The DEISA project – Network operation and support – first experiences

Ralph Niederberger and Olaf Mextorf

Research Centre Jülich, D -52425 Jülich, Germany
e-mail: {R.Niederberger/O.Mextorf}@fz-juelich.de

Abstract: The principal objective of DEISA[1,2], a consortium of major national supercomputing centres in Europe, is 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. The infrastructure, based on the tight coupling of homogeneous national supercomputers, provides a distributed supercomputing platform operating in multi-cluster mode. The production capability and capacity of the distributed platform will provide a substantial European added-value to the existing national infrastructures and will, in turn, be integrated into a larger heterogeneous Grid. The DEISA service activity "Network operation and support" defines, deploys and operates the network interconnect required by this infrastructure during the project lifetime and beyond. The paper will give an overview about the activities performed in this service activity, the experiences that have been gathered and the future tasks which have to be performed.

Key words: European distributed multi-cluster supercomputer, Grid infrastructure, GEANT IP Premium service, virtually dedicated bandwidth network interconnect

1. MOTIVATIONS

After a long period in which high-performance scientific computing drifted towards the use of "commodity off-the-shelf" hardware designed and optimised mainly for general purpose or commercial applications, a new change of paradigm has taken place. The emergence of the Japanese Earth Simulator (a vector architecture specially designed for scientific computing) [1] as the world's leading supercomputer, had triggered a profound revision of strategies and priorities in high performance computing (HPC). The United States launched several initiatives for stimulating the emergence of new, science-driven computer architectures, motivated by the simple goal of achieving maximum sustainable performance for scientific applications. The deployment of a few efficient and very high-performance platforms was planned to challenge the Japanese leadership. At the end of 2004, these new systems had been installed, so that the Earth Simulator, having been No. 1 for about 2.5 years in the TOP 500 supercomputer list, rated only no. 3 in November 2004 [2].

DEISA (Distributed European Infrastructure for Supercomputer Applications) [3] is a consortium of eleven major national supercomputing centres in Europe. The partners in this project consider it their mission and their responsibility to provide visionary leadership in the area of high-performance computing in Europe. Therefore, the consortium decided to create and operate a distributed terascale supercomputing facility, whose integrated power has been close to 30 teraflops in early 2005. The principal objective of this Integrated Infra-structure Initiative (I3) [4] is 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 appropriate software and with dedicated high-speed networks, represents a way open, in the immediate future, for innovative and creative thinking that enhances the impact of existing infrastructures. It is assumed that this approach will provide answers to some (if not all) very demanding requirements arising from the scientific community.

The DEISA infrastructure is based on the tight coupling of homogeneous[3] national supercomputers, to provide a distributed supercomputing platform operating in multi-cluster mode. The production capability of the distributed platform will provide a substantial European added-value to the existing national infrastructures. With the addition of other leading multi-teraflop supercomputing systems, the created hetero-geneous super-cluster will be a unique European supercom-

---

[1] DEISA (Distributed European Infrastructure for Supercomputer Applications) funded in part by the Eurpean Commission inder grant 508830.

[2] DEISA partners: BSC, Barcelona, Spain; CSC, Helsinki, Finland; CINECA, Bologna, Italy; EPCC, Edinburgh, UK; ECMWF, Reading, UK; Research Centre Jülich (FZJ), Jülich, Germany; HLRS, Stuttgart, Germany; IDRIS – CNRS, Orsay, France; LRZ, München, Germany; Rechenzentrum Garching (RZG), Garching, Germany; SARA, Amsterdam, The Netherlands.

[3] with respect to architecture and system software

puting platform as well as an extended heterogeneous super-computing grid. This distributed multi-cluster platform will, in turn, be integrated into the larger Grid.

Leading scientists across Europe will use the bundled supercomputing power and the related global data management infrastructures in a coherent and comfortable way. A special focus is set on grand-challenge applications from key scientific areas like material sciences, climate research, astrophysics, life sciences and fusion-oriented energy research. In 2005, the focus has been enhanced by defining the DEISA Extreme Computing initiative [5]. The DEISA research infrastructure will also be open, under certain conditions, to users of non-member organisations.

DEISA plans to collaborate with a large number of projects or institutions in Europe. The first priority is co-operation with other FP6 infrastructure projects in HPC (HPC-Europa [6]) or Grids (EGEE[7]). DEISA also intends to deploy very close collaborations with other projects emerging in the R&D area – in particular, projects funded in the "Solving Complex Problems with Grids" call – and with projects providing European Grid technologies (like UNICORE).

The United States is also deploying a technology-driven, grid-empowered infrastructure, TeraGrid [8]. In this project, five computing systems across the country are integrated to form a unique, single system platform using very high-bandwidth, dedicated networks (10-40 Gb/s) to guarantee highest performance. DEISA can be seen as the European counterpart to TeraGrid. Though implementation strategies are not identical, and TeraGrid uses Linux clusters whereas DEISA integrates high-end national supercomputers, very close collaboration with TeraGrid has been initiated.

