

Large scale HPC workflow management using PBS professional in a National Academic Computing Center



Outline

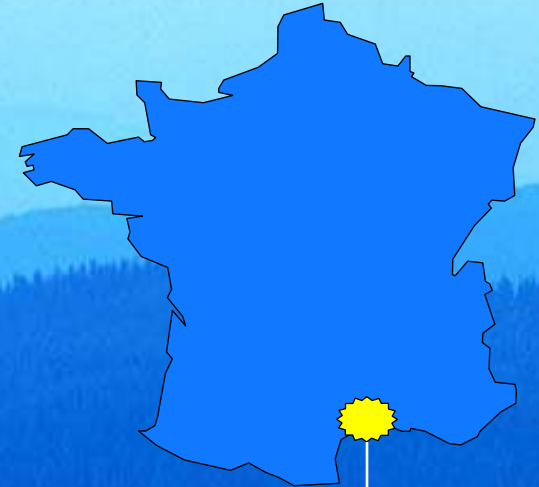
- **Presentation of CINES**
 - Missions / activities
 - Organisation of French HPC
 - CINES's HPC resources and services

- **Workflow Management**
 - Existing environment
 - PBSpro implementation
 - Statistics
 - Next

Outline

- **Presentation of CINES**
 - **Missions / activities**
 - **Organisation of French HPC**
 - **CINES's HPC resources and services**

CINES is supervised and funded by the Ministry of Higher Education and Research.



- **CINES is located in Montpellier (South of France)**
- **30 years serving the National Academic Research**
- **CINES provides the french public research community with computing resources and services.**
- **50 persons :**
(technicians, engineers and administratives)



2 MISSIONS

- ❖ **High Performance Computing**
- ❖ **Digital preservation**

Digital preservation

Since 2004 the CINES was given the mandate to provide long-term preservation capabilities for digital objects related to scientific and technical information

National agreement process:

CINES is in process of agreement by the “Archive Nationale”

International certification process:

CINES is one of the three pilot sites (UKDA, DNb) to test **European Certification Framework** for long term preservation supported by European Commission ➡ (iso certification)

Digital preservation

- Electronic PhD thesis

French Ministry of higher education referred CINES as the national center for long-term conservation of digital thesis



Digital preservation

- Electronic PhD thesis
- Digitized publications
- **TGE-ADONIS** : Multimedia documents for the Research Center on Oral Resources (CRDO)
- **Liber floridus** : medieval manuscripts of universities libraries Mazarine, St Geneviève, IRH, ...
- ...



Digital preservation

- Electronic PhD thesis
- Digitized publications
- Pedagogic multimedia

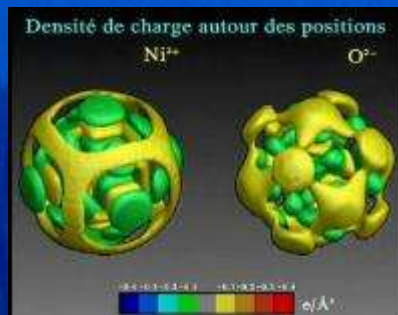
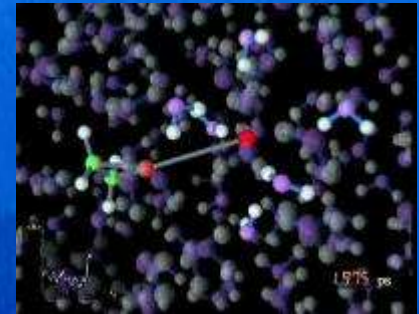


- **CANAL_U** : videos of University channels (for CERIMES)
-



Digital preservation

- Electronic PhD thesis
- Digitized publications
- Pedagogic multimedia
- Scientific datasets




A new domain directly linked to our mission in HPC


High Performance Computing


National HPC Center since 1980


- **Scalar, vector and parallel processing**
- **Accelerators : GPU, CELL, FPGA**


French HPC coordination





 CNRS
dépasser les frontières

 Conférence
des présidents
d'université

 INRIA

 cea


 Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
 MINISTÈRE
 DE L'ENSEIGNEMENT SUPÉRIEUR
 ET DE LA RECHERCHE

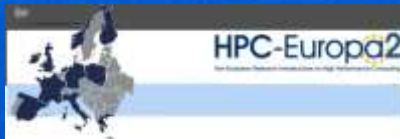
- French legal entity : « Société civile » created in 2007
- 5 shareholders

Main missions of



www.genci.fr

- Coordinate national academic supercomputing centers (civil activities)
- Promote the European HPC and participate to its organization



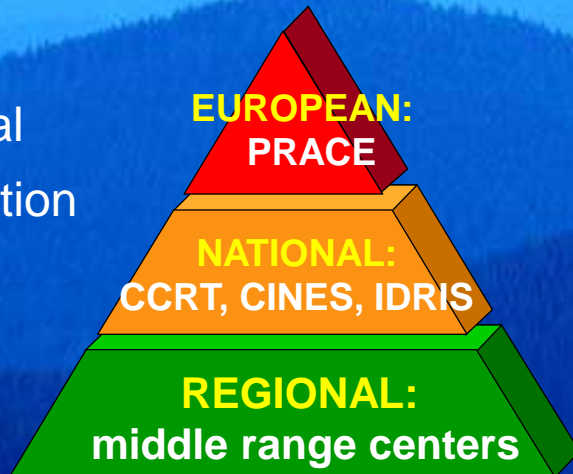
, European Exascale Software Initiative,...

