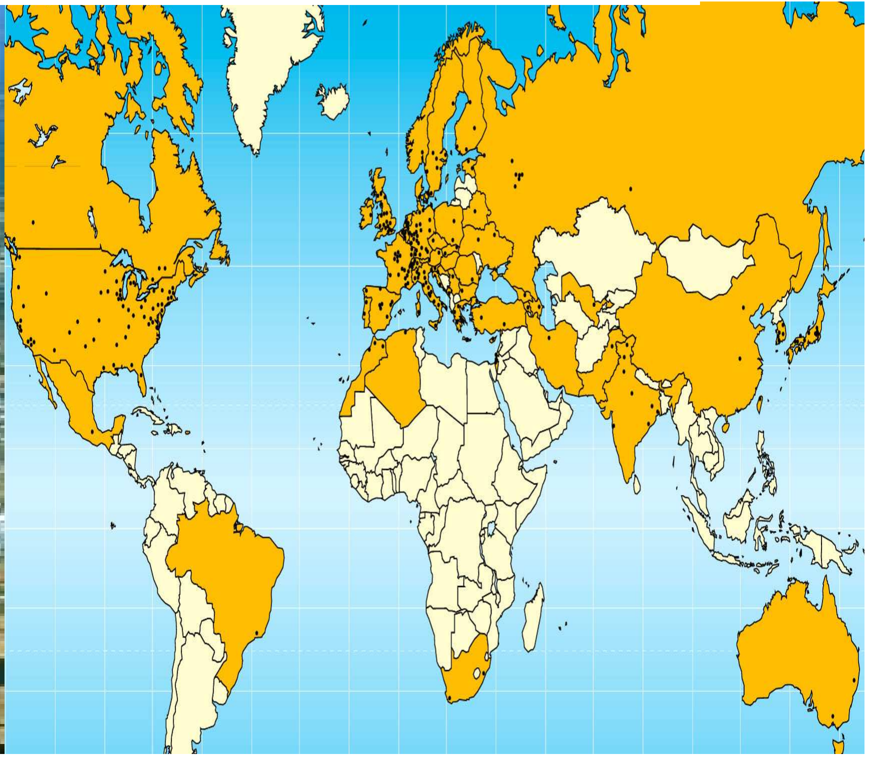
LHC Computing Model Perspective



Harvey B. Newman, Caltech

Data Analysis for Global HEP Collaborations

LCG Launch Workshop, CERN

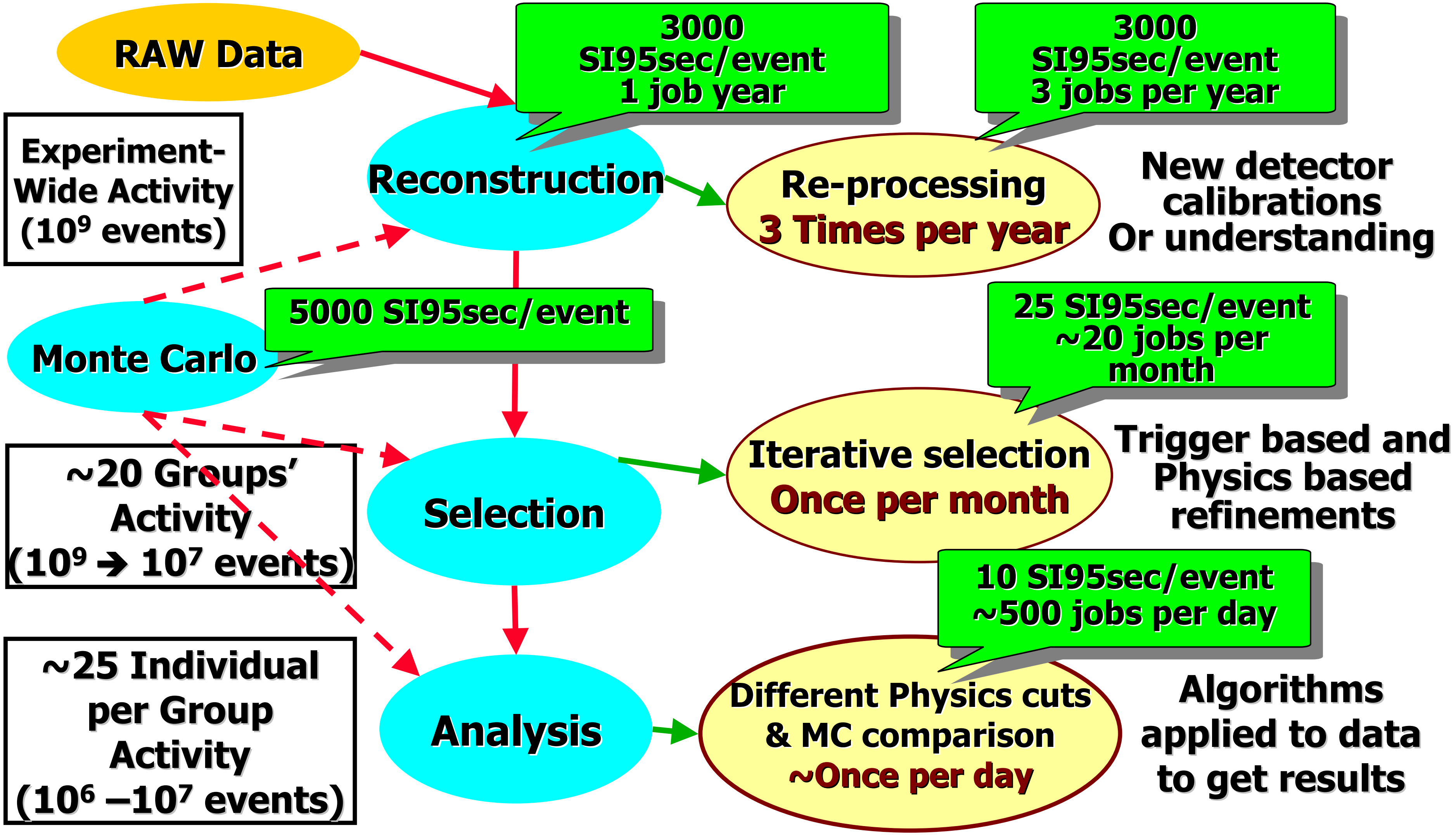
l3www.cern.ch/~newman/LHCCMPerspective_hbn031102.ppt

To Solve: the LHC “Data Problem”

- While the proposed LHC computing and data handling facilities are large by present-day standards,
  - ➔ They will not support FREE access, transport or processing for more than a minute part of the data
- **Technical Goals:** Ensure that the system is dimensioned, configured, managed and used “optimally”
- Specific Problems to be Explored. How to
  - ➔ Prioritise many hundreds of requests of local and remote communities, consistent with Collaboration policies
  - ➔ Develop Strategies to Simultaneously ensure: *Acceptable turnaround times; Efficient resource use*
  - ➔ Balance proximity to large computational and data handling facilities, against proximity to end users and more local resources (for frequently-accessed datasets)

MONARC: CMS Analysis Process

Hierarchy of Processes (Experiment, Analysis Groups,Individuals)



RAW Data

3000
SI95sec/event
1 job year

3000
SI95sec/event
3 jobs per year

Experiment-
Wide Activity
(10^9 events)

Reconstruction

Re-processing
3 Times per year

New detector
calibrations
Or understanding

Monte Carlo

5000 SI95sec/event

25 SI95sec/event
~20 jobs per
month

~20 Groups'
Activity
(10^9 ➔ 10^7 events)

Selection

Iterative selection
Once per month

Trigger based and
Physics based
refinements

10 SI95sec/event
~500 jobs per day

~25 Individual
per Group
Activity
(10^6 –10^7 events)

Analysis

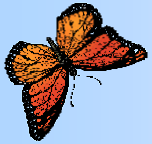
Different Physics cuts
& MC comparison
~Once per day

Algorithms
applied to data
to get results

Requirements Issues

Some significant aspects of the LHC Computing Models Need further study

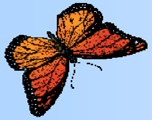
- A highly ordered analysis process: assumed relatively little re-reconstruction and event selection on demand
  - Restricted direct data flows from Tiers 0 and 1 to Tiers 3 and 4
- Efficiency of use of CPU and storage with a real workload
- Pressure to store more data
  - More data per Reconstructed Event
  - Higher DAQ recording rate
  - Simulated data: produced at many remote sites; eventually stored and accessed at CERN
- Tendency to greater CPU (as code and computers progress)
  - ~3000 SI95-sec to fully reconstruct (CMS ORCA Production)
  - To 20 SI95-sec to analyze
- B Physics: Samples of 1 to Several X 10^8 Events; MONARC CMS/ATLAS Studies assume typically 10^7 (aimed at high p_T physics)



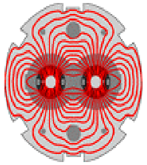
Role of Simulation for Distributed Systems

SIMULATIONS: Widely recognized as essential tools for the design, performance evaluation and optimisation of complex distributed systems

- From battlefields to agriculture; from the factory floor to telecommunications systems
- Very different from HEP "Monte Carlos"
  - ➔ "Time" intervals and interrupts are the essentials
- Simulations with an appropriate high level of abstraction are required to represent large systems with complex behavior
- Just started to be part of the HEP culture
  - ➔ Experience in trigger, online and tightly coupled computing systems: CERN CS2 models
  - ➔ *MONARC (Process-Oriented; Java Threads) Experience*
- *Simulation is vital to evaluate and optimize the LHC CM*
  - ➔ *And to design & optimise the Grid services themselves*



Some “Large” Grid Issues: to be Simulated and Studied



- Consistent transaction management
- Query (task completion time) estimation
- Queueing and co-scheduling strategies
- Load balancing (e.g. Self Organizing Neural Network)
- Error Recovery: Fallback and Redirection Strategies
- Strategy for use of tapes
- Extraction, transport and caching of physicists’ object-collections; Grid/Database Integration
- Policy-driven strategies for resource sharing among sites and activities; policy/capability tradeoffs
- Network Peformance and Problem Handling
  - ➔ Monitoring and Response to Bottlenecks
  - ➔ Configuration and Use of New-Technology Networks e.g. Dynamic Wavelength Scheduling or Switching
- **Fault-Tolerance, Performance of the Grid Services Architecture**

Transatlantic Net WG (HN, L. Price) Bandwidth Requirements [*]

	2001	2002	2003	2004	2005	2006
CMS	100	200	300	600	800	2500
ATLAS	50	100	300	600	800	2500
BaBar	300	600	1100	1600	2300	3000
CDF	100	300	400	2000	3000	6000
D0	400	1600	2400	3200	6400	8000
BTeV	20	40	100	200	300	500
DESY	100	180	210	240	270	300
US-CERN	310	622	1250	2500	5000	10000

[*] Installed BW. Maximum Link Occupancy 50% Assumed

The Network Challenge is Shared by Both Next- and Present Generation Experiments

See http://gate.hep.anl.gov/lprice/TAN

Gbps Network Issues & Challenges

Requirements for High Throughput

- *Packet Loss must be ~Zero (10^{-6} and Below for Large Flows)*
  - I.e. No "Commodity" networks
  - Need to track down packet loss
- No Local infrastructure bottlenecks
  - Gigabit Ethernet "clear paths" between selected host pairs needed now; To 10 Gbps Ethernet by ~2003 or 2004
- TCP/IP stack configuration and tuning Absolutely Required
  - Large Windows; Possibly Multiple Streams
  - New Concepts of *Fair Use* Must then be Developed
- Careful Router configuration; monitoring
  - Server and Client CPU, I/O and NIC throughput sufficient
- *End-to-end* monitoring and tracking of performance
- Close collaboration with local and "regional" network staffs

