



**International Committee for Future Accelerators (ICFA)
Standing Committee on Inter-Regional Connectivity (SCIC)
Chairperson: Professor Harvey Newman, Caltech**

**ICFA SCIC Report
Annexes
Networking for High Energy Physics**

On behalf of ICFA SCIC:

Harvey B. Newman newman@hep.caltech.edu

Azher Mughal azher@hep.caltech.edu

Dorian Kcira dkcira@caltech.edu

Justas Balcas justas.balcas@cern.ch

February 2017



Continental and Intercontinental Networks

| | |
|--|----|
| Annex 1: GEANT Status..... | 4 |
| Annex 2: Internet2 Overview and Status..... | 20 |
| Annex 3: ESnet Overview and Status | 36 |
| Annex 4: StarLight - An International/National Communications Exchange Facility..... | 53 |
| Annex 5: GLORIAD Status and Plan..... | 63 |
| Annex 6: Asia-Pacific Advanced Network (APAN) Status and plan..... | 72 |
| Annex 7: AmLight Project Status and Plan..... | 75 |

High Energy Physics Labs

| | |
|--|----|
| Annex 8: CERN Network Status and Plan..... | 88 |
| Annex 9: Fermilab Status and Plan | 92 |
| Annex 10: BNL Status and Plan..... | 97 |

National Networks

| | |
|---|-----|
| Annex 11: CANARIE (Canada) Status and Plan | 102 |
| Annex 12: SURFnet Status and Plan..... | 108 |
| Annex 13: GARR-X and GARR-X Progress (Italy) Status and Plan..... | 111 |
| Annex 14: CESNET2 (Czech Republic) Status and Plan..... | 115 |
| Annex 15: SANET (Slovakia) Status and Plan | 120 |
| Annex 16: PIONIER (Poland) Status and Plan | 121 |
| Annex 17: RoEduNet (Romania) Status and Plan..... | 126 |
| Annex 18: RENATER (France) Status Update | 132 |
| Annex 19: RNP (Brazil) Status Update and Plan..... | 134 |
| Annex 20: SPRACE (Brazil) Computing & Network Update | 144 |
| Annex 21: UERJ (Brazil) Tier2 Center Status and Plan | 157 |
| Annex 22: KREONET2 and KRLight (KOREA) Status and Plan..... | 159 |
| Annex 23: SINET4, SINET5 and HEPNet-J (Japan) Update..... | 166 |
| Annex 24: CERNET2 and CSTNET (China) Update | 169 |
| Annex 25: SingAREN (Singapore) Status | 174 |
| Annex 26: REUNA (Chile) Status | 177 |

Advanced Network Projects

| | |
|--|-----|
| Annex 27: A Next Generation Terabit/sec SDN Architecture and Data Intensive Applications for High Energy Physics and Exascale Science..... | 180 |
| Annex 28: AsyncStageOut - New component of the distributed data analysis system of CMS . | 197 |
| Annex 29: OSIRIS Open Storage Research Infrastructure | 200 |
| Annex 30: MonALISA Framework..... | 202 |

Annex 1: GEANT Status

Submitted by Cathrin Stover Cathrin.Stover@dante.org.uk

January 2017

The GÉANT project

The GÉANT project is a fundamental element of Europe's e-infrastructure, delivering the pan-European GÉANT network for scientific excellence, research, education and innovation. Through its integrated catalogue of connectivity, collaboration and identity services, GÉANT provides users with highly reliable, unconstrained access to computing, analysis, storage, applications and other resources, to ensure that Europe remains at the forefront of research.

Through interconnections with its 38 National Research and Education Network (NREN) partners, the GÉANT network is the largest and most advanced R&E network in the world, connecting over 50 million users at 10,000 institutions across Europe and supporting all scientific disciplines. The backbone network operates at speeds of up to 500Gbps and reaches over 100 national networks worldwide.

The network and associated services are co-funded by the European Commission through the GÉANT project (a collaboration of 38 partners consisting of the GÉANT organisation, 35 European NRENS and NORDUnet which represents the five Nordic countries).

Since its establishment over 20 years ago, the GÉANT network has developed progressively to ensure that European researchers lead international and global collaboration. Over 1000 terabytes of data is transferred via the GÉANT IP backbone every day. More than just an infrastructure for e-science, it stands as a positive example of European integration and collaboration.

GÉANT 2016 Highlights

The GÉANT network continues to deliver cost-effective and extremely high performance for all users. 2016 saw large increases in traffic across the core network, as well as almost 50% growth in global traffic, illustrating the value of GÉANT's European and international connectivity.

- Operational excellence: key to achieving 2016 objectives was to maintain the operational excellence of the established GÉANT services, whilst achieving significant economies on the costs of the backbone network.
- Traffic growth: core IP traffic increased by over 64%, which, when combined with dedicated services for large users, resulted in total traffic volume of 1425 Petabytes during 2016.
- Evolving the network: the second iteration of the network evolution plan was developed, with great progress made in such areas as fibre sharing, SDN and packet optical integration. Future work will provide greater clarity on how the GÉANT network will need to evolve to meet the demands of researchers in the future and to assist in the delivery of the European Open Science Cloud.
- GÉANT Testbed Service: 5 new GTS nodes deployed in Europe, supporting innovative uses of the network, and developed the GTS roadmap through to 2017.
- Regional connectivity: recommendations from the regional study completed in GN3plus were implemented, including a number of newly procured 10Gbps circuits in Southern

and Eastern Europe providing improved connectivity to the local NRENs at greatly reduced cost.

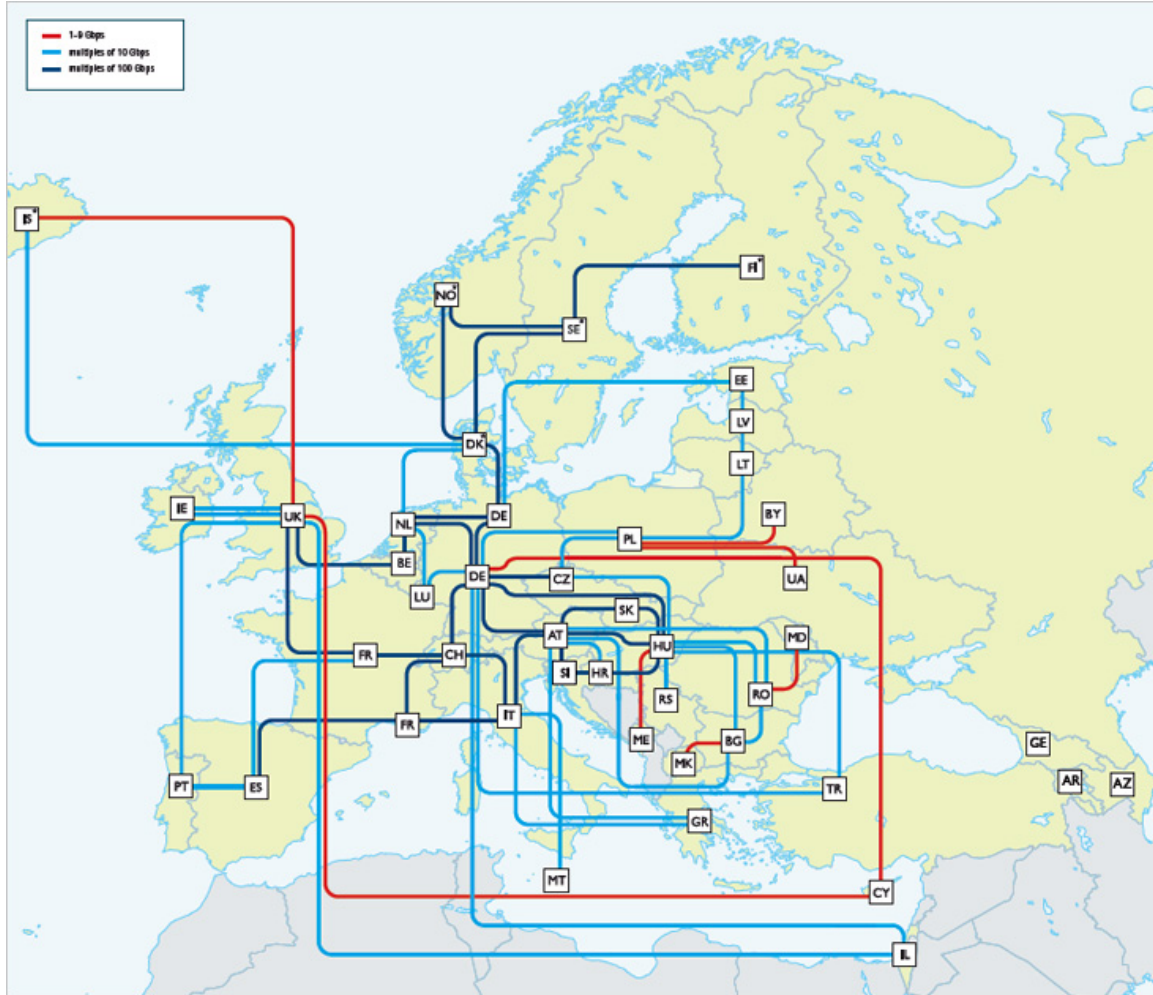


Figure 1: The European GEANT network in 2016

Support to CERN/LHC

In 2016, GÉANT provided support for to CERN/LHC on the following activities:

- The deployment of a second 100Gbps link between CERN and the Wigner centre in Budapest, Hungary.
- The expansion of LHCONE to Asia-Pacific, coordinating the deployment of a LHCONE VRF in the TEIN backbone and started discussions about the peering and transit policies within LHCONE. ThaiREN is the first Asian NREN to join LHCONE in the TEIN network.
- The inclusion of Belgium and Poland to the LHCONE network in Europe, and the starting of discussions to include Portugal in 2017.

2016 has been a pivotal year for the LHC-related network activities, mainly due to the dramatic increase in the data science production for the LHC experiments, which has exceeded the already high estimates for Run2 of LHC.

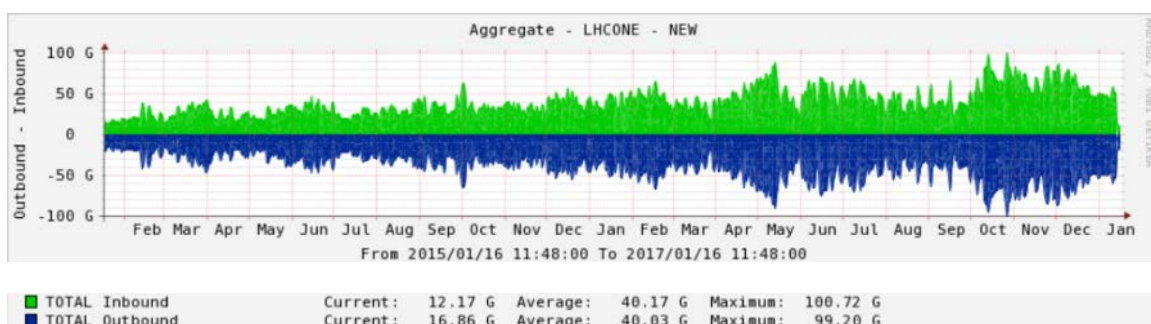


Figure 2: 2015-2016 LHCONE total traffic over GÉANT

Shown in the picture is the trend of the LHCONE traffic over the whole GÉANT backbone, spanning over 2015-2016. The picture shows massive increase in average traffic that happened in 2016 (plus 72% compared to the previous year), with peaks exceeding 100Gbit. During the year, GÉANT has also been very active in engaging with the LHC community in regions of the world not already connected to the LHCONE network, with particular attention to the Asia-Pacific area. The support to TEIN*CC in this field has brought to the creation of a LHCONE VRF onto the TEIN backbone, allowing the connection of ThaiREN as the first NREN to join LHCONE. In parallel, thanks to the continuous engagement with the other Asian partners, there's also been a remarkable advancement of the connectivity between the other networks active in the area, like APAN, ASGC and KISTI-KREONET, that now have an agreed plan of mutual peering in place. Discussions with TENET (South Africa) and REUNA (Chile) have started, in order for them to join the L3VPN.

At the same time, also the LHCONE connectors in Europe have increased, adding some new countries like Poland and Belgium, and paving the way for new connectors to come up in the following year.

GÉANT Services :

GÉANT offers a wide range of innovative services to enhance the experience of its network. Advanced connectivity, network support and access services make life easier for NRENs,

institutions, projects, researchers and students alike: GÉANT's portfolio comprises a range of aligned services that allow it to cater for the most varied and demanding needs.

The GÉANT network offers the high speed and huge capacity essential for the world's biggest science projects, routinely carrying more than a petabyte of information every day (that's 1 *million* gigabytes). The award-winning GÉANT network continues to set the standard for speed, service availability, security and reach, delivering levels of performance that commercial network providers can't provide.

The GÉANT and NREN networks underpin the work of a wide range of e-infrastructure and scientific research projects by providing a high performance, reliable and cost-effective communications platform across the research and education (R&E) community. Service options cover IP, dedicated private connections, virtual private networks and roaming options.

IP Networking The core of the GÉANT network is the world-class IP backbone. GÉANT IP provides general-purpose IP transit for national research and education networking (NREN) organisations and other approved research and education partners and providers.

Its function is to provide a private service for IP (internet protocol) traffic that is separated from general-purpose access to the internet. Offering users connection speeds of up to 100Gbps, GÉANT IP provides the essential communication service that supports inter-NREN connectivity.

In addition to the core IP networking services GÉANT offers users a range of specialised connectivity options. Many performance-critical services require guaranteed performance levels and additional security that is difficult to achieve through shared IP services. In particular, applications such as data centre backup and replication, real-time mission-critical services and broadcast quality video need the guaranteed bandwidth and low latency that only dedicated circuits separated from general IP traffic can offer.

Point-to-point services provide dedicated connectivity between two sites over the existing infrastructure without the cost and difficulty of building and managing a dedicated physical network. This type of connectivity can provide fixed latency between collaborating institutions, a high level of security and, if needed, guaranteed bandwidth of up to 100Gbps. Furthermore, the possibility of providing a L2 Ethernet channel end-to-end allows the use of network and transport protocol other than the classic TCP/IP, enabling users to experiment with new ways of using network connectivity. Good examples of such advanced use of the service are the SMARTfire and the InfiniCortex projects, using experimental streaming transport protocol and the InfiniBand network stack, respectively. There is also the long-lasting use in Radioastronomy and for the ITER project (linking the Cadarache facility in France to the Elios supercomputer in Rokkasho, Japan).

VPN Services Many projects may require teams across Europe to be able to collaborate effectively with enhanced privacy. By creating a Virtual Private Network (VPN), all sites on the VPN can communicate without the need to arrange separate physical networks, while benefiting from the privacy and security of a private infrastructure. GÉANT can provide VPNs between many sites over great distances within Europe and reach the USA (via Internet2 and ESnet), Canada (via CANARIE), and Asia.

Open Interconnectivity As international and public-private partnerships grow in importance within the R&E sector, so the high-performance, flexible and neutral interconnection points provided by the GÉANT Open service can offer new opportunities. Users can connect their own circuits – at 1Gbps, 10Gbps or 100Gbps – and can then request interconnections with any other participant.

GÉANT Testbeds Service The GÉANT Testbeds Service (GTS) delivers integrated virtual environments as 'testbeds' for the network research community. GTS is designed for researchers of advanced networking technologies to help support testing and development over a large-scale, dispersed environment. GTS can support multiple projects simultaneously and isolates them from each other and from the production GÉANT network to provide security and safety. This facility is leading the way in providing facilities to help develop the next generation of internet services.

eduroam provides 50 million students and researchers with access to thousands of wi-fi access points in over 70 countries using a single, secure login facility - making international collaboration much easier. Over 5 million international logins a day are enabled by eduroam.

Management and Support

The connectivity delivered by GÉANT is supported by a comprehensive range of network monitoring and management services. These optimise network performance by providing 24x7 monitoring across the GÉANT Service Area infrastructure, enabling fast identification and remedy of any faults on the network as well as providing powerful security to prevent and detect malicious attacks.

Users benefit from the range of GÉANT network monitoring, security and support services employed by NRENs to assure optimum performance for projects and institutions. The areas of tools and services in this group include performance measuring and monitoring, performance enhancement and security.

- **Performance measuring and monitoring** - Analysing performance in global research networks is complex since any single path might go through several domains – campus, local and national networks as well as the GÉANT backbone. Offering comprehensive multi-domain monitoring features, GÉANT's perfSONAR services allow users to access network performance metrics and perform network monitoring actions across multiple domains, ensuring that any source of congestion or outage on a point-to-point connection can be quickly and easily identified and addressed.
- **Performance enhancement** - The Performance Enhancement Response Team (PERT) provides an investigation and consulting service to academic and research users on their network performance issues. The service is achieved via eduPERT, the federated structure that combines the PERTs from the local institutions, NRENs and GÉANT and fosters knowledge-sharing across the GÉANT network community. eduPERT is part of GÉANT's commitment to helping users get the best performance from their connections.
- **Security** - In an online world, network security is of paramount importance. GÉANT takes a proactive approach to security to maintain the integrity of the network, implementing advanced defences that offer sophisticated handling of network incidents.
- Securing the GÉANT Service Area network elements through design and implementation of recommended access and usage policies.

- Building proactive security services based on incident databases, anomaly detection tools and common procedures for mitigation of denial-of-service attacks.
- Defining a common approach and processes for coordinating responses to security issues.

By providing strategies for incident prevention, detection and handling, the GÉANT security systems will allow users to keep network domains secure by monitoring traffic and routing information.

Trust, Identity & Security

GÉANT and its NREN partners provide technologies that build trust, promote security and support the use of online identities. This is an essential component of many infrastructure projects by bringing together services and users in a scalable, manageable and secure manner.

eduGAIN enables the trustworthy and secure exchange of authentication, authorisation and identity (AAI) information. It interconnects identity federations around the world, simplifying access to content, services and resources. eduGAIN provides a pan-European Web Single Sign On (Web SSO) (i.e. a single digital identity and password) to access all services provided by the participating federations and their affiliated service providers. This service is of special interest for distributed infrastructures or data archives, allowing data to be retained locally while researchers access data sets from different locations via a single sign on.

Clouds Services

Cloud services offer higher education and research organisations the opportunity to become more agile and provide their users with a wider range of IT services at a lower cost. GÉANT provides the platform for users to access cloud services and, through its cloud service catalogue, works with other e-infrastructure projects and commercial cloud service providers to help deliver innovative services to research and education institutions and their users.

GÉANT is actively helping national research and education networking organisations (NRENs) to deliver cloud services to their communities, with the right conditions of use. For example GÉANT and the NRENs have made arrangements with two of the biggest cloud service providers – Amazon and Microsoft to dramatically reduce the data connectivity costs associated with cloud computing - providing real cost savings to the R&E community. The goal is an attractive, well-balanced portfolio of cloud services to support European research and education.

OGF NSI protocol developments

The Network Service Interface (NSI) standard defines a suite of protocols that enable the seamless delivery of dynamic circuit provisioning around the world. NSI has been standardized within the Open Grid Forum (OGF) and aims to enable full interoperability between the different underlying dynamic circuit technologies in the R&E networks. The NSI working group has now released draft NSI CS v2.1 – This is backward compatible with CS v2.0. Work is ongoing in collaboration with the Global Lambda Integrated Facility (GLIF) to NSI-enable and extend the automated Global Open Lightpath Exchanges (auto-GOLEs). The full suite of NSI draft documents are available to download from here: https://redmine.ogf.org/dmsf/nsi-wg?folder_id=6526

SDN pilot Infrastructure

GÉANT is building an SDN pilot infrastructure based on Corsa white box switches deployed in 5 locations and interconnected by 10Gbps lambdas. The locations of the facilities are London, Paris, Amsterdam, Prague and Milan. Each location includes one Corsa DP2100 box which is shared with the GTS infrastructure. Sharing of the infrastructure is made possible by the Virtual Switch Context (VSC) methodology developed by Corsa and JRA2. Figure 1 below shows the underlying physical infrastructure available in GEANT for deployment of SDN pilot slices.

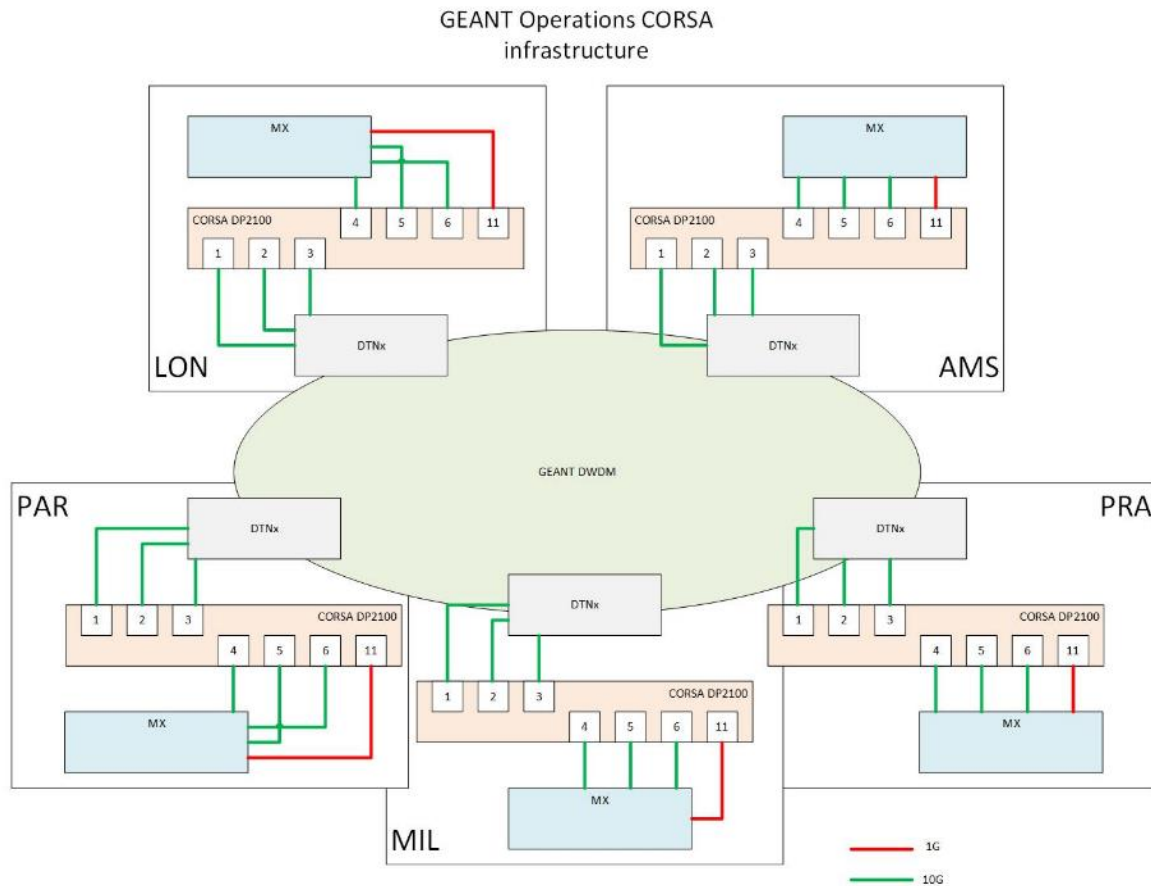


Figure 3: GÉANT Operations CORSA infrastructure

VSCs on Corsa boxes are OpenFlow enabled and a dedicated L3VPN is used for connectivity with the controller. Users gain access to the data plane of the SDN pilot infrastructure via their normal interconnect at their local GÉANT MX router.

The pilot service will initially be used to validate the L2 BoD SDN service use case in an operational environment. In this first use case the controller makes use of ONOS with the DYNPAC extension developed by JRA1. SDN BoD is NSI enabled. As additional use cases are completed (SDX and L3 SDN) these will also be piloted on this environment. The figure 2 below shows how the slice for SDN BoD piloting could be provisioned.

BOD SDN Pilot Slice Overview

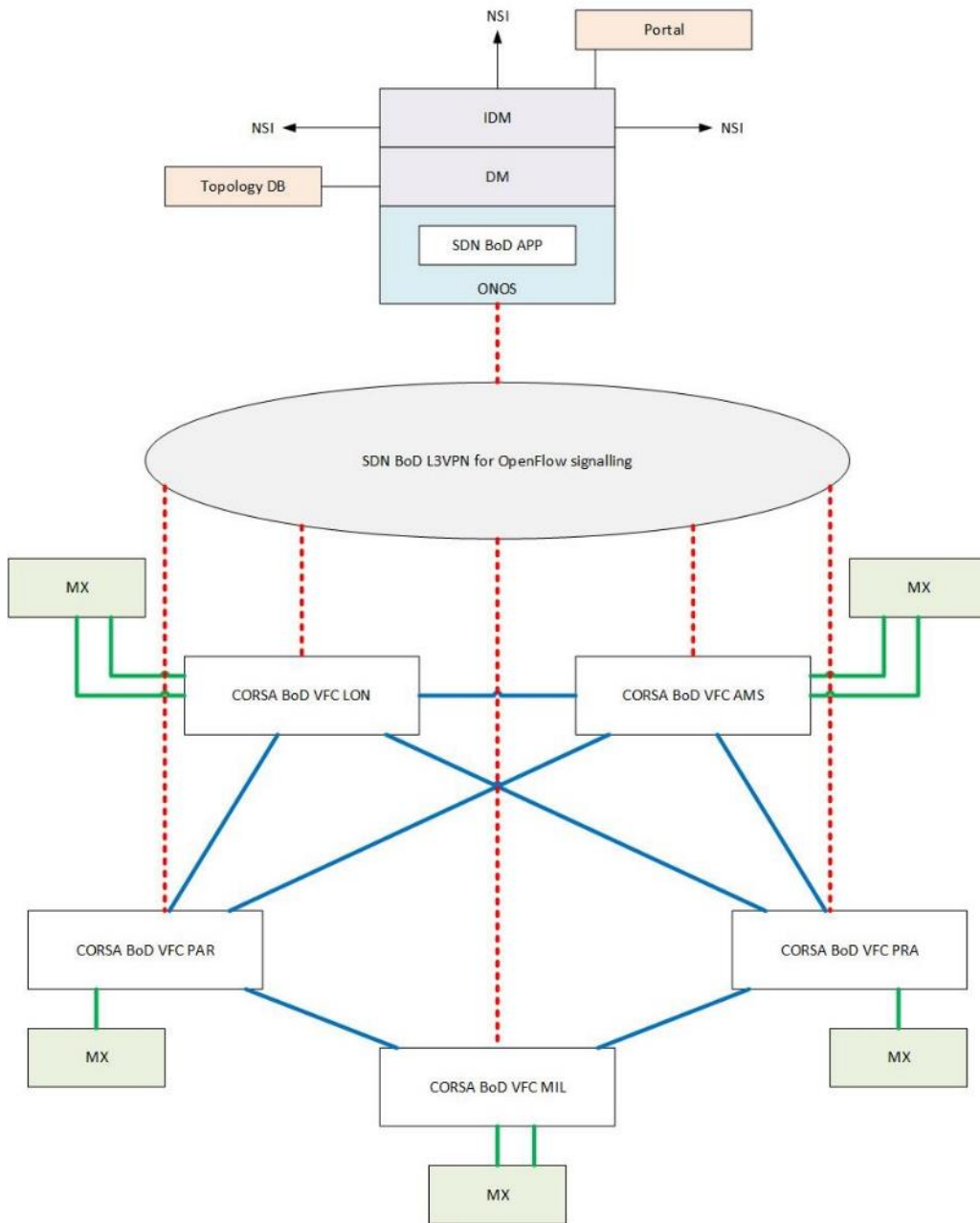


Figure 4: BOD SDN Pilot Service Overview

GÉANT International Connectivity:

GÉANT provides a significant number of connections to partner research and education networks beyond Europe.

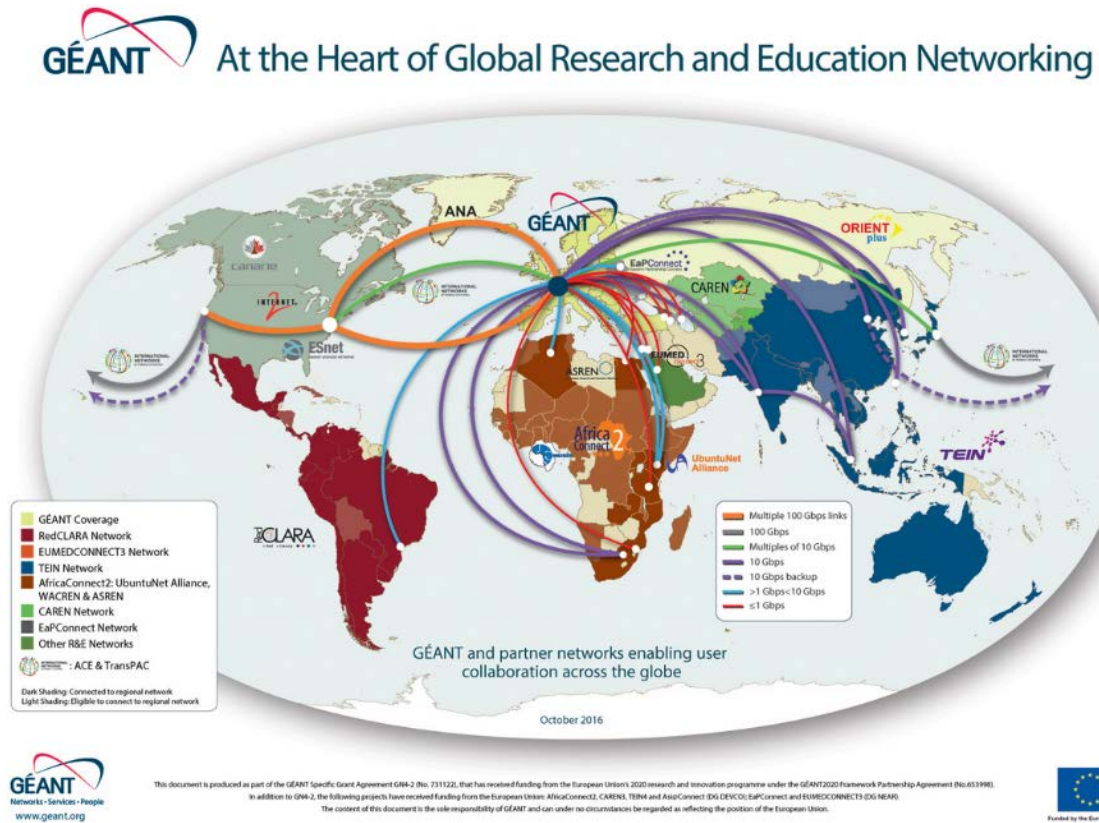


Figure 5: GÉANT at the Heart of Global Networking

During 2016, GÉANT collaborated with Indiana University and the African regional NRENs in the proposal and planning of the **NEAAR** (Networking European, African and American Researchers) project. NEAAR is an NSF funded project and follows on from the successful ACE project. NEAAR will provide 100G transatlantic capacity and will focus on training support to institutions connected to the African R&E networks.

Complementing the GÉANT network during 2016 are several GÉANT-managed and EU-funded projects, namely AfricaConnect 2, EUMEDCONNECT3, EaPConnect and CAREN3. The Asian-Pacific TEIN network, managed by TEIN* Cooperation Center, the RedCLARA network managed by RedCLARA, are also assisted by GÉANT and connected to its network. 2016 saw the shut-down of the C@ribNet network in the Caribbean.

The AfricaConnect project terminated on 6 May 2015. The AfricaConnect2 project started in June 2015. **AfricaConnect2** has a pan-African scope and is implemented through three geographically separate clusters. GÉANT, ASREN, WACREN and the UbuntuNet Alliance collaborate closely in the execution of AfricaConnect2. The main focus of activity in the Southern and Eastern, as well as Northern clusters has been the ongoing procurement of new or upgraded connectivity to the benefit of NRENs in these two regions. The first connectivity implemented through AfricaConnect2 is the improvement of Algerian connectivity to 2.5Gbps. In West and Central

Africa, the activity has focused on the preparation of the procurement for both connectivity and equipment, as well as ensuring that the recently created NRENs in the region will be ready to financially contribute to the project. Technical training and capacity building have been delivered across all three regions.

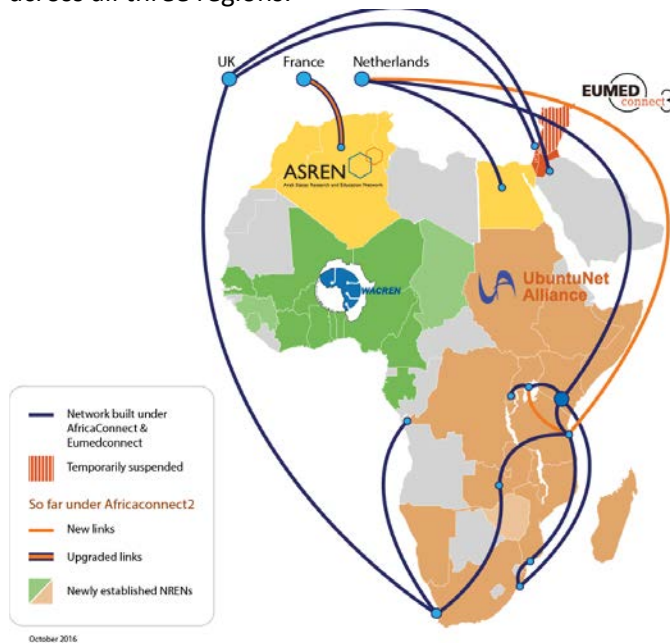


Figure 6: African R&E connectivity delivered through AfricaConnect

The EU-funded **EUMEDCONNECT** project supports research and education (R&E) networking for the Eastern Mediterranean (Jordan, Lebanon and Palestine). Now in its 3rd phase, it provides high-capacity international internet connectivity for academic and scientific collaborations. The project is run by the GÉANT organisation in partnership with ASREN, the local NRENs and the NRENs of Cyprus, France, Greece, Italy and Spain.

Through its earlier phases the EUMEDCONNECT programme also provided a regional R&E network for North Africa, with Algeria, Egypt, Morocco and Tunisia benefiting since 2004. In July 2015, the North African countries became partners in the new AfricaConnect2 project, also managed by GÉANT and ASREN. Close links with the EUMEDCONNECT3 community are being maintained as the two projects bring the African and Arab R&E communities together. Major users include educational and scientific collaborative projects with particular pertinence to problems affecting the Mediterranean region, such as desertification studies, tele-health, water shortage, sustainable farming and remote education.

In early 2016, the American University of Beirut (AUB) in Lebanon allocated 10 Mbps of its internet bandwidth to interconnect to the GÉANT network via ASREN's London hub for a pilot connection. The success of led to the capacity being upgraded to 320 Mbps in November and accelerated plans for an NREN in Lebanon (LERN). The EC has agreed to extend its share of project costs: this expected to lead to capacity upgrades in Jordan to support the SESAME synchrotron radiation facility as it enters production phase and to the re-connection of Palestine.

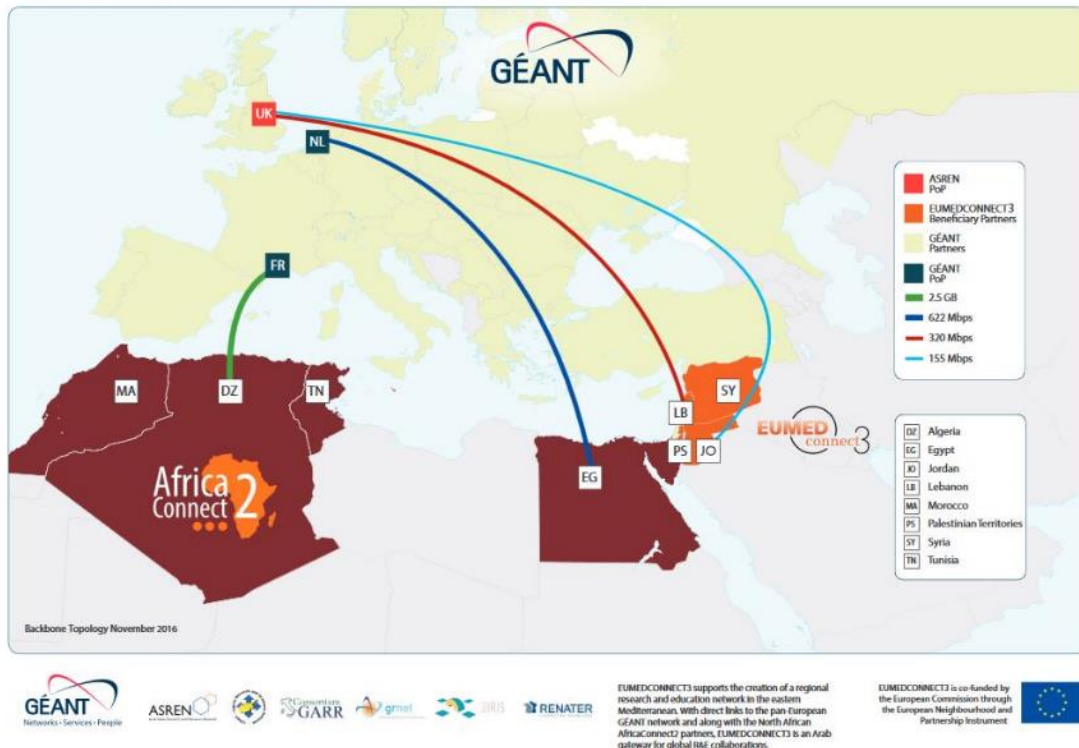


Figure 7: North African Connectivity to GÉANT

EAPConnect project was launched at a high-level ministerial event held in Luxembourg in June 2015 and officially commenced on 1 July 2015.

The project establishes and operates a high-capacity broadband internet network for research and education across the six Eastern Partnership (EAP) partner countries – Armenia, Azerbaijan, Belarus, Georgia, Moldova and Ukraine.

In 2016, the project procured connectivity to three South Caucasian countries: Georgia, Armenia and Azerbaijan, as well as Belarus.

Knowledge sharing and human capital building between the EaP project partners and the EU supporting partner NRENs took place in a designated project meeting with presentations by SURFnet, PSNC and LITNET describing and discussing their service portfolios and advising on possible options for the EaP region. To further progress the knowledge transfer of specific topics of interest to the EaP NRENs, the project held two successful workshops at SURFnet and PSNC. These topics included business strategy, innovation, eduroam, eduGAIN, security training, cloud services and support for education.

Researchers from Poland and Belarus exchanged experiences in collaborative robotics applications during a robotics workshop organised by the EaPConnect project at the end of

November 2016 at Poznań Supercomputing and Networking Center (PSNC), the operator of the Polish national research and education network (NREN). High-speed network connectivity for remote collaboration creates an opportunity to boost cooperation in solving challenges related to the application of complex robotic systems in key areas such as smart living spaces, precision farming or crisis management.

A first **Eastern Partnership E-infrastructure Conference**, EaPEC, was organised by EaPConnect and GRENA, the Georgian NREN, in Tbilisi, Georgia, on 6-7 October.

First regional **‘Enlighten Your Research’** contest – EYR@EAP 2016

Another highlight of EaPEC was the presentation of the Eastern Partnership region’s first ‘Enlighten Your Research’ (EYR) awards. This programme invites researchers to submit proposals that highlight how access to advanced networks, technologies and computation would significantly improve their research and discovery process, and was chosen by EaPConnect as a vehicle for engaging with research communities in the EaP region.

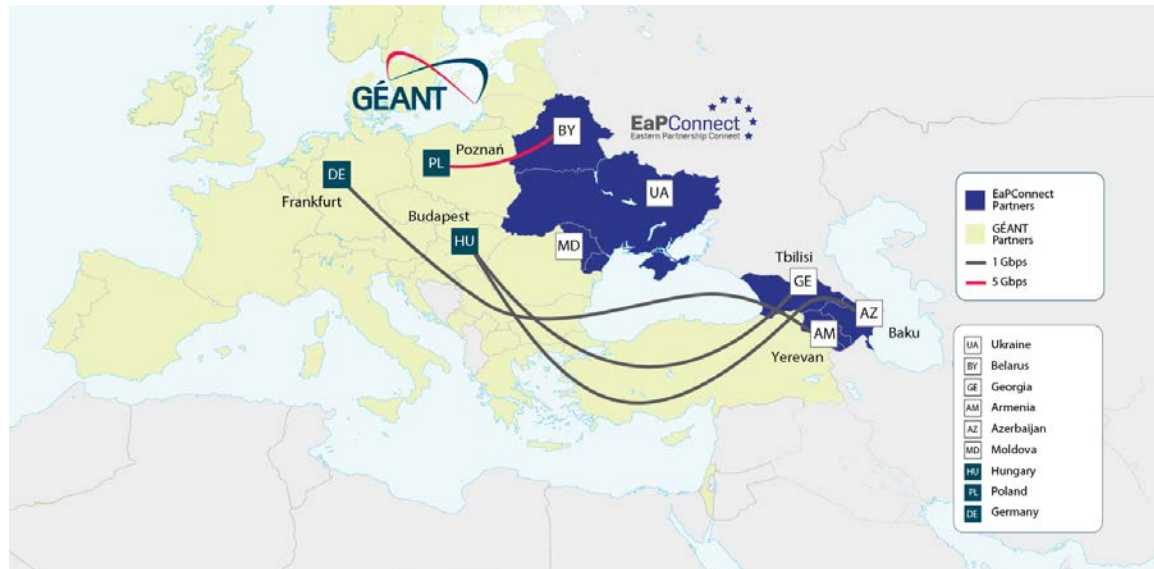


Figure 8: EAPConnect connectivity to GÉANT

The 3rd phase of the **CAREN project** in Central Asia started in July 2016 With 4.5M Euro initial EU co-funding the project will run up to 2019. CAREN3 is initially reconnecting Kyrgyzstan and Tajikistan where the governments have signed bilateral financing agreements with the EC. Kazakhstan, Turkmenistan and Uzbekistan are also eligible to join the project subject to EC approval and similar government financing agreements.

Connectivity for Kyrgyzstan and Tajikistan has been re-tendered and will be operational from early 2017. Existing and future collaborative projects span areas such as environmental monitoring, solar energy, telemedicine, the digitalisation of cultural heritage and e-learning which are to be re-started and further developed.

Following **TEIN4**, **Asi@Connect** marks the new five-year project phase of the TEIN initiative launched at the end of 2016 and managed by TEIN*CC in Korea. With substantial EC financial

support, increased effort and funding will go into developing new network services and capacity building, to extending the R&E footprint to additional Asia Pacific countries, promoting and supporting user communities to utilise R&E networking as well as to improving public internet access in least developed countries.

During 2016 new connectivity between **India** and GÉANT was established through links provided by NKN, initially at 2 x 5 Gbps capacities. This is additional to the TEIN4 project's EU-India connection which was upgraded from 2.5 to 10Gbps in 2016, with NKN co-financing the link as a TEIN project partner. These interconnections provide the highest R&E capacity for India's participation in high energy physics and supports other Asia/Europe collaborative programmes in areas such as Earth observation, climate research, food security, delivery of e-health and e-learning.

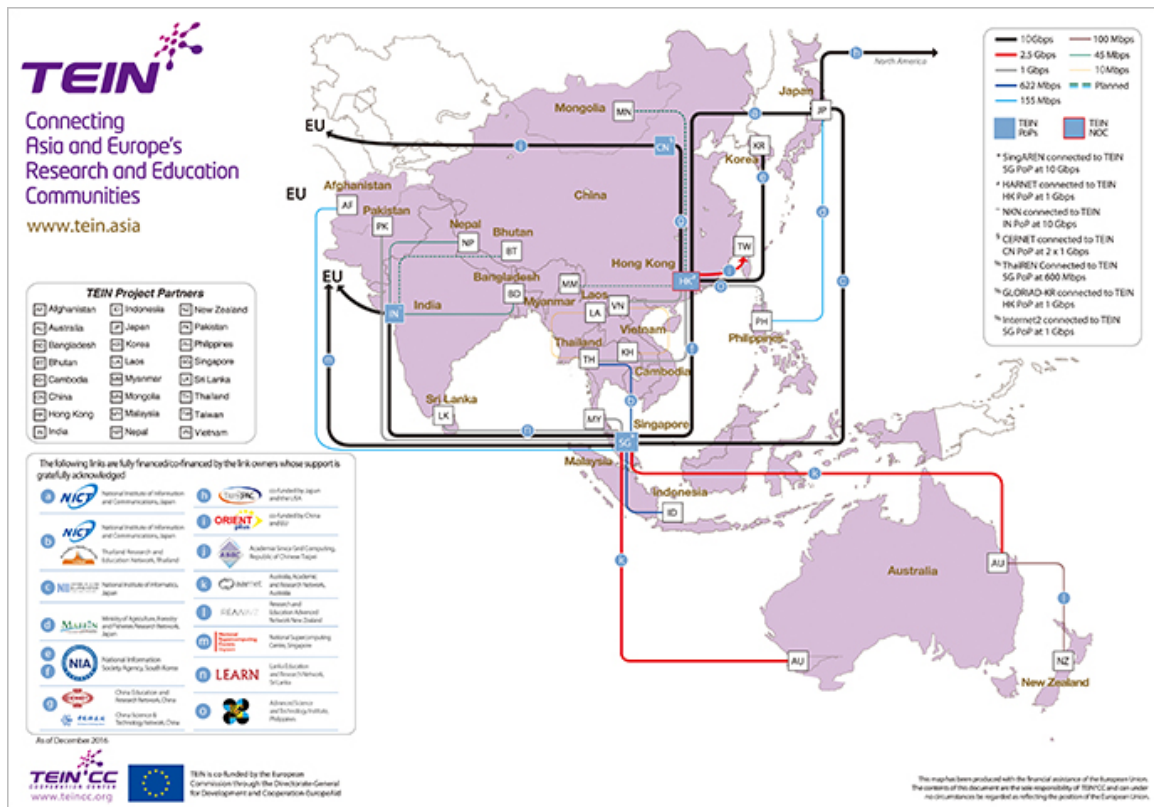


Figure 9: The TEIN network

MAGIC (Middleware for collaborative Applications and Global Virtual Communities) follows the success of the ELCIRA project. Led by RedCLARA, and with partners in Latin America, Europe, West and Central Africa, Eastern and Southern Africa, North Africa and the Middle East, Central Asia and Asia-Pacific, MAGIC aims to improve the ability of researchers and academics to collaborate globally. This is done by expanding eduroam and identity infrastructures around the world, and establishing middleware to enable NRENs to share services and real-time applications.

The Colaboratorio collaboration platform, developed by RedCLARA and which makes web-based collaboration tools including webconferencing, large file transfers and user group management, has been made available in other world regions including West and Central Africa and Central Asia.

MAGIC has created four Global Science Communities in the fields of e-health, biodiversity, the environment and remote instrumentation: These are supported by the project through the running of regular virtual meetings of participants from around the world to share experiences and knowledge. MAGIC also disseminates information on funding opportunities to these groups and more widely via the Colaboratorio. It also provides training in the use of the global collaboration tools implemented by the project.

MAGIC has a duration of two years from 1 May 2015 and will end in 2017.

The TransAfrican Network Development (**TANDEM**) project aims to create favourable conditions for WACREN, to enable it to draw maximum benefit from the forthcoming AfricaConnect2 project, and ensure its integration into the global Research and Education networking community and its long-term sustainability.

TANDEM enhances dialogue between WACREN and stakeholders in the West and Central Africa region through the creation of the PODWAG (Policy and Donors West and Central Africa Group) includes representatives of regional organisations, ministries, universities, research centres, international development organisations, and telecommunications operators and regulatory bodies. The purpose of the PODWAG is to provide a forum for discussion and progress on policy, funding sources and regulation issues relating to research and education networking.

TANDEM also engages with end users in the region. To do this the project has created National Focal Points, individuals who act as a liaison point between the NREN and end users. The Focal Points have assisted in the implementation of a survey of end user needs which will inform planning for services to be provided by WACREN and NRENs in the region beyond the end of TANDEM.

The project runs for 24 months from 1 May 2015 and will end in 2017. It is coordinated by the Institut de Recherche pour le Développement (France). Other partners in addition to GÉANT are WACREN (Ghana), RENATER (France), CIRAD (France), Brunel University (UK), UbuntuNet Alliance (Malawi) and RedCLARA (Uruguay).

International Connectivity: NRENs reached via GÉANT by World Region

AMERICAS & THE CARIBBEAN

- ARANDU (Paraguay)
- CANARIE (Canada)
- CEDIA (Ecuador)
- CoNARE (Costa Rica)
- CUDI (Mexico)
- ESNet (USA)
- INNOVA|RED (Argentina)
- Internet2 (USA)
- NISN – NASA (USA)
- NLR (USA)
- RAGIE (Guatemala)
- RAICES (El Salvador)
- RAU2 (Uruguay)
- REACCIUN2 (Venezuela)
- RedCyT (Panama)
- RENATA (Colombia)
- REUNA (Chile)
- RNP (Brazil)

EUROPE

- ACOnet (Austria)
- ARNES (Slovenia)
- BASNET (Belarus)
- BELNET (Belgium)
- CARNet (Croatia)
- CESNET (Czech Republic)
- CYPNET (Cyprus)
- DFN (Germany)
- EENet (Estonia)
- FCT (Portugal)
- GARR (Italy)
- GRNET (Greece)
- HEAnet (Ireland)
- Holy See Internet Office (Vatican City)
- BREN (Bulgaria)
- IUCC (Israel)
- Janet (United Kingdom)
- LITNET (Lithuania)
- NIIF/ HUNGARNET (Hungary)
- NORDUnet (Denmark, Finland, Iceland, Norway, Sweden)
- PSNC (Poland)
- RedIRIS (Spain)
- RENAM (Moldova)
- RENATER (France)
- RESTENA (Luxembourg)
- RoEduNet (Romania)
- SANET (Slovakia)
- SigmaNet (Latvia)
- SURFnet (The Netherlands)
- SWITCH (Switzerland)
- MARNET (FYR of Macedonia)
- ULAKBIM (Turkey)
- University of Andorra (Andorra)
- University of Malta (Malta)
- UoB/AMRES (Serbia & Montenegro)
- MREN (Serbia & Montenegro)
- URAN (Ukraine)

ASIA & OCEANIA

- AARNet (Australia)
- ASGC (Taiwan)
- BdREN (Bangladesh)
- CamREN (Cambodia)
- CERNET (China)
- CSTNET (China)
- HARNET (Hong Kong)
- INHERENT/ITB (Indonesia)
- JGN-X (Japan)
- KOREN (Korea)
- KRENA-AKNET (Kyrgyzstan)
- KREONET (Korea)
- LEARN (Sri Lanka)
- LERNET (Laos)
- MAFFIN (Japan)
- MYREN (Malaysia)
- NKN (India)
- NREN (Nepal)
- PERN2 (Pakistan)
- PREGINET (Philippines)
- REANNZ (New Zealand)
- SINET5 (Japan)
- SingAREN (Singapore)
- TARENA (Tajikistan)
- ThaiRen/ThaiSARN (Thailand)
- ThaiRen/UniNet (Thailand)
- TWAREN (Taiwan)
- VINAREN (Vietnam)

AFRICA & MIDDLE EAST

- ANKABUT (United Arab Emirates)
- ARN (Algeria)
- ASREN (Jordan)
- ENTSTINET (Egypt)
- EUN (Egypt)
- KENET (Kenya)
- LERN (Lebanon)
- MoRENet (Mozambique)
- QNREN (Qatar)
- RENU (Uganda)
- RwEdNet (Rwanda)
- SARInet (Saudi Arabia)
- TENET (South Africa)
- TERNET (Tanzania)
- ZAMREN (Zambia)

Annex 2: Internet2 Overview and Status

February 2017

<http://www.internet2.edu>

Submitted by John Hicks (jhicks@internet2.edu) and Matt Zekauskas (matt@internet2.edu)¹

Introduction:

The Internet2 community provides U.S. research and education (R&E) with a complete portfolio of advanced technologies to support the most demanding science collaborations, accelerate research discovery, and spark tomorrow's essential innovations.

The same community that played a seminal role in the creation of the modern Internet is providing a completely new portfolio of disruptive innovations to provide new dimensions of support for scientific research, once again.

Owned by U.S. research universities, Internet2 consists of more than 400 universities; corporations; government agencies; laboratories; and other national, regional and state research and education networks. It includes organizations representing more than 50 countries, and is governed by an executive Board of Trustees and strategic advisory councils.

Advanced Internet2 community technologies combine within the world's most sophisticated network platform with dynamic and tailored R&E cloud applications, and a trusted federated identity framework providing an integrated platform that supports the most advanced Big Data science collaborations. The platform serves as a national asset for R&E by offering a unique service delivery architecture that supports tailored commercial cloud services and provides an advanced environment for the community to create transformational applications for all of R&E.

The Internet2 Network – *New Dimensions of Science Support*

The same community that played a seminal role in the creation of the modern Internet is providing disruptive innovations to accelerate breakthroughs and transform the way the world works, again. The platform is open.

In late 2012, the next-generation Internet2 Network was launched over a groundbreaking new transcontinental 17,000 miles, 8.8 Tbps footprint utilizing 100G technologies and offering the first open, national-scale production network with Software Defined Networking (SDN) and OpenFlow standards. The old Internet2 infrastructure was designed to provide advanced production services as well as a platform for the development of completely new, experimental ideas and protocols. The Internet2 Network is the only community-owned, nationwide network in support of science, research and education in the U.S.

Community control and operation of the fundamental networking infrastructure provides the necessary scalability for members to efficiently provision the resources and services unique to R&E

¹ Based on contributions from Ryan Bass (bassr@internet2.edu), Ann Doyle (adoyle@internet2.edu), Dale Finkelson (dmf@internet2.edu), Ana Hunsinger (ana@internet2.edu), [Kenneth Klingenstein](mailto:Kenneth.Klingenstein@internet2.edu) (kjk@internet2.edu), John Krienke (jcwk@internet2.edu), Linda Roos (lroos@internet2.edu), Christian Todorov (ctodorov@internet2.edu), Rob Vietzke (rvietzke@internet2.edu), Stephen Wolff (swolff@internet2.edu), and Dean Woodbeck (woodbeck@internet2.edu)

such as bandwidth-intensive requirements for collaborative applications, distributed research experiments, Cloud-based data analysis and social networking.

Nearly 90,000 U.S. research university members, numerous corporate and government research laboratories, and many other institutions in the research, education and clinical communities connect to the Internet2 Network. The nationwide backbone network interconnects regional and state networks, which provide access to these institutions. In addition, the Internet2 Network has a broad set of interconnections with U.S. government national research networks and over 100 non-U.S. National Research and Education networks (NRENs) around the world providing global connectivity to Internet2 Network users.

To continue to meet the needs of the R&E community, Internet2 is transitioning from OpenFlow as their underlying backbone technology and moving forward with a MPLS solution. This will provide the community with the stability of an industry standard technology while still enabling advanced services such as dynamic layer 2 provisioning.

Internet2 Advanced Network Services:

The Internet2 Network offers a full range of advanced services providing the most complete toolkit for building specialized scientific collaborations.

Whether needs are reliable, high-capacity, best-in-breed building blocks for private networks, integration with the global R&E fabric, or the ability to customize services and operations for the most demanding Big Data and science requirements, Internet2's portfolio of advanced networking solutions is designed to meet the needs of all R&E users.

Dependable IP solutions—engineered specifically for research and education

Internet2's Advanced [Layer 3 Service](#) delivers specialized R&E network service, just for our community. Internet2 members take advantage of:

- A network *dedicated* to specialized R&E traffic
- A network *engineered* to allow wide reachability across the R&E community with abundant "headroom" optimized for peak performance. Average use and traffic congestion are not part of the discussion when it comes to network planning for this network. The network is engineered for the most^a demanding peak-plus-potential network traffic, ensuring optimal performance. The goal of this service is for users to never experience the dropping packets, jitter or other underperformance characteristics of commodity networks.
- A network that's highly reliable, and connects to a fabric of other national research and education networks around the globe
- Internet2 members have commercial traffic needs, too. Internet2's [TR-CPS](#): Provides higher quality peering to commercial content providers such as Amazon, Apple and Google
- Introduces IPv6 and multicast capabilities into commercial peering arrangements
- Allows unused capacity on R&E pipes with/for high value peers

Internet2's Advanced Layer 2 Service (AL2S) provides Global Ethernet Network flexibility. An effective and efficient wide area Ethernet services that allows CIOs and IT staff to support long-term or short-term global collaborations for data-intensive science or production services.

- Abundant bandwidth—free of policy or capacity restrictions—to support scientists' and researchers' "big data" networking needs

- Interconnections with global R&E and Global Optical Lightpath Exchange fabric enable Ethernet VLANs throughout the U.S. on Internet2, and around the world through partner networks
- The option to enable software-defined networking (SDN) through technologies such as OpenFlow for network innovators through a SDN Overlay network.
- Maximum value of Internet2 Network bandwidth, and better management of network traffic for network operators

SDN Overlay network - This new capability allows users of Internet2 to deploy their own controllers on the Internet2 network, giving them full control of a virtual network and enabling the extension of policy control. With this capability, Internet2 demonstrated the ability to quickly deploy a Layer 2 service across backbone networks, regional networks, campuses, and exchange points, all without a loss of local autonomy. Internet2 also demonstrated the ability to deploy customer controllers on the Internet2 backbone, starting with the GENI project.

Internet2's Advanced Layer 1 Service allows users to control their own network—*without building it*. The service is the most specialized and cost-effective way to build a custom, high-capacity network. And not just any network, but a state-of-the-art network at 10, 100, 400 gigabit—and eventually *1 terabit* speeds—with more access points than any other national R&E network, including paths through regions never served before. Internet2's national fiber network, optical system and network operations center (NOC) provide a set of leading edge resources and capabilities that offers the most reliable, high-capacity network solution. See [Internet2 Network Fees](#) for participation fees.

Internet2 Custom Network Solutions provides the network you need—*where you need it*.

If your organization has a special need that requires a different solution, Internet2 has extensive expertise in working with key industry, government, research and educational institutions to develop custom network solutions that integrate the right blend of Internet2 Advanced Layer 1, Layer 2 and Layer 3 services.

Whether you need dark fiber procurement to reach a new site; coordination with regional partners for last-mile connectivity; identification of collocation possibilities; hardware deployment; or a dedicated NOC to support the most demanding, on-going service levels—Internet2 has the right experience to design, build, integrate and operate your custom private network.

Internet2 Community Innovation Platform for U.S. Research & Economic Development

Creating new innovation opportunities begins with understanding what enabled them in the past.

A few decades ago, research and education (R&E) community innovation helped to create the Internet, spurring unprecedented technology development, applied research and data-intensive science capabilities like never before. These innovations transformed the global economy into an information-based powerhouse—turning an initial \$250 million total investment into an estimated \$1.4 *trillion* global market annually,* making the modern Internet and its applications one of the most transformative technologies of the 20th century.

Those seminal network investments put the R&E community “way out in front” of commercial markets by establishing a new, bandwidth-rich, large-scale and ubiquitous set of capabilities—on top of which network technology, research innovations and science-driven applications could flourish. This, in turn, led to the global transformation on which our current information-based economy is built.

The Internet2 community (including the same players and partnerships that launched the original Internet) believes investing in a new **Innovation Platform** will produce a new wave of transformation. A new breed of applications can fuel a new cycle of global economic development driven by U.S. R&E. Universities and research organizations will be better able to attract and nurture tomorrow’s innovators. Their technical innovations and scientific discoveries will create new markets and provide new solutions that move beyond R&E and positively affect mankind. Once again, R&E can lead the way in defining a path forward and strengthen the nation's position and global economy for decades to come.

Internet2 Research Solutions

At a time when every major research project involves remote resources—human and otherwise—Internet2 is committed to helping researchers and research managers focus on the *research itself*, rather than the provision of infrastructure. Internet2 members have access to a unique, nationwide high-performance network infrastructure that removes the boundaries of today's Internet and provides an unsurpassed testbed for research.

The CE (Community Engagement) and NS (Network Services) divisions are revitalizing research partnership and engagement, creating an integrated network and services architecture for Internet2, and implementing a roadmap that will keep the Internet2 community at the focal point of technology innovation and development. Contact rs@internet2.edu to find out more.

Research Support Center

The Internet2 Research Support Center works with you hand-in-hand to effectively integrate the research-enabling infrastructure and services of the Internet2 community into your own applications and environments, right down to the engineering, planning and pricing.

Research Wave Program

Have a project that would benefit from the world’s most advanced network research testbed? Request access on the Internet2 Network for specific, fixed-time projects and grant proposals. We think allowing researchers to experiment on a short-term basis is a great way of simulating innovation and transforming research.

Research Funding Support

Internet2 actively supports member research by providing advocacy with funding agencies, consultation, letters of Collaboration, grant preparation assistance and more.

Internet2 Receives NSF Grant to Support Small Colleges and Universities for Nationwide Research & Education Platform and Cyberinfrastructure

Internet2 NET+ Initiative

Leveraging the Internet2 Network and enabling services like InCommon federated identity management, the Internet2 NET+ team is developing a portfolio of service offerings that bring value to Internet2 members. The goal: To create services that are cost-effective, easy to access, simple to administer, and tailored to the unique needs of our community.

By applying the same open principles and community synergies that guided the creation of the Internet2 Network, the Internet2 NET+ initiative is creating a unified, integrated portfolio of **cloud** and **trust** solutions, blending both commercial services and community offerings. Using the infrastructure facilities of the **Internet2 Network** and the federated authentication and authorization services available through **Internet2's InCommon***, leaders have a dynamic and tailored technology toolkit to address new needs and challenges head-on.

Through Internet2 NET+, members collectively identify and vet community and industry cloud solutions that are (or could be) effective in meeting campus challenges, and have the potential to scale to benefit all member institutions' teaching, learning and research needs. Through a series of increasingly rigorous tests and focused **service phases**, members closely evaluate the viability of each proposed solution. Each service has its own **lifecycle**, and service phases mark the progress of each from idea through availability, and even on to retirement, when necessary.

The following are pointers to the NET+ service types, where you can learn more.

Video, Voice & Collaboration

Infrastructure & Platform

Software as a Service

Digital Content for Research & Education

Security & Identity

The Internet2 Trust and Identity Infrastructure:

Internet2 provides the community-built and community-driven trust and identity infrastructure that supports faculty and staff, researchers and scholars, and access to services across the U.S. and globally.

The Internet2 identity and access management model centers on InCommon, the identity management federation that provides the policy and technical backbone for secure interactions, and allows single sign-on convenience for individuals. Under this model, colleges and universities manage the identity infrastructure and provide access to services—such as collaboration tools, business applications, course management solutions, and others—in a secure and privacy-preserving way.

Internet2 also now provides the TIER program (Trust and Identity in Education and Research). TIER is both an open-source toolset and a campus practice set. The TIER software includes a packaged suite of identity and access management components with ongoing development and a regular cadence of improvements and upgrades. In addition to providing software, the community collaborates on key practice sets needed to ensure interoperability, usability and cross-organizational trust and security.

Our goal: ensuring that members of our community have access to the right services, at the right time, with the right protections and privacy considerations, while supporting easy collaboration globally.

Trust and Identity in Education and Research (TIER)

The TIER software includes a packaged suite of identity and access management components and APIs with ongoing development and a regular cadence of improvements and upgrades. Built by and for the community, ongoing collaboration on key practice sets needed to ensure interoperability, usability and cross-organizational trust and security.

InCommon Federation

Internet2's InCommon operates the identity trust federation for U.S. research and education, allowing for a secure and privacy-preserving trust fabric to enable the sharing of protected resources, and offering users single sign-on convenience.

InCommon Certificate Service

The InCommon Certificate Service provides U.S. higher education with unlimited certificates for one fixed annual fee, including SSL, extended validation, client (personal), and code signing certificates.

InCommon Assurance Program

The InCommon Assurance Program certifies campuses and non-profit sponsored partners and research organizations that meet the requirements of the InCommon Bronze and Silver assurance profiles (which are comparable to the NIST Levels of Assurance 1 and 2). These practices determine the confidence in the accuracy of a user's electronic identity and help mitigate risk for the service provider.

InCommon Multifactor Authentication Program

The InCommon Multifactor Authentication Program provides affordable solutions for various methods of achieving the additional security offered through using additional factors of authentication.

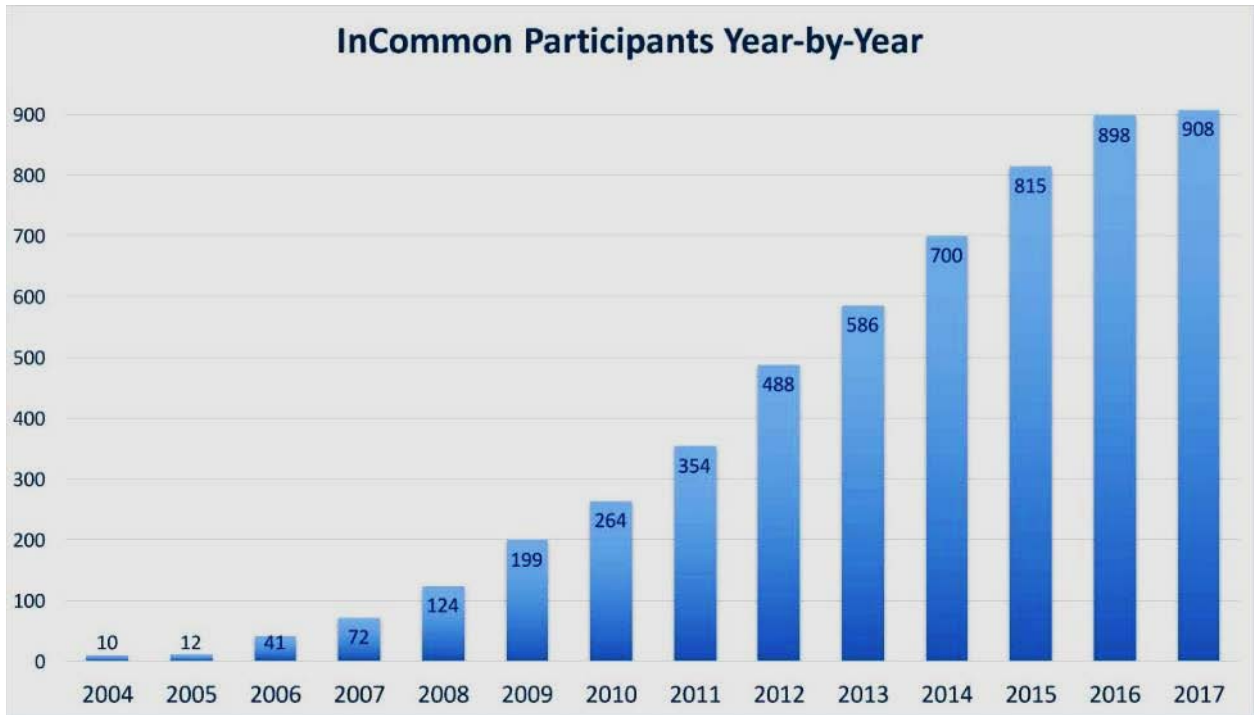


Figure 10: InCommon Growth

Shibboleth

An open-source project that provides single sign-on capabilities and allows sites to make informed authorization decisions for the individual access of protected online resources in a privacy-preserving manner.

Grouper

Handles groups and access management across applications and tracks information such as campus affiliations or roles.

COmanage

COmanage (Collaborative Organization Management) is a software platform that allows collaborative groups to streamline and manage the identity-oriented requirements of common collaboration tools.

eduPerson and eduOrg

eduPerson and eduOrg are LDAP schema designed to include widely-used person and organizational attributes in higher education.

eduroam

The eduroam service provides instant, authenticated and encrypted network access to the users of all participating institutions.

The Internet2 IP Network is built utilizing Juniper MX960 routers. The Internet2 IP Network supports IPv4, IPv6, and multicast protocols. As of January, of 2017, there were approximately 17,586 routes advertised for Research and Education IPv4 traffic, and 1,441 for IPv6.

The backbone is based on 100Gigabit Ethernet technology with 10 Gigabit Ethernet redundant paths for backup purposes.

Any Internet2 member organization may connect to the Internet2 IP Network through an Internet2 Connector. Members currently may also sponsor non-member organizations and statewide education networks – which typically connect primary, secondary schools, community colleges, libraries and museums - to use the Internet2 IP Network. The implementation of U.S. UCAN will offer connectivity to these community anchor institutions.

The Internet2 IP Network peers with the major U.S. national research networks, including DREN, ESnet4, NISN, NREN, and NWave. The Internet2 IP Network has direct peering with 24 non-US networks and via transit reaches over 88 National Research and Education Networks in Africa, Asia, Europe and Latin America and the Middle East and, in turn, provides transit between any non-U.S. networks.

Internet2 Advanced Layer 2 Ethernet Network:

Internet2's Advanced Layer 2 Service provides member organizations the ability to meet diverse requests from their most demanding users. Today, networks must enable different and distinct scientific big data exchanges across the country and globe, provide secure student and employee data transmissions into a cloud environment and support bandwidth-consuming multimedia—while supporting all user needs simultaneously. Internet2's Advanced Layer 2 Service now provides a scalable and flexible national network where members can build Layer 2 circuits between endpoints on the Internet2 Network and beyond to meet every user need using the OESS (Open Exchange Software Suite). OESS is a set of software used to configure and control dynamic (user-controlled) layer 2 virtual circuit (VLAN) networks on the Internet2 AL2S network. OESS provides sub-second circuit provisioning, automatic circuit failover, per-interface permissions, and automatic per-VLAN statistics. It includes a simple and user friendly web-based user interface as well as a web services API.

Internet2's Advanced Layer 2 Service is a cost-effective, highly reliable, advanced networking solution designed by and for the research and education community. The service offers:

- Efficient and effective network management, allowing CIOs and IT staff to support the varied needs of all their network users
- Dedicated bandwidth—free of policy or capacity restrictions—to support scientists' and researchers' big data networking needs
- Maximum value of Internet2 Network bandwidth, and better management of network traffic for network operators

Global Reach

The Internet2 Advanced Layer 2 Service will be operated as a distributed open exchange, allowing access not only to domestic endpoints, but also interconnectivity with global researchers through dozens of high-capacity intercontinental links to Europe, Asia, Africa, South America and other areas reachable through Internet2's global research and education network partnerships.

Connecting to Internet2 Advanced Layer 2 Service

Any organization can access Internet2's Advanced Layer 2 Service. Bandwidth can be delivered through an existing Internet2 connector port, or through a direct, dedicated Layer 2 port.

A dedicated Layer 2 port is most appropriate for organizations

- With significant bandwidth needs to support big data applications
- Who need full control of an open, policy-free network

Internet2 Advanced Layer 1 Network:

Internet2's new network provides over 8.8 terabits per second of dedicated capacity on a 16,000+ mile national footprint. North America's first nationwide 10G, 40G and 100G advanced wave network, it boasts a capacity of over 880 10 gigabit waves, providing researchers, campuses, private networks and corporate partners long-term, static, point-to-point and mesh wave capabilities to support their most demanding applications and research projects—at compelling cost-recovery fees.

Regional and provider-based partnerships extend domestic and global reach

Through partnerships with regional networks and telecommunications providers, Internet2 can act as a network integrator to create cost-effective access to wave capacity beyond the Internet2 Network in the United States and abroad. Internet2 will solicit the most cost-effective solution and contract with its partners to seamlessly provision circuits—optical carrier and digital signal—for any application that requires extending connectivity to any location on the globe.

Internet2 Wave Services provide global-scale aggregated rates—especially advantageous to demanding users that wish to extend their collaboration reach domestically and internationally. In addition to "lit" wave services, Internet2 can also develop dedicated fiber capacity to extend its backbone to sites requiring direct backbone extension.

What differentiates the Internet2 Network? Operational stability and operational transparency

Internet2 believes operational excellence comes through careful planning, communication with network users and continuous review and improvement of its practices. Internet2 has optimized its wave network for reliability and continuous service for demanding research, applications and private networks. The network is built for maximum uptime and employs the highest levels of communication and precision in its operational support procedures to assure best-in-class reliability.

Internet2 Network security: An inherent feature

Internet2 engineers understand the importance of network security to any national-scale infrastructure. Internet2 Network add/drop facilities are built in limited-access facilities with locked cages within secure facilities. Encryption services are available on individual waves and enhanced security protocols can be tailored to any network's needs.

Build your private network on Internet2's national infrastructure

Internet2's wave network is especially designed to support the build-out of multiple private networks for demanding research or domain science projects, advanced applications and government agency collaboration. In addition to Internet2's own research and education network services for the higher education community (which are built on top of Internet2 wave services), the Department of Energy's Energy Sciences Network (ESnet) and Internet2 have built optical waves to support connections to over 20 national research labs. The National Oceanic and Atmosphere Administration (NOAA) has also built its groundbreaking N-Wave national network on top of Internet2 wave infrastructure.

Internet2's network is ready to accommodate private networks of any scale throughout the United States—and through its partners, Internet2 is positioned to support private networks that extend the most advanced capabilities around the globe.

Technologies

The Internet2 community's significant upgrade of its national backbone network included the acquisition of nearly 16,000 miles of dark fiber. Built on Ciena's ActivFlex 6500 Packet-Optical Platform, the new 8.8 Terabit per second network is equipped with 100 Gbps optical backbone connections. The network provides services between nearly five dozen sites across the country. Level(3) provided the 8.8 terabit optical equipment complete with 24/7 monitoring and field support to these facilities. Level(3) also provides colocation facilities and power for the network.

Operations

Focused on the unique needs of the research and education community, the Internet2 Network provides transparent operations and is under constant evaluation and optimization by the community, to deliver leading edge network characteristics—such as adequate headroom—to meet the constantly evolving needs of high-speed research and collaboration.

Community input into the design and management of the network is done through the Network Architecture Operations and Policy program advisory group and the Network Technical Advisory Committee (NTAC).

Internet2 Network design, development, deployment, management and support are collaborations between the Internet2 Network Services staff and the [Internet2 Network Operations Center \(NOC\)](#). The Internet2 NOC is supported by the Indiana University [Global Network Operations Center \(GlobalNOC\)](#), which provides world-class production and research and development support services to Internet2 Network participants.

TransitRail-CPS:

Internet2 operates the TransitRail-Commercial Peering Service (TR-CPS), which provides strategic connectivity to portions of the public Internet and peering relationships through over 80 select entities. Internet2 members have commercial traffic needs, too. Internet2's commercial peering service offers a low-cost path with higher performance goals than commercial alternatives. TR-CPS provides high performance, low latency, and efficient (1 hop) access to some of the top content destinations in the world including: Google, Yahoo, NetFlix, and other commercial content providers. The service supports IPv4, IPv6 and multicast.

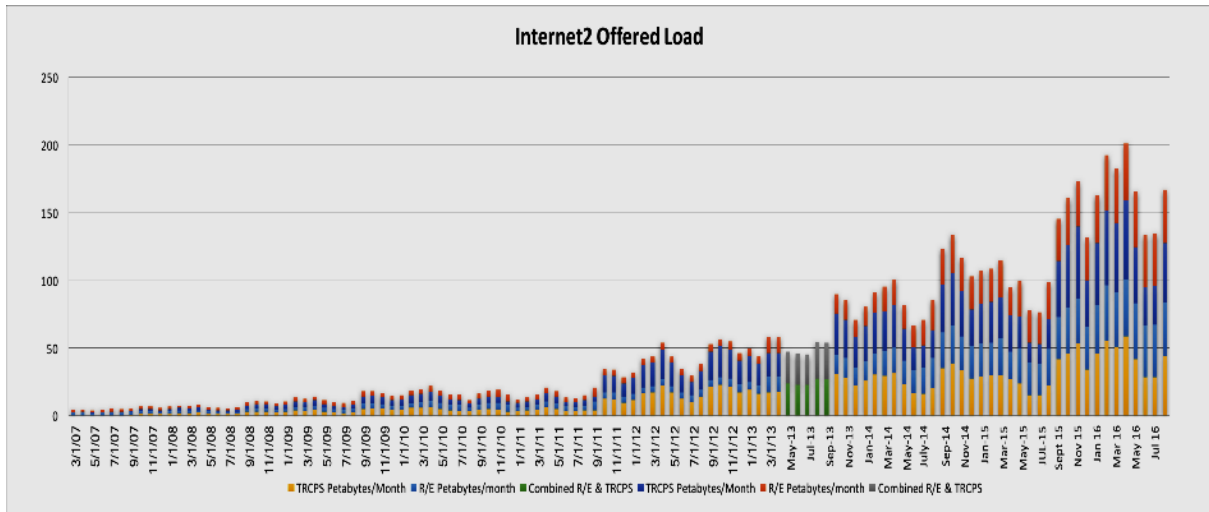


Figure 12: Internet2 Network Traffic

Internet2 Network Refresh

The Internet2 core network architecture is under active redevelopment and focused on the migration of AL2S and AL3S services to MPLS. While the goal is to preserve existing service capabilities, the 2017 architecture will also enable new ways to deliver services which may impact connectors.

Internet2 will build a uniform core network supporting AL2S and AL3S services that simplifies the capacity management on core network. We will improve connectivity to Internet Exchanges (IX/IXP) and remove the dependency on OpenFlow for AL2S and AL3S services. Internet2 will continue to provide advanced SDN capabilities through an overlay network connected through core.

The AL2S circuit service will remain in existence with OESS provisioning. The circuit service will be based on MPLS L2VPN and VPLS. However, path selection will be limited. The AL3S (TR-CPS, NET+) access to Layer 3 services will be expanded to all nodes. Conventional AL2S circuits providing Layer 3 services will be migrated to local nodes and OESS will no longer provision connectivity to Layer 3 exchanges.

The core Network will continue to provide persistent circuits to meet the needs of of the R&E community. Layer 2 and Layer3 access to IX / IXP, Cloud Services, and research collaborations (LHC et al) continue to be a priority.

Internet2 SDN Overlay network

The Overlay service contains 8 sites nationally. At each site, a Corsa switch is paired with a well provisioned server to provide virtualized data and control plane resources. These sites are interconnected over the MPLS based AL2S network in a partial mesh. Additionally, AL2S will be used to provide remote connectivity for those wishing to participate in a particular experiment or pilot.

While the Overlay Service is a supported and production service, its service expectations are not the same as that of AL2S. In particular, we fully expect from time to time controllers will induce or otherwise experience forwarding impairment or control plane anomalies. This service is designed to all participants to take engineering risk even if it impacts availability. The service will

be considered production with 24/7 monitoring and established operational procedures, but the process to operate a new slice on this service is simpler than in the prior iteration.

The SDN substrate will contain virtualized hardware and software resources. Initially this service will support OpenFlow 1.3 however any software based forwarding approach such as NFV or NDN can also be supported. Hardware components are interconnected with 10ge and sites are interconnected in a static partial mesh topology. Slices of the overlay network are provided by hardware virtualization.

Each slice that runs on this service will get its own set of virtual machines, data plane can control plane resources, and one gateway virtual machine. All resources on the control plane will be on a non-routed private network, with proxy access via the gateway VM. The data plane initially will be controlled using OpenFlow 1.3 with a single table pipeline. Within the core of the Overlay network, each slice has access to the full VLAN tag space from 0 to 4096. Partial tag availability is expected with each slice that has a layer2 interconnect with a campus or regional participant.

U.S. UCAN Initiative

[The United States Unified Community Anchor Network \(U.S. UCAN\)](#) is an Internet2 program working with regional research and education networks across the country to connect community anchor institutions—including schools, libraries, health care facilities and other public institutions—to advanced broadband capabilities.

Utilizing the Internet2 Network and in collaboration with regional research and education networks across the country, U.S. UCAN will enable anchor institutions to serve their communities with telemedicine, distance learning and other life-changing applications.

As of the summer of 2013, Internet2's [Sponsored Educational Group Participant \(SEGP\)](#) program is now formally U.S. UCAN. Internet2's Health and Life Sciences and K20 initiatives are also a part of U.S. UCAN. For more information on the K20 Initiative, visit its [social networking and collaboration website, Muse](#).

Network Performance Measurement and Monitoring

The perfSONAR development effort is a partnership between: Internet2, ESnet, Indiana University, and GÉANT to develop common standards for performance measurement frameworks and interoperable implementations of such frameworks. perfSONAR is an infrastructure for network performance monitoring, making it easier to solve end-to-end performance problems on paths crossing several networks. It contains a set of services delivering performance measurements in a federated environment. These services act as an intermediate layer, between the performance measurement tools and the diagnostic or visualization applications. This layer is aimed at making and exchanging performance measurements between networks, using well-defined protocols. At the present-time, multiple perfSONAR services exist, packaged into software bundles and deployed across major backbone networks such as ESnet, the GÉANT network, and the Internet2 Network. Many national and regional networks in and outside the United States, campus networks, and virtual organizations have also deployed this solution.

