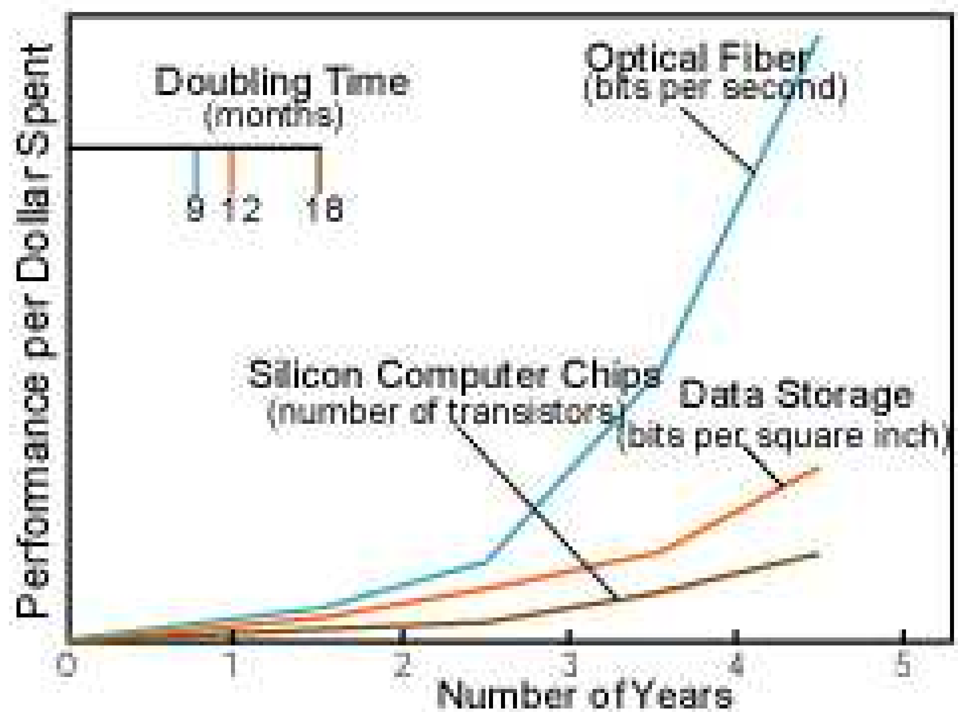


Why compute on a Grid?

- Network vs. computer performance
  - Computer speed doubles every 18 months
  - Network speed doubles every 9 months
  - Difference = order of magnitude per 5 years

!!!

- 1986 to 2000
  - Computers: x 500
  - Networks: x 340,000
- 2001 to 2010
  - Computers: x 60
  - Networks: x 4000



Moore's Law vs. storage improvements vs. optical improvements. Graph from **Scientific American** (Jan-2001) by Cleo Vilett, source Vined Khoslan, Kleiner, Caufield and Perkins.

The Grid World: Current Status

- Dozens of **major Grid projects** in scientific & technical computing/research & education
- **Considerable consensus** on key concepts and technologies
  - Open source Globus Toolkit™ a de facto standard for major protocols & services
  - Far from complete or perfect, but out there, evolving rapidly, and large tool/user base
- Industrial interest emerging rapidly
- Concepts map very well with HEP style of collaborating
  - Globally spread participation in experiments
  - Many funding sources
  - Widely spread expertise
  - 'transparent' access to data

Improve scientific result by

- **easy data access**
- **broad participation**

How do we solve problems? Q: is this valid for HEP?

- Communities committed to common goals
  - Virtual organizations map well to HEP collaborations
  - Teams with heterogeneous members & capabilities
- Distributed geographically and politically
  - No location/organization possesses all required skills and resources
- Adapt as a function of the situation
  - Adjust membership, reallocate responsibilities, renegotiate resources
- Online negotiation of access to services (dynamically):
  - who, what, why, when, how
- Establishment of applications and systems able to deliver multiple qualities of service
- Autonomic management of infrastructure elements
- Open, extensible, evolvable infrastructure

Grid Technology Area
Leveraging Grid R&D Projects

the globus project™
www.globus.org

DataGRID

crossgrid

GriPhyN
Data Intensive Science

PPDG

Trillium
GriPhyN PPDG iVDGL

iVDgL

DataTAG

Nordic Testbed for Wide Area Computing and Data Handling

GRID UK Particle Physics

INFN
Istituto Nazionale di Fisica Nucleare

Condor
High Throughput Computing

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

US projects

European projects

Grid Technology Area
Leveraging Grid R&D Projects

- significant R&D funding for Grid middleware
- risk of divergence
  → requires substantial coordination effort
  and interfacing work to HEP effort
- global grids need standards
- useful grids need stability
- hard to do this in the current state of maturity
  - Extensive testing and
    prototyping program required

We (HEP) feel we have no choice than to participate!!

