FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Technical Report

The DEISA Project's 10 Gb/s network infrastructure

Ralph Niederberger, Olaf Mextorf

FZJ-ZAM-IB-2007-08

July 2007

(last change: 19.07.2007)

The DEISA Project's 10 Gb/s network infrastructure

A design, management and operation overview

Ralph Niederberger, Olaf Mextorf
FZJ – Forschungszentrum Jülich GmbH, D - 52425 Jülich, Germany
e-mail: R.Niederberger@fz-juelich.de, O.Mextorf@fz-juelich.de

Abstract

The DEISA[1] [1] (Distributed European Infrastructure for Supercomputer Applications) project has been started in 2004 by eight leading national supercomputing centres in Europe. In 2005 the three additional sites BSC, LRZ and HLRS joined the DEISA consortium. DEISA currently deploys and operates a persistent, production quality, distributed supercomputing environment with continental scope that enables scientific discovery across a broad spectrum of science and technology, by enhancing and reinforcing European capabilities in the area of high performance computing. The existing national high-end platforms have been tightly coupled by a dedicated network. Innovative system and grid software allows easy access to all the different kinds of supercomputer architectures installed.

The currently used network infrastructure has evolved in several steps. Starting with virtually dedicated network interconnects (GÉANT IP Premium service [2][3]) in 2004 connecting four homogeneous[2] national supercomputers, to provide a distributed supercomputing platform operating in multi-cluster mode, in 2005 and 2006 most of the other sites have been connected to this 1 Gb/s network infrastructure. Within autumn 2005 a new network infrastructure was designed, which should allow a 10 Gb/s connectivity between all DEISA supercomputer systems. In mid of 2006 first contracts could be signed to implement a "proof of concept" DEISA network connecting the five supercomputer systems at FZJ, IDRIS, LRZ, RZG, and SARA with 10 Gb/s links. In a next step other sites have been connected to the DEISA 10 Gb/ backbone also. The first part of the paper describes the 10 Gb/s DEISA backbone design and management structure. The second part of the paper provides an overview about lessons learned from the network point of view, throughput values, operational issues and application experiences.

Keywords:

"European distributed multi-cluster supercomputer", "Grid infrastructure", "GÉANT2", "dedicated wavelength at an European scale", "large scale lambda setup"

Introduction

The European DEISA Integrated Infrastructure Initiative (I3) [4] is an infrastructure of infrastructures that devised an innovative strategy to enable the cooperative operation of existing national supercomputing infrastructures. This initiative led to the deployment and operation of a world class, persistent, production quality, distributed tera-scale supercomputing environment with continental scope enabling scientific discovery across a broad spectrum of science and technology.

---

[1] Funded in part by the European Commission under grant 508830, Consortium members: Barcelona Supercomputing Center (BSC), Barcelona, Spain; Consorzio Interuniversitario (CINECA), Bologna, Italy; Finnish Information Technology Centre for Science (CSC), Espoo, Finland; European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK; Edinburgh Parallel Computing Centre (EPCC), Edinburgh, UK; Institut du Développement et des Ressources en Informatique Scientifique (IDRIS-CNRS), Orsay, France; Forschungszentrum Jülich (FZJ), Jülich, Germany; High Performance Computing Center Stuttgart (HLRS), Stuttgart, Germany; Leibniz Computing Centre of the Bavarian Academy of Sciences and Humanities (LRZ), Garching, Germany; Rechenzentrum Garching of the Max Planck Society (RZG) , Garching, Germany; SARA Computing and Networking Services , Amsterdam, The Netherlands

[2] with respect to architecture and system software

To enable distributed computing there is a strong need for network connectivity between the involved eleven leading European High Performance Computing (HPC) centers in Europe with guaranteed capacity between the DEISA supercomputer systems. The DEISA network connectivity of the first phase [5] has been based on 1 Gb/s routed IP and MPLS tunnels and involved nine supercomputing centres in Europe (all DEISA sites except EPCC and HLRS). Connectivity was provided by direct links from DEISA sites to the local National Research and Education Networks (NREN). These have been interconnected by standard links to the European GÉANT network.

In order to scale up the capacity and the number of connected DEISA sites a "proof of concept" phase started in November 2006 in which five DEISA sites should be connected with 10 Gb/s to the DEISA backbone. This optical private network based on 10 Gigabit Ethernet (10GE) connections provided by the involved NRENs and GÉANT2 [6] has been designed to allow a future integration of the other DEISA sites with 10 Gb/s communication speed, too. Beginning in December 2006 the other DEISA locations have been connected to this backbone consecutively.

A DEISA project overview

Right from the start of the DEISA project the consortium followed a top-bottom strategic approach for the deployment and the evolution of the European tera-scale supercomputing environment. Only some very basic strategic requirements have influenced the deployment of the current infrastructure. First of all the necessity of fast deployment of the infrastructure has been essential. Also the coexistence of the European infrastructure with the national services had to be guaranteed, which require reliability and non-disruptive behavior. A third prerequisite has been user and application transparency, hiding complex grid technologies from users and minimizing application changes, because application development should not be strongly tied to an IT infrastructure. Designing an European supercomputing infrastructure cannot be done from scratch. All sites participating in DEISA had already installed a local supercomputing system before DEISA startup, which had to be integrated smoothly into the new infrastructure. Moreover having a project duration of four years it was quite clear from the beginning that there would be additional systems coming up to be integrated and older systems to be scrapped. Therefore DEISA decided to have a multi-layer infrastructure model.

