PROCEEDINGS OF SCIENCE

100 Gbps connection to the Large Hadron Collider Open Network Exchange (LHCONE), a virtual private network of the LHC Community for example the German collaborator DE-KIT

Bruno Hoeft[1]
Karlsruhe Institute of Technology
Hermann von Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: `bruno.hoeft@kit.edu`

Andreas Petzold
Karlsruhe Institute of Technology
Hermann von Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
E-mail: `andreas.petzold@kit.edu`

To provide the file server and the compute nodes with adequate network bandwidth the 10Gbps network of Karlsruhe Institute of Technology (KIT) [1] does not offer the required capacities. KIT is a collaborator of the Large Hadron Collider (LHC) [2] and their throughput requirements are not able to become realised with the current network capacities. The upgrade of the project GridKa/DE-KIT[2], the considerations leading to the upgrade as well as some of the problems undergone while walking through the 100Gbps network upgrade are discussed.

International Symposium on Grids and Clouds 2016
13-18 March 2016
Academia Sinica, Taipei, Taiwan

[1] Speaker
[2] GridKa/DE-KIT – The German WLCG-Tier-1 is named GridKa and within the network community DE-KIT. Throughout this document DE-KIT is used.

1.Introduction

Karlsruhe Institute of Technology (KIT) is the largest combined federal research and state education facility in Germany. The Steinbuch Centre for Computing (SCC) [3], one institute of KIT operates the German Worldwide Large Hadron Collider Grid – Tier1 (WLCG-Tier-1) [4] site and has been involved in designing and developing of LHCONE [5] from the very beginning. The virtual private network (VPN) LHCONE connects LHC collaborating sites, which are providing compute resources to CERN [6]. The upgrade of the DE-KIT local area network (LAN) as well as the wide area network (WAN) connections are the basis of this document. Since the SCC is the information technology center of KIT, the activities of SCC cover classical and specific tasks of a modern IT service center. The services of the IT service center as well as research and development aim particularly at a permanent and innovative optimization of the IT services. New network technologies and transfer capacities are evaluated and brought in production, if they are mature and justifiable. The involvement in 100Gbps transfer capacity reaches back to times before the final draft of the IEEE [7] 802.3ba [8] standard was approved in 2010.

2.Historic 100Gbps projects

The following paragraphs describe historic 100Gbps involvements.

2.1KIT and FZJ are connected with a bandwidth of 100Gbps

Already as early as 2010 a 100Gbps installation was deployed between Karlsruhe Institute of Technology (KIT) and the Research Center Jülich (ForschungsZentrum Jülich (FZJ)) [9]. Originally initiated by Deutsches ForschungsNetz (DFN) [10], the research partners were joined and supported by Huawei [11], GasLINE [12] and Cisco [13] as further participants. At each site up to max. 11 hosts each with a 10Gbps port were connected to a Cisco CRS-3 router. The routers at KIT and FZJ were connected with a 100Gbps bit uplink to a Huawei Dense Wavelength Division Multiplexing (DWDM). GasLINE contributed a standard fiber (ITU G.652 + G.655 [14]) to the testbed, to prove the 100Gbps readiness of their fiber. The two Huawei DWDMs were connected with a GasLINE fiber over a distance of approx. 440 km. Monitoring showed that a transfer rate of 96.5 Gbps could be achieved with a process between a client and server memory to memory throughput as shown in Figure 1.

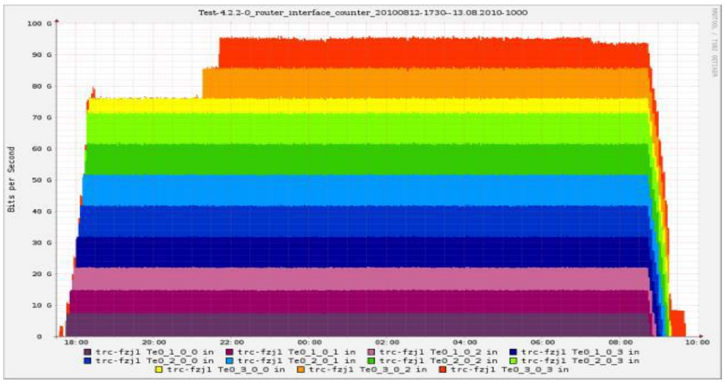

Figure 1 : aggregated throuput graph of 11 hosts

With a monitoring and measurement framework the most important IP network performance metrics could be recorded, e.g. one-way delay (OWD), packet loss and packet reordering. All relevant counters at the routers were captured as well. These combined information of the monitoring system indicated no packet losses nor any reordering of packets during the performed tests. Further details about this testbed and the results can be found at Terena 2011 proceedings under the title “DFN@100G – A Field Trial of 100G Technologies related to Real World Research Scenarios” [15]

