JMEA

Job Manager Enterprise Application

Thomas Soddemann, RZG

Overview

- RZG and DEISA
- DEISA and its resources
- Access to Resources in DEISA (s.a. next talk)
- Material Science and Plasma Physics Portal requirements
- JMEA

History of supercomputing at the RZG

1962: IBM 7090

1969: IBM 360/91

1979: Cray-1

1998: Cray T3E/816

1999: NEC SX-5/3C

2002/2003: IBM p690

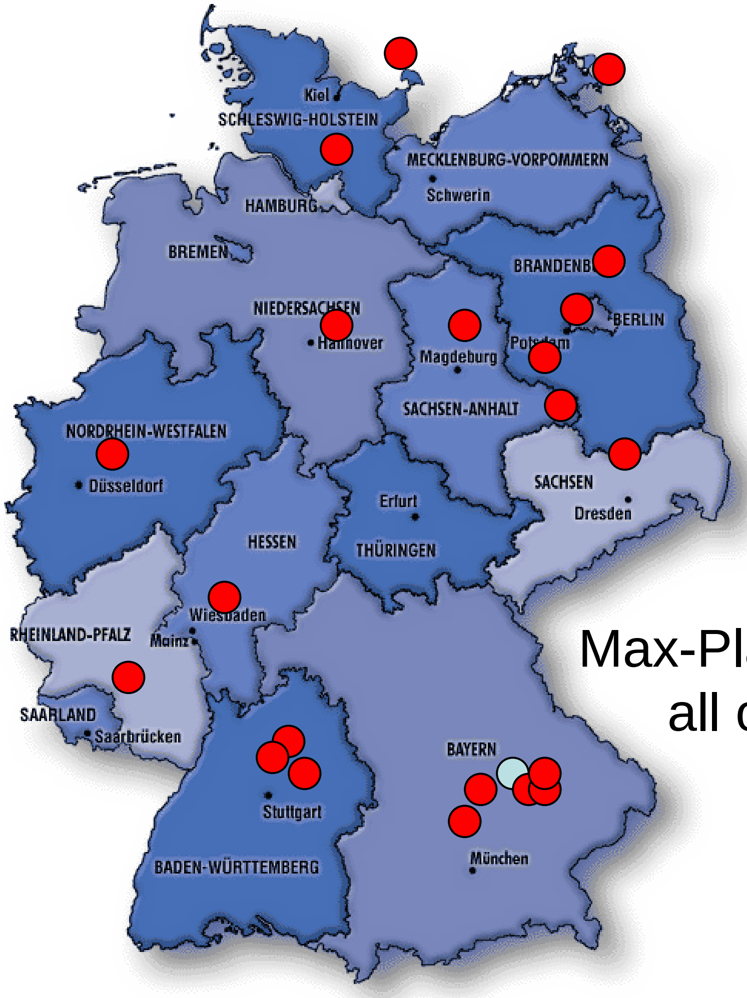
User community

- Supercomputing and Application Support

- Data management and long-term archives

- Data acquisition systems for fusion experiments

- Bioinformatics platform



Users from Max-Planck-Institutes all over Germany, Italy, and the Netherlands

DEISA – Distributed European Infrastructure for Supercomputing Applications

WWW.DEISA.ORG
RZG IPP
iris CNRS
epcc
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
BSC
CSC
sara
HLRS
lrz
European Centre for Medium-Range Weather Forecasts
Forschungszentrum Jülich
in der Helmholtz-Gemeinschaft
CINECA
Consorzio Interuniversitario

Partners

Institut du Développement et des Ressources en Informatique Scientifique, France

Forschungszentrum Jülich, Germany

Rechenzentrum Garching of the Max Planck Society, Germany

Consorzio Interuniversitario, Italy

Edinburgh Parallel Computing Centre, UK

SARA Computing and Networking Services, The Netherlands

Finnish Information Technology Center for Science, Finland

European Centre for Medium-Range Weather Forecasts, UK

High Performance Computing Center, Germany

Leibniz Computing Centre of the Bavarian Academy of Sciences and Humanities, Germany

Barcelona Supercomputing Center (BSC), Spain

DEISA

- DEISA is an European Supercomputing Service built on top of existing national services.

- DEISA deploys and operates a persistent, production quality, distributed supercomputing environment with continental scope

DEISA

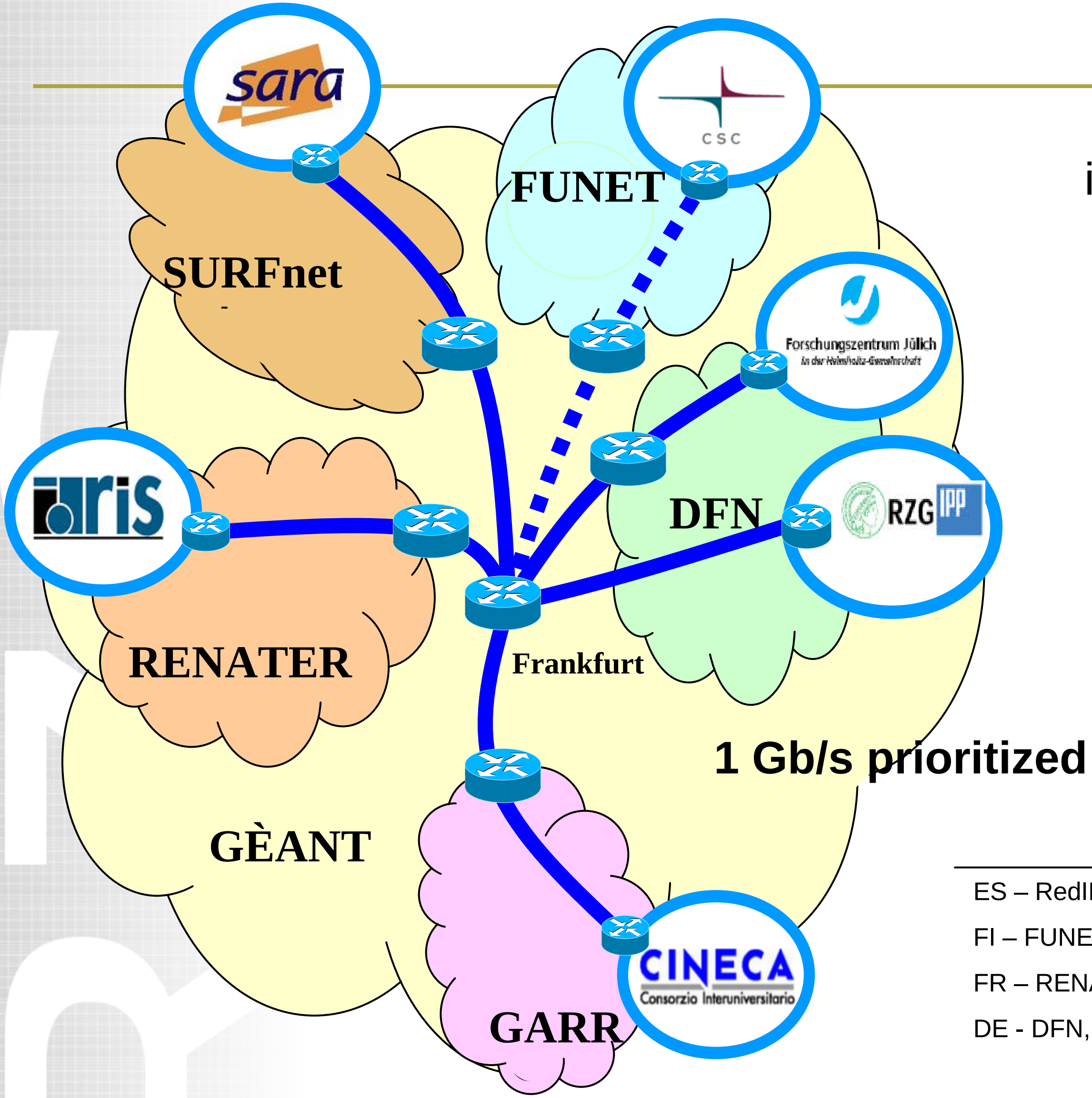
AIX distributed super-cluster

THE DEISA SUPERCOMPUTING GRID

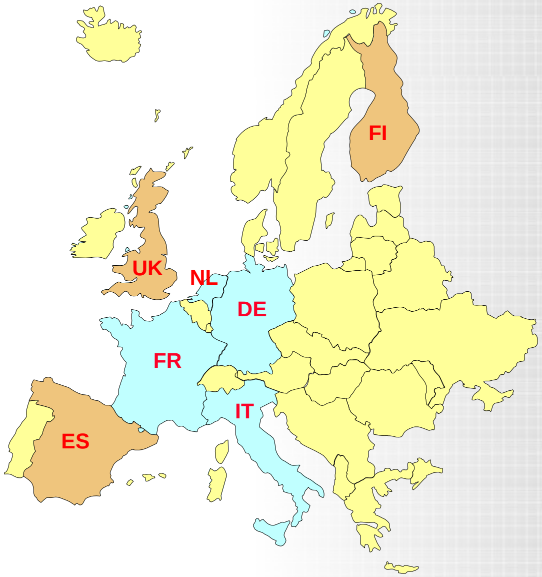
Linux systems (SGI, IBM, …)

Vector systems (NEC, …)

