UNICORE Deployment within the DEISA Supercomputing Grid Infrastructure

Distributed European Infrastructure for Supercomputing Applications

Luca Clementi
l.clementi@cineca.it
Michael Rambadt
m.rambadt@fz-juelich.de
Roger Menday
r.menday@fz-juelich.de
Johannes Reetz
johannes.reetz@rzg.mpg.de

Outline

- DEISA Infrastructure Overview
- Unicore architecture overview
- Unicore integration with DEISA infrastructure
  - UNICORE deployment
  - UNICORE User DataBase UUDB
  - USpace
  - LoadLeveler adaptations
- Final overview of the UNICORE deployment
- Conclusions

Distributed European Infrastructure for Supercomputing Applications

Distributed European Infrastructure for Supercomputing Applications

- The DEISA consortium
- Its hardware resources
- DEISA user-management
- Batch Scheduling System
- Shared File System
- User access to DEISA resources

The DEISA Consortium

Distributed European Infrastructure for Supercomputing Applications

DEISA is a consortium of leading national supercomputing centers that currently deploys and operates a persistent, production quality, distributed supercomputing environment with continental scope

DEISA is composed of:

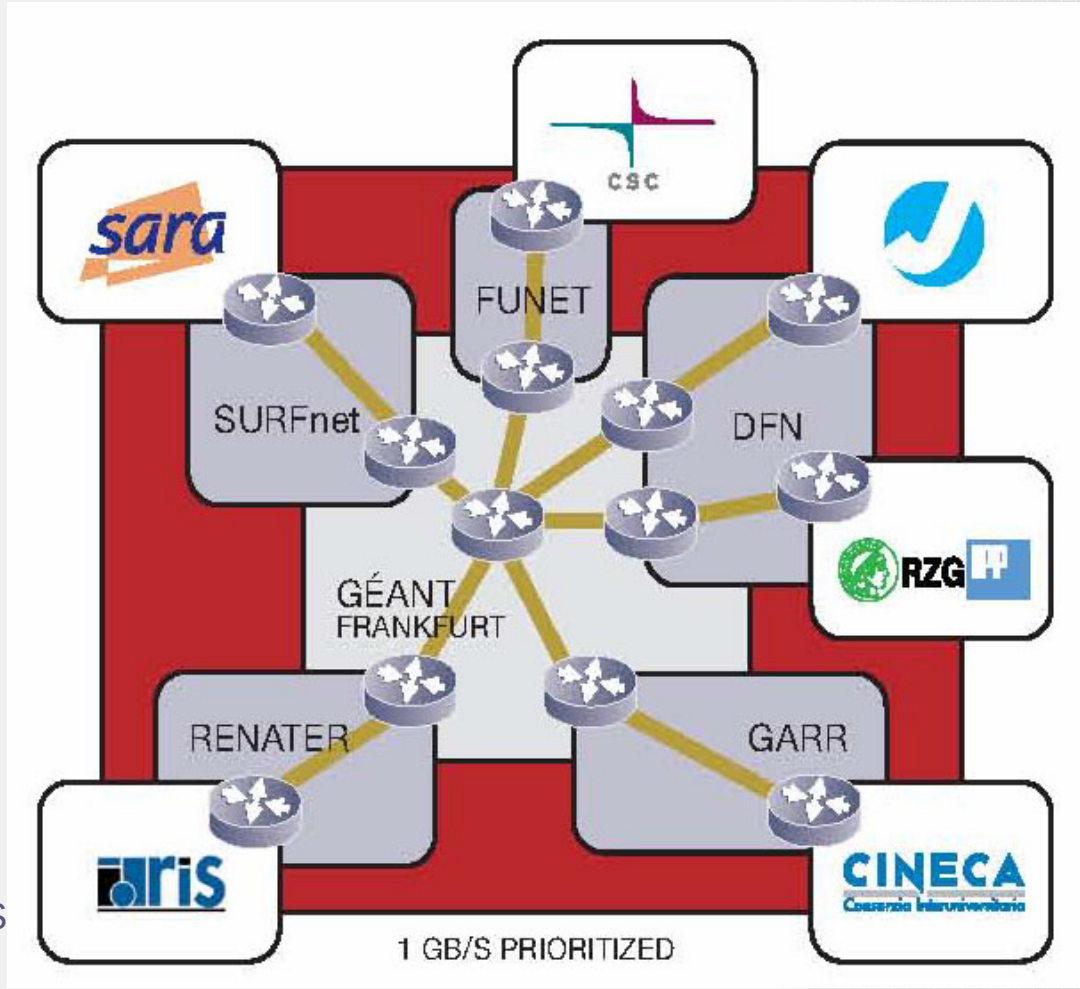
- homogenous infrastructure: strongly coupled distributed super-clusters based on IBM AIX OS and POWER microprocessor (FZJ, RZG, CINECA, CSC and IDRIS)
- Heterogeneous infrastructure: different Intel/Power/AMD computing clusters (SARA, BSC, LRZ, EPCC and ECMWF)

