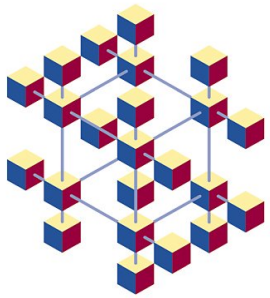
CALMIP (Calcul en Midi-Pyrénées) : pour ceux qui n'ont rien à ceux qui sont déjà bien équipés



calmip

EQUIP@MESO

Boris Dintrans (Président Comité de Programmes CALMIP)
Pierrette Babaresco* (Responsable Système)
Nicolas Renon* (Responsable Calcul Scientifique)
*DTSI - Université Paul Sabatier (nicolas.renon@univ-tlse3.fr)
http://www.calmip.cict.fr

- **CALMIP**
  - Structuration du Mésocentre de Calcul
  - Mode accès aux ressources
  - Les labos partenaires
- **Système de calcul**
  - Choix d'un système
  - Hardware / Software
- **Exploitation et Accompagnement**
  - Interface Utilisateur
  - Intégration projet de recherche
- **Perspectives**
  - EQUIP@MESO
  - ECA

Le Groupement Scientifique CALMIP : Historique

- **Fondé en 1994 par 17 Laboratoires de Recherche Publics en Région Midi-Pyrénées**
- **Soutien des 6 établissements universitaires toulousains**
  - Université Paul Sabatier (Sciences et Santé)
  - Institut National Polytechnique de Toulouse
  - Institut National des Sciences Appliquées
  - Université des Sciences Sociales
  - Université du Mirail , Lettres, Langues et Arts
- **Positionnement : Mésocentre de Calcul**
  - Promotion du calcul scientifique haute performance (contexte Multi-thématique)
  - Mise à disposition d'un environnement de Calcul Scientifique performant
    - Acquisition systèmes de calcul (contexte production)
    - Organisation de l'exploitation et du support aux utilisateurs

Le Groupement Scientifique CALMIP : Organisation

✓ Pilotage

Comité d'Orientation

6 Vice-Présidents des conseils scientifiques des établissements
Le Président du Comité de Programmes
Région
3 Représentants de la communauté des utilisateurs (Pôles de Recherche)
Représentants Pôles de compétitivité

✓Attribution des ressources :
✓ 2 AO par an
✓Animation scientifique

Comité de Programme

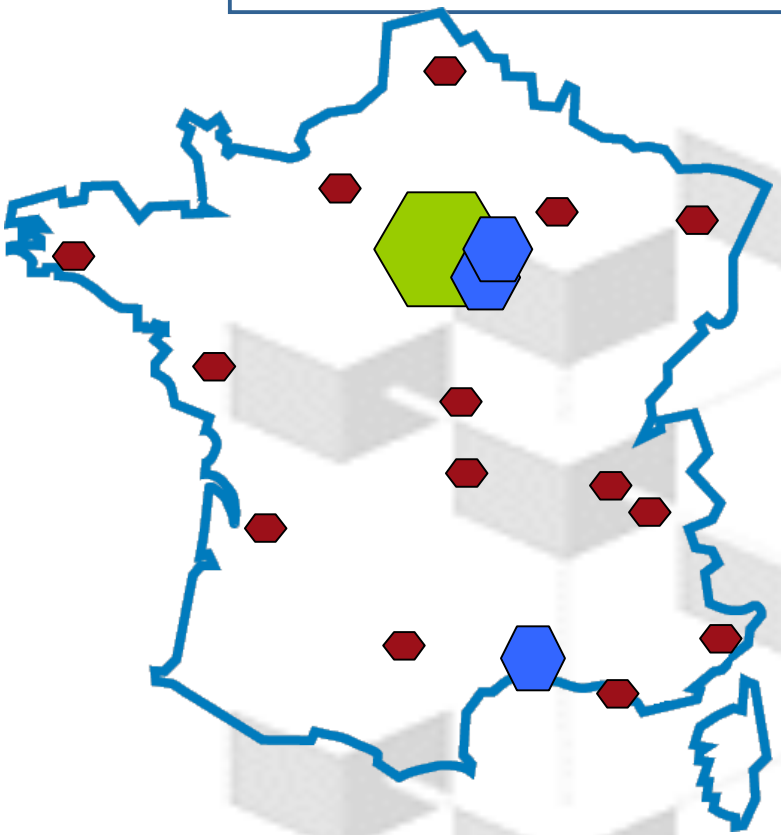
10 experts scientifiques issus des laboratoires
7 thématiques scientifiques

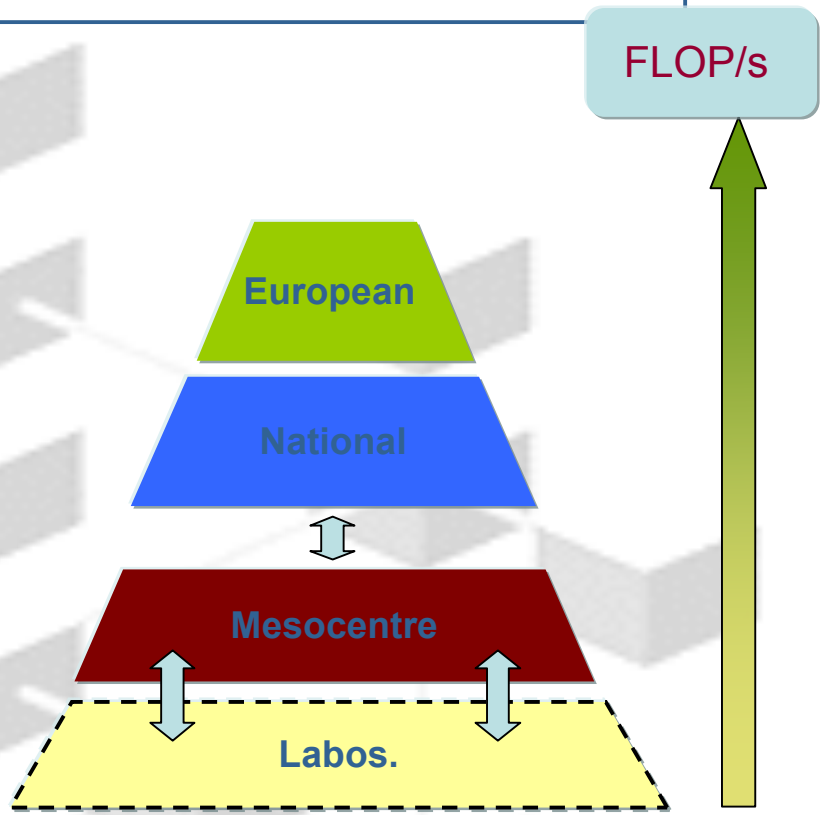
✓Support aux utilisateurs
✓Support projets de Recherche
✓Exploitation du supercalculateur

Université Paul Sabatier
D.T.S.I.

1 Ingénieur calcul scientifique
1,5 Ingénieurs système

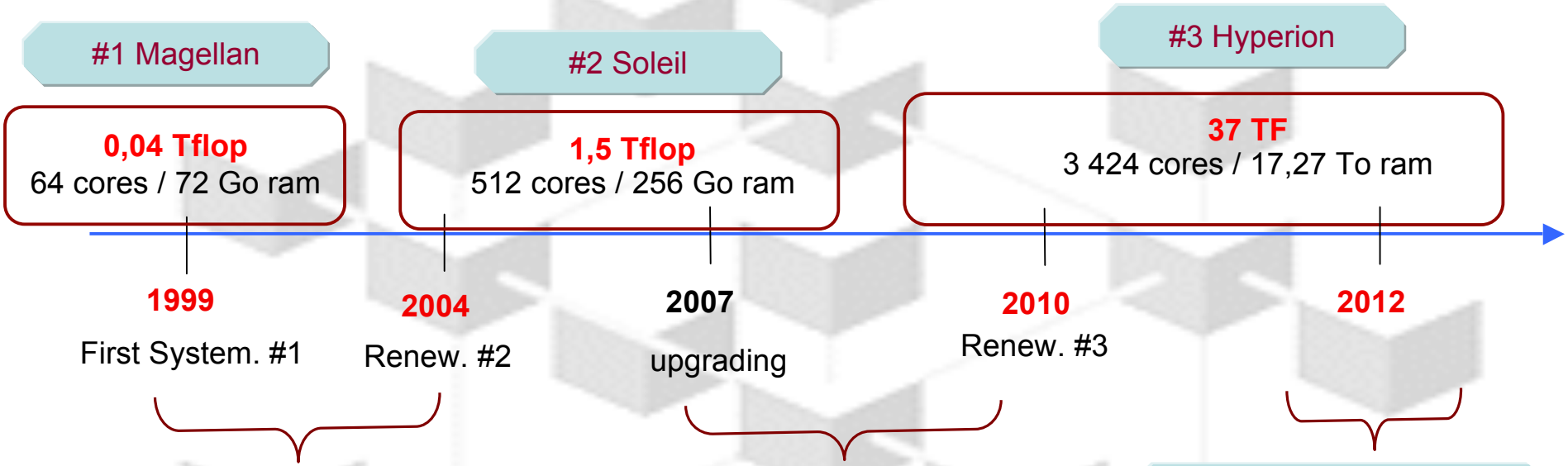
Positionnement CALMIP : Mésocentre





- ✓ **Mésocentre CALMIP :**
  - ✓ **Proximité**
  - ✓ **Contexte fort de Production**
  - ✓ **Multi-thématique**

2010 : 3rd system in production



National + University

Region(Local)+
National+University

AO EquipEX

15% computation system
devoted to SMEs

EQUIP@MESO :
10 mésocentres partenaires

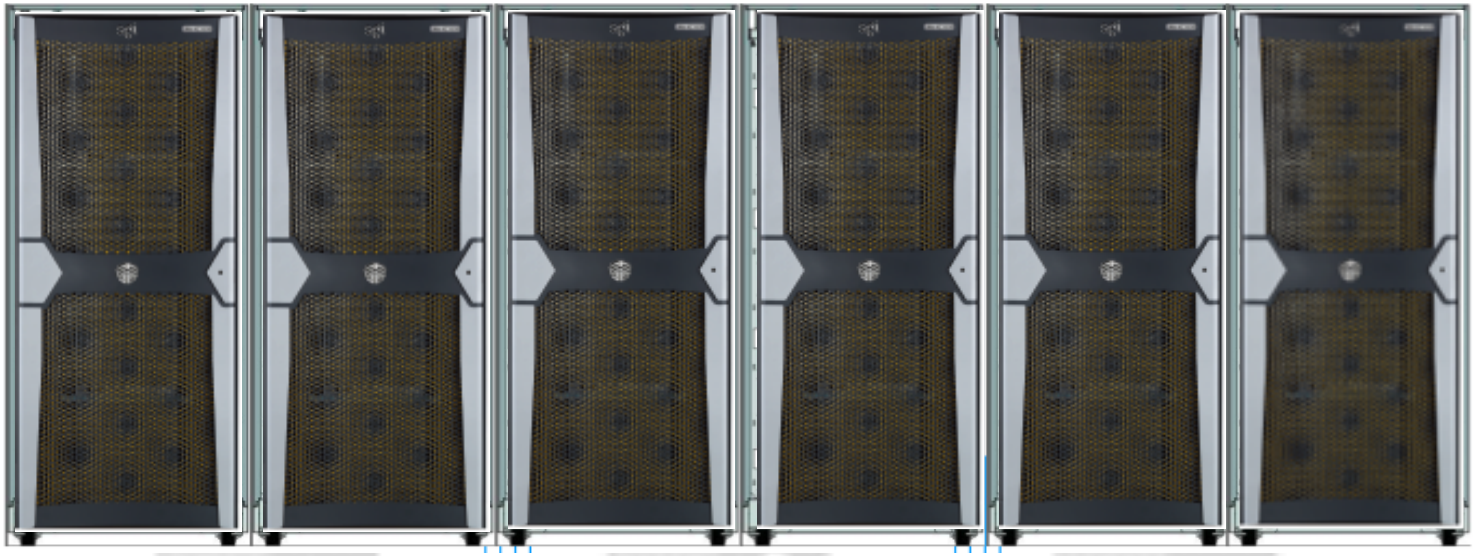
Fundings

EQUIP@MESO

‘HYPERION’, CALMIP computational system

- HYPERION (config avant EQUIP@MESO)
  - 2912 cores Nehalem Intel©
  - 33,57 TF peak
  - 223rd TOP 500 (november 2009)

