Introduction aux architectures parallèles et au supercalculateur Occigen

CINES/DCI - 30 janvier 2015

Programme

Déroulement de la journée de formation

- 9h30 - 12h00 : cours / TP
- 12h00 - 14h00 : repas
- 14h00 - 14h30 : cours
- 14h30 - 15h00 : visite de la salle machine
- 15h00 - 17h00 : cours / TP / QR

Introduction aux architectures parallèles et Occigen

Points Abordés

- Contexte HPC
- Architectures Matérielles
- Environnement logiciel du HPC
- Les moyens de calcul du CINES
- Occigen

HPC

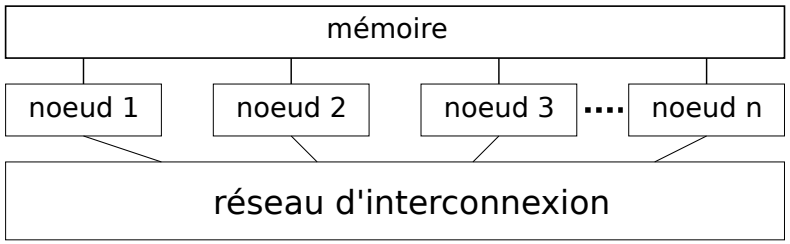
Définitions

- Agrégation de moyens de calcul permettant un gain en temps de restitution, en capacité de calcul et en disponibilité
- Mutualisation des infrastructures (bâtiments, énergie : climatisation, alimentation électrique)
- Simulation de phénomènes physiques, chimiques, ...etc
- De nombreux objectifs :
  - Une issue aux expériences
  - Comparaison avec des expériences, des observations
  - Outil de dimensionnement, d'investigation
  - ...

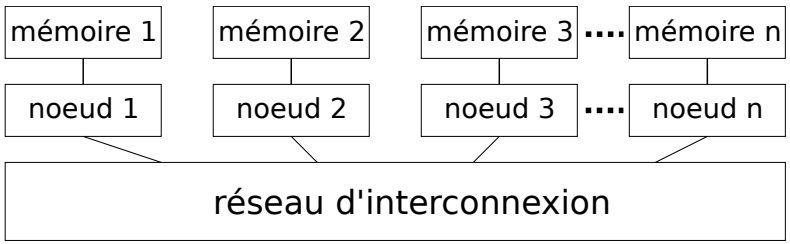
Architectures Matérielles

Différents types de machines parallèles

- Machines vectorielles (NEC)
- Machines scalaires
  - mémoire partagée (ex : UV)



  - mémoire distribuée (ex : BlueGene, Altix, XT, ...)



Architectures Matérielles

Glossaire

- cœur : unité de calcul
- processeur : composé des cœurs avec cache L1-2-3
- nœud, lame : ensemble de processeurs avec mémoire
- rack : ensemble de noeuds reliés par un un réseau d'interconnexion

Architectures Matérielles

Les différents éléments d'une machine parallèle : Processeurs / CPU

- UAL : Unité arithmétique et logique $\rightarrow + - */$ et $<=$)
- Registre : stocke les opérandes et résultats intermédiaires de calcul et les informations sur l'état du calcul (ex : numéro de l'itération)
- Mémoire cache : accès rapide aux données, mais capacité limitée
  - Cache L1 : le plus petit. Partitionné en 2 parties : une pour les instructions et une pour les données
  - Cache L2 et L3 (ou LLC) : données

Architectures Matérielles

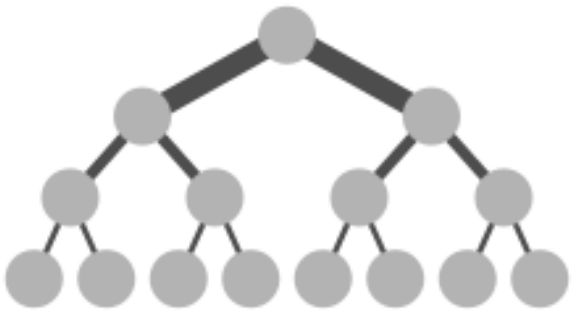
Les différents éléments d'une machine parallèle : Mémoire

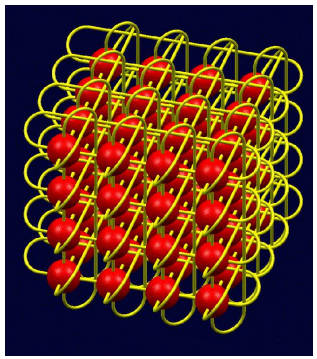
- composée de condensateurs prenant comme valeur 1 ou 0 et correspondant à un bit
- Mémoire statique et dynamique (DRAM mémoire centrale - SRAM cache)
- DR RAM - DDR SDRAM - DDR2/3/4 SDRAM
- débit d'information en GB/s
- sous forme de barettes mémoire (DIMM)
- latence (ex : 40 ns)