The Globus Toolkit

- Globus Toolkit is the source of many of the protocols described in "Grid architecture"
- Adopted by almost all major Grid projects worldwide as a source of infrastructure
- Open source, open architecture framework encourages community development
- Active R&D program continues to move technology forward
- Developers at ANL, USC/ISI, NCSA, LBNL, and other institutions
- Next steps:
  - Globus v3: implement toolkit using Web services (OGSA)
  - Service orientation to "virtualize" resources
    - Everything is a service

What is a "virtual" dataset?

- Tracking the <u>derivation</u> of experiment data with high fidelity
- Transparency with respect to location and materialization
  - Track all data assets
  - Accurately record how they were derived
  - Encapsulate the transformations that produce new data objects

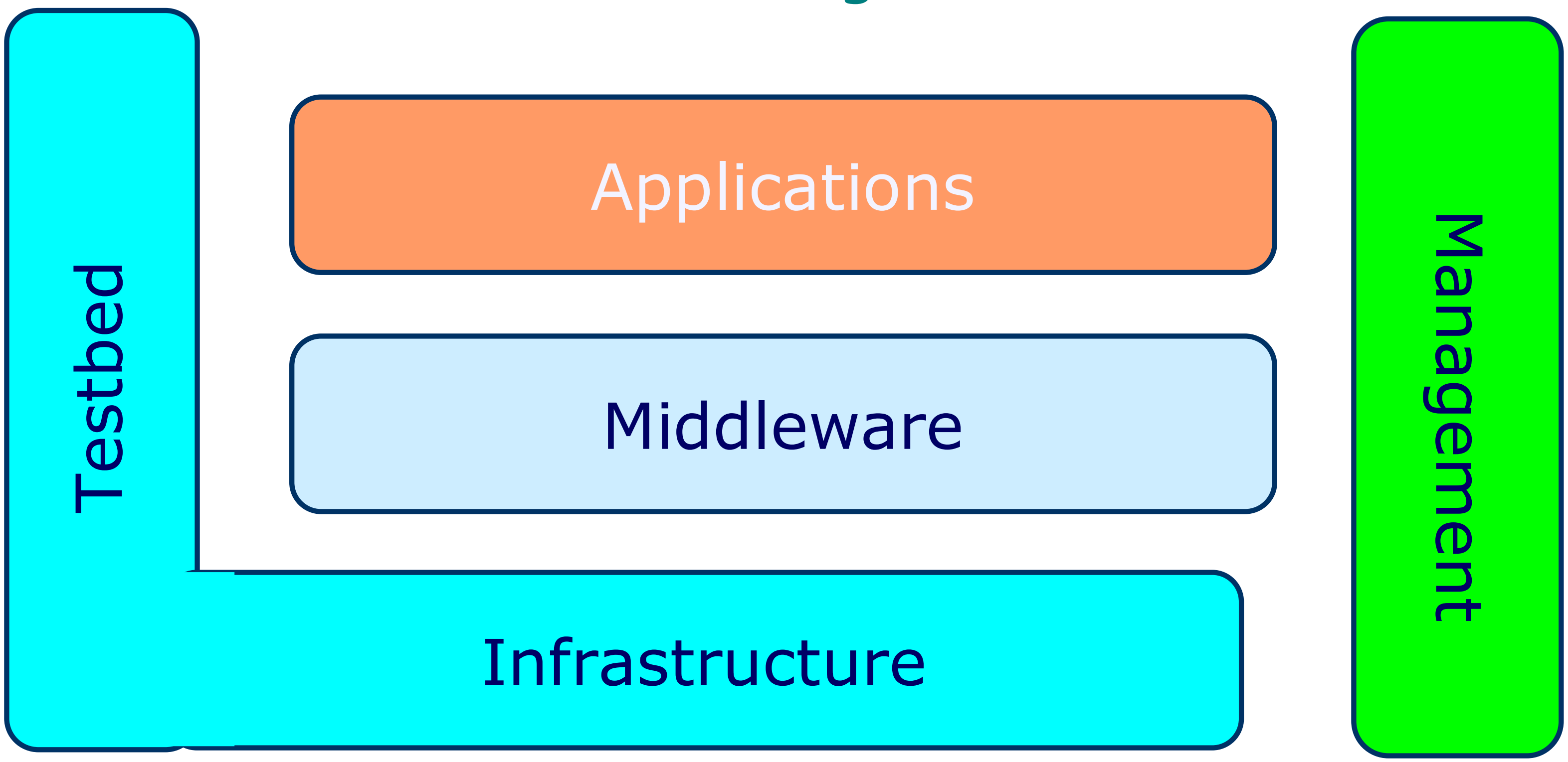
- Resulting data access possibilities are:
  1. **<u>Access data at storage site</u>**
  2. **<u>Copy dataset to requesting site</u>**
  3. **<u>Recreate dataset at requesting site</u>**