Altix ICE 8200 EX :



Calcul Haute Performance : TOP 500 List

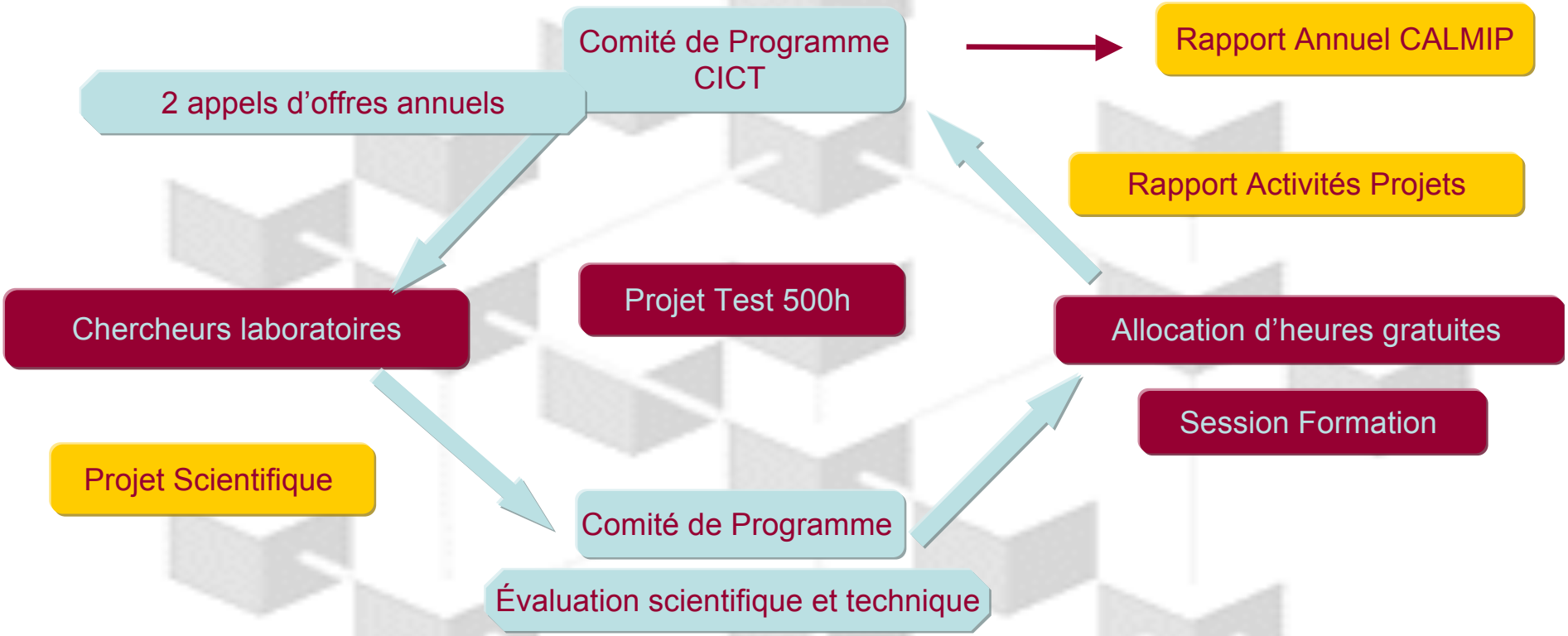
Top500 Novembre 2011

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010 NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.0
4	National Supercomputing Centre in Shenzhen (NSCS) China	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0
5	GSIC Center, Tokyo Institute of Technology Japan	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.6
6	DOE/NNSA/LANL/SNL United States	Cray XE6, Opteron 6136 8C 2.40GHz, Custom / 2011 Cray Inc.	142272	1110.00	1365.81	3980.0
7	NASA/Ames Research Center/NAS United States	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband / 2011 SGI	111104	1088.00	1315.33	4102.0
8	DOE/SC/LBNL/NERSC United States	Cray XE6, Opteron 6172 12C 2.10GHz, Custom / 2010 Cray Inc.	153408	1054.00	1288.63	2910.0
9	Commissariat a l'Energie Atomique (CEA) France	Bull bullx super-node S6010/S6030 / 2010 Bull	138368	1050.00	1254.55	4590.0
10	DOE/NNSA/LANL United States	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.0

12 MW

CURIE Machine « PRACE » (France)

CALMIP : Attribution des ressources pour la Recherche



- Prochain Appel d'offres
  - Mai 2012, (complément d'heure, nouveaux projets,…) ⇨ web : évaluation et attribution ressources jusqu'à fin année 2012

CALMIP : Les Labos utilisateurs

Pôle Science de la Matière :

CEMES - Centre d'Elaboration de Matériaux et d'Etudes Structurales (UPR 8011)

CIRIMAT - Centre Inter-universitaire de Recherche et d'ingénierie des Matériaux (UMR 5085)

IMRCP - Laboratoire des Interactions Moléculaires et Réactivité Chimique et Photochimique (UMR 5623)

LCC - Laboratoire de Chimie de Coordination (UPR 8241)

LNCMI - Laboratoire National des Champs Magnétiques Intenses (UPR 3228)

LCAR - Laboratoire Collisions Agrégats Réactivité (UMR 5589)

LCPQ - Laboratoire de Chimie et de Physique Quantiques (UMR 5626)

LPCNO - Laboratoire de Physique et Chimie des Nano-Objets (UMR 5215)

LPT - Laboratoire de Physique Théorique (UMR 5152)

Pôle Mathématiques Sciences et Technologies de l'Information et de l'Ingénierie :

ICA - Institut Clément Ader

IMFT - Institut de Mécanique des Fluides de Toulouse (UMR 5502)

IMT - Institut de Mathématiques de Toulouse (UMR 5219)

IRIT - Institut de Recherche en Informatique de Toulouse (UMR 5505)

LAAS - Laboratoire d'Analyse et d'Architecture des Systèmes (UPR 8001)

LGC - Laboratoire de GÃ©nie Chimique (UMR 5503)

LAPLACE - Laboratoire Plasma et Conversion d'Energie (UMR 5213)

Pôle Univers Planète Environnement Espace :

IRAP - Institut de Recherche en Astrophysique et Planétologie

CNRM/GAME - Centre National de Recherches Météorologiques (URA 1357)

LA - Laboratoire d'Aérologie (UMR 5560)

LEGOS - Laboratoire d'Etudes en Géophysique et Océanographie Spatiale (UMR 5566)

LMTG - Laboratoire des Mécanismes et Transferts en Géologie (UMR 5563)

Pôle Sciences du Vivant

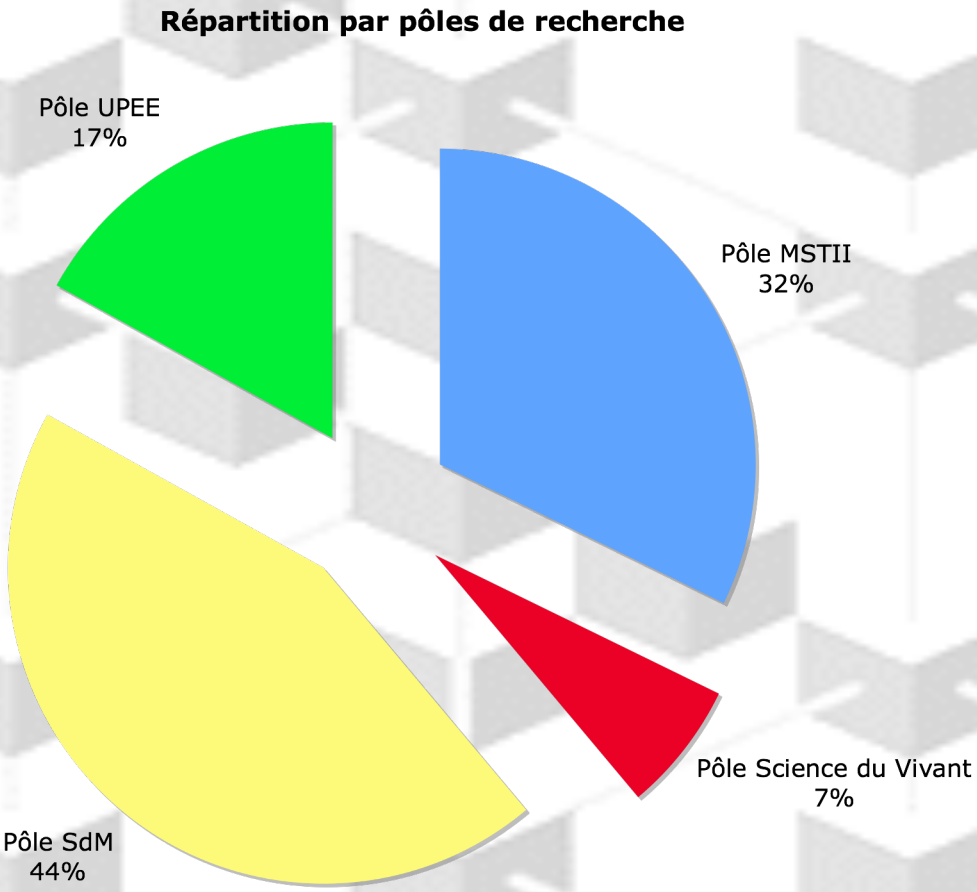
IPBS - Institut de Pharmacologie et de Biologie Structurale (UMR 5089)

LIPM - Laboratoire des Interactions Plantes Micro-organismes (UMR 2594)

