Présentation du mésocentre de calcul de l'université de Bourgogne

Jean-Jacques Gaillard

MPI et intégration SGE

Didier Rebeix

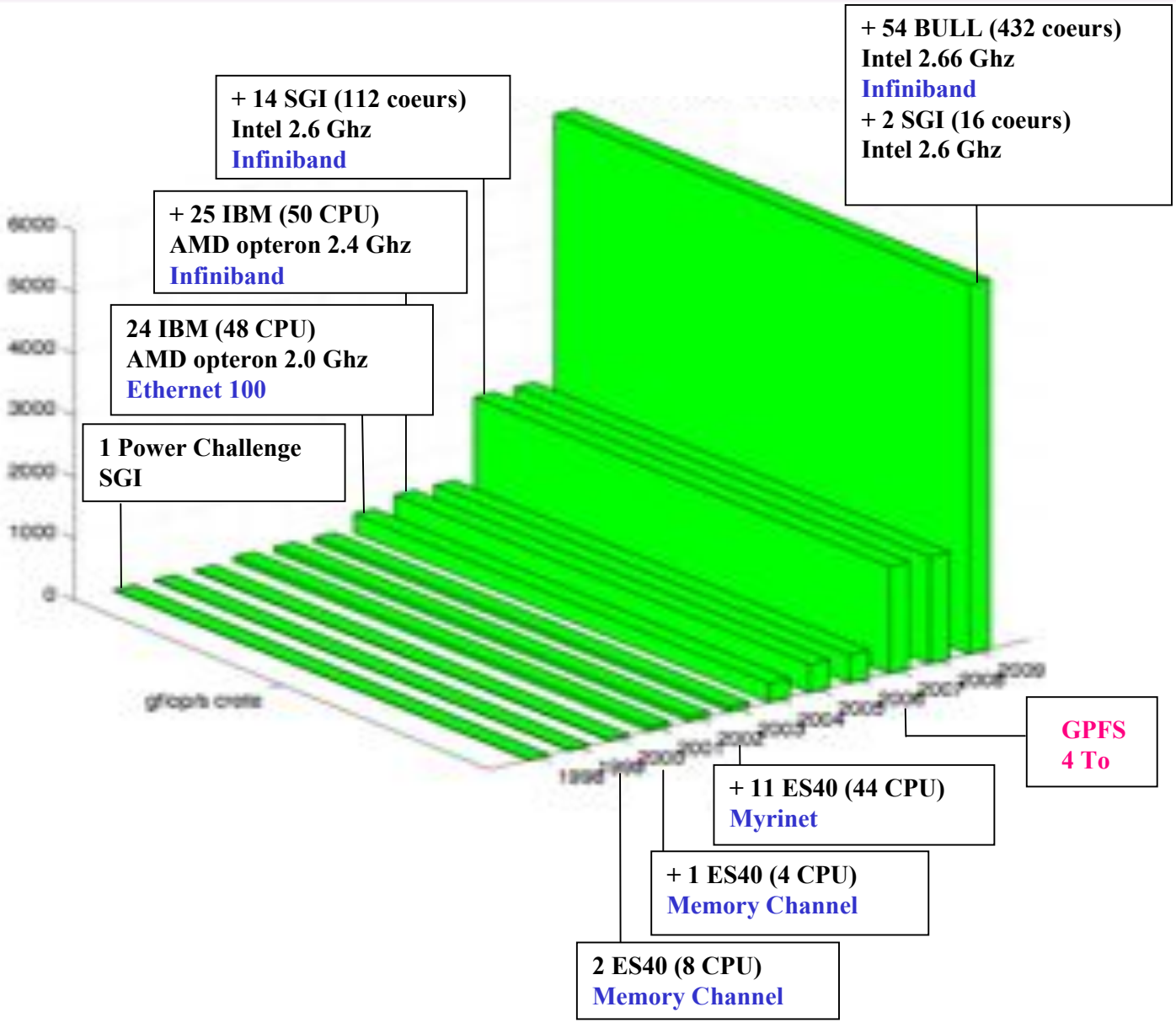
Retour utilisateur : Compilateurs, Précision

Olivier Politano

http://www.u-bourgogne.fr/cri-ccub

Réunion Mésocentres - Groupe Calcul IN2P3 Lyon - 10 juin 2010

Principales réalisations depuis 1998



Le cluster aujourd'hui

- 703 coeurs, 6.8 teraflop/s
- 60 teraoctets de stockage
- Réseaux Infiniband avec switch 96 ports SDR
- Sur 2 salles machines

Le mésocentre de calcul de l'uB

- 100 utilisateurs recherche / 1300 comptes avec enseignement

- Ouvert à tous les labos et composantes de l'université
  - ✓ 6 principaux laboratoires / instituts

- Centre de Calcul
  - ✓ 2.5 ETP messageries et LDAP étudiant.
  - ✓ 2.5 ETP calcul
  - ✓ 1 expert scientifique (MCF avec décharge de 75H eq TD).