The Network

Distributed European Infrastructure for Supercomputing Applications

All the sites are linked together by a dedicated network of 1Gbit/s

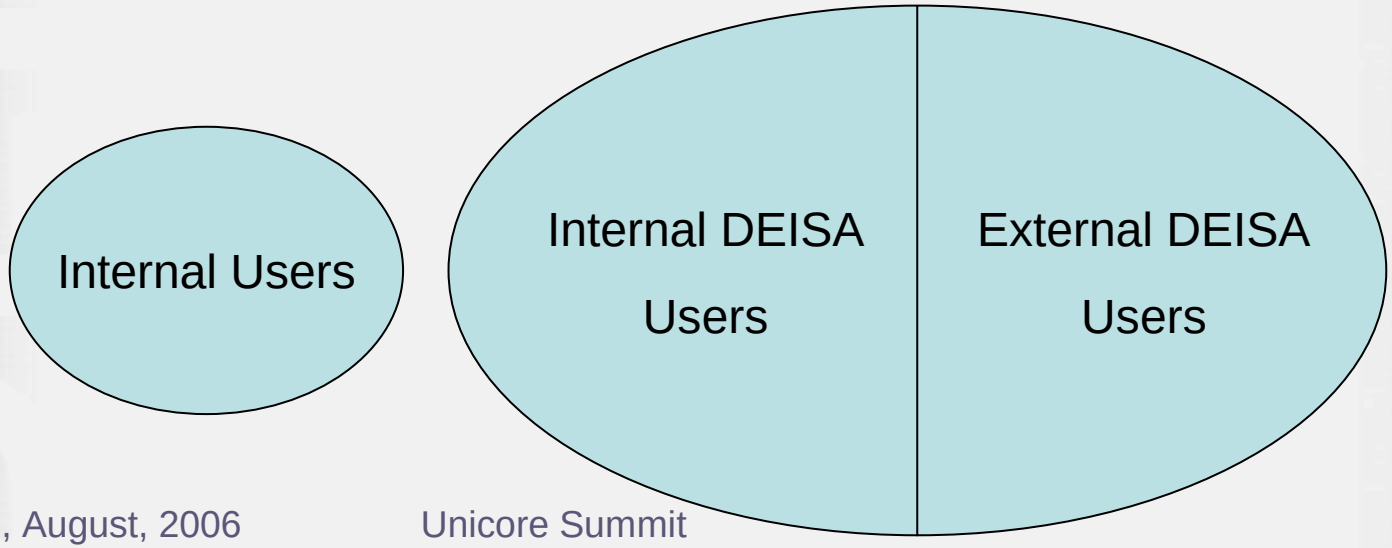
Provided by GENAT and the National Research Network



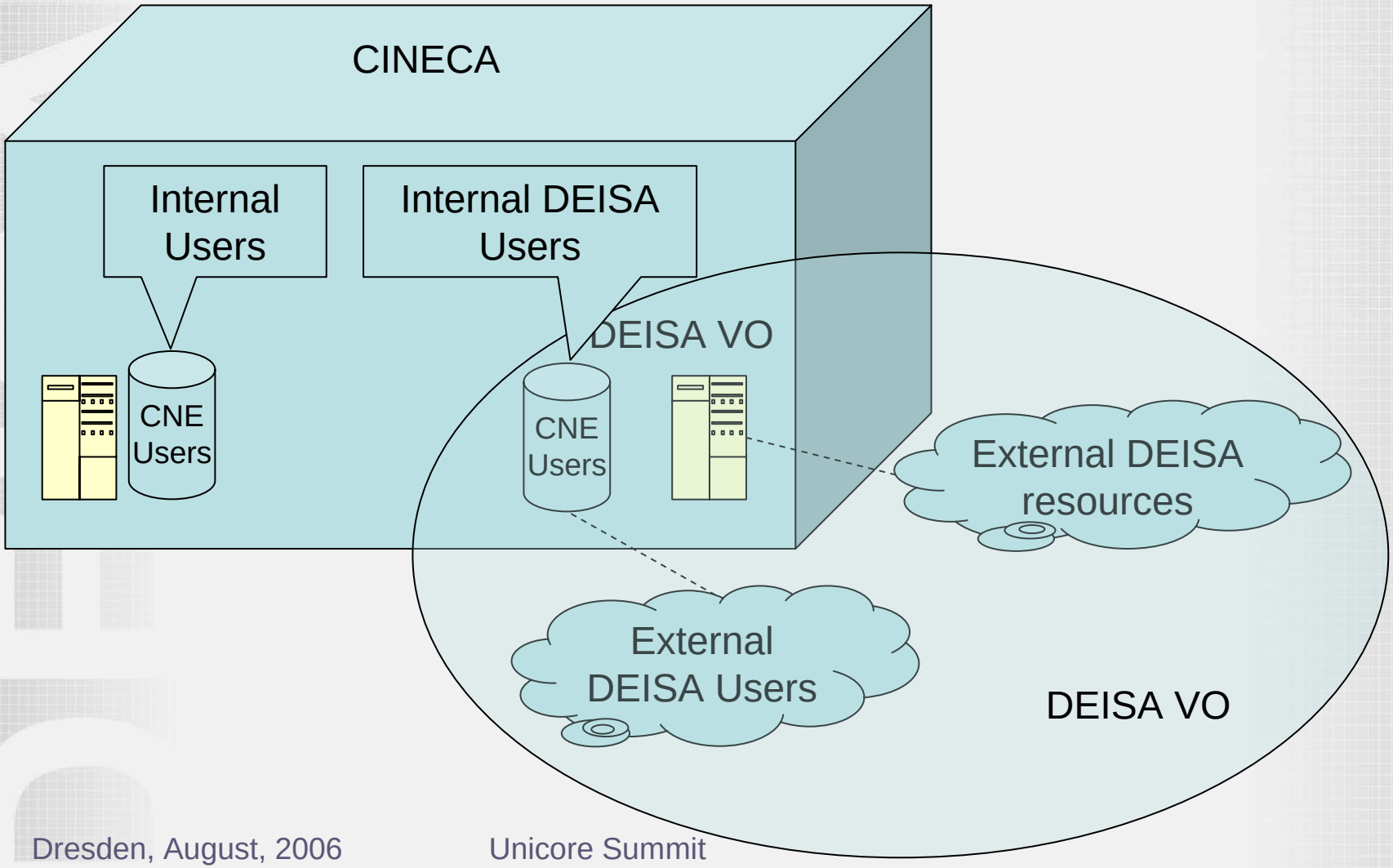
User Management

Every DEISA site can divide its users into three sets:

- Internal User: users that are not part of the DEISA VO
- DEISA User
  - Internal: users that belongs to the site
  - External: users that belongs an external site