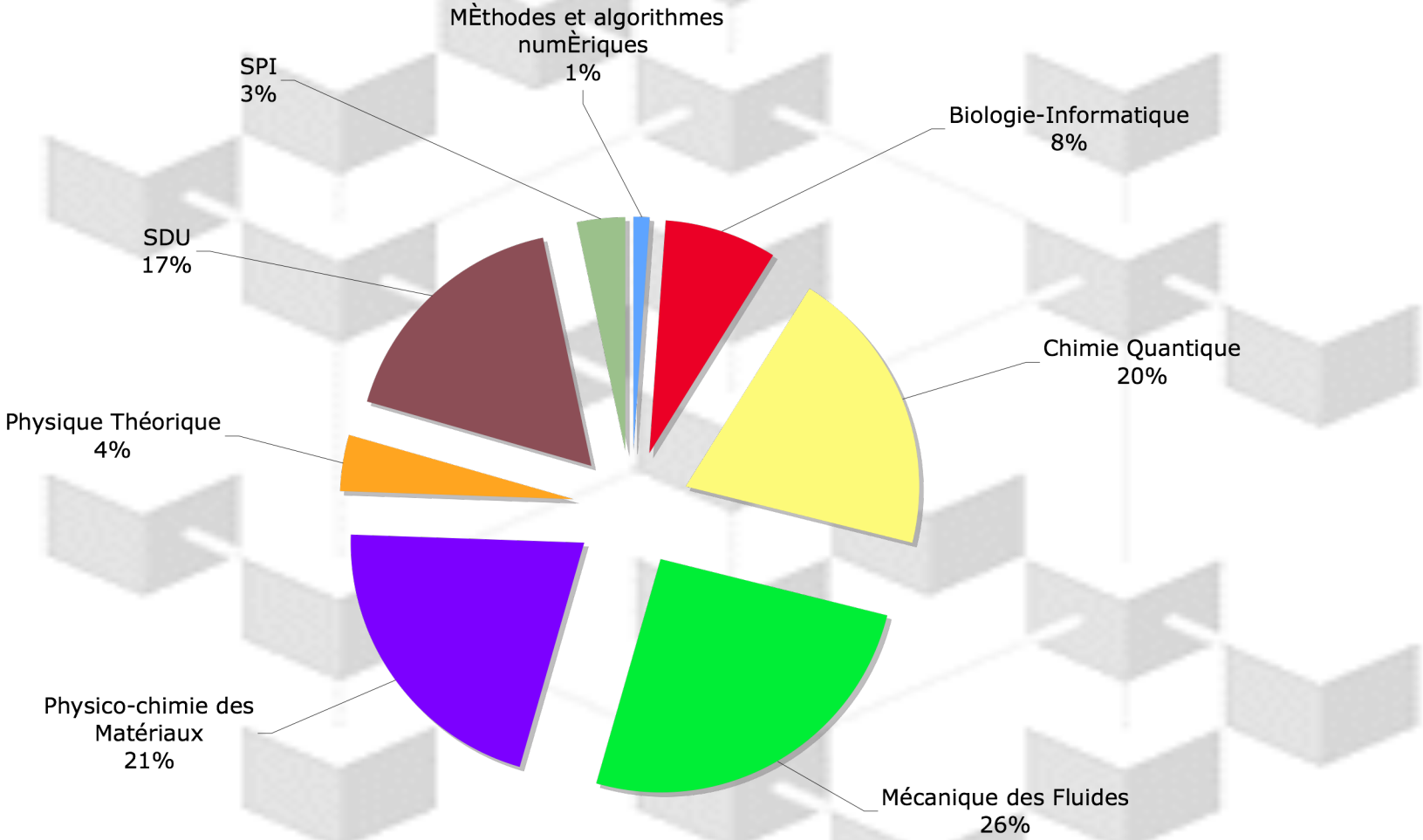
EDB - Evolution et Diversité Biologique (UMR 5174)

INSERM U563, dept. oncologie

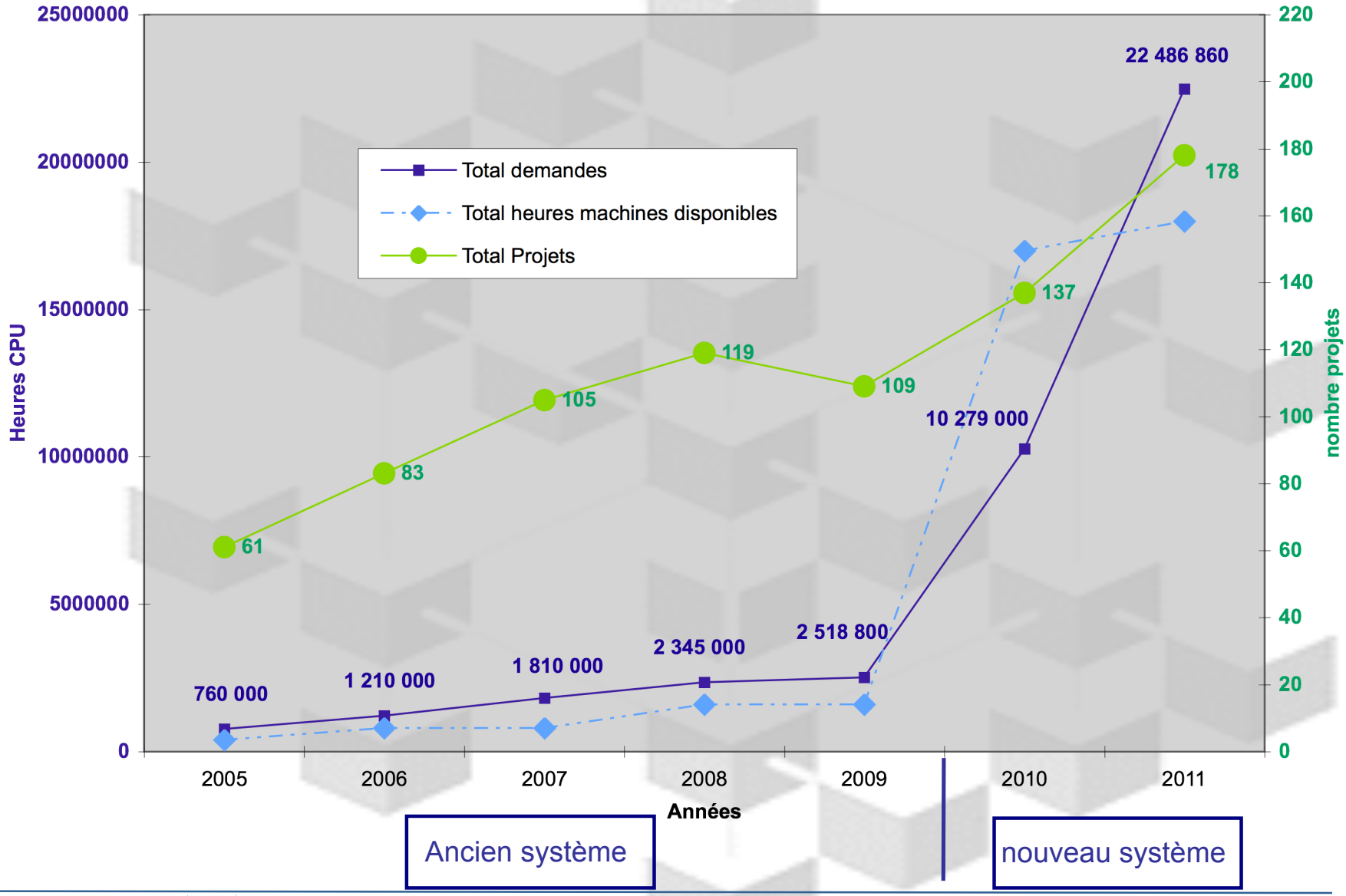
CALMIP : Attribution des ressources pour la Recherche



CALMIP : Attribution des ressources pour la Recherche

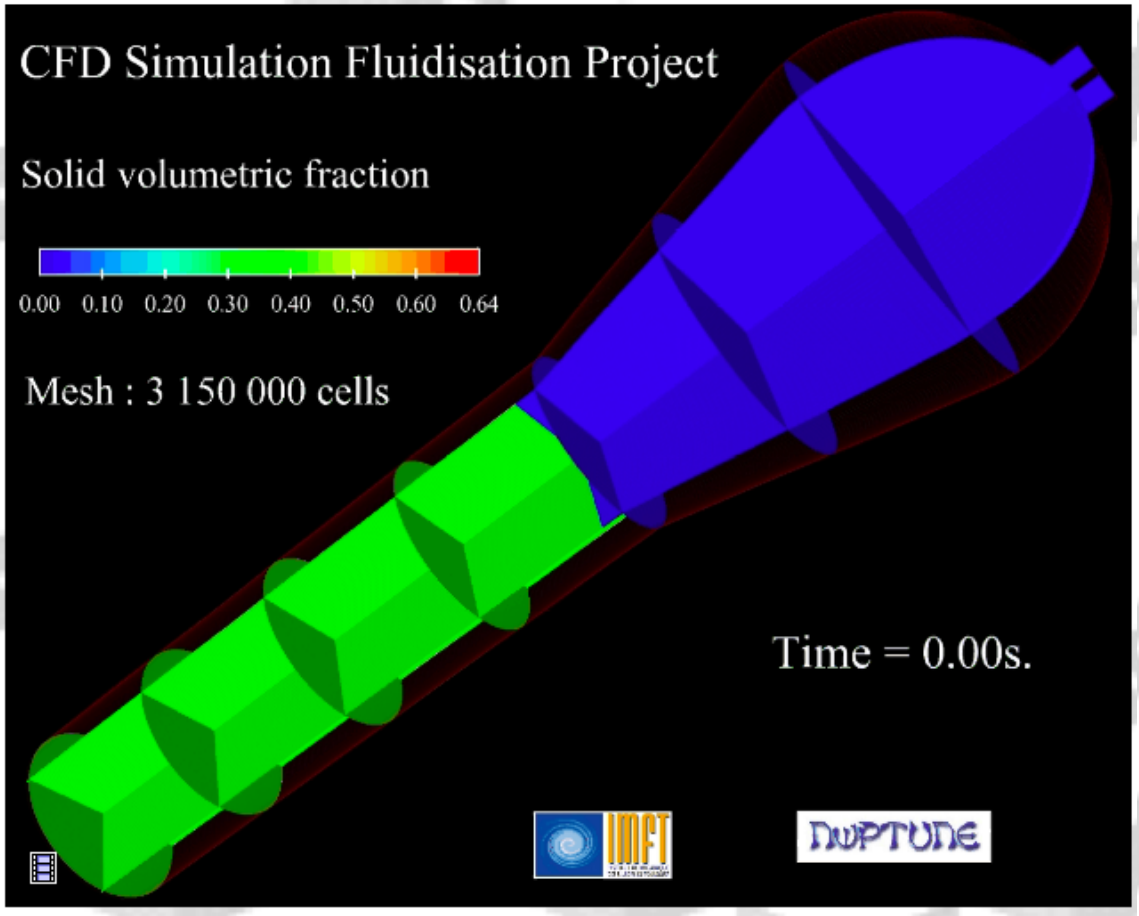


Evolution demande H_CPU/nombre_projets



Distributed memory example in CFD : Industrial fluidised bed

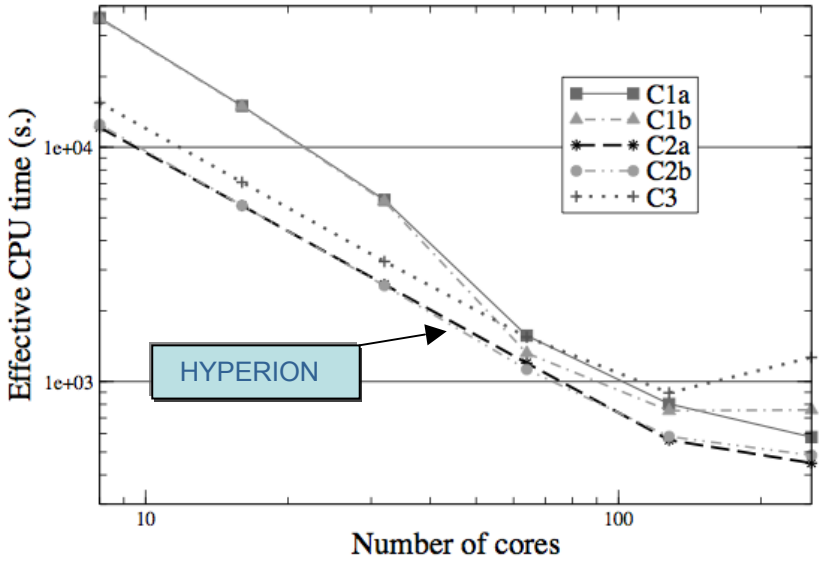
- Neptune_CFD : 3 000 000 cpu hours used in Y2010 on HYPERION
- Number of cores range in production : 68c - 512c

