

CONTRACT NUMBER 508830

DEISA
DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Annual activity report on network status and operation

Deliverable ID: DEISA-DSA1-4
Due date: April, 30th, 2007
Actual delivery date: May 28, 2007
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..... | 2 |
| 1. Introduction..... | 3 |
| 1.1 Executive Summary..... | 3 |
| 1.2 References and Applicable Documents | 4 |
| 1.3 Document Amendment Procedure | 4 |
| 1.4 List of Acronyms and Abbreviations | 4 |
| 2. Current status of the DEISA network infrastructure | 6 |
| 3. The design of the DEISA network backbone | 7 |
| 4. Architecture and Design decisions | 8 |
| 5. Conclusions | 10 |

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.

To enable distributed computing there is a strong need for network connectivity with guaranteed capacity between the DEISA supercomputer systems. The starting network connectivity of the DEISA project has been based on routed IP and MPLS tunnels and involved nine supercomputing centres in Europe (all DEISA sites except EPCC and HLRS). Each DEISA site had 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 a “proof of concept” phase in which five DEISA sites have been connected with 10 Gb/s to the DEISA backbone has been started. This optical private network based on 10 Gigabit Ethernet (10GE) has been connected to the older (phase 1) 1 Gb/s network infrastructure to allow connectivity between any DEISA partner independent of his connectivity bandwidth. The “proof of concept” phase was introduced to show that an easy upgrade to higher communications speeds is feasible and technologically reasonable.

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 of these homogeneous national supercomputers using *virtually* dedicated bandwidth network interconnects (GEANT IP Premium service [2]) to provide a distributed supercomputing platform operating in multi-cluster mode. In the next phase the DEISA sites BSC, CSC, ECMWF, LRZ, and SARA have been connected to the 1 Gb/s DEISA backbone. The network infrastructure has been operational for about 3 years without any major problems. Services to measure performance and monitor the status have been setup successfully.

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 mid of 2006 a “proof of concept”, DEISA phase 2, network infrastructure with 10 Gb/s throughput for the DEISA sites BSC, FZJ, LRZ, RZG, and SARA has been initiated. The network monitoring tools have been updated accordingly.

The new backbone has been 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 gives an overview about the current network status and layout, the activities done and goals reached throughout the reporting period May 2006 to April 2007.

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/>
- [4] TeraGrid open scientific discovery infrastructure, <http://www.teragrid.org/>
- [5] Ph.Andrews, M.Buechli, R.Harkness, R.Hatzky, Ch.Jordan, H.Lederer, R.Niederberger, A.Rimovsky, A.Schott, Th.Sodemann, 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/>

1.3 Document Amendment Procedure

1.4 List of Acronyms and Abbreviations

| | |
|---------------|--|
| BSC | Barcelona Supercomputing Center, Barcelona, Spain |
| CINECA | Consorzio Interuniversitario, Bologna, Italy |
| CSC | Finnish Information Technology Centre for Science, Espoo, Finland |
| DWDM | Dense Wavelength Division Multiplexing |
| ECMWF | European Centre for Medium-Range Weather Forecasts , Reading, UK |
| EPCC | Edinburgh Parallel Computing Centre, Edinburgh, UK |
| FZJ | Forschungszentrum Jülich, Jülich, Germany |
| 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 |
| HLRS | High Performance Computing Center Stuttgart, Stuttgart, Germany |

| | |
|--------------|--|
| IDRIS | Institut du Développement et des Ressources en Informatique Scientifique, Orsay, France |
| LRZ | Leibniz Computing Centre of the Bavarian Academy of Sciences and Humanities, Garching, Germany |
| MPLS | Multi Protocol Label Switching |
| NREN | National Research Network |
| PoP | Point of Presence |
| RZG | Rechenzentrum Garching of the Max Planck Society, Garching, Germany |
| SARA | SARA Computing and Networking Services, Amsterdam, The Netherlands |

2. Current status of the DEISA network infrastructure

After having integrated nine of the 11 DEISA partners, except EPCC and HLRS, into the “phase 1” DEISA 1 Gb/s network infrastructure the first main objective of reporting period May 2006 to April 2007 has been to operate this European wide network infrastructure in production quality without any major outages.