User Management



User Management

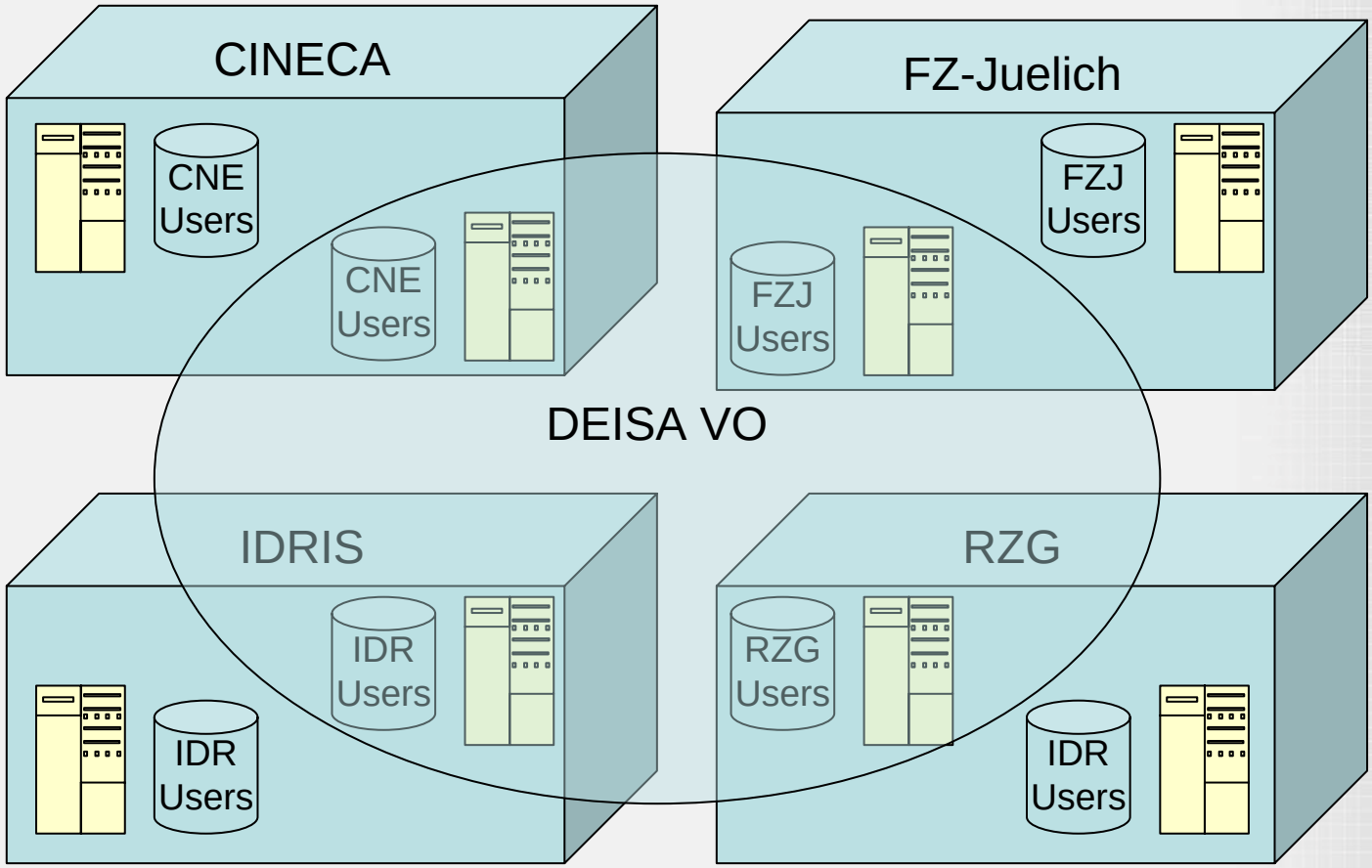
Distributed European Infrastructure for Supercomputing Applications

Naming rule for user names and reserved UID range for each partners

Site	acronym	first	last number
CINECA	cne	100000	199999
FZJ	fzj	200000	299999
IDRIS	idr	300000	399999

User account information published with a set of LDAP servers (one for each site).

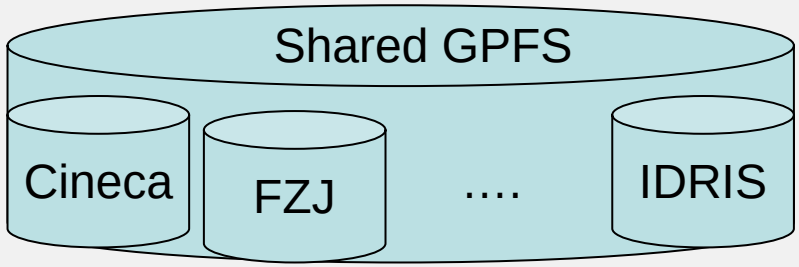
User Management



The Batch Scheduling System and the GPFS

Distributed European Infrastructure for Supercomputing Applications

- Shared High Performance File System over DEISA WAN for the DEISA homogeneous platform
- IBM General Parallel File System (2.3): cross mounted between all sites.



The Batch Scheduling System

Distributed European Infrastructure for Supercomputing Applications

AIX cluster -> LoadLeveler

IBM LoadLeveler (3.3): multi-clusters capability

Able to migrate submitted jobs to remote clusters

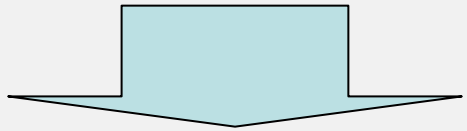
Common Production Environment: It is a standardization effort composed by a set of software tools like, shells, compilers, libraries, etc.

- They are present on each DEISA resources.
- They have the same release version and same configuration

DEISA Grid

Distributed European Infrastructure for Supercomputing Applications

- Common user account system
- GPFS
- LoadLeveler ‘multi-clusters’
- CPE



Migrated jobs find the same environment!

Tightly Integrated Grid

- Application can run all over DEISA resources without modifications

DEISA access method

Distributed European Infrastructure for Supercomputing Applications

Access method

- Unix Shell (only on the local site)
- Unicore
- Web portals (under development)