Deisa Network Status



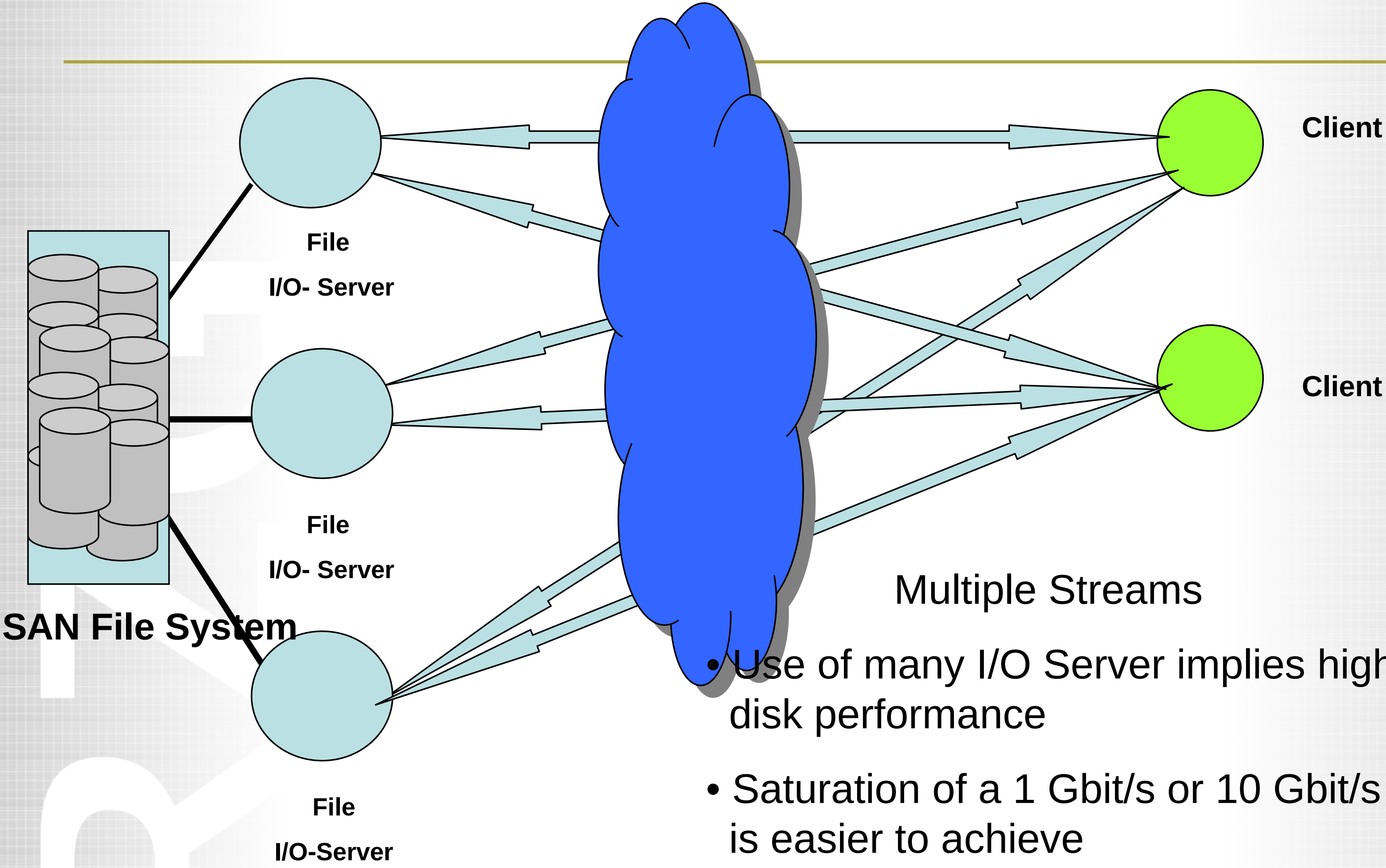
Dedicated network infrastructure using „Premium IP“



Participating NRENs

ES – RedIRIS, Spain	IT – GARR, Italy
FI – FUNET, Finland	NL – SURFNET, The Netherlands
FR – RENATER, France	UK – UKERNA/JANET, UK
DE - DFN, Germany	

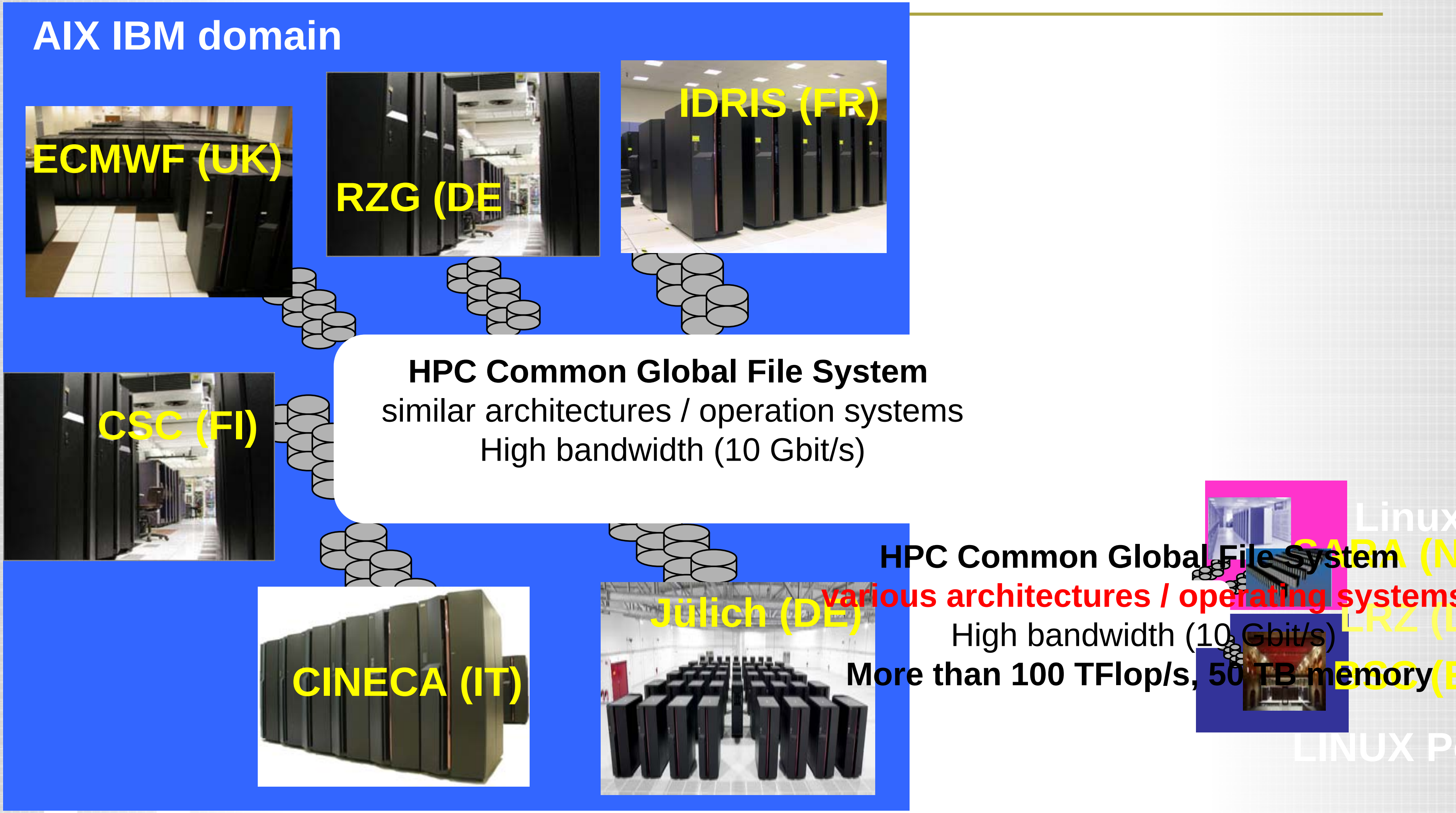
MC-GPFS Multiple Network Streams



Multiple Streams

- Use of many I/O Server implies high disk performance
- Saturation of a 1 Gbit/s or 10 Gbit/s line is easier to achieve

