UNICORE Deployment Within the DEISA Supercomputing Grid Infrastructure

Luca Clementi[1], Michael Rambadt[2], Roger Menday[2], and Johannes Reetz[3]

[1] CINECA
Via Magnanelli 6/3,
40033 Casalecchio di Reno, Italy
l.clementi@cineca.it
[2] Central Institute for Applied Mathematics
Forschungszentrum Jülich GmbH
D-52425 Jülich, Germany
{m.rambadt,r.menday}@fz-juelich.de
[3] Garching Computing Centre of the Max Planck Society
Max-Planck-Institute for Plasma Physics
D-85748 Garching, Germany
johannes.reetz@rzg.mpg.de

Abstract. DEISA is a consortium of leading national supercomputing centers that is building and operating a persistent distributed supercomputing environment with continental scope in Europe. To integrate their resources, the DEISA partners have adopted the most advanced middleware and applications currently available. The consortium decided to embrace UNICORE as a job submission interface for the DEISA grid infrastructure. UNICORE is the foremost European grid technology able to hide the complexity of the underlying resources providing a user-friendly graphical user interface for job submission. This paper presents the deployment solution and strategies implemented by DEISA in order to adapt UNICORE for their infrastructure.

Keywords: UNICORE, computational grid, middleware deployment, interoperability.

1 Introduction

The DEISA project [1] started in May 2004 with the goal of providing a persistent and production quality, distributed supercomputing environment. The members of the consortium want to improve the level of exploitation of their systems and, at the same time, to provide a higher Quality of Service to the users, being able to offer a larger joint resource pool [9]. When building such an infrastructure, the DEISA partners considered several applications and middleware technologies that are providing the functionalities necessary to integrate their high-performance computing systems.

The DEISA consortium decided to use UNICORE to establish a grid infrastructure. UNICORE is one of the leading grid middleware used in production in several supercomputing centers. It hides the complexity of the underlying systems and

W. Lehner et al. (Eds.): Euro-Par 2006 Workshops, LNCS 4375, pp. 264–273, 2007.

architectures and it provides a single sign-on mechanism based on X.509 certificates from a Public Key Infrastructure (PKI).

The DEISA partners adapted and customized the UNICORE architecture according to the specific DEISA requirements. This paper presents the deployment strategies adopted to incorporate UNICORE into the DEISA grid. The second and the third chapters describe the DEISA infrastructure and the UNICORE architecture. The fourth chapter presents a detailed analysis of the solutions and the adaptations put in place by the partners in order to deploy UNICORE without interference with their local policies. To conclude, the fifth chapter presents a summary and considerations regarding our experience with UNICORE.

2 DEISA

DEISA (Distributed European Infrastructure for Supercomputing Applications) is a consortium of leading national supercomputing centers that is funded by the 6th European Framework Program. It deploys a production-quality supercomputing environment by exploiting the grid paradigm.

The DEISA infrastructure has two levels of integration. The inner level comprises strongly coupled clusters running IBM AIX on IBM POWER systems located at CINECA, CSC, FZJ, IDRIS, and RZG. Due to the same operating system and batch scheduler in common, these coupled clusters establish a *homogenous* super-cluster. An outer level of *heterogeneous* supercomputer clusters comprises by the SGI ALTIX Linux cluster located at SARA, the IBM PowerPC Linux system at BSC, the SGI ALTIX Linux at LRZ [1, 2], plus other resources provided by EPCC and ECMWF.

All the DEISA sites are linked together via a dedicated network provided by GEANT and National Research and Education Network providers (NREN). Beside the provision of high-performance compute facilities, the DEISA consortium offers also services such as help desk, documentation, technical and scientific workshops.

2.1 Infrastructure Overview

To achieve a higher level of interoperability between the different resources, the partners decided to harmonize their user management systems and to establish a DEISA user administration system by deploying a distributed set of LDAP servers [3] used to propagate information about DEISA users from the user's home site to all the partner sites. A standardization of the naming schema for DEISA users, and the assignment of site-specific ranges of UIDs and GIDs, ensures that DEISA user accounts are replicable on every system belonging to the DEISA infrastructure.

A user who wants to use DEISA resources needs to apply for an account only at his home site. The user record information (user name, UID, GID, the subject of his certificate, etc.) propagates via LDAP from his home site to all the other DEISA sites.