2.2 SuperComputing [16] (SC) 2013 – 100Gbps from KIT to SC13 (Caltech-booth)

Three years later KIT was part of Caltech [17] 1 Terabit data transfer initiative to their booth at SC13 in Denver. The connection was a multi National Research and Education Network provider (NREN) collaboration as figure 2 displays. Several servers at KIT were connected to a Cisco Nexus 7000 and from there to DFN. DFN brought the light from Karlsruhe to Frankfurt via their DWDM equipment of ECI [18] and handed it over to Géant [19]. Géant then brought the light to Amsterdam where it was handed over to Netherlight [20]. Netherlight fed the light into the transatlantic link "ANA-100G" to Manlan of Internet2 [21] in New York. From Internet2 ESnet [22] took over and brought the light to the SC13 venue in Denver and there it was handed over to SCinet, the network provider of SC. SCinet brought it to the Caltech booth of SC13. Transferring LHC data files from the storage at KIT to the storage at the Caltech booth at the SC13 show floor was possible with a maximum data rate of 75Gbps, even though the transatlantic ANA-100G link only became available shortly after the SC13 show floor had opened.

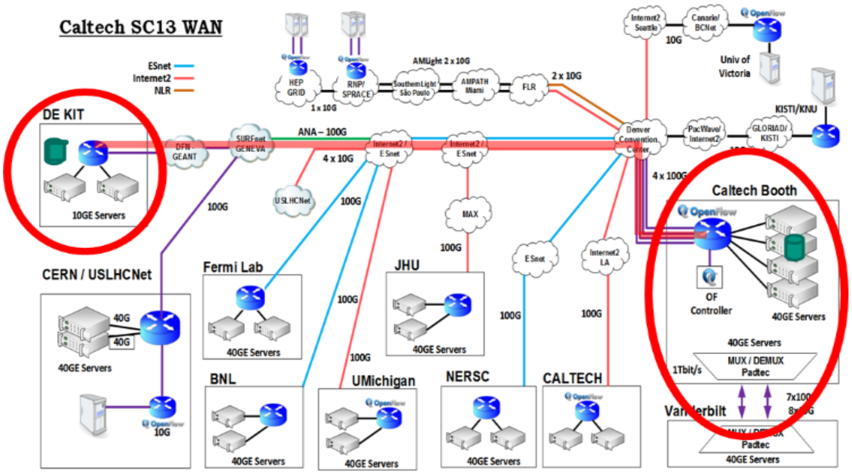

Figure 2 : multi NREN transatlantic light path

3.Introduction of 100Gbps networking at DE-KIT

3.1Current deployment of a 10Gbps based network

After this brief review of historic activities, this section describes our current deployment of 10Gbps WAN networking at DE-KIT shown in figure 3. Besides one 10Gbps direct link to CERN, DE-KIT has three 10Gbps direct connections to Worldwide Large Hadron Collider Grid – Tier1 (WLCG-Tier-1) centers in Europe and two dedicated links to WLCG-Tier-2 [23] centers in Europe, FZU [24] in Czech Republic (Prague) and a dedicated link to one WLCG-Tier-2 center (Poznan) in Poland; this last link connects the center in Poznan and two other Polish WLCG-Tier-2 centers. The direct link to CERN was frequently saturated above 75% utilization, i.e. a complete saturation of the link is foreseeable and an upgrade will soon be required.