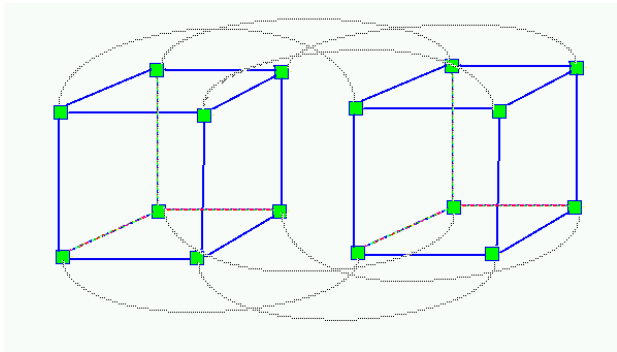
Architectures Matérielles

Les différents éléments d'une machine parallèle : Réseau

- caractérisé par :
  - sa bande passante : en Gbits/s, uni ou bi-directionnel (Ex : 56 Gb/s (IB 4x FDR), ...)
  - sa latence : temps en s de la transmission d'une donnée (Ex : 1.5 μs,...)
  - sa topologie : arborescence du réseau entre les nœuds ou racks (Ex : fat tree, tore 3d, hypercube, ...)



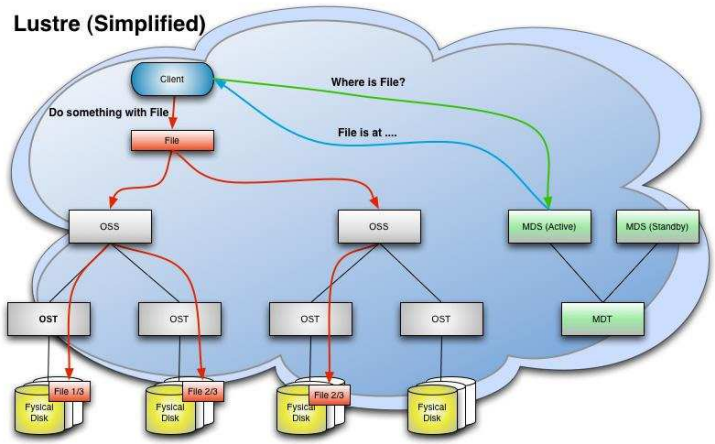




Architectures Matérielles

Les différents éléments d'une machine parallèle : Systèmes de fichiers

- séquentiel
  - NFS
- parallèle
  - Pour les I/O
  - Séparation des données et des méta-données
  - Aspect quantitatif : en nombre d'OST et d'OSS
  - Exemples :
    - Lustre
    - GPFS
    - ...



Architectures Matérielles

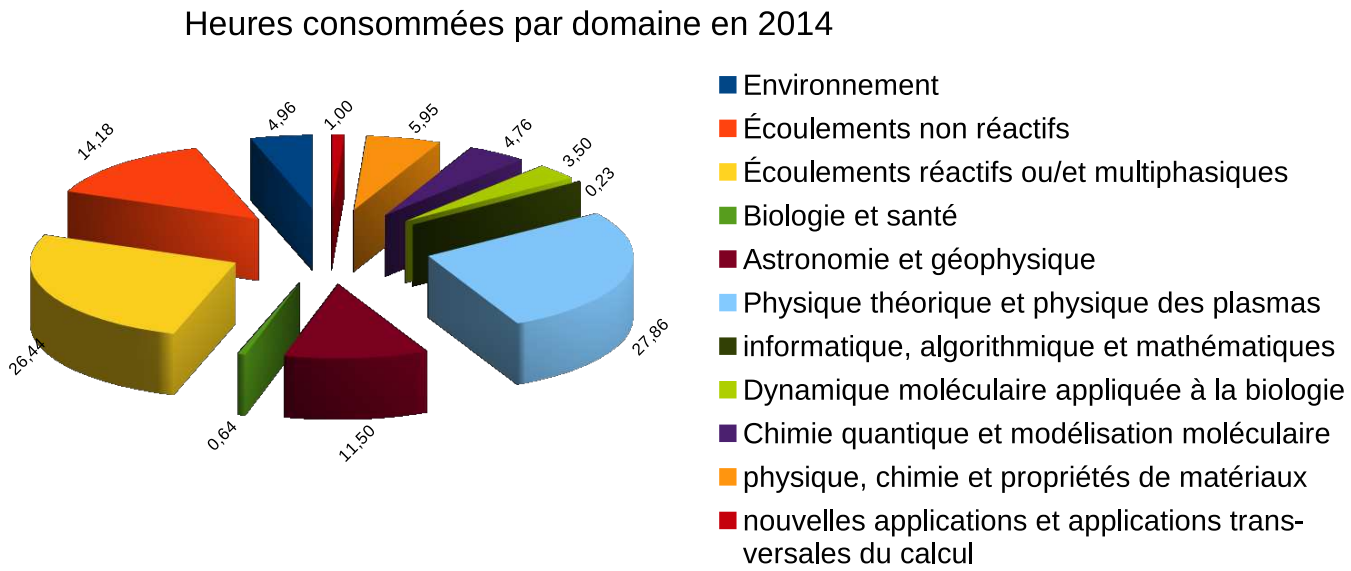
Les différents éléments d'une machine parallèle : Informations disponibles