Therefore the main focus of “Service Activity 1 – Network operation and Support” has been the definition, implementation, configuration, and optimisation of the old 1 Gb/s network infrastructure as well as the smooth integration of a new 10 Gb/s DEISA network infrastructure.

The provision of an adapted and extended availability and performance monitoring service for the old and later on the new infrastructure has been a further activity. These services have been integrated into the already existing web server at <http://wwwnet.deisa.fz-juelich.de> which is available from authorized sites and persons only.

A further area of investigation has been and will be in the future user support in any kind of network aspects as there are configuration issues, application setup, performance issues, fairness of network usage and security considerations.

Overall main objective of this service activity is providing a production quality network infrastructure during the whole project life cycle.

A major activity in the third year of operation has been the design and implementation of the new 10 Gb/s network infrastructure. Whereas the objective of SA1 has been to start a proof of concept phase for this 10 Gb/s network, the focus on eSA1 has been the extension of this preliminary phase to most of all partner sites. Therefore in the rest of this report the activities for both projects will be described combined.

Starting in July 2006 first negotiations with the NRENs and GÉANT2 have been started to realize the next phase of the DEISA network infrastructure. Unfortunately time delays caused by suppliers, (re)construction activities in the provider surroundings, installation and test activities allowed the installation of the central DEISA switch at DFN Frankfurt and the first links between German DEISA partners not before mid of November 2006. Nevertheless local tests have been started and configuration details could be planned before. The first non German DEISA partner SARA could be connected in January 2007. BSC followed at end of February 2007 after connectivity setup needed about one month because of the special setup going from Barcelona to Madrid and then back via France to Germany. In the meantime also the connectivity to HLRS could be initiated in February which allowed reaching this DEISA partner for the first time from all other sites via a dedicated DEISA network connection. In April 2007 the connectivity to IDRIS was setup so that now seven of the eleven partners are connected to the new 10 Gb/s infrastructure. These partners are BSC, FZJ, HLRS, IDRIS, LRZ, RZG, and SARA. The remaining sites CINECA, CSC and EPCC are planned to be connected within the May to July 2007 time period. The setup of connecting these seven DEISA partners with 10 Gb/s could be arranged only because of the additional budget available through eSA1. This budget will allow connecting the remaining sites in the future also. A 10 Gb/s connectivity to ECMWF has not been taken into consideration because of unforeseen cost issues for this special local provider link.

During the whole reporting period enhancements and upgrades of the monitoring facilities have been made. Nevertheless any new DEISA partner joining the new infrastructure demands configuration changes in network layout and hence monitoring setup.

Application programmers and supercomputer users have been educated concerning the features and drawbacks as well as pros and cons of the new infrastructure. This activity will outlast over the whole project life time.

Unfortunately we had to deal with some deviations from work program, which have been related to delays in network setup. Though the network layout has been designed and propagated long before the current reporting period, providers have not been ready to make available these links in time. As stated above, these delays have been related to provision of supplier equipment behind schedule, negotiation of schedules for reconfiguration and setup of links and equipment and especially because of testing dedicated wavelengths links not being standard in the past. Also this kind of new services provided by NRENs and GÉANT2 have not been state of the art until now and therefore led to delays because of unknown and unexpected problems (new service, new equipment, new contracts, new configurations and new interactions between partners (NREN, GEANT2, and DEISA staff). It is expected that these starting problems will be handled better in the future.

The definition of the future architecture and roadmap for a full, heterogeneous supercomputing Grid incorporating mostly all partners led to a new network strategy, which 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, the deployment of a limited “proof of concept” 10 Gb/s infrastructure with a restricted number of partners, optimizing as much as possible, should be started.