- Promote simulation and HPC in fundamental and industrial research
- Give access to its equipments to promote HPC :
HPC-PME initiative : GENCI - OSEO - INRIA

French academic Equipments for researchers



pyramidal
organization



National computing centers

- very large equipments for solving extremely complex problems
- free of charge resources allocated to scientific projects after evaluation by thematic national committees

www.edari.fr

- From 2007 to 2010 : 20 to 700 Tflops**

Main National Equipments



- **CCRT: CEA**

BULL : Novascale xeon system : 143 Tflops

NVIDIA : TESLA GPU system : 192 Tflops

- **CINES: Universities**

SGI : Altix ICE xeon system : 267 Tflops

- **IDRIS: CNRS**

IBM : SP / Power6 system : 68 Tflops

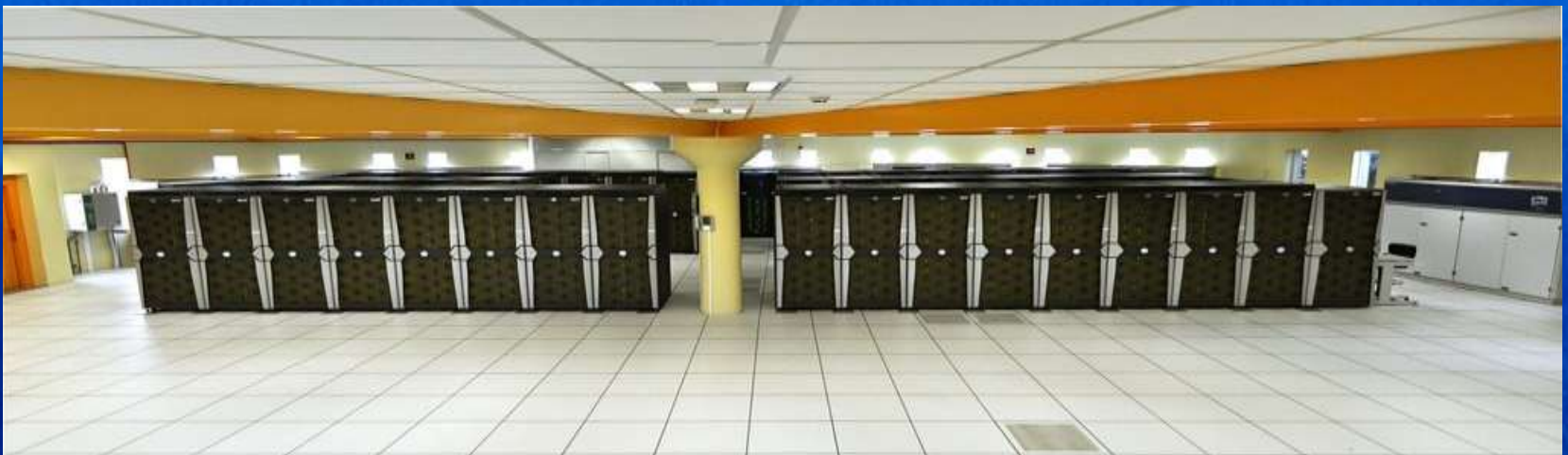
IBM : Blue Gene/P system : 139 Tflops

CINES national HPC resources

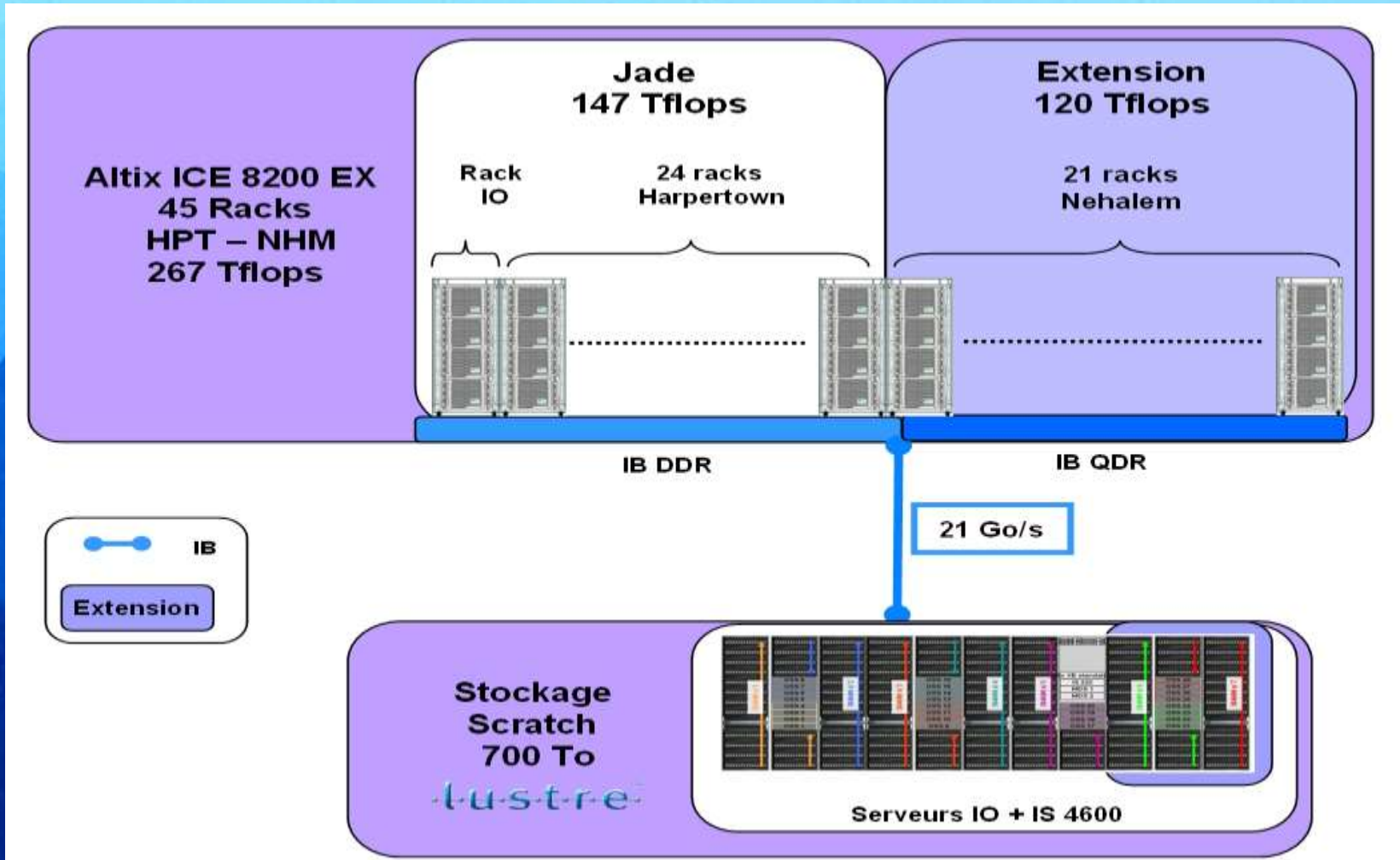
JADE : SGI Altix ICE 8200 EX

Linpack : 237 Tflops (n° 14 on Top500 , June 2010)

- Linux Suse
- SGI Tempo : Cluster administration
- **ALTAIR PBSpro 10.1.4** (patched) : Job scheduler
- Lustre : scratch files
- NFS / DMF : /home and archives

