

Networks and Grids for HEP Experiments

Harvey B Newman
California Institute of Technology
Pasadena, CA 91125, USA
newman@hep.caltech.edu

© Harvey B Newman, 2002

Scientific Exploration at High Energy Frontier

The major high energy physics experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures and symmetries that govern the nature of matter and spacetime. Among the principal goals is to find the mechanism responsible for mass in the universe, and the mysterious “Higgs” particles associated with mass generation.

The largest collaborations today, such as the CMS and ATLAS who are building experiments for CERN’s Large Hadron Collider (LHC) program, each encompass 2000 physicists from 150 institutions in more than 30 countries, and they each include 300-400 physicists in the US, from more than 30 universities as well as the major US HEP laboratories. Collaborations on this global scale would not have been attempted if the physicists could not plan on excellent networks: to interconnect the physics groups throughout the lifecycle of the experiment and, and to make possible the construction of Data Grids capable of providing access, process and analysis of massive datasets, rising from the Petabyte to the Exabyte scale within the next decade.

The current generation of experiments at SLAC (BaBar) and Fermilab (D0 and CDF) face similar challenges, and BaBar in particular has already accumulated datasets totaling more than 500 Terabytes.

HEP Challenges: at the Frontiers of Information Technology

Realizing the scientific wealth of these experiments presents new problems in data access, processing and distribution, and collaboration across national and international networks on a scale unprecedented in the history of science. The information technology challenges include:

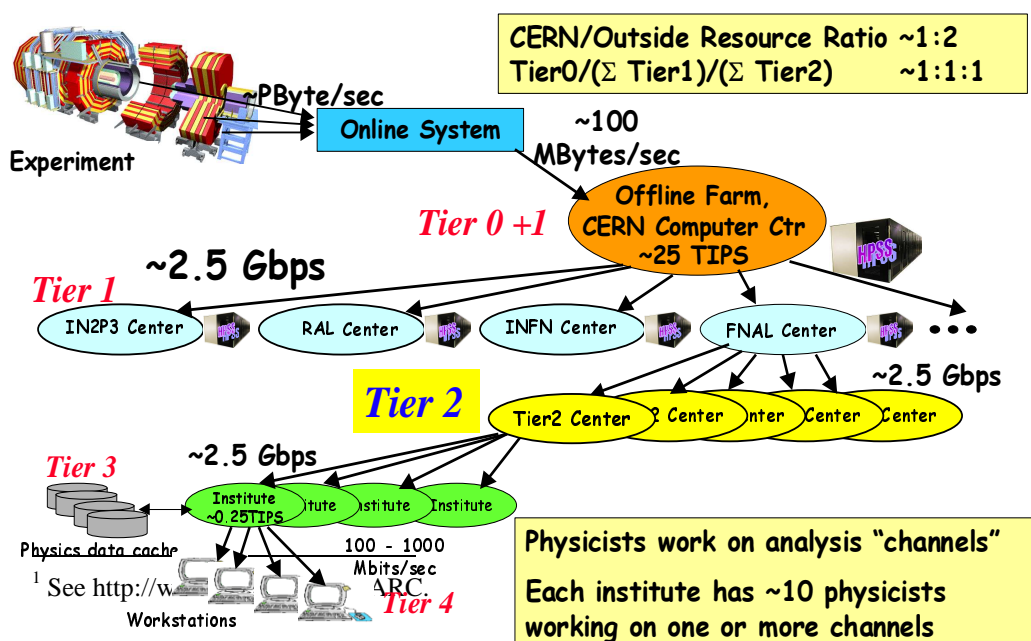
- Providing rapid access to data subsets drawn from massive data stores, rising from Petabytes (10^{15} bytes) in 2002 to ~100 Petabytes by 2007, and Exabytes (10^{18} bytes) by approximately 2012
- Providing secure, efficient and transparent managed access to heterogeneous worldwide-distributed computing and data handling resources, across an ensemble of networks of varying capability and reliability
- Tracking the state and usage patterns of computing and data resources in order to make rapid turnaround as well as efficient utilization of global resources possible
- Matching resource usage to policies set by the management of the experimental Collaborations over the long term
- Providing the collaborative infrastructure that will make it possible for physicists in all world regions to contribute effectively to the analysis and the physics results, including from their home institutions.

- Building regional, national, continental and transoceanic networks, with bandwidths rising from the Gigabit/sec to the Terabit/sec range over the next decade
- Integrating all of the above infrastructures to produce the first managed distributed systems serving “virtual organizations” on a global scale

Meeting the HEP Challenges: Data Grids as Managed Global Systems

In order to meet these challenges, the LHC experiments have adopted the “Data Grid Hierarchy” model (developed by Newman at Caltech and his collaborators in the MONARC¹ project) shown schematically in the figure below. This model shows data at the experiment is stored at the rate of 100 – 1500 Mbytes/sec throughout the year, resulting in many Petabytes per year of stored and processed binary data, accessed and processed repeatedly by the worldwide collaborations searching for new physics processes. Following initial processing and storage at the “Tier0” facility at the CERN laboratory site, the processed data is distributed over high speed networks to ~10 national “Tier1” centers in the US and the leading European and other countries. The data is further processed and analyzed and stored at approximately 50 “Tier2” regional centers, each serving a small to medium-sized country, or one region of a larger country (as in the US, UK and Italy). Data subsets are accessed from and further analyzed by physics groups using one of hundreds of “Tier3” workgroup servers and/or thousands of “Tier4” desktops.