EU DataGrid Project Objectives

- DataGrid is a project funded by European Union whose objective is *to exploit and build the next generation computing infrastructure providing intensive computation and analysis of shared large-scale databases*.
  - Start (Kick off) : Jan 1, 2001 End : Dec 31, 2003
- Applications/End Users Communities : HEP, Earth Observation, Biology
- Specific Project Objectives:
  - Middleware for fabric & grid management
  - Large scale testbed
  - Production quality demonstrations
  - To collaborate with and complement other European and US projects
  - Contribute to Open Standards and international bodies (GGF, Industry & Research forum)

EU DataGrid Working Areas

- The project is up and running!
  - All 21 partners are now contributing at contractual level
  - total of ~60 man years for first year
- The DataGrid project is divided in 12 Work Packages distributed in four Working Areas

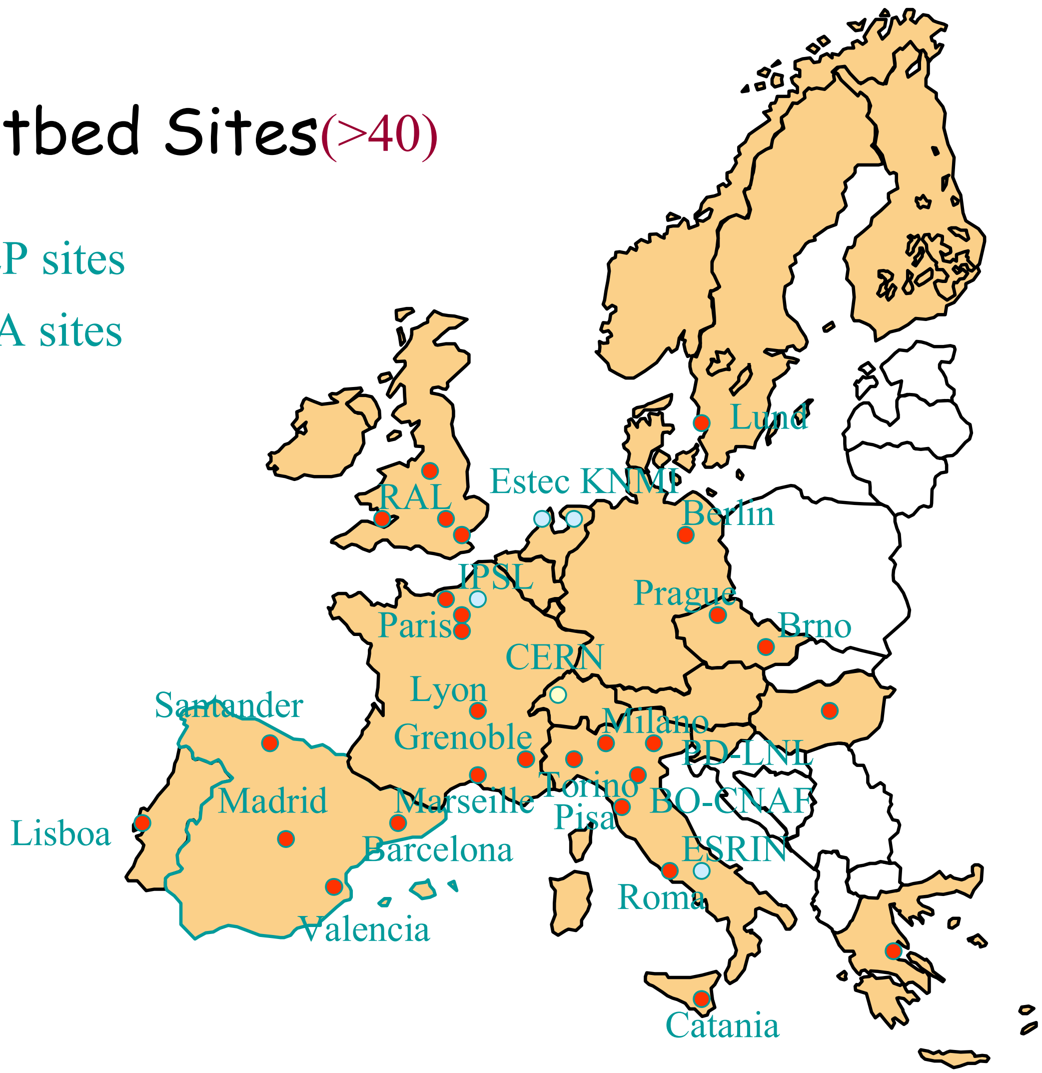


