100G Deployment@(DE-KIT)

Bruno Hoeft¹
Andreas Petzold¹

[1] Karlsruhe Institute of Technology (KIT), Steinbuch Centre for Computing (SCC), Hermann-von-Helmoltz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

E-mail: Bruno.Hoeft@kit.edu, Andreas.Petzold@kit.edu

Abstract. The Steinbuch Centre for Computing (SCC) at Karlsruhe Institute of Technology (KIT) has been involved fairly early in 100GE network technology. Initiated by DFN¹ (the German NREN), a first 100GE wide area network testbed over a distance of approx. 450 km was deployed between the national research organizations KIT and FZ-Jülich in 2010. Three years later in 2013, KIT joined the Caltech SuperComputing 2013 (SC13²) 100GE "show floor" initiative using the transatlantic ANA-100GE link to transfer LHC data from a storage at DE-KIT (GridKa) in Europe to hard disks at the show floor of SC13 in Denver (USA). The network infrastructure of KIT as well as of the German Tier-1 installation DE-KIT (GridKa), however, is still based on 10Gbps. As highlighted in the contribution "Status and Trends in Networking at LHC Tier1 Facilities" to CHEP 2012, proactive investment is required at the Tier-1 sites. Bandwidth requirements will grow beyond current capacity and the required upgrades are expected in 2015. In close cooperation with DFN, KIT drives the upgrade from 10GE to 100GE. The process is divided into several phases, due to upgrade costs and differing requirements in different parts of the network infrastructure. The requirements of the different phases as well as the planned topology will be described. Some of the obstacles we discovered during the deployment will be discussed and solutions or workarounds presented.

1. Introduction
KIT³ is the largest combined federal research and state education facility in Germany. The Steinbuch Centre for Computing (SCC)⁴ is the information technology centre of KIT. The activities of SCC comprise the classical and specific tasks of a modern IT service centre in science as well as in research and development aiming particularly at a permanent and innovative optimisation of the IT services.

During the last years, KIT/(SCC) has acquired a good understanding of both the handling of large data volumes and the accompanying requirements for high-throughput WAN access capabilities. Thus, KIT has been already involved at the very early deployment stage of the 100GE technology. In this paper, we describe the first trials of early 100GE technologies at SCC and our involvement in several testbeds of the new technology. With this well-founded experience, KIT plans to upgrade its network infrastructure to be able to cope with the increased network capacity requirements in the coming years. A major driving force for these upgrades are the increased requirements of GridKa⁵ (DE-KIT).

¹ https://www.dfn.de/
² http://sc13.supercomputing.org/
³ http://www.kit.edu/english/index.php
⁴ https://www.scc.kit.edu/en/index.php
⁵ http://www.gridka.de/cgi-bin/frame.pl?seite=/welcome.html
2. 100GE - Historical KIT Involvements
The first KIT involvement in 100GE installations goes back to 2010, when the 100GE IEEE standard was just to be ratified. The first trials were initiated by the German research network provider DFN with the following partners: Cisco, Huawei, Gasline, Research Centre Jülich, and KIT.

2.1. 2010 – KIT – Jülich (DFN/Huawei/Cisco/Gasline) [1]
Between Karlsruhe Institute of Technology (KIT) and Research Center Jülich (FZJ) 6, a 447 kilometer long fibre pair has been used for the trial. GasLINE, infrastructure provider for DFN’s national backbone, contributed this fibre pair to the trial. It has a typical quality (ITU G.652 + G.655) in comparison to the fibre pairs used for the national backbone.

Huawei Technologies contributed two Dense Wavelength Division Multiplexing (DWDM) systems, one in Karlsruhe at KIT and one in Jülich at FZJ, with 100GE transponders (see Fig. 1). In addition, Huawei placed optical amplifiers for boosting the signal at five optical amplifier locations along the fibre path.

Cisco Systems supplied a CRS-3 router at each endpoint, one for Karlsruhe and one for Jülich. Each of the CRS-3 routers was equipped with a 100 Gbps interface. The multivendor coupling of the 100GE interfaces was a challenge, since the 100 Gigabit Ethernet standard IEEE 802.3ba had only recently been approved and interoperability experiences had yet to be gained. At both locations, KIT and FZJ, testbed hosts were placed with 10 Gbps network interface cards for traffic generation. The traffic generated by these hosts was aggregated by the CRS-3 routers. Traffic volumes being representative for research and education networks were generated using this testbed infrastructure. Both at KIT and FZJ, nine dedicated hosts per location were used for traffic generation.

