October, 2001

Report of the Transatlantic Networking Committee

Flavia Donno, INFN Pisa (CDF)

Jim Leighton, LBNL (ESnet)

Olivier Martin, CERN

Richard Mount, SLAC

Boaz Klima, FNAL (D0)

Harvey Newman, California Institute of Technology (CMS) *

Ronald Poling, University of Minnesota (BTeV)

Lawrence Price, ANL (ATLAS)*

Wesley Smith, University of Wisconsin (ZEUS)

Charles Young, SLAC (*BABAR*)

* Co-chairs

A. Report Summary

This report presents the results of the Transatlantic Network Working Group, commissioned by the Joint Oversight Group (JOG) in which DOE and NSF oversee U.S. work on the Large Hadron Collider. The Working Group reviewed the present and future requirements for transatlantic networking over the next five years of the major HEP programs based at SLAC, Fermilab, DESY and CERN. The results include (1) a detailed bandwidth and funding plan for the US-CERN link, as that was flagged as an action item by the agencies, (2) a new set of estimated required bandwidths, which are larger than previous estimates for reasons explored in the report, and (3) a number of general conclusions on the implications for domestic US connectivity requirements. Given the rapid technology advances, and associated advances in distributed applications (including Grid systems) it is clear that periodic reviews of the requirements, both domestic and international, and an ongoing study of the evolution of networking and related technologies that affect HEP's use of networks will be needed.

Although this report on transatlantic networking for high energy physics was requested by the JOG, the Working Group included representatives of the other major HEP experiments with important European collaborators, including ZEUS, *BABAR*, D0, CMS, and BTeV. A significant finding is that the hierarchical, distributed access to computing and data envisioned under the name of "data grid" by the LHC experiments is already in use in an early form by current experiments, in particular *BABAR* and D0. Although the full grid toolkit is not yet available, these experiments have very important remote computing and storage centers abroad and need substantial links to make effective use of the resources provided by their collaborators. In fact, these two experiments are planning to transfer a larger *fraction* of their data across the transatlantic links than the LHC experiments are planning for. The LHC experiments are planning for the ability to use mature grid systems to keep the growth of network requirements well below what would be expected from the growth in raw data set size. The comparison between these presently running experiments and the future more complex and distributed LHC experiments, shows the LHC requirements to be rather small, rather than very ambitious when viewed in the technological environment of 1998 (just two years ago). This raises the possibility that LHC bandwidth usage could actually be larger than the present estimates in this report, especially if the cost per unit bandwidth falls faster than foreseen.

To round out the picture, ZEUS is a mature experiment that will continue its present model of concentrating computing and storage at DESY, while CDF also plans a concentrated computing model for Run II. BTeV is still becoming formed and has not fully planned its computing model.

The transatlantic bandwidth requirements, whose estimation constituted the largest part of the work of the Committee and which are summarized in Section B.4, show the requirements rising from 3,000 Megabits per second (Mbps) in 2002 to 23,000 Mbps in 2006, the year the LHC turns on. The question of what fraction of this substantial bandwidth must be installed and supplied by the U.S. program does not have a crisp answer. The countries supplying large computer centers abroad for currently running experiments are presently supplying virtually all the long-distance networking needed to make effective use of them. This pleasant circumstance has two potential problems. One is that the current connections to SLAC and FNAL, the source of the data for BaBar and D0, are not fast enough to make full use of the transatlantic links provided by foreign collaborators. The second is that the continuation (and required upgrades) of the needed links, while hoped for by the foreign collaborators, is not guaranteed, and might not be supplied for the duration of the experiments. Nevertheless, in projecting the needed size of the U.S.-CERN link, which is primarily supported by the U.S. program, the Committee has assumed that this emerging standard where remote collaborators will supply the needed long-haul networking will continue during the period considered in this report. The result is a projected bandwidth requirement on the U.S. CERN link of 622 Mbps in 2002, only 20% of the total, growing to 10,000 Mbps in 2006, close to 50% of the total, in an era when the LHC experiments are beginning to dominate and the U.S. has the major remote resources that must be connected by networks.

The problem that the networking between the U.S. and Europe, which is now provided by European collaborators, is not fully utilized because of limited bandwidth into U.S. laboratories is one that

needs close attention. This problem will grow with time, as reliance on the network increases. HEP experiments have moved faster than predicted to making use of the network as the principal, and in many cases the sole means of data movement. Although we were not able to do a full analysis, the Committee does provide an initial analysis of the rapid change from shipping data by tapes to shipping data over the network. There is also sufficient evidence that the accelerating growth of bandwidth usage is underway on the US-CERN link, notably by BaBar and Fermilab, reaching peak rates of 100 Mbps on the 155 Mbps link as of July 2001.

We conclude that the combination of economic tradeoffs with improved productivity of personnel and strong "sociological" incentives to provide resources at home, makes it unlikely that this trend will be reversed or even slowed. However, it is also true that the connection speeds that will be needed at accelerator and host laboratories, and at universities and laboratories that house Tier 2 centers, are greater than have been planned for by the U.S. backbone networks (ESnet for DOE laboratories and Internet2 for universities). It is imperative that the needs of the HEP experiments be communicated to the networks and made part of their planning. Similar planning will be needed on the European end, and for the field of Nuclear Physics.

Costs have been estimated using plausible extrapolations from current information and are summarized on p. 24. The costs are divided into those for the "network" costs of links and internet access (rising to $3M in 2006) and those for a variety of infrastructure costs, including routers and switches, monitoring systems, and a small number of personnel to support the network ($1.2M in 2006). We note that network costs are notoriously hard to predict 5 years in advance and that it has always been true that achievement of affordable networking has relied on intensive negotiation and comparison shopping, and that we have assumed continued success in cost savings in the estimates here.

Although these bandwidth and cost estimates are a substantial part of the results of this Committee, we also make five major recommendations. They are listed in Section C, and repeated here:

1. High performance networking is central to the success of international collaborations. Every effort should be made to provide the required transatlantic bandwidth, particularly that of the US-CERN link, as documented in this report for the period 2001-2006, and to continue to provide adequate bandwidth to keep pace with HEP's needs.
2. Arrangements must be made to ensure that the US-CERN connection and other international links used by HEP experiments are connected with the required high performance, to the networks that make the ultimate connections to the collaborating institutions in both regions. Similarly, arrangements should be made to ensure that key university or laboratory sites (e.g. housing Tier 2 or Tier B centers) are appropriately connected to their regional networks or to the national backbone networks. It is likely that ESnet and Internet 2 will not meet all needs within the U.S. without special arrangements, and some targeted funding by HEP should be foreseen.
3. Appropriate attention should be paid to providing high quality network service, and to monitoring and troubleshooting the network in order to maintain good performance.
4. The agencies should ensure that planning for networking (both domestic and international) to serve HEP experiments is reviewed at least annually by a committee with official representation from at least the major user experiments.
5. The committee found that there was a strong need for a technically oriented forum to promote information exchange and technical collaborations between the HEP community and network providers and technical experts, and the work of the forum should extend to identifying and making available standardized toolsets to ensure that the networks provide the needed end-to-end performance. The HENP Network Working Group was then formed in June 2001, which promises to meet these needs. The committee recommends that that this group be encouraged by the agencies and that technical representatives of laboratories and experiments be encouraged to participate in its work.

B. Report

B.1. Purpose of committee and charge

The Transatlantic Network Committee was created at the request of the LHC Joint Oversight Group, composed of representatives of the Department of Energy and the National Science Foundation, to assemble and review the projected needs for networking links across the Atlantic in support of the LHC experiments and other large international projects in high energy physics. Pursuant to the charge letter, which is reproduced in Appendix I of this report, members were nominated to the committee by the managements of major international experiments in which the U.S is involved, including ATLAS, CMS, CDF, D0, *BABAR*, BTeV, and ZEUS. In addition, representatives of ESnet and of the IT Division at CERN were included in the committee.

The specific charge to the committee was the following two points:

- ?? Estimate the bandwidth needed for the network connection to CERN to meet the needs of the U.S. HEP community over the next 5 years (2001-2005)
- ?? Estimate the cost of providing this bandwidth

In the course of the committee's deliberations, as summarized in this report, we found that in addition to the US-CERN link, other transatlantic links funded by European collaborators, domestic links to universities and laboratories, and the quality of the local infrastructure at each site, also play important roles in meeting HEP's overall networking needs.

B.2. History and background

B.2.1. Role of computer networking in HEP experiments

For 20 years, high energy physicists have relied on state-of-the-art computer networking to enable ever larger international collaborations. As computer networking became possible over wide areas, starting in the late 1970s, adoption by high energy physics came quickly. The practice, already prevalent, of conducting complex experiments through collaborations of groups from universities and laboratories that were often in different parts of the country created an immediate need for the new technology. In the 1980's, the connection to the emerging "HEPnet" became a required piece of infrastructure for any group serious about doing experiments in high energy physics. The network solved problems for the modest sized collaborations of the day, and in turn enabled the evolution of the ever-larger and increasingly international collaborations that were needed to mount major collider experiments and other complex instruments.

Collaborations on the scale of those now building LHC detectors could never have been attempted if they had not been able to expect the excellent international communications that make them possible.

Today, the network is needed for all aspects of collaborative work. Collaborators work together across the network to write proposals, produce and agree on designs of components and systems, collaborate on overall planning and integration of the detector, confer on all aspects of the device, including the final physics results, and provide information to collaborators and to the physics community and general public.

Data from experiments lives almost exclusively on the network. The data is acquired and written to disk or tape across at least the local area network of the host laboratory. It is moved to the places in the collaboration where it is needed for processing or detailed data analysis. And finally the resulting physics papers are written, edited, and reviewed on and across the network.

Because of its need for large distributed collaborations, HEP has traditionally led the demand for research networks among the sciences. As noted above, HEP developed its own national network in the early 1980s, until the multi-disciplinary networks supported by DOE and, for a time, by NSF emerged. Since the installation of national backbones, HEP and other sciences have generally received at least adequate support from the research network backbones. There have always been, however, specific network connections where HEP has found it necessary to support special capabilities that could not be supplied efficiently or capably enough through more general networks. A long-standing example is the link across the Atlantic to support U.S. experimenters at CERN, starting in the 1980s for L3 and other LEP experiments, and more recently for work on the LHC. These needs for links dedicated to HEP use are needed in special cases because HEP requirements can be large and can overwhelm those of researchers in other fields and, as in the case of the link to CERN, because regional networks do not give top priority to interregional connections.

B.2.2. LHCNet Brief History and Bandwidth

The US-CERN link was initiated in 1984 for the L3 and LEP experiments, and more recently the LHC program. After several technology generations, at speeds from 9.6 Kbps and up, the first leased digital line, between CERN and MIT at 64 Kbps, began service in 1989. In December 1995, the first all-IP T1 link began service and was upgraded periodically to meet HEP's needs. In April 1999, as a result of European deregulation of the telecommunications industry, we were able to move beyond T1 and E1 (2 Mbps) links to more cost effective services at 12 and then 20 Mbps[1] during 1999. In October 1999, after discussions between the Internet2 board, Caltech and CERN, CERN became the first full member of Internet2 outside the US.

The bandwidth levels reached in 1999 were funding limited: 12 Mbps (IP over ATM) starting in April 1999, with an upgrade 20 Mbps by October 1999. A call for tender for an upgrade to 45 Mbps (DS3), with a further upgrade to 155 Mbps (OC3) in October 2000, was issued by CERN in the Fall of 1999, and the tender was won by KPN/Qwest[2]. As part of the winning bid, Qwest offered to work with us to try to use the redundant backup path that protects every commercial optical fiber-based data transmission service (as a hot backup). If this is successful, we will have access to a second OC3 for transatlantic network research purposes, supporting the GriPhyN and PPDG Grid projects, by the end of 2001. Use of the second link will be at the cost of an additional New York – Chicago link, and an additional link to the STARTAP[3], and two additional routers and switches at CERN and at the Qwest Point of Presence (PoP) in Chicago. If the primary link fails, we would attempt an automatic switchover to the backup link using the second set of routers.

An upgrade to DS3 was completed in March 2000, allowing us to meet the combined needs of *BABAR*, Run2, and the LHC experiments, including the "data challenges" that began during the Spring and Summer of 2000 for CMS. We began the transition to OC3 in the Fall of 2000. The upgrade of the link, including the connection to the STARTAP peering point in Chicago at OC3 provided by Ameritech, was completed as of January 31, 2001. To complement the (required) bandwidth increases, upgraded new router and switch software and firmware were installed, to effectively manage the increasingly diverse range of applications, from large file transfers to a variety of realtime multimedia applications. It is planned to upgrade the transatlantic bandwidth from 155 Mbps (OC3) to 622 Mbps (OC12) in April 2002. In preparation for this upgrade a market survey has recently been performed, a Call for Tender went out this Summer, and the vendor will be selected by December 2001. We expect to connect the new OC12 link to the new StarLight[4] point of presence (PoP) in Chicago, which provides a neutral site for fee-free co-location of network equipment. Therefore, for our next upgrade, we expect to land at StarLight instead of the current location at the Qwest PoP in Chicago. Some new routing equipment also will be installed, to handle the increased bandwidth.

B.2.3. Expected model for data analysis in international collaborations

B.2.3.1. History

Until the last few years, analysis of data for extended collaborations was centered on the experimental site with only a few cases where major external centers were used. Up to the mid-1990s bulk data were transferred by shipping tapes, while networks were used for programs and conditions data. External analysis centers invariably served the local/national users only. Most successful examples involved staff (and often equipment) from the external center being placed at the experimental site to ensure the flow of tapes. All but the most successful examples caused the external analysis to be significantly disconnected from the collaboration mainstream. Nevertheless, this was the best practical solution in an era when affordable networks had capacity well below that needed to transport a significant fraction of the data.

B.2.3.2. Current Practice

This Committee found that a strikingly different model, relying almost exclusively on the network for data distribution to important remote centers, has now emerged as the dominant one. We discuss the new model by referring to the *BABAR* experiment (currently the experiment with the most data – 400 TB) as an example. Although the computing and analysis systems of CDF and D0 differ in detail from that of *BABAR*, all are based

[1] IP over ATM.

[2] Apart from the lowest cost, this choice had the advantage that Qwest is the supplier of the Abilene (Internet2) backbone in the US, and KPN currently supplies the TEN-155 academic and research network backbone in Europe.

[3] STARTAP is a designated location at which non-US networks may terminate their circuits in order to peer with the US networks including Internet2, ESnet, etc. STARTAP also serves as a point at which non-US networks may peer among themselves, without transiting US network links in conflict with "acceptable use policies" (AUP). See http://www.startap.net.

