

DEISA

# DEISA and GPFS in DEISA

Andreas Schott, Technical Coordinator  
RZG, [Andreas.Schott@rzg.mpg.de](mailto:Andreas.Schott@rzg.mpg.de)

[www.deisa.eu](http://www.deisa.eu)

**SP-XXL, Makena**  
January 13<sup>th</sup>, 2009



RI-222919



# Overview

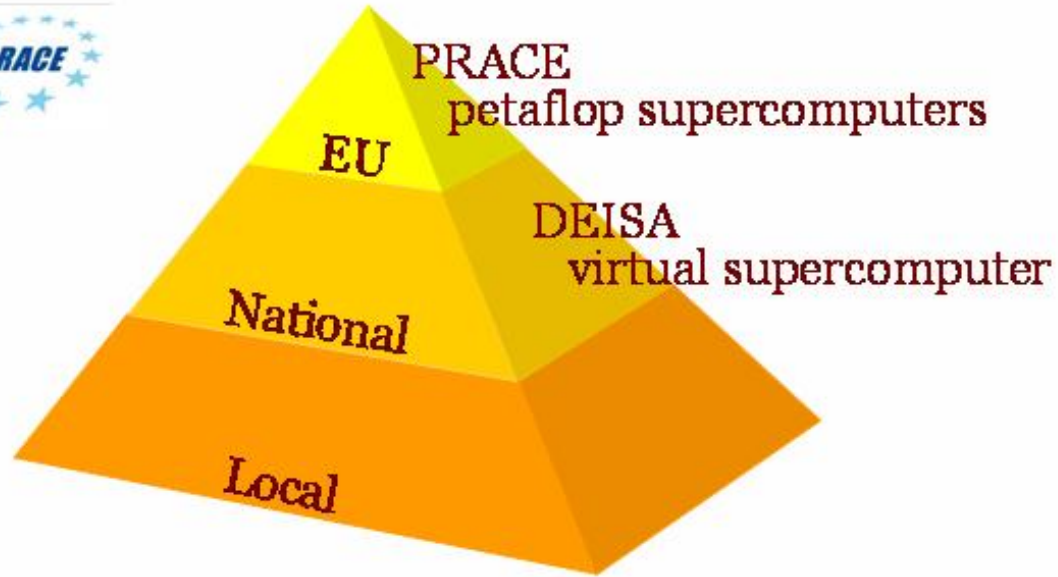
## Distributed European Infrastructure for Supercomputing Applications

- Resources on National Supercomputers
  - 11 Sites in 7 European Countries
- Connected mostly via 10Gbit/s by GÉANT
  - Central Switch in Frankfurt/Germany
- Providing Common Access
  - Certificates and Common Name Range
- Common Software and Middleware
  - Binaries, Libraries, Batch, Unicore, (GT4)
- User Environment and Application Support

# History and Future

- Planning from 2002
  - European Need for Supercomputing (ESFRI)
- Starting 2004 with 8 partners
  - CINECA (IT), CSC (FI), ECMWF (UK), EPCC (UK), FZJ (DE), IDRIS (FR), RZG (DE), SARA (NL)
- Extending 2005 to 11 partners
  - BSC (ES), HLRS (DE), LRZ (DE)
- Continuing 2008 with 11 partners
  - Project DEISA2
- Adding Associate Partners in 2008 and later
  - CEA (FR), CSCS (CH), PDC/KTH (SE), JRSS (RU)
- PRACE provides the legal entity for the persistent European Supercomputing based on DEISA technology

# new "petaflop" supercomputers



15



Mario Campolargo, OGF23, June 2008



# Essentials of DEISA2 Project

- Consolidation of the existing DEISA infrastructure
- Evolvement of this European infrastructure towards a robust and persistent European HPC ecosystem
- Enhancing the existing services, by deploying new services including support for European Virtual Communities, and by cooperating and collaborating with new European initiatives, especially PRACE
- DEISA2 as the vector for the integration of Tier-0 and Tier-1 systems in Europe
- To provide a lean and reliable turnkey operational solution for a persistent European HPC ecosystem
- Bridging worldwide HPC projects: To facilitate the support of international science communities with computational needs traversing existing political boundaries

## DEISA Key Strategies

- Deployment of global file systems across Europe
- Unified and seamless access to supercomputing resources throughout Europe
- Europe-wide support for enabling of grand challenge applications
- DEISA Extreme Computing Initiative DECI

# DEISA Extreme Computing Initiative

## DECI call 2005

51 proposals, 12 European countries involved, co-investigator from US)  
30 mio cpu-h requested  
29 proposals accepted, 12 mio cpu-h awarded (normalized to IBM P4+)

## DECI call 2006

41 proposals, 12 European countries involved  
co-investigators from N + S America, Asia (US, CA, AR, ISRAEL)  
28 mio cpu-h requested  
23 proposals accepted, 12 mio cpu-h awarded (normalized to IBM P4+)

## DECI call 2007

63 proposals, 14 European countries involved, co-investigators from  
N + S America, Asia, Australia (US, CA, BR, AR, ISRAEL, AUS)  
70 mio cpu-h requested  
45 proposals accepted, ~30 mio cpu-h awarded (normalized to IBM P4+)

## DECI call 2008

66 proposals, 15 European countries involved, co-investigators from  
N + S America, Asia, Australia  
134 mio cpu-h requested (normalized to IBM P4+)  
*Evaluation in progress*

