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

Technical Report

DEISA - cooperative extreme computing across Europe

Ralph Niederberger, Paolo Malfetti, Andreas Schott**, Achim Streit*

FZJ-ZAM-IB-2007-07

July 2007

(last change: 19.07.2007)

Preprint: to be published

(*) Consorzio Interuniversitario (CINECA) Bologna, Italy
(**) Rechenzentrum Garching, Germany

DEISA – cooperative extreme computing across Europe

Ralph Niederberger, Research Center Jülich, Germany, r.niederberger@fz-juelich.de
Paolo Malfetti, Consorzio Interuniversitario (CINECA) Bologna, Italy, p.malfetti@cineca.it
Andreas Schott, Rechenzentrum Garching, Germany, schott@rzg.mpg.de
Achim Streit, Research Center Jülich, Germany, a.streit@fz-juelich.de

Abstract

The European DEISA[1] project is an infrastructure of infrastructures – eleven of the leading European High Performance Computing (HPC) centers in Europe interconnected by a 10 Gb/s high speed network – that devised an innovative strategy to enable the cooperative operation of existing national supercomputing infrastructures. This initiative led to the deployment and operation of a world class, persistent, production quality, distributed tera-scale supercomputing environment with continental scope enabling scientific discovery across a broad spectrum of sciences and technologies.

The paper describes in detail the DEISA infrastructure and high speed network interconnect, its operation and management, as well as user support procedures. The second part explains the used Grid middleware and Grid security models as well as its integration into the local security policies. The third part of the paper gives a comprehensive overview about lessons learned. The paper closes with a summary and a vision on future DEISA project extension plans.

1. Introduction

The deployment and operation of a world class persistent, production quality, distributed tera-scale supercomputing environment providing extreme computing across Europe can not be handled by standard procedures of a single computing center. Furthermore a virtual European organization requires the establishment of virtual teams of remotely operating staff members interacting in close collaboration. DEISA is a supercomputing GRID-empowered e-Infrastructure, an European virtual organization that, integrates national High Performance Computing (HPC) infrastructures by means of modern grid technologies. DEISA main objective is to contribute to a significant enhancement of HPC capability and capacity in Europe for scientific and industrial communities.

The DEISA infrastructure requires the coordinated action of the different national supercomputing environments, services and staff members for both efficiency and performance of computing resources. Leading scientific users will have transparent access to this European pool of computing resources, therefore providing user support across national boundaries will be essential. The coordinated operation of the DEISA environment is tailored to enable new, ground breaking applications in computational sciences. These goals led to the DEISA Extreme Computing Initiative (DECI) started in 2005 which will be described in detail below.

2. The DEISA system and network infrastructure

The DEISA Consortium followed a top-bottom strategic approach for the deployment and the evolution of the infrastructure. The technologies chosen have been fully open and followed from the strategic requirements and the operational model of the DEISA virtual organization. Only some very basic strategic requirements have influenced

[1] Distributed European Infrastructure for Supercomputing Applications [DEISA]

This work is partially funded through the European DEISA project under grant FP6-508830. An extension to the DEISA project, namely eDEISA, is partially funded through the European DEISA project under grant FP6-031513. In this we do not distinguish DEISA activities from the eDEISA activities, given the fact that the latter is an extension of the former.

the deployment of the current infrastructure. First of all the necessity of fast deployment of the infrastructure has been essential. Also the coexistence of the European infrastructure with the national services had to be guaranteed, which requires reliability and non-disruptive behavior and approach. A third prerequisite has been user and application transparency, hiding complex grid technologies from users and minimizing application changes, because application development should not be strongly tied to an IT infrastructure.