It is possible to identify three types of users for every DEISA site:

1. Internal users: they have an account only on a single resource and their user records are not published via the DEISA LDAP servers.
2. Local DEISA users: as home site users, they belong to the DEISA site locally, and their user records are published via the DEISA LDAP server at their site. Hence, their user accounts are automatically replicated on all the other DEISA resources.

3. External DEISA users: they belong to other DEISA sites, but their user records have to be imported from the DEISA LDAP servers at their home site in order to replicate their accounts on the local systems.

All the resources belonging to the inner level of the DEISA infrastructure have been integrated also by means of a shared file system. The consortium has decided to use GPFS-MC (General Parallel File System-Multi Cluster) [4] to achieve a transparent high-performance data access over the Wide Area Network. All the shared instances of the GPFS-MC are mounted on specific paths beginning with /deisa/<site acronym>, and so they are accessible from all the DEISA sites in the same manner.

Finally, a grid-enabled version of the LoadLeveler Batch Scheduling System [5] from IBM has been adopted for the intra-cluster scheduling of jobs. This product allows that a job submitted, e.g., by a CINECA user to CINECA's IBM P5 cluster can be routed to another DEISA site, depending on the resource requirements of the job and the availability of appropriate resources at the other DEISA sites.

Thank to the shared file system and to the distributed user administration system, migrated jobs can be executed under the same UID used on the user's home site cluster; the ownership of files on the shared file system needs not to be translated.

Users can access DEISA resources via UNIX shell, UNICORE, or Web Portals. DEISA partners decided to adopt the shell access because it is still the most common user interfaces for UNIX systems, and because it is useful for debugging applications during the development phase. On the other side, UNICORE provides an abstract view of the underlying system with its powerful Graphical User Interface, and the single sign-on mechanism simplifies the access to the distributed DEISA resources.

In the next chapters, the integration of UNICORE with the other components of the DEISA infrastructure is explained in more detail.

3 UNICORE

UNICORE (UNiform Interface to Computing REsources) provides a seamless interface for preparing and submitting jobs to a wide variety of heterogeneous distributed computing resources and data storages. It supports users for running scientific and engineering applications in a heterogeneous Grid environment.

The UNICORE software has been developed in the UNICORE and UNICORE Plus [6, 12] projects funded by the German Ministry of Education and Science (BMBF) until the end of 2002. After that, its functionalities and its robustness were enhanced within the EU-funded projects EUROGRID [7], OpenmolGrid [8]. Since 2004, several supercomputing centers are employing UNICORE in production.

In UNICORE every job represented by Java based abstract job formulation, the so called Abstract Job Object (AJO). This gives the user the possibility to prepare jobs on an abstract level without having to know deep details of the target system. With the abstract formulation, the job can be submitted to different target architectures running different batch schedulers without significant changes.

3.1 UNICORE Components

UNICORE is designed as vertically integrated three-tier architecture. It provides client and server components. The server-side consists of the Gateway, Network Job Supervisor (NJS) including an Incarnation Database (IDB), a UNICORE User Database (UUDB), and the Target System Interface (TSI). All components (except the TSI) are written in Java allowing to install UNICORE on a large variety of operating systems.

3.1.1 UNICORE Client

The UNICORE client GUI is used for preparation, submission, monitoring, and administration of complex multi-site and multi-step jobs. It provides the user with an extensible application support, resource management of the target system and an integrated security mechanism.

Every submitted request (AJO) is signed using the personal X.509 certificate of the user. Thus, other UNICORE server components can perform authentication and authorization relying on the PKI in use. The client allows also performing data management and transfer provided by an intuitive GUI.

3.1.2 Gateway

The Gateway is the site's point of contact for all connections relative to a UNICORE site (Usite). It accepts SSL connections from clients and one or more NJSs, but only if the incoming certificate is signed by a trusted Certification Authority (CA). Moreover, it verifies if received AJOs have been signed with trusted and valid certificates. If the authentication is successful, the AJO is redirected to the corresponding NJS, otherwise it is rejected.

3.1.3 Network Job Supervisor (NJS)

The Network Job Supervisor (NJS) operates as a UNICORE scheduler and is responsible for the virtualization of the underlying resources. It receives/sends AJOs from/to the Gateway, translates them into concrete instances and sends them to the target system component, called Target System Interface (TSI) (see next paragraph). The NJS dispatches jobs to a dedicated target machine or cluster (Virtual site, Vsite), and handles dependencies and data transfers for complex workflows. It transfers the results of executed jobs from the target machine and forwards them via the Gateway to the UNICORE client.