2. THE EUROPEAN SUPERCOMPUTING INFRASTRUCTURE

DEISA is technology neutral, and no technology commitments of any kind have been made. The systems integrated into the DEISA infrastructure have been chosen mainly because of site specific strategic issues. Technology choices follow from the system capability to adapt to a pre-established operational model, and to provide real services to end users.

The DEISA supercomputing environment is sketched in Fig. 1. It has been structured into two layers. The "core" project deploys an "inner" research infrastructure, which is the distributed European supercomputing facility. This part of the project is strongly dependent on the deployment of an internal, dedicated high-bandwidth network interconnect between computing platforms, and is strongly focused in HPC. It uses those Grid and multi-cluster technologies that are relevant to achieve large-scale, tight-system integration of a super-cluster at the continental level. In this core infra-

structure, state-of-the-art Grid technologies are working transparently in the background and are not directly seen by the end users. Using the standard high-bandwidth connectivity of the supercomputer sites leads to increased complexity. These paths are also used by normal communications of the sites. It would have been necessary to reserve, simultaneously, networking resources in additional to the reservation of processors and storage resources needed. Software interfaces and accounting procedures to request communication paths, ad hoc or in advance, have not been available but those interfaces are the main focus of current networking projects just under consideration (*e.g.*, [9], [10]). Therefore, a dedicated network interconnect infrastructure was chosen. It allows unconstrained and undisturbed use of predefined communication bandwidth. This also simplifies the detection of bottlenecks and the evaluation of the interaction between the involved supercomputer systems, applications and the network itself.

The "extended" project deploys an "outer" research infrastructure, whose purpose is to efficiently interface the European distributed supercomputing facility to other supercomputing infrastructures in Europe, for example, through traditional middleware like Globus [11] and (or) UNICORE [12]. This requires the operation of usual Grid services with other partner HPC sites, infrastructures or Grid research projects. Here, the DEISA distributed facility is just an element in a larger Grid, and the relevant Grid services are those which provide added value to a larger HPC scientific community. Nevertheless, the focus is maintained on high-performance computing, and the services deployed are selected and identified on the basis of this criterion. These Grid services are supported by the standard high-bandwidth connectivity for general scientific users provided by the National Research and Education Networks (NRENs) and GÉANT.

The "core" supercomputing platform provides a single system image of an initially homogeneous super-cluster, constituted of several IBM P690, P690+ and P655+ national systems, as well as global data management at a continental scale. This super-cluster is homogeneous in the sense that the basic processing units are IBM processors and the operating system is AIX, but the national configurations integrated in this super-cluster are different. They mix 32 processor P690 and P690+ nodes, and 4-8 processor P655+ nodes. The Phase 1 initial core platform is composed of 3660 processors, each with a peak performance of at least 5 Gigaflops (FZJ, 1312 processors; IDRIS-CNRS, 1024 processors; RZG, 812 processors; CINECA, 512 processors).

The second generation Phase 2 core platform will benefit from several improvements. Along with additional IBM systems a heterogeneous extension has been started by the integration of huge Linux systems, the first one being the Linux system at SARA having been connected to DEISA in
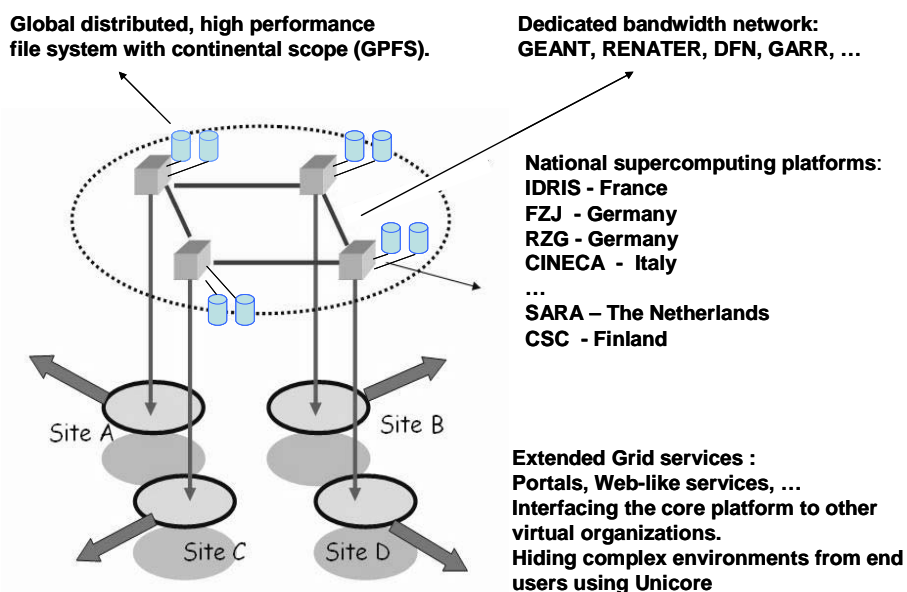
Global distributed, high performance file system with continental scope (GPFS).