[4] The Optical StarTap. See http://www.startap.net/starlight .

on the assumption that the network is an integral, but relatively costly, component of the data analysis environment. *BABAR* is implementing a model with Tier A, Tier B and Tier C centers. There are currently two Tier A centers, SLAC (the site of the experiment) and CCIN2P3 Lyon, France. The Tier A centers are at the disposal of *BABAR*, meaning that they can be exploited for any *BABAR* purpose and are "blind" to the institutional or national affiliation of any *BABAR* member. The CCIN2P3 center is ramping up this year. It already has a full micro-DST on disk. The likely future mode of operation is that the Tier A centers will each host 30 to 100% of all *BABAR* data in their robotic tape systems and will make mutually exclusive parts of all but the smallest analysis data sets available on disk. For example, 1999 and 2001 data might be on the SLAC disks and 2000 data on the CCIN2P3 disks.

Tier B centers generally offer regionally or nationally focused analysis facilities. Tier C centers are typically universities with significant local analysis facilities. To facilitate efficient analysis at the Tier B and Tier C centers, *BABAR* analysis data sets are being physically streamed. Currently 20 non-exclusive streams are used, significantly increasing the total data volume in mass storage, but probably increasing the effectiveness of disk caching at the larger centers.

Although *BABAR* analysis is based on Objects and mainly uses the Objectivity/DB database, data distribution is based on whole database files transferred using ftp-like utilities.

Simulated data are generated at various centers in the collaboration. Currently all are sent to SLAC. In the future some types of simulated data may not be sent to SLAC or any other site unless they are explicitly requested.

All *BABAR* data is transferred over networks. For example, newly reconstructed data are moved into the user-accessible "analysis federation" at SLAC twice a week and are also transferred by network to CCIN2P3 taking 2-3 days. The committee believes that this is the optimal solution in an era when networks, able to transmit raw and derived data once, have a comparable cost to the disk or CPUs or mass-storage systems in the analysis environment.

Although network costs have been dropping rapidly, full use of this model strains current network resources and analysis models are being constrained in order to minimize network costs. For the two running experiments most actively using the network for current data transfer, *BABAR* and D0, the major foreign partners are providing transatlantic links into the U.S. to match the substantial computing and storage facilities they are providing abroad. Although the future of these links is not guaranteed, they fit an emerging model where foreign partners on major experiments provide the long-haul networking needed for access to the host laboratory.

B.2.3.3. Comparison with Alternate Models

B.2.3.3.i. Why Not Centralize Almost Everything?

North America, Europe, Japan and many other regions are now interconnected by networks that make it almost as easy to log on to computers 10,000 miles away as to access local servers. Why not, for example, collect money from all *BABAR* collaborators and install all the computers and storage at SLAC? A common first thought is that that would be more cost-effective than a distributed approach.

Provided we take a narrow view, encompassing only one experiment and only the next few years, centralization may provide the best physics analysis capability. A broader view is required to see that there is really no alternative to the more distributed approach. In the broader view we note that collaboration on all the challenges of doing experiments is one of the great strengths of our field. HEP computing in the *BABAR*-RunII-LHC era is a scientific challenge in itself, and is a training ground for physicists and computer scientists whose computing skills are proving very valuable to their countries. Future world-scale experimental programs must be planned with explicit support for a collaborative environment that allows many nations to be full participants in the challenges of data analysis.

Today, an objective analysis of the benefits to national and local interests invariably and correctly leads to massive national and local investments whenever they are an effective alternative to sending a more modest amount of money to install systems at the site of an experiment. The result is an effective political decision by each funding agency to make every effort to avoid funding centralized systems in other countries.

One more variant on the centralized approach should be laid to rest. For US-based experiments, why not cut all the foreign network links and use the US funds saved to invest in computing at SLAC and Fermilab? *BABAR* has made a calculation of the value of foreign Tier A centers in reducing costs at SLAC. This will exceed $2M in FY2002. The total annual value of foreign centers to the US-based program is several times this and is greatly in excess of the estimated cost to the US of creating the required high-speed paths from SLAC and Fermilab to the landing points of lines funded by foreign collaborators.

B.2.3.3.ii. Why Not Send Tapes by FedEx?

Again, *BABAR* will be used as the illustration. Physical shipment of tapes was the only way to transfer bulk data until the late 1990s. Before the start of data taking in 1999, *BABAR* expected to use tapes to transport data to major remote centers and to individual universities.

An analysis of transatlantic network costs from 1987 to 1998 shows that capacity per unit cost took almost eight years to double. Since 1998, market forces have transformed the regulated transatlantic networking market and capacity per unit cost has almost doubled each two years. At the same time, Internet2 became a strong competitor to FedEx for intra-US data distribution to universities. As a result *BABAR* does not plan to use any tape-based data distribution, believing that the much closer collaborative environment offered by network distribution has a value more than offset by any additional cost. This view is particularly strongly felt by the European collaborators using their own national funds to transport *BABAR* data across the Atlantic.

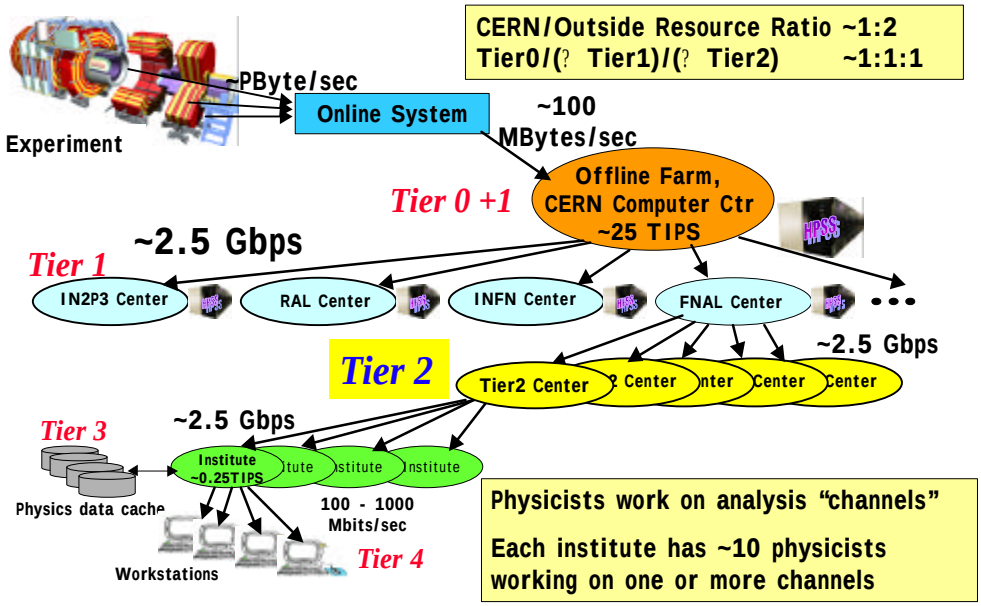
An attempt has been made to estimate the costs of using tapes to interchange data and simulations between SLAC and three Tier A centers in Europe at rates expected early in 2002. The estimated cost is about $800k per year as detailed in Appendix II. It is not the purpose of this estimate to claim that tapes are more expensive than networks, but it is claimed that the tapes, hardware and person power needed for a moderately responsive tape-based system have significant cost and that this cost is likely to be able to buy network capacity of more value to physics analysis.

B.2.3.4. LHC Era

The baseline model has a similar foundation to that of *BABAR* and other current experiments to the extent that the network is assumed to be an expensive resource whose use should be minimized. The LHC experiments have pioneered thinking about a grid-enabled hierarchical structure of storage and analysis centers, as shown in the CMS figure below, that extends and expands the model used presently in *BABAR*. The exact functions of Tier 1 and Tier A centers may differ, but the guiding principle will be to implement a network-use strategy dominated by careful planning so that network costs are kept close to the level below which analysis would be severely impacted. This level is likely to be relatively higher for the LHC experiments because the aggregate resources at each level in the hierarchy -- Tier 1, Tier 2, Tier 3 and Tier 4 -- will be comparable and large while the Tier 0 site at CERN will be modest as a percentage of the total, relative to currently running experiments.

Since the more local resources (desktop/Tier 4, Tier 3, Tier 2) will bring much of the analysis power, analysis-specific (or physicist-specific) event collections will have to flow over the network to these resources. This is expected to be considerably more dynamic and voluminous than the flows envisaged for current experiments in the next 5 years. Pre-emptive transfers will be used to improve responsiveness at the cost of some extra network traffic.

Fig. 1: LHC Model for Grid Hierarchy, as Planned by CMS



The multi-tier architecture must become more general and flexible so that it can accommodate the very large uncertainties in the relative costs of CPU, storage and networking while enabling physicists to work effectively in the face of data having unprecedented volume and complexity. Transparency and location independence are requirements – a need for individual physicists to understand and manipulate all the underlying transport and task-management systems would be crippling. It is quite probable that network unit costs will decline more rapidly than the costs for storage and computation, at least in the near-term future. It is obviously vital that the ability to exploit large bandwidth increases to improve the quality and speed of physics analysis be a fundamental part of the LHC analysis model.

The "Virtual Data" concept will be effectively implemented, at least for "production" object collections. Adequately precise cost estimation and accounting will be implemented for each element (network, CPU, storage, …) of the analysis environment. Optimizing the cost of instantiation and replication of data together with the location of task execution will allow maximum analysis effectiveness with minimal WAN use if WAN unit costs are high. If WAN unit costs decrease dramatically in comparison with other computing costs, the virtual data concept promises to automatically deliver greatly improved responsiveness from the distributed analysis facilities. In summary, baseline models minimize network costs, but resources will be more distributed than for current experiments. Added sophistication (e.g. virtual data) is needed to be competitive and efficient in a future where costs are very uncertain.

B.3. International Networking Requirements

As described above, the committee included representatives of the major international experiments of the next five years in which the U.S. is strongly involved. These include, in the chronological order in which experiments began (or will begin) taking data in their current configuration, ZEUS, *BABAR*, D0, CDF, CMS, ATLAS, and BTeV. In the following sections, we describe briefly each of these experiments and their international networking requirements. Note that for each experiment we give a table of required installed bandwidths, calculated as twice the required data transfer rate, in order to provide the usual practical utilization rate of 50% of the installed bandwidth.

B.3.1. ZEUS and DESY

B.3.1.1. Overview of DESY, HERA and the Experiments.

Located at the Deutsches Electron-Synchrotron (DESY) laboratory in Hamburg, Germany, the Hadron Elektron Ring Anlage (HERA) is the world's first electron-proton collider. It was approved in 1984 and had first collisions in October, 1991. Initially, HERA collided 27.5 GeV electrons and then positrons on 820 GeV protons. Beginning in 1998, 27.5 GeV electrons were collided with 920 GeV protons. After a mid-1999 switch back to positrons, HERA operated until September 2000. HERA will resume operations in Fall 2001 with a five-fold increase in luminosity and spin-rotators in place for polarized measurements.

There are 2500 Physicists involved in HERA experiments, the synchrotron light facility HASYLAB and different projects. This number is expected to remain relatively constant over the next 5 years. At present, the DESY laboratory has 1300 Lab Employees, Scientists, Engineers, and other personnel, which is also expected to be constant through 2005. There are 80 Collaborating Institutions. The HEP and NP experiments at HERA include the two general purpose ep collider experiments, ZEUS and H1, the HERMES experiment which uses the polarized electron beam on a fixed gas-jet target, and the HERA-B experiment which collides the proton beam halo with wires to study B-meson production.

B.3.1.2. Schedule and Data Estimates

The majority of the US HEP effort is on the ZEUS experiment, while there is significant NP participation in HERMES. The HERA program is expected to continue at least through 2005. The experimental program at DESY beyond 2005 will depend on the future of the TESLA linear collider proposal. Approval of this project at DESY could result in a significant increase in network requirements.

The ZEUS 1998-1999 electron running yielded a substantial sample from 17 pb^{-1} (24 million events). After a mid-1999 switch back to positrons, HERA operated until September 2000 with a total ZEUS integrated positron luminosity of 115 pb^{-1} (164 million events). After HERA resumes operations with increased luminosity polarized electron and positron running, the stated laboratory goal is 1 fb^{-1} by the end of 2005.

B.3.1.3. Computing and Analysis Usage

The DESY program produces 30 TB/year of raw data, 40 TB/year of reconstructed data, 20 TB/year of Monte Carlo data, 4 TB/year of DST (data + MC) data and 5 TB/year of engineering data. The total CPU power for reconstruction, batch-analysis and interactive processing is about 8000 SPECint95. All of this may increase at the rate of 20%/year. Currently access to DSTs and other data at DESY is at the rate of 20 TB/year, which is expected to increase to 40 TB/year.

B.3.1.4. International Bandwidth Requirements

The current bandwidth to DESY is carried to DESY via OC3/POS[5] as shown in figure 1. The US traffic travels via Frankfurt using 2 OC12/POS links to New York. At the moment, this is running at "best effort" only, meaning that the network traffic to and from DESY has standard priority on the shared link.

The estimated bandwidth needs are:

Table 1: ZEUS and other DESY required bandwidths to the U.S. in Mbps

	2001	*2002*	*2003*	*2004*	*2005*	*2006*
Interactive	**20**	**40**	**45**	**50**	**55**	**60**
Data Transfer	**30**	**50**	**60**	**70**	**80**	**90**
Total	**50**	**90**	**105**	**120**	**135**	**150**

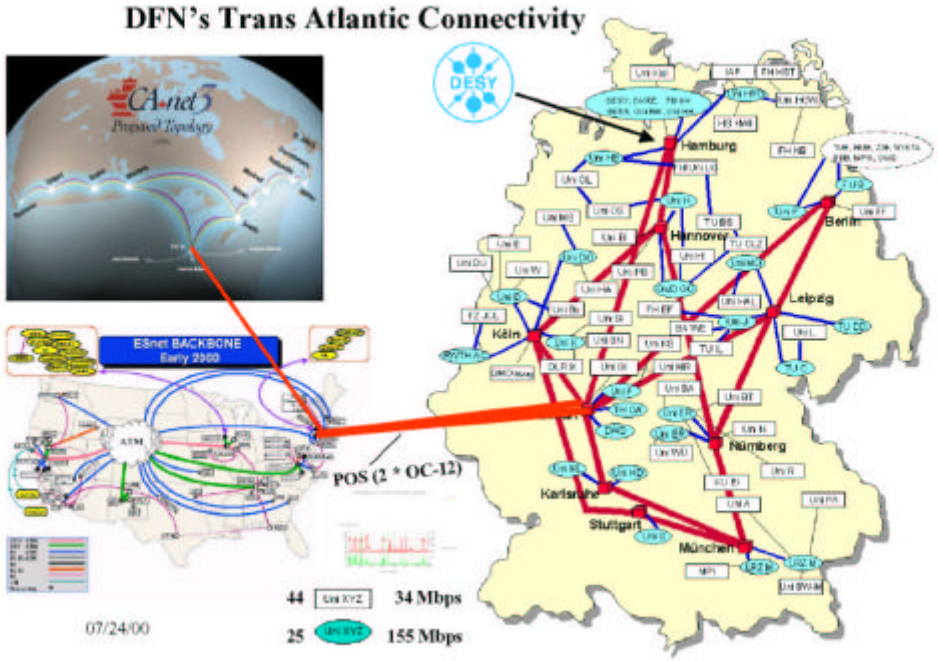


Fig. 2: DESY Network connectivity to the US and Canada.

