Production-Quality Grid Environments with UNICORE

D. Erwin, M. Rambadt, A. Streit, and Ph. Wieder

Central Institute for Applied Mathematics (ZAM)
Forschungszentrum Jülich (FZJ)
52425 Jülich, Germany

Abstract. The UNICORE Grid technology provides a seamless, secure, and intuitive access to distributed Grid resources. Since its initial funding in two German-funded research projects, UNICORE evolved to a full-grown and well-tested Grid middleware system. Today it is used in daily production at many supercomputing centers.
In this paper we present an overview on the UNICORE production environments at the John von Neumann-Institute for Computing and within the European DEISA project.

1 Introduction

End of 1998 the concept of “Grid computing” was introduced in the monograph “The Grid: Blueprint for a New Computing Infrastructure” by I. Foster and C. Kesselman [7]. Almost two years earlier, in 1997, the development of UNICORE - Uniform Interface to Computing Resources - was initiated to enable German supercomputer centers to provide their users with a *seamless, secure, and intuitive access* to their heterogeneous computing resources. Like in the case of the Globus Toolkit® [4] UNICORE was started before “Grid Computing” became the accepted new paradigm for distributed computing. At the beginning UNICORE was developed as a prototype software in the two German funded projects UNICORE[1] [1] and UNICORE Plus[2] [2]. Over the following years, in various European-funded projects, UNICORE evolved to a full-grown and well-tested Grid middleware system, which today is used in daily production at many supercomputing centers worldwide and became a solid basis for research projects like EUROGRID, OpenMolGRID, UniGrids, and the Japanese NaReGI project. In this paper we present an overview on the usage of UNICORE for production in the John von Neumann-Institute for Computing (NIC) at the Research Center Jülich and in the European DEISA project.

The remainder of this paper is structured as follows. In Section 2 UNICORE’s architecture and core features are described in more detail. Section 3 gives an overview on the usage of UNICORE in production and lessons learned. The paper ends with a brief conclusion.

2 The Architecture of UNICORE

Figure 1 shows the layered Grid architecture of UNICORE consisting of user, server and target system tier [11]. The implementation of all components shown is realized in Java. UNICORE meets the Open Grid Services Architecture (OGSA) [5] concept following the paradigm of ’Everything being a Service’. Indeed, an analysis has shown that the basic ideas behind UNICORE already realize this paradigm [13, 12].

[1] funded by BMBF grant 01 IR 703, duration: August 1997 - December 1999
[2] funded by BMBF grant 01 IR 001 A-D, duration: January 2000 - December 2002

2.1 User Tier

The UNICORE Client provides a graphical user interface to exploit the entire set of services offered by the underlying servers. The client communicates with the server tier by sending and receiving Abstract Job Objects (AJO) and file data via the UNICORE Protocol Layer (UPL) which is placed on top of the SSL protocol. The AJO is the realization of UNICORE's job model and central to UNICORE's philosophy of abstraction and seamlessness. It contains platform and site independent descriptions of computational and data related tasks, resource information and workflow specifications along with user and security information. AJOs are sent to the UNICORE Gateway in form of serialized and signed Java objects, followed by an optional stream of bytes if file data is to be transferred.