# DEISA Extreme Computing Initiative

**Involvements in projects from DECI calls 2005, 2006, 2007:**

**157** research institutes and universities

*from*

**15** European countries

Austria  
Italy  
Russia

Finland  
Netherlands  
Spain

France  
Poland  
Sweden

Germany  
Portugal  
Switzerland

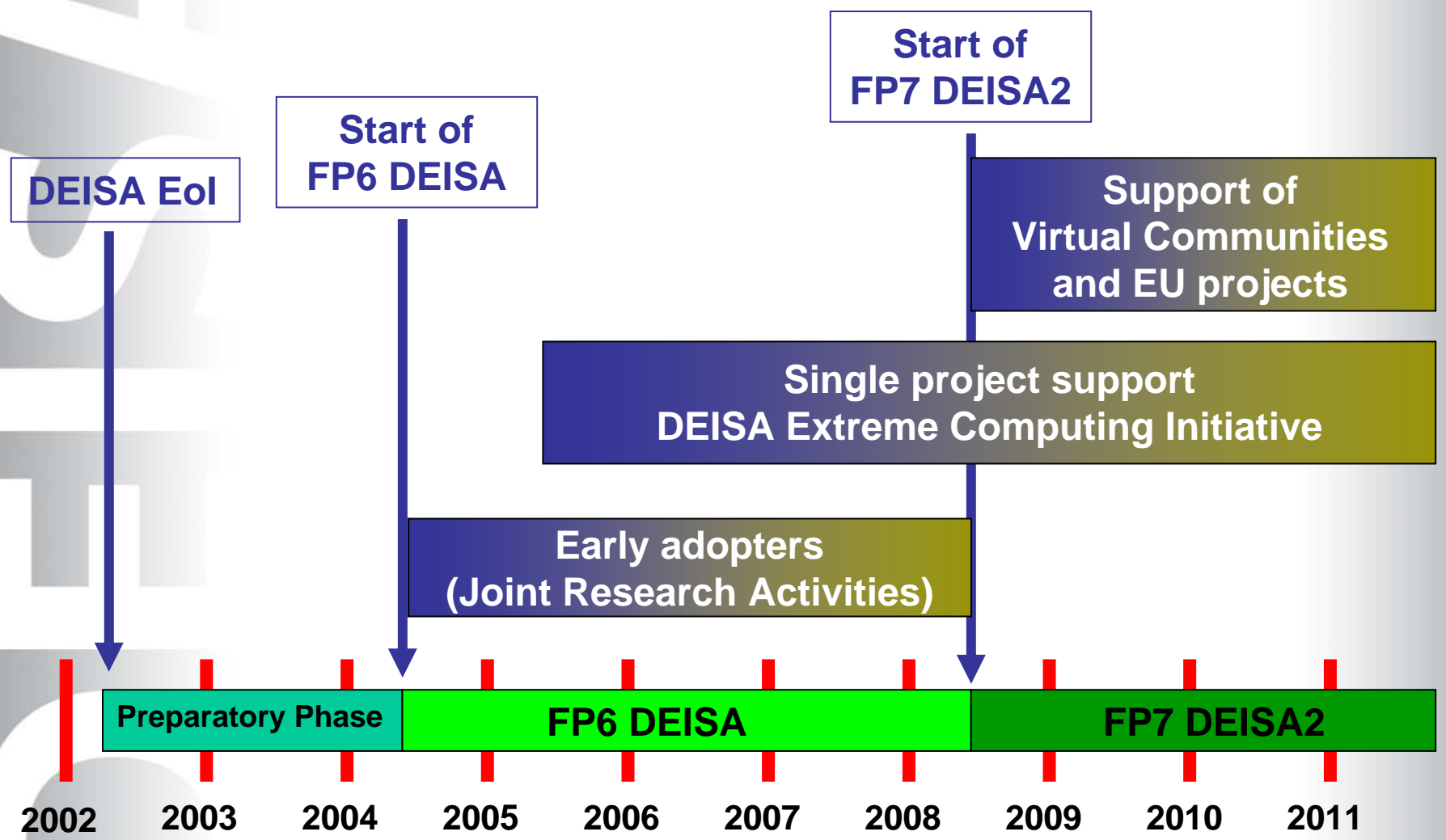
Hungary  
Romania  
UK

*with collaborators from*

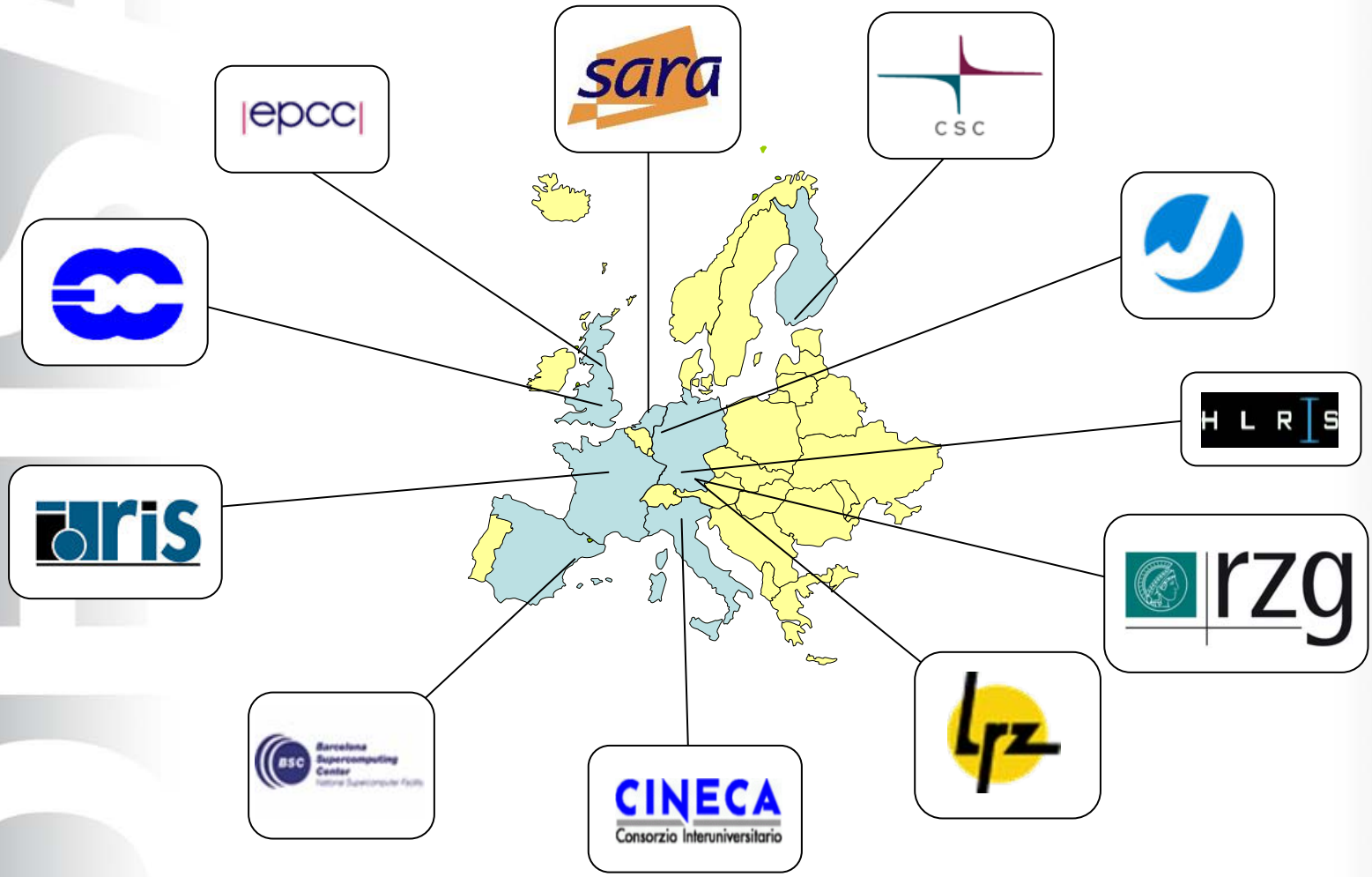
**four** other continents

North America, South America, Asia, Australia

# Evolution of User Categories in DEISA



# DEISA Partners



# Current Hardware in DEISA

- Cray XT-4/5 with Linux (CSC, EPCC)
- IBM-PowerPC with Linux (BSC)
- IBM-Power4 with AIX (EPCC)
- IBM-Power5 with AIX (CINECA)
- IBM-Power6 with AIX (ECMWF, FZJ, IDRIS, RZG)
- IBM-Power6 with Linux (SARA)
- IBM-BlueGene/P with Linux (FZJ, IDRIS, RZG)
- NEC-SX8 with Super-UX (HLRS)
- SGI-Altix with Linux (LRZ)

# Core Infrastructure and Services

## Dedicated High Speed Network

### Common AAA

- Single sign on
- Accounting/budgeting

### Global Data Management

- High performance remote I/O and data sharing with global file systems
- High performance transfers of large data sets

### User Operational Infrastructure

- Distributed Common Production Environment (DCPE)
- Job management service
- Common user support and help desk

### System Operational Infrastructure

- Common monitoring and information systems
- Common system operation

### Global Application Support

# Constraints

## Highest Priority: local stability

- 95% of the jobs are non-DEISA

## Ease of use for the users

- global transparent file access
  - `/deisa/{home,data,scratch}/<group>/<user>`
  - `$DEISA_HOME`, `$DEISA_DATA`, `$DEISA_SCRATCH`
- standardized software stack and use (DCPE)
  - libraries
  - binaries (shells, interpreters, compilers)
- unified access batch access (UNICORE)
- interactive login (new in DEISA2)

## Costs

- 850k€/year for the 10Gbit Network
- GPFS License and Maintenance

# GPFS in DEISA

Starting with GPFS v2.3 and v2.4

- not a real MC-GPFS
- no root squash
- pair-wise communication
- no flow control
- problems with latencies
- problems with lost packets
- avoided uid mapping