TCP Does Not Scale to the 1-10 Gbps Range

- New Technologies: Lambdas, MPLS, Lambda Switching
- Security and Firewall Performance

Tier0-Tier1 Link Requirements Estimate: Hoffmann Report 2001

◆1) Tier1 ⇆ Tier0 Data Flow for Analysis	0.5 - 1.0 Gbps
◆2) Tier2 ⇆ Tier0 Data Flow for Analysis	0.2 - 0.5 Gbps
◆3) Interactive Collaborative Sessions (30 Peak)	0.1 - 0.3 Gbps
◆4) Remote Interactive Sessions (30 Flows Peak)	0.1 - 0.2 Gbps
◆5) Individual (Tier3 or Tier4) data transfers (Limit to 10 Flows of 5 MBytes/sec each)	0.8 Gbps
TOTAL Per Tier0 - Tier1 Link	***<u>1.7 - 2.8 Gbps</u>***

◆<u>NOTE:</u>

➔ Adopted Baseline by the LHC Experiments;
Given in the Hoffmann Steering Committee Report:

- ❑ "1.5 - 3 Gbps per experiment"

Tier0-Tier1 BW Requirements Estimate: Hoffmann Report 2001

- *Scoped for 100Hz X 1 MB Data Recording (CMS and ATLAS)*
- Does Not Allow Fast Download to Tier3+4 of "Small" Object Collections
  - ➔ Example: Download 10^7 Events of AODs (10^4 Bytes Each) ➔ 100 GB; At 5 Mbytes/sec per person that's 6 Hours !
- Still a bottoms-up, static, and hence Conservative Model.
  - ➔ A Dynamic Grid system with Caching, Co-scheduling, and Pre-Emptive data movement may require greater bandwidth
  - ➔ Does Not Include "Virtual Data" operations: Derived Data Copies; DB and Data-description overheads
- Network Requirements will evolve as network technologies and prices advance

HENP Related Data Grid Projects

Projects

PPDG I	USA	DOE	$2M	1999-2001
GriPhyN	USA	NSF	$11.9M + $1.6M	2000-2005
EU DataGrid	EU	EC	€10M	2001-2004
PPDG II (CP)	*USA*	*DOE*	*$9.5M*	*2001-2004*
iVDGL	*USA*	*NSF*	*$13.7M + $2M*	*2001-2006*
DataTAG	*EU*	*EC*	*€4M*	*2002-2004*
GridPP	*UK*	*PPARC*	*>$15M*	*2001-2004*
LCG Phase1	*CERN*	*MS*	*30 MCHF*	*2002-2004*

Many Other Projects of interest to HENP

- Initiatives in US, UK, Italy, France, NL, Germany, Japan, …
- US and EU networking initiatives: *AMPATH, I2, DataTAG*
- US Distributed Terascale Facility:
  ($53M, 12 TeraFlops, 40 Gb/s network)

CMS Milestones: In Depth Design & Data Challenges 1999-2007

- Trigger (Filter) Studies: 1999-2001
- November 2000: Level 1 Trigger TDR (Completed)
  - ➔Large-scale productions for L1 trigger studies
- Dec 2002: DAQ TDR
  - ➔Continue High Level Trigger studies; Production at Prototype Tier0, Tier1s and Tier2s
- Dec 2003: Core Software and Computing TDR
  - ➔First large-scale Data Challenge (5%)
  - ➔Use full chain from online farms to production in Tier0, 1, 2 centers
- Dec 2004: Physics TDR
  - ➔Test physics performance, with large amount of data
  - ➔Verify technology choices with distributed analysis
- Dec 2004: Second large-scale Data Challenge (20%)
  - ➔Final test of scalability of the fully distributed CMS computing system before production system purchase
- Fall 2006: Computing, database and Grid systems in place. Commission for LHC Startup
- Apr. 2007: All Systems Ready for First LHC Runs

The LHC Distributed Computing Model: from Here Forward

Ongoing Study of the Model: Evolving with Experience and Advancing Technologies

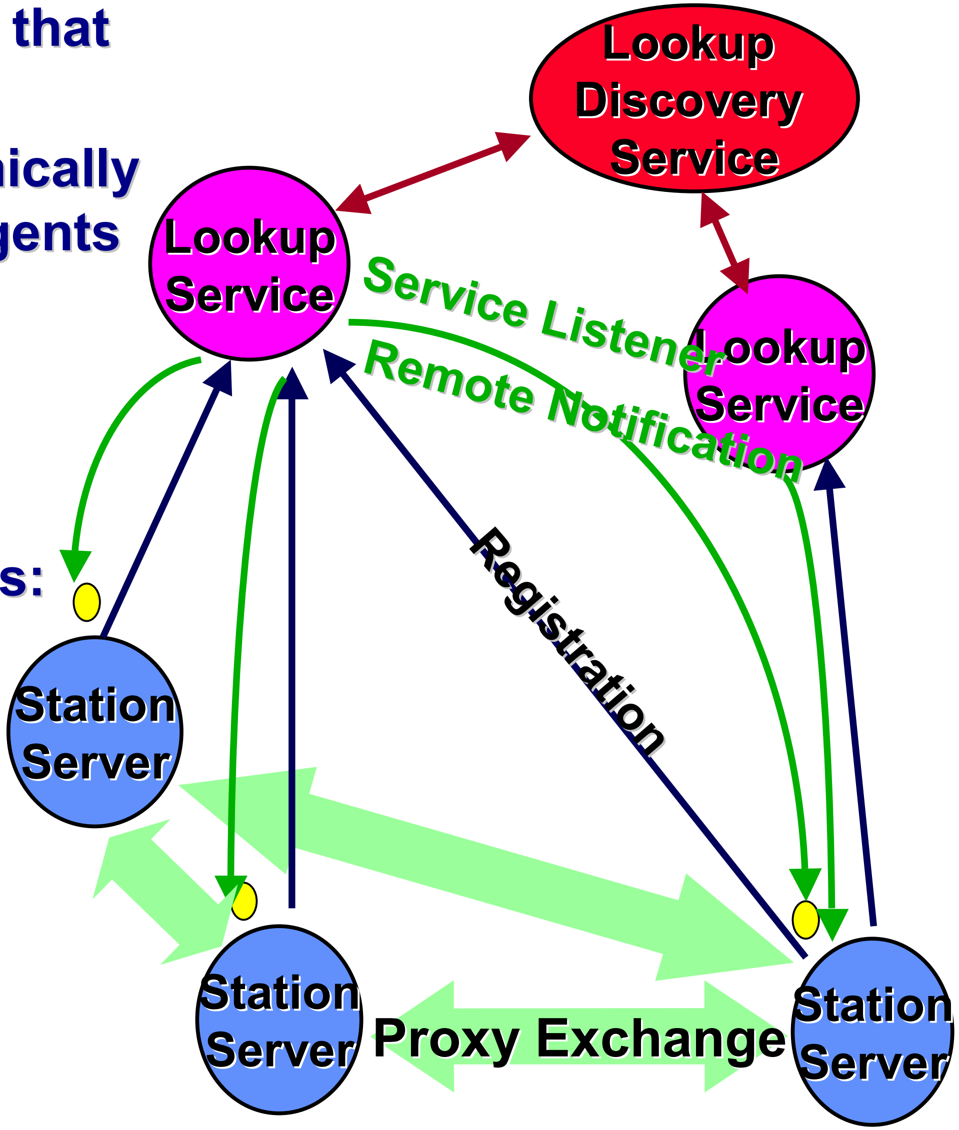
- Requirements
- Site components and architectures
- Networks: technology, scale, operations
- High Level Software Services architecture:
  - ❑ Scalable and resilient ➔ loosely coupled, adaptive, partly autonomous, e.g. agent-based
- Operational Modes (Develop a Common Understanding ?)
  - ❑ What are the technical goals + emphasis of the system *How is it intended to be used by the Collaboration ?*
  - ❑ e.g. What are guidelines and steps that make up the data access/processing/analysis policy and strategy

Note: Common services imply somewhat similar op. modes

Agent-Based Distributed Services: JINI Prototype (Caltech/Pakistan)

- Includes "Station Servers" (static) that host mobile "Dynamic Services"
- Servers are interconnected dynamically to form a fabric in which mobile agents travel, with a payload of physics analysis tasks
- Prototype is highly flexible and robust against network outages
- Adaptable to WSDL-based services: OGSA; and to many platforms
- The Design and Studies with this prototype use the MONARC Simulator, and build on SONN studies. See

http://home.cern.ch/clegrand/lia/



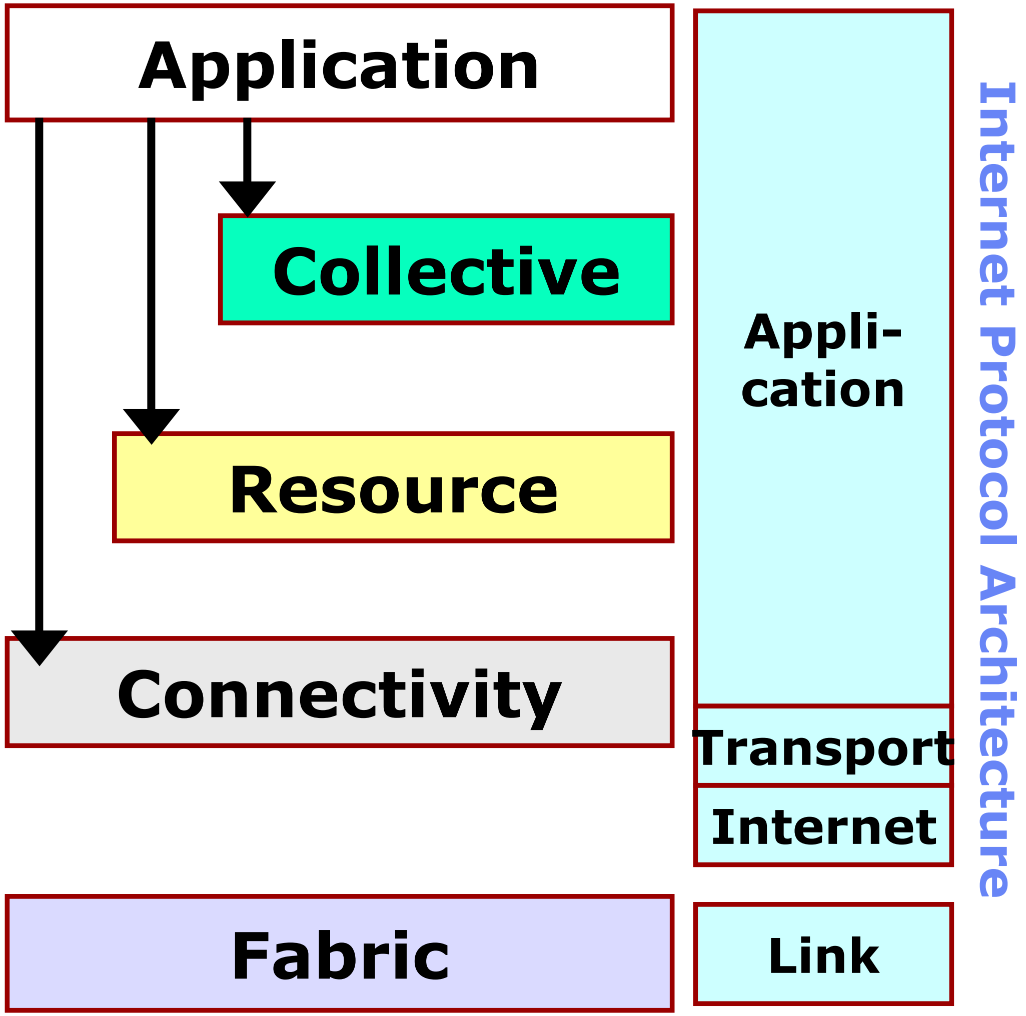
LHC Distributed CM: HENP Data Grids Versus Classical Grids