UNICORE

Distributed European Infrastructure for Supercomputing Applications

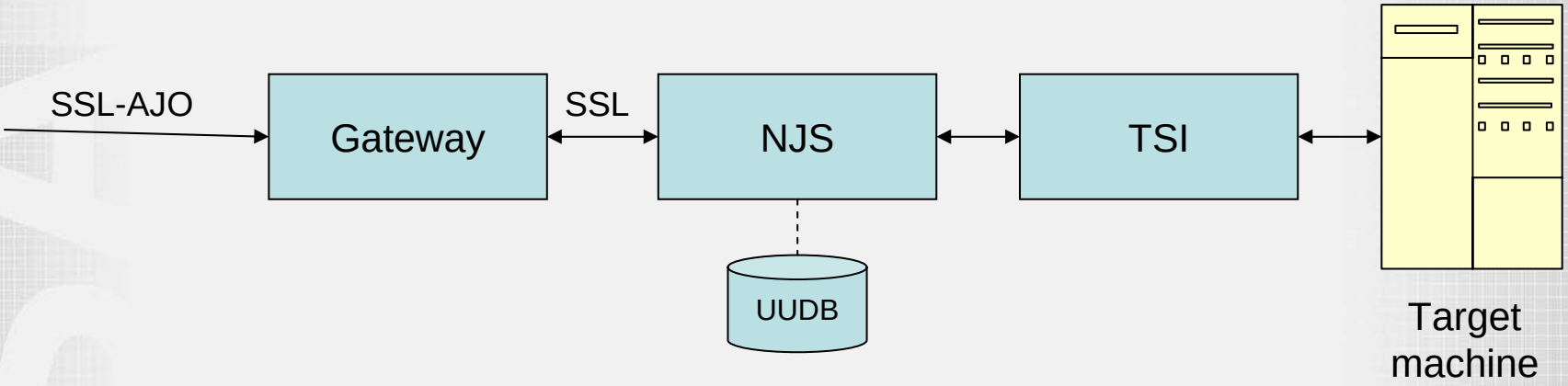
UNICORE (UNiform Interface to Computing REsources) provides a seamless interface for preparing and submitting jobs to a wide variety of heterogeneous distributed computing resources

- Job workflow management
- High abstraction level
- Single Sign-on based on PKI infrastructure

It is composed by:

- Client: it provides intuitive GUI for job and data management
- Gateway
- NJS
- TSI

UNICORE server components



Gateway: It is the entry point for all the incoming connection it is responsible to check for the validity of the SSL connection.

NJS: It is in charge for the translation of incoming AJO into concrete batch job translation. Mapping between user certificate and Unix user (UUDB)

TSI: It executes the command on the target machine.

Unicore deployment in DEISA

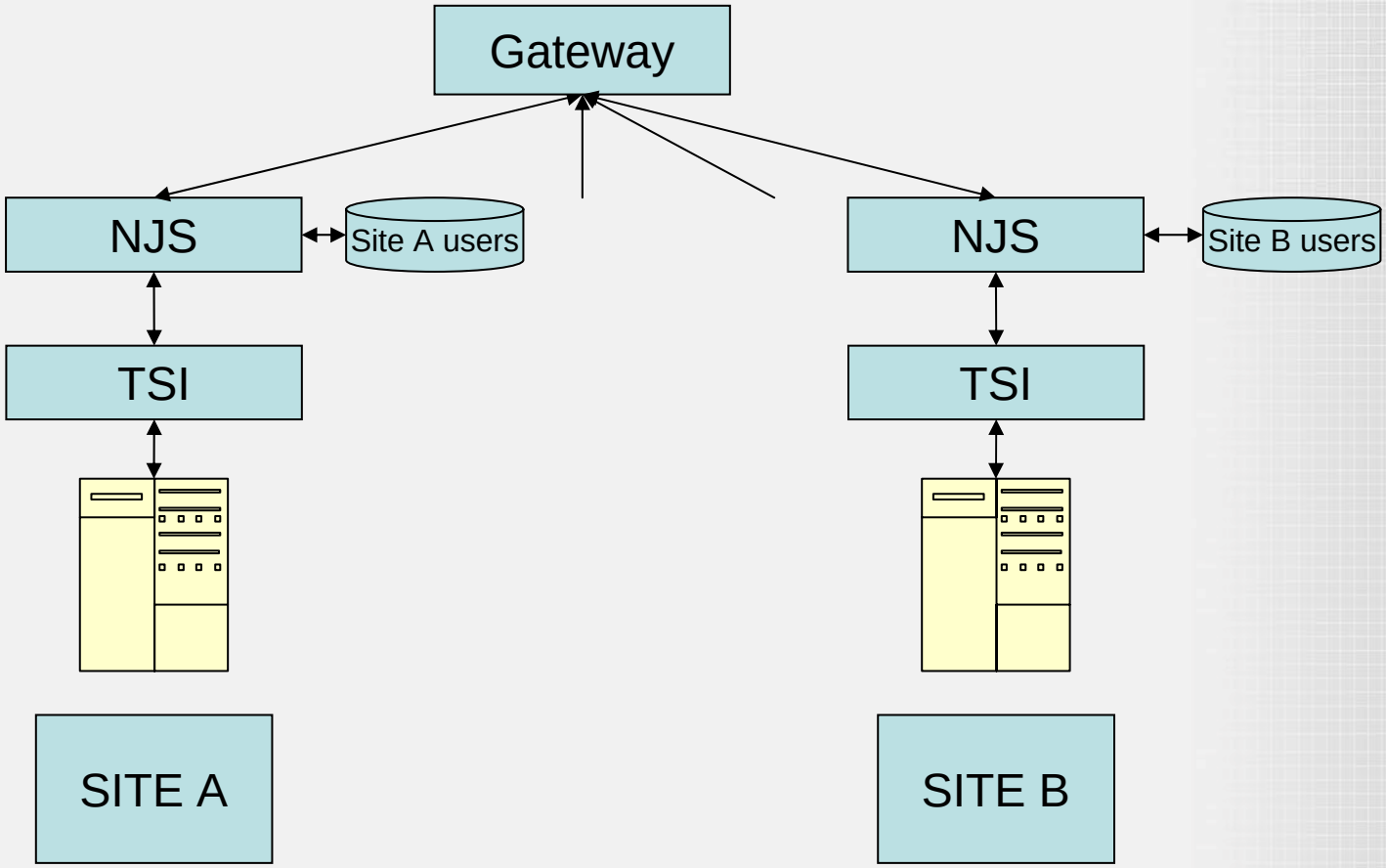
Distributed European Infrastructure for Supercomputing Applications

- Fully meshed interconnection between Gateways and NJSs
- DEISA UNICORE User DataBase
- GPFS and LL adaptation for UNICORE