The current DEISA supercomputing grid architecture meets these concerns by having an inner level, dealing with the deep integration and strongly coupled operation of similar, homogeneous IBM AIX clusters, which forms a *distributed European supercomputer* (called "*the AIX super-cluster*") and an outer level made of all heterogeneous systems. Such a grid of heterogeneous supercomputers and super-clusters is tight together in a looser federation pool of supercomputing resources by means of several top-level technologies layered one on top of the other: a 10 Gbps dedicated network, a wide area network shared file system, a production quality GRID middleware able to virtualise heterogeneous resources presenting them as homogeneous and an application virtualization layer. This top-level technologies GRID-empowered eInfrastructure now includes all the leading platforms in Europe including systems from IBM (different kinds, from SP systems, to Blue Gene/L, to cluster of PowerPCs), SGI and NEC, and Cray in the next future. A third level of integration will be visible in the future when external resources (compute, storage, data generation resources as telescopes, medical devices etc.) may be loosely connected to the existing DEISA infrastructure by means of the web (and not of a dedicated network), selected GRID middleware and other virtualization tools.

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

- FZJ-Jülich (Germany): P690 (32 processor nodes) architecture, incorporating 1312 processors, 8.9 teraflops peak.
- IDRIS-CNRS (France): Mixed P690/P690+ (32 processor nodes) and P655+ (4 processor nodes) architecture, incorporating 1024 processors, 6.7 teraflops peak.
- RZG–Garching (Germany): P690 architecture incorporating 896 processors, 4.6 teraflops peak.
- CINECA (Italy): P5-575 architecture incorporating 512 processors, 3,9 teraflops peak.
- CSC (Finland): P690 architecture incorporating 512 processors. 2.6 teraflops peak.
- BSC (Barcelona, Spain): 10240 Power PC 970 processors IBM Linux system, 94.2 teraflops peak.
- HLRS (Germany): NEC SX8 vector supercomputer, 576 processors, 12.7 teraflops peak.
- LRZ (Germany): SGI Altix Linux system, 4096 processors, 26.2 teraflops peak (> 60 teraflops peak in 2007 later this year)
- SARA (The Netherlands): SGI Altix 3700 Linux system, 416 Itanium-2 processors , 2.2 teraflops peak.
- ECMWF (International organization): Two P5-575+ clusters incorporating 2276 processors each, aggregated peak performance 33 teraflops.

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

- EPCC/HPCx (UK): P5-575 (16 processor nodes) incorporating 2560 processors, 15.36 teraflops peak.

DEISA uses an internal network provided by GEANT and the National research networks (NRNs) that connects the above supercomputers of the DEISA partners (not the sites) and offers reserved bandwidth. This internal network exists, of course, in addition to the standard Internet connectivity that each national supercomputer centre offers.

The deployment of the dedicated DEISA network infrastructure has proceeded in several steps, following the evolutions of the national and European research network infrastructures and the adoption of the infrastructure by the user communities. After four core partners had verified in a proof of concept phase that the DEISA concept is sustainable, an extension to the other sites was started.

In the first phase all centres have been connected via a virtual dedicated 1 Gb/s network provided by the National Research and Education Networks (NRENs) and the multi-gigabit pan-European data communications network ([GEANT2]). Two years of stable operation have proven the reliability of this concept.

An intermediate phase connecting the 1 Gb/s Phase 1 network and the future star-like configuration 10 Gb/s Phase2 network has already been initiated. This Phase 2 infrastructure will operate at 10 Gb/s between all DEISA sites. The upgrade is driven primarily by the availability of the new GEANT2 and local NREN infrastructures currently under construction. The Phase 2 is the most ambitious phase, where technological requirements and application needs challenge the limits of what providers can offer to the supercomputer sites. The starting design of the future DEISA network backbone is a so-called star network with 10 Gb/s bandwidth from DEISA sites all over Europe to a central switch located at Frankfurt/Germany. An expansion to a decentralized design with backup paths is considered for the future.

A close collaboration between the DEISA network team and staff members of the NRENs and GEANT2 as well as a close interaction with the administrators of the supercomputer systems guarantees optimum performance of the DEISA network to meet the needs of the user communities.

The current layout of the DEISA network is shown in figure 1 below.