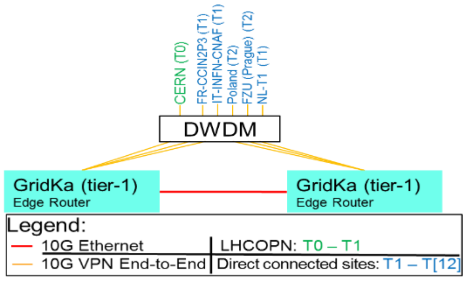

Figure 3 : dedicated connections between DE-KIT and European WLCG-Tier-[1,2]

3.2Géant is motivating to migrate direct connections to LHCONE

In 2014 Géant suggested to move the WLCG-Tier-1 to WLCG-Tier-1 traffic from the dedicated connections between the WLCG-Tier-1s to LHCONE for a test period of three months only (April to June 2014). Three European NRENs (the German NREN DFN, the Italian NREN

GARR, the French NREN Renater) joined the effort and doubled the LHCONE connection capacity of the participating WLCG-Tier-1 centers IT-INFN-CNAF, FR-CCIN2P3 as well as DE-KIT for this test period. Neither during the merging from the dedicated connections between the centers to LHCONE, nor during the three-month test period of the routed LHCONE Virtual Private Network (VPN) connection a significant packet error rate was measured or any other end user complaints became evident. The red line in figure 4 shows the transmission date before and after the migration of the traffic to LHCONE. There was no visible interruption of the traffic. During the test period the incoming data rate to DE-KIT from LHCONE increased by about one third and peaks above 15Gbps were measured for short periods.

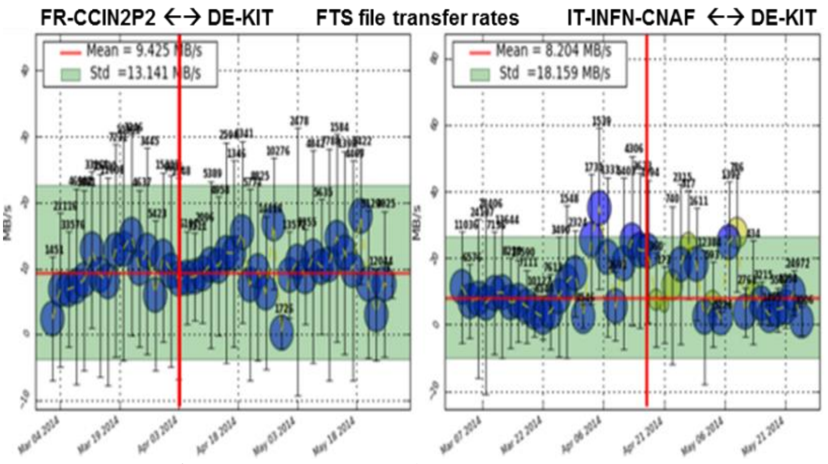

Figure 4 : transfer migration to routed LHCONE

3.3DE-KIT WAN upgrade

In 2015, during a revision of the dedicated DE-KIT connections to WLCG-Tier-1 sites, it became obvious that here were also actions required, since the capacity usage, not only of the link DE-KIT to CERN, as mentioned before, but of every link was often saturated beyond 75% of the link capacity. Therefore, an upgrade of the 10Gbps WLCG-Tier-x links to the WLCG-Tier-1 DE-KIT had become necessary. One option was to deploy a second 10Gbps link to all WLCG-Tiers directly connected to DE-KIT. Discussions with DFN highlighted another option: upgrading the DE-KIT LHCONE link into a high capacity connection of a 100Gbps link. For reasons of redundancy two 100Gbps connections were deployed, but entitled to use only 50% of the capacity. This upgrade was not an exclusive action of DE-KIT, but a combined effort of three European