2.1.1. UDP 7 and TCP 8 tests
With a software-based packet generator mgen 9 deployed on a linux host, UDP packets were injected into a routing loop. Another technique even closer to a real production environment was TCP-based. At KIT and FJZ, dedicated hosts were deployed for this testbed. Streams over the 10GE interfaces of the hosts were injected with iperf.

---

6 http://www.fz-juelich.de/portal/EN/Home/home_node.html
7 https://www.ietf.org/rfc/rfc768.txt
8 https://www.ietf.org/rfc/rfc793.txt
9 http://www.nrl.navy.mil/itd/ncs/products/mgen
2.1.1. UDP tests

During the UDP test, a maximum bit rate of 99.6 Gbps was achieved on the 100GE link as shown in Fig. 2. The figure shows the bit rate as it was measured on a 100GE router interface, when 10 different UDP/IP flows were injected, each with a bit rate of 78.74 Mbps and TTL=254. Each flow was scheduled to start with a delay of 10 minutes after the previous flow was started.

2.1.1.2. TCP tests

The TCP load tests were conducted using the application iperf. Nine hosts at KIT and nine hosts at FZJ acted as iperf clients and iperf servers, respectively. During a preparation phase, individual pairs of hosts at KIT and FZJ were used in order to find each host’s performance limitation and verify the tuned kernel parameters. Due to variation of the hardware specification of the testbed machines, the throughput of the iperf TCP streams varied between 5 Gbps and 9.5 Gbps, even after the kernels’ TCP protocol tuning. Afterwards, as many host pairs as possible were used simultaneously. The sum of the sustained streams measured between the nine nodes on each site was up to 79 Gbps. After converting the two measurement hosts of DFN at both sites to data source/destination hosts, the throughput was increased to 95 Gbps as shown in Fig. 3. There was a gap of appr. 5 Gbps to saturate, but there were no additional hosts with 10 Gbps interfaces available for the testbed.

The result demonstrates the very stable behaviour of the whole network including routers and optical transmission systems during the full test time frame of more than 14 hours. In particular, each individual TCP stream reached the maximum TCP throughput given by the server’s limits, without any interference between each individual TCP stream.

In total, all tests - even those lasting for several hours - showed a very stable behaviour. During the tests with a duration of 18 hours, the accumulated transferred data volume was 770 Terabyte. A stable utilisation at 95% of the theoretically reachable 100 Gbps was achieved over several hours. The availability accomplished matched the expectation in a production environment. During the UDP test, a single stream limitation of 92 Gbps was discovered. Cisco acknowledged that this limitation was caused by limitation of the ASIC and stated that the First Customer Shipment will be equipped with a newly designed ASIC with 100GE line speed capability.

2.2. Cisco NEXUS 7000 M2 Linecard EFT

In 2012, Cisco invited KIT to participate in the Early Field Test (EFT) of the not yet released Nexus M2 Linecard. Two N7K-M202CF-22L line cards, each with two 100GE interfaces, and two line cards with 24 port 10GE interfaces N7K-M224XP-23L were supplied by Cisco. The 14 hosts with 38 10GE interfaces were divided over two separate virtual routers (VDC) and two of the 100GE ports were assigned to each VDC as shown in Fig.4. The VDCs were separated by 10km (fibre on a drum). A throughput of 200Gbps was achieved. Besides the high throughput, the M2 line card was able to load the full internet routing table successfully. During the tests, a limitation of 12.75Gbps per stream was found, and Cisco acknowledged that this limitation was due to the ASICs of the line card.

---

10 http://sourceforge.net/projects/iperf2
2.3. KIT – TU Dresden

In 2013, KIT and the Technical University of Dresden were connected via 100GE. Over the link, a throughput of 97.63Gbps could be achieved with iperf streams between 14 hosts involved on each site.

2.4. KIT – CERN\(^{11}\)

At the end of 2013, KIT and CERN were connected through Amsterdam and it was possible to arrange tests between these sites. LHC data read at KIT from storage a throughput of 100Gbps could be achieved. At CERN, approx. 1/3 of the data were written to disk and 2/3 were dropped to /dev/null.