[5] Packet Over Sonet, the current de facto standard for a high speed network link. OC3 denotes a raw bandwidth capacity of 155 Mbps, and OC12 denotes 622 Mbps.

B.3.2. *BABAR*

B.3.2.1. Overview of the *BABAR* experiment

The *BABAR* experiment is located at the asymmetric e^+e^- storage ring PEP-II at SLAC. One of its primary physics goals is the study of CP violation in B mesons. The *BABAR* detector consists of a silicon vertex tracker, a drift chamber, a particle identification system, a CsI electromagnetic calorimeter, and a magnet with an instrumented flux return. The *BABAR* Collaboration consists of approximately 600 physicists and engineers from 73 institutions in 9 countries. Approximately half the physicists are affiliated with U.S. institutions, and the other half with European institutions.

B.3.2.2. Schedule and Data Estimates

PEP-II and *BABAR* started operation in the summer of 1999. A peak luminosity of ~3 x 10^{33} cm^{-2} sec^{-1} was achieved in the first run ending in 2000, and *BABAR* accumulated ~25 fb^{-1} of data. The peak luminosity is expected to increase gradually to over ~1 x 10^{34} cm^{-2} sec^{-1} by 2005. Projections of year-by-year integrated luminosity depend on details of the running schedule as well as deployment scenarios of various PEP-II and *BABAR* improvements. They all give ~500 fb^{-1} by 2005. The following table is a typical profile from 2001 to 2006. Projections beyond 2005 are comparatively more uncertain until improvement and operational plans have been developed.

Table 2: BABAR luminosity profile

	2001	***2002***	***2003***	***2004***	***2005***	***2006***
Peak Luminosity (10^{33} cm^{-2} sec^{-1})	**4.0**	**5.0**	**8.0**	**11**	**13**	**16**
Yearly Integrated Luminosity (fb^{-1})	**40**	**50**	**85**	**110**	**130**	**130**
Total Luminosity (fb^{-1})	**65**	**115**	**200**	**310**	**440**	**570**

The effective cross section in *BABAR* of fully reconstructed events written to data store is ~12 nb.[6] This is made up of ~5 nb of hadronic events, and ~4 nb of calibration events (mostly Bhabha and ?-pair), the rest being other physics channels (such as ?-pair and 2-?) and background. The average event size is ~800 KB for a data volume of 9.6 TB per fb-1. We therefore expect 380 TB of new beam data in 2001, and a total of 620 TB including data from 2000.

BABAR plans to generate ~80 x 10^6 Monte Carlo events per month, corresponding to 770 TB per year, by late 2001. It provides a simulated data sample comparable in luminosity to the beam data. This Monte Carlo effort is distributed over approximately ten sites, with ~50% from RAL in the U.K. and another ~10% from elsewhere in Europe. The European production rate is therefore 460 TB per year.

BABAR plans to continue this 1:1 ratio of Monte Carlo to beam luminosity for later years. While the total Monte Carlo data volume is expected to scale simply with integrated luminosity from PEP-II, the mix of simulated events is expected to evolve. As analyses turn to rarer processes and/or more precise determination of parameters, we anticipate generating more Monte Carlo events per unit luminosity for these physics signals. This tends to increase the Monte Carlo sample. On the other hand, subtle detector effects that become more important with increasing statistics are difficult if not impossible to simulate properly. These will have to be obtained from beam data. This tends to reduce the Monte Carlo sample. We have assumed here that these two effects cancel, and use a simple scaling with luminosity.

B.3.2.3. Computing and Analysis Model

All the data acquired during a particular year are reconstructed that year. Furthermore, the continued refinements in reconstruction algorithms as well as calibration constants lead to yearly reconstructions of all previously acquired data. For example, we expect to log and reconstruct 40 fb^{-1} of new data in year 2001. We

[6] BABAR is similar to other e^+e^- experiments in having an inclusive trigger. The subset of triggered events that pass a subsequent off-line filter is written to data store. It is anticipated that this effective cross section will not change with luminosity, unlike what is generally expected for hadron colliders. .

also expect to reconstruct the 25 fb^{-1} of data acquired in year 2000. Thus, the total load in 2001 is given by their sum of 65 fb^{-1}. Similarly, we expect total loads of 115, 200, 310 and 440 fb^{-1} in years 2002, 2003, 2004 and 2005 respectively.

Analysis centers in *BABAR* are divided into three groups. Tier A centers represent major long-term commitments by their respective host countries. They are open to all *BABAR* collaborators, and each center hosts a substantial fraction of the entire *BABAR* data set. These data samples are expected to be largely non-overlapping. CC-IN2P3 will be fully operational as a Tier A center by the end of 2001, joining SLAC as the 2nd Tier A center. Tier B centers are regional in nature. Access is restricted to collaborators in that region. While they may contribute significantly to Monte Carlo production, Tier B centers are not expected to be major repositories for beam or simulated data. Tier C sites are typically universities. Resources are sized for the local group.

BABAR plans to transfer the data once from the source to their eventual repository. It will then be necessary during the analysis phase to move the applications to the data. This is more restrictive than some of the longer term Grid plans where the application and/or the data can migrate dynamically; however, the *BABAR* approach has more predictable network requirements. The following estimates are driven by transferring beam data from SLAC to CC-IN2P3, and by transferring Monte Carlo events produced in Europe back to SLAC.

B.3.2.4. International Bandwidth Requirements

BABAR expects to run ~9 months per year; thus the new beam data will be spread over this period. However, the data from previous years are not tied to the accelerator schedule. For the purposes of estimating network requirements, we will assume that the total data volume is produced uniformly over the entire year. The average data volume is 1.7 TB per day in 2001.

This data have to be distributed from SLAC to all its collaborating institutions. The primary overseas site at this time is CC-IN2P3 in Lyon, France. Assuming a factor of 2 between network bandwidth and actual realizable payload results in a 320Mbps requirement for 2001 from SLAC to CC-IN2P3. France will provide this bandwidth for 2001.

The projections in Table 2 for later years are based on the integrated luminosity figures in Table 1. We expect additional Tier A site(s) to be established in 2003 and beyond. This bandwidth requirement is not expected to change as a result, as any data that needs to be transferred should cross the Atlantic only once. The multiple Tier A sites in Europe may have to transfer data among themselves. However, the destination may not be only CC-IN2P3 in later years. At this time, there are no definite arrangements on cost sharing with European collaborators for 2002 and beyond.

Table 3: BABAR beam data volume and bandwidth requirement

	2001	***2002***	***2003***	***2004***	***2005***	***2006***
Total Beam Data Volume (PB)	**0.6**	**1.1**	**2.0**	**3.0**	**4.3**	**5.5**
Average Data Volume (TB per day)	**1.7**	**3.0**	**5.3**	**8.1**	**12**	**15**
Bandwidth Requirement (Mbps)	**320**	**560**	**1000**	**1500**	**2200**	**2700**

Monte Carlo production is assumed to be spread over the entire year. Production centers in Europe are expected to generate 1.3 TB of Monte Carlo data per day. Using the same factor of two to estimate necessary bandwidth yields a 230Mbps requirement for 2001. Note that this traffic goes in the opposite direction to the beam data, with the bulk of the data going from RAL to SLAC, so that this data flow does not generate additional bandwidth requirements.

Projections for later years in Table 3 are based on the expected integrated luminosity, and the fraction of the Monte Carlo production in Europe remaining unchanged. We have used a constant 60% in our projections, because there are no definitive profiles available. If Monte Carlo production shifts significantly to the U.S., it will reduce the bandwidth requirement from Europe to the U.S. It will simultaneously increase the bandwidth requirement from the U.S. to Europe as the simulated data must be made available in European Tier A centers.

Table 4: BABAR Monte Carlo data volume and bandwidth requirement

	2001	2002	2003	2004	2005	2006
Simulated Data Transferred (PB)	0.46	0.82	1.4	2.2	3.1	4.1
Average Data Volume (TB per day)	1.3	2.2	4.0	6.0	8.5	11
Bandwidth Requirement (Mbps)	230	410	720	1100	1500	2000

The average load from interactive use is expected to grow steadily over the next few years. Unlike bulk data transfer, it is expected to have large fluctuations. Peak loads are no more than a few percent of the bandwidth required for data transfer. However, good interactive response imposes additional quality of service requirements not necessary for bulk transfers.

In summary, *BABAR* expects to move ~1.7 TB per day from the U.S. to Europe, and ~1.3 TB per day from Europe to the U.S. in 2001. The corresponding transatlantic network bandwidth requirements are 0.32 and 0.23 Gbps respectively. These numbers are expected to increase steadily to 2.2 and 1.5 Gbps by 2005, with slower increases after that. Interactive usage is characterized by modest bandwidth and good response requirements.

B.3.3. Dzero

B.3.3.1. Overview of the Dzero Experiment

The Dzero collaboration adopted a distributed data processing model in which Monte Carlo data (MC) and beam data (Data) are processed at Fermilab and at big computing farms that are located mainly in Europe. Currently the Fermilab farm dedicated to processing Dzero data consists of approximately 200 750MHz-equivalent PC processors. Other farms offsite that are already actively generating MC events, located mainly in Europe, consist of about 600 similar PC processors in total. The number of processors is expected to increase significantly over the next several years.

B.3.3.2. Computing and Analysis Model

The data that have been acquired and reconstructed at Fermilab must be distributed from Fermilab to all the collaborating institutions, currently 72 in 17 countries (8 in Europe). The primary overseas sites at this time are NIKHEF in Amsterdam, the Netherlands (currently having ~200 processors), Lancaster University in Lancaster, United Kingdom (200), IN2P3 in Lyon, France (~50), and Institute of Physics (Academy of Sciences) in Prague, Czech Republic (~30). Simulated data are produced at many sites, and must be transferred to Fermilab for central archiving. It is expected that a sustained data transfer rate, after taking into account down times, overheads, etc., of about 1 TB per day is required in 2001 for each direction on a transatlantic link. The requirement is expected to double approximately every year.

All the data acquired during a particular year are reconstructed that year. Furthermore, the continued refinements in reconstruction algorithms as well as calibration constants lead to yearly reconstructions of all previously acquired data. In some cases, specific data streams are reconstructed more than once.

B.3.3.3. Schedule and Data Estimates

Dzero is expected to accumulate ~2 fb^{-1} of data in 2001-2003 (Run 2a), with most of the data collected in the second year and third year. After several months of shutdown to upgrade the accelerator and the detector, Dzero is expected to accumulate ~15 fb^{-1} of data in 2004-2007 (Run 2b), and roughly 5 fb^{-1} every year starting in 2005. The peak luminosity as well as yearly integrated luminosity are expected to increase with time.

The trigger rate in run 2a is expected to be 50 (20) Hz peak (DC). The number of events written to tape is going to be about 2 million per day. The average event size is expected to be about 300 KB for a data volume of about 300 TB during Run 2a. In Run 2b the luminosity is expected to increase by a factor of about 4; the trigger rate will increase accordingly. As the design of the upgraded detector for Run 2b is still in its initial phases, we

cannot comment on the event size and will assume no increase.

While we currently plan on reconstructing most of the data at Fermilab, we foresee the need to distribute it to collaborating institutions as well as to re-reconstruct some portion (specific streams) at offsite centers. The average rate of Data transferred from Fermilab to Europe (and back) is therefore expected to be about 3 TB per day in 2002.

In order to have a sample of simulated events comparable to the beam data sample, Dzero plans to generate ~(1-2) x 10^8 MC events per year during 2002 at offsite centers. We plan on generating the MC events, run them through the simulation of the detector, and process them with the official Dzero reconstruction program. Processing one MC event through the entire chain currently takes ~2min on one CPU processor. The output event size is about 1.2 MB depending on physics process type and luminosity.

This effort is currently distributed over 6 sites, with ~85% in Europe. The average rate of simulated data transferred from Europe to Fermilab is therefore expected to be about 3 TB per day in 2002. We also expect MC data transfer from the archival center at Fermilab to collaborating institutions, including the big centers. We currently estimate it to be about 1 TB per day in 2002.

B.3.3.4. International Bandwidth Requirements

The average rate of Data and MC transferred from Fermilab to Europe and back in 2002 is expected to be about 7 TB per day. Estimating the needs for international networking requirements for Dzero for the next several years is hard. Our current best estimate is the following:

Table 5: Dzero data volume and bandwidth requirement

	2001	*2002*	*2003*	*2004*	*2005*	*2006*
Simulated Data Transferred (TB per day)	**0.5**	**2.5**	**3.7**	**5.0**	**10**	**12**
Average Data Volume (TB per day)	**2**	**7**	**10**	**14**	**28**	**35**
Bandwidth Requirement (Mbps)	**400**	**1600**	**2400**	**3200**	**6400**	**8000**

Note that Holland and the U.K. are presently providing substantial transatlantic links to handle the data transfer associated with their computing centers for Dzero. They hope, but have not committed to continue and upgrade these links during the lifetime of the experiment to meet the projected growth in requirements. Although the transatlantic links may be provided by foreign partners, U.S. networks must plan to provide appropriate connections to these foreign links, as well as to FNAL and other Dzero institutions.

B.3.4. CDF

B.3.4.1. Overview of the CDF experiment

The Run 2 Tevatron Collider at Fermilab will operate at the energy frontier by providing proton-antiproton collisions at 2 TeV in the center-of-mass, peak luminosities of $5x10^{32}cm^{-2}$ sec^{-1} and bunch spacing as short as 132 ns.

The purpose of the CDF Upgrade Project (CDF II) is to upgrade the CDF detector to exploit the physics opportunities available with an upgraded Tevatron Collider for Run 2.

The CDF II detector will carry out a broad physics program that addresses the most fundamental questions in modern high-energy physics. This program includes the search for the Higgs boson and the origin of the electroweak symmetry breaking, studies of the W and Z bosons, precision measurements of the top quark mass and its decay properties, QCD studies, precision studies of the CKM matrix elements, studies of particles containing a b quark, CP violation in the B sector and Bs mixing. CDF II will also search for the Higgs boson and the origin of the electroweak symmetry breaking.

The CDF II detector employs three precision tracking chambers in a large solenoidal magnet: the Silicon micro-vertex detector (SVXII), the Intermediate Silicon Layers (ISL) and the Central Outer Tracker (COT). Surrounding the magnet coil is the electromagnetic calorimeter, which is in turn surrounded by the hadronic calorimeter and muon chambers.

The CDF Collaboration consists of 525 physicists from 52 institutions representing 11 countries. About 170 physicists belong to European institutions.

B.3.4.2. Schedule and Data Estimates

CDF II started operations around April 2001. The initial instantaneous luminosity during the first operation run was about 0.6 ? 10^{30} cm^{-2} sec^{-1}. After the first tuning period, expected to be completed by the end of 2001, CDF will accumulate data for two or three years (Run 2a) in production mode for an integrated luminosity of 2 fb^{-1}.

