

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Provision of the “proof of concept” 10 Gb/s DEISA network
infrastructure

Deliverable ID: DEISA-DSA1-2
Due date: August, 31st, 2006
Actual delivery date: November 24, 2006
Lead contractor for this deliverable: FZ-Jülich, Germany

Project start date: May 1st, 2004
Duration: 4 years

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Content

Table of Content.....	3
1. Introduction.....	4
1.1 Executive Summary.....	4
1.2 References and Applicable Documents	5
1.3 Document Amendment Procedure	5
1.4 List of Acronyms and Abbreviations	5
2. The second phase DEISA “proof of concept” network infrastructure.....	6
3. Architecture and Design decisions	7
4. Implementation of the DEISA backbone.....	8
5. Future directions.....	9
6. Status of the “proof of concept” 10 Gb/s infrastructure.....	10
7. Implementation specifics at DEISA sites	11
8. First test results (iperf)	13
9. Conclusions	14

1. Introduction

DEISA [1] is a consortium of leading national supercomputing centres¹ (DEISA sites) that deploys and operates a persistent, heterogeneous, production quality, distributed supercomputing environment with continental scope and tera-scale performance. The purpose of this FP6 funded research infrastructure is to enable scientific discovery across a broad spectrum of science and technology areas, by enhancing and reinforcing European capabilities in the area of high performance computing. This becomes possible through a deep integration of existing national high-end platforms, tightly coupled by a dedicated network and supported by innovative system and grid software (source: www.deisa.org).

To enable distributed computing there is a strong need for network connectivity with guaranteed capacity between the DEISA supercomputer systems. The current network connectivity is based on routed IP and MPLS tunnels and involves nine supercomputing centres in Europe (all DEISA sites except EPCC and HLRS, which will participate in the phase 2 network also). Currently each DEISA site has a dedicated Gigabit Ethernet (GE) connection to its local National Research and Education Network (NREN).

In order to scale up the capacity and the number of connected DEISA sites the proposal has been to start a “proof of concept” phase in which five DEISA sites will be connected with 10 Gb/s to the DEISA backbone. This optical private network based on GE and 10 Gigabit Ethernet (10GE) connections provided by the involved NRENs and GÉANT2 should be designed to allow a future integration of additional DEISA sites with higher communication speed, if adequate budget will be available. The proposed backbone allows those “proof of concept” DEISA locations to transmit traffic at peak capacity (up to 10 Gbps) without interfering with non-DEISA traffic.

1.1 Executive Summary

The DEISA Service Activity 1 – Network Operation and Support is responsible for deploying the high performance network infrastructure for DEISA. In phase 1 of the project the main task has been the deployment of the infrastructure especially for the four DEISA “Proof of concept” sites at CINECA (Italy), IDRIS (France), RZG (Germany) and FZJ (Germany). The infrastructure has been based on the tight coupling - using *virtually* dedicated bandwidth network interconnects (GEANT IP Premium service [2]) - of these four homogeneous national supercomputers, to provide a distributed supercomputing platform operating in multi-cluster mode. In the next phase the remaining DEISA sites (five out of seven remaining, except HLRS and EPCC) have been connected to the 1 Gb/s DEISA backbone. The network infrastructure is fully operational including services to measure performance and monitor the status.

¹ 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

Starting in the middle of 2005 a small team of network experts has been setup to design a DEISA phase 2 network infrastructure for some of the DEISA sites (proof of concept). The new backbone will be based on fibre links provided by GÉANT2 [3] and the involved NRENs, allowing 10 Gb/s throughput connectivity.

The new 10 Gb/s backbone has been designed to allow easy connectivity between 1Gb/s and 10 Gb/s DEISA sites as well as easy upgrade of 1 Gb/s sites to 10 Gb/s. This document describes in detail the design principles and current status of this new 10 Gb/s network infrastructure, to which the phase 2 “proof of concept” sites are connected to, as well as the intercommunication with the other DEISA sites which are connected with 1 Gb/s only.