Outreach efforts have focused on delivering products capable of providing both diagnostics and regular monitoring for large international scientific efforts including the Large Hadron Collider (LHC) project. The LHC project has several ongoing experimental components, ATLAS and CMS being two of the largest, containing researchers working globally and utilizing R&E networks to move massive amounts of experimental data.

Internet2 International Activities:

Through partnerships with the world’s major national and multi-national research and education networks, Internet2 Network users have access to research, university, school, hospital and other institutions in 96 countries.

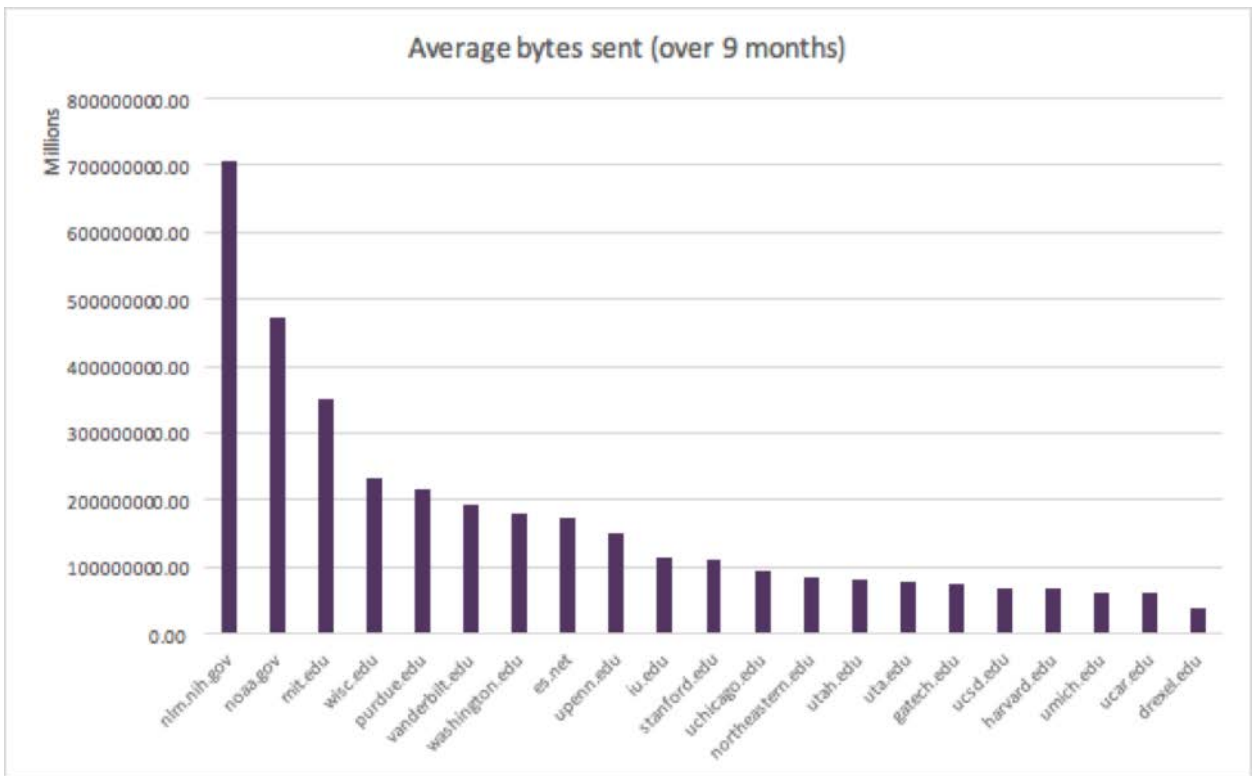


Figure 4 – International traffic 2016

ANA-300G

The Internet2 community is leading the dissemination of new technologies and applications on a global scale to advance R&E’s heritage of benefitting the global economy, and greater mankind. By empowering the most advanced research from universities and institutions worldwide, the growing resource demands of specialized, international “team-research” projects are being met and important new breakthroughs are being accelerated to make all our lives healthier, safer, and better.

Laying the groundwork for global scientists to collaborate in new ways, Internet2, in coordination with six of the world's leading R&E networks and two commercial partners, deployed the world’s first intercontinental 100Gbps transatlantic link between North America and Europe in 2013. The 100 Gbps link, called the Advanced North Atlantic 100G Pilot project (ANA-100G) helped determine the operational requirements needed to effectively run 100 Gbps wavelengths between

North America and Europe, modeling the same owned infrastructure approaches used terrestrially in Europe and North America to create bandwidth abundance and encourage utilization. The project also aimed to advance two significant trends in research: science is increasingly data driven, with datasets from large-scale experiments mounting into the tera-scale level; and increasing dependence on international collaborations, with researchers around the globe expecting immediate access to the datasets—and colleagues. Since 2013, the ANA-100G project was expanded to include 2 more 100G circuits along with terrestrial redundancy in the US and Europe. This moves the project into a production phase.

The high-energy physics experiments based at the ITER fusion reactor in France and the Large Hadron Collider (LHC) in Switzerland—often called “the world’s largest science experiment”—are two key examples of this trend, known as The Fourth Paradigm. ITER is a large-scale global team-science experiment aiming to produce commercial energy from fusion and provide us all with a cleaner, safer, and unlimited source of energy. And, through high-energy experiments with thousands of participating physicists around the world, the LHC is significantly advanced our understanding of the composition of the universe—producing the Nobel Prize-winning discovery of the Higgs Boson particle in 2012.

The ANA-100G link with subsequent expansion to 300G has already demonstrated ultra-high speed and reliability in exchanging scientific big data. Transfers between collaborators throughout Europe and their peers in the United States and Canada now take only a few minutes as opposed to many hours over the public Internet. In addition, the operation of this first 100 Gbps link across the Atlantic Ocean provides the cornerstone for the global deployment of cutting-edge networking technologies, such as software-defined networking, enabling the development of new applications that will, in turn, encourage economic growth. Advancements of this scale illustrate the Internet2 community’s global leadership—providing entirely new capabilities for international science and discovery and making impacts far beyond academia.

Internet2 Network International Connectivity:

Outside the US, connectivity for the Internet2 Network is provided by Internet2’s international partners, by the National Science Foundation’s (NSF) International Research Network Connections (IRNC) program.

NEAAR

- PI: Jennifer Schopf
- Partners: GÉANT, ASREN, WACREN, and TENET
- The Networks for European, American, and African Research (NEAAR) collaboration is a powerful, cross organizational project that will provide services and bandwidth connecting researchers in the US with their counterparts in Europe and Africa. This project will have an immediate impact on the research environment and supports future application and technology advances.

TransPAC4

- PI: Jennifer Schopf
- Partners: APAN, TEIN3, NICT-Japan, NII- Japan, CERNET – China, DLT, and several other institutions.
- 10Gbps connection possibly moving to 100Gbps from the U.S. to Asia provided through this NSF funding. Connections continuing to South and South East Asia. Partnership with TEIN3 network provides connections to Europe, resulting in a global network (TP3-

TEIN3-GÉANT-ACE). Connection to the TAJ network provides second link to Asia, Europe and North Africa

Backbone: AmLight Express and Protect (ExP)

- PI: Julio Ibarra
- Partners: Association for Universities for Research in Astronomy (AURA, USA), Cooperación Latino Americana de Redes Avanzadas (RedCLARA: Latin America), Internet2 (USA), Red Universitaria Nacional (REUNA:Chile), Academic Network at São Paulo (ANSP: São Paulo, Brazil), Rede Nacional de Ensino e Pesquisa (RNP: Brazil), Canadian Advanced Research and Education Network (CANARIE, Canada), the Florida Lambda Rail (FLR, USA) and Florida International University (AtlanticWave and AMPATH, USA).
- In response to the network requirements of these U.S.-Latin America collaborative science research communities, the AmLight Express and Protect (ExP) implements a hybrid network strategy that combines optical spectrum (Express) and leased capacity (Protect) that builds a reliable, leading-edge diverse network infrastructure for research and education.

SXTRANSPORT Pacific Islands Research and Education Network

- PI: Dave Lassner
- Partners: University of Hawaii (UHNet), AARNET, REANNZ, CENIC, and Pacific Wave
- The project has two main thrusts. First, it leverages prior NSF investments and mature international partnerships to maintain support for the current resilient production 2 x 40Gbps submarine fiber connections from Australia (AARNet) and New Zealand (REANNZ) to the U.S. and, in 2016, upgrades these connections to 2 x 100Gbps.

Internet2 China Program

The goal of this program is to jointly promote interoperability, seamless networking and collaboration, and enhanced interaction between U.S. and Chinese CIOs and researchers regarding critical and timely network research and “clean slate” efforts. This is accomplished through multi-year collaboration, building upon successful CANS (Chinese-American Network Symposium) momentum and enhancing NSF’s larger efforts in promoting global networking and ensuring that these developments are available and robust enough to enhance and ensure increased scientific collaboration, continued deployment of the US satellite campuses in China, and international exchange of ideas across many disciplines.

The Chinese-American Networking Symposium (CANS) is an annual event that focuses on ways to enhance and increase collaborative opportunities for Internet2 and its members working with Chinese universities and researchers, building upon the recent and long-term interaction and engagement between Internet2, CERNET, CSTNET, and the Chinese Academy of Sciences that has led to enhanced relationships, coordination, information-sharing, and improved services. The most recent CANS event was hosted by Rice University in Houston, Texas, from Oct. 17-19, 2016.

Annex 3: ESnet Overview and Status

Submitted by Brian Tierney, bltierney@es.net, Inder Monga, imonga@es.net, Jon Dugan, jdugan@es.net, Patty Giuntoli, [pmgiuntoli@es.net](mailto:pmgiantoli@es.net), Chin Guok, chin@es.net, Lauren Rotman, Lauren@es.net, Patrick Dorn, dorn@es.net

January 2017

Summary

The Energy Sciences Network (ESnet) provides a high-performance network that spans the US and key European locations. The network is built and optimized in support of national and international science data transport. Funded by the U.S. Department of Energy's Office of Science (DOE SC) and managed by Lawrence Berkeley National Laboratory, ESnet's mission is to accelerate scientific discovery. All of its capabilities are engineered for that purpose.

ESnet deployed its fifth generation network infrastructure in 2011/2012, called ESnet5. This network has expanded in several ways in 2016 to continue to meet evolving DOE SC mission requirements. In 2016, ESnet added 100G more to its core backbone capacity, as well as expanded from 10G to 100G at key peering points. Additionally, ESnet established new peering capability at US (Seattle Internet Exchange, SIX) and European (Amsterdam Internet Exchange, AMSIX). ESnet also continued to grow support for DOE supported research at U.S. universities, including carrying LHC data between specific U.S. Tier-2 universities, and from U.S. Tier-2 universities to European collaborators. ESnet6, the next generation of ESnet, is well into its research and development phase; several architectures are being explored to meet expected data transport growth anticipated over the next 10-12 years.

DOE SC is one of the largest supporters of basic research in the physical sciences in the U.S. It directly supports the research of roughly 27,000 scientists, postdocs and graduate students, and operates major scientific facilities at DOE laboratories that interact with national and international research and education (R&E) communities, as well as other federal agencies and industry.

ESnet provides direct connections to more than 50 SC research sites and their users, including the entire National Laboratory system, its supercomputing facilities, its major scientific instruments, and specific U.S. universities. ESnet also connects to more than 200 research and commercial networks, permitting tens of thousands of DOE-funded scientists around the world to collaborate productively.

As of January 2017, ESnet transported approximately 64 petabytes (PB) per month— nearly doubling in volume compared to the 38PB transported in January 2016. This number has grown roughly 72% each year for over 20 years [see Figure 1]. ESnet's architecture consists of 100 Gbps backbone links with the capability of scaling to 44 x 100 Gbps on the ESnet5 optical network. In 2015, ESnet's backbone extended across the Atlantic into Europe via 3 x 100 Gbps and a 40 Gbps leased links, to serve the networking requirements for the U.S.-based high-energy physics (HEP) community.

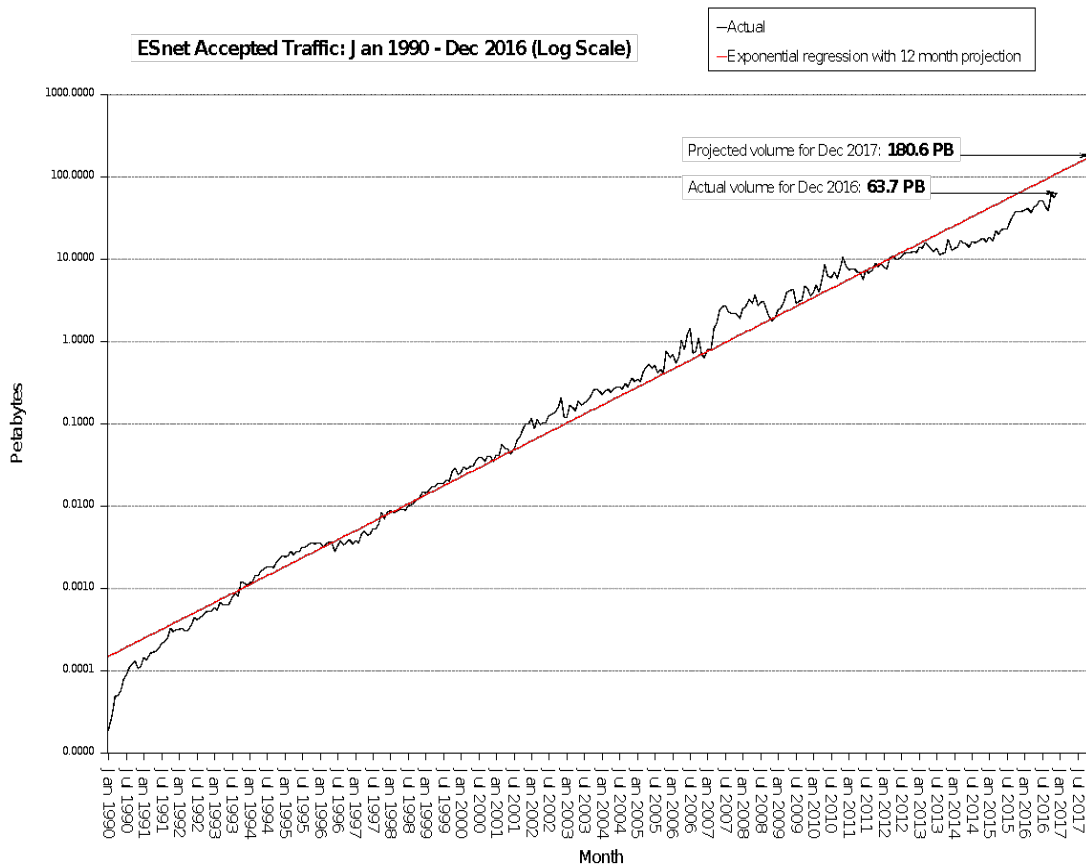


Figure 13: A graph highlighting ESnet’s traffic growth since 1990 through December 2016.

Current ESnet Architecture

The ESnet network has two distinct network layers. The lower layer is an optical network that provides point-to-point circuits between endpoints. The upper layer is a routed network, which is built on top of the point-to-point optical circuits. The routed network implementation is consistent across the backbone while the implementation of the optical layer varies across three geographic regions: the United States, Europe, and the transatlantic regions.

Optical Layer

Most of the U.S. optical layer is furnished via an ESnet-Internet2 collaboration. Through this agreement, Internet2 and ESnet have partnered with Ciena to deploy its ActivFlex 6500 Packet-Optical Platform equipped with WaveLogic™ coherent optical processors on a U.S.-wide backbone of dark fiber. Internet2 and ESnet equally share 8.8 Terabits of total optical capacity on this platform.

The optical layer supporting the European ring is the result of an ESnet-GEANT collaboration. Through this agreement, the GEANT Association is providing ESnet with 100Gbps circuits on their Infinera-based optical network.

The transatlantic portion of the optical layer consists of circuits purchased from multiple telecommunications carriers on diverse sub-sea cables that provide a robust set of circuits with diverse optical systems, right-of-way, and business relationships.

Routed Layer

The 100Gbps-routed layer is built on top of the optical network. It uses Alcatel Lucent 7750-SR12 Service routers at all 100G hubs. Juniper MX, and Juniper M series routers are used for 10 and 1G connections.

The routed layer supports general IP transit to labs, multiple virtual overlay networks, and point-to-point virtual circuits. The virtual circuits are provided via the OSCARS dynamic circuit provisioning system, which supports the ability to schedule and reserve guaranteed bandwidth virtual circuits, based on Multiprotocol Label Switch (MPLS) technology.

Network Topology

The routed network is constructed using point-to-point optical circuits interconnected in a series of rings to provide a cost-effective and reliable core. This architecture continues to be successful in providing a core network availability of greater than 99.99% and a site availability of 99.9% or greater. Most of the large science labs that are connected via metropolitan area network (MAN) rings experienced 100% availability in 2015 and 2016. As each site increases its dependency on the network including cloud services, redundant connections are being added to the end-sites, improving service reliability.

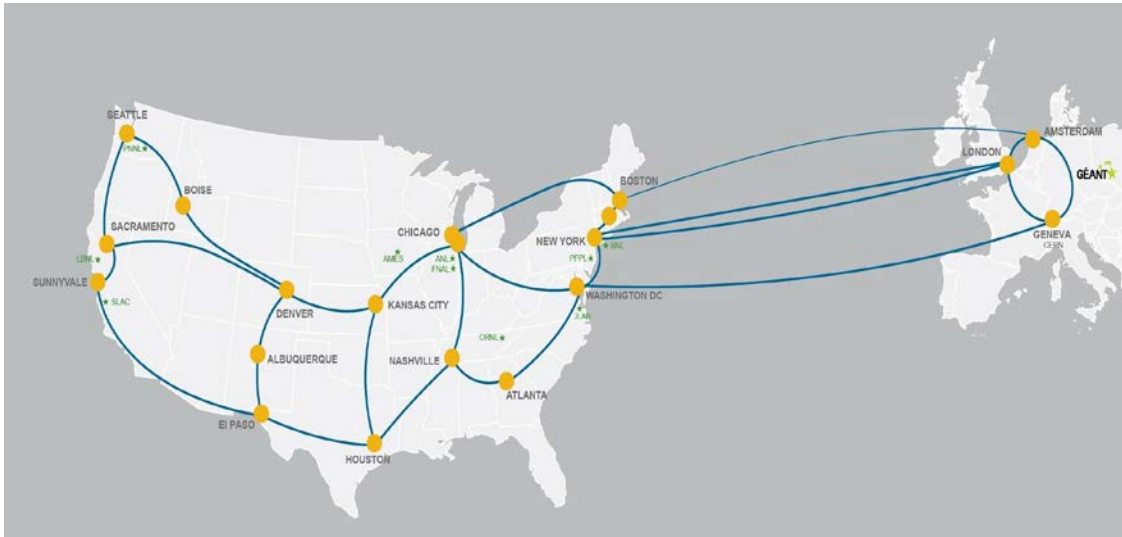


Figure 14: Current ESnet network topology

Beginning in 2015 ESnet expanded our support of the DOE SC HEP mission by providing LHCONE service to selected LHC computing centers at US Universities at DOE SC HEP's request. ESnet's LHCONE service supports transferring LHC data between participants, and to other LHC centers around the world. US-ATLAS and US-CMS have appointed Experiment Site Coordinators to oversee and manage the ESnet LHCONE Service to university participants. Information about the LHCONE service to universities can be found on the MyESnet Portal under LHCONE collaborations²

² <https://my.es.net/collaborations/lhcone>

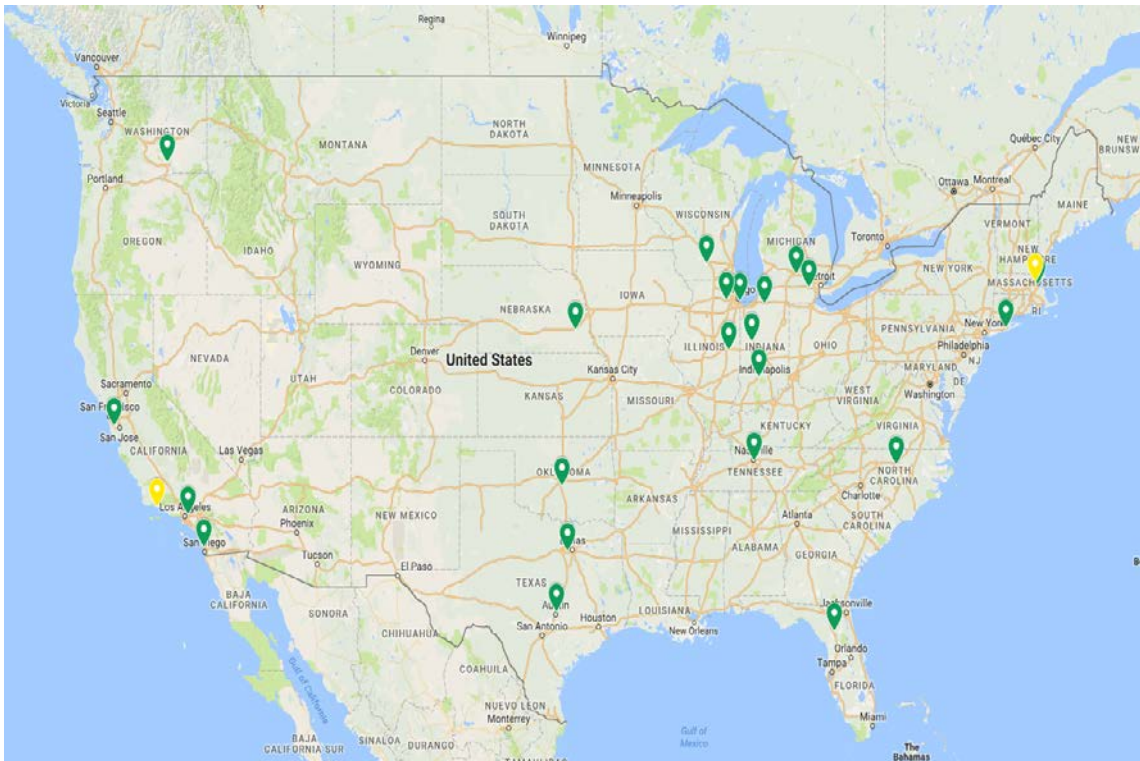


Figure 15: Map of U.S. universities connecting to ESnet’s LHCONE Service. Green locations mark the universities who are using service. Yellow locations are universities who are in progress of connecting to the service.

Last Mile

The strategy for connecting the DOE science labs reliably is to build a metropolitan or regional ring architecture that effectively places the labs directly on the ESnet core network. The strategy for connecting to other Research and Education networks is to co-locate ESnet core nodes at Open Exchange points.

ESnet Capabilities

In all its activities, ESnet strives to be at the leading edge of scientific networking practice, and to develop and champion solutions that advance the state of the art in distributed science. Leveraging its requirements workshops, collaboration with ESnet site network coordinators and its peers around the world, ESnet has developed a portfolio of production capabilities to meet its users current and near future needs.

Maintaining and offering a consistent set of capabilities for science collaborations delivered by independent but cooperating organizations worldwide, is a critical feature of ESnet’s research ecosystem and for which all of its infrastructure and offerings are engineered. ESnet offerings include:

- *Network Services*
 - IP circuits
 - Virtual circuits
 - User oriented overlay networks
 - Performance Measurement and Monitoring Tools
- *Science Engagement Services*
 - A dedicated Science Engagement Team to provide scientists and engineers with

- consulting support in data transfer, network architecture, and performance measurement
- Visualization tools
- *Advanced Network Testbed and Research*
 - Support for industry-leading research activities and demonstrations leveraging our collaboration with researchers and network testbeds that we operate as a service to the larger network research community

Network Services

IP connectivity

ESnet provides routed IP connectivity to its sites for several different purposes. ESnet's routed offering is engineered to provide the highest level of support for demanding science applications, in particular high-performance TCP-based data transfers that are intolerant of packet loss. In addition to high-performance science connectivity, the ESnet routed IP offering provides commodity Internet connectivity through a mixture of peering and commercial transit offerings.

Virtual Circuit Connectivity – OSCARS

Traditional shared IP networks are not always able to handle the large and sustained bursts of data some experiments produce without disrupting other traffic on the network and they cannot assure the deterministic quality that is sometimes required for experiments involving remote instrumentation. To this end, ESnet offers a multi-domain virtual circuit capability through its On-demand Secure Circuits and Advance Reservation System (OSCARS).

This open source, software application allows users to create and reserve virtual circuits on demand or in advance. Using the Network Services Interface (NSI), API, OSCARS gives users and ESnet engineers the ability to engineer, manage and automate virtual circuits over the network based on the specific needs of their work with scientific instruments, computation, and collaborations. The automation of this complex process has reduced circuit setup time to minutes; previously it had often taken weeks, especially in the inter-domain case. Today, OSCARS circuits carry about 20% of ESnet's annual traffic (figure 4).

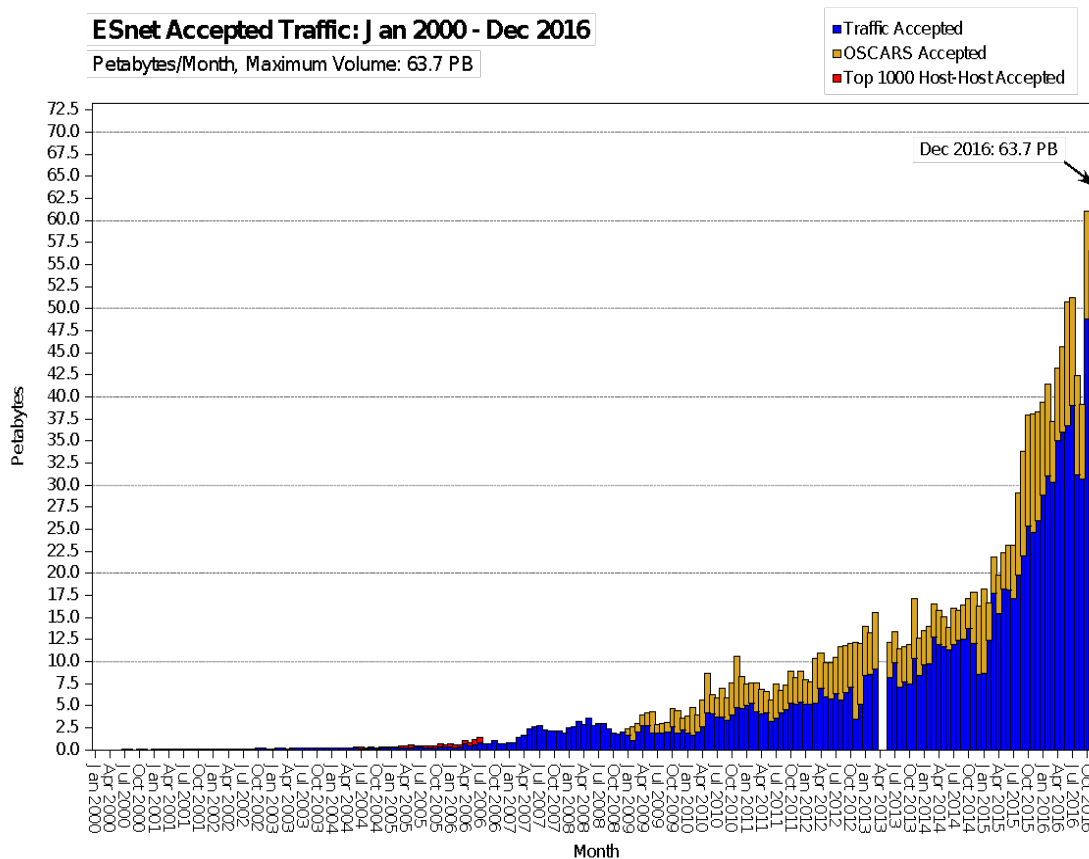


Figure 16: A depiction of the proportion of OSCARS traffic on ESnet since 2009.

OSCARS was awarded the prestigious R&D magazine R&D100 award in 2013.³

ESnet staff co-chaired the effort to standardize the Network Services Interface (NSI) in the Open Grid Forum (OGF). NSI is a protocol through which applications and middleware components can request dedicated circuits across multiple domains. Eventually it will have the ability to exchange performance information, configure overlays and obtain other services.

In 2016, the OSCARS software development team initiated an extensive re-engineering effort to improve service capabilities including adding multi-point connections, automatic re-routing, and modernizing the user interface. The effort also addresses potential security issues, and improves the build / deployment cycle. The resulting new version of the software, called v1.0, is undergoing final testing and will be deployed in early 2017.

User-Oriented Overlay Networks

ESnet supports virtual private networks for multiple collaborations. The virtual private networks provide traffic isolation, separation and in some instances guaranteed capacity. Some of the virtual private networks are just across ESnet, while others are provided in close collaboration with other research and engineering networks to offer VPN services to global science collaborations (such as LHCONE).

³ <http://www.rdmag.com/award-winners/2013/08/allowing-research-anywhere>

Performance Measurement and Monitoring Tools

Several R&E network partners including ESnet, Internet2, GEANT and RNP (and others) have developed perfSONAR, a platform and protocol suite for end-to-end network performance measurement and monitoring. Using perfSONAR nodes and services, engineers can quickly identify when network problems are affecting performance, isolating problems to a single domain so that corrections can be made quickly. ESnet has equipped its own network with perfSONAR infrastructure, and works with its sites and users to deploy additional nodes. perfSONAR has been widely successful and influential in the R&E network community—more than 2000 registered perfSONAR hosts have been deployed on 400 networks and in more than 50 countries.

The screenshot shows the perfSONAR Lookup Service Directory interface. It features a search bar at the top left, a browser section with a tree view of services (e.g., BWCTL Server, OWAMP Server, NDT Server), a service information table, a host information table, and a world map showing the global distribution of nodes. The service information table has columns for Service Name, Addresses, Geographic Location, Communities, Version, and Custom. The host information table has columns for Host Name, Hardware, System Info, Toolkit Version, and Communities. The world map shows a dense distribution of red dots representing nodes across all major continents.

Figure 17: perfSONAR Lookup service interface that allows users to search for public test nodes. <http://stats.es.net/ServicesDirectory/>.

In 2016 the perfSONAR team was focused on a new release (v4.0). The first release candidate for v4.0 was made public in November 2016, and the final release is planned for February 2017. Some highlighted features for the 4.0 release includes:

- A new scheduling infrastructure has been introduced in perfSONAR version 4.0 called pScheduler. pScheduler is a complete replacement for BWCTL with a number of new features requested by the community over the years.
- This release contains completely new graphs to view the measurement results, which are based on the ESnet, react charting libraries. The new graphs now stack different metrics for easier comparison without overloading axes. They also include a number of performance improvements over previous iterations of the graphs
- perfSONAR now officially supports CentOS 7 and Debian 8
- An all-new GUI for developing mesh configurations has been developed which doesn't require hand editing of the configuration file.

More information on the perfSONAR and perfSONAR releases can be found at <http://www.perfsonar.net/>.

Science Engagement Services

Dedicated Science Engagement Team

The ESnet Science Engagement Team provides consulting support in data transfer, network architecture, and performance measurement. In addition, ESnet staff work directly with scientists and site network and system administrators to help resolve a variety of network problems and help them proactively develop solutions for current and future applications. ESnet provides consulting ranging from time-critical troubleshooting performance problems to architectural and technology recommendations for network upgrades and performance enhancements. In 2016, the Science Engagement Team comprised four members and will look to expand engagement services to a broader number of domains and collaborations. This will be accomplished through various community partnerships within DOE program management, other NRENs, ESnet site coordinators and facility user services teams.

To forge strong ties to community and users, ESnet invests considerable energy in education, outreach, and advocacy. Examples of workshops that focus on education and advocacy include the Operating Innovative Networks workshop series (OIN, www.oinworkshop.com), the CrossConnects Workshop Series⁴, and the DOE Office of Science and ESnet Requirements Reviews⁵.

In a typical year, ESnet staff participate in more than 100 community and stakeholder meetings and give 60-90 technical presentations including tutorials, workshops, and information sessions to help DOE scientists and IT professionals effectively use network capabilities to accelerate discovery.

ESnet has made critical contributions in the development of a strategy for network architecture called the Science DMZ that enable scientists and IT staff to resolve bulk data transport issues at the local campus level. This architecture has become increasingly popular due to the improved end-to-end performance that results from its implementation and ESnet has partnered with Indiana University and Internet2 to deliver educational outreach in the form of the Operating Innovative Networks (OIN - www.oinworkshop.com) workshop series. After being self-funded for over twelve months, this workshop series was extended through a grant from the National Science Foundation. Over the last 6 years, the National Science Foundation (NSF) funded around 120 projects to implement some variant of the Science DMZ.

Requirements Analysis Process

ESnet's architecture and capabilities are tailored for its users, whose needs are derived through a formal requirements gathering process. Periodic workshops focus on each of the six science communities funded by SC. These workshops analyze and discuss scientific case studies rather than technologies. With this method, ESnet derives the science communities networking requirements by understanding how specific disciplines produce knowledge. This successful approach has served as a model for other research and education (R&E) organizations.

While very successful, the program has remained largely unchanged for ten years while the science ESnet serves has revolutionized over that time. Over the past year, ESnet has received feedback from some review participants that highlight potential weaknesses in the current process. New methodologies for analyzing qualitative and quantitative requirements has also greatly improved through innovations by industry.

⁴ <https://www.es.net/science-engagement/programs-and-workshops/crossconnects-workshop-series/>

⁵ <http://es.net/science-engagement/science-requirements-reviews/requirements-review-reports/>

In Q3 2016, ESnet Science Engagement embarked on a project to evaluate its current requirements review process in partnership with an external consultant to help collect and analyze opinions from multiple stakeholders, through professionally designed data collection mechanisms built for the targeted Review participants, ESnet is looking to strengthen the program. Once the evaluation is complete, ESnet will adjust its program and roll out its revised review process by end of Q2 2017.

- For a summary of ESnet requirements workshops, visit:
<http://www.es.net/about/science-requirements/>
- For a summary of ESnet presentations, visit:
<http://www.es.net/news-and-publications/publications-and-presentations/>

Visualization Tools: MyESnet Portal

While perfSONAR has helped users make significant progress in debugging network issues, DOE researchers and IT staff have requested better visualization and support for contextualizing wide-area network usage and patterns. In 2011, ESnet launched MyESnet (my.es.net), a portal that consolidates multiple real-time network visualization tools in a single, well-designed interface for science communities as well as IT experts. A recent development of the portal includes a real-time visualization of utilization on the backbone network (Figure 6).

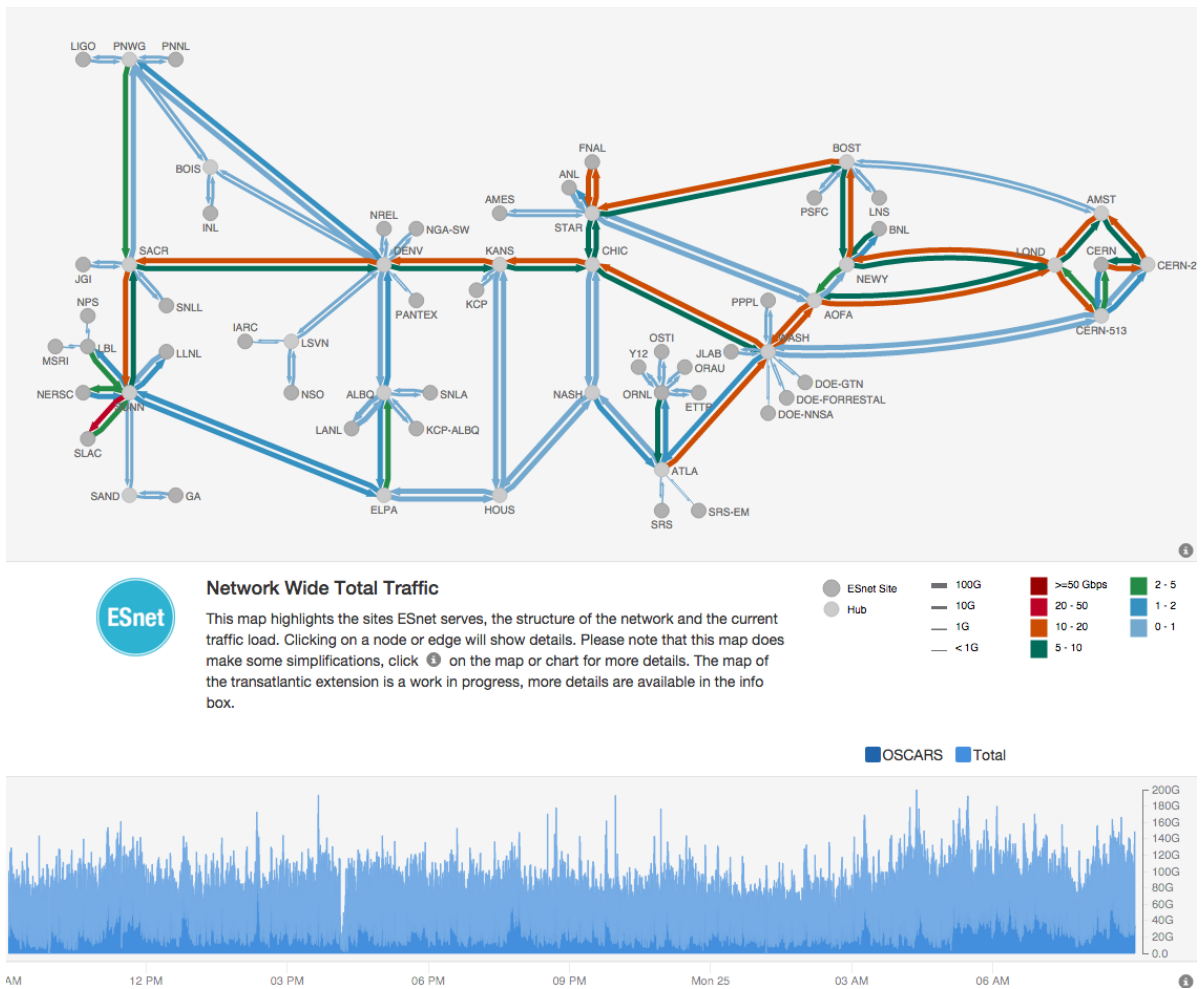


Figure 18: Interactive map showing real-time network utilization on the ESnet backbone and connected sites.

The map presents a simplified, high-level view of the network by combining and simplifying some aspects. One simplification is that connectivity for each site is shown as a single pair of lines that show the traffic traversing all connections to the site. This simplification means that this map does not show the fault tolerance, which has been carefully engineered to provide the sites with highly available connectivity. The second simplification is that for any pair of nodes shown there is only a single pair of lines representing all of the connections between that pair of nodes. This map is the first of several envisioned by the team, which plans to create additional maps that will focus on highlighting specific aspects of the network.

MyESnet provides customized dashboards for each connected site (Figure 7), including visualizations for network utilization, flow analysis, and relevant network outages. my.es.net includes integration with perfSONAR and OSCARS and it will ultimately be able to display information about the science impact of major network flows.

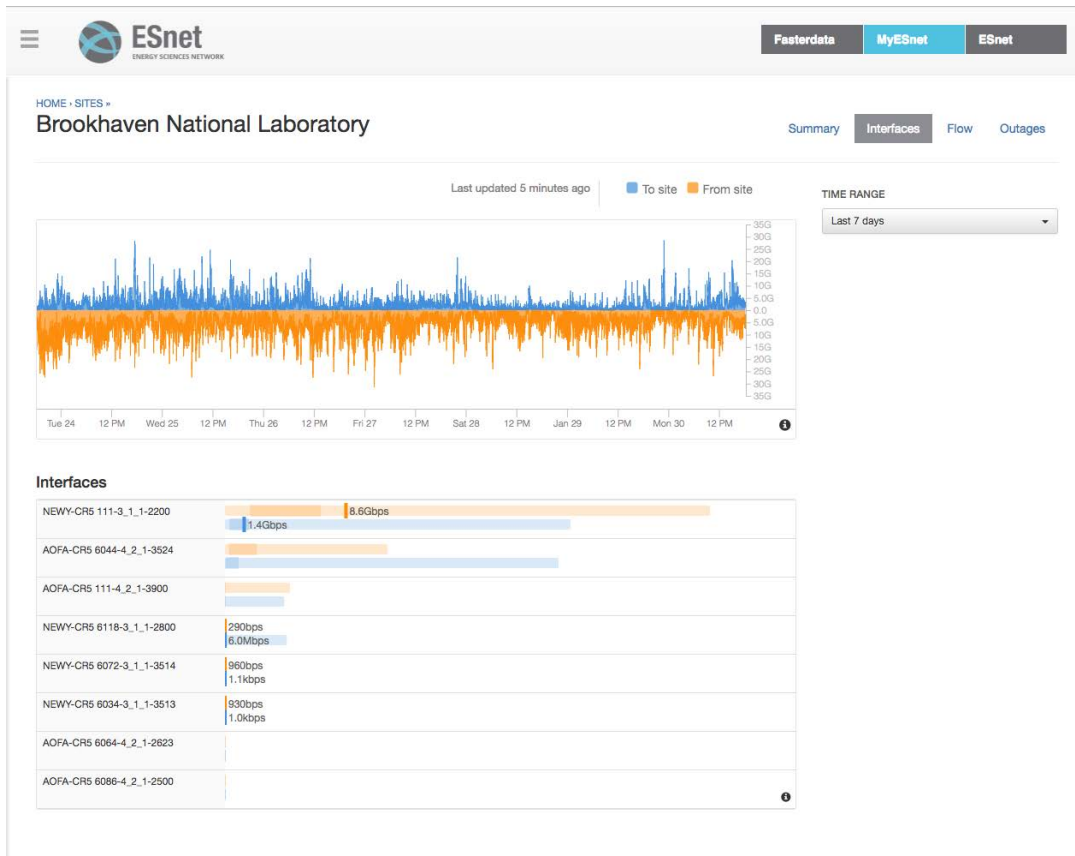


Figure 19: Example of site information available via MyESnet showing traffic over the last 7 days.

The MyESnet portal has been popular across the R&E community and as a result a number of organizations have asked us to provide access to the code as open source. We have always been open to this idea but have had concerns about the usability of the bulk of the portal code in other contexts. To address this issue we decided that rather than release the portal code itself we would work to factor out the most reusable components into libraries that could be reused. This resulted in the release of three open source packages that implement the reusable portions of the portal:

- Time Series Charts (<http://software.es.net/react-timeseries-charts/>)
- Network Diagrams (<http://software.es.net/react-network-diagrams/>)
- Pond, a time series data abstraction and manipulation library (<http://software.es.net/pond/>)

The Time Series Charts library contains a set of modular charting components used for building flexible interactive charts. It was built for React.js from the ground up, specifically to visualize time series data and network traffic data in particular. Low-level elements are constructed using d3.js, while higher-level composability is provided by React. Charts can be stacked as rows, overlaid on top of each other, or any combination, all in a highly declarative manner. The Network Diagrams library contains an initial set of React circuit drawing and network mapping components which are used within the ESnet Portal, but are not tied to ESnet, or even to network visualization. Pond provides time-based data structures and processing within ESnet tools. For data structures, it unifies the use of time ranges, events and time series. For processing, it provides a chained operations interface to aggregate and collect streams of events. We are still developing Pond as it integrates further into our code, so it may change or be incomplete in parts.

Advanced Network Testbed and Research

Industry-leading Research Activities and Demonstrations

ESnet's research and development activities are integral to our mission of advancing the capabilities of today's networking technologies to better serve the science requirements of the future. We continually investigate and test the services, protocols, routing techniques, and tools necessary to meet the expanding needs of our user community of DOE scientists. Since 1985, in addition to running an excellent production Internet backbone that links DOE researchers and user facilities, ESnet has been engaged in developing and rolling out technology solutions and services such as OSCARS and perfSONAR. ESnet has established many ongoing collaborations with other research and education networks and actively contributes to forums or collaborations including:

- Open Grid Forum ([OGF](#))
- Global Lambda Integrated Facility ([GLIF](#))
- Global Network Architecture ([GNA](#))

The following are some of the highlights of research efforts and accomplishments in the last year.

National 100G Software Defined Networking Testbed

ESnet operates a several network testbeds, available to researchers from DOE, universities, and industry. The 100G testbed, logically separate from ESnet's production network, allows for potentially disruptive wide-area network research in areas of new network protocols, including alternatives to TCP/IP; automatic classification of large bulk data flows; and high-throughput middleware and application development. It is a realistic national-scale environment for innovative network research. A high-latency environment such as seen in production transcontinental networks would be impossible for most researchers to create in their labs. The testbed includes mechanisms to make it easy for researchers to manage their custom host images, including tools to create, modify, save, and restore test environments. It offers researchers maximum flexibility as they get "super-user" to all hosts, "bare metal" host access, their own boot image with root access, ability to install custom operating systems on hosts, a controlled environment that supports reproducible results and complete control of every packet on the network. The testbed also has connections to two ExoGENI racks, one as NERSC in Berkeley, and one at StarLight in Chicago.

Software Defined Networking Testbed Hardware Update

The ESnet Software Defined Networking (SDN) testbed (see Fig 8) deployed in 2015 has undergone a switch hardware refresh to the Corsa Gen2 DP series switches. With the upgrade to these switches, functions such as hardware virtualization, resource slicing, and service isolation are easily facilitated by the switches overlay/underlay architecture

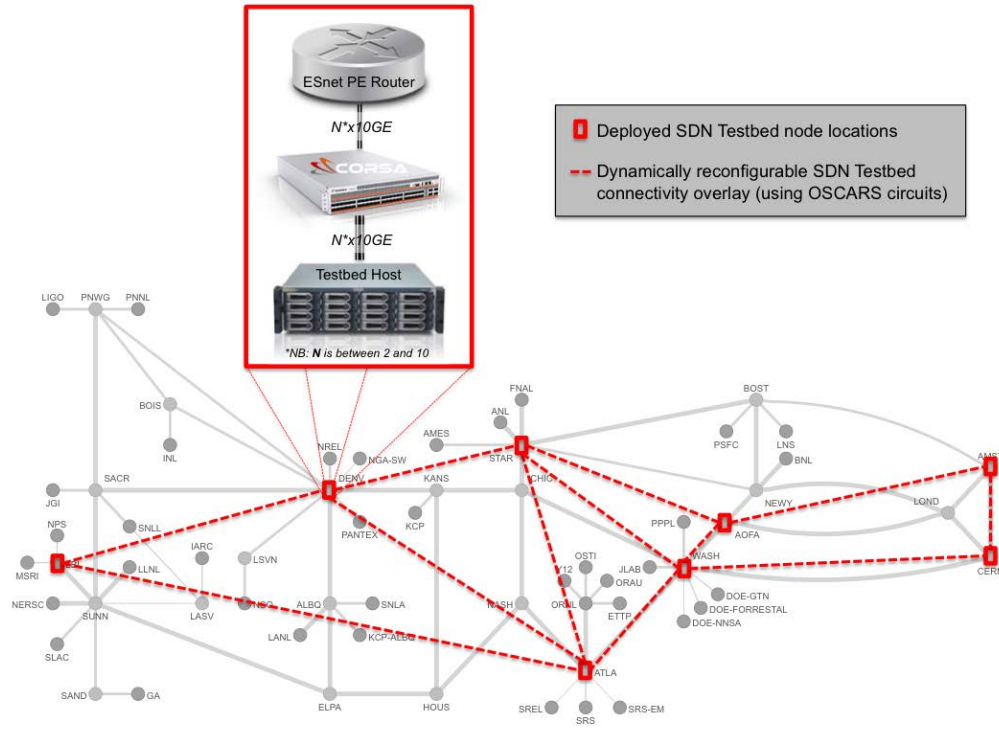


Figure 20: Diagram of ESnet's SDN Testbed.

Figure 8 shows the SDN portion of the testbed, based on OpenFlow v1.3 “white boxes” from the networking startup company Corsa (www.corsa.com). The new testbed will support a range of SDN experiments, including Software Defined Exchanges (SDX), SDN-based security and traffic analysis, dynamic multi-point VPNs, and much more.

ESnet held an SDN workshop in July 2015 at the Lawrence Berkeley National Laboratory. The workshop brought representatives from academia, industry, government, and the R&E community into a dialogue of SDN realities and desirable futures, with a focus on the government and R&E community application requirements. The overall goal of this workshop was to identify the challenges and opportunities to operationalize Software Defined Networks, Layer 1/2/3 Exchanges and Infrastructures (SDN/SDX/SDI) for research and education in the 2-3 year timeframe.

Dark Fiber Testbeds

ESnet’s dark fiber test bed enables opportunities to test and research new network architectures and technologies. Possibilities include potentially disruptive technologies that are incompatible with existing optical systems, requiring dedicated fiber, such as dynamic optical or packet-optical systems and high-speed networking greater than 100Gbps. Researchers share the cost of installing new hardware, colocation space and miscellaneous costs such as shipping.

Intent-based Networking Research

Intent-based networking is a novel research area, particularly focused on network users to provide user-specific networks for user-intended service. Its goal is to provide an easy to use English-based language to communicate user requirements to underlying network architectures.

ESnet has developed a proof-of-concept system called INDIRA (Intelligent Network Deployment Intent Renderer Application) to translate user needs into network commands, currently instigating path allocation (using NSI) and transferring files (using Globus). INDIRA uses natural language processing and semantic RDF graphs to understand, interact, and create the required network services, interacting with SDN northbound interfaces to enable what-is-called Intent-Based Networking for Scientific Networks. For example, high-level user queries such as “For project1 transfer files between LBL and BNL”, is translated into network provisioning commands, setting up links between data transfer nodes at LBL and BNL, and instantiating the data transfer. INDIRA can create multiple network scenarios for multiple applications, configure multiple network tools and collects information on user-application needs for future analysis of network usage.

Figure 9 shows the stages involved in a service using intent from start until completion. The intent language is parsed into network parameters, and deployed into the network using SDN. The renderer takes inputs from multiple data files, such as user profiles and topology details to automate the conversion.

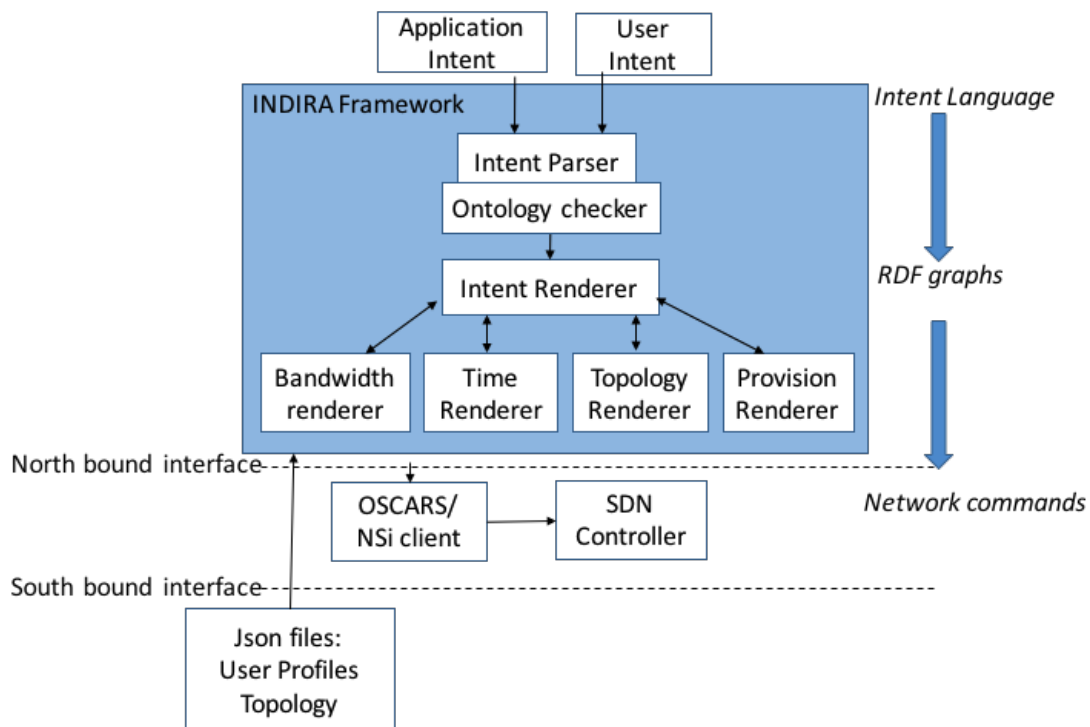


Figure 21: INDIRA architecture.

In its current state, INDIRA has made two significant accomplishments: (1) an innovative method of using natural language processing or ontologies to improve network intent capture, and (2) integrating multiple network tools to satisfy high-level user QoS/QoE requirements. INDIRA uses ontology engineering with minimal implementation of AI reasoning, assesses user-network needs, topology awareness, bandwidth permissions and profile checking. This was successfully demonstrated at the SC16 conference, where we did a conversion of ‘English’ to network API calls, enabling multiple tools such as NSI and Globus to perform the intents. INDIRA was able to enable multiple service quality for user transfers, which are not otherwise possible by using these tools individually. We are hopeful that this method will start a new research direction in intent-based networking, focusing on end-user needs and simplifying networks for them. This method also brings a one-command configuration process, as opposed to using multiple lines of code to configure networks, an approach still widely used by other intent networking projects.

With this new approach, applications can specify what they want, rather than how. Using a local knowledgebase to render intent into actionable set of network commands, intent verbs and nouns are converted to network variables. In addition, we developed a feedback loop, communicating to users, the state of network and administrative policies, to refine intent. This is important for self-discovery, programmability and automation. We are in the process of enhancing INDIRA with further network provisioning tools, automating path provisioning and publishing results to the network research community

ESnet's Participation in SC16

For the SC conference in Salt Lake City, UT. ESnet, in partnership with CenturyLink provided five 100 Gbps circuits to the convention center in support of various community demonstrations. The network links were used for a number of demonstrations between booths on the exhibition floor and sites around the world.

- A Caltech group showcased network path-building and flow optimizations using SDN and intelligent traffic engineering techniques, built on top of a 100G OpenFlow ring at the show site and connected to remote sites including the Pacific Research Platform (PRP).
- The Multicore-Aware Data Transfer Middleware (MDTM) project demonstrated the advantages of MDTM in fully utilizing the multi-core system resources, in particular with NUMA architecture using their mdtmFTP data transfer tool.
- The Naval Research Lab's demonstrated large-scale remote data access, a dynamic pipelined distributed processing framework and Software Defined Networking (SDN) enabled automation between distant operating locations. The demonstration included a complex video processing workflow, that could be rapidly redeployed as needed to satisfy varying processing needs and leverage available distributed resources in a way that is relevant to emerging intelligence data processing challenges.
- An ANL collaboration used the ESnet SDN testbed to demonstrate a new traffic management algorithm that takes advantage of the SDN controller's ability to monitor flow statistics and push flow control logic down to the network fabric in real-time.
- SLAC demonstrated the transfer of many small files (1 million 1MByte files) with and without TLS encryption, motivated by LCLS-II's need for semi real-time transfer and access to the data acquisition system's output.
- Aspera demonstrated the next generation of Aspera FASP, which is about optimizing both the WAN and Storage to provide maximum utilization of available resources to transfer data as fast as possible. FASP operates over regular IP, supports AES-GCM encryption, and works on commodity Intel hardware.

ESnet6

The drive to build the next evolution of the ESnet backbone network comes with the primary mission "To enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools". To this end, efforts to design ESnet6 is underway with the expectation of rolling it out in the 2019-2020 timeframe.

Formalizing Requirements

To understand the design parameters for ESnet6, several tasks were taken to formalize requirements from which design decisions would be made. These included understanding what capabilities were needed as well as raw capacity requirements.

The capability requirements were primarily focused on three efforts:

- Documenting scientific workflows which provided an understanding of how our customers (the scientists) use the network
- Developing a portfolio of services that would support the various documented scientific workflows
- Formalizing technical requirements needed to facilitate the defined services

To understand capacity requirements, a bandwidth usage projection effort was undertaken to provide guidance for usage within the 2020 and 2025 timeframes. The basis of this investigation revolved around ESnet’s general growth trend of about 10x every 47 months since 1990. The result of this exercise were capacity maps such as the one seen in Figure 10.

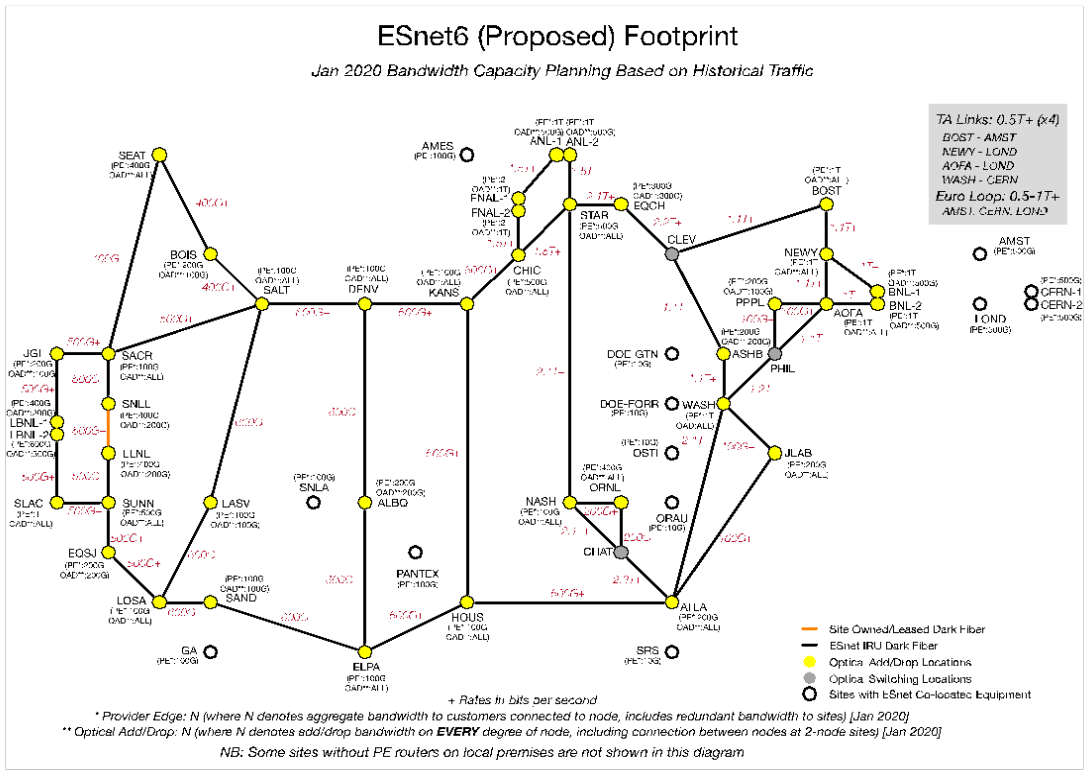


Figure 22: ESnet Proposed Capacity Map for 2020

Architecture Exploration

With the mindset of building a network using next-generation technology, it was determined that a 12-month R&D investigatory period was necessary understand the technology landscape for different features and functions, and cost and supportability. The initial study focused on six architectures (see Fig 11) to frame the technology research scope, with the aim to eventually eliminate or converge remaining architectures picking the best of breed technologies.

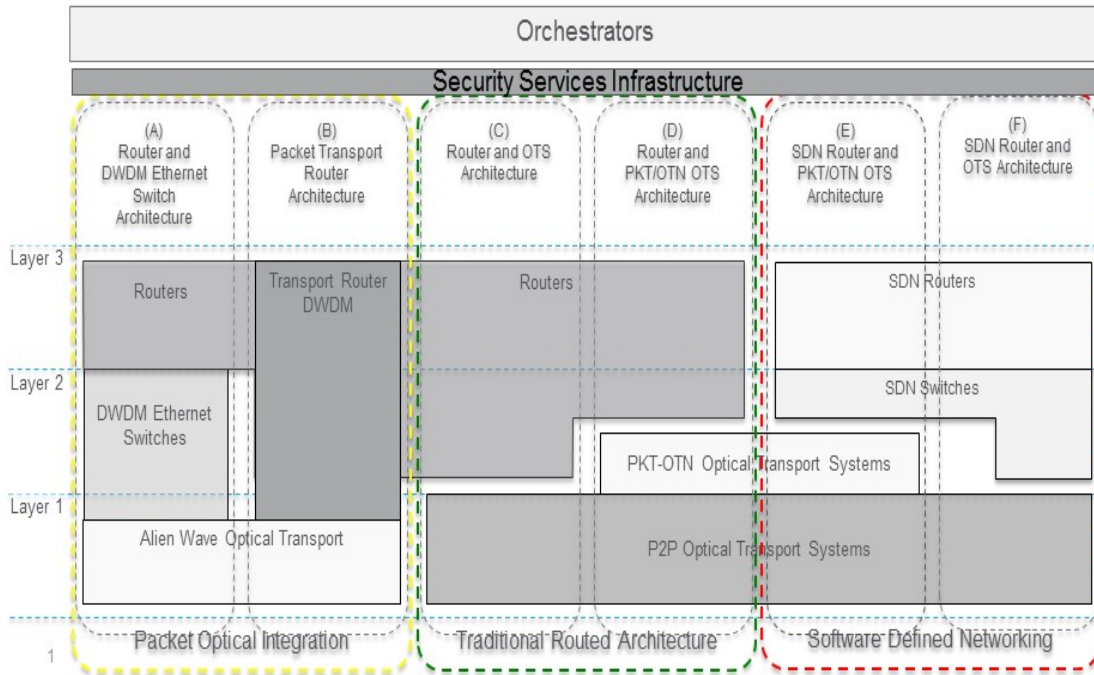


Figure 23: ESnet6 Initial Architecture Study-Architecture and Technologies Matrix

The results of this study are expected in 2H2017 when the formal research period concludes.

Conclusion

ESnet's mission is to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools. Our vision is that scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data and that collaborations at every scale, in every domain, will have the information and tools they need to achieve maximum benefit from scientific facilities, global networks, and emerging network capabilities. **ESnet continues its work to foster the partnerships and pioneer the technologies necessary to ensure that these transformations occur.**

Annex 4: StarLight - An International/National Communications Exchange Facility

Submitted by

Joe Mambretti, Northwestern University

Maxine D. Brown, University of Illinois at Chicago

Thomas A. DeFanti, University of California, San Diego

January 2017

Introduction:

The StarLight International/National Communications Exchange Facility, which is located on the Chicago campus of Northwestern University <<http://www.startap.net/starlight/>>, was prototyped in 2000 and established as a production facility in 2001, when it became the world's first international all-optical exchange, along with its partner, NetherLight, in Amsterdam. Today, StarLight is the highest capacity exchange in the world, supporting multiple 100 Gbps networks (including support for more than 40 individual 100 Gbps paths) as well as over 100 10 Gbps paths in the metro area, regionally, nationally, and internationally. Since its inception, StarLight management and engineering experts have been working directly in partnership with national and international academic research communities to create a leading-edge facility in support of advanced data-intensive e-Science applications, architectures, services, technologies, network performance measurement and analysis, and in-depth computing and networking technology evaluations. In addition, the StarLight facility has been, and continues to be, a major innovator and implementer of architecture and techniques related to highly flexible networking – *programmable networking* (now termed Software Defined Networking or SDN), including SDN exchanges (SDXs) and Software Defined Infrastructure (SDI) extensions. The StarLight facility constitutes an innovation platform that enables researchers to design, develop, empirically test, and demonstrate next-generation services, capabilities, protocols, standards, and technologies. At any given time, StarLight typically supports between 20 and 30 national and international research testbeds.

The StarLight facility supports multiple National Research & Education Networks, all major Federal agency networks (serving as as the Midwest NGIX), regional networks, state-wide networks, specialized science networks, and more major international, national, regional and local network research testbeds (such as iGENI, the International Global Environment for Network Innovations, the national GENI environment, and NSF cloud testbeds such as Chameleon) than any other exchange facility in the world. StarLight has fiber and circuits from many vendors, including AboveNet, AT&T, Cogent, Global Crossing, Level3, CenturyLink, RCN, Lightower, Zayo Group, CrownHub, and Sunesys.

StarWave was introduced in 2010, an NSF-funded multi-100 Gbps communications exchange facility for data-intensive scientific research that is located within the StarLight Facility. Core components of StarWave include high-performance foundation switches based on 100 Gbps standards that have been finalized by international standards organizations, including the IEEE (for 100 GE), the ITU (for 100 Gbps switching), and the OIF for interconnections. This project was undertaken by the Metropolitan Research and Education Network (MREN), an advanced seven-state regional R&E network that exclusively focuses on supporting data-intensive science research in partnership with the StarLight consortium and several major research universities and national laboratories in the region, including Argonne National Laboratory, Fermi National Accelerator Laboratory and the National Center for Supercomputing Applications at University of Illinois at Urbana-Champaign.

StarWave leverages several 100 Gbps wide-area network infrastructure projects currently under development by U.S. and international Federal agencies and interconnects with those fabrics, including:

- **CANARIE**, Canadian national 100 Gbps backbone network;
- **U.S. DOE ESnet5 (Soon ESnet6)**, ESnet's fifth-generation 100 Gbps production network, which refers to both ESnet and ESnet SDN;
- **U.S. DOE ESnet 100 Gbps Testbed**
- **Internet2** national 100 Gbps backbone;
- **PacificWave/Pacific NorthWest GigaPoP**
- **Global Research Platform Network (GRPnet)**
- **KREONET**: The international Korean R&E 100 Gbps network
- **KREONET SD-WAN**: The international Korean R&E SDN network
- **Metropolitan Research and Education Network (MREN)**
- **I-Wave**: A National Center for Supercomputing Applications Network
- **UICnet**: a 100 Gbps network connecting the University of Illinois at Chicago campus
- **Research On Demand Network** (An international Ciena 100 Gbps testbed)
- **Petascale Science Network Testbed**
- **OMNIPoP**
- **StarLight International Exchange Network**

StarLight users include a global scientific community involved in networking, computational science, network science, data management, visualization and computing research using IP-over-lambda networks. StarLight also supports experimental protocol and middleware research of high-performance application provisioning of lightpaths over optical networks. Many of its experimental testbeds are supported in partnership with Northwestern University's International Center for Advanced Internet Research (iCAIR), whose research laboratory is located across the street from StarLight and is directly connected with fiber capable of supporting 360 Gbps. iCAIR has a partnership with the NASA Goddard Space Flight Center, the Laboratory for Advanced Computing at the University of Chicago (LAC), MREN, the StarLight consortium, the Open Data Commons consortium, the Open Science Data Cloud (OSDC), the Open Cloud Consortium (OCC), CalTech, and multiple national and international partners to investigate novel architectures, technologies (including new protocols), and techniques for data-intensive science based on 100 Gbps capabilities. As part of this research, iCAIR and OSDC have conducted experimental investigations using a novel cloud technology on the OSDC national testbed, which is supported by the NSF and the OCC. In November 2015, these research partners staged more than 15 100 Gbps demonstrations at SC15, working with the SCinet WAN group to provision four dedicated 100 Gbps paths from StarLight to the SC15 showfloor in Austin, Texas, two dedicated 100 Gbps paths from the Naval Research Lab in Washington DC, and a 100 Gbps over ESnet from NASA Goddard Space Flight Center in Greenbelt, Maryland, through the MidAtlantic Crossroads Exchange (MAX) near Washington DC, using private fiber. On the conference show floor, a 100 Gbps ring was established that interconnected four booths via SCinet: iCAIR/OCC, CalTech, University of Michigan, and Dell. At SC15, StarLight also supported the University of Chicago-iCAIR demonstration of a Bioinformatics Software Defined Network Exchange (SDX) and large-scale transfers of National Oceanographic and Atmospheric Administration (NOAA) data from the OCC repository.

StarLight supports several major NSF project, the Global Environment for Network Innovations (GENI), as well as a related project, iGENI – international GENI – which is defining, designing, and implementing a globally distributed network research infrastructure. iGENI is being integrated

with current and emerging GENI resources, extending GENI internationally, to: (a) expand the variety of resources, especially controllable transport services, available to GENI researchers, (b) add additional capabilities, (c) make GENI available to worldwide research communities, and (d) provide an experimental platform to demonstrate its capabilities for supporting experiments.

Another GENI supported project is designing, implementing and operating a GENI Software Defined Networking Exchange (SDX) (aka Multi Services Exchange), an initiative established Sept-Oct 2013. This project is designing and implementing key software and hardware components of a layer 2 SDN/OpenFlow exchange (SDX) between GENI layer 2 network resources and other research networks. The project has implemented capabilities to request and receive resources from the exchange that are fully integrated with GENI standard interfaces, such as the GENI clearinghouse, the GENI AM API, GENI stitching AMs, and GENI Commercial Software Defined Exchange Point. This initiative demonstrated SDX capabilities at multiple national and international workshops and conferences, including the GENI Engineering Conferences and the SC conferences.

StarLight also supports an NSFCloud project, the Chameleon Cloud national testbed, which is based on a 100 Gbps national backbone with regional 100 Gbps network extensions. StarLight cloud research activities include investigations of techniques for Software Defined Infrastructure (SDI).

A recent grant, awarded through the NSF International Research Network Connections program (IRNC), provides support for the design, implementation and operation of an international SDX to support global data-intensive science. This IRNC SDX initiative is designing, developing and implementing highly advanced, diverse, reliable, persistent, and secure networking services, architectures, and technologies, to enable scientists, engineers, and educators to optimally access and utilize resources in North America, South America, Asia, South Asia (including India), Australia, New Zealand, Europe, the Middle East, North Africa, and other sites around the world. The StarLight International/National SDX is being developed to ensure continued innovation and development of advanced networking services and technologies, and will be interconnected to other IRNC supported SDXs as well as national SDXs. One SDX supported is the international BioInformatics SDX.

StarLight also supports the Pacific Research Platform (PRP) an NSF-funded leading-edge distributed research infrastructure, which interconnects the Science DMZs of dozens of top research institutions through three advanced networks: CENIC's California Research & Education Network (CalREN), the Department of Energy's Energy Science Network (ESnet), and Pacific Wave. Currently, plans are underway to extend this platform to additional sites, including sites supported by the StarLight Software Defined Network Exchange (SDX) in Chicago. Connectivity is also anticipated for UC San Francisco, UC Santa Barbara, UC Merced, the National Center for Atmospheric Research (NCAR), the University of Amsterdam in the Netherlands, Montana State University, the University of Hawaii, and the University of Illinois at Chicago, through StarLight. Science DMZs are designed to create secure network enclaves for data-intensive science and high-speed data transport. This regional research platform will interconnect many campuses' dedicated research networks to create a secure, seamless fabric that maximizes end-to-end performance and ease of use, and minimizes administrative difficulty.

The StarLight consortium is also participating in the recently announced initiative Global Research Platform (GRP), which will extend the PRP architecture to additional sites around the world. One component of this fabric is the CENIC/PacificWave/PacificNorthwest GigaPoP series of planned SDX facilities and the new PNWGP↔StarLight 100 Gbps connections, including the PacificWave/PNWGP↔StarLight Global Research Platform Network (GRPnet).

The capabilities of the StarLight exchange are showcased during major national and international workshops, conferences and other forums, especially the International Conference for High Performance Computing, Networking, Storage and Analysis (SC conferences). In November 2016, at SC16, the StarLight consortium with its national and international partners participated in the SCinet Network Research Exhibition (NRE) showcases and experiments, including staging 42 sets of national and international 100 Gbps demonstrations using SDN, SDX, and SDI architecture and technologies.

StarLight International R&E Networks

- **ANA-300G:** Advanced North Atlantic 100Gbps Ring (200G) for Research & Education
- **ASGCNet:** Academia Sinica Grid Computing Network, Taiwan
- **AmLight / AMPATH / Florida LambdaRail:** Americas Lightpaths, AMPATH, Florida LambdaRail, U.S.
- **CANARIE:** Canada's Advanced Research and Innovation Network, Canada
- **CERNET:** China Education and Research Network, China
- **CERNnet:** European Organization for Nuclear Research network (AS-513 at StarLight), CERN
- **CESNET / CzechLight:** Czech Educational and Scientific Network and CzechLight, Czech Republic
- **CSTNET:** The Chinese Science and Technology Network
- **ENSTInet:** Egyptian R&E network
- **GARR** Italian Academic & Research Network
- **GÉANT:** Pan-European network infrastructure serving Europe's Research & Education community
- **GLORIAD:** Global Ring Network for Advanced Applications Development, Global
- **GLORIAD-ENSTINET:** GLORIAD and Egyptian National Science and Technology Information Network, Egypt
- **GRNET** Greek Research and Technology Network, Greece
- **KREONet2 / KRLight:** GLORIAD and Korea Research Environment Open Network and KRLight, Korea
- **KREONet SD WAN** network
- **LHCONE:** Large Hadron Collider (LHC) Open Network Environment (Tiers 2 and 3), Global
- **LHCONE GARR:** LHCONE (LHC Open Network Environment) on the GARR (Italian Academic & Research Network) infrastructure)
- **LHCONE GÉANT:** LHCONE (LHC Open Network Environment) on the GÉANT infrastructure
- **LHCOPN:** Large Hadron Collider (LHC) Optical Private Network (Tier 1), Global
- **Open Transit:** Open Transit Internet, France Telecom, France
- **SINET** Japanese science network
- **StarLight Network:** StarLight/STARTAP facility private network for interconnectivity, U.S.
- **SingAREN** Singapore international Advanced Research and Education Network and SingLight, Singapore
- **SURFnet / NetherLight:** Dutch Higher Education and Research Network and NetherLight, The Netherlands
- **TWAREN / TaiwanLight:** Taiwan Advanced Research and Education Network, Taiwan
- **UNAMnet:** National Autonomous University of Mexico Research and Education Network, Mexico

- **USLHCNet:** U.S. Large Hadron Collider (LHC) Network, U.S.
- **XENON** network

StarLight National R&E Production Networks

- **DREN:** U.S. Department of Defense, Defense Research and Engineering Network, U.S.
- **ESnet SDN:** ESnet Science Data Network (SDN), U.S.
- **ESnet5:** U.S. Department of Energy Office of Science, Energy Sciences Network (IP network), U.S.
- **Global Research Platform Network (GRPnet)** – a 100 Gbps network)
- **Internet2 AL2S:** Internet2 100GE, Layer 2 connection, SDN and Science DMZ, U.S.
- **Internet2 ION:** Internet2 Interoperable On-demand Network (ION), U.S.
- **Internet2 IP Network:** Internet2 Research & Education Network, U.S.
- **Internet2 NDDI/OS3E:** Internet2 Network Development and Deployment Initiative (NDDI) / Open Science, Scholarship and Services Exchange (OS3E), U.S.
- **Internet2 TR-CPS:** Internet2 TransitRail-Commercial Peering Service (TR-CPS), U.S.
- **N-WAVE:** NOAA (National Oceanic and Atmospheric Administration) Research Network, U.S.
- **NIH Network:** U.S. Department of Health and Human Services, National Institutes of Health, U.S.
- **NISN:** NASA Integrated Services Network, U.S.
- **PacificWave/Pacific Northwest GigaPoP Network** (a 100 Gbps network, including the Western Region Network)
- **USGSnet:** United States Geological Survey Network, U.S.
- **XSEDE** Network

StarLight Selected National R&E Testbed Networks

- **AutoGOLE Testbed**
- **Advanced International Network Research Testbed**
- **Ciena International Research On Demand 100 Gbps Testbed**
- **Chameleon Cloud Testbed Network**
- **Chameleon Cloud Staging Testbed Network**
- **Cybera SDN/OpenFlow Testbed**
- **ESnet 100G Testbed:** U.S. Department of Energy ESnet 100G national testbed, U.S.
- **ExoGENI International Testbed:** A GENI-ESnet-SURnet-University of Amsterdam-iCAIR Ciena Research Labs collaborative project exploring SDN and 100 Gbps and 40 Gbps paths
- **FELIX: FEderated Test-beds for Large-scale Infrastructure eXperiments,** a common framework for requesting, monitoring and managing slices provisioned over distributed and distant Future Internet experimental facilities in Europe and Japan
- **Future Internet Research Experiment Testbed (FIRE)**
- **GEMnet:** NTT Global Enhanced Multifunctional Network, Japan
- **GENI Experiment Engine (GEE)**
- **GENI Mesoscale network**
- **GLIF NSI Testbed:** GLIF Network Services Interface testbed, Global
- **Grid'5000 Testbed**
- **GpENI** Network research testbed
- **HECN:** NASA Goddard Space Flight Center High-End Computer Network (HECN) Network, U.S.
- **HPDMnet:** Global research consortium's High Performance Digital Media Network, Global
- **International GENI Testbed (iGENI)**

- **ICN Testbed**
- **IOFnet:** International OpenFlow Testbed Network, Global
- **JGN-X:** Japan Gigabit Network – eXtreme testbed, Japan
- **LHC P2P Testbed**
- **MEICAN** – A DCN Life-Cycle Management Platform testbed
- **NASA GSFC Testbed**
- **NRL Testbed**
- **OMNInet:** U.S. research consortium’s Optical Metro Network Initiative (OMNInet), U.S.
- **Open Science Data Cloud Network:** Open Science Data Cloud national research testbed, U.S.
- **Petascale Science Testbed:** Multi-institutional national 100 Gbps testbed, U.S.
- **Research On Demand 100 Gbps Testbed:** Ciena Research Labs 100 Gbps international testbed
- **SeaCloud – International testbed**
- **Smart Applications On Virtual Infrastructure (SAVI) Cloud Testbed:** Canadian cloud national distributed application platform testbed for creating and delivering future internet applications
- **ToMaTo International Testbed:** International testbed for experimenting with topology-oriented control framework for virtual networking, managed by the University of Kaiserslautern
- **TransCloud Network:** Dynamic network / Dynamic cloud resources network testbed, U.S.
- **TMRTSnet:** International transcoding migration testbed network
- **Virtual Transit Service Network Research Testbed**
- **V-Node Testbed**

StarLight Other National / Regional / Local R&E Production Networks

- **MREN:** Midwest consortium regional Metropolitan Research and Education Network (MREN), U.S.
- **BOREAS:** Midwest consortium Broadband Optical Research, Education and Sciences Network, U.S.
- **CCAnet:** Columbia College Advanced Network, U.S.
- **CMHnet:** Children’s Memorial Hospital Network, U.S.
- **DePaul Advanced Network**
- **DOE/ESnet 100 Gbps CHI-Express:** Chicago Metropolitan Area Network project connecting ESnet in Chicago to StarLight, Argonne National Laboratory, Fermi National Acceleratory Laboratory, U.S.
- **Fermi LightPath:** Fermi National Accelerator Laboratory Lightpath, U.S.
- **Loyola University** advanced network
- **MPEAnet:** Metropolitan Pier and Exhibition Authority Network
- **Northern Tier Network Consortium:** Research and education network serving institutions in the upper-northwestern states, U.S.
- **NorthernWave:** A peering exchange facility connecting Pacific Wave in Seattle with StarLight in Chicago
- **Oak Ridge Network:** Oak Ridge National Laboratory advanced network, U.S.
- **OMNIPoP:** CIC Regional network, U.S.
- **PNWGPnet:** Pacific NorthWest GigaPoP Network, U.S.
- **Southern Light Rail:** Southern Light Rail (SRL) regional research network, U.S.
- **UILnet:** Private optical fiber network connecting the Digital Manufacturing and Design Innovation Institute

- **University of Chicago private optical fiber based 100 Gbps network**
- **UIC network**
- **Western Region Network**
- **XSEDE:** Extreme Science and Engineering Discovery Environment (XSEDE), U.S.

StarLight State R&E Networks

- **HealthNet Connect:** Iowa Health System's HealthNet connect network, Iowa
- **I-Light:** State of Indiana's optical fiber network for higher education, Indiana
- **ICCN:** University of Illinois Intercampus Communications Network, Illinois
- **ICCnet:** Illinois Cloud Consortium program, Illinois
- **InterCampus Network** (formerly part of I-WIRE), a 100 Gbps network interconnecting StarLight the University of Illinois at Chicago, and the University of Chicago on private fiber.
- **Illinois Century Network:** Illinois Century Network Research and Education Network, Illinois, based on a 100 Gbps service at StarLight
- **Illinois RuralHealthNet:** Illinois Rural HealthNet, Illinois
- **IllinoisWave:** National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign, Illinois, (NCSA/UIUC/NCSA) has four 100 Gbps paths to Chicago, including one terminating on the StarWave facility in support of the NSF Blue Waters petascale computing facility, implemented on the Illinois campus-to-campus optical fiber network (ICCN)
- **MERIT:** Merit Network, Michigan
- **MiLR:** Michigan LambdaRail (MiLR), Michigan
- **MONON100:** Indiana 100Gbps Research and Education Network
- **NIUnet:** Northern Illinois University R&E Network, Illinois
- **OARNET:** Ohio Academic Resources Network, Ohio
- **PeoriaNEXT:** Peoria NEXT for regional economic development, Illinois
- **Southern Illinois University Advanced Network**
- **UMNnet:** University of Minnesota Network, Minnesota
- **UNDnet:** Notre Dame University R&E Network, Indiana
- **University of Chicago Advanced Network**
- **WiscWave / WiscNet:** University of Wisconsin–Madison optical network; Wisconsin Research and Education Network, Wisconsin

StarLight Commercial Networks

- **AboveNet**
- **AT&T**
- **CenturyLink**
- **Cogent Communications**
- **CrownHub**
- **Global Crossing**
- **Level3 Communications**
- **RCN**
- **Lightower**
- **Sunesys**
- **Windstream**
- **Vorizon**
- **Zayo Group**

StarLight Accessible International and National R&E Production Networks

While not directly connected to StarLight, these networks are closely integrated via other R&E networks.