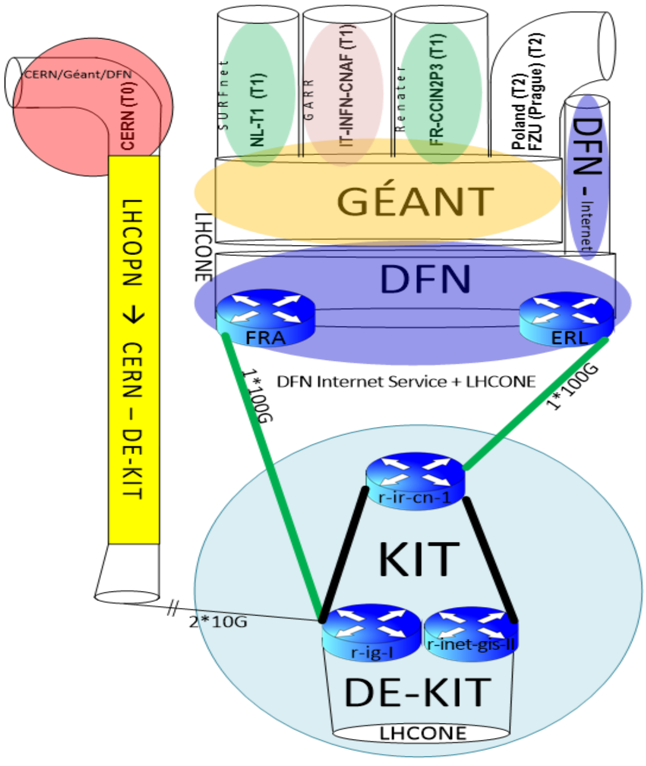

Figure 5 : DE-KIT uplink to LHCONE

NRENs (DFN, GARR, Renater) as well as the WLCG-Tier-1 centers of France FR-CCIN2P3, Italy IT-INFN-CNAF, Germany DE-KIT to join LHCONE with a 100Gbps uplink. The dedicated 10Gpbs links between the European WLCG-Tier-1 centers were supposed to build the backup links of the sites to reach CERN. This had now to be covered by LHCONE. It is now the NRENs responsibility to make sure that LHCONE provides sufficient bandwidth in case it has to cover the backup between CERN and one of the three WLCG-Tier-1 sites. The other dedicated links to the European WLGC-Tier-2 sites were migrated to LHCONE, too. Figure 5 shows the new 100Gpbs WAN network layout of DE-KIT. Our plans were based on the following calculation: the previous usage of LHCONE specific traffic was 15 Gbps, the dedicated direct links to the WLGC-Tier-1 centers 'contributed' 8Gpbs usage each (24 Gbps) and the general purpose internet traffic of DE-KIT was 10Gbps. We rounded this to approx. 45Gbps and took into account the increase of the data rates foreseen for the “LHC run period 2”, where the representatives of the relevant experiments had stated that approximately 25% more traffic should be expected. To anticipate a data rate of 55Gbps, seemed to be a realistic aim for the year 2016.

4.Upgrade to 100Gbps at KIT

KIT itself has its own 100Gbps deployment. The two main campuses of KIT with a distance of 7 miles by air are equipped with two edge routers at each location connected to two different NRENs.

The first upgrade planning included the redundant edge router at each location of KIT. Upgrading only campus north (CN) would include 6 ports that were due for an upgrade - two ports connecting to the NREN DFN at KIT-CN and 4 ports between the aforementioned redundant edge routers, as shown in figure 6.

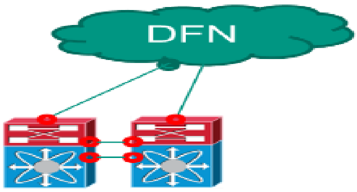

Figure 6 : first proposed 100Gbps upgrade including connection of KIT to DFN

A 100Gbps upgrade over the complete routing equipment of KIT would include 74 port upgrades and moreover 20 100Gbps transponder ports for the DWDM connections between the two campuses. The DWDM connects the two KIT campuses with two redundant and crossing free rented dark fiber, one with a distance of 16 miles, the other one with 21 miles.

One edge router at CN and one at CS connected to different NRENs were considered sufficiently redundant. The limited financial budget of KIT for the 100Gbps migration was a second motivation to reduce the number of edge routers to one at each main campus. The redundant core routers at each campus should also be reduced to one core router. This presupposition decreases the number of ports considered for the upgrade to 44 router ports and the transponder ports to 16 as displayed in figure 7.