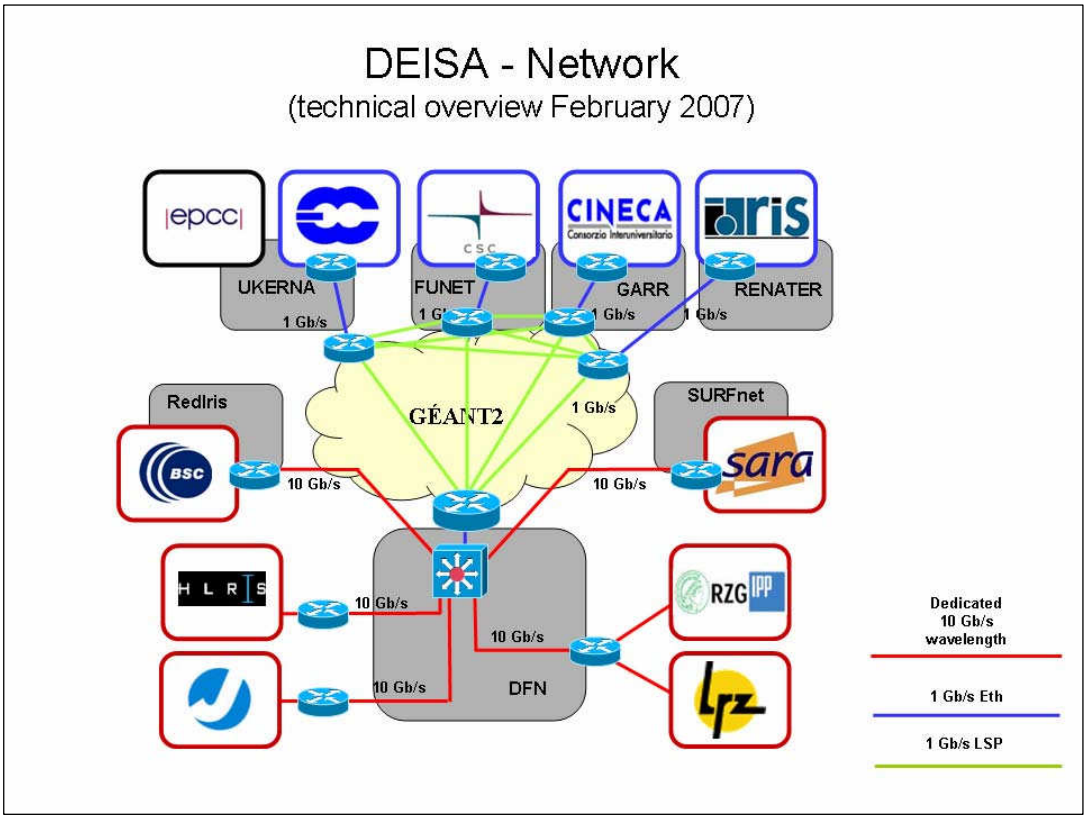


Figure 1: DEISA Network (technical overview February 2007)

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

An application running on one site can access data files previously "exported" from other sites as if they were local files. Therefore, it does not matter in which site the application is executed, and applications can be moved across sites transparently to the user. Though the concept of deploying network file systems is an old one, GPFS provides high performance remote access needed for high performance computing. In 2005 the TeraGrid project [Teragrid] has shown that on their 30 Gb/s TeraGrid network in the USA, GPFS is able to achieve about 27 Gb/s network

throughput when accessing remote data via GPFS. This proved that this software is capable of taking full advantage of underlying high performance networks. Therefore DEISA intended from the beginning to set up such a kind of high speed interconnect.

3. DEISA and its grid middleware

The strong integration of IBM AIX systems aims at providing a single system image of a distributed supercomputer. This is fully transparent to end users, which will access the super-cluster through the site in which they have a login.

The fundamental purpose of the AIX super-cluster operation is running bigger and more demanding applications than the ones that can be run today on each national cluster. One possibility of doing this would be to "grid enable" the application so that it can run on more than one platform. However, this strategy – that requires a modification of the application - does not really work for tightly coupled parallel applications. In this case, the finite signal propagation velocity induces MPI communication latencies in a wide area network that are intolerable for high performance computing.

DEISA adopts a different strategy, based on load balancing the computational workload across national borders. Huge, demanding applications are run by reorganizing the global operation in order to allocate substantial resources in one site. Therefore they run "as such" with no modification. Smaller jobs will be rerouted to other systems making room for large applications with huge amount of CPU, memory or disk resources.