The current architecture of the DEISA supercomputing grid meets these concerns by having an inner level, dealing with the deep integration and strongly coupled operation of similar, homogeneous IBM AIX clusters, which form a *distributed European supercomputer*.

An outer level of non-IBM AIX-cluster systems completes the DEISA infrastructure. This grid of supercomputers and super-clusters was tightened together in a looser federation of heterogeneous supercomputing resources. Currently this supercomputing grid includes all the leading platforms in Europe including systems from IBM, SGI, NEC and Cray. A third level of integration will be visible in the future when external resources (compute, storage, databases, data generation resources as telescopes, medical devices etc.) may be loosely connected to the existing DEISA infrastructure.

The DEISA partners contribute a significant amount of their national supercomputing resources (of the order of 10% or more) to a globally managed European resource pool. The leading supercomputing platforms in Europe participating to the DEISA resource pool (from March 2007) are:

- FZJ-Jülich (Germany): P690 (32 processor nodes) architecture, incorporating 1312 processors. Peak performance is 8.9 Teraflops.
- IDRIS-CNRS (France): Mixed P60 and P655+ (4 processor nodes) architecture, incorporating 1024 processors. Peak performance is 6.7 Teraflops.
- RZG–Garching (Germany): P690 architecture with 896 processors. Peak performance 4.6 Teraflops.
- CINECA (Italy): P690 architecture incorporating 512 processors. Peak performance is 2.6 Teraflops.
- CSC (Finland). P P690 architecture incorporating 512 processors. Peak performance is 2.2 Teraflops.
- BSC (Barcelona, Spain): 10240 processors IBM PowerPC with Linux, peak performance 94 Teraflops.
- HLRS (Germany): NEC SX8 vector supercomputer, 576 processors, 9.2 TB memory space, 180 TB disk space. Peak performance is 12.67 Teraflops.
- LRZ (Germany): SGI Altix, 4096 processors and 26,2 Teraflops peak (> 60 Teraflops end of 2007)
- SARA (The Netherlands): SGI Altix 3700 Linux system, 416 Itanium-2 processors, 832 GB main memory, peak performance 2.2 teraflops.

- ECMWF (International organization): Two 690+ clusters with 68 32-way nodes each, with an aggregated performance of 33 Teraflops peak.

The HPCx system connected via EPCC, UK is planned to be integrated in May 2007 also.

The IBM AIX systems listed above are running IBM's GPFS (Global Parallel File System, [7]) as a cluster file system. IBM has incorporated wide area network functionality in GPFS, enabling the deployment of *distributed* global file systems. This is the basic integration technology of the AIX super-cluster currently used.

An application running on one site can access data files previously "exported" from other sites as if they were local files. Therefore, it does not matter in which site the application is executed, and applications can be moved across sites transparently to the user. Though the concept of deploying network file systems is an old one, GPFS provides high performance remote access needed for high performance computing. In 2005 the TeraGrid project [8] in the United States has shown that on their 30 Gb/s TeraGrid network in the USA, GPFS is able to achieve about 27 Gb/s network throughput when accessing remote data via this file system. This proved that the software is capable of taking full advantage of underlying high performance networks. DEISA, which can be seen as a counterpart to TeraGrid, intended from the beginning to offer such a kind of high speed interconnect to applications like GPFS. At the Supercomputing 2005 conference at Seattle an interaction between both projects, DEISA and TeraGrid, could be initiated which allowed seamless transparent access to data between Europe and the United States without any user interaction [9]. The achieved throughput was only dependent on the number of installed I/O servers and requesting client applications.

The DEISA project which started in 2004 had the mission to provide visionary leadership in the area of high performance computing in Europe. After three years of successful operation the distributed terascale supercomputing facility, which had an aggregated power of close to 30 teraflops in early 2005, provides more than 200 teraflops aggregated CPU power in March 2007 with more than 23.000 processors. The principal objective of the project has been to advance computational science in leading scientific and industrial disciplines by deploying an innovative Grid-empowered infrastructure to enhance and reinforce High Performance Computing in Europe. This super-cluster with the appropriate software and the dedicated high speed networks, represents a way, open in the immediate future, for innovative and creative thinking that enhances the impact of existing infrastructures. The integration of this distributed multi-cluster platform into the larger worldwide Grid infrastructure, third level or shell mentioned above, has already been started and will be extended further more in the near future.

Leading scientists across Europe are using the current supercomputing infrastructures with so called grand challenge applications from scientific key areas like material sciences, climate research, astrophysics, life sciences and fusion oriented energy research. In 2005 the focus of DEISA has been enhanced by defining the DEISA Extreme Computing initiative [10]. In 2005/2006 about 29 DECI applications have been in production many of them requiring more than 1 Million CPU hours. In 2006/2007 additional 23 applications could be supported having additional applications in stand by position if some of the others get into delay. The DEISA research infrastructure will also be open, under certain conditions, to users of non-member organisations.

DEISA collaborates with a large number of projects and institutions in Europe. The first priority is the cooperation with other FP6 infrastructure projects in HPC or Grids especially HPC-Europe [11] and EGEE [12]. Last but not least a close collaboration with GÉANT2 has been the driving force for the evolution of the second phase DEISA network infrastructure.

