100G Ethernet in the wild – first experiences

Journal of Physics: Conference Series **331** (2011) 052007 doi:10.1088/1742-6596/331/5/052007

100G Ethernet in the wild – first experiences

Autor: Bruno Hoeft, Karlsruhe Inst. of Technology (KIT/SCC), Bruno.Hoeft@kit.edu
Co-Authors:
Robert Stoy, Deutsches Forschungsnetz (DFN), stoy@noc.dfn.de
Frank Schröder, Deutsches Forschungsnetz (DFN), frank@noc.dfn.de
Aurelie Reymund, Karlsruhe Inst. of Technology (KIT/SCC), Aurelie.Reymund@kit.edu
Ralf Niederberger, Forschungszentrum Jülich (FZJ), r.niederberger@fz-juelich.de
Olaf Mextorf, Forschungszentrum Jülich (FZJ), o.mextorf@fz-juelich.de
Sabine Werner, Forschungszentrum Jülich (FZJ), s.werner@fz-juelich.de

Abstract:

A 100 Gigabit Testbed was established in a collaboration of 6 partners. Three industry partners have contributed the fiber infrastructure, the DWDM equipment, as well as the required routers. 447 kilometer was the distance of the wide area testbed established in collaboration with the German NREN DFN between Karlsruhe Institute of Technology and Forschungszentrum Jülich. Before starting, DFN assured the quality of the fiber infrastructure, the operation of the DWDM systems at both locations, as well as the connection of the routers to this WAN link with a bandwidth of 100GE. 12*10GE interfaces were available at each site for connecting the local testnodes to the routers. A monitoring and measurement framework was installed for recording the most important IP network performance metrics, among them the One Way Delay (OWD) and its Variation, Packet Loss and Packet Reordering. The delay measurements were conducted between the GPS time synchronized Hades[1] measurement nodes at each location. Additionally all relevant counters at the routers have been recorded using a SNMP based Network Manangement Station and supplemented special command line interface output gathering and parsing scripts. The interfaces statistics were stored in 60 second intervals. The aim of the testbed was to demonstrate a failure-free transmission of one or more IP datastreams over 100GE during the whole period of 4 weeks.This included the evaluation of the 100 Gbit/s optical transmission system, the 100GE interfaces between the routers and the optical system, and the evaluation of a sustained 100GE transmission as well as the evaluation of the use of 100GE in a production like environment. The evaluation included a circulated (in a routing loop) tunable load between 1 and 100 Gbit/s, measurement of transmission quality of TCP and UDP datastreams between the endsystems, measurements of one way latency, a ramping up data transmission from approx. 8 Gbit/s up to 96 Gbit/s.

1. Introduction

100GE is a new defined standard at IEEE (802.3ba) [2]. This standard was ratified at 17th of June 2010. Some vendors deployed this new standard before its ratification for testing purposes already in their pre-production products. Besides the German National Research and Education Network (DFN), providing the communications network for science and research in Germany (X-WiN), and the three commercial project partners GasLine, Huawei, Cisco, the Karlsruhe Institute of Technology (KIT) and Forschungszentrum Jülich (FZJ) as academical partners participated in the project.

Karlsruhe Institute of Technology (KIT), a new founded organisation at 1th of October 2009, bundles the missions of both precursory institutions: A university of the state of Baden-Wuerttemberg with teaching and research tasks and a large-scale research institution of the Helmholtz Association conducting program-oriented provident research on behalf of the Federal Republic of Germany. Within these missions, KIT is operating along the three strategic fields of action of research, teaching, and innovation. Whereas the Forschungszentrum (research centre) Jülich pursues cutting-edge interdisciplinary research on solving the grand challenges facing society in the fields of health, energy and the environment, and also information technologies. In combination with its two key competencies – physics and supercomputing – work at Jülich focuses on both long-term, fundamental and multidisciplinary contributions to science and technology as well as on specific technological

applications. Regarding supercomputing FZJ currently holds position 5 in the top 500 HPC list with the IBM JUGENE - Blue Gene/P Solution[3] in September 2009.