The DEISA Super Cluster in 2005/2006

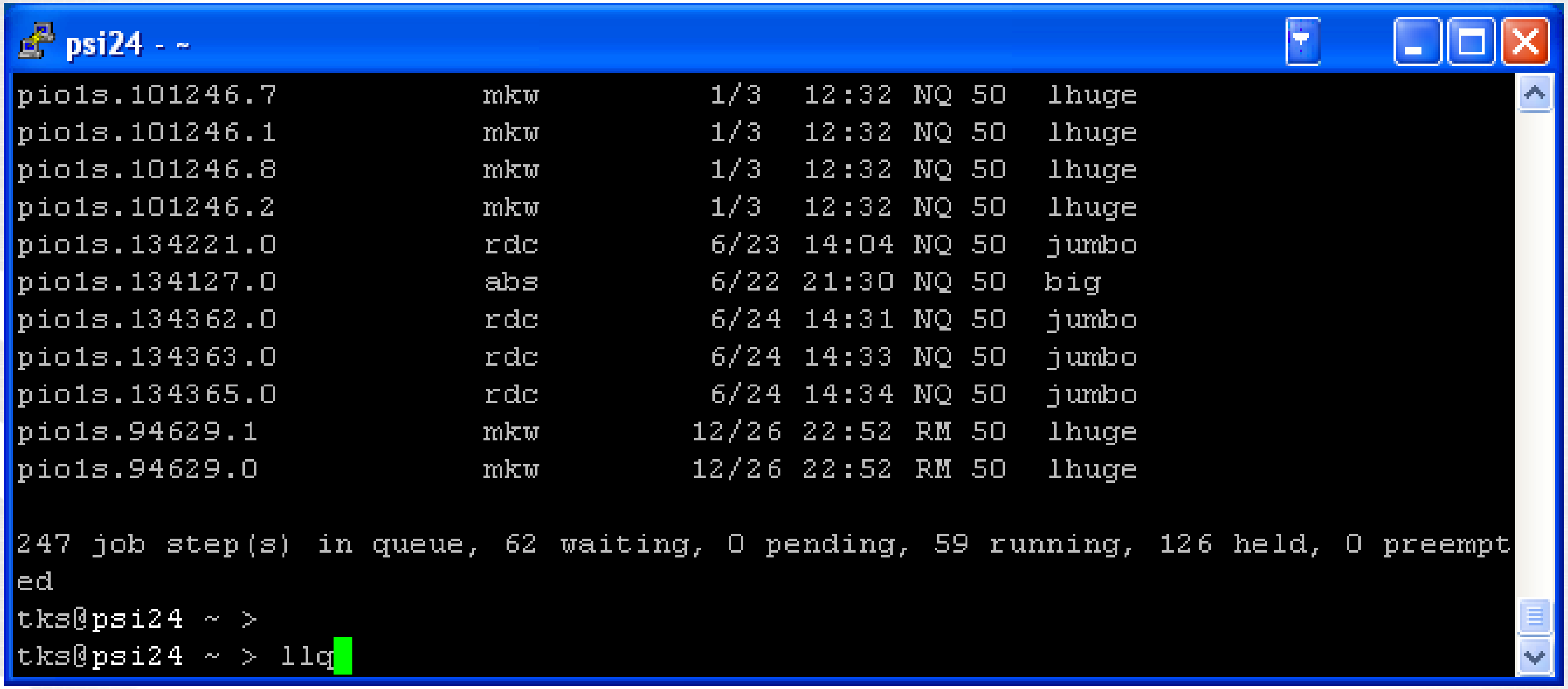


HPC Common Global File System
similar architectures / operation systems
High bandwidth (10 Gbit/s)

HPC Common Global File System
various architectures / operating systems
High bandwidth (10 Gbit/s)
More than 100 TFlop/s, 50 TB memory

Ways of accessing resources: CLI

ssh, qsub/llsubmit, qstat/llq, …



```
pio1s.101246.7          mkw          1/3   12:32 NQ 50   lhuge
pio1s.101246.1          mkw          1/3   12:32 NQ 50   lhuge
pio1s.101246.8          mkw          1/3   12:32 NQ 50   lhuge
pio1s.101246.2          mkw          1/3   12:32 NQ 50   lhuge
pio1s.134221.0          rdc          6/23 14:04 NQ 50   jumbo
pio1s.134127.0          abs          6/22 21:30 NQ 50   big
pio1s.134362.0          rdc          6/24 14:31 NQ 50   jumbo
pio1s.134363.0          rdc          6/24 14:33 NQ 50   jumbo
pio1s.134365.0          rdc          6/24 14:34 NQ 50   jumbo
pio1s.94629.1           mkw         12/26 22:52 RM 50   lhuge
pio1s.94629.0           mkw         12/26 22:52 RM 50   lhuge

247 job step(s) in queue, 62 waiting, 0 pending, 59 running, 126 held, 0 preempt
ed
tks@psi24 ~ >
tks@psi24 ~ > llq
```

Local Resource Management

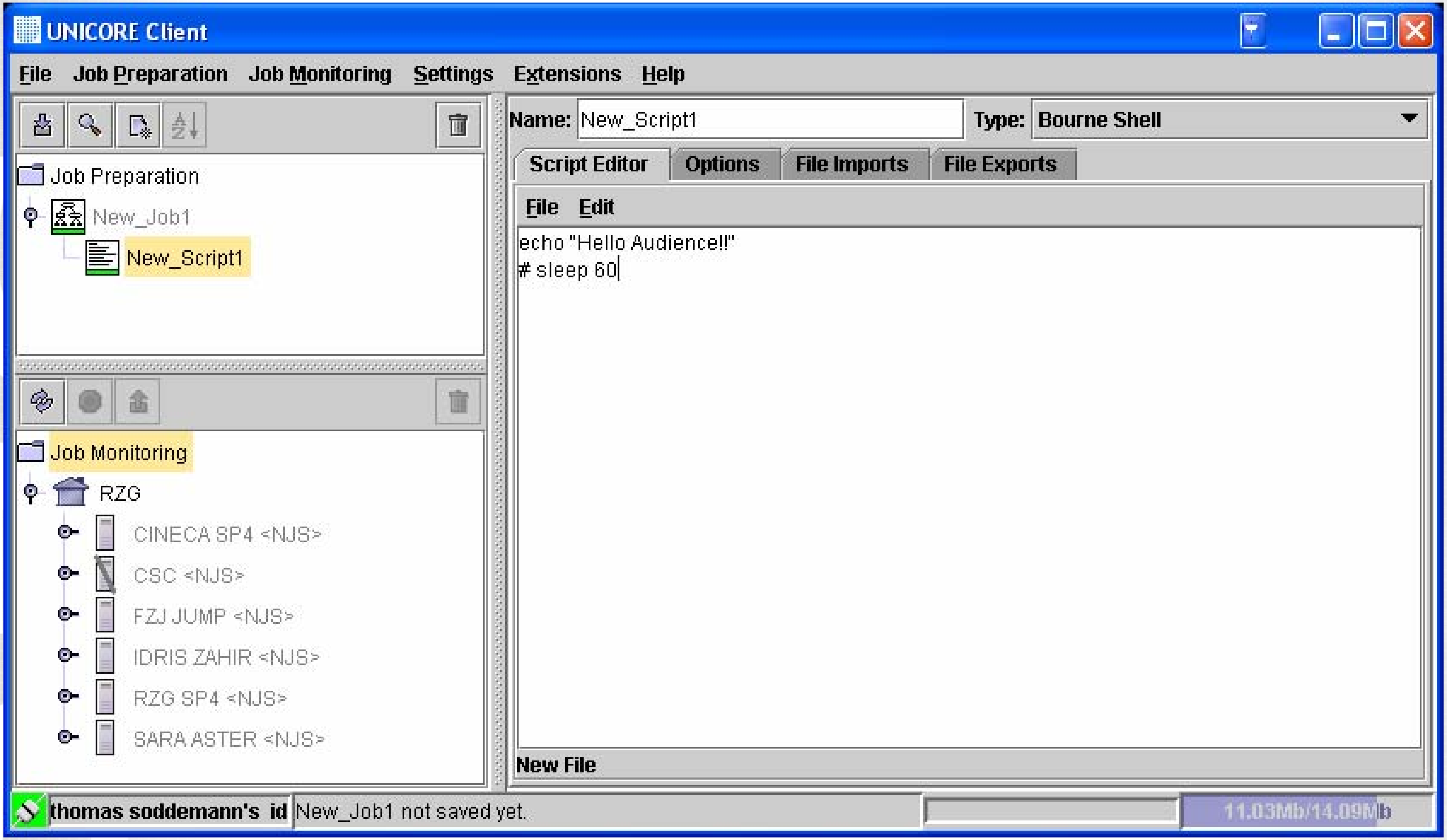
- Load Leveler
- LSF
- OpenPBS/PBSpro
- Sun Grid Engine
- Torque

- MC-LoadLeveler

Obvious disadvantages:

- Separate batch script for each environment
- No job rerouting

Ways of accessing resources: Rich Client Solution



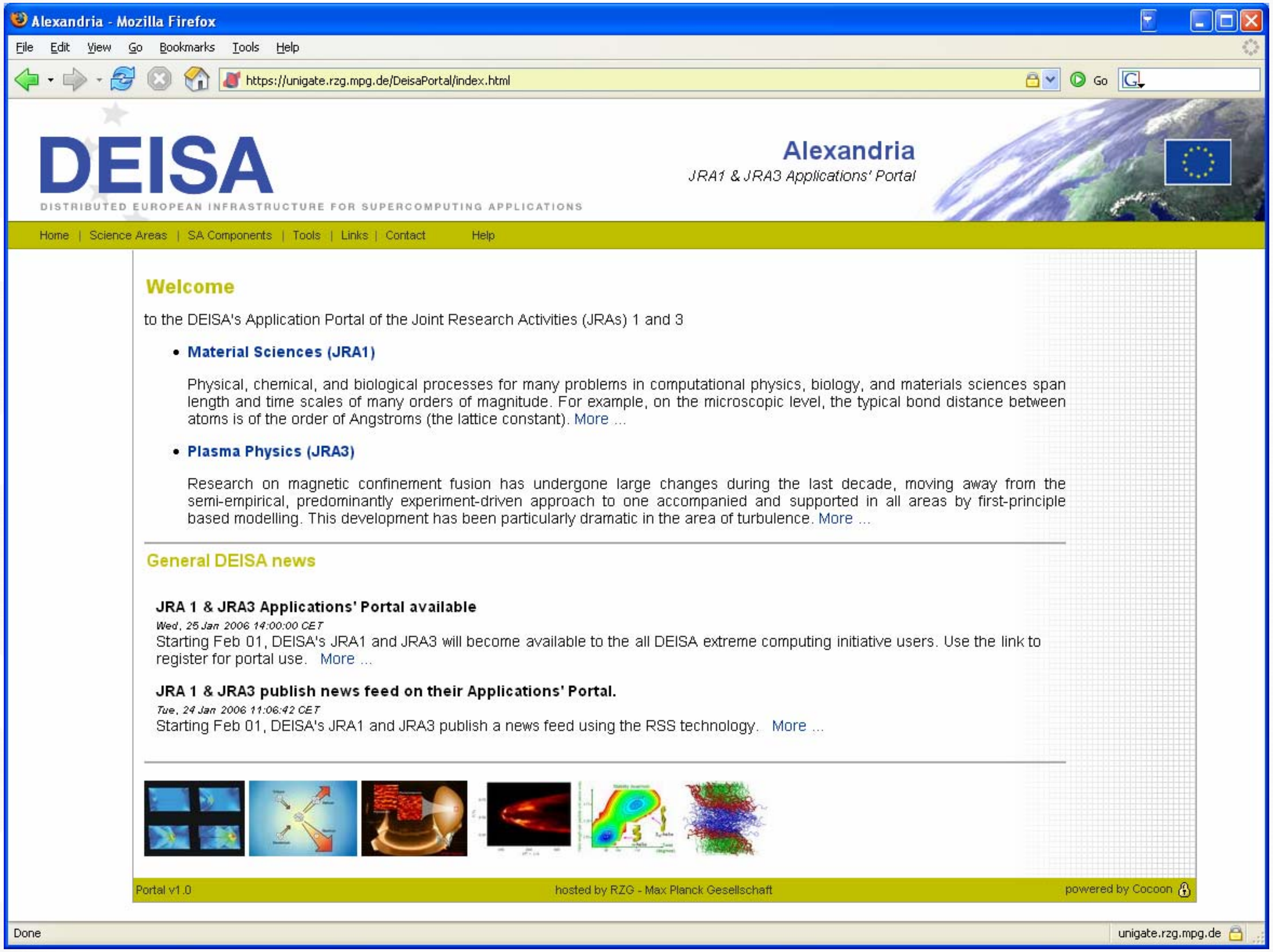
Name: New_Script1 Type: Bourne Shell

```
echo "Hello Audience!!"
# sleep 60
```

New File

thomas soddemann's id | New_Job1 not saved yet. | 11.03Mb/14.09Mb

Ways of accessing resources: Web Portal Solution


Alexandria - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://unigate.rzg.mpg.de/DeisaPortal/index.html
Go
DEISA
DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR SUPERCOMPUTING APPLICATIONS
Alexandria
JRA1 & JRA3 Applications' Portal
Home | Science Areas | SA Components | Tools | Links | Contact Help

Welcome

to the DEISA's Application Portal of the Joint Research Activities (JRAs) 1 and 3

- **Material Sciences (JRA1)**

  Physical, chemical, and biological processes for many problems in computational physics, biology, and materials sciences span length and time scales of many orders of magnitude. For example, on the microscopic level, the typical bond distance between atoms is of the order of Angstroms (the lattice constant). More ...

- **Plasma Physics (JRA3)**

  Research on magnetic confinement fusion has undergone large changes during the last decade, moving away from the semi-empirical, predominantly experiment-driven approach to one accompanied and supported in all areas by first-principle based modelling. This development has been particularly dramatic in the area of turbulence. More ...

General DEISA news

JRA 1 & JRA3 Applications' Portal available

Wed, 25 Jan 2006 14:00:00 CET

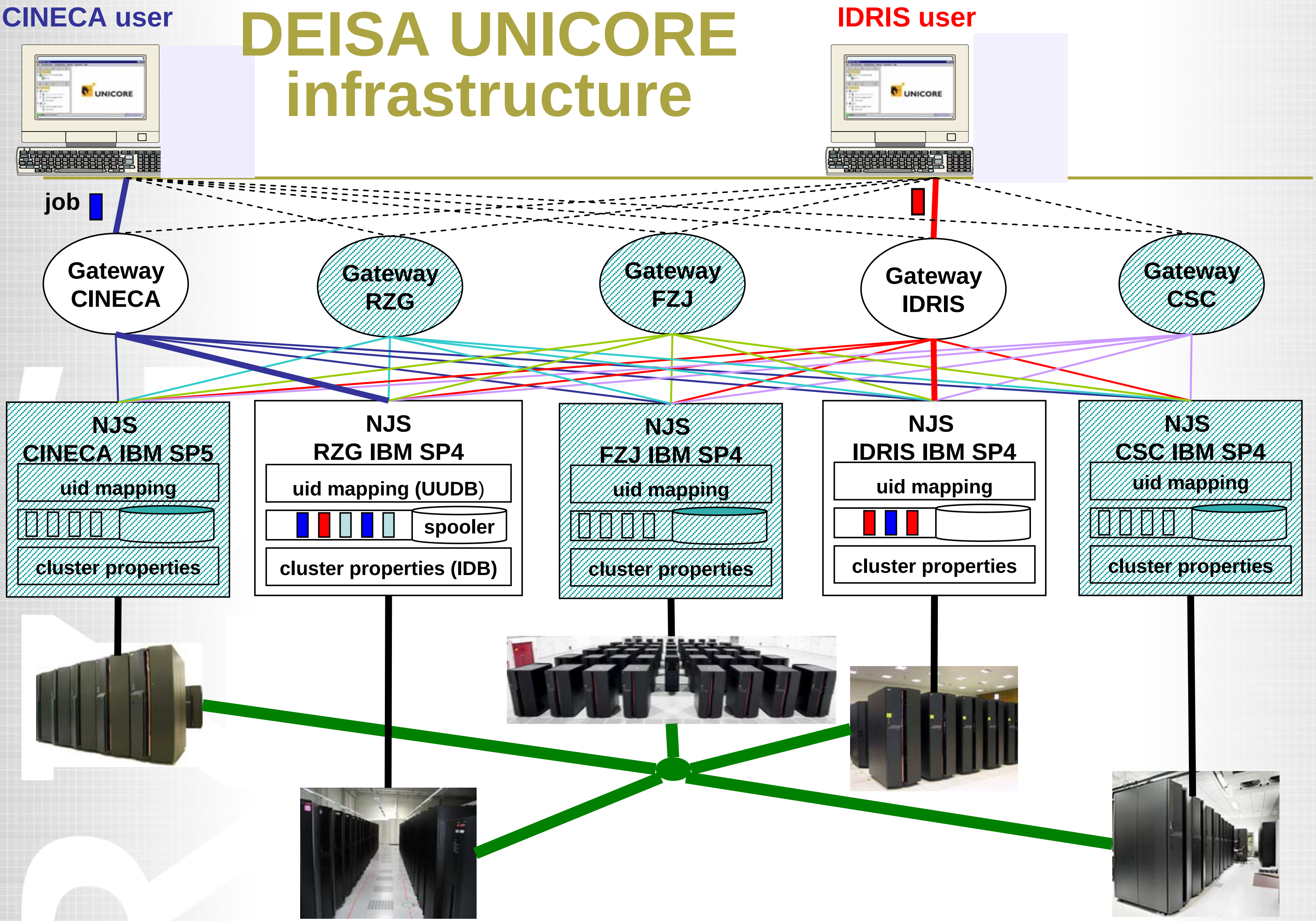
Starting Feb 01, DEISA's JRA1 and JRA3 will become available to the all DEISA extreme computing initiative users. Use the link to register for portal use. More ...

JRA 1 & JRA3 publish news feed on their Applications' Portal.

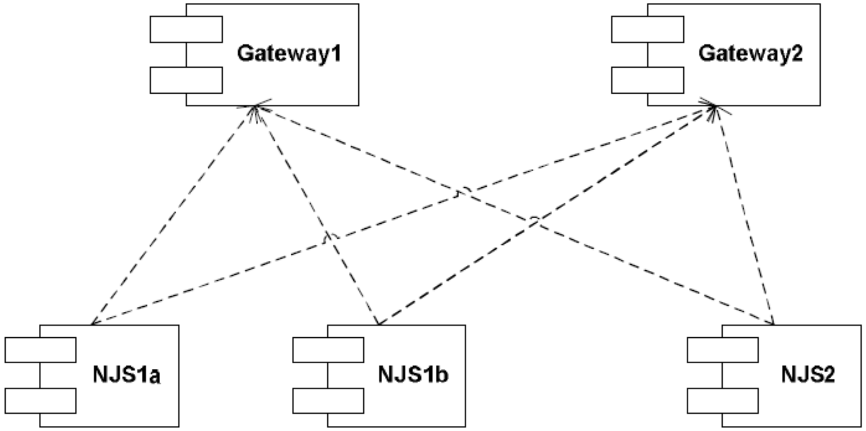
Tue, 24 Jan 2006 11:06:42 CET

Starting Feb 01, DEISA's JRA1 and JRA3 publish a news feed using the RSS technology. More ...