Matthias Kasemann, FNAL and CERN, July 30, 2002

EU DataGrid Testbed

Testbed Sites(>40)

- HEP sites
- ESA sites



- Dubna
- Moscow

Nordic Testbed for Wide Area Computing and Data Handling

Another Grid Implementation: NorduNet

- Launched in spring 2001, with the aim of creating a Grid infrastructure in the Nordic countries.
- Partners from Denmark, Norway, Sweden, and Finland.
- Powered mainly by ATLAS groups (Lund, Copenhagen, Stockholm, Uppsala, Oslo).
- Relatively short term project – ends in October 2002.
- Relies on very limited human resources (3 full–time researchers, few part–time ones) with funding from NorduNet2.

US CMS T2 Prototypes and Test-beds

Tier-1 and Tier-2 Prototypes and Test-beds operational

University Wisconsin at Madison

Condor
High Throughput Computing

"Thousands of CPUs!"

University of Florida

iVDgL

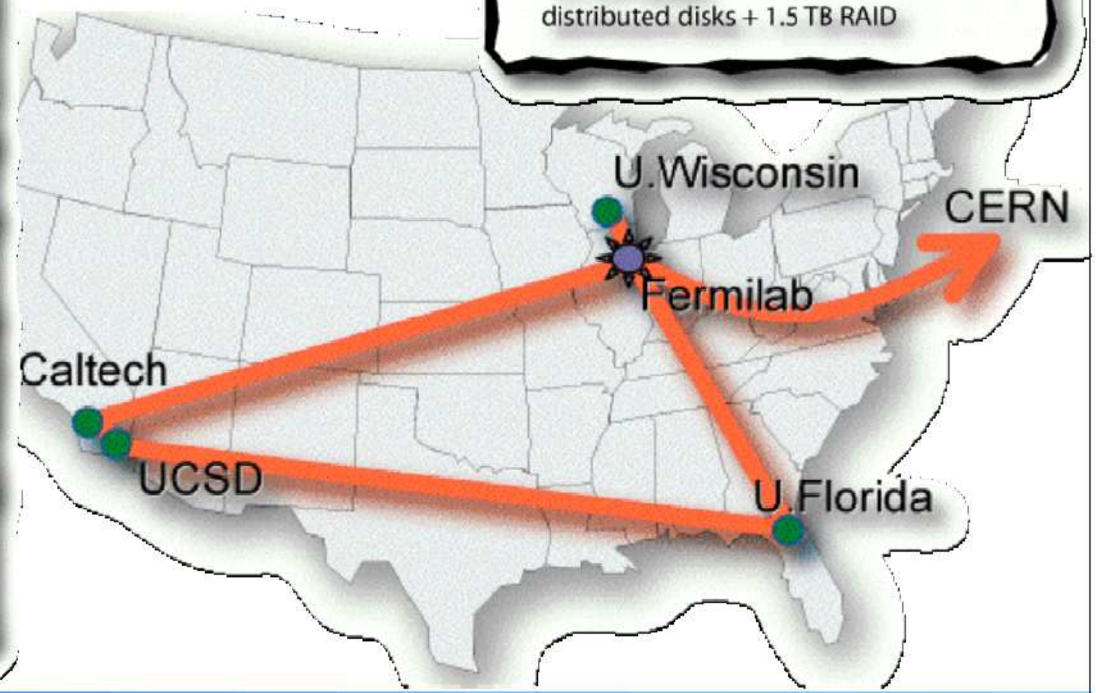
2nd U.S. CMS prototype Tier-2
72 CPU nodes
distributed disks + 1.5 TB RAID