The abstract definition of the Job is translated to a concrete job in the NJS with the help of the Incarnation Database (IDB). The IDB contains all target system specific information regarding computing resources typology and availability of applications. Therefore, each NJS has a dedicated IDB that describes its specific target system.

Finally, the NJS implements the UNICORE security model for user authorization. All public user certificates are stored in the UNICORE User Database (UUDB) and they are mapped with an account existing on the target system. Every time the NJS receives an AJO, it checks if the signer's certificate is present in the UUDB, and on success, the job is forwarded to the target system and assigned to the corresponding user account.

3.1.4 Target System Interface (TSI)

The TSI running on the target machine is the interface to the batch scheduler. It comprises a set of Perl libraries that implements the specific target system commands for job submission, status query, file handling, etc. A variety of TSI implementations are available for different batch schedulers and operating systems, e.g., LoadLeveler under AIX, LSF, PBS-Pro, and CCS.

4 UNICORE and DEISA Infrastructure Integration

The main design guidelines for the DEISA grid can be summarized as follows:

- The grid middleware has to present the different target architectures and resources with a seamless view hiding all the underlying complexity
- Users need to be able to address the various DEISA resources without perceiving the complexity behind them
- Grid middleware needs to provide reliability to software and hardware failures
- As far as possible, the deployment of the DEISA grid infrastructure should not interfere with the local site policies and security requirements. The middleware must be easily adaptable to the local site procedures and policies.

The following paragraphs demonstrate how UNICORE respects these principles.

4.1 UNICORE Deployment for DEISA

According to the conventional UNICORE deployment pattern, every site represents a separate Usite: every NJS, located at one site, connects only to its site Gateway. A virtual organization (VO) could provide several Gateways, whereas each represents the entry point to a separate site. The advantage of this pattern is a better scaling due to the decoupling of the UNICORE deployment at the VO member sites. On the other hand, the composition of the VO is not pervasive in this case, and has to be defined on the client-side using a list of references to all the member gateways.

Due to the limited amount of DEISA members and an efficient cooperation of all the DEISA sites, the consortium decided that all the UNICORE Gateways are to be connected to all NJSs and vice versa. Exploiting the UNICORE dynamic Vsite registrations feature a fully meshed UNICORE infrastructure was built up to allow for the job submission via all Gateways to all target systems. Figure 1 depicts the DEISA UNICORE infrastructure showing five DEISA sites as an example. As a result the DEISA UNICORE infrastructure provides a distributed set of DEISA access points. Each represents an entry point to the whole infrastructure. Preferentially, DEISA users shall use the Gateway at their home site for submitting jobs to any DEISA target system.

The benefits of the deployment pattern shown, where all the NJSs are *jointly connected* to one or more gateways are:

- Pervasive visibility of all the available NJSs in DEISA at every single Gateway
- As an expression of corporate identity, all the DEISA sites are committing to maintain the persistent availability and accessibility of their DEISA resources via any Gateway of the DEISA partner sites.