1.2 References and Applicable Documents

- [1] Distributed European Infrastructure for Supercomputer Applications,
<http://www.deisa.org>
- [2] GÉANT - GÉANT/Dante description of the Premium IP service,
<http://www.dante.net/server/show/nav.00700a003>
- [3] GÉANT2 Home page,
<http://www.geant2.net/>

1.3 Document Amendment Procedure

1.4 List of Acronyms and Abbreviations

DWDM	Dense Wavelength Division Multiplexing
GÉANT	multi-gigabit pan-European data communications network administrated and operated by Dante
GÉANT2	the seventh generation of pan-European research and education network, successor of GÉANT
MPLS	Multi Protocol Label Switching
NREN	National Research Network
PoP	Point of Presence

2. The second phase DEISA “proof of concept” network infrastructure

After integrating three additional partners, BSC, LRZ and HLRS, in May 1st, 2005 into the DEISA consortium and the definition of a further future architecture and roadmap for a full, heterogeneous supercomputing Grid incorporating the new partners, a new network strategy was adopted at PM12:

- A first priority was given to guarantee the connectivity of all (new and initial) partners at 1 Gb/s, for the whole of the project duration. This means, in particular, reserving the funding needed for the operation of the 1 Gb/s network infrastructure connecting all partners until April 2008.
- With the remaining funds, deployment of a limited “proof of concept” 10 Gb/s infrastructure with a restricted number of partners, optimizing as much as possible the impact on the DEISA Grid.

The “proof of concept” sites chosen in this phase have been FZJ, IDRIS, RZG, and SARA. This decision was based primarily on the near future availability of the corresponding NREN infrastructures and the financial feasibility. Because of the possibility to provide network connectivity for LRZ also with only a small amount of additional budget, it was decided to include this DEISA site also already in the “proof of concept” phase.

During a next phase it is planned to interconnect all eleven DEISA sites (financed by eDEISA Budget) mostly with 10 GE in seven countries throughout Europe to the DEISA backbone. The sites, their location and involved NRENs are listed in Table 1. All these NRENs are connected via the European wide GÉANT2 network provided by DANTE.

Name of DEISA host	DEISA Location	NREN involved
Institut du Développement et des Ressources en Informatique Scientifique (IDRIS-CNRS)	Orsay, France	Renater
Forschungszentrum Jülich (FZJ)	Jülich, Germany	DFN
Rechenzentrum Garching of the Max Planck Society (RZG)	Garching, Germany	DFN
Consorzio Interuniversitario (CINECA)	Bologna, Italy	GARR
Finnish Information Technology Centre for Science (CSC)	Espoo, Finland	NORUnet
SARA Computing and Networking Services	Amsterdam, The Netherlands	SURFnet
Leibniz Computing Centre of the Bavarian Academy of Sciences and Humanities (LRZ)	Garching, Germany	DFN
Barcelona Supercomputing Center (BSC)	Barcelona, Spain	RedIRIS
European Centre for Medium-Range Weather Forecasts (ECMWF)	Reading, UK	UKERNA
High Performance Computing Center Stuttgart (HLRS)	Stuttgart, Germany	DFN
Edinburgh Parallel Computing Centre (EPCC)	Edinburgh, UK	UKERNA

Table 1: DEISA sites and DEISA locations

The switch/routers of the DEISA supercomputing sites will have 10GE connections to the DEISA backbone. However, at some sites, the single supercomputer system nodes will be connected with 1GE connections to this DEISA site local traffic conglomerating

switch/router only. The principle local network configurations depend on supercomputer system characteristics, I/O node and file system configurations and performance. E.g. it doesn't make any sense to connect 4000 nodes with 10 Gb/s each to a central switch which is connected to the DEISA backbone with one 10 Gb/s link only. In order to obtain maximum throughput on a 10 Gb/s network connection system interrupts should be minimized. Therefore it is planned to support jumbo frames, i.e. 9180 byte frames.

The DEISA backbone is based on a non-blocking infrastructure. Transfers between pairs of DEISA locations do not impact each other in terms of performance. Protection against node and link failures as well as downtime due to maintenance is considered to be a part of the network design. It will increase the service availability of the DEISA backbone.