GasLINE[4] one major European fiber infrastructure provider with over 9.500 km fiber infrastructure only in Germany supported the 100GE testbed by providing the 447 km fiber infrastructure between the two locations at Karlsruhe and Jülich. For the 100G testbed the standard optical fiber infrastructure of GasLINE with no special specifications has been used. The infrastructure is divided into 7 parts by optical fiber amplifier for regenerating the signal. The motivation of GasLINE is to evaluate their standard infrastructure for the 100GE readiness.

2. 100GE Testbed Design and Commissioning

Huawei delivered to the testbed a preproduction Dense Wavelength Division Multiplexing (DWDM) system. One DWDM system is located at KIT and the other at FZJ. The 100 Gbit/s Wavelength Conversion Board, the transponder board, connects the two DWDM systems through the optical transmission system (447 km distance) with 2 Wavelength each 56 Gbps. As a 100 GE client interface Huawei has certified a Finisar CFP. It is the Finisar's 100GE LR4 CFP optical transceivers which is compliant with the CFP Multi-Source Agreement[5] (MSA) and supports the 100GBASE-LR4 (4x25G) optical interface standardized in the IEEE P802.3ba[6] draft document. CFP is the acronym for „100 Gigabit Small Form Factor Pluggable". The leading C stands for roman 100.

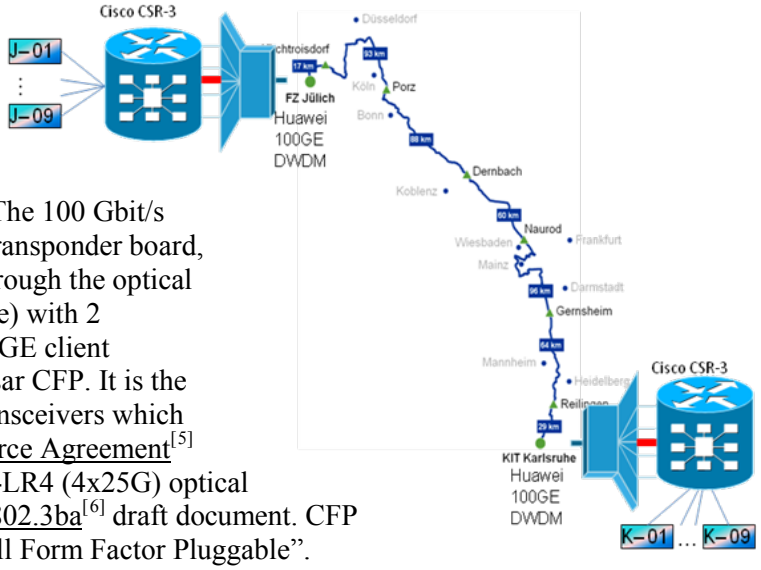

Fig. 1: Topology of 100G Testbed

An FTB 85100G[7] packet blazer of the vendor Exfo has been connected with a Finisar 100GBASE-LR4 (4x25G) CFP to the Finisar CFP with the same specification at the DWDM client site. For certifying the network protocol layer-2 quality of the installation an end to end long term test over 30 hours has been executed. The packets have been sent from the FTB 85100G via the CFP to the Huawei DWDM at Karlsruhe and from there to the DWDM system at Jülich. From Jülich the traffic has been sent back through a hardware loop at the DWDM system to Karlsruhe and further to the client interface of the FTB 85100G. During the 30h test period no CRC error or any packet loss appeared. Besides the readiness of the DWDM system this test has shown as well that the quality of the fiber connection Karlsruhe and Jülich is outstanding.