# GPFS in DEISA

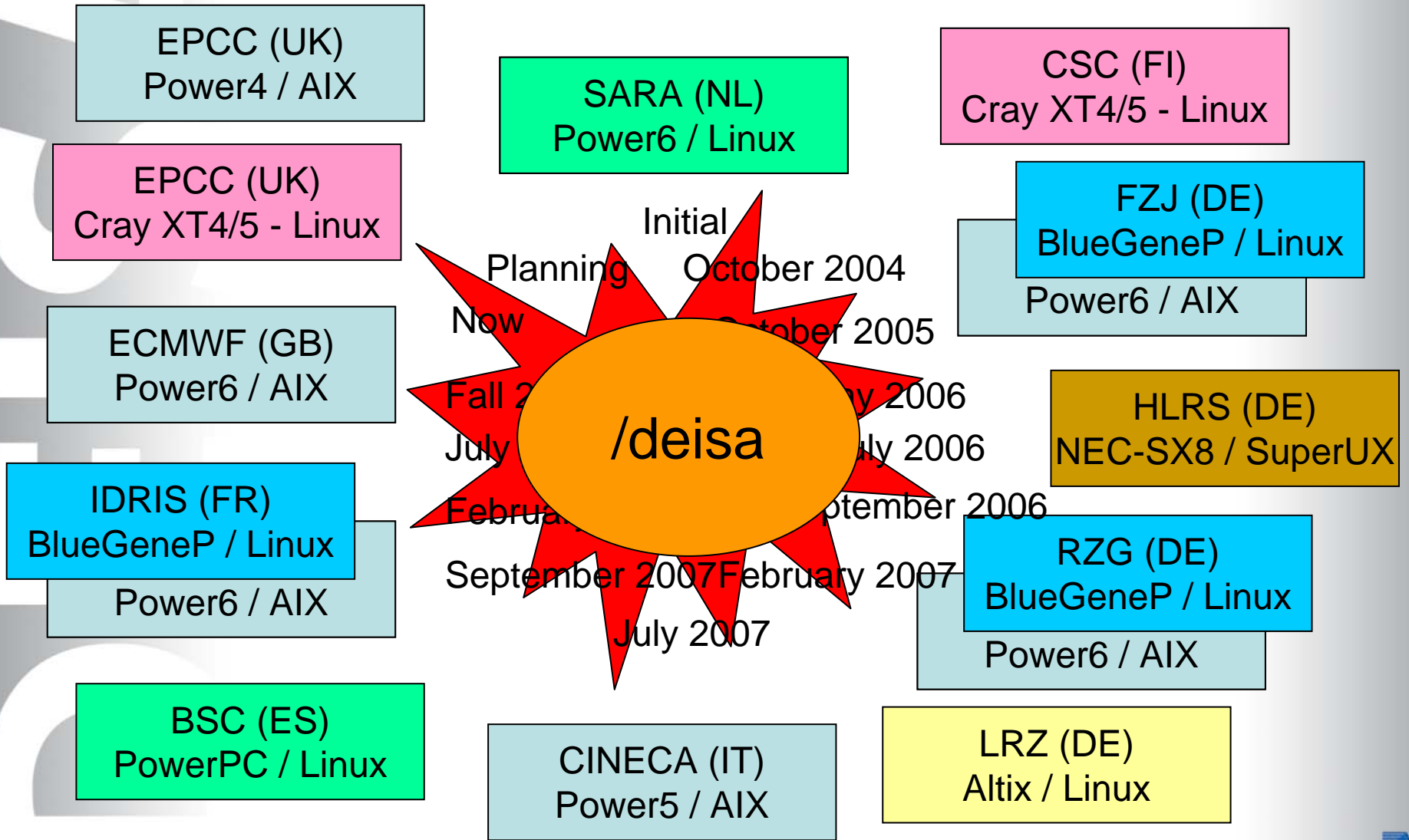
## Upgrade to GPFS v3.1

- incompatibility with v2
- lost integration of some partners
- much more stable
- new problems
- critical situation beginning of 2008
- planning for complete restructuring
- improvements by upgrade to 10Gbit/s network