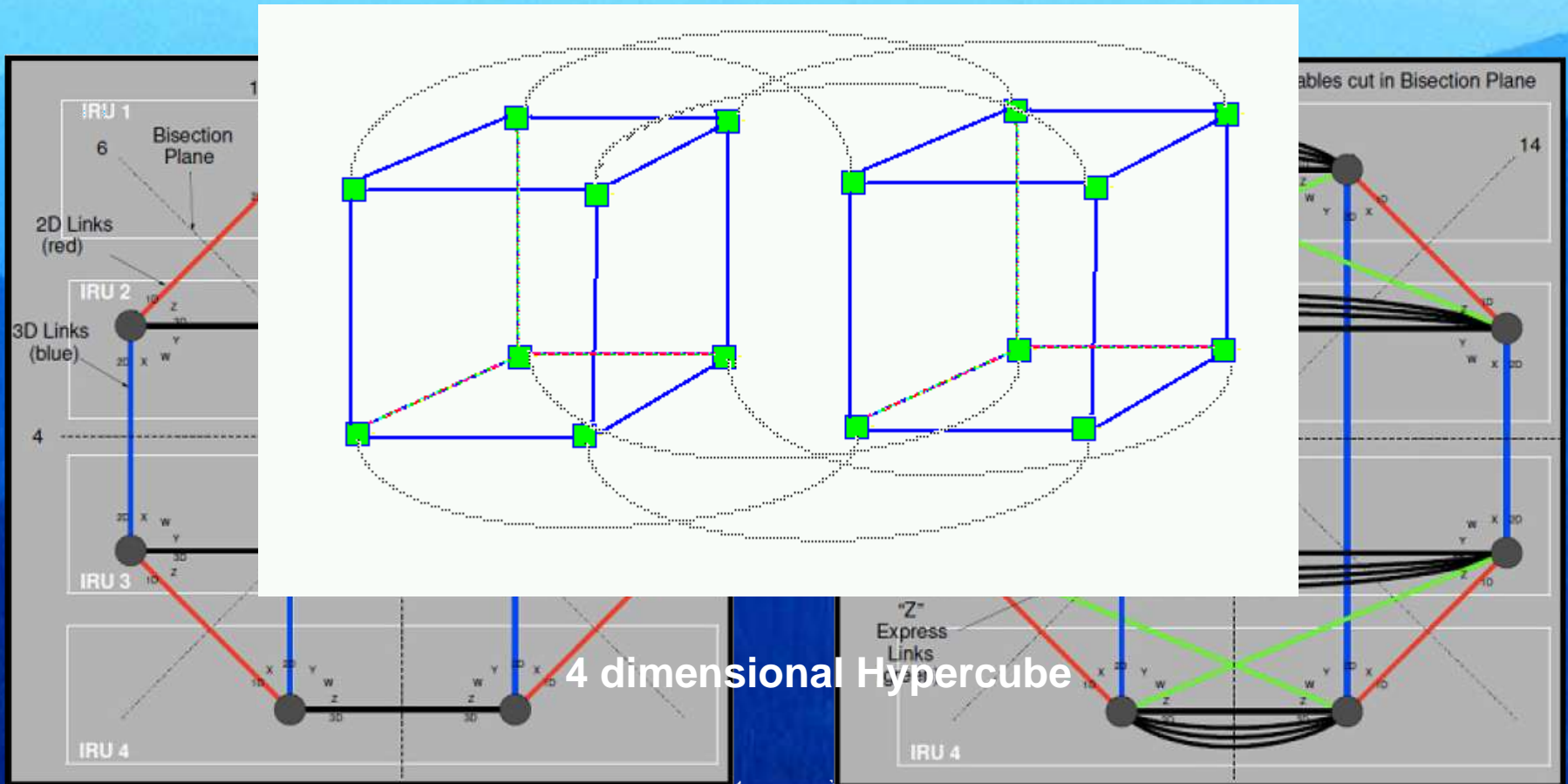


JADE



JADE Infiniband hypercube topology

Increase the bisectional bandwidth within the same rack



CINES HPC systems Computing resources

Pre/post processing resources

IBM P1600

- ❑ 5 nodes P575 16cores: Power5 /32 GB
Infiniband DDR + GPFS : **0,5 Tflops**
- ❑ 4 nodes P755 32cores: Power7 /128GB,
Infiniband DDR + GPFS : **3 Tflops**



Bull Novascale

- ❑ 20 nodes R422, 2 sockets quad core / 8Go
- ❑ 24 nodes R422-E1 + Tesla S1070 GPU,