- Informations concernant l'architecture matérielle de la machine dans : /proc
  - CPU : /proc/cpuinfo : fréquence, type, taille du cache, ...
  - mémoire : /proc/meminfo : totale, swap, ...
  - FS : /proc/fs : type, version, ...

Environnement logiciel HPC

Codes de calcul

Simulation numérique dans la plupart des domaines scientifiques



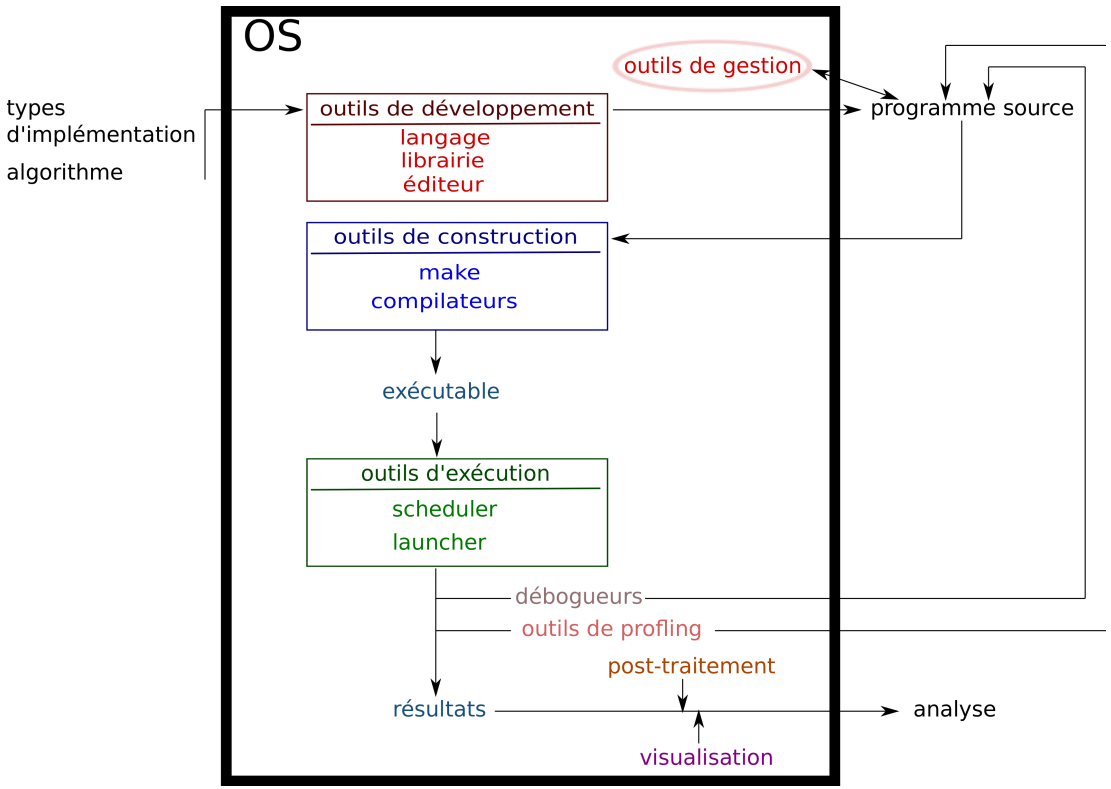
Environnement logiciel HPC

Environnement logiciel : codes de calcul disponibles

- Domaines scientifiques variés : dynamique moléculaire, chimie, biologie, mécanique des fluides, ...etc
- Exemples :
  - Namd
  - Abinit
  - Gromacs
  - Vasp
  - Fluent
  - OpenFoam
  - ...