In 2004 an upgrade of the experiment is foreseen to improve detector performance and make it compatible with the higher luminosity running conditions of Run 2b (2004-2007). The laboratory has in fact issued a memorandum stating a goal of providing a total integrated luminosity of 15 fb^{-1} per experiment by 2007.

In the table we give a luminosity profile from 2001 to 2006.

Table 6: FNAL Run II Luminosity Projection

	2001	***2002***	***2003***	***2004***	***2005***	***2006***
Luminosity ($10^{32}cm^{-2}$ sec^{-1})	**1**	**8**	**12**	**20**	**40**	**50**
Integrated Luminosity (fb^{-1})	**0.1**	**1.0**	**2.0**	**3.0**	**6.3**	**10.5**

For Run 2a, the average raw event size is about 250 KB while the reconstructed event (DST) size is about 135 KB. The DSTs are divided in about 40 streams, the primary data sets, and written on permanent storage. Together with the DSTs, the reduced Physics Analysis Data Set (PAD) is also produced. It contains enough info for a good portion of the data analysis. These are the data sets most institutions are interested in. They will very likely travel over the net. The typical PAD event size is of about 60KB.

Average trigger rates for the recorded events are estimated to be 30Hz (= 10^9 events/year). So the expected total data volume for Run 2a is of about 1PB.

For Run 2b, the data acquisition rate will increase up to 100Hz and we can expect to collect three times the amount of data collected for Run 2a.

B.3.4.3. Computing and Analysis Model

CDF has chosen a centralized computing model where all data are stored at FNAL in the central computing and storage facility. The Raw data coming from the detector are divided into 8 different streams by the Level 3 trigger and sent via optical fiber to the Feynman Computer Center (FCC) and stored on tapes.

All events are then reconstructed on a PC farm and the DST data is divided in 40 streams constituting the primary data set. Together with the DSTs the PAD data set is produced and stored on tape. The total volume of PADs for Run 2a is about 200TB. All data are available to the collaboration for processing directly at FCC.

The Physics Groups will produce finer selections of events in PAD format: the secondary data set. These will reside on disk. The space available for those will be about 20TB.

The final users will work on PADs available on disk or tape to produce finer selections, the tertiary data set, or n-tuples. The n-tuples can be copied on local desktops or at the remote institutions.

To realize this analysis chain a good analysis facility has been setup at FCC. It is based on a robotic tape library with a total capacity of more than 1PB, a pool of medium-size SMP UNIX servers with a good connectivity to the robotic library and a pool of disks for a total of 80TB of space in a SAN[7] configuration. Remote institutions are encouraged to contribute to the FNAL computing facility and do their analysis at FNAL.

[7] Storage Area Network.

The European remote institutions of the CDF collaboration are located in Italy, Germany, Spain and Switzerland and UK. These institutions are not planning to replicate the entire PAD data sample at their site. For the UK, most of the data brought over from FNAL on tapes will be placed on a central storage/processing facility at RAL (about 20 TB). The other UK groups will then run production/stripping jobs, and will send the parts of the data specific to their analysis needs to their home disk areas (about 2-5 TB at each institution). The systems at home institutes ought to have sufficient I/O and computing power to churn through a 50-100 GB n-tuple in about 10 minutes. The n-tuples will be mostly generated locally for the UK institutions. The UK groups would like to be able to access the latest data samples over the net in a timely manner.

For the other European institutions, the analysis model is slightly different. N-tuples of size 1-100 GB will be mostly generated at FNAL and copied back onto remote machines for final analysis. Some PAD data samples will still be copied locally. Depending on the analysis type, the PAD data set required could vary between 100 GB-1 TB, as for the high P_T selection, up to 1-10TB, as for low P_T.

The computing and analysis model just summarized could change for Run 2b, to a more distributed model, because the data volumes and required CPU power for analysis will increase substantially. The network requirements also could change drastically.

B.3.4.4. CDF Network Bandwidth Requirement

In order to guarantee that remote institutions are able to work in an efficient and competitive manner, the transatlantic network bandwidth available for CDF should be sufficient to allow good interactive response, and a transfer time for 50-100 GB files of the order of 1 hour (or less) for Run 2a. Assuming that there are about 20 active users at a given time at remote institutions such as Italy and UK and 2 to 5 users in the other countries, an estimate of the network needs is given in Mbps in the table below for the years 2001-2003. This is not dedicated bandwidth, but it should be available on demand.The current CDF data and event model was developed in the context of quite poor performance of the network to Europe and Japan. In this kind of environment the most efficient method of analysis requires significant central resources that remote collaborators can use by logging in remotely to the central systems. In this way even users with poor bandwidth connections can at least take advantage of significant central compute resources.

In the future it is expected that the CDF collaboration will be presented with a proposal to incorporate CDF and many collaborating institutions into the Grid computing architecture. This model is less centralised and relies more heavily on network resources in order to transfer data sets and metadata so that the entire CDF-Grid can remain apprized of the resources available at all participating sites. In particular we are expecting approximately 2000-2500 Mbps continuous transfer rates and data volume movements to Europe, corresponding to approximately 200 TB of data transferred per year.

Table 7: Network bandwidth needed on-demand for CDF RunIIA/B in Mbps

2001	*2002*	*2003*	*2004*	*2005*	*2006*
50	**150**	**200**	**2000**	**3000**	**6000**

B.3.5. CMS

B.3.5.1. Overview of CMS

The Compact Muon Solenoid (CMS) collaboration currently includes 1850 physicists and engineers from 144 institutes in 31 countries, including 390 physicists from 38 institutions in the US, and is expected to grow in size significantly by the start of LHC operation. CMS will exploit the full range of TeV-scale physics made accessible by the LHC. CMS will use its high resolution crystal calorimeter, muon, tracker and hadron calorimeter systems to study the Higgs sector, search for supersymmetry and a host of other hypothesized new particles, and detect other signs of new physics including the onset of effects from extra dimensions of space-time, by precisely measuring rare events containing electrons, muons, photons, jets and missing energy.

B.3.5.2. Schedule and Data Estimates

CMS, situated at Point 5 at the LHC, will begin data taking at the LHC's pilot run, scheduled in April 2006, followed by the first physics run at the LHC starting in August 2006.

A brief list of the major CMS data production and analysis milestones is given below.

??	Nov 2000	Level 1 Trigger Technical Design Report (TDR): Completed.
??	Nov 2002	DAQ System TDR
??	Dec 2002	5% Complexity[8] Data Challenge Complete
??	Dec 2002	Delivery of the core software for preparation of the Core Computing and Software, and the Physics TDRs
??	Dec 2003	Core Computing and Software TDR
??	Dec 2004	Physics TDR: Large scale interactive analysis; using GEANT4;
??	Dec 2004	20% Data Challenge Complete; Final test before purchasing the production systems in 2005-7.
??	Dec 2005	Computing, database and Grid systems in place ready for the LHC Pilot Run; fully operational computing systems at 20% capacity
??	Mid-2006	All systems (software, computing, networking, remote collaboration) ready for the first LHC Physics Run (August 2006 – February 2007)
??	Dec 2007	Computing systems fully operational at 100% of the baseline capacity planned for the first year of LHC operation at design luminosity

The CMS online systems will filter the 10^9 interactions per sec produced by the LHC using dedicated hardware in the first level trigger (output 75 kHz), and farms of ~1000 online processors to reduce the recorded rate to 100 MB/sec of raw data. This will result in ~10 PB of accumulated raw and processed data stored in the first year of operation, in 2006. The data volume is expected to increase rapidly in subsequent years, so that the accumulated data volume will reach 1 Exabyte[9] by around 2015. This data is to be accessed, analyzed and processed using a Data Grid system that extends over a worldwide ensemble of national, regional and local facilities[10] as the physicists' reconstruction, calibration and search strategies and algorithms evolve over time.

B.3.5.3. Computing and Analysis Model; Software and Grid R&D Status

CMS has been a leader in the development of the "Tiered" computing model adopted by all four LHC experiments[11], shown in the figure below. The bandwidths shown in the figure are the baseline requirements for a single LHC experiment (CMS or ATLAS), corresponding to the US-CERN link reaching 10 Gbps by 2006. These requirements and the likely need for greater bandwidths on the major national and transoceanic links are discussed further elsewhere in this report.

Members of CMS in the US and in Europe, working with PPDG, the EU DataGrid and GriPhyN projects have produced the Grid Data Management Pilot[12] system (GDMP). GDMP is now used to support large scale distributed production of simulated and reconstructed events among an increasing number of sites (currently 10) in the US, Europe and Asia. Production cycles are scheduled two to three times per year, in support of the development of the experiment's high-level trigger and Physics Reconstruction and Selection (PRS) activities, and for studying in depth the detector's capabilities for discovering new physics. The Spring 2001 production cycle is using approximately 1000 CPUs, leading to an estimated 20 Terabytes of data stored.

Beyond distributed production services, CMS will extract, transport, coalesce, process and analyze selected Terabyte-scale object-collections from distributed Petabyte-scale data stores managed by Object Database systems (ODBMS). CMS software is well-advanced in this direction, and is fully functional in this mode over local area networks. A detailed design to integrate the CMS frameworks with Grid software[13], to allow CMS physicists to carry out their data analysis efficiently over wide area networks, also is underway. This mode of transparent access to manageable subsets of massive data stores has no analog in current practice by running

[8] The number of major processing and data handling components (boxes), relative to the Tier0 system to be fully commissioned at CERN by 2007.

[9] 1 Exabyte = 10^{18} Bytes

[10] This system is being developed by a coordinated effort of the major Grid projects in the US and Europe, including the Particle Physics Data Grid (See http://www.ppdg.net), the Grid Physics Network (See http://www.griphyn.org) and the EU DataGrid Project (see http://www.eu-datagrid.org).

[11] As mentioned in the "Report of the Steering Group of the LHC Computing Review" (see CERN/LHCC/2001-004 at http://lhc-computing-review-public.web.cern.ch/lhc-computing-review-public/Public/Report_final.PDF) the requirement for each Tier0-Tier1 link is 1.5 – 3 Gbps, with the Tier1-Tier2 links being similar. This report also considered the effects of higher bandwidths, such as 10 Gbps Ethernet integrated with OC-192 WAN links (see http://www.10gea.com).

[12] See http://cmsdoc.cern.ch/cms/grid/ on GDMP Version 1.2.2 (April 2001).

[13] See "CMS Virtual Data Requirements" (June 2001) http://kholtman.home.cern.ch/kholtman/tmp/cmsreqsv11.pdf.

HEP experiments, but will be important to allow physicists at sites around the world to fully exploit LHC physics starting in 2006.

CMS has completed the second major object-oriented software development cycle, the "functional prototype" phase, in the development of its software frameworks (COBRA[14], CARF), reconstruction code (ORCA, now in its fifth major release) and its interactive graphics analysis environment (IGUANA). Persistent objects are handled using Objectivity as the baseline ODBMS, and transparent access by users to a distributed federation of objects, where each user is able to seamlessly use a private schema. The third major software development cycle, leading to "fully functional software" is now underway.

B.3.5.4. International Bandwidth Requirements

For the LHC Computing Review, CMS developed a breakdown of the baseline needs for data throughput and installed bandwidth on each of its Tier 0-Tier 1 links, summarized below. This is based on the LHC Computing Models and simulation studies of the MONARC project[15] at CERN, extended to consider the data flows to/from Tier2s and a modest amount of data transfer to Tier 3s (institute workgroup servers) and Tier 4s (desktops)[16]. As mentioned elsewhere in this report, this level of bandwidth does not take the effects of running the ensemble of facilities as an integrated Grid system into account.

1)	Tier 1 ? Tier 0 Data Flow for Analysis	0.5 - 1.0 Gbps
2)	Tier 2 ? Tier 0 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 (Tier 3 or Tier 4) data transfers Limit to 10 Flows of 5 Mbytes/sec each	0.8 Gbps
	TOTAL Per Tier 0 - Tier 1 Link	1.7 - 2.8 Gbps

Based on the major milestones summarized above, and the experience gained in CMS distributed production and distributed analysis prototypes, the estimated baseline bandwidth needs between CERN and the Tier 1 center at FNAL (in Mbps) are[17]:

Table 8: CMS Bandwidth Requirements in Mbps

	2001	***2002***	***2003***	***2004***	***2005***	***2006***
Installed BW	**100**	**200**	**300**	**600**	**800**	**2500**

B.3.6. ATLAS

B.3.6.1. Overview of ATLAS

The Large Hadron Collider (LHC) at CERN will open a new frontier in particle physics due to its higher collision energy and luminosity compared to existing accelerators. The guiding principle in designing the ATLAS experiment, one of the two major LHC experiments, has been maximizing the discovery potential for new physics such as Higgs bosons and supersymmetric particles, while keeping the capability of high-accuracy measurements of known objects such as heavy quarks and gauge bosons. The central goal is to search for understanding of the mechanism that generates the masses of the W and Z gauge bosons and of quarks and leptons.

[14] Coherent Object-Oriented Base for Reconstruction, Analysis and Simulation. See http://cobra.web.cern.ch/cobra/).

[15] "Models of Networked Analysis at Regional Centers". See www.cern.ch/MONARC. The final report of the project's Phase 2 (March 2000) may be found at this Web site.

[16] These estimates have been adopted by the LHC Computing Review as an estimate of the bandwidth required for each of the LHC experiments, between the CERN Tier0 and each Tier1 center.

[17] CMS as a whole also has significant needs for networks to Asia, where Tier2 centers are planned in (at least)_India, Pakistan and China. However there is no expectation that this will substantially impact the network bandwidth requirements to and from the US.

ATLAS is being constructed by 1850 collaborators in 150 institutes from 32 countries around the world. The ATLAS detector consists of an inner tracking system with silicon pixels, silicon strips and a transition radiation tracker (TRT); a liquid argon electromagnetic and forward calorimeter; a scintillating tile hadronic calorimeter; a muon spectrometer, and a trigger and data acquisition system. There are superconducting solenoid and toroid magnets to allow sign determinations and momentum measurements for charged particle products of the collisions. U.S. groups, from 33 U.S. universities and national laboratories, are involved in almost all of these components of the ATLAS detector.

B.3.6.2. Schedule and Data Estimates

ATLAS data will be divided into classes according to the degree of processing that has taken place and the frequency of access that is expected during analysis. Raw data estimates are based on a provisional average event size of 2 MB, though it is hope that can be reduced with initial experience of calorimeter reconstruction. Trigger rates for recorded events are estimated to be 100 Hz during initial running in 2006, growing to 270 and 400 Hz by the end of 2007 and 2008 respectively. Thus raw data sets (based on 10^7 seconds per year) will be approximately 2, 5.4, and 8 PB for the first three years of data taking.

Only small samples of raw data will be transferred to the U.S. Tier 1 center. The data sets to be transferred are the Event Summary Data (estimated at 0.5 MB per event), Analysis Object Data (10 KB per event) and event tag metadata and smaller data sets. Using these numbers with the estimates above gives 0.5, 1.4, and 2 PB for the main analysis data in the first three years. It is expected that these numbers set a scale, but that total data transfer will be at least 5 times these numbers.