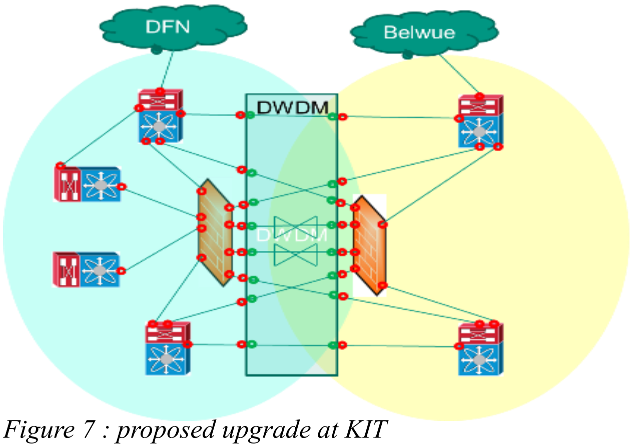

Figure 7 : proposed upgrade at KIT

The upgrade will have to be split into different parts and divided over years to cover the costs of the upgrade in KIT corresponding with infrastructure budgets of the years to come. The first production 100Gbps interface has already been deployed in 2013 between the SCC project Large Scale Data Facility (LSDF) [25] at KIT and BioQuant [26] at University Heidelberg. Towards the end of 2014 a 100Gbps interface was deployed to connect DE-KIT to LHCONE.

Also in 2014 a market survey was started to evaluate suppliers of different edge routers. The aim was to find the one best fitting for KIT. A team was investigating the information provided by various vendors and compared them with KIT's requirements: are the required protocols including at least bgp, ecmp, ospf, pbr, ipv4 and ipv6 supported by the router as well as IEEE 802.1q (vlan)? Is there sufficient router memory to accommodate at least two copies of the current internet ipv4 and ipv6 bgp table with 570.000 entries and the number of the linespeed 100Gbps ports? Do the costs quoted for the router include a maintenance contract? Considering all these questions led to the conclusion that the Brocade MLXe router was the best option for KIT.

However, there were more points to be discussed and the decision making process took longer than expected. The order of the edge router KIT-CN was split into two parts. The first part was ordered and has already been delivered. The ordering process of the second part has not been finished. We hope to complete the deployment process of the edge router KIT-CN in Q2/2016.

There are two major points that must be dealt with in regard to their adaption – the simply network management protocol (snmp) and sflow, a method for monitoring traffic in switched and routed networks as defined in RFC 3176. The snmp data retrieving for the cacti [27] monitoring program must be changed and deployed with the Brocade router templates. Packet flow monitoring is deployed at KIT to identify e.g. the source of malicious packets and to identify communication processes down to a certain layer-3 interface. Nfsen and nfdump, two public domain software packets are currently configured for netflow data. The netflow protocol is defined at RFC 3954 and is deployed at network components of the vendor Cisco. The sflow protocol is offered by Brocade and is supported within nfsen and nfdump. The support of both flow protocol formats will need to be implemented in the network traffic monitoring tools.

Towards the end of 2016 we intend to start the ordering process for the edge router KIT Campus South (KIT-CS) and the 100Gpbs upgrade of the DWDM equipment connecting KIT-CS and KIT-CN. The sequence of the following steps has not been finally decided yet: the upgrade of the KIT core router or the upgrade of the KIT security system including the firewall.

After the final deployment of the KIT edge router the redundant LHCONE uplinks directly connected to DFN for the project DE-KIT will undergo another change. The DE-KIT edge routers will move half a step back from the edge. The DE-KIT edge router will keep two 100Gbps uplinks. One uplink will connect the DE-KIT edge router to LHCONE directly via the DFN uplink, whereas the second (redundant) uplink will be connected via the KIT edge router to DFN.

5.DE-KIT core upgrade to 100Gbps

In 2016 the project DE-KIT will execute several upgrades in the core and edge area. In the core area the number of routers will be reduced from 6 to 4 routers - this will diminish the complexity as well. One new router - a Cisco Nexus 7700 (r-grid145) - will replace three Cisco catalyst 6509 routers and can provide a fully non-blocking throughput for each port and the possibility of a wide variation of different transfer capacities (1/10/40/100 Gbps), if necessary. Even though the new router requires less rack space it covers the same number of ports as the previous routers and still offers quite a number of additional ports for later use. One additional Nexus 7700 will be placed in the DE-KIT core and replace the r-grid2 a catalyst 6509 router. An interconnecting link with the capacity of 100Gbps will be deployed between the new Nexus 7700 routers. The new topology is shown in figure 8, with a pair of redundant edge routers connected to the NREN DFN on one side and on the other side to the Core network of DE-KIT. Between the Edge and the Core network a firewall has been built. The other two core routers (r-grid3 and r-grid6) which are Nexus 7010 routers stay on 10Gbps technology. We will replace only the hardware (linecards) of types that come close to the end of their support life cycle as announced by Cisco. The interconnecting links will be built with a multiple 10Gbps link between the two routers and their link to the new Nexus 7700.