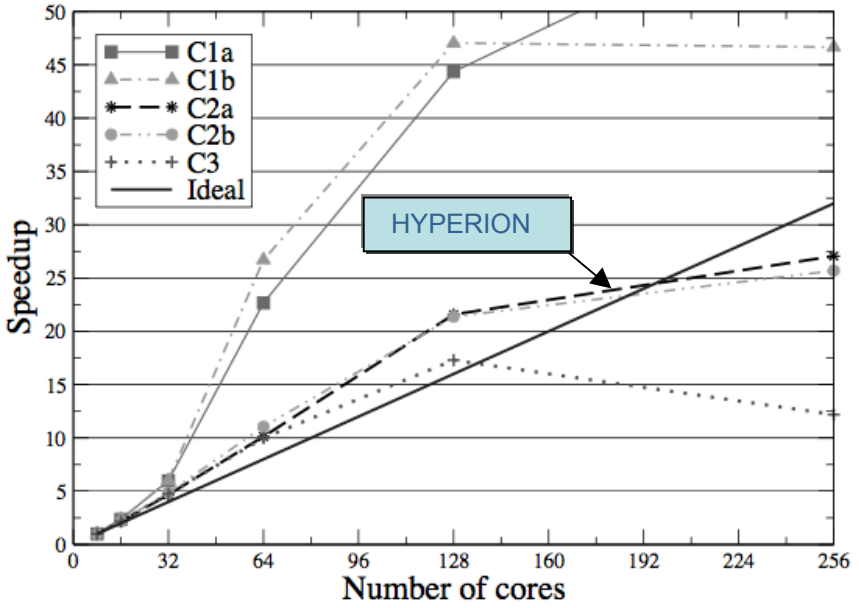


- Courtesy of : O Simonin, H. Neau, Laviéville - Institut de Mécanique des fluides de Toulouse - Université de Toulouse/ CNRS

Distributed memory example in CFD : Industrial fluidized bed

- Time-to-solution and speed-up





- C1 a : Altix ICE Harpertown, Intel_MPI
- C1 b : Altix ICE Harpertown, MPT
- C2 a : 'HYPERION' Altix ICE NHM, Intel MPI
- C2 b : 'HYPERION' Altix ICE NHM, MPT
- C3 : Cluster IB AMD Shanghaï, OpenMPI

- Courtesy of : O Simonin, H. Neau, Laviéville - Institut de Mécanique des fluides de Toulouse - Université de Toulouse/ CNRS

- Example in Life Science
  - Fish population fragmentation studies
    - Counting in fish in two Rivers near by Toulouse
  - fish population fragmentation studies
  - 12 000 cpu hour used
  - mostly data set exploring (code in C! Good !)

Chevaine

Gandoise

Vairon

- CALMIP
  - Structuration du Mésocentre de Calcul
  - Les labos partenaires
  - Quelques chiffres
- **Système de calcul**
  - **Choix d'un système**
  - **Hardware / Software**
- Exploitation et Accompagnement
  - Interface Utilisateur
  - Intégration projet de recherche
- Perspectives
  - EQUIP@MESO
  - ECA

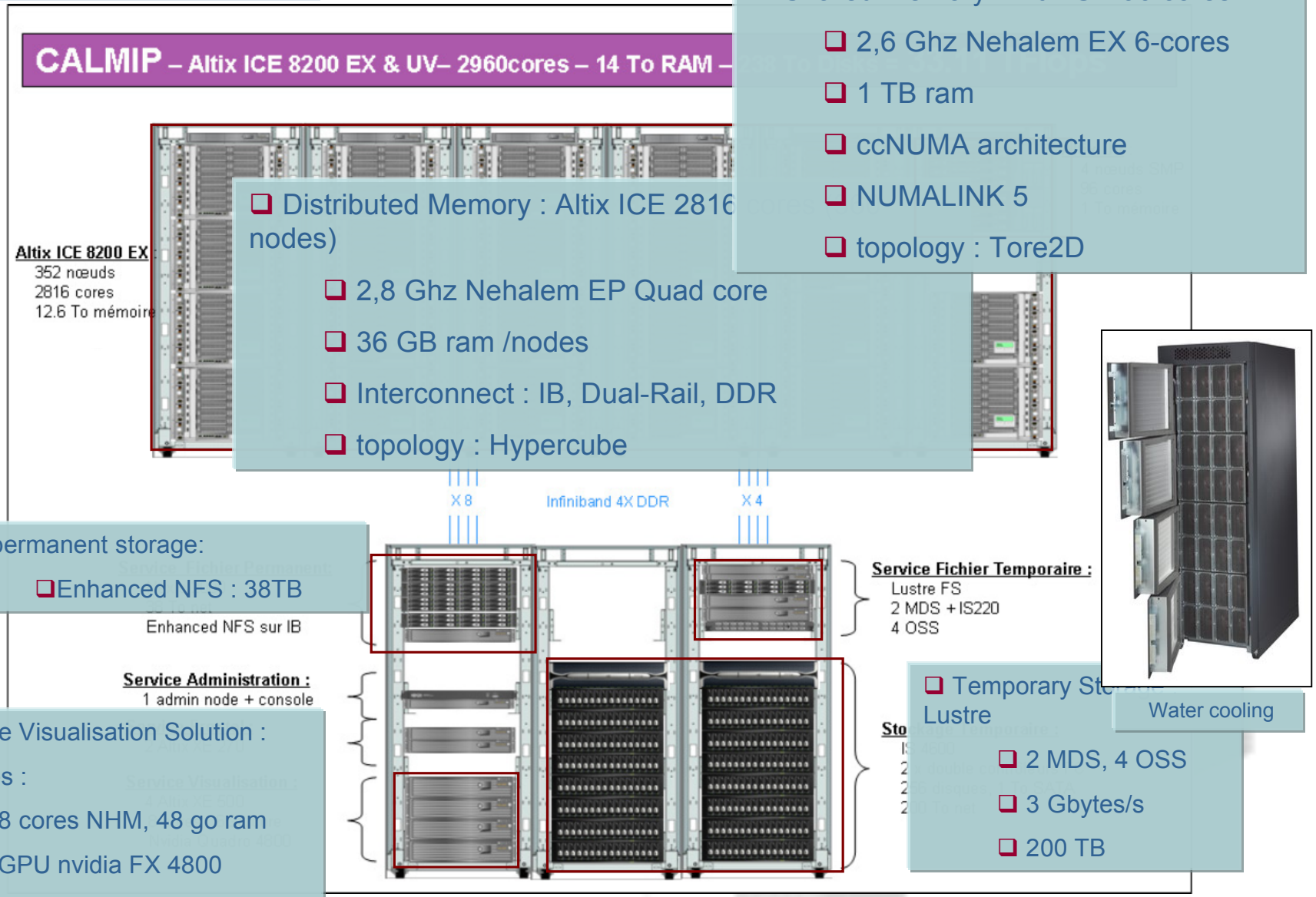
CALMIP : Processus de renouvellement 2008-2009

- Troisième génération de supercalculateur
- Critères techniques:
  - Qualité:
    - Système performant pour tirer les applications vers le haut
    - Critères environnementaux: consommation électrique, mode de refroidissement, …
  - Quantité:
    - Pallier la pénurie de ressources et anticiper leur augmentation
- Adhésion de la communauté scientifique:
  - Faire participer les utilisateurs à ce choix.
- Procédure : dialogue Compétitif

CALMIP : Processus de renouvellement 2008-2009

- **Groupe de test de performance : 10 benchs, 11 codes, 20 personnes, 11 Labos**
  - Participants
    - Ch. Lepetit (LCC), R. Poteau (LPCNO), L. Maron (LPCNO), F. Jolibois (LPCNO)
      - code GAUSSIAN (Chimie Quantique), OpenMP
    - H. Tang (CEMES), I. Gerber (LPCNO)
      - VASP (Physico-Chimie de la Matière), MPI
    - J. Czaplicki (IPBS), P. Arnaud (LCC) et I. Andre (LISBP)
      - AMBER (Dynamique Moléculaire), MPI
    - B. Dintrans (LATT, OBS-MIP)
      - Pencil Code (Astrophysique), MPI
    - A. Pedrono (IMFT) H. Neau (IMFT)
      - JADIM(dev.), NEPTUNE, (Mécanique des Fluides), MPI
    - S. Capponi (LPT-IRSAMC)
      - DO36(dev.) (Physique Théorique), OpenMP
    - A. Scemama (LCPQ – IRSAMC)
      - QmcCHEM(dev.) (Physique théorique et Chimie Quantique), MPI
    - E. Courcelles (LIPM) Génomique-Bioinformatique ProdomAlign (dev.) OpenMP
    - I. Touche (LGC) Algorithmique - Filtrage Particulaire (Dev.), MPI

HYPERION (2009-2013)



- Shared memory : Altix UV 96 cores
  - 2,6 Ghz Nehalem EX 6-cores
  - 1 TB ram
  - ccNUMA architecture
  - NUMALINK 5
  - topology : Tore2D
- Distributed Memory : Altix ICE 2816 nodes)
  - 2,8 Ghz Nehalem EP Quad core
  - 36 GB ram /nodes
  - Interconnect : IB, Dual-Rail, DDR
  - topology : Hypercube
- permanent storage:
  - Enhanced NFS : 38TB
- Temporary St… Lustre
  - 2 MDS, 4 OSS
  - 3 Gbytes/s
  - 200 TB
- Remote Visualisation Solution :
- 4 nodes :
  - 8 cores NHM, 48 go ram
  - GPU nvidia FX 4800
- Virtual GL/ Turbo VNC

Water cooling

CALMIP : Software environnement

- Profils codes
  - Langage : Fortran (77,90), C, C++
  - Parallèle : MPI (70%), OpenMP (30%)
  - Mémoire partagée : jusqu'à 600 Go (OpenMP)
- Environnement de développement
  - Compilo Intel (v10, 11, 12) (+ gnu)
  - MPI : MPT (MPI SGI®, Intel MPI, Open MPI)
  - Librairies scientifiques
    - MKL intel® (BLAS, LAPACK, Scalapack, …)
    - Petsc, MUMPS,…
  - Deboggeur // : ddt
  - Tunning code
    - Intel Trace Analyzer
    - Profileur code parallèle
    - Outils placement processus