- **ACE** network (US↔EU)
- **AARNet-SXTransPORT**: Australian Advanced Research Network in cooperation with Southern Cross Cable Networks, Australia
- **CENIC**: Corporation for Education Network Initiatives in California, U.S.
- **CERN/TIFR**: CERN/Tata Institute of Fundamental Research (TIFR), CERN/India
- **CUDI**: Corporación Universitaria para el Desarrollo de Internet, Mexico
- **HARNET**: Hong Kong Academic and Research Network, Hong Kong
- **i2CAT-CATLight**: Internet2 Catalonia, Barcelona
- **IceLink**: Transatlantic polar network linking the U.S., Canada, Greenland, and the five Nordic countries (Iceland, Denmark, Norway, Sweden, Finland), Global
- **Innova-Red**: Argentina Research and Education Network, Argentina
- **JANET**: United Kingdom Education and Research Network, U.K.
- **KAUST**: King Abdullah University of Science and Technology, Saudi Arabia
- **KOREN - APII - JGN-X**: Asia Pacific Information Infrastructure Testbed (APII) managed by Korea Advanced Research Network (KOREN) and the Japan Gigabit Network – eXtreme testbed (JGN-X), Asian
- **KyaTera-Fapesp**: Sao Paulo state research funding foundation (Fapesp) optical network (KyaTera), Brazil
- **NORDUnet / NorthernLight**: Nordic research community Nordic and international network, in collaboration with Denmark, Finland, Iceland, Norway and Sweden and NorthernLight, Nordic Countries
- **PIONIER**: Polish optical network NREN, Poland
- **QNREN** –Qatar R&E Network
- **RBnet** Russia NREN
- **RedCLARA**: Latin American Cooperation of Advanced Networks, Latin America
- **REANNZ** New Zealand NREN
- **RENATER** France NREN
- **REUNA**: Chile Research and Education Network, Chile
- **RNP / SouthernLight**: Brazil National Education and Research Network and SouthernLight, Brazil
- **RNP-Ipe**: Brazilian National Education and Research Network multi-gigabit backbone, Brazil
- **RNP-CPqD-GIGA**: RNP/CPqD experimental optical network, Brazil
- **RUNET**
- **Pacific Wave TransPAC Trans-Pacific Network**: Integrated 100Gbps trans-pacific layer 1, 2 and 3–network
- **TEIN** –EU Orient R&E Network
- **ThaiREN**
- **TransPAC4 / APAN**: Asia-U.S. High-Performance International Networking and the Asia Pacific Advanced Network
- **WIDE** – Japanese network

StarLight organizers are founding participants of the Global Lambda Integrated Facility (GLIF). StarLight is a GLIF Open Lightpath Exchange (GOLE) <<http://www.glif.is/resources/>>. GOLEs, operated by GLIF participants, are comprised of interoperable and interconnected equipment that is capable of terminating lambdas and performing lightpath switching. Using this approach, different lambdas and L2 paths can be connected together, and end-to-end lightpaths established

over them. Normally, GOLEs must interconnect at least two autonomous optical domains in order to be provided with this designation.



Acknowledgment

StarLight continues to be developed by the International Center for Advanced Internet Research (iCAIR) at Northwestern University, the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago (UIC), the California Institute for Telecommunications and Information Technology (Calit2) / Qualcomm Institute at the University of California, San Diego, and the Mathematics and Computer Science Division at Argonne National Laboratory, in partnership with Canada’s CANARIE and the Netherlands’ SURFnet.

NSF funding to Tom DeFanti (PI) and Maxine Brown (co-PI) established STAR TAP (NSF ANI-9712283, Apr 1997-Mar 2000) and StarLight (NSF SCI-9980480, May 2000-April 2005, and OCI-0229642, Oct 2002-Sept 2006). NSF IRNC funding to DeFanti (PI), Brown (co-PI), Tajana Rosing (co-PI) and Joe Mambretti (co-PI) provided a portion of international network engineering support at StarLight (NSF OCI-0962997 for the period July 2010 – August 2015).

NSF funding to Northwestern University and Metropolitan Research and Education Network (MREN), Joe Mambretti (PI) and Linda Winkler (co-PI), supported the development of StarWave (NSF OIA-0963095, September 2010 – August 2013).

NSF funding through the GENI program provided support for the GENI projects noted, including for the GENI prototype Software Defined Network Exchange (SDX). A recent grant awarded

through the NSF International Research Network Connections program (IRNC) provides support for the design, implementation and operation of an international SDX to support global data-intensive science (April 2015-May 2020).

Annex 5: GLORIAD Status and Plan

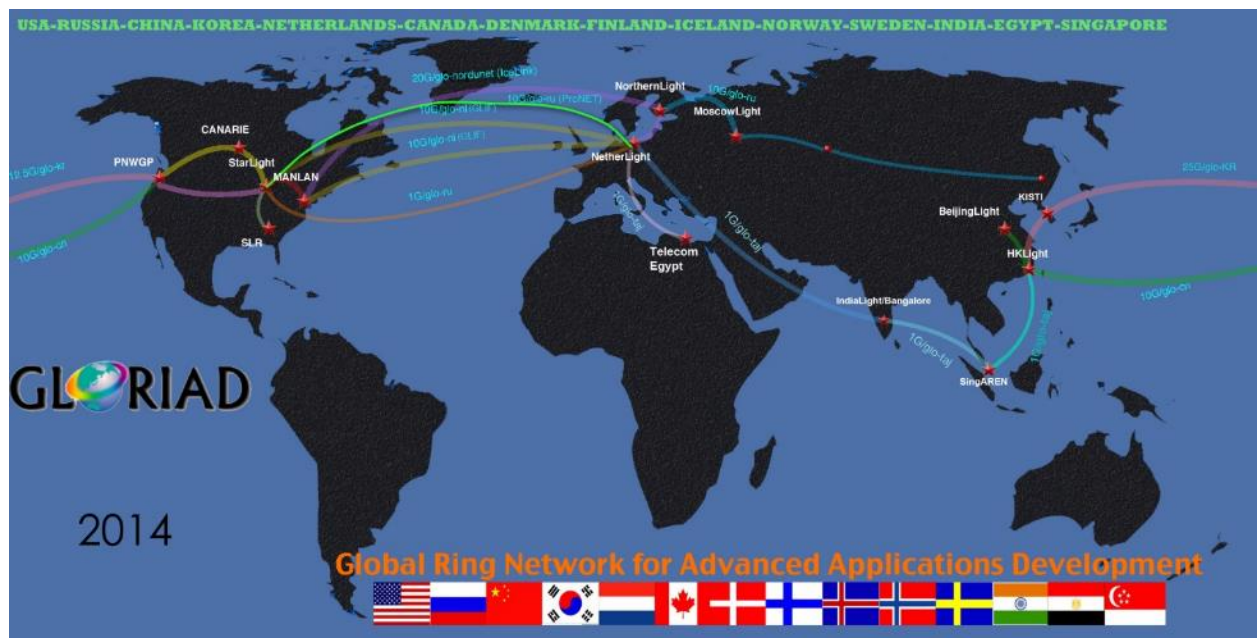
Submitted by Greg Cole (gcole@GLORIAD.org), Susie Baker (sbaker@GLORIAD.org)
University of Tennessee, Knoxville
February 2015

Introduction to GLORIAD

The Global Ring Network for Advanced Applications Development (GLORIAD) is a network-enabled community providing advanced collaborative services to scientists, educators and students around the world. GLORIAD's operations and governance are decentralized, serving to foster the continuous, flexible, community-driven innovation required for modern science infrastructure. Its evolution since its inception in the late 1990s has focused on core principles:

- Build on existing GLORIAD infrastructure to meet new and more advanced needs by an ever growing community;
- Promote an open access community-managed approach to global cyberinfrastructure; improving database and data accuracy to provide solid metrics on utilization, performance and security;
- Broaden U.S. access to underserved science communities;
- Increase the cyberinfrastructure awareness and proficiency of end-users (in the U.S. and elsewhere) through efficient targeted information dissemination and continuing education;
- Connect R&E communities regionally in collaboration with public/private partners; and
- Include new R&E networks that have been largely inaccessible to the US R&E community.

GLORIAD's hybrid network of multiple technologies provides up to 12.5 Gbps, and allows us to provide appropriate and dedicated services to all of our network users according to their research



and education needs. Rich bandwidth and redundant network paths provide excellent reliability.

Figure 24: Latest GLORIAD Network service map. This map includes the recent Korea upgrades (25G to US), the new 10G trans-atlantic link (deployed late 2014); the faded arcs to India and Singapore indicate non-operational status of those links (but plans to re-deploy).

GLORIAD has the capability to link to or peer with all international advanced computing networks. The network is composed of hybrid technologies that serve the most diverse user group. Network services include Layer 1 “optical” lambda lightpaths for “big science” projects and experimental networking research, Layer 2 “switched” Ethernet services especially well-suited for fundamental capacity building and reliability/redundancy, and Layer 3 “routed” high-bandwidth service with peering capability based on the needs of the users.

GLORIAD’s daily traffic typically involves over a million significant flows and over 14 terabytes of data. GLORIAD measures performance through a series of automated queries to determine such factors as “top users,” applications, volume of traffic, and packet loss, which can hinder performance. Every second of every day GLORIAD serves a user community distributed across over 15 million research and education IP addresses.

GLORIAD Network Upgrades in 2014

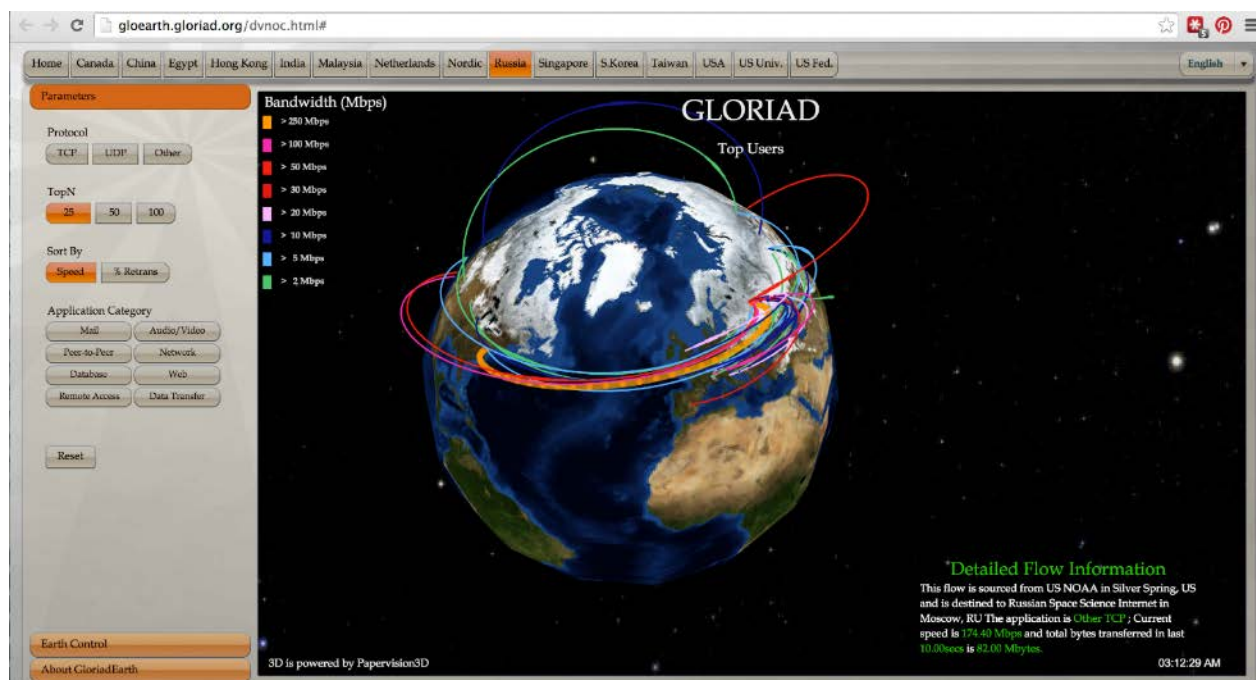


Figure 25: GLORIAD’s upgrade of Russia Circuit with RUNnet Network Flow supports >250Mbps transfer between US NOAA and Russian Space Science Internet (RSSI) System.

Summary of Network Upgrades

- New 10GE circuit follows AC-1 submarine cable from Amsterdam/NetherLight to Chicago/StarLight via NYC, PHL, DCA, PIT, CLE, for direct peering with US via Layer2 connections (VLANs), while their peering with GLORIAD on Layer3 serves as a backup.
- This main router in Amsterdam is connected to NetherLight exchange and is backed up with 3x1GE, consolidated from previous 1GE connections to Russia, India and Egypt and will carry most of the peerings over the Atlantic by NORDUnet, SURFnet/Netherlight and CANARIE.
- With upgrade of US-Russia link to 10G, GLORIAD established a PoP at Vancis facility (formerly SARA) at Science Park in Amsterdam.
- Established a 10GE connection to Russian R&E equipment co-located in Amsterdam, extended by RUNnet via shared 10GE to Moscow (see Figure 2. GLORIADEarth).
- Established a connection with Cisco Telepresence between KISTI and GLORIAD-US.
- Peering was turned-down with NLR in Chicago and Seattle due to network shutdown.

- CANARIE switched one of the GE circuits (on NLR path) between Seattle and Chicago used by GLORIAD, to a new path provided by Internet2.
- GLORIAD's primary connection (VLAN on 10G) between Seattle and Chicago switched from NLR to Northern Wave.
- CSTnet started peering via new 4x1GE links with most US R&E networks.
- Deployed new circuit to allow CSTnet to participate in OpenFlow/SDN testbeds.
- CSTnet has established equipment presence in the US, installing routers in Seattle and Chicago, Juniper MX480 and MX80 respectively.
- GLORIAD-US and GLORIAD-CN provisioned 4 new GE circuits, for CSTnet use, from HKG to Seattle on the jointly operated 10G link.
- CSTnet connected at 10GE through StarLight in Chicago and PacificWave in Seattle, along with backup VLAN connection on Northern Wave between Seattle and Chicago for CSTnet.
- US-Qatar capacity available over new 10GE GLORIAD Trans-Atlantic link supporting Qatar R&E peerings (working with Qatar NREN, QNREN).
- Egypt R&E traffic (working with Egyptian NRENs, ENSTInet and EUNet) provided over 1.2 Gbps capacity to NetherLight (Amsterdam) and 1.0 Gbps to GLORIAD facilities at StarLight (Chicago).

US-China Infrastructure Improvements

GLORIAD-CN partner CSTnet has established equipment presence in the US in early 2014. They installed routers in Seattle and Chicago, Juniper MX480 and MX80 respectively. GLORIAD-US and GLORIAD-CN provisioned 4 new GE circuits from HKG to Seattle on the jointly operated 10G link. These circuits are for CSTnet use. Also, GLORIAD-US provisioned new cross-connects to CSTnet equipment at both locations and re-allocated 2 x 1GE between Seattle and Chicago for CSTnet use and donated by CANARIE. CSTnet has connected at 10GE at both Starlight in Chicago and at Pacific Wave in Seattle. We also worked with PNWGP on creating a VLAN on Northern Wave as a backup connection between Seattle and Chicago for CSTnet. CSTnet started peering via new 4 x 1GE links with most US R&E networks by late Summer 2014.

US-Korea Infrastructure Improvements

October 2013: GLORIAD-US and KISTI established 10GE connection in Hong Kong. GLORIAD-US equipment is housed in KISTI rack. GLORIAD-KR/KISTI and KNU (Kyungbuk National University) jointly participated in a demo, hosted by Caltech, at Denver SC13 event. A 10G dedicated path from KRLight Seattle PoP collocated with Pacific Wave in Westin Bld. was provisioned to Denver SC13 showroom. In Year 4 (April 2014), we also established a connection with Cisco Telepresence between KISTI and GLORIAD-US. GLORIAD-US Chief Network Engineer visited KISTI in May 2014 and had discussions about upcoming projects, including possible use of the new Trans-Atlantic 10G circuit for connecting KISTI directly with CERN and other EU networks in Amsterdam. KISTI team visited Tennessee in July 2014.

New Trans-Atlantic Link for US-Russia/Gulf Region

In late 2014, GLORIAD completed a major project milestone by opening its exchange in Amsterdam at the Vancis/SURFSara facility – built primarily to handle GLORIAD's new 10G US-Russia IRNC link. The GLORIAD-AMS node is designed in a similar fashion to

exchange points in Chicago and Seattle and includes a multi-10GE port and multi-1GE port capable Force10 switch/router. Also included are nprobe server for collecting traffic data/ performance measurements and a PerfSONAR server for link monitoring.

The new 10GE trans-Atlantic link has been provisioned via Global Netwave and Level3 between Chicago and Amsterdam, and is backed up by several GE connections from Amsterdam to Chicago, provided by SurfNET/Netherlight, NORDUnet and CANARIE. Primary connection for the exchange in Amsterdam is 10GE to Netherlight optical switching gear, where many European and Gulf region networks have connections as well; we expect a broad range of peerings to be established in the early 2015. Connection to Russian R&E networks and their equipment at near-by NIKHEF facility in Amsterdam has been upgraded from 1GE to 10GE and now in use via a 10GE connection to Moscow, in collaboration with Russia's RUNNET. Presence of Layer2/3 equipment in Amsterdam will also allow provisioning of direct circuits (via VLAN) for partner networks to connect to exchanges in Chicago and Seattle via Layer2 and allow them to directly peer with US R&E networks, providing better and greater control over routing world-wide. Soon planned (in early January, 2015) is the migration of the US-Egypt link (STM-4/622Mbps) to Amsterdam (currently it is terminated in Chicago), and upgrade of that path from Amsterdam to Chicago to a full GE to serve as a backup link and for experimentation with dynamic circuit provisioning, openflow and other special uses. In addition to the main router, a small switch was installed to aggregate 1GE server connections, as well as a remote access KVM system and a mac-mini for network equipment console connections. A small out-of-band subnet with accompanying IP-transit was acquired from the facility for the additional/redundant access in case of outages or other issues with the equipment. Contract with the facility (Vancis) also includes remote hands as a way of support for the node on a continuous basis.

Measurement Updates in 2014: GLORIAD *InSight* System

GLORIAD's network measurement, monitoring and cybersecurity system, InSight, was completed and put into operation during this program year. This was a substantial achievement and leverages strong partnerships with Cisco (who has provided advice, technical assistance, and nearly \$1M in equipment for our measurement instrumentation and computational cluster) and Qosient (the company behind the open-source Argus network measurement software used as the fundamental base later of InSight).

GLORIAD InSight is an open-source, flow-level passive measurement, analysis and visualization system. It uses the open-source Argus flow-monitoring system as its primary network activity data source – deployed as network probes in Chicago and Seattle to live-analyze traffic from GLORIAD circuits.

Argus provides near real-time comprehensive multi-model, multi-layer, bi-directional network activity monitoring data designed to support network operations, performance and security management. Argus provides structured data models, and metrics for network entities such as L2 and L3 addresses, autonomous systems, overlay identifiers, tunnel identifiers, service and application identifiers, as well as flow oriented utilization, transactional reachability, connectivity, availability, throughput, demand, load, loss and packet dynamics metrics, that can be used to describe complex application, system and path behaviors.

GLORIAD has coupled distributed Argus data generation and collection to its own data transport, processing, and storage technology (built on top of a cluster of powerful Cisco-provided UCS hardware), using ZeroMQ, ElasticSearch, and other open-source technologies, to provide an advanced network situational awareness capability, supporting global network operations, fast coordinated troubleshooting and fault mitigation for the R&E network community.

The system has been built leveraging GLORIAD’s 15-year experience protecting privacy-sensitive flow data, in transit and on storage media (with no breach nor compromise); the system is built on key privacy-sensitive mechanisms regarding data elements observed, stored (primary and backups), in transit and shared.

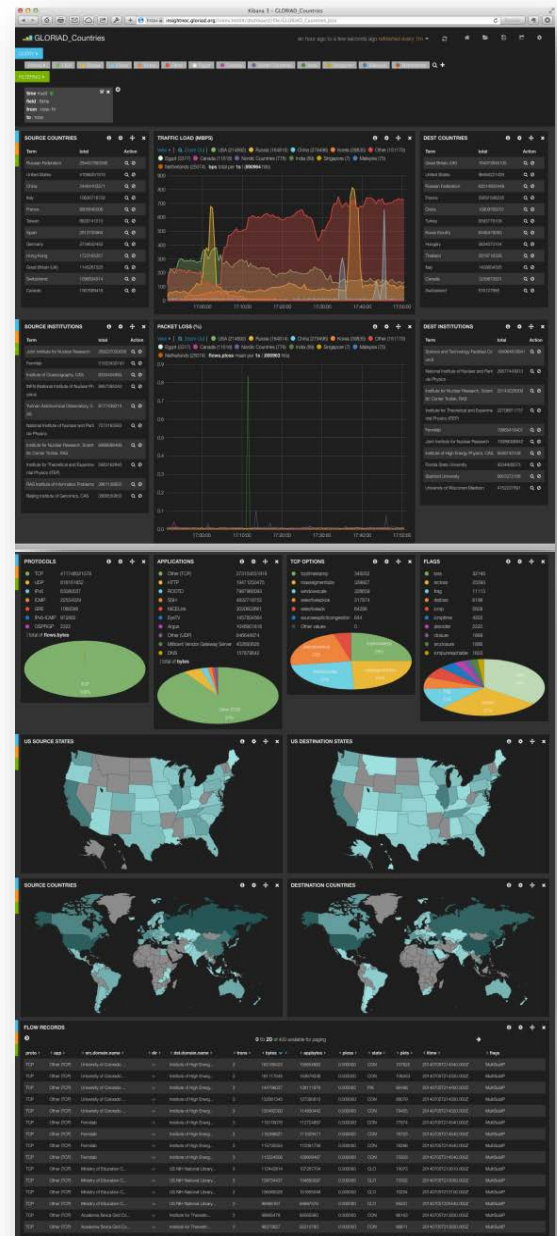


Figure 26: GLORIAD InSight System. An analytic dashboard illustrating near-real-time updates of live traffic flows and including country-level summarization, average packet loss, protocol, application and various performance flags, source/destination by US state and country and individual flow records.

The InSight family of tools is designed to support a variety of implementation scenarios. This family includes the primary InSight product (used for GLORIAD measurement and monitoring), InSight-LAN (running on GLORIAD’s departmental (UT) network today and integrating syslog and other local data sources), InSight-History (updated daily,

provides aggregated summary data since GLORIAD began operating (as MIRnet) in 1998. Publicly accessible URL is <https://insight.gloriad.org/history/>. InSight-Public (full InSight but eliminating all privacy-sensitive data), InSight-Home (running on home networks of several testers today), and InSight-School (similar to InSight-Home but designed for use on K-12 school networks and running on small Intel NUC boxes and Mac Minis). Finally, InSight-NOC (used currently by the GLORIAD operations team) is identical to InSight but with extensions to ingest data from other network devices (e.g., SNMP counters on routers and switches) and various log data sources, integration with the RT ticketing system and tools to assist operations teams. All use the same underlying open-source technologies and differ only in deployment target and in minor, subtle details. For example, large networks require separate probes and servers for scalable data flow; smaller ones combine Argus probe with data store, analytic and web service functionality in one box. InSight for large network and institutional use focuses exclusively on R&E traffic and thus integrates closely with GLORIAD's Global Science Registry; InSight Home and InSight School products provide overview of all (i.e., R&E + commodity) traffic and thus use a different model for assigning flows to organizational domains.

An early version of the system was presented at FloCon 2014 in February 2014 in Charleston, SC where it received much interest. The primary GLORIAD/InSight system was put in production in June 2014 with the Home InSight product deployed at about the same time.

Technologies

The multi-layered *InSight* system is founded on Argus networking monitoring and measurement software and builds on other open-source tools such as MySQL, SQLite, Elasticsearch and Kibana (for extremely fast search and analysis), Perl and ZeroMQ for fast/easy messaging. An open API will permit versions of the system to be built with Python and Ruby. Network measurement devices run on commodity hardware running Unix (FreeBSD and Linux) as do associated analysis platforms.

System Overview

InSight's roots are in a system developed and continuously maintained by the US GLORIAD team since 1999 - with enormous help by China's CSTnet and Korea's KISTI. It begins with a collection of (MySQL) databases describing networks, institutions, geo coordinates, ASnums, VLANs and mappings of R&E addresses to institutions, ASnums, etc. The same MySQL system also tracks large network flows across GLORIAD infrastructure in the US (large flow is any single flow > 100K bytes (not so large by today's standards)) - almost 2 billion records (with a million new records added each day) since GLORIAD began as "MIRnet" in 1999. The system describing R&E institutions is particularly important. Known as the "Science Registry," this database provides descriptive metadata on over 14,000 institutions around the world - and incorporates the very flexible "Dublin Core" standard for easy/flexible addition of new metadata elements. All IP addresses are mapped to records in the Science Registry, enabling GLORIAD to very openly share information about utilization and performance while avoiding serious privacy-sensitive issues related to IP addresses.

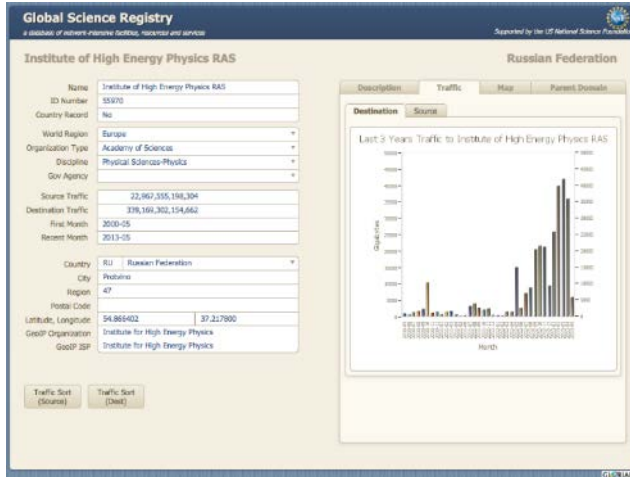


Figure 27: GLORIAD's Science Registry has been developed since 1998 as a means of describing over 14,000 science and education institutions seen as users of the GLORIAD (and other R&E)

Since July of 2012, GLORIAD has been using Argus - a very powerful, open-source bi-directional flow monitor which provides a rich variety of data elements per flow (over 100 metrics per flow) - and with live updates every 5 seconds on each flow. GLORIAD has deployed Argus nodes (using simple commodity Unix Intel boxes (currently FreeBSD although Linux works as well) at its exchanges in Chicago and Seattle and collects the argus data (over 200 million records a day) at its home facility in Knoxville, TN. The live updates are processed utilizing "stream analytic" techniques to assign metadata while the records are still "in flight" (i.e., before being written to disk) and then distributed to various processes for further analysis, live reporting and visualization

and, finally, to disk. This permits the InSight system to provide near-real-time reporting on all traffic flows - with flows showing up on InSight analytic displays within 10-15 seconds of actual appearance on the network. Network engineers can now troubleshoot performance and security-related issues by observing such metrics as throughput, packet loss, jitter - and related DNS/ICMP events - in real-time traffic crawls -- an an individual flow level. This level of "insight" would have previously required hours of painstaking work with tcpdump/wireshark (or would not have been done at all).

A distributed "farm" of "evented" Perl/POE processes do most of the live analysis - an important one pushes metadata assigned records to the Elasticsearch cluster which provides the extraordinarily fast search/analysis for InSight - satisfying most queries in less than 30 milliseconds (and generally involving millions of records). We currently push almost 1,000 records per second into Elasticsearch, which easily handles the indexing and simultaneous searching. Argus and Elasticsearch are the two core technologies on which InSight is built - although it would be impossible to overstate the importance of event-driven Perl/POE applications, ZeroMQ for messaging and of the MySQL and SQLite supporting technologies which drive the critical metadata labeling. Another core technology used is MaxMind's GeoIP database - used for assigning geo-location attributes to all flows as well as other attributions such as country code, AS number, etc. While the core GeoIP database is available for free, GLORIAD subscribes to their commercial service for the improved data quality.

The scale of the global GLORIAD network requires heavy server infrastructure - which has been generously provided by Cisco in the form of 6 blade servers and other equipment - with the servers providing anywhere from 8-32 cores each and from 64G - 256G RAM (and about 40 terabytes of disk space). These machines, all running FreeBSD O/S and the ZFS file system, drive the various Perl/POE-based analytic processes, metadata assignment, database and, importantly, provide an 11-node Elasticsearch cluster (running on the 6 machines). Despite the heavy workload, the machines currently run at less than 2% capacity - providing plenty of room as GLORIAD grows its network capacity, services and analytics.

It is important to note that the exact same system has been deployed on a small network of roughly 70 devices - providing the exact same functionality - and all running on a single 8-year-old (2006) Macintosh with 2 core, 4G RAM and running FreeBSD 10. This machine also runs at less than 2%

load monitoring a network link of 100Mbps - and utilizes all the same components as the larger GLORIAD installation - i.e., it runs Argus, Elasticsearch, ZeroMQ, MySQL and the various Perl/POE based analytics - and, again, on a single (rather old) machine.

Utilization

The network's broad importance is evidenced by the over 8,000 partnerships GLORIAD serves every day. In any 24-hour period, there are tens of thousands of research and education domains involved in over 1,400,000 significant flows involving more than 100 countries that are exchanging over 4 terabytes of data.

GLORIAD measures utilization to illustrate such factors as top institutional users, applications and volume of traffic – as well as packet loss that so dramatically affects performance. Over the past 14 years, we have seen a dramatic increase in users and utilization: from a handful of countries using the network in 1999 to serving over 100 countries in 2012. Furthermore, in the recent past up to 20,000 distinct domains have been detected by our system, which served by the GLORIAD network.

A detailed appendix describing utilization of the GLORIAD-US Infrastructure is available at the following URL: <http://www.gloriad.org/gloriad.annual.report.pdfs.zip>. The appendix is composed of 17 pdf documents. The first describes overall utilization for GLORIAD as a whole and the remaining documents describe utilization by each of the 16 countries affiliated with the GLORIAD project. The US report – at 175 pages – is the most detailed with many institutional summary reports. Each of the appendices includes traffic analysis by communicating countries, autonomous systems (e.g., ASnums), types of organizations (e.g., university, government, academies of science, etc.), top institutional users, scientific discipline, network protocol, network application (and a special set of analyses describing PerfSonar traffic – active measurement data used by R&E networks around the world), and, finally, a detailed analysis of traffic to/from heaviest institutional users.

Many new types of analysis are offered in this appendix and reflect the completion of GLORIAD's new InSight measurement, monitoring and cybersecurity system. The graphic above illustrates total monthly traffic volume (scale is in terabytes) since NSF first began funding in 1999 (project known as "MIRnet" at that time). Total traffic volume since the US team began keeping records in June 1999 is 13.4 petabytes; 10.3 petabytes of that total (77%) has been transmitted since the current IRNC ProNET award began in August 2010. Given the focus of GLORIAD (since its inception as MIRnet) on connecting individual science users, the over 2,000 pages of graphics, tables and narrative richly illustrate use of the GLORIAD infrastructure and its continuing growth.

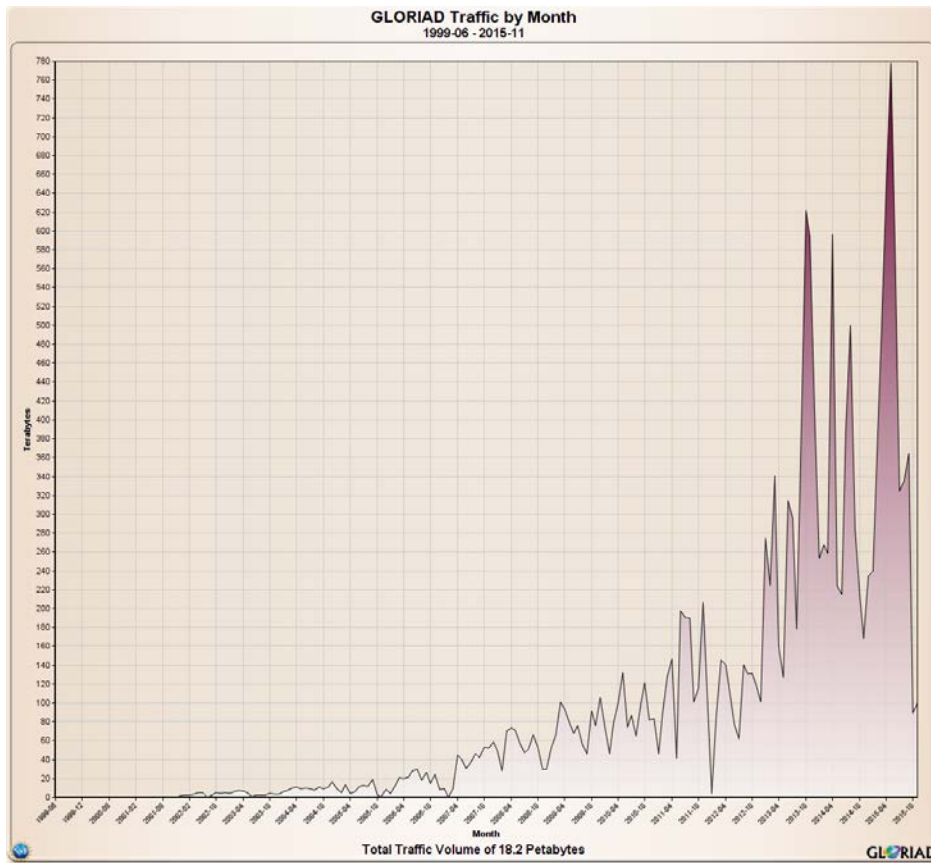


Figure 28: Total GLORIAD traffic volume since the US team began keeping records in June 1999 (through October 2015) is 18.2 petabytes

International Partners

The United States, China and Russia jointly launched GLORIAD in 2004. The mission of GLORIAD in 2005 saw the network expand to Korea, Canada and the Netherlands; and in 2006 the five Nordic countries of Denmark, Sweden, Finland, Norway and Iceland joined GLORIAD, bringing together the world's top networking experts into one distributed, virtual network operating system (dvNOC). With the Taj expansion in 2009, India, Egypt, Singapore, Greenland and, later, Malaysia were added to the GLORIAD family of partners.

Future Plans

- Complete work on new links from Qatar and Kuwait.
- Continue updating and refining the Global Science Registry. In addition to identifying the "big sources" of science data transiting GLORIAD links, we will tag NSF-funded projects for "big facilities" (e.g., data repositories, instruments, etc.). We plan to identify a few of these also for other US government research users (e.g., DOE, NASA, NOAA, NIH, etc.).
- Work with CSTnet on completing deployment of their routing equipment in US.
- Work with CSTnet regarding the next Research Exchange from CNIC (early 2015).
- Host sabbatical for KISTI/Korea senior staff member, Dr. Seunghae Kim.
- Continue work on InSight capabilities (especially geared towards operationalizing near real-time network telemetry) and on "open sourcing" the InSight software base.

Annex 6: Asia-Pacific Advanced Network (APAN) Status and plan

*Submitted by Marcus Buchhorn (markus@intersect.org.au)
February 2015*

Introduction:

APAN (the Asia Pacific Advanced Network) refers to both the organisation representing its members, and to the backbone network that connects the research and education networks of its member countries/economies to each other and to other research networks around the world. APAN members are the organisations representing research and education network interests in the countries/economies of Asia and Oceania.

The focus of APAN is around its so-called “5C’s”: Community, Continuity, Coordination and Collaboration, build across Connectivity. APAN was established in 1995 and has worked continuously since then, leveraging a wide range of investments and contributions, and is not tied to any single project or grant initiative. The Principal Objects of APAN (APAN’s Charter) are set out at: <http://www.apan.net/home/aboutapan/objects.pdf>

APAN Ltd is the legally incorporated not-for-profit Association created to coordinate developments and interactions among its members and with peer international organisations, in both network technology and applications. A major focus of APAN is the user or application community, which is enabled by the efforts of the technical community. It remains a key driver in promoting and facilitating network-enabled research collaboration; knowledge discovery; telehealth; food and water security, and natural disaster mitigation.

APAN also has a commitment to fostering and mentoring the next generation of network engineers and application specialists in the Asia Pacific Region. It runs a Fellowship Program that encourages participation at APAN meetings as well as organising training and building relationships. APAN organises two large meetings each year, typically 300+ attendees, at locations across the region. These provide the venue for its members and other interested participants to come together in working groups, BoFs, committees, plenary sessions, masterclasses and other meetings to review progress, demonstrate advances in technology and applications, and make plans for future activities. Between meetings the working groups run their own occasional workshops and other meetings, which APAN Ltd is seeking to increasingly support. APAN works closely not just with NRENs and NREN partnerships but also funding and development agencies such as the World Bank and UNESCO to identify joint opportunities for community improvements through NREN development and use.

Connectivity

APAN Ltd does not itself own any links, yet, but helps to coordinate the investment efforts of its member link-owners for international R&E connectivity, and relies upon a common Acceptable Use Policy approach to enable peering amongst the many link owners. The major backbone links connecting the Asia-Pacific with Europe and North America are largely funded by members in the region, together with significant contributions from the US (e.g. TransPAC, Glorid and others) and the European Commission (TEIN and others) through various cost-sharing models and strategies. It is not clear how the financial difficulties of the last few years in Europe and the US will impact future investments in the region, which itself has not remained unscathed by the global downturn.

With rapid evolution and deployment of new circuits, keeping the network topology diagrams up to date and in a usable form is time consuming and challenging. APAN has commenced a project

to develop what we hope will be an easier to maintain interactive map, where the user can specify the type of information that he/she wishes to visualise. A prototype of this is available at <http://www.nav6.usm.my/apan/>. Depending on how the project develops, the visual toolset could be expanded to cover routing and management policies as well as performance monitoring. The TERENA Compendium group is adapting their information model to align with the mapping work, to allow easier and greater contributions of link information that can then be published in various forms.

The current topology is summarised in the map below (a high resolution version is available at: http://www.jp.apan.net/NOC/apan-topology_original1.jpg)



Many of the larger economies in the region have more than one NREN, driven in part by various domestic priorities, sometimes by the funding contributions from different ministries or departments. Some focus more on the ‘high-end’ while others tackle the ‘breadth-first’ problems of connecting large numbers of institutions. The latter often limits the peak bandwidth but ensures greater equity of access. These apply to both research and education initiatives. Some also have responsibility for commodity Internet access for their members, and seek to grow their capacity for both R&E and commodity through linked deals.

Recent developments

Apart from the infrastructure changes summarised below, APAN has also become more engaged with infrastructure providers above the network layers. A new Cloud working group was kicked off during 2014 to collate, summarise, review and potentially provide access to a range of cloud-infrastructure developments within the region. International federation of such services is an area of great interest, based on some of the experience with the Australian Research Cloud and others being deployed. In parallel, a new Identity and Access Management Task Force was established, bringing together expertise from identity federations in Australia, New Zealand, Japan, Malaysia and others. This Task Force is within a project-based framework, unlike working groups, and is

developing work plans for the development, deployment and interconnection of identity federations across the APAN members. This will work closely into the proposed MAGIC proposal (led by REDCLARA, seeking EU funding) and potentially other related projects. The other significant change is the establishment of an Internet-of-Things working group, based initially around the former Sensor Network working group and their collaboration with the Agriculture working group. On the network infrastructure, over the last twelve months there have been ongoing domestic upgrades across the region, or planning for them, as well as regular, but less frequent, international upgrades. The more developed NRENs are in the progress of moving some domestic links from 10G to 100G, some via N*10G, via 40G or via 100G. At the other end of the scale some NRENs are only just now approaching 1G domestic connectivity. Internationally the bandwidths are even more diverse, and include some satellite links for particularly remote areas or specific locations such as educational institutions not otherwise served by their (emerging sometimes) NRENs. The table below broadly summarises the current situation across the region; it is missing a large number of economies, but as the compendium and mapping efforts increase and merge it will become more complete.

| Economy | Domestic | International |
|-------------|-------------------|-----------------------------------|
| Australia | n*100G + 10G | 2x2.5G to Asia, 2x40G (R&E) to US |
| Bangladesh | 1-10G | 45M |
| China | 10G-100G | Multiple 1G and 10G links |
| Hong Kong | 1-10G | Multiple 155M-10G |
| India | 1G-10G | 2.5G |
| Indonesia | 100M | 622M |
| Japan | Multiple <1G-100G | 1.5M(satellite) to multiple 10G |
| Korea | Multiple 10G | Multiple 10G |
| Sri Lanka | 1-500M | 45M-> 1G |
| Malaysia | 1G | 100M-622M |
| Nepal | | 45M |
| New Zealand | 1G-10G | 1G->40G |
| Philippines | 1G-10G | Multiple 155M-1G |
| Pakistan | 1G-n*10G | 155M |
| Singapore | 1G-10G | Multiple 155M-10G |
| Thailand | 1G | 310M-1G |
| Taiwan | 10+G->100G | Multiple 2.5-10G to Asia, US, EU |
| Vietnam | 30M-1G | 622M |

International, subsea capacities have been steadily increasing over the last twelve months, almost entirely due to improved signalling on existing cables. A number of proposed new cables are still in financial limbo and awaiting decisions to go ahead, but 100G single-wavelength links across the Pacific are now technically, if not yet financially, feasible for much of eastern Asia, including Japan, China, Korea, Singapore, Taiwan and Australia.

Annex 7: AmLight Project Status and Plan

(Links to Latin America)

Center for Internet Augmented Research and Assessment (CIARA)
at Florida International University

Submitted by Heidi Alvarez (heidi@fiu.edu) and Julio Ibarra (Julio@fiu.edu), FIU
January 2017

Americas Lightpaths (AmLight)

AmLight Express and Protect (AmLight-ExP)

The AmLight Express and Protect (ExP⁶) project implements a hybrid network strategy that combines optical spectrum (Express) and leased capacity (Protect) that builds a reliable, leading-edge diverse network infrastructure for research and education. Researchers will be able to leverage the resources of AmLight ExP to foster network innovation and to address increasing network services requirements between the U.S. and the nations in South America.

AmLight ExP (NSF ACI-1451018⁷) is a reliable, leading-edge infrastructure for research and education. The total bandwidth provided by AmLight ExP between the U.S. and South America is expected to grow to more than 680 Gibabits per second in aggregate capacity between 2015 and 2020. This serves as a flexible inter-regional infrastructure, enabling communities of scientists to expand their research, education, and learning activities uniquely empowered through access to unlit optical spectrum on submarine cables, and through AmLight ExP's use of dynamic circuits in a production environment.

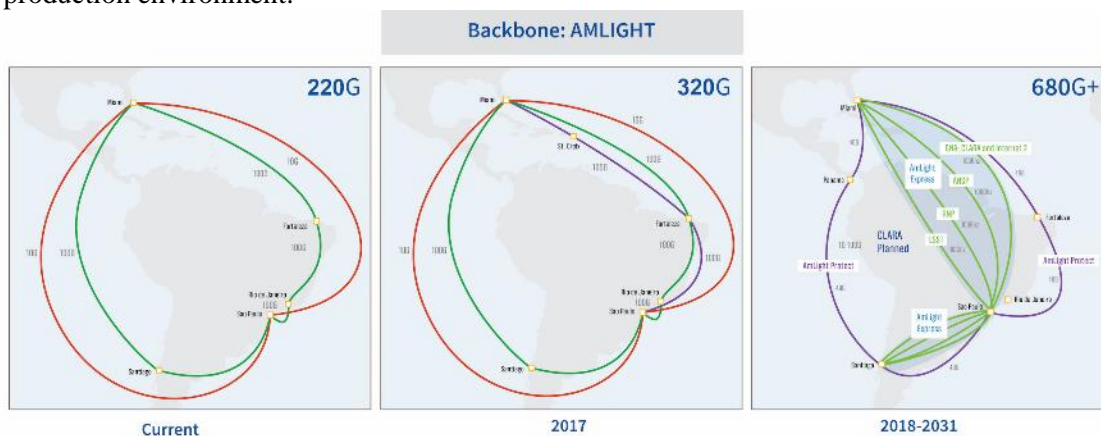


Figure 29 AmLight ExP network topology phases

The AmLight-ExP project builds upon the achievements of the highly successful Americas Lightpaths (ACI-0963053⁸) project, in cooperative partnership with the Latin American Cooperation of Advanced Networks (RedCLARA-Cooperación Latino Americana de Redes Avanzadas⁹), the Academic Network at São Paulo (ANSP¹⁰), the Rede Nacional de Ensino e

⁶ <http://amlight.net/>

⁷ NSF Award #1451018, IRNC: Backbone: AmLight Express and Protect (ExP): https://www.nsf.gov/awardsearch/showAward?AWD_ID=1451018&HistoricalAwards=false

⁸ NSF Award #0963053, IRNC-ProNet: Americas Lightpaths: Increasing the Rate of Discovery and Enhancing Education across the Americas: https://www.nsf.gov/awardsearch/showAward?AWD_ID=0963053&HistoricalAwards=false

⁹ RedCLARA develops and operates advanced Internet network in 13 Latin American countries: <http://www.redclara.net/index.php/en/>

¹⁰ ANSP provides connectivity to the top R&E institutions, facilities and researchers in the State of São Paulo: <http://www.ansp.br/index.php/us/>

Pesquisa (RNP¹¹), the Association of Universities for Research in Astronomy (AURA¹²), the National University Network - Chile (REUNA-Red Universitaria Nacional¹³) and in further cooperation with Florida LambdaRail (FLR¹⁴), Internet2 (I2¹⁵), as well as numerous other research and education networks, of national or global scope. AmLight-ExP is continuing to evolve a rational network infrastructure, designed to provide scalable multi-gigabit bandwidth and services, using both leased capacity and spectrum, supporting the needs of U.S.-Western Hemisphere research and education communities in a manner that fosters the evolving nature of discovery and scholarship.

The following major goals form the structure of the AmLight-ExP project:

- Implement an infrastructure that interconnects North America to key aggregation points in South and Central America (Brazil, Chile, Panama).
- Evolve the connections into a reliable, flexible and efficient infrastructure.
- Facilitate effective peering between NRENs and communities of interest through a distributed exchange model.
- Meet or exceed the requirements of the science drivers.
- Facilitate outreach to researchers and network operators.

AmLight serves as an open instrument for collaboration, interconnecting existing points of aggregation, and providing a means to leverage collaborative purchasing and network operation in order to effectively maximize the benefits of all investors, and to manage the NSF investment in the context of international partnerships.

Improvement of AmLight ExP network connectivity

New connectivity in South America

On July 2016, two 100 Gbps connections between São Paulo and Miami were activated, which expand the capacity to accommodate new applications and science drivers, such as LSST and LHC. The links, which go through submarine cables in the Atlantic and Pacific oceans, are maintained by the AmLight consortium.

It's worth mentioning that to evaluate the link, AmLight engineers employed a traffic generator in Miami. By using this approach, the engineers were able to generate more than 90Gbps of steady traffic, and with the use of OpenFlow flow entries, redirected such traffic to the international link. Thus, the equipment and circuits were evaluated in both directions (download and upload traffic). Figures 2 and 3 show the 100G links utilization during the process of validation.

¹¹ RNP operates the national research and education network and several networks in Brazil, providing access to around 400 institutions in the fields of Higher Education, Research, Health and Culture throughout the country: <https://www.rnp.br/en>

¹² (AURA) is a consortium of 42 US institutions and 5 international affiliates that operates world-class astronomical observatories: <http://www.aura-astronomy.org/>

¹³ National Network for Research and Education in Chile (NREN), and is currently made up of 34 institutions

¹⁴ FLR is the regional optical network of Florida, formed as a consortium of 12 universities

¹⁵ Internet2 is a consortium of leading US research universities working in partnership with industry and government to develop and deploy advanced network applications and technologies.

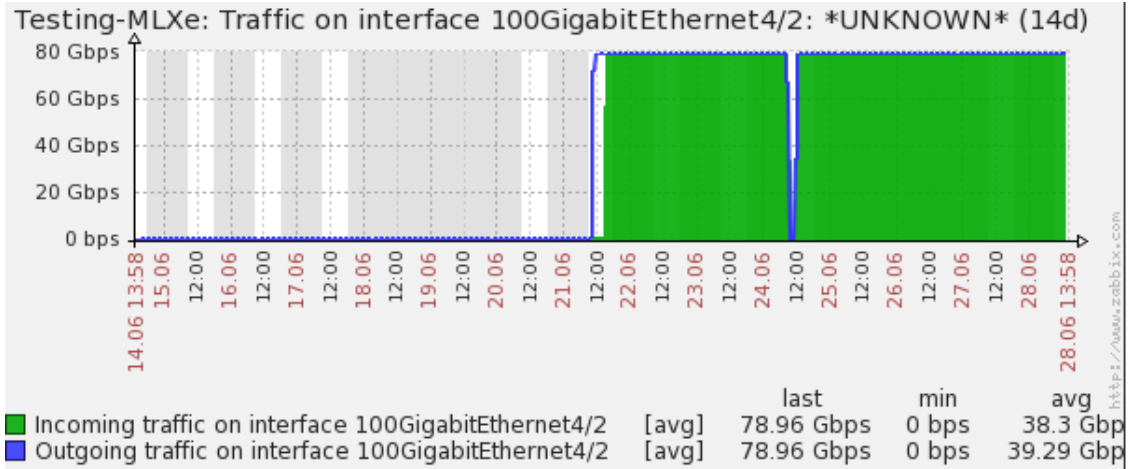


Figure 30 Pacific 100G Link being tested after its installation

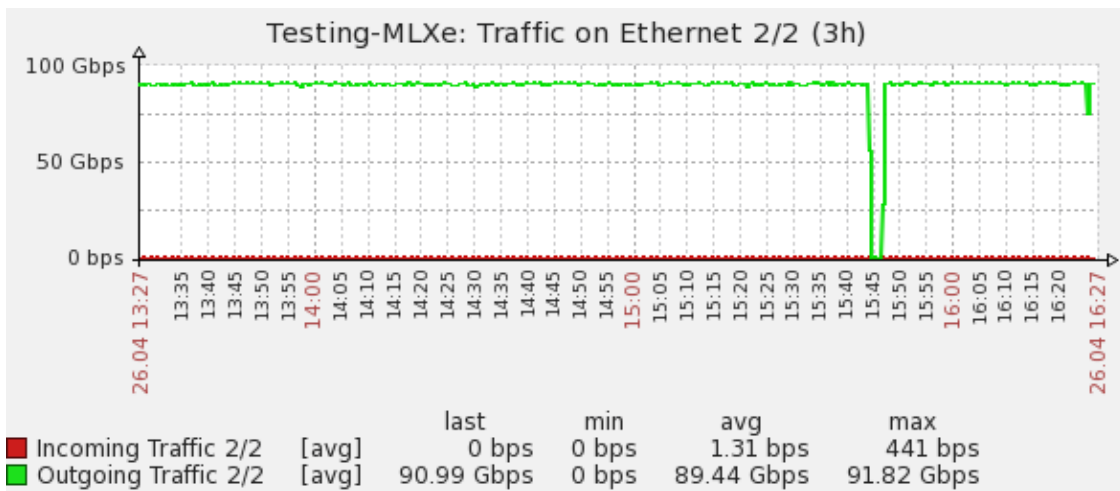


Figure 31 Testing the 100G link between Miami and Brazil

One of the main challenges for the activation of this high-performance infrastructure was cleaning the optical fiber cables in the land connections, since any vestige of dirt and oiliness in the interface between the fibers may melt with the heat propagated by data traffic, damaging the physical integrity of the optical fiber.

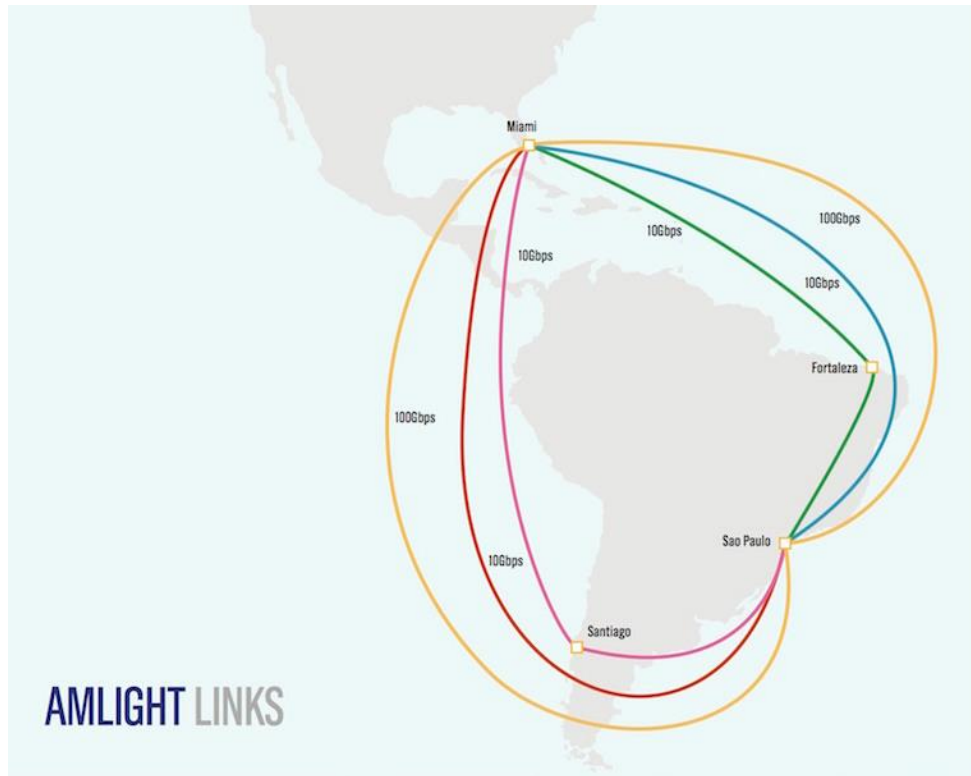


Figure 32 AmLight ExP Infrastructure 2016

The 100 Gbps international connections set new high-performance connectivity parameters in the Americas, and enable opportunities for scientific collaboration. One of the benefitted initiatives will be an international astronomy project, the Large Synoptic Survey Telescope (LSST¹⁶), which features the participation of 50 Brazilian researchers. The LSST is a telescope under construction in Cerro Pachón, in Chile, and is expected to enter into operation in 2022. It shall be able to map almost half of the sky for a ten-year period.

AmLight supports network virtualization

AmLight uses a hybrid approach, comprised of legacy (VLAN) and SDN/OpenFlow.

The main benefits from the 2014 SDN implementation are network virtualization and programmability.

¹⁶ The goal of the Large Synoptic Survey Telescope (LSST) project is to conduct a 10-year survey of the sky that will deliver a 200 petabyte set of images and data products. <https://www.lsst.org/about>

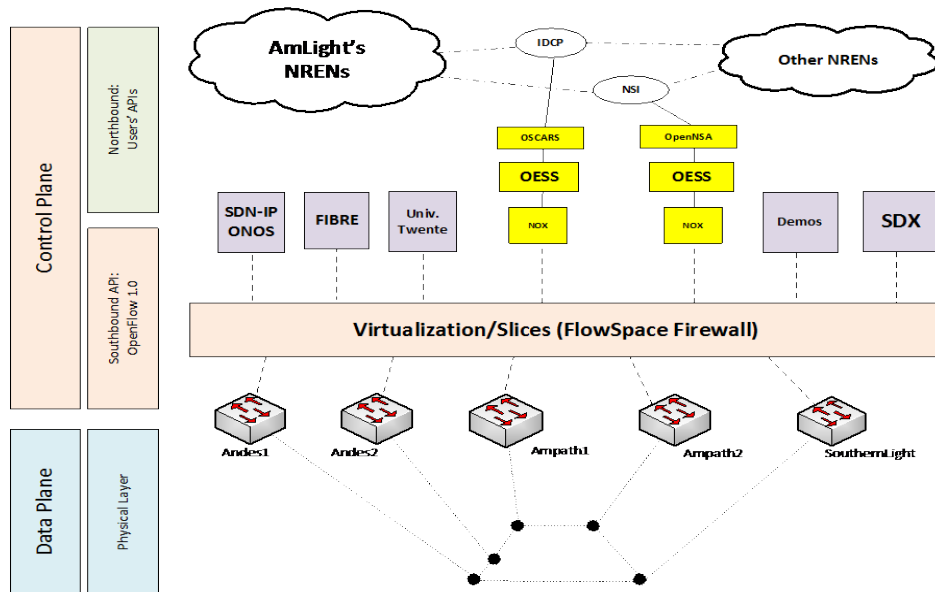


Figure 33 AmLight SDN stack

Figure 5 represents the current SDN stack at AmLight after deploying OpenFlow 1.0, Flow Space Firewall (FSFW¹⁷) and SDN applications. On the bottom are the OpenFlow devices and links connecting them. The dashed lines between devices and the Flow Space Firewall and, between Flow Space Firewall and SDN applications (represented by purple boxes) represent the OpenFlow sessions established. In this configuration in the SDN stack, FSFW acts as a proxy between the physical layer (represented by OpenFlow devices and links) and the control layer, represented by SDN applications.

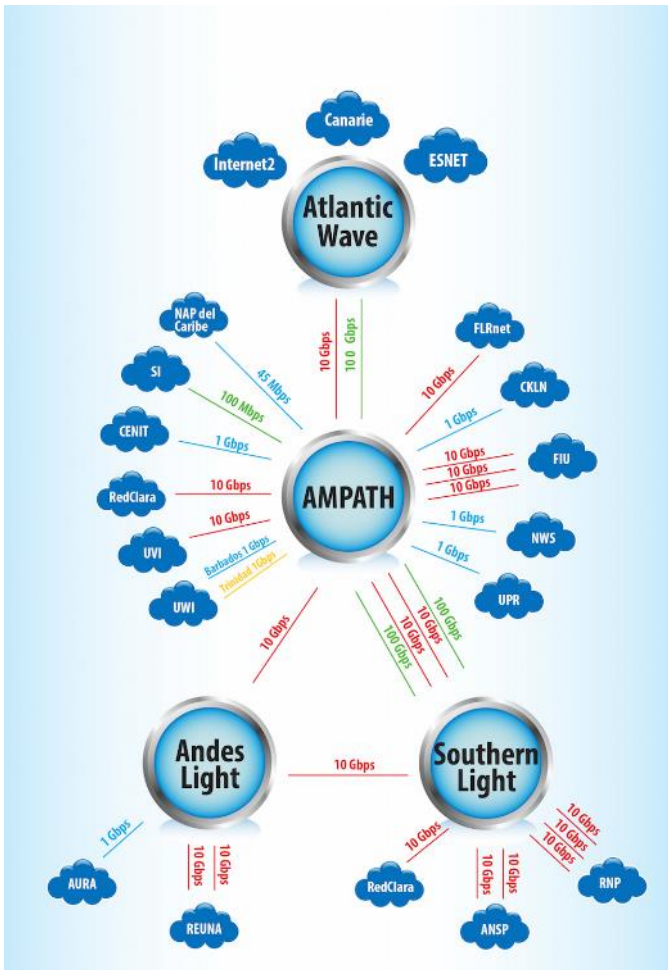
To describe how AmLight supports network virtualization, we refer to Figure 5: the FSFW manages what each SDN application can do to OpenFlow devices. It is important to observe that from the perspective of the data plane, all flows are handled in the same way. OpenFlow switches are not aware of multiple controllers, and all flow entries are inserted in the same table, as part of the same data plane. In this case, because SDN applications can send OpenFlow messages to the OpenFlow switches (assuming they were allowed by the FSFW), the OpenFlow agent inside each switch is responsible for interpreting those messages and reacting in the proper way (sending an error, installing the flow, sending a reply, etc.).

AmLight hosted six network testbeds, where two were production testbeds and four were experimental. AmLight's users or engineers used the experimental network testbeds to try different applications and forwarding approaches. Some details are provided below:

- In partnership with RNP, a FIBRE (*Future Internet testbeds / experimentation between BRazil and Europe*¹⁸) island was installed at AMPATH and AmLight was used to connect to other islands in Brazil. The idea of this testbed was to understand what kind of requirements overlay SDN applications have and what should be changed on production transport networks to support them.

¹⁷ FlowSpace Firewall Application a floodlight based controller allowing multiple controllers to talk to a single switch, but can not interact with each others flow space (hence FlowSpace Firewall) <http://globalnoc.iu.edu/sdn/fsfw.html>

¹⁸ The FIBRE testbed is as research facility constructed in the scope of a former project funded by the 2010 Brazil-EU Coordinated Call in ICT: <http://fibre.org.br/about/what-is-fibre/>



- In partnership with On.Lab¹⁹, the Open Network Operational System – ONOS²⁰ – and its application SDN-IP were installed at AmLight. The idea behind this testbed was to understand how OpenFlow could be used for IP traffic.
- In partnership with Internet2, a network testbed was hosted to demonstrate five different SDN islands interconnected using BGP and ONOS/SDN-IP application. This demonstration happened during the Internet2 Global Summit 2015.
- In partnership with University of Twente, an assessment of some OpenFlow devices was made to understand how OpenFlow could be used as a measurement protocol. Details about the tests were reported on the “Assessing the Quality of Flow Measurements from OpenFlow Devices” paper²¹, which was accepted for the Traffic Monitoring and Analysis workshop 2016.

Future work

AmLight plans to keep testing different SDN network devices to find a solution that works with OpenFlow 1.3+, and fulfills some requirements, such as buffer size and the number of 100G ports.

In addition, AutoGOLE²² will foster effective peering with international and domestic R&E networks. Activities involve continuing participation in the AutoGOLE project, and continuing to enhance peering globally by using and contributing to the improvement of the ONOS SDN-IP controller.

AmLight, AMPATH and SouthernLight participate in the AutoGOLE working group. AmLight will facilitate the use of SDXs to improve the provisioning process for the AutoGOLE community.

AMPATH International Exchange Point

As a facility, AMPATH²³ located in Miami, Florida, is the premiere International Exchange Point (IXP) serving network-enabled U.S.-Latin America and Caribbean science research and education communities. AMPATH provides wide bandwidth network services for U.S. and international research and education networks to extend participation to underrepresented groups in Latin America and the Caribbean.

Figure 34 AMPATH Connection Map

¹⁹ Open Networking Lab (ON.Lab) is a nonprofit organization dedicated to developing tools and platforms and building open source communities to realize the full potential of SDN: <http://onlab.us/>

²⁰ The Open Network Operating System (ONOS) is a software defined networking (SDN): <http://onosproject.org/>

²¹ Hendriks, L.; Schmidt, R.; Sadre, R.; Bezerra, J.; Pras, A.. Assessing the Quality of Flow Measurements from OpenFlow Devices. 8th International Workshop on Traffic Monitoring and Analysis. Belgium, 2016: <http://amlight.net/wp-content/uploads/2015/04/tma2016-final34.pdf>

²² AutoGOLE (Auto- GLIF Open Lightpath Exchange): <https://www.glif.is/>

²³ <http://www.ampath.net/>

AMPATH operates as a major research facility recognized by the U.S. NSF, supporting international e-science.

AMPATH provides its international connectors with access to U.S. production and experimental backbone networks, such as Internet2, ESnet, Florida LambdaRail, etc., to facilitate international science research and education collaborations.

AMPATH provides a scalable, redundant Ethernet switching fabric to all users. Standard interface configurations include support for jumbo frames. AMPATH offers collocation space for connectors that would like to house their equipment in Miami, in its facility at the NAP of the Americas. International connectors are able to peer and exchange traffic with U.S. national R&E backbone networks, and international R&E networks using dedicated or non-dedicated bandwidth, with support for IPv4 and IPv6 unicast and multicast. High-availability is provided through multiple diverse paths using layer2 vlans and layer3 routed connectivity.

AMPATH operates as an open exchange peering fabric, facilitating its connectors to establish peering connections with networks connected at the exchange point. AMPATH extends its peering fabric through the AtlanticWave²⁴ and implements best practices for international exchange points. AMPATH is a participant of the Global Lambda Integrated Facility (GLIF) as an Open Lightpath Exchange (GOLE), and supports hybrid services and dynamic provisioning through NSI and OSCARS protocols. Also, AMPATH fully supports OpenFlow, offering more possibilities to users and researchers.

Atlantic Wave-SDX

AtlanticWave-SDX description

Demand is growing to develop the capability to support end-to-end services, capable of spanning multiple Software Defined Networking (SDN) domains. SDN deployments that cross multiple domains continue to be constructed manually, involving significant coordination and effort by network operators. Moreover, the demand for more intelligent network services to support the evolving science research and education activities between the U.S. and South America are increasing; these network services, which include dynamic provisioning of end-to-end multi-domain layer2 circuits, and network programmability, are needed to foster innovation for application developers, and to increase efficiency for network operators. AtlanticWave-SDX is a response to the demand for more intelligent network services to foster innovation and to increase network efficiency.

Florida International University (FIU) and the Georgia Institute of Technology (GT) are implementing AtlanticWave-SDX²⁵: A distributed experimental Software-Defined Exchange (SDX), supporting research, experimental deployments, prototyping and interoperability testing, on national and international scales. A Software-Defined Exchange (SDX) will provide a capability to prototype an OpenFlow network where members of each Internet peering fabric could exchange traffic based in different layers of abstraction.

AtlanticWave-SDX (NSF ACI-1451024²⁶) is comprised of two components: (1) a network infrastructure development component to bridge 100G of network capacity between Research and Education (R&E) backbone networks in the U.S. and South America; and (2) an innovation component to build a distributed intercontinental experimental SDX between the U.S. and South

²⁴ AtlanticWave is an international peering fabric interconnecting: US, Canada, Europe, and South America. With distributed IP peering points in New York, Washington D.C., Atlanta, Miami, and Sao Paulo: <http://atlanticwave.net/>

²⁵ <http://www.atlanticwave-sdx.net/>

²⁶ NSF Award #1451024, IRNC: RXP: AtlanticWave-Software Defined Exchange: A Distributed Intercontinental Experimental Software Defined Exchange (SDX): https://www.nsf.gov/awardsearch/showAward?AWD_ID=1451024&HistoricalAwards=false

America, by leveraging open exchange point resources at SoX (Atlanta), AMPATH (Miami), and Southern Light (São Paulo, Brazil).

Goals

The following four major goals form the structure of the AtlanticWave-SDX project:

Building a distributed experimental SDX between the U.S. and S. America: AtlanticWave-SDX couples the R&D expertise on Software Defined Networking (SDN) and SDX at GT, with the open exchange point and network operations capabilities at AMPATH and SoX to prototype and operate a distributed experimental SDX. The strategy in AtlanticWave-SDX includes standing up SDXs at AMPATH and SoX, then extending the experimental SDX to the Southern Light Open Exchange Point in São Paulo, Brazil, in collaboration with ANSP and RNP.

Enhancing a platform of network innovation capabilities at AMPATH and SoX: Multiple international R&E backbone networks from S. America terminate their connections at AMPATH with support from the IRNC AmLight project. The AmLight EXP proposal in the IRNC Backbone category requires spectrum as a service at AMPATH to support greater capacity connections from S. America. In collaboration with Florida LambdaRail (FLR), AtlanticWave-SDX proposes a design and implementation plan to support spectrum services and to integrate them into a distributed open SDX fabric.

Leveraging network infrastructure between the U.S. and S. America: AtlanticWave-SDX proposes a highly leveraged strategy, building upon the infrastructure established by the current and previous IRNC awards (ACI-0963053 , OCI-0441095) to FIU for U.S.-Latin America connectivity; and leveraging the SDN and GENI racks resources at GT and FIU.

Providing leadership and coordination in the experimental SDX community: AtlanticWave-SDX will establish a unique proving ground for the design and operation of SDX-based peering architectures on an international scale. The team will work with operators and researchers, including other IRNC projects, to solicit input and cooperation on these designs and make the resulting software tools available.

AtlanticWave-SDX Controller Overview

A Software-Defined Internet Exchange Point (SDX) is a new technology that has many, often conflicting definitions. The AtlanticWave-SDX team hopes to cement the definition to refer to two specific aspects: first, a SDN-based dataplane that connects various networks together, and second, participant networks are able to define forwarding policy within the exchange point by way of a configuration API. With such an API, participants can make forwarding decisions based on more information than simply destination IP address, as is the case with traditional Internet Exchange Points (IXPs).

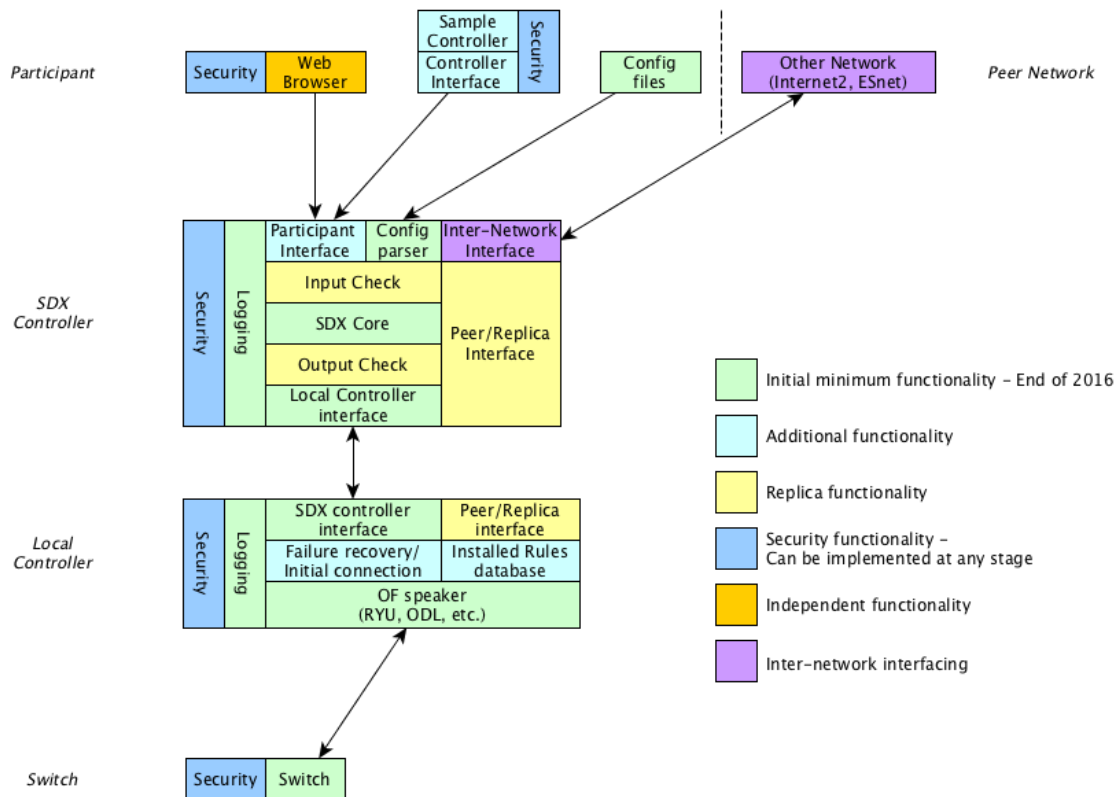


Figure 35 AtlanticWave-SDX Controller Design

Figure 7 shows the AtlanticWave-SDX Controller design. The AtlanticWave-SDX controller is still under development. It will support research, experimental deployments, prototyping and interoperability testing, on national and international scales. Two demonstrations were made, one during the I2 Tech Exchange Conference in Miami in September 2016, and an updated version was presented at the Super Computing (SC16) Conference in Salt Lake City in November 2016. The SDX controller source code is available at GitHub²⁷.

Further, a Distributed SDX is a new style of SDX, wherein participant networks can connect to the SDX at multiple locations, to multiple switches, potentially owned and operated by multiple organizations in different countries as shown on the Figure 8 below.

²⁷ <https://github.com/atlanticwave-sdx/atlanticwave-proto>

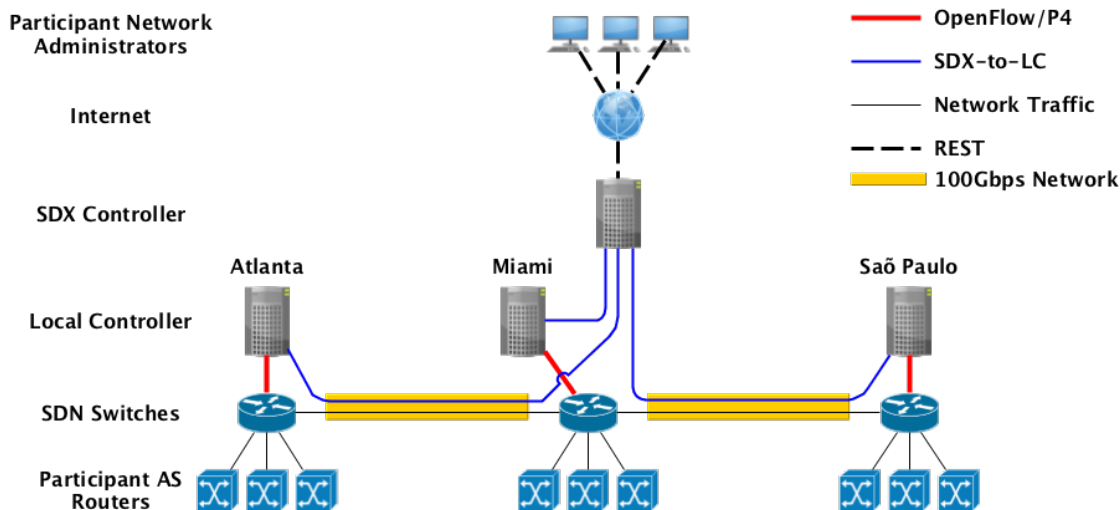


Figure 36 AtlanticWave-SDX Controller Architecture

Challenges of developing an SDX controller to domain scientists and network engineers

Having a distributed SDX leads to a number of challenges. Having centralized control of a dataplane in a datacenter is reasonable, while having centralized control across states, or even countries leads to challenges with control distribution, recovery, latency of control signal propagation, in-band vs. out-of-band control, etc. Having different organizations manage different sites can lead to other challenges, including handling heterogeneous switching hardware with different capabilities. Finally, with different hardware, different control protocols are possible.

To these ends, Georgia Tech and FIU are creating a distributed SDX, spanning Atlanta, Miami, and São Paulo as part of the AtlanticWave-SDX IRNC project. Beyond the challenges associated with a distributed SDX, we have found two challenges that drive our research agenda: (1) Available 100Gb OpenFlow-controllable hardware that supports all OpenFlow 1.3 functionality on any match-action table. This provides the opportunity to discover alternative network designs in order to provide better functionality to participants; and (2) Defining an API for participants to use, determining how much abstraction is necessary to mask any heterogeneity of hardware resources, while providing a level of functionality that is not present in traditional IXPs.

Supercomputing Conference 2016 (SC16) November 13-18

During the Super Computing 2016, network engineers from AmLight, Academic Network of São Paulo (ANSP), State University of São Paulo (UNESP), Florida LambdaRail and California Institute of Technology leveraged the AmLight-ExP network to demonstrate the new 200G bandwidth capacity between the U.S. and South America. A total of 176Gbps of aggregated traffic was measured between São Paulo, Miami and Salt Lake City/ Utah, where the Super Computing 2016 was hosted. The Figure 9 below shows a sustained flow rate of ~85Gbps, with peaks of 97Gbps, generated using the [FDT software](http://monalisa.caltech.edu/FDT/)²⁸.

²⁸ Fast Data Transfer (FDT) is an Application for Efficient Data Transfers, which is capable of reading and writing at disk speed over wide area networks (with standard TCP). It is written in Java, runs on all major platforms and it is easy to use. <http://monalisa.caltech.edu/FDT/>

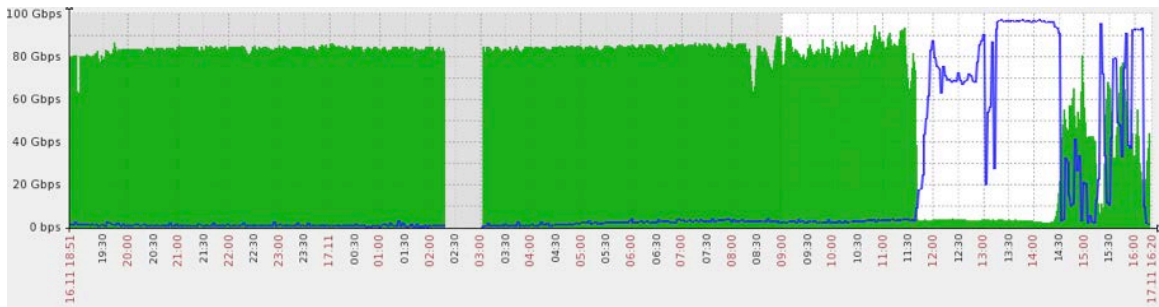


Figure 37 Demonstrated traffic in the Atlantic 100Gbps link

For this demonstration, two high-end servers were installed in São Paulo (one at São Paulo Research and Analysis Center (SPRACE)/UNESP and one at ANSP) and one server was installed at AMPATH/NAP of the Americas, Miami, FL. An addition, Caltech provided a few 100G servers.

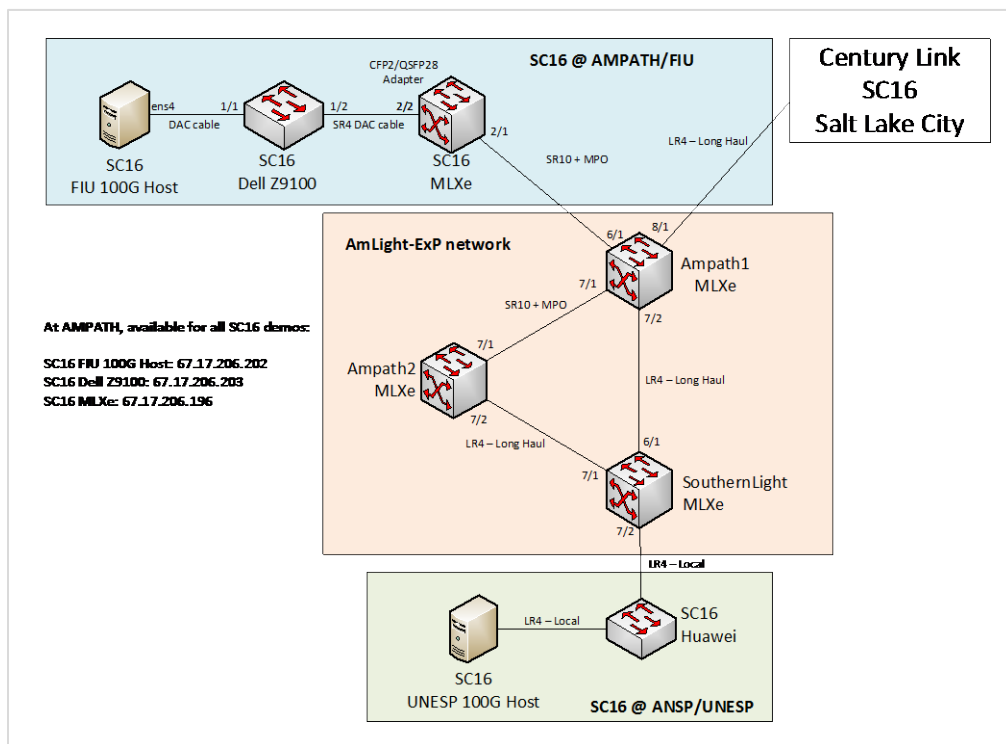


Figure 38 Network Infrastructure deployed in Miami (AMPATH & AmLight)

Figure 39 shows the network infrastructure used for the demonstration at SC16. FIU, ANSP, and UNESP teams demonstrated an end-to-end transmission from Latin America to the U.S. using 100G links, showcased for the first time using the network infrastructure provided by the AmLight-ExP consortium.