DEISA global job scheduling within the "*the AIX super-cluster*" can be done via the Multi Cluster Loadleveler provided by IBM. This facility allows the definition of job queues, rerouting facility (jobs, not already started) and status information. Here it is possible to load balance DEISA supercomputer systems and mainly freeing huge systems by rerouting small jobs to smaller systems.

The other benefit of the AIX super-cluster comes from the possibility of transparently sharing data through GPFS. European data repositories that require frequent updates – like bio-informatics databases, for example – can be established in one site and accessed by all the others.

DEISA has a special service activity (SA3: Resource Management) to deploy and operate generic Grid Middleware needed for the operation of the DEISA supercomputing Grid infrastructure. The services provided include "basic services", which enable local or extended batch schedulers and other cluster-features to simplify user access to the DEISA infrastructure. These basic services are enhanced by advanced services which allow resource and network monitoring as well as information services and global management of the distributed resources. Examples for these services are harmonization of national job management strategies, deployment, test und update of middleware like UNICORE and Globus. Though most of these services are standard services in supercomputer environments, they have to be adapted to European distributed infrastructures. Similar adaptation activities have to be done in the batch scheduling software area.

DEISA has a "first generation" co-allocation service based on LSF Multi Cluster from Platform Computing which was extremely dependent on this particular technology provider. This new service allows a persistent integration of other future resources. eSA3 (the extension of SA3) focuses on deploying a "second generation" co-allocation service, that is vendor independent to enhance the co-allocation feature by extending the service to resources other than CPUs. Here the implementation of other types of co-allocation, as there are meta-job co-allocation or scheduled workflow co-allocation, may be promising.

User transparency is a necessity (users should not be aware of complex grid technologies) and applications transparency (minimal intrusion on applications, which, being part of the corporate wealth of research organizations, should not be strongly tied to an IT infrastructure). For this reason the UNICORE software is used as the middleware in the DEISA infrastructure to access the heterogeneous set of computing resources and managing workflow applications. Furthermore, in order to achieve interoperability with other leading Grid infrastructures in the world, new middleware is being evaluated, to decide about the best possible usage in the deployment of the infrastructure. Currently the DEISA partners are evaluating the current versions of Unicore 6, and Globus Toolkit 4 (GTK4). Only

fully production quality (with RAS features) middleware will be retained and integrated into the production environment. As the specifications of OGSA and related standards are likely to evolve also middleware interoperability needs to be ensured.

Nowadays, leading scientific applications analyze or produce large amounts of data. Some of these applications also need the computational capacities offered by the DEISA Grid. With GPFS DEISA has a very efficient tool for global data management, well adapted to High Performance Computing – but this technology does not cover all the global data management requirements. First of all, not all computing systems in DEISA can be integrated into the existing Global File Systems. Moreover, because of limited space on the DEISA global file systems, large amounts of data cannot be stored for an infinitely long time, and as a consequence data can not always be directly accessible from the applications running on the DEISA facilities. Before processing data they have to be transferred to a DEISA global file system or a local scratch system. Also, at the end of an application run, output data may have to be transferred to other storage facilities e.g. mass storage facilities of DEISA partners. Therefore DEISA has deployed a second high performance file transfer service based on striped GridFTP [GridFTP], which is also capable of taking advantage of the full network bandwidth for individual transfers. Last but not least, global file systems allow different applications to share the same data, but the opposite service is also needed: an application that needs to access a distributed dataset. Therefore the DEISA global data management roadmap focuses on the complementary objective of providing high performance access to distributed data sets, by enabling database management software like OGSA-DAI [OGSA-DAI] or grid storage software like SRB [SRB]. Moreover Grid based data transfers and Grid enabled data registration systems will provide DEISA users with facilities to retrieve and store data in Hierarchical Storage Management (HSM) facilities at DEISA sites and to register files independent of their physical location having global file names translated through registries and catalogues. Additionally it is planned to provide an uniform grid enabled access to specialized, in-house developed, or legacy databases by Grid enabled database access services independent of locations and formats of the databases. DEISA definitely will expand its data management capabilities in order to stay attractive for “grand challenge” applications.