Another important issue for the reliable operation of the applications using the DEISA backbone is the ability to monitor the DEISA backbone. Therefore, a 24/7 centrally operated monitoring system will be implemented in future for the DEISA backbone.

This monitoring system will provide information about the current and past network performance using a graphical user interface and will be able to collect and manage data from different administrative domains, i.e. from local networking equipment as well as the DEISA backbone infrastructure.

3. Architecture and Design decisions

The DEISA backbone will be an Optical Private Network (OPN) built from components of the NREN and GÉANT2 transmission platforms. I.e. the links between DEISA locations are reserved for DEISA usage only. Each DEISA location will have a 1GE or 10 GE link to the DEISA backbone. This link provides the DEISA hosts with access to the DEISA backbone, which comprises part of the GÉANT2 and NREN infrastructure.

The 10GE connections will be carried directly over the transmission system. Gigabit Ethernet connections are transported over the switched SDH/SONET network using the Generic Framing Procedure (GFP) and virtual concatenation (VC4-7v).

A high-level network overview of all DEISA Partners with their connecting NRENs is shown in Figure 1.

In order to create the DEISA backbone a number of switches need to be deployed. First, there will be one central switch (may be more decentralized switches in a later eDEISA stage) deployed in Frankfurt at the PoP of DFN, the German NREN. Secondly, at each DEISA location a layer 2/3 switch will be deployed in order to connect the DEISA hosts to the DEISA backbone. Monitoring workstations will be connected to some of these local switches as well. The design will allow the integration of additional secondary links between these switches in the future for traffic path changes in case the primary links fail.

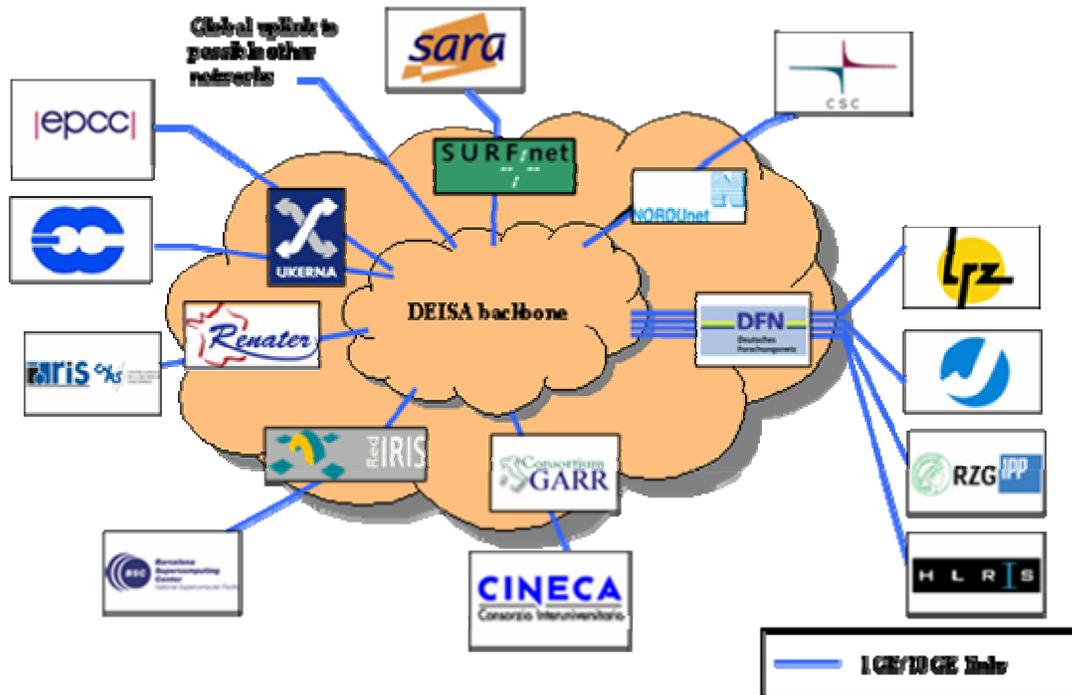


Figure 1: Structure of the DEISA backbone

4. Implementation of the DEISA backbone

The implementation of the DEISA backbone is based on a star topology. Other topologies have been discussed also but would lead to much higher costs. The current star topology consists of a single central switch/router.