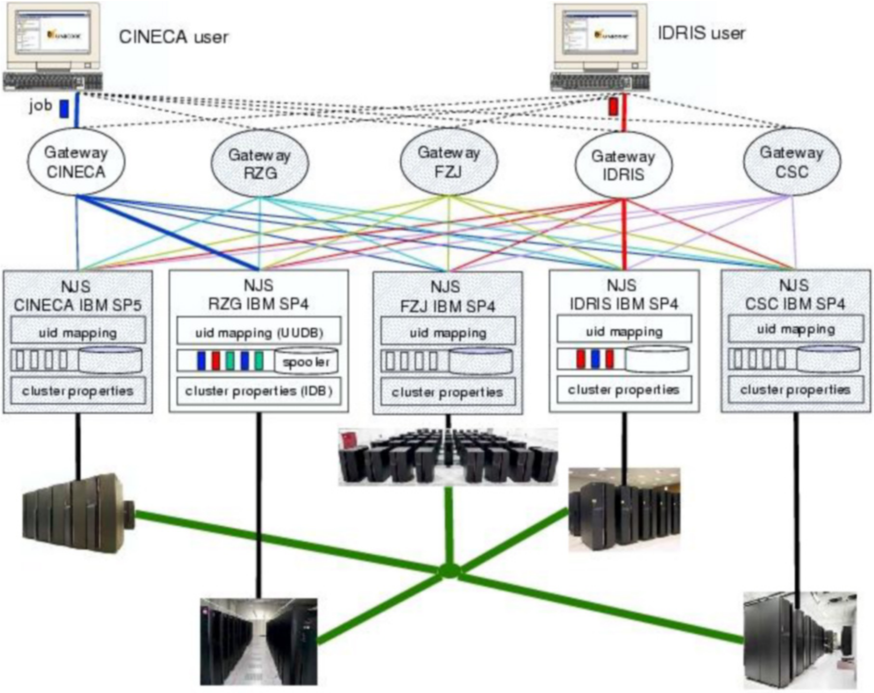

Fig. 1. Schema of the UNICORE deployment at DEISA for the homogeneous infrastructure

- On the client-side, DEISA users need only to validate the issuer of the server certificate of their home site Gateway. It is usually not necessary to import certificates of trusted CAs from other countries to authenticate other Gateways.
- The DEISA NJSs can be enabled to communicate with each other allowing to employ implementations of an alternative file transfers mechanism (AFT). This is useful particularly if the file system of a site does not share GPFS-MC.

The benefits of the *fully meshed deployment* pattern with multiple gateways are:

- Reliability of service: if the DEISA gateway of the user's home site is down or heavily loaded, other DEISA gateways can substitute it completely. In order to access another Gateway, the user must of course import the certificate of the trusted issuer of the Gateway's certificate into his client's truststore.
- Load balancing on the multiple Gateways: provided an appropriate information system is available, it is possible to select a gateway according to the actual load or number of the available DEISA gateways.

Additional UNICORE gateways can be added to the infrastructure in cooperation with the DEISA partners. As a proof of concept, the DEISA UNICORE infrastructure is now composed of 10 sites, and every site is allowing access to its resources through a UNICORE Gateway and NJS, configured as explained in this paragraph.

This fully meshed infrastructure remains firewall friendly and flexible in terms of deployment. Each site has to open only two ports in the site firewall (DMZ or intranet firewall); one is the Gateway port that has to be opened for the whole outside world for inbound connections. While the NJS port needs to be opened only for inbound connections coming from the other DEISA Gateways.

4.1.1 UNICORE User Management for DEISA

In order to get the user and server certificates needed for UNICORE a set of root Certification Authorities (CA) are required. DEISA decided to use only CAs accredited at the EUGridPMA [10]. These CAs issue user and server certificates in compliance with the minimal requirements of the EUGridPMA [11]. Users who want to submit jobs via UNICORE need to obtain a DEISA user account and a EUGridPMA compliant personal certificate signed by their national EUGridPMA member CA.

As an improvement of the UNICORE authorization system the DEISA consortium requested to implement a modification in the UUDB internal management. The standard UUDB implementation maps the complete public part of the user's certificate to a user's account, while the modified DEISA UUDB checks only whether the Distinguished Name (DN) of the certificate used to sign the AJO is present in the UUDB. With support of the UNICORE developers, the implementation of the UUDB authorization mechanism has be adapted accordingly.

4.1.1.1 Security implications. For the DEISA infrastructure, checking only for the DN it is not a security risk since DEISA relies on the joint PKIs of the EUGridPMA member CAs that ensure the uniqueness of the DN within their joint domain. Therefore, a DEISA UNICORE Gateway will never authenticate a certificate issued by a suspected CA.

The benefits of the modified implementation of the UUDB are that

- the DEISA user management system needs to store and exchange only the DN
- there is no need to update the content of the UUDB when a user certificate has been reissued, because the DN of the reissued certificate remains unchanged.

Furthermore the DEISA consortium is considering the adoption of some Globus components (GRAM and GridFTP) [15] for its infrastructure which are based on the Grid Security Infrastructure (GSI) security protocol [16]. GSI authentication is based on PKI and the credentials needed for the authentication are the DN of the user's certificate with the corresponding user name. Hence, the integration of the GSI authentication system with the DEISA infrastructure will be straightforward due to the similarities with the current UNICORE UUDB.