Americas Africa Research and eduCation Lightpaths (AARCLight)

Florida International University and AmLight consortium partners are planning, designing, and defining a strategy for high capacity connectivity research and education network connectivity between the US and West Africa, called AARCLight: Americas Africa Research and eduCation Lightpaths (NSF ACI-1638990²⁹). Science is being conducted in an era of information abundance. Sharing science resources, such as data, instrumentation, technology, and best practices, across

²⁹ NSF Award #1638990, IRNC: Backbone: Americas Africa Research and eduCation Lightpaths (AARCLight): https://www.nsf.gov/awardsearch/showAward?AWD_ID=1638990&HistoricalAwards=false

national borders, can promote expanded scientific inquiry, and has the potential to advance discovery. Linking the U.S. and the nations of Africa's researcher and education communities is an increasingly strategic priority. Africa offers research and education communities with unique biological, environmental, geological, anthropological, and cultural resources. Research challenges in atmospheric and geosciences, materials sciences, tropical diseases, biology, astronomy, and other disciplines will benefit by enhancing the technological and social connections between the research and education communities of the US and Africa.

The planning project is largely based on the planned availability of submarine cable spectrum for use by research and education communities. It creates an unprecedented opportunity for the stakeholders in the U.S., Africa, and Brazil to coordinate planning efforts to strategically make use of the offered spectrum towards serving the broadest communities of interest in research and education.

High Energy Physics Labs

Annex 8: CERN Network Status and Plan

Edoardo Martelli (Edoardo.Martelli@cern.ch)

January 2017

CERN Computer Network

The CERN Computer Network interconnects all the network capable computing devices used at CERN: desktop PCs, server farms, wireless devices, down to the LHC control devices and sensors, and the detector data acquisition systems. IT/CS is the group in the CERN IT department that takes care of the engineering, deployment and operations of the whole network.

The CERN Computer Network provides IPv4 and IPv6 connectivity over 1,10,40,100Gbps Ethernet.

The bandwidth demand has increased to unprecedented levels during 2016, thanks to the exceptional performance of the LHC. The four LHC Experiments have produced more data than foreseen, which could still be transported without any problem from the pits to the CERN datacentres and then forwarded to the WLCG sites for storage and analysis.

To cope with the increasing volume of traffic, the IT department and the LHC experiments have kept improving the capacity of their networks. The LHC Experiments have upgraded their links from the Data Acquisition farms to the IT data centre in Geneva, to cope with the larger streams of data that the detectors generate. The bandwidth is now 120G for Atlas, 160G for CMS, 80G for Alice and 20G for LHCb, for a total of 380Gbps.

2016 has seen the completion of the Wigner datacentre (Budapest, HU). This datacentre is an extension of the datacentre in Geneva; the two centres are interconnected with two 100Gbps links and MPLS networking. An additional 100Gbps link was ordered at the end of 2015, but due to some tendering issue on the local loop, it hasn't been delivered yet. The problem has been solved and the installation has already started and will be completed in January 2017.

Connectivity to European HEP sites has been improved with the doubling of the LHCOPN links to five WLCG Tier1s.

Architecture

The CERN network is divided in several domains, each of them dedicated to a specific purpose. The overall architecture and the different domains are depicted in Figure 40.

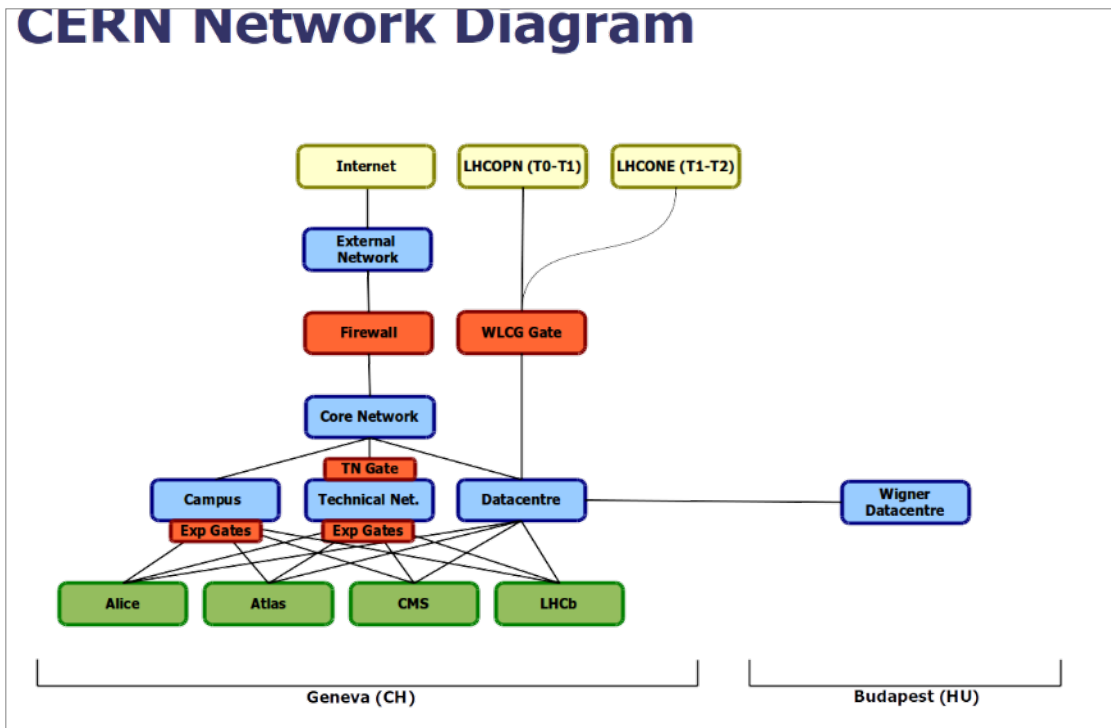


Figure 39: CERN network domains

During 2016 the architecture of the datacentre has been changed in order to remove bottlenecks and optimize resources. Before it consisted of two domains (LCG for physics data and GPN for generic services), implemented with two different physical networks. In 2016 the two physical networks have been merged in one and the two domains implemented with Virtual Routing instances. With the delivery of the 3rd 100Gbps link Geneva-Budapest, also the Wigner datacentre will be integrated in the unified infrastructure.

The LHC Computing Grid network (LCG) domain serves the LCG Tier0 servers farm (CPUs, disks, tapes) in the main computer centre. It is directly connected with the Data Acquisition systems of the LHC Experiments and also to the eleven Tier1 centres via the LHCOPN and to the largest Tier2s and Tier3s via the LHCONE.

The Campus Network (or General Purpose Network, GPN) domain provides connectivity to the generic users and services. It interconnects all the campus offices and control centres, the WIFI access points and most of the generic IP devices.

The Technical Network domain connects all the control devices of the LHC (access control, cryogenic system control, etc.). It's a highly secured network and it's separated from all the other areas by dedicated firewalls. The Technical Network is also connected with the control devices of all the LHC Experiments.

The four LHC Experiments (ALICE, ATLAS, CMS and LHCb) have each one a dedicated network for their control and data acquisition devices.

The Firewall area is the main gateway to the Internet.

The External Network area connects to the major Research and Education backbones of Europe and North America; it also connects to the generic Internet.

Bandwidth capacity

The capacity of the datacentre backbone is 9.6 Tbps non-blocking and it may double during 2017-2018, if budget allows it. The LCG servers farm is connected to the Tier1s with aggregated capacity of 300Gbps. The connection to LHCONE is 200Gbps.

The Campus backbone has a non-blocking capacity of 3.2 Tbps.

The primary firewall system consists of two systems capable of passing 40Gbps traffic each. One system serves the LCG domain, the other one the rest of CERN. Each one can backup the other.

External connectivity to Research and Education networks has an aggregated capacity of 180Gbps and is provided by ESnet, GEANT, NORDUnet, RENATER, SURFnet, SWITCH.

External connectivity to the general purpose Internet has an aggregated capacity of 30Gbps and is provided by three commercial ISPs and peerings at CIXP (CERN Internet Exchange Point).

Datacentre

CERN has launched a market survey for the supply of the next generation high end switch-routers. The switch-routers are requested to provide a large number of 100Gbps Ethernet ports and be ready to support 400Gbps Ethernet ports. Among other requirements, the selected equipment must be able to provide an open standard Ethernet fabric.

The tender process will be completed before the end of 2017; the intent is to use the selected equipment to implement the doubling of the datacentre network switching capacity.

IPv6

The deployment of IPv6 connectivity was completed in 2013 and has been opened to all the Campus and Datacentre users in 2014. IPv6 adoption is increasing thanks to several initiatives taken to push IPv6 compatibility of the WLCG applications and services.

Security

All the different areas are secured with Access Control Lists (ACLs) applied at the interconnecting routers. Those ACLs are automatically generated and deployed by the CERN Network Management System, according to the policies provided by the CERN Security Team and the connectivity requirements of the users.

The Technical Network and the Experiments are not reachable from the Internet.

All the plugs to the network are secured, with checks made of the credentials of the machines and of the users that try to connect.

WIFI

CERN has awarded to Aruba HPE a contract to provide the next generation WIFI equipment that will be used to extend the WIFI coverage and advanced features to the whole CERN campus. The deployment has already started with a pilot covering the IT buildings. The full coverage should be achieved within two years.

Business continuity

The CERN IT department has been working on the implementation of a business-continuity project. The intent is to reduce the impact of a potential major disastrous event that could impair for a long time its main datacentre in Geneva, where all the network connections and equipment are located. CERN is building a small network hub building in its French site, few kilometres away from the main datacentre. The network hub building will be completed in the third quarter 2017. Part of the most critical network connections and equipment will be moved to the new building soon afterwards.

SDN

CERN collaborates with Brocade Networks in a Openlab project on openflow. The project aims to enhance the Brocade Flow Optimizer framework to specific support CERN use cases. The case currently being developed is a system that can intelligently mirror the traffic of the CERN central firewall to an IDS, without overloading it with the large LHC data transfers.

CERN has also started exploring the capabilities of white box switches, with the intent of standardizing its network provisioning framework to an open Network Operating System.

Annex 9: Fermilab Status and Plan

Submitted by Phil DeMar (demar@fnal.gov)

February, 2017

In 2016, Fermilab's major production network initiatives focused on completing the migration of the Laboratory's perimeter infrastructure over to 100GE technology, as well as providing incremental upgrade of the Laboratory's data center network capacity with additional 100GE links. In the area of network research, the high performance Multicore-aware Data Transfer Middleware file transfer tool (mdtmFTP) was rolled into evaluation phase by a number of science disciplines. In addition, a Laboratory-developed packet capture & GPU-based analysis system was enhanced to provide deep packet inspection capabilities.

Network Perimeter (WAN) Upgrades:

Since 2015, the Laboratory has supported two 100GE network connections to ESnet. However, in the initial configuration, the links operated in primary and failover mode, effectively providing a 100gb/s of off-site bandwidth for the Laboratory's science data movement. The 100GE WAN connections were dedicated to HEP-specific overlay networks such as the LHCOPN and LHCONE, as well as WAN-related network R&D activities. In addition, general internet traffic was routed via a 2x10GE connection to ESnet. A significant portion of the network traffic using the general internet path was science data movement to/from sites that didn't have connectivity to HEP-specific overlay networks such as the LHCOPN and LHCONE. This configuration provided the Laboratory with an effective off-site bandwidth capacity of 120Gb/s.

In 2016, the Laboratory completed the migration of its WAN services over to its 100GE-based infrastructure. The result is both 100GE WAN links are now fully utilized to carry all types of production Laboratory network traffic. Science data movement to special purpose overlay networks is logically isolated from general internet traffic by Virtual Routing & Forwarding (VRF) technology. R&D traffic for network research projects is similarly isolated on a separate VRF.

Finally, the internal network links to the perimeter infrastructure for both the CMS Tier-1 facility and the network research test bed have been upgraded to 100GE as well. Figure 1 shows the current Laboratory network perimeter configuration.

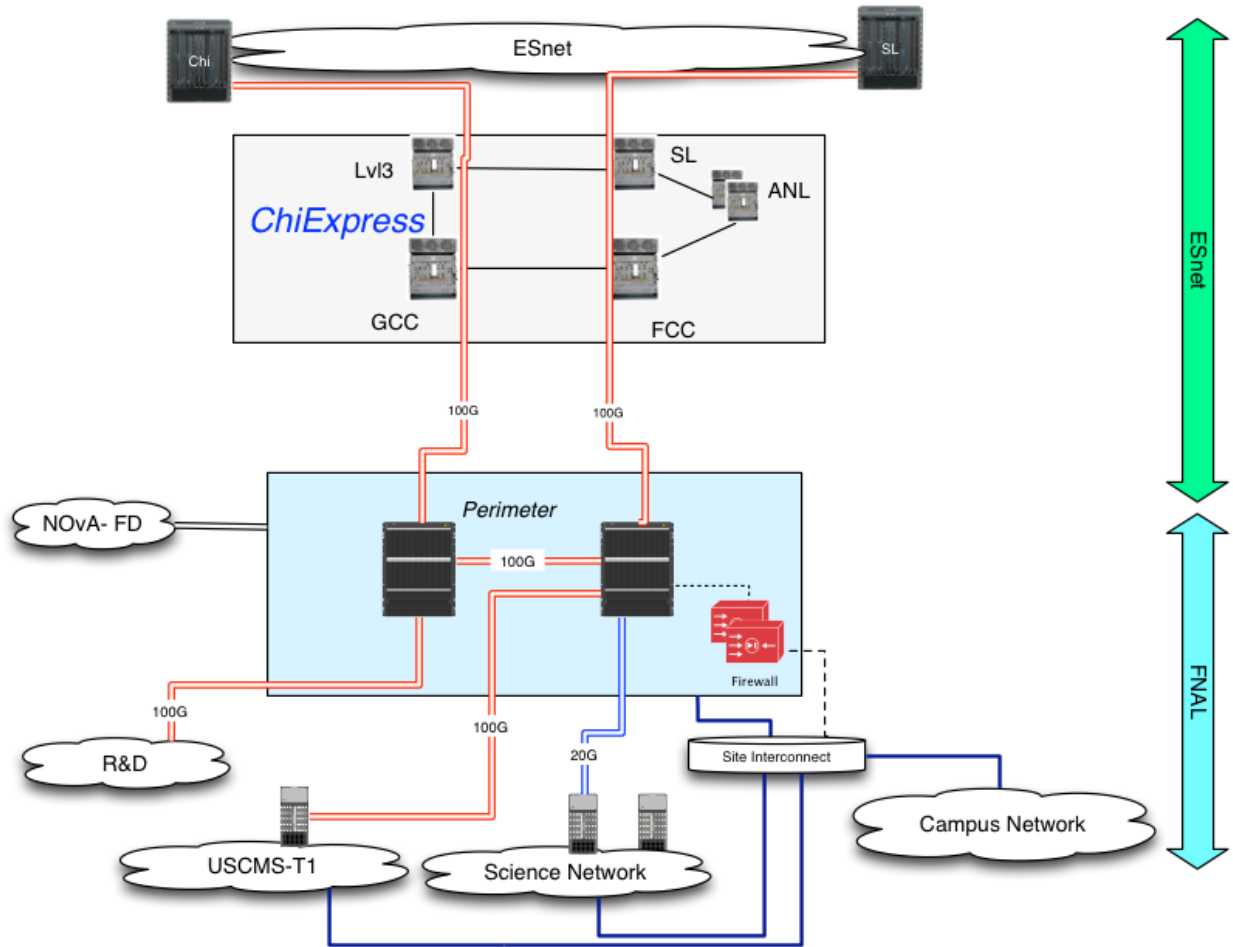


Figure 40: 2016 FNAL Network Perimeter Configuration

Data Center Networks:

The Laboratory has two main data centers, each supporting multiple computer rooms. The US-CMS Tier-1 Center has its own network infrastructure, with computer systems distributed across four different computer rooms, spread across the two data centers. The general Laboratory data center network is similarly spread across the same four computer rooms. Network equipment and links are upgraded on an annual basis to keep up with switch port and network bandwidth demands.

In 2016, the CMS Tier-1 network received major network upgrades, both in terms of equipment and network bandwidth. Two new Cisco Nexus 7010s were deployed as central aggregation devices for the Tier-1 network. The two aggregation switches are interconnected at 4 x 100GE. Existing distribution switches in the four computer rooms are connected to the aggregation devices at 100GE, or in several cases by 'n' x 10GE connections. The latter situations arise with legacy network switches that can't support 100GE in a cost-effective way. These devices will be phased out over the next several years, and replaced with 100GE-capable infrastructure. The number of 10GE-connected systems within the CMS Tier-1 has risen to ~250. Figure 2 depicts the current configuration of the CMS Tier-1 network.

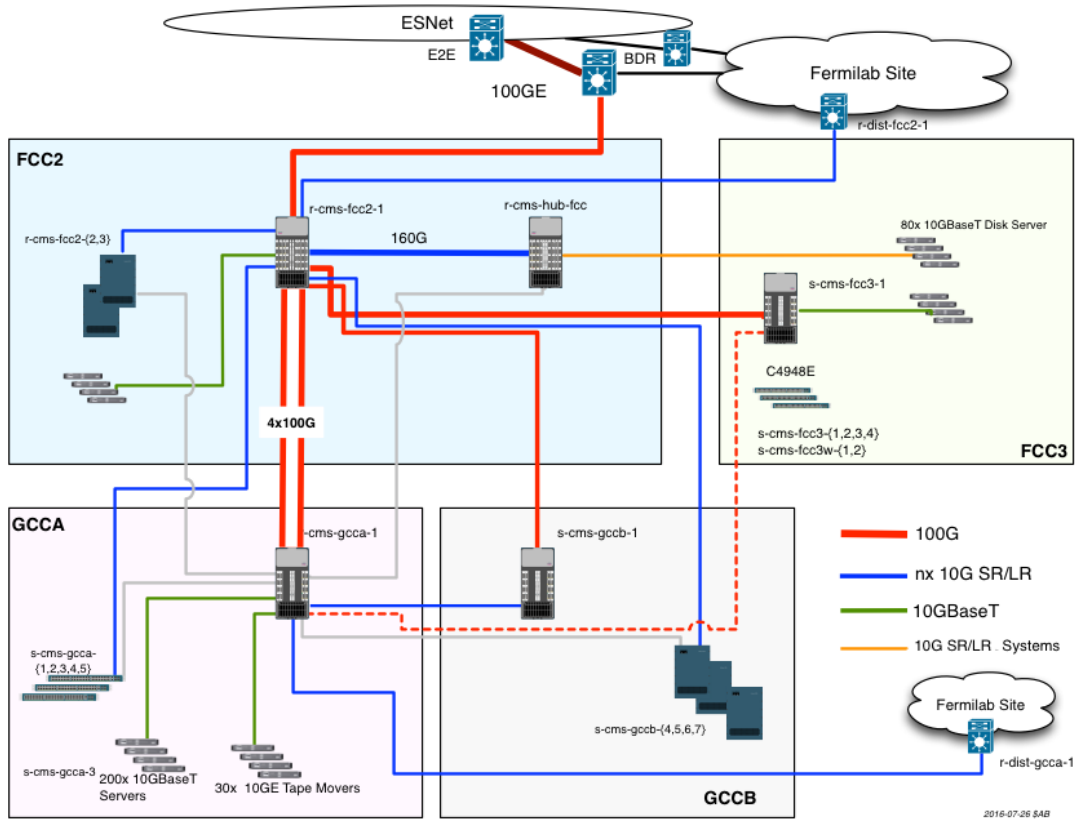


Figure 41: 2016 CMS Tier-1 Network (at FNAL)

In 2016, the general data center network was similarly upgraded with new aggregation devices, and additional 100GE links. Two Cisco Nexus 7710s were deployed as central aggregation devices, one in each of the Laboratory’s two data centers. The aggregation devices are interconnected at 2 x 100GE. Cisco Fabric Path provides a layer-2 fabric between the two new aggregation devices and legacy aggregation devices in the other computer rooms. The layer-2 fabric path enables relocation of systems between computer rooms without necessitating readdressing to a different subnet. WAN access from the general data center network is via the site’s core network infrastructure, and currently is 4 x 10GE. WAN access from the general data center network to discipline-specific overlay networks such as LHCONE and point-to-point circuits is via ‘n’ x 10GE direct (bypass) connections to the Laboratory’s perimeter network infrastructure. The bypass links are expected to be upgraded to 100GE in the coming year. Figure 3 displays the current configuration of the general data center network.

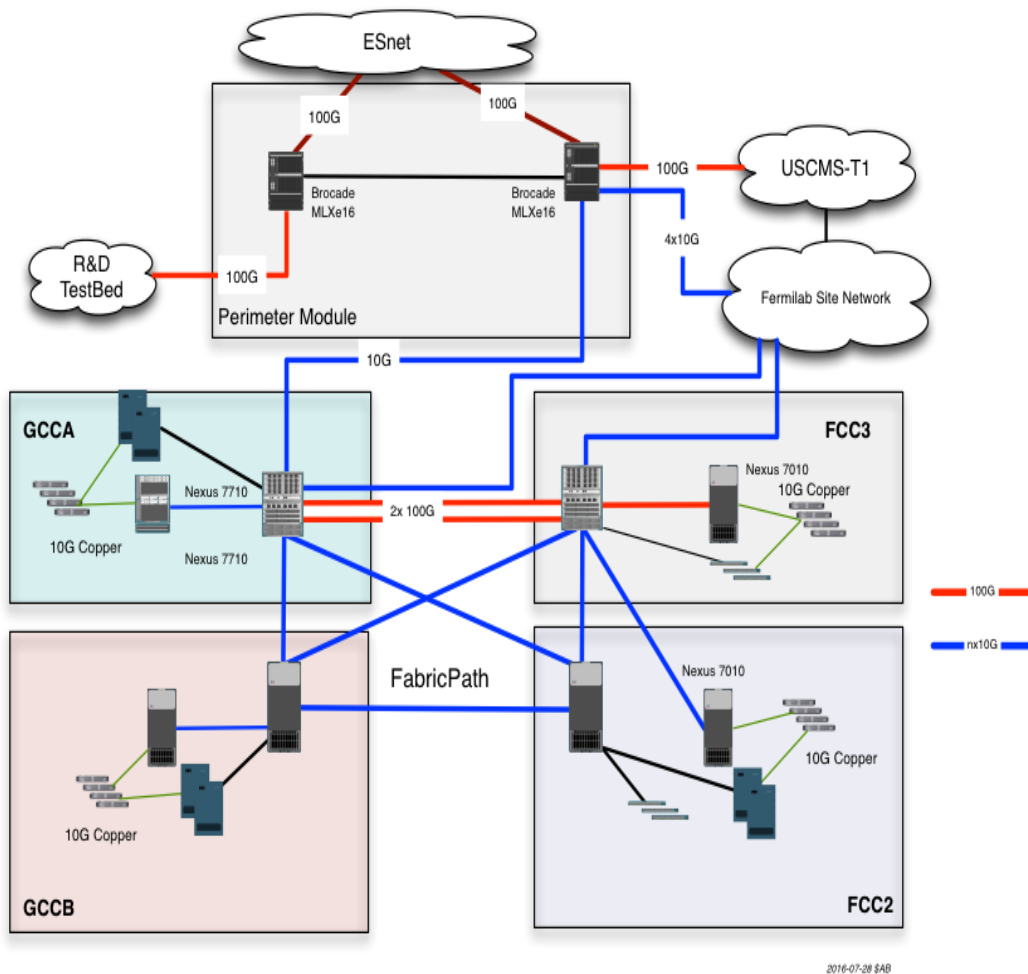


Figure 42: 2016 FNAL General Data Center Network

Network Research Activities:

In 2016, the Fermilab-led Multicore-Aware Data Transfer Middleware (MDTM) project successfully produced a release version of its high performance data transfer tool *mdtmFTP*. MDTM utilizes a pipelined I/O-centric design to optimize scheduling of data transfer network & disk I/O threads on multicore platforms. The *mdtmFTP* data transfer tool is particularly efficient on high performance systems with 40GE or 100GE NICs. It also implements a large virtual file mechanism to deal with Lots of Small Files (LoSF) data transfer scenarios, significantly exceeding the LoSF transfer performance of the current generation of file transfer tools. MDTM is undergoing evaluation within the Fusion community, with deployments in Korea, Singapore, and a number of US locations.

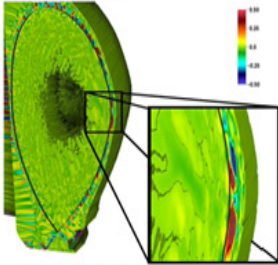
The MDTM development team has also prototyped and demonstrated a real-time scientific data streaming capability in *mdtmFTP*. At Super Computing (2016), FNAL, ORNL, and the Singapore A*Star Computation Resource Center used the ADIOS data management middleware and *mdtmFTP* to demonstrate a distributed data processing workflow, involving real-time analysis at FNAL of a Fusion simulation being conducted at Singapore. Figure 4 shows the remote analysis as demonstrated at Super Computing.

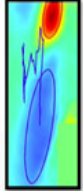
PowerPoint Slide Show - [3C16Demo-NSTX]

Remote Fusion Experiment Data Analysis Through Wide-Area Network

Abstract:
 We demonstrate remote data processing capability of large and high-throughput science experiment through cross-Pacific wide area networks and show how we can manage science workflow executions remotely by using ORNL ADIOS data management system and FNAL mdtmFTP data transfer system.

In this demo, we show a fusion data processing workflow, called Gas Puff Imaging (GPI) analysis, to detect and trace blob movements during fusion experiment. We send GPI data streams from Singapore to Fermilab for near-real time analysis. ADIOS manages analysis workflows and mdtmFTP transports stream data.


 Blobs in fusion reaction (EPST project)


 Blob trajectory




Figure 43: mdtmFTP Data Streaming Demo at Super Computing 2016
 The MDTM project web page is available at <https://mdtm.fnal.gov>.

Annex 10: BNL Status and Plan

Submitted by John Bigrow (big@bnl.gov)

January, 2017

Introduction:

Over the last year, Brookhaven National Laboratory Network Engineering has continued to enhance and refine our high-performance networking capabilities. New services and collaborations have been attached into both the “Science DMZ” and “High Performance Core” (HPC) areas. As a point of reference, the HPC has been rebranded as the “Science Core” (SC) and will be described with this new acronym even though there are no functional differences. Both the RHIC and ATLAS collaborations have historically driven the bandwidth demands at BNL. However, we have several new facilities and experiments coming on-line. The exact bandwidth demands of these facilities are unknown at this time, but could become substantial in the near future as they reach operational status. The Simons Foundation, National Synchrotron Light Source (NSLSII), Center for Functional Nanomaterials (CFN) and Computational Science Initiative (CSI) all have connectivity into the BNL “Science DMZ” and / or the SC enclaves. In turn, these enclaves can support direct access to our high-bandwidth Wide Area Network (WAN) links.

In order to provide additional functionality and capacity, the ATLAS and RHIC collaborations are being re-architected to remove any direct campus network connectivity and re-provision them as a totally separate and dedicated “Science DMZ” enclave. Once completed, this new RHIC / ATLAS architecture will support any additional bandwidth demands. Furthermore, this enhanced architecture will support the use of Internet Protocol version 6 (IPv6) for connectivity between the global ATLAS data centers and facilities.

The primary focus of this update will center on the three major high-bandwidth areas of the BNL network infrastructure; namely: the “Science Core” (SC), our evolving “Science DMZ”, and the BNL Wide Area Networking (WAN) capacities. On a last minute late note, BNL is in the preliminary stages of requesting an additional 100 gb/sec link to Manhattan from our provider the Energy Sciences network (ESnet). Besides providing additional capacity, this third 100 gb/sec link will greatly enhance our resiliency in failure mode operation. Since this additional WAN request is in the early stages of discussion and evaluation, no decision has been reached as of yet.

Wide Area Networking (WAN):

As discussed in previous site updates, BNL is currently provisioned with two 100 gb/sec WAN connections from our service provider ESnet. The underlying technology used for these circuits is Dense Wave Division Multiplexing (DWDM). Each of these links traverse disparate paths between the BNL campus and termination points in Manhattan. One circuit follows a southerly path traversing Long Island while the other fiber optic cable takes a northerly route from BNL to terminate in another New York City location. Both circuits are optically protected with sub-second failover capabilities that are supported in the ESnet owned Infinera DWDM equipment. At the network layer, each of these 100 gb/sec links are configured in a redundant failover configuration

using the Border Gateway Protocol (BGP) with plumbed Multi-Protocol Label Switching (MPLS) technologies. These MPLS circuits provide direct connectivity to CERN and the LHCOPN with three circuits; and to the LHCONE network with two circuits. The latest addition to this configuration is a dedicated link between BNL and Argonne National Laboratory (ANL).

During the last year, the ATLAS group at BNL participated in an evaluation with Amazon Web Services (AWS) for accessing commodity CPU cycles in the Amazon cloud. From the networking angle, BNL provisioned a dedicated MPLS-based circuit into the AWS cloud. The ATLAS collaboration used this link to provision on-demand CPU jobs at several of the Amazon data centers. This experimental service was so successful that ESnet has offered it as a production service for all the National Laboratories with both East and West coast connectivity into the AWS data facilities. It is anticipated that as more and more scientific and administrative support services become cloud based, the BNL WAN architecture will evolve to keep pace with these computing technologies in support of our users.

BNL continues to exploit the capabilities of the Border Gateway Protocol (BGP) for both our “Science DMZ” and SC network areas. We typically implement a private Autonomous System (AS) number inside each enclave or special service area. By utilizing this feature, we can precisely control network advertisements into each unique enclave. As an example; we use the private AS 65441 to peer with our “Black Hole Routers” which are part of our cyber security infrastructure. These route prefixes are used internally to drop Internet traffic that is considered malicious. Since these prefixes are tagged with private AS numbers we filter them out of our external BGP advertisements with our provider. This preserves the integrity of our external route peerings.

Science DMZ:

Also mentioned in previous site update reports, the BNL “Science DMZ” is a purpose-built networking infrastructure to support high-bandwidth WAN-centric collaborations such as RHIC and ATLAS. In furthering the use of the “Science DMZ” architecture, these two experiments are being re-engineered to be totally contained within a unique “Science DMZ” enclave. This is being done by removing any remaining direct connectivity from RHIC / ATLAS to the BNL campus network. Additional servers, firewalls, and other appliances have been purchased, installed and configured to support this new architecture. Additional services that needed to be setup include; the Domain Name Service (DNS) with a separate zone, both SSH and HTTP proxies. Lastly, two Juniper SRX-5400 series firewalls have been procured specifically to support the RHIC / ATLAS enclave in this enhanced “Science DMZ” architecture.

A secondary, yet vital, focus of this re-architecting is to support the IPv6 protocol. RHIC and ATLAS have performance requirements for implementing and using the IPv6 protocol throughout their global experimental facilities. The current Policy Based Routing (PBR) requirements of the ATLAS collaboration made this re-architecting a necessity in order to support IPv6 for these two collaborations without the additional expense of purchasing new equipment that would otherwise be needed to support IPv6 in the main BNL campus network. Unfortunately, this re-architecting increased the complexity of the PBR exponentially to support these additional capabilities. PBR support will be both a long term financial liability since only very high end equipment can support this feature; and will also be a constant source of frustration for the network engineering staff to

change or update. In the future, ATLAS may want to reconsider the long-term costs of supporting the PBR functionality which only provides minimal security for the facilities within this collaboration.

Figure 1 below is a greatly simplified diagram of the BNL perimeter network and “Science DMZ” enclaves. The twin Juniper MX2010 routers are configured in a complementary redundant mode. The MX-2010 units provide exterior BGP connectivity and perform the Policy Based Routing (PBR) functions. Since they are redundant, only one Juniper unit is required to be operational to maintain “Science DMZ” and site Internet traffic. As shown, other enclaves such as RHIC / ATLAS are redundantly attached to the Arista “Science DMZ” layer-3 switches, again with their own unique AS number. There is more than sufficient capacity to allow additional enclaves to connect to the “Science DMZ”.

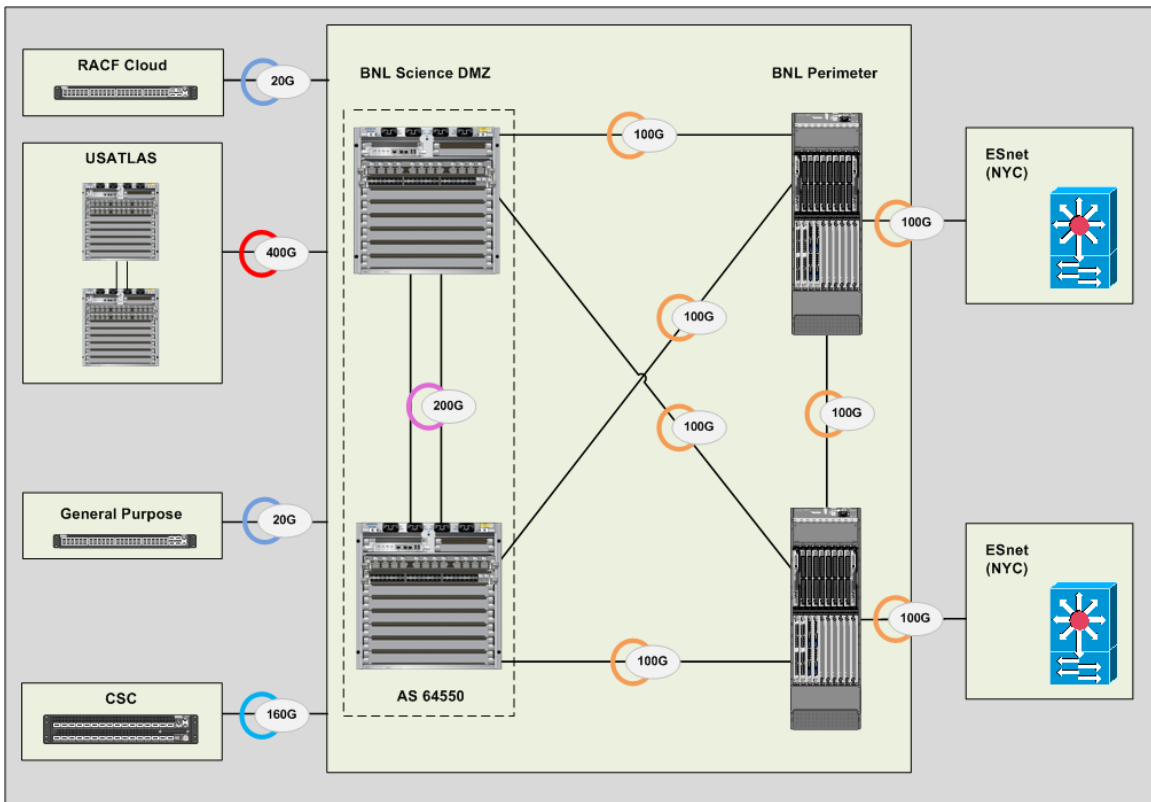


Figure 44: BNL perimeter network and “Science DMZ” enclaves

Science Core:

Unlike the “Science DMZ” which provides WAN-centric connectivity, The SC core supports internal intra-campus high-bandwidth connectivity. Throughout the BNL campus there are many scientific experiments which are physically far removed from the primary BNL data center. This physical isolation makes access to any high-bandwidth services difficult and expensive. The SC was engineered to provide a solution to this problem. An individual collaboration only need fund the necessary fiber optic installation from their campus locations into the BNL data center that houses the SC core. Once in the data center, they are physically and logically attached the SC core

for localized high-bandwidth connectivity. Additionally any SC core enclave can be BGP peered to the “Science DMZ” for WAN connectivity if necessary.

Figure 2 shows the SC core architecture that is implemented at BNL. Notice that each experiment enclave is allocated a private AS number as part of their BGP configuration. Each individual AS is peered back to the main HPC core routers and the “Science DMZ” if necessary.

RACF SC LAN Services Customers v2.0

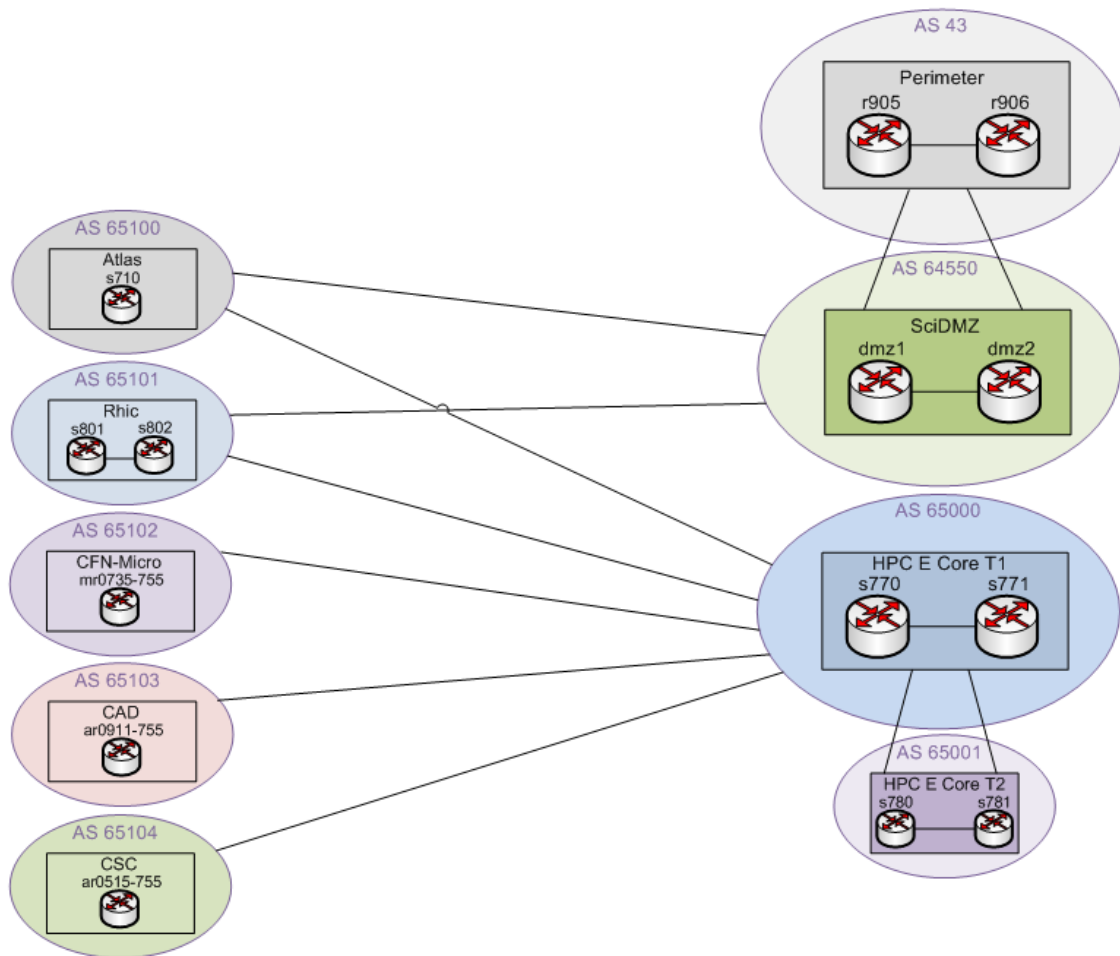


Figure 45: SC core architecture that is implemented at BNL

Summary:

BNL continues to expand our networking capabilities in our SC core, “Science DMZ” and the WAN perimeter network areas. As we evolve these architectures, they will continue to grow to support our scientific mission for many years to come.

For additional information on Brookhaven National Laboratory (BNL), the following link is a good starting point: <https://www.bnl.gov>

National Networks

Annex 11: CANARIE (Canada) Status and Plan

Submitted by *Randy Sobie* (rsobie@uvic.ca)

January 2017

This report describes the status and plans of the Canadian network infrastructure used for particle physics research in 2016. HEPnet/Canada³⁰ (<http://hepnetcanada.ca>) coordinates the network, working together with the network providers, and the Canadian universities and laboratories. We describe the status of the CANARIE network infrastructure, the ATLAS Tier-1 and Tier-2 centres in Canada, and our other network projects.

CANARIE

Twelve provincial and territorial network partners, together with CANARIE, collectively form Canada's National Research and Education Network (NREN). Canadians at universities, colleges, research institutes, hospitals, and government laboratories rely on this ultra high-speed network to collaborate in data-intensive, cutting-edge research and innovation within Canada and with colleagues in over 100 countries.

Beyond the network, CANARIE funds and promotes reusable research software tools to accelerate scientific discovery. CANARIE also supports Research Data Canada as it leads national research data management initiatives, and through the Canadian Access Federation, provides identity management services that enable secure, ubiquitous connectivity and content access to the academic community. To boost commercialization in Canada's technology sector, CANARIE offers cloud resources to startups through its DAIR service, and links a powerful community of public and private sector partners in the Centre of Excellence in Next Generation Networks (CENGN).

CANARIE Infrastructure

CANARIE continued its vision of building fibre infrastructure from coast-to-coast. The nationwide fibre network allows CANARIE to deliver high-bandwidth services for data intensive applications, high performance computing centres and research facilities. CANARIE started the fibre system extension into eastern Canada in 2016 and will complete the fibre build in 2017. The new system brings significantly increased network capacity in eastern Canada, providing a potential linkage to connect researchers in Europe through a cable landing point in Halifax, Canada.

In 2016 CANARIE completed the decommissioning of its SONET infrastructure and fully deployed 100Gbps redundant core IP network. Also CANARIE operationalized L2 VPLS services offering highly resilient, dedicated point-to-point or point-to-multipoint services for data intensive applications. The redundant nature of the L2 VPLS technology increases the availability of the network services.

CANARIE Network Services

CANARIE offers a number of network services such as: R&E IP Service, Content Delivery Service, p2p Connection Service.

R&E IP Service

IP network service remains the biggest workhorse of CANARIE's service offerings. The 100Gbps redundant infrastructure significantly improves the performance and reliability of our network

³⁰ Additional information can be obtained by contacting Dr. R. Sobie, Director of HEPNET/Canada (rsobie@uvic.ca)

services, which means fewer disruptions and delays for researchers. With a number of internal segments that link the routers in a partial-mesh topology, the IP core network provides full and equal support for IPv4 and IPv6 unicast and multicast routing. As well, the network consists of external network segments that extend to international R&E exchanges: Pacific Wave in Seattle, StarLight in Chicago, and Manhattan Landing (MANLAN) in New York. Through these exchange points CANARIE reaches National R&E networks around the world.

Content Delivery Service

Content Delivery Service provides institutional users with high-speed access to major content providers, like Amazon, Microsoft, Google, Yahoo, Facebook and Box.net. It has become an important service of CANARIE service offerings. The Content Delivery Service IP network, which is logically separated from the CANARIE R&E IP Network, links to SIX (Seattle, WA), Pacific Wave (Seattle, WA), TorIX (Toronto, ON) and NYIIX (New York City, NY) to source content. With increasing demand of accessing content, CANARIE started establishing direct peering with a few big content providers.

p2p Connection Service

As stated previously, CANARIE p2p Connection Service is standardized in L2 VPLS technology. Servicing up to a full 100Gbps connection, a p2p connection can be delivered directly into researchers' equipment, connecting high-performance computing centres or research facilities across Canada or in other continents. TRIUMF and Canadian Tier 2 centres utilize this service to support LHCOPN and LHCONE infrastructure.

International Connectivity

CANARIE, partnering with Internet2, NORDUnet, SURFnet, and GEANT provisioned 3 x 100Gbps links, called ANA-300G, between five open exchange points at both sides of the North Atlantic. The ANA-300G enables researchers and scientists to transfer data between North America and Europe at speeds that were previously only possible within the continents. In 2016, ANA operation team started a plan of re-terminating one of the 100Gbps links from New York to the CANARIE PoP in Montréal. The move would offer full diversity of physical landing points in North American. The new Montréal link allows direct traffic from Canada to Europe, providing even greater flexibility and network diversity to access research data and content.

ATLAS Tier 1 Computing Centre at the TRIUMF Laboratory

TRIUMF, Canada's National Laboratory for Nuclear and Particle Research operates a Tier-1 (T1) Computing Centre for the ATLAS experiment in Canada. The TRIUMF Centre is linked to the LHC Worldwide Computing Grid (WLCG) and provides an interface to a grid of computing resources at universities across Canada.

In July 2005, CANARIE signed a Memorandum of Understanding (MOU) with HEPnet/Canada, ATLAS Canada and TRIUMF to provide the high-energy physics community with a dedicated 10G circuit across Canada and initial 5G lightpath to the CERN Tier-0 (T0) Centre. This lightpath became active in December 2006.

Each T1 site must use a small or series of small publicly routable Classless Inter-Domain Routing (CIDR) blocks as only traffic from the Large Hadron Collider Optical Private Network (LHCOPN) address space is allowed to flow over the network. Exchange of routing information is performed using Border Gateway Protocol (BGP) at the T1 and T0 institutions. This circuit was replaced in 2015 with a MPLS VPN circuit that will connect with the ANA-300 transatlantic link at ManLan in New York City. The T1 to T0 circuit is provisioned with 10 Gbps with the potential for this to

be increased easily within the CANARIE network in their more flexible MPLS system and across the ANA-300 Network.

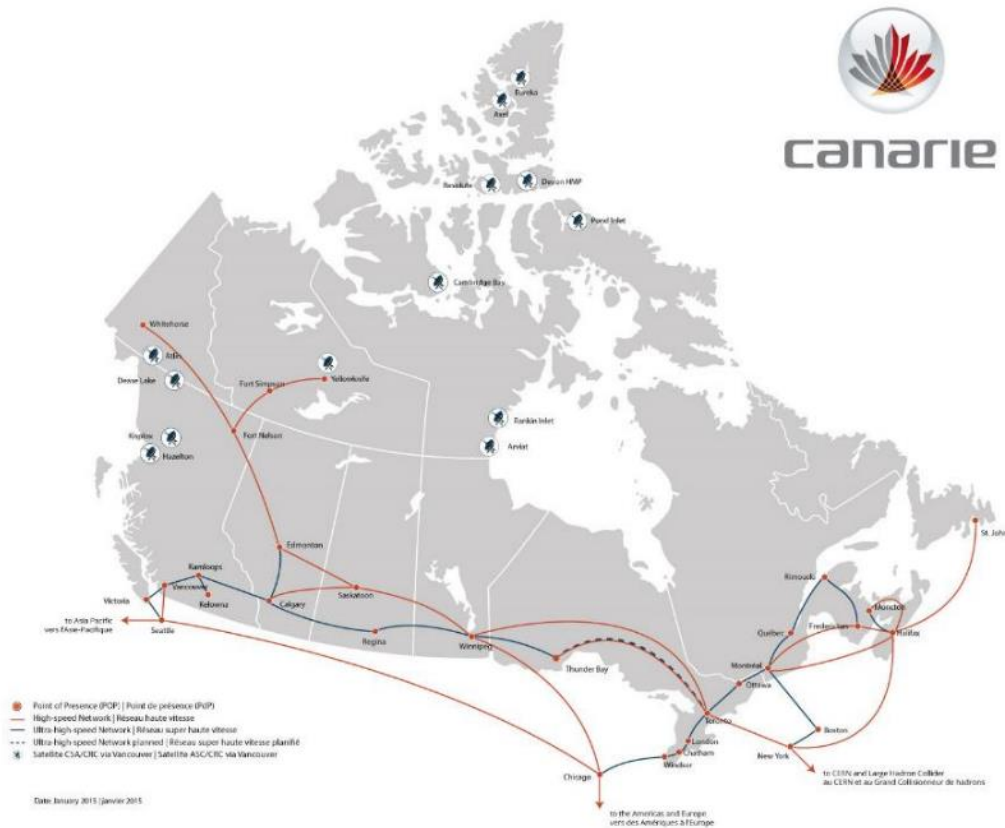


Figure 46: The CANARIE network in 2016

TRIUMF has one physical 10G LHCOPN connection to BCNET/CANARIE with two dedicated VLANs going from TRIUMF to CERN. The VLANs are designed to provide redundancy for Tier-1 site, and CANARIE achieves this by distributing them on a different set of geographically separated backbone routes. In case both VLAN are down LHCONE will be used as a last network resource to get to CERN.

LHCONE Network in Canada

The LHCONE network is deployed as purpose built VRF overlay network for use by Canadian T1 and T2 sites spanning all CANARIE Juniper MX series routers. All Canadian T2s and the T1 were connected to this service at 10G by December 2012. The LHCONE serves HEP connectivity requirements within Canada in addition to providing connectivity to international sites. This approach allowed us to eliminate a number of point-to-point circuits between TRIUMF and the T2s. The CANARIE LHCONE VRF has international peerings with Internet2, ESnet and GEANT at ManLan (New York City) and Starlight (Chicago), and with ESnet, Internet 2 at PacificWave (Seattle).

During the development of the LHCONE in Canada the decision was taken early to avoid any type of Policy Based Routing (PBR) on the advice of the Canadian R&E networking community. We concluded that the best option for sites without dedicated hardware for the LHCONE would be to create a VRF at the site in a configuration shown in Figure 2. The Site Local LHCONE VRF is created on the available data centre edge equipment and configured to peer with the CANARIE LHCONE VRF. The Campus Router is set as the default route for the site Local LHCONE VRF to obtain regular R&E IP network connectivity. The CANARIE VRF is configured to accept only pre-negotiated subnets that are known to contain LHC equipment. The Site Local LHCONE VRF does not re-advertise the remote LHCONE routes to the Campus Router to avoid any asymmetric routes.

The LHCONE L3VPN was very stable and successful accounting for a significant fraction of the total traffic across CANARIE. A contributing factor to the stability was the initial design that avoids PBR and instead uses VRFs at each LHCONE site where dedicated equipment was not available (see Figure 3). LHCONE bandwidth across CANARIE was provisioned at 2x10G circuits in 2014 and the LHCONE overlay network was moved onto the new CANARIE 100G IP network in 2015. Canadian T2s are beginning to observe occasional bottlenecks with 10G links into the LHCONE, particularly in the case where the LHCONE uplink is shared with other non-HEP projects. Investigation for adding additional capacity is underway.

The BelleII experiment located at the KEK Laboratory in Tsukuba in Japan is developing its computing infrastructure. The computing requirements are expected to be comparable to a smaller LHC experiment. Currently, there are two BelleII sites in Canada, at UVic and McGill. The BelleII experiment joined the LHCONE in 2015. At the moment traffic on the LHCONE from BelleII is modest in Canada but expected to grow over the next few years.

The Canadian LHCONE network was very busy in 2016. The top plot in figure 4 shows the network traffic from the Tier-1 facility at TRIUMF. The lower three plots show the traffic to and from Canada from ESNET, Internet2 and GEANT. Canada connects to GEANT from Montreal and Toronto; the plot only shows the traffic via Montreal (the traffic via Toronto is comparable).

Other activities

IPv6

With the advent of personalized cell phones and tablets, the IPv4 address range is becoming more and more limited. IPv6 with its vast address space provides a solution, but the whole affected network infrastructure outside and inside the universities and laboratories must be IPv6 enabled before worker nodes and workstations can enable IPv6 or enable IPv4 and IPv6 dual stack.

During early 2016 UVic enabled IPv6 on their perfSonar nodes, fully enabling in and outgoing perfSonar related traffic by Summer 2016, and offering help and advice for IPv6 configuration to other sites since. In September 2016 the WLCG Management Board approved the IPv6 deployment plan. Dual stack availability is mandatory for the Tier0 and Tier1s by April 2017. By April 2018 dual stack should be available in production for the Tier0 and Tier1s. By the end of Run2 a large number of sites should have migrated their storage to IPv6. At the end of 2016 TRIUMF successfully completed setup of dual stack PerfsonarPS, advertising IPV6 address space to LHCOPN and LHCONE. TRIUMF is working on their dual stack storage configuration in the timeline provided by WLCG.

Network monitoring

HEPnet Canada deployed 10G capable perfSonar boxes to all sites during 2011. In 2014, perfSonar became a mandatory installation for all ATLAS sites. In addition Canadian HPC centres and research networks adopted a similar system for monitoring their network. The HEPnet nodes are participating in a central mesh including all ATLAS sites worldwide. A sub mesh including only Canadian sites will be configured, however it was not found necessary in the past. These nodes have been crucial in identifying networking problems in a timely manner on Canadian as well as on American networks.

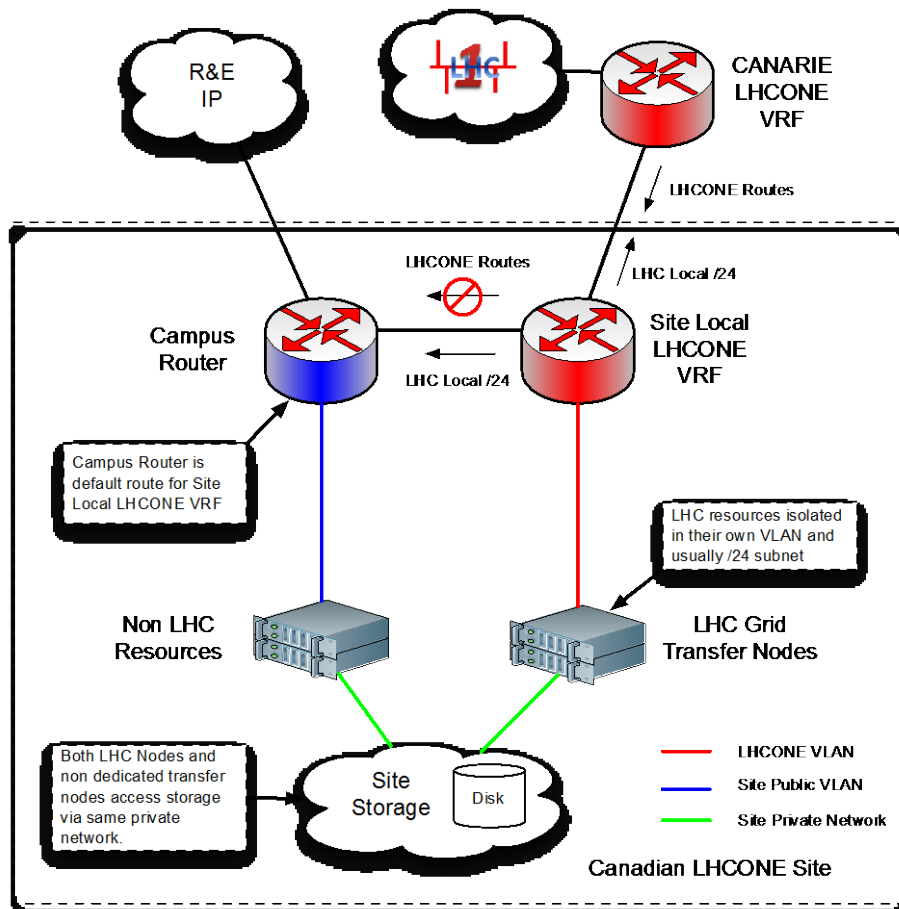


Figure 47: LHCONE network design at a typical Canadian HEP site.

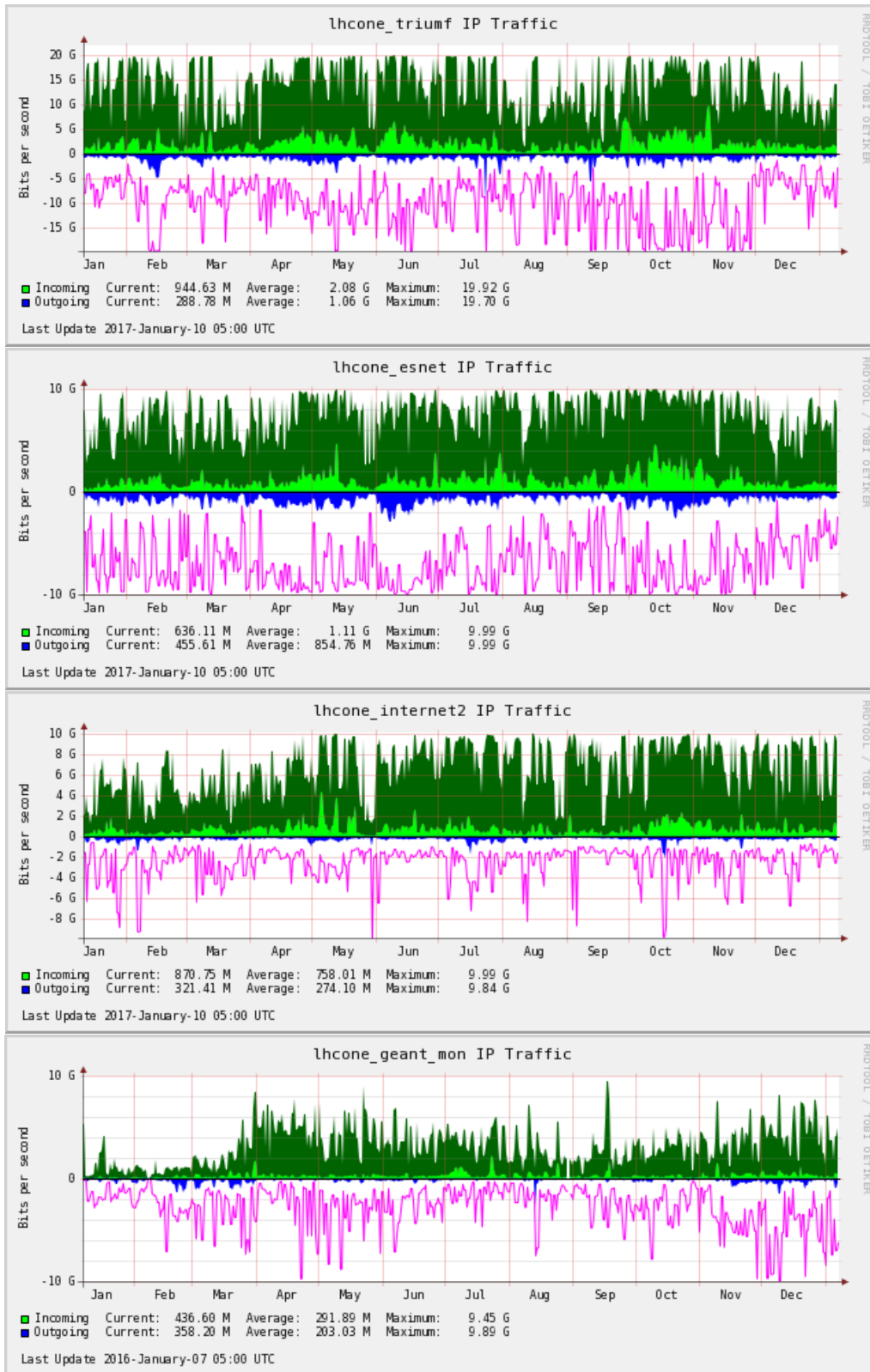


Figure 4: LHCONE traffic from TRIUMF and Canadian traffic to ESNET, Internet2 and GEANT (Montreal).

Annex 12: SURFnet Status and Plan

Submitted by Alexander Vanden Hil (alexander.vandenhil@surfnet.nl), Gerben van Malenstein (gerben.vanmalenstein@surfnet.nl), Migiel de Vos (migiel@surfnet.nl)
February 2017

Introduction:

National backbone – SURFnet8

In 2016 SURFnet prepared for the realisation of our next generation network, SURFnet8. We successfully procured³¹ a new photonic layer that will be delivered by ECI Telecom³². This optical layer ensures growing capacity demand, will serve the migration of existing services and forms the basis of the new service layer – to be procured in 2017 – as well. All optical channels in SURFnet8 will be operating at 100G from the start and alien waves will be possible a standard network feature.

For SURFnet8, we envision automation of the various technology domains: optical layer, service layer and Network Function Virtualisation. By having open APIs available for these layers, not only can we deliver specific technology domain services faster, but this also allows for orchestration of all technology domains, depicted in Figure 1.

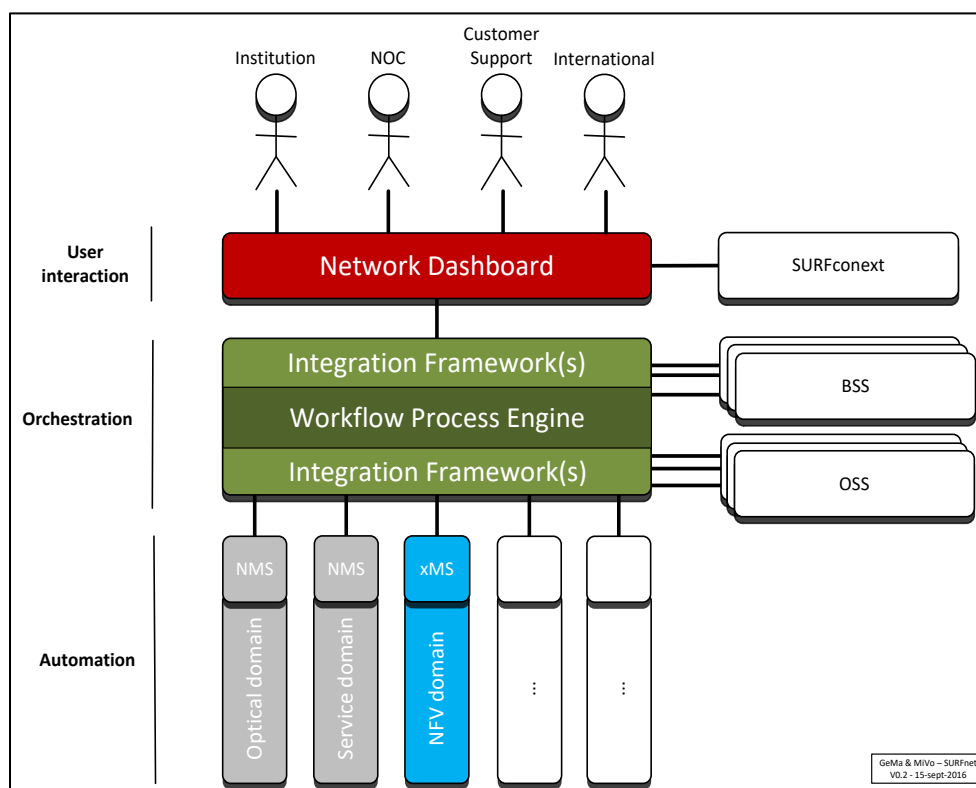


Figure 48: Orchestration of technology domains

³¹

(Dutch)

<https://www.tenderned.nl/tenderned-web/aankondiging/detail/samenvatting/akid/bd29f372398e1a50f45311b108999c0d/pageId/D909B/huidige>

[menu/aankondiging/cid/319041/cvp/join](https://www.tenderned.nl/tenderned-web/aankondiging/detail/samenvatting/akid/bd29f372398e1a50f45311b108999c0d/pageId/D909B/huidige)

³² <https://www.surf.nl/en/news/2016/12/eci-delivers-new-photonic-layer-for-surfnet-network.html>

All user interaction with the network and network functions will be facilitated by a so-called Dashboard interface.

Network Function Virtualisation

With NFV, the functions conventionally performed by specialised on-site network equipment are migrated to (off-site) virtual environments, so they can be used in a flexible and scalable manner. Examples of these network functions include rate limiting, firewalling, intrusion detection, switching, routing, monitoring, and so on.

We foresee NFV to play a key role in SURFnet8, our next generation production network. Therefore, we are actively researching and testing this technology: NFV in combination with Service Function Chaining (SFC) has been demonstrated³³ successfully at SC'16 in Salt Lake City, Utah.

International

While SURFnet operates several Cross Border Fibers (CBFs) internationally, research was conducted on spectrum sharing with PSNC on the Hamburg-Frankfurt-Geneva dark fiber. This enabled us to deliver fully geographically disjoint services from Amsterdam to Geneva.

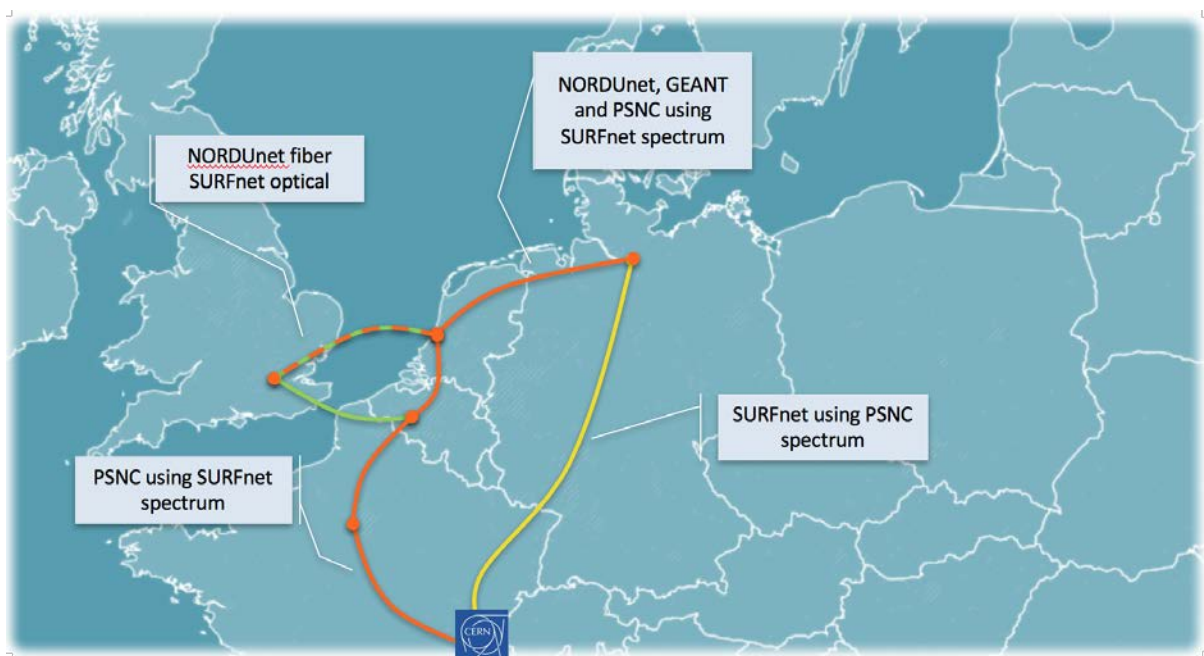


Figure 49: Spectrum sharing SURFnet-PSNC

On the trans-Atlantic section, SURFnet has decommissioned all of our OC192s, resulting in 100G-only circuits, via the ANA-300G consortium, see Figure 3.

³³ <https://blog.surf.nl/en/demonstrating-network-functions-with-virtual-reality/>



Figure 50: ANA-300 Consortium and connectivity

The Automated GOLE, depicted in Figure 4, enables automated setup of multi-domain network services. For this environment, we choose to use MEICAN from RNP as the world-wide provisioning dashboard for NOCs³⁴.



Figure 51: Automated GOLE

³⁴

<https://indico.cern.ch/event/527372/contributions/2288042/attachments/1339832/2017195/LHCONE.NSI.AGOLE.update.pdf>

Annex 13: GARR-X and GARR-X Progress (Italy) Status and Plan

January 2017
planning@garr.it

Introduction:

During 2016, the GARR network continued its evolution, planning the extension to centre-north regions of Italy of the advanced features successfully tested in the GARR-X Progress project during 2015.

Public tenders were launched to procure dark fibres with long term contracts to connect user sites to GARR-X PoPs in the centre-north. This is a long term investment that will permit to expand the network capacity in the future in a flexible way, maintaining the full control of fibres and equipment independently from the telecommunication operators. Cost sustainability has been one of the keywords that guided the evolution strategies of the GARR network over the years, containing maintenance and operational costs in view of the infrastructural investments achieved with GARR ordinary funds.

The procedure was completed at the end of 2016, and it is expected that the first realisations will be in service during late spring of 2017. The project includes new dark fibre local loops for 120 user sites, new optical transmission equipment and new IP/MPLS router parts, in order to provide from day one 500Gbps (but ready to grow up to 1Tbps) on the transmission backbone and 100Gbps between IP routers.

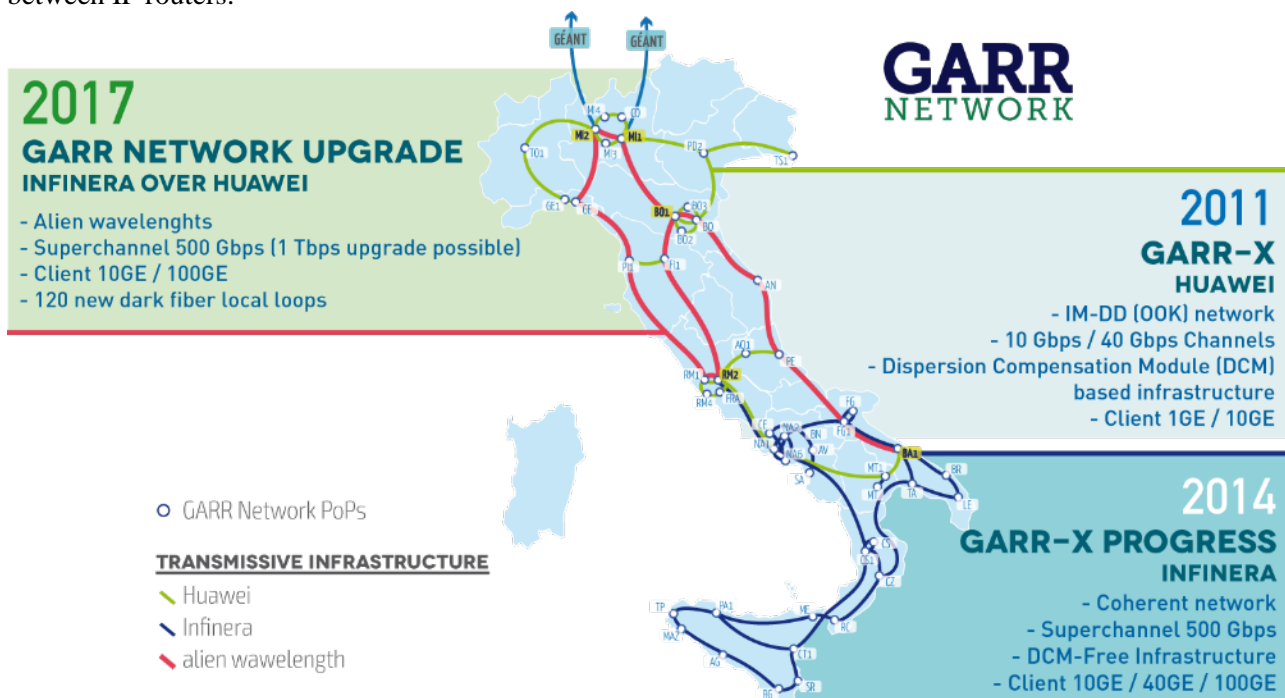


Figure 52 - GARR-X network upgrade, 2017

Alien wavelenghts: our allies on the optical network

Optical networks sometimes encounter limitations in terms of flexibility in service provisioning and ability to follow the swift evolutions of optical communications. These limitations can be overcome with the integration of heterogeneous optical platforms. Thanks to this approach, it is possible to provide next-generation transmission services over the existing transport and

regeneration equipment. At the same time, the integration allows for the evolution of the optical network by means of targeted interventions, while ensuring service reliability.

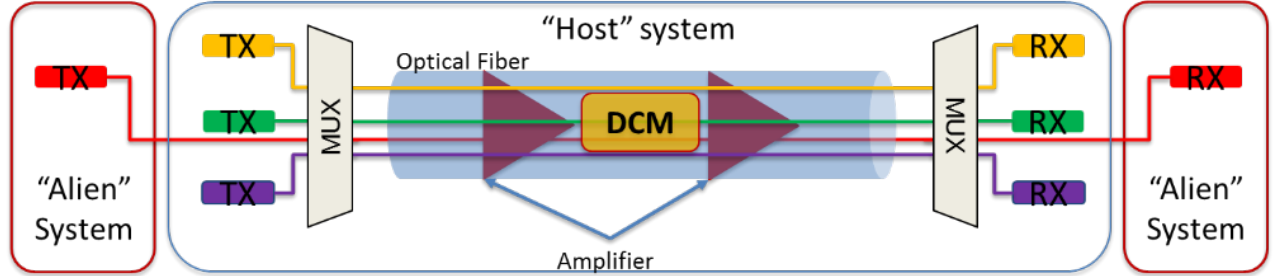


Figure 53 - Alien wavelengths block diagram

GARR did an extensive field testing of the alien wavelength technique, a hybrid solution based on the transmission and reception of optical signals generated by an infrastructure different from the one providing transport and regeneration. These optical platforms deeply interoperate, as the operational functionalities of the transit nodes (i.e. multiplexing, optical switching, routing, and amplification) must act in the same way both on native and alien signals.

Thanks to the field test results, GARR is now planning to use the alien wavelength technique to provide 100G Ethernet client services delivered by means of the new Infinera equipment on the main backbone nodes of the Huawei infrastructure, in the northern and central part of Italy.

GARR-X and the INFN Tier1/Tier2 network (the Italian LHCONE)

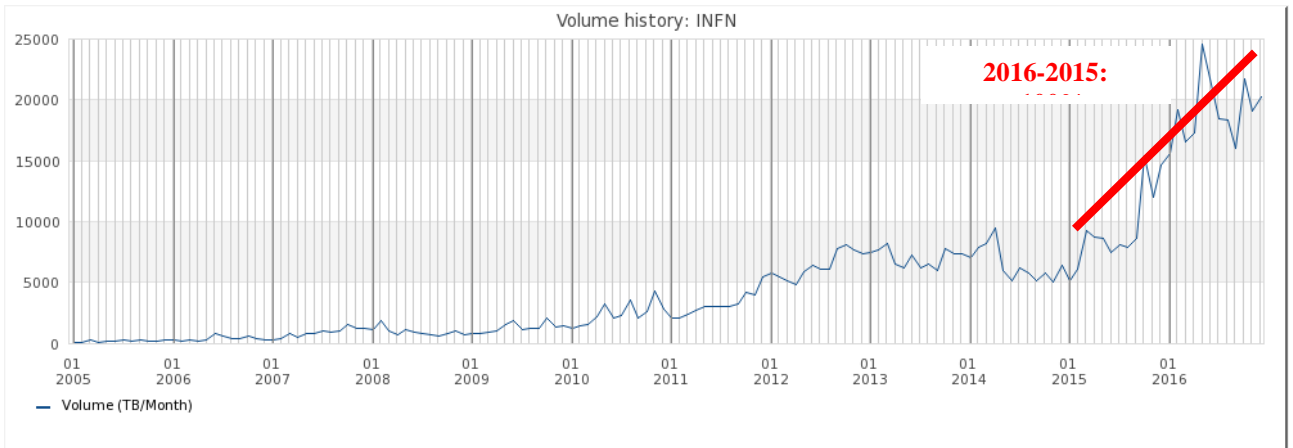


Figure 54 - aggregate volume (IN+OUT, TB/month) of all INFN sites

During 2016 the overall traffic volume generated by INFN sites (mostly dominated by LHC data transfers) has seen a 100% rise compared to 2015, in connection with the restart of LHC operations in April. This huge increase was perceived on the whole GARR network, urging for the backbone network upgrade to be completed in the expected timeframe, in order to accommodate for further increases.

The Italian LHCONE is a Virtual Private Network built over the IP network and using the GARR-X capacity, interconnecting the Italian Tier2s each other and with INFN-CNAF Tier1, and interconnected with the rest of LHCONE via a dedicated GÈANT connection running over a 100Gbps link. The current LHCONE GÈANT implementation is interconnecting at the European level the Italian, French, German, Spanish (among others) LHC sites and offering interconnections with the rest of the European, North American, and worldwide centres. The total network access capacity of the Italian LHC sites amounts to 190Gbps, while the total capacity of international links available to LHC data transfers amounts to 250Gbps.

| Site | Present physical connection |
|------------------------|---|
| INFN-CNAF Bologna (T1) | 6x10G |
| INFN-Bari (T2) | 10G |
| INFN-Catania (T2) | 2x10G |
| INFN-Frascati (T2) | 10G |
| INFN-Legnaro (T2) | 2x10G (shared with general purpose traffic) |
| INFN-Napoli (T2) | 20G |
| INFN-Milano (T2) | 10G |
| INFN-Pisa (T2) | 2x10G (shared with general purpose traffic) |
| INFN-Roma1 (T2) | 10G |
| INFN-Torino (T2) | 10G |

Figure 55 - Tier1 / Tier2 network connectivity as of December 2016 (LHCONE sites in green)
 From the LHCOPN point of view, the capacity available for the Italian Tier1 traffic was upgraded in July from 20Gbit/s to 40Gbit/s, using three 10Gbit/s to CERN via GÈANT circuit services, and another 10Gbit/s to CERN built over a “cross border fiber” link across Italy, Switzerland. The LHCOPN backup path to CERN uses the LHCONE capacity. The capacity upgrade was immediately used by CNAF data transfers from CERN.

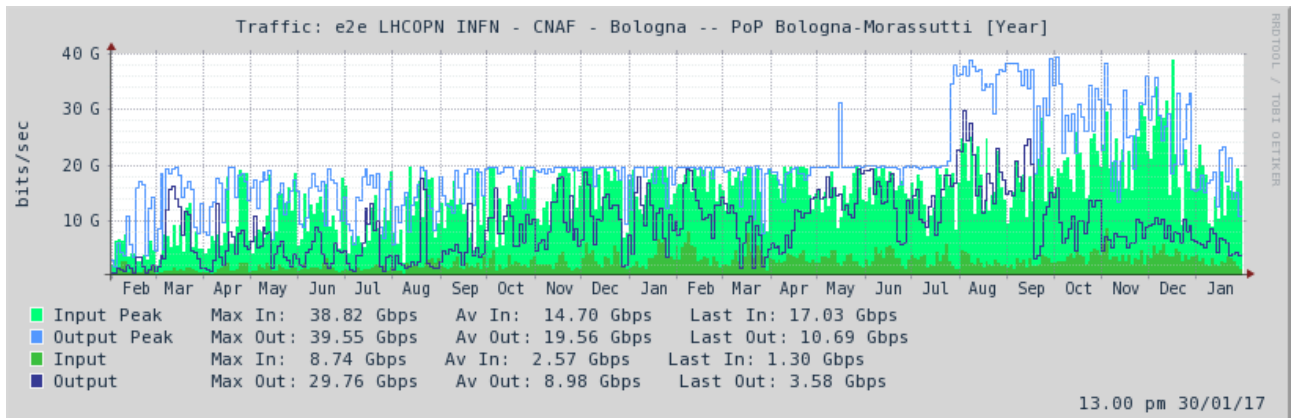


Figure 56 - LHCOPN network traffic seen by CNAF

Tier-1 farm extension

The foreseen needs for CPU power is expected to grow considerably in the next few years, hence a strong interest in testing the usage of remote resources to (dynamically) extend the Tier-1 farm has arisen. The INFN-CNAF team started during 2016 to experiment a remote extension of the Tier-1 datacenter to the INFN-Bari datacenter, similar to the CERN-Wigner extension. The goal

was to offer the experiment a direct and transparent access to Bari resources from CNAF. Jobs were expected to access data the same way as at CNAF, making them unaware of the extension, dispatching jobs to the remote datacenter and providing a cache system to speed up the response time.

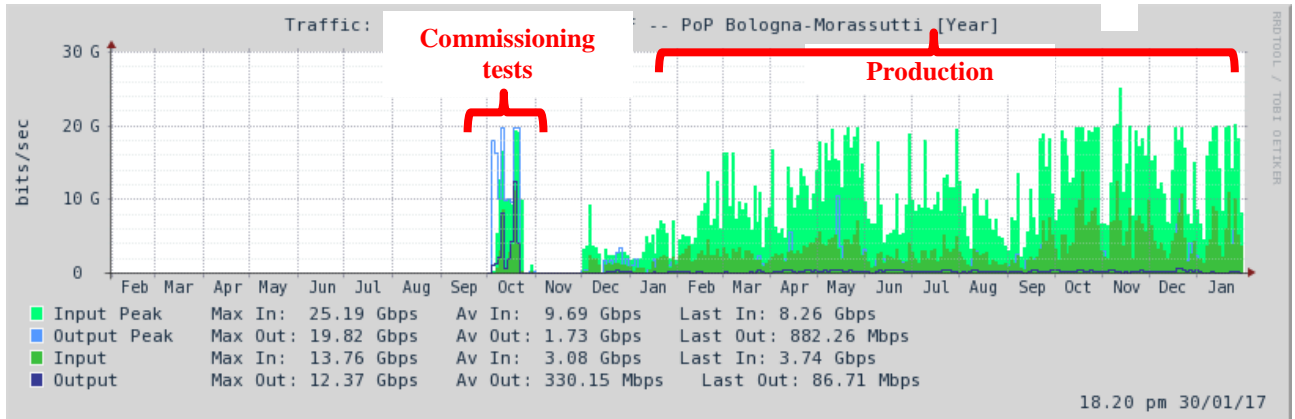


Figure 57 - Tier-1 extension network usage

GARR provided for this project a dedicated 2x10Gbit/s L3VPN from Bologna to Bari (9ms RTT), that proved successful and was subject to intense usage, allowing to develop the system and to gain experience.