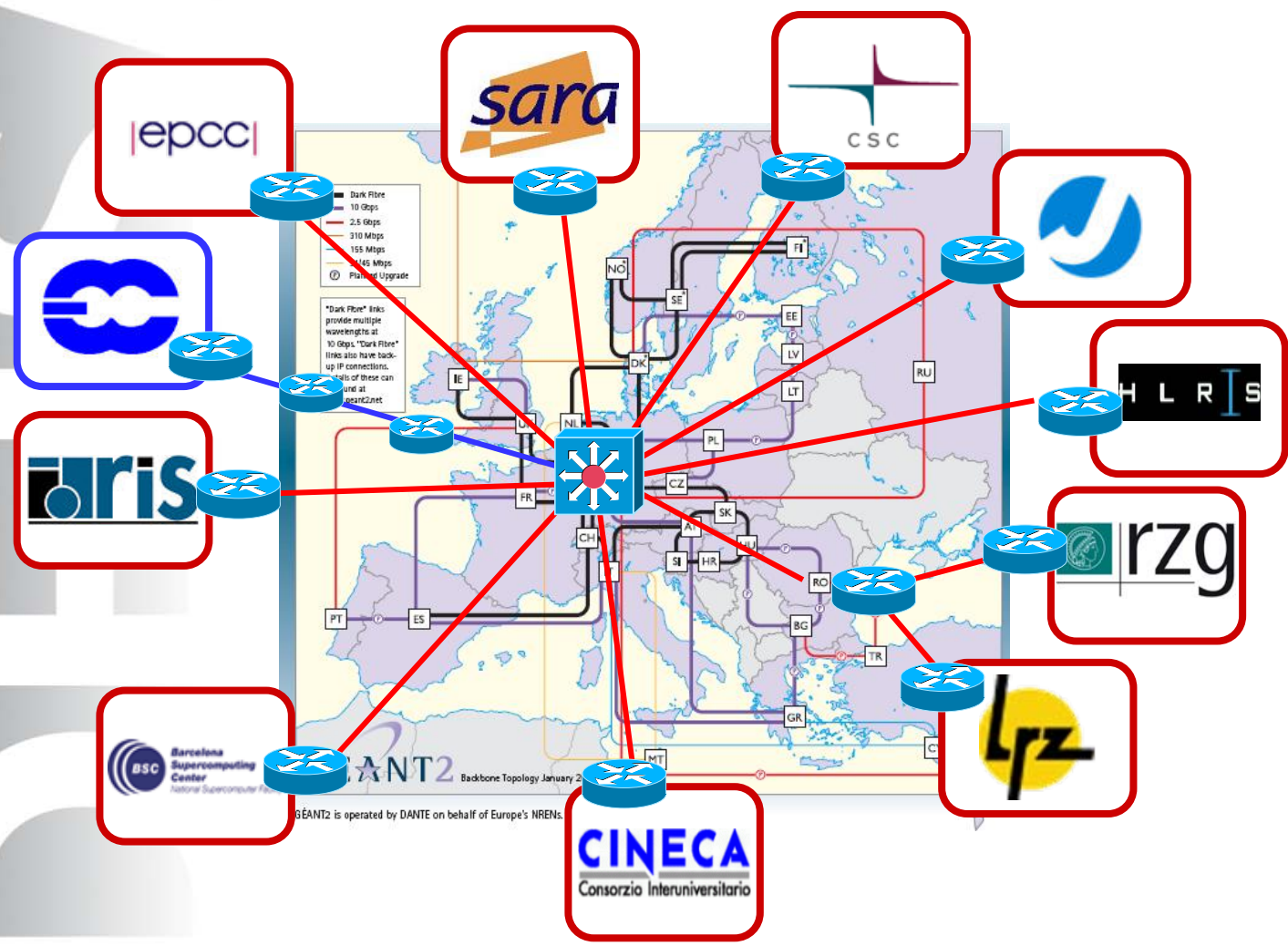
## Upgrade to GPFS v3.2

- more stability

# GPFS Integration Timeline



# DEISA – Network Logical View

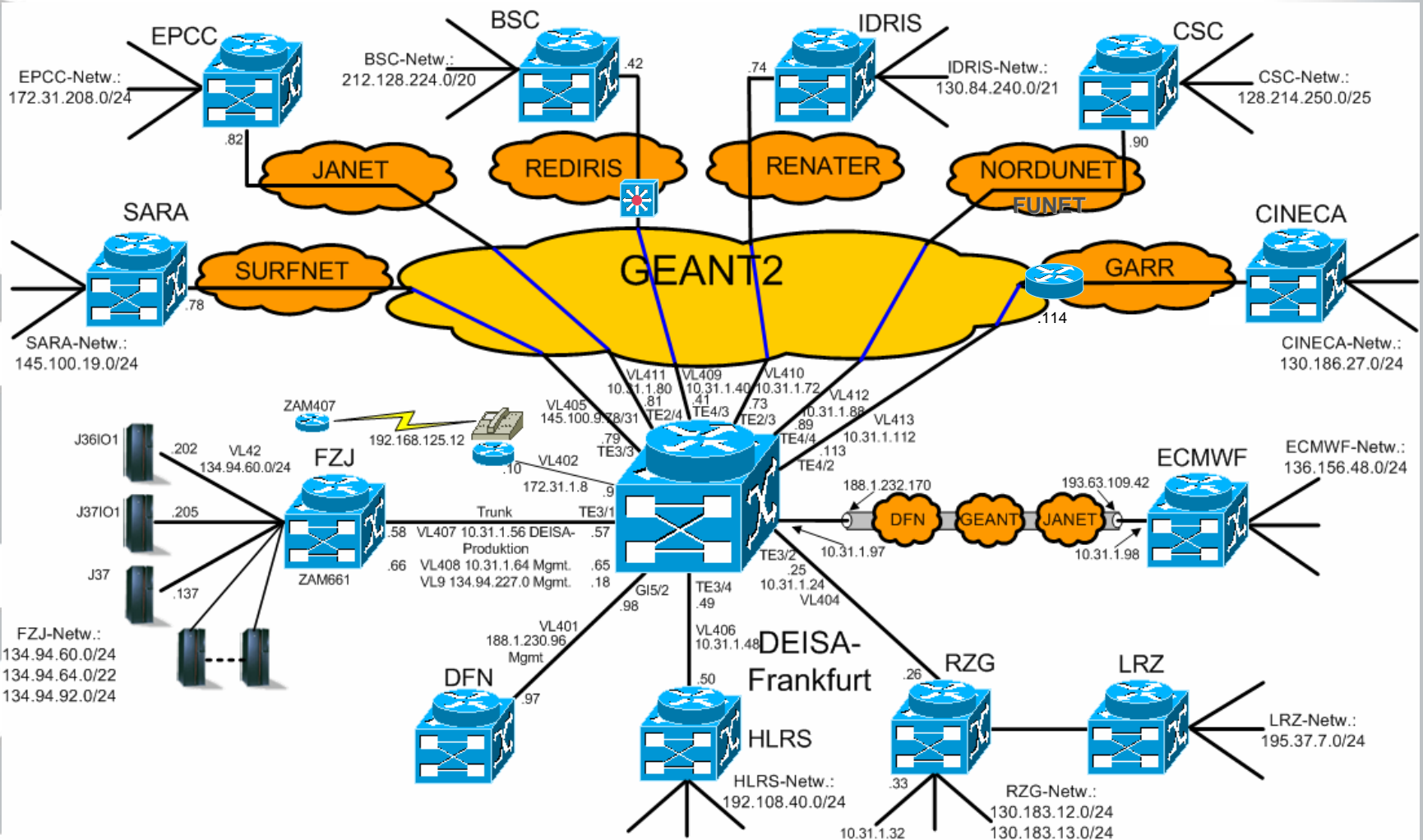


- NRENs involved:
- DFN
  - Nordunet/FUNET
  - GARR
  - RedIRIS
  - RENATER
  - SURFnet
  - UKERNA