After the infrastructure up to the DWDM client interface at KIT and FZJ had been installed and tested, the installation of a preproduction 4 slot CSR-3 Cisco router was initiated. CSR-3 provided some new features relevant for the 100GE test. These features included an upgraded backplane capacity with the new 4 Fabric Cards to 140Gbps, a Modular Service Card (MSC) with a capacity of 140Gbps and a Physical Layer Interface Module (PLIM) supporting a 100GE CFP. The Cisco certified CFP 100GBase-LR4 is from the vendor Opnext.
After booting the CSR-3 router successfully at KIT and FZJ, the Cisco router has been attached to the Huawei DWDM.The interfaces were connected, but the interlinks between the DWDM client interface and the router at both locations KIT and FZJ could not be activated. The specifications were checked again, but since the preproduction CFPs of both vendors are 100GBASE-LR4 (4x25G) they should have recognized their lights respectively. At the end a one vendor CFP environment was installed; at KIT an Opnext CFP at the Huawei DWDM client interface and an Opnext CFP at the Cisco router PLIM and at FZJ a Finisar CFP at the Huawei DWDM client interface and a Finisar CFP at the Cisco router PLIM. With this constellation a first light between the Cisco CRS-3 routers could be seen, at the end the 100GE interfaces could be brought up. Both CFP vendors were acutely aware of this issue and fortunately with the now released production version the vendor incompatability has been solved.

3. Monitoring and traffic injector nodes

A traffic injector, ti-kit1 was connected to the router at KIT in order to inject scalable UDP traffic load into a routing loop over the 100G link. The nodes ms-kit/fzj1 have been used for measuring packet loss and packet reordering using iperf UDP data transfers at 5 Gbit/s, and additionally they recorded the RTT every second for the purpose of availablity measurements. The nodes hades-kit1/fzj1 were GPS time synchronized measurement nodes used for permanent measurements of one way delay and its variation. The router interfaces were monitored through SNMP in 60 seconds intervals from a central Network Management Station (NMS) that is not shown in the picture. The NMS stored the results in an extended RRD database with extended round robin time of 6 months.

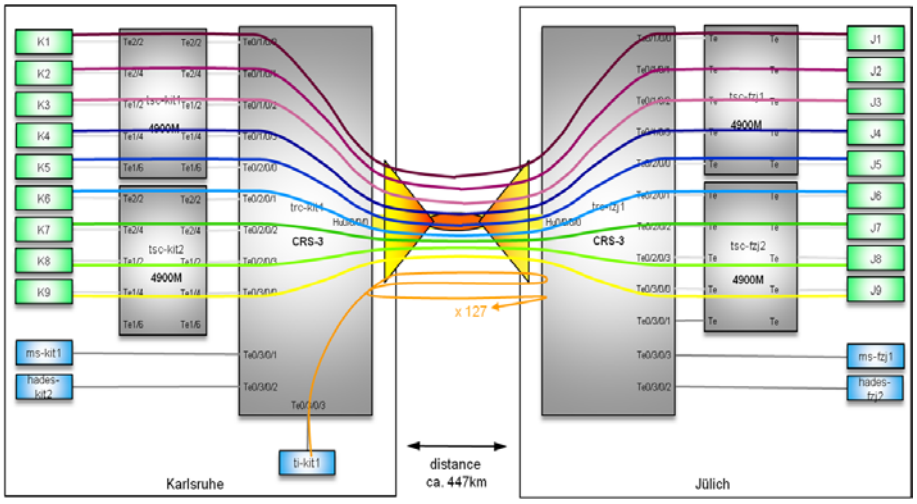

Fig. 2 : Topology incl. data streams (TCP and UDP)

Additionally the NMS recorded all other router information by using the routers command line interface in 5 min intervals, special post processing scripts had been developed for the analysis of these traffic relevant data. The loction of the traffic injector as well as the measurment nodes are shown in Figure 2.