The DEISA network backbone design

At start of the DEISA project in 2004 it has been very expensive to install a 10 Gb/s network infrastructure across Europe. NRENs and GÉANT have not been capable of technically providing cheap links across national boundaries. Therefore the DEISA consortium decided to start with a 1 Gb/s network infrastructure using "virtually dedicated" links between DEISA sites. Being aware that a fully-meshed topology involving point-to-point connections across all platforms can hardly be considered, because of involved costs, GÉANT proposed the use of a PREMIUM IP service provided via the normal GÉANT infrastructure. This design led to a full meshed 1 Gb/s network infrastructure between involved NRENs of participating DEISA sites. DEISA started the implementation as a "proof of concept" phase including only four sites to deeply analyse the communication behaviour and to prepare a full mesh to all sites in future. A multilayer switch had been installed at every DEISA site to which the local supercomputer system had been connected by one or more Gigabit-Ethernet interfaces

(Etherchannel). With this design and using the Premium IP service an upper-bounded one-way delay, upper-bounded Instantaneous Packet Delay Variation (IPDV), no packet loss due to congestion and guaranteed capacity as well as a throughput capacity of 1 Gb/s could be guaranteed. After careful analysis the "proof of concept" phase could be completed successfully and a 1 Gb/s network infrastructure to the remaining sites has been established. The network phase 1 is now working for two years in a stable production status without larger disruptions and failures. The motivation for implementing a "dedicated" network infrastructure for DEISA has been discussed in an earlier Terena paper in detail [5]. As well the security impacts using a "virtual dedicated" network across NRENs and GÉANT have been discussed here.

In July 2005 a first DEISA-GÉANT workshop was held in Munich to discuss a 10 Gb/s network infrastructure for DEISA. Here a small working group had been initiated which considered design, backup scenarios, and scenarios for potential future extensions concerning bandwidth and number of participants.

In the beginning of 2006 it became technically feasible to start negotiations with all the providers involved. The negotiations have been facilitated because of the availability of the new GÉANT2 infrastructure allowing multiple wavelengths across the installed GÉANT2 footprint. The DWDM design of the European infrastructure provided the ability to have additional wavelengths without major additional technical efforts. It became obvious to design the new DEISA network in a star like fashion with a central DEISA multilayer switch located somewhere central in Europe. Contracts could be signed with the German NREN DFN to provide a central placement of the DEISA-10GE-switch at a housing company at Frankfurt, Germany, where DFN- and GEANT2-PoPs were already established, because four of the DEISA partners are located in Germany. Leaf layer 2/3 switches have been installed at all DEISA sites. A direct connectivity between these switches, that is wavelength across NRENs and GÉANT2 DWDM equipment, offers short delays, simple management and low total cost of ownership. A Cisco Catalyst 6500 switch has been chosen as heart of the DEISA network infrastructure. The installed network links are 10 Gb/s Ethernet. The 10GE-interfaces which have XENPACK-10GB-LR optics at Frankfurt have been configured with FlowControl and JumboFrames enabled. They do not issue nor accept BPDUs. The links between central switch and leafs use private network addresses as defined in RFC 1918 [13]. An uplink to the worldwide INTERNET is currently not configured, because of the private DEISA network infrastructure usage, but it is intended to interconnect the central DEISA switch to the European Grid infrastructure in future. Access lists installed on every interface of the central switch allow a secure network environment. Any DEISA site is open to install additional firewall(s) or packet filters at their local site to ensure enhanced security requirements. Figure 1 depicts the current network infrastructure. The remaining sites CINECA, CSC and EPCC which currently use the old 1 Gb/s phase 1 infrastructure because of infrastructural and technical delays will be connected within the next three month. This will complete the phase 2 DEISA network installation.

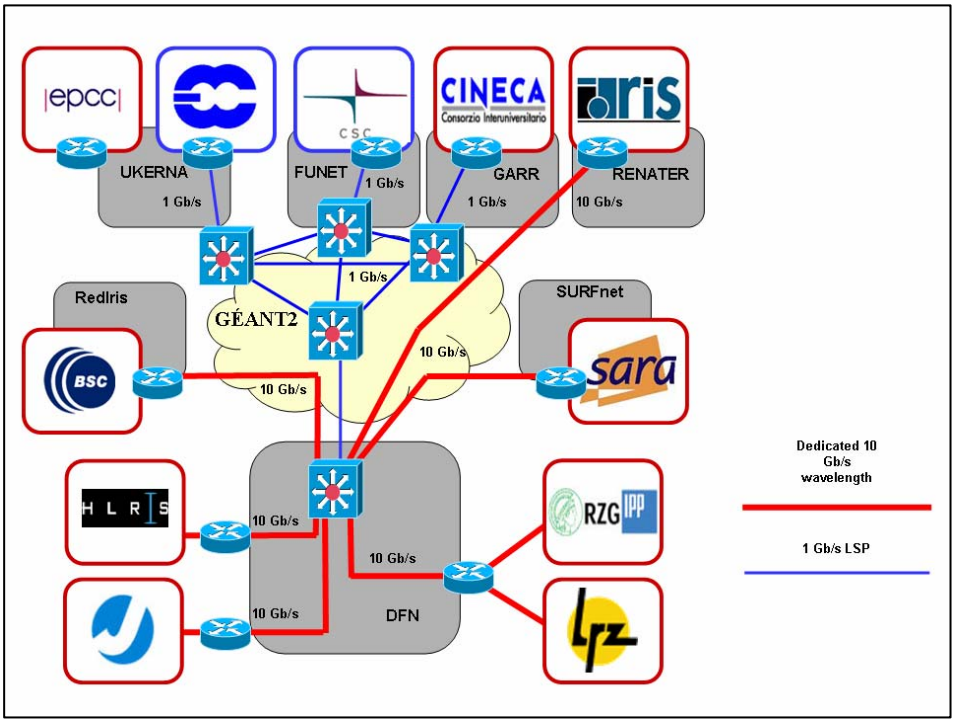

Figure 1: DEISA Network (technical overview April 2007)

Backup scenarios have been considered but not yet installed via cross boarder fibres between neighbouring NRENs. A backup scenario using GÉANT2 fibre footprint is under discussion also. A backup path being essential for most of the production Grid infrastructures today has not been of first priority in DEISA until now. The partner supercomputer systems are working well as stand alone systems also, if a network link is not available. They simply don't schedule jobs which require remote data if the necessary links are down.