- Facturation pour les labos

- Budget : 160 K€/an investissement et fonctionnement.

- Environ 70 publications / an

La commission Calcul Numérique Intensif

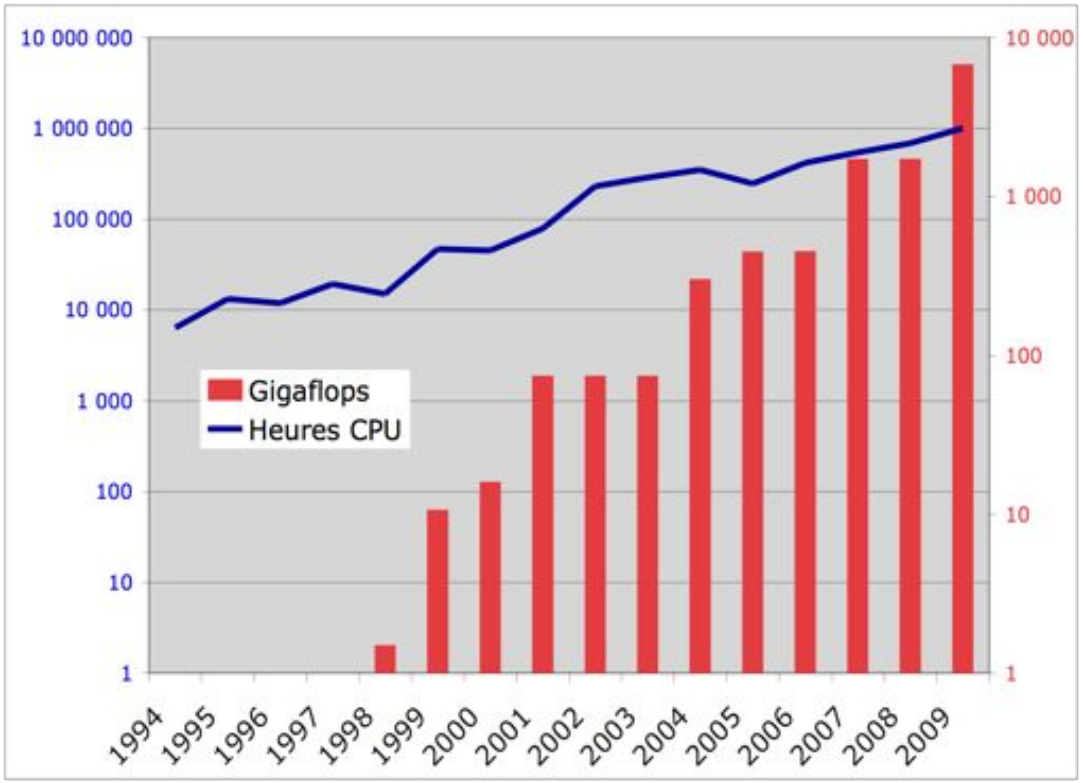
- Direction
  - ✓ Jean-Jacques Gaillard et Olivier Politano

- Membres
  - ✓ directrice du CRI + responsable système
  - ✓ au moins 1 représentant de chaque laboratoire utilisateur

- Role
  - ✓ enquête annuelle auprès des utilisateurs
  - ✓ reçoit les fournisseurs, établit les cahiers des charges.
  - ✓ conseille et guide le CRI-CCUB pour investissements (matériels, logiciels,…)

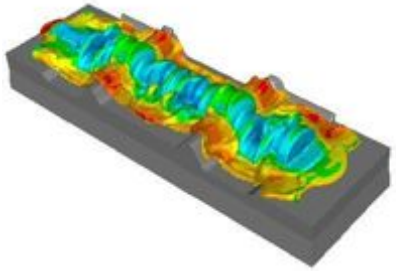
Statistiques

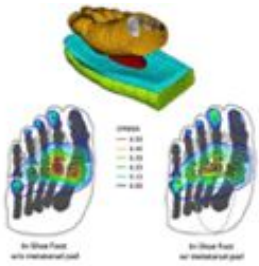
- 2008 : 683 000 h
- 2009 : 1 000 000 h réalisées
- 2010 : 1 190 000 h sur 5 mois => 3 000 000 ???