Belle II data challenge

During 2016 the Belle II experiment performed a new data challenge test, involving KEK, the house of the SuperKEKB collider near Tsukuba in Japan, PNNL, near Seattle in the USA and two Italian sites, INFN-CNAF and INFN Napoli. The data challenge used the LHCONE network paths available between Japan, USA and Europe (especially the new 2x10Gbit/s Tokyo-London link) and was meant to stress the network infrastructure and software stack between the major Belle II sites in both ways. The tests were a success, and demonstrated a huge improvement thanks to the lower delay guaranteed by the new network path and the preparedness of the Italian Belle-II sites for the expected network requirements.

About GARR:

GARR is a non-for-profit organization that plans, implements and operates the Italian Research and Academic Network. GARR is an independent legal entity from 2003, whose founding members are the Italian universities, and most prominent research organizations in Italy, i.e. the National Research Council (CNR), the National Entity for Energy and Environment (ENEA), the National Institute for Nuclear Physics (INFN). All major scientific, academic and cultural institutions in Italy connect to the GARR network.

Annex 14: CESNET2 (Czech Republic) Status and Plan

<http://www.ces.net>

January 2017

Submitted by Jiří Navrátil (jiri@cesnet.cz)

The following report summarizes changes in the CESNET e-infrastructure and its services for HEP in the Czech Republic in 2016.

CESNET:

CESNET, the Czech National Research and Education Network provider, is an association of legal entities formed by all universities of the Czech Republic and the Academy of Sciences of the Czech Republic (ASCR). Its main goals are the operation and development of the Czech national e-infrastructure for research, experimental development and innovation, and research and development of advanced network technologies and applications. The e-infrastructure includes national communication infrastructure (CESNET2 network), national grid infrastructure (NGI), high capacity distributed storage system, applications and tools for effective co-operation among distributed users and teams as well as instruments and services for controlling access to e-infrastructure resources, instruments for ensuring security of communication and data protection, CESNET is also research organization with wide spectrum of applied research activities and it also takes an active part in consortia of several international research projects in the information and communication technologies area.

e-Infrastructure Status 2016:

1) CESNET2 Network

CESNET 2 is hybrid communication network based on over 6000km of leased optical fibers deployed with transmission systems DWDM³⁵ operated by the CESNET. The part of this infrastructure is based on “lighted fiber” services (typically CL DWDM lines).

The CESNET2 optical DWDM backbone operates two different types of the DWDM technology, the main DWDM system is based on the Cisco ONS 15454 MSTP technology with the ROADM³⁶ and open DWDM systems based on CzechLight family (CL) programmable devices (this technology is developed by the CESNET research project). The ONS 15454 MSTP ROADM technology is based on the 80-channel wavelength crossconnect cards. Dedicated wavelength paths are provisioned by the central management system and can be created on-demand by user requests. The ONS 15454 MSTP DWDM system supports alien wavelength transport and Ethernet L2 over DWDM point-to-point and multipoint VLAN or QinQ (using the XPonder card) services. This system is designed to support 100 Gbps wavelength.

Open DWDM based on CL devices creates complementary part of network to the ONS 15454 MSTP system as cost effective solutions also including single fiber bidirectional operation. This

³⁵ Dense Wavelength Division Multiplexing

³⁶ A **Reconfigurable Optical Add-Drop Multiplexer**. See http://en.wikipedia.org/wiki/Reconfigurable_optical_add-drop_multiplexer

³ WSS - **Wavelength selective switching**. See https://en.wikipedia.org/wiki/Wavelength_selective_switching.

system operates the 40-channel plan and includes ROADM and WSS³⁷ technology including flexible optical spectrum allocation. These optical channels are typically terminated on the pluggable DWDM optics with the DOM support installed in the end equipment (routers and switches). It also offers another option to convert DWDM signal to grey optics using optics converters. The actual DWDM optical topology is shown on Figure 1.

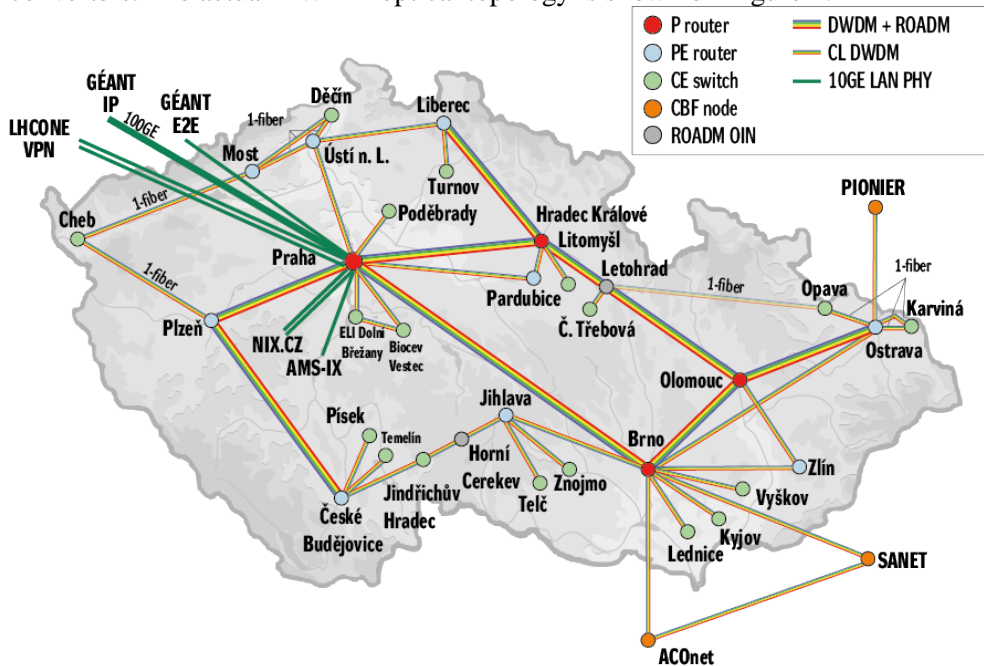


Figure 58: CESNET2 DWDM optical topology

The most of our optical DWDM PoPs are connected by diverse optical lines, that in the combination of different DWDM technology increases its availability and reliability. Optical backbone also provide photonic services allowing transfer of advanced signals as precise time, ultra-stable frequency or sensing signals.

Above the optical transmission layer is IP/MPLS network layer. The main DWDM ring is equipped by the core P routers in Praha (dual PoP), Brno, Olomouc and Hradec Králové. In the other PoPs are installed access PE routers, which provide all the functionality and services of backbone network (MPLS³⁸, EoMPLS, QoS, IPv4/IPv6 unicast and multicast routing and NetFlow statistics. See details in Figure 2.

³⁸ Short for **Multiprotocol Label Switching**, an IETF initiative that integrates Layer 2 information about network links (bandwidth, latency, utilization) into Layer 3 (IP) within a particular autonomous system--or ISP--in order to simplify and improve IP-packet exchange.

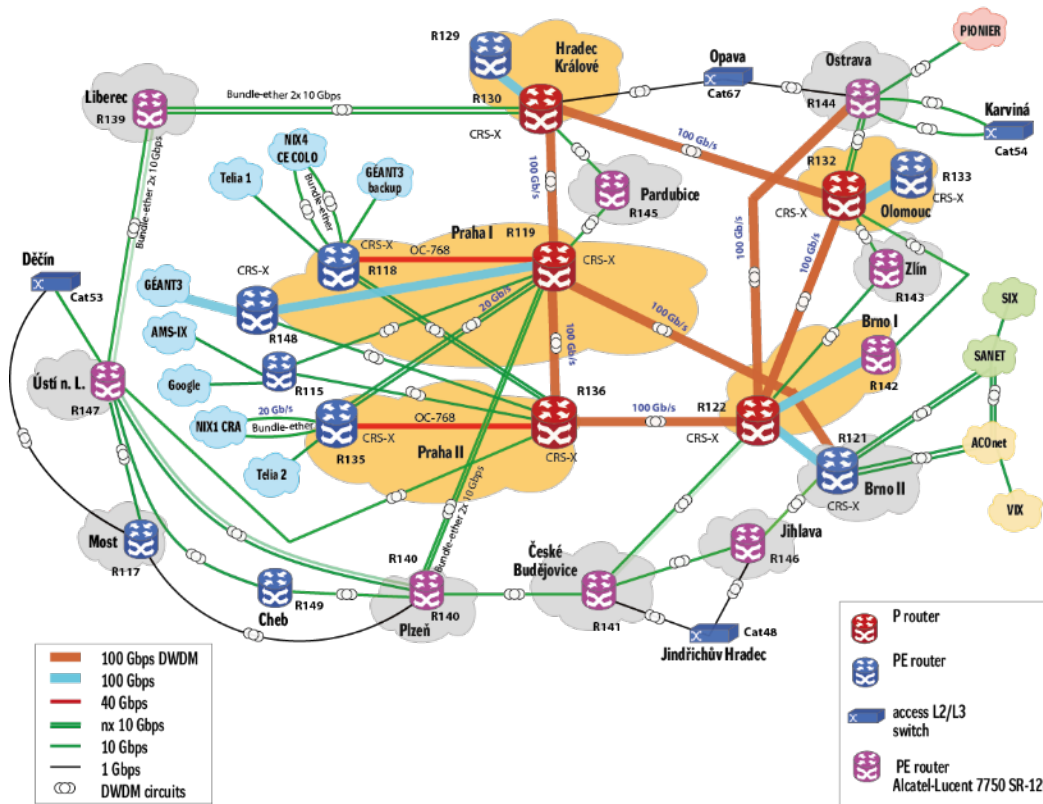


Figure 59: MPLS topology

The CESNET still keeps very high level of an external connectivity with academic partners and commercial ISPs. Currently CESNET has:

- 100 Gbps to Géant
- 20 Gbps connectivity to LHCONE infrastructure (see <http://lhcone.net>),
- 10 Gbps for commodity traffic
- 10 Gbps to NetherLight for projects within the framework of the GLIF initiative
- 10 Gbps to AMS-IX

Crossborder connections:

- 20 Gbps to SANET (the NREN of the Slovak Republic) and SIX
- 20 Gbps to ACONet (the NREN of Austria) and VIX, including precise time transmission
- 10 Gbps to PIONIER (the NREN of Poland)

2x20 Gbps to the Czech Neutral Internet Exchange (NIX.CZ)

2) NGI

The National Grid Infrastructure represented by MetaCentrum operates and manages distributed computing infrastructure consisting of computing and storage resources owned by CESNET as well

as many co-operative centers within the Czech Republic (see Figure 3). MetaCentrum is responsible for building the NGI and its integration to related international activities, especially in the European Union (EGI). MetaCentrum is actively involved in many international Grid projects such as EGI Engage, INDIGO-DataCloud, ELIXIR EXCELERATE, AARC, MAGIC, EOSC. MetaCentrum provides grid, cloud and map-reduce environments. The system provided by MetaCentrum is flexible enough to fully integrate any computing capacities to get significant higher computing power for research.



Figure 3: National grid Infrastructure

In 2016, the computing power managed by MetaCentrum is presented more than 13 500 CPUs in national infrastructure and 3 700 CPUs in EGI. For its users MetaCentrum provides temporary and semi-permanent storage capacity which was upgraded to 4 PB in national infrastructure and 3,7 PB in EGI.

3) The Distributed Storage

The high capacity distributed storage system consists of three geographically distributed nodes, each of them consisting of a Hierarchical Storage Management (HSM) system. The first storage node in Plzeň was put into operation at the beginning of 2012; the nodes in Jihlava and Brno have started their operation in the second half of 2013. Moreover, at the end of 2013, the capacity of the first site (in Plzeň) was significantly upgraded, too. The result of that is increasing of whole raw capacity of the distributed storage system to approximately 21 PB. All three storage sites are in routine operation. The storage facilities are accessible via a plethora of protocols and they are used mainly for backups, archiving and sharing data originating from a wide range of research and educational activities of the R&D community in the Czech Republic. Geographical replications among sites have been introduced in 2016 as the most significant new service. Over 70% of overall capacity of the storage facilities is currently utilized by 197 user groups (virtual organizations).

4) Collaboration Infrastructure

Multimedia collaboration infrastructure offers PostHD transmissions systems, videoconference, webconference, streaming and IP telephony services. For individual users and teams, infrastructure

offers virtual rooms on geographically distributed HD MCUs with option to record and stream the meetings. VC infrastructure peering uses GDS or basic IP based dialing. Webconference system based on Adobe Connect offers connection of up to 200 users in multiple virtual rooms. Events could be also streamed using streaming servers infrastructure and delivered to thousands of viewers. Connected organizations are also interconnecting their PBX into IP telephony network and utilize IP connectivity for voice calls.

Support of HEP

The particle physics community in the Czech Republic is actively involved in the LHC, RHIC and KEKB accelerator experiments, in Fermilab Intensity Frontier experiment, reactor neutrino experiments as well as in astrophysics experiments. The most important HEP groups in the Czech Republic are:

1. Institute of Physics, ASCR, Prague,
2. Nuclear Physics Institute, ASCR, Řež,
3. Faculty of Mathematics and Physics, Charles University, Prague,
4. Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague.
5. Palacký University, Olomouc

In addition to standard IP connectivity, these institutes are connected to the CESNET with 1 or 10 GE lambdas.

The Institute of Physics hosts a Tier2 center for LHC experiments ATLAS and ALICE. Some storage servers are also located in the Nuclear Physics Institute in Řež. The total pledged capacity for the WLCG was 16500 HEPSPEC2006 units and 2630 TB of disk space in 2015. Another 10000 HEPSPEC2006 units were available for the D0 experiment. Belle experiment and astroparticle physics experiment Pierre Auger Observatory. They used mostly the NGI/EGI sites for simulations and tests of production framework with a total capacity 920 jobslots.

Special international connection which was in the past provided by CESNET to HEP via dedicated links to the Tier2 Centre in Karlsruhe, Fermilab (NOvA experiment) and to BNL (STAR experiment and later for connection to the BNL's ATLAS Tier1 center) were replaced by LHCONE infrastructure.

Plans for the future:

In near future, following changes are expected.

1. Increasing capacity for the access to the CESNET2 backbone by adding 100GE/40GE network interfaces to the IP/MPLS routers. Further extension of advanced high-speed network services.
2. Provisioning of ultrastable optical frequency via photonic service for NPI, Řež
3. Continuing work on extending the SDN paradigms mainly towards the optical layer.
4. The renewal of capacity and further upgrade of the distributed computational infrastructure of MetaCentre and continuous development of its advanced services.
5. Further extension and modernization of high capacity distributed storage system and continuous development of its advanced services.
6. Further modernization of the infrastructure for team collaboration.

Annex 15: SANET (Slovakia) Status and Plan

Submitted by Tibor Weis (tibor@tuzvo.sk)

January 2017

SANET:

The SANET network infrastructure in Slovakia covers all Slovak Universities and institutions of the Slovak Academy of Science in 37 towns. The network is completely built up on single mode dark fiber leased among the cities. The network is configured as several rings providing full redundancy.



Figure 60: SANET Network Map 2016

In 2016 SANET completed major national infrastructure upgrade, which provides its members with N x 100 GE capacity. Transport is provided by Infinera CloudXpress CX100E point-to-point DWDM system (initially supporting 2 x 100GE bandwidth) which is connected to 3.2 Tbps TRILL switches (Huawei CE8860, based on BCM Tomahawk ASICs). TRILL enables SANET to have all links always in forwarding state and provides IP-like shortest-path routing for L2 ethernet packets, optimal utilization of all core links and fast convergence - without the need for MPLS.

In the near future, SANET is planning to install Infinera DWDM system on additional links in order to establish fully resilient N x 100GE backbone.

Annex 16: PIONIER (Poland) Status and Plan

Submitted by Marek Bazyly (bazyly@man.poznan.pl)

Poznan

January 2015

Introduction:

Poznan Supercomputing and Networking Center is the operator of the polish optical national research network, PIONIER. The network (*Figure 62* and *Figure 63*) consist of approx. 6500 km of dark fiber inside Poland and has international connections to research and education networks in Germany, the Czech Republic, Slovakia, Ukraine, Belarus, Lithuania and Russia. PIONIER dark fiber network interconnects 26 major university centers in Poland (Poznań, Warszawa, Kraków, Wrocław, Łódź, Gdańsk, Katowice, Szczecin, Toruń, Bydgoszcz, Zielona Góra, Białystok, Lublin, Częstochowa, Kielce, Koszalin, Radom, Rzeszów, Opole, Olsztyn, Puławy, Suwałki, Gorzów, Elbląg, Zamość and Bielsko-Biała).

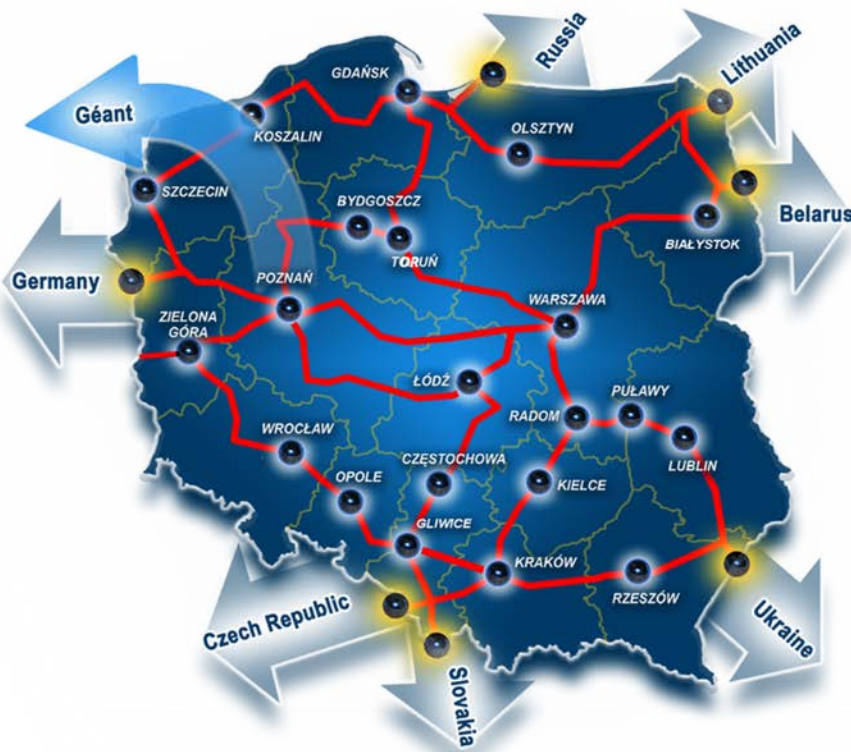


Figure 61: National Research and Education Network - PIONIER (topology).

Each fiber line of PIONIER (with exception of Poznań-Łódź and Warszawa-Toruń line) consists of several fiber pairs. The available fiber types include G.652 standard single mode fibers and G.655 NZDS fiber.

PSNC has a DWDM transmission system ADVA running on all its fiber lines. The system is equipped with cards for one or two 10Gbps wavelengths. Each of the wavelengths can carry either

STM-64 or 10GE LAN-PHY signals. On each link one wavelength is used for Internet traffic while the other one is used for direct interconnections between High Performance Computer centers and for experiments.

Recently DWDM system was extended with ROADM support (colorless, directionless, contentionless) and additional transponder modules. In each of the university cities PSNC has a MPLS LSRs (Foundry NetIron XMR 16000/8000, Juniper MX960) equipped with 10GE and 1GE interfaces. The connections between the backbone switches are always 10Gigabit Ethernet (LAN-PHY or WAN-PHY) over ADVA DWDM links.

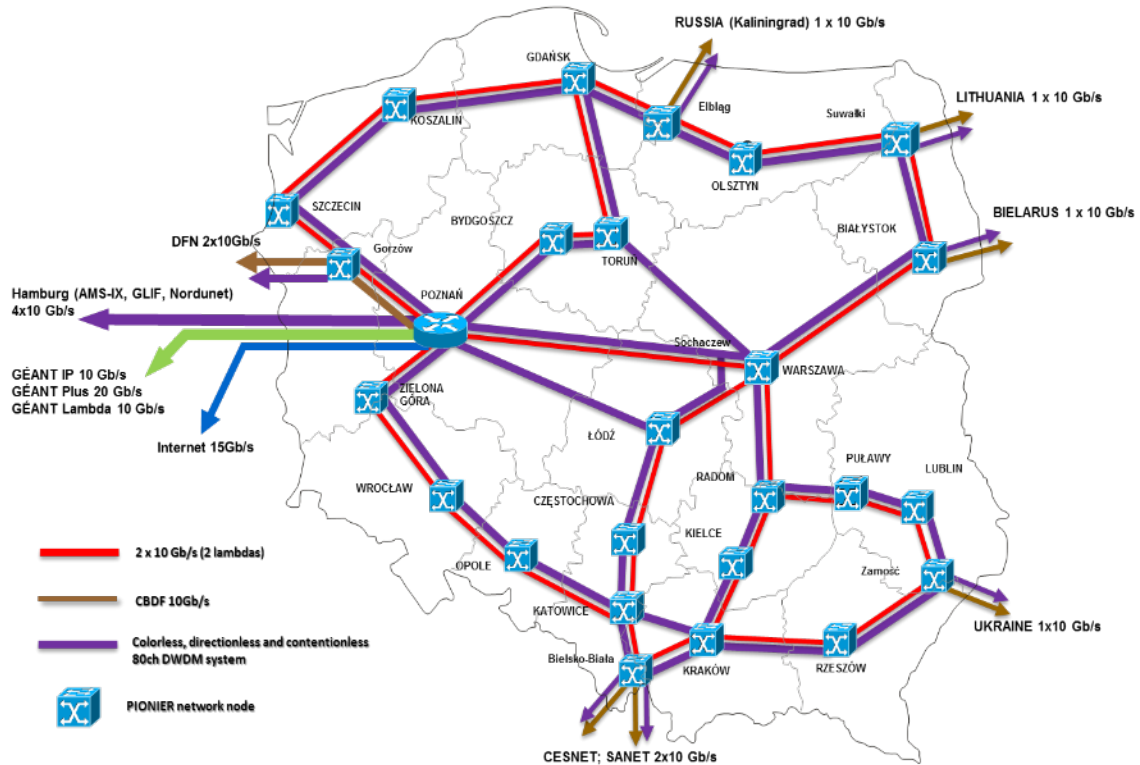


Figure 62: National Research and Education Network - PIONIER (transmission).



Figure 63: PIONIER Network in Hamburg.

The PIONIER network DWDM system is also installed on fiber line to Hamburg (Figure 64). The DWDM system capacity allows to carry 40 channels up to 10Gbps each. In Hamburg PIONIER is directly connected to the national research networks SURFNet and NORDUnet. PIONIER is also a member of an international virtual organization GLIF, the Global Lambda Integrated Facility. Additionally PIONIER is a member of AMSIX exchange with 10GE access interface.

The PIONIER network provides high speed transmission for academic and research community in Poland. The MPLS technology available in entire network, allows to implement flexible services like VPLS, Virtual Leased Lines and L3VPNs.

In Poznan there are located routers for international IP connectivity (Juniper T320 series and MX960 series). The routers provides redundant and reliable access to Internet for polish academic and research entities. PIONIER maintains two Network Operating Centers. The main NOC is located in Poznan and the backup NOC (operating in standby mode) is located in Łódź. The transmission performance is constantly monitored with dedicated servers. In case of any problem the Agilent N2X router tester can be used to measure the transmission parameters.

Most of the HPC Centers in Poland are involved in EGEE project. As Regional Operation Center in EGEE we cooperate with other EGEE partners from Central Europe region sharing experiences and knowledge related to the grid computing and massive storages which are developed and maintained in scope of the project.

Basing on this knowledge and using our national PIONIER Network, POLTIER2 consortium has been created as "Polish computing system Tier2 for LHC experiments". HPC centers from Poznan

(PSNC), Warsaw (ICM) and Cracow (CYFRONET) established logical Tier2 layer for LHC experiments using dedicated VLAN links between these centers based on the PIONIER infrastructure. POLTIER2 network is then connected via cross-border dark fiber link Poznan-Karlsruhe which gives us a high bandwidth (10Gbps) network connection between Polish Tier2 and Tier1 center in Karlsruhe.

Internal dedicated links between polish computing centers and dedicated link between PIONIER and DFN network, gives us possibility of huge, stable and reliable data transfers on the Tier2-Tier1 path required by LHC experiments (Figure 65)

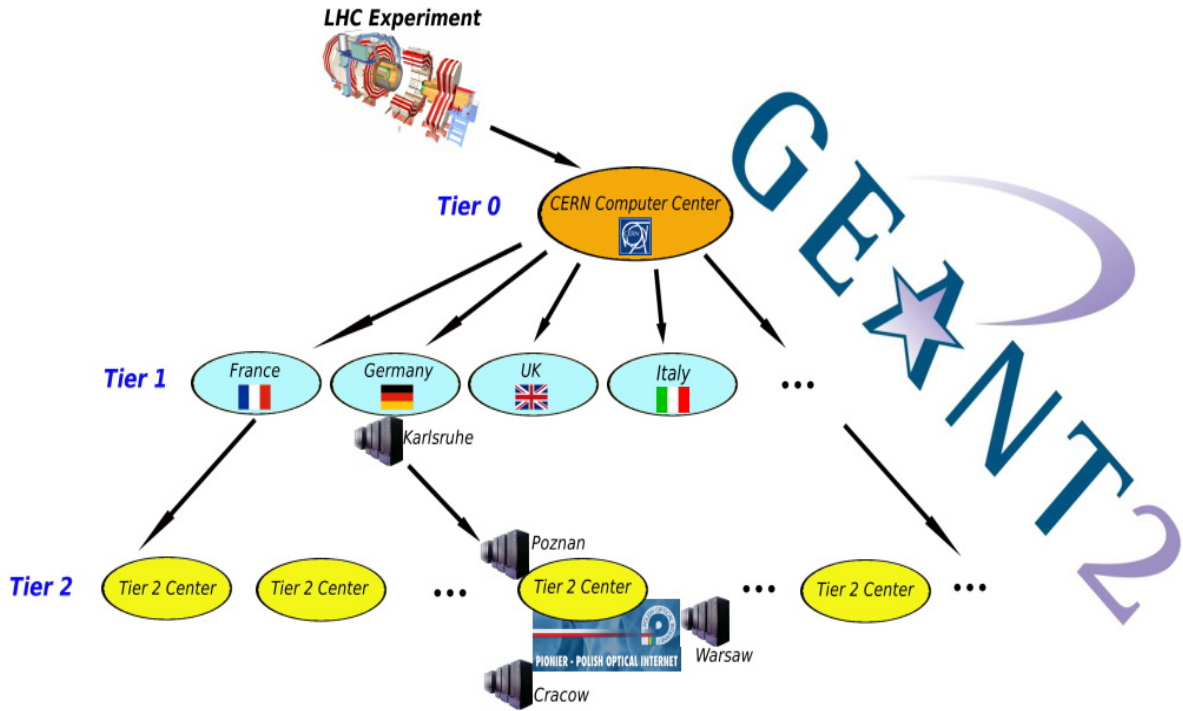


Figure 64: Tier1 - Tier2 Paths.

PSNC provides access to PRACE project network for polish HPC centers. The PIONIER network is connected to PRACE switch located in Frankfurt /Main with 10GE lambda provided by GEANT (Dante). (Figure 66)



Figure 65: PIONIER network is connected to PRACE.

Annex 17: RoEduNet (Romania) Status and Plan

*Submitted by Dr. Octavian RUSU (octavian@roedu.net), CEO RoEduNet,
Paul Gasner (paul@roedu.net),
January 2015*

Introduction:

Networking services for research and education in Romania are provided by RoEduNet, the Romanian NREN since 1998 when the organization has been officially established through Romanian government decision 515/1998. RoEduNet, as Romanian NREN, is member of the GÉANT consortium (www.geant.net), TERENA (www.terena.org) and CEENet (www.ceenet.org) associations. Starting with 2006 there is a unique network infrastructure in Romania, to serve research and education communities, fact officially established through the government decision 1609/2008.

Networking services for academic and research community in Romania were provided before RoEduNet by some Universities within the country, Universities that host the main Network Operation Centers (NOC) of the current RoEduNet infrastructure. The most important milestones that lead to Romanian NREN – RoEduNet (using a bottom-up approach) are:

1. 1990 – joint research project to introduce e-mail service by Politehnica University in Bucharest and Technische Universität Darmstadt Germany;
2. 1991-1992 – joint research project between Politehnica University Bucharest and Deutsches Forschungsnetz (DFN - Verein) – first dedicated international connection Bucharest – Darmstadt;
3. 1993 – First national dedicated connections from Iasi and Cluj-Napoca to Bucharest (costs supported by Universities);
4. 1996 – The network is recognized by the Ministry of Education and was defined in the strategy as RDIS (Romanian Higher Education Network);
5. 1998 – The official born of RoEduNet as institution and the network for research and education under the auspices of Ministry of Education;
6. 2008 – RoEduNet2 project implemented for all national links and about 90% of the regional links (local loops to be installed) at the end of the year;
7. December 2009 – administrative reorganization of the RoEduNet - establish the Agency ARNIEC/RoEduNet in charge for the management of the Romanian NREN, the NREN status of the Agency was clearly stated in the government decision;
8. 2009 – RoEduNet2 DWDM network in intensive testing in the first month of the year and into production state starting with February.

According to its statute, Romanian NREN (through its network named RoEduNet) provides data communication services for research and education in Romania and provides connectivity to the GÉANT network and to the Internet for research and academic community within the country. Also, RoEduNet facilitates research in its own right in the field of data communication, participating into research projects and providing experimental test beds to implement new services and advanced network technologies.

NREN status in January 2015

Communication infrastructure

The communication infrastructure of Romanian NREN is based on dark fiber that belongs to the state owned company Telecomunicatii CFR and DWDM equipment owned by Agency ARNIEC/RoEduNet. The DWDM network is called RoEduNet2.

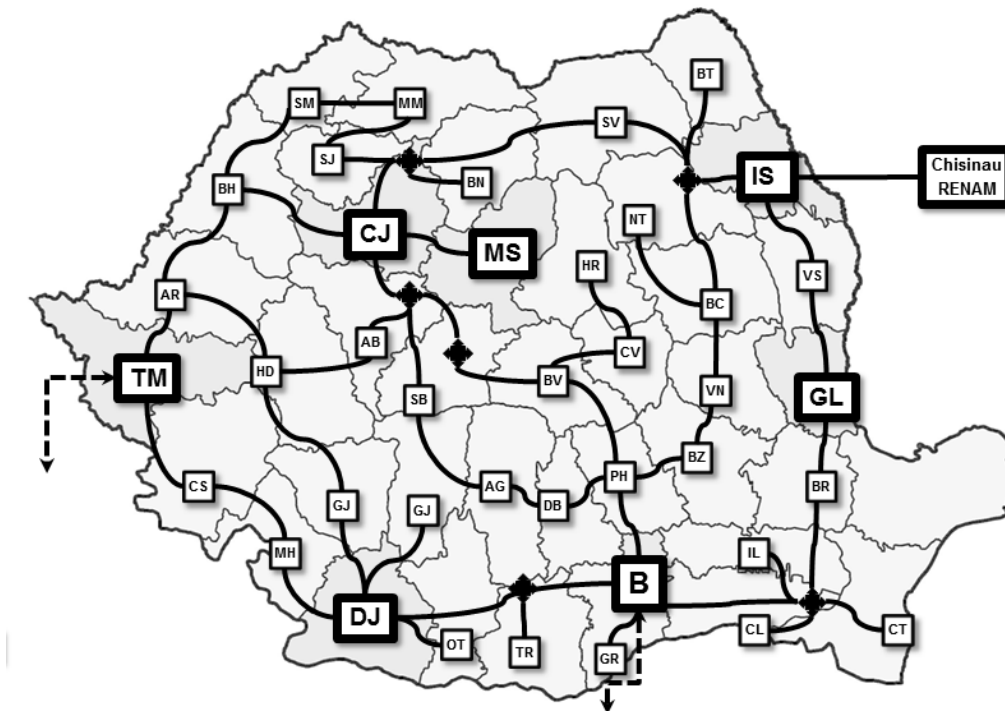


Figure 66: DWDM network including extensions installed in 2011.

The total length of the fiber is 5636 km and connects 40 counties capitals out of 41. Dark fiber footprint of RoEduNet2 network is presented in Figure 67. There is one county capital not covered because there is no fiber installed there, this city is connected using 100 Mbps leased line. The fiber is lighted up using DWDM technology. DWDM equipment are installed in a number of 56 sites, as follows:

- Reconfigurable Optical Add and Drop Multiplexers (ROADM) are installed in all locations where there are more than two directions: 18 sites. This technology was used to allow dynamic reconfiguration of the lambda topology for the whole network. All sites installed in the main NOCs of RoEduNet use also ROADMs to allow the expansion of the network in the metropolitan area in the main cities. This will enable RoEduNet to provide in the future lambdas on demand for the institutions connected to these main NOCs as well as future opportunities for scientific research on optical switching.
- Optical Add and Drop Multiplexers (OADM) are installed in all cities where there are only two fiber directions available: 36 sites. All OADM sites are connected using usually two 10 Gbps lambdas to two main NOCs (this arrangement provides backup for all institutions connected through OADM sites)
- Optical amplifiers sites are used for fiber segments that exceed 100 km, a number of 15 such sites are installed.

It should be noted that as a result of a NATO Science Project an extension of the RoEduNet2 network was installed to Chisinau, Republic of Moldova, to provide GÉANT connectivity for RENAM, the Moldavian NREN. As a result of this project the first cross border fiber connection is installed and RENAM has access to the RoEduNet2 network to install their lambdas to the GÉANT POP in Bucharest. Also, there are another two sites prepared to host cross borders with Serbia (Timisoara) and Bulgaria (Giurgiu).

There are four types of lambdas installed in the network:

- 100 Gbps Ethernet: six lambdas, from Bucharest to NOCs in Iasi, Cluj, Timisoara, Galati, Tg. Mures, Craiova, and Magurele;

- 10 Gbps Ethernet: 35 lambdas, RoEduNet is using 19 of them to interconnect the main NOCs providing, at least, two directions backup for each NOC;
- 3 STM-64 lambdas used by Telecomunicatii CFR;
- 54 lambdas of 10 x 1Gbps Ethernet. These lambdas provide 540 circuits of 1 Gbps, 432 circuits are fully equipped; almost each site is able to host another 2 circuits if necessary (sites at the end of branches have circuits only to the branching point). RoEduNet is using a number of 216 1 Gbps circuits for its POPs. It should be noted that each POP is connected using 1 Gbps circuits to two NOCs including the branches.

The calculated metric, according with TERENA compendium for the whole RoEduNet2 network is 386857 Gbps*km calculated by summing all circuits bandwidth multiplied by each circuit length. In fact, the total metric of the DWDM network 240037 Gbps*km but a number of lambdas are used by our partner in the project Telecomunicatii CFR.

The topology of the communication circuits used by RoEduNet is presented in Figure 68. There are 100 Gbps circuits marked with purple color, red lines are 10 Gbps Ethernet backbone circuits and blue lines are 10 Gbps lambdas (10 Gbps Ethernet or 10x1 Gbps Ethernet circuits) to connect the POPs to the NOCs. The connection from Iasi to Chisinau (to connect RENAM in Chisinau) was put into production starting with May 2010.

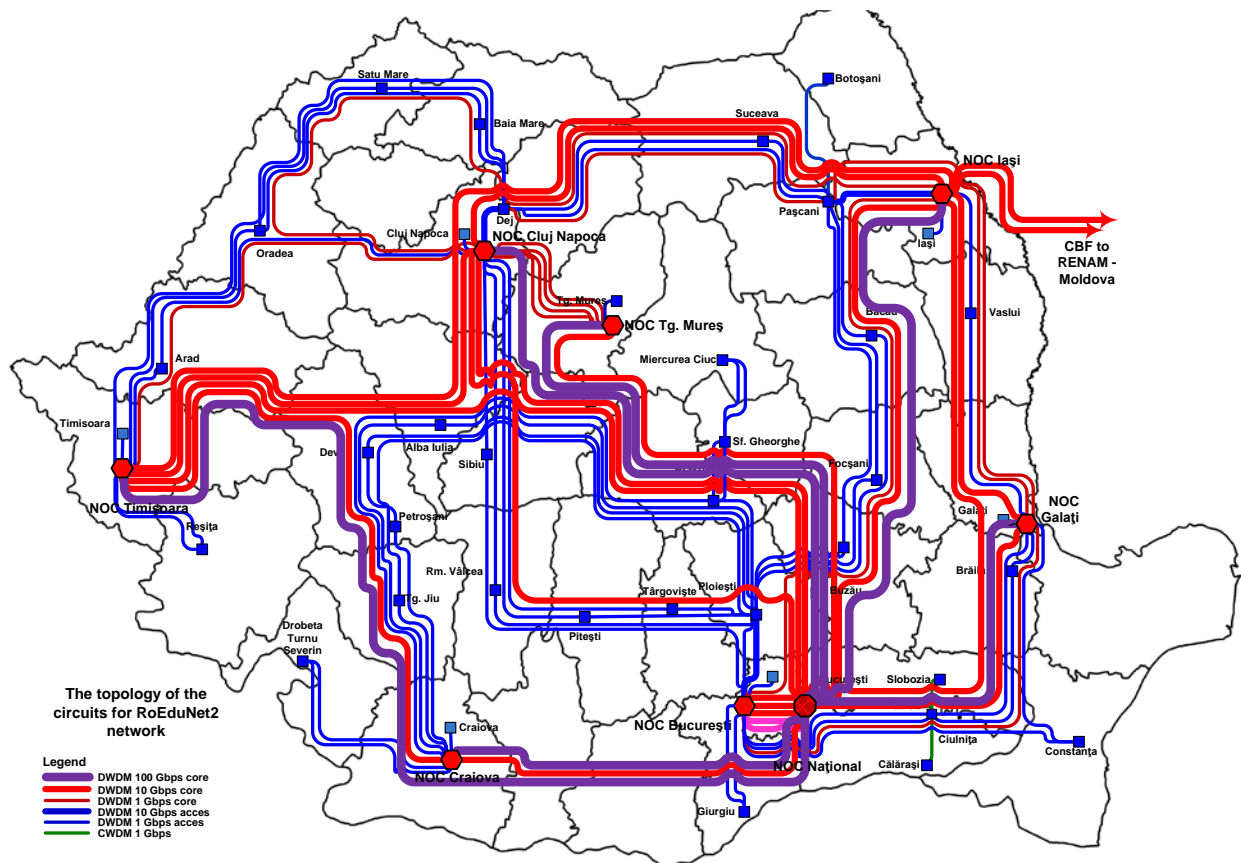


Figure 67: RoEduNet2 network: the topology of the lambdas for backbone and access networks.

Apart from RoEduNet2 network all POPs are connected also using leased lines, circuits of 100 Mbps leased from other telecom operators in Romania. All of these lines are used for backup.

International connectivity for Romanian NREN consists of three 10 Gbps circuits: one for the connection to the Bucharest POP of the GÉANT network (hosted by RoEduNet in the data center of the National NOC), one to Level3 POP and the last one to Cogent POP in Bucharest. The connections with Level3 and Cogent were installed as part of the Global IP Services - commercial traffic for NRENs negotiated by DANTE, the operator of the GÉANT network.

To minimize the commercial traffic to/from the network, RoEduNet installed more than ten peering connections with Romanian ISPs. Also, RoEduNet is present in Romanian Internet Exchanges such as InterLAN, RoNIX and Balcan-IX. It should be noted that the traffic through these connections is about two times greater than the traffic through the international links. The medium value of the total external traffic of the RoEduNet network (including peering, GEANT and Global IP Services) is around 50 Gbps.

The backbone of the network has been massively upgraded in 2012 by installing 100 Gbps lambdas for all NOCs in the country as presented in Figure 69. Because there are so many installed circuits it is difficult to have a clear view of the backbone network. The backbone consists of four rings as presented in Figure 69:

- one small ring in Bucharest linking on two paths the National node with Bucharest node with three 10 Gbps lambdas and one 100 Gbps circuit;
- three rings in the country each one connecting two regional nodes connected to national node using one 100 Gbps lambda and at least one 10 Gbps lambda (smaller nodes are connected using one 10 Gbps lambda, bigger nodes – Iasi, Cluj Napoca and Timisoara are connected with two 10 Gbps lambdas to the national node). It should be noted that all three rings are interconnected each other not only in the national node, they are connected also to each other through bigger nodes.

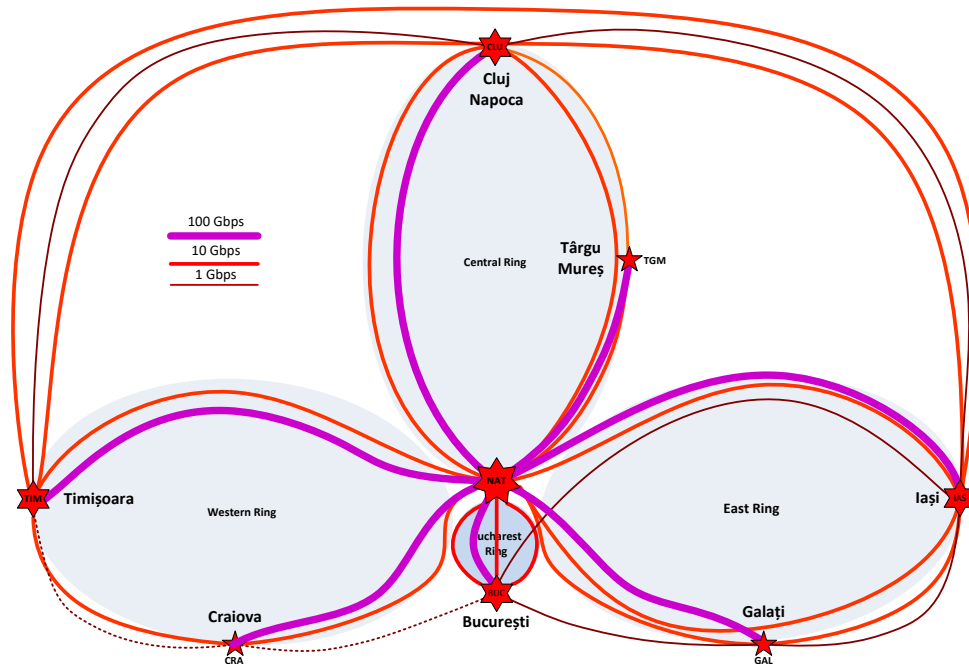


Figure 68. RoEduNet2 - backbone topology

An important location for the research community in Romania is Magurele, near Bucharest (not on the map), where are located the most important GRID sites and the upcoming ELI project. In 2013 Magurele POP was connected at 10 Gbps using dark fiber and lighted up with Ciena DWDM

equipment integrated in the backbone of the network. It should be noted that in 2014 the link was upgraded to 100 Gbps.

Another major upgrade in 2014 is the replacement of core routers in all NOCs with Cisco CRS – CRS 4/S in Bucharest and Tg. Mures and CRS-8/S-B for Iasi, Cluj-Napoca, Craiova, Galati, Timisoara, Magurele.

Networking services

A full range of IP services are provided to all connected institutions, including IPv6. Due to the legal issues, VoIP is not provided as a service but it is used for communications between NOCs within RoEduNet. Also, for some research projects support for VoIP traffic has been offered to the universities and other institutions involved into this project.

Special circuits were installed and tested for the Romanian GRID community (Romanian Tier 2 Federation). VRF technology is used to separate this traffic and connect to GEANT LHCONE VRF network. At this moment there are five 100 or 10 Gbps connections in production service for GRID communities (directly connected to the NREN infrastructure):

1. RoEduNet to IFIN to connect the GRID sites and VOs from Institute for Space Sciences (ISS) ISS (ALICE / AliEn) and “Horia Hulubei” National Institute of RD for Physics and Nuclear Engineering (IFIN-HH): NIHAM (ALICE / AliEn); NIHAM (ALICE / GLITE), RO-02-NIPNE (ATLAS, HONE / GLITE), RO-07-NIPNE (ALICE, ATLAS, LHCb, SEE, SEEGRID / GLITE) and RO-11-NIPNE (LHCb) – 100 Gbps (10x10 Gbps);
2. RoEduNet to “Politehnica” University in Bucharest to connect the site RO-03-UPB, UPB (ALICE / AliEn) – 10 Gbps;
3. RoEduNet to “Alexandru Ioan Cuza” University in Iasi to connect the cluster from the Digital Communications Department (RO-16-UAIC) – 10 Gbps;
4. RoEduNet to Technical University from Cluj Napoca to connect the GRID cluster of the University part of the SEE-GRID project – 10 Gbps;
5. RoEduNet to ITIM Cluj Napoca to connect the GRID clusters within Romanian Tier 2 Federation – 10 Gbps.

Other GRID sites in Romania are connected using gigabit links: National Institute of RD in Informatics from Bucharest (ICI) RO-01-ICI (LHCb, BIOMED, SEE, SEEGRID, GLITE) and Romanian Space Agency.

A list of the available services, except standard IP services like DNS, mail relay, routing and switching is presented below:

1. Centralized Helpdesk within NREN available in the normal working hours (9 a.m. to 5 p.m.)
2. Trouble Ticket system accessible for all connected institutions in operation;
3. Computer Security Incident Response Team – RoCSIRT:
 - a. established in 2008;
 - b. entered in operational state in January 2009;
 - c. accredited by Trusted Introducer (<http://www.trusted-introducer.org>) in August 2009;
 - d. first official participation at Terena CSIRT task force (TF-CSIRT) meeting in September 2009, official partner since then.
 - e. FIRST (<http://first.org/>) membership affiliation process started in 2013.
4. Centralized management software for the entire network, including the management of the DWDM equipment and establish the VNOC (Virtual NOC) structure for the network operation and introduction of the new services; network management is based on open source software and HP OpenView;

5. Public access to all the traffic graphs for all equipment within all NOC's (using weathermap and NMIS);
6. Looking glass available to all users worldwide;
7. Digital certificates services for connected institutions, using TERENA TCS (<http://www.terena.org/activities/tcs/>)
8. eduroam - the service is available at the national level, the nationwide RADIUS server being deployed and registered upwards, 5 universities are offering wireless access for eduroam participants;
9. Videoconferencing service using dedicated equipment (110 HD Tandberg terminals and 2 MCU units acting as gatekeepers – project implemented in 2010);
10. VoIP services for all connected institutions, so far two Universities have full connectivity and there is installed service in all POPs to be used in the relation between these and the Ministry of Education;
11. Secure intranet for the Ministry of Education to accommodate online services provided by this authority for all schools and high schools in Romania;
12. Online Archive Mirroring Server – available since 2002, now is offering 6TB of mirrored content (70 mirrors, 6 million files) to connected institutions. Estimated traffic value is around 1TB/day.
13. Hosting and operation of Anelis infrastructure – national hardware infrastructure for scientific national repository and access to scientific information and documentation literature for all universities and research institutes.

Research activities

RoEduNet is a member of the GÉANT consortium and has been involved in the activities of the GN3plus project contributing to a several number of tasks especially on security area.

Agency ARNIEC/RoEduNet is the coordinator of another four POSDRU projects (financed through structural funds) and organizing networking and computer related courses for the employers of schools and high schools for a better use of IT labs already installed in those institutions.

Each year RoEduNet organizes an international conference, which now (2014) is at the 13th edition. The purpose of this conference is to bring together the academics and the companies from Romania and from abroad for a discussion about computer networking, in its technical, social and strategic aspects, with a special focus on directions and applications in education and research. The 13th edition was organized as joint event with RENAM, Moldova and the proceedings of the 13th RoEduNet conference are published in IEEE Explore. Next year (2015) conference is proposed to be organized together with in Craiova, hosted by the University of Craiova.

Annex 18: RENATER (France) Status Update

RENATER, 23 rue Daviel 75013 Paris – France

Submitted by Frederic Loui (frederic.loui@renater.fr), Sabine Jaume (sabine.jaume@renater.fr), Patrick Donath (patrick.donath@renater.fr)

February 2015

Update on the Research Networks in France

There is a unique network infrastructure in France, to serve Research and Education communities. This infrastructure is made on one side, of a national backbone operated by GIP RENATER, and on the other side of a collection of regional/metropolitan networks, connected to RENATER backbone. RENATER is providing both international connectivity mostly through the pan-European backbone GEANT, and commodity traffic through two dedicated access to a worldwide IP Transit network and two accesses to major Internet exchange points in Paris : SFINX (operated by RENATER) and FranceIX.

Key updates

Changes done in 2014 are mainly dedicated HEP research projects, involving LHC. 7x10GE wavelength have been deployed between T1/T2/T3 LHC and LHCONE data center. The total capacity in the French Area at HEP disposal has reached 190 GE (19 Wavelengths). Layer 3 view of HEP in France is shown in the Figure 73.

In term of services, 6VPE has been enabled within LHCONE FRANCE, in order to cope with the IPv4 address shortage. This is due to the tremendous numbers of Virtual Machines within T1/T2/T3. This has been done in synchronization with all of the NRENs and GEANT.

In terms of traffic consumption, the total aggregated traffic related to RENATER has reached 410 PBytes (IN+OUT).

In terms of incoming traffic, the proportion related to HEP traffic has reached 30% of the total incoming traffic $((410/2)*0,30) \Rightarrow 61,5$ PBytes.

In terms of outgoing traffic, the proportion related to HEP traffic has reached 37% of the total outgoing traffic $((410/2)*0,37) \Rightarrow 75,85$ PBytes.

These numbers represent minimum values, as the proportion of traffic to and from the GÉANT network is not taken into account. HEP prefixes are also advertised within the GEANT IP and is part of the R&E IP basic IP traffic, which corresponds to approximately 102,5 PBytes IN+OUT).

In 2015, the planned evolutions are:

- Deployment of the new 100GE capable transmission equipment.
- Deployment of the new 100GE capable routers in the major route PARIS-LYON-MARSEILLE within the core.

However, considering the current evolution of HEP traffic level and also the amount of RAW data created by the upgraded accelerator T0, French T1/T2 sites might upgrade their ports to 100GE this year.

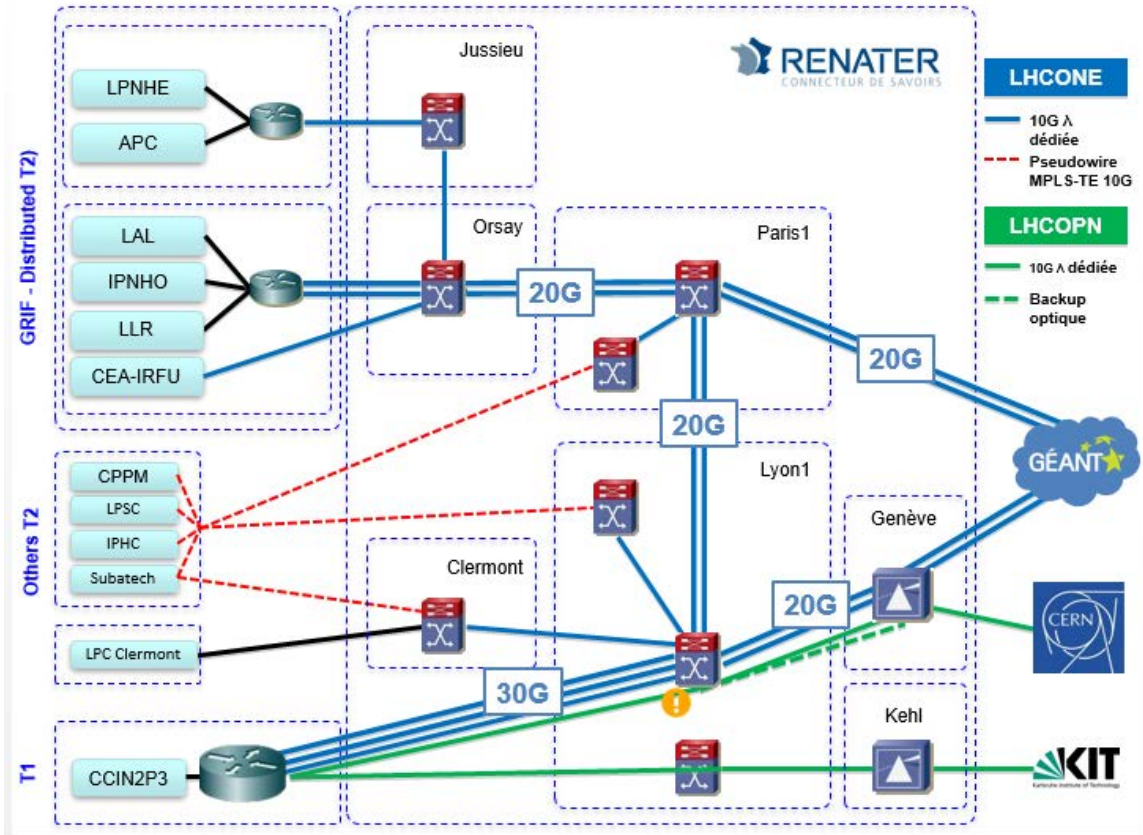


Figure 69: French architecture dedicated to HEP in France.

Annex 19: RNP (Brazil) Status Update and Plan

Submitted by Michael Stanton, RNP (michael@rnp.br)

January, 2017

RNP national backbone network (Ipê network)

The Brazilian National Research and Education Network – RNP – commissioned its Phase 6 Ipê backbone network in May, 2011, which suffered further significant alterations in 2016, leading to the topology shown in Figure 1.

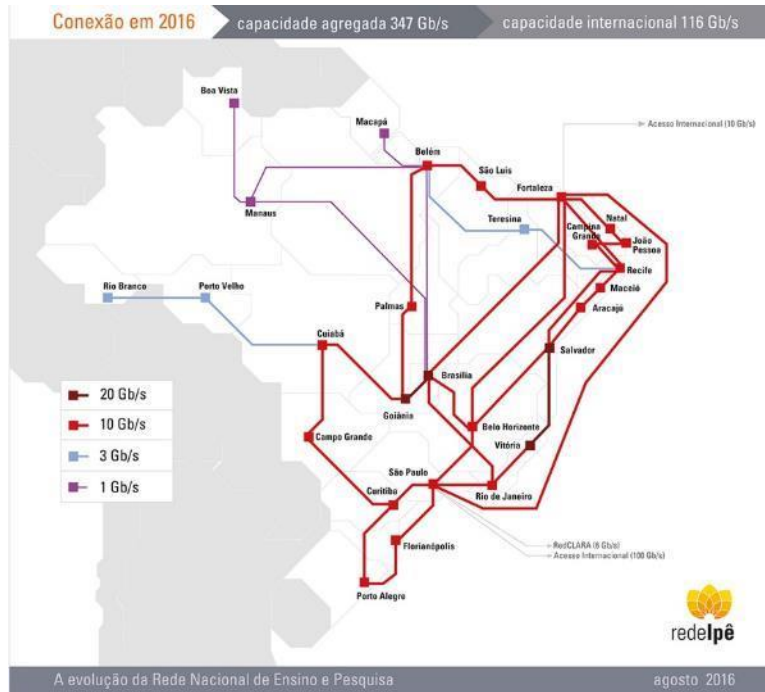


Figure 70: Configuration of the Phase 6 Ipê backbone of RNP in 3Q 2016

Major novelties include (1) the upgrade to 1 Gbps of the links from Belém to Macapá and to Manaus, and from Manaus to Boavista; (2) the removal of the landing point in Rio de Janeiro of the SAC submarine cable between Fortaleza and São Paulo, resulting in a direct 10 Gbps backbone link between these two cities; and (3) the decommissioning of the link between Boavista and Fortaleza which used the GlobeNet international submarine cable.

The discussion of the future of the national infrastructure used by RNP has continued, especially in relation to the requirement to support 100G lambdas starting in 2017, and scalable capacity in the medium and long-term (see below). Thus RNP is seeking to acquire long-term rights in terrestrial fiber infrastructure to meet its coming needs, and has already signed a cooperative agreement with CHESF (Companhia Hidro Elétrica do São Francisco, an electrical energy generating company, which owns extensive optical fiber (OPGW) assets in the Northeast of Brazil, and belongs to the federal government-controlled Eletrobrás group. By this agreement, RNP and CHESF share the use of CHESF optical fibers, using equipment acquired by RNP. In principle RNP acquires the right to use 50% of the available optical spectrum on the shared fibers. Similar agreements are being sought with other optical fiber owners in other parts of Brazil, most urgently in the Southeast and South, which will enable the transition by 2019 to a scalable backbone

structure between Fortaleza, in the state of Ceará, and Porto Alegre, in the state of Rio Grande do Sul, as well as a ring connecting the major cities in the Southeast and Center-West regions, which shown as 100 Gbps links in Figure 2.

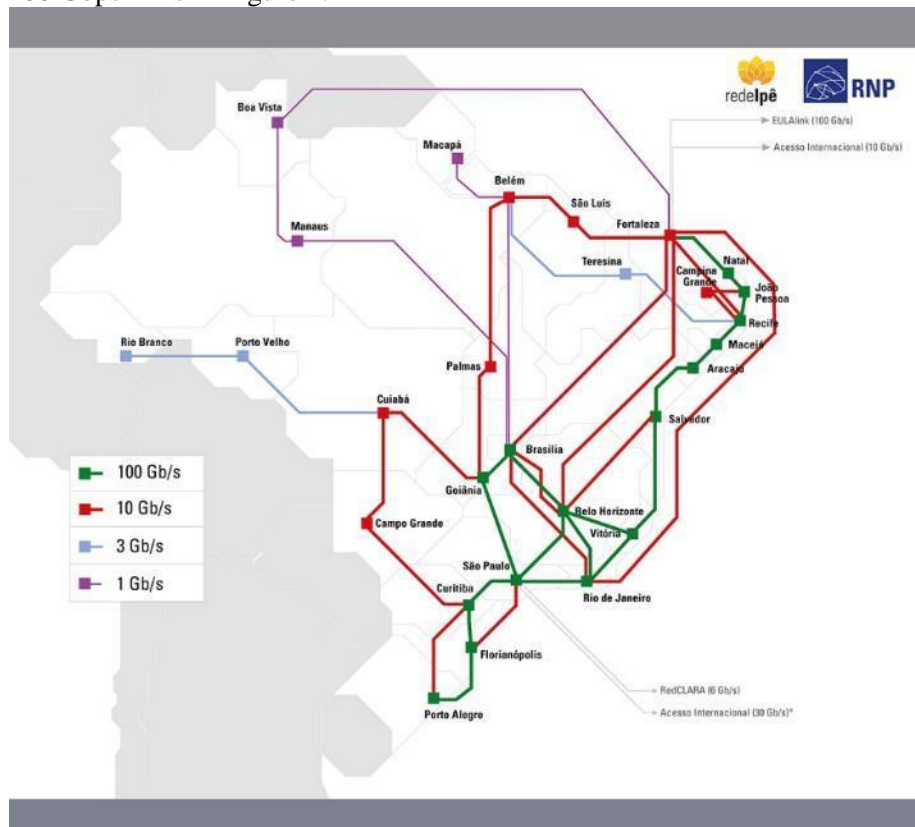


Figure 71: Expected configuration of the Phase 7 Ipê backbone of RNP by 2019

Subfluvial network infrastructure in Amazonia

As mentioned in 2016, the most significant novelty affecting connectivity in the Amazon region is the partnership with the Brazilian Army, called Amazônia Conectada (Amazonia Connected), <http://www.amazoniaconectada.eb.mil.br/eng/>, whose objective to install *Infovias* (data highways) along the courses of the following rivers in western Amazonia: Negro, Solimões (or Upper Amazon), Juruá, Purus and Madeira, with a total length of over 7000 km, as shown in Figure N+2. Work has already begun on installing the first of these infovias along the Solimões, which extends upstream from Manaus to the border with Peru and Colombia, a distance of over 900 km. The first section was laid between the cities of Coari and Tefé, a distance of around 240 km, and was concluded in April, 2016 (some photographs available in Figure N+3). The next two sections to be installed will be between Manaus and Coari, and between Tefé and Tabatinga, respectively.



Figure 72: Proposed Infovias (data highways) of the Amazonia Connected programme (red lines indicate existing fiber; blue lines are proposed sub-fluvial fiber)



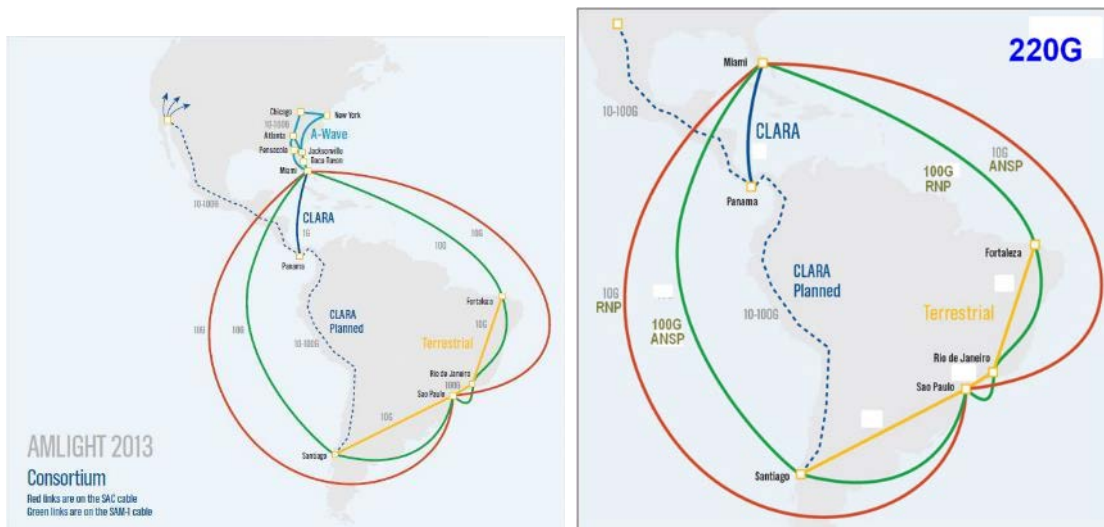
Figure 73: Cable-laying between Coari and Tefé on the Solimões: photos of the equipment used in 2016 (Images from <http://www.amazoniaconectada.eb.mil.br/eng/>)
Above: prow and stern of cable laying barge and tugboats;
Below left: deck view with one of the reels of cable;
Below right: paying out the cable from the stern

International connectivity for Brazil

Since 2009, connectivity has been provided between Brazil and the US A (Miami) using multiple 10G links by what is now called the Amlight consortium (RNP-ANSP-FIU) (see www.amlight.net/). Starting in 2009, this consortium provided 20G of aggregate capacity, using 2 independent submarine cable links, for use by RNP and the state network in São Paulo, ANSP, which is funded by FAPESP. The RedCLARA network (see below) also uses a fraction of this capacity. The US partner is Florida International University in Miami, which has received funding from NSF for providing connectivity to Brazil through awards from successive versions (2004, 2009, 2014) of the International Research Network Connections (IRNC) program. AMPATH, the Miami exchange point, is connected to other US and international networks using FLR (Florida Lambda Rail) and Internet2's infrastructure.

In 2013, the capacity was doubled, through the use of 4 independent submarine cables (shown in Fig N+4 (a)). Before this upgrade, the Brazil endpoint was located in São Paulo, but since then intermediate points have been activated on one of the cables in the cities of Fortaleza and Rio de Janeiro, providing redundant connectivity to existing terrestrial links of RNP's Ipê backbone, through the incorporation of the cable segments between São Paulo, Rio de Janeiro and Fortaleza. The landing in Rio de Janeiro was suspended in 2016, as mentioned in the first section of this report. Additionally, the one of the cables on the Pacific route also included a landing in Santiago, which is used by ANSP and the Chilean national network, REUNA.

In 2016, a significant change was made to this configuration with the upgrading to 100 Gbps of the two 10 Gbps links already in use on the LANutilus cable, shown in green in Figure N+4 (b), with expected intermediate landings in Fortaleza and Rio de Janeiro (for RNP) and Santiago (for ANSP). Thus the aggregate capacity available to RNP and ANSP will increase by a factor of 5.5, from 40 to 220 Gbps. Further extensions to this capacity are expected in 2017, with the commissioning of the Monet cable between Florida and Santos (see below).



$4 \times 10\text{G} = 40\text{G}$ (from 2013)

$2 \times 10\text{G} + 2 \times 100\text{G} = 220\text{G}$ (from 2016)

Figure 74: Links between Brazil and the US since early 2013 (courtesy of Chip Cox)



Figure 75: RedCLARA topology and capacities in 2016

The other international links are provided by the RedCLARA network, the Latin American regional academic network, which had received funding from the EC between 2004 and 2012. The current network topology is shown in Figure N+5, and shows the current situation after the 4Q2015 increase from 2.5 to 10 Gbps of the RedCLARA backbone, and the inclusion of a backbone node in Bogotá, Colombia. It should be noted that the connections shown in this figure as the triangle US-BR-CL in fact use the Amlight consortium infrastructure mentioned above.

The considerable upgrades in the RedClara backbone and its extra-regional connectivity have led to a significant alteration in the management of how network bandwidth is accessed. A description of these changes is provided in an Addendum included at the end of this Appendix.

New submarine cables

A number of new generation (100G+ technology) submarine cable systems connecting Brazil to other continents have been announced, some of which will have the explicit participation of academic user communities. Some of these cables will operate on a shared ownership basis, where fiber spectrum is allocated to different owners, to be illuminated as the owner sees fit.

The first 100 Gbps cable, AMX-1, belonging to América Móvil of México, has already been in use since 2014, but is not widely used by companies outside the América Móvil umbrella. The remaining cables, which have been proposed and in some cases partially built, are summarized in Table 1 and depicted in Figure N+6.

Table 1: Future submarine cable systems announced to enter in service by 2018
(Sources: various)

| Cable | Owners | Year ready for service | Capacity | Length (km) | Landing points in Brazil | Other countries served |
|---|---|------------------------|----------|-------------|------------------------------|--------------------------|
| Monet | Google, Antel, Angola Cables, Algar Telecom | 2017 | 64 Tb/s | 10,556 | Fortaleza (branch) Santos | USA (Boca Ratón, FL) |
| Seabras-1 | Seaborn Networks | 2017 | 72 Tb/s | 10,500 | Fortaleza (branch) Santos | USA (New York) |
| South Atlantic Cable System (SACS) | Angola Cables | 2018 | 40 Tb/s | 6,165 | Fortaleza | Angola (Luanda) |
| Tannat | Google, Antel | 2018 | 90 Tb/s | 2,000 | Santos | Uruguay (Maldonado) |
| South Atlantic Inter Link (SAIL, formerly CBCS) | Camtel, China Unicom | 2018 | 32 Tb/s | 5,900 | Fortaleza | Cameroun (Kribi) |
| BRUSA | Telefonica | 2018 | | 11.000 | Fortaleza, Rio de Janeiro | USA (Virginia Beach, VA) |
| Ellalink | Telebras, IslaLink | 2019 | 48 Tb/s | 9,501 | Fortaleza (branch) Santos | Portugal (Sines) |



Figure 76: Expected new cables in the South Atlantic, showing topology and landing points (CMR = Cameroon)

The Ellalink (Brazil-Europe) cable, now expected to be delivered in 2019, is being built by Eulalink, a joint venture between Telebras (Brazilian telco) and IslaLink (Spanish telco), to connect

Sines, Portugal, to Fortaleza and Santos. This will be the first cable directly connecting Europe with South America since Atlantis II, a non-Internet era cable delivered in 2000. Ellalink will create new international routes from South America, which will not touch North America. Two other transatlantic cables between Brazil and Africa, SAIL (formerly CBCS) to Cameroon and SACS to Angola, will also make possible new international routes avoiding North America, both to Africa itself and also to Europe, Asia and Oceania. There are also two new cables expected to connect Brazil and the US by 2017, providing direct links to Santos, with branch access to Fortaleza.

Some of these new cables will be of great use for data intensive science projects, which require the transmission of huge volumes of data around the world. These could include the transport of observational data from optical telescopes and radiotelescopes in the Atacama Desert of Chile, as well as earth observation data to and from South America and even radioastronomy data between the SKA observatory in South Africa and destinations in the Americas, both North and South.

In order to meet some of these needs, some progress has been made to acquire long-term access to new generation cables for science and education uses. The LSST (Large Synoptic Survey Telescope), under construction in Chile, will transport its data from South America using the Monet cable between Santos and the USA, and the LSST project, financed by the National Science Foundation, is acquiring spectrum on one of the fiber pairs, permitting the installation of up to 6 100G lambdas. This capacity will be divided between LSST demands, and also those of the Brazilian networks RNP and ANSP.

There is also considerable interest in Europe and South America in acquiring long-term access to scalable capacity for science and education traffic between the two continents. This may be implemented by acquisition of optical spectrum on the fibers of a submarine cable, which may be lit up on demand by the user organizations. Money is being made available, both by the European Commission and by South American academic networks, for the purpose of acquiring spectrum sufficient for about 45 lambdas on a direct transatlantic submarine cable and the building of a terrestrial access network in South America, which should form the future core of the RedCLARA network in this continent. The complete project is called BELLA (Building Europe Link to Latin America), and the terrestrial network is called BELLA-T. The lead organizations in the two continents are GÉANT and RedCLARA. A tender was initiated in 2016 to select a suitable submarine cable route for this investment, between Portugal and northeast Brazil. The probable candidate is the Ellalink cable, to be built by 2019 by Eulalink, which has the advantage of using the shortest path between the two endpoints.

A second new submarine cable, this time between Boca Ratón, Florida, and Santos (São Paulo) is already under construction. The cable, called Monet, has as its stakeholders Google, and three telcos: Algar Telecom (Brasil), Antel (Uruguay) and Angola Cables (Angola). Monet will also have an intermediate landing point in Fortaleza. This cable is expected to be operational in 2017. There will also be a stakeholder interest of the academic community in this cable, with the acquisition by the LSST (Large Synoptic Survey Telescope) project, using funding from NSF, of optical spectrum on one of the fiber pairs, to support six 100G lambdas, and there is a proposal for 2/3 of this capacity to be used by the Brazilian networks, RNP and ANSP, in exchange for other infrastructure support in Brazil or neighbouring countries for the LSST project (see www.lsst.org/lsst). The remaining 1/3 is to be used partly by LSST and partly by the Internet2 and RedClara networks to support their interconnection.

With the exception of Tannat, which may be seen as an extension of Monet to Uruguay, where Antel is one of the Monet investors, these new cables all originate in, pass through or possess a branch to Fortaleza, making this city an ideal location for an International Exchange Point for the South Atlantic region.

BELLA-T and the future RNP backbone

The business plan for the South American networks, coordinated by RedClara, includes building BELLA-T, the terrestrial network linking the cable landing in Fortaleza to the national networks in several countries in the continent. A possible configuration for BELLA-T is shown in Figure N+7.

As the submarine link from Europe to Fortaleza will consist of multiple 100G lambdas, it is imagined that BELLA-T will be a scalable optical network with a similar structure. This will require that the participating national networks of Brazil, Argentina, Chile, Peru, Ecuador and Colombia gain access to 100G capable fiber systems and make available for common use the requisite number of lambdas, in accordance with the needs of the user communities in Europe and South America. The number of lambdas is expected to increase throughout the lifetime of the submarine cable link. The estimated length, without redundancy, of the BELLA-T segment in each of these countries is shown in Table 2. Naturally, it is expected that steps will be taken in each country to provide redundancy of the national segment, to increase availability of the network.

Table 2: Lengths of national segments of BELLA-T optical infrastructure

| Country | Brazil | Argentina | Chile | Peru | Ecuador | Colombia | TOTAL |
|--------------|---------|-----------|---------|---------|---------|----------|-----------------|
| route length | 6223 km | 2500 km | 2000 km | 2594 km | 1330 km | 1803 km | 16450 km |

Within each country, a $N \times 100G$ lambda infrastructure will provide support for both the RedClara regional network and the national backbone network, with usually few nodes (perhaps only one) of the regional network in each country.

In the case of Brazil, the probable BELLA-T optical infrastructure will extend from Fortaleza to the southern border with Argentina. Such an infrastructure is currently being planned, in collaboration with a number of existing holders of fiber assets. Figure N+8 shows the probable topology of this infrastructure.

RNP’s national backbone network will also be a user of this infrastructure, as explained in the first section of this report, since the first two 100 Gbps international links from the USA to Brazil already arrived in São Paulo in 2016, and the arrival of such high capacity will require a correspondingly high capacity national backhaul network in order for the international link to be adequately utilized.



Figure 77: planned BELLA-T access network in South America

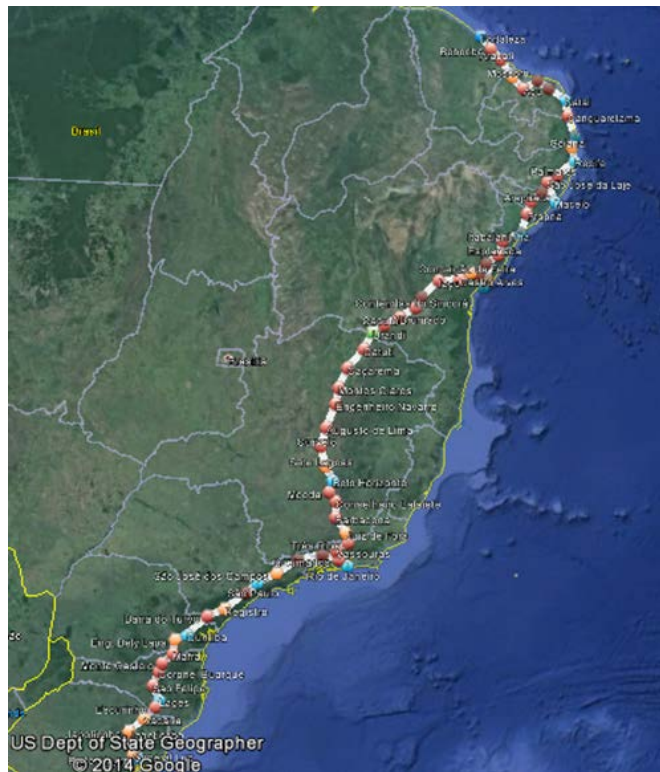


Figure 78: Probable geographical location of the BELLA-T infrastructure in Brazil
(Courtesy of Eduard Grizendi)

Addendum:

International bandwidth management in RedClara

Submitted by Marco Teixeira, RNP (marco.teixeira@rnp.br), Lead Engineer at RedClara

January, 2017

1. Motivation

The NEG (Network Engineering Group) of RedClara has implemented a new routing scheme on the backbone, meeting the following requirements:

- Each NREN will be able to use all the available Link Access Bandwidth on the access links to communicate within Latin America.
- Each NREN will have an upper limit, called the External Bandwidth, to traffic that is sent and received to/from outside Latin America, to communicate with Internet2, Geant or any other academic peering. This External Bandwidth will be used for billing purposes.

2. 2017 Bandwidth chart

| Country | NREN | Link Access Bandwidth | External Bandwidth |
|-------------|-----------|-----------------------|--------------------|
| Argentina | INNOVARED | 10 Gbps | 500 Mbps |
| Brazil | RNP | 10 Gbps | 4 Gbps |
| Chile | REUNA | 10 Gbps | 500 Mbps |
| Colômbia | RENATA | 10 Gbps | 500 Mbps |
| Costa Rica | CONARE | 2 Gbps | 500 Mbps |
| El Salvador | RAICES | 100 Mbps | 100 Mbps |
| Equador | CEDIA | 600 Mbps | 300 Mbps |
| Guatemala | RAGIE | 100 Mbps | 100 Mbps |
| México | CUDI | 300 Mbps | 200 Mbps |
| Paraguai | ARANDU | 100 Mbps | 100 Mbps |
| Uruguai | RAU2 | 300 Mbps | 155 Mbps |
| Venezuela | REACCIUN | 100 Mbps | 100 Mbps |

Obs: The L.A. bandwidth is the maximum capacity for the physical link, an NREN can reach up to this limit if the External bandwidth is unused. If an NREN is using part of its link capacity for External bandwidth, the capacity available for Latin American traffic will be reduced accordingly.

Annex 20: SPRACE (Brazil) Computing & Network Update

São Paulo Research and Analysis Center & UNESP Center for Scientific Computing

São Paulo State University (UNESP)

Submitted by Rogerio L. Iope <mailto:cavanaugh@phys.ufl.edu>

(rogerio.iope@cern.ch*)*

on behalf of the SPRACE Collaboration

<http://www.sprace.org.br/SPRACE/>

“Last updated: February 2017”

Introduction

Since it was first commissioned in 2004 as a regional analysis center to carry out Monte Carlo simulations for the Fermilab’s DZero Collaboration, the São Paulo Research Analysis Center (SPRACE) has made remarkable progresses in its processing power, storage capacity, and network connectivity. Starting in 2004 with a single main server connected to a shared Layer 2 access switch with a 100Mbps uplink to a core router, both the SPRACE cluster and its networking infrastructure experienced a noteworthy growth over the late 12 years. Among several other projects in partnership with leading global companies, such as Intel and Huawei, the SPRACE research and engineering teams operate a well-established Tier-2 class center for the CERN’s Compact Muon Solenoid Collaboration, which is fully integrated to the Worldwide LHC Computing Grid, and the GridUNESP system, one of the largest grid infrastructures in Latin America fully dedicated to scientific research. These remarkable achievements have only been possible due to the collaborative efforts provided by the Academic Network at São Paulo (ANSP), the Brazilian National Research and Education Network (RNP), and the support from the ICFA SCIC members, namely Caltech, Florida International University, and Fermilab.

ANSP leads the provisioning of advanced network connectivity to São Paulo State public and private universities, government and research institutes since the beginning of the Internet in Brazil³⁹. It provides connectivity to more than fifty institutions, which are responsible for more than 40% of all the Brazilian science production. ANSP is in a constant process of updating end-to-end connections across its network backbone in close collaboration with RNP and the AmLight Consortium lead by the Florida International University’s Center for Internet Augmented Research and Assessment (CIARA). ANSP was established in 1989 as a special program of FAPESP, the State of São Paulo Research Foundation - one of the main funding agencies for scientific and technological research in the country. ANSP administered the first neutral Internet exchange point in Brazil, a secure environment where São Paulo state universities and research centers, Internet access providers and local telecom companies started exchanging traffic, storing data and keeping secure information, among other functional features. ANSP headquarters had been located on the third floor of FAPESP (São Paulo Research Foundation) building until 2002, sharing the same space with the Foundation's data center. In 2002 it was transferred to the Verizon Terremark data center (recently acquired by Equinix), the so-called ‘NAP do Brasil’, located in Barueri, a small city in the São Paulo metropolitan area.

The São Paulo Research and Analysis Center

³⁹ For an interesting review of the history of the development of *networking in Brazil*, see M. A. Stanton, *Non-Commercial Networking in Brazil*. In: *INET’93, San Francisco, 08/93*. Available at: <http://www.ic.uff.br/~michael/pubs/inet93.ps>

The São Paulo Research and Analysis Center (SPRACE) was established in 2003 to support the effort of particle physicists from the São Paulo state to get involved in major high energy physics experiments. SPRACE research group took part of the DZero collaboration at Fermilab from 1999 to the end of the experiment, and became member of the CERN CMS collaboration in 2004. The main contribution of the SPRACE group to these collaborations has been related to the processing and analysis of the data collected by the experiments. The research team has been exploring two important branches of fundamental science: the search for new phenomena beyond the Standard Model and the physics of quark-gluon plasma produced in heavy-ion collisions.

The SPRACE team also includes computing engineers that have been operating a Tier-2 class site (BR-SP-SPRACE) of the Worldwide LHC Computing Grid (WLCG) together with the United States Tier-2s at Caltech, Florida, MIT, Nebraska, Purdue, UCSD, Vanderbilt, and Wisconsin. BR-SP-SPRACE is an official Tier-2 in Latin America and its operation has been a remarkable success, being among the most reliable Tier-2s of the whole collaboration. The expertise acquired by the group with the deployment and operation of the SPRACE cluster resulted in an important spin-off: the deployment of the first campus grid in Latin America – GridUNESP – that operates in close association with the US Open Science Grid (OSG) infrastructure.