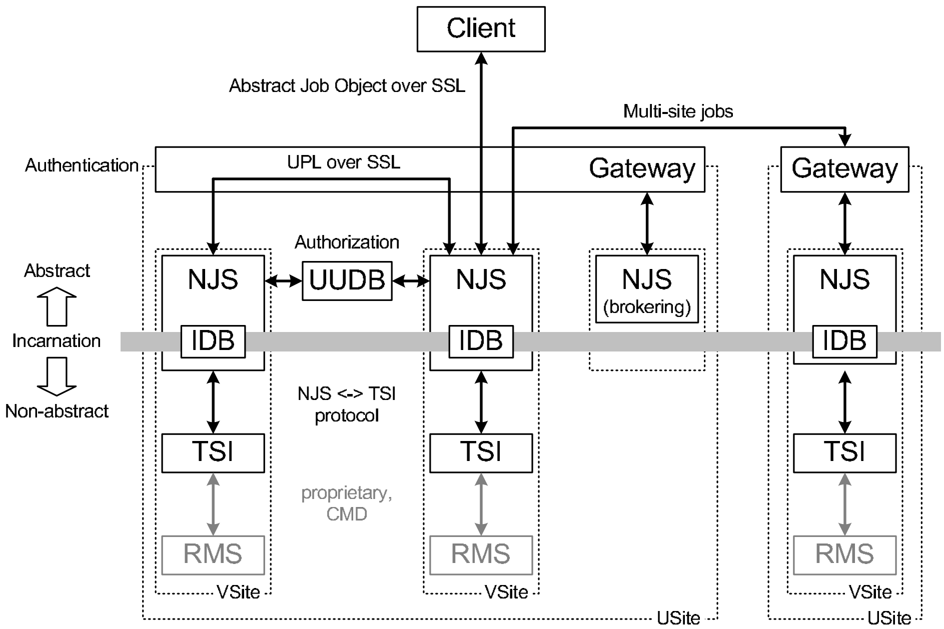

Fig. 1. The UNICORE architecture.

The UNICORE client assists the user in creating complex, interdependent jobs that can be executed on any UNICORE site (Usite) without requiring any modifications. A UNICORE job, more precisely a job group, may recursively contain other job groups and/or tasks and may also contain dependencies between job groups to generate job workflows. Besides the description of a job as a set of one or more directed a-cyclic graphs, conditional and repetitive execution of job groups or tasks are also included. For the monitoring of jobs, their status is available at each level of recursion down to the individual task. Detailed log information is available to analyze potential error conditions. At the end of the execution of the job it is possible to retrieve the stdout and stderr output of the job. Data management functions like import, export, and transfer are available through the GUI as explicit tasks. This allows the user to specify data transfer from one target system to another (e. g. for workflows), from or to the local workstation before or after the execution of a job, or to store data permanently in archives.

The previously described features already provide an effective tool to use resources of different computing centers both for capacity or capability computing, but many scientists and engineers use application packages. For applications without a graphical user interface, a tool kit simplifies the development of a custom built UNICORE plug-in. Over the years many plug-ins were developed, so that plug-ins already exist for many standard scientific applications, as e. g. for CPMD (Car-Parrinello Molecular Dynamics), Fluent, Gaussian, or MSC Nastran.

2.2 Server Tier

The server tier contains the Gateway and the Network Job Supervisor (NJS). The Gateway controls the access to a Usite and acts as the secure entry point accepting and authenticating UPL requests. A Usite identifies the participating organization (e. g. a supercomputing center) to the Grid with a symbolic name that resolves into the URL of the Gateway. An organization may be part of multiple Grids offering the same or different resources to different communities. The Gateway forwards incoming requests to the underlying Network Job Supervisor (NJS) of a virtual site (Vsite) for further processing. The NJS represents resources with a uniform user mapping scheme and no boundaries like firewalls between them.

A Vsite identifies a particular set of resources at a Usite and is controlled by a NJS. A Vsite may consist of a single supercomputer, e. g. a IBM p690 System with LoadLeveler, or a Linux cluster with PBS as resource management system. The flexibility of this concept supports different system architectures and gives the organization full control over its resources. Note that there can be more than one Vsite inside each Usite as depicted in Figure 1.

The NJS is responsible for the virtualization of the underlying resources by mapping the abstract job on a specific target system. This process is called "incarnation" and makes use of the Incarnation Database (IDB). System-specific data are stored in the IDB describing the software and hardware infrastructure of the target system. Among others, the available resources like software, incarnation of abstract commands (standard UNIX commands like rm, cp, ...) and site-specific administrative information are stored. In addition to the incarnation the NJS processes workflow descriptions included in an AJO, performs pre- and post-staging of files and authorizes the user via the UNICORE User Database (UUDB). Typically the Gateway and NJS are running on dedicated secure systems behind a firewall, although the Gateway could be placed outside a firewall or in a demilitarized zone.