Environnement logiciel HPC

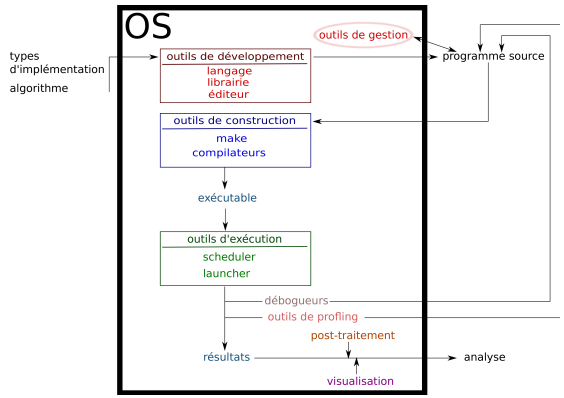
Environnement des codes de calcul



Environnement logiciel HPC

Environnement système

- Système d'exploitation (OS) : SLES, Bullx SCS, CENTOS, ...
- Environnement shell : bash, csh, tcsh, ksh
- Outil de script : shell, python, perl, ruby
- Manipulation de fichiers et de chaînes de caractères : sed, cat, head, awk
- Chargement d'environnement logiciel : module



Environnement logiciel HPC

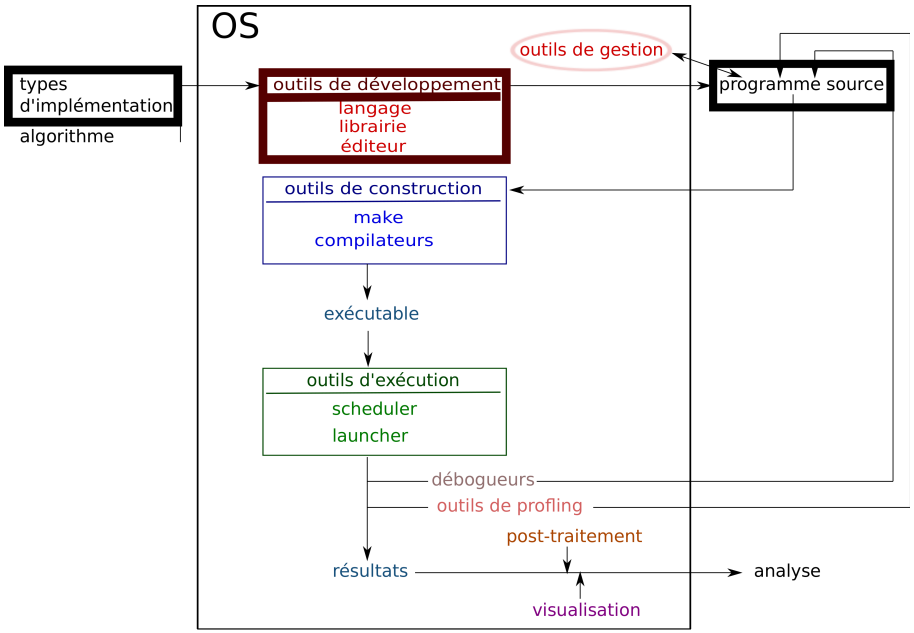
Types de licences utilisées par les codes de calcul

- commerciale
  - jeton : flottant ($\Rightarrow$ serveur de licences, ex : abaqus, fluent), relatif à une machine
  - globale : relative à un groupe de personnes (ex : vasp, adf, wien2k)
- publique
  - GPL, LGPL, GFDL
  - exemple GPL : gromacs, abinit

Environnement logiciel HPC

Les outils de développement des codes de calcul

- Les langages de programmation
  - C/C++, fortran
  - python
  - OpenCL, Cuda
  - coarrayfortran, chapel
- Types d'implémentation
  - Séquentiel
  - OpenMP
  - MPI : SPMD, ...
  - Accélérateurs
  - Hybride : MPI/OpenMP, MPI/GPU
- Éditeurs de texte : emacs, nedit, vim



Environnement logiciel HPC

Environnement logiciel : compilation

- Compilateurs
  - intel : icpc, icc, ifort
  - gnu : g++, gcc, gfortran
  - version par commande -v (ex : icc -v)
  - informations sur les options : man (ex : man icc)
- Wrappers MPI
  - BullxMPI : mpiCC, mpicc, mpif90
  - IntelMPI : mpicc, mpiicpc, mpiifort (intel) ; mpigxx, mpigcc, mpif90 (gnu)
  - OpenMPI : mpic++, mpicc, mpif90 (intel)
- OpenMP
  - intel : -openmp
  - gnu : -fopenmp