The location of the central DEISA switch has been discussed several times, to be at one special DEISA site, within one preferred NREN PoP or in a central location of GÉANT2. Discussing financial, contractual, technical and operational possibility, it was decided to locate the central DEISA switch within the PoP of the German NREN DFN at Frankfurt, because Frankfurt is a geographically central point in the DEISA topology. From here four German sites can be accessed easily. Additionally, the GÉANT2 PoP is located within the same building in Frankfurt, Germany, so that the connectivity from the other 7 DEISA sites via GÉANT2 to this central switch could be provided without any problems. The main focus for an optimal solution has been to reduce the number of required wavelengths, and hence, the reduction of the overall transmission costs.

This central switch is in principle a single point of failure. However, this risk will be significantly reduced (or removed) by using a redundant switch that contains multiple power supplies and control modules.

At each DEISA location a DEISA edge switch will be deployed, which will be connected to the central switch in Frankfurt via local NREN and GÉANT2 wavelengths. In future the DEISA edge switches can be connected with two diversely routed links to the central switch for backup purposes if needed.

The location of the central switch allows an easy integration of connections between the DEISA project in Europe and other projects worldwide. The US TeraGrid project had

been connected for test purposes and interoperability demonstrations at the Supercomputing Conference SC05 in Seattle in autumn 2005 to the 1 Gb/s backbone. Future similar tests could be arranged easily. Having implemented the 10 Gb/s DEISA backbone it should be possible to add site/project connections by just adding another link to this backbone switch via GÉANT2.

5. Future directions

The use of Cross Border Fibre (CBF)² will potentially lead to shorter backup routes and may therefore economise the DEISA Optical Private Network (OPN). The current network model will be capable to adapt to future plans of European high performance computing installations.

Although the next phase of the DEISA backbone will be constructed as a star topology future evolutions of the backbone have to be kept in mind. An initial design could be a star network with a single switch. In this case there is redundancy within this switching equipment (backup control and media modules, dual power supplies, etc.). A backup connection can be integrated for each link. A possible evolution is an extended core network, which is not a full mesh, with multiple "central" switches at different locations and a fast network (multiple 10 GE) connecting them (see Figure 2). DEISA sites could be connected with their local switches to these central switches, i.e. the DEISA backbone. Using alternate NREN links or CBFs backup links can be used. In other words, a distributed core infrastructure can be arranged. Using this distributed core infrastructure other projects like TeraGrid could be connected in a more convenient way.

It has to be clarified here, that this future network, having all DEISA sites connected with 10 Gb/s, can not be financed by DEISA budget. Also the extended version with multiple backbone switches, as sketched in figure 2, can be realized with eDEISA budget only and only if prices drop down dramatically in the future.

² Cross-border dark fibre (CBF) within this context are direct connections between neighbouring NRENs, which allow to connect NREN A and NREN B without using the GÉANT or GÉANT2 network in between.