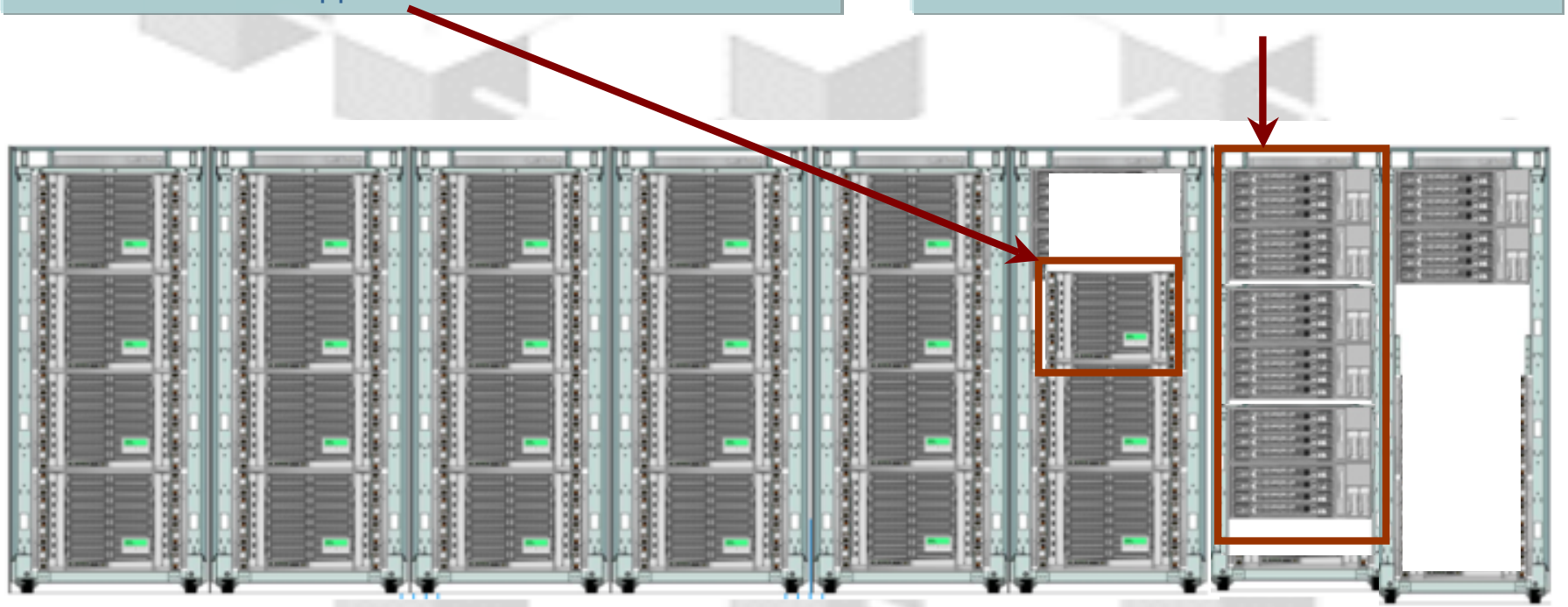
- codes utilisés :
  - VASP
  - SIESTA
  - NEPTUNE
  - Gaussian
  - AMBER
  - SYMPHONIE
  - WRF
  - ABAQUS
  - RADIOSS
  - OPENFOAM
  - MFIX
  - VINA
  - JADIM
  - ORCA
  - codes utilisateurs (dev.), etc …

- OS : SUSE SLES 11 SP2
- Gestionnaire de batch : PBS Pro
- Outil d'administration système TEMPO (SGI®)

Projet EQUIP@MESO (AO EQUIPEX) : UPGRADING

Projet EQUIP@MESO (AO EQUIPEX) : UPGRADING

- Distributed Memory : 16 nœuds supplémentaires nodes)
  - 2,8 Ghz Nehalem EP Quad core
  - 36 GB ram /nodes
  - 128 cores supplémentaires
- Shared memory : Altix UV 384 cores
  - 48 processeurs WestmereEX 8-cores
  - 3 To RAM
  - ccNUMA architecture



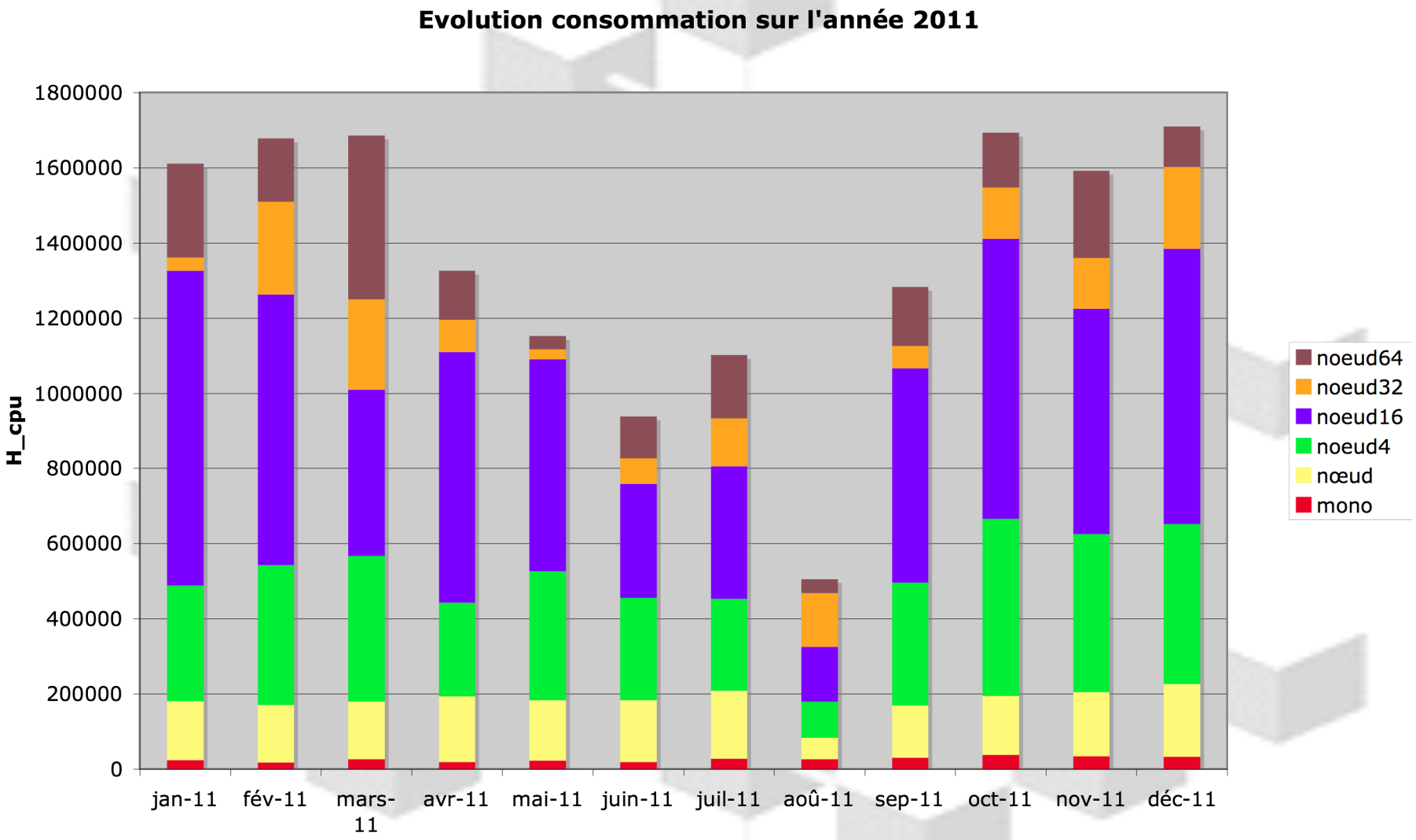
En Production

- CALMIP
  - Structuration du Mésocentre de Calcul
  - Les labos partenaires
  - Quelques chiffres
- Système de calcul
  - Choix d'un système
  - Hardware / Software
- **Exploitation et Accompagnement**
  - **Interface Utilisateur**
  - **Intégration projet de recherche**
- Perspectives
  - EQUIP@MESO
  - ECA

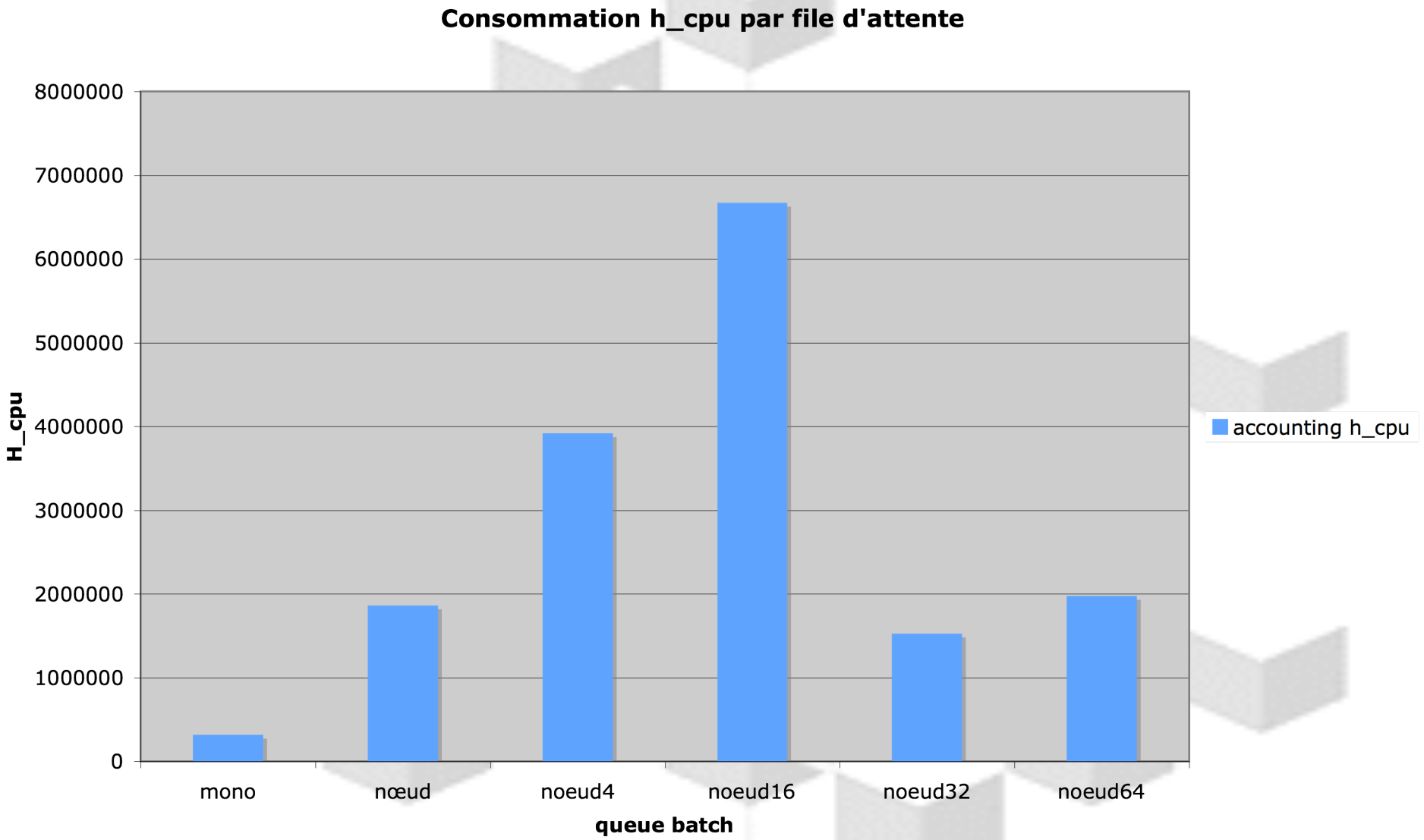
Exploitation

Queue	Nombre de cores	Nombre de nœuds	Walltime	Remarque
mono	< 8	1	400h	Non exclusif -non HT
noeud	8	1	300h	Exclusif - HT
noeud4	9 - 32	2 - 4	200h	Exclusif - HT
noeud16	33 - 128	5 - 16	150h	Exclusif - HT
noeud32	129 - 256	17 - 32	48h	Exclusif - HT
noeud64	257 - 512	33 - 64	36h	Exclusif -HT
UV*	1-384	1	240h	Non-exclusif - non HT
graphique	8 max.	1	2h	Non exclusif - HT