Environnement logiciel HPC

Environnement logiciel : compilation

- Exemples de commande
  - Série
    - création des fichiers objets : icc -c hello.c
    - création de l'exécutable : icc -o hello hello.o
  - OpenMP
    - icc -openmp -c hello_omp.c
    - icc -openmp -o hello_omp hello_omp.o
  - MPI
    - mpicc -c hello_mpi.c
    - mpicc -o hello_mpt hello_mpi.o
  - Hybride : MPI/OpenMP
    - mpicc -openmp -c hello_hyb.c
    - mpicc -openmp -o hello_hyb hello_hyb.o

Environnement logiciel HPC

Environnement logiciel : compilation

- Besoins : configuration, automatisation
  - Création de librairies
  - Dépendance entre les fonctions et fichiers
  - Création de plusieurs exécutables
- Outils
  - Configure
    - permet de tester la présence de fichiers ou librairies
    - permet de choisir les librairies utilisées par l'exécutable (-h)
    - créer les makefile
    - une fois la commande configure exécutée : make
    - exemple

```
./configure --with-fft=mkl --enable-shared --disable-static --enable-mpi\
--enable-double --prefix=/........?.......... CC=mpicc\
CFLAGS="-O3 ..?.." F77=mpif90 LIBS="-mkl" --program-suffix=_d
```

  - cmake
    - même principe que configure
    - mode graphique : commande ccmake

Environnement logiciel HPC

Environnement logiciel : compilation

- Outils
  - Makefile

```
FC      = mpif90
FFLAGS  = -O3
INCS    = -I......?......
LD      = ${FC}
LDFLAGS = ${FFLAGS}
LIBS    = -L/.....................?.............\
          -lmkl_intel_lp64 -lmkl_sequential -lmkl_core
OBJECTS  = distmod.o numerics.o mod1i.o  dddot1.o dddot2.o dddot3.o\
           csetup.o  state.o
EXEC    = x.mod1i
${EXEC}   : ${OBJECTS}
            ${LD} ${LDFLAGS} -o $@ ${OBJECTS} ${LIBS}
.SUFFIXES : .f .o
.f.o      :
            ${FC} ${FFLAGS} ${INCS} -c $<
clean     :
            @rm -f ${OBJECTS} *.mod ${EXEC}
```

  - installation et contrôle de l'installation : make install, make check
- Outils d'aide à la résolution d'erreurs d'installation, de compilation
  - Dépendances et fichiers include : -I $\Rightarrow$ find, grep
  - Link : -L $\Rightarrow$ nm, find, grep

Occigen

Environnement logiciel : développement

- Outils existants disponibles
  - Librairies MPI : bullxmpi (BULL), IntelMPI
  - différences dans les performances (différents algorithmes)
- Librairies mathématiques
  - MKL (blas, lapack, scalapack, fftw)
  - FFTW3
  - Numpy, Scipy
  - PETSC
- Librairies I/O
  - Netcdf
  - HDF5
- Outils de gestion de versions de codes (Filtrage IP) : git, svn

Occigen

Environnement logiciel : compilation

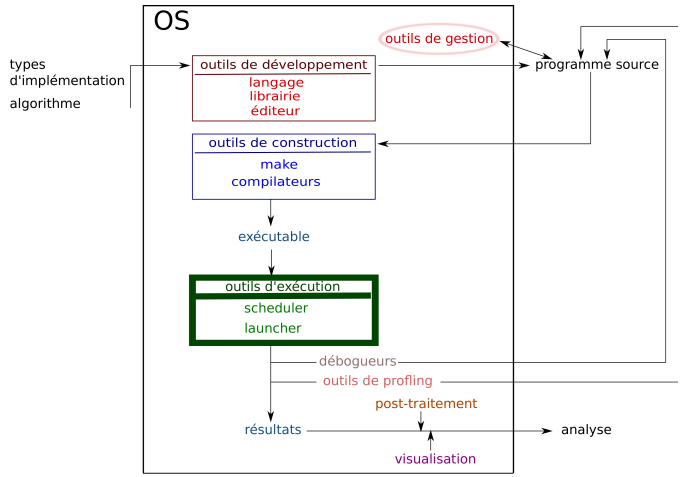
- Module : permet l'utilisation des librairies et logiciels installés sous /opt/software
  - module avail : affiche tous les modules disponibles sur Occigen
  - module load : définit les variables d'environnement d'une librairie ou d'un code nécessaire à son utilisation
  - module rm : décharge un module
  - module purge : décharge tous les modules
  - module list : affiche les modules qui sont chargés dans votre environnement
  - module show : affiche :
    - les variables d'environnement d'une librairie ou d'un code
    - les modules prérequis, les conflits
    - les chemins vers les librairies et binaires