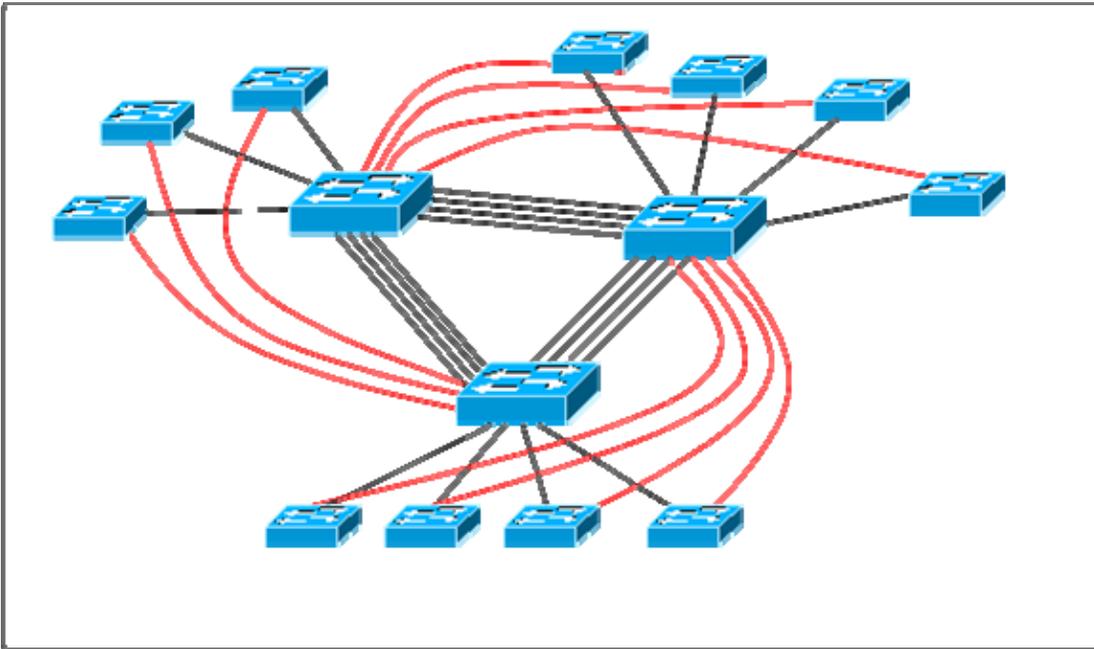


Figure 2: Extended core network with backup links from DEISA sites (Central switches locates within GEANT2 PoPs or NREN PoPs, peripheral switches located at DEISA sites)

6. Status of the “proof of concept” 10 Gb/s infrastructure

Since several months the DEISA SA1 team is in discussion with the NRENs and Dante as the organisational provider of GÉANT2 to foster the evolution of the DEISA 10 Gb/s infrastructure test bed. Unfortunately delivery and installation of equipment as well as testing of fibre infrastructure requires several weeks. The provision of a site to site wavelength link within Germany takes about 8 weeks approximately. Dante states that a NREN to NREN connection (wavelength) is projected with a 12 weeks delay after order acceptance, because of ordering equipment delays and test of the new link. Therefore a first phase 10 Gb/s network backbone has become available only at begin of November 2006. Currently the central switch at DFN PoP in Frankfurt has been installed. The DEISA sites FZJ, RZG and LRZ have been connected to the network infrastructure. SARA will be connected in the near future. Current estimate is end of November, because of delays in the delivery of the GÉANT2 Alcatel DWDM equipment. Unfortunately because of some additional delays in provisioning links within France NREN, Renater, it is assumed that IDRIS can not be connected before February 2006.

Fortunately, the eDEISA contract has been signed in the last days. The eDEISA budget allows connecting the other six DEISA sites with 10 Gb/s to the new DEISA infrastructure. Therefore it is possible to start with the provisioning of links to these sites also. Stuttgart and BSC have made agreements to be connected as soon as possible after signing of the eDEISA contract. DANTE and DFN acted here without any official contract and ordered already relevant equipment, so that it will be possible to connect HLRS, Stuttgart, within November and BSC, Barcelona, within mid of December to the 10 Gb/s infrastructure (though this is within the eDEISA and not the DEISA contract). Especially early connectivity of HLRS to the 10 Gb/s network infrastructure is very important, because this site did not participate to the 1 Gb/s network backbone. Being one of the last connected to the DEISA infrastructure they will be some of the first to have a high speed interconnect to DEISA.

7. Implementation specifics at DEISA sites

It is planned to connect five of the DEISA sites within the “proof of Concept” phase to the DEISA 10 Gb/s infrastructure as described above. The other sites had to remain connected to the 1 Gb/s infrastructure until additional budget could be assigned.

Because of the eDEISA contract it has become feasible to connect all sites to the new DEISA backbone. Unfortunately this has to be accomplished within evolutionary steps dependent of availability of NREN network infrastructures and implementation costs. Therefore an intermediate transmission phase has to be introduced where some sites are connected to the new backbone whereas others are connected to the old (1Gb/s) DEISA infrastructure. This will be accomplished by having a 1 Gb/s link between this both infrastructures at Frankfurt. Because of connecting more and more sites to the 10 Gb/s infrastructure in the near future, it is assumed that no additional bandwidth will be needed.