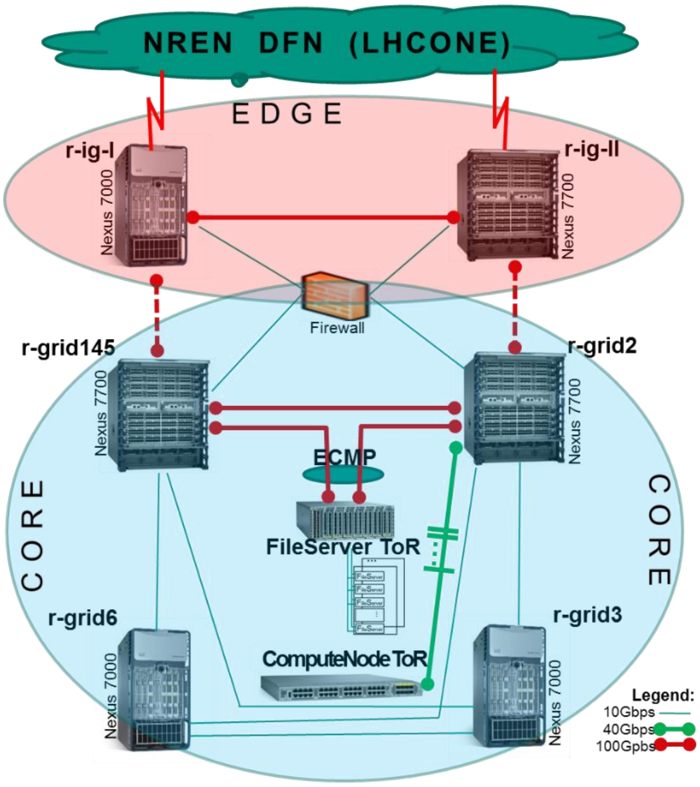

Figure 8 : tentative network sketch of DE-KIT

For the deployment of the compute nodes Fabric Extender [28] as a Top of Rack (ToR) switch will be placed in each rack. The uplink to the Fabric Extender will be a multiple 40Gbps, since 100Gbps uplinks are not available. The 40Gbps uplink can be realized with the fiber infrastructure currently deployed at DE-KIT using a bidirectional QSFP transceiver. The Fabric Extender provides 48 1/10Gbps interfaces to the compute nodes. A more radical change is, of course, the deployment of "file server top of the rack" switches instead of connecting the file server directly to the core router. Each switch can be modularly assembled with 10/40/100Gpbs interfaces. One option is to connect 3 neighboring file server racks via 40Gbps copper cables to one switch and connect this switch with 100Gbps redundant uplinks each to one of the new router nexus 7700. For the redundant uplink one of the protocols MLAG (multi chassis link aggregation), ECMP [29] or Fabric Path [30] will be deployed.

In the edge area one Nexus 7000 with 100Gbps interfaces has already been deployed. This router will be equipped with one additional 2 port 100Gpbs linecard. The second edge router – at present a catalyst 6509 - will be replaced by a Nexus 7700. This replacement will only be realized when the 100Gbps M3 linecard becomes available. The M3 linecard will be able to accommodate the full internet routing table and will include deep buffers for each 100Gbps port. This router will then enable the second 100Gpbs VPN uplink to LHCONE and the 100GE router interlink between the edge and the core router will be completed.

6.Conclusion

In summary DE-KIT has accrued from the early cooperation between KIT (then still Forschungszentrum Karlsruhe GmbH) and CERN. In close cooperation with the international partners in research and business a reliable network has been set up. LHCONE and LHCOPN are directs results of the association with CERN. The ongoing and continuous cooperation has led to a state of the art network which is outstanding within the research community.

The 100 Gbps upgrade as well as the network changes that are bound to be executed in 2016 and 2017 are vital as a solid basis for installation and buildup of the storage and compute nodes that will be acquired soon. Together these implementations will enable DE-KIT to fulfill their commitments required for the LHC Run 2 as well as other responsibilities.

References

[1] Karlsruhe Institute of Technology, Research and Education Institute of Germany at State of Baden-Würtemberg: http://www.kit.edu.