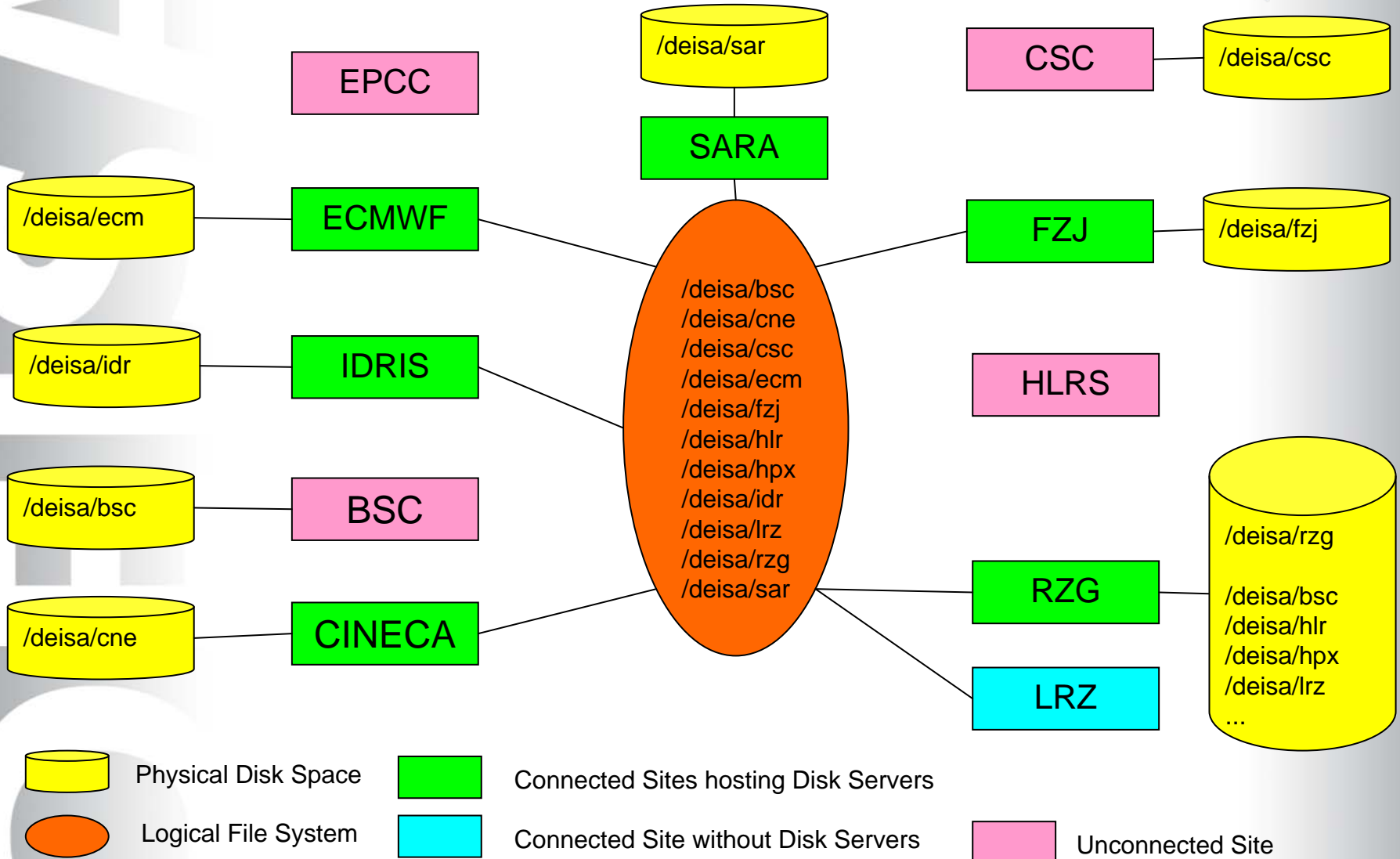
Dedicated  
10 Gb/s  
wavelength

1 Gb/s LSP  
GRE-Tunnel

# DEISA – Network Technical View



# Overall GPFS Configuration



# Planning / Considered Options

- Treatment of BG and alike systems
  - gateway nodes
- Centralized GPFS Servers
  - with mirroring
  - more unique HW and performance
  - better support
- Intermediate Clients
  - as local exporters (NFS)
- pNFS
  - pNFS access to GPFS
  - Panache (WAN-caching GPFS)
  - pNFS on other local parallel filesystems (Lustre)
- other options ?