Users can also access DEISA resources using an interactive shell where the authentication is performed by means of username and password. In order to avoid the distribution of users password hash, the DEISA partner decided not to publish this information on the DEISA LDAP servers. Therefore, DEISA users are allowed for interactive access only on the home site system. By this approach, no security sensible user information is published on the DEISA LDAP servers.

4.2 Integration of UNICORE with DEISA Batch Scheduling Systems and GPFS Multi Cluster

A key feature of the DEISA infrastructure is the capability of migrating submitted job to other clusters using LoadLeveler (LL). Migrated jobs can use the DEISA wide shared file system to access data, or write output and debugging information to the user directory allowing a user to monitor the execution status of his application. In order to keep this functionality several modifications have been performed to its standard configuration.

4.2.1 Adaptations Regarding LoadLeveler

To accomplish this goal all the LL instances running on the homogeneous POWER/AIX clusters needed to be configured in a coherent way. This was achieved defining a set of mandatory parameters that have to be specified by a user to describe the resource requirements of his job to be executable on DEISA resources. These parameters are *total tasks*, *threads per task*, *wall clock limit*, *data memory limit*, *stack memory limit*, and a keyword notifying that the job has to be handled by the underlying queuing system as an explicit DEISA job.

The DEISA partners agreed that this set of parameters is adequate for all the cluster configurations and Batch Scheduling Systems currently employed at all the DEISA sites (LL, PBSPro, LSF, and Torque).

The standard version of UNICORE allows for specifying mainly up to six, partly different parameters of an abstract job. Since these parameters can not be mapped directly to those identified to be relevant for DEISA, it was decided to use only four of the original parameters (*total tasks*, *threads per task*, *wall clock limit*, *data memory limit*) and to specify the remaining two parameters as environment variables in an adequate way. UNICORE allows specifying additional environment variables for each submitted job. To take the additional environment variables into account, the TSI, in particular the script that creates the submission command for the LL, has been easily adapted.

4.2.2 Adaptations Regarding GPFS-MC

When submitting a job, UNICORE creates a temporary working directory (called USPACE) at the target system where, among others, batch scripts, input, output and error files are placed. At first, DEISA partners did not require a common path for USPACE. The USPACE was simply located on a site local file system. However, when submitted UNICORE jobs are to be migrated to other clusters by the Multi-Cluster LoadLeveler, the USPACE at the originating cluster needs to be transparent.

As a solution, the different partners have decided to configure UNICORE in order to use a common USPACE path on the GPFS-MC. In this way, UNICORE jobs submitted to the homogeneous super-cluster have always a consistent reference to the USPACE and thus to the files needed by the NJS at the originating site for monitoring the job status and fetching the output. The implementation of this solution required the modification of some TSI scripts.

4.3 The Final Picture

The figure 2 shows how the various components of the DEISA Grid interact. GPFS-MC provides a common shared storage resource available for the different DEISA resources. The Multi-Cluster LoadLeveler enables to address local and remote computational resources. The interactive usage of a Shell provides access to the local instance of LL while UNICORE allows additionally to access remote instances of LL.