- Grid projects have been a step forward for HEP and LHC: a path to meet the “LHC Computing” challenges
  - But: the differences between HENP Grids and classical Grids are not yet fully appreciated
- The original Computational and Data Grid concepts are largely stateless, open systems: known to be scalable
  - ➔ Analogous to the Web
- The classical Grid architecture has a number of implicit assumptions
  - ➔ The ability to locate and schedule suitable resources, within a tolerably short time (i.e. resource richness)
  - ➔ Short transactions; Relatively simple failure modes
- HEP Grids are data-intensive and resource constrained
  - ➔ Long transactions; some long queues
  - ➔ Schedule conflicts; policy decisions; task redirection
  - ➔ A Lot of global system state to be monitored+tracked

Upcoming Grid Challenges: Secure Workflow Management and Optimization

- Maintaining a *Global View* of Resources and System State
  - End-to-end System Monitoring
  - Adaptive Learning: new paradigms for execution optimization (eventually automated)
- *Workflow Management,* Balancing Policy Versus Moment-to-moment Capability to Complete Tasks
  - Balance High Levels of Usage of Limited Resources Against Better Turnaround Times for Priority Jobs
  - *Goal-Oriented; Steering* Requests According to (Yet to be Developed) Metrics
- Robust Grid Transactions In a Multi-User Environment
- Realtime Error Detection, Recovery
  - Handling User-Grid Interactions: Guidelines; Agents
- Building Higher Level Services, and an Integrated User Environment for the Above

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

More info: www.globus.org/research/papers/anatomy.pdf

HENP Grid Architecture: Layers Above the Collective Layer

- Physicists' Application Codes
  - Reconstruction, Calibration, Analysis
- Experiments' Software Framework Layer
  - Modular and Grid-aware: Architecture able to interact effectively with the lower layers (above)
- Grid Applications Layer

  (Parameters and algorithms that govern system operations)
  - Policy and priority metrics
  - Workflow evaluation metrics
  - Task-Site Coupling proximity metrics
- Global End-to-End System Services Layer
  - Monitoring and Tracking Component performance
  - Workflow monitoring and evaluation mechanisms
  - Error recovery and redirection mechanisms
  - System self-monitoring, evaluation and optimisation mechanisms

The Evolution of Global Grid Standards

- GGF4 (Feb. 2002): Presentation of the OGSA (Draft) See http://www.globus.org/research/papers/ogsa.pdf
  - ☐ Uniform Grid Services are defined
  - ☐ Defines standard mechanisms for creating, naming and discovering transient Grid services
  - ☐ Defines Web-service (WSDL) interfaces, conventions and mechanisms to build the basic services
    - ➔ As required for composing sophisticated distributed systems
  - ☐ Expresses the intent to provide higher level standard services: for distributed data management; workflow; auditing; instrumentation and monitoring; problem determination for distributed computing, security protocol mapping
- Adoption of the Web-services approach by a broad range of major industrial players, most notably IBM

The Evolution of Grid Standards and the LHC/HENP Grid Task

- The emergence of a standard Web-services based architecture (OGSA) is a major step forward
- But we have to consider a number of practical factors:
  - ☐ Schedule of Emerging Standards relative to the LHC Experiments' Schedule and Milestones
  - ☐ Availability and functionality of standard services as a function of time
  - ☐ Extent and scope of the standard services
    - *Basic services will be standardized*
    - *Industry will compete over tools and higher level services built on top of the basic services*
    - *Major vendors are not in the business of vertically integrated applications (for the community)*
- Question at GGF4: Who builds the distributed system, with sufficient intelligence and functionality to meet our needs ?
  - ☐ Answer: *You Do.*

The LHC "Computing Problem" and Grid R&D/Deployment Strategy

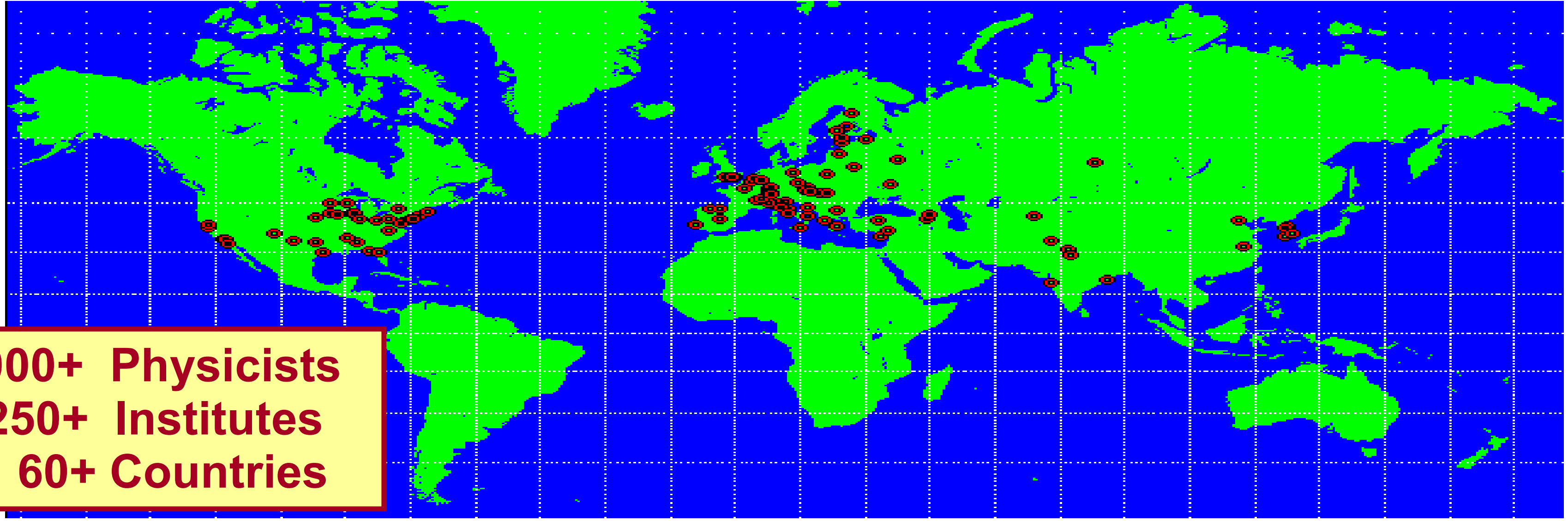
- Focus on End-to-End integration and deployment of experiment applications with existing and emerging Grid services
    - *Including the E2E and Grid Applications Layers*
- Collaborative development of Grid middleware and extensions between application and middleware groups
    - Leading to pragmatic and acceptable-risk solutions
- Grid technologies and services need to be deployed in production (24x7) environments
    - Meeting experiments' Milestones
    - With stressful performance needs
    - Services that work; increasing functionality at each stage

  as an integral part of the development process
- *We need to adopt common basic security and information infrastructures, and basic components soon*
- *Move on to tackle the LHC "Computing Problem" as a whole*
    - *Develop the network-distributed data analysis and collaborative systems*
    - *To meet the needs of the global LHC Collaborations*

Some Extra Slides Follow

Computing Challenges: Petabyes, Petaflops, Global VOs

- ➔ **Geographical dispersion:** of people and resources
- ➔ **Complexity:** the detector and the LHC environment
- ➔ **Scale:** Tens of Petabytes per year of data



5000+ Physicists
250+ Institutes
60+ Countries

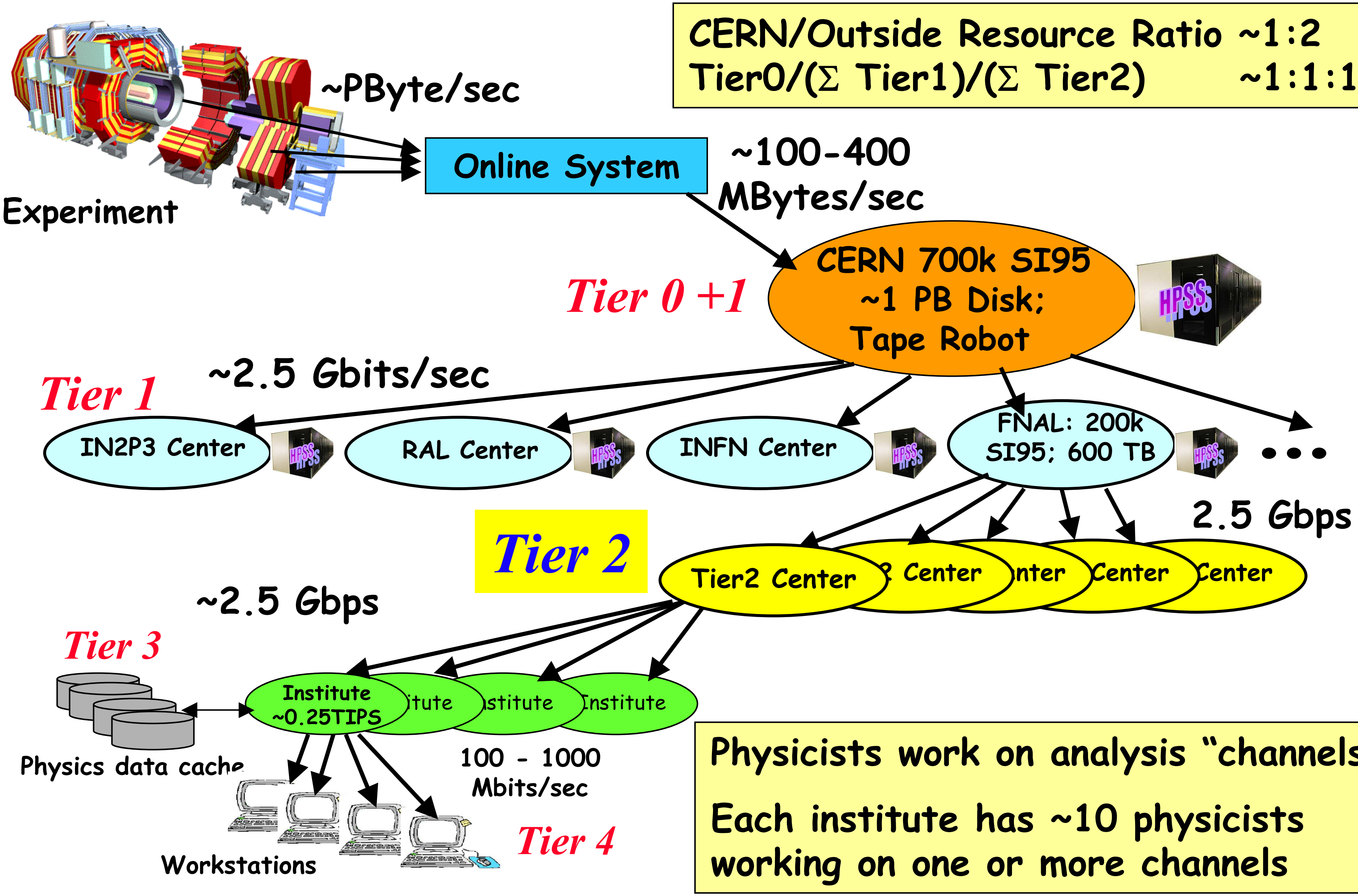
Major challenges associated with:

Communication and collaboration at a distance
Managing globally distributed computing & data resources
Remote software development and physics analysis
R&D: New Forms of Distributed Systems: Data Grids

LHC Data Grid Hierarchy

CERN/Outside Resource Ratio ~1:2

Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1



Physicists work on analysis "channels"

Each institute has ~10 physicists working on one or more channels

Why Worldwide Computing? Regional Center Concept

- Maximize total funding resources to meet the total computing and data handling needs
- An N-Tiered Model: for fair-shared access for Physicists everywhere
  - ➔ Smaller size, greater control as N increases
- Utilize all intellectual resources, & expertise in *all time zones*
  - ➔ Involving students and physicists at home universities and labs
- Greater flexibility to pursue different physics interests, priorities, and resource allocation strategies by region
  - ➔ And/or by Common Interest: physics topics, subdetectors,…
- Manage the System’s Complexity
  - ➔ Partitioning facility tasks, to manage & focus resources
- Efficient use of network: higher throughput
  - ➔ Per Flow: Local > regional > national > international

MONARC: Project at CERN

Models Of Networked Analysis
At Regional Centers

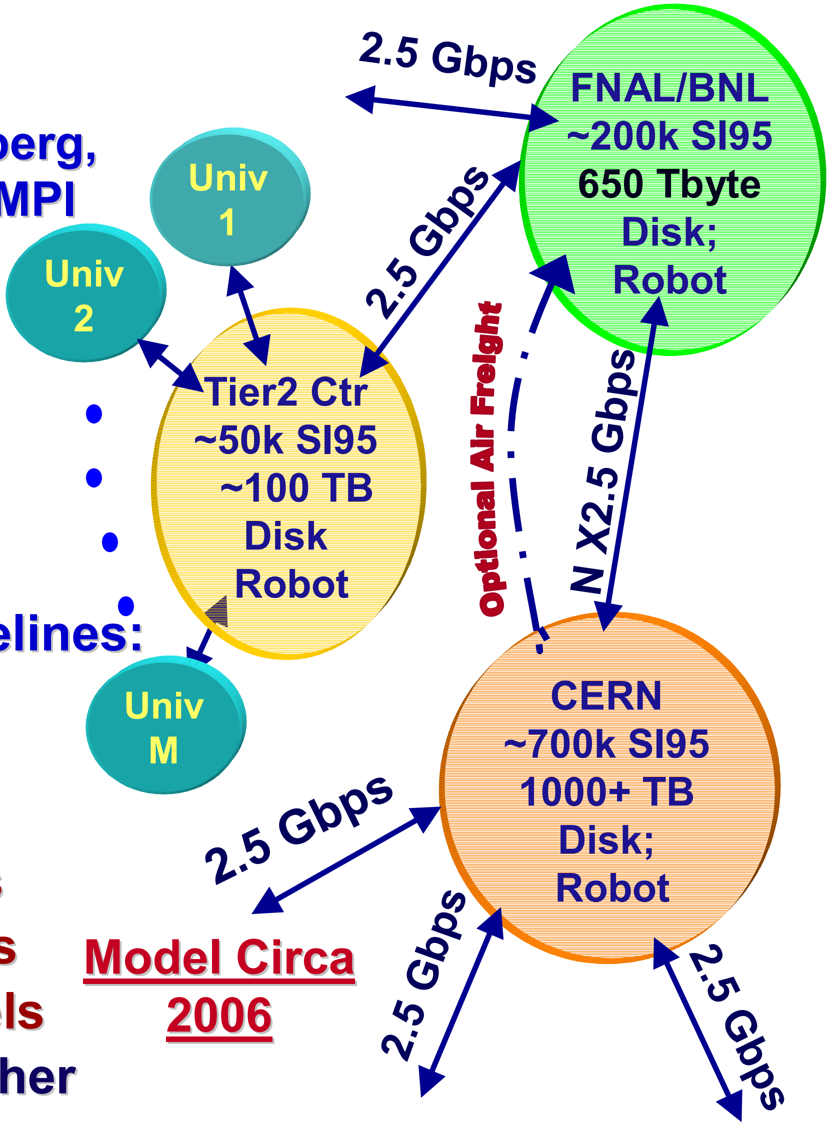
Caltech, CERN, Columbia, FNAL, Heidelberg, Helsinki, INFN, IN2P3, KEK, Marseilles, MPI Munich, Orsay, Oxford, Tufts

PROJECT GOALS ACHIEVED

- ➔ Developed LHC "Baseline Models"
- ➔ Specified the main parameters characterizing the Model's performance: throughputs, latencies
- ➔ Established resource requirement baselines: Computing, Data handling, Networks

TECHNICAL GOALS

- ➔ Defined the baseline Analysis Process
- ➔ Defined RC Architectures and Services
- ➔ Provided Guidelines for the final Models
- ➔ Provided a Simulation Toolset for Further Model studies



FNAL/BNL
~200k SI95
650 Tbyte
Disk;
Robot

Univ
1

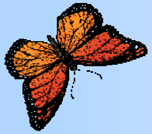
Univ
2

Tier2 Ctr
~50k SI95
~100 TB
Disk
Robot

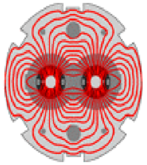
Univ
M

CERN
~700k SI95
1000+ TB
Disk;
Robot

Model Circa 2006



MONARC History



Spring 1998	First Distributed Center Models (Bunn; Von Praun)
6/1998	Presentation to LCB; Project Assignment Plan
Summer 1998	MONARC Project Startup (ATLAS, CMS, LHCb)
9 - 10/1998	Project Execution Plan; Approved by LCB
1/1999	First Analysis Process to be Modeled
2/1999	First Java Based Simulation Models (I. Legrand)
Spring 1999	Java2 Based Simulations; GUI
4/99; 8/99; 12/99	Regional Centre Representative Meetings
6/1999	Mid-Project Progress Report Including MONARC Baseline Models
9/1999	Validation of MONARC Simulation on Testbeds Reports at LCB Workshop (HN, I. Legrand)
1/2000	Phase 3 Letter of Intent (4 LHC Experiments)
2/2000	Papers and Presentations at CHEP2000: D385, F148, D127, D235, C113, C169
3/2000	Phase 2 Report
Spring 2000	New Tools: SNMP-based Monitoring; S.O.M.
5/2000	Phase 3 Simulation of ORCA4 Production; Begin Studies with Tapes
Spring 2000	MONARC Model Recognized by Hoffmann WWC Panel; Basis of Data Grid Efforts in US and Europe

MONARC Key Features for a Successful Project

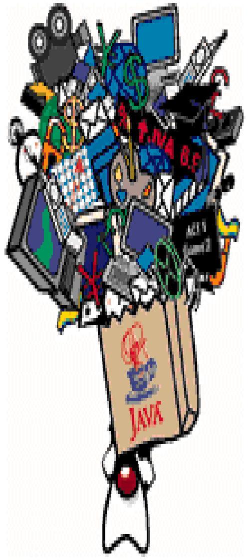
- The broad based nature of the collaboration: LHC experiments, regional representatives, covering different local conditions and a range of estimated financial means
- The choice of the process-oriented discrete event simulation approach backed up by testbeds, allowing to simulate accurately
    - ➔ a complex set of networked Tier0/Tier1/Tier2 Centres
    - ➔ the analysis process: a dynamic workload of reconstruction and analysis jobs submitted to job schedulers, and then to multi-tasking compute and data servers
    - ➔ the behavior of key elements of the system, such as distributed database servers and networks
- The design of the simulation system, with an appropriate level of abstraction, allowing it to be CPU and memory-efficient
- The use of prototyping on the testbeds to ensure the simulation is capable of providing accurate results
- Organization into four technical working groups
- Incorporation of the Regional Centres Committee

"MONARC" Simulations and LHC CM Development

- ***Major Steps***
  - ❑ Conceptualize, profile and parameterize workloads and their time-behaviors
  - ❑ Develop and parameterize schemes for task prioritization, coupling tasks to sites
  - ❑ Simulate individual Grid services & transaction behavior
  - ❑ Develop/test error recovery and fallback strategies
    - ➔ Handle an increasingly rich set of "situations" (failures) as the Grid system and workload scales
- Learn from experiments' Data Challenge Milestones
- Also study: Grid-Enabled User Analysis Environments

Design Considerations of the MONARC Simulation System

F148



- This simulation project is based on Java2(TM) technology which provides adequate tools for developing a flexible and distributed process oriented simulation. Java has built-in multi-thread support for concurrent processing, which can be used for simulation purposes by providing a dedicated scheduling mechanism.
- The distributed objects support (through RMI or CORBA) can be used on distributed simulations, or for an environment in which parts of the system are simulated and interfaced through such a mechanism with other parts which actually are running the real application.

A *PROCESS ORIENTED APPROACH* for discrete event simulation is well-suited to describe concurrent running tasks

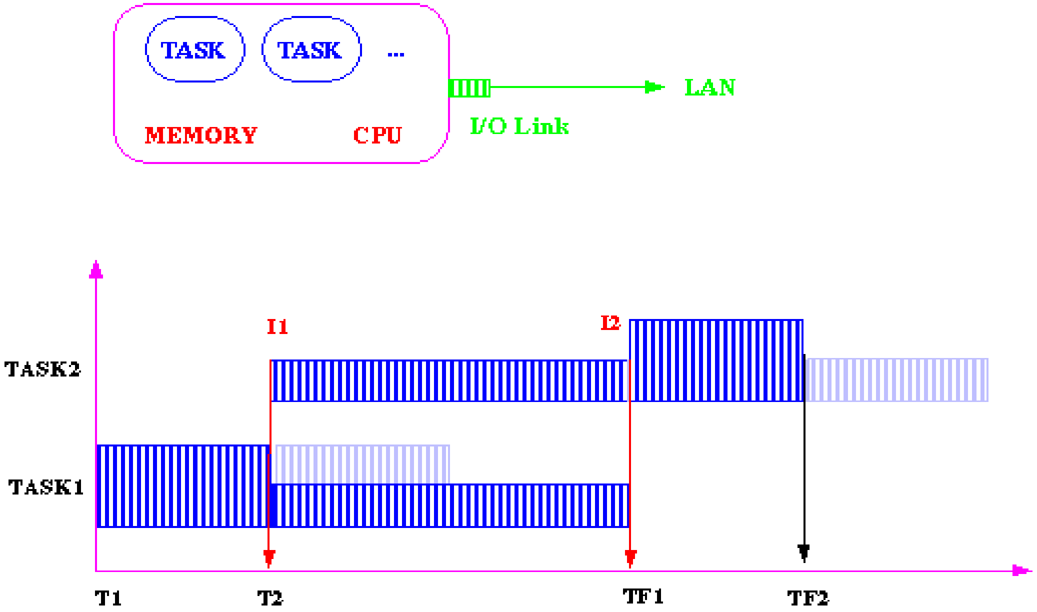
& "Active objects" (having an execution thread, a program counter, stack...) provide an easy way to map the structure of a set of distributed running programs into the simulation environment.