California prototype Tier-2

Caltech

UCSD

40 Duals
TB Storage Servers
Gbit Ethernet switches



EU Data Grid projects: Future Plans

- Expand and consolidate testbed operations
  - Improve the distribution, maintenance and support process
  - Understand, refine Grid operations
- Evolve architecture and software on the basis of Testbed usage and feedback from users
  - Make the various Grid efforts interoperable
    - Defined Project: Grid Laboratory Uniform Environment (GLUE)
  - Adapt to Globus Web services interfaces and components
- Prepare for second test bed in autumn 2002 in close collaboration with LHC Computing Grid project (LCG)
- Enhance synergy between EU and US projects
- Promote early standards adoption with participation to relevant bodies

Computing for the LHC experiments

A new Project has been setup at CERN:
the **LHC Grid Computing Project (LCG)**

The first phase of the project: 2002-2005

- preparing the prototype computing environment, including
  - support for applications – libraries, tools, frameworks, common developments, …..
  - global grid computing service
- Shared funding by Regional Centers, CERN, Contributions
- Grid software developments by national and regional Grid projects

Phase 2: 2005-2007
construction and operation of the initial LHC Computing Service

LCG: Steps towards LHC computing

- Prepare and deploy the LHC Computing Environment
  - Applications - provide the common components, tools and infrastructure for the physics application software
  - Computing system – fabric, grid, global analysis system
  - Deployment – foster collaboration and coherence
  - Not just another grid technology project
- Validate the software by participating in Data Challenges using the progressively more complex Grid Prototype
  - Phase 1 - 50% model production grid in 2004
- Produce a TDR for full system to be built in Phase 2
  - Software performance impacts on size and cost of production facility
  - Analysis models impact on exploitation of production grid
- Maintain opportunities for reuse of deliverables outside LHC experimental programme

LCG: Applications Activity Areas

- Application software infrastructure
  - physics software development environment, standard libraries, development tools
- Common frameworks for simulation and analysis
  - Development and integration of toolkits & components
- Support for physics applications
  - Development, support of common software tools & frameworks
- Adaptation of Physics Applications to Grid environment
- Object persistency and data management tools
  - Event data, metadata, conditions data, analysis objects,

Potential common LHC software (1/2)

Data persistency	High priority item
Simulation tools	Important part
Detector description, model	Description tools, geometry model
Conditions database	In addition to event persistency
Data dictionary	Key need for common service
Interactive frameworks	What do we want, have, need
Statistical analysis	Tools, interfaces, integration
Visualization	Tools, interfaces, integration
Physics packages	Important area but scope unclear
Framework services	If common framework is too optimistic…
C++ class libraries	Standard foundation libraries

Potential common LHC software (2/2)

Event processing framework	Long term
Distributed analysis	Application layer over grid
Distributed production	Application layer over grid
Small scale persistency	Simple persistency tools
Software testing	Together with Software management
Software distribution	From central 'Program Library' to convenient broad distribution
OO language usage	C++, Java (..?) roles in the future
Benchmarking suite	Comprehensive suite for LCG software
Online notebooks	Long term

Summary of Computing Capacity Required for all LHC Experiments in 2007

source: CERN/LHCC/2001-004 - Report of the LHC Computing Review - 20 Fe

(ATLAS with 270Hz trigger)