Unicore deployment in DEISA

Distributed European Infrastructure for Supercomputing Applications

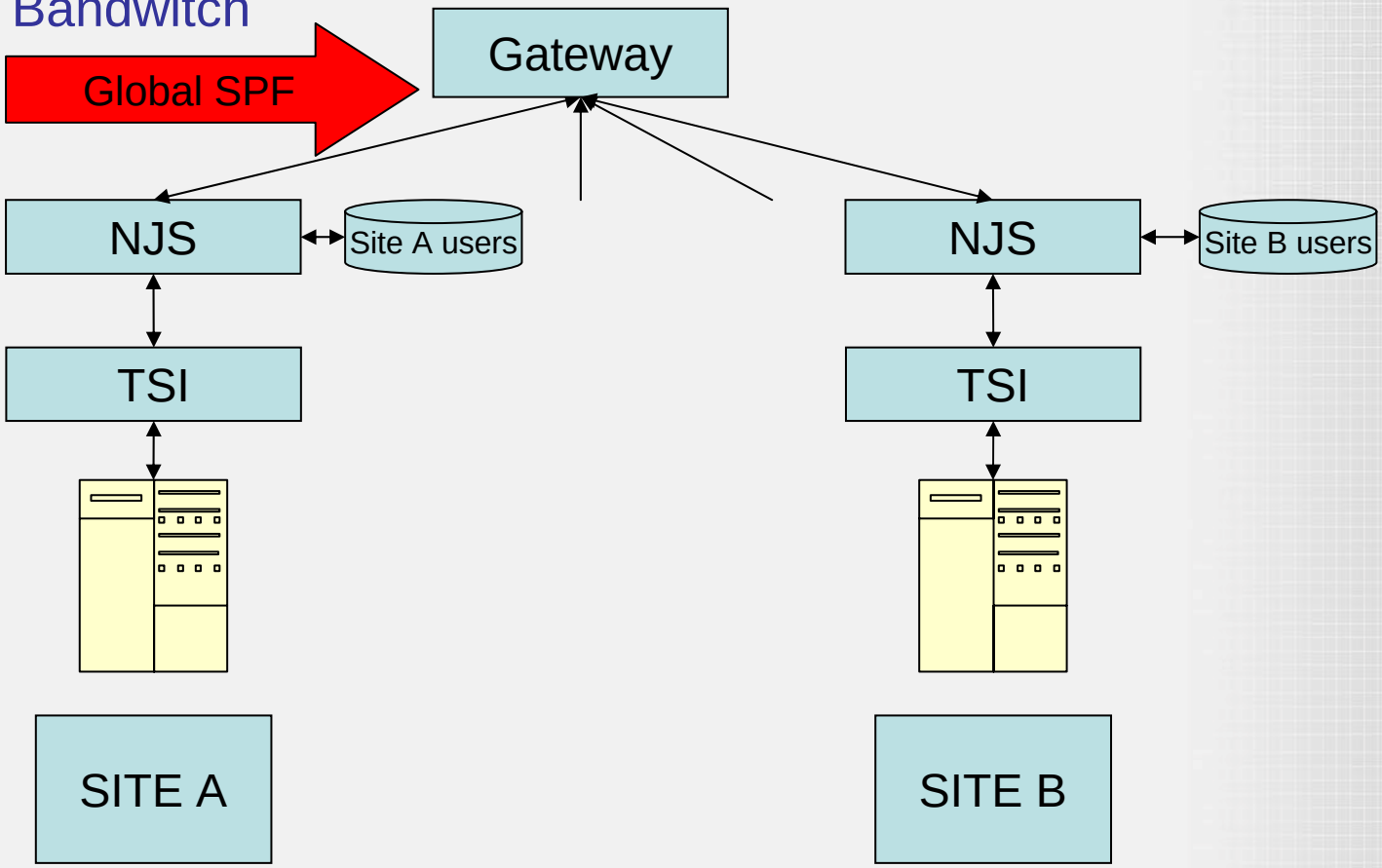
Standard UNICORE deployment



Unicore deployment in DEISA

Problem:

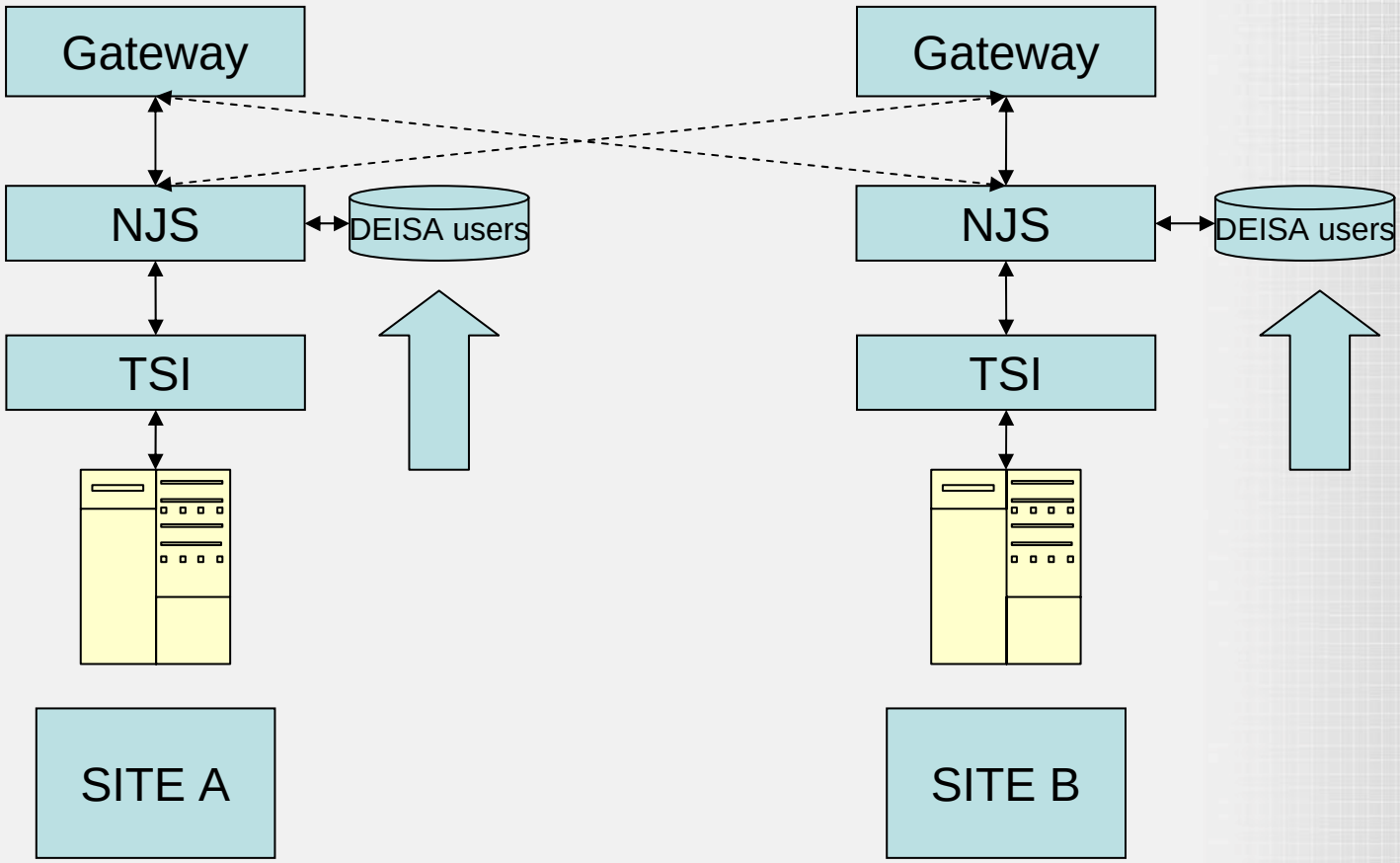
- Sigle Point of Failure
- Bandwitch



Unicore deployment in DEISA

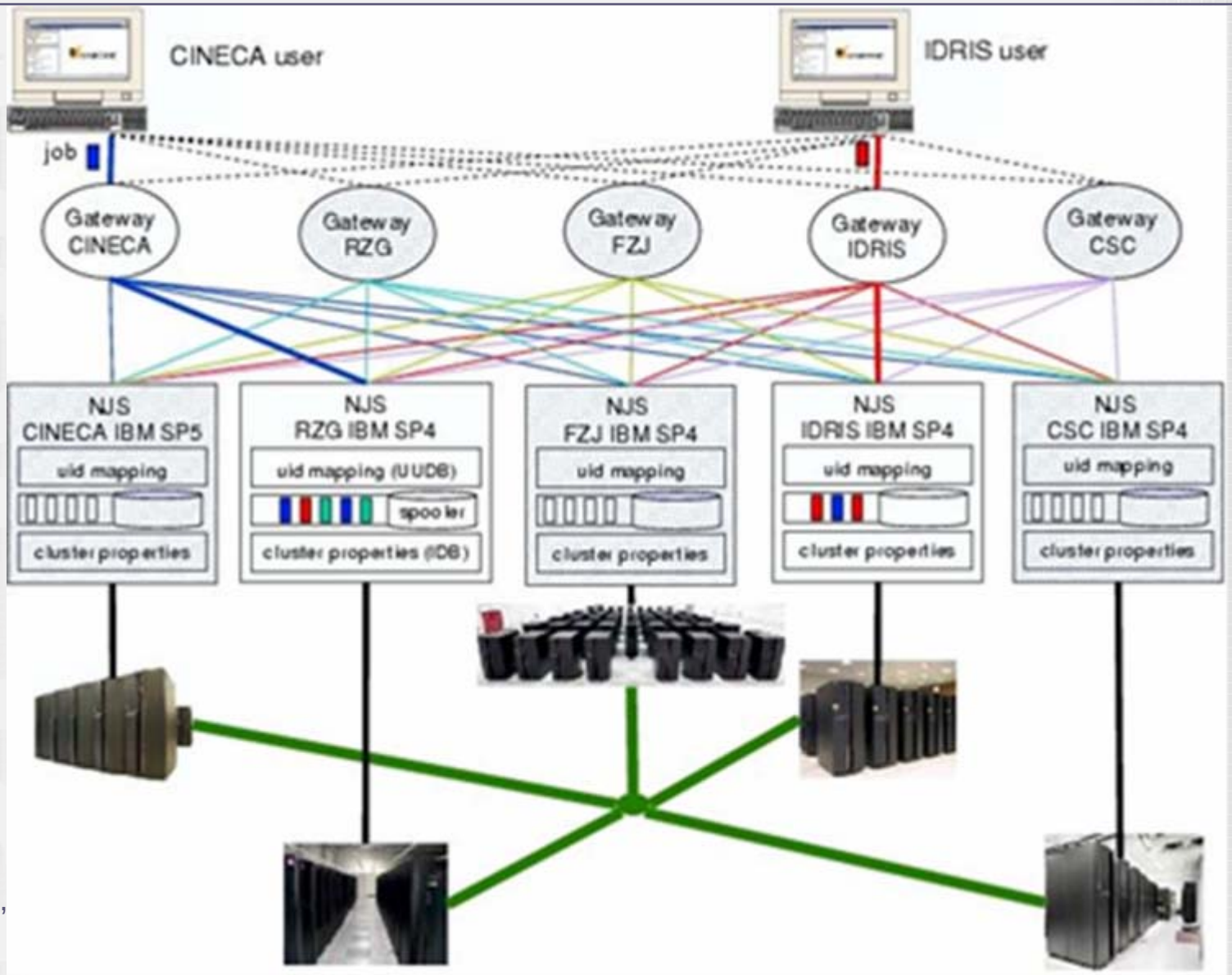
Distributed European Infrastructure for Supercomputing Applications

DEISA UNICORE deployment



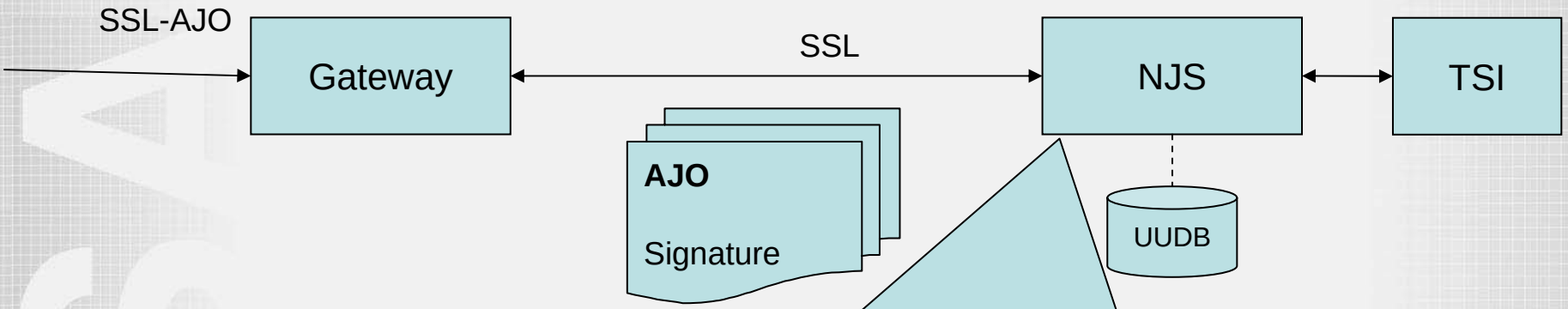
Unicore deployment in DEISA

Distributed European Infrastructure for Supercomputing Applications



DEISA UNICORE User DataBase (UUDB)