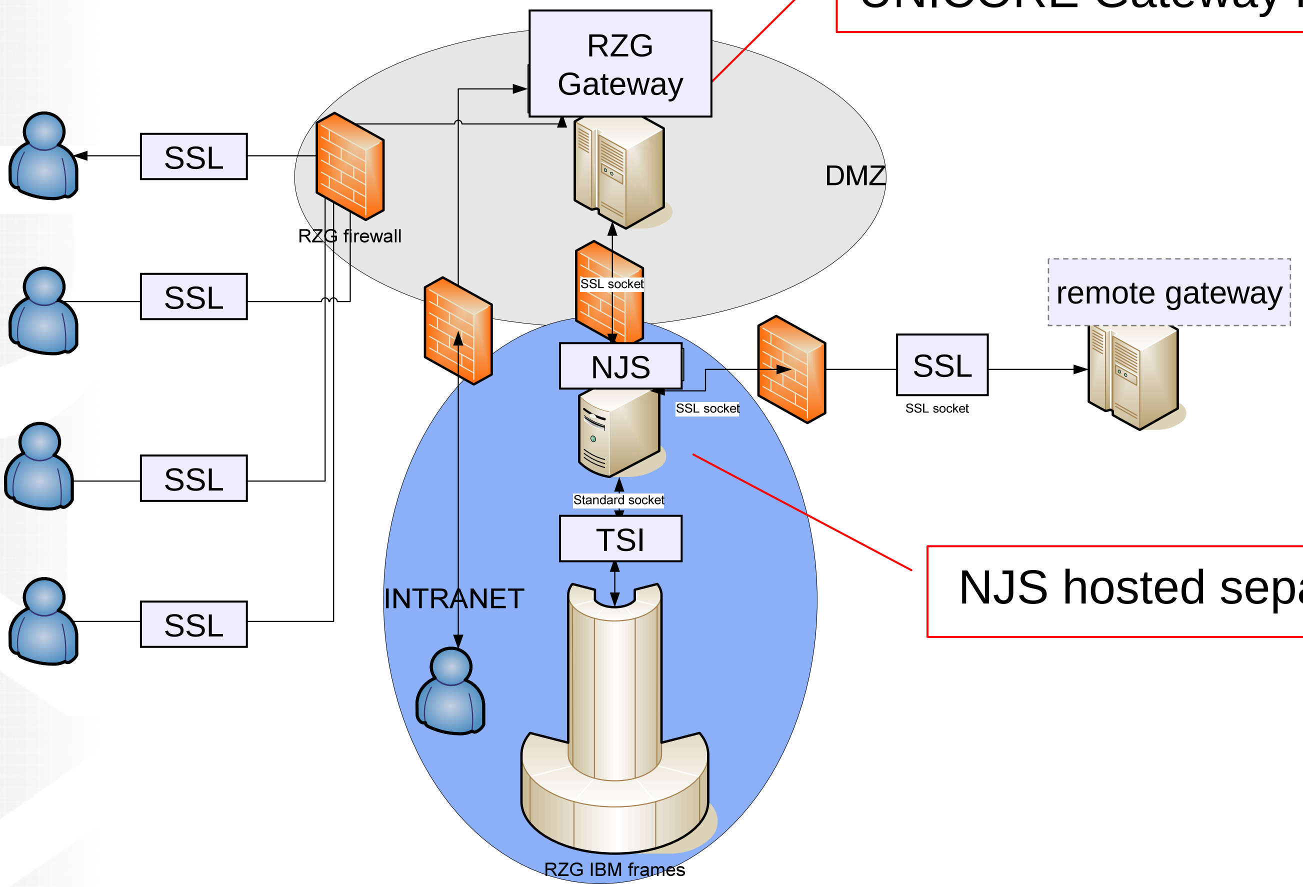
DEISA UNICORE infrastructure



UNICORE deployment in DEISA



UNICORE Configuration (RZG)



UNICORE Gateway in DMZ

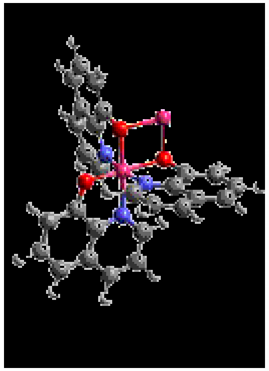
NJS hosted separately

DEISA Research Activities

JRA1 – Material Sciences

CPMD

CP2K



JRA3 – Plasma Physics

TORB

Requirements for a Portal Solution

- Compute Job Handling
  (submit, cancel, hold, status, …)
  => components holding job information

- Session management

- File staging support
- Remote file system access
- Database access

- User Administration (auth*)

Job Manager

Session Manager

Persistence Manager

Identity Manager

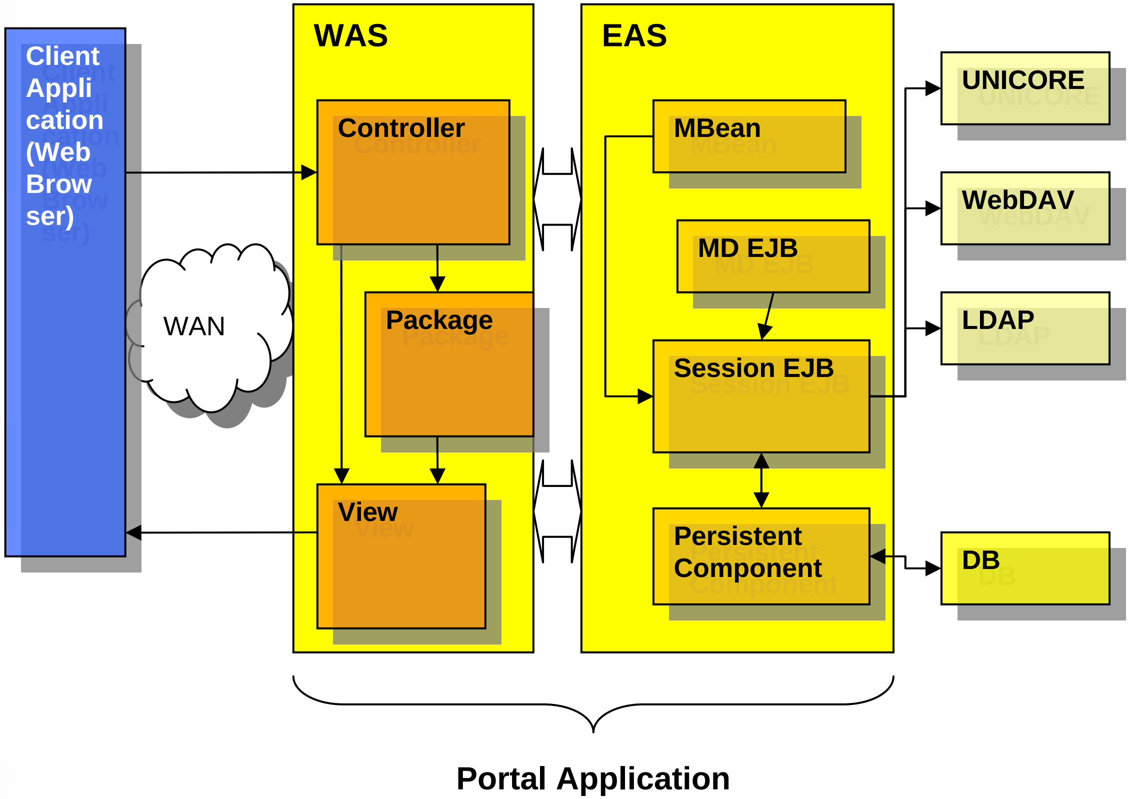
Advantages of Portal Applications

- Give the possibility to hide the complexity of Grid Infrastructures (sensible simplifications vs. mystification MS approach)
- Can give the impression of direct use of an application