- ❑ Infiniband DDR +Lustre



BACKBONE 10 Gb/e

SGI ICE 8200 EX : 267 Tflops

- ❑ 1536 nodes bi-proc quad core (12288 cores)
Xeon 3 GHz (Harpertown), 32 GB/node
- ❑ 1344 nodes bi-proc quad core (10752 cores)
Xeon 2.8 GHz (Nehalem), 36 GB/node
- ❑ Infiniband DDR/QDR, 700 TB (Lustre)



Storage resources



File servers

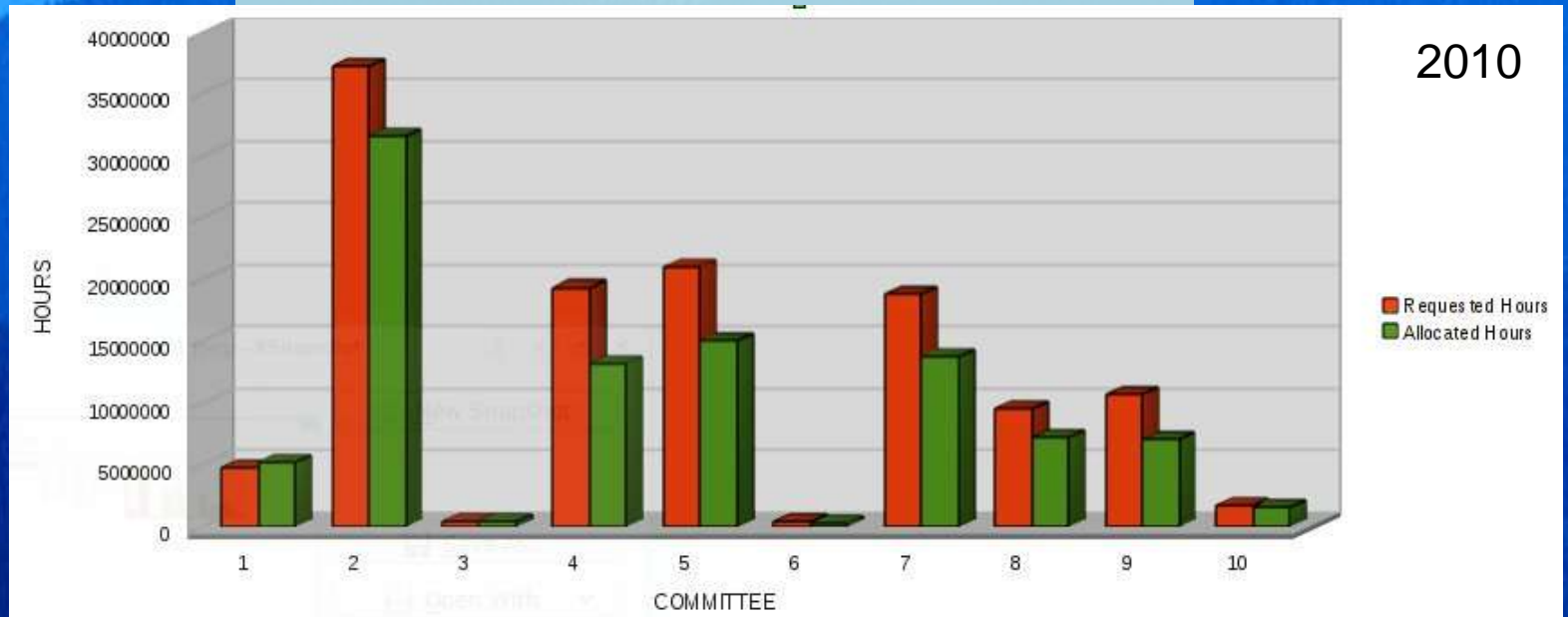
- ❑ SGI Altix 450 16 Montecito cores / 64 GB
500 To LSI disks storage (DMF)
- ❑ 2 NAS FAS250 : /home