Nevertheless in future Grid environments computer systems will be tied together even more so that backup paths become necessary. Therefore it is planned to redesign the central DEISA network and install potentially three switches connected by a high performance ring infrastructure across Europe (30-40 Gb/s) and connect the DEISA sites to two of these by cross boarder fibres (dual homing). This allows easy rerouting in case of link disruption because of parallel connectivity. This design would also enhance overall network bandwidth between sites by load balancing traffic or splitting traffic dependent on destination. A schematic overview of the future backup design is depicted in figure 2.

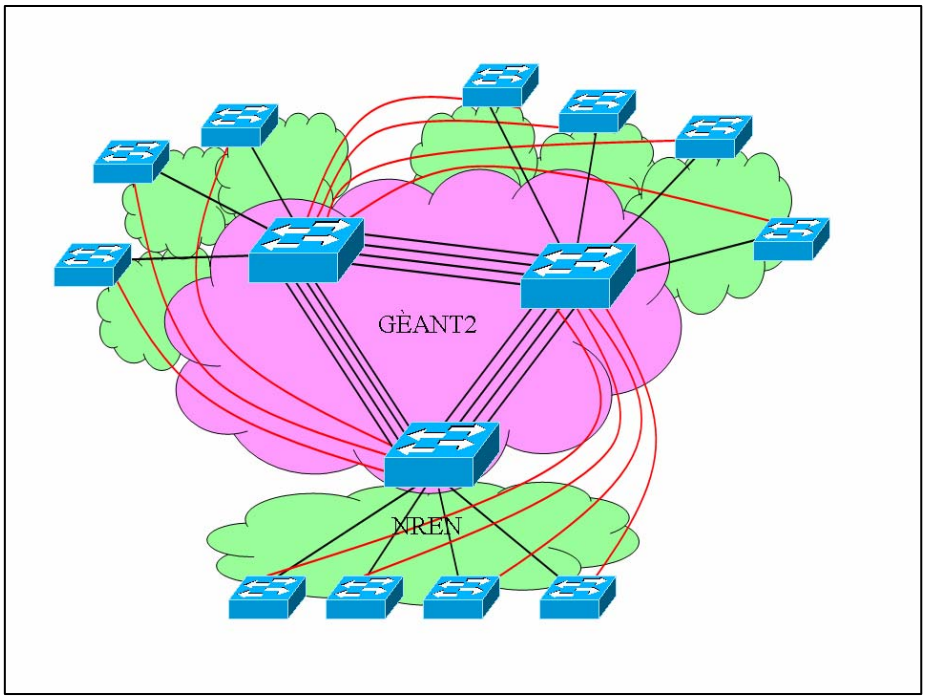

Figure 2: A schematic view of a possible future DEISA Backup Configuration

Managing a 10 Gb/s European project network infrastructure

The management of a dedicated private 10 Gb/s WAN infrastructure connecting supercomputers across Europe does not differ much from general LAN management. The main difference comes from the fact that multiple local organisations, NRENs, national fibre providers and international network providers are involved. Therefore a multi domain network management system has to be developed, that allows to operate the infrastructure in a 24/7 fashion, to identify errors, to provide solutions and to operate an advisory service from the network point of view. Integrating the DEISA backbone into the normal network management procedures is a permanent activity. Though the infrastructure can be seen as only another small sub network of the local network that has to be monitored, the WAN character of this environment requires additional activities. In a local environment a link interruption can be handled by controlling cables and switch interfaces, changing equipment and retesting functionality. In principle this is the same for an European WAN. The difference arises when the error is located somewhere within the infrastructure. A fibre cable may have been cut on the way from Barcelona to Frankfurt. Because of the "simple" nature of the DEISA network to be managed, consisting of one central switch and only leafs at the DEISA partner sites, the intermediate equipment will be hidden. There are many fibre sections in between where the problem could be located. Patch panels at NREN sites, GÉANT and local sites, WAN DWDM equipment including intermediate wavelength amplifiers may be the source of the problem. This implies involvement of many staff members of different organizations (local site, cable provider, lower layer service providers, NRENs, GÉANT2, DEISA network management staff) acting together for problem solution. This personal communication infrastructure has to be organized, centralized, managed and updated continuously.

Beneath normal network monitoring based on standards like SNMP, RMON, Netflow/ipfix and syslog providing information about network utilization, logging, accounting, alarming, error recovery and problem determination, active probing procedures based on IPERF [14] and ping have been installed to log available TCP and UDP throughput dependent on system and network load in a 24 hours 7 days a week fashion. A special web site has been set up which provides this information to DEISA administrators and users. As argued above a close collaboration between the DEISA network team and staff members of the NRENs and GEANT2 as well as a close interaction with the administrators of the supercomputer systems guarantees optimum performance of the DEISA network to meet the needs of the user communities. A close interaction with the DEISA "Operation team", responsible for any software or hardware changes within DEISA as well as any issues related to an undisturbed operation of the supercomputer systems, assures the production quality of the overall DEISA environment.

Operation experiences, Throughput values, et al.