Dedicated bandwidth network: GEANT, RENATER, DFN, GARR, …

National supercomputing platforms:
IDRIS - France
FZJ - Germany
RZG - Germany
CINECA - Italy
…
SARA – The Netherlands
CSC - Finland

Extended Grid services :
Portals, Web-like services, …
Interfacing the core platform to other virtual organizations.
Hiding complex environments from end users using Unicore

Fig. 1. The DEISA supercomputing environment

May 2005. Other leading national systems are scheduled to be integrated into the core platform.

3. THE "CORE" PROJECT INTEGRATION STRATEGY AND OPERATIONAL MODEL

The core infrastructure relies on software that operates underneath standard Grid middleware like Unicore and/or the Globus Toolkit. The distributed file system used provides global data management and high-performance distributed supercomputing at a continental scale.

Global distributed file systems are sophisticated software environments necessary – but not sufficient – to provide a single system image in a clustered platform. They avoid data replication at each computing node, because they allow data to be shared by all nodes. The initial Phase 1 facility, being a homogeneous IBM system, uses a simple extension of IBM's GPFS (Global Parallel File System) [13]. A supercomputing cluster operating in a DEISA site is composed of a huge amount of autonomous computing nodes each with their own operating system, a shared large disk space and an internal, often proprietary, network connecting all these elements.

To deploy a European supercomputing cluster, DEISA just extends this idea to a continental scale by adding one extra layer of dedicated network connectivity across computing nodes residing in different national sites, and by deploying a global file system with continental scope. Users will not see cluster internals. They will have a unified view of the whole environment. Data sets can be accessed on all

computing nodes independent of their location and transparent to the user. Therefore, a user does not need to know on which nodes its application is executed and where the data resides.
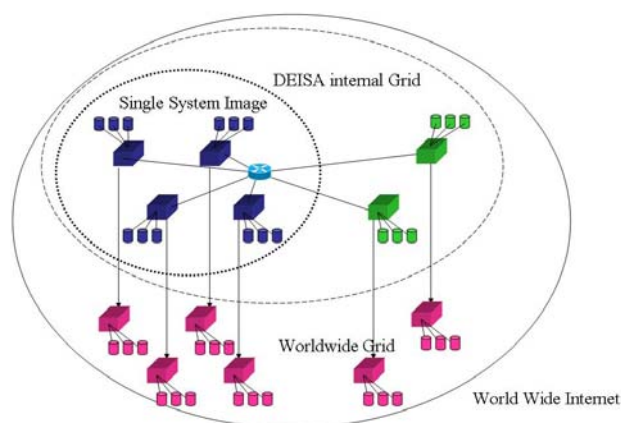

Fig. 2. DEISA and the grid

The "virtual" European centre has been structured like a real one. First of all, a number of classic trans-national Service Activities, each one lead by a different site, have been provided: Network Operation and Support (led by FZJ), Data Management with Global File Systems (RZG), Middleware (CINECA), Applications and User Support (IDRIS) and Security (SARA). A number of advisory committees work in co-operation with the project management: a users committee

for quality of service feedback, a scientific committee for advice on scientific policies and resource management and an advisory committee on technology issues.

4. BENEFITS FROM DEISA'S OPERATIONAL MODEL

The operational model described and sketched in Figure 2 above allows the co-operative operation of user groups across national borders. The DEISA global data management offers trans-national scientific projects management of significant amounts of data and computing cycles without knowledge of the specifics of internal implementation.

The co-ordinated operation of several national computing platforms enables the execution of leading-edge applications that require huge amounts of computing power. Rather than doing meta-computing and deploying the application over several platforms (which requires a subtle co-allocation of resources and is dependent on bandwidth, delay and jitter), DEISA can allocate a large fraction of the resources of one site to a single job by relying on job migration, namely, by transmitting jobs from one site to another and so freeing resources at one site. The workload can be balanced on a European computing resource pool, allowing the operation of national services even if very demanding DEISA jobs are running on one particular site. Because of the availability of a high-performance distributed file system, explicitly moving the data used by the applications can be avoided. Due to the DEISA operation model with transparent file access, applications can run "as is" and do not need to be adapted to a Grid to operate at an extreme performance level. The "virtual" European supercomputing centre" is not very different from a "real" one.

In high-performance-capability computing, the DEISA infrastructure can deal with distributed multi-physics, multi-scale, multi-component applications involving coupled software modules modelling different physical phenomena. So, leading-edge grid applications distributed by design, which can benefit from a heterogeneous super-cluster, will manifest different levels of parallelism, but with moderate amounts of data flow at the highest (distributed) level.