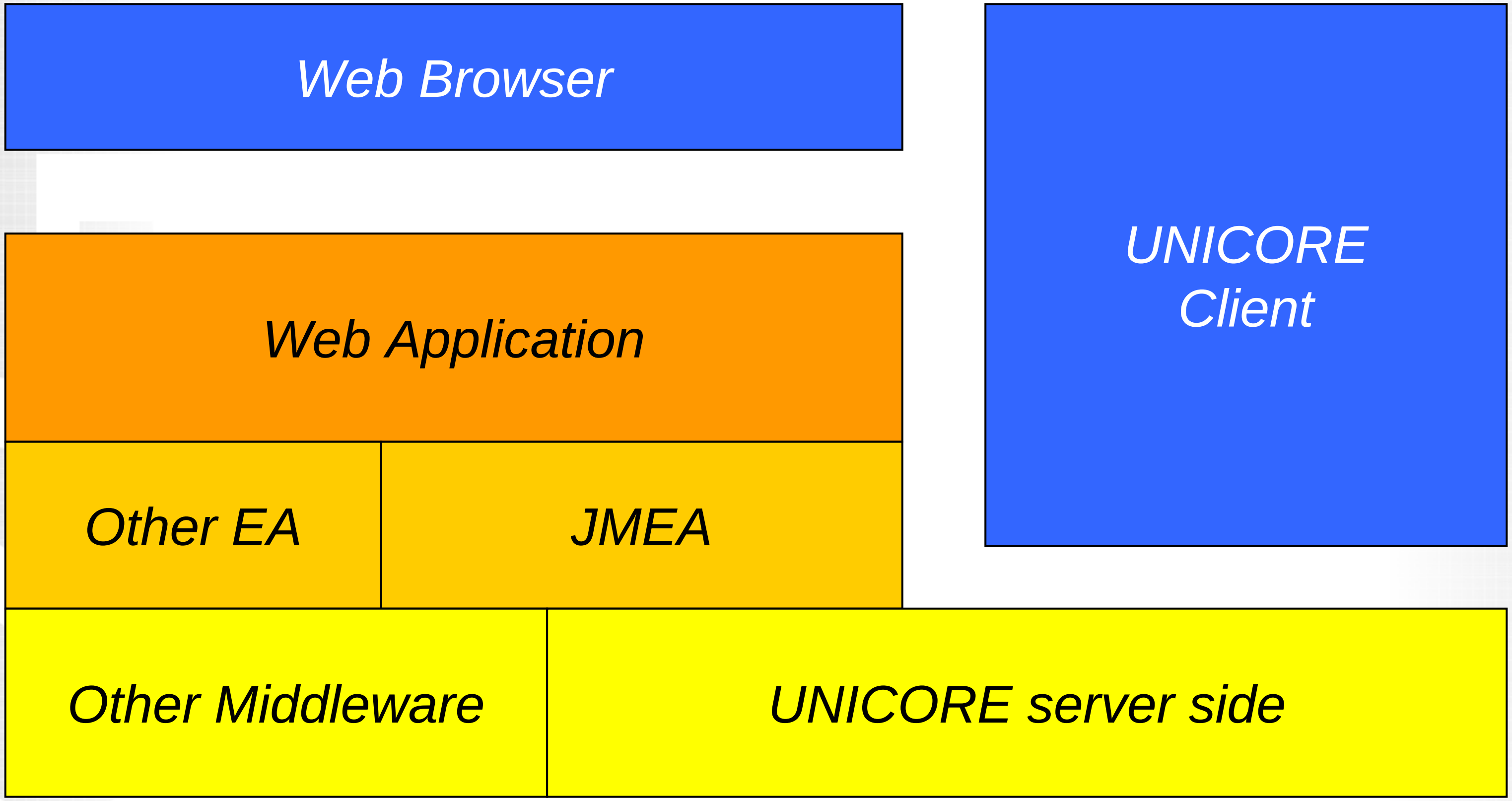
Application Service Provider -> ASP

- Can be accessed from almost everywhere

Architecture of a Web Based Enterprise Solution



Architecture of a Web Based Enterprise Solution



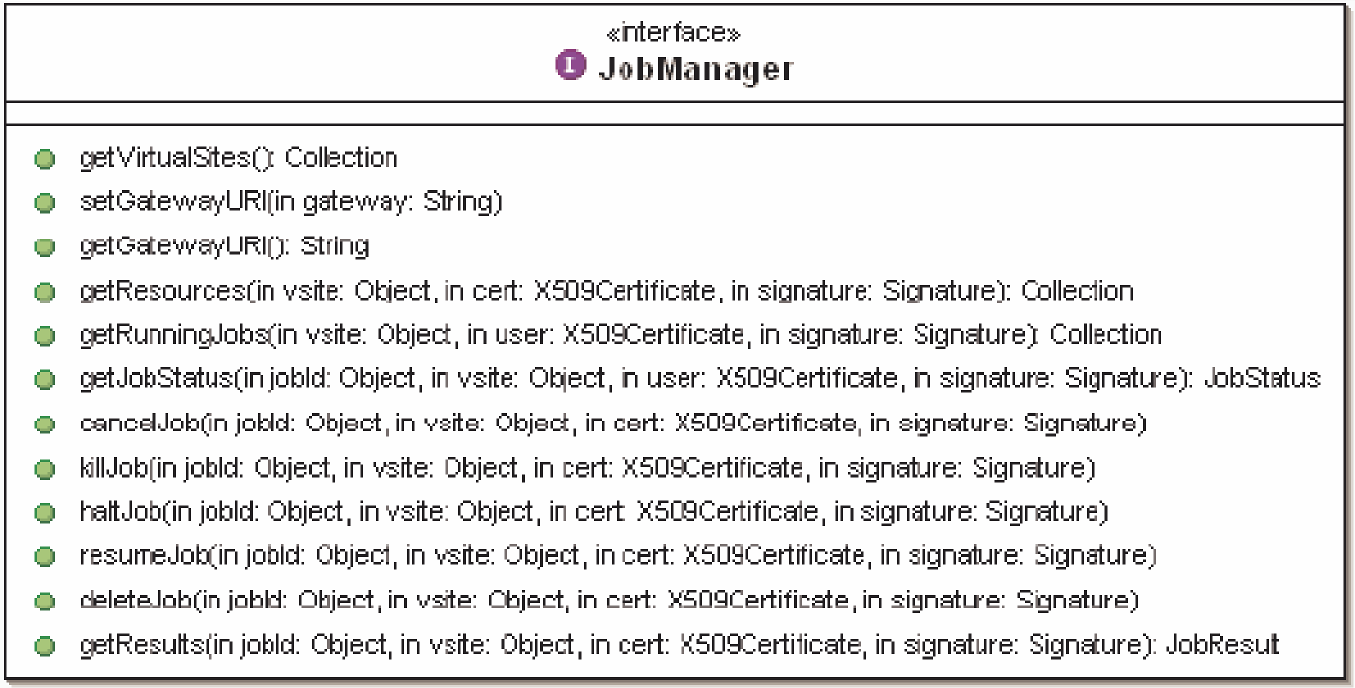
JobManager Methods

- submit (submitting the job request),
- cancel (canceling a job request which is not being executed)
- delete (delete a finished job request),
- kill (kill a running job request)
- halt (halt a job which is being executed)
- resume (resume a previously halted job).

- Publish information about
  - resources
  - status of jobs
  - fetch console output.

- In addition
  - Support of Proxy Certificates
  - Support of Explicit Trust delegation

The JobManager Interface



«interface»
JobManager

- getVirtualSites(): Collection
- setGatewayURI(in gateway: String)
- getGatewayURI(): String
- getResources(in vsite: Object, in cert: X509Certificate, in signature: Signature): Collection
- getRunningJobs(in vsite: Object, in user: X509Certificate, in signature: Signature): Collection
- getJobStatus(in jobId: Object, in vsite: Object, in user: X509Certificate, in signature: Signature): JobStatus
- cancelJob(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature)
- killJob(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature)
- haltJob(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature)
- resumeJob(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature)
- deleteJob(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature)
- getResults(in jobId: Object, in vsite: Object, in cert: X509Certificate, in signature: Signature): JobResult

Allows different implementations:

- UNICORE (primary target)
- Globus
- …

Arcon Library disadvantages (Multi User Application)

In order to avoid race conditions in multi threaded applications, one should

- omit static variables

  unless they are used for communication between the threads and their access is synchronized.

- synchronized access must not lead to a performance bottlenecks

- Keep in mind that the application might be clustered

Arcon Library disadvantages (Multi User Application)

Arcon Library defines:

- `outcome_dir` which specifies the directory, where streamed files will be stored
- `buffer_size` which reflects the buffer size for connections
- `always_poll` which tells if request are always asynchronous or not.

The abstract class Connection implements three static variables:

- `keep_open` which defines if the next connection is kept open after use.
- `compression` which tells if the transmission should be compressed or not.
- `encrypt` which defines whether the next retrieved connection should be encrypted communication or not.

Arcon Library disadvantages (Multi User Application)

Further disadvantages

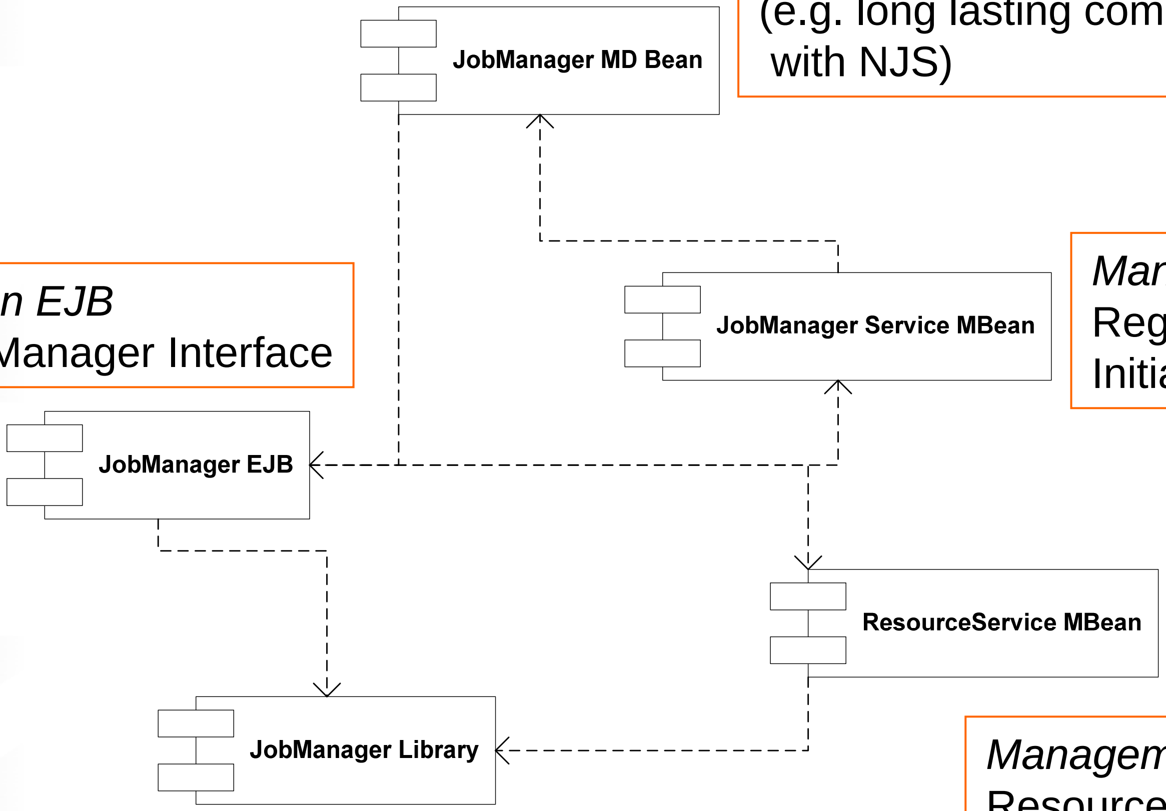
- Proprietary Logging
- Exceptions used for control flow rather than for error handling
- Missing support for ETD (at least in the official release)