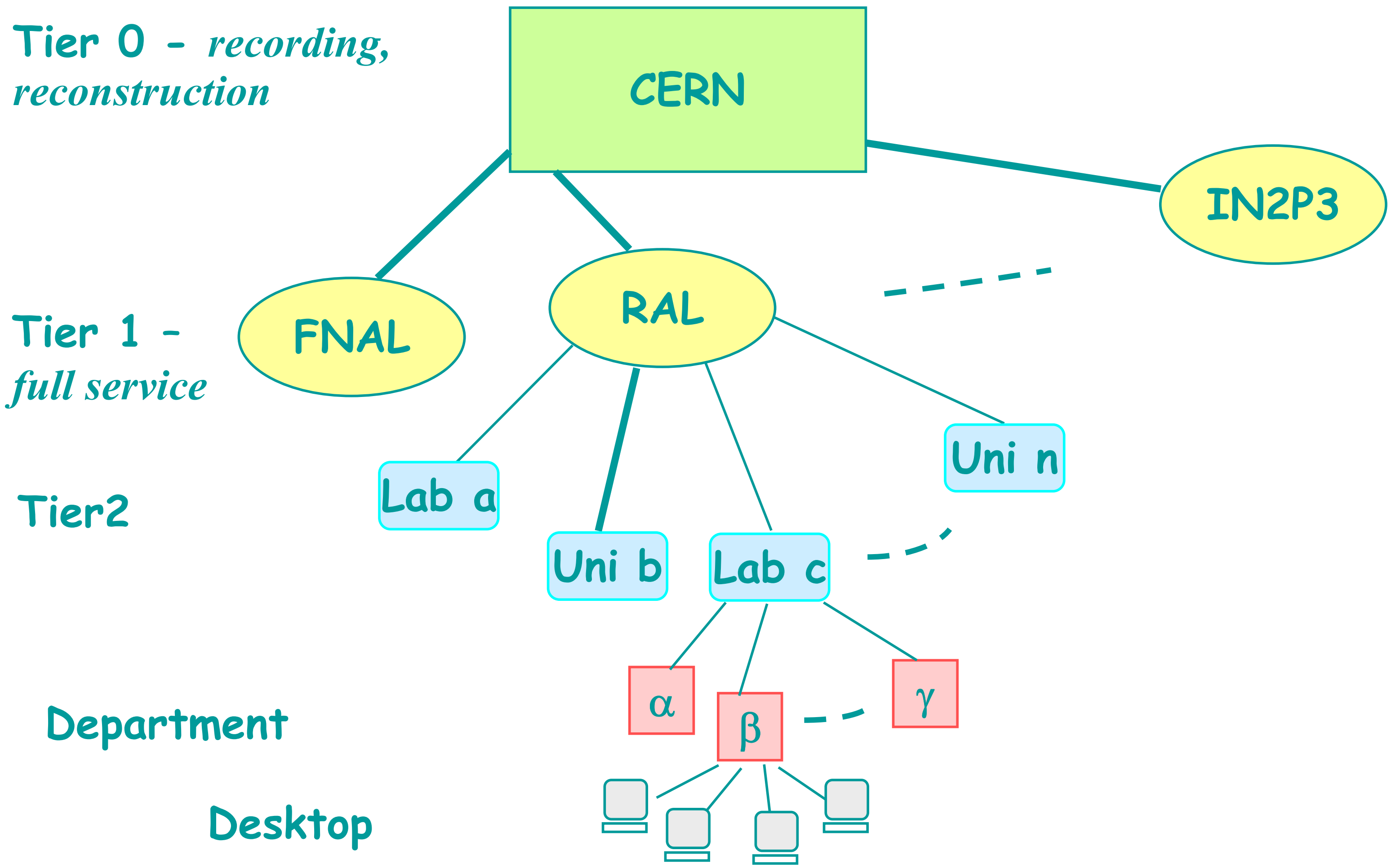
Today: ~100 Si95/box
In 2007: 800 Si95/box

	---------- CERN ----------			Regional	Grand
	Tier 0	**Tier 1**	**Total**	**Centres**	**Total**
Processing (K SI95)	1,727	832	2,559	4,974	7,533
Disk (PB)	1.2	1.2	2.4	8.7	11.1
Magnetic tape (PB)	16.3	1.2	17.6	20.3	37.9

LHC experiments foresee (Funding dictates) –

- Worldwide distributed computing system
- Small fraction of the analysis at CERN
- Batch analysis – using 12-20 large regional centers
  - how to use the resources efficiently
  - establishing and maintaining a uniform physics environment
- Data exchange and interactive analysis involving tens of smaller regional centers, universities, labs

The MONARC Multi-Tier Model (1999)