2.5. KIT – SC13 Caltech\(^{12}\) booth [2]

KIT was part of the institutions (CERN/USLHCNet, FermiLab, UMichigan, BNL,...) contributing data streams to the show floor initiative of Caltech at SuperComputing13 in Denver. A multi National Research and Education Network (NREN) provider link was established. DFN brought the light of KIT from Karlsruhe to Frankfurt, handing it over to Géant. Géant peered in Amsterdam with Netherlight. Here was the landing point of the transatlantic link ANA-100G over which the light was transported to Internet2 at New York MANLAN. From here, ESnet brought the light to SCNet in Denver. Eventually, the light was dropped at the show floor booth of Caltech. At KIT, 20 servers were installed and dedicated to this initiative. Each of them was equipped with a 10GE NIC. At the Caltech SC13 show floor booth, three systems with 40GE NICs were receiving and storing the LHC data to disk. All systems of the Caltech show floor booth were equipped with SSDs. The transatlantic ANA-100G link was down before SC13. The sea cable was broken. Due to weather and sea conditions, the

\(^{11}\) http://home.web.cern.ch/

\(^{12}\) http://www.caltech.edu/
ANA-100G was repaired just in time for the show floor opening of SC13. With Fast Data Transfer (FDT)\textsuperscript{13}, LHC data could be read from storage at KIT and written to storage at the show floor with a transfer rate of up to 75Gbps. For this WAN transfer, several factors needed to be taken care of.

Besides the network equipment over the long distance and the controller card of the storage server, more investigations were necessary, e.g.: in the CPU capacity of each host, in the PCIe and SATA bus of each host, as well as into the disks and their controllers. Only by this thorough investigation we found that the first RAID controller for SSDs which was developed just before SC13 was not performing well together with the available SSDs. After removing the RAID controller and restructuring the echo stream system, the above-mentioned throughput rate could be achieved.

3. 100GE Deployment at KIT

At CHEP 2012, Andrey Bobyshev in a jointly written talk “Status and Trends in Networking at LHC Tier1 Facilities”\textsuperscript{3} already proposed that LHC Tier-1 centres will find it necessary to upgrade their network capacity to the next scale. At KIT, we are currently at the first stage of upgrading. During our involvement in the various 100GE testbeds we have earned valuable lessons and the 100GE WAN connection upgrade of two projects started already. A general upgrade to the next capacity scale will be a natural next step.

3.1. Projects

The 100GE deployment at KIT has started already. For two projects, 100GE connections are already installed. The KIT campus upgrade is still pending.

3.1.1. LSDF\textsuperscript{14} KIT – Uni Heidelberg BioQuant\textsuperscript{15}

In 2013, the first project production uplink was upgraded to 100GE. The data sources at BioQuant are the NIKON Imaging Center and the Hamamatsu TIGA (Tissue Imaging and Analysis). They generate massive amounts of data. These data are transferred from Heidelberg to Karlsruhe for long-term archival. In order to for the archival process to keep up with the large data volume produced at BioQuant, a new high capacity link was required.

\textsuperscript{13} http://iopscience.iop.org/1742-6596/331/5/052014/pdf/1742-6596_331_5_052014.pdf

\textsuperscript{14} https://www.scc.kit.edu/forschung/lsdf.php

\textsuperscript{15} http://www.bioquant.uni-heidelberg.de/
3.1.2. GridKa LHCONE\textsuperscript{16}

At the GridKa (DE-KIT) LHC Tier-1 centre, the necessity of a capacity upgrade had been apparent already for some time. However, it was necessary to evaluate two upgrade scenarios where either the existing point-to-point links would be kept or a new solution better suited to the new technology would be deployed.

3.1.2.1. Prior status

DE-KIT was connected with only 10GE links. Four 10GE LHCOPN\textsuperscript{17} were dedicated point-to-point links, one to the Tier-0 site CERN as well as one to each of the Tier-1 sites IT-INFN-CNAF, FR-CCIN2P3 and NL-T1, including the functionality of a backup link for the case of a non-operational Tier-1 Tier-0 link. There were two additional dedicated point-to-point links, one 10GE link to the Tier-2 site FZU (Prague) and one 1GE link to Poznan in Poland, from where it was distributed to the three polish Tier-2s. In addition, there were two redundant 10GE links to LHCONE serving the sites connected to LHCONE and one 10GE link to the General Purpose Internet (GPI) for all other sites.