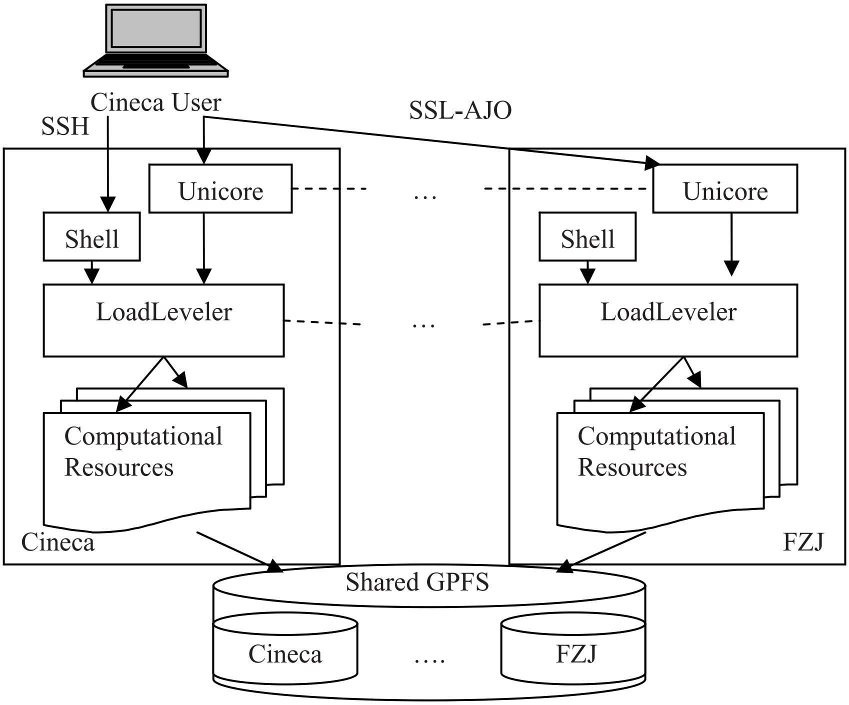

Fig. 2. A general schema of the DEISA Infrastructure representing only two sites

5 Conclusions

UNICORE has proven to be a suitable solution for a complex environment like the DEISA infrastructure. Its high customization level has been shown to be essential for its deployment in accordance with DEISA requirements. Its open source characteristic allows the DEISA partners to modify the default behaviors. Its wide adoption and its long-standing production usage guarantee the Quality of Service that is required for the DEISA infrastructure.

The functionalities of UNICORE within the DEISA consortium have been demonstrated at the IST conference of 2004 where a demonstration showed the seamless usage of the homogeneous resources by means of UNICORE. The number of DEISA users is rising constantly and the commitment to serve their aims with

UNICORE will remain part of the consortium goal. Moreover, UNICORE will play a central role for the integration of the heterogeneous clusters considering its capability to support various system platforms and batch schedulers.

Finally, GRIP [13] and UniGrids [14], two EU funded research projects have developed UNICORE extensions for interoperability with Globus [15]. These enhancements allow DEISA to integrate Globus components, such as GridFTP and GRAM, into the DEISA UNICORE infrastructure.

Acknowledgments. A special thank to all the DEISA partners that contributed their solutions and ideas to the current deployments and configurations.

References

1. http://www.deisa.org - Distributed European Infrastructure for Supercomputing Appications
2. DEISA Primer. DEISA Consortium, Version 1.2 (02/2006)
3. K. Zeilenga, and OpenLDAP foundation: Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map. RFC 4510, (06/2006)
4. Frank Schmuck, Roger Haskin: GPFS: A Shared-Disk File System for Large Computing Clusters. Proceedings of the FAST, Monterey, (01/2002)
5. LoadLeveler for AIX 5L and Linux V3.3.1 Using and Administering. IBM (11/2005)
6. D. Erwin (Ed.): UNICORE Plus Final Report - Uniform Interface to Computing Resources. Forschungszentrum, Julich, (2003)
7. K. Nowinski, B. Lesyng, M. Niezgódka, P. Bala: Project EUROGRID. Proceeding of the PIONIER 2001, Poznan (2001)
8. S. Sild, U. Maran, M. Romberg, B. Schuller, E. Benfenati: OpenMolGRID: Using Automated Workflows in GRID Computing Environment. Proceedings of the European Grid Conference 2005, Amsterdam, (02/2005)
9. I. Foster, C. Kesselman (Eds.). The Grid 2: Blueprint for a New Computing Infrastructure. Morgan Kaufmann Publishers Inc. San Fransisco (2004)
10. www.eugridpma.org
11. Profile for Traditional X.509 Public Key Certification Authorities with secured infrastructure. EUGridPMAa, Version 4.0 (10/2005)
12. D. W. Erwin, D. F. Snelling: UNICORE: A grid computing environment. Proceedings of Euro-Par 2001, Springer, Machester (08/2001)
13. Michael Rambadt, Philipp Wieder. UNICORE - Globus: Interoperability of Grid Infrastructures. Proceedings of the Cray User Group 2002, Manchester (05/2002)
14. http://www.unigrids.org/ - Uniform Interface to Grid Services
15. I. Foster: Globus Toolkit Version 4: Software for Service-Oriented Systems. International Conference on Network and Parallel Computing, Springer LNCS, (2005)
16. R. Butler, D. Engert, I. Foster, C. Kesselman, S. Tuecke, J. Volmer, V. Welch: A National-Scale Authentication Infrastructure. IEEE Computer, 33(12):60-66, (2000)