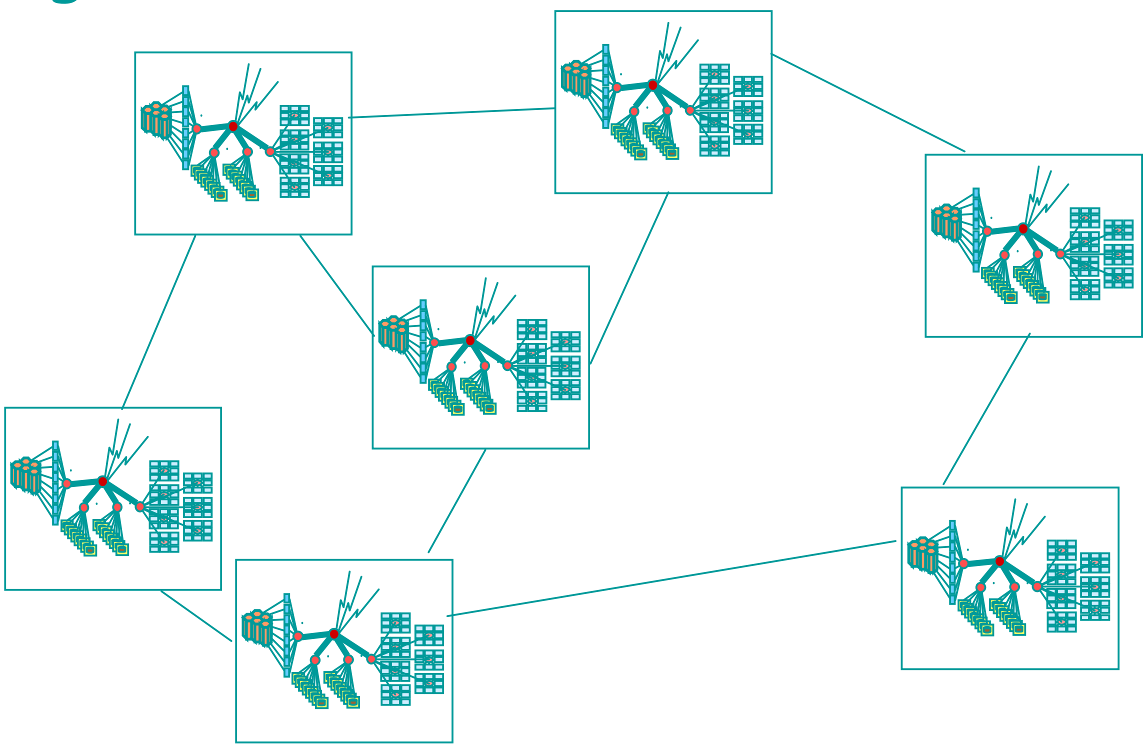
Matthias Kasemann, FNAL and CERN, July 30, 2002