Library : 2 x IBM TS3500

- 2000 cartridges
- 7 readers Jaguar 3
- 7 readers LTO 4

Annual National resource allocation campaign : eDARI

- 1 - ENVIRONMENT
- 2 - COMPUTATIONAL FLUID DYNAMICS
- 3 - BIOMEDICAL SIMULATION & HEALTH APPLICATIONS
- 4 - ASTROPHYSICS & GEOPHYSICS
- 5 - THEORETICAL PHYSICS & PLASMA PHYSICS
- 6 - COMPUTER SCIENCES & MATHEMATICS
- 7 - MOLECULAR SYSTEMS & BIOLOGY
- 8 - QUANTUM CHEMISTRY & MOLECULAR DYNAMICS
- 9 - PHYSICS & MATERIAL SCIENCE
- 10 - NEW APPLICATIONS & TRANSVERSE APPLICATIONS



Services for users

- **User support and HPC expertise**

CINES offers expertise in profiling, optimisation and parallelization

- **Data visualization**

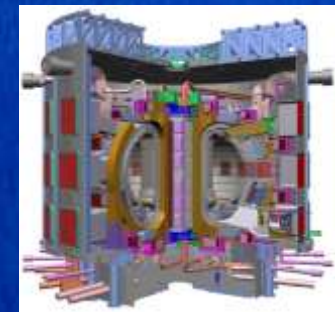
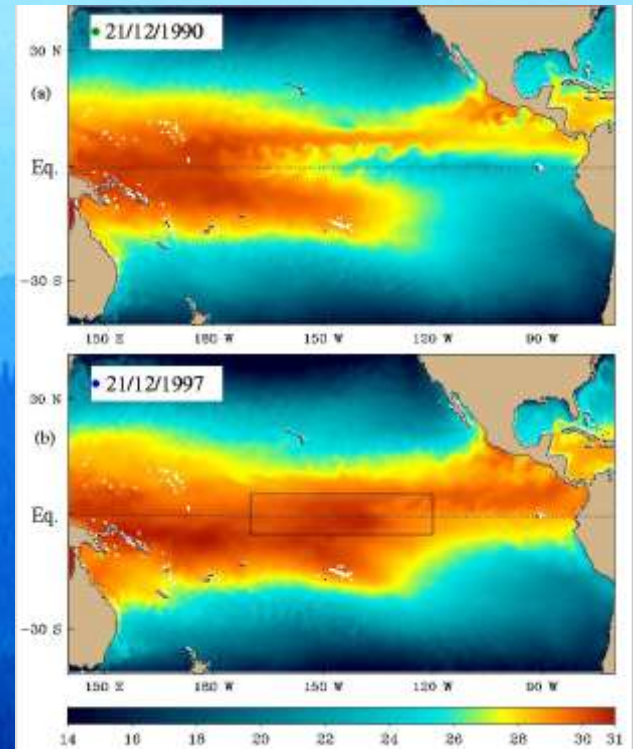
Hardware, Softwares and expertise

- **Trainings and workshops**

Programming, MPI, OpenMP, ...