5. NETWORK OPERATION AND SUPPORT

The main task of the DEISA service activity, "Network operation and support", is the definition and deployment of a dedicated network required by this infrastructure, in close collaboration with the European network providers. Further-more, the operation of the dedicated network infrastructure as a production network has to be guaranteed throughout the project lifetime and beyond. Optimisation and upgrade of

the network including tests of novel network infrastructures will be important tasks of this service activity. Deployment of network monitoring procedures and services to monitor the network and send alarms in case of malfunctions, including information about network utilisation, throughput, one-way delay, jitter, communication errors and error recovery, will be additional tasks.

5.1. Management of the DEISA network infrastructure

The connection of the supercomputer systems to the national/international network infrastructure provided by GÉANT and the NRENs has been based primarily on predefined connectivity constraints given by the providers as well as the locally established and configured infrastructure so that a smooth production environment for the single system image could be guaranteed.

The integration of a virtual supercomputer infra-structure, single system image into an existing local supercomputer environment running in production mode requires a close collaboration with the supercomputer system administrators and the operations team. Knowledge of specific supercomputer communication internals had to be achieved to guarantee optimum performance of the local system as well as optimum performance in distributed mode of operation and network throughput. Changes and upgrades in production environments are very delicate operations, because of the necessity of pre-serving the quality of services.

A main task is to analyse the different implementations of network interfaces, so as to specify the very best solution for every installation site. This implies a steady contact to the supercomputer and networking equipment vendors, to remain informed about new supported interfaces and protocols or to trigger the development of techniques and protocols needed in future. This task will be of importance all over the project life time because additional partners with different supercomputer environments have to be added and the bandwidth upgrade from 1 Gb/s to 10 Gb/s or more has to be carefully planed and safely implemented.

Monitoring procedures have been installed as required in every production environment, which inform about network infrastructure utilisation, occurrences of communication errors, error recovery and so on. These monitoring facilities are available 24 hours 7 days a week for the whole inter-connect from machine to machine. This 24/7 alarming facility is independent of to-be-defined reaction times in the case of an error. Given the special nature of the DEISA infrastructure, consisting of site specific as well as national and international network components, a close interaction between DEISA and GÉANT/NREN staff had to be established. In future, a network support centre will be set up where system admin-

istrators as well as users can report problems or get help concerning communication questions. This support centre has to localise, analyse, extract and eliminate reported problems all over the project life time. Successful operation requires deep knowledge of all protocols layers of the ISO-OSI reference model, including specifics of applications and services used throughout DEISA.

A special Web site is the central contact point giving information whom to contact within the DEISA network team if any problems arise. Similarly a NOC has been setup handling network problems, dealing with alarm generation and status monitoring. So the DEISA project accommodates its role as a unique single European project.

5.2. Motivating a "dedicated" infrastructure

Computing architectures of all kinds have today a profound mismatch between the processors theoretical capabilities to process data ("peak performance") and what happens in real, production systems ("sustained performances"). Sustained performance depends on how efficiently the processors can be fed with data to process. The sustained performance of a computing system is therefore essentially determined today by the internal network interconnect bandwidths (processor to memory, processor to processor, node to node, and national cluster to national cluster in the DEISA core infrastructure). Therefore dedicated, high-bandwidth networks or networks providing bandwidth-on-demand are needed for high performance in world-class production systems.

Developing and deploying coupled applications – enabled to take advantage of multiple computing platforms – is an activity at the frontiers of software engineering, to which DEISA will pay undivided attention. But this class of meta-computing applications will not, for some time, be able to provide an important production workload. As the national supercomputers of the DEISA nodes get bigger and bigger, it is even possible that they will never contribute an important production workload. Nevertheless, these applications must be fully supported, but they are not sufficient to provide a complete justification for a dedicated high-bandwidth interconnect.

GPFS can be tuned to take full advantage of the dedicated network, for high performance, reliability and quality-of-service. Indeed, the access to data is transparent, but data must, nevertheless, be moved to and from the computing platforms. High performance is crucial here, to prevent sustained access of remote files from producing computational bottlenecks.

Because of GPFS, it is not necessary to transfer executables and data files in order to move a job from one execution platform to another. This opens the way to a simple and efficient global management of a European pool of computing resources. Site A, for example, can transfer a substantial num-

ber of its small memory jobs to site B, in order to be able to allocate important resources to a big and demanding job coming from DEISA users that are sharing resources on the whole core platform. Enabling very demanding simulations by job re-routing to other sites is, contrarily to meta-computing, likely to contribute substantially to a production workload. This mode of operation can hardly be imagined without a network of trust on top of dedicated network interconnects.

The current national and international interconnects are based upon dedicated bandwidths to national providers (NRENs). A predefined bandwidth is supported for national and international communication purposes.