During 2016, the SPRACE Tier-2 facility remained among the top most available and reliable computing infrastructure of the whole WLCG. According to the Tier-2 Availability and Reliability Report published each month by the collaboration, along 2016 the availability and reliability of SPRACE Compute and Storage Elements remained above the minimum threshold, established to be at 80% level. The infrastructure suffered a few sporadic falls because of instabilities in the network and planned shutdowns for middleware and software updates. These marks are steadily measured all over the year by the Site Readiness Monitoring framework, developed by the CMS Collaboration to assess their sites readiness. It consists of three main ingredients:

- Site Availability Monitoring (SAM) tests to check site specific services
- Monitoring jobs to test the data processing work
- Monitoring data transfers to evaluate the capability of replicating data between sites

Site readiness is assessed by means of a set of defined metrics using the results of these tests. This factor determines whether a site can be safely used by users from anywhere in the world to perform CMS data analyses and Monte Carlo simulations.

The volume of data routed to our farm as a function of time, through the CMS Tier-1s and Tier-2s data transfer links, is shown in Figure 1. One can observe the evolution of the transfer rate during a period of 52 weeks, ranging from the beginning of December 2015 to the end of November 2016, and the cumulative volume of weekly transfers with an average of 22.59 TB, reaching a peak of more than 105 TB of files transferred in only one week.

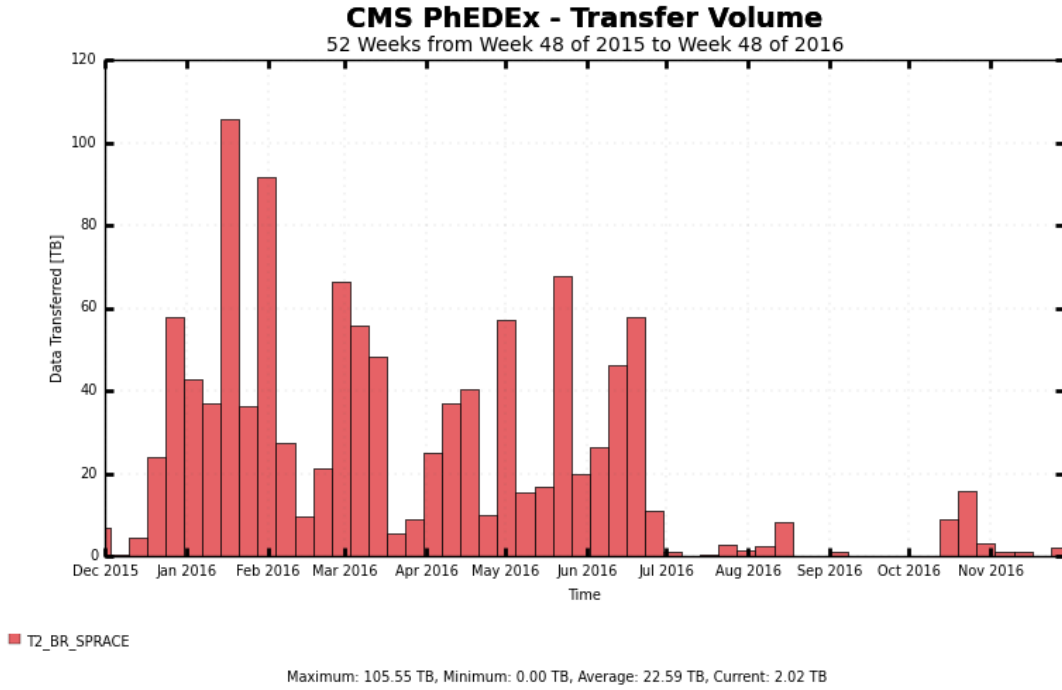


Figure 1: Volume of data transferred to SPRACE storage in 2016

The cumulative graph shown in Figure 2 shows that a total volume of more than 1 PB of data has been transferred to the SPRACE storage during this period.

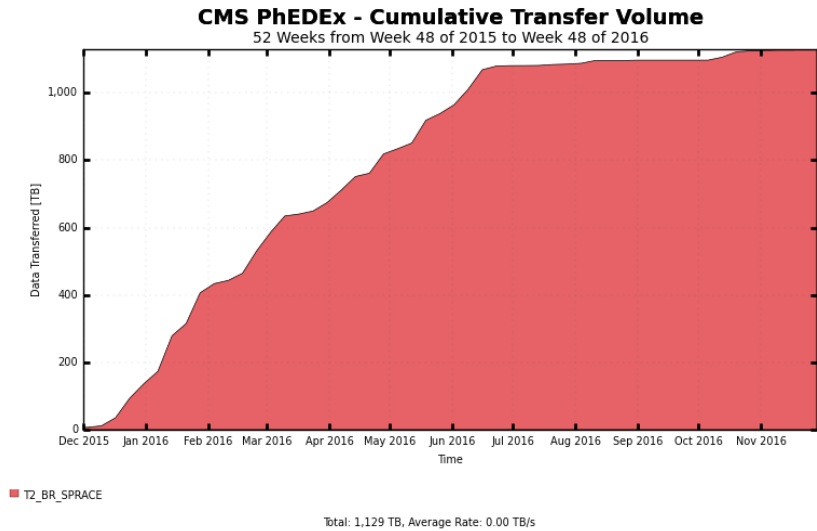


Figure 2: Cumulative data transferred to SPRACE storage in 2016

The same cumulative graph shown in previous figure is shown in more detail in Figure 3, which includes the contribution in data transfer due to each CMS site.

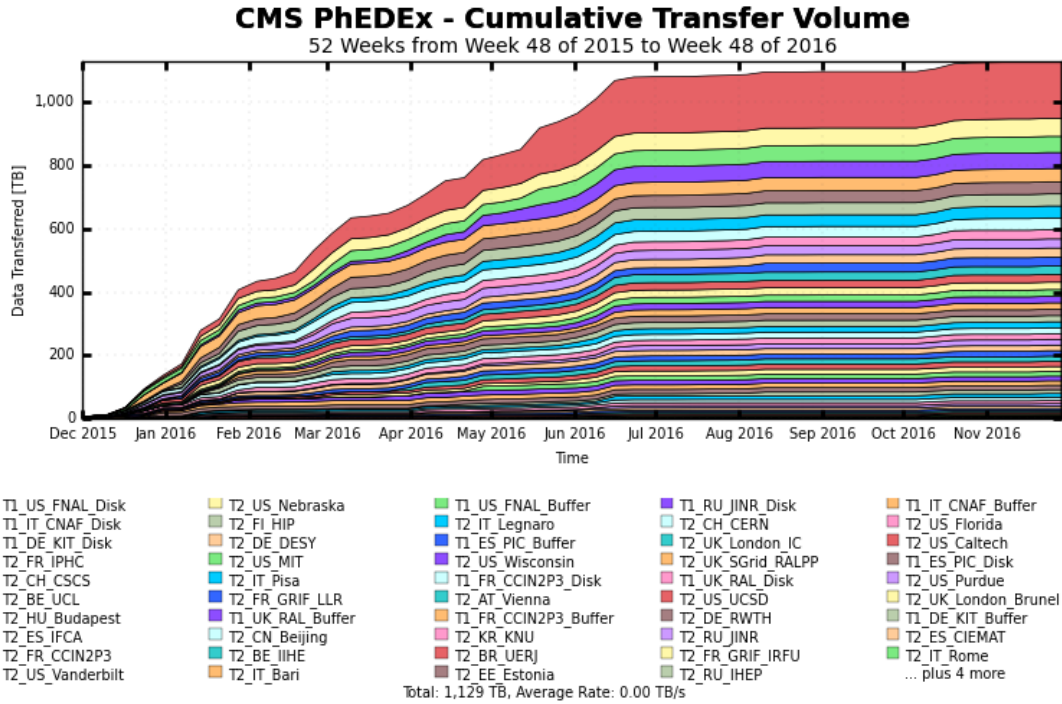


Figure 3: Cumulative data transferred to SPRACE storage in 2016 (by each CMS site)

Besides serving the CMS experiment, by offering processing power and storage space, the SPRACE cluster is also used - with a lower priority level - by several other scientific experiments, which are organized into Virtual Organizations (VO) inside the Open Science Grid collaboration. Figure 4 shows the usage of the SPRACE cluster by some of these virtual organizations, where one can see a twelve-month period utilization, from November 29, 2015 to November 28, 2016 (each bar represents the number of hours spent on jobs per month). We may notice that there was a predominance of the CMS VO, as expected.

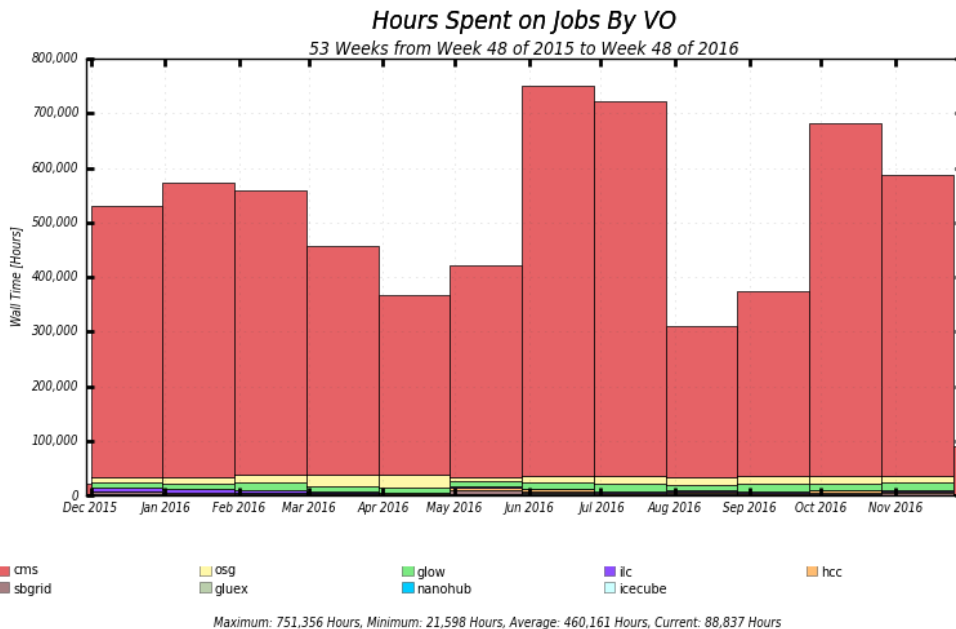


Figure 4: SPRACE processing resources utilization by VO (bar chart)

A different view of the data presented on the previous graph is shown in Figure 5, in which we can see the overall utilization of SPRACE processing resources in the same period (Nov 2015 to Nov 2016). During this period the SPRACE computing infrastructure delivered more than 6.4 million hours of data processing to distinct scientific collaborations, from which around 6 millions has been dedicated to the CMS experiment. In 2015 the number of CPU-hours used by the collaboration was 4.686.366, so we see an increase of 28% this year. We think this should be credited to the 16 new worker nodes we bought in the second semester of 2015.

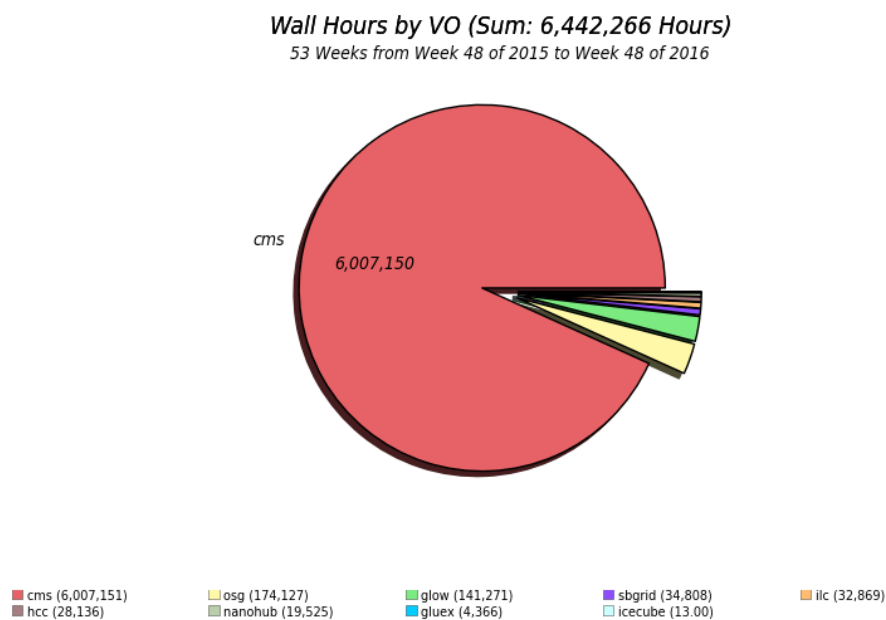


Figure 5: SPRACE processing resources utilization by VO (pie chart)

More recently, SPRACE became also involved in the field of scientific instrumentation, by leveraging a team of electrical engineers which are contributing to the Level1 Tracking Trigger (L1TT) system development proposed by Fermilab and other partners for the future CMS Phase II upgrade, when the Large Hadron Collider (LHC) will have an increase in its collisions luminosity. The SPRACE instrumentation team has been working in the demonstration and test setup for the Fermilab AM+FPGA proposal. The system is all based on a general purpose FPGA carrier board developed by Fermilab for ATCA shelves.

UNESP Center for Scientific Computing - The GridUNESP project

In 2009, the São Paulo State University began operating GridUNESP, one of the largest multi-campus grid infrastructures in Latin America fully dedicated to scientific research. With 24 campuses distributed throughout São Paulo, the São Paulo State University (UNESP) is the second largest university in Brazil. To accommodate the computing resources acquired by means of the GridUNESP project, in 2008 the UNESP Center for Scientific Computing (CSC) was established, which now operates two main clusters: SPRACE - dedicated to CERN; and GridUNESP - dedicated to university researchers and students. Both share a common underlying infrastructure, provided by the CSC datacenter, which was built in 2009.

GridUNESP has been designed to take advantage of the distributed scientific research at UNESP. The grid consists of a central cluster located in the CSC datacenter in São Paulo, with seven secondary clusters at different UNESP campuses spread over the state. The secondary clusters are used as auxiliary resources that complement the main HPC cluster.

From the very beginning, GridUNESP established a formal partnership with the OSG. The GridUNESP VO (virtual organization) was established in December 2009, and it was the first OSG VO located outside of the US. As a multipurpose VO, it includes projects from astronomy, biology and biophysics, biomedical engineering, chemistry, computer science, the geosciences, materials science, meteorology, and physics. This partnership enables GridUNESP to use the OSG middleware stack to integrate its computational resources and share them with other institutions. GridUNESP can run jobs from several different US VOs. In 2010, the UNESP CSC engineering team organized the ‘São Paulo OSG School’ with the full support of the OSG education, outreach, and training team, and brought five OSG experts to Brazil. Attendees had the opportunity to learn how to leverage their research using grid resources and tools.

The number of researchers using GridUNESP has been growing, as can be seen in Figure 6. In its seven years of operation, GridUNESP has processed more than 7 million jobs and logged over 50 million CPU hours. External researchers can also use the HPC cluster, provided that they are affiliated with or connected to UNESP researchers. A major upgrade of the GridUNESP physical infrastructure is ongoing and should be finished in February 2017.

Another important milestone was the establishment of a certificate authority. Together with ANSP engineering team, UNESP CSC engineers deployed an official Certificate Authority for the state of São Paulo, known as the ANSPGridCA. Its accreditation was approved by TAGPMA (The Americas Grid Policy Management Authority) in April 2012 and by IGTF (Interoperable Global Trust Federation) in August 2012, and included in the TERENA Academic CA Repository (TACAR) in October 2012. ANSPGridCA began issuing GridUNESP certificates in march 2013. After nine months of testing, certificate production was open to the entire academic community of São Paulo. Local researchers no longer need to rely on certificate authority from abroad.

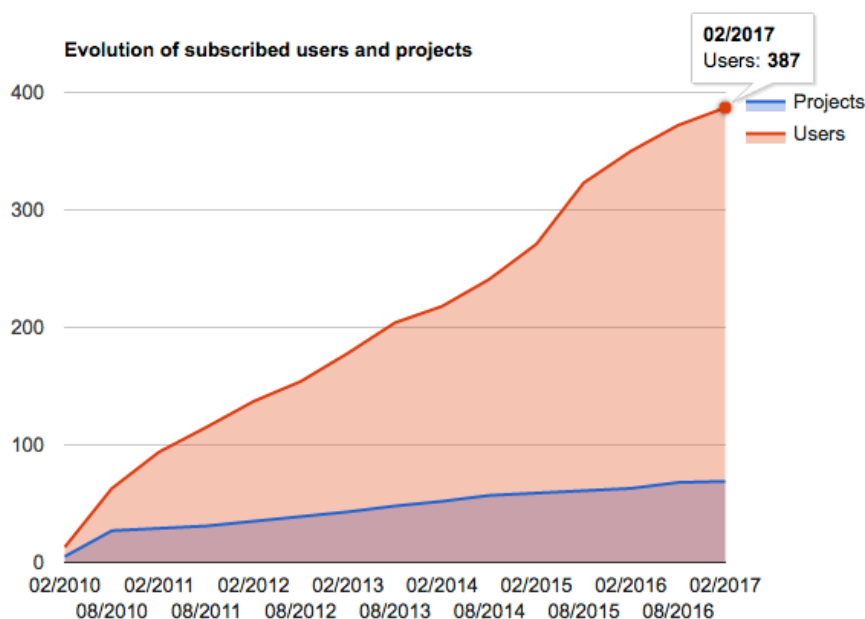


Figure 6: Evolution of the number of GridUNESP subscribed users and projects

UNESP CSC has been establishing partnerships with leading global companies. In 2014 the team was awarded with a grant from Intel to establish a ‘Parallel Computing Center’, aiming to include the SPRACE computing team in the R&D efforts necessary to adapt high-energy physics software tools to the modern computing architectures that support multi-threading, vectorization, and other parallel processing techniques, to make data processing more cost-effective. The work has been focused on testing vector-coprocessor prototypes in hybrid parallel computing systems and analyzing the performance of the next generation of Intel Xeon Phi coprocessor (Knights Landing). From Dec. 2014 to Nov. 2016, the activities were closely related to the development of Geant-V, the next generation of the Geant simulation engine, which will include massive parallelism natively. UNESP CSC has also been awarded with another grant from Intel and became an Intel Modern Code Partner, being responsible for training courses, workshops and providing technical consultancy in parallel programming for Xeon and Xeon Phi architectures both in Brazil and in other countries (so far including Colombia, France, Germany and Portugal).

By the end of 2015 UNESP CSC team established a partnership with another company, the chinese telecom equipment maker Huawei. The project is related to the development of new open source services, tools and methods to leverage the integration of SDN with cloud computing technologies to promote site orchestration among virtualized servers, storage systems and subnets to successfully co-schedule CPU, storage and network resources on a wide area environment. In order to validate research ideas in a real environment instead of a simulated one, the project includes the build up and deployment of a simple but scalable SDN WAN testbed for development and experimentation, with computing and SDN-based network resources spread over three geographically distant sites: Unesp in São Paulo, Brazil, Caltech in California, USA, and CERN in Geneva, Switzerland. The proposed testbed will also allow us to explore the performance of data transfers over continental and transoceanic distances on the scale of hundreds of gigabits/sec using state-of-the-art servers and high-performance transfer tools such as Caltech FDT. Researchers working on this testbed may instantiate and program virtual topologies consisting of virtual machines (VMs), programmable switch datapaths, and virtual network links based on Ethernet standards. The services, tools and methods to be developed are intended to interface seamlessly to the adaptive network operating system and tools being developed by ESnet, Fermilab, Caltech and others, to ensure that authorized applications achieve full throughput. The project also includes the development of a new OpenFlow controller entirely built from the ground and fully open source, to be used in various scenarios and in various network environments. UNESP CSC developers are working hard on the establishment of a broad community of developers that could help on the development of new network applications on top of this new controller.

According to the 2016 BrandZ™ ‘Top 100 Most Valuable Global Brands’ ranking and report⁴⁰ compiled by WPP, a world leader in marketing communications, Huawei and Intel are ranked at positions 50 and 51, respectively.

Recent progresses - New 100 Gbps links

In the late years, the São Paulo State University network engineering team, responsible for the University’s commodity network (UnespNet), has made extensive investments in the network infrastructure that connects the university core resources to the academic providers ANSP and

⁴⁰ http://wppbaz.com/admin/uploads/files/BZ_Global_2016_Report.pdf

RNP. As a result of the negotiation for loaning new dark fibers and the upgrade of the corresponding network equipments, a new link has been established between the UNESP central datacenter at the Office of the President building, in São Paulo downtown, to the academic access point at NAP of Brazil, located in Barueri (~ 50 Km away). Luckily, UNESP CSC datacenter is right in the middle of the path, and a joint collaboration between UNESP and ANSP engineering teams resulted in the acquisition of dark fibers and WDM-based optical equipments in such a way that both the University commodity network and the SPRACE network could share the same fiber cable on the path to ANSP headquarters. Due to the WDM technology, SPRACE network can use distinct wavelengths, so the traffic its cluster generates do not disturb the University commodity network.

In 2014, the SPRACE team invested efforts on the deployment of a new 100G link from the CSC datacenter to ANSP headquarters using this new dark fiber. However, prices for both optical equipments and network switches operating at 100G were extremely high, well beyond the available budget for such new deployment. A workaround was to deploy a 100G link on the line side, but divided on the client side into two 40G channels plus two 10G channels. With this arrangement, SPRACE network link up to ANSP could be easily upgraded from 10G to 40G, and extra channels became available for experiments. The partnership with Huawei in the beginning of 2016 provided us with extra budget, and the Brazilian optical company Padtec offered a very special price for the new optical 100G transponders, so by the end of 2016 it was also possible to deploy a second 100G channel over the same fiber, using a different wavelength. Huawei switches with 40G and 100G ports have been acquired in the context of the SDN project, which enabled the utilization of the new 100G channels. Figure 7 shows a sketch of the network channels that connects UNESP CSC to ANSP core routers. To our knowledge, no other university or research center in Latin America has an infrastructure that can compare to the one that is now available to the SPRACE Tier-2 center and the remaining HPC resources installed at UNESP CSC datacenter.

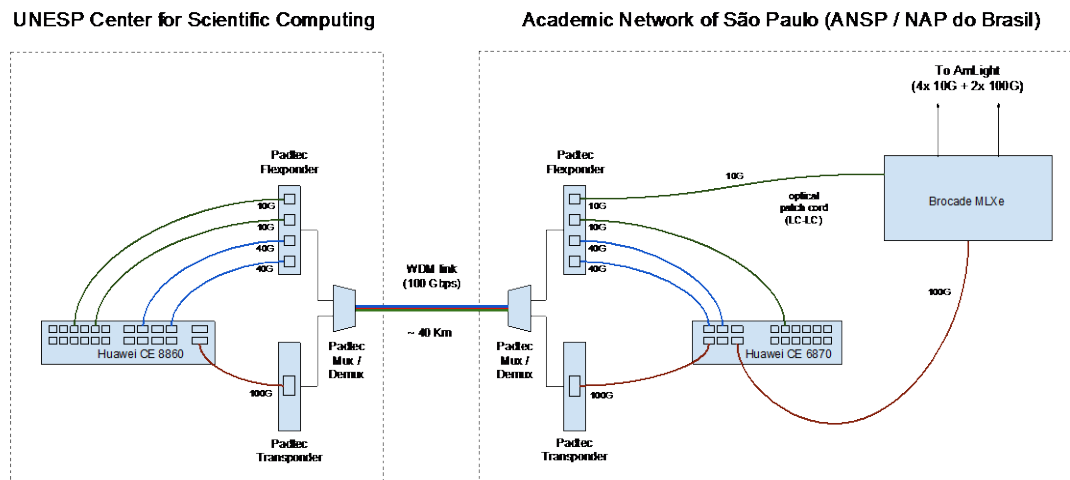


Figure 7: A sketch of the link between UNESP CSC to ANSP headquarters

More or less at the same time, the AmLight Consortium was working hard to activate the first 100G link of the AmLight-ExP project. It became ready in April 2016, connecting Miami, FL and Sao Paulo, Brazil (106ms round-trip delay) through the Atlantic side of the South America continent. A second 100G link was deployed in August 2016, connecting the same endpoints but going through the Pacific side of the continent. The AmLight Consortium is a group of not-for-profit universities, state, national and regional research and education networks including the AmLight

ExP project at Florida International University, RNP, ANSP, RedClara, REUNA, FLR, AURA, Latin American Nautilus, and Internet2.

Participation in the Supercomputing Conference 2016

During the SC'16 in November, the FIU, RNP, ANSP and UNESP teams, together with Caltech and partners, exercised the 'Atlantic side' 100G connection between São Paulo and Miami, activated in April 2016 to expand the international output of the Brazilian academic network. These new interconnections are part of the AmLight Express and Protect project ([NSF Award#ACI-1451018](#)), founded by the National Science Foundation (NSF), by the São Paulo Research Foundation (FAPESP) - by means of its academic network arm (ANSP) -, and by the Brazilian National Research and Education Network (RNP).

A pair of high-end data transfers nodes (DTNs), manufactured by Huawei with Mellanox 100G NICs and Intel NVMe SSD cards installed at UNESP datacenter in São Paulo, as well as OpenFlow-enabled Huawei switches and WDM appliances at UNESP and ANSP were used during the exercise. During the demonstration days, UNESP team also presented its new OpenFlow controller, Kytos, a modular, event-based open-source OpenFlow controller being developed in Python with a simple asynchronous architecture to facilitate the exchange of messages between its core and users applications. This development is part of a 3-year R&D cooperation agreement between UNESP and Huawei, which also includes experiments on the interoperability of distinct SDN controllers and the integration of SDN and cloud technologies.

As a result of a joint effort, during SC'16 in Salt Lake City FIU, ANSP, and UNESP teams showcased for the first time an end-to-end transmission from Latin America to the U.S. using a 100G link. This network infrastructure, provided by the AmLight Express and Protect project (AmLight-ExP), implements a hybrid network strategy that combines optical spectrum (Express) and leased capacity (Protect) that builds a reliable, leading-edge diverse network infrastructure for the research and education community.

Transfers between Brazil and U.S. achieved more than 100 Gbps of aggregated traffic when using simultaneously the two 100G paths on the Atlantic and Pacific sides. A lot of effort has been concentrated on tuning transfers between Sao Paulo and Miami (NAP of the Americas), and interesting results have been achieved. During the first experiments, the team was able to sustain ~85G on the Atlantic link for several hours, as can be seen on Figures 8 and 9.

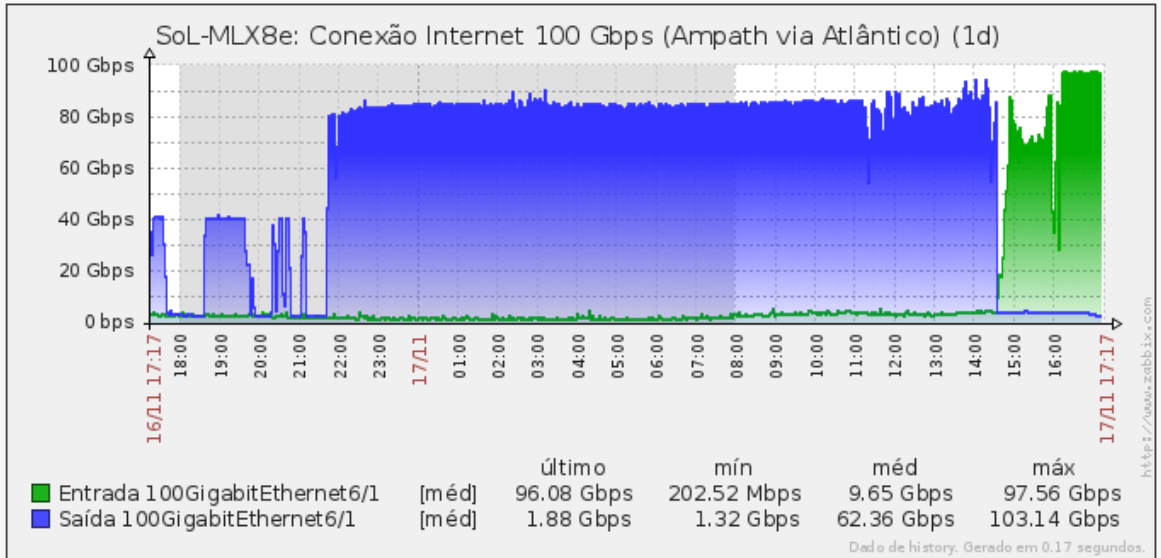


Figure 8: Stable network transfers sustained for more than 15 hours (seen from ANSP)

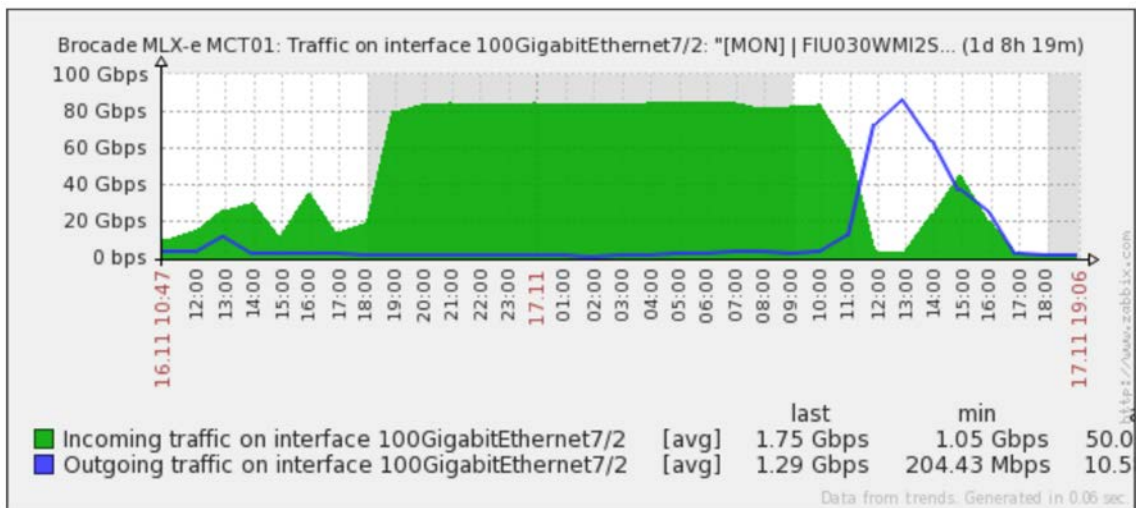


Figure 9: Stable network transfers sustained for more than 15 hours (seen from AmLight)

After initial tests, new careful experiments have been performed, first with transfers going from São Paulo to Miami (Figure 10) and then from Miami to São Paulo (Figure 11). In this later experiment we have maintained an extremely stable transfer, with minimum transfer rate of 95.86Gpbs and maximum of 97.56Gpbs for more than 1 hour, using Caltech tool known as Fast Data Transfer (using up to 600 simultaneous threads).

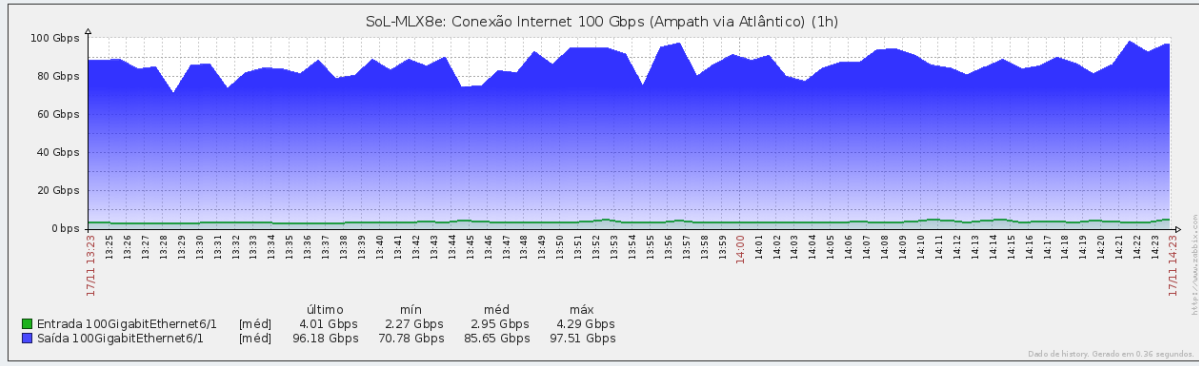


Figure 10: Data transfer from São Paulo to Miami

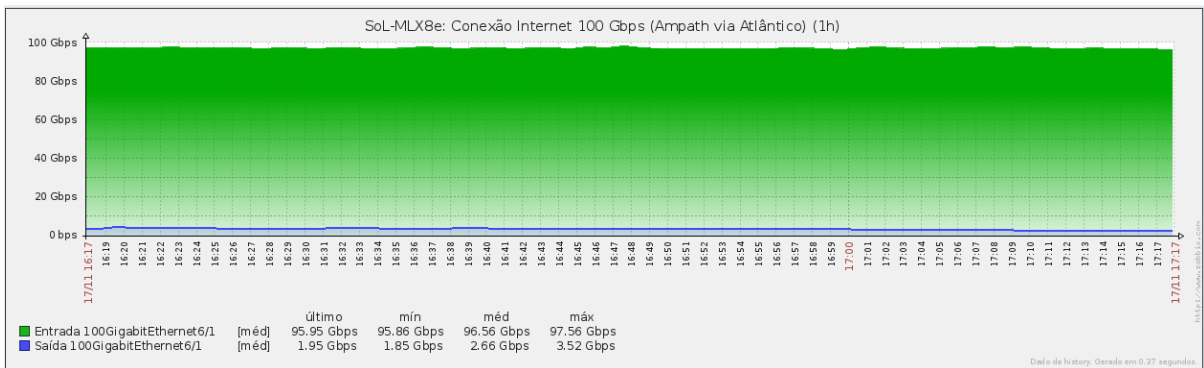


Figure 11: Data transfer from Miami to São Paulo

We used two high-end servers installed in Sao Paulo (one at SPRACE/Unesp and the other at ANSP) provided by Huawei and a third server installed at AmPATH/NAP of the Americas, provided by the AmLight team. A couple of vlans have been deployed by means of AmLight OESS on both Atlantic and Pacific 100G links (2370, 2372 and 2374 through the new 100G Atlantic link, and 2371, 2373 and 2375 through the new 100G Pacific link). Figures 12 and 13 illustrate the deployment.

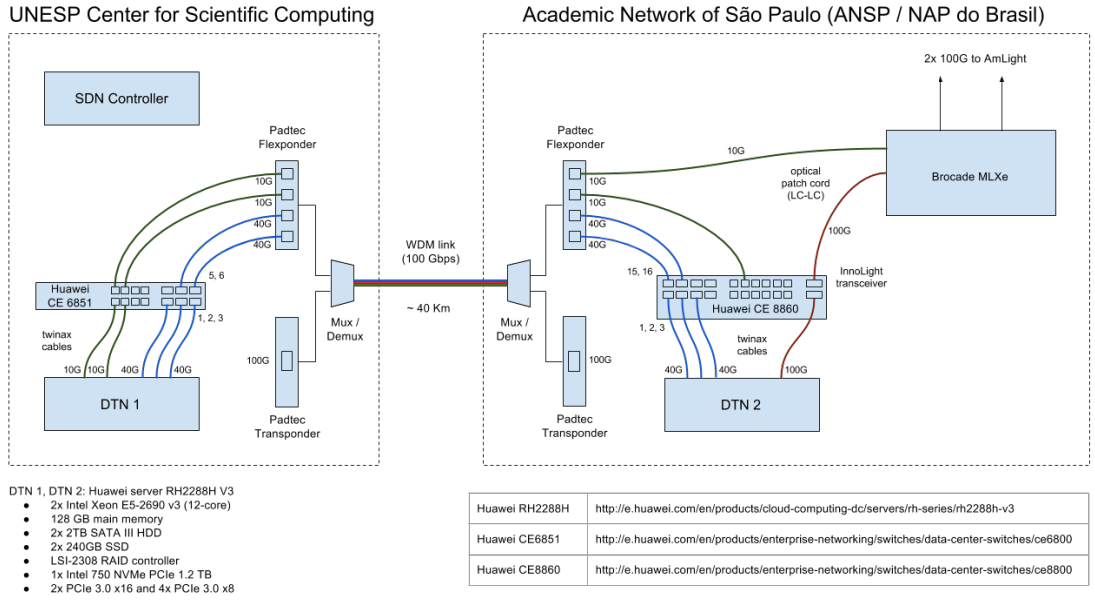


Figure 12: Server/switch configuration in São Paulo (Unesp, ANSP)

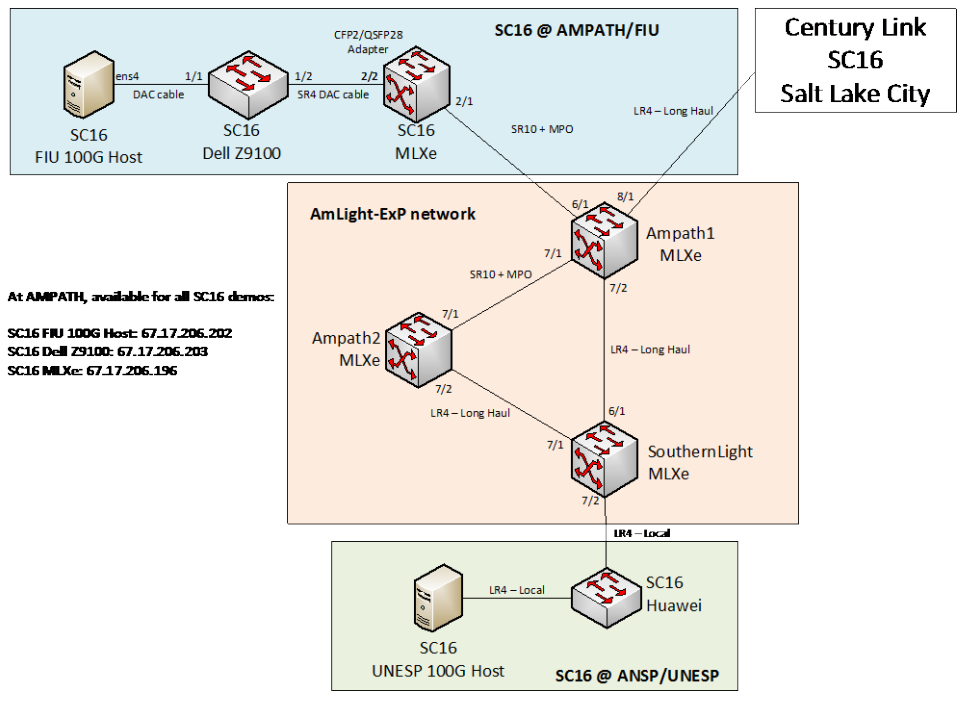


Figure 13: Network configuration (UNESP, ANSP, AMPATH, FIU)

During the demonstration days, the UNESP CSC team also showcased the Kytos controller for the very first time [https://github.com/kytos]. The Kytos project aims to build a 100% open source framework to orchestrate Software Defined Networks (SDN). The controller was tested during the

conference on a testbed infrastructure based on Huawei equipments built at SPRACE facilities in Sao Paulo.

A hierarchical topology has been built on two physical hosts using KVM-based virtual machines running CentOS, connected to several OVS-based virtual switches, as can be seen on Figure 14. The switches were connected to each other using virtual and physical connections. All switches were under the control of Kytos controller using openflow 1.0, which in turn was located at the SC16 showfloor. The Kytos team used this infrastructure to exercise data transfers among all the virtual servers.

During the demo, the controller was able to manage flows passing through the 100 GbE network interfaces. The controller is available at Github (github.com/kytos) under an MIT license.

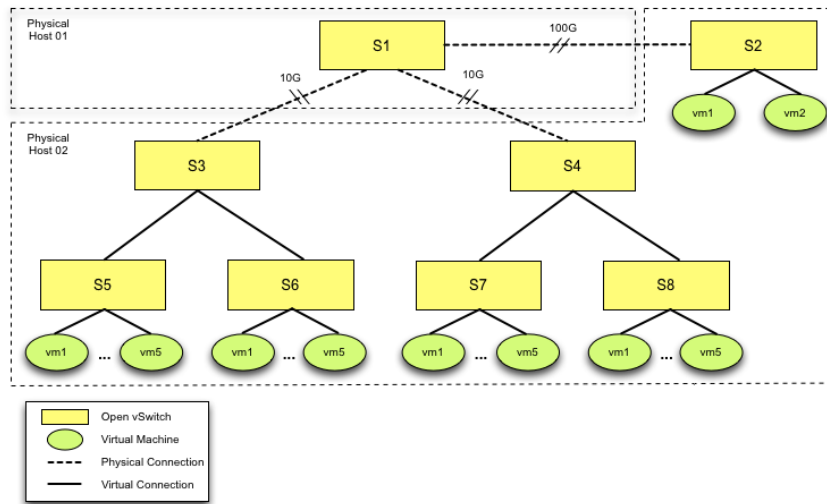


Figure 14: Hierarchical topology of virtual servers and switches created on 2 real servers

Annex 21: UERJ (Brazil) Tier2 Center Status and Plan

Submitted by Alberto Santoro (alberto.santoro@cern.ch) and

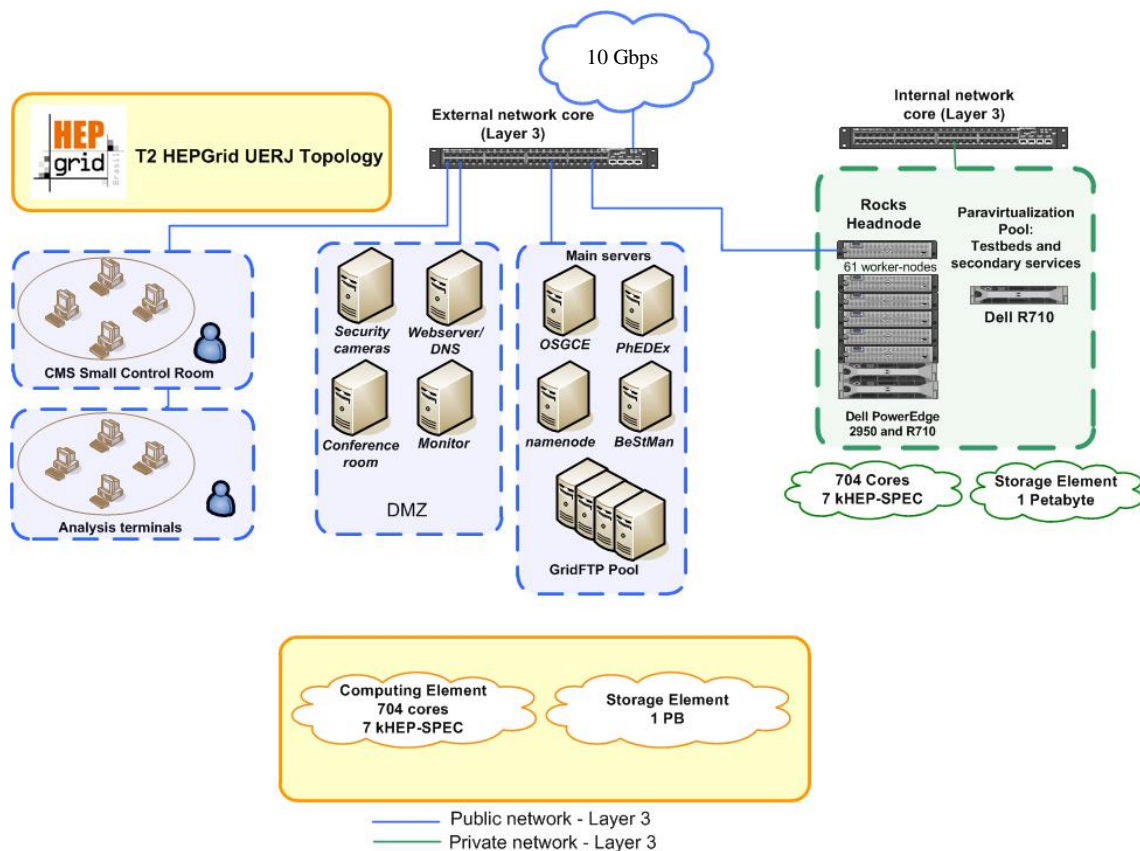
Eduardo Revoredo (erevored@cern.ch)

January 2017

Introduction:

Since February 2016 our cluster network connectivity has been provided by Redecomep-Rio¹, a joint initiative between FAPERJ² and RNP³ financed with funds from FINEP⁴. Redecomep-Rio also provided connectivity to our campus university and its network has already implemented a high-speed network infrastructure (10 Gbps now and 100 Gbps in the future) for education, science, technology and government institutions in Rio de Janeiro, interconnecting more than 80 teaching and research units in the metropolitan region of Rio de Janeiro through 300 kilometers of optical fibers (DWDM technology support). Right now, in this configuration our maximum transfer speed is 10 Gbps.

Currently Tier-2 UERJ contains 704 cores and 1 PB of storage space, distributed among 61 worker nodes. Our topology is described below in more detail.



Our Tier-2 will carry on with analysis tasks that we have performed on the previous years. For CMS physics, we will continue support officially the Forward physics, B-Physics and Higgs physics. The Tier2 is working full time for CMS and eventually receiving jobs from other collaborations, even if they're out of the HEP community (a small number), like for example Artificial Intelligence jobs from computer science department at UERJ.

Since the 2015 started a big economical and political crises in Brazil but mainly in Rio de Janeiro State. The consequences affected directly the University of the State of Rio de Janeiro and FAPERJ the main financial support agency for Science. As the last example the no-break of our Tier-2 broke and only in the last month we got a new unit donated by the Olympic Committee. Then we expect to resume the operation of our Tier-2 in March 2017.

1. Redecomep-Rio - More information at: <http://redecomep.rnp.br/?consorcio=2>

2. FAPERJ (Carlos Chagas Filho Foundation for Supporting Research in the State of Rio de Janeiro) is a public foundation linked to the Rio de Janeiro State Department for Science and Technology, the aim of which is to stimulate research and foster the kind of scientific and technological activities necessary for the socio-cultural development of the state of Rio de Janeiro, Brazil.

3. RNP - Brazilian National Research and Educational Network. The main goal of this initiative is to implement high-speed networks in the Brazilian metropolitan areas served by RNP Points of Presence.

4. FINEP - Funding Authority for Studies and Projects is an organization of the Brazilian federal government, devoted to funding of science and technology in the country.

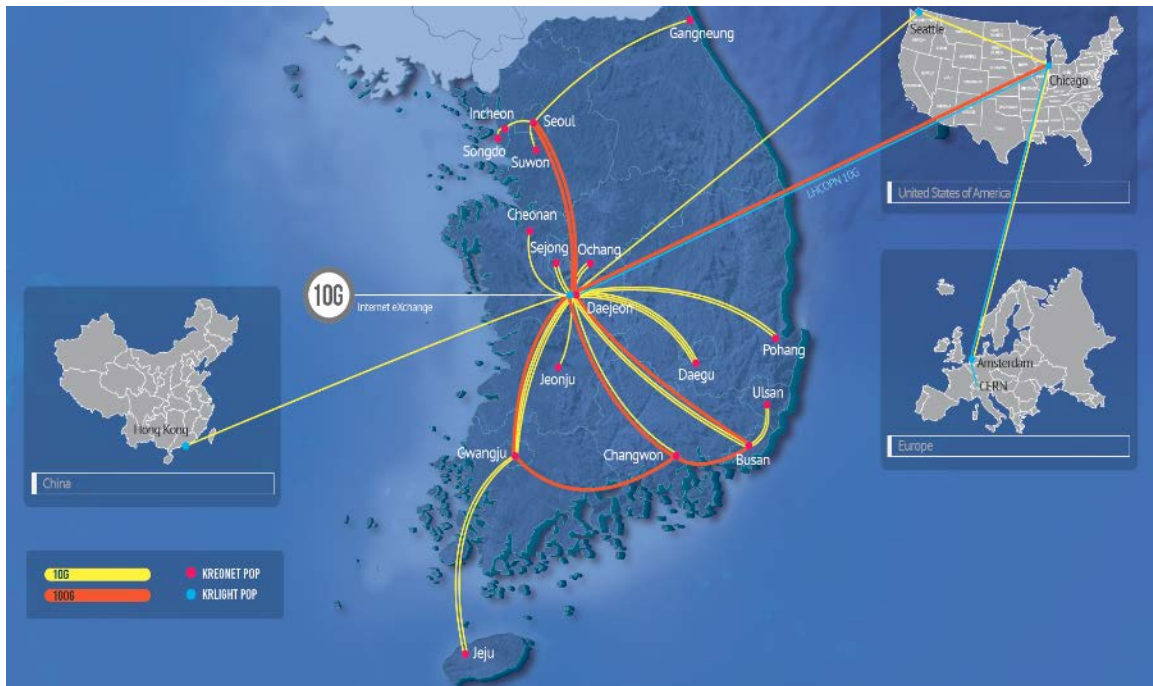
More information at: <http://www.finep.gov.br/>

Annex 22: KREONET2 and KRLight (KOREA) Status and Plan

Submitted by Buseung Cho (bscho@kisti.re.kr)
February 2017

Introduction

KREONET (Korea Research Environment Open NETwork) is Korea's national science and research network designed, built and operated by KISTI (Korea Institute of Science and Technology Information) through funded by Government of Korea, MISP (Ministry of Science, ICT and Future Planning) since 1988. Most scientific and research institutions in Korea has connected to KREONET. Now about 200 institutions are KREONET member and more than 300 institutions use KREONET. KREONet2 means all international connectivity of KREONET. KREONET/KREONet2 provide high quality network services using state-of-the-art networking technology for the science and research community including university, national laboratory, hospital, library, and so on in Korea. In order to best meet user's requirement, KREONET/KREONet2 has provided advanced high quality network services to the users. In 2016, KREONET and GLORIAD would be upgraded into 100Gbps high performance network.



KREONET Backbone (2016)

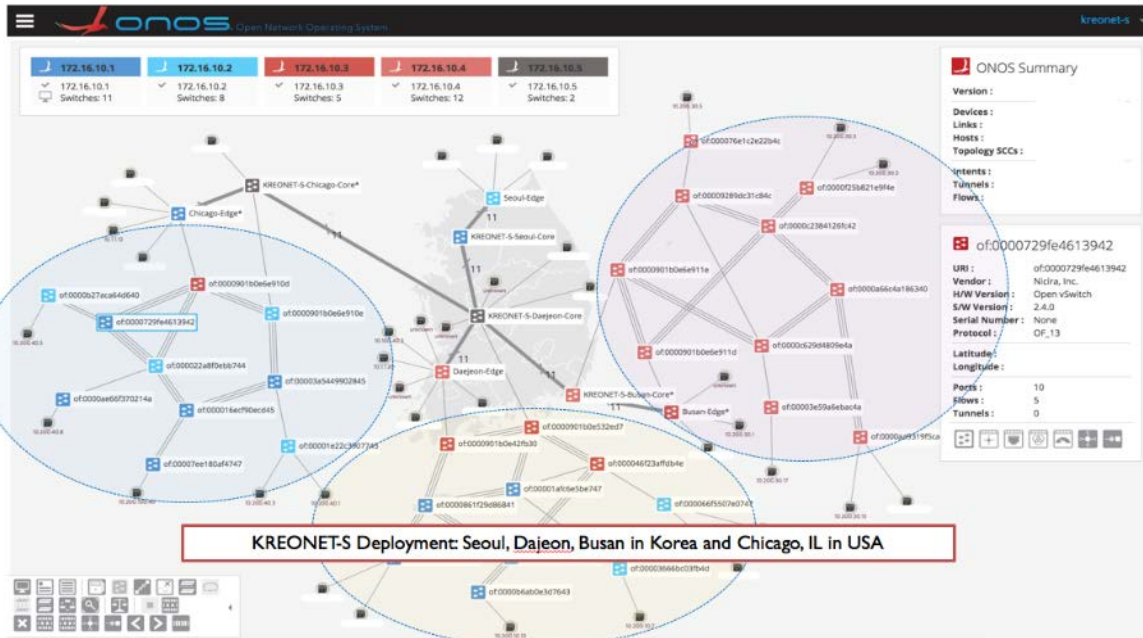
KREONET/KREONet2

KREONET has built 17 regional GigaPoPs, covered Korean peninsula, and enabled 40G/100G transmission technology using ROADM and Dark Fiber. In Daeduck Science Park of Daejeon City, there is SuperSIREn as a regional network that connected several

national research institutes like KAIST (Korea Advanced Institute of Science & Technology), KISTI (Korea Institute of Science and Technology Information), KRIBB (Korea Research Institute of Bioscience and Biotechnology), KBSI (Korea Basic Science Institute), KIGAM (Korean Institute of Geoscience and Mineral Resources), KARI (Korea Aerospace Research Institute), CNU (Chungnam National University). KREONET users connect to 17 GigaPoPs of KREONET, along with 1G/40G/100G Ethernet. Also KREONET has been a service network of KISTI's Supercomputer since 1988, which is national Supercomputer in Korea.

The architecture of KREONET is both packet and circuit switching network and it enables hybrid network service. It provides combined network service for KREONET user: end-to-end lightpath provisioning service and traditional IP connection service. Particularly end-to-end lightpath is dedicated communication channels between KREONET users with 1/10Gbps and 40/100Gbps connection for advance scientific users. Lightpath is suitable network service for data-intensive applications such as high energy physics, radio-astronomy, meteorology/climatology, the International Thermonuclear Experimental Reactor (ITER), geological sciences, and global e-Science applications, requires high performance and secure network service. KREONET also offers OPN (Optical Private Network) services. To be able to establish lightpaths dynamically, dynamic lightpath allocation systems such as DynamicKL (developed by KISTI) and OSCARS@KREONET have been deployed.

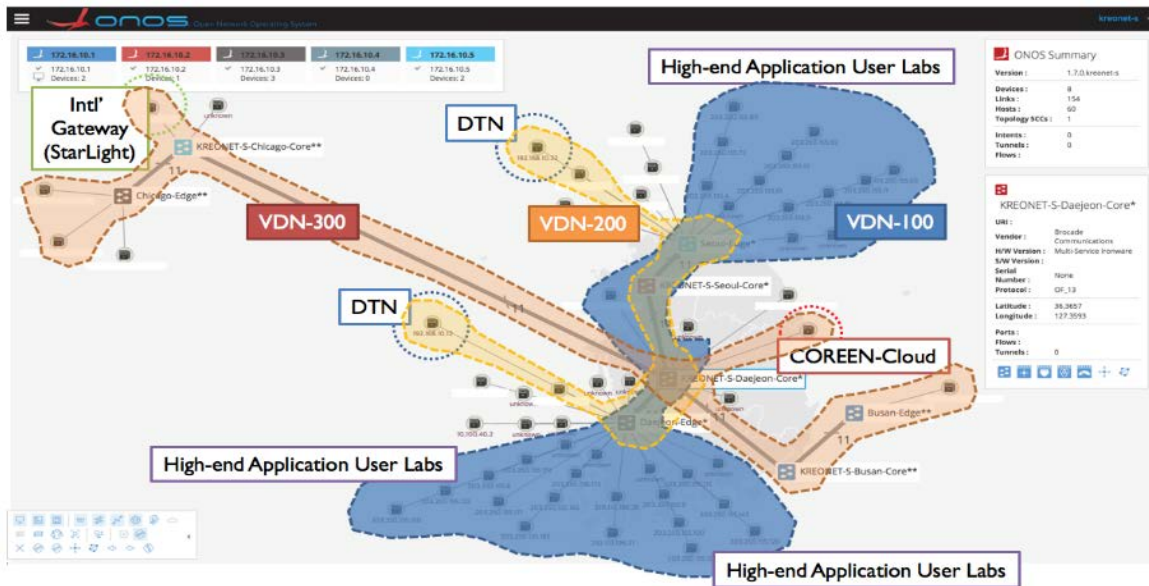
Since 2010, KREONET has provided KREONET Dynamic Ethernet Service based on the Next Generation Ethernet (NGE) technology such as PBB (Provider Backbone Bridge)/PBB-TE (Provider Backbone Bridges Traffic Engineering)/PLSB (Provider Link State Bridging)/G.8032 (Ethernet Ring Protection Switching). As well as the lightpath based on OTN (Optical Transport Network) switching/provisioning and SONET/SDH channeling, KREONET Dynamic Ethernet Service can provide dedicated high performance network and aggregation service based on E-LAN, E-LINE or E-TREE of the Carrier Ethernet service.



KREONET-S (International) SD-WAN Deployment

KREONET expanded its Software-Defined Wide Area Network (KREONET-S) infrastructure from two locations (Daejeon and Seoul in Korea) in 2015 to four locations with two new cities SDN-enabled (Busan in Korea and Chicago, IL in USA) in 2016. KREONET-S is designed to implement a nationwide and international virtual programmable network infrastructure based on ONOS (Open Network Operating System) distributed control platform over a large-scale wide-area SDN, providing end-to-end SD-WAN production services for advanced research and applications demanding specific time-to-research and time-to-collaborations in particular. As for the international KREONET-S deployment in 2016, KREONET-S made a long-distance SD-WAN connection to StarLight facility in the US over 100Gbps optical fiber, enabling high-end research partners and organizations to optimally access resources in North America, South America, Asia, South Asia, Australia, New Zealand, Europe, and other sites around the world.

Furthermore, among the primary building blocks of KREONET-S, Virtual Dedicate Network (VDN) system was enhanced in 2016 (as an SD-WAN application component) from its prototype version made in 2015, in order to provide following new features: 1) dynamic on-demand (international) virtual network generation, 2) logically isolated group network creation with dedicated network bandwidths (up to 100Gbps), 3) ONOS-based event (e.g., link up/down) detection and recovery, 4) user-friendly and GUI-based intuitive virtual network manipulations (create/update/delete), 5) user-oriented virtual network topology visibility and REST APIs. Based on the above SD-WAN technologies, KREONET-S is able to provision virtually isolated group networks for (international) high-end application scientists, researchers, and user sites up to 100Gbps, while providing dedicated virtual networks for massive science data transfer nodes (DTNs), as well as high-performance networking for 1-hop access of international research network gateway (StarLight facility) and cloud system gateway (e.g., COREEN platform), as shown in the following figure.



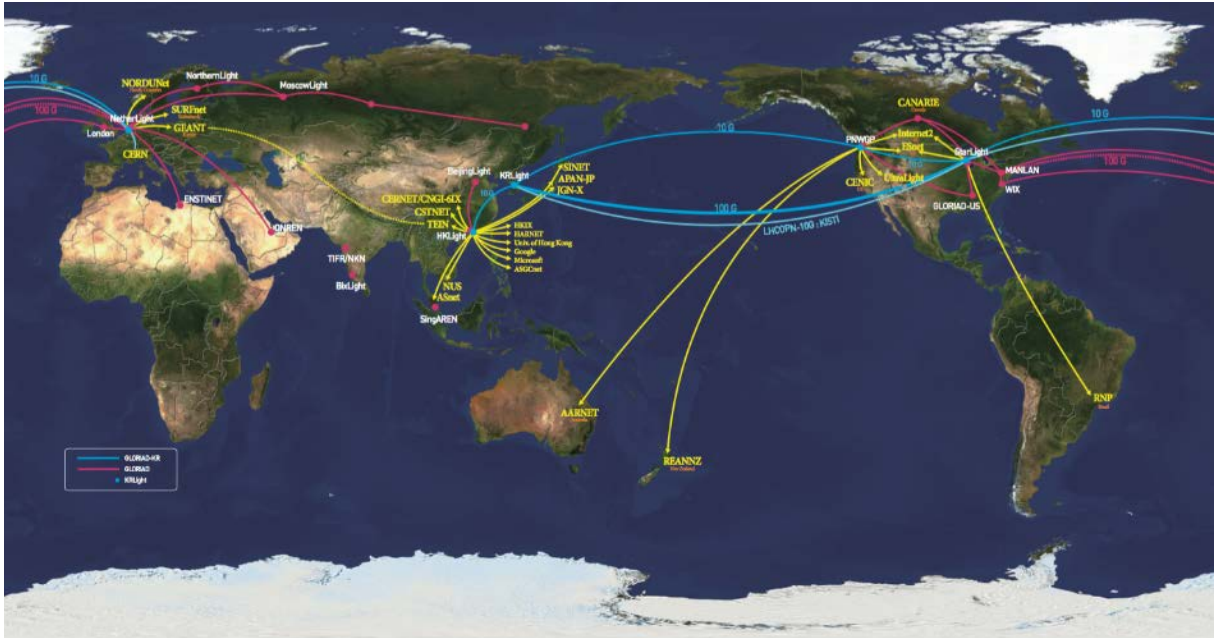
KREONET-S Virtual Dedicate Network (VDN) System and its Use Cases

GLORIAD-KR

KISTI joined the GLORIAD program as the fourth core member, funded by Korean government in 2005, under the name of GLORIAD-KR (GLORIAD-Korea). The major role of GLORIAD-KR is to develop 10Gbps hybrid networks and advanced technologies as a part of global ring network. In order to support collaborative advanced applications, GLORIAD-KR had been operating two 10Gbps packet and optical hybrid international networks since 2005. In 2016, totally new GLORIAD-KR infrastructure would be built including 100Gbps trans-pacific international link.

All GLORIAD-KR international links are as below.

1. Daejeon, KR ~ Chicago, US : 1*100Gbps unprotection link (Primary) + 1*10Gbps unprotection link (Backup)
2. Daejeon, KR ~ Seattle, US : 1*10Gbps unprotection link (Primary) + 1*2.5G unprotectionl ink (Backup)
3. Daejeon, KR ~ Hong Kong, CN : 1*10Gbps unprotection link (Primary) + 1*2.5G unprotectionl ink (Backup)
4. Chicago, US ~ Amsterdam, NL : 1*10Gbps unprotection link (Primary) + 1*10Gbps unprotection link (Backup)
5. Seattle, US ~ Chicago, US : 1*10Gbps unprotection link (Primary)



GLORIAD-KR and GLORIAD (2016)

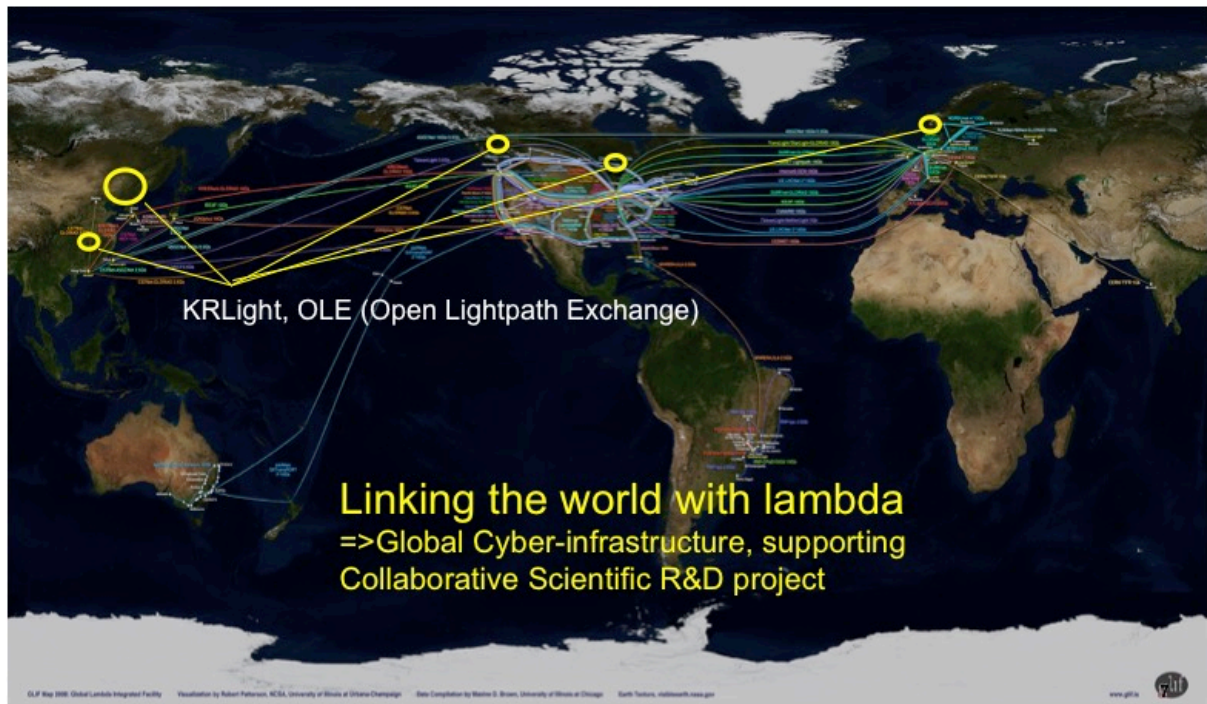
KRLight

KRLight is one of the GOLEs (GLIF Open Lightpath Exchange), lambda’s exchange point of GLIF, in Korea. KRLight is comprised of GLORIAD-KR international links and optical network equipments such as Ciena OME6500, Juniper MX960 etc and provide end-to-end lightpath from Korea to other countries, and vice versa. KRLight is interconnected with other GOLEs, NRENs (National Research and Educational Networks), and institutions in 4 regions, Daejeon (Korea), Seattle/Chicago (USA), Hong Kong (China) and Amsterdam (Netherland). Advanced application researchers and students can make use of KRLight infrastructure that provides them with both circuit oriented connections and packet oriented connections for high performance and large-scale researches in Korea.

KRLight directly connects to international partners including GOLEs, NRENs and research organizations like below.

| | Direct Connection | Direct Peering |
|-----------------------------------|--|---|
| GOLE | Pacific Wave, StarLight, NetherLight, HKOEP | Pacific Wave, StarLight |
| NREN | CANARIE, KREONET/KREONet2, CERNET, CSTNET, ESnet, JGN-X | CANARIE, Internet2, ESnet, TWAREN, CERNET, CSTNET, KREONET / KREONet2, AARnet, NLR, KAREN, JGN-X, NORDUnet, JGN-X, GEANT |
| Research Network and Organization | GLORIAD-US, HK-IX, CERN, NUS, CNGIX-6IX, CUHK, Hong Kong Univ., ASGC, Google, CERN | MREN, CERN, CENIC, UltraLight, GLORIAD-US, ASGC, CUHK, HARNET, APAN-JP, Microsoft, ASNet, Hurricane Electric, Google, CNGIX-6IX, NUS, Hong Kong Univ. |

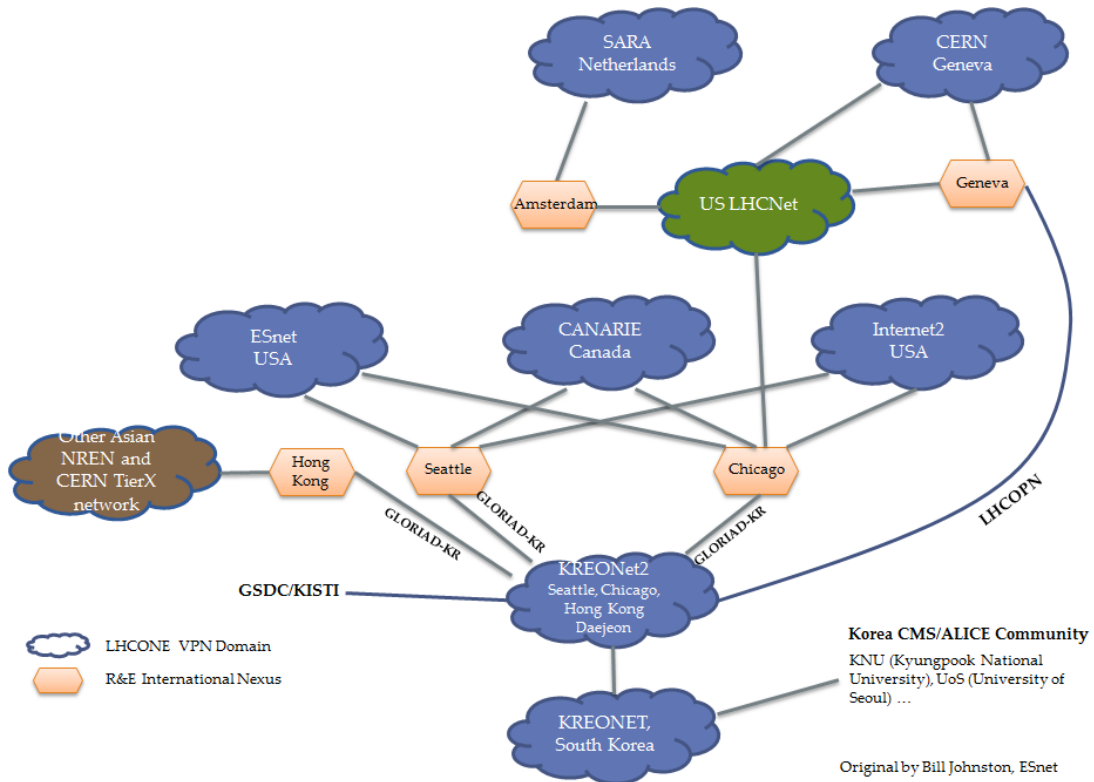
KRLight connectivity (2016)



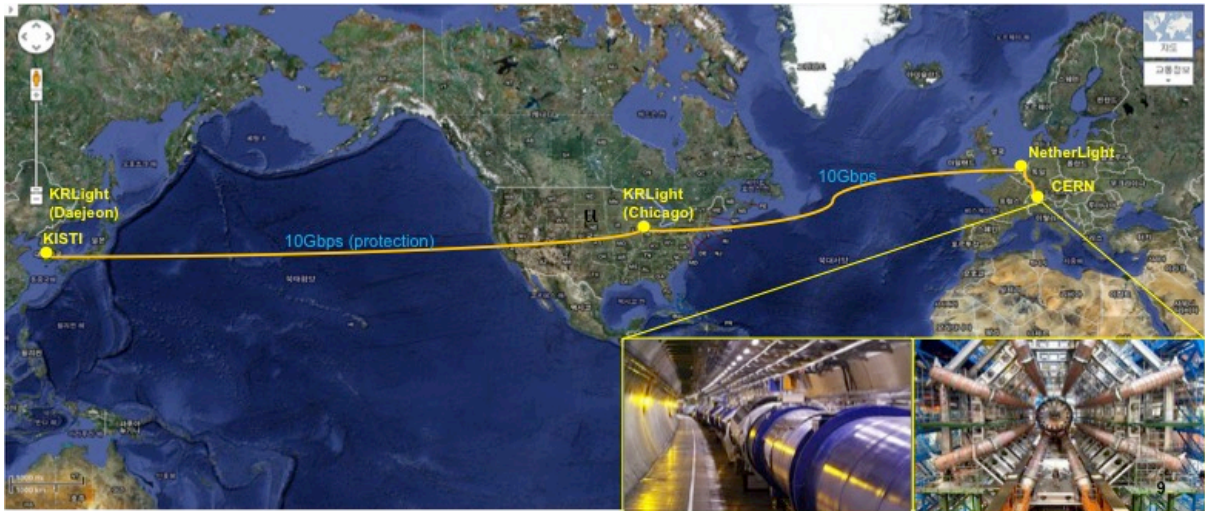
KRLight in GLIF

HEP (High Energy Physics) on KREONET

KREONET has supported HEP (High Energy Physics) community and institution in Korea. KNU (Kyungpook National University) connected to KREONET with 10Gigabit connection, in order to collaborate with CERN, Caltech, and so on. Also UOS (University of Seoul), Korea University, SNU (Seoul National University) has 1Gigabit connection with KREONET for collaboration with CERN, KEK, BNL etc. GSDC (Global Science experimental Data hub Center), connected to KREONET with 2*10Gbps in KISTI promotes data intensive researcher related to KEK, Fermilab, BNL (BROOKHAVEN National Laboratory), LIGO and ALICE of CERN. Also in 2013 GSDC has become a CERN LHC Alice Tier 1 Center based on dedicate 10G link between KISTI and CERN. KREONET/KRLight has provided n*1Gbps or 10Gbps lightpath for several demonstrations about large-scale data transmission related to HEP such as SC (Supercomputing Conference) to KREONET users. In 2016, Along with upgrading GLORIAD-KR infrastructure, KREONet2 and GLORIAD will enable LHCONE network that could connect to LHCONE in ESnet, CANARIE, Internet2 and USLHCNet. It will enhance the performance of HEP network on KREONET.



LHCONE design concept on KREONet2 and GLORIAD



10G LHCOPN between GSDC/KISTI and CERN

Annex 23: SINET4, SINET5 and HEPNet-J (Japan) Update

Submitted by Soh Suzuki (soh.suzuki@kek.jp)

January 2017

Introduction:

National Institute of Informatics (NII) in Japan has been operating SINET that is the primary NREN for High Energy Physics laboratories in Japan. The migration from SINET4 to SINET5 was held in February and March 2015. All prefectures have one Data Center at least and it accepts 100G-LR4, 40G-LR4, 10G-LR, and 1000Base-LX and doesn't accept the SR family. The data center is far than the reach of LR from most of universities, so typically WDM are placed in front of the border switch of the university. Many HEP laboratories in domestic university are connected to HEPnet-J that is implemented as a VPN group on SINET. Before the migration, there was a dedicated port on the SINET switch in each university to connect HEPnet-J, but SINET5 has removed them. At January and February, we re-arranged the physical connection in all sites of HEPnet-J. Some of them got new 10G links on border switches of the university (Fig. 1).

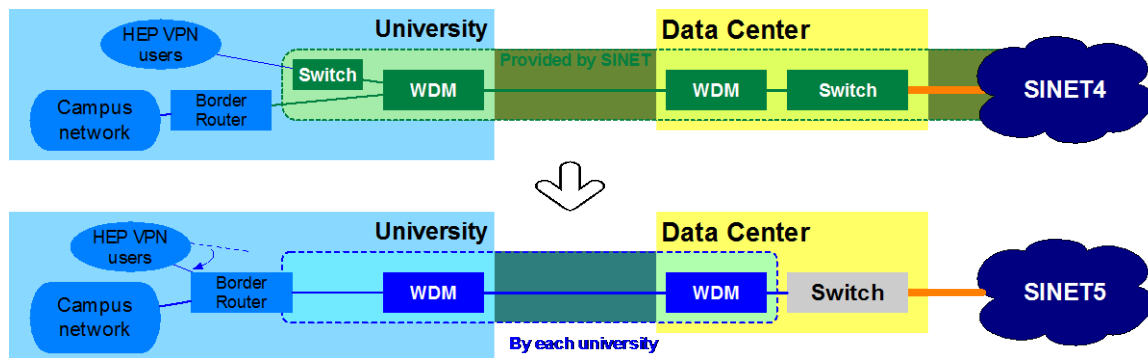


Fig. 1: SINET5 doesn't provide access links to any universities and most of HEP VPN users will get allocation on border routers of campus networks.

SINET4 had four international links, three links of 10G (LAX, MANLAN, WIX) and 10G to Singapore. 10G to MANLAN was mostly used for the traffic between Japan and Europe. SINET5 upgraded the link to LAX by 100G, and replaced the link to WIX by 2 links of 10G to London (Fig. 2). This new direct path reduces RTT between Japan and sites in EU and provides better performance for data transmission (

Fig. 3).

To take advantage of high speed international links, the LHCONE connectivity is indispensable for HEP computing sites. SINET had two international VRF groups for HEP (Fig. 4). The first one is for LHCONE. It connected the LHCONE instance in GÉANT and ATLAS Tier-2 in ICEPP, the university of Tokyo. At the migration from SINET4 to SINET5, the peering at WIX was transplanted into London. The second one connected KEK, PNNL via ESnet, and sites in HEPnet Canada via CANet. As KEK got the permission to connect LHCONE and the central computing system (KEKCC) was upgraded and got the policy routing functionality since Sep 2016, these two VRF groups can be simply unified as a LHCONE instance in SINET (Fig. 5). This unification took about a week because of unexpected trouble during the work. The unification improved the data transfer speed about two or three times faster between KEK and sites in LHCONE. For example, from INFN Napoli to KEK was only 3Gbps, but now it has achieved 8.8Gbps.

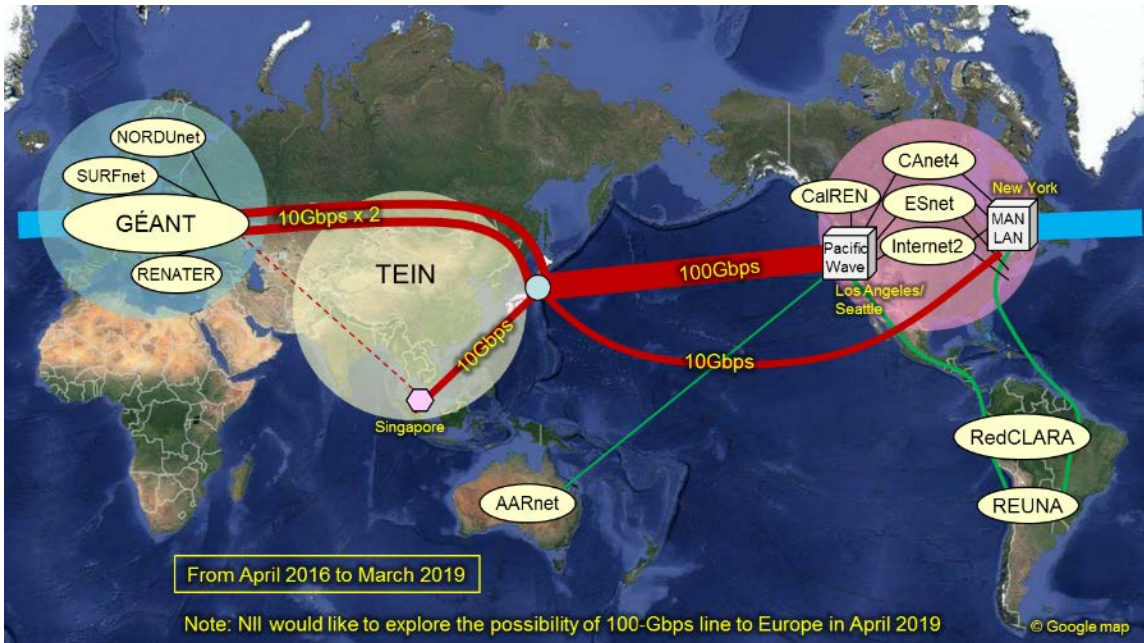


Fig. 2: The international link upgrade of SINET5. The bandwidth to U.S. will be upgraded to 100Gbps and a direct link to Europe will be newly introduced.

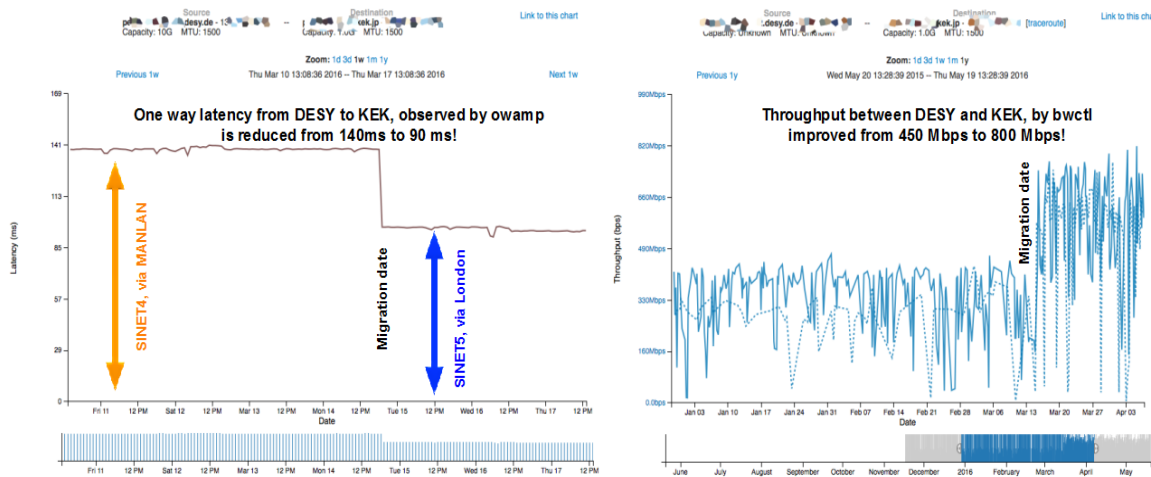


Fig. 3: History graphs of perfSONAR show improvements of direct link from Japan to London.