- **Participation in projects**

European projects, national and international collaborations



Outline

- **Presentation of CINES**
 - Missions
 - Organisation of French HPC
 - CINES HPC resources and services

- **Workflow Management**
 - Existing environment
 - PBSpro implementation
 - Statistics
 - Next

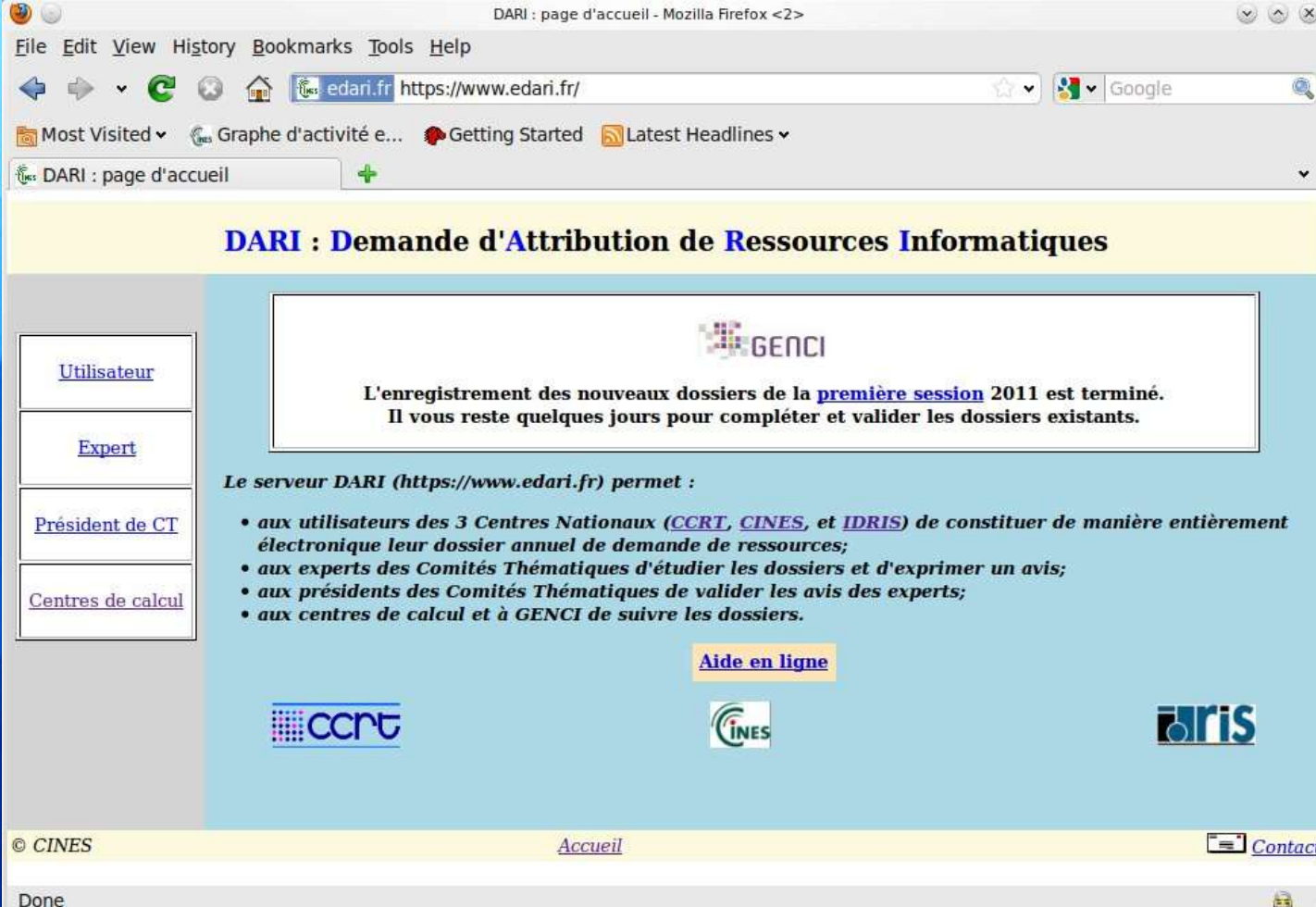
Existing multi platform tools

- Project management
- Workload monitoring
- Jobs monitoring
- Jobs statistics
- Accounting process