3.1.2.2. Motivation

The motivation for the link upgrade resulted from the following reasons: The 10GE LHCOPN link to CERN was at the capacity limit over many hours even during LS1, when the accelerator at CERN produced no data. The point-to-point connection between DE-KIT and three other European LHC Tier-1 sites, the LHCOPN backup links, were not filled completely, but the projected Tier-1-to-Tier-1 data rates of the experiment during Run2 will require an upgrade of the links. The LHCONE link with two times 10GE was only filled to three quarters. However, according to the estimates for Run2, double the capacity will just be enough to fulfill the demands of the LHC experiments. As data rates to other sites connected to neither LHCOPN nor LHCONE are also expected to increase significantly, an upgrade of the link to the GPI will be required.

3.1.2.3. Actions taken already

An upgrade of the dedicated link DE-KIT to CERN with two times 10GE for LHCOPN is already deployed. During 2014, a Géant-initiated trial over three months was executed, including the NRENs DFN, GARR, RENATER to connect three of the European Tier-1 sites (IT-INFN-CNAF, FR-CCIN2P3, and DE-KIT) via LHCONE. For this trial, the capacity of the LHCONE link at DE-KIT was upgraded to 30GE as were the links at the other participating Tier-1s. This migration of the inter-Tier-1 communication for the trial went fairly smooth, as well as the data exchange of the sites during the trial. Even from the economical point of view, this migration is to be favoured compared to the upgrade of the dedicated point-to-point links. With the successful conclusion of the trial, the decision was taken to migrate the inter-Tier-1 communication of IT-INFN-CNAF, FR-CCIN2P3, DE-KIT and NL-T1 to LHCONE. Based on the experiences gained during the trial period, the final migration was executed without any problems. DE-KIT is now connected with redundant 80GE connections to LHCONE. Before that, the point-to-point links between the Tier-1s provided the backup link for the Tier-1-to-Tier-0 link. This backup link is routed via LHCONE. The NRENs are now responsible for

\textsuperscript{16} https://www.scc.kit.edu/forschung/lsdf.php

\textsuperscript{17} https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome
providing enough capacity to certify the backup links for the Tier-1-Tier-0 connectivity. The connection between KIT and DFN to the GPI has been upgraded from 10GE to 20GE. Two activities are still pending, the migration of the link to Poland and to FZU (Prague) to LHCONE.

3.1.2.4. Proposed final deployment
The KIT provider edge (PE) router will be replaced with a router offering 100GE interfaces and will serve both the GPI and LHCONE links via this connection. The DE-KIT (GridKa) edge router will move from the PE router to a KIT project router only and receive 100GE connections from the KIT PE router. The only remaining external connection at the DE-KIT router will be the 20GE LHCOPN connection to CERN.

3.2. 100GE Upgrade scenario for KIT
The network equipment of KIT consists of 4 DWDM systems, 20 layer-3 routers, 1800 switches, 1600 wireless access points and approximately 55,000 network ports. The network upgrade at KIT has to be designed to include all the different layers even if in the first step the upgrade will mainly include the layer-3 routers. Besides the two projects already equipped with 100GE links, analysis has to determine the most relevant network segment of KIT for the 100GE upgrade.

3.2.1. Deployment parameter
(Remark: During this paragraph, only a count of potential interfaces to be upgraded is provided. No sophisticated analyses of infrastructure or topology changes are included)
To get an idea of the size of the KIT 100GE upgrade, a closer look at the different upgrade phases is required:

- The first phase includes only the KIT-CN PE router, a VSS system, built out of two router chassis installed in two different locations at KIT-CN with two connections (redundant and active/active) to the provider, two router interlinks, and two links to the edge router of the DE-KIT Tier-1 centre. In total, there are eight interfaces requiring 100GE upgrades.
- The second phase will include KIT-CN and KIT-CS. The two campuses are approx. 10km apart. Two DWDM systems are deployed at each site. The two PE routers (VSS) of KIT-CN are already covered in the first phase, and the two PE routers (VSS) of KIT-CS will be part of the second phase. The second phase includes a total of four routers with 22 100GE interfaces and four 100GE transponders.
- During the third phase, the KIT core router and the firewall will be addressed. The interfaces to be upgraded are 70 100GE interfaces and 20 100GE transponder interfaces for an intercampus connect.
- In last phase, the different projects with dedicated routers need to be included. For each project, six additional interfaces requiring 100GE upgrade are estimated.