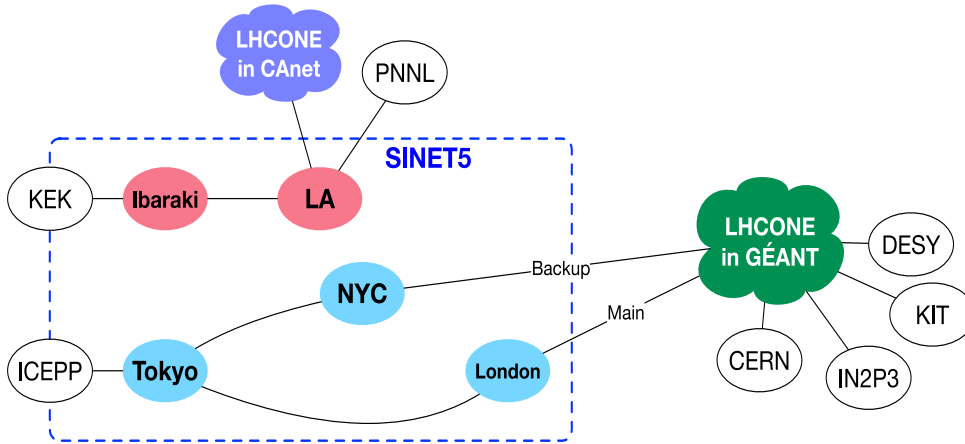


Fig. 4: Before the migration of LHCONE related VPNs in SINET. ICEPP was accessible via LHCONE, but KEK was not so.

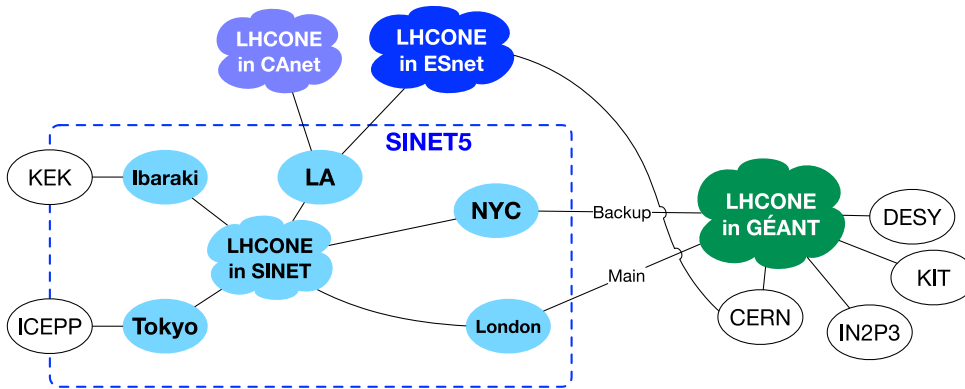


Fig. 5: After the migration, both ICEPP and KEK are accessible from LHCONE.

Annex 24: CERNET2 and CSTNET (China) Update

Submitted by Gang Chen (Gang.Chen@ihep.ac.cn)

IHEP Beijing

January 2017

CSTNET

China Science and Technology Network, or CSTNet is an academic network system operated by Chinese Academy of Sciences.

Chinese Academy of Sciences, CAS, is the leading research organization in China conducting researches in most areas of basic science and technology as well as strategic advanced technologies. It comprises more than 100 research institutes, three universities throughout the country. These institutions are home to about 130 national key labs and engineering centers as well as nearly 200 CAS key labs and engineering centers. Altogether, CAS comprises 1,000 sites and stations across the country.

CSTNet has one core center and 12 regional branch centers, CSTNet provides Internet services to meet the requirements from more than sixty thousands CAS scientists. CSTNet also provides services to many research organizations out of CAS and tollaly connecting 370 Institutes and one million end users around China. The whole CSTNet supports IPv4 and IPv6. The links among major cities are mainly 2.5 Gbps. The middle size cities and some important scientific labs are

connected with access networks of 155Mbps-2.5Gbps in CSTNet. Figure 1 shows the latest status of CSTNet backbone.

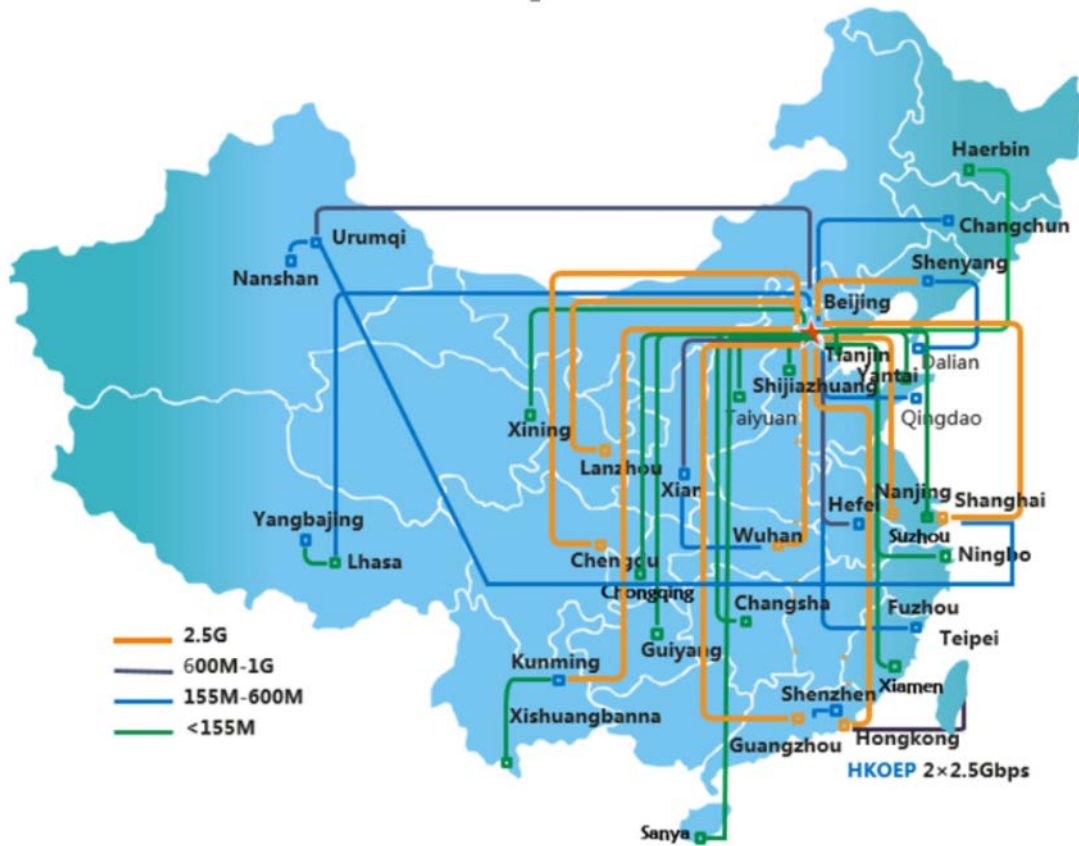


Figure 79: CSTNet domestic backbone.

In 2016, CSTNet upgraded the Seattle to Chicago link from 1Gbps to 10Gbps (Figure 2).

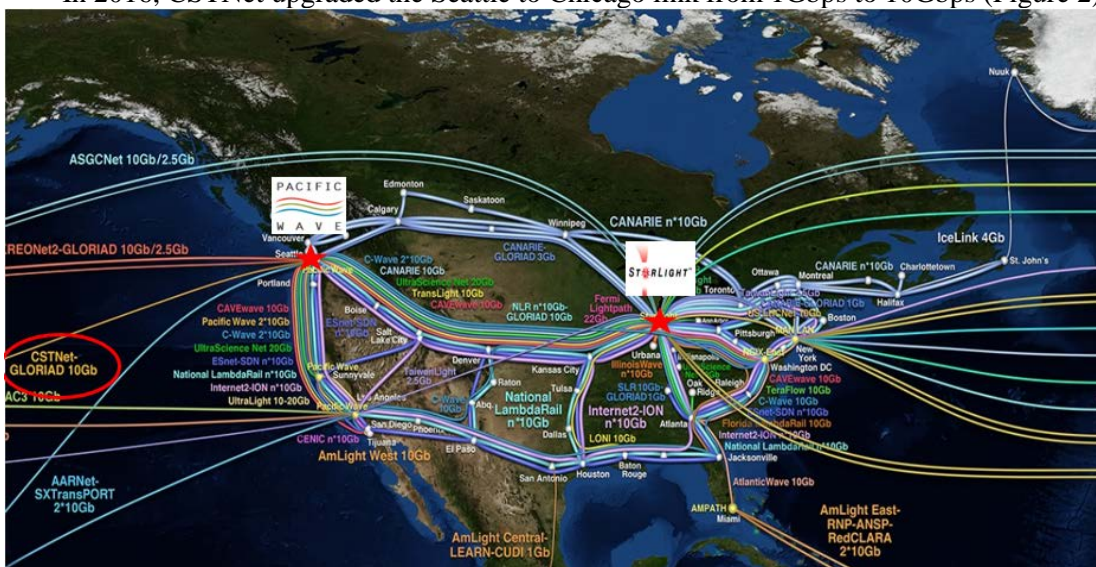


Figure 80: New exchange point in US

CSTNet shares with CERNET the 10Gbps link between Beijing and Europe (ORIENT+) for the China-Europe traffic. ORIENT+ is the major academic network link and the important infrastructure for WLCG and other collaborations between China and Europe. Gloriad link from Hong Kong to the US is 10 Gbps, but the link from Beijing to Hong Kong is still 2X2.5 Gbps. Figure 3 shows the Domestic & International Interconnections of CSTNet.

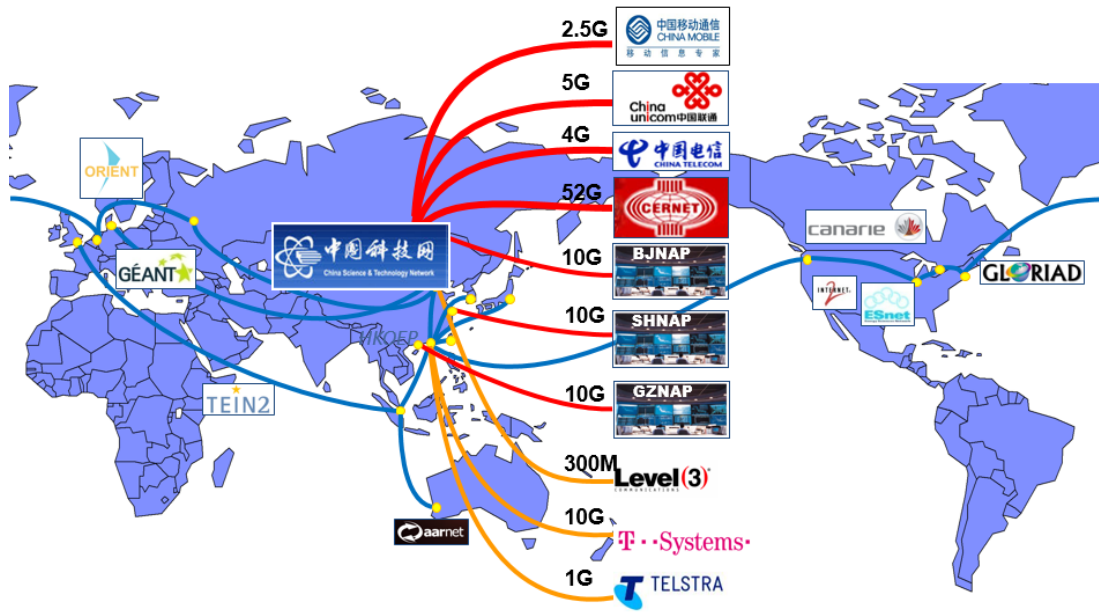


Figure 81: CSTNet Domestic & International Interconnections.

As one of the two important academic networks in China CSTNet offer services to users as followings:

CSTNet provides advanced scientific network services including network management cloud service, network security cloud service, unified communication service, collaboration environment service, network research and experimentation service, etc.

CSTNet provides network security protections including security monitoring and situation awareness, malicious code in-depth analysis, emergency response on information security, security assessment and reinforcement, safety training, etc.

CERNET

CERNET, or China Education and Research Network is the largest academic network in the country. The backbone of CERNET is 10~100Gbps with 38 PoPs in 36 cities and over 2600 universities/institutes/organizations connected, and the total number of CERNET users is more than 25 million.

CERNET has ranked in the one of the largest international academic networks not only in the transportation level but also in the capacity. CERNET is also the first 100Gbps enabled backbone IP network in China.

CERNET has established the international network exchange center in Beijing and Hong Kong, connecting with the United States, Europe and the Asian-pacific region. CERNET has 13 domestic exchange centers all over the country.

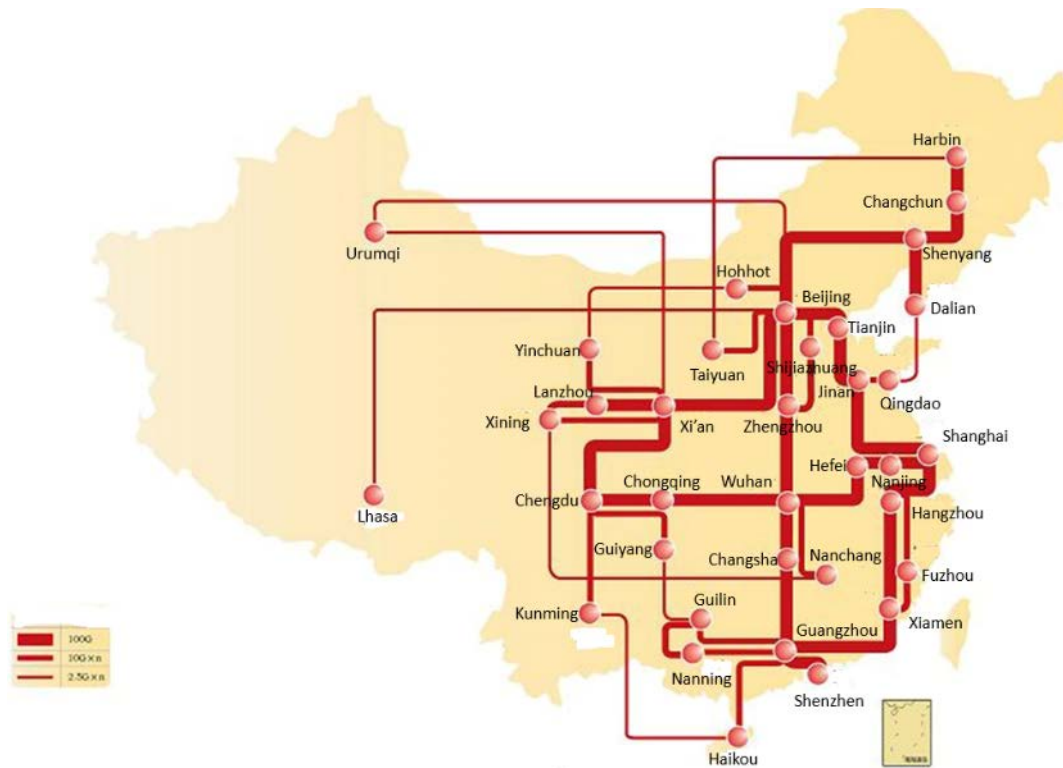


Figure 82: Domestic Backbone of CERNET.

CERNET2

CERNET2 is the second generation of the Education and Research Network in China and one of the Chinese Next Generation Internet (CNGI) core networks.

The backbone of the CERNET2 is 2.5~10Gbps with 25 PoPs in 20 cities and over 600 universities/institutes/organizations connected, and the total number of CERNET2 users (native IPv6) is more than 5 million. CERNET2 has good connections through the Chinese Next Generation Internet Exchange Center (CNGI-6IX). The CNGI-6IX locates in Tsinghua University and links to North America, Europe and Asia-Pacific next generation network with high-speed bandwidth.

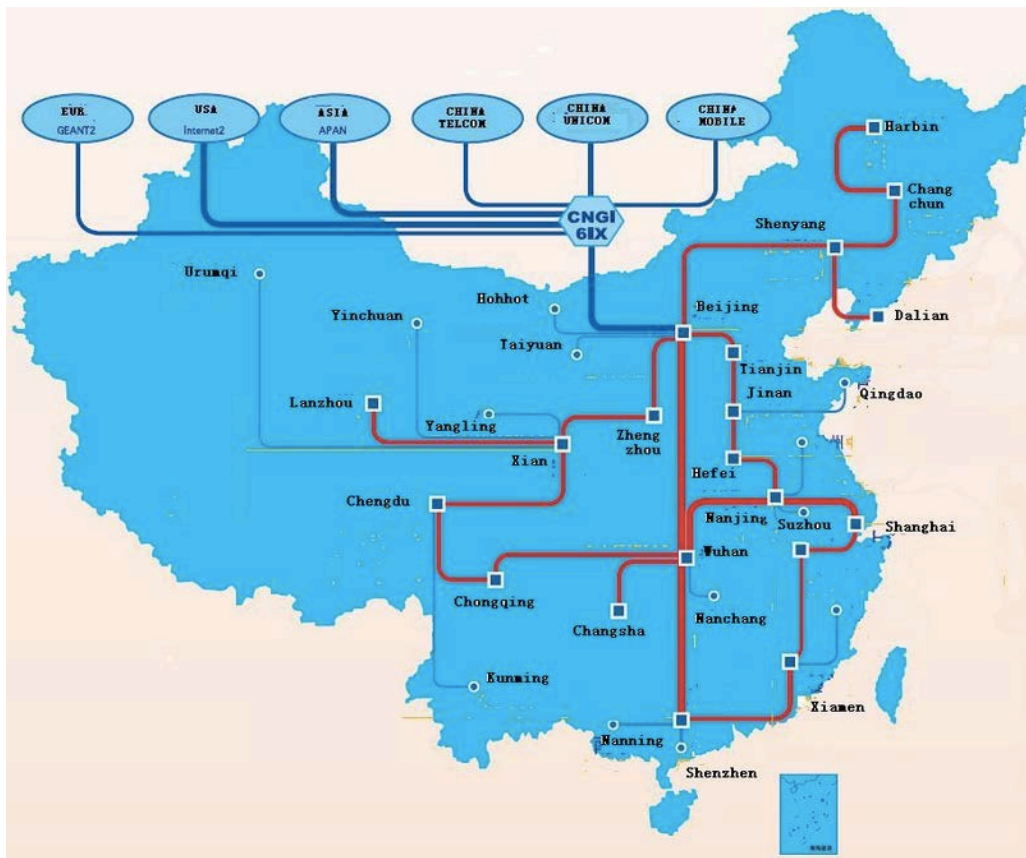


Figure 83: CERNET2 core networks.

Other Activities

In 2016, many tools and services for R&E have been deployed in China for the global network performance monitoring and measurement. A number of perfSONAR hosts have been launched in China and joined into different network performance measurement clouds. At the same time, CSTNet and CERNet have started to deploy the federation platform to support EDUROAM, there are 16 institutes and 36 universities are supporting EDUROAM now in China.

Annex 25: SingAREN (Singapore) Status

Submitted by: Ong Bin Lay, ongbl@singaren.net.sg
February 2017

Introduction:

Singapore Advanced Research and Education Network (SingAREN) is Singapore's national research and education network. It is the sole provider of local and international networks dedicated for serving the Research and Education (R&E) community in Singapore. SingAREN's members consist of the Institutions of Higher Learning, Government and network industry players. Some of the R&E activities that are supported by SingAREN are: E-Learning, video-conferences and research data management across international boundaries. SingAREN is also the Roaming Operator for **eduroam** for the R&E community in Singapore.

SingAREN has launched the Federated Identity Management (FIM) service, termed as **Singapore Access Federation (SGAF)** and **Database Mirroring Service** dbmirror.singaren.net.sg in 2016.

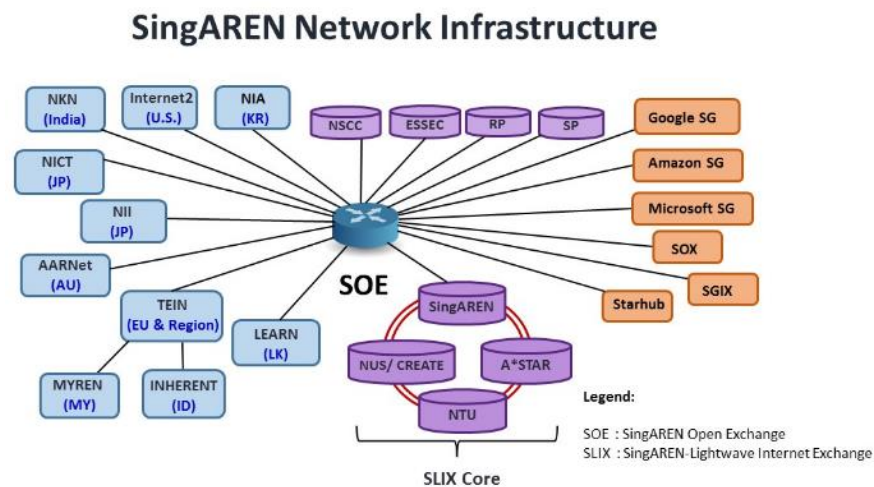


Figure 841: SingAREN Connectivity

SingAREN is pleased to announce the launch of **SingAREN Open Exchange (SOE)**. The objective of SOE is to establish a **resilient and open Internet Exchange in Singapore** to benefit the domestic and international research and education communities.

SOE is co-funded by **National Supercomputing Centre (NSCC) Singapore**, which provides state-of-the-art petascale High Performance Computing (HPC) facility, multi-petabyte data storage and multi-Gigabit speed global connectivity.

With SOE, domestic and international institutions can connect to Singapore at a bandwidth of **up to 100Gbps**, thus enabling research projects with higher bandwidth connectivity requirements. Through connecting to SOE, it facilitates high-speed data transfer, improves routing efficiency, and reduces latency.

SOE serves as a **single point of presence (PoP)** in Singapore for establishing interconnections in Singapore & internationally. SingAREN is connected at 100Gbps to National Supercomputing Centre Singapore (NSCC). On the international connectivity, SingAREN is connected directly to

Internet2 (U.S.A), *GÉANT* (Europe), NII and NICT (Japan), AARNet (Australia), LEARN (Sri Lanka), NIA (South Korea), NKN (India) and via TEIN to other Asia countries. Please refer to [Connectivity](#) for details.

SingAREN has direct 10Gbps connection and peering with Amazon Web Services, Google and Microsoft in Singapore.

For SingAREN's connectivity, please refer to <https://www.singaren.net.sg/connectivity.php>.

2 Applications of SOE Network.

The SOE network has been used by Singapore's R&E community for various applications:

- i. SingAREN facilitates large data transfers for the **GenomeAsia 100K Project** (<http://www.genomeasia100k.com>) through its resilient international links & high-speed fiber network. The project aims to sequence 100,000 genomes from various South, North, and East Asia populations, with the objective to accelerate precision medicine and clinical application for Asian patients by leveraging new information and understanding from the 100,000 genomes.

Case Study for GenomeAsia 100K project:
[“Precision Medicine: Improving how diseases are treated”](#)

- ii. SingAREN enables researchers to have high-speed access to overseas genomic research databases, thus resulting in a tremendous improvement in the genomic data download time.

Case Study for Cancer Science Institute of Singapore (CSI Singapore):
[“Enabling Cancer Genomics Research with Advanced R&E Networks”](#)

- iii. Support to A*CRC (Singapore) for Network Demonstration at **Supercomputing Conference (SC16):**

SingAREN was one of the collaboration partners, together with other international organizations (TEIN, GEANT, PIONEER, RENATER, Internet2, SCinet), in enabling A*CRC's **InfiniCortex Demonstration**. The countries involved in the network demonstration were Singapore, Poland, France and U.S. A global WAN Infiniband network was set up with 4 distinct subnets:

- NSCC InfiniCloud, Singapore;
- Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), Poland;
- A*CRC, Singapore + University of Reims Champagne-Ardenne (URCA), France + George Washington University (GWU), U.S. ;
- Stony Brook University (SBU), U.S.

3 Launch of New Services - SingAREN

3.1 Identity Federation

SingAREN has launched the Federated Identity Management (FIM) service, termed as **Singapore Access Federation (SGAF)**, to Singapore's R&E community on 14 July 2016. SGAF is a **federated authentication and authorization system** to enable scalable, trusted collaborations among Singapore's research and education community. <http://www.singaren.net.sg/SGAF.php>

These are the service providers for SGAF at the moment:

- **Cloud Computational Services** - Enables authorized institutional users to request and provision Virtual Machine (VM) resources for research and academic purposes.
- **National Supercomputing Centre (NSCC) Singapore user portal**
- **FileSender** - A web based application that allows authenticated users to securely and easily send arbitrarily large files to other users.

http://www.singaren.net.sg/library/newsroom/20160719_SGAF_Launch_event_14July16.pdf

3.2 Database Mirroring Service

SingAREN has launched the **Database Mirroring Service** dbmirror.singaren.net.sg in April 2016. It is a one-stop portal which mirrors major overseas scientific databases in Singapore, thus enabling collaborations by the research communities in Singapore and overseas. The database mirroring service is made available to Singapore's Research and Education community, in particular SingAREN members with needs for regular and frequent access to large scientific databases that are hosted overseas.

http://www.singaren.net.sg/library/newsroom/20160429_Database_Mirroring_service_launch.pdf

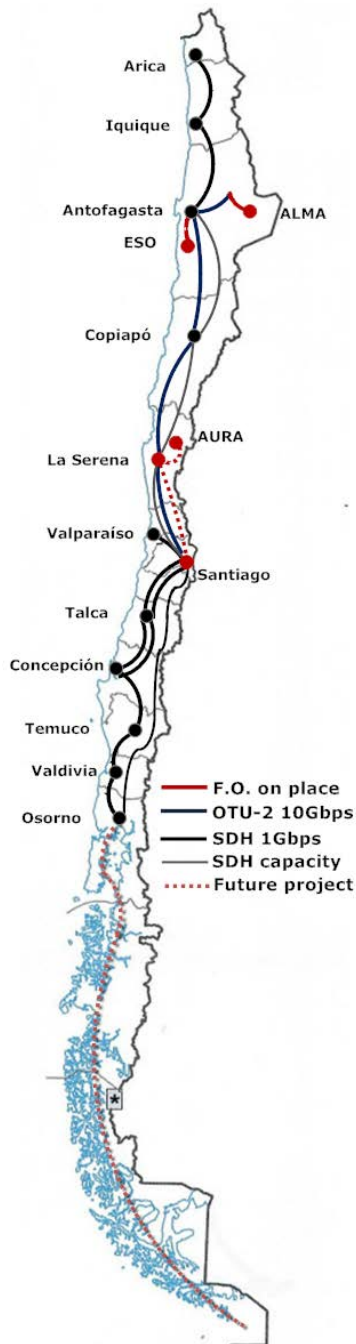
4 Future Plans

SingAREN will provide more value-added services to benefit Singapore's research and education community.

SingAREN will seek to support international network collaborations through engagement with our partners to connect at SOE for establishing interconnections with other countries.

Annex 26: REUNA (Chile) Status

Submitted by Sandra Jaque (sjaque@reuna.cl), Albert Astudillo (aastudil@reuna.cl)
January 2015



In the frame of its Strategic Plan 2013-2017, REUNA is working to have a full national photonic network based on a DWDM backbone which will allow the Corporation to have the capacity needed by the research and education community and also a robust network infrastructure. With that goal in mind, during 2013 and 2014 we have made improvements, and integrated new projects:

Improvements 2013-2014:

- A second 1Gbps over 500Kms from Santiago to the south, involving three backbone nodes: Santiago, Talca and Concepción.
- A backup capacity, 1Gbps over 900Kms between Santiago and Osorno. This also allows to have a ring along this path.



ALMA network project:

REUNA collaborates with ALMA to deploy a DWDM network solution from the summit to their facilities in Santiago. The solution consist in 150kms of new fiber between summit to the nearest City (Calama) equipped with a DWDM solution, from there - over an OTU-4 wavelength - to the nearest PoP of REUNA backbone (Antofagasta)

and from this PoP the data is transported to Santiago.

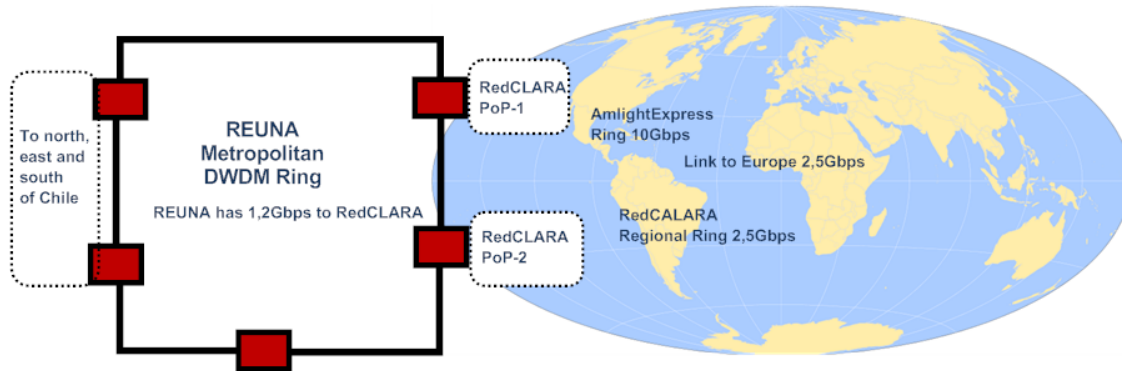
Future Improvements:

REUNA will continue seeking to have direct access to fiber along the country, on that way; during 2015 will be implementing its first long distance fiber path. Around 700Kms of fiber that will be lighted with coherent lambdas at 100Gbps, this infrastructure allows REUNA to gives the communication solution needed by the LSST telescope.

Another relevant opportunity happening in Chile is that the government has decided to support the deployment of a network infrastructure to the southern of Chile. Over 3000kms that will be cover mainly with submarine fiber. Today those cities are reached mainly by Argentina or satellite solutions. For REUNA the project is very relevant because will allow to reach the research and education community with the capacity needed by them.

International connectivity

REUNA is a founding member of RedCLARA and connected to this LatinAmerican NREN backbone since 2005. Currently REUNA connects over a ring of dark fiber to the two PoPs of RedCLARA in Santiago. The ring has a mixture of DWDM and Ethernet technologies and we are working to improve it, so the goal is to have a full DWDM optical path switch over the ring. REUNA is also connected in Santiago by RedCLARA to Amlight links facilities. Currently, in term of contracted bandwidth, we have 1,2Gbps activated but physically the links are 10Gbps.



References:

REUNA: www.reuna.cl.

ALMA: <http://www.almaobservatory.org> Network project: <http://goo.gl/uxtu0g>

LSST: <http://www.lsst.org/lsst/> Network project: <http://goo.gl/vWXbst>

RedCLARA: <http://www.redclara.net/>

Amlight: <http://www.amlight.net/>

Advanced Network Projects

Annex 27: A Next Generation Terabit/sec SDN Architecture and Data Intensive Applications for High Energy Physics and Exascale Science

By Harvey Newman (newman@hep.caltech.edu)

2016

INTRODUCTION

The largest data- and network-intensive programs today, from the Upgraded High Luminosity Large Hadron Collider (HL LHC) program (A Large Ion Collider Experiment, n.d.), to the LSST (The Large Synoptic Survey Telescope, n.d.) and SKA (Square Kilometre Array, n.d.) astrophysics surveys and many other data-intensive areas of science, face unprecedented challenges in global data distribution, processing, access and analysis, and in the coordinated use of massive but still limited CPU, storage and network resources.

In response to these challenges Caltech together with CENIC, ESnet, Fermilab, Starlight/iCAIR, AmLight/FIU, SPRACE/UNESP, Yale and other key laboratory, university and industry partners has designed and is developing the first stages of an SDN-driven Next Generation Integrated Architecture (NGenIA) for HEP and global scale science (Next Generation Exascale Network Integrated Architecture for HEP and Global Science, n.d.). While the initial focus will be on the challenging LHC use case, the products developed will be general, and apply to many fields of data intensive science.

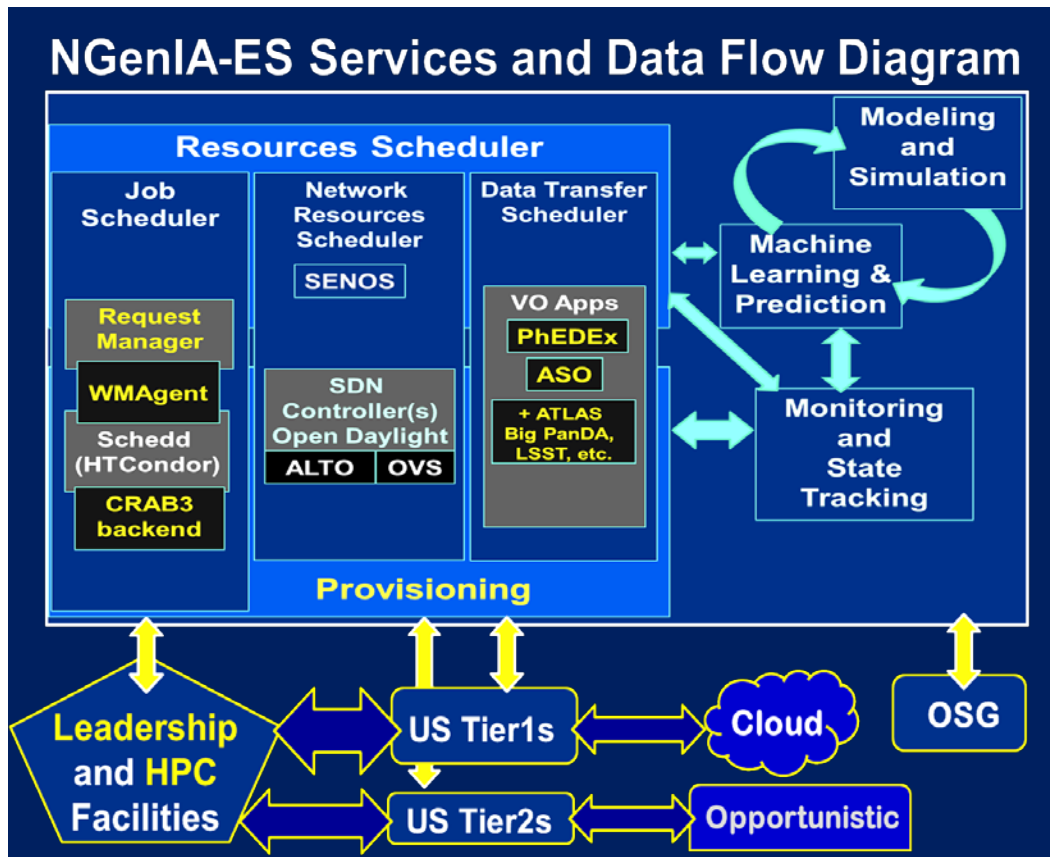


Figure 85. The NGenIA-ES network-integrated system under development for the LHC experiments, bringing the Leadership and other HPC facilities into the HEP data intensive ecosystem.

A recent extension (NGenIA-ES) of this program (illustrated in Figure 93) to be demonstrated at SC16 through a set of balanced state of the art SDN-managed flows fully matching a Terabit/sec complex network driven by the latest data transfer nodes (DTNs), is petabyte-scale transactions among HEP Tier1 and Tier2 sites and the edge systems of the pre-exascale and exascale HPC facilities now deployed, and being planned over the next technology generations, at the leadership DOE and NSF centers. This forms the basis of the progressive integration of such facilities into the data and computationally intensive ecosystem of the LHC and other major HEP physics programs.

This paper reviews the multifaceted recent work of the team, and summarizes the earlier results shown at SC15. The SC16 demonstrations include a fundamentally new concept of “consistent network operations,” where stable load balanced workflows crossing optimally chosen network paths, up to preset *high water marks* to accommodate other traffic, are provided by autonomous site-resident services, dynamically interacting with network-resident services, in response to requests from the science programs’ principal data distribution and management systems. This is being implemented in support of petascale workflows using:

- Protocol agnostic (Open vSwitch-based) (**Open vSwitch project, n.d.**) traffic shaping services at the site egress that will provide stable, predictable data transfer rates, and auto-configuration of data transfer nodes,
- A substantially extended OpenDaylight (**The OpenDaylight Platform, n.d.**) controller using a unified control plane programming framework, consisting of novel components including the Application Level Traffic Optimization (ALTO) Protocol, a min-max fair resource allocation algorithm-set providing flow control and load balancing in the network core, a data-driven function store for high-level, change-oblivious SDN programming, and a data-path oblivious high-level programming framework.
- Smart middleware to interface to SDN-orchestrated data flows over network paths with guaranteed bandwidth all the way to a set of high performance data transfer nodes (DTNs),
- Machine learning for identification of the key variables controlling the system’s throughput and stability, and for overall system optimization.

The accumulated knowledge of this development program also will serve to inform the design of the following generations of distributed petabit/sec systems, including continental scale instruments such as SKA, and the exascale computing systems of the next decade harnessing zettabyte datasets.

The team’s demonstrations this year also will include:

- High speed LHC data classification using a compact multi-GPU server coupled to rapid database traversal and deep learning methods, as well as other machine learning applications.
- A Virtual Reality “CMS.VR.v1” event display of the CMS experiment at the LHC, developed together with NovaVR LLC, and FNAL for the CMS Collaboration. The display and software are based on NVIDIA graphics equipment and the Unity3d gaming engine, and versions for Oculus Rift and HTC Vive systems have already been developed. The VR project, initially scoped for public outreach is planned to be powered by AI methods in upcoming implementations, and used for data monitoring and verification in the experiment. The VR experience includes both a virtual tour of the CMS experiment, and embeds a set of CMS’ publicly available data events. This is an immersive experience where one

can navigate inside the detector and view the tracks, vertices and other energy deposits from the LHC proton-proton collisions.

Supercomputing 2015: A Pilot with Terabit/sec Transactions for LCFs in the Pre-Exascale Era

During SC15 in Austin an international team of Caltech, SPRACE Sao Paulo and the Univ. of Michigan, together with teams from FIU, Vanderbilt and support from vendors including Dell, Mangstor, Mellanox, QLogic, SGI and Spirent worked to demonstrate large data flow transfers across a highly intelligent SDN network. The networks for this work were supported by SCinet Network Research Exhibition (NRE), ESnet, Century Link, CENIC and Pacific Wave.

The global picture of SC15 demonstrations involving various WAN paths is shown in Figure 2. The SDN demonstrations revolved around an OpenFlow ring connecting seven different booths and the WAN connections to Caltech and other campuses in the Pacific Research Platform, Michigan, Starlight, CERN and Sao Paulo. Some of the WAN connections were built using NSI (Kudoh, Roberts, & Monga, 2013) dynamic circuits and stitched together to form end-to-end paths using a custom SDN application. All of the remote switches were controlled by a single controller in the Caltech booth on the show floor. The results were presented at the 2015 INDIS (INDIS Innovating the Network for Data-Intensive Science, n.d.) and SDN (ESNet: Software Defined Networking (SDN) Workshop, n.d.) workshops during the conference. A paper (al., 2015) was published in IEEE ACM.

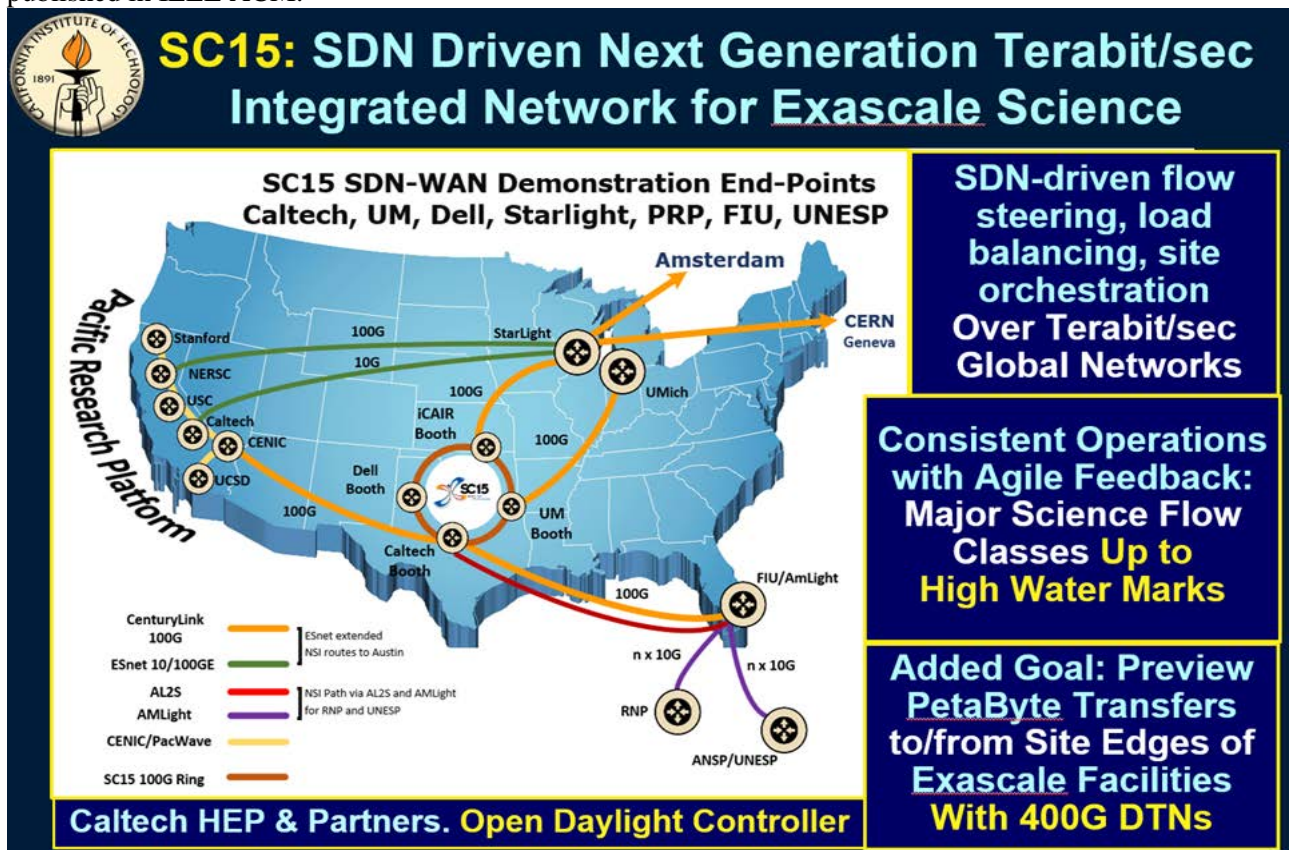


Figure 86. Global picture of the Exascale NGENIA prototype demonstration at SC15.

While terabit/sec aggregate flows were already achieved during SC14, the SC15 demonstrations included the use of a large set (29) of the very first 100 Gbps network interfaces on servers, as needed to support the large data transactions foreseen with the Argonne Leadership Computing

Facility (ALCF) (Argonne Leadership Computing Facility, n.d.) in the pre-exascale era using Theta and subsequently Aurora in 2016-19, on the way to exascale operations with petabyte data transactions. The deployment at SC15 included several Dell and Supermicro servers each capable of stable bidirectional flows to and from a single port of greater than 100Gbps (illustrated in Figure 3) and stable aggregate flows per server of > 300 Gbps using four network interfaces and Caltech's FDT (Fast Data Transfer (FDT), n.d.). The overall throughput capability deployed in one rack at SC15 was 1.5 Terabits/sec, matching (and exceeding) the local network connections.

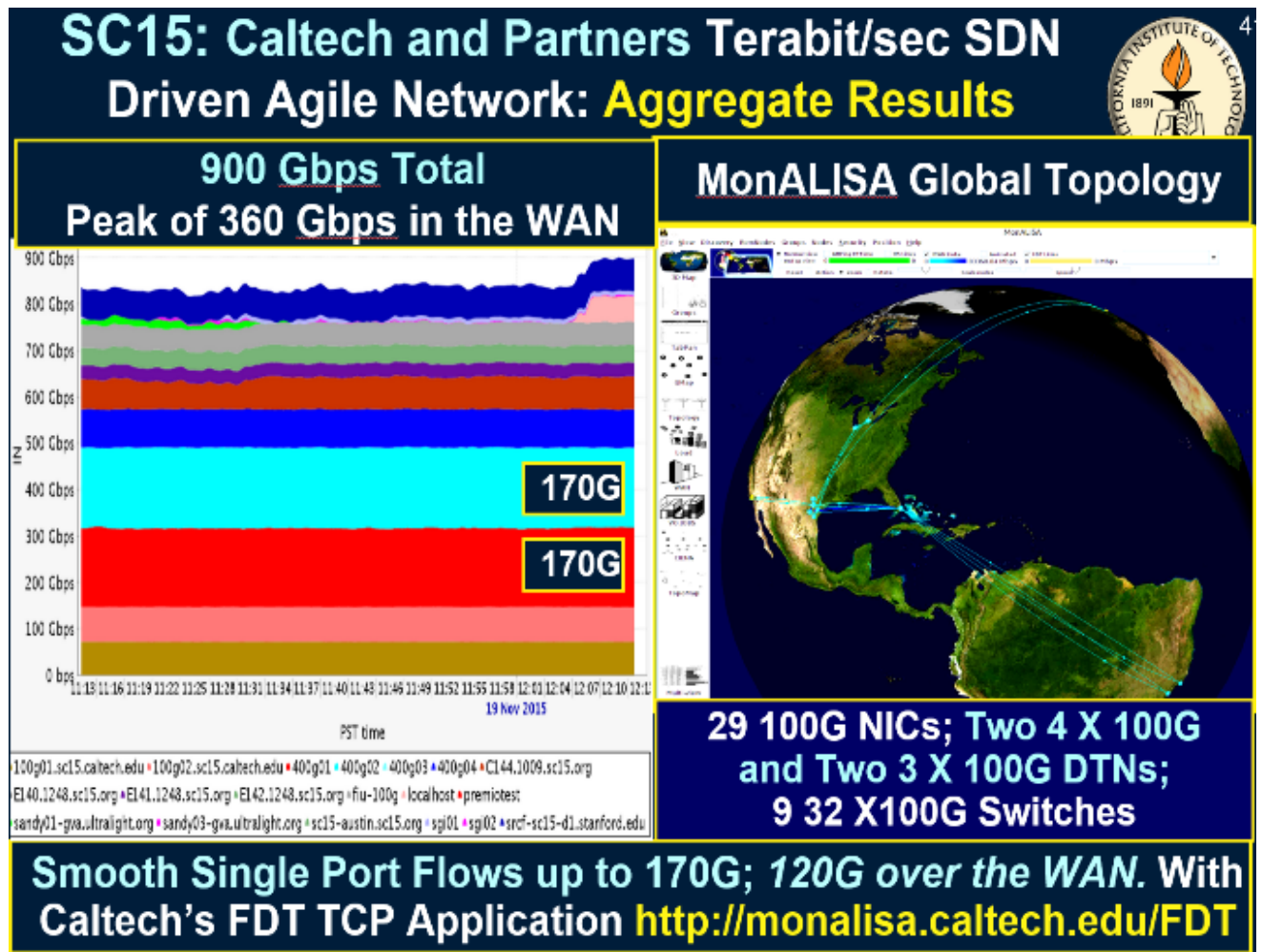


Figure 87. Examples of record throughputs and network monitoring at the SC15 conference using Caltech's MonALISA monitoring and FDT. The results achieved are appropriate for supporting sub-petabyte transactions in the pre-exascale era, as part of the operational model discussed in the text.

The success of the SC15 demonstrations led to the design, commissioning and testing of servers for SC16 capable of 1 Terabit/sec, appropriate for use with exascale-era computing systems and next (to next) generations of networks, together with Dell, EchoStreams, 2CRSI, Mellanox, HGST and other partners.

Network-Endsite Flow Rate Limits up to 100G Wire Speed with Open vSwitch

Following SC15, the team developed the use of Open vSwitch to manage “orchestration” of operations among the end sites and the network, under software control.

A key component of the control software is Open Virtual Switch (OVS) (Open vSwitch project, n.d.), which can make end-hosts appear as a switch. Available as part of the major standard Linux distributions, and supporting standard, well-established protocols for internal management, security, monitoring, and automated control for traffic flows, OVS enables network automation via programmatic interfaces, especially for the virtualized environments, and thus becomes an integral part of our control software framework. Following developments and patches of the latest OVS traffic control module versions by the team, OVS was shown to perform extremely well, with the ability to stably limit traffic rates at any level up to 100 Gbps wire speed, with very low CPU overhead, as illustrated in Figure 4.

In achieving smooth high rate flows, especially flows at the rates shown in the figure, it is important to note that transfers do not involve files per se, but rather data buffers sent and received by Caltech’s FDT (Fast Data Transfer (FDT), n.d.) application. FDT is able to decompose any structure of multiple files into buffers whose dimension is adapted to the I/O capability of the end-systems involved in a transfer, and send the buffers to the network at a rate compatible with the real-time capacity of the network path which is monitored (using MonALISA (MONitoring Agents using a Large Integrated Services Architecture (MonALISA), n.d.)) in real-time, resulting in smooth “impedance matched” flows. The file structure is then restored to its original form at the destination.

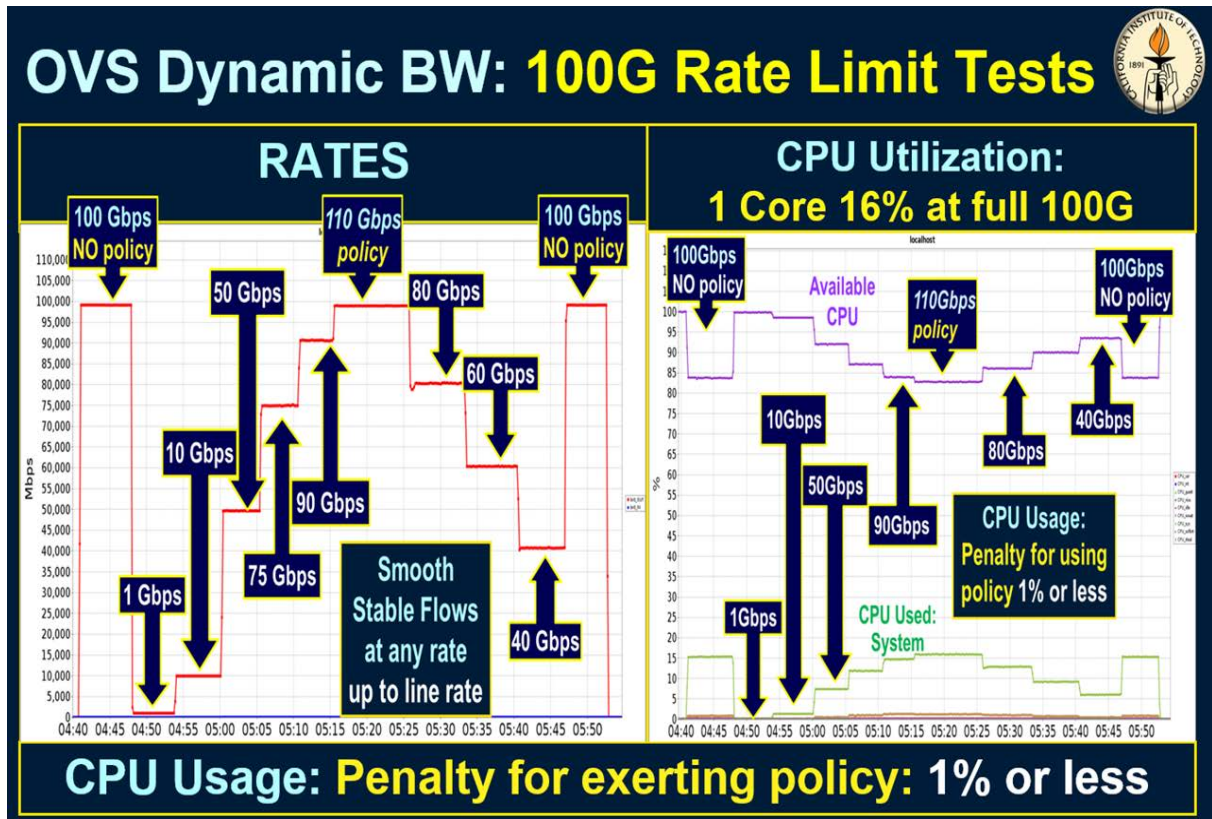


Figure 88. Tests of Open vSwitch to dynamically control bandwidth between servers with 100G interfaces. The ability to smoothly control the flows using OVS and FDT at any level up to wire

speed is shown on the left, and the very small added CPU load resulting from exerting a rate limit(1% or less) is shown on the right.

Next Generation SDN for Consistent Operations with Load-balanced High Throughput Data Flows

Building on the preceding OVS software component, we designed the second key component: a unified control plane programming framework for a next generation “consistent operations” SDN paradigm. The goal of the paradigm is to achieve orchestration of multiple, stable high throughput flows among the end sites, supporting the large set of HEP and other data flows without impeding other network traffic (from other science programs or general purpose traffic), following the NGENIA-ES architecture described above, and through high-level SDN programming.

End-to-end Flow Control, Path Assignments and Bandwidth Allocations

Figure 97 shows two key components involved in establishing an optimized set of high throughput flows among multiple end sites, within the constraints: (1) Open vSwitch (OVS) (Open vSwitch project, n.d.) to stably rate limit the flows at the edges, and (2) Application Layer Traffic Optimization (ALTO) in OpenDaylight (The OpenDaylight Platform, n.d.) for end-to-end optimal path creation, coupled to flow metering and high watermarks set in the network core. We have teamed up with Richard Yang’s computer science group at Yale University (Yale University, n.d.) for the integration of ALTO modules with CMS’ PhEDEx (V. Lapadatescu, T. Wildish et al, 2014) and ASO (J.Balcas, 2016) (Riahi et al., 2015) applications. This methodology will query the replica tuples for a given set of datasets. Once the end nodes with datasets are identified then the MonALISA scheduler using ALTO modules will create multi-domain end to end transfer paths, assign each transfer to a given path, and allocate defined bandwidth levels to each transfer.

The allocations are subsequently be adjusted using OpenFlow flow-metering functions in response to requests from user applications or controllers, or by the requirements of the NGENIA-ES system itself, as described in the previous section. The flow-metering in the network core will be fed back to the OVS instances at the edges, and changes will be applied at a rate consistent with the smooth progress of end-to-end flows.

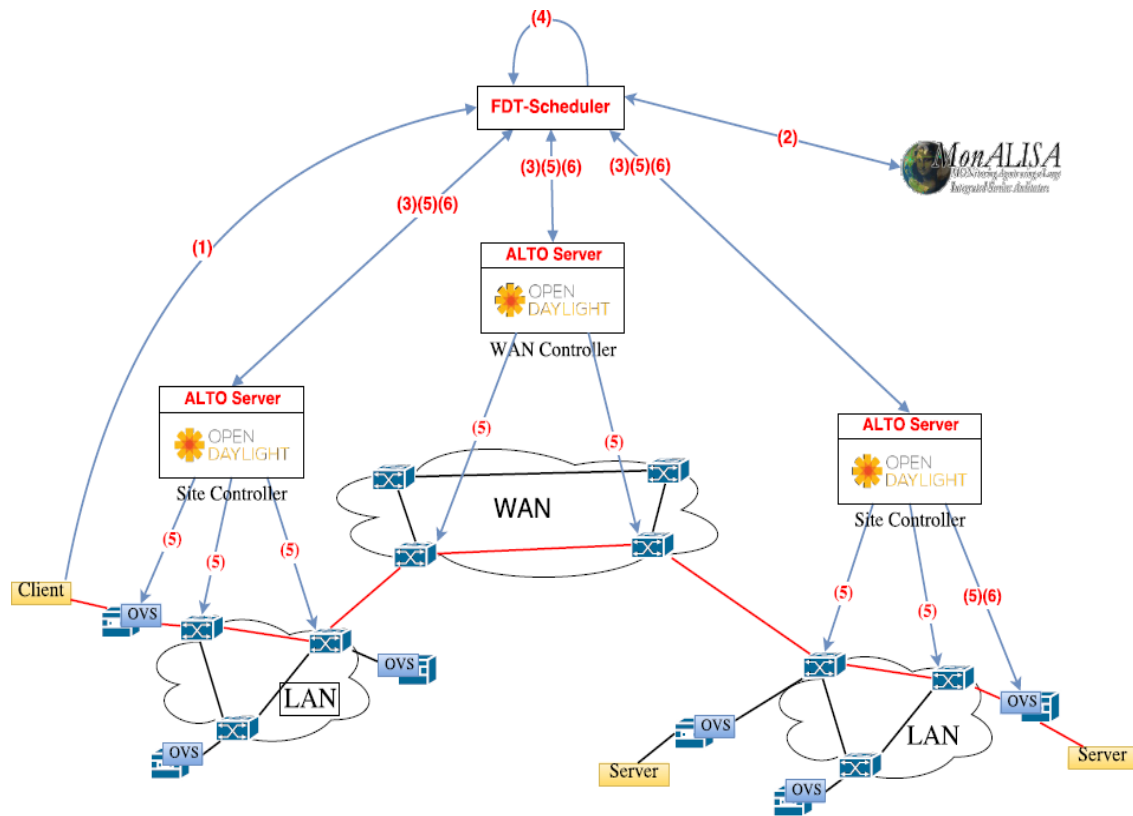


Figure 89. CMS file transfers using ALTO and MonALISA FDT Schedulers.

The Yale/CS team working with Caltech has begun to develop an iterative “Max-Min Fair Resource Allocation” (MRFA) scheduling solution to the resource allocation problem whose aim is to minimize the maximal time to complete a transfer subject to a set of constraints, applied to the NGENIA-ES case. The Yale team has extensive research and experience with techniques for optimizing inter-datacenter transfers across a complex wide area network topology (Neely, 2010) (Chiang, 2006) (E. L. Lawler, 1993) (R. K. Ahuja) (Gallager, 2001) and with practical solutions and extensions of the well-studied “Max-Min Fair Sharing” (MMF) problem (X. Lu, 2015) (E. Danna, 2012) (Shavitt, 2008), as well as frontline SDN developments in the OpenDaylight framework. The constraints include the priority of each class of flows, expressed in terms of upper and lower limits on the allocated bandwidth between the source and destination for each transfer, and the capacity (or maximum sustainable aggregate throughput in practice) of each link in the network.

Additional objectives of the dataset transfer scheduler are to (1) fully utilize the network resources, i.e., the bandwidth, so that all dataset transfers can be completed in a timely manner; (2) allocate network resources fairly so that no transfer request suffers starvation, leading to an excessively long time to completion of the transfer, and (3) load balance among the sites and the network paths crossing a complex network topology so that no site and no network link is oversubscribed, and (4) keep the total bandwidth allocated to the flows under its management below prescribed high water marks, to accommodate other network traffic.

Orchestration Among Multiple Host Groups With Diverse Paths and Policies

The concept of NGENIA-ES’s end-to-end orchestration of data flows involving multiple host groups at many sites, multiple diverse network paths among them, and diverse policies governing

the path setup and prioritization of flows, is illustrated in Figure 98. Diverse network paths are constructed to support sets of flows, each of which may be assigned bandwidth individually or in groups according to the priorities and policies within and among the science programs.

In order to meet the demands and adapt to changing network conditions in real-time, the NGenIA-ES system will respond to (1) requests from high level applications (PhEDEX and ASO are shown as examples), (2) shell script commands, (3) other upstream SDN controllers. Adjustment of the allocations can be triggered by (1) new requests, (2) real-time feedback based on the progress of transfers, (3) network state changes or error conditions, (4) proactive load-balancing operations, or (5) orchestration operations imposed by controllers or emerging network operating systems (the example of ESnet’s SENOS (SENSE Kick-Off Meeting, n.d.) is shown) that manage and stabilize operations in the wide area network core.

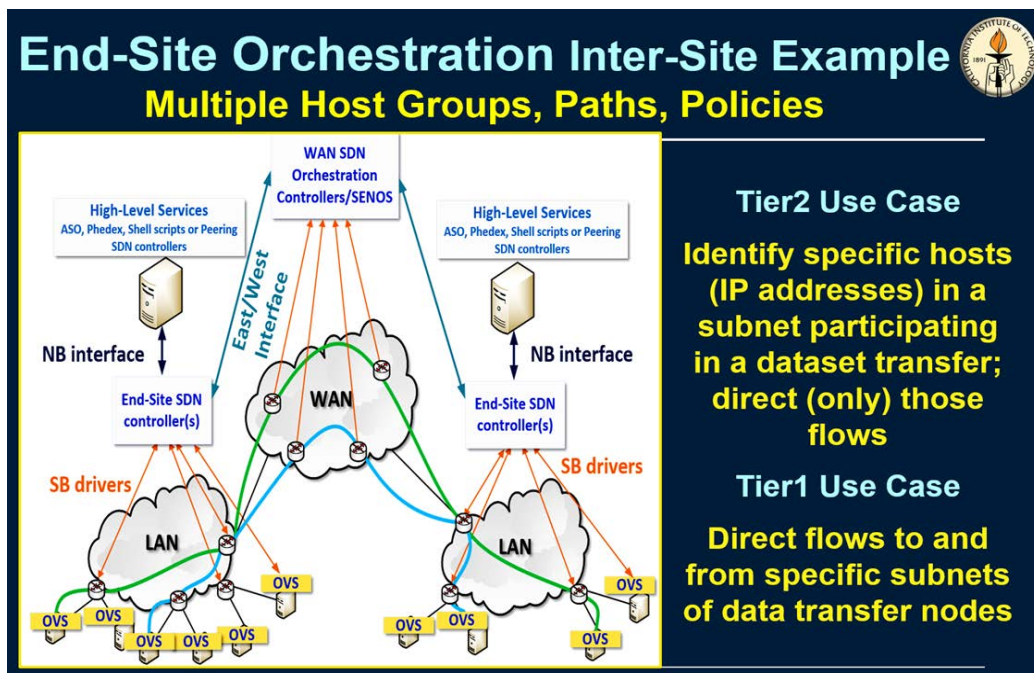


Figure 90 Illustration of construction of network paths and management of flows among multiple clusters at multiple sites using a set of OVS instances, SDN switches and controllers.

Pervasive monitoring and tracking of operations supporting the orchestration functions will be provided by Caltech’s MonALISA (MONitoring Agents using a Large Integrated Services Architecture (MonALISA), n.d.) monitoring system. As SDN frameworks such as OpenDaylight (The OpenDaylight Platform, n.d.) mature, it is expected that some of the monitoring functions may be migrated to standard SDN services.

Caltech State of the Art SDN and DTN Testbed

The developments described above are being carried out on a testbed based at Caltech, with extensions to StarLight, Michigan and other sites. Development of this testbed, which will be an integral part of our SC16 demonstrations. The testbed began at SC13, came to the fore in SC15 through our working in partnership with the Sao Paulo team of S. Novaes et al. and has since been under further development, including the incorporation of ALTO and high-level SDN programming, in cooperation with the Yale team.

The testbed currently involves 13 network switches with many 10G, 40G and 100G ports and 16 servers, several of which are capable of full 100G and higher throughput. Most of the switches and

servers shown have been obtained through the NSF DYNES (J. Zurawski, E. Boyd et al., 2011) , ANSE (LHCONE Point-to-Point Service Workshop, December 2012, n.d.) and CHOPIN (CHOPIN project, n.d.) projects, or through donations by the manufacturers. An example of the automatically discovered real-time topology of the SDN testbed, using the Caltech OpenDaylight controller, showing the ports and interconnections and some of the flows installed using OpenFlow, is shown in Figure 99.

A Next Generation Terabit/sec Network and Applications Architecture at SC16

The terabit demonstrations at SC16 are designed based on two completely different transport platforms using either standard TCP/IP protocol or the RoCE [41] using RDMA. Both of these demonstrations are an integral part of the larger scale demonstrations of the new SDN-driven network paradigm described in earlier sections, and integrated in the SDN controller that will install Terabit/sec flows. The network layout on the SC16 floor and connections SCinet and the metro- and wide area network connections are shown schematically in Figure 100.

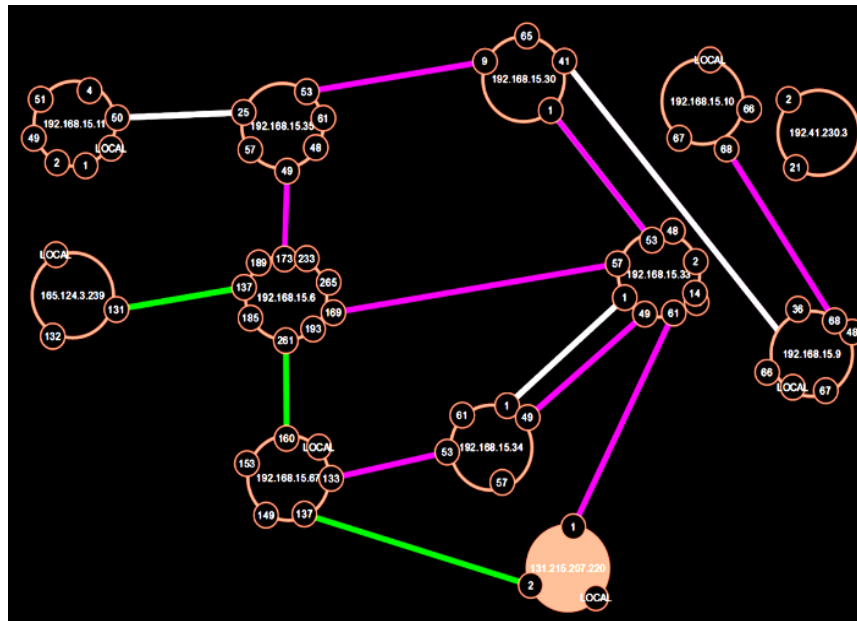


Figure 91 An example of the auto-discovered topology of the Caltech SDN + DTN development testbed.

The first Terabit network is between the StarLight (Data Intensive Science) and the Caltech booths which will highlight the CPU bypass using data transfer from user’s application memory address space to the network. A single SuperMicro server with 10 x 100GE Mellanox NICs and connected through a Dell Z9100 switch will be used to generate RoCE traffic to Caltech Booth. A second Terabit/sec demonstration will utilize four Dell R930 servers in order to highlight TCP/IP based data transfer between the two Caltech booths as well as exchanging large data flows with the remote sites. Caltech’s FDT application will be used to transfer data at Terabit among two pair of servers. Both demos will be interconnected across the OTN Terabit infrastructure provided by Ciena, Cisco,

⁴¹ RDMA over Converged Ethernet (RoCE): http://www.mellanox.com/page/products_dyn?product_family=79

Coriant and deployed with the help of SCinet. The Caltech SDN controller will govern and steer the traffic flows among the booths and across the whole SDN infrastructure.

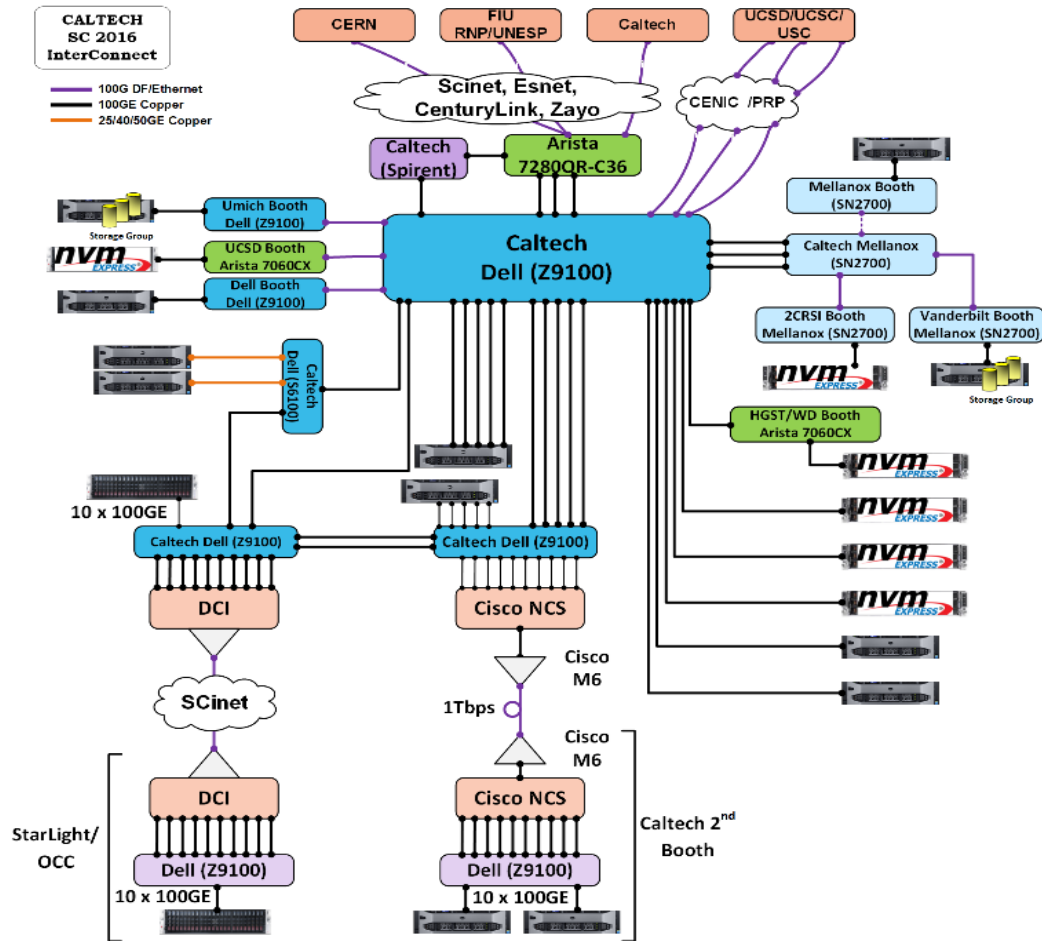


Figure 92 SC16 terabit/sec network layout showing the Caltech, Starlight and partner booths, servers and external connection

The FIU, RNP, ANSP and UNESP teams will exercise two new 100 Gbps connections between São Paulo and Miami, activated in July 2016 to expand the international output of the Brazilian academic network. These new interconnections are part of the AmLight Express and Protect project (Award #1451018), founded by the National Science Foundation (NSF), by the São Paulo Research Foundation (FAPESP) - by means of its academic network arm (ANSP) -, and by the Brazilian National Research and Education Network (RNP). A pair of high-end data transfer nodes (DTNs), manufactured by the Chinese company Huawei with Mellanox 100G NICs and Intel NVMe SSD cards installed at UNESP datacenter in São Paulo, as well as OpenFlow-enabled Huawei switches and WDM appliances at UNESP and ANSP will be used during the exercise. UNESP team will also preview its new OpenFlow controller, Kytos, a modular, event-based open-source OpenFlow controller being developed in Python with a simple asynchronous architecture to facilitate the exchange of messages between its core and users applications. This development is part of a 3-year R&D cooperation agreement between UNESP and Huawei, which also includes experiments on the interoperability of distinct SDN controllers and the integration of SDN and cloud technologies.

Integration of the New Network Paradigm with LHC Data Management Applications at SC16

As mentioned above, the new network paradigm mapping high throughput, load balanced transfers over multiple paths in a complex Terabit/sec network that will be a focus of our SC16 demonstrations, is being integrated with the main data management tools of CMS, namely PhEDEx (V. Lapadatescu, T. Wildish et al, 2014) and ASO (Riahi et al., 2015). PhEDEx is the principal data transfer management tool for CMS. It gathers dataset transfer requests from users and from automated components of the CMS computing and data handling system, schedules the transfers to each destination site, then transfers the data in a reliable and robust manner, and reports the results back to a central database.

ANSE (Advanced Network Services for Experiments) (LHCONE Point-to-Point Service Workshop, December 2012, n.d.) is an NSF funded project, which has aimed to incorporate advanced network-aware tools into the mainstream production workflows of LHC's two largest experiments: ATLAS and CMS. The goal of ANSE has been to improve the overall working efficiency of the experiments, by allowing for more deterministic times to completion for a designated set of data transfers, through the use of end-to-end dynamic virtual circuits with guaranteed bandwidth. For CMS this translates to the integration of bandwidth provisioning capabilities in PhEDEx and ASO, its data transfer management tools. PhEDEx controls the large scale of data flows over wide area networks, on average handling 6 PB of data per week which is spread over 170 sites.

PhEDEx has successfully managed large scale data transfers for CMS for over 10 years now, and continues to be a workhorse of the experiments' operations model. The ANSE project has been working closely with PhEDEx since 2013 to this end, integrating network-awareness and dynamic allocations of OSCARS (C. Guok, 2006) and NSI (INDIS Innovating the Network for Data-Intensive Science, n.d.) circuits into PhEDEx by enabling it to create, use, and destroy virtual network circuits dynamically using ESnet's OSCARS (C. Guok, 2006) and NSI (INDIS Innovating the Network for Data-Intensive Science, n.d.) protocols to improve transfer performance. The first results have already been reported (J. Zurawski, E. Boyd et al., 2011) (V. Lapadatescu, T. Wildish et al, 2014) (J.Balacas, 2016), showing that PhEDEx can now transparently switch to using a network circuit when one is available, including across the LHCONE virtual routing and forwarding (VRF) (LJC Open Network Environment, n.d.) fabric supporting operations involving the LHC Tier2 as well as Tier1 facilities.

The second major application of CMS, AsyncStageOut (ASO) (Riahi et al., 2015), has evolved to a highly adaptable service for managing users' output files from analysis jobs. ASO currently transfers 600k output files daily, which are spread over multiple compute nodes. Through our work in ANSE in 2014-16, we have extended the work with PhEDEx implemented dynamic circuit functionality in PhEDEx and with AsyncStageOut. AsyncStageOut is now capable of communicating with PhEDEx, and making smart use of dynamic circuits, creating one when it's worth doing so: for transfers of a sufficient data volume, duration, and priority. The two systems now communicate with each other and take decisions by checking the current workload and past transfer history (J.Balacas, 2016).

During SC16 the Caltech team together with partners will present a new scheduler system deployed for multidomain end to end transfers, assigning transfers to paths and an adjustable bandwidth allocation for each transfer using a new FDT scheduler, being developed in the context of the DOE SDN NGenIA and SENSE projects. As described above, the new scheduler will combine limits at the edge with OVS, and in the core with ALTO as part of the team's Open Daylight controller, and the edge- and core-limits will be coordinated for a consistent end-to-end outcome. A proof-of-concept prototype of the new methods has already been worked out as of this writing. During SC16 the Caltech team will show how both systems (PhEDEx and ASO) can exploit the new paradigm to achieve better overall throughput in complex multiuser situations, by working with better defined

time-to-completion for each of the transfers in progress, and those planned for the immediate future. From this prototype a refactored and pre-release ready version is planned to be ready by Q2 2017. Ongoing development of the new paradigm includes work on: (1) cost algorithms for mapping flows onto paths, (2) topology presentation of key information from network core services to client applications, (3) development of effective metrics for optimal path assignment, bandwidth allocation profiles, and the time-profile for allocation adjustments, to achieve high throughput stable solutions for the ensemble of transfers among the sites, while making good use of the sites' computing and storage resources.

It is important to note that the new scheduler system and network controls (Q. Xiang, Traffic Optimization for ExaScale Science Applications, 2016) (Q. Xiang, Traffic Optimization for ExaScale Science Applications draft, 2016) are neither CMS- or high energy physics-specific, but they can be used by data intensive applications involving large scale data transfers operations in any field of science.

Deep Recurrent Neural Net Distributed Training for High Energy Physics Event Classification

The purpose of this part of our SC16 demonstrations, carried out by the Caltech team with the support of NVidia, is to illustrate how such systems coupled to the latest deep learning techniques may be used to train deep neural networks on large HEP datasets. Part of the future vision is also to migrate the use of these trained models to real-time filtering of the data online as it is taken, allowing for a richer recorded dataset that extends the reach of the LHC experiments in our searches for new physics processes, within the feasible bounds of computation and storage.

We will use state of the art deep learning neural networks running on a Supermicro server hosting eight NVidia GTX 1080 GPUs and an attached storage array, to demonstrate the rapid training of deep neural nets models on a large set of simulated high energy physics ``events'', which are the result of proton-proton collisions in the LHC as observed in CMS⁴². To accelerate the process, we will work with a compact data form consisting of the energy-momentum four vectors of the particles, jets and missing energy vector (see Appendix A) for each of the events.

A complementary thrust of this demonstration is to illustrate the emerging use of the latest deep learning techniques in data intensive science, using Recurrent Neural Networks (RNN) (Recurrent Neural Networks, n.d.) with Long Short Term Memory (LSTM) cells (Long Short Term Memory cells, n.d.), applied to the high energy physics use case. While the field of HEP has developed its own advanced culture of extracting the presence of new signals from amidst massive numbers of ``background'' events coming from known processes, and has progressed from early forms of feed-forward neural nets (Neural Networks, n.d.) to so-called Boosted Decision Trees (Boosted Decision Trees (in HEP) in Practice, n.d.), the resurgence of deep learning nets such as RNNs holds the promise of further accelerating the time to discovery in our field.

The specific choice of RNNs with LSTM cells over alternative deep learning methods is motivated by the progress in Natural Language Processing (NLP) (Deep Learning for Natural Language Processing, n.d.) brought by these techniques. A key feature of RNNs is their ability to deal with variable input sequences, making them well-adapted to extract and learn the context and meaning of written text. RNNs also have been successful in image characterization (captioning), classification of high level features such as the sentiment expressed by a text, machine translation between languages, and frame-by-frame classification of video sequences. The use of LSTM cells in RNNs is an important source of their recent successes, resulting from a more sophisticated structure interlinking successive modules in the network. This special structure allows the RNN to

⁴² An introduction to LHC physics and the nature of events, in a nutshell, is provided in Appendix A.

handle long-range correlations among the inputs (such as long term memory when dealing with a time-sequence).

We will use several categories of simulated physics processes (Pythia generators, n.d.) (Madgraph, n.d.), where the simulated particles are passed through a simplified representation of the CMS experiment (Delphes, n.d.), resulting in a series of particle 4-vectors giving their momenta and energies. These will be transformed and fed to a deep learning library (Keras deep learning library, n.d.) with the ability to do the necessary mathematical manipulations of arrays efficiently (Theano software, n.d.) (TensorFlow, n.d.). Our models will have various architectures including a novel layer type developed for the purpose of training on particle physics 4-vectors.

We will visualize the evolution of the classification performance of our model as the training goes, using a separate test dataset running in parallel to the training itself on another server. We will visualize the convergence of the model resulting from the training using tools such as the NVIDIA Deep Learning GPU Training System (DIGITS) (NVIDIA Deep Learning GPU Training System, n.d.) or TensorBoard (TensorBoard, n.d.). The events that are placed in each category will be shown using a well-known event display package (Event Display WorkBook, n.d.). in a mosaic of categories, together with histograms of high level features that are known to be discriminate among the various event types.

GPU-Accelerated Indexing and Querying HEP Big Data with Analytics

This part of our SC16 program, complementing the application presented in the previous section and carried out by the Caltech team with the support of Echostreams, NVidia, Orange Labs (Silicon Valley) and SQream, will illustrate how we can reduce the time to discovery by rapidly scanning, indexing, filtering and analyzing large LHC datasets using a compact multi-GPU server system coupled to a unique GPU-accelerated database, and backed by a high throughput data store.

We will use the SQream DB Big Data Analytics SQL database (SQream DB Big Data Analytics SQL database, n.d.) running on a CocoLink Klimax server with eight Titan X GPUs and a 10 Terabyte attached array of 36 SSDs, to demonstrate the rapid traversal and selection of billions of LHC collision events as observed in CMS, showing the rate of data injection and the speed of queries for retrieving filtered data.

The data will be converted from the native ROOT format (ROOT Data Analysis Framework used by high energy physicists, n.d.) into SQL queries using the CSV format (Comma-Separated Values files, n.d.) which are suitable for data injection into the SQream DB. Performance of the queries will be visualized as a function of the data volume and the filters applied. The results of the queries will be visualized through a set of histograms displaying the characteristics of the selected data extracted from the database.

Broad HPC and Science Relevance

Fast and efficient data distribution and access, as required by distributed scientific instrumentation such as the LHC experiments' computing and storage infrastructures, and paid analysis of this data leading to physics discoveries, rely on the smooth interplay of many components. On top of the raw network capacity, the network architecture, switching equipment features and performance, end-system I/O architecture, the transfer applications and data management system software need to be tuned, and to some extent co-designed and co-developed, for frictionless operation. This has become increasingly challenging as the volumes and transfer rates required by the target science programs continue to grow exponentially, with a beyond-Moore's Law slope.

As noted above there are many areas of science that are beginning to generate large amounts of data that need to be shared and accessed seamlessly across multiple institutions, with novel compute

and data intensive applications to extract groundbreaking knowledge from the data. The infrastructure and applications we are demonstrating pave the way towards methodologies and solutions that can serve many different domains of science, and serve as a working roadmap for other larger scale future programs to explore. Certainly research universities and laboratories will want to understand what is possible and what will be required to ensure that they can effectively support their researchers, as they collaborate in data-intensive programs that span many campuses and many countries, crossing the boundaries among science and engineering disciplines.