Logiciels installés

- Chimie : Gaussian, Vasp, Gamess, Gromacs...
- Physique/méca : Comsol, Forge, Castem, Abaqus
- Climatologie : WRF,...
- Calcul formel : Maple, Matlab, Mathematica
- Stat/analyse de données : SAS, Matlab, R
- Visualisation graphique : Amira, VMD
- Calcul parallèle : MPI, Fortran, C
- Bureautique scientifique : TEX





Formations via la mission doctorale.

- LINUX
- FORTRAN
- Parallèle MPI
- BATCH
- MATLAB

Problématique / Projets

- Stockage haute performance
  - ✓ Capacité d'au moins 30 To
  - ✓ Performances i/o
  - ✓ FS: lustre, GPFS, PanFS, NFS ou PNFS
- Climatisation
  - ✓ Actuellement 30 + 20 Kw
- Alimentation électrique insuffisante pour un développement futur.
- Salle inadaptée.

Projet : Nouvelle salle machine

- Achat climatisation 50 kw provisoire
- Salle trop petite
- On en cherche une de 100 m² extensible

- Refroidissement adapté : refroidissement liquide des portes AR
  - ✓souci d'efficacité, on refroidit les baies pas la salle
  - ✓plus économe
  - ✓moins de gaz effet de serre

- Et une alimentation électrique adéquate
  - ✓ extensible jusqu'à 400 KW ?
  - ✓ onduleur ou groupe tournant ?

Présentation du mésocentre de calcul de l'université de Bourgogne

Jean-Jacques Gaillard

MPI et intégration SGE

Didier Rebeix

Retour utilisateur : Compilateurs, Précision

Olivier Politano

MPI un choix difficile ?

- Une grande variété de piles mpi :
  - ✓ MPI constructeurs
  - ✓ MPI propriétaires : Intel MPI, HP MPI, ...
  - ✓ MPI génériques : mvapitch, openmpi, mpich, lammpi, …
- Diversité des logiciels et des compilateurs
- Certains logiciels propriétaires imposent une pile MPI (Forge, Abaqus, ...)
- Certains logiciels sont sensibles au compilateur choisi (gnu, Intel, PGI ?)
- Les piles MPI sont sensibles au compilateur choisi
- Les piles MPI ont des interfaces différentes (surtout arguments de mpirun)

MPI un choix difficile ?

- Nécessité de faire cohabiter différentes piles MPI.
- Nécessité de faire cohabiter différentes versions des piles MPI en fonction du compilateur et de ses versions
- Nécessité d'unifier les différentes interfaces

MPI un choix difficile ?

- Historique des piles MPI au CCUB :

  - ✓ Pré 2007 :
    - MPI constructeur : Infinicon (switch IB silverstorm)
  - ✓ 2007/2008 :
    - installation d'un nouveau switch IB voltaire
    - MPI constructeur : Voltaire OFED + Voltaire MPI (basé sur mvapitch)
    - Problèmes de stabilité et de « scaling » avec certains logiciels (VASP, WRF, …)
  - ✓ 2009/2010 :
    - Installation HP MPI pour Forge et Abaqus
    - Migration progressive vers OFED + openmpi

MPI un choix difficile ?

- Openmpi :

  - ✓ Opensource
  - ✓ Modulaire
  - ✓ Reprend le meilleur des autres piles MPI
  - ✓ Bonne intégration avec SGE (accounting temps CPU)
  - ✓ Compilation « facile », génération de rpms
  - ✓ Stable, performant, gratuit

MPI un choix difficile ?

- Cohabitation des différents piles MPI
  - ✓ Installation dans /opt/$MPI_TYPE/$MPI_VERSION/$COMPILER/$COMPILER_VERSION

- Unification des interfaces des différentes piles MPI :
  - ✓ Des scripts d'environnements (PATH)
    - Pour le choix du compilateur
    - Pour le choix de la pile MPI
  - ✓ Un « wrapper » pour mpirun

SGE

- Migration de LSF à SGE en 2007.

- Présentation rapide :
  - ✓ SGE : Sun Grid Engine
  - ✓ Gestionnaire de batch
  - ✓ Développé par Sun Microsystems (racheté par Oracle)
  - ✓ Open source (Pour combien de temps encore ?)