As mentioned earlier there will be different local network configurations at the divers DEISA sites. E.g. at FZJ the local switch has 10 Gb/s network connections to each of the two relevant DEISA GPFS I/O nodes. An additional 10 Gb/s connection has been installed to a “test” node of the JUMP supercomputer cluster. All other nodes are connected with 1 Gb/s network interfaces to the local DEISA switch (see Figure 3). This configuration allows to test the difference between a direct 1 Gb/s connectivity (using all the interfaces connected to the DEISA local switch) of a node to the DEISA backbone (that is using all the interfaces as installed in figure 3) and alternatively a gateway connectivity via the internal IBM 16 Gb/s Federation Switch Network and a directly to DEISA connected 10 Gb/s gateway node (see Figure 4), where the 1 Gb/s interfaces have been disabled or used for other internal non-DEISA connections. Here the connectivity from all supercomputer nodes to DEISA is realized by using this single Gateway node.

The difference is, that in figure 3 all nodes can communicate in parallel with a maximum throughput of 1 Gb/s throughput each, but with a summarized load of 10 Gb/s. In figure 4 all nodes can communicate in parallel with a maximum of 10 Gb/s per link and also with a summarized maximum of 10 Gb/s. Dependent on the applications used the first or second version would be better.

At RZG the two GPFS I/O nodes are connect with 10 Gb/s to the local RZG DEISA switch acting as gateways for the other compute nodes. This RZG switch has been connected with 10 Gb/s to the DEISA backbone. So this configuration allows a conglomerated access of 10 Gb/s to RZG supercomputer nodes. Each node itself can get data with a maximum of 1 Gb/s.

LRZ has installed at each of its ALTIX nodes a 10 Gb/s interface, which will be connected to a local switch. This switch is connected to the DEISA RZG local switch with 10 Gb/s allowing 10 Gb/s throughput to every node of the Altix system.

Because of “proof of concept” considerations it was decided to connect both sites, RZG and LRZ, with one 10 Gb/s link to the DEISA central switch only. If required this configuration can be substituted by a 10 Gb/s link for each of both sites in the future without any design changes if necessary after eDEISA has been started. The following figure illustrates the current network design.

