

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

Provision of a high performance network infrastructure
(all sites)

Deliverable ID: DEISA-DSA1-1.2
Due date : April, 22nd, 2005
Actual delivery date: May 15, 2004
Lead contractor for this deliverable: FZ-Jülich, Germany

Project start date : May 1st, 2004
Duration: 5 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	
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	3
1.3 Document Amendment Procedure	3
1.4 List of Acronyms and Abbreviations	3
2. High performance network infrastructure	5
2.1 Requirements.....	5
2.2 The DEISA network infrastructure	5
2.3 Distinguishing between DEISA and general network traffic.....	7
2.4 Testing the DEISA network infrastructure	7
3. Network security considerations for DEISA.....	8
4. Network demonstration	9

1. Introduction

1.1 Executive Summary

The Service Activity 1 – Network Operation and Support is responsible for deploying the high performance network infrastructure for DEISA [1]. In phase 1 of the project the main task has been the deployment of the infrastructure for all DEISA sites. The infrastructure is based on the tight coupling - using virtually dedicated bandwidth network interconnects (GEANT Premium IP service [2]¹) of all DEISA sites supercomputer systems, to provide a distributed supercomputing platform operating in multi-cluster mode.

The network infrastructure is in part fully operational including services to measure performance and monitor the status.

This document describes the DEISA network infrastructure in detail including the used IP addressing scheme and the resulting security implications according to the ISO-OSI reference model layer 1-4.

1.2 References and Applicable Documents

- [1] Distributed European Infrastructure for Supercomputer Applications, <http://www.deisa.org>
- [2] GÉANT/Dante description of the Premium IP service, <http://www.dante.net/server/show/nav.00700a003>
- [3] IPERF – Network performance test program developed by NLANR (National Laboratory for Applied Network Research, www.nlanr.net).

1.3 Document Amendment Procedure

1.4 List of Acronyms and Abbreviations

CWDM	Coarse Wavelength Division Multiplexing
DSCP	Differentiated services code point
GPFS	Global Parallel File System
HiPPI	High Performance Parallel Interface
ISO	International Standards Organisation
NREN	National Research Network
OSI	Open Systems Interconnection
PoP	Point of Presence

¹ Premium IP is a service that offers network priority over other traffic on the GÉANT network.

2. High performance network infrastructure

2.1 Requirements

The full deployment of the dedicated DEISA [1] network infrastructure will proceed in several steps, following the evolutions of the national and European research network infrastructures, and the adoption of the infrastructure by the user communities. During the first phase of the project, an initial high performance network infrastructure has been deployed as a “proof of concept” connecting four supercomputer systems at CINECA (Italy), FZJ (Germany), IDRIS (France) and RZG (Germany). The network infrastructure has been implemented with virtual dedicated 1 Gb/s links. In April 2005 the fifth supercomputer system located at SARA (The Netherlands) has been connected. The DEISA “dedicated” network uses services of the national research networks RENATER, GARR, SURFnet and DFN and the European research network GÉANT. SARA is the first site not using an IBM supercomputer system that has been connected to DEISA and this is the first step on the way to heterogeneity. Based on the results seen within the proof of concept phase planning for the future extension of the network to the DEISA sites CSC in Finland and ECMWF in UK has been started. Contrary to the plans (deliverable DSA1-1.2) not all sites have been connected until now.

Because of the relocation of the supercomputer centre of CSC, Espoo, Finland, a connection to DEISA was not appropriate. The connection will be implemented as soon as the relocation has been finished. This will presumably be done in May or June 2005. A connection of the other partner ECMWF to the DEISA backbone was not expedient since currently no applications with DEISA network requirements have been installed until now. This site will be connected as soon as appropriate. So DSA1-1.2 “Provision of a high performance network infrastructure (all sites)” has been delivered in detail only. Connectivity of all sites has not been appropriate until now.

2.2 The DEISA network infrastructure

The DEISA dedicated network involves the five sites CINECA, FZJ, IDRIS, RZG and SARA currently. Each site is connected with a 1 Gb/s communication link (LSP - label-switched-path) via its local NREN to the GEANT backbone. An expansion to site CSC at Espoo, Finland will follow in the near future (May/June 2005). The logical configuration of the DEISA “dedicated” network infrastructure is sketched in Figure 1 below.