- Intégration MPI et OpenMP dans SGE
  - ✓ MPI : Le wrapper mpirun convertit le PE_HOSTFILE au format machinefile de mpirun
  - ✓ OpenMP : dans le starter :

    export OMP_NUM_THREADS=$NSLOTS

SGE

- Limitations rencontrées :

  ✓ Pas de système natif pour limiter le nombre de CPU par utilisateurs

  ✓ Calcul du nombre de processeurs fait dans le prologue, exit(99) si le nombre dépasse la limite configurée.

SGE

- Solution : un metascheduler ?
  - ✓ Des files d'attentes de soumission sans slots
  - ✓ Des files d'exécution correspondantes
  - ✓ Un daemon qui déplace (qalter) les jobs en files d'attente vers les files d'exécution.
    - Impose les limites
    - Corrige les erreurs courantes des utilisateurs (mauvais choix de file)

- Problèmes
  - ✓ Le metascheduler et le scheduler peuvent entrer en conflit
  - ✓ Les informations retournées par qstat sont faussées (state, ...)

SGE

- Metascheduler : une bonne idée ?

- Changer/modifier le scheduler de SGE

Présentation du mésocentre de calcul de l'université de Bourgogne

Jean-Jacques Gaillard

MPI et intégration SGE

Didier Rebeix

Retour utilisateur : Compilateurs, Précision

Olivier Politano

Retour utilisateur : Compilateurs, Précision

- un compilateur historique : Portland Group
  - ✓ imposé par Gaussian pour recompiler les sources
  - ✓ en 2003 meilleures performances pour les AMD opterons
- depuis 2010 : compilateur intel
- Pb manque de formation des utilisateurs

Comparaison Portland - Intel - Gnu

- Code de dynamique Moléculaire caractéristique d'un code utilisateur

http://www.fisica.uniud.it/~ercolessi/md/f90/

- Critères :

  ✓ Temps d'exécution

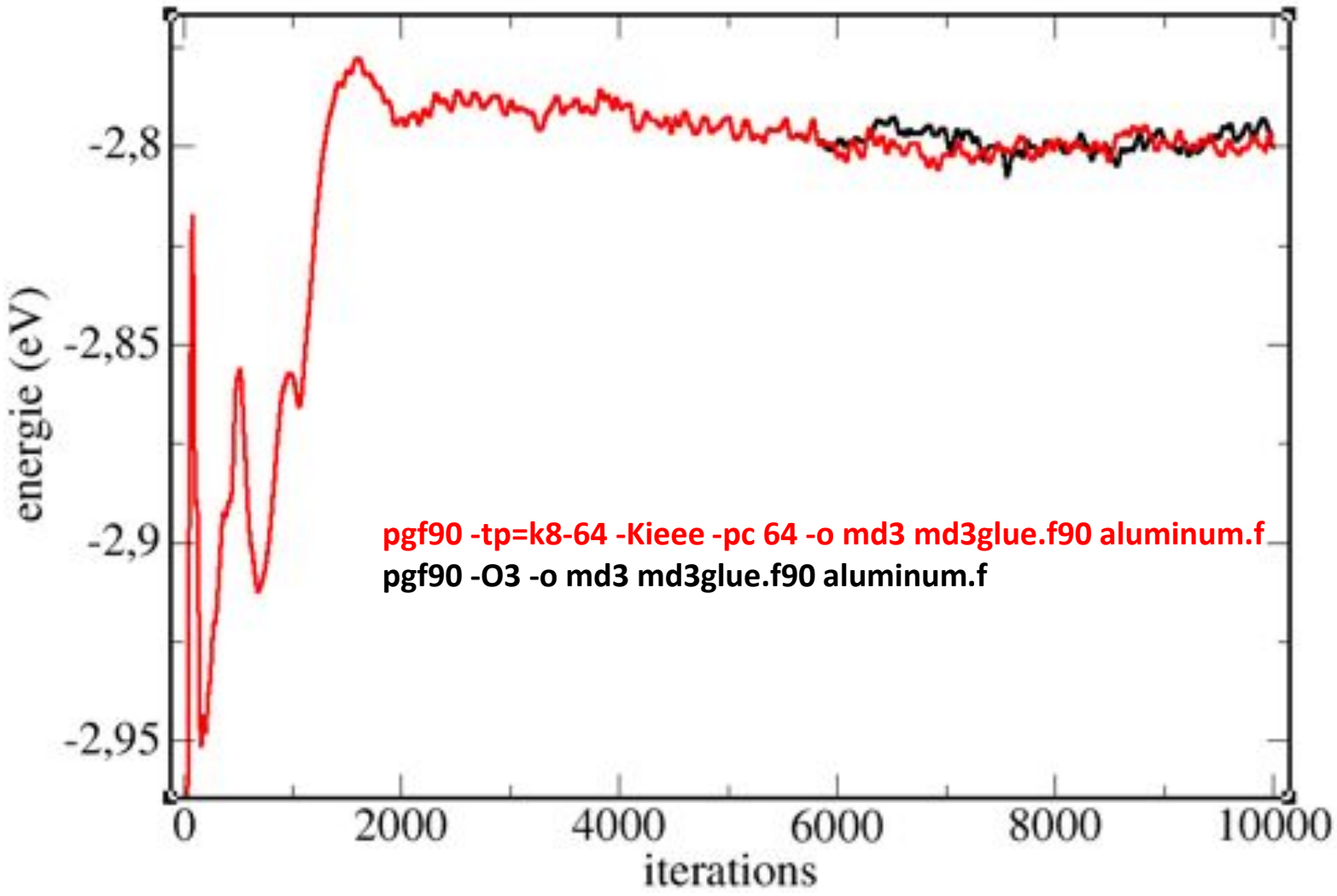
  ✓ Erreur sur Energie moyenne du système