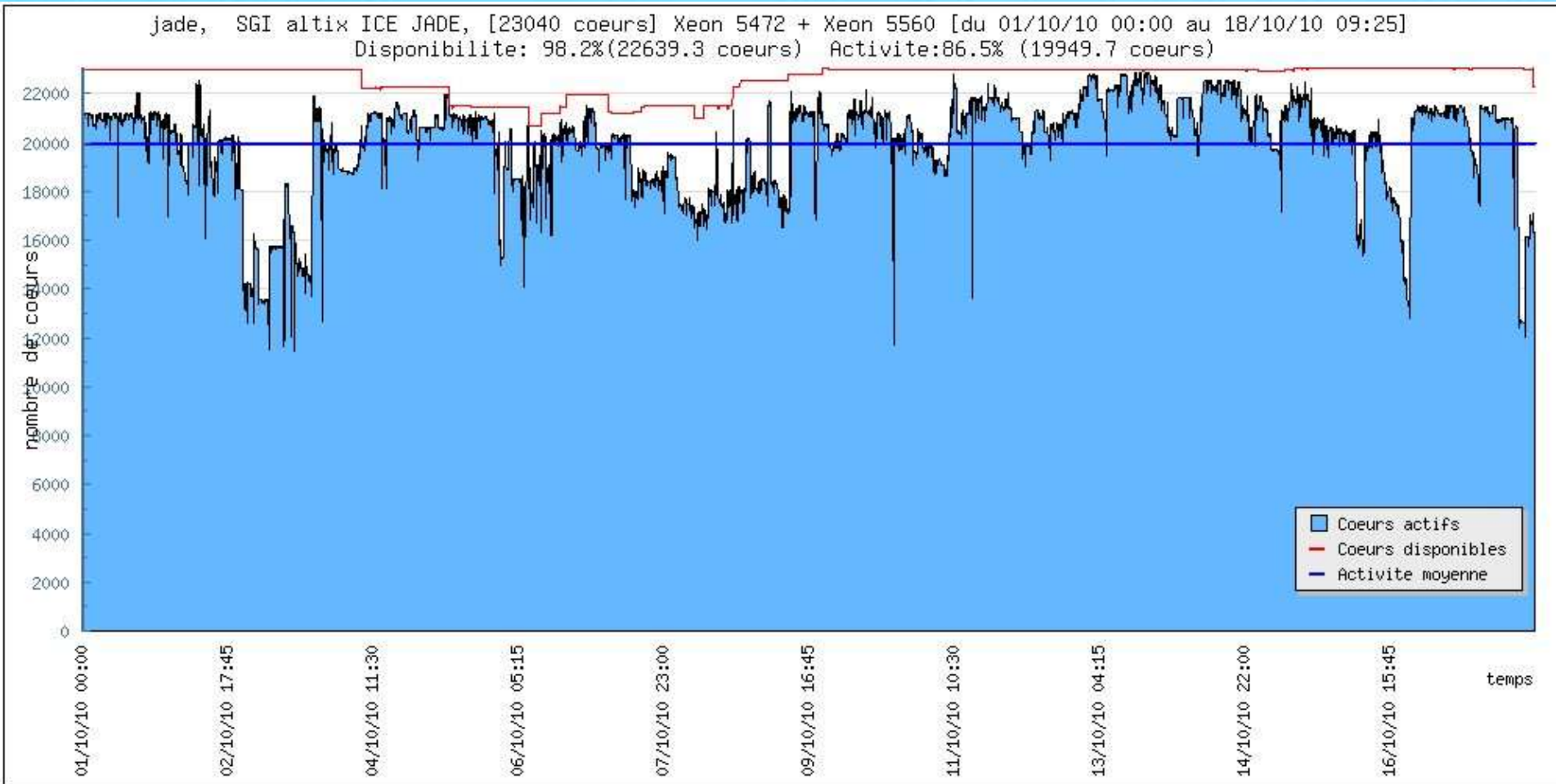
Today only PBSpro installed from « PBS GridWorks suite »

Project Management : eDARI



The screenshot shows a Mozilla Firefox browser window displaying the eDARI website. The browser's address bar shows the URL <https://www.edari.fr/>. The website's main heading is "DARI : Demande d'Attribution de Ressources Informatiques". A central message box contains the GENCI logo and the text: "L'enregistrement des nouveaux dossiers de la [première session](#) 2011 est terminé. Il vous reste quelques jours pour compléter et valider les dossiers existants." Below this, a section titled "Le serveur DARI (<https://www.edari.fr>) permet :" lists three bullet points: "aux utilisateurs des 3 Centres Nationaux ([CCRT](#), [CINES](#), et [IDRIS](#)) de constituer de manière entièrement électronique leur dossier annuel de demande de ressources;" "aux experts des Comités Thématiques d'étudier les dossiers et d'exprimer un avis;" and "aux présidents des Comités Thématiques de valider les avis des experts;" and "aux centres de calcul et à GENCI de suivre les dossiers." A yellow button labeled "Aide en ligne" is positioned below the list. At the bottom of the page, there are logos for CCRT, CINES, and IDRIS, along with a "Contact" link. The footer includes "© CINES", "Accueil", and "Contact". The browser's status bar at the bottom shows "Done".

Machines workload and availability (admin)