2.3 Target System Tier

The Target System Interface (TSI) implements the interface to the underlying supercomputer with its resource management system. It is a stateless daemon running on the target system and interfacing with the local resource manager realized either by a batch system like PBS, a batch system emulation on top of e. g. Linux, or a Grid resource manager like Globus' GRAM [6, 9].

2.4 Single Sign-On

The UNICORE security model relies on the usage of permanent X.509 certificates issued by a trusted Certification Authority (CA) and SSL based communication across 'insecure' networks. Certificates are used to provide a single sign-on in the client. The client unlocks the user's keystore when it is first started, so that no further password requests are handed to the user. All authentication and authorization is done on the basis of the user certificate. At each UNICORE site user certificates are mapped to local accounts (standard UNIX uid/gid), which may be different at each site, due to existing naming conventions. The sites retain full control over the acceptance of users based on the identity of the individual – the distinguished name – or other information that might be contained in the certificate. UNICORE can handle multiple user certificates, i. e. it permits a client to be part of multiple, disjoint Grids. It is also possible to specify project accounts in the client allowing users to select different accounts for different projects on one execution system or to assume different roles with different privileges.

The private key in the certificate is used to sign each job and all included sub-jobs during the transit from the client to sites and between sites. This protects against tampering while the job is transmitted over insecure internet connections and it allows to verify the identity of the owner at the receiving end, without having to trust the intermediate sites which forwarded the job.

3 UNICORE-Based Production Environments

3.1 Production System on Jump

Since July 2004 UNICORE is established as production software to access the supercomputer resources of the John von Neumann-Institute for Computing (NIC) at the Research Center Jülich. These are the 1312-processor IBM p690 cluster (Jump) [8], the Cray SV1 vector machine, and a new Cray XD1 cluster system. As an alternative to the standard SSH login, UNICORE provides an intuitive and easy way for submitting batch jobs to the systems. The academic and industrial users come from all over Germany and from parts of Europe. The applications come from a broad field of domains, e. g. astrophysics, quantumphysics, medicine, biology, chemistry, and climate research, just to name the largest user communities. In the first five month of 2005 31.51% of the used CPU-cycles of the Jump system where used by UNICORE jobs, details for each month are in Table 1.

month	used CPU-cycles of UNICORE jobs
January 2005	30.45%
February 2005	30.50%
March 2005	27.07%
April 2005	29.74%
May 2005	39.13%

Table 1. Usage Statistics of UNICORE.

A dedicated, pre-configured UNICORE client with all required certificates and accessible Vsites is available for download. This alleviates the installation and configuration process significantly. Furthermore, an online installation guide including a certificate assistant, an user manual, and example jobs help users getting started.

To provide the NIC-users with adequate certificates and to ease the process of requesting and receiving a certificate, a certificate authority (CA) was established. User certificate requests are generated in the client and have to be send to the CA. Since introduction of UNICORE at NIC, more than 120 active users requested a UNICORE user certificate.

A mailing list serves as a direct link of the users to UNICORE developers in the Research Center Jülich. The list allows to post problems, bug reports, and feature requests. This input is helpful in enhancing UNICORE with new features and services, in solving problems, identifying and correcting bugs, and influences new releases of UNICORE available at SourceForge [15].

3.2 DEISA – Distributed European Infrastructure for Scientific Applications

Traditionally, the provision of high performance computing resources to researchers has been the objective and mission of national HPC centers. On the one hand, there is an increasing global competition between Europe, USA, and Japan with growing demands for compute resources at the highest performance level, and on the other hand stagnant or even shrinking budgets. To stay competitive major investments are needed every two years – an innovation cycle that even the most prosperous countries have difficulties to fund.

To advance science in Europe, eight leading European HPC centers devised an innovative strategy to build a Distributed European Infrastructure for Scientific Applications (DEISA)[3] [3]. The centers join in building and operating a tera-scale supercomputing facility. This becomes possible through deep integration of existing national high-end platforms, tightly coupled by a dedicated network and supported by innovative system and Grid software. The resulting virtual distributed supercomputer has the capability for natural growth in all dimensions without singular procurements at the European level. Advances in network technology and the resulting increase in bandwidth and lower latency virtually shrink the distance between the nodes in the distributed super-cluster. Furthermore, DEISA can expand horizontally by adding new systems, new architectures, and new partners thus increasing the capabilities and attractiveness of the infrastructure in a non-disruptive way.