The schedule for ATLAS is indicated by the sample milestones that follow:

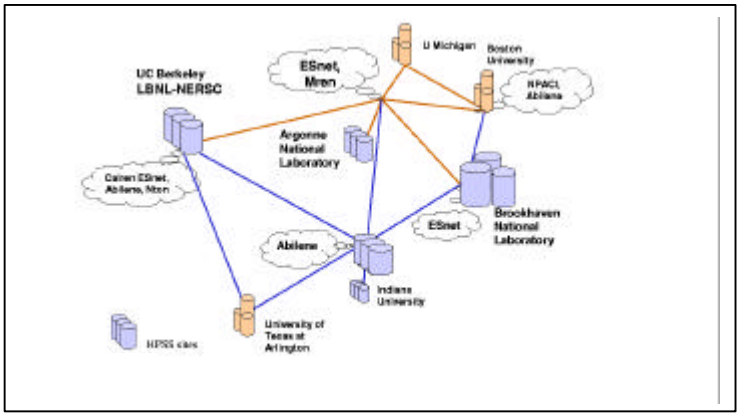
??	Tier 1 Storage Prototype	10/1/2001
??	Mock Data Challenge (MDC) 0 Completed	12/12/2001
??	MDC 1 Completed	7/30/2002
??	Computing Tech. Design Report Finished	11/29/2002
??	Tier 1 Upgrade (for MDC2)	12/31/2002
??	Tier 1 Large Scale Test; MDC2 (25% of turnon capacity)	9/30/2003
??	Full software chain in real environment	7/30/2004
??	Full Database infrastructure available	12/31/2004
??	20% Processing Farm Prototype	9/30/2005
??	Tier 1 Full scale Operation	10/2/2006

B.3.6.3. Computing and Analysis Model

ATLAS will follow the hierarchical LHC computing model, with raw data archived at the Tier 0 center at CERN. Current planning provides for six Tier 1 centers, constituting major national or regional computing and data staging centers. The U.S. center, at Brookhaven National Laboratory, will be the largest of the Tier 1 centers corresponding to the status of the U.S. as the largest national participant in ATLAS. Current planning for the Tier 1 center at the turn on of ATLAS data taking provides capacities of 209K SPECint95 CPU, 365 Tbytes disk, and 1.8 Pbyte tape storage. The U.S. Tier 1 center will be complemented by five Tier 2 centers located at different points in the U.S., each of which will have capacities nominally 20% of the Tier 1 center.

International networking will be needed in the first instance to move data from CERN to the Tier 1 center at BNL. As discussed above, it is expected that at least 2.5 PB of all types of data will be transferred per year at turn on in 2006, growing with time at least by a factor of 3 when the LHC reaches design luminosity in 2007 or 2008 and continuing to grow after that. Lower speed connections directly to CERN and other international collaborators will be needed from the Tier 2 centers and from individual institutions in the U.S. Although the aggregate data rates to international sites are dominated by the Tier 0 to Tier 1 traffic, it is also a strong requirement that the international links are well connected to ESnet, Abilene, and other U.S. networks that may carry traffic to our laboratories and universities. The degree of connectivity needed is illustrated by the figure of the U.S. ATLAS testbed already in operation.

Fig 3: Current U.S. ATLAS Grid testbed



B.3.6.4. International Bandwidth Requirements

Analysis of the operations of the Tier 1 and Tier 2 centers as a function of time leads to the following table of bandwidth requirements. The transatlantic requirement is show in the line for Tier 1 – CERN, while the Tier 2 centers sum the required bandwidth from the Tier 2 centers to Tier 1 and to the regional users of the Tier 2 centers.

Table 9: U.S. ATLAS Required Bandwidths to Regional Centers in Mbps

Link to:	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*
Tier 1 - CERN	50	100	200	300	600	2500
Tier 2a	50	150	300	600	1200	2500
Tier 2b		150	300	600	1200	2500
Tier 2 c-e				600	1200	2500

B.3.7. BTeV

B.3.7.1. Overview of BTeV

BTeV is a third-generation B-physics experiment that will operate at the Fermilab Tevatron in the era of the LHC. The BTeV detector is designed as a two-arm spectrometer, taking advantage of the unique kinematics of forward production in?pp and the enormous cross section for producing b-flavored mesons and baryons to probe questions that will be inaccessible at the SLAC and KEK asymmetric B-factories. BTeV achieves its physics reach with a very advanced detector design, highlighted by a separated-vertex pixel-detector trigger that achieves efficiencies of 50-74% for critical decay modes. Additional features of the detector that are key to its reach are an excellent lead-tungstate calorimeter inspired by that of CMS, particle identification by RICH, excellent tracking and muon identification, and a data acquisition system capable of reading out 1 kHz of B events.

The BTeV collaboration currently consists of 30 groups and ~115 physicists, with approximately one quarter of collaborators located in Europe or Asia. It is expected that the membership of the collaboration will grow steadily in the next several years to a level that is comparable to CDF or D0, and much smaller than ATLAS or CMS. The data-sample size is expected to be about ~2 Pbyte per year, comparable to the LHC experiments. Therefore we expect BTeV to be a major network user beginning in the last third of this decade.

B.3.7.2. Schedule and data Estimates

Since BTeV will run after the completion of Tevatron Run 2b, it is not expected to have a major impact on

international and domestic networking within the time horizon of this report. During the next five years BTeV requirements will gradually ramp up, driven by extensive simulations, analysis of test-beam data, frequent use of videoconferencing and other collaboration tools, and the development of structures and procedures for management of code and data in preparation for initial running. It is expected that sizable Monte Carlo samples will be transferred between collaborators in the US and Italy during the current calendar year. While quantitative estimates are difficult at this time, the recent experience of other major projects leads to the projected average bandwidth requirements through 2006 that are shown in the table below.

Table 10: BTeV Required Bandwidth to Europe in Mbps

2001	*2002*	*2003*	*2004*	*2005*	*2006*
20	**40**	**100**	**200**	**300**	**500**

B.4. Summary of requirements

This section summarizes the bandwidth needs-projections, technology choices, and approximate projected costs for the major transoceanic links and domestic links to the US HEP labs, as well as links to universities.

Beyond the simple requirement of adequate bandwidth for large file transfers of large files, physicists in DOE/HEP's (and NSF/EPP's) major programs require (1) an integrated set of local, regional, national and international networks able to interoperate seamlessly, without bottlenecks, (2) networks that will accommodate and satisfy the latency and jitter requirements of realtime applications coexisting with high-bandwidth data transfers, (3) network and user software that will work together to manage the bandwidth effectively, and (4) a suite of videoconferencing, shared application and other high-level tools for remote collaboration that will make data analysis from the US (and from other remote sites) effective. For US physicists involved in experiments at CERN, keeping a strong focus of physics activities in the US has led to the requirement for ``collaboratories'' at the US HEP labs and at some universities (e.g. the Tier1 and Tier2 center sites), where US physicists can participate in the running and monitoring of the experiment.

Two levels of bandwidth are foreseen:

?? Baseline requirements based on a relatively conservative view of technology evolution, sufficient for the experiments' needs for distributed data access, processing, delivery and analysis, in a strictly managed environment.

?? Requirements to satisfy the needs for access to significant data subsets (0.1 to 1 TB scale) on-demand, by individuals and small working groups, and for automated data movements within a unified Grid environment, and for such new working methods as widespread "persistent" remote collaboration throughout the working day. This corresponds to a best guesstimate (a medium-optimistic view) of the evolution of technology and cost-performance over the next five years.

The baseline bandwidth needs of the major experiments, introduced in the previous sections and discussed extensively by our Working Group, are summarized in the following table, in Mbps. The table sums the bandwidth requirements for reference and then shows the transatlantic bandwidth which is not expected to be supplied by foreign collaborators as the bandwidth expected to be needed on the U.S.-CERN connection. The table gives the projected bandwidth requirements for each major experiment to the experimental site, where the largest computational and data handling facilities are or will be located. In the case of the LHC experiments, the US ATLAS and US CMS contingents will share network access to the CERN laboratory with ALICE, LHCb and the non-LHC programs at CERN. The US-CERN link also carries a significant portion of the network traffic to Europe, for *BABAR*, Run2 and other major HEP programs, so that the US-CERN bandwidth requirement is expected to reach the 10 Gbps range by 2006, if not earlier.

The figures in the table are the baseline requirements for installed bandwidth, where we assume, based on experience in Internet2 and elsewhere that stable network operation requires a network occupancy at 50% or less. The projected requirements for the actual sustained data rate for an experiment, averaged over the long term, are thus one-half of the figures shown[18,19].

[18] Discussions with Qwest on their next generation network plans have indicated that the maximum sustained occupancy of an IP-based wide area network in the Gbps range should be 40%. Given the uncertainties in the baseline requirements figures, and the newness of IP networks at these speeds, we have decided to keep the figures in the table based on 50% occupancy.

Table 11: Summary of installed bandwidth requirements (in Mbps) by experiment

	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
Dzero	400	1600	2400	3200	6400	8000
BTeV	20	40	100	200	300	500
DESY	100	180	210	240	270	300
Total Bandwidth	1070	3020	4810	8440	13870	22800
US-CERN BW	155-310	622	1250	2500	5000	10000

It should be noted that the table represents minimum requirements, in a strictly managed working environment. We have averaged among experiments and reduced some estimates presented to our Committee, to come up with the baselines. Specifically, we have not included support for the following activities in the table:

?? Sufficient bandwidth for *BABAR* to reprocess and analyze data at a rate substantially larger than the data taking rate, during peak periods.

?? Data distribution, processing and access by ATLAS at a rate corresponding to the trigger rate foreseen: 270 Hz at 10^{33} luminosity, and 400 Hz at the design luminosity of 10^{34}, though these rates are projected for times after that covered in the table**.** Up to the present, the Hoffmann Review on LHC Computing has assumed a reference trigger rate of 100 Hz for both ATLAS and CMS.

?? The impact of other programs, such as NuMI and the rest of the Fermilab fixed target program, and the recently proposed CLEO-c.

Although there is a strong trend towards more reliance on the network, we note that there are significant differences in the fraction of total data assumed to be transferred across the network, reflecting significant differences in Computing Models among the experiments. The CDF model for Run 2a is relatively highly centralized with only modest amounts of data being sent overseas via networks. The CDF numbers for Run 2b reflect a general expectation that data grid systems will be employed in those later years in a much more decentralized model. The D0 model, both for real and simulated data, is substantially more distributed already for Run 2a. The large numbers for D0 and also for CDF in Run 2b, starting in 2004, are notably larger than the LHC experiments in the same years. Although D0 has serious plans starting with Run 2a, it should still be noted as a caution that neither Run 2 experiment has experience with this model of distributing data. The LHC experiments' models are structurally in between the CDF and D0 Models, in that the data flow across networks is assumed to be reduced by providing efficient access to subsets of the data through the use of Object Databases coupled to data grid systems. This implies a large software and database development effort by the experiments over the next few years,[20] but will keep the network bandwidth reasonable if it is successful.

Several of the experiments listed would benefit substantially from additional bandwidth, leading to better use of the available resources and greater working efficiency. Greater bandwidths would allow freer access to the data and processed results by individuals and small workgroups, more extensive data distribution among the sites, and greater flexibility in treating the ensemble of site facilities for data analysis as a coherently managed "Grid" of resources.

[19] It should also be noted that the peak needs for the largest current users of the network, *BABAR* and CMS are actually higher than half the figure in the table. Greater bandwidths would reduce the file transfer times from several hours to the 1-2 hour range, where the probability of successful completion of the transfer would approach 100%.

[20] If this software effort is less than fully successful, a likely consequence would be increased network needs for the LHC program.

The purpose of the baseline figures is to establish a bandwidth level at which experiments will be able to function, within a moderate budget envelope. In case technology evolution and market forces lead to lower unit prices than assumed in the baseline, it is recommended that the experiments be allowed to optimize their network installations and modes of operation to achieve greater working efficiency in extracting physics results.

The figures above for *BABAR* roughly correspond to the needs for connectivity into the SLAC site, since large-scale data transfer needs between SLAC and IN2P3, as well as other Regional Center sites, dominate those needs[21]. In the case of FNAL, one has to consider the sum of the D0, CDF, and BTeV needs, a major part of the CMS needs, and the needs of MINOS and other FNAL programs. At BNL the needs for the RHIC program as well as ATLAS need to be supported

We note that the figures above are roughly consistent with earlier network requirements estimates, including those of the ICFA-NTF[22] (1997-8) and the needs recognized by the ICFA-SCIC (1998), although in detail they are larger. This reflects the rapid advance of network technology over the last five years, the outlook for rapid reductions in the unit price of bandwidth between 2001 and 2006[23], greater network-awareness of the major HEP collaborations, and greater understanding in detail of the data analysis-related tasks to be performed.

B.5. Meeting the requirements

B.5.1. Baseline Model for Bandwidth Evolution

We have used the baseline requirements given above to derive the required connectivity to each HEP laboratory (using ESnet) as a function of time. We have expressed the requirements in term of the current standard optical link services available on the market: OC3 (155 Mbps) through OC192 (10 Gbps). In general the next higher level of bandwidth service costs approximately 2.5 times the cost of the service below it[24], and this is reflected in the entries chosen in the table. We also show the transoceanic links used by HEP for which the US (DOE/MICS, DOE/HEP and NSF/CISE through the Eurolink program) currently provide funding.

Table 12: Summary of Installed Bandwidth Requirements by Laboratory

	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**

Using the estimated bandwidths shown in the table efficiently will require an aggressive R&D program, and the installation of new routing and switching equipment, as well as large data servers able to read and write data at speeds corresponding to the bandwidths in the table. The challenge at Fermilab appears to be quite significant where the aggregate rate of reading and writing data across networks is estimated to reach 1 Gbyte/sec before 2005.

Each of the HEP labs will need to be connected, at a speed comparable to the entry given in the table above, to one or more sites which are the termination points of transoceanic links. This includes the US-CERN link as well as links for *BABAR* and Run2 funded by European partners to the national networks in France (RENATER), Italy (GARRNet) and the Netherlands (SURFNet). The connectivity to ANL and LBNL, which

[21] Note that the bandwidth required to reprocess data in shorter periods than the time it takes to acquire the data in the first place are not included in these "baseline" figures.

[22] See the ICFA-NTF Requirements WG Report (May 1998) at http://l3www.cern.ch/~newman/icfareq98.html.

[23] Estimates of the yearly cost reduction of bandwidth range from 25 to 50%. The current precipitous drop in prices is expected to moderate in the coming years, as vendors must recoup their investments (a national aggregate of hundreds of billions of dollars) in the build out of national and transoceanic fiber infrastructures. See for example "Busted by Broadband", Time magazine, March 26, 2001.