The traffic generating and receiving nodes at KIT and FZJ have not been equal. At both locations 9 nodes with 10GE interfaces each have been dedicated for the testbed. At FZJ nodes of three different hardware specifications have been deployed. All nodes have been based on an Intel architecture with the same Myricom Myri-10G Dual-Protocol NIC (10G-PCIE-8A) network adapter, but with different processors; three nodes with a Dual Core AMD Opteron™ Processor 265 with 1,8GHz, one node with an Intel® Xeon™ CPU 3.00GHz and five nodes with an Intel® Core™ i7 CPU860 @ 2.80GHz. At KIT all nine nodes have been of the same type: Dell™ PowerEdge R510 with Intel® Xeon™ X5560 Processor 2,8GHz and a Broadcom® NetXtreme™ II 57711 PCIe x8 network adapter with dual sfp+ 10GE transceivers. The nodes have been equipped with one 10GE multimode sfp+ transceiver only. Since the supported 10GE interfaces at the CRS-3 have been single mode XenPAK only, a conversion was needed. This media conversion between single mode transceiver and multi mode transceiver was realized with two Cisco Catalyst 4900M at each location. Each of the Catalyst 4900M was populated with 12 ports 10 Gigabit Ethernet (X2). With this configuration each “odd” port could be deployed as a multimode connect to a node and the following “even” port as the corresponding single mode connection to the CRS-3 router.

4. Challenge : increasing error rate

Shortly after the first test was finished, a significant high error rate was observed. A standard ping showed a packet loss of approx. 2%. Since this was not acceptable, an investigation was started immediately. The high error rate appeared first at the “in” interface at the CRS-3 router at FZJ. The high error rate shifted during the investigation to the “in” interface of the CRS-3 at KIT. Further investigation caused a link flapping which could not be resolved for several days. Huawei found and resolved one network issue between the two DWDM systems. One second issue was found at the sender side of the CFP transceiver at CRS-3 at KIT. This transmitter showed a minor bit error, identifiable by a moderate (CRC and “input error”) packet error with a ratio in the order of 10^{-7} at the receiving CFP interface at CRS-3 at FZJ. This error ratio corresponded to an error rate of less than 1 error packet per second (0.8333..) by a full sustained 100 Gbps interface rate, which would not be acceptable for a production environment, but did not have further impacts on test results. (One node with a 10GE interface will be affected only by a loss of a packet every 10 seconds.)

5. First experience - UDP and TCP traffic over 100GE

The initial testplan for the 100G testbed included UDP load on a routing loop over the 100GE link between the two routers. Followed by TCP load generated by iperf client/server processes and a mixture of TCP and UDP traffic with MTU sizes of 1500 byte and 9000 byte were included in the testplan. The roadmap during the actual test period was refined and adapted to the requirements. In the following passages the different tests will be elaborately described as well as the discovered surprises.

For the first test, after the first day of 100GE light between KIT and FZJ, a routing loop was created and UDP traffic was injected by a software load generator running at "ti-kit1". To achieve the explicit setting of the TTL value greater than 64, the linux kernel at "ti-kit1" had to be modified, since the max. TTL value was limited to 64 by default. The TTL value was explicitly set to 254, this kept each packet 127 times in the loop until the counter had been subtracted to 0 and the router dropped the packet. This allowed filling a 100GE link with a traffic injector connected via a 1GE interface only. This behavior can be calculated as follows: MTU (1500 byte) * bit of a byte (8) * (ramping up from 1 to 70.400) packets/sec results in 844.800.000bit or 805.66Mbps. Including the loop (127) in the calculation the 100Gbps are reached closely: 107.289.600.000bps (99,92Gbps). This is, what the theory would imply, but in reality the injected single UDP/IP flow could not grow bigger than 92.2Gbps. The limiting factor is located at the asics of the MSC of the Cisco CRS-3 router. In real life there is currently no single source emitting a load in a single stream close to 100G, but Cisco pledged to eliminate this limitation in a new revision of the MSC which will be launched together with the 100GE PLIM.

The next UDP load generated two UDP flows. The ramp up of the flow started from 0 to as close as possible to a full saturation of the 100GE link (2 * 35200 packets / MTU 1500 byte). The ramping up was realized with an add-on by 0.5Gbps with a duration of 2 minutes per step. The total duration of this test was calculated with approximately 3.5 hours. The resulting traffic stream could saturate the 100GE interface with 99.4Gbps. This interface load was only exceeded by one UDP test with 10 parallel running loops with a total maximum interface load of 99.6Gbps sustained over a period of 2 hours. Between the standard MTU size of 1500 byte per packet and jumbo frame MTU size of 9000 byte per packet no difference regarding the sustained throughput could be seen. By reducing the MTU size to 128 byte, the maximum packet rate was 72.1 million packets per second with a corresponding throughput rate of 87.3Gbps. With the reduced MTU size the error packet rate increased by approx. factor 10 but this did not affect the error ratio, the error ratio remained the same.