Exploitation



Exploitation

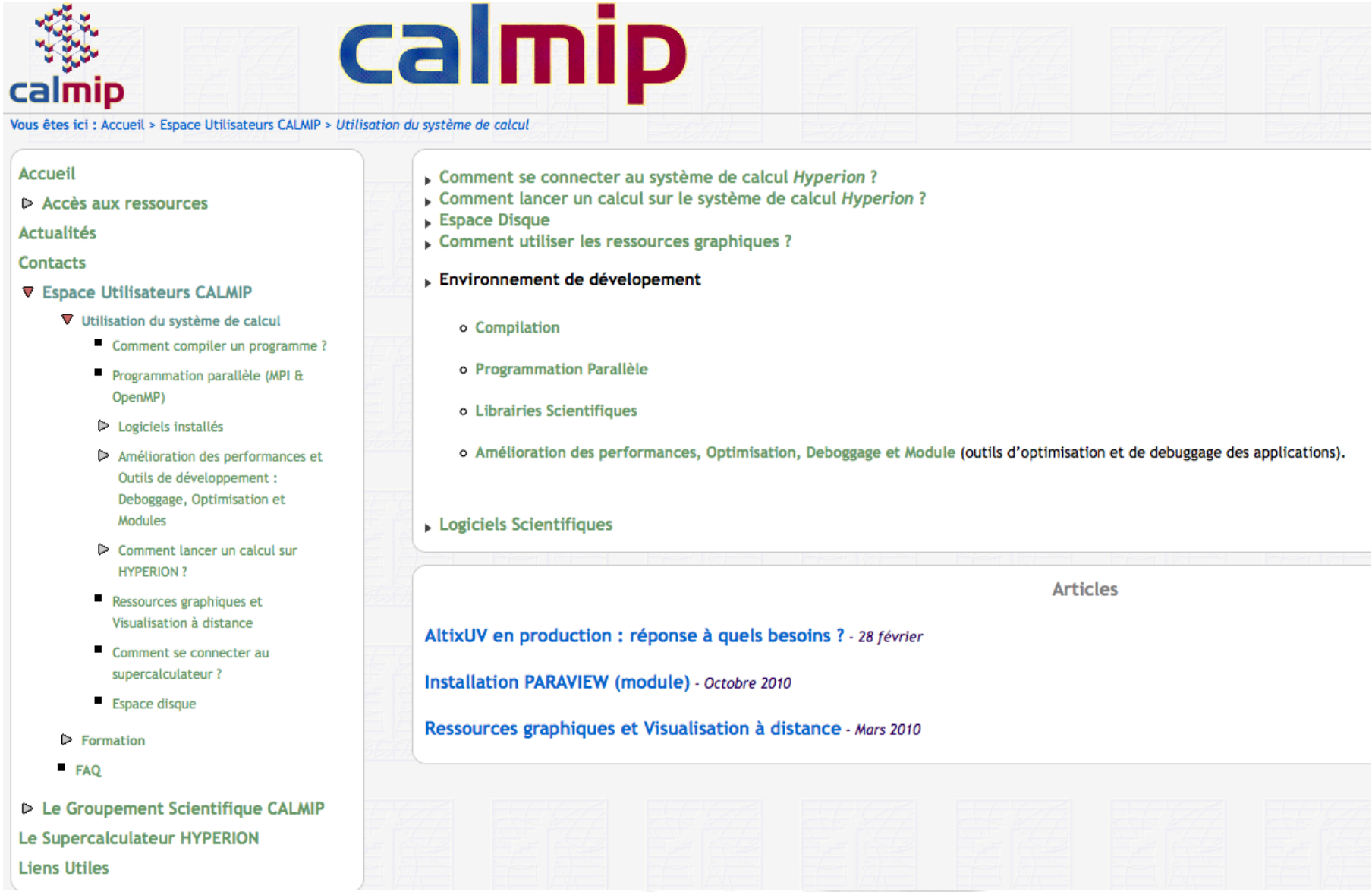


CALMIP : Interface/ Accompagnement utilisateur

- Base :
  - accueil / discussion / échanges problématique recherche
  - synergie(on essaye!) Ingé. Calcul / ASR dans les labos
- Interface 3 niveaux
  - Niveau 1 : prise en main opérationnel du système, les calculs tournent
    - Site WEB (compilation, batch, debuggeur) - Mise en autonomie
  - Niveau 2 : le code tourne, tourne-t-il de façon optimale ?
    - Outil de d'analyse, profiling, debugger
    - Optimisation des performances
  - Niveau 3 : Collaboration long terme/ récurrente
    - intégration dans le projet de recherche
    - Conseil expertise

- Formations :
  - Formation récurrente : prise en main système CALMIP / Portage Optimisation des codes
    - 2011 : 3 sessions, 50 participants
  - Formations Ponctuelles (HMPP)
  - On s'appuie sur : CUTIS, Groupe Calcul (National), Cerfacs, EQUIP@MESO

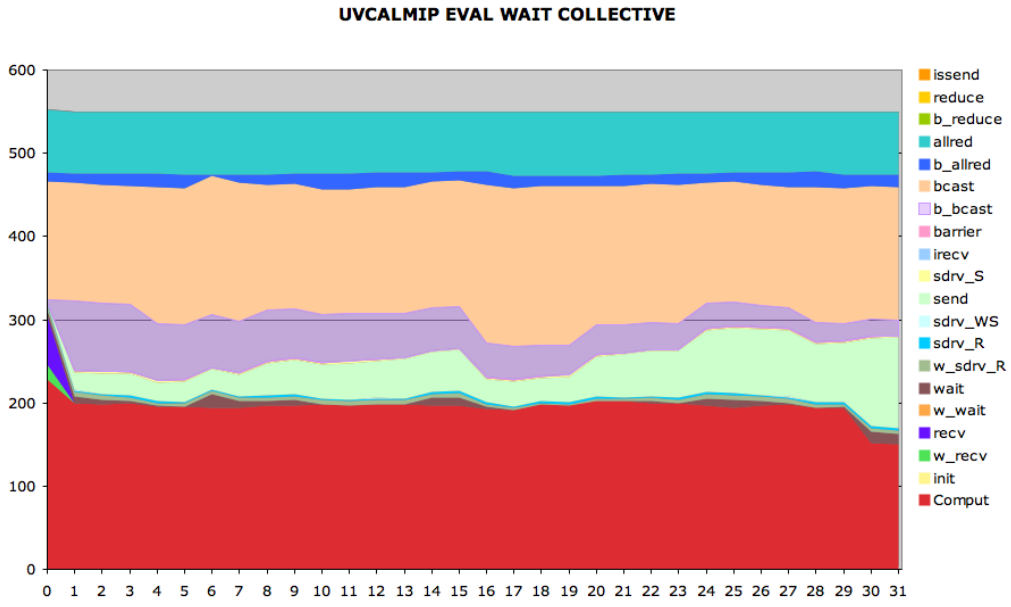
CALMIP : Interface/ Accompagnement utilisateur Niveau 1



CALMIP : Interface/ Accompagnement utilisateur Niveau 2

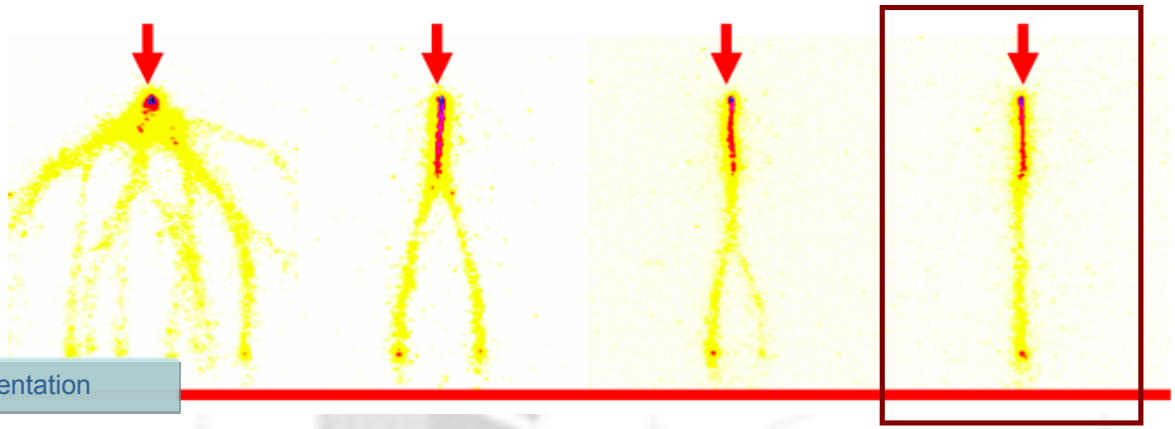
Les principaux outils d'optimisation disponibles sur HYPERION sont :

- NUMATools : outils propres aux architectures Altix : meilleures performances des jobs
  - exemple utilisation
- Accélérer les communications MPI : Perfboost
- Outil d'aide à l'analyse du parallèlisme des codes :
  - outil d'analyse de code MPI Intel(c) Trace Collector et Intel Trace Analyser
  - MPI_INSIDE
- PerfSuite : Analyse de performance des codes :
  - savoir facilement quel(s) est(sont) la(es) zone(s) du