  ✓ Paranoia : codes C et fortran pour l'arithmétique des nombres flottants.
  http://www.netlib.org/paranoia/

- Référence :

  ✓ pgf90 -tp=k8-64 -Kieee -pc 64 -o md3 md3glue.f90 aluminum.f

Comparaison d'options Portland (pgf90)



Compilateur Portland group

pgf90 7.2-2 64-bit target on x86-64 Linux -tp penryn-64

pgf90	t/t_{ref}	$\Delta_{max}(\%)$	Paranoia
-tp=k8-64 -Kieee -pc 64	1.00	0.0	1 alerte
-tp=x64 -Kieee -pc 64	0.96	0.0	1 alerte
-tp=x64 -Kieee -pc 64 -Mbounds	2.36	0.0	1 alerte
-O0	0.98	0.4	1 alerte
-O1	0.94	0.4	1 alerte
-O2	0.39	0.4	1 alerte
-O3	0.39	0.4	1 alerte
-O3 -Kieee -pc 64	0.39	0.0	1 alerte
-fast -Mipa=fast	0.24	0.3	1 alerte
-fast -Mipa=fast -Kieee -pc 64	0.24	0.3	1 alerte
-fast -Mvect -Kieee -pc 64	0.24	0.3	1 alerte

Intel(R) Fortran Intel(R) 64 Compiler Professional for applications running on Intel(R) 64, Version 11.1

ifort	t/t_{ref}	Δ_{max}(%)	Paranoia (Failure, Serious, Defect, Flaw)
-fp-model precise -pc 64	0.38	0.0	
-fp-model precise -pc 64 -check bounds	0.70	0.0	
-O0	1.03	0.0	
-O1	0.29	0.0	3, 7, 3,1
-O2	0.28	0.4	4, 7, 4, 2
-O3	0.26	0.4	4, 7, 4, 2
-O3 -fp-model precise -pc 64	0.26	0.0	
-fast	0.25	0.5	4, 7, 3, 2
-fp-model precise -pc 64 -fast	0.26	0.3	0, 0, 0, 1
-O3 -fltconsistency -mieee-fp -fp_port -pc 64	0.56	0.3	0, 0, 0, 1

-O0 ne désactive pas toutes les optimisations !

Compilateur GNU : gfortran v. 4.1.2

gfortran	t/t_{ref}	$\Delta_{max}(\%)$	Paranoia
-mieee-fp	1.02	0.0	
-O0	1.02	0.0	
-O1	0.29	0.0	
-O2	0.26	0.4	
-O3	0.23	0.4	
-O3 -mieee-fp	0.23	0.0	
-fast-math	1.02	0.5	1 alerte

Conclusions

- Sur cet exemple précis, les 3 compilateurs présentent des performances et résultats similaires en utilisant le bon jeu d'option (ifort user manual =3824 pages !!!).
- Par défaut ifort privilégie la rapidité au détriment de la précision. gfortran et pgf90 ont une approche plus conservatrice.
- Voir les options utilisées dans le specbench
- Avoir un regard critique sur ses résultats. Résultats dépendants des applications => nombreux tests indispensables

Recommandations :

- Approche conservative et se méfier des options par défaut
- Respecter la norme ieee

REMERCIEMENTS

- Soutien financier du CS (BQR), de l'uB et de la Région Bourgogne (HCP/CP) et de FEDER

- Soutien financier (pour le nouveau cluster) des laboratoires ICB, ICMUB, CRC.
- Nos collègues du CCUB
- Serviware