Though each of the four installations in the “proof of concept” phase uses an IBM supercomputer composed of a huge number of P690, P690+ and P655 processors, each installation has a different internal configuration. At SARA an SGI Altix system has been installed leading now as a first step to an heterogeneous environment. Nodes with 1, 2, 4 and up to 32 processors each use Fast-Ethernet, Gigabit-Ethernet, HiPPI, Fiberchannel, Federation Switch and other communication techniques to transfer data. These network technologies will be used for high internal communication throughput, file access and other communication needs as backup, archiving and research projects. At some installations additional GigE-interfaces have been installed on some of the supercomputer nodes for DEISA purposes, for example FZ-Jülich installed a dedicated 1 Gb/s Ethernet adapter on each of its 41 nodes to support DEISA applications. Other

installations use already existing interfaces and configure them with “alias” addresses (additional addresses).

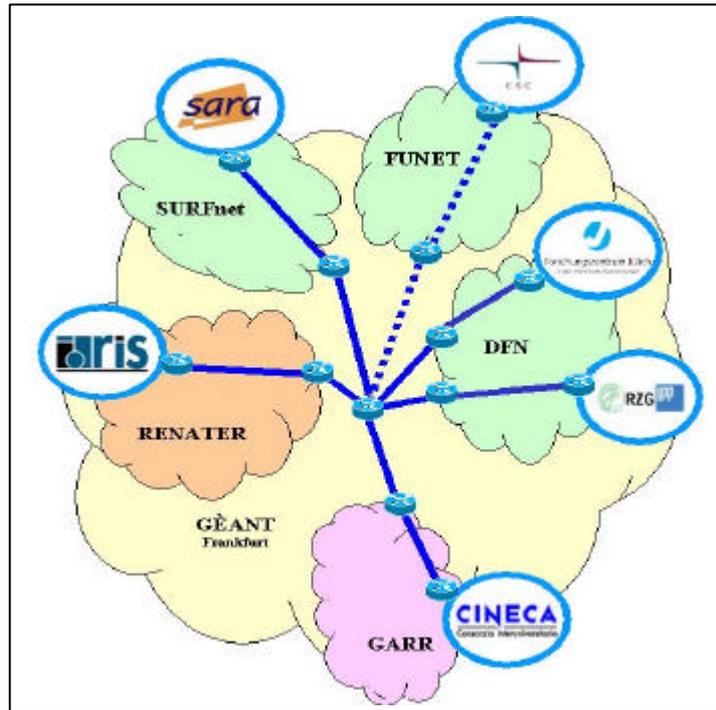


Figure 1: “Dedicated” DEISA network infrastructure

The so called DEISA interfaces (or second addresses on dual used interfaces) have got addresses within the subnets of the local DEISA installations. If a supercomputer application at IDRIS, RZG, Jülich, CINECA or SARA wants to send data to the supercomputer at another DEISA site it has to use one of these addresses. The routing tables of the supercomputers have been configured to guarantee this kind of different communication paths. The interfaces used within the supercomputer systems for DEISA purposes have been connected to special Ethernet switches/routers, which are connected to the communication equipment of the local NRENs. Depending on the PoP² of the NREN this connections are realized by local Ethernet connections (RZG, IDRIS), special CWDM equipment (CINECA), dark fibre (FZJ) or a dedicated light path (SARA). Depending on local requirements connections may change over time because e.g. additional fibres will be available or additional communication needs lead to additional fibre requirements or new NREN-PoPs or a change of location of NREN-PoPs.

The DEISA dedicated backbone will be implemented with GEANT’s “Premium IP service” [2]. DEISA packets will be flagged with a special format DSCP field (0x’46’)³. So prioritized traffic can be extracted and transmitted with better performance within the DEISA backbone.

² PoP – Point of Presence, The location where NRENs have installed their routers and switches to connect organizations, universities etc.

³ DSCP - Differentiated Services (DiffServ) is a model in which traffic is treated by intermediate systems with relative priorities based on the type of services (ToS) field within every IP packet. The six most significant bits of the DiffServ field are called “Differentiated services code point” (DSCP)

2.3 Distinguishing between DEISA and general network traffic

DEISA uses the above described internal network provided by GEANT and the National research networks (NRENs), which offer reserved bandwidth, to connect the supercomputers of the DEISA sites (not the sites itself). This internal network exists, of course, in addition to the standard Internet connectivity that each national supercomputer centre offers.

Since every installation supercomputer system uses two communication links one for normal and one for DEISA specific traffic, this dual attached configuration requires to distinguish between these two traffic classes. This will be done by using different addresses for the two traffic classes as well as special configurations within each component of the infrastructure. Within the NREN and GEANT network the DSCP flag distinguishes the two traffic classes. No traffic with this flags can be inserted into the normal communication paths of the national backbones. The local NREN routers will delete such traffic. Otherwise traffic, which doesn't have a source or destination address within the DEISA address space, can not be inserted from other interfaces than those local interfaces of the DEISA sites. If by any means traffic has been generated and transported via the NREN or GEANT network, this traffic will be deleted at least when arriving at the NREN routers at DEISA sites.

