A New Generation of Networks and Computing Models for High Energy Physics in the LHC Era

H Newman
Professor of Physics, California Institute of Technology
1200 East California Boulevard, Pasadena CA 91125, USA
E-mail: newman@hep.caltech.edu

Abstract.
Wide area networks of increasing end-to-end capacity and capability are vital for every phase of high energy physicists' work. Our bandwidth usage, and the typical capacity of the major national backbones and intercontinental links used by our field have progressed by a factor of several hundred times over the past decade. With the opening of the LHC era in 2009-10 and the prospects for discoveries in the upcoming LHC run, the outlook is for a continuation or an acceleration of these trends using next generation networks over the next few years. Responding to the need to rapidly distribute and access datasets of tens to hundreds of terabytes drawn from multi-petabyte data stores, high energy physicists working with network engineers and computer scientists are learning to use long range networks effectively on an increasing scale, and aggregate flows reaching the 100 Gbps range have been observed. The progress of the LHC, and the unprecedented ability of the experiments to produce results rapidly using worldwide distributed data processing and analysis has sparked major, emerging changes in the LHC Computing Models, which are moving from the classic hierarchical model designed a decade ago to more agile peer-to-peer-like models that make more effective use of the resources at Tier2 and Tier3 sites located throughout the world. A new requirements working group has gauged the needs of Tier2 centers, and charged the LHCOPN group that runs the network interconnecting the LHC Tier1s with designing a new architecture interconnecting the Tier2s. As seen from the perspective of ICFA's Standing Committee on Inter-regional Connectivity (SCIC), the Digital Divide that separates physicists in several regions of the developing world from those in the developed world remains acute, although many countries have made major advances through the rapid installation of modern network infrastructures. A case in point is Africa, where a new round of undersea cables promises to transform the continent.

1. Introduction
Wide area networks are vital for every phase of high energy physicists' work. With the opening of the LHC era in the Spring of 2010 and the remarkable progress of the LHC accelerator and experiments ever since, the field's reliance on wide area networks of sufficient, and rapidly increasing end-to-end capacity and capability has intensified, following a 30 year trajectory, marked by:

1 To whom any correspondence should be addressed.
• **Exponential growth in capacity and bandwidth use.** Our bandwidth usage, and the typical capacity of the major national backbones and intercontinental links used by our field have each progressed by a factor of 300-1000 over the past decade. Major examples include the growth in capacity of US LHCNet supporting transatlantic traffic among the US Tier1s by a factor of $6 \times 10^6$ from 9.6 kbits/sec (kbps) to 60 gigabits/sec (Gbps) over the last 25 years [1], and the remarkably steady growth of incoming network traffic measured on the ESnet, by a factor of 10 every 47 months on average since 1992 [2], where the monthly ESnet traffic volume is expected to reach more than 17 petabytes by the end of 2011². Recent studies of bandwidth requirements and traffic growth indicate that these trends will continue over the next 5 years, spurred on by the rapid progress of the LHC physics program and encouraged by the transition to the next generation of network technologies with link speeds of 40 and 100 Gbps, but at a pace that may be limited by the available funding.

• **A sustained ability to use ever-larger continental and transoceanic networks effectively using high throughput transfers, and dynamic network channels.** Responding to need to rapidly distribute datasets of tens to hundreds of Terabytes drawn from multi-Petabyte data stores among the hundreds of Tier1, Tier2 and Tier3 centers located at sites throughout the world, physicists working with network engineers and computer scientists have learned to transfer datasets at rates of 1-10 Gbps, and higher in some cases, with peak aggregate transfer rates among the sites reaching 100 Gbps peaks during the Fall of 2010. Some network teams supporting HEP also have learned to allocate and schedule bandwidth channels dynamically crossing multiple network domains [3]. These developments have made high energy physicists and the network engineers and computer scientists working with them, among the leading developers as well as the leading users of global networks.

• **The need for highly reliable networks.** The goal for network service availability has been set by the WLCG and more recently by the US DOE at 99.95% [4], which corresponds to a downtime of at most 4 hours per year. This impressive figure has been met and exceed in continental infrastructures such as ESNet [2]. Achieving this high level of service availability on transoceanic network links each of which typically has an uptime of 98.5 - 99% is more difficult, but has been accomplished in US LHCNet, for example, through the use of virtualized network channels spread across multiple underwater cables following diverse physical paths, with automated failover using mesh restoration at the optical layer, coupled to a highly reliable monitoring system [5] that maintains the network topology and protects the highest priority services by automated reconfiguration even in the presence of multiple link outages, in addition to equipment redundancy, and redundancy of its points of presence (in cooperation with ESNet).