Multitasking Processing Model

➔ Assign active tasks (CPU, I/O, network) to Java threads
➔ Concurrent running tasks share resources (CPU, memory, I/O)

"Interrupt" driven scheme:

For each new task or when one task is finished, an interrupt is generated and all "times to completion" are recomputed.



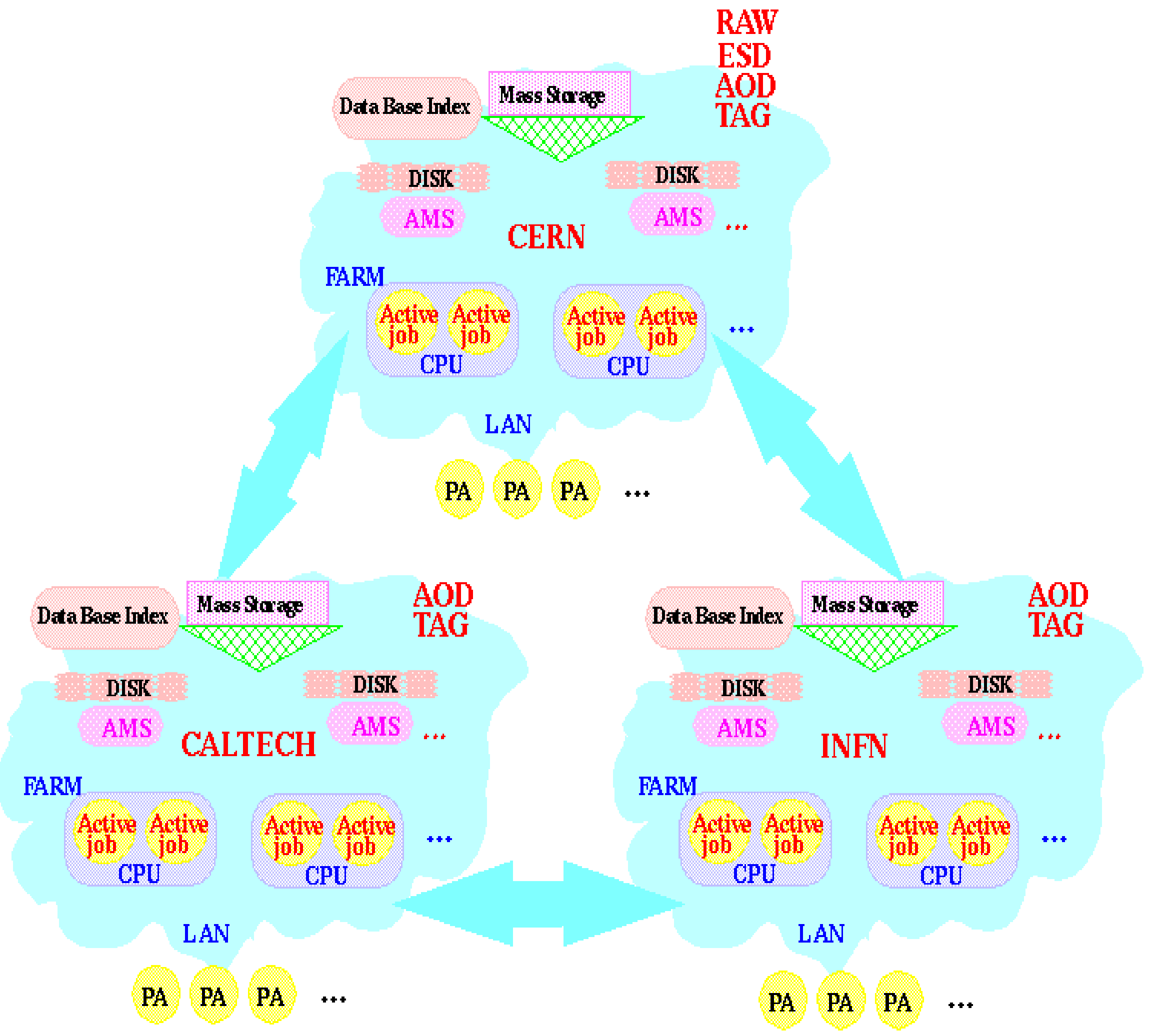
It provides:

An efficient mechanism to simulate multitask processing

An easy way to apply different load balancing schemes

Example : Physics Analysis at Regional Centres

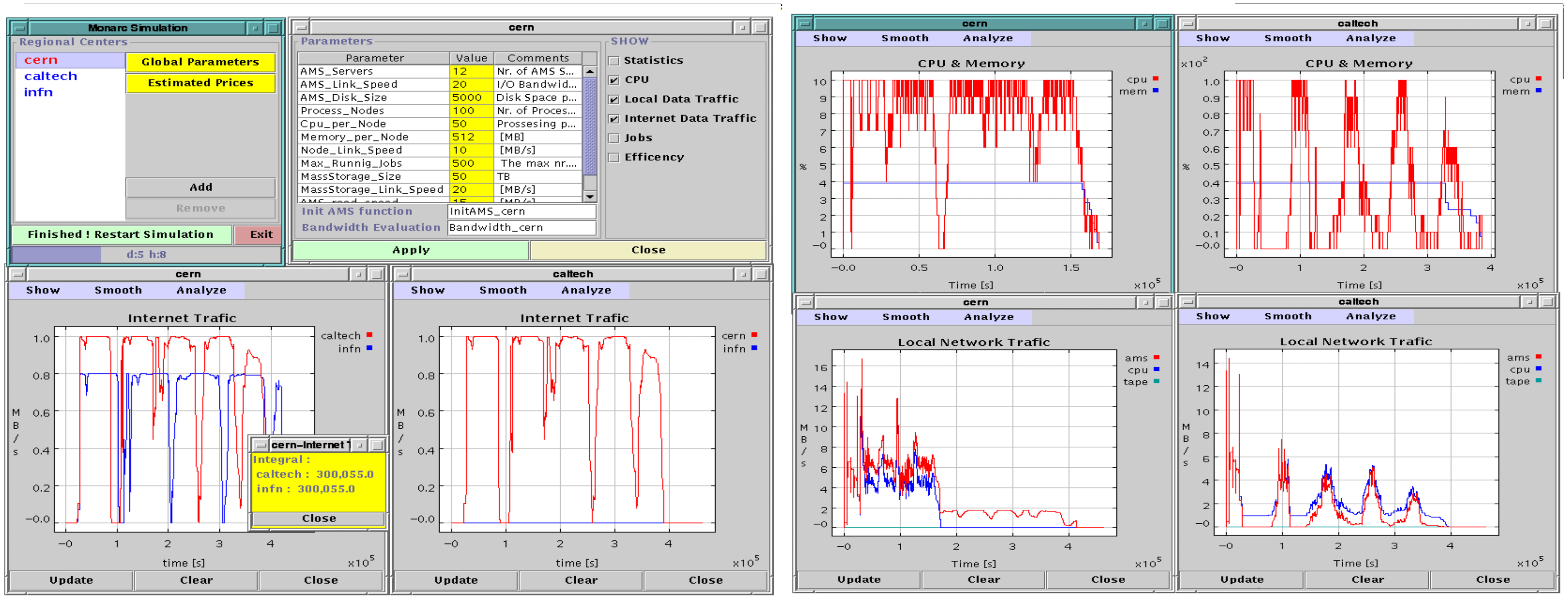
- Similar data processing jobs are performed in each of several RCs
- There is profile of jobs, each submitted to a job scheduler
- Each Centre has “TAG” and “AOD” databases replicated.
- Main Centre provides “ESD” and “RAW” data
- Each job processes AOD data, and also a a fraction of ESD and RAW data.



Modeling and Simulation: MONARC System

- ➢ **Modelling and understanding networked regional center configurations, their performance and limitations, is essential for the design of large scale distributed systems.**
- ❖ **The simulation system developed in MONARC (Models Of Networked Analysis At Regional Centers), based on a process oriented approach to discrete event simulation using Java$^{(TM)}$ technology, provides a scalable tool for realistic modelling of large scale distributed systems.**

SIMULATION of Complex Distributed Systems



Monarc Simulation

Regional Centers: cern, caltech, infn

cern — Parameters

Parameter	Value	Comments
AMS_Servers	12	Nr. of AMS S...
AMS_Link_Speed	20	I/O Bandwid...
AMS_Disk_Size	5000	Disk Space p...
Process_Nodes	100	Nr. of Proces...
Cpu_per_Node	50	Prossesing p...
Memory_per_Node	512	[MB]
Node_Link_Speed	10	[MB/s]
Max_Runnig_Jobs	500	The max nr....
MassStorage_Size	50	TB
MassStorage_Link_Speed	20	[MB/s]

Init AMS function: InitAMS_cern

Bandwidth Evaluation: Bandwidth_cern

SHOW
- ☐ Statistics
- ☑ CPU
- ☑ Local Data Traffic
- ☑ Internet Data Traffic
- ☐ Jobs
- ☐ Efficency

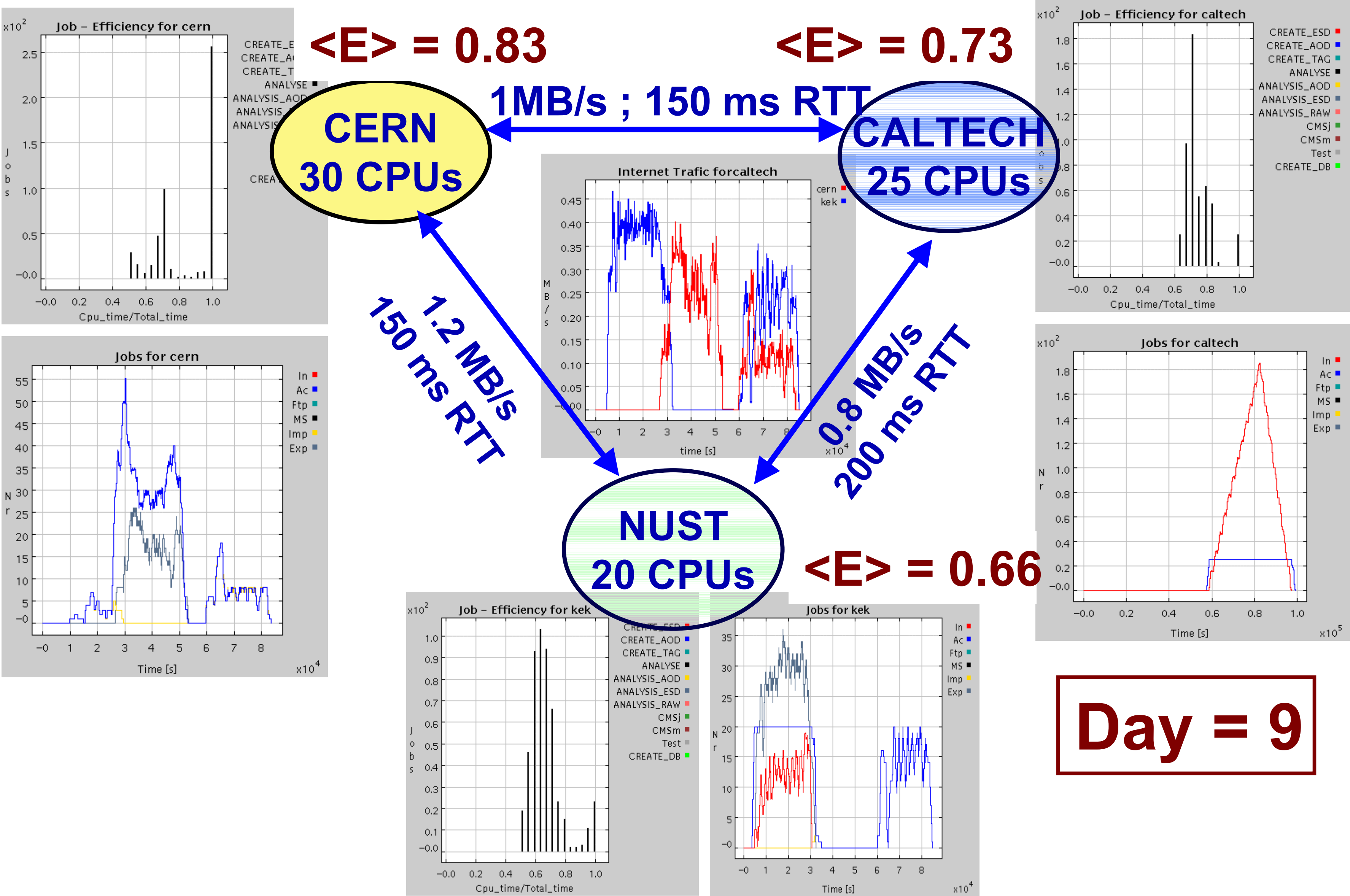
Finished ! Restart Simulation

d:5 h:8

cern-Internet T

Integral :
caltech : 300,055.0
infn : 300,055.0

MONARC SONN: 3 Regional Centres Learning to Export Jobs (Day 9)