Differentiation between the different services (applications) and priorities is done on a very rough basis. Furthermore, it is not guaranteed that a special predefined bandwidth can be assured since backbone paths are not implemented in an adequate manner. Packets are transmitted on a best effort basis. This situation has an even larger impact on international lines because of the increased line lengths.

This framework leads to bandwidth sharing with competing applications using the same communication paths on the international provider lines. Though GÉANT currently provides a very good international communication throughput, often international communication lines support lower communication throughput than the one leased by research organisations from their national NRENs. Even in the presence of the full bandwidth of the international links, the overall communication throughput is reduced.

Currently, it is not possible to establish network connections ad hoc, using automated bandwidth reservation mechanisms within the international networking environment, which would allow an immediate usage of remote supercomputer resources. In a meta-computing context, it is necessary to reserve supercomputer resources in advance at particular predefined times (batch model). The simultaneous reservation of networking resources to achieve undisturbed communication between involved processes (processors) on the distributed supercomputer systems or to access remote storage space leads to a significantly increased complexity for the reservation task. Software interfaces and accounting procedures to request such ad hoc communication paths haven't been developed until now in an international context.

Therefore a dedicated environment is a fundamental requirement to evaluate the interaction between the involved supercomputer systems, applications and the networking environment. Since bottlenecks concerning the offered services can be found within the computing as well as communication environment, the dedicated environment allows the removal of bottlenecks by the network team, which are locatable in a reproducible environment only. Since a real, dedicated infrastructure would have stressed the project budget extremely, a virtual solution based on Premium IP (see chapter 5.4) has been chosen.

5.3. The security model

For a decade, the security model of a firewall has become familiar to secure the local network against remote attacks of hackers. Every institution has resources which they do not want to share with external unauthorised users. Security policies have been defined by all institutions which may differ slightly from other ones. Every additional connection to the outside world opens a new entrance to the resources of the local institution. IT security officers are aware of these additional risks and try to keep these communication "backdoors" small to limit the risk of potential hacking attacks. The problem which arises is that this new links challenging new science models, also use high bandwidth communication links. These links often can not be processed by the installed firewall systems, since they would have to be upgraded with additional high bandwidth communication interfaces, additional or more powerful central processing units and often also with new system software (beta software) allowing enhanced features. These changes to the security systems are very expensive, being on the front end of security system research, and undesired for a production environment. So other security design models have to be found.

The security model of a single system image super-computer spread across Europe differs drastically from a Grid security model. In a Grid environment, communication relationships vary extremely during time. Ports used may change with every connection. Within a single system image, we can assume a net of trust". We know the involved partners in advance. They do not change over time. We have a closed system group (known IP and MAC addresses). So knowing the involved communication partners allows the configuration of this communication relationship in advance in the local router/switch systems, so that access lists can be used to secure connections across these communication paths. We do not have to use expensive high performance firewalls. A cost saving router, functioning as a static firewall system, can be used.

Another problem often thrown into discussion is the fact that a system could have been hacked. Including external systems, the remote supercomputers, into the local security zone by allowing uncontrolled connectivity, decreases the local security level to the level of the outside system. But this argument does not take into consideration that all users of the remote system are nevertheless allowed to login into the local system because of the single system image model. So having real firewalls with restricted access rights for the external supercomputer systems would give no additional security because every authorised user may open up a tunnelled connection between the two systems, remote and local, giving access to the local supercomputer system. So the main point here is that every installation has to have

trust in the security model of the other DEISA partners. A "net of trust" has to be assigned.

Though DEISA assumes a net of trust and additionally assigns router ACLs to regulate communications, a main task of the service activity "Network operation and support" will be searching for alternative, more secure solutions as *e.g.*, dynamic firewalls to minimise open ports, load-balancing firewalls to allow cost saving high throughput and additional security models to realise secure communication in a single system-image environment spread across different administrative security domains.

5.4. The "Phase 1" network

The DEISA consortium is, of course, aware that a fully-meshed topology involving point-to-point connections across all platforms can hardly be considered, because of the costs of international communication lines. A star topology design having a switch at one location connecting all other platforms directly to the switch would be preferable, because of the obvious symmetry of all computing platforms in terms of connection bandwidth. A slightly less optimal solution is to implement a ring network having PoPs (SubPoPs) connecting every DEISA site to the ring. This solution would be sub-optimal because of bottlenecks, but acceptable. A line topology having the sites as pearls on a chain is definitely not adequate, because of so many induced bottlenecks.

Our strategy was to implement a "proof of concept" phase – which included only four sites – with a star topology, and at the same time to conduct a deeper analysis of a start topology for Phase 2.

A switch/router has been installed at every DEISA site to which the local supercomputer system has been connected. Preferably, the provided local interface is one or more Gigabit-Ethernet (Etherchannel) and will be 10 Gigabit-Ethernet at phase 2. Ethernet interfaces are recommended, to facilitate the upgrade of the network facility.