2. **Network Requirements in the LHC Era**

The exponential growth trend of HEP's network traffic and the corresponding network capacity needs are expected to continue over the next five years, as the LHC integrated luminosity and the data volumes stored by the LHC experiments continue to grow. Following a 2009 network requirements study led by the DOE in 2009 [6], the roadmap for transatlantic network capacity in US LHCNet for example calls for an increase from 60 Gbps in 2010 to 400 Gbps in 2015. This is in addition to an

² It should be noted that the ESnet traffic recorded in May 2010, shortly after the LHC started up at 7 TeV, was close to 9 petabytes, which was above the long term trendline; during the heavy ion LHC run in November, the data flow exceeded 10 petabytes. Interestingly, a significant part of this flow was genomics data. Given the rapid progress of the LHC and the changes in Computing Models discussed in this report, and the evolution of the volume of genomics data, it remains to be seen if the traffic in 2011 will follow or exceed the historical trend.
expected rise in transatlantic capacity to be provided by the GEANT pan-European network [7], and by the US NSF-funded ACE project [8].

It is noteworthy that this roadmap predates the observed rates of network usage since LHC startup, which have substantially exceeded the "design levels" foreseen and tested during the LHC startup preparations in 2009, as well as the observed intensity in the use of Tier2 and Tier3 resources, and the emergent changes in the LHC Computing Models reviewed in the following sections. One example of the unexpectedly large worldwide data throughput, measured at the ATLAS Tier1 sites [9], is shown in Figure 1. As shown in the figure the daily throughput reached a peak of more than 7 GBytes/sec in the Spring of 2010, with a similar peak over a longer period during the recent heavy ion run in November.

As mentioned by F. Gianotti at the ICHEP conference [10], short term peaks of more than 10 GBytes/sec were observed, compared to a "design value" of 2 Gbytes/sec, and it was the reliability of the networks in spite of the unexpected load that enabled the LHC results to be prepared at remarkable speed, including the data taken just days before the presentation. Figure 2 shows the aggregate and Tier1-Tier2 dataflows measured in CMS using their PhEDEx tool [11]. A daily peak exceeding 5 Gbytes/sec was observed in September, with extended periods above 3 Gbytes/sec. Tier2 traffic had short peaks of 2 Gbytes/sec, with Tier2-Tier2 traffic making up 25% of it, with peaks of 50% during reprocessing and redistribution periods.

Another recent development (also during the CHEP 2010 conference itself) is the rapid progress and outstanding performance of the LHC, which met and exceeded its 2010 peak luminosity goals, reaching 2% of design proton-proton luminosity by mid-October. This has led to more ambitious plans for a 2011 and a possible extension of the upcoming run into 2012, including raising the LHC energy to a likely 8 TeV, increasing the focusing at the interaction regions, and decreasing the inter-bunch spacing so as to raise the luminosity by another order of magnitude. This has created the prospect that the integrated luminosity during 2011 will exceed the estimates made in 2009 by a factor of 2 to 8, and bring the long sought Standard Model Higgs, supersymmetry, as well as a wide range of hypothesized exotic new physics scenarios within the discovery reach of both ATLAS and CMS during the upcoming run.
As a result, we expect that the intensity of the physicists’ work, already at a high level in 2010, will undergo a further substantial increase throughout 2011, with a potential continuation through 2012. In addition to the scaling up with the integrated luminosity, the larger currents and smaller transverse sizes of the bunches is expected to lead to an increase in the number of interactions per crossing, leading to a substantial increase in the complexity of the events stored, and with it a larger event size as well as an increase in the time needed to construct the events. All in, we can expect the growth in network requirements during the upcoming run to significantly exceed the estimates developed prior to LHC startup in 2009. If the run extends into 2012, the 2009 roadmaps will have to be further modified as they presume a yearlong shutdown during that year, along with relatively slow or no growth in network requirements in 2012 compared to 2011.

3. Meeting the Needs in 2011-15 and the Transition to Next Generation Networks

Meeting HEP’s network needs over the next 5 years will include the use of next generation Ethernet and optical network technologies, both in wide area and local networks, with link speeds of 40 and 100 Gbps [12] replacing the current standards with maximum speeds of 10 Gbps. The transition to the next generation is being spurred on the growth of traffic throughout the global Internet, by typically 40-65% per year in the developed world, and more than 100% in some sectors of the developing world [13]. The increased loading on transatlantic networks in particular has led to the prospect [14] that the capacity on the existing undersea cables will be exhausted by the end of 2013, unless the transition to at least 40 Gbps is made before then on many transatlantic routes, with a transition to 100 Gbps links just a few years later. 40 Gbps links are already in production use in SURFNet, that Dutch national network. The increasing reach and maturity of 100 Gbps technologies has led many of the major research and education networks, including Internet2 and ESnet in the US, and GEANT as well as leading national networks to Europe, to plan to move directly to 100 Gbps in the next 1-3 years.
followed shortly by the use of 100 Gbps links in advanced transoceanic networks including US LHCNet [15] (see Figure 3 for US LHCNet's current planning scenario for 2015).