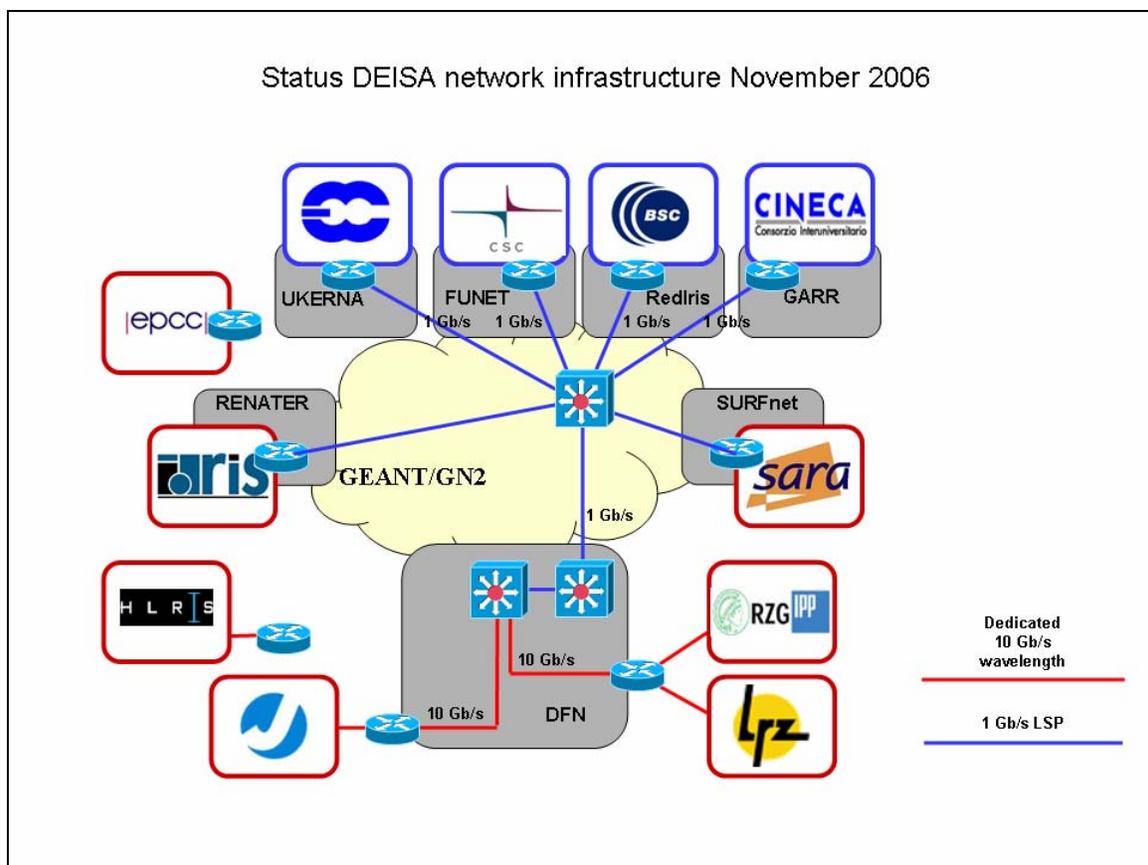


Figure 5: DEISA phase 2 “proof of concept” network design November 2006

8. First test results (iperf)

The DEISA 10 Gb/s switch at Frankfurt has been installed within the last days. Therefore throughput values could not be tested until now. We are just in the phase of configuration. A detailed discussion on throughput values reached, network options used and application behaviours analyzed will be given in the next report.

Local tests at FZJ with the DEISA 10 Gb/s interfaces and network components (switches) have shown that the throughput values that can be reached per supercomputer interface are less than 50 % of the theoretical maximum. Connecting two IBM supercomputer nodes at FZJ we got about 2.5 – 3.0 Gb/s throughput using an ordinary MTU size of 1500. These I/O nodes have been equipped with 4 Power4+ CPUs each.

By using Jumbo frames (9000 byte MTU size), a maximum of 3.5 to 4.0 Gb/s between these nodes could be measured.

As these nodes are involved in our production environment, there has been done no tuning of network parameters currently. The tests have been done using the iperf program that we already used for monitoring and tuning of the 1 Gb/s DEISA network.

The test results seen so far reflect similar tests that have been done by numerous independent network testers all over the world. Current 10 Gb/s Ethernet network interfaces cannot be fed up to the theoretical throughput by current operating systems.

Nevertheless this fact does not influence the overall importance of the 10 Gb/s infrastructure for DEISA. DEISA relies on GPFS which does file transfer in a parallel manner, allowing a number of I/O nodes to transfer data in parallel. Therefore a 10 Gb/s link can be fed without any problem by parallel streams. These parallel streams can result from one application generating these multiple streams e.g. GridFTP as well as from a distributed application running on several parallel nodes and therefore using additional network interfaces in parallel. Additionally a supercomputer system is normally used by many users in parallel. All these users can generate independent network load to different supercomputer systems in Europe in parallel. Multiplexing these streams from one supercomputer system to the other DEISA sites across Europe requires an adequate 10 Gb/s backbone infrastructure.

We will look into these scenarios in more detail in the future. Detailed information concerning these tests will be given in the next report.

9. Conclusions

This document set out the general architectural principles for the second DEISA “proof of concept” networking phase. It has sketched the network design and the operational procedures. Details of the further evolution of the DEISA backbone will be summarized in future SA1 reports.

The distributed character of the optical networks deployed by GÉANT2 and by its partner NRENs will allow DEISA and in fact any future HPC initiative to flexibly adapt to the changing needs of the capability computing discipline in Europe as well as in global co-operation.