[24] The cost of the highest speed links may be disproportionately large at the upper end of the range of commercially available links (OC192 at the time of this writing). This is because of the special engineering required for their implementation, and/or for the relatively rare expertise required for vendors to support them. Routers and switches for these links (such as the Juniper M160 equipped with OC192 ports today) also are relatively expensive.

are involved in FNAL Run 2 and will be Tier 2 centers for ATLAS, are expected to be covered by the existing links planned by ESnet: OC48 now, rising to OC192 well before 2005.

Universities are assumed to be connected to the HEP labs, and to the US-CERN and other transoceanic links, using the Internet2 backbone Abilene, and later Internet2's next-generation backbone, together with ESnet. Considering the aggregate bandwidth required between ESnet and the Internet2 backbone, connection at the major peering points needs to be at least OC48 by 2002, and OC192 by 2005 if not earlier[25].

?? Beyond the requirements for raw wide area network bandwidth, we also need to consider that effective use of this bandwidth requires:Periodic upgrades of the routing and switching equipment at each laboratory and university site.

?? Timely upgrades of the US Regional Networks connecting to the Internet2 backbone.

?? Timely upgrades of campus local area network infrastructures, and/or the installation of special links to the Regional Network.

?? The use of standardized tools and methods to make effective use of the available bandwidth[26,27]. These include software tools for tuning the TCP/IP protocol stack for high throughput, monitoring tools to monitor WAN status and overall performance, and possibly QoS (Quality of Service) settings designed to manage the relative priorities of traffic flows while meeting the latency and jitter requirements of interactive and realtime applications.

Because of the relatively large bandwidth requirements for some university groups, such as those hosting Tier 2 centers for the LHC or Tier B centers for *BABAR*, these groups will have to develop effective coordination and an ongoing working relationship with their campus network organizations. In cases where the network usage level associated with a Tier 2 (or Tier B) center represents a substantial fraction of the overall campus usage, DOE/NSF funding may be required for part of the connection to the Regional Network, and/or for network equipment (such as a router) or network equipment-upgrades needed specifically to support the center[28].

Connectivity between the Tier 1 and Tier 2 centers, over ESnet, Internet2 and the US Regional networks, would have to be scaled appropriately to the bandwidths and levels of network service given above. In the LHC hierarchical model, the required connectivity between a Tier 1 and a Tier 2 in the US (as well as in Europe and Japan) has been estimated to be in the 0.6 Gbps to 2.5 Gbps range. Realistically, the bandwidth available between Tier 1s and Tier 2s will depend on the state of development of Internet2 and the Regional nets; and on the Regional nets for the Tier 2 to Tier 3 connections. As indicated above, data transfers of 0.1 TB – 10 TB, which are relevant for large scale HEP data analysis by both large and small groups, and individuals in some cases, will require links in the range of 1 Gbps and up.

B.5.2. Funding Plan for the US-CERN Link

In response to one of the main elements of the charge of our Committee, a detailed funding plan for the US-CERN link was developed by Caltech and reviewed by the Committee. This follows the LHC experiments' and the ICFA-NTF's and ICFA-SCIC's determination of future network requirements over the last four years. This plan has been designed to meet the baseline needs for US-CERN transatlantic networking: both for the LHC experiments and for DOE/HEP's other major programs.[29] Similar plans will have to be developed for some of the principal domestic links, in the context of ESnet, plus any costs for links to Tier 2 centers (discussed briefly

---

[25] Matching ESnet capabilities to the next-generation Internet2 backbone speeds would require that ESnet's major links should reach OC192 by 2003; considerably earlier than is indicated by our HEP baseline requirements estimates alone.

[26] See for example the Web100 Project, at http://www.web100.org/, "a research effort aimed at bringing data transmission rates of 100 megabits per second to the desktops of researchers" funded by NSF. It should be noted that extending these methods to throughputs above 1 Gbps, as required for HENP programs in future, will require additional developments, and/or the use of protocols other than standard TCP/IP.

[27] Such standardized tools for high throughput, performance monitoring and tracking do not yet exist, but are the subject of future development by the HENP Network Working Group. This group first met in June 2001 (see http://www.transpac.org/meeting.html) and decided to work in the framework of an Internet2 Working Group.

[28] As an example of possible costs of this type: following discussions with the NSF/CISE networking directorate, (W.~Decker) in 1999, the first GriPhyN proposal foresaw an average of $ 300,000 per year for connections from Tier2 centers to Abilene.

[29] The discussions in the "Hoffmann" Review of LHC Computing at CERN have confirmed the recent estimates for network requirements at the start of LHC operation: the aggregate requirement into CERN is estimated to reach several times 10 Gbps by 2006. This would include a 10 Gbps link between the US and CERN. This is expected to be just adequate to accommodate 2.5 Gbps of dedicated bandwidth (corresponding to approximately 1.5 Gbps of sustained throughput for data) each for CMS and ATLAS.

above).

The basis for determining the bandwidth requirements and cost parameters in the plan is summarized as follows:

1. Follow the known bandwidth trends on the US-CERN link, as well as other larger scale research and education networks. These trends are remarkably consistent between ESnet, use of the CERN link, DFN in Germany, and the academic portion of ``the Internet" in other countries. The rate of bandwidth growth was very close to 100 % per year during the late 1990's, although the rate of growth has accelerated in 2000 to approximately 150 % per year on Internet2, and greater growth rates are currently observed in some national academic and research networks outside the US[30].

2. Ramp up early in 2002 to 622 Mbps, in time to support *BABAR* and RHIC running with higher data rates and larger event samples to analyze than in 2001, and the startup period of Run 2, along with the large simulated event productions for CMS data challenges, and Data Grid developments for the LHC.

   The ramp-up to OC3 brought economies of scale in 2001, as this is the typical bandwidth unit in the US (and increasingly across the Atlantic) for medium-sized wide area network customers.[31] Budgetary prices from vendors in the Spring of 2001 indicated that the cost of OC12 (4 times more in bandwidth) is 2.5 times the cost of OC3.

3. Following more conservative, longer term trends, we plan for a bandwidth increase by a factor very close to two (a 100% increase) each year from 2003-2006 (as shown in the tables given above). This proposed rate of growth also is consistent with the longer term trends of HEP network usage over the last decade, as derived by the ICFA-NTF and ESnet, of a factor of 10 in bandwidth every 3 to 4 years.

4. Reach 10 Gbps by 2005 or 2006. This is expected to meet the baseline needs of CMS and ATLAS, the revised bandwidth estimates from *BABAR*[32] resulting from the exceptional performance of the accelerator and the detector, the estimated use by Run2 experiments (assumed to be similar to *BABAR*), and most of the other transatlantic network use by HEP[33]. This turns out to be generally consistent with following the general bandwidth trend mentioned above.

5. For 2001-2002 assume a cost decrease of 45% corresponding to the middle of the range of cost decreases this year predicted by vendors and industry sources. Assume a cost decrease by 37% per year from 2002 onward, or a factor of 2 every 1.5 years. Although data link costs have recently fallen more rapidly, they have often fallen much more slowly in the past. There is also hard evidence that the national build out of optical fiber-cable infrastructure is over, and that vendors will have to focus on recovering their large investments in building this infrastructure, to survive the next few years[34]. We therefore use this "Moore's Law" decrease as a long term average, following the cost evolution in other fields of information technology. Note that in reality the cost decreases are likely to be more sporadic, with periods of relatively stable prices followed by precipitous drops[35].

6. Assume relatively modest support for network engineering[36,] at a level evolving from 1 to 2 FTEs, at an annual cost of $145k per FTE (including indirect costs), and an annual salary index of 5%. These FTE

[30] Examples include Japan, Chile, and eastern China.

[31] The typical bandwidth on the major links in international academic and research networks was 622 Mbps (OC-12) in 2000, and is expected to move to 2.5 Gbps (OC-48) in 2001. Domestic link bandwidths are often higher, as Abilene, vBNS and the B-WIN German network have been at 2.5 Gbps and are planning to move to 10 Gbps.

[32] Reference: L. Cottrell, ESnet International Meeting, Kyoto, July 2000.

[33] Taking into account that some of the traffic for BABAR and Run2 will be carried on links funded by the European partners in these experiments.

[34] See "Busted by Broadband", Time Magazine March 26, 2001 at http://www.time.com/time/magazine/article/0,9171,1101010326-102923,00.html .

[35] For example, following the commissioning of a new generation transoceanic cable with substantially larger capacity than previous-generation cables. See for example: http://www.globalcrossing.com/network/net_ac2.htm .

[36] For tasks including network configuration, monitoring, troubleshooting, installations and upgrades at vendor points of presence; some help with hardware installations and/or configuration and trouble shooting at various HEP sites. Work in coordination with the CERN External Networking Group, the Internet2 and Regional Network engineering teams, and the groups developing and operating Data Grids.

support levels and costs per FTE are low compared to the cost of network engineers on the open market. Also assume modest support at the level of 1.5 FTEs for VRVS, which is the minimum level required for the current level of usage according to recent experience[37.]

7. Reach a constant DOE funding level during LHC operation from 2006 onwards. This assumes that non-DOE sources contribute a substantial portion of the overall transatlantic link cost, as is the case now. In 2007 and beyond, it is assumed that the bandwidth will continue to increase, but at a reduced rate corresponding to continued funding at the 2006 level, in 2006 constant dollars.

The proposed funding profile, which represents the best current estimate of what is required to meet HEP's (and some NP) network needs, is summarized in the following plot. Apart from the charges for leasing the transatlantic link, there are smaller but significant charges for "Infrastructure" which includes the required network hardware (routers, switches, interfaces), rental costs for placing and maintaining a rack in the telecom vendor's point of presence, connections to the general purpose Internet, a modest amount for salaries of network support engineers, and maintenance (24 hour/7 day per week/4 hour response time) at the termination point of the link. We emphasize that prediction of the actual costs after 2002-3 is very difficult and uncertain. Quoted prices depend strongly on market conditions as they develop and it is sometimes possible to negotiate special prices on an *ad hoc* basis. The costs shown here assume reasonable success in these negotiations.

The budgets below do not include explicit funding for "research" links at higher bandwidth, which are based on the use of a wavelength (or a subchannel within one wavelength) on a fiber carrying many wavelengths[38]. In contrast to the traditional services which, until now, are based fully on redundant SONET or SDH[39] links, the wavelength services are typically "unprotected". This means that wavelength-based services today do not come equipped with a "hot backup" line that cuts in immediately in case of a failure of the primary link. For this reason, the wavelength-based services are typically 50-70% less expensive than fully redundant SONET or SDH services. Use of these services must however be restricted to network and Grid R&D use, and for use as a way to accommodate peak bandwidth demands above the baseline, because the usual "production" level of reliability (typically 99.9% for SONET/SDH links) is not guaranteed; an outage could take several days to a week to restore in some cases. Such a research link at OC-48, between CERN and the US, is the subject of the DataTAG proposal recently approved by the European Union (see Section B.5.3 below).The costs for the "Infrastructure" listed above are broken down by category and year in the following figure and table (in millions of dollars):

B.5.3. Potential contributions from foreign partners

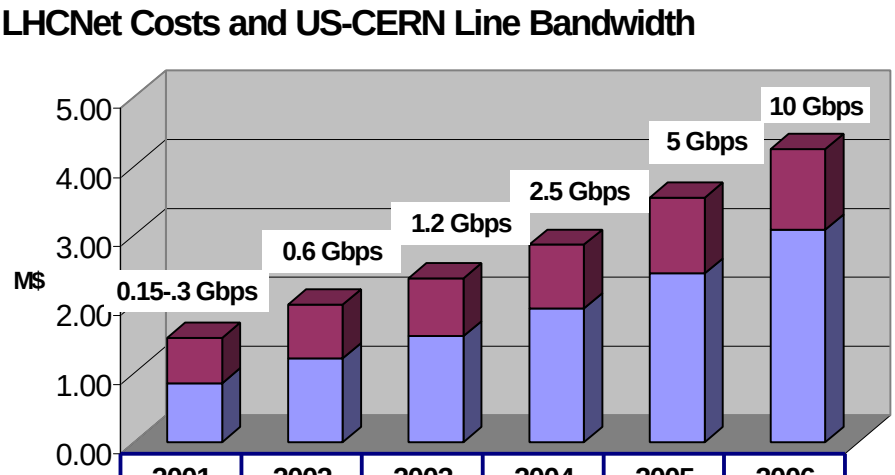
On the European side the partners in the U.S.-CERN link are IN2P3 (CCPN/Lyon), WHO (World Health Organization) and CERN, contributing to the CERN-Chicago link at the level of 400KCHF, 200KCHF and 350KCHF respectively, hence a total of 950KCHF, i.e. approximately 550 K$.

The WHO contribution is unlikely to change as they already contribute more than an amount corresponding to the bandwidth they use. The IN2P3 contribution does not include the very high cost of the local loop between Lyon and Geneva (i.e. 500KCHF) and although the link has recently been upgraded to 155Mbps, it is not at all clear what will happen next year (i.e. June 2002) when the existing Telecom contract needs to be renewed; here are some options:

[37] Significantly expanded use by HENP would require additional funding for support. We assume here that any large scale use of VRVS outside of HENP will be funded by sources other than HENP.

[38] Using Wavelength Division Multiplexing (WDM). For an explanation see for example: ftp://ftp.netlab.ohio-state.edu/pub/jain/courses/cis788-99/h_5opt2.pdf

[39] SONET and SDH are a set of related standards for synchronous data transmission over fiber optic networks. SONET is short for Synchronous Optical NETwork and SDH is an acronym for Synchronous Digital Hierarchy. SONET is the United States version of the standard published by the American National Standards Institute (ANSI), while SDH is the international version of the standard published by the International Telecommunications Union (ITU). See http://www.techfest.com/networking/wan/sonet.htm.