![Figure 3 A US LHCNet scenario for transatlantic networking in 2015. The "OTU-4" links shown are 100 Gbps links following the ITU's Optical Transport Network standards [16].]

4. Emergence of a New Generation of LHC Computing Models

The experience of 2010 showed that the simplified hierarchical Computing Model developed for the LHC experiments during the MONARC project [17] in 1998-2000, where the focus is "pushed" from the Tier0 at CERN to the Tier1s which subsequently push selected and/or reduced datasets to the Tier2s for further processing and analysis was often not very efficient. Handling all of the data at the Tier1s before forwarding it to the Tier2s introduced undue delays in ATLAS, for example, and both ATLAS and CMS found that this sometimes led to unnecessarily large data flows in that many of the transferred datasets were seldom, if ever, referenced. The need for the LHC experiments to improve the responsiveness of their Computing Models to the physicists' needs, along with the Models overall efficiency, was highlighted by the experiments' unprecedented ability to produce a full range of results in hours to days, first shown at the ICHEP conference in Paris in July, rather than weeks to years as in previous experimental programs.

In ATLAS, although the "push" model was initially quite successful, with millions of jobs run every week at hundreds of sites, where the number of copies and data types sent promptly to Tier2s was frequently adjusted in April and May, it was found that the Tier2 sites were filling up too rapidly, including datasets that were rarely if ever used. In June a transition to a new peer-to-peer-like dynamic

---

3 It should be mentioned that the ALICE experiment responded to these needs early, by building a highly structured object-data and ROOT based operational model, with many steps in the workflow monitored and automated with the help of Caltech's MonALISA system [17].
data placement scheme called "PD2P" was tried [18], where a dataset is subscribed to a Tier2 as soon as any user needs it, if there are no copies other than at a Tier1. While the data is being transferred user jobs are sent to the Tier1 where the data already resides, and additional dataset replicas are sent to additional Tier2s if needed, as judged by the backlog of jobs using the dataset. As shown in Figure 4, the US Tier2 sites in ATLAS experienced a steep exponential rise in the data stored on disk prior to the use of the PD2P scheme, and the rise was much slower from then on, in spite of the continued rapid rise of LHC luminosity.

Figure 4 Cumulative data stored at US ATLAS Tier2 sites in 2010
(Courtesy Kaushik De, UT Arlington)

In CMS, as illustrated in Figure 5, the hierarchy was entirely "flattened", meaning that a Tier2 site (as well as some Tier3s) could request a dataset, and pull data from any Tier1 or any Tier2, very much as in a peer-to-peer model. Space allocations including centrally managed space, as well as allocations by physics topic, and some space for local usage by users of a given Tier2. The number of copies of a given dataset is adjusted to give users accessing data at the various Tier2s comparable equal performance. The cumulative volume of data transferred by CMS between Tier1s and Tier2s and among Tier2s in the 12 months prior to this conference is shown in Figure 6.

Figure 5 (Left) The classic "push" Model developed by MONARC; (Right) The peer-to-peer-like Model used by CMS in 2009-2010 where a Tier2 can request a dataset from any other Tier1 or Tier2.
CMS also has been experimenting with an "Any Data, Anywhere" prototype architecture developed at Nebraska for distributed data access [19] based on the Xrootd protocol and software developed at SLAC, aimed at making small and medium scale end-user analysis as easy as possible, where CMS data objects are streamed as needed over the network directly to ROOT. The architecture is similar to the current data management architecture of the ALICE experiment, and follows the work by ALICE and the ROOT team to identify the leaves needed within ROOT trees, and to do intelligent object-based read-ahead to increase the efficiency of data access and processing. The global architecture and local access scheme with automatic streaming input from one or remote sites, shown in Figure 7, are now implemented at Nebraska, Caltech, UCSD, Purdue, Wisconsin and Bari, and the Fermilab Tier1 as well as several other Tier2 and Tier3 are also participating.