The installation, configuration and operation of a European WAN will not happen without any pitfalls, surprises and new experiences. The lessons learned, local network and system configuration parameters used, throughput values seen between benchmark applications from the network point of view (Iperf) and real user applications (GPFS, GridFTP, …) [15][16] as well as user acceptance and usability will be of overall importance for other organisations planning to install similar infrastructures. It has not been expected to measure a 10 Gb/s throughput for a single stream application. But the question arose what the overall aggregated performance of the underlying network will be and how a single flow/application can be optimized.

Optimized network throughput can be achieved by applications only if the systems have been configured accordingly. A good source for configuration information can be found at the "advanced networking" web page of the Pittsburgh Supercomputing Center [17]. Mathis et al. provide detailed information which networking parameters to alter and adjust. Unfortunately these setting change from OS version to OS version. Though most of them are obvious to be altered, others depend on settings at communication partner's site. They will have an effect only if the corresponding site changes parameters also. Others depend on remote setting in the way that they increase performance if also set on partners site and degrade performance if not. Since many of these parameters can be set only system wide and not interface specific they influence also local communications. E.g. high default buffer settings allow a large bandwidth delay product, but are inefficient if thousands of local connections have to be initiated. The values proposed within DEISA environment are shown in figure 3.

ipqmaxlen=2048	tcp_ecn=0	udp_pmtu_discover=1
rfc1323=1	tcp_mssdflt=1460	udp_recvspace=655360
rfc2414=1	tcp_newreno=1	udp_sendspace=655360
sack=1	tcp_nodelayack=0	use_isno=0
sb_max=20971520	tcp_pmtu_discover=1	
thewall=1572864	tcp_recvspace=2621440	
	tcp_sendspace=2621440	

Figure 3: Optimised network options used at the DEISA supercomputer sites

After having installed and tuned the DEISA 10 Gb/s network we measured performance of the underlying infrastructure. It has been obvious that we will not see throughput values measured in an IBM laboratory test environment which has been about 5.6 Gb/s with an 10 Gb/s interface (10GBaseLR, Typ 7040-5719) in an P5-system with RIO3 drawers. The performance results have shown that throughput varies dependent on operating system version and hardware used, local configurations and load of the system (we are running in production). So we have seen at FZJ about 4.8 Gb/s IPERF throughput to localhost which implies only using internal CPU and no network hardware, 4.3 Gb/s between 2 nodes of the FZJ system using the supercomputer internal IBM Federation switch and about 3 Gb/s between 2 nodes of the FZJ system using the 10 Gb/s GE interfaces connected by a CISCO Catalyst 6500 switch. Similar results can be measured at RZG, but 20-25 % less, which arises from slower CPUs within the RZG systems. These values can also be verified by testing IPERF throughput between a node in FZJ and another in RZG, both connected by a 10 Gb/s Ethernet interface to the DEISA-network. Throughput values reach up to 2.4 Gb/s. Here we started the client at FZJ and the server at RZG with the following commands:

Client: root@fzj::>iperf -i1 -w16000K -f m -c rzg

Server: root@rzg::>iperf –w16000k –s –D

System buffers have been aligned accordingly and jumbo frames have been used. The bandwidth delay product requires about 10 Gb/s * 13 ms ≈16 MB. As mentioned before the systems used for these tests have been in production, so the measured throughput values varied dependent on system load. Sometimes intermediate values of about 3.2 Gb/s could be seen checking throughput in 1 second intervals using the "-i1" option of IPERF.

Of course those performance values may not be seen in "real" applications. The standard applications mostly used within DEISA are GPFS and GridFTP. Both of them allow use of parallel streams to optimize throughput. Using 4 parallel sessions with GridFTP a file can be transferred with a throughput of about 2.7 Gb/s from local FZJ site's /etc/dev0 to remote RZG site's /dev/null. Using /etc/dev0 and /etc/dev/null bypasses any delays because of disc access.

Network Design and its influence on DEISA applications

Several heterogeneous supercomputer systems with different operating systems and system internal network configurations and various LAN designs are used within the DEISA backbone infrastructure. The local connectivity to the DEISA edge switches may differ also. I/O nodes are connected with one 1 Gb/s or multiple 1 Gb/s (channelling) or one 10 Gb/s connection to DEISA. Dependent on these local network designs different throughput values will be measured, because the involved nodes are providing different services and are therefore strained with potential additional work load. The following figures describe some of these scenarios to be found within DEISA:

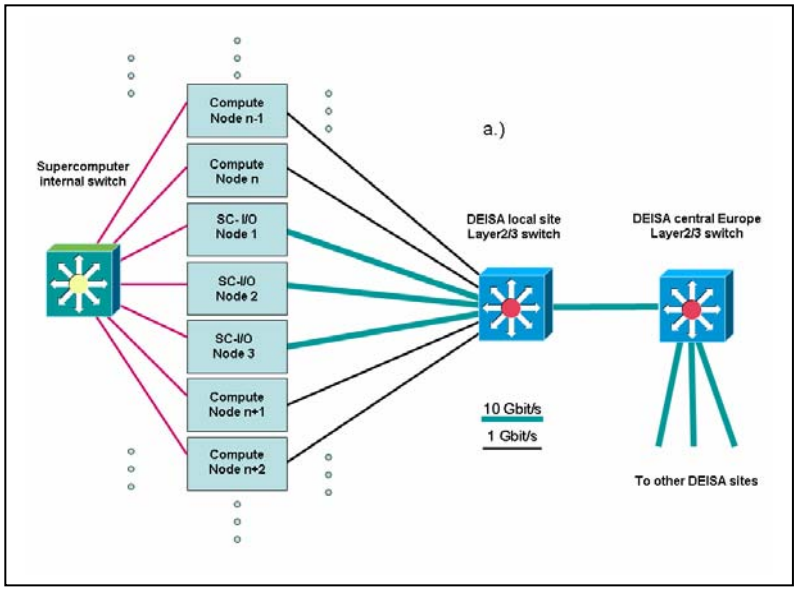

Figure 4 a

- Figure 4 a.) describes a setup with compute nodes connected by 1 Gb/s and I/O nodes connected by 10 Gb/s for optimal access to data. The internal supercomputer switch is used for internal communications only. The setup allows many connections to be started in parallel with 1 Gb/s each. Whether a remote application can access the data stored at the I/O nodes with more than 1 Gb/s depends on the remote supercomputer network configuration.