The first part of the next test phase was an adjustment of the iperf client and server nodes at KIT and Jülich for bringing them up to speed and identifying the performance of each node. For this a good performing reference node was pointed out at each site and the nodes were separately measured against this reference node. Each end node was configured to deliver the highest possible output performance regarding WAN connectivity. The iperf performance of the nodes varied between 4.9Gbps and 9.8Gbps.
At the second part of the TCP test phase iperf streams were started, synchronized by crontab, from the 9 nodes at each site and at the routers aggregated over the 100GE link. With the 9 streams at FZJ the client (source) and at KIT the server (destination) close to 80Gpbs (79.6Gbps) could be realized constantly. The same nodes did send the 9 streams in the opposite direction from KIT to FZJ. The first 8 streams were started and added their load to the 100GE link. The last (the 9.ed) stream reduced the bandwidth of stream 6 to 8, so that the total bandwidth including the 9.ed stream was equal to the 8 streams before. Since each additional stream was initiated with a slight delay it was possible to identify that the last initiated stream was causing the "reduction". Further investigation showed that it was not down to one of the nodes nor to the 100GE interface. One of the Cisco Catalyst 4900M, used as media converter, had a faulty asic. Fortunately the faulty asic did only affect data flow in the direction from KIT to FZJ.

A red line can be identified at the top of the graphs produced at KIT as well as at FZJ, e.g. Fig. 3. This red line is oscillating up and down (every 5 min.). Looking closer at it, it can be identified that when the graph is increasing up at KIT, the graph is decreasing at FZJ. This line is part of the

ms-[kit/fzj]1 measurement functions identifying unconditional high packet drop rate. This function is not able to send and receive packets at the same time, so packets need to be sent in one direction only. That's why it alters the direction every five minutes from KIT to FZJ and the other way around.