5. Responding to New Computing Models: Network and Other Requirements for Tier2s

Realizing the rapid evolution of the ATLAS and CMS Computing Models, with greater emphasis on agile data operations at many Tier2 and Tier3 to increase the working efficiency of physicists doing analysis, two workshops were held in June 2010 to consider the corresponding new network and other requirements. CERN and Caltech organized an LHCOPN Workshop on Transatlantic Networking for the LHC [20], and the WLCG and the LHC experiments hosted a workshop on Data Access and Management in Amsterdam [21]. The main themes of the workshops included improving connectivity to the Tier2s, and the use of the network to facilitate data analysis.
One outcome of the workshops was the formation of a Requirements Working Group led by Kors Bos (ATLAS) and including representatives of both the LHC experiments and the network providers [22], charged with investigating near-future network requirements, with a focus on the Tier2s. In September, the group issued its report [23], which listed the Tier1s and Tier2s in order of size, characterized the Tier2s in three classes, and gave the corresponding data throughput requirements, summarized in Table 1 from the report. The report noted that of the 10 Tier1 and 77 Tier2 sites listed by size, the top half would qualify as being in the Leadership class. The corresponding aggregate network provisioning requirement at the sites is thus clearly in the Terabit per second range. Note that the report did not explicitly consider long term network requirements evolution, although it noted that the sites classified as Minimal would be expected to migrate to the Nominal class, and those in the Nominal class would in some cases migrate to the Leadership class. The report also notes that the costs of meeting the network requirements need to be understood, and accommodated in budgets from now on, along with CPU power and storage.

| Category  | Throughput   | Target                                                                 |
|-----------|--------------|-------------------------------------------------------------------------|
| Minimal   | 1 Gbps       | Small Tier-2 installations: at the minimal connection speed a Tier-2 will be able to function, but will not be able to provide users with the same flexibility and quality of service. |
| Nominal   | 5 Gbps       | Normal Sized Tier-2 installations: at nominal connection speeds the samples can be updated in reasonable time and the Tier-2 storage can be updated at regular intervals. |
| Leadership| 10 Gbps and greater | Large Tier-2 installations: leadership Tier-2 facilities are significant analysis facilities supporting large numbers of analysis users. The high connection speed allows the large local storage to be updated and samples provided to several individual users working simultaneously. |

Table 1 Tier2 classifications defined by the Requirements Working Group [23]. Throughput refers to user throughput, so that the corresponding network provisioning requirements are estimated to be approximately 2, 10 and 20 Gbps respectively for the three classes.
6. Developing an Architecture to Serve Tier2 Network Needs

The LHCOPN [24] has been tasked with coming up with an architectural design to meet the Tier2s' needs. As of this writing, ideas for the new architecture have been submitted in the form of white papers, and will be considered at an upcoming LHCOPN meeting in January 2011. One concept put forward (by D. Foster, CERN/IT), sketched in Figure 8, is an open, federated, extensible "infrastructure of infrastructures", interconnecting open exchange points modeled after the GLIF Open Lambda Exchanges (GOLE) [25] and regional exchange points. All parties serving the LHC program, including members of the LHCOPN, research and education networks, mission oriented networks, cross-border dark fiber links and some commercial providers would be encouraged to connect. The growth of such a system is conceived to be organic, based on need and capability and using all opportunities to connect, in order to meet the needs. It is also realized that with no pre-determined central funding, the challenges will be adequate funding, interoperability, end-to-end coordinated operations bringing with it the need for global "system-level" monitoring.

7. Dynamic Circuit Networks

A major theme over the last six years has been the development of "hybrid networks" incorporating co-called lightpaths that allow the large flows typical of high volume LHC data transfers to coexist with the smaller flows characteristic of general-purpose networking, by allocating a defined bandwidth capacity using optical channels and/or dedicated Ethernet links to the large flows. The use of lightpaths, whether implemented at the optical layer (as in US LHCNet) or at the link layer (as in Internet2 and ESnet) also enables quality of service guarantees, including adequate bandwidth, low latency and/or jitter, and high availability through automated fallback over alternate network paths in case of link outages. Dynamic circuits have been developed over the same period, to enable demands for large flows to be scheduled, prioritized and managed. Major dynamic circuit projects include ESnet's OSCARS [26], Internet2's Dynamic Circuit Network (DCN) [27], and GEANT's AutoBAHN [28]. US LHCNet interworks with all of these and provides additional dynamic services at the optical layer across the Atlantic for high availability, through the use of its multi-service Ciena optical multiplexers. Dynamic circuits are used routinely in production to support most of the large volume traffic at the US Tier1 sites at Fermilab and BNL using LambdaStation [29] and Terapaths [30] software respectively, together with many Tier2 sites. A view of many of the networks and sites participating in the use of dynamic circuits across multiple administrative is given, from the Internet2 perspective, in Figure 9.
In 2010 a collaborative network middleware project was formed among Fermilab, Brookhaven and the University of Delaware to develop an End Site Control Plane Subsystem (ESCPs) [31], which is designed to address challenges such as cross-domain control plane signaling and interoperability, authentication and authorization in a grid environment, topology discovery, and dynamic circuit status tracking. ESCPS was presented by P. Demar during the CHEP 2010 conference.