The successful use of this global ensemble of systems to meet the experiments’ scientific goals depends on the development of Data Grids capable of managing and marshalling the “Tier-N” resources, and supporting collaborative software development by groups of varying sizes spread across the globe. The modes of usage and prioritization of tasks must be done in such a way that the physicists’ requests for data and processed results are handled in a reasonable turnaround time, and at the same time the Collaboration’s resources are used efficiently. The GriPhyN, PPDG, iVDGL, EU Datagrid, DataTAG, LHC Computing Grid and national European Grid projects are working together, in multi-year R&D programs, to develop the necessary Grid systems. The DataTAG project is also working to address some of the network R&D issues.



The data rates and network bandwidths shown in the figures are a very conservative “baseline” formulated using a 1999-2000 evolutionary view of network technologies.

In order to build a “survivable”, flexible distributed system, much larger bandwidths are required, so that the typical transactions, drawing 1 to 10 Terabyte and eventually 100 Terabyte subsamples from the multi-Petabyte data stores, can be completed in 1 to 10 minutes. Completing these transactions in a few minutes (rather than hours) is necessary to avoid the inherently fragile state that would result if hundreds to thousands of requests were left pending for long periods, and to avoid the bottleneck that would result from tens and then hundreds of such “data-intensive” requests per day (each still representing a very small fraction of the stored data). It is important to note that transactions on this scale correspond to data throughputs across networks of 10 Gbps to 1 Tbps for 10 minute transactions, and up to 10 Tbps (more than the current capacity of a fully instrumented fiber circa 2002) for 1 minute transactions.

In order to fully understand the potential of these applications to overwhelm future planned networks, we note that the binary (compacted) data stored is pre-filtered by a factor of 10^6 to 10^7 by the “Online System” (a large cluster of hundreds to thousands of CPUs that filter the data in real time). This realtime filtering, though traditional, runs a certain risk of throwing away data from subtly new interactions that do not pre-conceived existing or hypothesized theories. The basic problem is to find new interactions from the particle collisions, down to the level of a few interactions per year out of 10^{16} produced. A direct attack on this data analysis and reduction problem, analyzing every event in some depth, is beyond the current and foreseen states of network and computing technologies.

US universities and laboratories engaged in high energy physics have had a leading role in these developments. The BaBar experiment at SLAC has the largest accumulated data store today, and is among the largest users of national and international networks. The US contingent of the CMS experiment, including Caltech, Florida and Fermilab in particular, has led the development of the LHC distributed computing model and has had a leading role in the development, operation and planning for HEP’s international networks over the last 20 years, in collaboration with LBNL, SLAC, ANL and FNAL, and more recently CERN and STARLIGHT. US physicists in the ATLAS project also have contributed to these efforts, led by the University of Michigan, Indiana and the Argonne, Berkeley and Brookhaven national labs. Caltech and UCSD recently deployed the first prototype Tier2 center in CMS, split between Caltech/CACR and SDSC, and are working closely with UC Davis, Riverside and UCLA in California, as well as the university of Florida on a diverse set of physics studies aimed at searching for the Higgs particles and supersymmetry and optimizing the performance of the CMS detector. Caltech and UCSD also will use the TeraGrid, and prototype Tier1 centers at FNAL and CERN, to meet the continuing needs for simulated particle interaction “events”.

Plans are already being developed to put “last mile fiber” in place between Caltech/CACR and 818 W. 7th St., and to use OC192 wavelengths on these fiber strands for HEP applications starting in the Spring of 2003. A similar initiative is underway to link Fermilab to STARLIGHT in Chicago. But for the longer term, dark fibers are the preferred solution for meeting the most data intensive needs of these groups in high energy physics. Such facilities are currently planned for Fermilab, which will soon be connected to STARLIGHT using dark fibers.

Relevance of Meeting These Challenges for Future Networks and Society

The HEP (or HENP, for high energy and nuclear physics) problems are the most data-intensive known. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals, better calibrate the detector and better reconstruct the quantities of interest (energies and decay vertices of particles such as electrons, photons and muons, as well as jets of particles from quarks and gluons). The globally distributed ensemble of facilities, while large by any standard, is less than the physicists require to do their work in an unbridled way. There is thus a need, and a drive to solve the problem of managing global resources in an optimal way, in order to maximize the potential of the major experiments for breakthrough discoveries.

In order to meet these technical goals, priorities have to be set, the system has to be managed and monitored globally end-to-end, and a new mode of "human-Grid" interactions has to be developed and deployed so that the physicists, as well as the Grid system itself, can learn to operate optimally to maximize the workflow through the system. Developing an effective set of tradeoffs between high levels of resource utilization, rapid turnaround time, and matching resource usage profiles to the policy of each scientific collaboration over the long term presents new challenges (new in scale and complexity) for distributed systems.

A new scalable Grid agent-based monitoring architecture, a Grid-enabled Data Analysis Environment, and new optimization algorithms coupled to Grid simulations are all under development in the HEP community.

Successful construction of network and Grid systems able to serve the global HEP and other scientific communities with data-intensive needs could have wide-ranging effects on research, industrial and commercial operations. Intelligent, resilient, self-aware systems able to support a large volume of robust Terabyte and larger transactions, to adapt to a changing workload, and to match the use of resources to policies would provide a strong foundation for the distributed data-intensive business processes of the multinational corporations of the future.

Development of the new generation of systems of this kind could also lead to new modes of interaction between people and "persistent information" in their daily lives. Learning to provide, manage and absorb this information and in a persistent, collaborative environment would have a profound transformational effect on our society, in ways that cannot be imagined given the limited network and nascent Grid technologies currently available to the world population.