Environnement logiciel HPC

L'environnement d'exécution

- Le gestionnaire de travaux : SLURM
- Paramètres de lancement de travaux :
  - temps
  - nombre de nœuds
  - nombre de cœurs par nœud ou total
  - nombre de processus MPI
  - nombre de threads OpenMP
  - quantité de mémoire
  - nom du job
  - nom des fichiers de sorties et chemin
  - email

```
#!/bin/bash
#SBATCH -o /home/.../out.hello.%j.%N
#SBATCH -D /home/.../test
#SBATCH -J test_hello
#SBATCH --ntasks=32
#SBATCH --nodes=2
#SBATCH --time=08:00:00
...
```



Occigen

Environnement logiciel : exécution

- Le gestionnaire de travaux SLURM
  - Paramètres principaux du script SLURM : nombre de nœuds, de processus MPI et temps

```
#!/bin/bash
#SBATCH -J job_name
#SBATCH --nodes=2
#SBATCH --ntasks=48
#SBATCH --ntasks-per-node=24
#SBATCH --threads-per-core=1
#SBATCH --time=00:01:00
module purge
module load intel/15.0.0.090 bullxmpi/1.2.8.3
srun  --mpi=pmi2 -K1 --resv-ports \
-n $SLURM_NTASKS ./mon_prog-mpi
```

Occigen

Environnement logiciel : exécution

- Le gestionnaire de travaux SLURM
  - Jobs en modes dépeuplé et hybride

```
#!/bin/bash
#SBATCH --nodes=2
#SBATCH --ntasks=24
#SBATCH --ntasks-per-node=12
#SBATCH --threads-per-core=1
#SBATCH --mail-type=end
#SBATCH --mail-user=name@server
#SBATCH -J thello
#SBATCH --time=01:00:00
module purge
module load intel/15.0.0.090 bullxmpi/1.2.8.3
srun --mpi=pmi2 -K1 --resv-ports \
-n $SLURM_NTASKS abinit < input.abinit.files
```

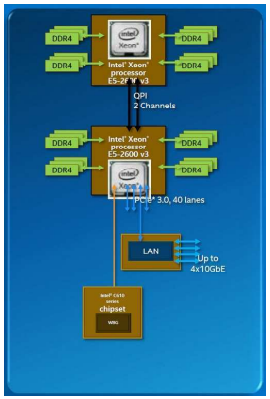
```
#!/bin/bash
#SBATCH --nodes=2
#SBATCH --ntasks=8
#SBATCH --ntasks-per-node=4
#SBATCH --threads-per-core=1
#SBATCH --cpus-per-task=6
#SBATCH -J thello
#SBATCH --time=01:00:00
module purge
module load intel/15.0.0.090 bullxmpi/1.2.8.3
export OMP_NUM_THREADS=6
export KMP_AFFINITY=granularity=fine,compact,1,0
srun --mpi=pmi2 -K1 -m block:block \
-c 6 --resv-ports -n $SLURM_NTASKS ./mpi_omp_prog
```

Occigen

L'environnement d'exécution

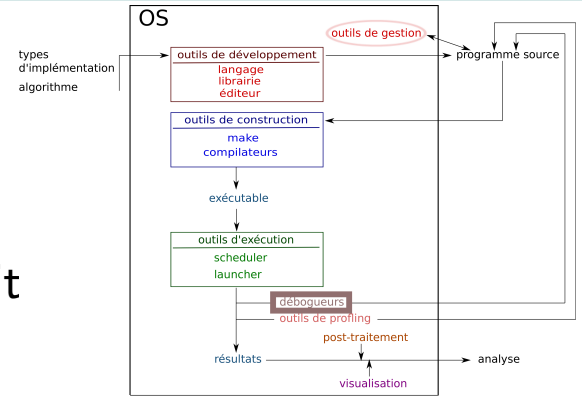
- Les commandes de lancement
  - OpenMP :
    export OMP_NUM_THREADS=24
    export KMP_AFFINITY=compact,1,0
    ./exe
  - BullxMPI :
    srun –mpi=pmi2 -K1 –resv-ports -n $SLURM_NTASKS ./exe

core 01	core ..	core 06
L1C	L1C	L1C
L2C	L2C	L2C
L3 Cache		
L2C	L2C	L2C
L1C	L1C	L1C
core 07	core ..	core 12