CERN
30 CPUs

<E> = 0.83

CALTECH
25 CPUs

<E> = 0.73

NUST
20 CPUs

<E> = 0.66

1MB/s ; 150 ms RTT

1.2 MB/s
150 ms RTT

0.8 MB/s
200 ms RTT

Day = 9

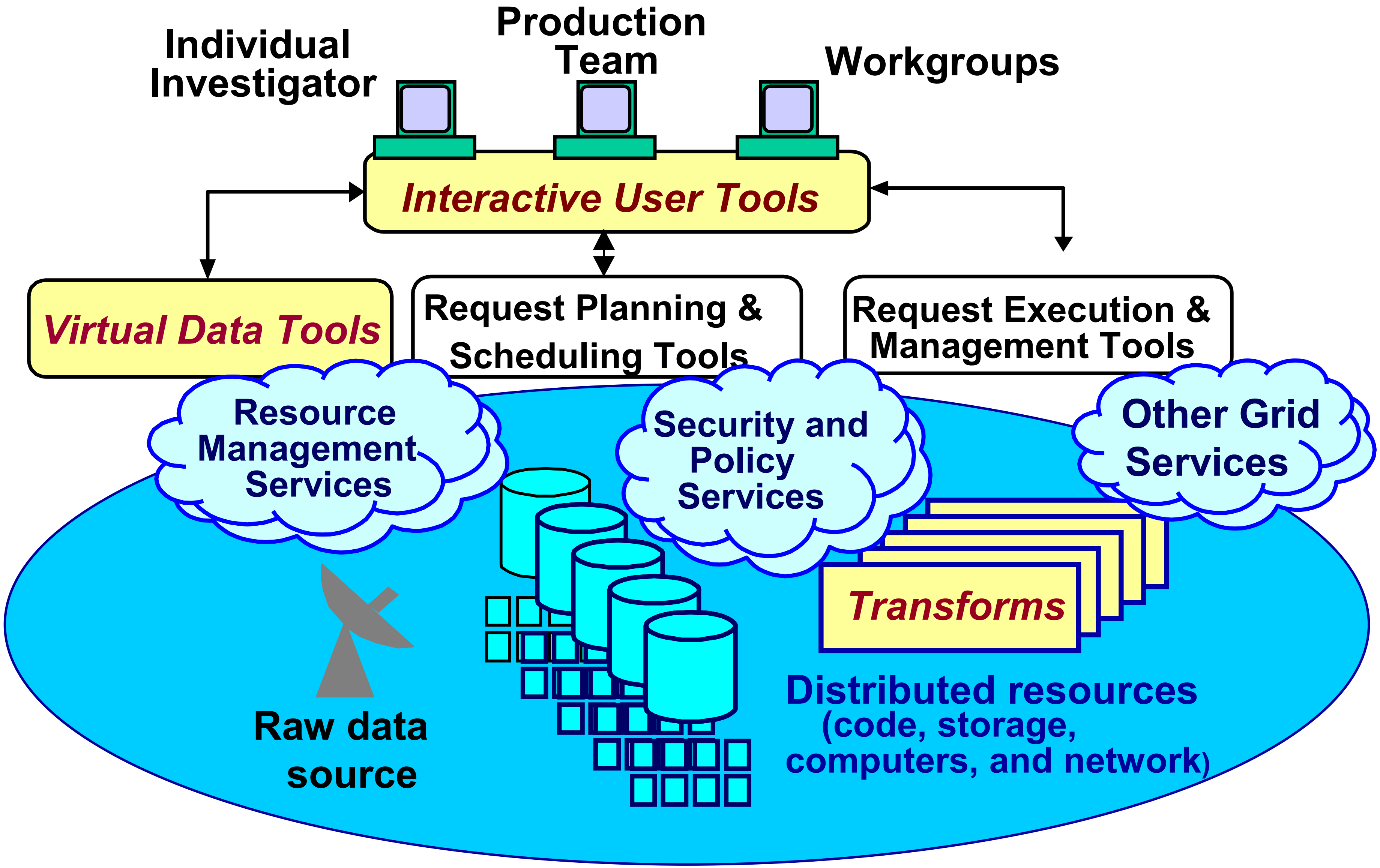
Links Required to US Labs and Transatlantic [*]

	2001	2002	2003	2004	2005	2006
SLAC	OC12	2 X OC12	2 X OC12	OC48	OC48	2 X OC48
BNL	OC12	2 X OC12	2 X OC12	OC48	OC48	2 X OC48
FNAL	OC12	OC48	2 X OC48	OC192	OC192	2 X OC192
US-CERN	2 X OC3	OC12	2 X OC12	OC48	2 X OC48	OC192
US-DESY	OC3	2 X OC3	2 X OC3	2 X OC3	2 X OC3	OC12

[*] Maximum Link Occupancy 50% Assumed

May Indicate N X OC192 Required Into CERN By 2007

GriPhyN: PetaScale Virtual Data Grids



Individual Investigator

Production Team

Workgroups

Interactive User Tools

Virtual Data Tools

Request Planning & Scheduling Tools

Request Execution & Management Tools

Resource Management Services

Security and Policy Services

Other Grid Services

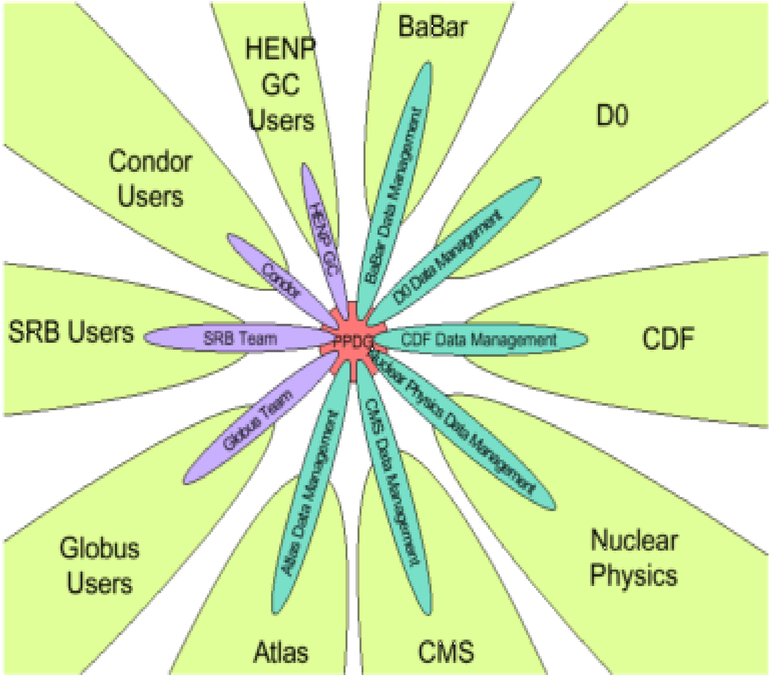
Transforms

Raw data source

Distributed resources (code, storage, computers, and network)

Particle Physics Data Grid Collaboratory Pilot (2001-2003)

"The PPDG Collaboratory Pilot will develop, evaluate and deliver vitally needed Grid-enabled tools for data-intensive collaboration in particle and nuclear physics. Novel mechanisms and policies will be vertically integrated with Grid Middleware, experiment-specific applications and computing resources to provide effective end-to-end capability."



Computer Science Program of Work

- ❐ **CS1: Job Description Language**
- ❐ **CS2: Schedule and Manage Data Processing and Placement Activities**
- ❐ **CS3 Monitoring and Status Reporting**
- ❐ **CS4 Storage Resource Management**
- ❐ **CS5 Reliable Replication Services**
- ❐ **CS6 High Performance Robust File Transfer Services**
- ❐ **CS7 Collect/Document Current Experiment Practices and Potential Generalizations…**
- ❐ **CS9 Authent., Authorization, Security**
- ❐ **CS10 End-to-End Apps. & Testbeds**

PPDG: Focus and Foundations

- **TECHNICAL FOCUS:** *End-to-End Applications & Integrated Production Systems,* With
  - Robust Data Replication
  - Intelligent Job Placement and Scheduling
  - Management of Storage Resources
  - Monitoring and Information Global Services
- **METHODOLOGY: Deploy Systems Useful to the Experiments**
  - In 24 X 7 Production Environments, with Stressful Requirements
  - With Increasing Functionality at Each Round
- **STANDARD Grid Middleware Components Integrated as they Emerge**