Application requirements on delay and jitter have not been specified until now. Delay and jitter depend on the hardware used within the international network infrastructure, on the design of the infrastructure itself, and on the characteristics of the applications used and implemented within the DEISA service activities. A main task for the future will be to evaluate delay and jitter dependencies against application requirements. Nevertheless minimum jitter and delay will be inevitable.

The deployment of the dedicated DEISA network infrastructure proceeds in several steps and follows the evolution of the national and European research network infrastructures and the adoption of the infrastructure by the user communities. During the "proof-of-concept" a 1 Gb/s network infrastructure connecting the four supercomputer systems at CINECA, FZJ, IDRIS and RZG has been implemented. After

the four core partners had verified that the DEISA concept was sustainable and an extension to include other sites has been scheduled. In the meantime, a fifth site, SARA has been connected. CSC will be connected soon. Additional sites which will be connected in the future are ECMWF, BSC, and LRZ, so further NRENs are also involved.

The "phase 1" network has been implemented by using the Premium IP service [14] provided by GÉANT and the national NRENs, DFN in Germany, RENATER in France and GARR in Italy. This service is comparable to virtual leased lines.

Premium IP is a service that offers some network traffic classes priority over other traffic. It gives priority over all other services, such as Best Efforts (BE) and Less Than Best Efforts (LBE). During times of congestion, Premium IP traffic receives a guaranteed level of network performance. This can be particularly useful for real-time applications, such as Voice over IP (VoIP) and videoconferencing, and especially the GPFS traffic within DEISA.

Data packets that are sent using the Premium IP service will experience no congestion in the network regardless of the load of the other traffic classes. As a result, delay and packet loss are kept to a minimum.

In order to effectively send and receive data using Premium IP, each network on the end-to-end path travelled by

The offered Premium IP service provides upper-bounded one-way delay, upper-bounded Instantaneous Packet Delay Variation (IPDV), no packet loss due to congestion and guaranteed capacity. For DEISA a throughput capacity of 1 Gb/s is guaranteed.

The available network capacity for Premium IP traffic is currently monitored for each link. The TF-NGN Performance Monitoring activity [16] is currently working on a monitoring infrastructure to measure one-way delay, IPDV, and packet loss. As soon as this monitoring information is available it will be included into the DEISA monitoring procedures.

Prior usage of the Premium IP service has to be negotiated with GÉANT. Additionally every Premium IP packet has to be tagged with a special DSCP[4] value. Any packets tagged as Premium IP without prior reservation will be considered unauthorised and will be re-tagged as Best Effort by GÉANT.

The DEISA infrastructure topology within GÉANT has been designed as a mesh of LSPs[5] between NRENs. The local DEISA sites network connectivity has been implemented using additional connections to the local NRENs. Normal site traffic will not use these lines, which implies that no interferences have to be considered. The different implementations are based on real additional lines (fibres) or WDM
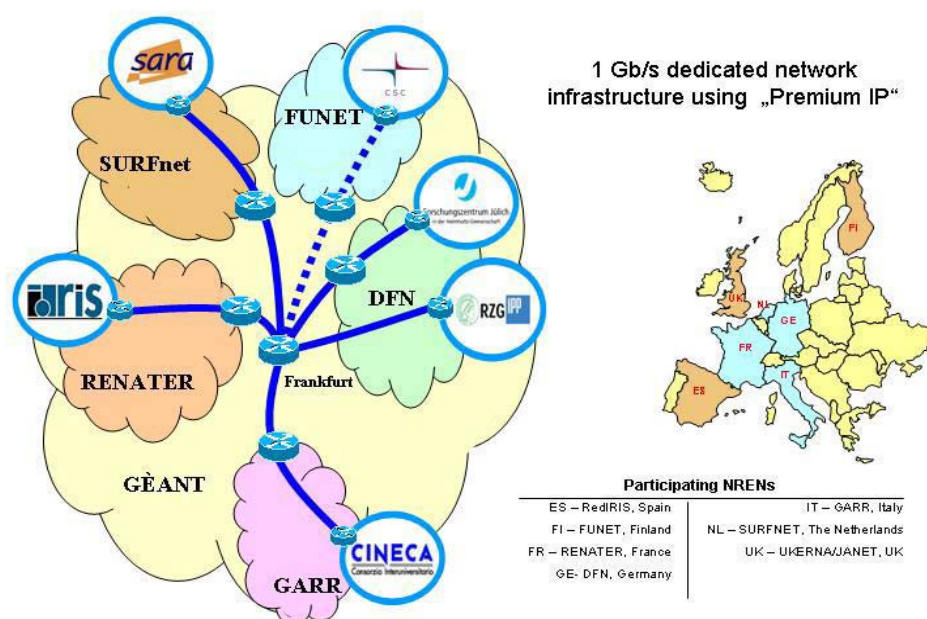

Fig. 3. European supercomputer cluster (Network infrastructure view, June 205)

the data must support the same Premium IP service model. As GÉANT interconnects all the European National Research and Education Networks (NRENs) involved in DEISA, it has implemented the Premium IP service model defined by the SEQUIN project [15].