One of the major projects extending dynamic circuits to many campuses is the Dynamic Network System (DYNES) project [32] funded by the US NSF, led by Internet2 together with Caltech, Michigan and Vanderbilt. DYNES will extend the use of dynamic circuits to at least 40 US universities and 11 Internet2 connectors, together with partner sites in Latin America, Europe and Asia. DYNES is based on a "hybrid" packet and circuit architecture composed of Internet2's ION service and extensions over regional and state networks to US campuses. It will connect with transoceanic (ACE, USLHCNet), European (GÉANT), Asian (SINET3 [33]) and Latin American (RNP and ANSP) R&E networks. It will build on existing key open source software components that have already been individually field-tested and hardened in part by the PIs: DCN Software Suite (OSCARS/DRAGON [34]), perfSONAR [35], the UltraLight Linux kernel, FDT [36], FDT/dCache, and FDT/Hadoop. The DYNES team will partner with the LHC and astrophysics communities, OSG, and Worldwide LHC Computing Grid (WLCG) to deliver these capabilities to the LHC experiments as well as others, broadening existing Grid computing systems by promoting the network to a reliable, high performance, actively managed component. The DYNES architecture is illustrated in Figure 10.
As shown in the figure, the OSCARS software handles user requests, inter-domain interactions, and scheduling of resources. The IDCP defines the format of messages that exchange information about local network topology, check resource availability, and coordinate circuit creation between networks. OSCARS works by accepting incoming requests from a user authenticated by a username and password or X.509 certificate. The user may request a circuit using (1) a web page interface or (2) through an external application that speaks the IDC protocol (such as LambdaStation or TeraPaths). A path is then calculated and the request is forwarded down the chain until each network has had a chance to verify the resources. Communication between domains is secured using X.509 certificates to identify the networks.

The wide acceptance and adoption of dynamic network services crossing multiple administrative domains as part of hybrid networks, and architectures such as the one involving open exchange points shown in Figure 8, will depend on the use and some further development of de facto standard circuit services, including those mentioned above and some others. It will also likely require the development of generic network service standards, to facilitate global inter-working among the different network circuit protocol and service stacks developed by several groups. Work on a standard Network Services Framework to support network service requests and management through a Network Service Interface (NSI) that allows an application or network provider to request network services from other network providers is now underway in the Open Grid Forum [37].

8. The Digital Divide in the Scientific Community and the ICFA SCIC

The rapid advance of network requirements, network capacity and advanced technologies in the developed world, some of which has been summarized above, makes it clear that it is as urgent as ever to lessen the so-called Digital Divide that separates scientists in developing regions from the rest of the world. This is particularly relevant for the LHC program, where physicists from all world regions require substantial network resources if they are to participate effectively from their home institutions, and to make effective use of the Tier3 and in some cases Tier2 facilities that they have taken great pain to deploy.
Since 2002, the Standing Committee on Inter-regional Connectivity (SCIC) [38], a technical panel of ICFA, has worked to monitor and measure the state and evolution of the world's networks serving the scientific community, to track and participate in state of the art network developments, and above to work towards eliminating the Digital Divide wherever possible, as reported annually in an extensive series of reports to ICFA. The SCIC's work on the Digital Divide includes low impact monitoring of network availability, packet loss rates and achievable throughput through its Monitoring Working Group led by R. Cottrell (SLAC); following the field of networking and disseminating news on events, trends, and technology roadmaps; gathering and sharing reports from continental and national research and education networks throughout the world to track and compare progress while identifying any major issues in many countries and regions, facilitating a worldwide comparison while demonstrating what some countries have achieved under difficult economic conditions; sharing information on and encouraging deployment of modern cost-effective network infrastructures and site facilities, holding workshops and disseminating tools for effective high throughput data transfers and advanced network monitoring; and providing on-site help in the development of cost effective network and grid infrastructures in less-developed regions.

The SCIC Monitoring working group uses are two complementary toolsets [39]: (1) PingER which uses the ubiquitous "ping" utility available on most modern hosts [40]. PingER provides low intrusiveness (usually ~100bits/s per host pair monitored) RTT, loss, jitter, and reachability. The low intrusiveness enables the method to be very effective for measuring regions and hosts with poor connectivity. Since the ping server is pre-installed on all remote hosts of interest, minimal support is needed for the remote host (no software to install, no account needed etc.), and (2) PerfSONAR is used to measure high network and application throughput between hosts with excellent connections, as can be found at HEP accelerator sites and Tier1 and most Tier2 sites, major grid sites, and major academic and research sites in N. America, Japan and Europe. The method can be quite intrusive (for each remote host being monitored from a monitoring host, it can utilize hundreds of Mbits/s or more for ten seconds to a minute, each hour). This method, which can be quite intrusive, provides expectations of throughput achievable at the network and application levels, as well as information on how to achieve it, and trouble-shooting information.