CMS Production: Event Simulation and Reconstruction

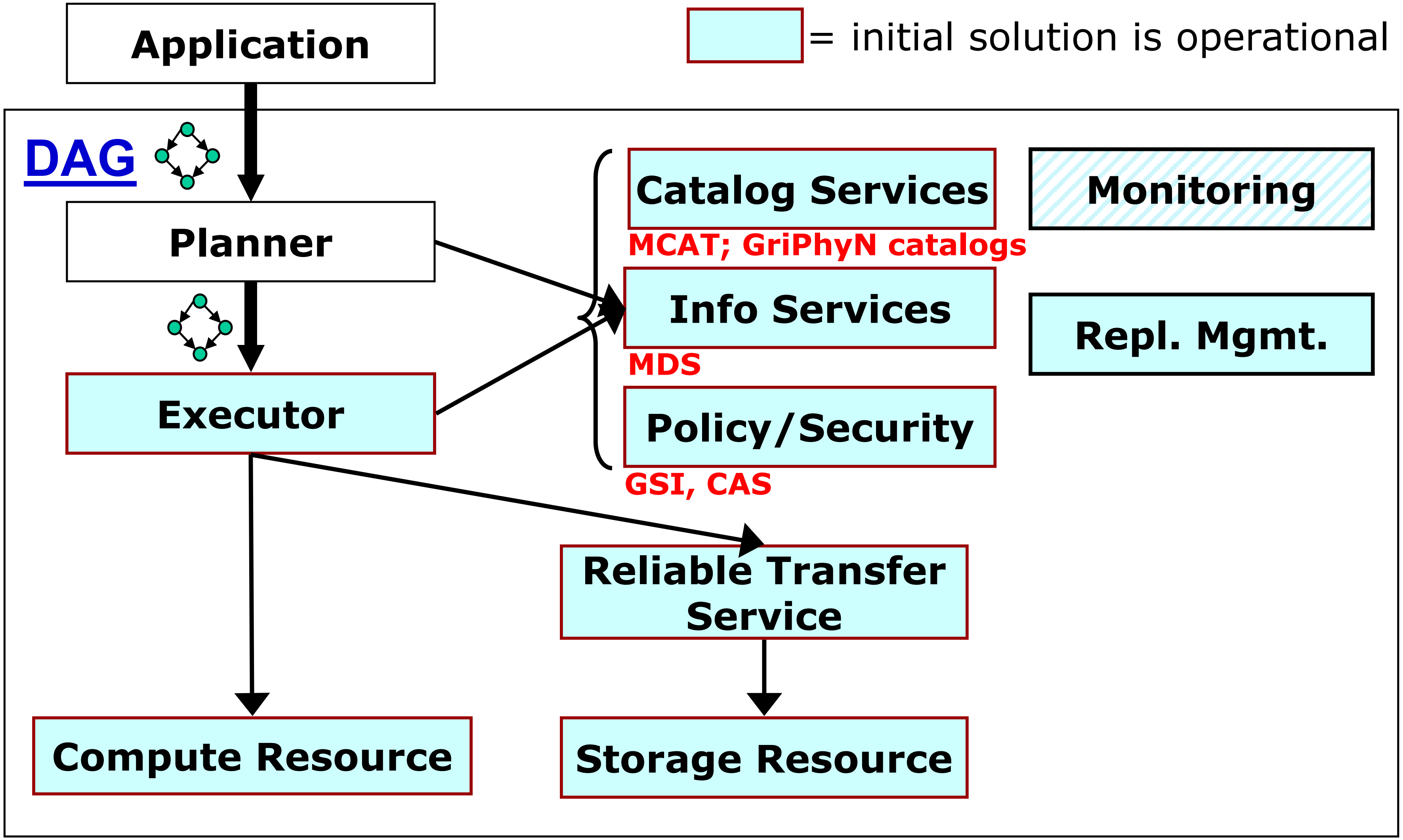
	Simulation	Digitization		GDMP	Common Prod. tools (IMPALA)
		No PU	PU		
CERN	*Fully operational*			✓	✓
FNAL				✓	✓
Moscow				✓	In progress
INFN				✓	✓
Caltech				✓	✓
UCSD				✓	✓
UFL				✓	✓
Imperial College				✓	✓
Bristol				✓	✓
Wisconsin				✓	✓
IN2P3				✓	✓
Helsinki				✓	✓

Worldwide Production at 12 Sites

"Grid-Enabled" **Automated**

GriPhyN/PPDG Data Grid Architecture

🟦 = initial solution is operational



Application

DAG

Planner

Executor

Catalog Services

MCAT; GriPhyN catalogs

Info Services

MDS

Policy/Security

GSI, CAS

Monitoring

Repl. Mgmt.

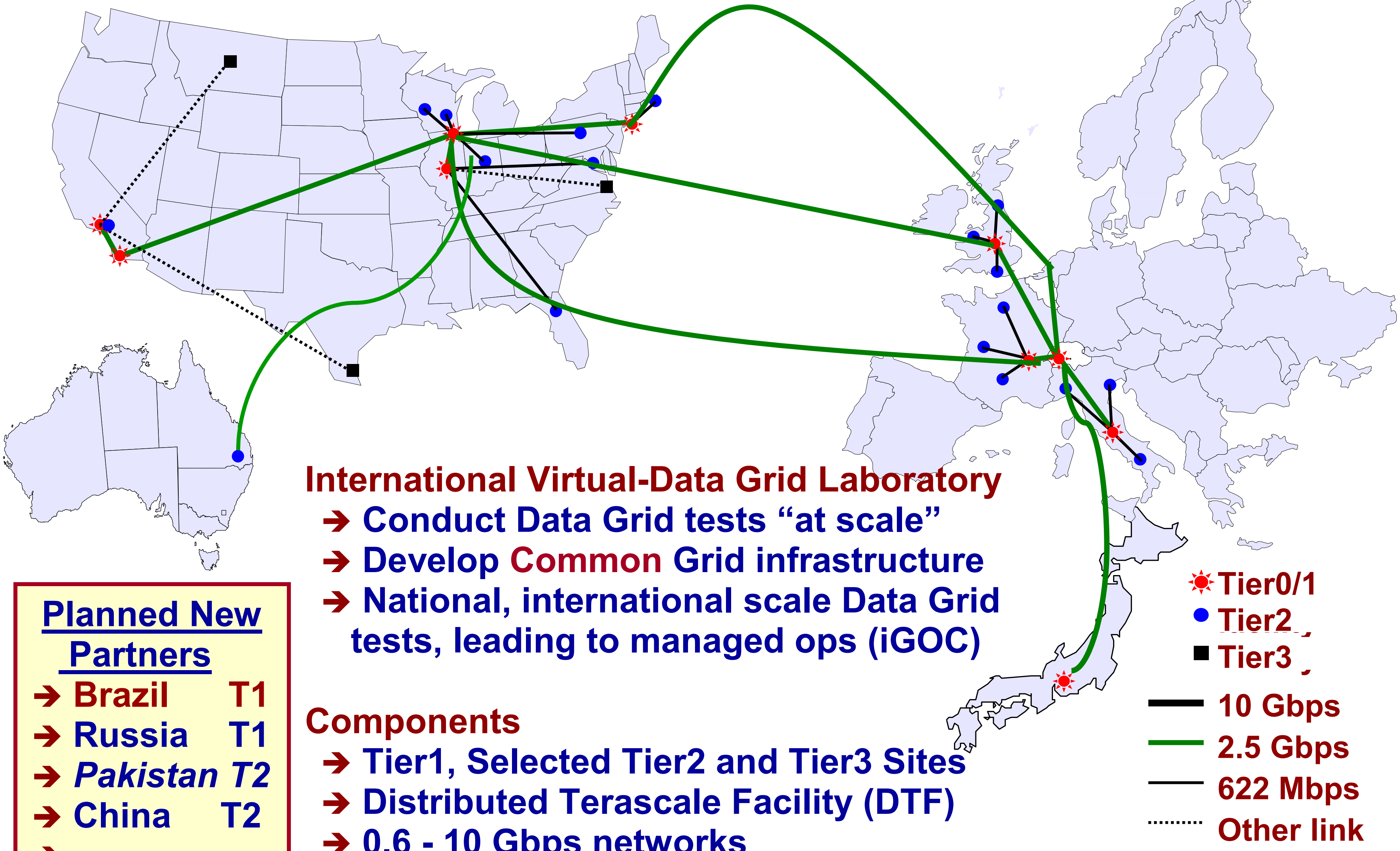
Reliable Transfer Service

Compute Resource

Storage Resource

Ian Foster, Carl Kesselman, Miron Livny, Mike Wilde, others

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

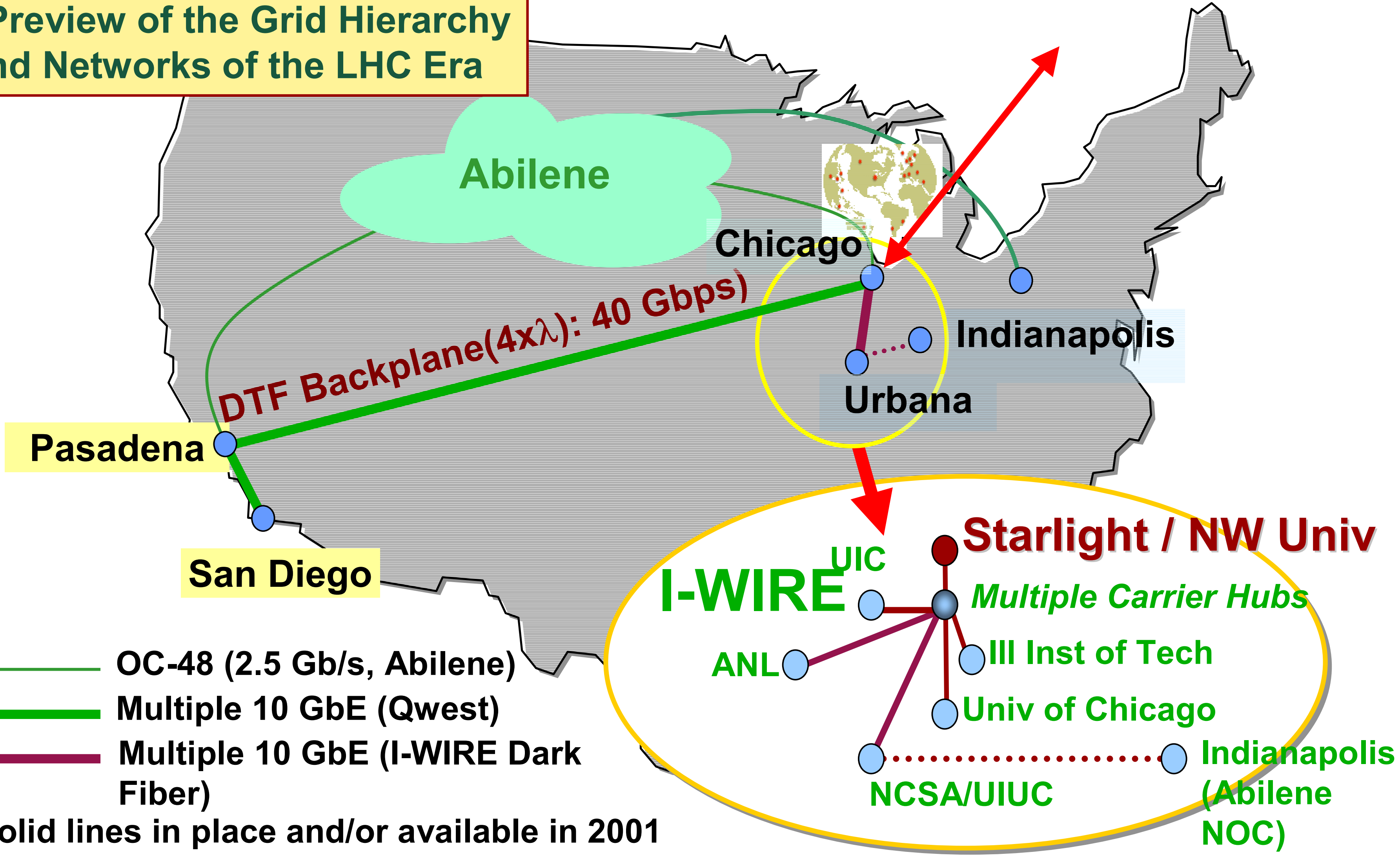
- ➔ Tier1, Selected Tier2 and Tier3 Sites
- ➔ Distributed Terascale Facility (DTF)
- ➔ 0.6 - 10 Gbps networks