Occigen

Les outils de débogage

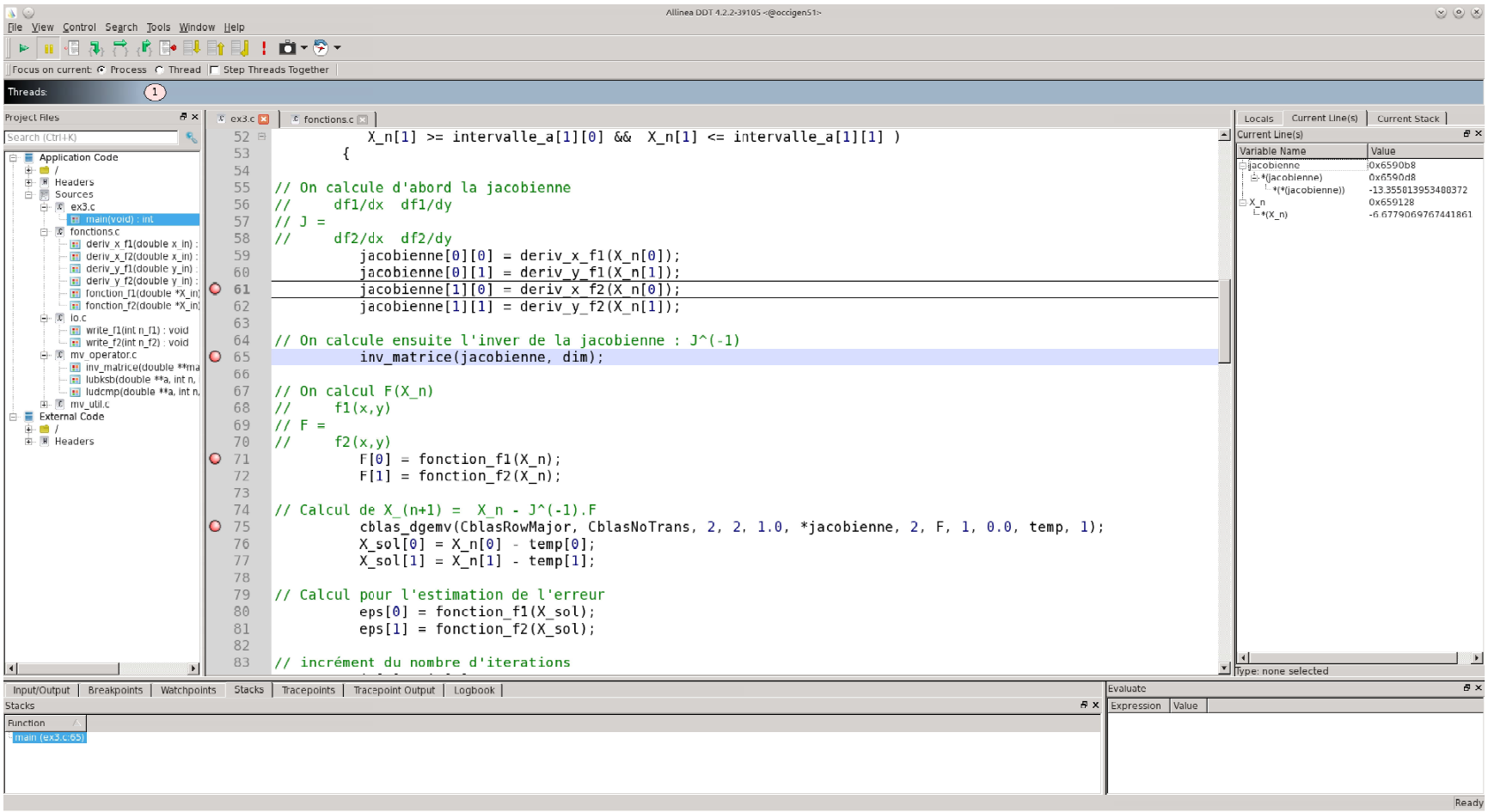


- Les flags de compilation
  - intel
    - fortran : -g -O0 -check all -traceback
    - C/C++ : -g -O0 -traceback -check-uninit
  - gnu
    - fortran : -g -O0 -fbacktrace -fbounds-check -ffpe-trap=zero,underflow,overflow,invalid
    - C/C++ : -g -Wuninitialized -O -fbounds-check -ftrapv
- Suivi des jobs sur les nœuds de calcul : top, strace -p, gstack
- Logiciels :
  - libre : gdb, valgrind
  - Bull : padb
  - commercial : ddt, totalview