3.2.2. Deployment timeline
The workload as well the cost of the upgrade needs to be distributed over several years.
- In 2013, a 100GE link over a distance of approximately 50km was installed for the LSDF project connecting KIT and BIOQant, a partner organization at University of Heidelberg.
- In 2014, the German Tier-1 DE-KIT/GridKa has received a 100GE link to LHCONE.
- In 2015, the next deployment step will be to connect KIT-CN to DFN.
- In 2016, the next major deployment step will be to connect the provider edge router of KIT-CS to Belwue and connect the two edge routers with 100GE. At this stage, the upgrade of the
A DWDM system connecting KIT-CN and KIT-CS with 100GE transponder line cards needs to be included as well.

- The core routers and the security system have to be included in the following upgrade step, however, there is currently no estimation when this is about to happen.

### 3.2.3. Market evaluation

After the first investigations in the 100GE upgrade it became clear that an upgrade will not be possible without replacing the current routers at KIT. There are no line cards available with a 100GE interface capacity fitting the KIT PE/core routers for offering the interface capacity transition or upgrade from 10GE to 100GE. For the upgrade, at least the PE routers have to be replaced and probably some other routers as well. Currently, our PE router is from Cisco. Therefore, Cisco was the first to be approached in this issue and three options for a replacement were presented:

- **Catalyst 6800**, the follow-up model of Catalyst 6500, but the 100GE line card is not yet available.
- **Nexus 7000/7700**, the M2 line card is available, but the release of the follow-up M3 line card, allowing a wire speed single-stream throughput, is still pending.
- **ASR9000** series router is considered as a provider engine and could serve as a PE router.

However, since routers have to be replaced and a redesign of the network structure as well as the topology will anyway be necessary as part of the upgrade, the decision was taken to approach other vendors and initiate a market survey. The required specifications include the number of 100GE ports: The interoperability with the current environment including the DWDM, support of the current distributed firewall, the support and interoperability with IP[v4/v6] unicast/multicast, OSPF[v3], BGP[v4] and hardware-based IPv[v4/6] ACL filtering.

Some vendors were invited to present their solutions, but the decision is still open.

### 4. Next Steps

The above-mentioned market survey will not be finalized soon and will take a lot of time and effort. At the same time, we are working on plans to restructure the network layout, e.g. if the necessity arises for two PE routers and core routers for redundancy independently at each KIT campus. However, having gone through all these efforts, KIT will be equipped with a network with the required capacity for several years to come. Of course, there are still other changes and amendments to be considered, e.g. to include SDN capabilities, yet the bandwidth upgrade will be in place and will need no further attention.

### 5. Conclusion

Even though 100 Gbps is available today, upgrading a whole institution to 100 Gbps has to be planned very carefully. To begin with, the process must be defined in close cooperation with a vendor, since several specifics of a vendor have to be included in the local network design. Therefore, starting a market survey as part of the 100Gbps upgrade process is one of the challenges you will have to face. If you are willing to deal with a lot of uncertainties and do not mind changing long discussed plans now and then, you should give it a try. At KIT, we had a lot of obstacles to overcome, however, it was worth doing it.

### 6. References

[1] DFN@100G – A Field Trial of 100G Technologies related to Real World Research Scenarios; URL: https://tnc2011.terena.org/getfile/33

[2] SC13 Data Movement WAN and ShowFloor
   [http://indico.cern.ch/event/269840/contribution/8/material/slides/1.pdf](http://indico.cern.ch/event/269840/contribution/8/material/slides/1.pdf)

[3] Status and Trends in Networking at LHC Tier1 Facilities
   [http://iopscience.iop.org/1742-6596/396/4/042011/pdf/1742-6596_396_4_042011.pdf](http://iopscience.iop.org/1742-6596/396/4/042011/pdf/1742-6596_396_4_042011.pdf)