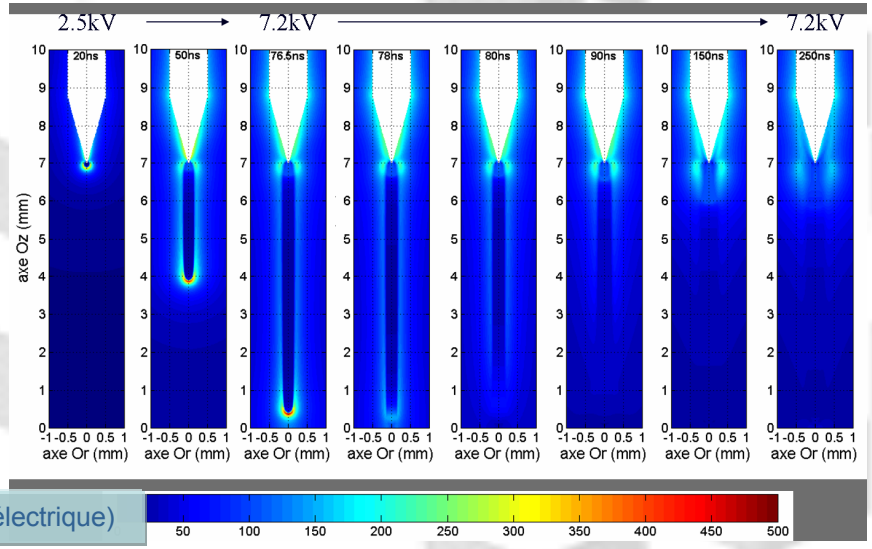


CALMIP : Interface/ Accompagnement utilisateur Niveau 3

- Exemple collaboration:
  - Projet CALMIP p0604 - Physique des Plasma - Laboratoire LAPLACE - O. Eichwald, O. Ducasse

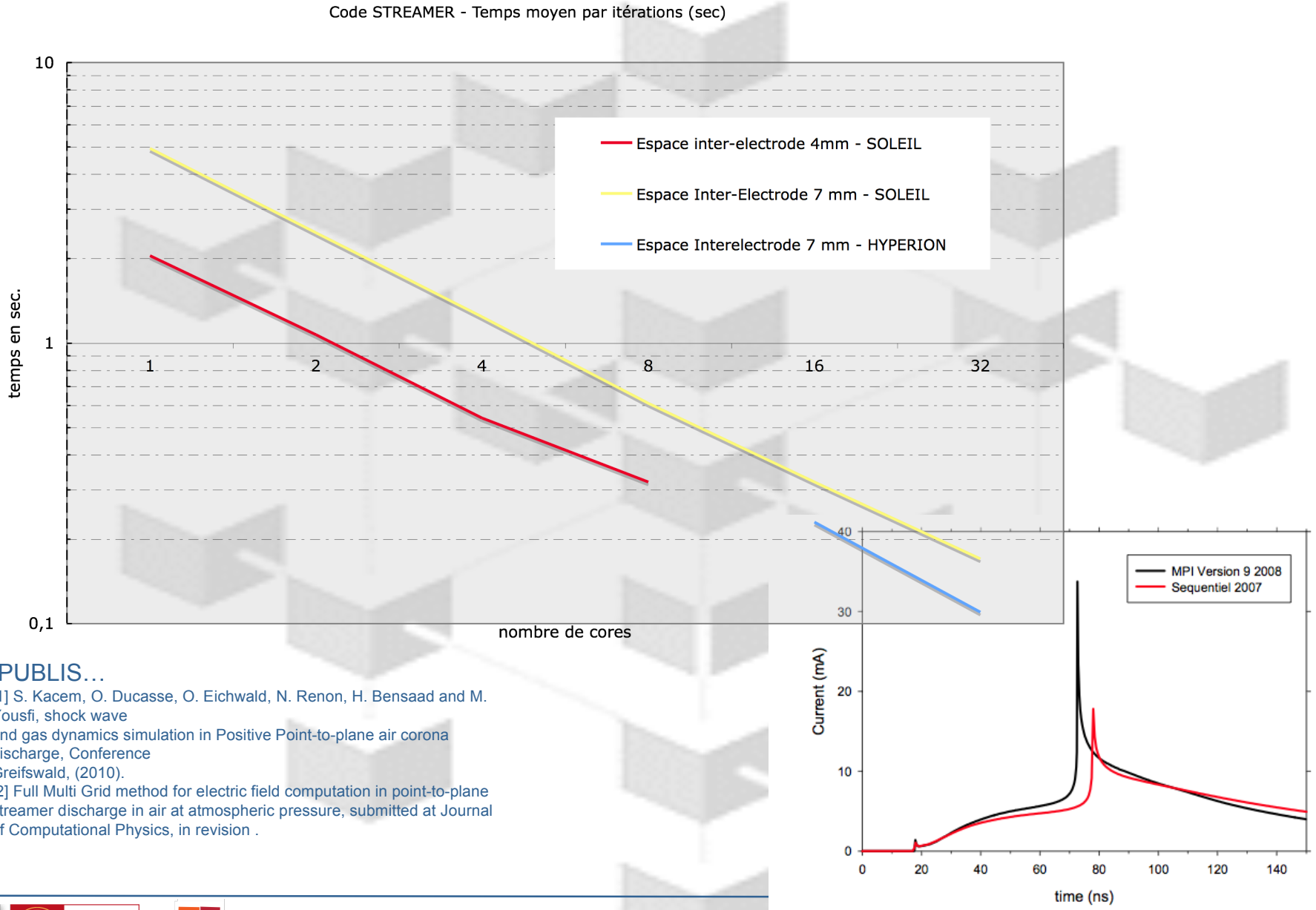


Expérimentation



Simulation (E - champ électrique)

CALMIP : Interface/ Accompagnement utilisateur Niveau 3



PUBLIS…

[1] S. Kacem, O. Ducasse, O. Eichwald, N. Renon, H. Bensaad and M. Yousfi, shock wave
and gas dynamics simulation in Positive Point-to-plane air corona discharge, Conference
Greifswald, (2010).

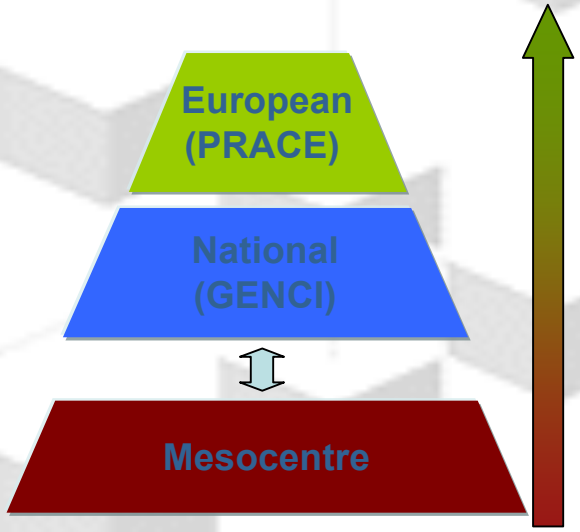
[2] Full Multi Grid method for electric field computation in point-to-plane streamer discharge in air at atmospheric pressure, submitted at Journal of Computational Physics, in revision .

- CALMIP
  - Structuration du Mésocentre de Calcul
  - Les labos partenaires
  - Quelques chiffres
- Système de calcul
  - Choix d'un système
  - Hardware / Software
- Exploitation et Accompagnement
  - Interface Utilisateur
  - Intégration projet de recherche
- **Perspectives**
  - **EQUIP@MESO**
  - **ECA**

CALMIP Partenaire EQUIP@MESO

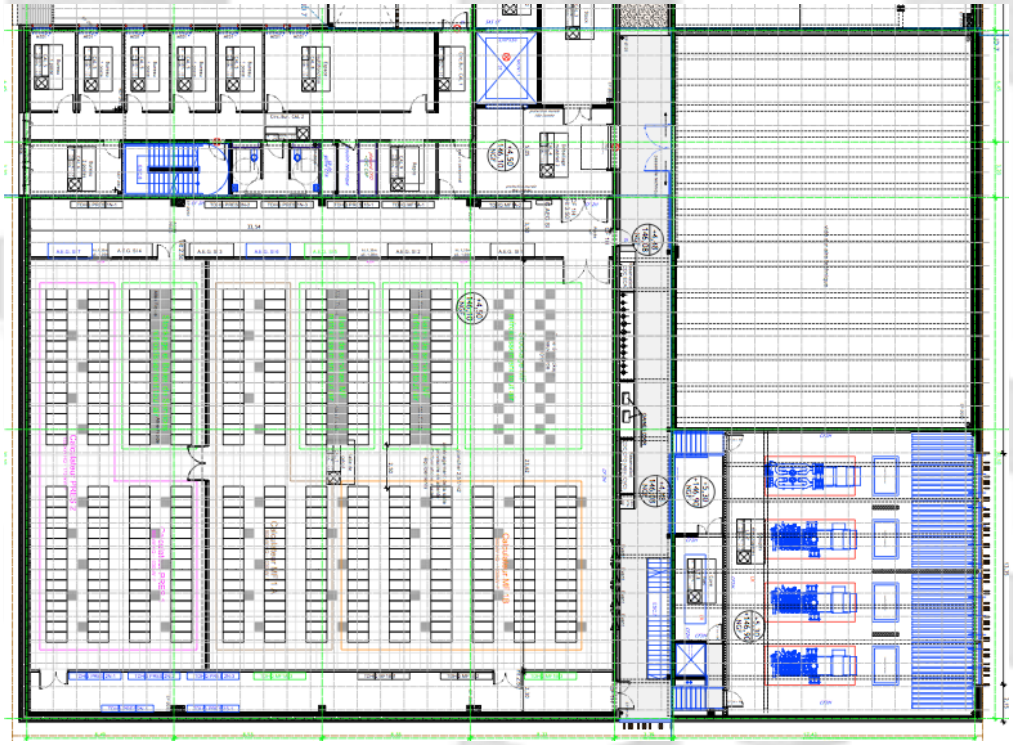
- Projet EQUIP@MESO
  - 10 mésocentres partenaires, coordination GENCI (Grand équipements Nationaux en Calcul Intensif)
  - Objectif : mise en niveau/amplification de 10 mésocentres (maillage mésocentre)
  - Animation scientifique : conférence EQUIP@MESO 2012 « Chimie et Sciences de la vie : de la simulation numérique au HPC »
    - Conférencier CALMIP : M. Caffarel/A. Scemama (LCPQ-IRSAMC) (projet CALMIP p0510 - Méthodes Monte-Carlo Quantiques pour les Molécules) :

      « Pyramide HPC : du labo au Pétaflop en passant par les mésocentres »