	2001	2002	2003	2004	2005	2006
Infrastructure (M$)	0.66	0.78	0.84	0.92	1.10	1.17
TA Link and Internet Access Costs (M$)	0.85	1.22	1.53	1.94	2.44	3.08

Fig. 4: LHCnet Bandwidth and Costs

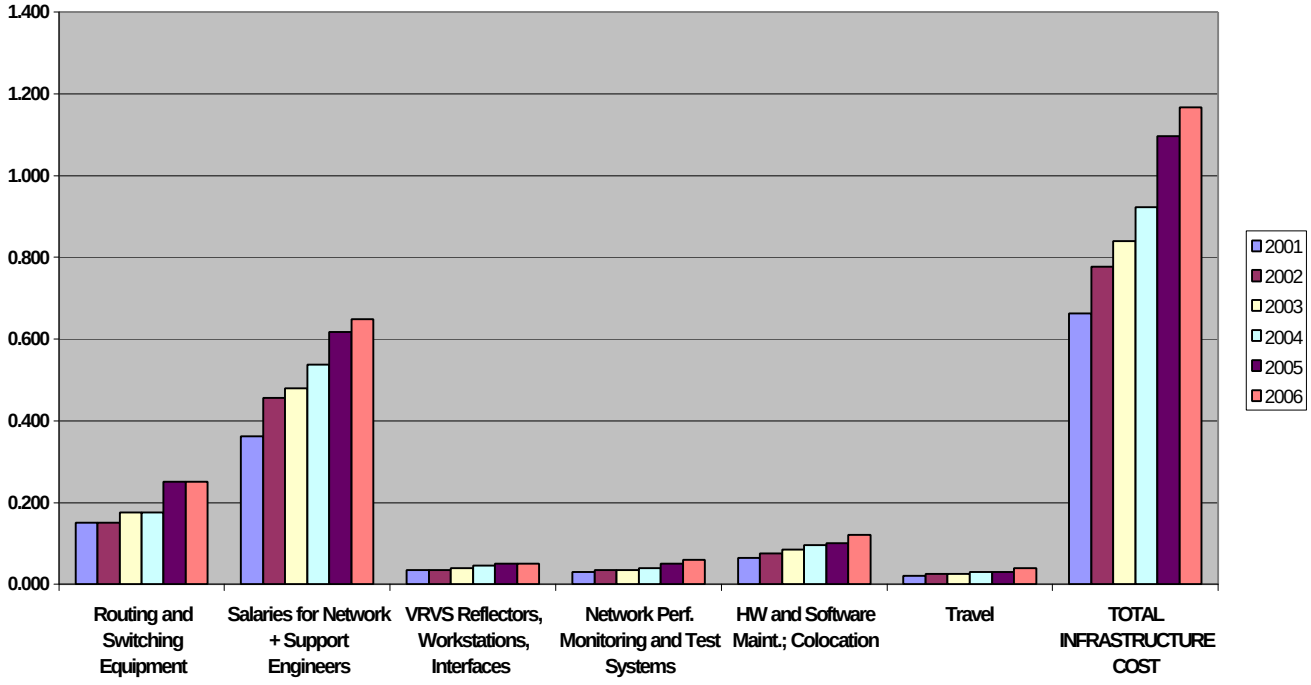


Fig. 5: LHCnet Infrastructure Costs

	2001	2002	2003	2004	2005	2006
Routing and Switching Equipment	**0.150**	**0.150**	**0.175**	**0.175**	**0.250**	**0.250**
Salaries for Network + Support Engineers	**0.363**	**0.457**	**0.480**	**0.537**	**0.617**	**0.648**
VRVS Reflectors, Workstations, Interfaces	**0.035**	**0.035**	**0.040**	**0.045**	**0.050**	**0.050**
Network Perf. Monitoring and Test Systems	**0.030**	**0.035**	**0.035**	**0.040**	**0.050**	**0.060**
HW and Software Maint.; Colocation	**0.065**	**0.075**	**0.085**	**0.095**	**0.100**	**0.120**
Travel	**0.020**	**0.025**	**0.025**	**0.030**	**0.030**	**0.040**
TOTAL INFRASTRUCTURE COST (M$)	**0.663**	**0.777**	**0.840**	**0.922**	**1.097**	**1.168**

1: Make a call for tender and negotiate more advantageous conditions, e.g. OC12c (622Mbps) or more, at the same price, in which case one would expect IN2P3 to contribute additional money to the CERN-USA link.

2: Integrate the upgraded Lyon-Geneva circuit into RENATER. One could also expect additional contributions from RENATER.

3: Discontinue the Lyon-Geneva link (i.e. full integration of IN2P3 into RENATER).

The CERN contribution is not planned to change before 2005, however, the CERN management agreed to use the NSF award (3 times 350K$ for the period 2000-2003) to gradually increase the bandwidth on the CERN-USA link.

One cannot exclude that other partners may join, however, this is unlikely to happen.

The DataTAG proposal, recently accepted for funding by the European Union, will provide bandwidth for data grid testbeds and may be used in conjunction with some of the bandwidth on the U.S.-CERN connection.

B.6. International Connections to US and European Networks

Transatlantic links are only part of the story. The links considered above must connect appropriately to the networks on each side of the ocean that actually connect to the institutions of the experiments. In this section, we give a brief overview of those networks and some issues involved.

B.6.1. ESnet

The Energy Sciences Network, or ESnet, is a high-speed data communications network serving thousands of U.S. Department of Energy (DOE) scientists and their collaborators worldwide. An early provider of high-bandwidth, reliable network services, ESnet gives researchers at national laboratories, universities, and other institutions the collaborative capabilities needed to address some of the world's most important scientific challenges (http://www.es.net/).

As a mission-oriented effort, ESnet is organized to provide the best possible networking for DOE programs. Many of the participating DOE programs rely fundamentally on the capabilities of the network to enable their research functions. For them, advances in network capabilities translate directly into advances in research capabilities. This vital connection has led ESnet to pursue the following goals:

?? Reliable, production-quality network services with capabilities based on leading-edge technology.

?? Close coupling to the Department's programmatic requirements.

?? Ongoing improvements in network services and related applications targeted at the rapidly evolving and growing needs of the programs.

?? Highly leveraged interaction and coordination with ESnet sites to optimize service, performance, and resources.

?? Effective interagency and international coordination and cooperation.

ESnet provides advanced computer networking for the DOE science mission and other DOE missions. Managed

and operated by the ESnet staff at Lawrence Berkeley National Laboratory, ESnet currently provides direct connections to more than 30 major DOE sites at speeds up to 622 megabits per second (Mbps). Connectivity to the global Internet is maintained through interconnection ("peering") arrangements with more than 100 other Internet service providers. The ESnet service footprint is shown in the figure below.

Funded principally by DOE's Office of Science, ESnet allows DOE scientists to use unique DOE research facilities and computing resources independent of time and location with state-of-the-art performance levels. ESnet supports science in ways that extend from the mundane, such as communication via e-mail, to the data- and bandwidth-intensive, such as the rapid distribution of enormous volumes of scientific data to researchers around the world.

The ESnet Project enjoys an excellent working relationship with both its technical and program user communities. This relationship is maintained through three user committees:

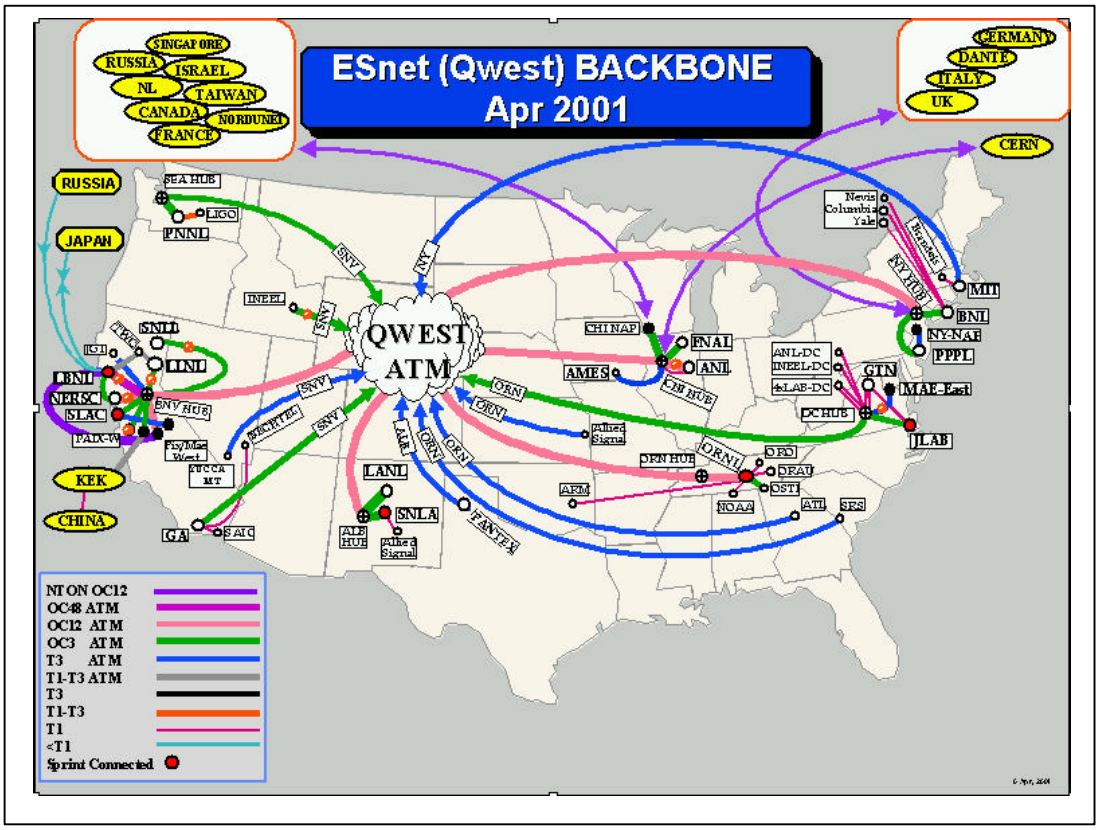
?? The ESnet Steering Committee (ESSC) deals with ESnet's strategy, policy, operational requirements, and priorities as established by principal investigators representing DOE program areas.

?? The ESnet Coordinating Committee (ESCC) deals with the associated technical issues affecting the sites and the backbone.

?? The ESnet Research Support Committee (ESRSC) deals with the requirements and technical issues related to ESnet testbed activities.

These three committees sponsor special-interest task forces and working groups that study issues, establish consensus, and share recommendations and information. An important element of DOE research is collaboration among teams of researchers located around the world. ESnet enhances the effectiveness of these scientists' work by providing a rich interconnectivity to the "outside" world. The network includes interconnections to many other U.S. networks, as well as several direct connections to international sites and networks. A recent emphasis has been enhancing interconnectivity to U.S. universities with "peering" interconnects to Abilene (the Internet2 backbone network) at speeds up to OC12 (622 Mbps).

Fig. 6: ESnet Backbone Structure



B.6.2. Internet 2 and universities

Internet2[40] is a project managed by UCAID, the University Consortium for Advanced Internet Development.

The Internet2 backbone Abilene, on which most of HEP's universities depend for their national network connectivity, along with the member university sites, is shown below. Internet2 is being led by a consortium of 185 universities working in partnership with 75 corporations in industry and 40 non-profit companies and government labs, with the mission to "develop and deploy advanced network applications and technologies for research and higher education, accelerating the creation of tomorrow's Internet."

The primary goals of Internet2 are to:

?? Create a leading edge network capability for the national research community
?? Enable revolutionary Internet applications
?? Ensure the rapid transfer of new network services and applications to the broader Internet community.

Fig. 7: Internet2 Abilene Network

The Internet2 board includes the Chairs of four Advisory Councils, on Applications Strategy[41], Network Planning, Network Research Liaison, and Industry Liaison

Internet2 Working Groups are the primary mechanism for the design, development, implementation and documentation of advanced network services. Through Internet2 Working Groups and initiatives, Internet2 members are collaborating on:

?? Advanced Applications
?? Middleware
?? New Networking Capabilities

[40] See http://www.internet2.edu.

[41] H. Newman has been appointed to this council, and represents HENP interests on it. As of September 2001 he is also the HENP contact for the Internet2 End-to-End Initiative, discussed in this section.

?? Advanced Network Infrastructure
?? Partnerships and alliances

Internet2 Applications Working Groups are currently active in the health sciences, arts and humanities, along with technology-oriented groups on Digital Video[42], Network Storage and Voice over IP and an HENP Applications Working Group[43] is now being formed, with the agreement of the Internet2.

The Internet2 Middleware Initiative covers a number of issues that are synergistic with the Grid R&D and other related efforts in which HEP is involved. These include authentication, identification, authorization, directories and security issues.

The Internet End-to-End Initiative[44] (E2E) has the goal "To create a predictable, and well-supported environment in which Internet2 campus network users have routinely successful experiences in their development and use of advanced Internet applications, by focusing resources and efforts on improving performance problem detection and resolution throughout campus, regional, and national networking infrastructures." This initiative clearly is synergistic with HEP's need to develop and deploy standardized methods and toolsets, to enable the major HEP programs and the at-large HEP community to use the available networks most effectively.

The Internet2 architectural concept is a hierarchy in which regional networks and some members access the I2 backbone through "GigaPoPs"[45]. The current locations of the GigaPoPs (as of May 2001) are shown in the following figure:

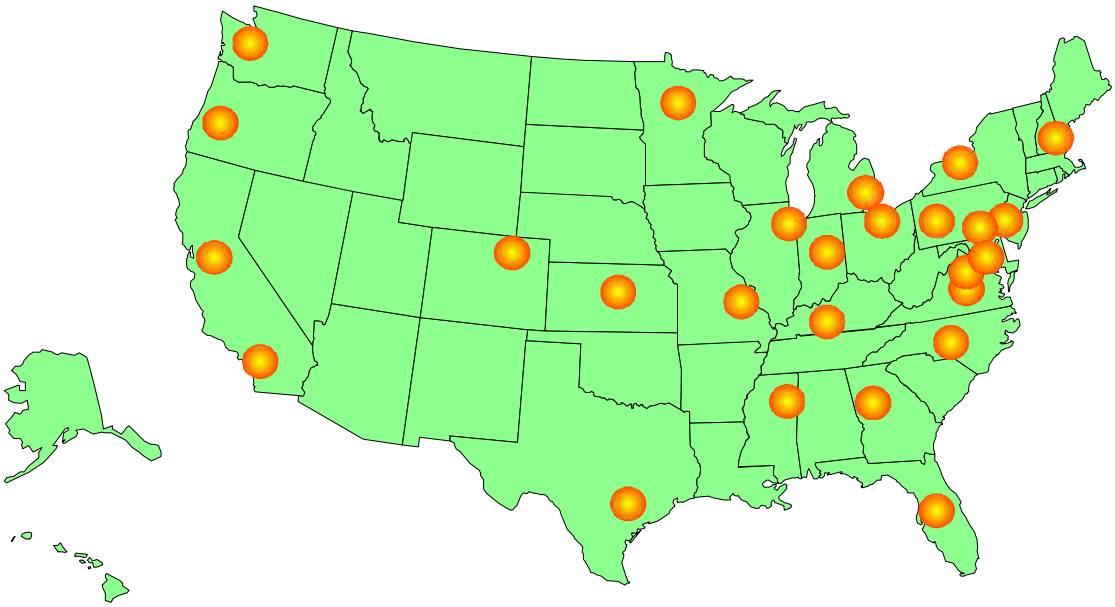


Fig. 8: Internet2 GigaPOPs

In planning to meet HEP's needs we have implicitly assumed that the current Internet2 backbones and GigaPoPs will be followed by a next-generation national network infrastructure interconnecting the US research universities with sufficient bandwidth (at 10 Gbps and up). HEP needs to work in support of this goal, along with the corresponding ESnet upgrades to provide the required high speed connections to the HEP laboratories.

[42] Including HENP's VRVS system as one of the core technologies, as part of the "Internet2 Commons" Videoconferencing Implementation Plan (April 2001).

[43] Following the HENP Network WG Users' Meeting at Indiana University, June 1-2, 2001, organized by R.~Gardner (Indiana) and H.~Newman (Caltech).

[44] See www.internet2.edu /e2e . The design document for the E2E Initiative is now available for comment at: http://www.internet2.edu/e2eperf/papers/End-to-End-Perf-Design-Paper.pdf .