By using the UNICORE technology, the four core partners of the projects have coupled their systems using virtually dedicated 1 Gbit/s connections. All other sites will follow in the next step. The DEISA super-cluster currently consists of over 4000 IBM Power 4 processors and 416 SGI processors with an aggregated peak performance of about 22 teraflops. UNICORE provides the seamless, secure and intuitive access to the super-cluster.

The Research Center Jülich is one of the DEISA core partners and is responsible for introducing UNICORE as Grid middleware at all partner sites and for providing support to local UNICORE administrators.

In the following we describe the DEISA architecture. Note, a detailed description of UNICORE's architecture and server components can be found in Section 2 and in particular in Figure 1. All DEISA partners have installed the UNICORE server components Gateway, NJS, TSI, and UUDB to access the local supercomputer resources of each site via UNICORE. Figure 2 shows the DEISA UNICORE configuration of the core production environment. For clarity only four sites are shown. At each site, a Gateway exists as an access to the DEISA infrastructure. The NJSs are not only registered to their local Gateway, but to all other Gateways at the partner sites as well. Local security measures like firewall configurations need to consider this, by permitting access to all DEISA users and NJSs. This fully connected architecture has several advantages. If one Gateway has a high load, access to the high performance supercomputers through DEISA is not limited. Due to the fully connected architecture, no single point of failure exists and the flexibility is increased.

The DEISA partners operate different supercomputer architectures, which are all accessible through UNICORE. Initially all partners with IBM p690 clusters are connected to one large virtual

[3] funded by EC grant FP6-508803, duration: May 2004 - April 2009

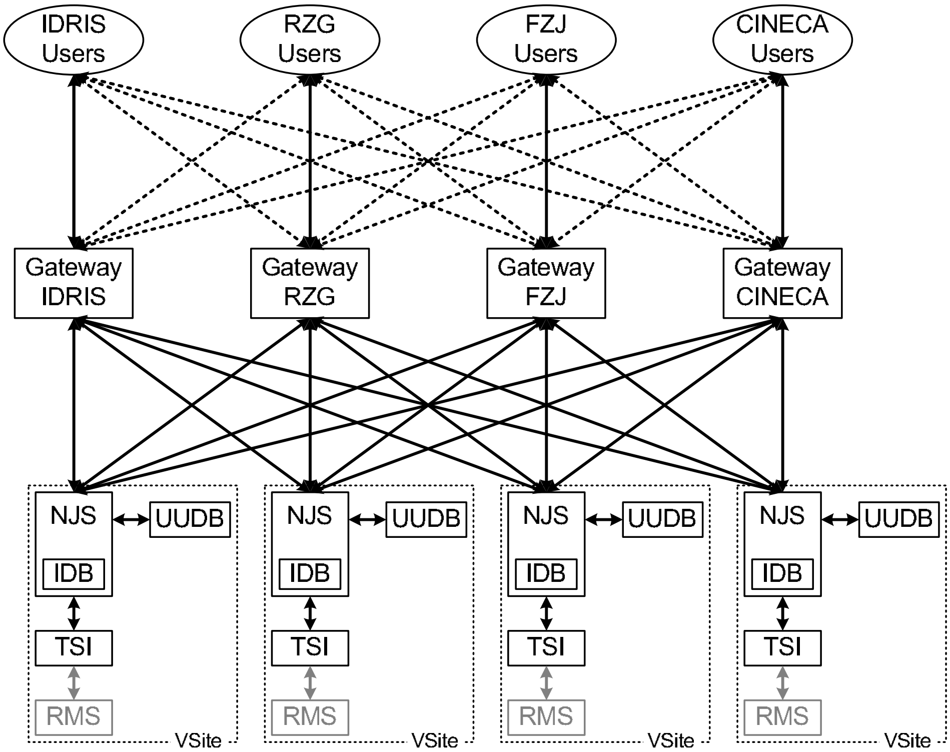

Fig. 2. The DEISA architecture of the core production environment.

supercomputer. In a second step other supercomputers of different variety are connected to DEISA, making the virtual supercomputer heterogeneous. UNICORE can handle this, as it is designed to serve such heterogeneous architectures in a seamless, secure, and intuitive way.