Distributed European Infrastructure for Supercomputing Applications



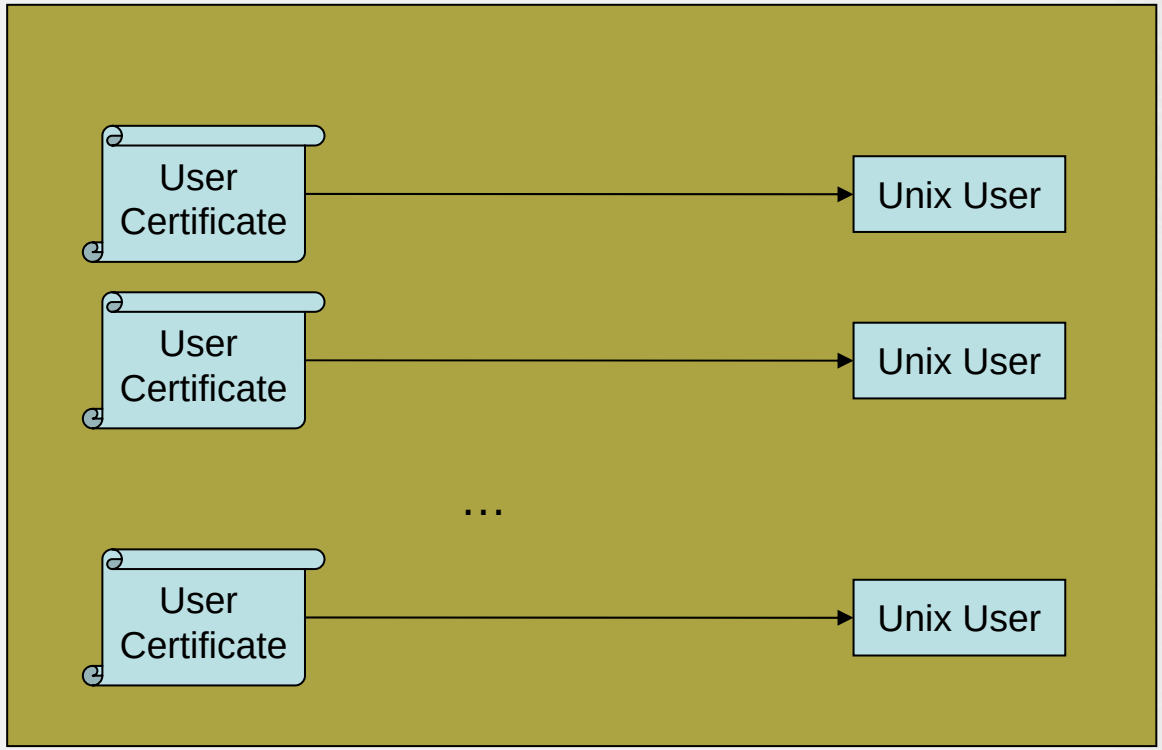
NJS checks that:

- The signature has been generated with the user's certificate
- The certificate is not expired
- The certificate has been issued by a trusted CA
- The certificate is present in the UUDB
  - Standard UUDB: the certificate is exactly the same
  - DEISA UUDB: the DN of the certificate is the same

DEISA UNICORE User DataBase (UUDB)

Distributed European Infrastructure for Supercomputing Applications

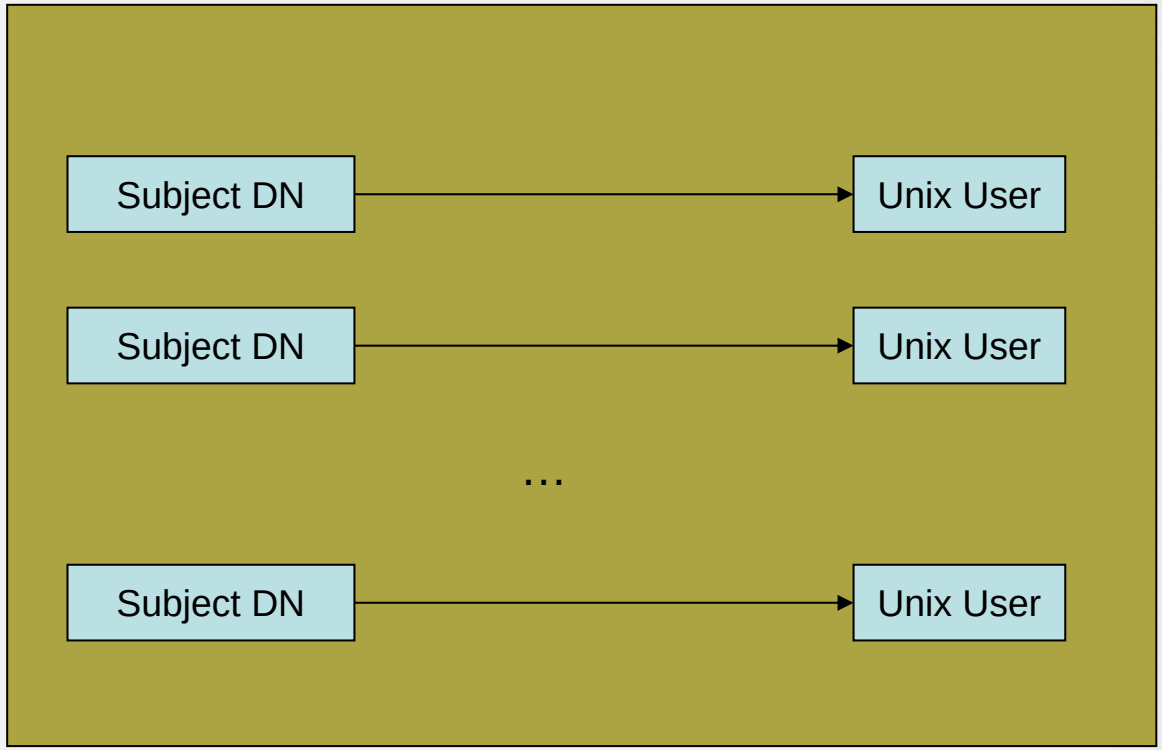
Standard UNICORE UUDB



DEISA UNICORE User DataBase (UUDB)

Distributed European Infrastructure for Supercomputing Applications

DEISA UNICORE UUDB



DEISA UNICORE User DataBase (UUDB)

Distributed European Infrastructure for Supercomputing Applications

This poses some restrictions on how to choose the set of trusted CAs by DEISA:

The DN of user's certificate released by the DEISA root CAs shall never be equal.