# GPFS-Configurations: LRZ

SGI: Altix 4700 (Montecito)

OS: SLES10 SP1

GPFS: 3.1.0.20

Local FS: CXFS

DEISA FS: none

Internal Network:

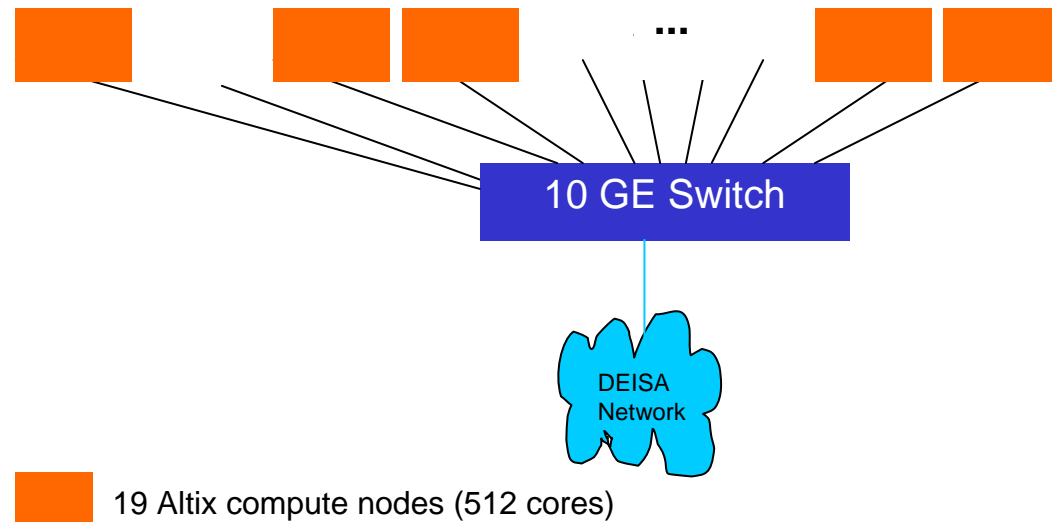
10Gbit/s

External Network:

10Gbit/s

DEISA Uplink:

10Gbit/s



# GPFS-Configurations: CSC

Cray: XT4 (Opteron)

OS: Catamount

GPFS: 3.1.0.16

Local FS: Lustre

DEISA FS: MC-GPFS

PPC, AIX 5.3, GPFS 3.1.0.14

Internal Network:

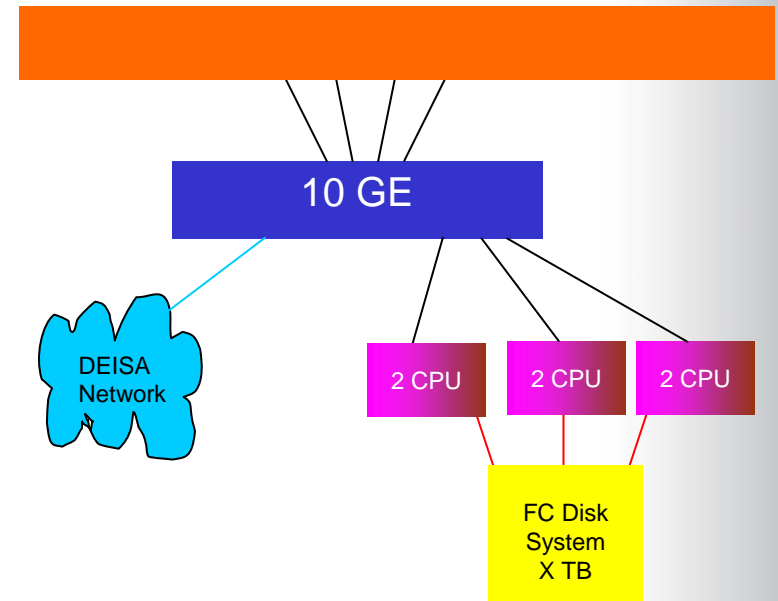
10Gbit/s

External Network:

10Gbit/s (n for XT4)

DEISA Uplink:

10Gbit/s



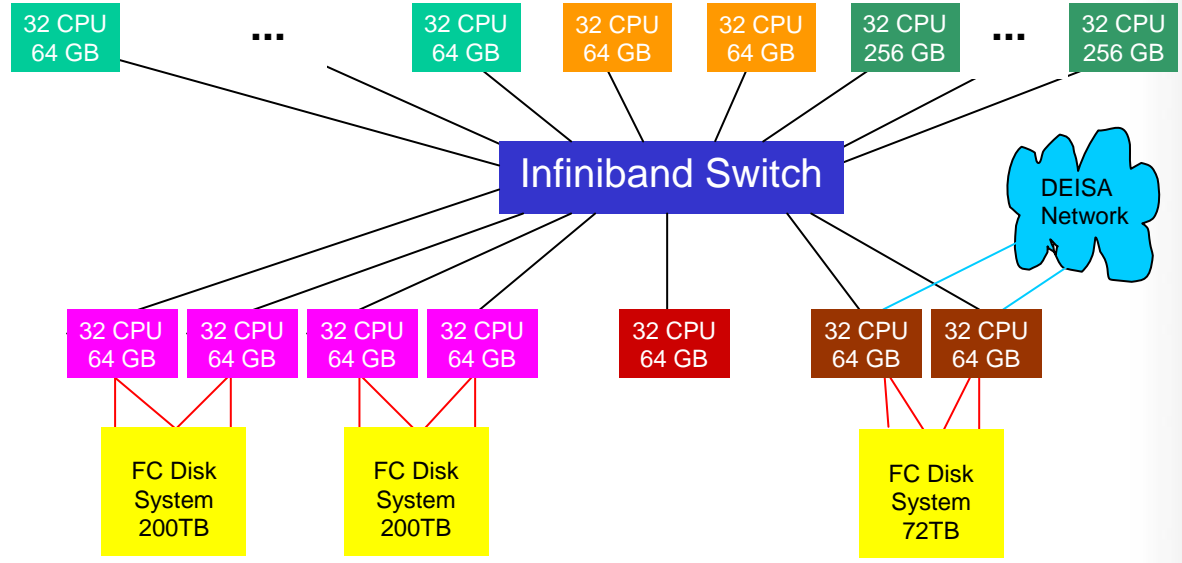
 Cray XT4

 3 nodes with 1 MC-GPFS for DEISA

# GPFS-Configuration: RZG

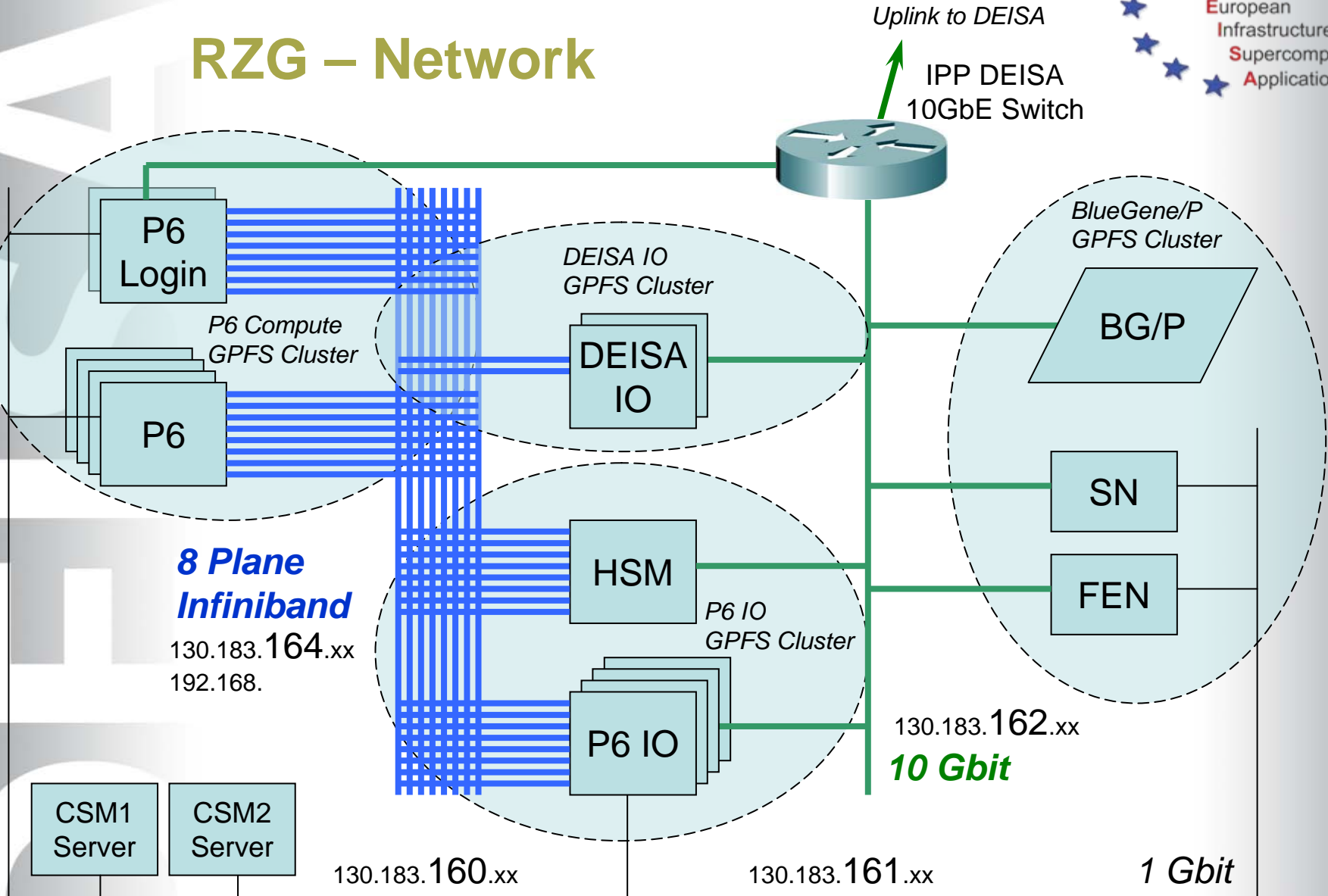
IBM: Power6  
 OS: AIX 5.3  
 5300-08  
 GPFS: 3.2.1.6  
 Local FS: NSD  
 DEISA FS: NSD

Internal Network:  
 Infiniband  
 External Network:  
 10Gbit/s  
 DEISA Uplink  
 10Gbit/s



- 70 + 130 compute nodes
- 2 login node
- HSM 1 backup node
- I/O 4 nodes in two racks with 3 local GPFS file systems
- I/O 2 nodes in two racks with 1 MC-GPFS for DEISA

# RZG – Network



# DEISA - Essentials

- DEISA has established a systematic network of cooperation of major supercomputing centres in Europe and operates a powerful supercomputing grid on top of national services.
- DEISA provides a collaborative environment for capability computing and data management, and also support for application enabling for complex and extreme computing.
- DEISA paves the way towards usage of the most adequate supercomputer architectures and towards usage of the most powerful supercomputers, for the benefit of European researchers and the advancement of science.
- Many of the services deployed will enable the efficient integration of future European petascale systems into a seamless European HPC ecosystem.

# Questions? DEISA



Distributed European Infrastructure for Supercomputing Applications