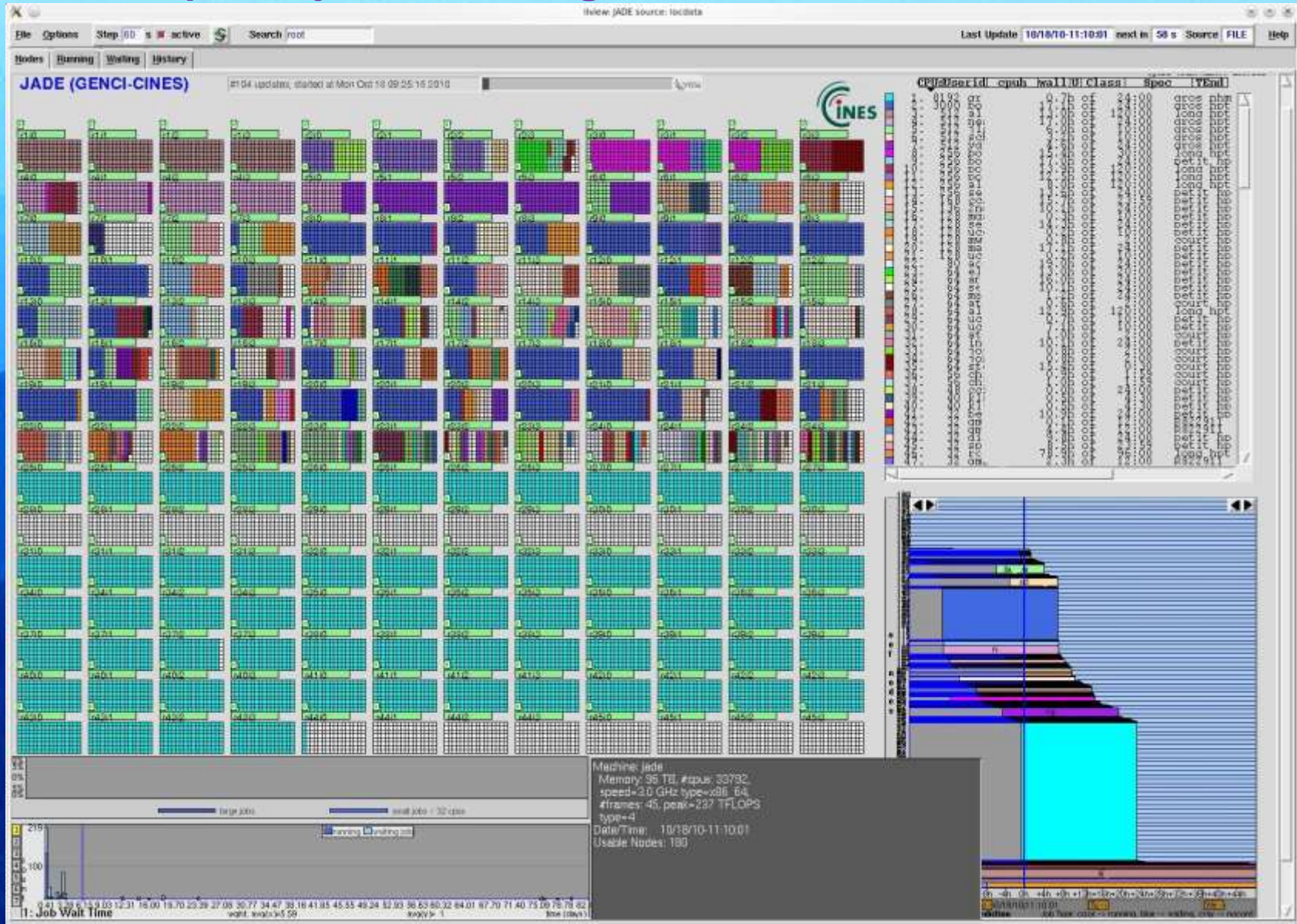


Machines workload and availability (users)

Type	Nom	Etat	Processeurs	Charge	actif		en attente				Dernière mise à jour
					job	cpu	job	cpu	job	cpu	
IBM P1600 Power5+	anakin	↑	68	32 %	3	22	0	0	0	0	le 18/10 à 11:00
SGI altix ICE JADE Xeon 5472 + Xeon 5560	jade	↑	22264	91 %	176	20472	168	259	126868	187416	le 18/10 à 11:00

Available on <http://www.cines.fr>

Graphical job monitoring: Iview (Juelich Supercomputing Center) (ADMIN)



« PBS professional » on JADE

- **Selected as « Job Scheduler » within the JADE procurement**
- **Installed since 2008 (v 9.1 → v 10.1)**
- **Several adjustments , RFE and bugs corrections to fit our needs**
 - **node selection based on topology**
 - **accounting feature**
 - **large-job management**
 - **...**
- **PBS is configured in order to facilitate resources access to users**
 - **no need to know the « queue » configuration**
 - **only «walltime» & «number of nodes/cores» are required**
 - **specify more to enforce specific selection or placement**

« PBS professional » on JADE

- 2 logical partitions to manage the machine heterogeneity
- Individual casting of applications, based on efficiency (profiling)
 - Largest and most efficient jobs routed to « Nehalem »
 - All other routed to « Harpertown »
- One PBSpro «**HOOK**» to adapt PBS to our needs
 - Automatic allocation of job resources and priority
 - Batch access control to Users
 - Machine selection
 - users authorization control
 - parameters controls
- Dynamic node pools
- Users/job priority : Backfill (Top1), Fairshare, Formula
- Large jobs are privileged

PBSpro usage: some key numbers

- Almost 1 million jobs treated since september 2008
- Up to 8000 cores per job usually managed by PBSpro
- Jobs walltime up to
 - 24 hours (standard)
 - 120 hours (only if checkpointed)
- Most of cpu hours consumed by
 - jobs over than 512 cores on « Harpertown » partition
 - jobs over than 2048 cores on « Nehalem » partition
- Waiting cores in queue = 10 times the machine size
- Machine's workload close to 90% every month

PBSpro : next

- **Upgrade to V 11**
- **Improve job priority management : FAIRSHARE + FORMULA**
- **Improve workload with « scheduled machine maintenance »**
- **Manage «Service Licence Agreement » for users**
- ...

Questions ?