[45] GigaPoP: A "Gigabit capacity point of presence". The Internet2 engineering architecture and the structure and function of the GigaPoPs is (still, as of June 2001) described accurately in: http://www.internet2.edu/html/97engineering.html .

B.6.3. Impact of International Links on Network Infrastructure

The global networking research needs of the HEP create both a benefit and challenge for the program. The HEP worldwide networking requirements benefit from the global nature of the Internet, as it would be extremely difficult and expensive to provide the networking needed on a program-dedicated basis. On the other hand, this creates a challenge for the program in that its requirements must be met making use of facilities and resources that are not completely under its control. For example, networking between US sites and those in Europe depends on transatlantic links largely funded by European entities (the Chicago-CERN link is the notable exception).

In general, the exchange of traffic between research networks is done with the tacit assumption that overall there is equity in benefit, and there is little demand for adjustment through a balance of payments. However, the extensive and large projected future networking requirements of the HEP (as documented in this report) may alter this situation. The demands to be placed on the various research networking components of the collaborating enterprises may prove to be too large and grow too fast to be expected to be absorbed on a "business as usual" basis. This situation is examined in more detail below.

B.6.3.1. European National Research Networks

Most countries in Europe provide a national research network. Networks of particular interest to the HEP include the following:

	Country	Organization	Network	US Bandwidth
1.	Germany	DFN-Verein	B-WiN/G-WiN	2xOC12
2.	Italy	INFN	GARR-B/GARR-G	OC12
3.	UK	UKERNA	SuperJANET	OC12
4.	Netherlands	SURFnet bv	SURFnet5	2x1 Gbps
5.	Nordic Countries	NORDUnet Board	NORDUnet	1.4 Gbps total

Funding: Most of the European national research networks are funded through a combination of national government funding and/or institutional fees. There are typically no direct fees for exchange of peer traffic.

International Links: Currently both DFN and INFN contract with DANTE for their transatlantic networking links. ESnet interconnects with G-WiN, GARR, and JANET in New York and contributes to the cost of the international links via a contract with DANTE and by funding the international peering points in NY and Chicago.

ESnet also interconnects with a large number of international peers in Chicago via an existing OC3 shared interconnect to the Chicago NAP.

B.6.3.2. Future Plans:

Currently it appears that there will be continued expansion of the capacity of the above international links into NY. A new high-speed interconnect point called STARLIGHT (mentioned above) is also emerging in the Chicago area which will emphasize optical interconnects.

B.6.3.3. Pan-European Research Network

Within Europe a number of national and international research networks carry HEP network traffic. The major pan-European network is managed by the DANTE organization and is now involved in a major upgrade to a version called GEANT using 10 Gbps trunking that is expected to be in operation during the latter part of 2001. In general, this is a backbone network that interconnects National Research and Educations Networks (NREN), rather than specific end sites.

Funding: The network is funded, in part, by fees from the NRENs that connect to it.

International Links: DANTE provides a number of OC3 and OC12 circuits between Europe and NY, some of which are for its own activities, other of which are "subcontracted" to specific NRENs.

B.6.3.4. US Research Networks

The two major research networks that carry HEP traffic within the US are Abilene and ESnet. Abilene interconnects virtually all of the major academic institutions within the US, while ESnet interconnects all the

major DOE National Laboratories and research facilities.

Funding: Abilene is funded via institution connect fees, as well as through very substantial donation, including over $500M of bandwidth over a five year period from Qwest. The initial donation period will expire in early 2003, and the project management is actively developing a plan for continued operation using a new state of the art production and research backbone after that time. Although the likelihood seems small, failure to extend the program would have a very serious impact to the HEP program, as virtually all access to academic collaborators would have to be rebuilt.

ESnet is centrally funded via the DOE Office of Science Mathematics Information and Computer Sciences (MICS) office, and provides networking support to five major program offices within the Office of Science. It provides networking support to virtually all the major DOE laboratory research sites.

International Links: Neither Abilene or ESnet are currently directly responsible for any transatlantic links. Both have international peering at both Chicago and NY. ESnet contributes funding directly to DANTE for its use of transatlantic capacity.

B.6.3.5. Challenges

The above summary points out several potential areas of concern:

1. ESnet future funding may not be adequate to meet the growth demands of the HEP. Current projections for the next few annual budget cycles show ESnet at a "flat" budget. Although some performance growth can still be expected due to anticipated reductions in the unit cost of bandwidth, it is unrealistic to expect that such savings will be sufficient to match the normal growth in traffic (a canonical 100% per year) and the projected rapid growth of HEP network usage.

The potential incremental financial impact to ESnet is very difficult to estimate. The HEP requirements given herein are stated as aggregate bandwidth per site or experiment, which is insufficient information to engineer a responsive network architecture within a limited funding envelope. In addition, projecting the benefit of anticipated pricing reductions is complicated by the fact that ESnet costs are dependent on contractual negotiations and are only driven by commercial pricing as a second-order effect (which is, in turn, relatively unpredictable). However, experience would indicate that an incremental adjustment of $ 1.5-3 M/year would allow ESnet to meet most or all of its actual new HEP requirements.

2. Access to academic institutions will depend heavily on the continued operation of Abilene and/or its follow-on. Currently, Abilene is very well provisioned in bandwidth with a number of lightly loaded OC48 trunks, and is likely to be easily capable of carrying the projected HEP traffic loads over the next 1-2 years. However, beyond that timeframe the traffic loads may be sufficiently high to cause noticeable impact on both the Abilene follow-on as well as the local access arrangements (GigaPops) and the campus of the major academic collaborators.

A related issue is the impact of the projected bandwidth requirements on campus infrastructure and access. Both can be relatively expensive and the impact of the HEP requirements on either or both may not be negligible. It may at times be requested that the HEP program pay incremental costs on a campus associated with particularly intensive requirements for the HEP program, especially at universities that host Tier 2 centers or other special facilities.

3. International links are currently heavily subsidized by the European NRENs and/or DANTE. Although ESnet does contribute to some degree to its associated international link costs, Abilene does not. Should this state of affairs change such that the generous underwriting of the international links by non-US organizations is eliminated, it would have serious impact on the HEP program outlined here.

4. Access to international institutions: this is an area of only moderate concern for the future. For major international institutions there can be the expectation that the institutions will be responsible for meeting the networking requirements of HEP. A major exception is CERN, where the experiment is sited outside the U.S. However the impact and cost of US connectivity to CERN is addressed elsewhere in this document.

B.7. Recent Developments in Requirements and Technology

Over the last year, a number of factors affecting the technical requirements and costs have changed. While unit costs for data links have fallen, the data volumes of major experiments (*BABAR* now; LHC in future), and the bandwidth estimates as a function of time for *BABAR* and DOE's present-generation experiments as well as the LHC have increased substantially. The near-term milestones for LHC trigger and detector performance studies, as well as the advent of Data Grid concepts, have advanced the need to engineer, test, and bring into production higher bandwidth on the major links to HEP labs, and the US-CERN link. The baseline bandwidth profile over

time for the US-CERN link is 622 Mbps (OC12) in 2002, rising to 10 Gbps (OC192) by 2006. Taking into account the multiple experiments served by the US HEP labs, the bandwidth requirements for SLAC, FNAL and BNL are OC12 this year, 2 X OC12 or OC48 (for FNAL) next year, evolving quickly to 2 X OC48 or OC192 by 2005 or 2006, if not earlier.

In terms of technology, this fits well with vendor plans to market data channels using "wavelengths" in wavelength division multiplexing systems at 2.5 Gbps (OC48) starting this year, and 10 Gbps in the next 12-24 months, on transoceanic links.

The telecommunications industry may soon transition to photonic switching in its core infrastructure, from the current technology based on electronic switching of signals. This means that within the next 2-4 years, it will be substantially more cost effective to work at 2.5 Gbps, then 10 Gbps, than to try to buy data links at lower speeds based on the current core technologies. This is analogous to the current situation where the OC3 (155 Mbps) US-CERN linki has brought us up to the lower end of the range of standard fiber-based electronically switched data transmission rates on that link.
Beyond the baseline scenario summarized in this letter, a number of possible changes are being discussed in research organizations and industry that may make the scale and modes of use of networks substantially different from our current concept. These include:

?? Greater bandwidth per wavelength (40 Gbps); expected to appear as a commercial product within the next two years.

?? More wavelengths per fiber compared to current products with ~100 wavelengths (1 Terabit/sec) per fiber[46].

?? The delivery of wavelengths, at 2.5, 10 and eventually 40 Gbps, by the vendors directly to customer sites.

?? Integration between OC192 WAN links and 10 Gigabit/sec Ethernet, to provide greater end-to-end network transparency[47].

?? ASICs to support multi-Gbps data transmission at low cost.

?? New Internet software architectures designed to manage the broadcast and distribution of data on a large scale in quasi-realtime.

These advances are not part of the current baseline but are being continuously tracked, to ensure that our usage scenarios, requirements estimates and plans for evolving the network infrastructure and bandwidth remain current and cost effective.

C. Recommendations

On the basis of the committee's work summarized above, the committee makes the following recommendations:

1. High performance networking is central to the success of international collaborations. Every effort should be made to provide the required transatlantic bandwidth, particularly that of the US-CERN link, as documented in this report for the period 2001-2006, and to continue to provide adequate bandwidth to keep pace with HEP's needs.
2. Arrangements must be made to ensure that the US-CERN connection and other international links used by HEP experiments are connected with the required high performance, to the networks that make the ultimate connections to the collaborating institutions in both regions. Similarly, arrangements should be made to ensure that key university or laboratory sites (e.g. housing Tier 2 or Tier B centers) are appropriately connected to their regional networks or to the national backbone networks. It is likely that ESnet and Internet2 will not meet all needs within the U.S. without special arrangements, and some targeted funding

[46] While more wavelengths and higher bandwidths per wavelength are likely to exist, it should be kept in mind that the usable spectral range, and the required inter-wavelength gap will limit the capacity of a single fiber to at most 10 Tbps using currently-known optical-fiber transmission technology. At the time of this writing, 3 Tbps has been achieved with a single fiber in the lab.

[47] As discussed at the Web site of the 10 Gigabit Ethernet Alliance (see http://www.10gea.com/) the 10 Gigabit Ethernet standard, which is designed to span Metropolitan Area Network (MAN) distances of 40 miles, is expected to be fully completed, on schedule, in March 2002. Unlike Gigabit Ethernet, this new standard has been designed with integration between LANs and WANs in mind. Updated information on the status and projected costs of 10 Gigabit Ethernet (May 2001) may be found at http://www.10gea.com/Tech-whitepapers.htm and http://www.10gea.com/NR%20-%20Presentations.HTM .

by HEP should be foreseen.

3. Appropriate attention should be paid to providing high quality network service and to monitoring the network and maintaining good performance.
4. The agencies should ensure that planning for networking (both domestic and international) to serve HEP experiments is reviewed at least annually by a committee with official representation from at least the major user experiments.
5. The committee found that there was a strong need for a technically oriented forum to promote information exchange and technical collaborations between the HEP community and network providers and technical experts, and the work of the forum should extend to identifying and making available standardized toolsets to ensure that the networks provide the needed end-to-end performance. The HENP Network Working Group was then formed in June 2001, which promises to meet these needs. The committee recommends that that this group be encouraged by the agencies and that technical representatives of laboratories and experiments be encouraged to participate in its work.

D. Appendix I: Charge Letter

U.S. Department of Energy
and the
National Science Foundation

JOINT OVERSIGHT GROUP

June 22, 2000

Dr. Daniel Green
Fermi National Accelerator Laboratory
P.O. Box 500 MS-205
Batavia, IL 60510

Dr. Harvey Newman
256-48 HEP, Caltech
1200 E. California Blvd.
Pasadena, CA 91125

Dear Drs. Green and Newman:

We are well aware that substantial planning and R&D is underway to produce the computing infrastructure needed for U.S. physicists to effectively participate in the analysis of the LHC data. A cornerstone of this computing infrastructure will be the network link(s) between the U.S. and CERN. We are concerned that the networking needs of the computing activity may not be fully understood.

We want to create a Working Group of appropriate technical experts to assemble and review the projected networking needs related to the link to CERN. We would like you to appoint a person representing the U.S. CMS project who will serve, together with a person representing the U.S. ATLAS project, as co-chairs of this Working Group. The co-chairs should secure appropriate representation from the other large U.S. collaborations (e.g. CDF, D0, BaBar, Zeus) which – at least potentially – can make beneficial use of the CERN link, at least one additional representative from the university community, and other technical experts as desired. Consideration should be given to inviting CERN and ESNet representatives as observers based on their anticipated involvement at the two ends of a Trans-Atlantic link, and the NSF has useful in-house technical expertise, which can also be tapped.

The Charge to the Working Group is as follows:

1. Estimate the bandwidth needed for the network connection to CERN to meet the needs of the U.S. HEP community over the next 5 years (2001-2005).
2. Estimate the cost of providing this bandwidth.

We are very much aware that both parts of the Charge involve large uncertainties, but we need the best information we can get to plan for the future.

We would appreciate a response by late summer.

The working Group should keep Glen Crawford (DOE) and Marvin Goldberg (NSF) informed of their activities.

Sincerely,

Dr. John R. O'Fallon
Co-chair
U.S. LHC Joint Oversight Group
Department of Energy

Dr. John Lightbody, Jr.
Co-chair
U.S. LHC Joint Oversight Group
National Science Foundation

cc: Matthias Kasemann, U.S. CMS Acting Project Manager for Software and Computing
Ken Stanfield, Associate Director, Fermi National Accelerator Laboratory

Appendix II: Spreadsheet calculations for Section B.2.3.3.ii

Financial Model for Tape-Based Data Exchange		
Assume:	SLAC distributes data to three Tier A centers at 1TB/day	
	SLAC receives simulated data from three centers at 1TB/day	
	This is a continuous data rate of ~120 Mbits/s in and ~120 Mbits/s out. 60 Gbyte STK 9940 tapes and drives	
	Twice-weekly Fedex data exchange	
For 1 TB/day bi-directional transfer		
	Tapes written/read per day (100% efficient tape use)	66.7
	Tapes written/read per day (75% efficient tape use)	88.9
	Drive-hours (100% efficient use)	111.1
	Drive-hours (50% efficient use)	222.2
	Tape drives required at SLAC (assume 24x7 operation)	5
	Tape drives required at one Tier A (assume 24x7 operation)	2
	Data-aide personpower at SLAC	1.5
	Expert personpower at SLAC Data-aide personpower at one Tier A	0.25 1
	Expert personpower at one Tier A	0.2
Annual Costs:		
	Tapes (assume 2-months worth with 2-year lifetime)	$101,667
	Drives, fiberchannel, servers etc. purchase+maint (assume 3-year lifetime)	$189,673
	People at SLAC (salary+benefits+overhead) People in Europe (salary + benefits + overhead)	$183,750 $210,000
	Fedex (assume 30% discount on published rates)	$117,499
Total Annual Costs		**$802,589**