---

[4] DSCP = Differentiated Services Code Point. In QoS, a modification of the type of service byte. Six bits of this byte are being reallocated for use as the DSCP field, where each DSCP specifies a particular per-hop behaviour that is applied to a packet.

[5] LSP = Label Switched Path

equipment depending on the locally involved and installed hardware.

The first connection tests showed that some reconfigurations were indispensable. As expected, redefinition of network options at the supercomputer nodes to get optimum performance was required according to the "bandwidth delay product" [17], having round trip delays of above 19 milliseconds. Rearrangements of supercomputer partitions adding additional CPUs, working on network load, led also to increased network throughput. Last but not least, some switch equipment had to be replaced by other vendors to circumvent some software bugs only observable in this special "WAN" environment. After having reconfigured the infrastructure accordingly, a throughput of about 900 Mb/s between the supercomputer sites at FZJ and RZG and between FZJ and IDRIS could be reached using the "iperf" [18] program developed by NLANR. Real applications, *e.g.*, ftp, scp or http surely will not be able to cope with these throughput values. However, this does not apply to GPFS. GPFS reads a dataset in parallel from a predefined number of I/O nodes and therefore generates multiple streams of data out of the system to a remote site. This implies that a summarising effect will take place, which leads to increased throughput related to a single file. Teragrid used this effect to fill up a network pipe of 30 Gb/s capacity using 40 I/O servers and 64 GPFS clients. First local DEISA tests showed already that a 10 Gb/s net-

planned to have a substantial upgrade of throughput capacity up to 10 Gb/s or more to some sites. Due to the diversity of the involved supercomputer systems and their internal configuration, the time of upgrade to the proposed bandwidths at the different locations may vary slightly. The upgrade will be driven by application needs and by the availability of the next generation GÉANT [19] and NREN infrastructures, *e.g.* X-WiN in Germany. First meetings and discussions with the providers have already been scheduled. This phase will be the most challenging phase, where technological requirements and application needs are at the upper limit of what providers can offer to a single supercomputer site.

The "phase 2" DEISA network will also be adapted to the deployment of the complete supercomputing grid, incorporating a number of leading heterogeneous platforms in Europe. Close collaboration between the DEISA network team and staff members of the NRENs, 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 and to achieve the ambitious goals of the project.

Since this network upgrade correlates in timing with DEISA phase 2, where the heterogeneous grid extension will be introduced, there will be an additional complexity within the network infrastructure of the general project.

ipqmaxlen=512	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=6553600	tcp_pmtu_discover=1	
thewall=1572864	tcp_recvspace=2621440	
	tcp_sendspace=2621440	

Fig. 4. Optimised network options used at the DEISA supercomputer sites

work connection can be fully utilised, at a minimum. To achieve this throughput in a WAN environment, additional local network rearrangements at the DEISA supercomputer systems have to be made in the future.

The "phase 1" network infrastructure is expected to operate for a maximum of 18 months. It is planned that the "phase 2" infrastructure will operate at 10 Gb/s or more.

5.5. The "Phase 2" network

At the beginning of the second phase, additional super-computing centres will be added, most likely with 1 Gb/s connection bandwidth, where adequate. In parallel, it is

The inner research infrastructure will overlap with the outer research infrastructure. The European Grid infrastructure will become a reality.

5.6. Future enhancements and co-operation

The DEISA project is aware of the many other activities started in parallel all over the world. As described earlier, DEISA can be seen as the European version of TeraGrid. So it is natural to have close collaboration with this project. First contacts have already been established to find research areas where both projects can interact and complement each other. It is planned to show a first demonstration of inter-

action between both projects at the Supercomputing 2005 conference.

Another project being of great interest for DEISA is the EU project EGEE. Joint conferences have already been organised. In the future, additional interaction will be scheduled.

The GGF[6] has initiated a lot of working and research groups addressing optical infrastructures for Grids [20] and networking issues for Grid infrastructures [21], raising questions and trying to give answers to some of them. These are also relevant for the design of a virtual European supercomputer. A task force/research group working on dynamic firewalls is about to be initiated. The outcome of these activities will be of interest to DEISA also.

Flexible bandwidth broker architectures with the goal to incorporate the network as a manageable resource into a Grid resource management infrastructure have been considered for some years [22]. The Path Allocation in Backbone Networks [23], funded by the German NREN (DFN), explored advanced mechanisms for providing service guarantees under the particular constraints of an IP-based backbone. Additionally, networking projects researching bandwidth-on-demand protocols and reservation systems for bandwidth resources have started in 2004, *e.g.*, VIOLA [9] and Mupbed [10], and will be of great interest for the networking group of DEISA. Contrary to the LHC project at CERN [24] where continuous streams of experimental data have to be stored on remote sites and therefore to be transferred over a WAN, the communication traffic of DEISA will be more ad hoc. So, having scheduled networking resources and privileged traffic classes would circumvent dedicated communication lines hired in advance and give much more latitude for fruitful project enhancements.