Planned New Partners

- ➔ Brazil T1
- ➔ Russia T1
- ➔ *Pakistan T2*
- ➔ China T2
- ➔ …

TeraGrid (www.teragrid.org)
NCSA, ANL, SDSC, Caltech

A Preview of the Grid Hierarchy and Networks of the LHC Era



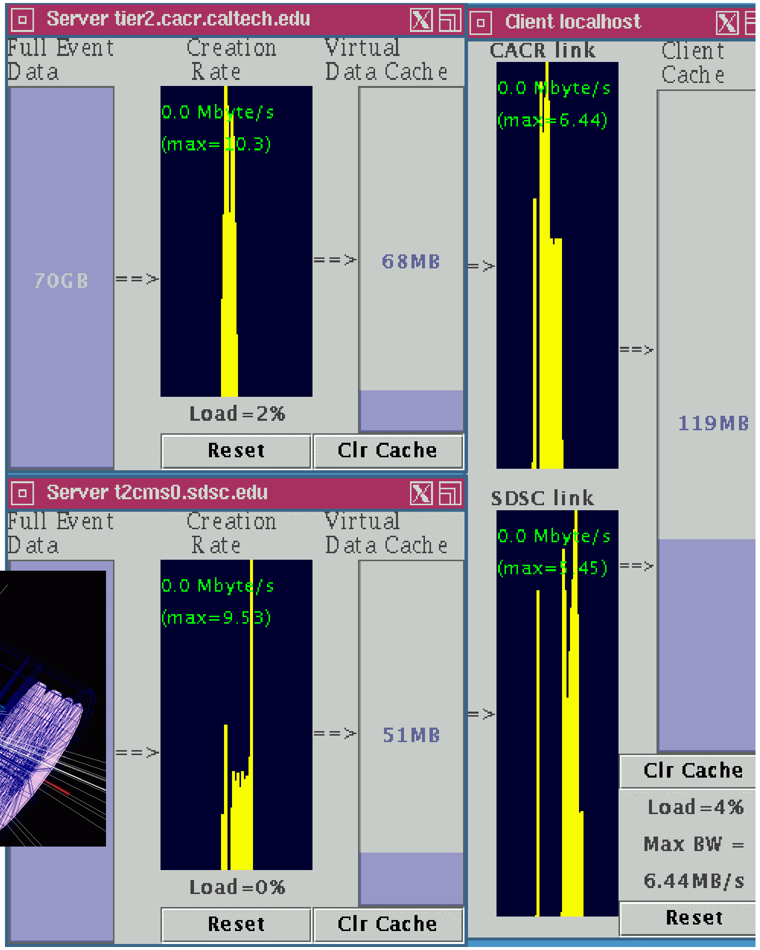
- ▬ OC-48 (2.5 Gb/s, Abilene)
- ▬ Multiple 10 GbE (Qwest)
- ▬ Multiple 10 GbE (I-WIRE Dark Fiber)

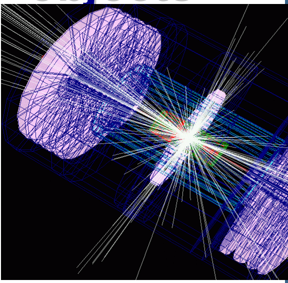
- Solid lines in place and/or available in 2001
- Dashed I-WIRE lines planned for Summer 2002

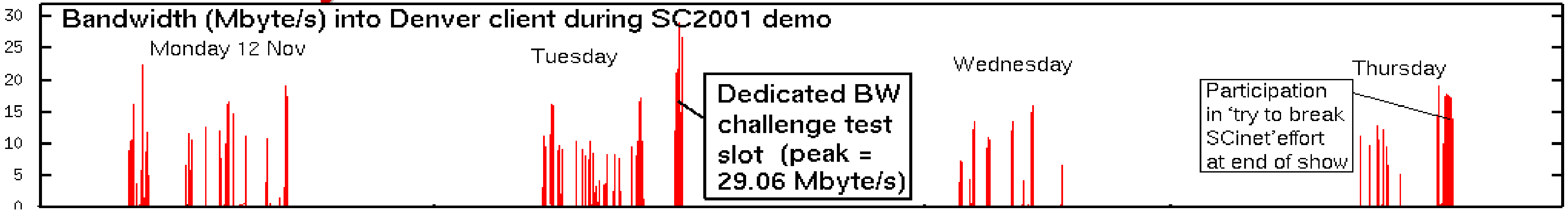
Source: Charlie Catlett, Argonne

Grid-enabled Data Analysis: SC2001 Demo by K. Holtman, J. Bunn (CMS/Caltech)

- Demonstration of the use of Virtual Data technology for interactive CMS physics analysis at Supercomputing 2001, Denver
  - ➔ Interactive subsetting and analysis of 144,000 CMS QCD events (105 GB)
  - ➔ Tier 4 workstation (Denver) gets data from two tier 2 servers (Caltech and San Diego)
- Prototype tool showing feasibility of these CMS computing model concepts:
  - ➔ Navigates from tag data to full event data
  - ➔ Transparently accesses `virtual' objects through Grid-API
  - ➔ Reconstructs On-Demand *(=Virtual Data materialisation)*
  - ➔ Integrates object persistency layer and grid layer
- Peak throughput achieved: 29.1 Mbyte/s; 78% efficiency on 3 Fast Ethernet Ports



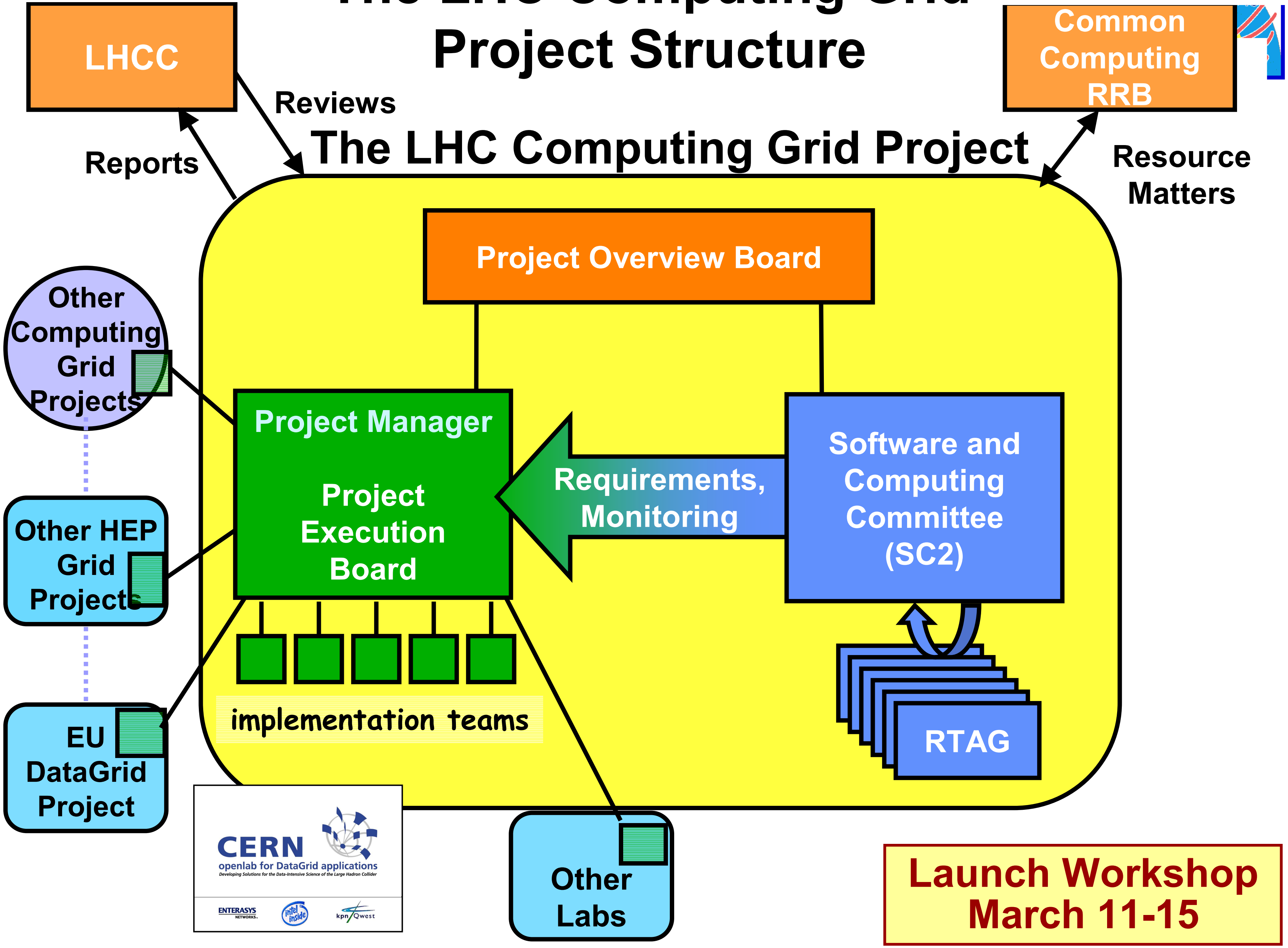




Dedicated BW challenge test slot (peak = 29.06 Mbyte/s)

Participation in 'try to break SCinet' effort at end of show

The LHC Computing Grid Project Structure



CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider
ENTERASYS NETWORKS
intel inside
kpnQwest

Launch Workshop
March 11-15

Grid R&D: Focal Areas

- ***Development of Grid-Enabled User Analysis Environments***
  - **Web Services (OGSA based) for ubiquitous, platform and OS-independent data (and code) access**
  - **Analysis Portals for Event Visualization, Data Processing and Analysis**
- ***Simulations for Systems Modeling, Optimization***
  - **For example: the MONARC System**
- ***Globally Scalable Agent-Based Realtime Information Marshalling Systems***
  - **For the next-generation challenge of Dynamic Grid design and operations**
  - **Self-learning (e.g. SONN) optimization**
  - **Simulation enhanced: to monitor, track and forward predict site, network and global system state**
- ***1-10 Gbps Networking development and deployment***
  - **Work with DataTAG, the TeraGrid, STARLIGHT, Abilene, the iVDGL, iGOC, HENP Internet2 WG, Internet2 E2E**
- ***Global Collaboratory Development: e.g. VRVS, Virtual Access Grid***

Jean-Marie BROM (IReS)
jmb@193.48.90.136/h261
5.6 f/s 213 kb/s (0%)
☐ mute ☑ color

Geoff Hall
ghall@155.198.211.150/h261
8.0 f/s 130 kb/s (0%)
☐ mute ☑ color

Guenter Fluegge
fluegge@137.226.33.77/h261
7.9 f/s 104 kb/s (0.1%)
☐ mute ☑ color

Rino Castaldi ()
castaldi@193.205.77.230/h261
7.5 f/s 206 kb/s (2.5%)
☐ mute ☑ color

Joe Incandela (University of Cal
incandel@131.225.235.137/h26
8.1 f/s 101 kb/s (0%)
☐ mute ☑ color

Room 40-R-D10 (CERN)
CERN@137.138.77.179/h261
8.0 f/s 33 kb/s (0%)
☐ mute ☑ color

http://www.vrvs.org/

SUN
"CMS TRACKING S.C."
16:28:02
YOUR TIME (GMT +1)
UNTIL
17:58

Cern 40-R-D10 Room	Cern Guenter Fluegge	Infn Rino Castaldi	Rutherford Geoff Hall
Cern JEAN-MARIE BROM	Caltech Joe Incandela	Caltech Philippe Galvez	Cern danek kotlinski

VRVS Team (@CERN)
...gge
...ROM (IReS)
...ski PSI
...()
...10 (CERN)
...a (University of California,

mute mute
50
0

Participants
40-R-D10 Room
Geoff Hall
Gregory Denis
Guenter Fluegge
JEAN-MARIE BROM
Joe Incandela
Rino Castaldi
danek kotlinski(2)

My office ROOM
Controlling Camera: 1
<--- PAN --->
TILT
Zoom: 3

Warning: Applet Window

9352 Hosts;
5369 Registered Users in 63 Countries
42 (7 I2) Reflectors
Annual Growth 2.5X