MONARC report: http://home.cern.ch/~barone/monarc/RCArchitecture.html

Building a Grid for LHC

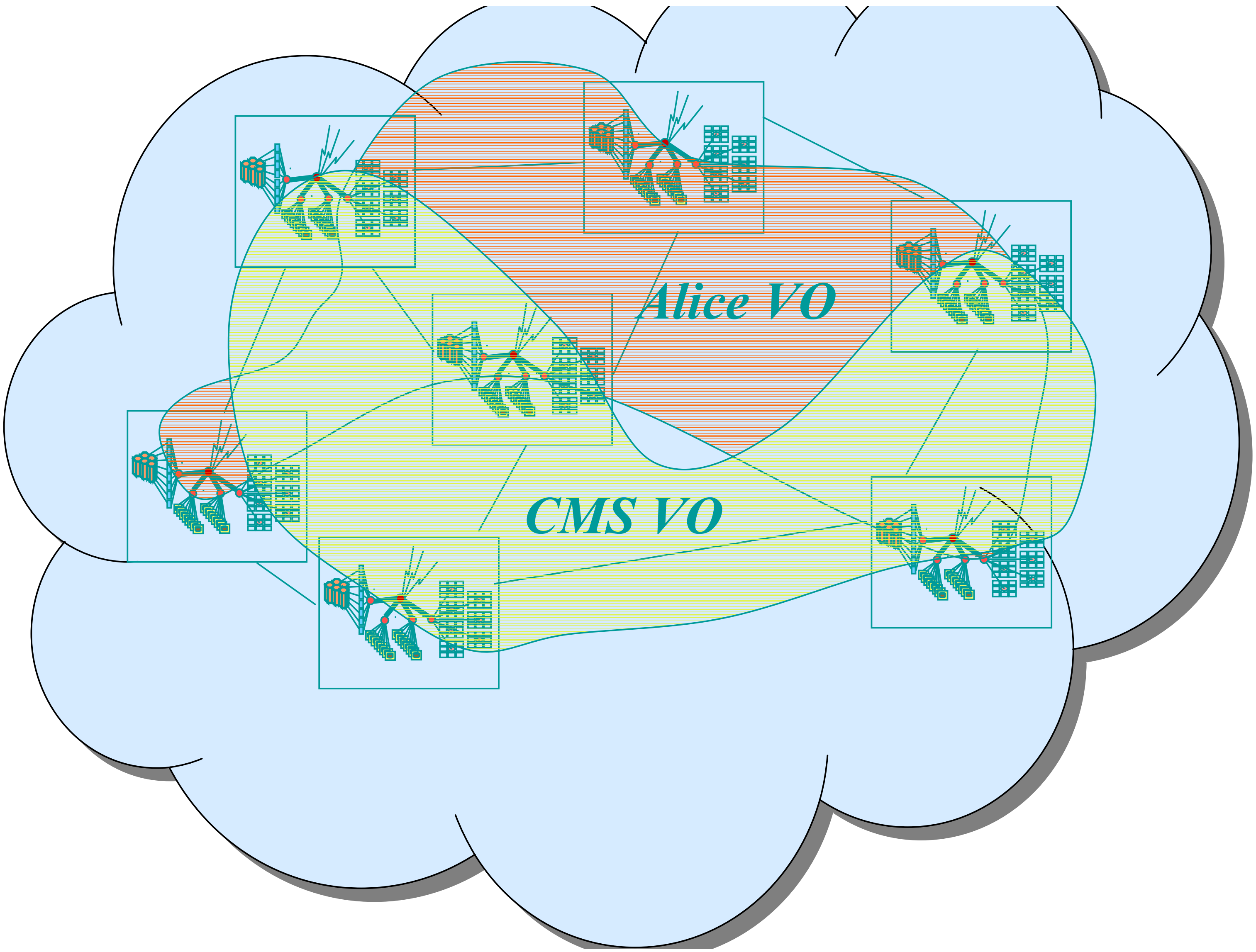
Matthias Kasemann, FNAL and CERN, July 30, 2002

Collaborating Computer Centers



Building a Grid for LHC

→The "virtual" LHC Computing Center



Matthias Kasemann, FNAL and CERN, July 30, 2002

The “virtual” LHC Computing Center

- The aim is to build
  - a general computing service
  - for a very large user population
  - of independently-minded scientists
  - using a large number of independently managed sites
- This is NOT a collection of sites providing pre-defined services
  - it is the user’s job that defines the service
  - it is current research interests that define the workload
  - it is the workload that defines the data distribution
- DEMAND - Unpredictable & Chaotic
- But the SERVICE had better be Available & Reliable

LCG: We need to use Grid Technology

- Supplied and maintained by the "Grid projects"
  - Current status:
    - Work to get the first "production" data intensive grids going as user services
    - Establish long-term support and maintenance model
    - Find balance between new functionality and stability

- For LHC we must deploy (and participate in) a GLOBAL COMPUTING GRID
  - essential to have compatible middleware & grid infrastructure across all sites
  - better – have identical middleware

Funding arguments for HEP computing

- "*Because we need it*" may not bring as far enough!
- HEP seen as a ground-breaker in computing
  - initiator of the Web
  - track record of exploiting leading edge computing
  - effective global collaborations
  - real need – for data as well as computation
  - one of the few application areas with real cross-border data needs
- LHC in sync with
  -- emergence of Grid technology
  -- explosion of network bandwidth available
- The LCG project must deliver on Phase 1 for LHC -
  and show the relevance for other sciences
- We are getting funding because of the relevance for other sciences, engineering, business --
  keeping things general, main-line must remain a high priority

Conclusions

- Existing experiments cannot perform analysis without substantial resources
  - It is easier to collect and operate the in a distributed way (using Grid ideas and technology)
- Example: ATLAS guiding principles (true for all LHC experiments):
  - Every physicist in ATLAS must have the best possible **access to the data** necessary for the analysis, **irrespective of his/her location**.
  - The access to the data should be transparent and efficient.
  - We should **profit from resources** (money, manpower and hardware) available in the different countries.
  - We should **benefit from the outcome of the Grid projects**.
- The leading role and the massive participation of high-energy physics is based on the assumption that the Grid will form the basis of the LHC computing, it better does work.
  - This needs an extensive prototyping and testing program.