Conclusions and Path Forward

Starting at SC14 we have shown what is achievable with state-of-the-art components and applications including Terabit/s data movement between nodes both at the exhibition floor as well as to and from several LHC computing sites reachable over 100G WAN infrastructures, coupled to intelligent path construction and flow distributions over multiple paths. We have also advanced the state of the art in the management and optimization of flows with SDN methods at layers 1, 2 and 3 and advanced transfer protocols, integrated with some of the mainstream global data management applications of the LHC experiments.

Our SC15 demonstrations further advanced the state of the art, moving these methods closer to production readiness for the LHC and other use cases, exploiting the latest SDN Open Daylight releases and related methods (such as OVS), and spreading the knowledge to a growing circle of SDN developers in the HEP community. The scale also advanced to the Terabit/sec range through the use of many first-generation 100GE network interface cards with bidirectional flows of 170 Gbps from a single port.

Starting in 2015, a new application focus has been the use of the Leadership and other major HPC facilities now in operation, and those planned by the US DOE and NSF and other agencies, which are projected to reach the 200 petaflop level by 2019 and 1 Exaflop by approximately 2023. Developing the means to use these facilities effectively in concert with the HEP and other laboratories holding exabytes of data, is a pivotal development for the science communities. One focal point of the program described above is the development of an operational model of petabyte transactions involving DTNs at the HPC facility site-edge, to make this possible.

Our SC16 demonstrations will present a major step forward towards a new generation of intelligent networks and applications with a new scale and scope, as well as data intensive operations and applications with exascale computing systems, including: (1) A new “consistent operations” paradigm and programming environment for complex networks interlinking major research facilities and science teams, wherein protocol-agnostic edge-control and core-control services and the science programs’ data management applications cooperate to allocate high bandwidth, load balanced high throughput flows over selected paths, optimized through deep learning methods, (2) A substantially extended OpenDaylight controller using a unified multilevel control plane programming framework to drive the new network paradigm, (3) More advanced integration functions with the data management applications of the CMS experiment, (4) A new Terabit/sec network complex interconnecting eight booths on the show floor with many 100GE local and wide area connections to remote sites on 4 continents, along with the latest generation of DTNs driving 100-1000 Gbps flows across the complex network, (5) Novel deep learning and database architectures and methods for rapid training on, and traversal of LHC data, driving high throughput event classification and characterization by use of multi-GPU systems backed by a high throughput SSD data stores, and (6) A new immersive VR experience including a virtual tour of the CMS experiment at the LHC and an inside-out exploration of LHC collision data in the experiment.

SC15, ANSE, SDN NGenIA, SENSE and SC16 Team Members

An Academic Network at Sao Paulo (ANSP): Jorge Marcos, Luis Lopez

California Institute of Technology (Caltech): Azher Mughal, Dorian Kcira, Harvey Newman, Iosif Legrand, Ramiro Voicu, Maria Spiropulu, Jean-Roch Vlimant, Justas Balcas, Wayne Hendricks

CERN: Edoardo Martelli, David Foster

Colorado State University: Christos Papadopoulos, Susmit Shannigrahi

Echostreams: Andy Lee, Gene Lee

ESnet: Brian Tierney, Eric Pouyoul, Inder Monga, Chin Guok, Greg Bell, Bill Johnston

Imperial College London: David Colling, Duncan Rand, Simon Fayer

Internet2: Eric Boyd

Fermilab: Phil DeMar, Wenji Wu, Panagiotis Spentzouris

Florida International University/AmLight: Julio Ibarra, Heidi Morgan, Jeronimo Bezerra

Michigan State University: Andrew Keen

Northeastern University: Edmund Yeh, Ran Liu

Northwestern University: James Chen, Joe Mambretti

Rede Nacional de Ensino e Pesquisa (RNP): Alex Moura, Gustavo Dias, Leandro Ciuffo, Michael Stanton

SURFnet: Gerben von Malenstein

Systems Networking Lab at Tongji and Yale University: Mingming Chen, Shenshen Chen, Haizhou Du, Kai Gao, Chen Gu, Christopher Leet, Geng Li, Xiao Lin, Charles Proctor, Yichen Qian, May Wang, Tony Wang, Qiao Xiang, and Y. R. Yang

Universidade Estadual Paulista (UNESP): Sergio Novaes, Rogerio Iope, Beraldo Leal, Artur Baruchi, Raphael Cbe, Marcio Costa, Diego Oliveira, Carlos Eduardo Santos

Universidade Federal do Rio de Janeiro: Gustavo Pavani

University of Michigan (UMICH): Robert Ball, Roy Hockett, Shawn Mckee

University of Texas at Arlington (UTA): Kaushik De

Vanderbilt University: Alan Tackett, Andrew Melo, Paul Sheldon

Engagement and Partnerships

The progress of the project pilots mentioned above have been made possible through our strong collaboration and partnership with colleagues and groups both at research institutions and in industry. Some of our most important partners for this project are listed below.

Academic Partners: Pacific Research Platform (PRP) (includes UCSD, Stanford, UCLA, UC Berkeley and many other campuses throughout California), Fermi National Accelerator Laboratory, Brookhaven National Lab, Lawrence Berkeley National Lab, CERN, SPRACE and GridUNESP (Sao Paulo), UERJ (Rio de Janeiro), University of Michigan, Florida International University, Vanderbilt University, University of Florida, KISTI (Korea), Yale University.

Research and Education Network Partners: ESNet, Internet2, CENIC (California), FLR (Florida), AmLight (Miami), SURFNet (Amsterdam), MiLR (Michigan), Academic Network of So Paulo (ANSP), RNP (Brazil), KREONET (Korea).

Industry Partners:

(1) Networks: Arista, Ciena, Century Link, Coriant, Brocade Networks, Dell, Inventec, Spirent, Cisco, Mellanox, QLogic;

(2) Storage Systems: Samsung, HGST, Mangstor

- (3) Server Systems and Integrators: EchoStreams, 2CRSI, Dell, Supermicro
- (4) GPU systems for machine learning: NVidia, Orange Labs (Silicon Valley), SQream
- (5) Virtual Reality: NovaVR LLC.

ACKNOWLEDGMENT

A key factor in the progress and success of the work presented here has been the support and engagement of the DOE Offices of Advanced Scientific Computing and High Energy Physics, and the NSF Directorate for Computer & Information Science and Engineering (CISE). We thank the agencies for the following grants, under which much of this work was carried out:

Caltech Team:

- SDN NGenIA, DOE/ASCR project ID 000219898
- SENSE, FNAL PO #626507 under DOE award # DE-AC02-07CH11359
- ANSE – NSF award # 1246133
- CHOPIN – NSF award # 1341024
- CISCO – Award # 2014-128271
- Tier2 – NSF award # 1120138
- OLIMPS - DOE award # DE-SC0007346 (through 7/14/14)
- US LHCNet - DOE # DE-AC02-07CH11359 (through 5/31/15)

Sao Paulo UNESP Team:

- São Paulo Research Foundation (FAPESP) under Grant # 2013/01907-0
- Huawei do Brasil Telecomunicações Ltda. under Grant Fundunesp # 2481-2015

Yale Team:

- NSF: under award# CC*IIE-1440745, award# CNS-1218457
- Google: under Google Faculty Research Award (2015)
- Huawei: under Huawei Research Award (2014-2015)

We also thank the SCinet network team for their strong support over the many past and upcoming editions of Supercomputing exhibition, culminating in Terabit/sec intelligent SDN infrastructures and new generations of data intensive applications and methods.

Appendix A. LHC Physics and Events in a Nutshell

The potential for discoveries in physics at the Large Hadron Collider derives from its combination of high beam energy and “luminosity” (intensity), where the physicists can find and study rare events with key features that indicate the presence of new processes, beyond those predicted by the well-established Standard Model of particle physics, in the data.

The LHC, which has reached and exceeded its design luminosity this year, now produces up to 1 GHz of high energy collisions of protons at the points where the bunches of protons beams stored in the LHC ring collide, including at the center of the CMS experiment. CMS functions in essence as a high speed electronic camera that takes “snapshots” of the results of each collision, called “events”. Each event consists of the digitized data representing the coordinates in space and/or the energy deposited in the various layers of the particle detection instruments that make up the

experiment, each of which is designed to detect, measure, and help identify the particles of various types.

The events are filtered so that only the potentially most interesting subsets are kept for further analysis, to limit the computing time required. The digitized data, whether from actual events in the LHC or simulated events used to test a wide variety of new physics models, is then “reconstructed”, and our demonstrations it is further reduced to a compact form that includes the energy, momentum and a set of flags for each of the jets, electrons, photons, muons and missing energy, in the event.

The types of particles produced are predominantly “hadrons” which are particles that interact strongly, such as pions, kaons, and protons, as well as electrons, photons and muons, each of which has its own characteristic pattern as seen in the layers of the experiment. The overall characteristic pattern in high energy events usually includes “jets” of hadrons resulting from the quarks and gluons produced in the collision, and some “missing energy” that is the result of neutrinos or other weakly interacting particles that escape and carry away energy without being detected in the experiment. An event with a large missing energy transverse to the beam line, which is relatively rare, is one of the principal signs of new physics, since heavy weakly interacting particles would lead to this signature, and many new physics scenarios predict such particles.

Annex 28: AsyncStageOut - New component of the distributed data analysis system of CMS

By Justas Balcas (justas.balcas@cern.ch)

January 2016

Introduction

AsyncStageOut (ASO) is a new component of the distributed data analysis system of CMS, CRAB3 [1], designed for managing users' data. It addresses a major weakness of the previous model, namely that mass storage of output data was part of the job execution resulting in inefficient use of job slots and an unacceptable failure rate at the end of the jobs. ASO used FTS3 [2] to schedule and execute the transfers between the storage elements of the source and destination sites. It has evolved from a limited prototype to a highly adaptable service, which manages and monitors the user file placement and bookkeeping.

A prototype (Figure 1 and Figure 2) was created end of October in 2015, which we used during Super Computing 15 [3]. ASO has only one agent which manages all transfers between nodes and new component was implemented. New component features:

- * Queue mechanism for new component which uses third part copy (FDTCP);
- * Checks site local configuration and if FDT is available;
- * Groups multiple files into one transfer queue file;
- * If circuit is available, use it for transfers.

Current prototype showed missing features, weakness of ASO and new component implementation:

- * ASO is grouping small files together with big;
- * There is no information in ASO implemented about transfer queue, size, throughput of new, idle, ongoing, finished transfers and it is hard to make decision whether it should go on a circuit or not.
- * So far circuits IPs must be defined in site local configuration.

We are currently working on production ready version, which will have all missing features implemented and circuit request will be done together with PhEDEx implementation [4].

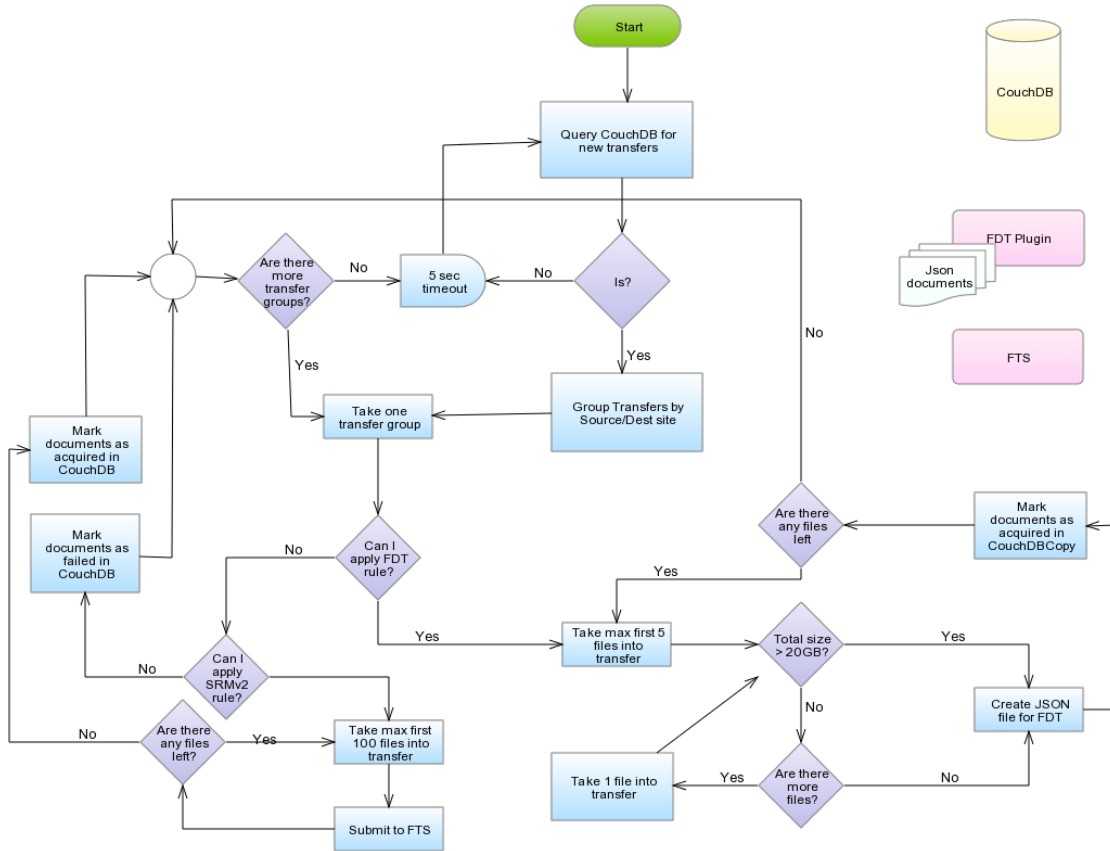


Figure 93: ASO work flow with new component

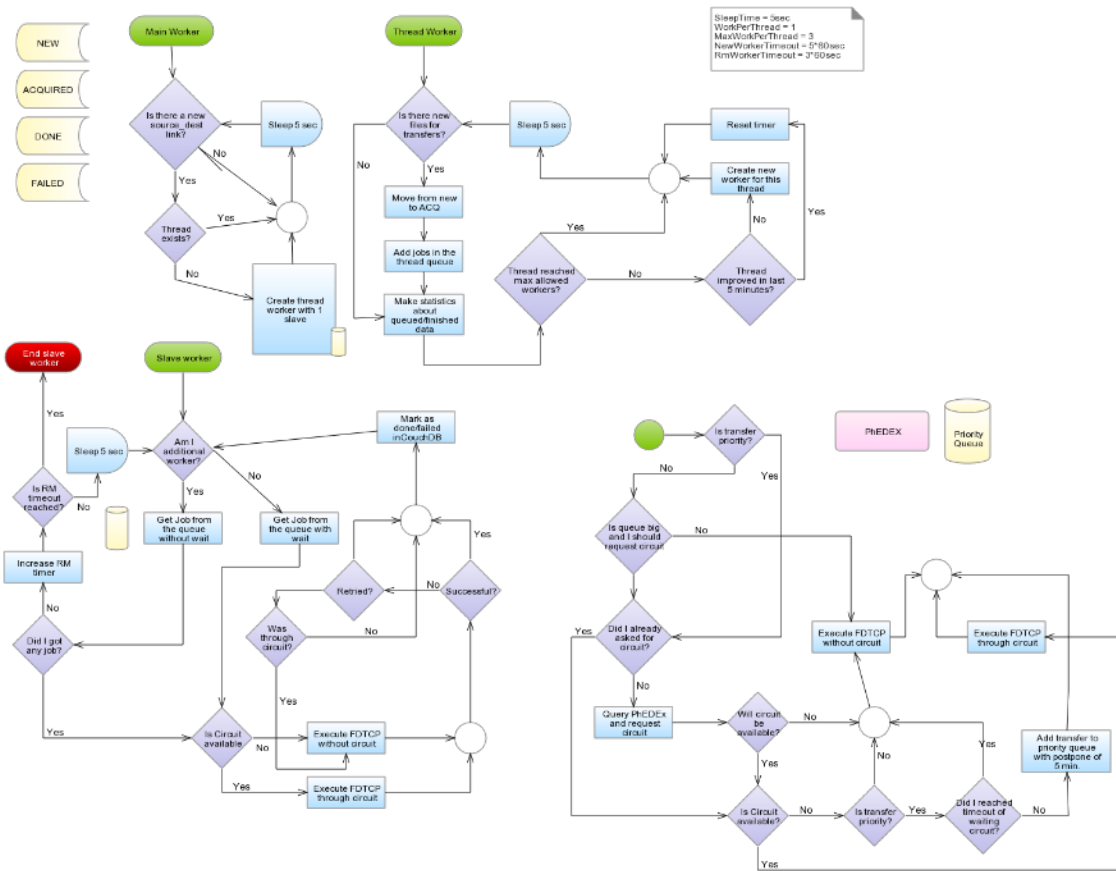


Figure 94: ASO FDT component flow view

References

- [1] Mascheroni M et al. CMS Distributed Data Analysis with CRAB3. Proceedings of CHEP' 15 to be published by IOP J. Phys. Conf. Ser.
- [2] Ayllon A A et al. 2014 FTS3: New Data Movement Service for WLCG. J. Phys.: Conf. Ser. 513 032081
- [3] Super Computing 15 <http://sc15.supercomputing.org/>
- [4] Wildish T et al. Virtual Circuits in PhEDEx, an update from the ANSE project. Proceedings of CHEP' 15 to be published by IOP J. Phys. Conf. Ser.

Annex 29: OSIRIS Open Storage Research Infrastructure

By Shawn Mckee (smckee@umich.edu)

January 2017

Introduction

OSiRIS is a collaboration between University of Michigan, Wayne State University, Michigan State University, and Indiana University that started in 2016. Our goal is to provide transparent, high-performance access to the same storage infrastructure from well-connected locations on any of our campuses. We intend to enable this via a combination of network discovery, monitoring and management tools and through the creative use of Ceph features.

Project member sites are linked at a minimum of 10Gb/s and up to 80Gb/s between U-M and MSU. The project deployment includes a multi-site Ceph cluster with current capacity of 5.2 PB(raw), perfSonar-Periscope network monitoring nodes at each site and at client sites, and AAA services to bridge Ceph and federated authentication networks such as InCommon and eduGain.

Future technical augmentations include the use of Software Defined Networking to orchestrate network flows, both for the Ceph storage inter-site communication and to optimize simultaneous science domain use of the infrastructure. Additionally the project intends to incorporate data lifecycle management automation, relying upon the software defined storage aspects of Ceph to capture relevant meta-data as appropriate for each science domain.

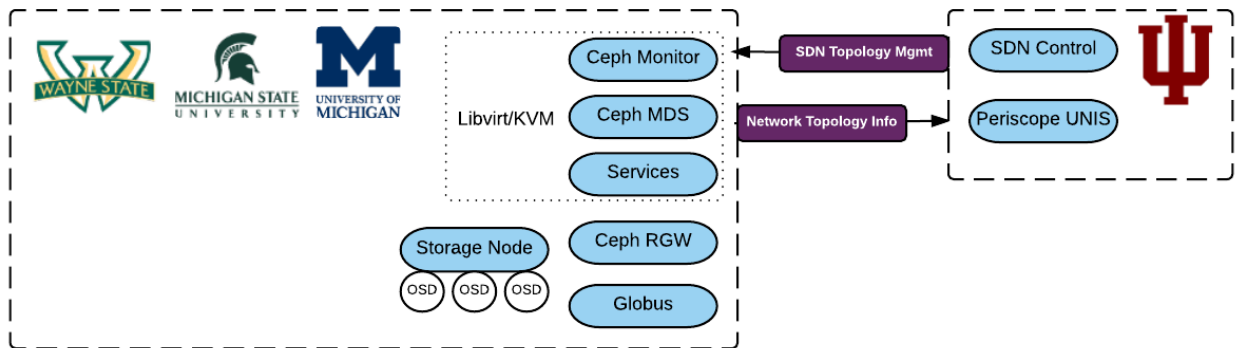


Figure 95: Overview of OSiRIS site structure

Our AAA services are built on a bearer token architecture known as OSiRIS Access Assertions. The system is still in development and we have an ongoing engagement with the CTSC to vet our design. The goal is to enable researchers to access OSiRIS with their existing institutional credentials through a variety of federations.

The project will provide storage services to multiple science domains. How they access our storage will vary by project needs - Ceph has options to mount a POSIX fs, provide an S3 compatible object gateway, provide kernel block devices, or be used directly as an object store.

Our projected roadmap is:

- Year 1: High-energy Physics (ATLAS), High-resolution Ocean Modeling (both now ongoing)
- Year 2: Biosocial Methods and Population Studies, Aquatic Bio-Geochemistry
- Year 3: Neurodegenerative Disease Studies

- Year 4: Statistical Genetics, Genomics and Bioinformatics
- Year 5: Remaining participants, New Science Domains

We are currently working on supporting ATLAS via the S3 gateway and are in process of load-testing our S3 gateways from ATLAS compute farms at ANL and BNL. We also plan to provide ATLAS support via Ceph GridFTP plugins.

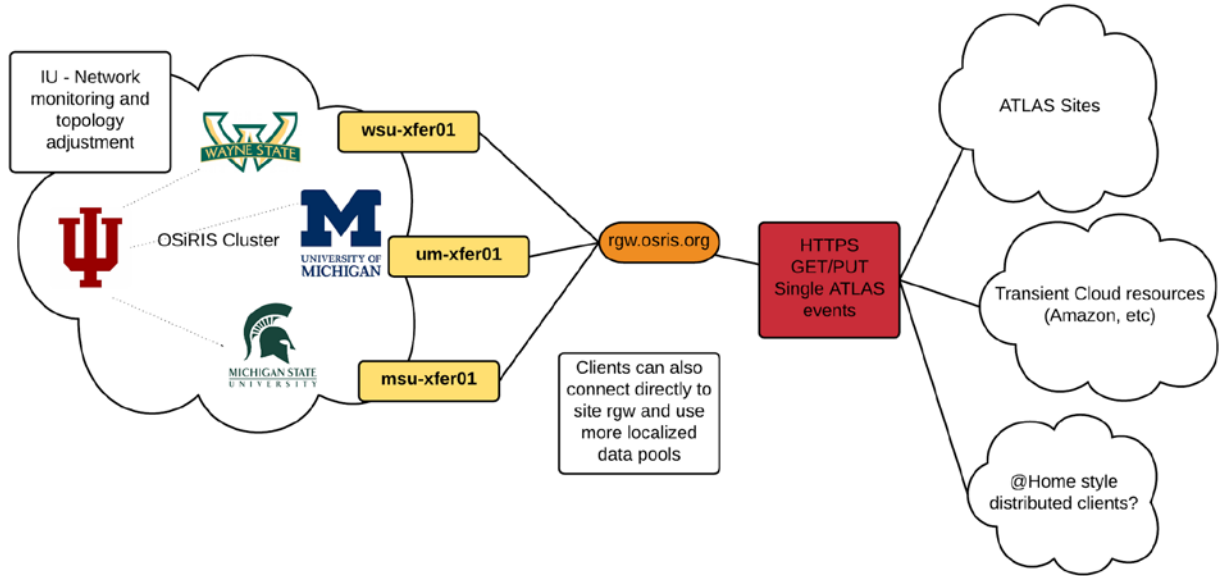


Figure 96: ATLAS usage of OSiRIS S3

For high-energy physics and ATLAS in particular, the advantage of using a public resource like OSiRIS is enabling a wider range of computing resources to process the data and removing the storage management burden. One of our project goals is to provide stable and ready-made storage infrastructure for big-data projects like ATLAS. We hope that a resource like OSiRIS can enable researchers to worry less about creating dedicated and potentially expensive storage infrastructures and instead leverage our resource to store and collaborate on their data.

Annex 30: MonALISA Framework

Submitted by:

*Ramiro Voicu (ramiro.voicu@cern.ch), Iosif Legrand (Iosif.Legrand@cern.ch), Caltech
February 2015*

Introduction

MonALISA (Monitoring Agents in A Large Integrated Services Architecture) (<http://monalisa.caltech.edu>) is a globally scalable framework of services developed by Caltech to monitor and help manage and optimize the operational performance of grids, networks and running applications in real-time. MonALISA is currently used in several large scale HEP communities and grid systems including CMS, ALICE, ATLAS, and the Russian LCG sites. It was actively used to monitor the entire US LHCNet infrastructure and the network services until the end of 2014. MonALISA also is used to monitor, control and administer all of the Seevogh Research Network⁴³ video conference streaming reflectors, and to help manage and optimize their interconnections.

As of this writing, more than 370 MonALISA services are running throughout the world. These services monitor more than 55,000 compute servers, ~2,000 xrootd servers, and tens of thousands concurrent jobs. More than 5 million persistent parameters are currently monitored in near-real time with an aggregate update rate of approximately 30,000 parameters per second. In addition it collects more than 100 million volatile parameters per day which are used to generate aggregate values or only to check if different services or jobs are working correctly. It is also used to monitor and perform end to end performance measurements on more than 18,000 network connections.

The monitoring information collected is used in a variety of higher-level services that provide optimized grid job-scheduling services, dynamically optimized connectivity among the SeeVogh reflectors, and the best available end-to-end network path for large file transfers. Global MonALISA repositories are used by many communities to aggregate information from many sites, to properly organize them for the users and to keep long term histories. During the last year, the repository system served more than 50 million user-requests.

MonALISA System Design

The MonALISA system is designed as an ensemble of autonomous self-describing agent-based subsystems which are registered as dynamic services. These services are able to collaborate and cooperate in performing a wide range of distributed information-gathering and processing tasks. An agent-based architecture of this kind is well-adapted to the operation and management of large scale grids, by providing global optimization services capable of orchestrating computing, storage and network resources to support complex workflows. By monitoring the state of the grid-sites and their network connections end-to-end in real time, the MonALISA services are able to rapidly detect, help diagnose and in many cases mitigate problem conditions, thereby increasing the overall reliability and manageability of the grid.

The MonALISA architecture, presented in Figure 105, is based on four layers of global services. The network of Lookup Discovery Services (LUS) provides dynamic registration and discovery for all other services and agents. Each MonALISA service executes many monitoring tasks in parallel through the use of a multithreaded execution engine, and uses a variety of loosely coupled agents to analyze the collected information in real time.

⁴³ See <http://research.seevogh.com/> and <http://www.ezuze.com/>

The secure layer of Proxy services, shown in the figure, provides an intelligent multiplexing of the information requested by clients or other services. It can also be used as an Access Control Enforcement layer.

As has been demonstrated in round-the-clock operation over the last six years, the system integrates easily with a wide variety of existing monitoring tools and procedures, and is able to provide this information in a customized, self-describing way to any other set of services or clients.

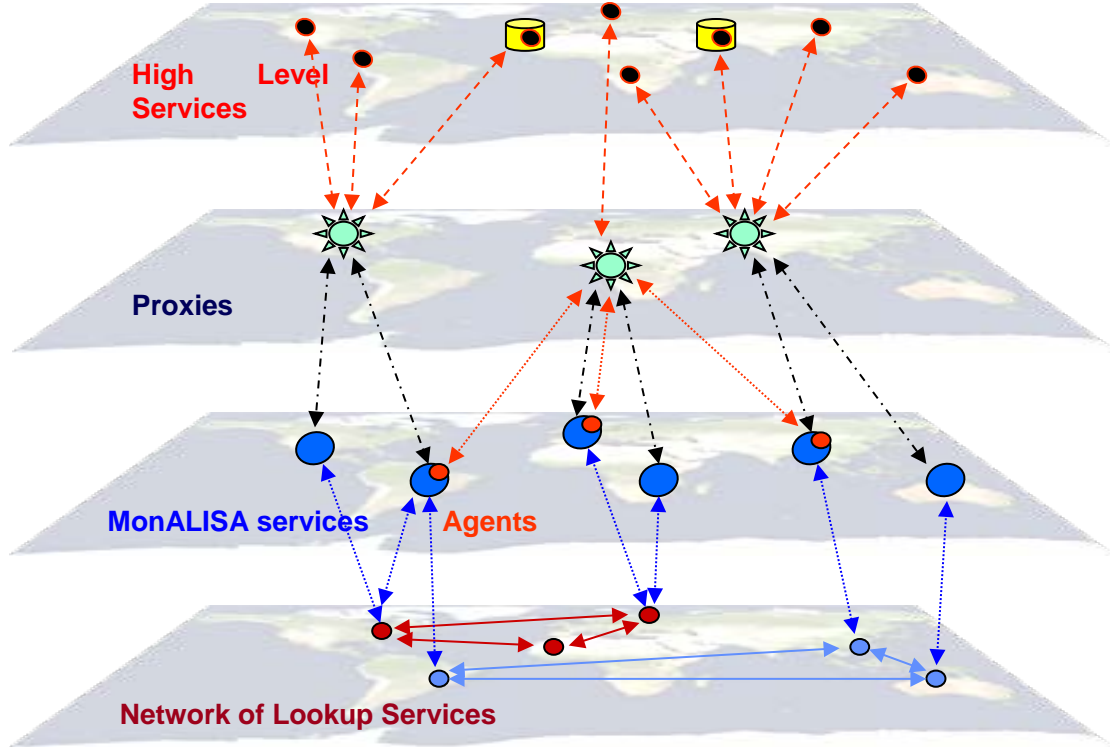


Figure 97: The four layers, main services and components of the MonALISA framework.

MonALISA Deployment in Grids

The MonALISA services currently deployed are used by the HEP community to monitor computing resources, running jobs and applications, different Grid services and network traffic.

MonALISA and its APIs are currently used by a wide range of grid applications in the High Energy Physics community:

For CMS MonALISA is used to collect and transport all the monitoring data from the running jobs to the central CMS dashboard. This is based on ApMon, a MonALISA API that allows monitoring any type of customized information for physics jobs and complete monitoring of the computer the job is executed. The system is used by all the job submission tools for analysis jobs (CRAB), production jobs (ProdAgent) and the Tier0 submission application for the main production activities at CERN. The MonALISA system monitors detailed information on how the jobs are submitted to different systems, the resources consumed, and how the execution is progressing in real-time. It also records errors or component failures during this entire process.

The MonALISA services have been used in production for CMS since 2009, without any problems. MonALISA collected more than 200 billion parameters since Spring 2009, with sustained rates of more than 1,000 parameters per second and peaks of 8,000 parameters per second, without stressing the system.

In ALICE MonALISA is used to provide complete monitoring for their entire offline system, which is based on the “ALIEN” software. Here MonALISA is used to monitor jobs, computing facilities, all the storage systems, experiment-specific services and all the data transfers. ALIEN's application plugins make use of MonALISA monitoring sensors to report application-specific data during execution. This information can be used by the submitter to follow the application's progress and the computer resources acquired, or it can be used by the framework itself to optimize the (re)scheduling decisions.

MonALISA also provides accounting of the resources used. Analysis elements, such as the xrootd servers and clients are instrumented with MonALISA APIs, and this near real-time information is used for load balancing during parallel interactive analysis. ALICE extensively uses MonALISA's ability to react to alarm conditions and rapidly take appropriate action, specifically to restart services which do not work correctly, and to control the overall submission of production jobs. The monitoring information is used as an automatic feedback from different user communities, and it can be used by users or system administrators to understand how the system is functioning, and to detect problems. MonALISA is used in production for ALICE offline computing since 2007. Since then, it has collected more than 250 billion persistent monitoring parameters from all the ALICE sites and stored them into the central MonALISA repository for ALICE computing. The total number of monitoring values collected for ALICE from all the MonALISA services, including the volatile data, is ~ 25 times larger, namely more than 6 trillion parameters.

In Figure 106 Global view of monitoring ALICE grid sites and the connectivity between them. Figure 106 we present a global view of monitoring all the ALICE grid sites and the quality of connectivity between them.

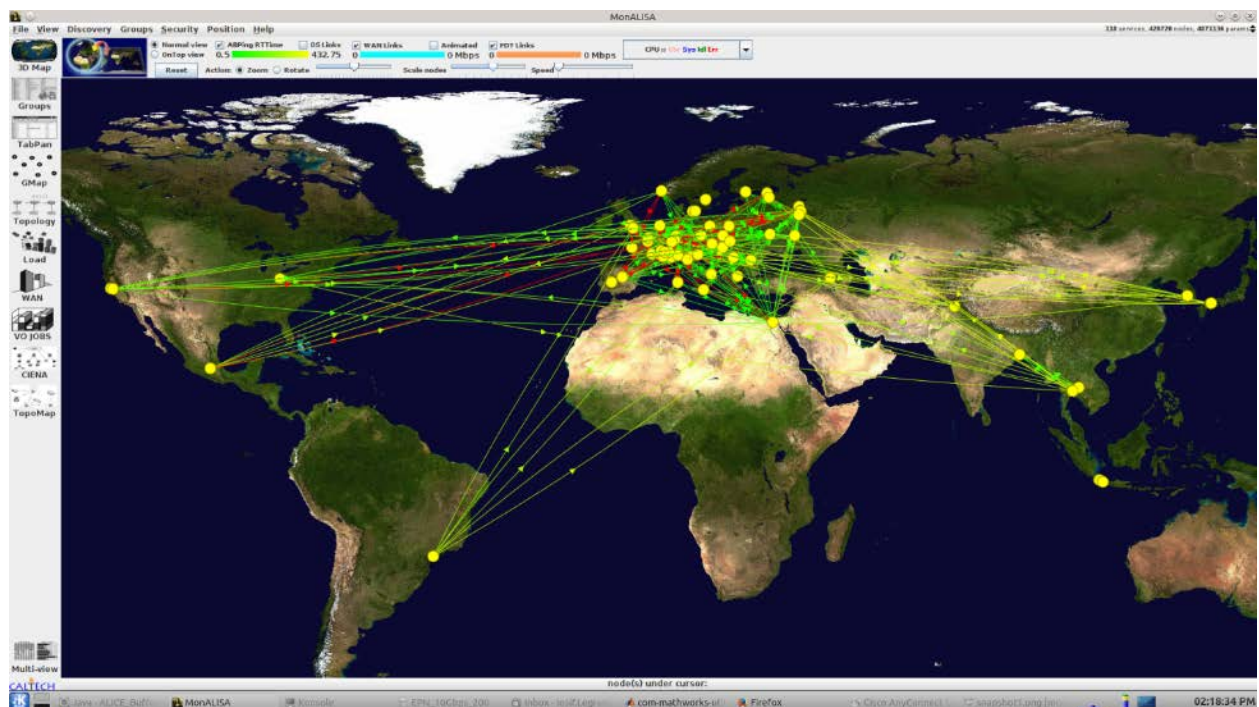


Figure 98 Global view of monitoring ALICE grid sites and the connectivity between them.

CMS and ATLAS are using the MonALISA system to monitor all the xrootd servers, and each experiment developed its own set of repositories to keep and present the monitoring information. For network monitoring the MonALISA system allows one to collect, display and analyze a complete set of measurements, and to correlate these measurements from different sites in order to

present global pictures of the network state and performance, including: WAN topology, delay in each segment, and an accurate measure of the available bandwidth between any two sites. As described in the main body of this document, in the previous Annex and the following section, these particular functions used extensively in US LHCNet operations, including its dynamic circuit services.

MonALISA Network Monitoring and Management

In order to build a coherent set of network management services, it is very important to collect in information about the network traffic volume and its quality in near real-time, and analyze the major flows and the topology of connectivity. MonALISA's high level of performance and its MTBF measured in years allows US LHCNet to respond reliably and rapidly to any link outage, and automatically reconfigure the network as needed according to a pre-defined set of policies. As shown in years of transatlantic network field-operations, this significantly reduces the effective downtime resulting from an outage, boosts the overall operational efficiency, and reduces US LHCNet's manpower requirements and cost of operations relative to non-automated systems.

Access to both real-time and historical data, as provided by MonALISA, also is important for developing services able to predict the usage pattern, to aid in efficiently allocating resources “globally” across a set of network links.

A large set of MonALISA monitoring modules has been developed to collect specific network information or to interface it with existing monitoring tools, including:

- SNMP modules for passive traffic measurements
- Active network measurements using simple ping-like measurements
- Tracepath-like measurements to generate the global topology of a wide area network
- Interfaces with the well-known monitoring tools MRTG, RRD, IPBM, PIPEs
- Data Transfer Applications such as GridFTP, xrootd, FDT
- Modules to collect dynamic NetFlow / Sflow information
- Available Bandwidth measurements using tools like pathload
- Dedicated modules for TL1 interfaces with CIENA's CD/CIs, optical switches (GlimmerGlass and Calient)

These modules have been field-proven to function with a very high level of reliability over the last seven years.

The way in which MonALISA is able to construct the overall topology of a complex wide area network, based on the delay on each network segment determined by tracepath-like measurements from each site to all other sites, is illustrated in Figure 107. The combined information from all the sites allows one to detect asymmetric routing, route instability or links with performance problems. For global applications, such as distributing large data files to many grid sites, this information is used to define the set of optimized replication paths.

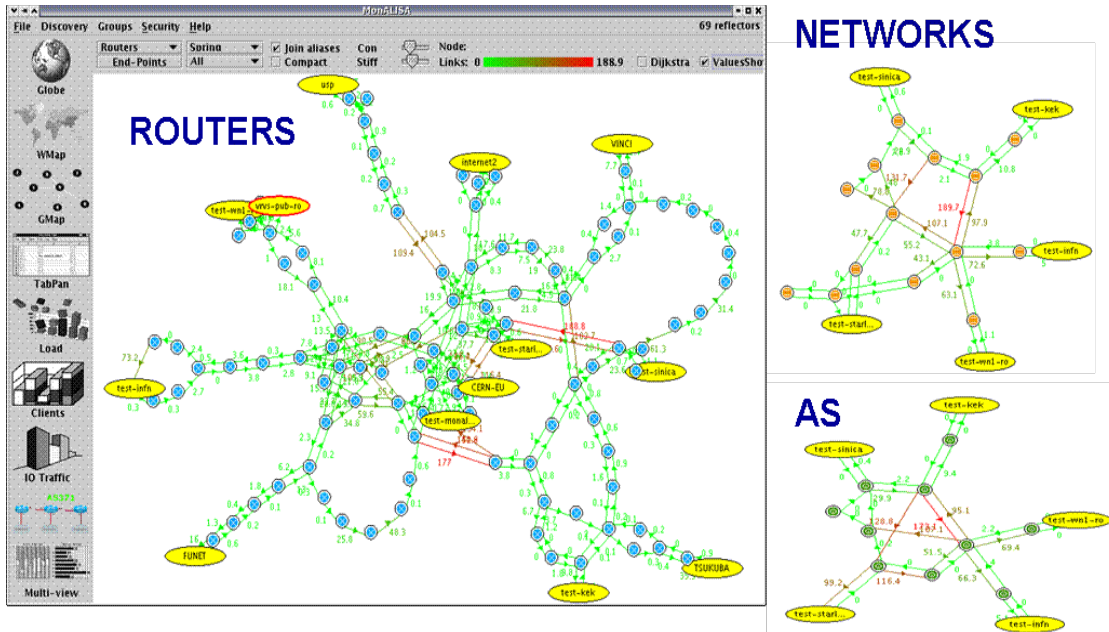


Figure 99: MonALISA real time view of the topology of WANs used by HEP. A view of all the routers, or just the network or “autonomous system” identifiers can be shown.

Specialized TL1 modules are used to monitor the power on Optical Switches and to present the topology. The MonALISA framework allows one to securely configure many such devices from a single GUI, to see the state of each link in real time, and to have historical plots for the state and activity on each link. It is also easy to manually create a path using the GUI. In Figure 108 we show the MonALISA GUI that is used to monitor the topology on Layer 0/1 connections and the state and optical power of the links.

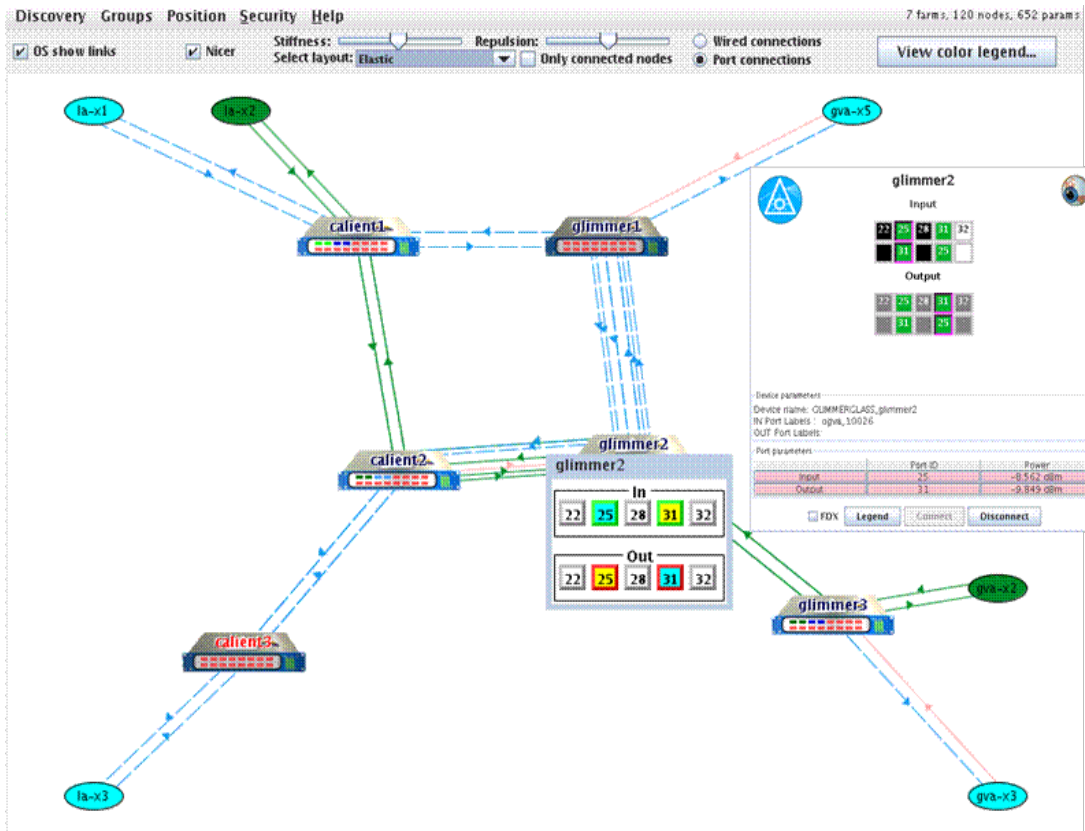
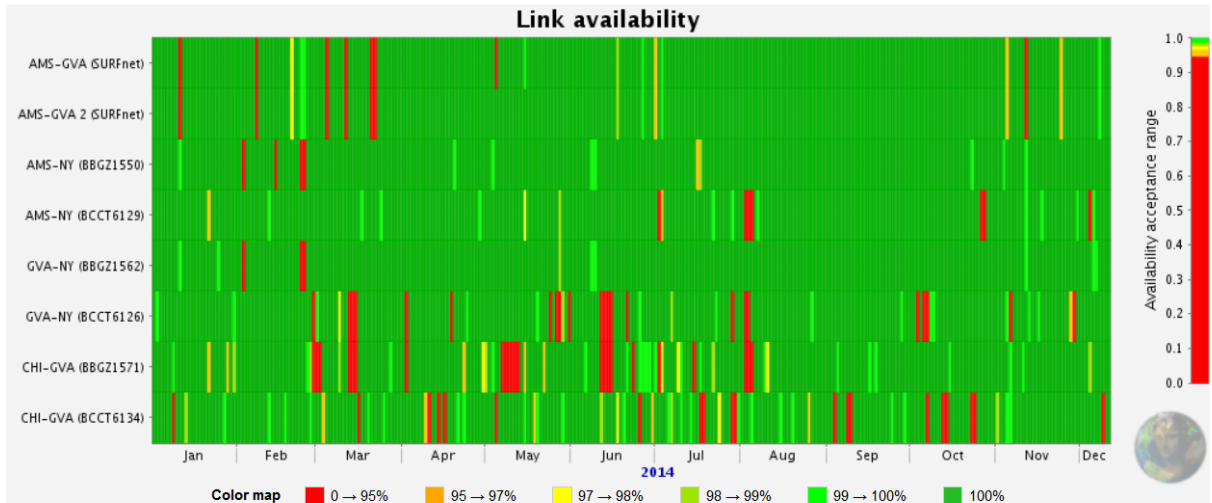


Figure 100: Monitoring and autonomous control for optical switches and optical links.

Monitoring US LHCNet

MonALISA was used to provide reliable, 24 X 7 X 365 real-time monitoring of the US LHCNet infrastructure until the end of the project in December 2014. At each point of presence (GVA, AMS, CHI, NYC) we used a MonALISA service to monitor the links, the network equipment and the peering with other networks. Each major link is monitored at both ends from two independent MonALISA services (the local one and the one at the PoP at the other end of the link). MonALISA services keep the history of all the measurements locally, while global aggregation values are kept in a MonALISA repository to provide a long term history. Dedicated TL1 modules for the Ciena CoreDirector were developed to collect specific information on these devices' topology, dynamic circuits and operational status.

Link Status. MonALISA monitors the status of all US LHCNet WAN links and peering connections. For the Force10 switches we use SNMP and for the Ciena CoreDirectors we use the TL1 interface. The repository analyzes the status information from all the distributed measurements, for each segment, to generate reliable status information. Measurements are done every ~30s and the full history is kept in the repository database. The system allows one to transparently change the way a WAN is operated (via Ciena CoreDirectors and/or the Force10 switch-routers) and keeps a consistent history. Figure 109 shows the panel that allows one to analyze the links' availability for any time interval.



| Statistics | | | | | | |
|---------------------|-------------------|-------------------|-----------------|---------|-----------------|--|
| Link name | Data | | Monitoring | | Link | |
| | Starts | Ends | Availability(%) | Gaps | Availability(%) | |
| AMS-GVA (SURFnet) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 27s | 99.66% | |
| AMS-GVA 2 (SURFnet) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 27s | 99.70% | |
| AMS-NY (BBGZ1550) | 01 Jan 2014 06:20 | 12 Dec 2014 06:19 | 99.99% | 38m 13s | 99.59% | |
| AMS-NY (BCCT6129) | 01 Jan 2014 06:20 | 12 Dec 2014 06:19 | 99.99% | 38m 13s | 99.09% | |
| GVA-NY (BBGZ1562) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 24s | 99.64% | |
| GVA-NY (BCCT6126) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 24s | 97.11% | |
| CHI-GVA (BBGZ1571) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 27s | 96.28% | |
| CHI-GVA (BCCT6134) | 01 Jan 2014 06:19 | 12 Dec 2014 06:19 | 99.99% | 29m 27s | 98.55% | |

Figure 101: Monitoring the status of US LHCNet's major links for all of the past year. Highly granular information also is available via this interface.

Traffic Monitoring. We monitor the total traffic on all the Force10 ports and on the Ethernet ports on the CIENA CoreDirectors. The traffic on Ciena virtual circuits is also monitored by dedicated modules in MonALISA. Different aggregated views are presented: total Tier0 – Tier1 traffic during the last year (shown in Figure 110 and Figure 111) total traffic on all the circuits inside US LHCNet, as well as integrated traffic over any time interval (Figure 112 and Figure 113).

Looking more closely at Figure 112, we observe large number of short bursts close to 9 Gbps, when a transfer reaches the maximum capacity of an OC-192 link, and many other peaks of 5-7 Gbps. While most monitoring systems do not register such peaks, usually reporting longer term averages, the MonALISA monitoring service clearly shows that such peaks, using all or most of the link capacity for minutes to hours, do occur.

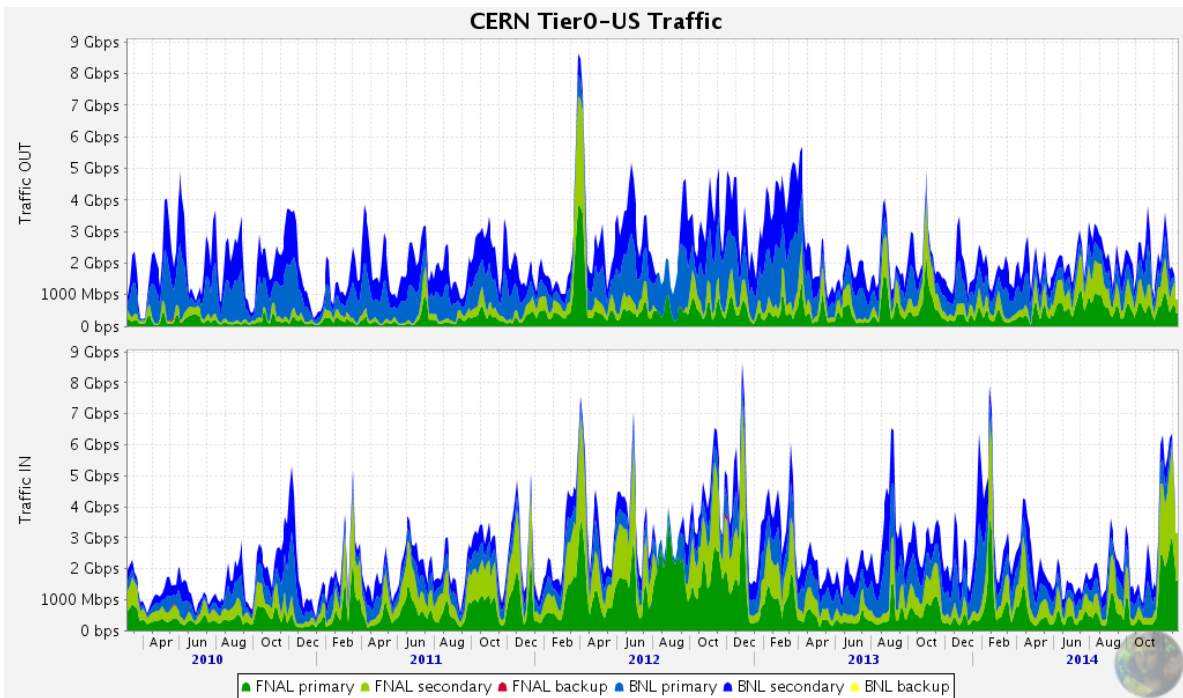


Figure 102: Traffic history for the Tier0-Tier1 circuits in US LHCNet.

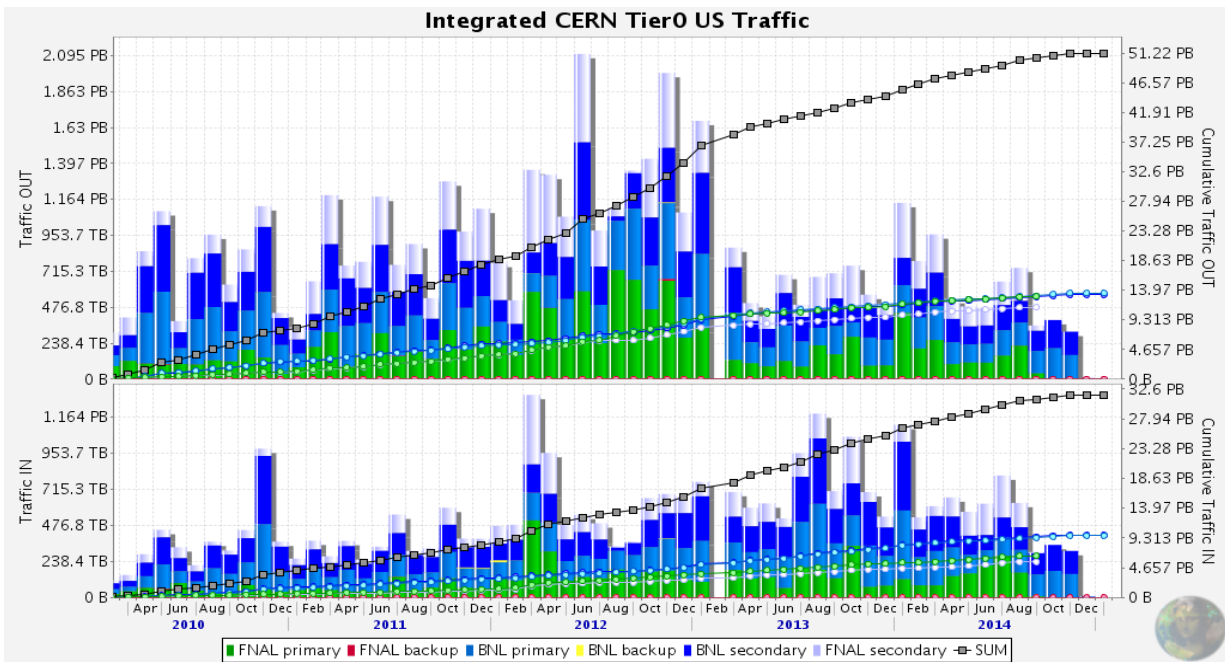


Figure 103: Integrated traffic for Tier0-Tier1 circuits.

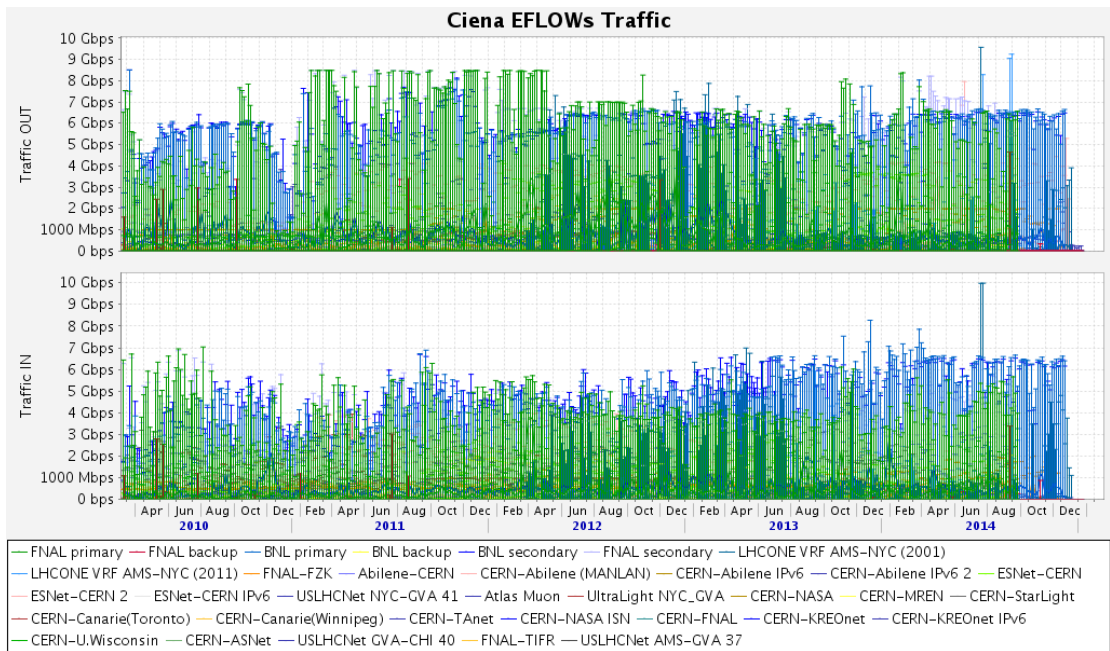


Figure 104 Traffic history for all the circuits in US LHCNet. A large number of 9 Gbps peaks, and a greater number of peaks of 5 to 7 Gbps are observed

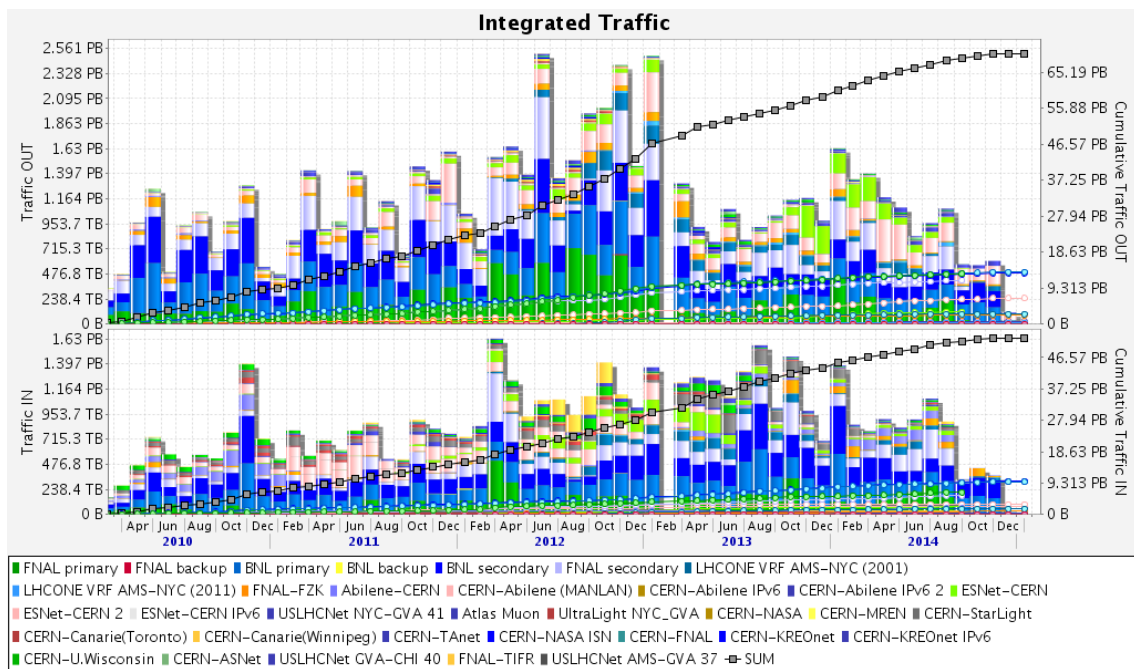


Figure 105: Integrated traffic on all circuits in US LHCNet.

Alarms and Notification. The operational status for the Force10 ports and all the Ciena CD/CI alarms are recorded by the MonALISA services. The alarms are analyzed and SMS/email notifications are generated based on different error conditions. We also “monitor” the services used to collect monitoring information. A global repository for all these alarms is available on the

MonALISA servers, which allows one to select and sort the alarms based on different conditions. Figure 114 presents the panel which allows one to analyze the alarms from the entire system.

| CIENA Alarms for USLHCNet | | | | |
|---------------------------|------------------|----------------|--|---------|
| Date (GMT) | Site | Node IP | Alarm | Remarks |
| last 2 months | CHI | | OCN | Filter |
| 10.01.2014 01:01 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,RFI-L,NSA,2014-01-10,01-01-34,,:\^Line RFI\^,^ | |
| 10.01.2014 01:01 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,RFI-L,NSA,2014-01-10,01-01-34,,:\^Line RFI\^,^ | |
| 10.01.2014 01:01 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,RFI-L,NSA,2014-01-10,01-01-34,,:\^Line RFI\^,^ | |
| 09.01.2014 23:15 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2014-01-09,23-16-00,,:\^Line AIS\^,^ | |
| 09.01.2014 04:24 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:MN,RFI-L,NSA,2014-01-09,04-25-21,,:\^Line RFI\^,^ | |
| 19.12.2013 02:06 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,BERSD,SA,2013-12-19,02-07-27,,:\^Signal degrade against 10E-9... | |
| 18.12.2013 09:00 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,BERSD,NSA,2013-12-18,09-00-40,,:\^Signal degrade against 10E-9... | |
| 18.12.2013 05:10 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,LOF,NSA,2013-12-18,05-10-36,,:\^Loss of frame\^,^ | |
| 16.12.2013 08:34 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2013-12-16,08-34-46,,:\^Line AIS\^,^ | |
| 13.12.2013 04:56 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,BERSD,SA,2013-12-13,04-57-04,,:\^Signal degrade against 10E-9... | |
| 13.12.2013 03:51 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,LOF,NSA,2013-12-13,03-51-22,,:\^Loss of frame\^,^ | |
| 11.12.2013 02:36 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,BERSD,SA,2013-12-11,02-36-34,,:\^Signal degrade against 10E-9... | |
| 11.12.2013 00:58 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,BERSD,SA,2013-12-11,00-59-22,,:\^Signal degrade against 10E-9... | |
| 11.12.2013 00:54 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2013-12-11,00-54-45,,:\^Line AIS\^,^ | |
| 06.12.2013 08:42 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2013-12-06,08-42-58,,:\^Line AIS\^,^ | |
| 05.12.2013 07:42 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:MN,LOS,NSA,2013-12-05,07-43-27,,:\^Loss of signal\^,^ | |
| 05.12.2013 06:32 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:MN,LOS,NSA,2013-12-05,06-32-51,,:\^Loss of signal\^,^ | |
| 05.12.2013 05:14 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:MN,LOS,NSA,2013-12-05,05-14-20,,:\^Loss of signal\^,^ | |
| 05.12.2013 05:14 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,LOF,NSA,2013-12-05,05-14-40,,:\^Loss of frame\^,^ | |
| 04.12.2013 09:28 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2013-12-04,09-28-30,,:\^Line AIS\^,^ | |
| 22.11.2013 16:19 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,AIS-L,NSA,2013-11-22,16-19-41,,:\^Line AIS\^,^ | |
| 21.11.2013 20:29 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:CR,LOS,SA,2013-11-21,20-30-18,,:\^Loss of signal\^,^ | |
| 21.11.2013 20:27 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-2-2,OCN:MN,LOS,NSA,2013-11-21,20-27-49,,:\^Loss of signal\^,^ | |
| 15.11.2013 05:08 | CHI_USLHCNET_CDS | 192.65.196.172 | \^1-A-1-1,OCN:MN,RFI-L,NSA,2013-11-15,05-08-46,,:\^Line RFI\^,^ | |
| TOTAL : 24 alarms | | | | |

Figure 106: Global repository for the Ciena CoreDirector alarms. MonALISA provides a user friendly interface to sort and analyze these alarms as needed.

Network Topology. For the Core Director nodes, MonALISA provides real-time information for the OSRP connections with all the attributes for the SONET links (illustrated in Figure 115).

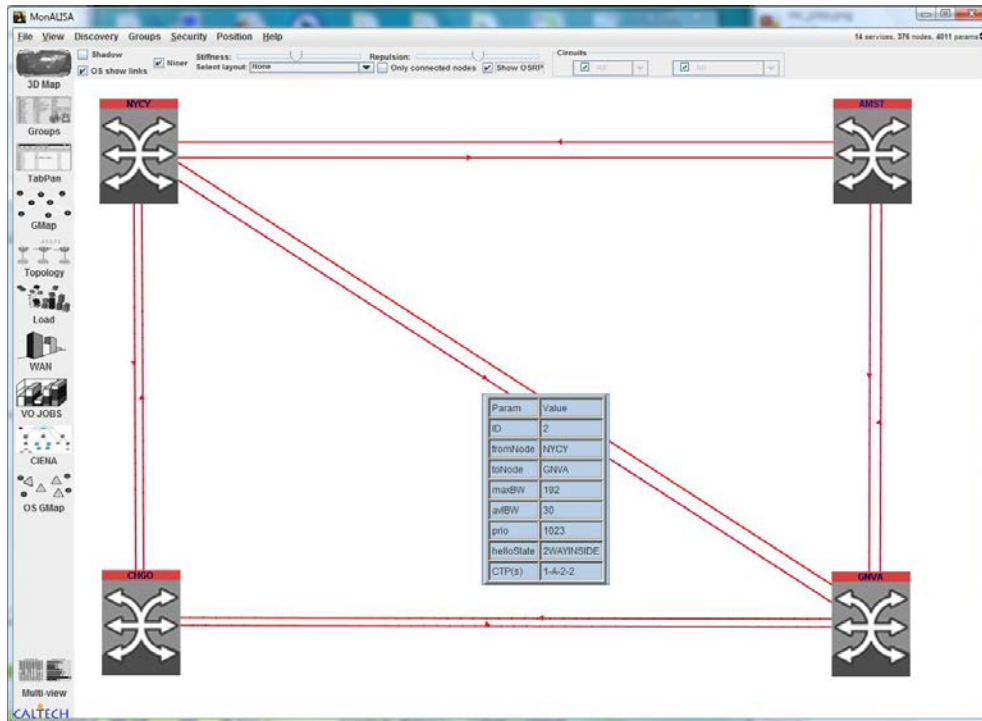


Figure 107: The Physical topology for the Ciena SONET links

Topology of dynamic circuits. The topology of all the circuits created in the entire network is presented in real time in the MonALISA interactive client. This panel allows one to select any set of circuits, and it presents how they are mapped onto the physical network, with all their attributes (as shown in Figure 116).

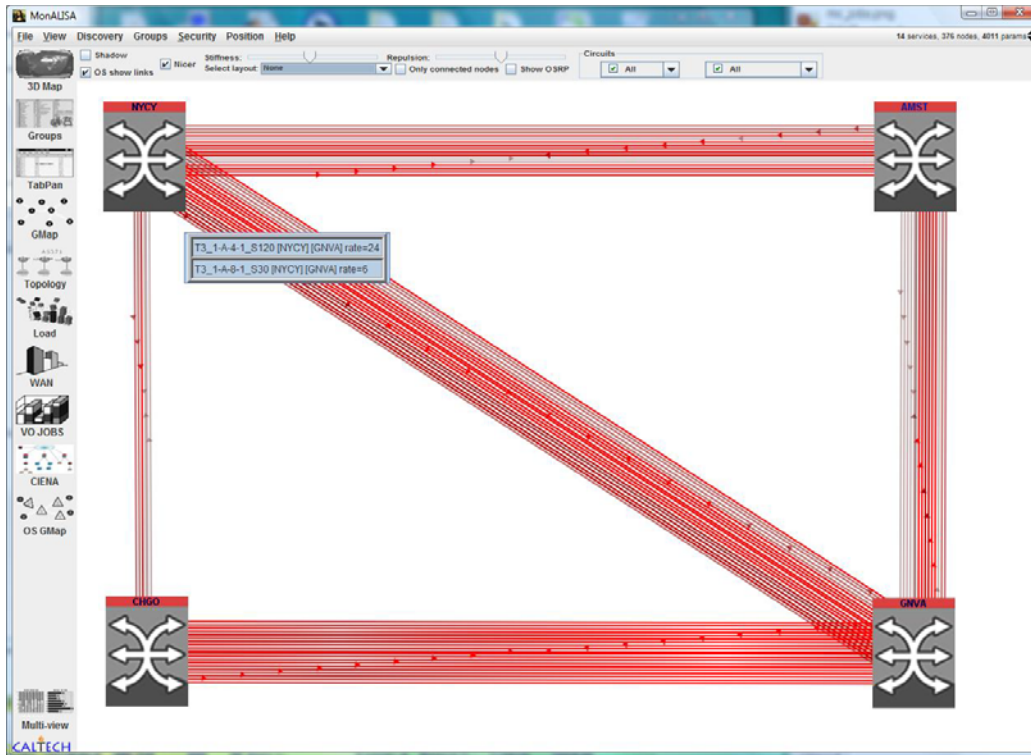


Figure 108: The topology of the dynamic circuits in the entire network and their attributes, presented in compact graphical form.

Development of an automated network management system

The MonALISA team has started the development of an automated network management system using the Telescent Light switch optical patch panel⁴⁴, under a series of DOE/OASCR SBIR grants to Telescent, Inc. In the first phase of this work a prototype Telescent optical switch was deployed at Caltech and then at CERN for a two weeks period and was integrated into a test environment for the US LHCNet infrastructure (Figure 117).

⁴⁴ www.telescent.com. The US LHCNet and MonALISA teams also have years of experience with other such "Layer 0" optical patch panels, such the smaller MEMs-based Glimmerglass and Calient switches.



Figure 109: The Telescent optical switch was integrated into a test environment in the US LHCNet infrastructure at CERN.

A set of MonALISA dedicated prototype modules were developed to monitor and control the Telescent optical patch panel. These modules are based on a set of java based APIs that are used to communicate with the switch firmware. The monitoring modules are currently used to get the connectivity matrix for the switch. The control modules can send reconfiguration commands to the switch.

For the development and testing of a global management system, the Telescent switch was logically divided into four sub-switches and each one was monitored and controlled independently by a MonALISA service. The topology for such distributed setups is currently done using configuration files for each switching unit, but it can be extended to use the RFID information from the Telescent switch as soon as this is available. Figure 118 presents the topology GUI in the MonALISA framework for global systems. It presents all the connected links and the interfaces used for each device.

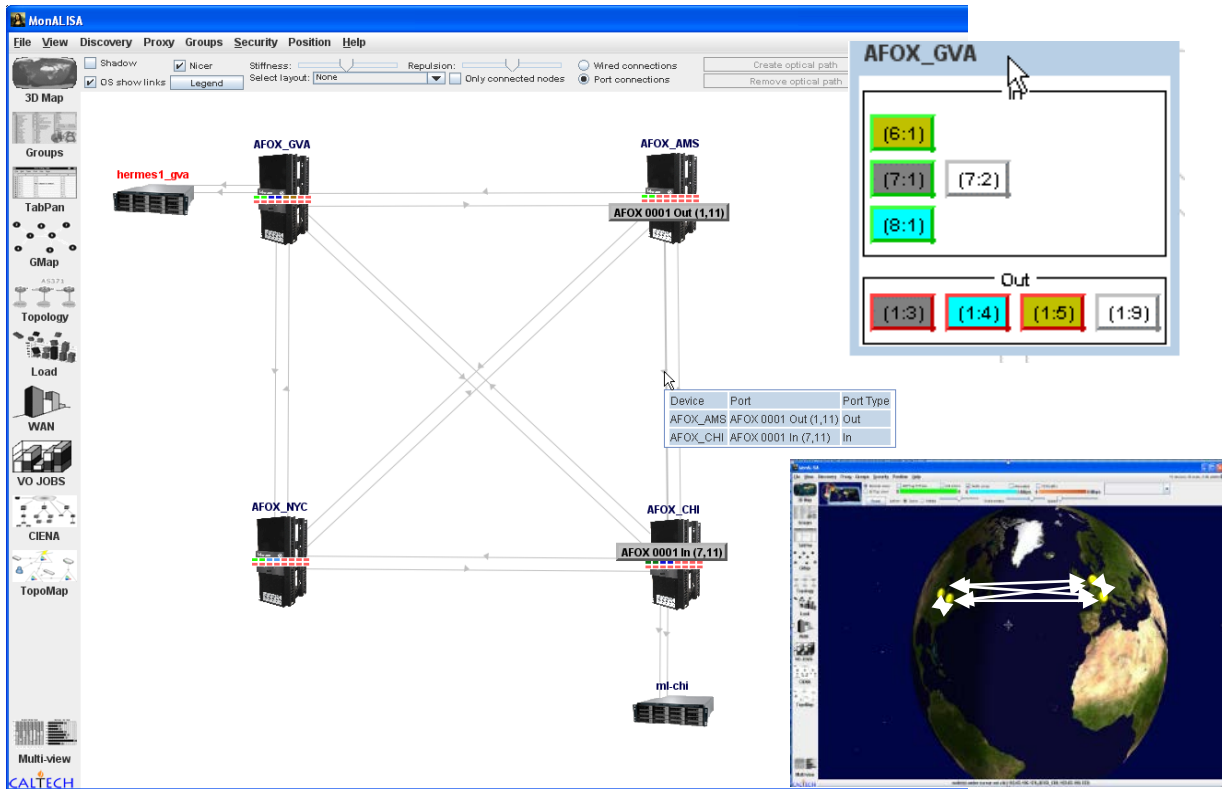


Figure 110: The MonALISA GUI presenting global topology for four Telescent sub-switches.

We used this setup to test different reconfigurations using four (virtual) switches. Two servers were connected to the switch (ml-chi and hermes3-gva) and we used a set of cross connects to simulate the connectivity between the virtual switches. The Fast Data Transfer⁴⁵ (FDT) application was used to send data between the two systems. In Figure 119 we present topology reconfiguration and the traffic between the two servers. As soon as the reconfiguration was done, the FDT transfer recovered. The reconfiguration time for this test is quite long because all the operations are done on the same physical switch and it requires two fiber connections per link. The switching time per fiber will also improve in the next versions of the Telescent switches.

To build a global network management system, we need to integrate several different types of network devices with complex interconnection topologies. Dedicated MonALISA modes are under development to provide status and connectivity information for different types of routers and switches. This information is used together with the Layer 0 connectivity maps from the optical switches. Figure 120 presents a simulation of the US LHCNet topology which includes several types of network devices. These global views are used to develop higher level services capable of taking automatic action and generate the reconfigurations maps when we detect failures in connectivity or network equipment.

⁴⁵ FDT web page : <http://monalisa.cern.ch/FDT>

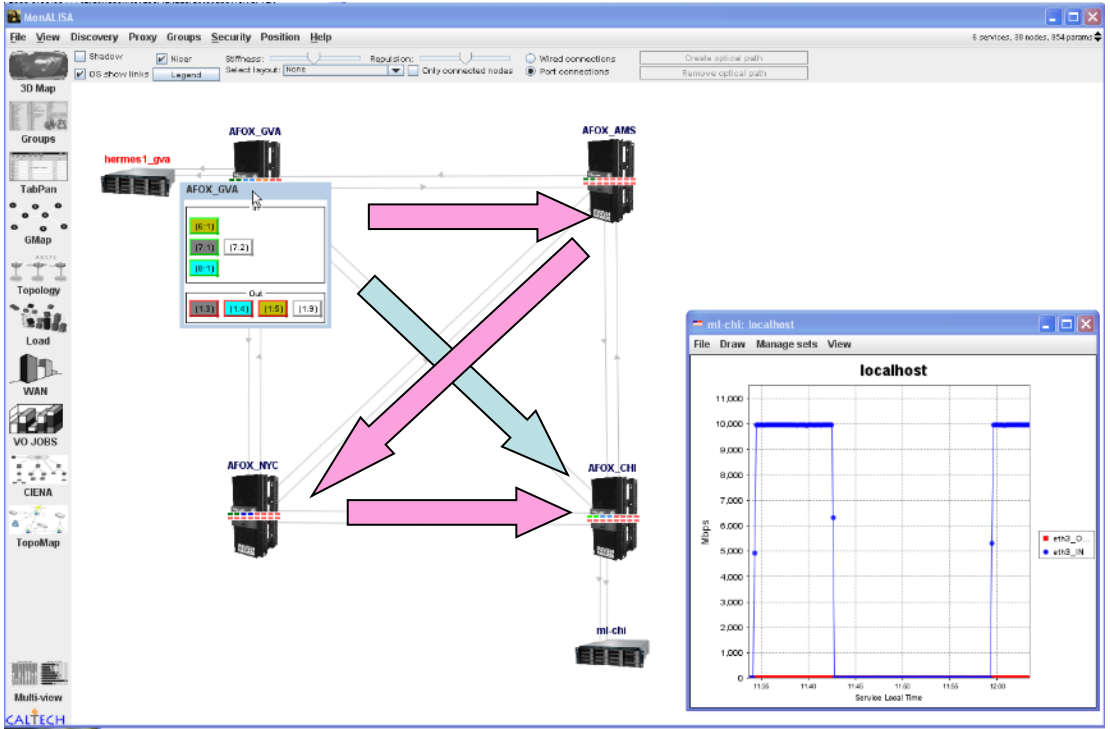


Figure 111: A global reconfiguration for the four sub-switches. Initially the two systems were connected using the direct link (blue arrow). The network was then reconfigured to connect the two systems using all four sub systems (pink arrows). The data transfer between the two servers recovered once the reconfiguration was done.

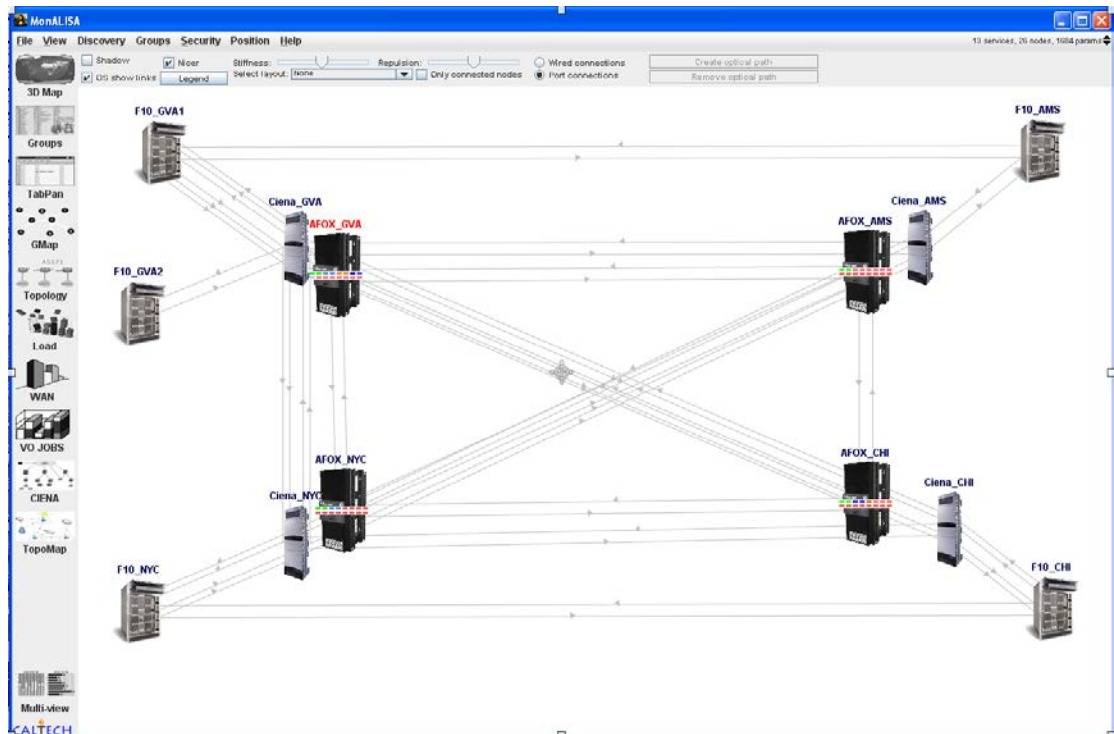


Figure 112. A simulation of the global topology for US LHCNet.

Monitoring OpenFlow US LHCNet testbed

Starting in 2013 we developed a set of dedicated monitoring modules for the Olimps/FloodLight OpenFlow controller. The modules use the REST/JSON interface to query the controller, which exports the information about the topology, ports and flows utilization. The information is aggregated inside the MonALISA monitoring service and is then presented to the client. Figure 121 shows the GUI for the OpenFlow monitoring of multipath flows.

This was used in demonstrations of OpenFlow-driven dynamic flow management across multiple network paths among the US LHCNet points of presence using Multipath TCP, in the context of the DOE/OASCR OlimPS project. This work, demonstrated at the 2013 GLIF, TERENA and Supercomputing conferences, has led to a persistent transatlantic OpenFlow testbed (see Figure 121) that is being used to help drive OpenFlow developments, notably developments in dynamic network circuits and flow control, of direct interest to the LHC experiments through the ANSE project and LHCONE.

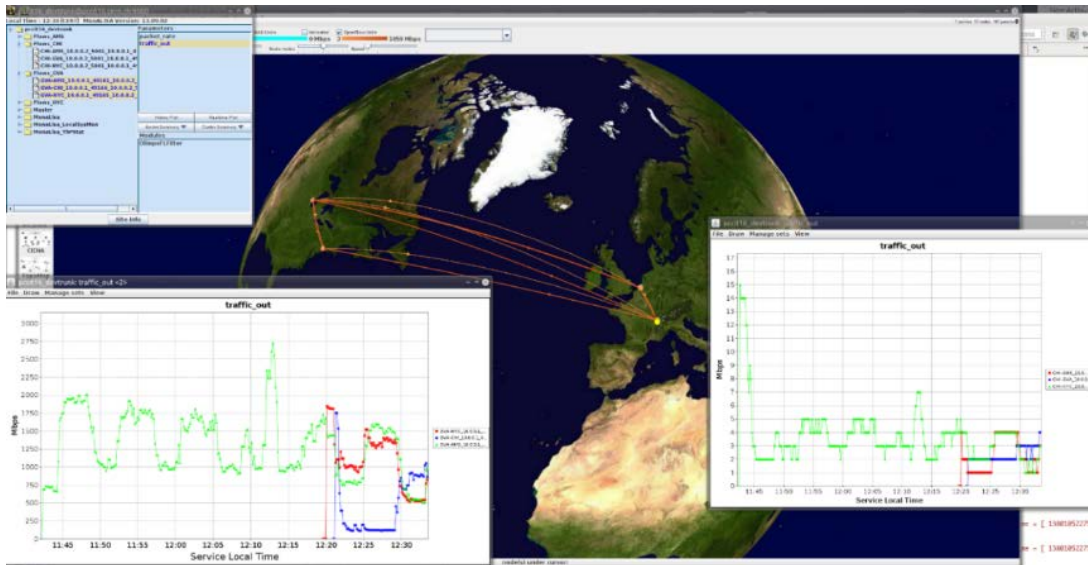


Figure 113. The MonALISA 3D GUI presenting the OpenFlow testbed in all four US LHCNet PoPs. The two history plots represent the active flows in the OpenFlow testbed in Geneva and Chicago.