[2] Large Hadron Collider (LHC) at CERN, a high energy physik particle accelorator, accelorating particles close to light speed: http://home.cern/topics/large-hadron-collider.

[3] Steinbuch Center for Computing (SCC), the IT Research and Service Center: https://www.scc.kit.edu/en/index.php.

[4] Worldwide Large Hadron Collider Grid (WLCG), a world wide collaboration of sites providing compute- storage- and archive-resources to LHC: http://wlcg.web.cern.ch/.

[5] LHCONE, Large Hadron Collider Open Network Exchange, a distributed private network connecting LHC collaborating sites: http://lhcone.web.cern.ch/.

[6] Conseil Européen pour la Recherche Nucléaire (CERN), the European Organization for Nuclear Research: http://cern.ch/about.

[7] Institute of Electrical and Electronics Engineers (IEEE), leading standards development organization: https://www.ieee.org/index.html.

[8] 802.3ba, an IEEE 40/100 Gbps transceiver standard, available since June 2010.

[9] Forschungszentrum Jülich (FZJ), a research center with the focus of the areas of energy and environment, information and brain research: http://www.fz-juelich.de/portal/EN/Home/home_node.html.

[10] Deutsche ForschungsNetz (DFN), the german research and education network: http://www.dfn.de.

[11] Huawei, a chinese vendor supporting the testbed with optical 100Gbps DWDM and routing equipmemt : http://carrier.huawei.com/en/products/fixed-network.

[12] GasLINE, optical fiber network (10.000km in germany): http://www.gasline.de.

[13] Cisco, vendor, supplied router and switches to the testbed: http://cisco.com.

[14] International Telecommunication Union (ITU), standartisation organisation, G.655 : Characteristics of a non-zero dispersion-shifted single-mode optical fibre and cable: https://www.itu.int/en/Pages/default.aspx, G.652 : Characteristics of a single-mode optical fibre and cable.

[15] Bruno Hoeft, Andreas Hanemann, Robert Stoy, DFN@100G – A Field Trial of 100G Technologies related to Real, Trans european research and education networking association (Terena): https://tnc2011.terena.org/getfile/336, Terena Network Conference (TNC) 2011.

[16] The Super Computing Conferece Series: http://supercomputing.org.

[17] California Institute of Technology (Caltech), University at Pasadena, California, USA: https://www.caltech.edu.

[18] ECI, Vendor of optical fiber equipment: http://www.ecitele.com.

[19] Géant - trans european research and education network: http://www.geant.org.

[20] NetherLight - open light path exchange in Amsterdam: https://netherlight.net.

[21] Internet2 - USA national resaerch and education network provider: http://www.internet2.edu.

[22] ESnet - USA national research and education network provider: https://www.es.net.

[23] MONARC model, is based on a hirarchical tier model: https://cds.cern.ch/record/510694/files/Phase2Report.pdf, Models of Networked Analysis at Regional Centres for LHC Experiments.

[24] Dagmar Adamova1, Jiri Chudoba2, Marek Elias, WLCG Tier-2 site in Prague: a little bit of history, current status and future perspectives, Institute of Physics AS CR (FZU) at Prague, Czech Republic: https://indico.cern.ch/event/258092/contributions/1588570/attachments/454212/629622/adamova_acat_version3.pdf.

[25] LSDF, Large scale Data Facility: http://wiki.scc.kit.edu/lsdf/index.php/About_the_LSDF.

[26] BioQant, Quantitative Analysis of Molecular and Cellular Biosystems: http://www.bioquant.uni-heidelberg.de/.

[27] cacti, is a complete network graphing solution: http://cacti.net.

[28] Fabric Extender (FEX), Cisco proprietary technology delivers an extensible and a scalable fabric based on the standard IEEE 802.1BR (Bridge Port Extension): http://www.cisco.com/c/en/us/solutions/data-center-virtualization/fabric-extender-technology-fex-technology/index.html.

[29] ECMP, Equal Cost Multiple Paths: http://www.ieee802.org/1/pages/802.1bp.html.

[30] Fabric Path, Cisco proprietar - based on trill: http://www.cisco.com/c/dam/en/us/products/collateral/switches/nexus-7000-series-switches/at_a_glance_c45-605626.pdf.