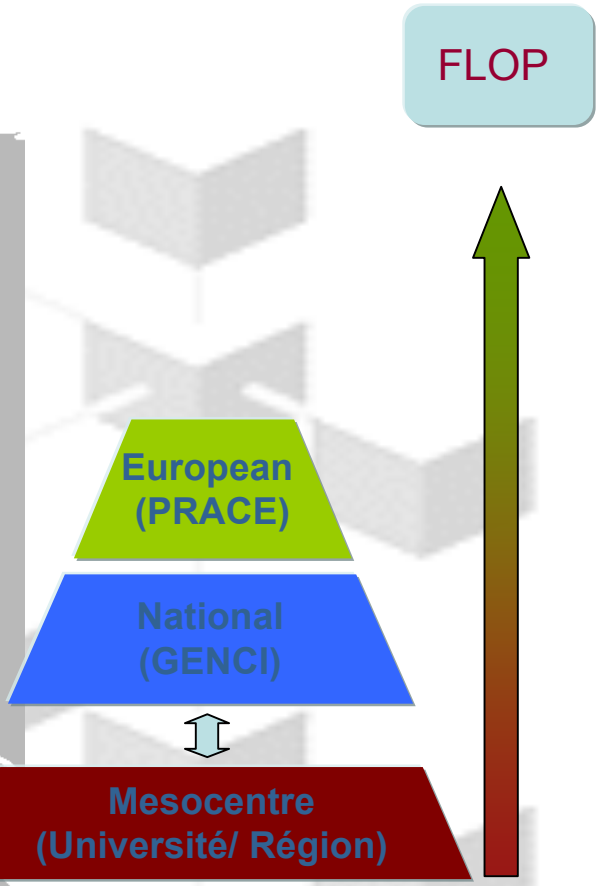


CALMIP : Renouvellement système pour 2014

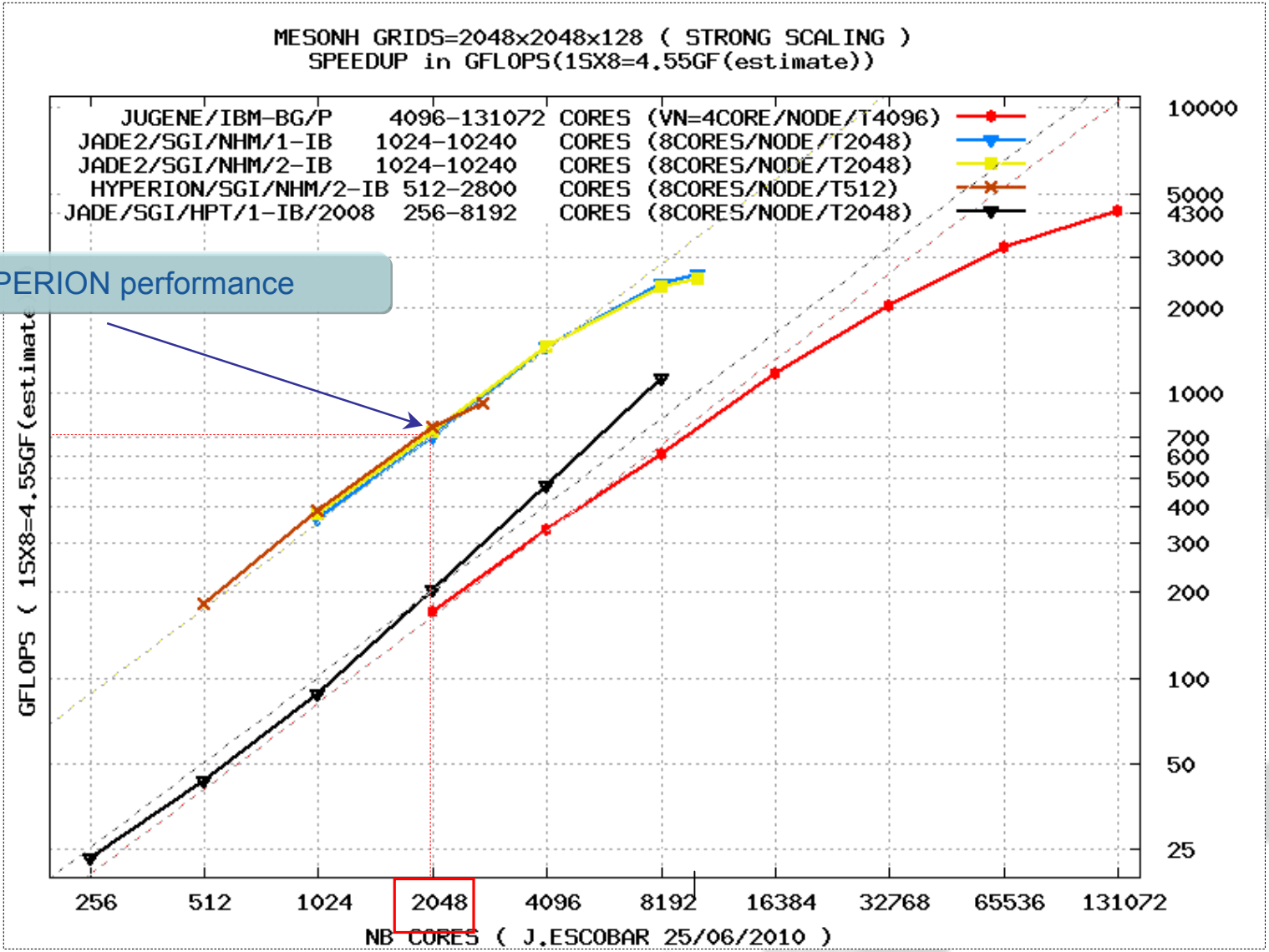
- Renouvellement d'HYPERION (Procédure dialogue 2012-2013)
  - Fonds CPER 2007 - 2013
  - Installation Espace Clément Ader (Partie PCI) :
    - Infrastructure d'accueil Mutualisée avec Météo-France
  - Production visée : 2014



HPC en Europe (Partnership for Advanced Computing in Europe)



MesoNH Performance on HYPERION



HYPERION performance

❑ Courtesy of : JP Pinty & J. Escobar Laboratoire Aérologie - Observatoire Midi-Pyrénées Université Paul Sabatier/ CNRS

Code STREAMER : Equation du problème et Algorithme

$$\frac{\partial n_s}{\partial t} + \vec{\nabla}.n_s\vec{v}_s\left(\frac{E}{N}\right) = \sigma_s\left(\frac{E}{N}\right) \qquad s = e, p, n \qquad (1)$$

Bilan particule pour chaque espèce (equation de transport ou equation de continuité)

$$n_s\vec{v}_s = n_s\mu_s\left(\frac{E}{N}\right)\vec{E} - D_s\left(\frac{E}{N}\right)\vec{\nabla}n_s \qquad s = e, p, n \qquad (2)$$

Approximation dérive-diffusion du transport de la quantité de mouvement

$$\Delta V = \frac{q_e}{\varepsilon_0}(n_p - n_e - n_n) \qquad (3)$$

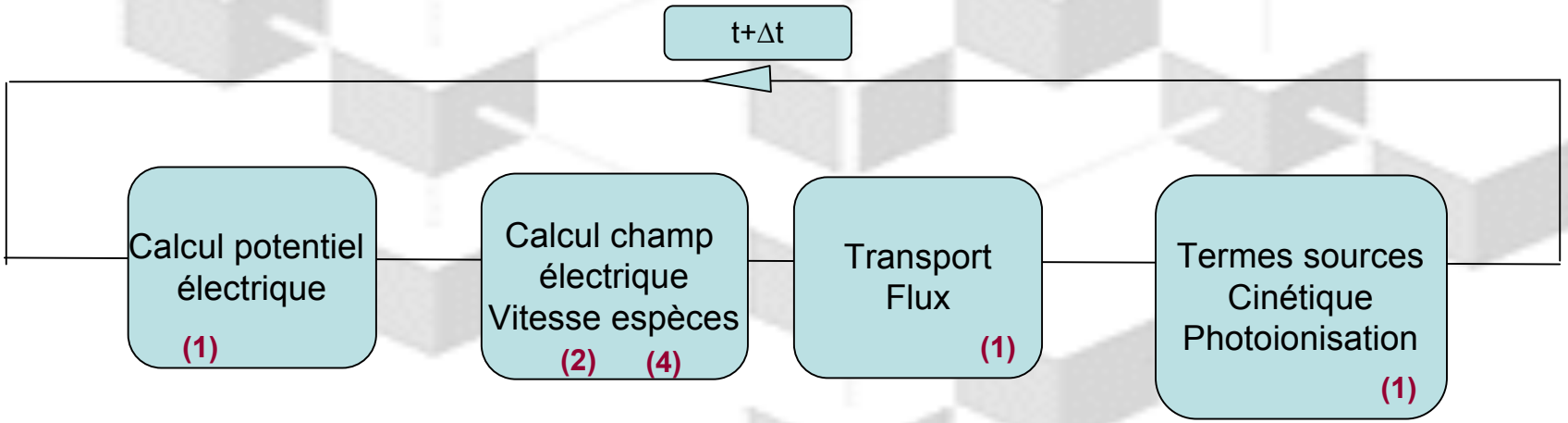
Equation de poisson calcul potentiel electrique

$$\vec{E} = -\vec{\nabla}V \qquad (4)$$

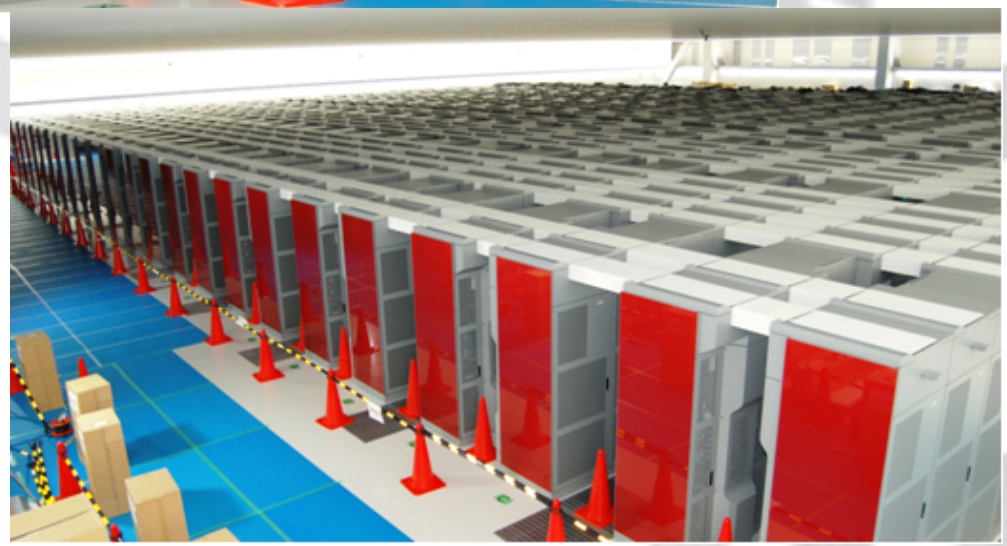
Relation champ-potentiel électrique

Domaine Ω semi-infini => condition dirichlet (plan) + Neuman (gradient nul)

Discrétisation Volume fini (2D axis) /Schéma en temps explicite



Japon : K-computer (Y2011), 580 000 cores



- Fujitsu RIKEN 8 PFlop (Kobe Japan)
- efficiency 93%
- 548 k core Sparc
- 10 Mw, 825 Mflop/W
- Single socket, 8C socket, 8 flop per core
- 6D tore Interconnect

CALMIP : Processus de renouvellement 2008-2009

❑ Bilan de la procédure

- ❑ Long mais nécessaire et extrêmement bénéfique pour la communauté CALMIP
- ❑ Échanges directs avec les candidats dans un cadre administratif clair
- ❑ Participation des utilisateurs (préfigure une exploitation réussie…) :
  - ❑ Performance et adhésion (facilement exploitable)
  - ❑ Evolutions techniques (programme fonctionnel) : au plus près des besoins
  - ❑ Enrichissement : Visualisation à distance