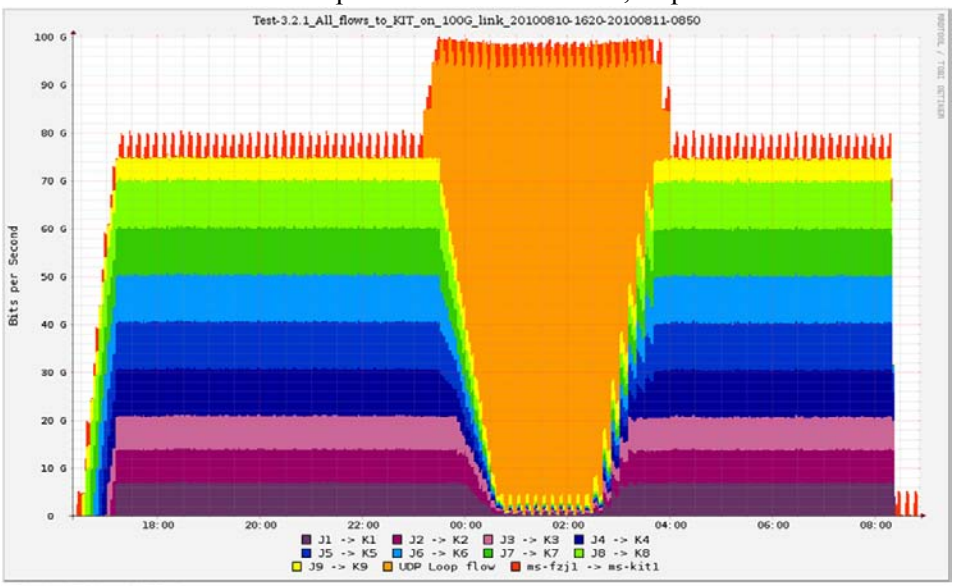

Fig. 3 : TCP and UDP data streams

The next part of the test phase included TCP streams with additional UDP traffic. 9 TCP streams were first initiated. They filled 75Gbps of the 100GE link. Approximately five hours later a UDP stream was added. The UDP stream was ramping up from 0 to close to line speed which took approx. 1.5 hours. Over a period of 2 hours the UDP load was stable. The ramping down afterwards took again a period of 1.5 hours, till only the TCP streams were left. Looking more closely, it can be seen that the UDP stream does not only fill up the remaining capacity, but when getting larger than the remaining capacity, the UDP stream interferes with the TCP streams. A TCP stream does react to an overloaded and congested network. A dropped TCP packet reduces the TCP stream by half (dividing TCP sliding windows[8] by half). The TCP load increases by adding approx.5% to the current capacity but only by successful delivered and acknowledged packets. This indicates a much slower increase than the decrease. The UDP protocol is ignorant to dropped packets, the flow control or with other words the control of correct packet transmission is realized at the application layer. An uncontrolled UDP stream can force TCP flows to reduce their capacity. This behavior is demonstrated at Fig. 3. The UDP traffic is forcing the TCP streams not only to shrink, but it forced the TCP streams more or less to diminish to a very negligible something very close to "0" for the 2 hours the UDP traffic was injected with full capacity. After the 2 hours, the UDP stream injection was reduced, the TCP streams increased to 75 Gbps. Even while the load at the 100GE interface at the router was exceeded there were no additional packet errors, CRC, or packet drops at the router interfaces recognized.

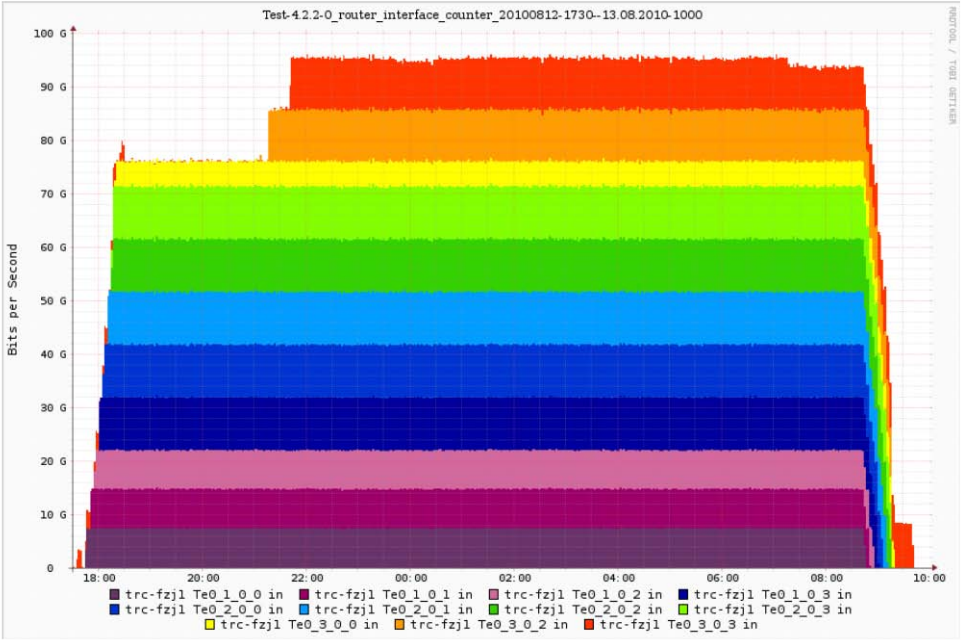

Fig-4: 11 TCP streams

The last test part was designed after the active test period was already started. Even if each of the nine nodes could deliver network traffic of up to 9.5 Gbps, it would not be enough to fill a 100GE link. The situation in the testbed was even worse: the capacity of the nodes was varying widely (zam404: 6 Gbit/s, zam405: 7.3