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. Unfortunately, because of delays at RENATER, IDRIS could not be integrated in the proof of concept and therefore was substituted by BSC.

3. The design of the DEISA network backbone

The switch/routers of the DEISA supercomputing sites have been installed with 10GE interfaces connected to the DEISA backbone, i.e. the central DEISA switch at DFN Frankfurt using dedicated wavelengths provided by local NRENs. However, at some sites, the single supercomputer system nodes have been 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 site local DEISA edge 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 the links have been configured 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.

Table 1 gives an overview about DEISA partners, their locations, and NRENs involved.

| 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 | NORDUnet |
| 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

4. Architecture and Design decisions

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 DEISA backbone has been implemented as an Optical Private Network (OPN) built from components of the NREN and GÉANT2 transmission platforms which uses DWDM equipment from Alcatel to provide dedicated wavelength across Europe. I.e. the links between DEISA locations are dedicated wavelengths reserved for DEISA usage only. Each DEISA location has a 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.

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

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 the four German sites FZJ, HLRS, LRZ, and RZG 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.

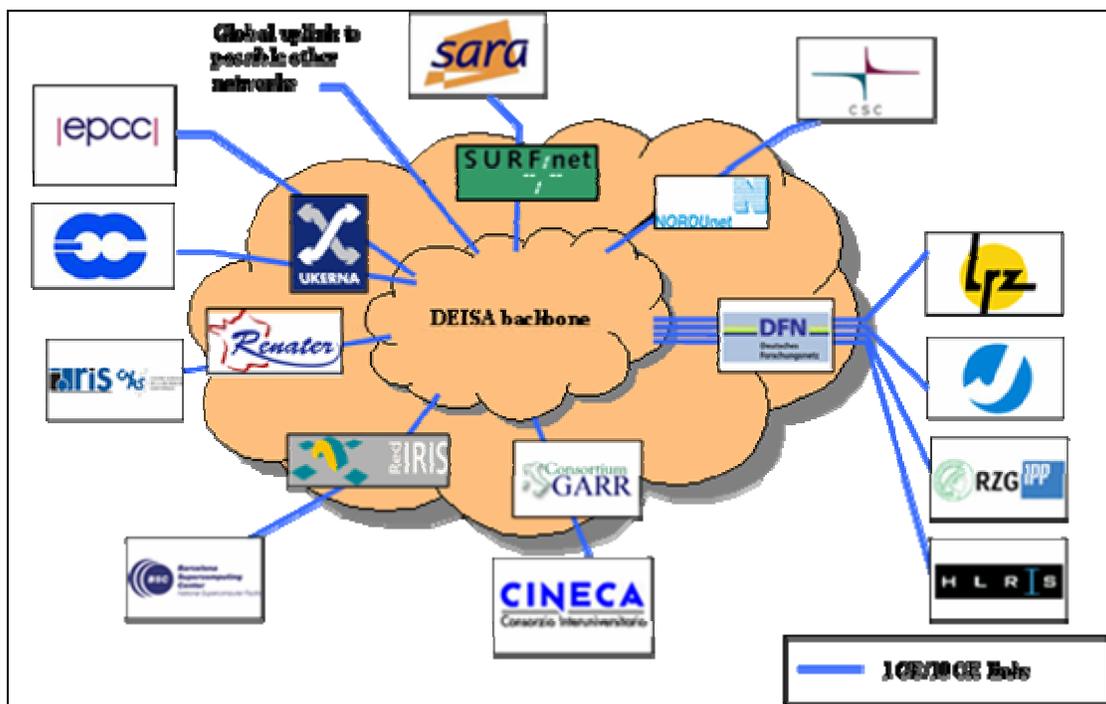


Figure 1: Structure of the DEISA backbone

At each DEISA location a DEISA edge switch has been deployed, which is 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 [4] 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 [5]. Future similar tests could be arranged easily. Having implemented the 10 Gb/s DEISA backbone it is easily possible to add site/project connections to DEISA by routing them through GÉANT2 to the central switch at Frankfurt. If high throughput is needed for some connections via GÉANT2, just adding another link to this backbone switch is required.