- Within Figure 4 b.) all nodes of the supercomputer system are connected with 1 Gb/s to the 10 Gb/s network infrastructure. The DEISA backbone allows many 1 Gb/s streams to cross the network in parallel. No stream can be initiated which exceeds the local 1 Gb/s link bottleneck. The internal network will not be used for DEISA purposes.

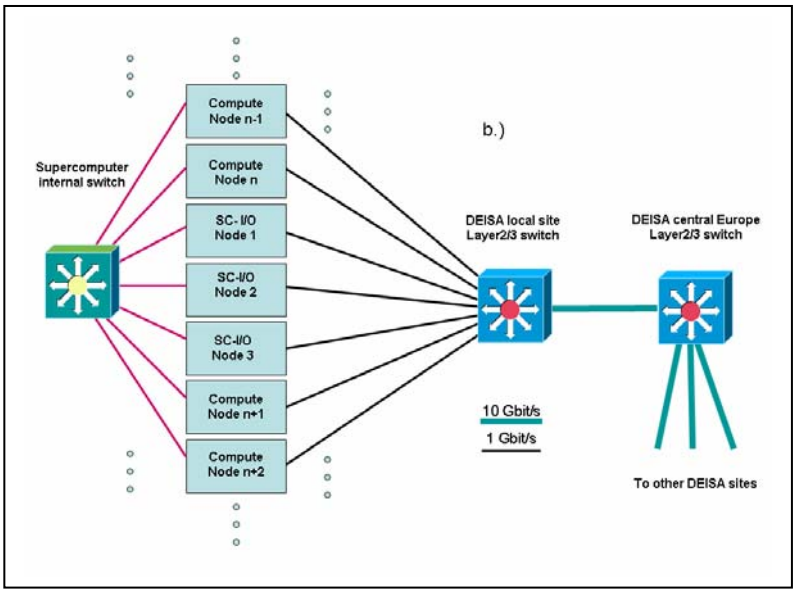

Figure 4 b

- Figure 4 c.) describes a scenario where a special node provides a 10 Gb/s link from the supercomputer to the DEISA backbone working as a gateway node. All communications from I/O nodes and compute nodes must traverse the gateway node via the internal supercomputer network. This setup allows an easy integration of the supercomputer system into the DEISA infrastructure and potentially enables a single 10 Gb/s application stream from every node to the DEISA backbone, but is highly dependent on the power of the gateway system, acting as a "software" router. At FZJ we saw a P4 node consuming approximately one CPU per 1 Gb/s forwarding from internal Federation network to 10GE. That rate was not linear scaling to higher throughput and didn't reach the measured IPERF bandwidth numbers for 10GE usage without forwarding.

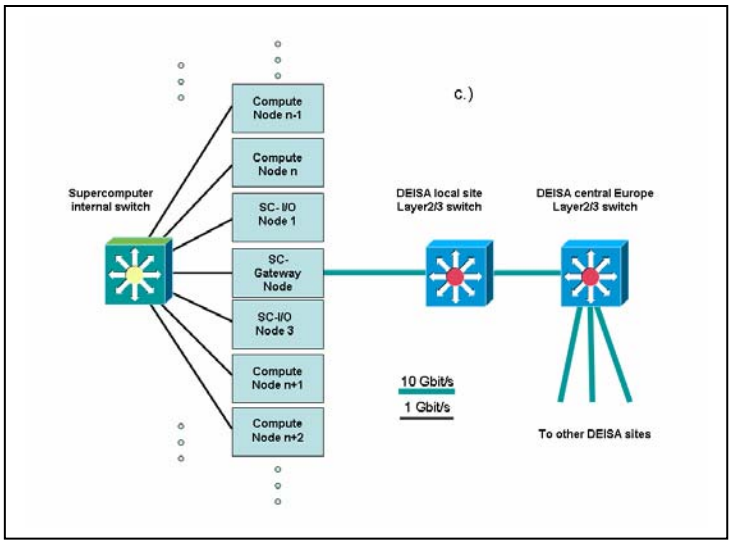

Figure 4 c

The scenario in Figure 4 d.) differs from 4 c.) in the way that there is no special gateway node. A FrondEnd or Login node provides access to the supercomputer system. The same node will also be used for none DEISA purposes. Being the cheapest and simplest connectivity setup it implies disadvantages because of the heavy load which will be generated at the login node. Often this node will be also used for interactive purposes, so that users will not be really satisfied because of overload of the login node. Depending on the configuration this setup can be used to only access the system (really FrontEnd) which implies the supercomputer compute nodes are not directly reachable from DEISA or the FrontEnd node also provides gateway functionality giving access to I/O and/or compute nodes.