The DEISA internal network is structured in the following manner:

Research Center Jülich:	134.94.60.0/24 and 134.94.92.0/24
IPP Garching:	130.183.12.0/23
CINECA Bologna:	130.186.27.0/24
IDRIS Orsay:	130.84.240.0/21
SARA Amsterdam:	145.100.19.0/24

2.4 Testing the DEISA network infrastructure

After having deployed the 1 Gb/s network infrastructure, the main task has been configuring and testing the quality and throughput of the connections between the supercomputer systems of the "proof of concept" sites. Further on stability and reliability of the network infrastructure was tested and verified. The performance was optimized especially to support GPFS. It has been shown that instead of the default configurations of the IBM supercomputer systems especially the network option parameters have to be changed to get this optimized performance. The proposed changes to default system parameters have been documented at the DEISA network web site described below. Since these default options may vary from OS version to OS version and different options influence each other this tests have to be verified after each new system upgrade. After optimizing these network configuration parameters data communication with the network performance test tool *iperf* [3] led to tcp and udp throughput values of up to 940 Mb/s between remote DEISA supercomputer systems.

GPFS, the Global Parallel File System, has been deployed at the core sites. Network performance is extremely critical to GPFS, both with respect to latency and bandwidth. After tuning the network parameters, it was demonstrated that GPFS is capable of saturating the gigabit link. For example, GPFS at Research Center Jülich (FZJ) can deliver more than 10 Gbit/s of data locally via the internal Federation switch. It can easily deliver 880 Mbit/s via the network (1 Gbit/s Ethernet interface). Depending on the filesystem layout GByte/s transfers require parallel streams. This has to be considered in

the internal system configuration, especially in phase 2 of the project when higher bandwidths will be deployed.

A web server at FZJ (wwwnet.deisa.fz-juelich.de) has been set up which generates first statistical reports of the router/switch interfaces where the supercomputer systems have been connected to. Additionally the server requests restricted web information from the NREN and GEANT switches and routers, extracts the DEISA relevant parts and makes this information available to DEISA.

The web server already provides first information concerning network throughput and stability and enables the network administration to monitor network utilization, throughput and stability of the dedicated DEISA network and its components. Additionally the reachability of the DEISA supercomputer nodes from the network point of view will be monitored. Furthermore network topology maps of connected sites have been added. Further links provide information about the connected supercomputer systems, configurations, documentations and system contacts as well as network contacts at the different DEISA installations.

The web server has been made available to DEISA "proof of concept" sites and networking staff at GEANT and NRENs. Currently it acts as an information point for the joint DEISA/NRENs/GEANT network team only. Later on, the information relevant to DEISA users will be publicly available. Additional information will be restricted to authorized persons only.

3. Network security considerations for DEISA

The DEISA project is structured into different service activities. Many of these activities will have some overlap within the area of their tasks and milestones. One of these overlaps will be the security area. Since DEISA has its own service activity "SA5 - Security", it has to be precised which kind of security has to be realized by which service activity. Some of this different flavours of security are system security, data security (ownership, integrity, ...) user security (system access, ldap info, certificates) and network security.

Firewalls, working mainly on layer 3 and 4 of the ISO-OSI reference model, have not been addressed directly in the DEISA project. High speed and dynamic firewalls are of main interest in current research projects and therefore are currently not available. So another security model was realized.

DEISA assumes a net of trust. Every user having access to a supercomputer resource at one DEISA installation is assumed trustworthy to other DEISA sites. Since packets traversing the DEISA backbone can only be initiated from a DEISA supercomputer, where only reliable users have access to this source is thought of as a secure one. If an attack is initiated from one DEISA partner to another partner, this attack has to be seen as insider attack and has to be managed by both security teams of these sites. Therefore a security help desk has been organized in SA5.

To prevent attacks coming from other non DEISA sites, firewalls at the "normal" traffic communication lines will be used.

Packets traversing DEISA communication paths will be checked by access lists configured on NREN switches/routers as well as on the local DEISA sites switches/routers. So an additional security barrier has been established denying packets not deleted by external backbone routers (NRENs).

4. Network demonstration

The operational network will be demonstrated at the Project review Meeting in June. There, live data transfers, synthetic ones with the network performance test program iperf as well as real application streams, e.g. GPFS, will be shown. The network utilization depending on setting of network options will be shown too.