The PingER data and results extend back to the start of 1995. They thus provide a valuable history of Internet performance. Thanks in part to strong collaborations with ICTP Trieste and Pakistan's National University of Science and Technology, as well as CERN and DESY, PingER coverage has expanded to ~70 monitoring nodes in ~22 countries, that monitor over 900 remote nodes at over 750 sites in over 165 countries, including 169 sites in 50 African countries. These countries contain over 98% of the world's population and over 99% of the online users of the Internet. PingER's coverage as of December 2010 is shown in Figure 11.

To visualize the yearly trends, and any major changes in network performance, that typically happen in a stepwise fashion, Figure 12 shows the trend lines in the achievable throughput in each region, derived from the well-known Mathis formula [40].
Figure 11: Locations of PingER monitoring and remote sites as of Dec 2010. Red sites are monitoring sites, blue sites are beacons that are monitored by most monitoring sites, and green sites are remote sites that are monitored by one or more monitoring sites.

Figure 12: Derived throughput in kbits/sec from SLAC to each world region. The yellow line is to help show the typical exponential growth in achievable throughput in the developed world. By drawing horizontal lines parallel to the x-axis from the October 2010 points on the lines shown for each region, and finding the intercept with the line for Europe (extrapolating that line backward if needed), one can estimate how many years behind each region is, relative to Europe.
The slow increase for Europe in Figure 12 is partially an artifact of the difficulty of accurately measuring loss with a relatively small number of pings (14,400 pings/month at 10 pings/30 minute interval, i.e. a loss of one packet ~ 1/10,000 loss rate). Looking at the data one can see that at the present rate, East Asia and Australasia will catch up with Europe in the next few years. Russia, Latin America and the Middle East are approximately 5-6 years behind Europe but are catching up. South East Asia is about 9 years behind but with a comparable growth rate to Europe, while South Asia and Central Asia are about 12-14 years while also maintaining a comparable growth rate (so that the Digital Divide is not being reduced). Africa is now 18 years behind Europe, and even worse it has been falling further behind. If one extrapolates the trend lines for Africa and Europe then at the current rate Africa’s throughput will be 40 times worse than Europe’s by 2020.

Several examples of countries that have transformed themselves through the rapid installation of modern telecommunications infrastructures (Brazil, Slovakia and Romania; among many others reviewed in the ICFA SCIC annual reports [38] were presented at CHEP 2010. Figure 13 summarizes the status and developments in Romania, which has installed a close to state of the art national research and education network, and increased its national network capacity by a remarkable factor of more than $10^4$ times over the last 8 years.

![RoEduNet2 (ROMANIA)](image)

RoEduNet2 (ROMANIA) > 10,000X Since 2002: Pan-European "Role of Science in the Information Society” Ministerial Meeting with HEP Bucharest

Octavian Rusu

4240 Km Dark Fiber
600 Km with WDM
38 10G + 41 1G Waves
56 Sites

Separate Optical Control Plane (Nortel);
No regeneration:
Up to 1000 km spans

Cross Border Dark Fiber to Moldova

2001 – RoEduNet joined GEANT as partner
2006 – RoEduNet2 project approved
2007 – New modern data centers in Bucharest: National NOC and Bucharest NOC
2007 – More than 40 new routers installed in network, layer 3 of network completely upgraded
2008 – August – GEANT POP installed in Bucharest: 10 Gbps to GEANT, 2.5 Gbps committed
2008 – December – RoEduNet2 network in production
2010 – 1st CBF from Romania installed: Iasi – Chisinau (Moldova) DWDM segment operational

www.ces.net/events/2010/cef/p/rusu.pdf

Figure 13 The status and historical development of the Romanian national research and education network (Courtesy Octavian Rusu, Rodedunet2)

One of the most troubling aspects of the Digital Divide issue is the lack of relative progress in vast regions of Africa, a continent of 1 billion people beset by poverty, political unrest, lack of energy, lack of basic infrastructure and training, political unrest and corruption, and the ravages of disease (especially AIDs). In past years, major developments aimed at bringing much greater bandwidth to Africa via modern undersea cables have failed due to the lack of commercial viability, due in part to the difficulty of bringing the bandwidth to a sufficient number of institutions. However, as summarized in Figure 15, a new round of undersea cables [42] has come to both coasts of Africa,
including some spurred on by the 2010 World Cup being held in South Africa. While only time will tell, the prospect is that at least some of these ventures, including several already in operation, will succeed commercially, with a potential capacity increase of 1000 times. Other major factors contributing to the success are a vast ensemble of fiber infrastructures (noted in the figure), and the formation of the “UbuntuNet Alliance” [43] that brings together 12 existing and emerging national research and education networks in Eastern and Southern Africa.