The current network layout status is show in figure 2 below.

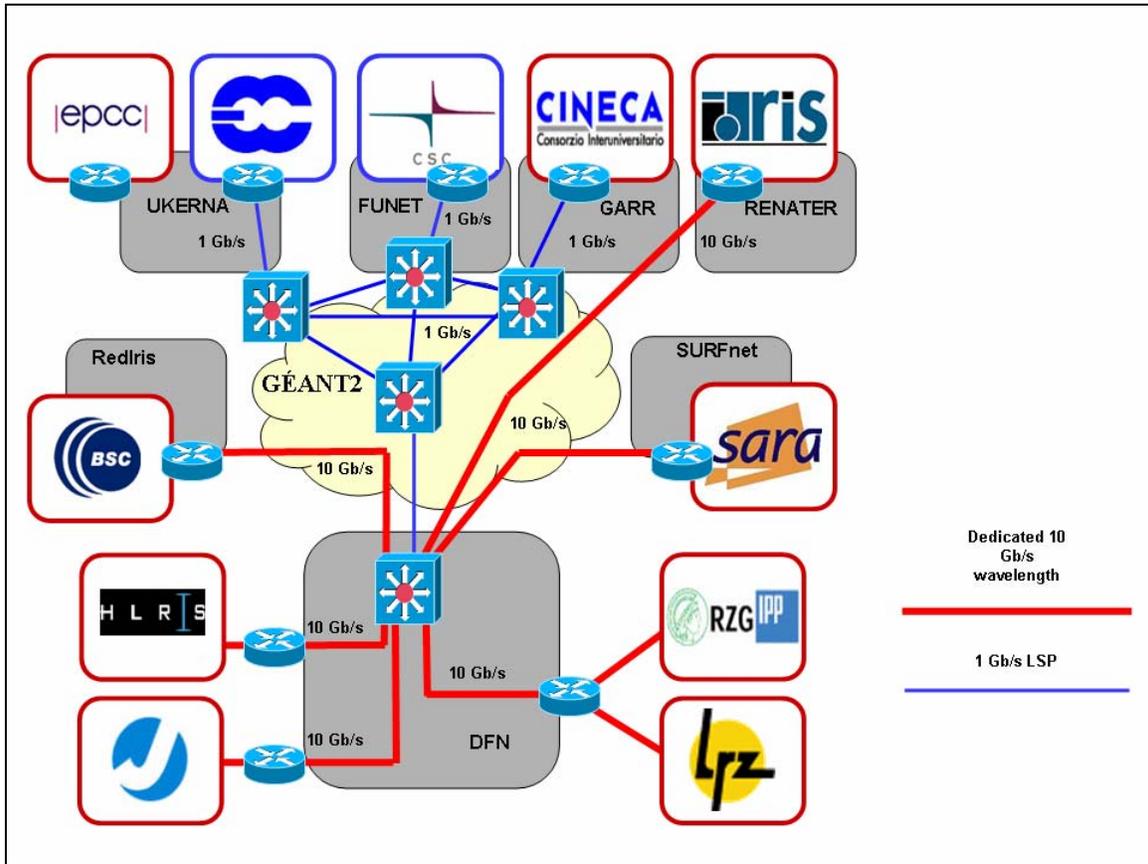


Figure 2: Current status of DEISA network

5. Conclusions

This document set out the general architectural principles for the second DEISA “proof of concept” networking phase. It enumerated the activities done within the last reporting period and gave an overview about and described the status of the current DEISA network infrastructure. Details of the further evolution of the DEISA backbone will be summarized in future SA1 and eSA1 reports.

The milestones set for DEISA SA1 have been reached though unforeseen delays had to be mastered. These obstructions did not lead to any major problems, but required some changes in time schedules only. At end of April SA1 is on schedule again and the future extension to the rest of the DEISA partners can be started.