New UNICORE JobManager library

- Need for a new job management library implementation for UNICORE
- Arcon partly code reused/refactored
- Reimplementation of problematic parts
- ETD support
- Proxy Certificate Support
- Thread safe

Job Management Enterprise Application - JMEA

Message Driven EJB
Performs asynchronous task
(e.g. long lasting communication
with NJS)

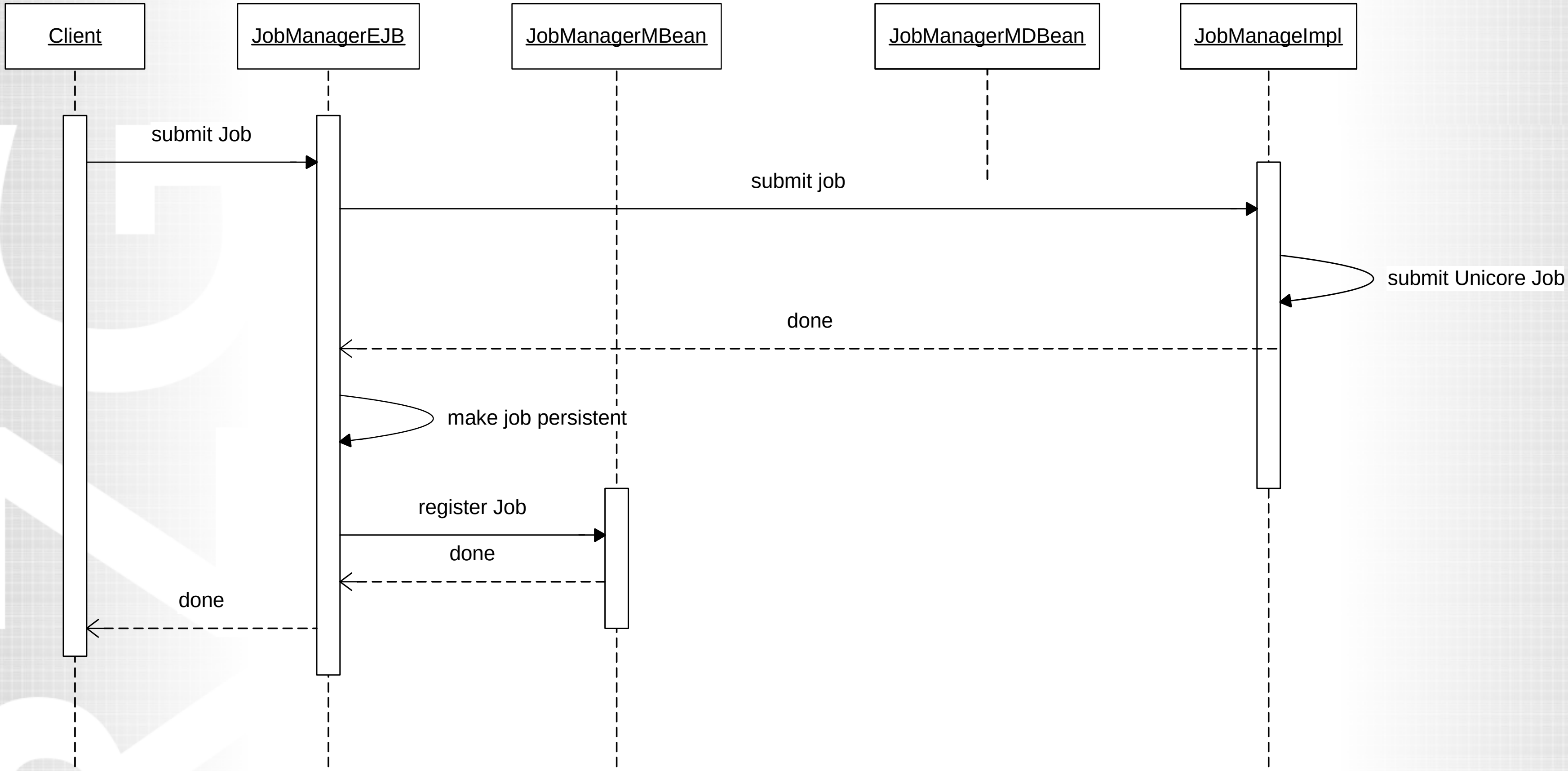


Stateless Session EJB
Implements JobManager Interface

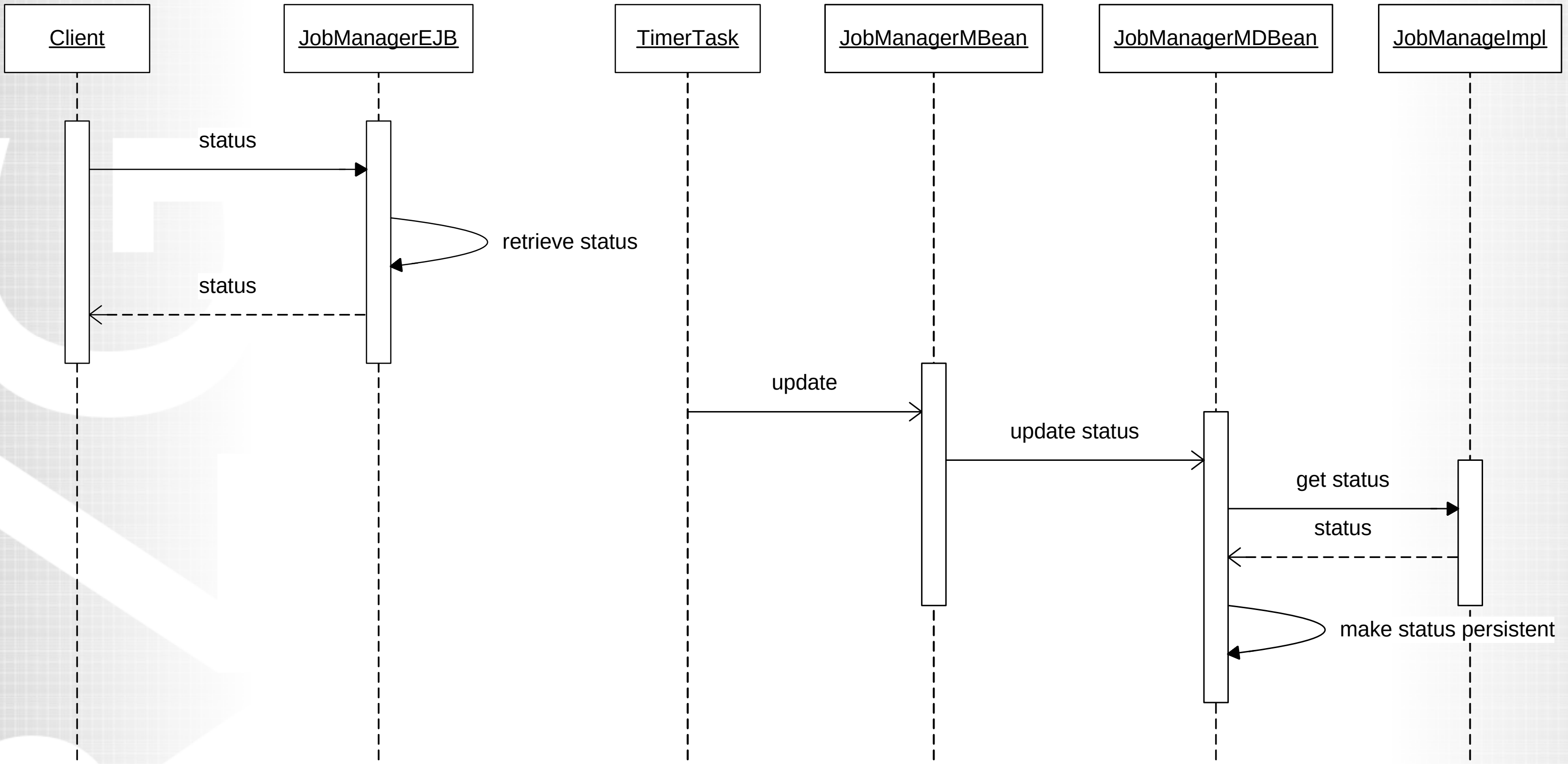
Management Bean
Registry for Jobs
Initiates DB updates