9. Summary and Conclusions

The capacity and capability of the networks used and in some cases operated by the HEP community continue to advance, keeping pace with the exponentially growing needs. The capacity and reliability of the major continental and transoceanic networks has been one key factor supporting the remarkable ability of the LHC experiments to rapidly obtain a full range of physics results through worldwide data distribution, access and analysis involving hundreds of Tier1, Tier2 and Tier3 facilities. But along with the early successes has come the realization that more agile and efficient, and complex Computing Models, including more intensive use of the available Tier2 and Tier3 resources will be needed in the upcoming LHC run, as the experiments move from the rediscovery of the Standard Model to the discovery phase.

Figure 14 New undersea cables linking the coasts of Africa to Europe, India, Pakistan, and the Middle East.

Along with the new emerging Models come a new set of requirements, and the LHCOPN which has constructed and operates the core networks serving the Tier1s has been charged to develop a new architecture to adequately serve the Tier2s (and eventually the Tier3 sites as well). In fact the experiments and the network providers need to work together from now on, as the Models and the scale and nature of the architecture will continue to evolve.
In terms of scale, the transition to the new generation of networks and associated standards, with link speeds up to 100 Gbps, has already begun and will be well advanced by 2013 in continental networks, and by approximately 2015 if not sooner in transoceanic networks, driven by the exponential growth of global internet traffic. In terms of the quality of networks, dynamic network circuit services are taking hold, and will be a major direction by which the networks supporting HEP can schedule and manage high volume network flows, while continuing to support the multitude of smaller flows making up general purpose Internet traffic.

With the rise in scale and capability of networks in the developed world, the Digital Divide issue is as acute as ever, including in our own community. Reducing or eliminating the Divide is crucial if physicists in the developing world are to realize their dreams of participating as partners in the impending discoveries, including students and physicists working from their home institutions. While our community can help in many ways, the major successes have come through national, regional, metropolitan and campus initiatives, public and private, that have transformed nations through the development and deployment of modern network infrastructures, coupled to technical, educational and political programs that target enabling the population to use the existing technologies and tools effectively, and develop the next generation of technologies.

Among the growing number of national success examples, 2010 has brought renewed hope for the modernization of networks serving the African continent, and with it the hope for a rapidly growing African physics community in the coming decade.

References

[1] Further discussion of these trends, and a worldview of the research and education networks and advanced network projects supporting the HEP community may be found in the annual report of the ICFA Standing Committee on Inter-regional Connectivity at http://cern.ch/icfa-scic.
[2] S. Cotter, ESnet Update, Summer 2010 ESCC Meeting, Columbus, Ohio, http://www.es.net/pub/esnet-doc/ESCC-Meeting-Summer-2010/ESCC_v1.pdf
[3] See for example: http://www.internet2.edu/network/dc/, http://www.controlplane.net/, and http://www.es.net/pub/esnet-doc/DANMS08_1569141354_Guok_et-al.pdf
[4] DOE Office of HEP Request for Information on Transatlantic Networking, August 2010.
[5] MonALISA (Monitoring Agents using a Large Integrated Services Architecture) by I. Legrand et al. http://monalisa.caltech.edu
[6] HEP Science Network Requirements, Office of HEP Workshop, August 27-28 2010, Final Report. http://www.es.net/pub/esnet-doc/HEP-Net-Req-Workshop-2009-Final-Report.pdf
[7] GÉANT: http://www.geant.net/
[8] ACE: America Connects to Europe Project, J. Williams et al., Indiana University. http://internationalnetworking.iu.edu/ACE
[9] H. Ma, The ATLAS Experiment: Status, Detector Performance and First Physics Results, Conference on LHC First Data Conference, University of Michigan, Ann Arbor, Dec. 2010
[10] F. Gianotti, ATLAS Status and Highlights, ICHEP 2010 Conference, Paris, July 2010. http://webcast.in2p3.fr/2010/ichep/index.php?video=gianotti.ram
[11] PhEDEx - CMS Data Transfers. https://cmsweb.cern.ch/phedex/
[12] For information about 100G Ethernet, see http://www.100ge.info/. Information on the Optical Transport Network (OTN) standards of the ITU can be found in IEEE Communications, Volume 48, Issue 9, September 2010, pp. 44-75. http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=5560574
[13] Telegeography CommsUpdate, October 6, 2010, Global internet traffic growth remains strong in 2010. http://www.telegeography.com/cu/article.php?article_id=34653
See for example Telegeography CommsUpdate, June 22, 2009, *Trans-Atlantic bandwidth - the hangover lingers*. http://www.telegeography.com/cu/article.php?article_id=28963

US LHCNet: http://lhcnet.caltech.edu/

See J-P Elbers and Neil Peers, OTN-Compatible 40 GbE and 100 GbE Interfaces, ADVA AG Optical Networking, http://www.ieee802.org/3/hssg/public/nov07/elbers_01_1107.pdf

The MONARC Project (Models of Networked Analysis at Regional Centers for LHC Experiments, 1998-2000. http://monarc.web.cern.ch/MONARC/

Kaushik De, UT Arlington, *Data Caching Instead of Data Placement*, reported at ATLAS Software Week, Fall 2010.