In December 2004 a first successful UNICORE demonstration between the four DEISA core sites FZJ (Research Center Jülich, Germany), RZG (Computing Center Garching, Germany), CINECA (Italian Interuniversity Consortium, Italy) and IDRIS (Institute for Development and Resources in Intensive Scientific Computing, France) was given. Different parts of a distributed astrophysical application were generated and submitted with UNICORE to all four sites.

The experience and knowledge of the researchers, developers, users, and administrators in working with UNICORE in the DEISA project on a large production platform will be used as useful input for future developments of the UNICORE technology. A close synchronization with the UniGrids project [16] is foreseen.

3.3 Lessons Learned

The deployment of new software to be used in production has to offer added value to the users, otherwise the new software will not be accepted. Hence, users have to be stimulated to use the software. They have to be encouraged to use Grid technology for applications, computations, data transfers, and access to resources, to make their applications Grid-aware, and to consider Grid technology for their daily problem solving.

The obvious prerequisite is that the software provides the necessary functions users require. However, the beginning of the UNICORE development showed that the sole fulfillment of these requirements if not sufficient. Although all requested functions were present, only a small number of users used UNICORE for their daily work.

The transition from prototype to production requires:

- High quality of the software, especially high reliability and resilience.
- Help for the users to overcome initial hurdles and permanent assistance to users in case of problems.
- Permanent, 24/7 availability of the infrastructure.
- A long term commitment for continuous development and support.

As soon as the Research Center Jülich implemented and gave the commitment for these additional aspects through initiating and supporting the UNICORE@SourceForge [15] initiative, the use of UNICORE increased significantly as the statistics confirm.

UNICORE and Grid technology in general have made the step away from being used just for demos towards real work in production environments. The driving forces for this effect in our center and the DEISA infrastructure are the above described features of the technology. Through the single sign-on mechanism users do no longer have to remember different account names and passwords for different machines and projects. The easy-to-use and intuitive graphical client allows users to define complex workflows with data staging, data- or time-dependencies between sub-jobs and involving different applications in a more comfortable and sophisticated way as with previously used tools of the local resource management system. Specifying data dependencies, both data-parallel and data-sequential jobs can be executed, enabling e. g. workflows of applications, where the output of one step is used as input for the next step. According to user feedback this feature is very valuable for applications from quantum computing and bio-molecular science. Once jobs are fully defined including an appropriate resource assignment, they can be saved for later usage. By re-loading previously stored jobs in the client, users are able to submit one specific job or workflow multiple times, only with e. g. different input data or input parameters in order to conduct parameter studies.

A detailed analysis of the jobs shows that the increase in resource consumption of UNICORE jobs (cf. Table 1) is due to the submission of very large jobs through UNICORE. These jobs are both large in CPU requirements as well as in run time. Typically the resources are used for parameter studies with simulation codes. It can be observed that users combine several iterations and steps (about 20 to 30) within a single UNICORE workflow with data- and/or time-dependencies.

Several projects from various domains of science make use of UNICORE to access the compute resources at Research Center Jülich. Among these projects are:

- QCD simulations with light quark flavors (elementary particle physics)
- Finite temperature meson correlation functions (elementary particle physics)
- Non-leptonic kaon decays in lattice QCD (elementary particle physics)
- Nucleon matrix elements from overlap fermions (elementary particle physics)
- Electronic and optical properties of capped silicon and germanium nanocrystallites (material science)
- Visco-elastic shear flow (fluid science)
- Small scale structure of the universe (astro physics)

4 Conclusion

In this paper we presented the usage of UNICORE in production-quality Grid environments. UNICORE – Uniform Interface to Computing Resources – provides a *seamless, secure and intuitive* access to distributed Grid resources. Initially developed in two German projects, the software evolved from prototype status to a full-grown and well-tested Grid system, which is today used in daily production at many supercomputing centers in Europe.