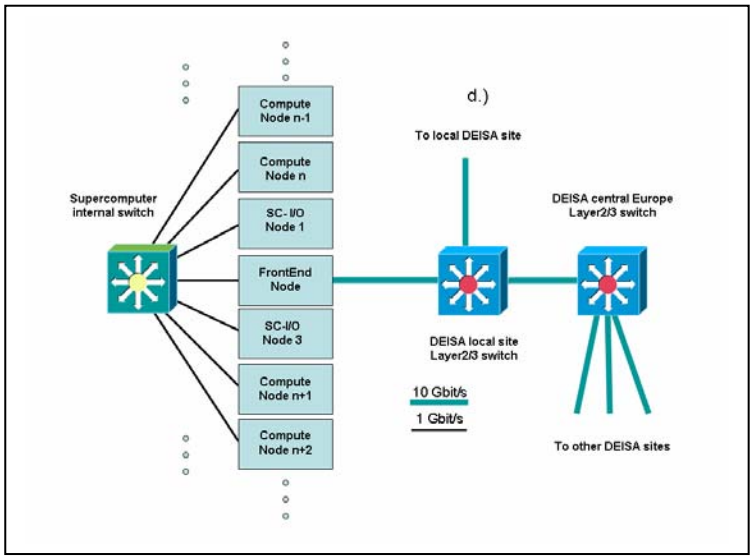

Figure 4 d

A more detailed technical view about an exemplary lambda setup at FZJ is shown in figure 5.). Here 2 I/O nodes are connected to the FZJ local DEISA leaf switch with 10 Gb/s. A third node is configured as Gateway node and connected with 10 Gb/s also to DEISA. All other supercomputer nodes, named compute nodes are connected by 1 Gb/s Ethernet interfaces. For test purposes these nodes can be configured to directly communicate to DEISA or alternatively use the 10 Gb/s gateway node. Any mixed setup is also possible only dependent on local compute node configuration and FZJ local DEISA leaf switch setup. The DEISA leaf switch has one 10 Gb/s Long Range Interface which is connected to the local NREN DFN DWDM equipment. Using DFN's national fibre footprint one special, meanwhile optical protected, wavelength is reserved for DEISA. This wavelength is terminated at the DEISA central switch in Frankfurt, Germany. All other sites have a similar setup. The only difference for non German sites is that their wavelengths are transparently prolonged through the European GÉANT2 network which has been sketched in the right lower corner of Figure 5.). Figure 5.) doesn't show the normal links to the local FZJ network where all the other FZJ hosts are located, nor the "normal" Internet connection of FZJ, which can be access via the login node at the left lower corner of this picture.

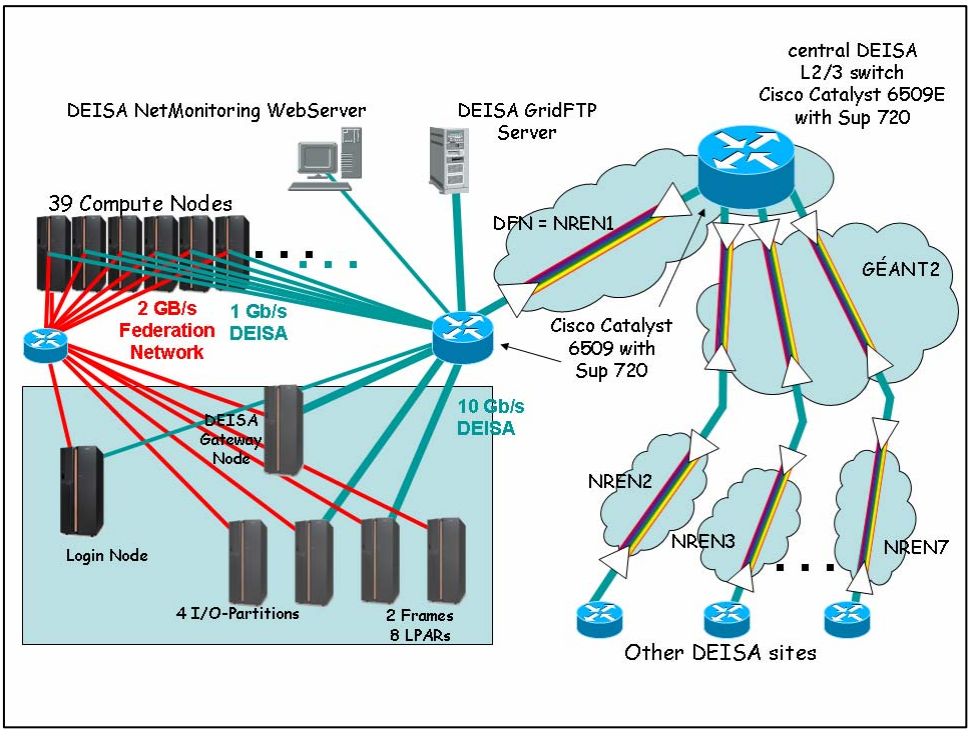

Figure 5: Schematic view of DEISA lambda setup

Lessons learned

The operation of an infrastructure like DEISA leads to new management problems not seen before. Managing a supercomputer system or a number of locally installed cluster systems differs heavily from an European supercomputer infrastructure where staff members dealing with the same problem are thousands of miles away. There is no short cut, going to the office next door, just checking if we agree on some option settings within a software component. Within a virtual organization every small modification has to be check by all partners over and over again. Installing new software components requires checking with all participants, if any dependencies exist. Scheduling of tasks, installations, maintenance, network infrastructure changes and others have to be agreed on. Often a task needs much more of time than estimated. Someone has to deal with those issues.