Last but not least, networking security will be of increasing interest because of a static configured security model failing when additional communication relationships will show up. A dynamic firewall design will be indispensable. Here we see some projects coming up currently also, where DEISA will look onto with great interest.

6. CONCLUSION

DEISA intends to contribute to the significant enhancement of capabilities and capacities of high-performance computing (HPC) in Europe by the integration of leading national supercomputing infrastructures. To realise this mission, it deploys and operates the distributed multi-terascale European computing platform based on the strong coupling of existing national supercomputers and operates as a virtual European supercomputing centre. Interfacing the DEISA research infrastructure with the rest of the European IT infrastructure by adopting Grid technologies contributes to the development of an extended, heterogeneous Grid computing environment for HPC in Europe. To achieve these challenging scientific goals, future national and international research networks will be indispensable prerequisites. Cost-efficient high-performance interconnects including future bandwidth-on-demand services will be the basis. NRENs and GN2 are on the way to do their job. Having a world-class, high-bandwidth multi-protocol network infrastructure within Europe will push forward science to new dimensions.

References

[1] The Japanese Earth Simulator, http://www.es.jamstec.go.jp/esc/eng/index.html

[2] TOP 500 supercomputer sites, http://www.top500.org

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

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

[5] DEISA Extreme Computing Initiative, http://www.deisa.org/grid/initiative.php

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

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

[8] TeraGrid, homepage, http://www.teragrid.org

[9] VIOLA – Vertically Integrated Optical Testbed for Large Applications, http://www.viola-testbed.de

[10] Mupbed – Multi-Partner European Test Beds for Research Networking, http://www.ist-mupbed.org

[11] The Globus alliance, http://www.globus.org/

[12] D.Erwin Ed., Unicore Plus Final Report, Unicore Interface to Computing Resources, ISBN 300-011592-7, 2003

[13] 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

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

[15] DANTE, SEQUIN Project – SErvice QUality across Independently managed Networks, http://archive.dante.net/sequin/

[16] N. Simar, Performance Monitoring – TF-NGN Meeting, Lissabon, Sept. 2004, http://www.dante.net/upload/pdf/NS-TF-NGN-Sept_04.pdf

[17] M. Mathis, R. Reddy, J. Mahdavi, Advanced Network computing – Enabling High Performance Data Transfers, http://www.psc.edu/networking/projects/tcptune/, March 2005

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

[19] About GÉANT2, http://www.dante.net/server/show/conWebDoc.1202

[20] D. Simeonidou, R. Nejabati (Editors), Optical Network Infrastructure for Grid, GFD-I.036, http://forge.gridforum.org/projects/ghpn-rg/, August 2004

---

[6] GGF = Global Grid Forum, http://www.ggf.org

[21] V. Sander, Networking Issues for Grid Infrastructure, http://www.ggf.org/documents/GWD-I-E/GFD-I.037.pdf, Nov. 2004

[22] V. Sander, Design and Evaluation of a Bandwidth Broker that Provides Network Quality of Service for Grid Applications, NIC Series Volume 16, Forschungszentrum Jülich, ISBN 3-00-010002-4, 2003

[23] M. Fidler, W. Klimala, V. Sander, Path Allocation in Back-bone Networks: Project Report, http://webdoc.sub.gwdg.de/ebook/ah/dfn/PAB.pdf, June 2004

[24] The Large Hadron Collider Project, http://lhc.cern/lhc/general/gen_info.htm

RALPH NIEDERBERGER studied Computer Science at the University of Bonn and got his graduate in 1987. He is a Senior Research Associate at the John von Neumann Institute for Computing and the Central Institute for Applied Mathematics of the Research Center Jülich in Germany. Over the last years he worked in several networking projects combining supercomputer systems over WAN networks. Currently Mr. Niederberger is the leader of the service activity 1, Network operation and support, of the DEISA project. Main interests of Ralph Niederberger are High-Speed-Communications, Internet-Security (Firewalls/Intrusion detection), Network Management and Administration.

OLAF MEXTORF is a Senior Network Engineer at the John von Neumann Institute for Computing and the Central Institute for Applied Mathematics of the research Center Jülich in Germany. With more than 15 years experience in networking Olaf Mextorf is responsible for the design, deployment and monitoring of the campus network at the Research Center Juelich, one of Europe's biggest research facilities. Furthermore he accounts for the high speed and legacy networking of Juelich's Supercomputers and is engaged in scientific projects like the German VIOLA optical testbed for advanced network services and Europe's DEISA infrastructure for a continental-scope supercomputing environment.