Management Bean
Resource information cache

Job Submission Sequence



Job Status Request Sequence



Facts on JMEA

Advantages

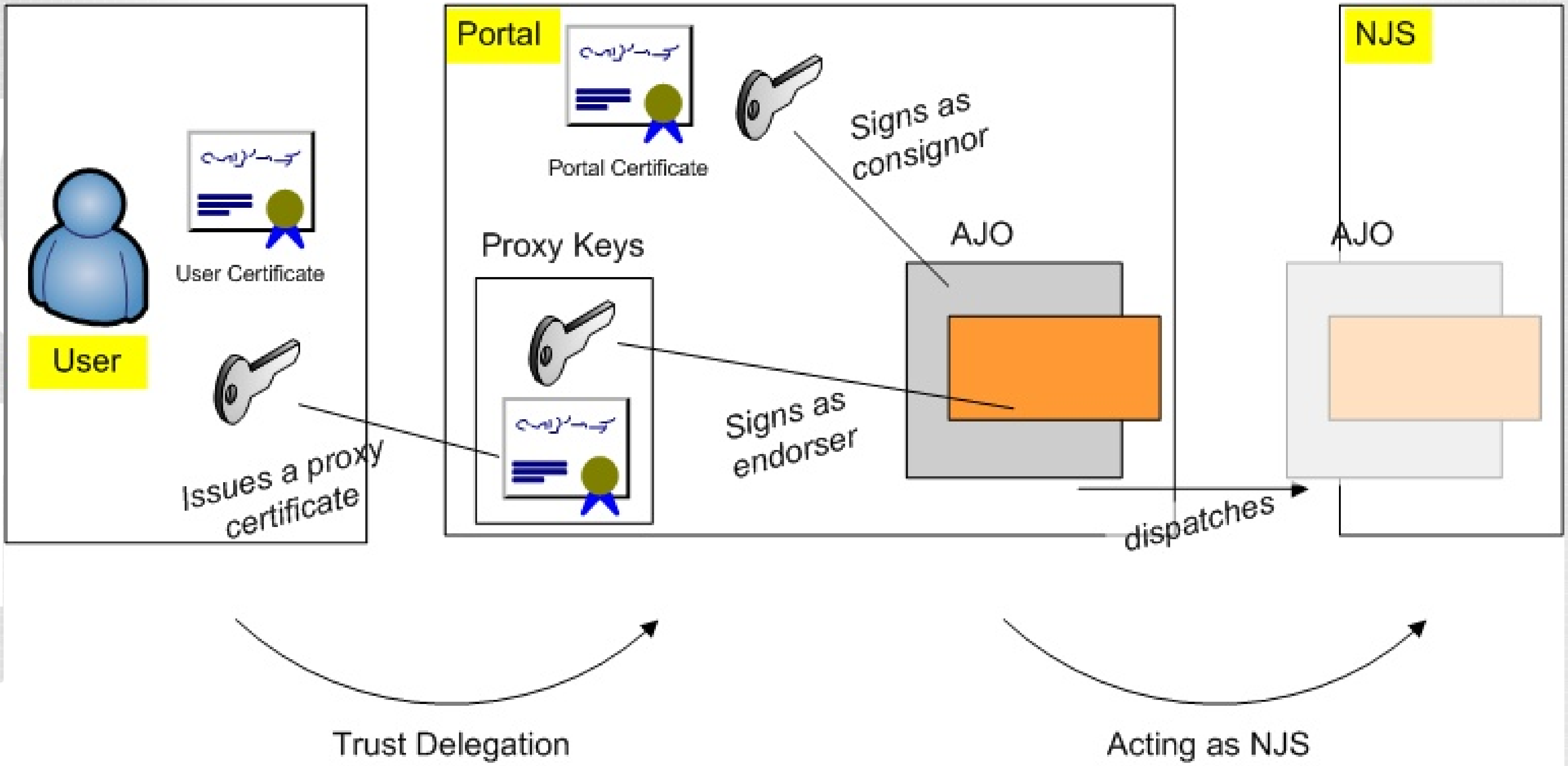
- Responding fast to client requests
- Scalable (number of client requests)
  - Implies scalable database and container infrastructure
- fault tolerant (to some extend)

Disadvantages

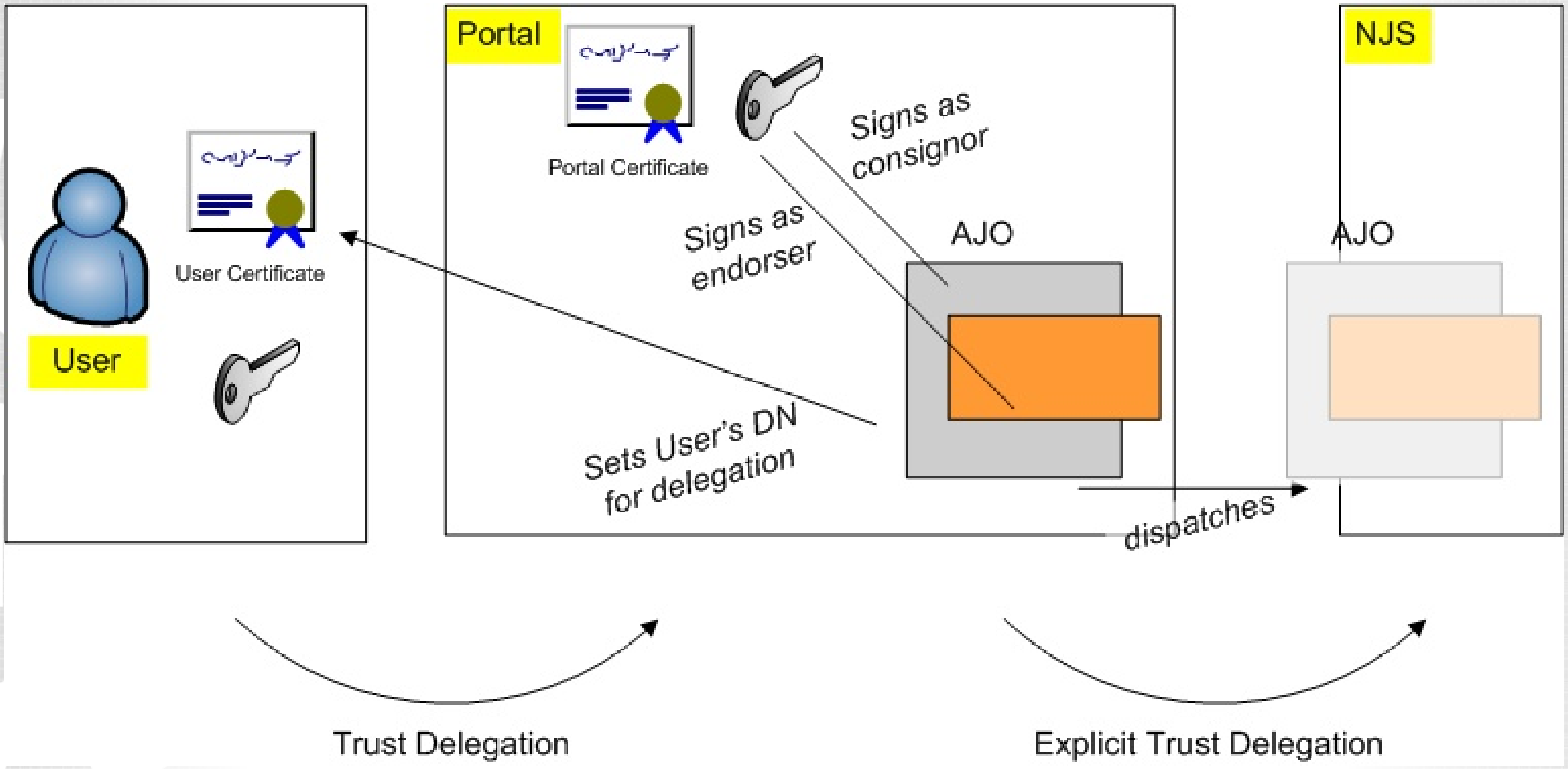
- Risk of delivering outdated data
- No support UNICORE alternate file transfer

  The web application has files transferred independently (GPFS, CIFS)

Security Solutions: Proxy Certificate Approach



Security Solutions: Explicit Trust Delegation



Possible UNICORE5 improvements

AJO

- Default constructors for all classes (unless it does not make sense)
  - Background: Persistence

ETD

- Use of X500Principals instead of whole X509 certificates as user attribute

- Allowing more “direct“ requests for an ETD agent
  - E.g. for Resource information

Conclusion

- JMEA is an EA which proves to work with UNICORE5 in DEISA
- It has all basic features implemented which are needed for successful job management
- It is designed to work with the JRA1/JRA3 Web application
- It can be used in a different context (OMII/GridSAM)
- But, it does not provide a standards based interface

- WS-GRAM (4.0 and 4.1) support is being developed
- UNICORE6 support is hopefully given with WS-GRAM 4.2 support

Thanks

Contributors to the JRA1/JRA3 endeavors:

Johannes Reetz, RZG, Garching, Germany
Daniel Frank, RZG, Garching, Germany

Discussion Partners (GridSAM)
Stephen McGough, IC, London, England

Researchers/Beta-Testers
Matthias Krack, ETH, Zürich, Switzerland
Peter Coveney, UCL, London, England