Gbit/s, zam406: 7.9 Gbit/s, zam770: 9.8 Gbit/s, zam771: 9.3 Gbit/s, zam772: 9.3 Gbit/s, zam773: 8.3 Gbit/s, zam774: 9.2 Gbit/s, zam669: 4.9 Gbit/s). Taking this together it sums up to approximately 75 Gbps. At KIT and FZJ two test nodes of DFN, the "ms-[kit/fzj]1" and "hades-[kit/fzj]2" were installed. These two nodes were configured as iperf client/server nodes additionally, so that their traffic would be added to the 100GE load. The client/server processes of the 9 original test nodes were initiated first. After all streams were running, a stable load was recognized over the 100GE link of about 75.4 Gpbs. 2.5 hours later the streams of the 2 DFN nodes were started. The first additional data stream was exchanged between the two nodes hades-[kit/fzj]2 with a constant data stream of about 9.7 Gbps. The second data stream between the nodes ms-[kit/fzj]1 was not stable. This stream was varying between 8 and 9 Gbps. The reason in this case was not the capability of the node, but a sensitive reaction of both nodes to packet losses. The varying at the last stream at the sensitive node of packet losses was indicating that with 95 Gbps the maximum capacity of the link was reached and the bandwidth was utilized. Even under this condition no additional packet loss or any interface or protocol error could be discovered.

6. Résumé:

After solving the first difficulties during the startup period of the 100G testbed a solid basis for the tests to be executed had been established. The fiber infrastructure offered by GasLINE provided a high quality and turned out to be definitely 100G ready. Realizing a single vendor CFP environment between the Huawei client CFP and the 100 GE CFP PLIM interface at the Cisco CRS-3 at KIT and FZJ the 100 GE link could be brought up without problems. Both CFP vendors were informed about this interoperability issue and both, Finisar and Opnext, promised to solve this issue before the first CFP will be launched at the market.
Since the 100G testbed was mainly equipped with preproduction equipment and only one of the CFP vendors were certified by Cisco and Huawei, developers of both vendors supported the installation period. After first light was initiated and seen between KIT and FZJ and all interfaces were activated, a quite stable test environment could be established, and only at specific points further intervention of the developers was required. The source of slight errors at the receiver interface at the CRS-3 router at FZJ could be identified. Through additional tests it could be certified that this error did not disturb the executed tests nor did it introduce inaccuracy in the test results.

As a result of this testbed it can be summarized, that the tested fiber infrastructure, which is of the same kind as the fibers used in the X-Win production network, as well as the used 100GE components of the participating vendors are well prepared to be used in a future 100G X-WiN environment.

[1] **Hades Active Delay Evaluation System (HADES)**
http://kb.pert.geant.net/PERTKB/HadesTool
[2] **Ratification of IEEE 802.3ba standard of 40Gb/s and 100Gb/s Ethernet**
http://standards.ieee.org/news/2010/ratification8023ba.html
[3] **HPC TOP500 List of 2009**
http://www.top500.org/system/9899
[4] **GasLINE Telekommunikationsnetzgesellschaft deutscher Gasversorgungsunternehmen mbH & Co. KG**
http://www.gasline.de/gasline-english/home.htm
[5] **CFP MSA (Muti Source Agreement)**
http://www.cfp-msa.org/index.html
[6] **IEEE 802.3ba standard of 40Gb/s and 100Gb/s Ethernet**
https://sbwsweb.ieee.org/ecustomercme_enu/start.swe?SWECmd=GotoView&src=0&Join=n&SWEView=Catalog+View+%28eSales%29_Main_JournalMags_IEEE&mem_type=Customer&HideNew=N&SWEHo=sbwsweb.ieee.org&SWETS=1292187620
[7] **EXFO -- 100G/40G Ethernet Test Module - FTB-85100G Packet Blazer**
http://www.exfo.com/en/products/Products.aspx?Id=440
[8] **TRANSMISSION CONTROL PROTOCOL -- RFC: 793**
http://tools.ietf.org/html/rfc793