The Applications and User Support Service Activity, both with the Enabling Key Applications Service Activity, is in charge of all actions that will enable or enhance the access to the DEISA supercomputing resources and their impact on computational sciences. It provides direct support to the major scientific initiatives of DEISA and helps users to run new challenging scientific applications, as there are large, demanding applications, running in parallel on several hundreds or thousands of processors in one specific site or multi-site Grid applications, running concurrently on several systems, so that each component is executed on the most appropriate platform as well as applications running at one site using data sets distributed over the whole infrastructure and multiple applications running at several sites sharing common data repositories. Additionally portals and Web interfaces used to hide complex environments from end users and to facilitate the access to a supercomputing environment to non-traditional user communities have to be supported.

To achieve these objectives, several activities have been deployed. The DEISA Common Production Environment (CPE) is running on all platforms of the infrastructure, with a very high coherency across the homogeneous super-clusters and a weaker one across heterogeneous platforms. DEISA CPE has been designed as a combination of software components (shells, compilers, libraries, tools and applications) available on each site and an interface to access these in a uniform way, despite the local differences in installation and location. CPE is automatically monitored checking its behavior continuously and identifying unexpected problems. User support also includes documentations on access and usage of the DEISA supercomputing environment as well as installing a decentralized Help Desk. Last but not least training sessions are organized to enable fast development of user skills and know-how for the efficient utilization of the DEISA infrastructure.

Enabling new challenging supercomputing applications is of key importance to advance computational sciences in Europe in the supercomputing area. The DEISA Extreme Computing Initiative (DECI) has been launched in 2005 to enhance DEISA's impact on science and technology. 29 applications have been selected on scientific excellence, innovation and relevance by a collaboration of HPC national evaluation committees and have been in operation in the 2005 – 2006 time frame with an aggregated number of 9.5 Mio CPU hours. A second European Call for Extreme Computing Proposals has been closed some month ago. 23 of these projects have been selected for operation in 2006 – 2007 time frame (aggregated 11.3 Mio CPU hours). 5 additional projects are in hold state and may be started as soon as computing time is available. A detailed list of the projects selected for both calls is available at “http://www.deisa.org/applications/”. The main purpose of the DECI initiative is to enable a number of “grand challenge” applications in all areas of science and technology. These leading, ground breaking applications must deal

with complex, demanding and innovative simulations that would not be possible without the DEISA infrastructure, and which benefit from the exceptional resources provided by the Consortium. These DEISA applications are expected to have requirements that cannot be fulfilled by the national services alone. To support these applications a special Applications Task Force, a team of leading experts in high performance and Grid computing, has been constituted. Its main task is helping users to enable the adoption of the DEISA Grid infrastructure and to enable a number of those Grand Challenge applications in all areas of science and technology. Therefore the Applications Task Force deals with all aspects of Extreme Computing projects from the European scientific community in its full complexity at all necessary levels, in order to enhance the impact on computational sciences.

Because of the application enabling activities provided by DEISA, a large number of scientific user codes, representing common applications for different scientific areas, have become available, and can be used to establish an European benchmark suite. These codes will be chosen to ensure a comprehensive coverage of major scientific areas as well as all the functionalities to be evaluated. These benchmarks will be focused on providing an efficient instrument for the acquisition of future European supercomputers. The benchmarking activity also includes the preparation of low-level tests that complement the user code tests.

4. VO and security

The DEISA grid infrastructure forms a conglomerate of diverse security policies. Within this virtual organization users need transparent access to the DEISA Grid infrastructure with single sign-on facilities. Vice versa partners need control on usage of their resources. These facilities, commonly referred to as Authentication, Authorization and Accounting services, must be trusted by all sites to protect their local environment against unauthorized access. Because of non direct contacts between users and remote DEISA sites dispatch services, a global administration had to be developed. Within DEISA a user only needs to contact a local administrator to get a DEISA POSIX (uid/gid) account. The user information will be stored into a LDAP-services database which allows to update local information consecutively every day on all DEISA systems in a secure manner.