Though all these things can be handled by e-mail mostly, it is nevertheless mandatory to have regular phone and video conferences, writing minutes and checking for completion of tasks. Additionally it is often necessary to have agreed on strict rules for processing if any disagreements arise. Those dissents are mainly found among others in security policy issues, scheduling of software installation and upgrades, budget issues for needed components and other issues. From the network point of view the optimal network setup requires extensive discussions. Often simple network redesigns will lead to time-consuming discussions. Local priorities often contradict optimal solutions. Especially in a multi to multi network scenario (DEISA has 11 sites) a design has to be chosen, which arises from the lowest common denominator. Things have to be configured that allow a good performance for all sites, which may not be the optimal solution for every connection setup.

Future enhancements, co-operations and summary

DEISA is aware of all the activities around Europe and measures itself only upon its mission to enable new science in Europe. Many new activities will come up with the European FP7 program. DEISA has to find its position in these upcoming activities. DEISA is open for new grand challenges and is prepared to be part of the future European HPC infrastructure. Currently T1 High Performance Centres (HPC) are integrated into DEISA, but the upcoming European T0-HPC centres will be of major importance to DEISA. DEISA sees its role as an integrating part of those centres being the global player. Even if T0 centres should not be fully integrated into the DEISA infrastructure, a close collaboration will be indispensable. Therefore DEISA plans to integrate these sites into its backbone infrastructure implying further evolution of the network in future.

DEISA feels confident that its challenging goals are highly dependent on the future development of future national and international networks. Future HPC centres may be able to operate and serve its users only if an adequate networking connectivity can be provided. Researchers are spread across Europe, but distances become shorter. These scientists work together more and more looking for resources where ever possible. In future it will become irrelevant where the supercomputer system is located. Connectivity will be the driving force.

Cost-efficient high-performance interconnects need to be provided. Future bandwidth-on-demand services will allow the exact reservation of network links with appropriate bandwidth, quality and availability. GÉANT2 and the European NRENs have done a good job until now, but they will have to evolve further on. A European world-class, high-bandwidth, multi protocol network infrastructure will push forward science to new dimensions.

Three years of successful DEISA operation have shown that the concept implemented in DEISA proceeded very well. This does not preclude that organizational structures of DEISA may change over time. But the general idea of DEISA will sustain. The next steps within DEISA will be to establish an efficient organization embracing all relevant HPC organizations in Europe. Being a central player within European HPC, DEISA intends to contribute to a global eInfrastructure for science and technology furthermore. Integrating leading supercomputing platforms with Grid technologies and reinforcing capability with shared petascale systems is needed to open the way to new research dimensions. The new DEISA 10 Gb/s European network infrastructure provides the fundament for the DEISA supercomputer infrastructure through which DEISA will pave the way for further scientific computing. The DEISA supercomputing infrastructure, inviting leading scientific research areas to operate on future grand challenges, provides a vision of future European Grid computing.

References

[1] Distributed European Infrastructure for Supercomputer Applications, http://www.deisa.org

[2] GÉANT home page, http://www.geant.net

[3] GÉANT - GÉANT/Dante description of the Premium IP service, http://www.dante.net/server/show/nav.00700a003

[4] FP6 European Research Infrastructures, http://www.cordis.lu/infrastructures/home.html

[5] R.Niederberger, O.Mextorf, The DEISA Project – Network Operation and Support - First experiences, "Selected Papers from the TERENA Networking Conference 2005", June 2005, (ISBN 9077559094)

[6] GÉANT2 home page, http://www.geant2.net/

[7] GPFS: A Shared-Disk File System for Large Computing Clusters, F.Schmuck, R.Haskin, Proceedings of the Conference on File and Storage Technologies, 28–30 January 2002, Monterey, CA, pp. 231–244., http://www.almaden.ibm.com/StorageSystems/file_systems/GPFS/Fast02.pdf

[8] TeraGrid open scientific discovery infrastructure, http://www.teragrid.org/

[9] Ph.Andrews, M.Buechli, R.Harkness, R.Hatzky, Ch.Jordan, H.Lederer, R.Niederberger, A.Rimovsky, A.Schott, Th.Soddemann, V.Springel: Exploring the Hyper-grid idea with grand challenge applications: The DEISA-TERAGRID interoperability demonstration, Clade 2006 Workshop/ HPDC-15, June 2006, Paris, France, http://www-unix.mcs.anl.gov/~bair/CLADE2006/

[10] DEISA Extreme Computing Initative, June 2006, http://www.deisa.org/grid/initiative.php

[11] HPC-Europa – Pan-European Research Infrastructure on High Performance Computing, http://www.hpc-europa.org/

[12] The Enabling Grids for E-sciencE (EGEE) project, http://public.eu-egee.org/

[13] Y. Rekhter, B. Moskowitz, D. Karrenberg, G.J.de Groot, E. Lear, Address Allocation for Private Internets, ftp://ftp.rfc-editor.org/in-notes/rfc1918.txt

[14] Iperf - Version 2.0.1, The National Laboratory for Applied Network Research (NLANR), Distributed Application support team, http://dast.nlanr.net/Projects/Iperf/

[15] W. Allcock, GridFTP: Protocol Extensions to FTP for the Grid, Open Grid Forum document, http://www.ogf.org/documents/GFD.20.pdf and I. Mandrichenko, GridFTP Protocol Improvements, Open Grid Forum document, http://www.ogf.org/documents/GFD.21.pdf

[16] I. Mandrichenko, W. Allcock, T.Perelmutov, GridFTP v2 Protocol Description, Open Grid Forum document, http://www.ogf.org/documents/GFD.47.pdf

[17] M.Mathis, R.Reddy, J.Mahdavi, Advanced Network computing – Enabling High Performance Data Transfers, http://www.psc.edu/networking/projects/tcptune/, Nov 2006