B. Bockelman et al., *CMS Xrootd Architecture*, https://twiki.cern.ch/twiki/bin/view/Main/CmsXrootdArchitecture

A. Barczyk (Caltech), D. Foster (CERN) et al., http://indico.cern.ch/conferenceTimeTable.py?confId=88883#20100610.detailed

*Jamboree on Evolution of WLCG Data & Storage Management* http://indico.cern.ch/contributionDisplay.py?sessionId=0&contribId=11&confId=92416

Working Group members: K. Bos (ATLAS, Chair), H. Newman (Caltech, US LHCNet), W. Johnston (ESNet), J. Sobieski (NORDunet), K. Ullmann (DFN, DANTE), D. Foster (CERN), I. Fisk (CMS), A. Barczyk (Caltech, US LHCNet), E. Boyd (Internet2)

K. Bos et al., Networking Requirements Working Group Report, September 2010.

The LHC Optical Private Network. http://lhcopn.web.cern.ch/lhcopn/

GLIF Open Lambda Exchanges. http://www.glif.is/resources/

ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS), Chin Guok et al., http://www.es.net/oscars/

Internet2 Dynamic Circuit Network (DCN): http://www.internet2.edu/network/dc/ and its ION Service http://www.internet2.edu/ion/

GÉANT2 Automated Bandwidth Allocation across Heterogeneous Networks (AutoBAHN) system. http://www.geant2.net/server/show/nav,756

Lambda Station (Fermilab and Caltech), http://www.lambdadstation.org/

Terapaths (BNL, Michigan and SLAC), https://www.racf.bnl.gov/terapaths/project/introduction

ESCPs (Fermilab, BNL and Delaware), A. Bobyshev et al. 2010, *A collaborative network middleware project by Lambda Station, TeraPaths, and Phoebus*, J. Phys.: Conf. Ser. 219 062034, http://iopscience.iop.org/1742-6596/219/6/062034

DYNES, http://www.internet2.edu/dynes

SINET3: Japan's third generation Science Information Network, http://www.sinet.ad.jp/ Case study examples of use and the architecture of SINET3 as well as SINET4 that will start in April 2011 may be found at: http://www.sinet.ad.jp/document/sinet_doc/pamphlet/2010-04

DRAGON: Dynamic Resource Allocation via GMPLS Optical Networks, J. Sobieski et al., http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/WebHome

PerfSONAR: PERformance Service Oriented Network monitoring ARchitecture, http://www.perfsonar.net/

FDT: Fast Data Transfer, I. Legrand et al. (Caltech), http://monalisa.cern.ch/FDT/

Network Services Framework, NSI Working Group, J. A. Garcia-Espin et al., Grid Working Draft, August 3, 2010. https://forge.gridforum.org/sf/sfmain/do/downloadAttachment/projects.ggf-editor/tracker.submit_ggf_draft/artif6449?id=atch4908

ICFA Standing Committee on Inter-regional Connectivity. http://icfa-scic.web.cern.ch/icfa-scic/

The ICFA SCIC Monitoring Working Group Reports may be found at http://www.slac.stanford.edu/xorg/icfa/scic-netmon/

The PingER (Ping end-to-end reporting) Project, R. Cottrell et al., http://www-iepm.slac.stanford.edu/pinger/
[41] M. Mathis, J. Semke, J. Mahdavi, T. Ott, The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm, Computer Communication Review, volume 27, number 3, July 1997.

[42] http://manypossibilities.net/african-undersea-cables/ Version 23, September 2010. SLAC's IEPM project has been monitoring the impact of the Seacom undersea cable on round trip times and therefore achievable throughput in Africa: see https://confluence.slac.stanford.edu/display/IEPM/New+E.+Coast+of+Africa+Fibre

[43] UbuntuNet Alliance for Research and Education Networking. http://www.ubuntunet.net/

Acknowledgements

The author gratefully acknowledges the work of many colleagues in the LHC experiments, the ICFA SCIC, the US LHCNet team, CERN IT Department, Fermilab, BNL and SLAC, and the many networks and projects cited, as well as the support of the US Department of Energy (DOE Grant DE-FG02-08ER41559) and the US National Science Foundation (NSF Grants PHY-0427110 and PHY-0612805), that made this report possible.