A secure single sign-on is realized via X.509 certificates for authentication and authorization. DEISA trusts the certificates issued by the Certificate Authorities (CAs) accredited by the EuGridPMA, one of the members of the IGTF, a worldwide federation of CAs. This guarantees uniqueness of the certificates. Matching of uids and X.509 certificates allows the deployed Grid middleware to decide which services may be accessed. Because of the availability of the LDAP-information in all locations an XML-based database has been established which holds and presents all the relevant information for accounting. Aggregated reports will be created on resource usage by individual users and projects on a monthly basis.

The security of the DEISA infrastructure depends on the trust provided by the installed middleware that operates the services and on the security services that are used by the middleware as well as by the dedicated nature of the DEISA 10 Gb/s network infrastructure. Security issues related to networking below ISO/OSI layer 5, which is transport, network, link and physical layers are very low, because of the dedicated nature of the network. Connections established or packets inserted into existing streams can be done by already known individuals residing on DEISA hosts, assuming no DEISA system has been hacked. Nevertheless an insider threat attack could be started. Because of this the CERT[2] teams of all organizations have to work closely together and have to exchange any kind of security incidences as soon as possible. A mutual cooperative trustfulness concerning vulnerability assessment will be indispensable. Having had no security incidence within DEISA during the whole lifetime of the DEISA project attests that this faith into each other has been justified.

5. Management and operation

Operation and management within DEISA is done through the definition of a number of executive and working teams managing, delegating, and operating the required tasks. First of all the DEISA Executive Committee (DEC), constituted by managing staff members of all participating supercomputing centers, deals with higher level project

[2] CERT = Computer Emergency Response Team

management issues as technological strategies, scientific management, provision of computational resources, relations with other organizations and European projects etc.
The main focus of DEISA having a persistent and sustained supercomputing infrastructure required the participation of national scientific councils that establish national policies and provide funding for the national supercomputing services. The DEISA Policy committee is the board where representatives of these organizations can decide on major strategic issues e.g. strategic orientation of DEISA, models for service provisioning and global resource management. Additional advisory committees act as external consulting boards on technology issues (ATC), scientific issues (ASC), allocation of the DEISA computational resources, external analysis of the operation and quality of service assurance (AUC).
The DEISA Infrastructure Management Committee (DIMC) is the executive extension of the DEC. It acts as a direction team for the virtual supercomputing centre and its heterogeneous Grid extensions.
Underneath these committees the service activities (SAs) and joint research activities (JRAs) are responsible for the deployment and operation of the infrastructure. Currently the following SAs and JRAs are defined and operating:

Service Activities

SA1: Network Operation and Support
eSA1: Extended Network Infrastructure
SA2: Data Management with Global File Systems
eSA2: Operation of the Grid Infrastructure
SA3: Resource Management
eSA3: Extended Resource Management
SA4: Applications and User Support
eSA4: Application Enabling
SA5: Security
eSA5: User Interfaces

Joint Research Activities

JA1: Material Sciences
JA2: Cosmology
JA3: Plasma Physics
JA4: Life Sciences
JA5: Industrial CFD
JA6: Coupled Applications
JA7: Access to Resources in Heterogeneous Environments

The organizational structure of these service and joint research activities are depicted in figure 2.