At the John von Neumann-Institute for Computing, Research Center Jülich, many users submit their batch jobs through UNICORE to the 1312-processor 8.9 TFlop/s IBM p690 cluster, the Cray SV1 vector machine, and a new Cray XD1 cluster system. Leading European HPC centers joined in the project DEISA to build and operate a distributed European Grid supercomputing infrastructure for scientific applications with multi tera-scale performance and production quality, similar to the TeraGrid project [14]. Within the DEISA project UNICORE is used as the Grid middleware to connect all sites to the infrastructure.

The future of UNICORE is promising and follows the trend of "Everything being a Service", as it follows the Open Grid Service Architecture (OGSA) [5]. In this context, the UniGrids[4] project [16] continues previous efforts in integrating the Web Services and UNICORE technology to enhance UNICORE to an architecture of loosely-coupled components while keeping its "end-to-end" nature. To this end UNICORE/GS will be developed, which makes UNICORE compliant with the Web Services Resource Framework (WS-RF) [10].

Since May 2004 the UNICORE software is available as open source under a BSD licence from SourceForge for download [15]. Numerous contributors from all over the world, e. g. Norway, Poland, China and Russia checked-in their developments and until April 2005 more than 3100 downloads of UNICORE were counted.

References

1. D. Erwin (Ed.). *UNICORE - Uniformes Interface für Computing Ressourcen, final project report (in German)*. 2000.
2. D. Erwin (Ed.). *UNICORE Plus Final Report - Uniform Interface to Computing Resources*. Forschungszentrum Jülich, 2003.
3. DEISA - Distributed European Infrastructure for Supercomputing Applications. `http://www.deisa.org`.
4. I. Foster and C. Kesselman. Globus: A Metacomputing Infrastructure Toolkit. *International Journal on Supercomputer Applications*, 11(2):115–128, 1997.
5. I. Foster, C. Kesselmann, J. M. Nick, and S. Tuecke. The Physiology of the Grid. In F. Berman, G. C. Fox, and A. J. G. Hey, editors, *Grid Computing*, pages 217–249. John Wiley & Sons Ltd, 2003.
6. Globus: Research in Resource Management. `http://www.globus.org/research/resource-management.html`.
7. I. Foster, C. Kesselman (Eds.). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers Inc. San Fransisco, 1999.
8. Jump - Juelich Multi Processor, 8,9 TFlop/s IBM p690 eServer Cluster. `http://jumpdoc.fz-juelich.de`.
9. R. Menday and Ph. Wieder. GRIP: The Evolution of UNICORE towards a Service-Oriented Grid. In *Proc. of the 3rd Cracow Grid Workshop (CGW'03)*, pages 142–150, 2003.
10. OASIS Web Services Resource Framework (WSRF). `http://www.oasis-open.org/committees/wsrf`.
11. M. Romberg. The UNICORE Grid Infrastructure. *Scientific Programming*, 10(2):149–157, 2002.
12. D. Snelling. UNICORE and the Open Grid Services Architecture. In F. Berman, G. Fox, and T. Hey, editor, *Grid Computing: Making The Global Infrastructure a Reality*, pages 701–712. John Wiley & Sons, 2003.
13. D. Snelling, S. van den Berghe, G. von Laszweski, Ph. Wieder, D. Breuer, J. MacLaren, D. Nicole, and H.-Ch. Hoppe. A UNICORE Globus Interoperability Layer. *Computing and Informatics*, 21:399–411, 2002.
14. TeraGrid. `http://www.teragrid.org/`.
15. UNICORE at SourceForge. `http://unicore.sourceforge.net`.
16. UniGrids - Uniform Access to Grid Services. `http://www.unigrids.org`.

[4] funded by EC grant IST-2002-004279, duration: July 2004 - June 2006