Advantages:

- Only the DN needs to be exchanged instead of the entire certificate
- If a certificate expires there is no need to modify the UUDB
- Easier integration with Globus authentication system (GSI Grid-mapfile)

DEISA Batch Scheduling System

Distributed European Infrastructure for Supercomputing Applications

DEISA partners agreed on a common set of parameters to describe a job:

- total tasks
- threads per task
- wall clock limit
- CPU limit
- data memory limit
- stack memory limit

Using a work-around (environmental variables) UNICORE can handle to the scheduler these parameters.

DEISA GPFS

Distributed European Infrastructure for Supercomputing Applications

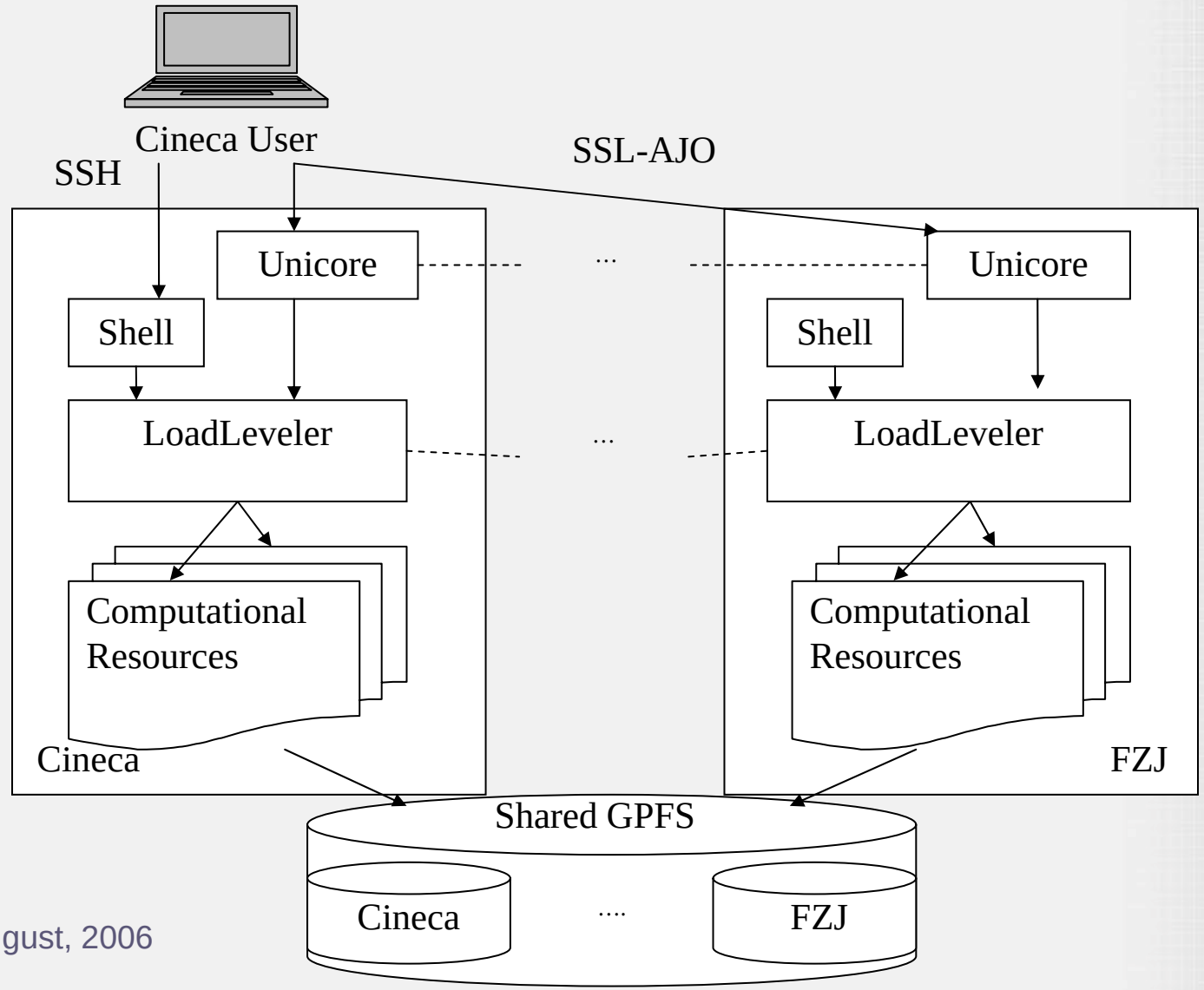
UNICORE uses temporary working directory called USPACE to place standard output and input

USPACE has been placed under GPFS file system with a common naming path shared by all sites.

A job submitted with UNICORE and migrated with LoadLeveler will find its files on the remote cluster.

DEISA Final Picture

Distributed European Infrastructure for Supercomputing Applications



Conclusions

Distributed European Infrastructure for Supercomputing Applications

- Flexible Architecture
  - Fault tolerant configuration
  - Transparent to underling technologies
- Integration of the heterogeneous platform
  - Meta scheduler that sits on the top of LoadLeveler and other schedulers
  - It can be already done by UNICORE
- Interoperability with Globus
  - Some Globus components will be deployed in DEISA (GRID-FTP, GRAM, and MDS)
  - Several interfaces for interoperability with Globus developed in European projects (GRIP, UniGrids)

Acknowledgments

Thanks to all the DEISA partners for their ideas.

And…

Thank you for your attention

http://www.deisa.org