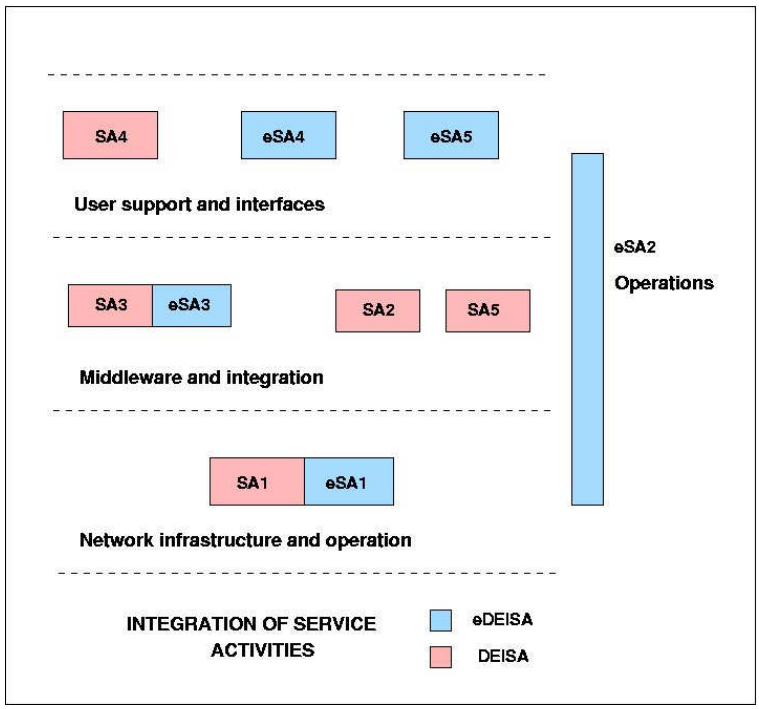
Beneath these organizational and management structures the eSA2 activity, named Operation Team, is responsible for a smooth operation of the research infrastructure.

The formation of the Operation Team is described next as well as the definition and setup of its responsibilities and work flows, by which the quality of the services provided through DEISA is assured.

The operation of an European wide virtual organization has to be done mainly by the involved service activities and local supercomputer persons in charge (local administrators, network and security experts). Although this works quite well in projects with a small number of participating organizations, an overall coordination team acting as link between the real organizations has to be established within greater projects. This team, coordinating local to global operations, providing guidelines for the overall virtual organization including task scheduling as maintenance outage, coordinated software upgrade, defining global security policies and advanced user administration, has turned out to be indispensable. Though all participating partners are always willing to pave the way for an advanced European wide virtual supercomputing infrastructure, nevertheless local priorities controvert these visions. At this point the operation team has to coordinate these contradictions and provide a way to overcome these congestions and advice local staff members to act accordingly.

Therefore the operation team is constituted of highly experienced system administrators of each site, who are able to coordinate the operational environments existing at each site as well as the interactions influencing the operational behavior of the infrastructure between the different SAs in DEISA. Beneath the overall operation it also acts as an emergency team, in cases where a quick reaction on problems arising in the DEISA infrastructure is needed.

Figure 2:
Organization of service and joint research activities



All service activities have their own well defined communication and operating channels concerning their own activity. But e.g. the setup of the Global File System GPFS by service activity SA2 requires specific network settings. Most of these parameters have been exchanged directly between SA1, providing the network, and SA2, responsible for the file system. Additionally there is the need of changing configurations in the firewall, which are seen as security critical by some sites. Therefore often decisions and actions of one service activity have an influence on another service activity. Thus coordination effort between different service activities is required on a case by case basis. The Operation Team is the instance to establish this.

Though all the DEISA sites operate in principle autonomously, the actions at one site may have an influence at another site. When a site is going into maintenance, its part of the Global File System will not be available at any other site. So no jobs of users of that site may run on another site as their DEISA home and data directories are not available. This requires a dismount of the file system and a closing of the batch system for the affected users. All these actions have to be done via a cross site resource coordination, which is done by video and phone conferences, document exchange, e-mail exchange. Different mailing lists are provided for general operational issues and special maintenance information.

A trouble ticket system allows keeping track of problems until a solution has been provided. Additionally historical information concerning solved problems can be accessed.

The software component Multi Cluster - Global Parallel File System (MC-GPFS) is highly dependent on reachability of remote file systems. Though IBM has incorporated the multi cluster feature into GPFS, problems arose often from these software components because of having remote file systems connected across WANs and not only locally. Therefore the operations team maintenance mailing list has become an intrinsic component of file system management.

The RMIS (Resource Management Information System) system is used to deliver up to date and complete resource management information of the supercomputing systems. It provides information from remote sites to system administrators and end users. The functionalities needed by DEISA include a secure access to the Grid and to resources, job submission and control capability (mainly with [UNICORE]), job rerouting capability (manual operation using relevant information), brokering capability (using relevant information), resource co-allocation capability, an advance reservation capability and an accounting capability. The information is provided based on the Ganglia monitoring tool [Ganglia] coupled to the MDS2/Globus [MDS2] component. Ganglia provides scalable monitoring of distributed systems and large-scale clusters. It relies on a multicast-based listen/announce protocol to monitor state within clusters and uses a tree of point-to-point connections amongst representative cluster nodes to federate clusters and aggregate their state. The monitored information will be fed into the Globus Toolkit [GTK] 2.x information system, MDS2, which provides a standard mechanism for publishing and discovering resource status and configuration information via a uniform, flexible interface to data collected by Ganglia.

An INCA system implementation provided by service activity SA4 provides an overview of and manages the DEISA common production environment (CPE) (for CPE see below). INCA is specifically designed to periodically run a

collection of validation scripts, called *reporters*, with the purpose of collecting the version of the software installed (*version reporters*) and the availability and the correct operation of this software (*unit reporters*). The collected information is then cached on the INCA server, and can be archived to produce a historical representation of the status of the resources of a grid. The architecture of INCA is composed of a centralized *server and clients*, installed on each resource to be validated. Web based data consumers display the results. The installed system displays all software components used in the DEISA production environment. Software administrators and users are able to check via the "Common Production Environment" Inca status page where their application software can be run because of available compilers libraries etc.

6. Lessons learned

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

Though all these things can be handled by e-mail mostly, it is nevertheless mandatory to have regular phone or video conferences, writing minutes and checking for completion of tasks. Additionally it is often necessary to have agreed on strict rules for processing if any disagreements arise. Those dissents are mainly found among others in security policy issues, scheduling of software installation and upgrades, budget issues for needed components.

For these purposes the operation team has been established in DEISA. The planning and coordination of tasks, forwarding of information, power of decision and "managing" in general are prerequisites without which a production quality European wide infrastructure cannot be implemented. Establishing this team has simplified work extremely and it should be recommended to anyone dealing with those kinds of infrastructures not to start without adequate structures.

7. Summary

Three years of DEISA production have shown that the concept implemented in DEISA has succeeded very well. DEISA aimed at deploying a persistent basic European infrastructure for general purpose high performance computing. DEISA intends to adapt to new FP7[3] strategies. This does not preclude that organizational structures of DEISA may change because of merging with new HPC initiatives. But the general idea of DEISA will be sustained. The main next challenge will be to establish an efficient organization embracing all relevant HPC organizations in Europe. Being a central player within European HPC initiatives, DEISA intends to contribute to a global eInfrastructure for science and technology furthermore. Integrating leading supercomputing platforms with Grid technologies and reinforcing capability with shared petascale systems is needed to open the way to new research dimensions in Europe.

8. References

[DEISA] The DEISA Project home page, Distributed European Infrastructure for Supercomputing Applications, http://www.deisa.org

[EGEE] Homepage of EGEE, The Enabling Grids for E-sciencE project , http://www.eu-egee.org/

[Ganglia] Ganglia is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids, http://ganglia.sourceforge.net

[GEANT2] GÉANT2 is the seventh generation of pan-European research and education network, http://www.geant2.net

[3] FP7 = Seventh Research Framework Programme of the European Union (EU)

[GPFS] IBM's General Parallel File System, http://www-03.ibm.com/systems/clusters/software/gpfs.html

GridFTP GridFTP file transfer protocol, https://forge.gridforum.org/projects/gridftp-wg

[GTK] The Globus Toolkit, http://www.globus.org

[INCA] User level grid monitoring, http://inca.sdsc.edu/

[MDS2] The Monitoring and Discovery Service (MDS) is the information services component of the Globus Toolkit, http://www.globus.org/toolkit/docs/2.4/mds/

[OGSA-DAI] Open Grid Services Architecture Data Access and Integration (OGSA-DAI) is a Globus project, http://dev.globus.org/wiki/OGSA-DAI

[SRB] Storage Resource Broker of SDSC, http://www.sdsc.edu/srb/index.php/Main_Page

[Teragrid] TeraGrid is an open scientific discovery infrastructure combining leadership class resources in the US, http://www.teragrid.org/

[UNICORE] UNICORE (Uniform Interface to Computing Resources) is a ready-to-run Grid system including client and server software, http://www.unicore.org/