Occigen

Les outils de débogage

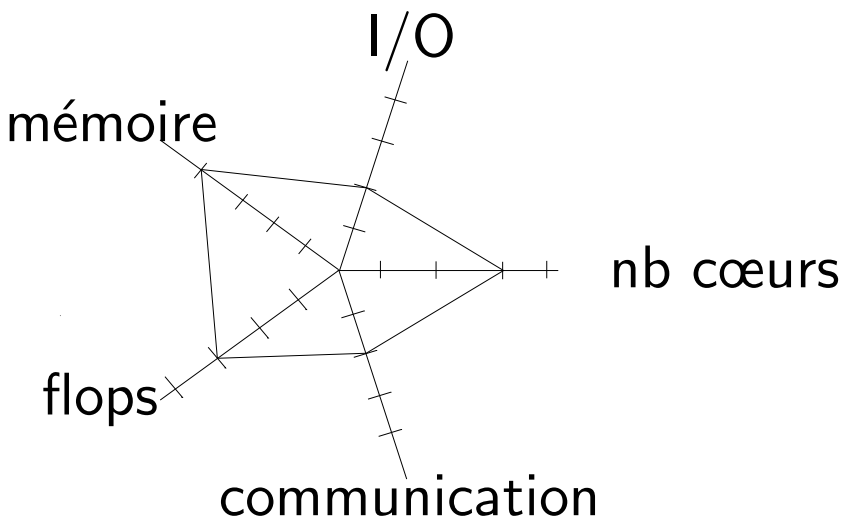
- DDT

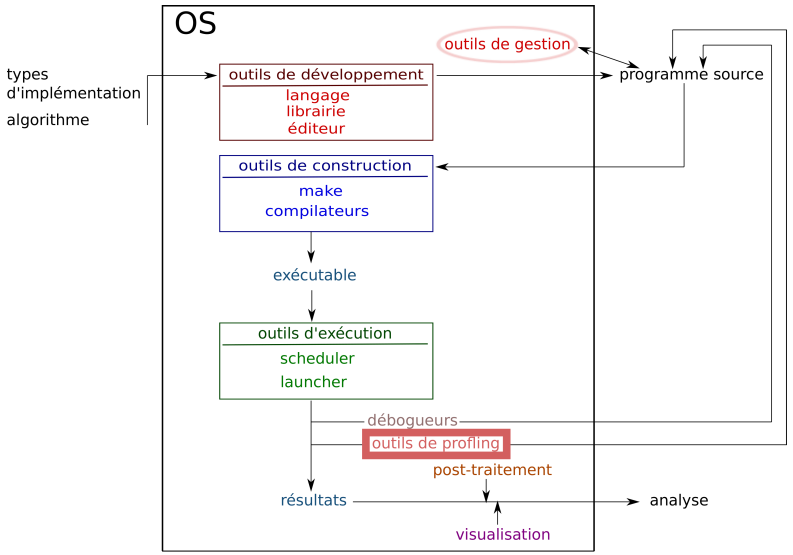


Occigen

Les outils d'analyse de performance, d'optimisation

- Domaines d'optimisation



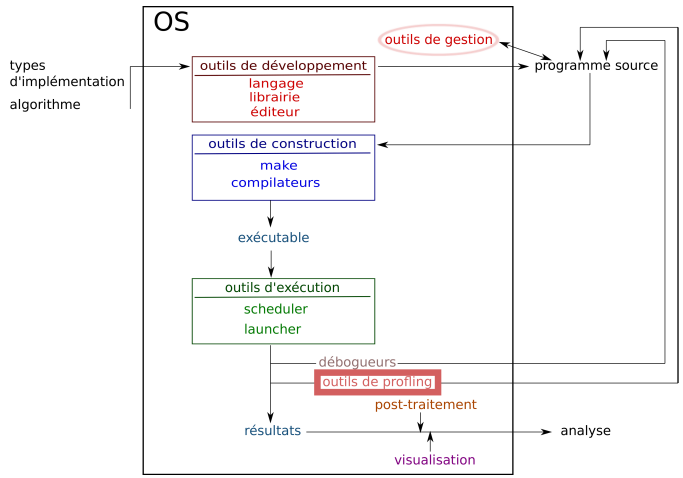


- Courbes de scaling :
  - weak : taille du problème varie en fonction du nombre de cœurs
  - strong : taille fixe du problème

Occigen

Les outils d'analyse de performance, d'optimisation : par type d'analyse

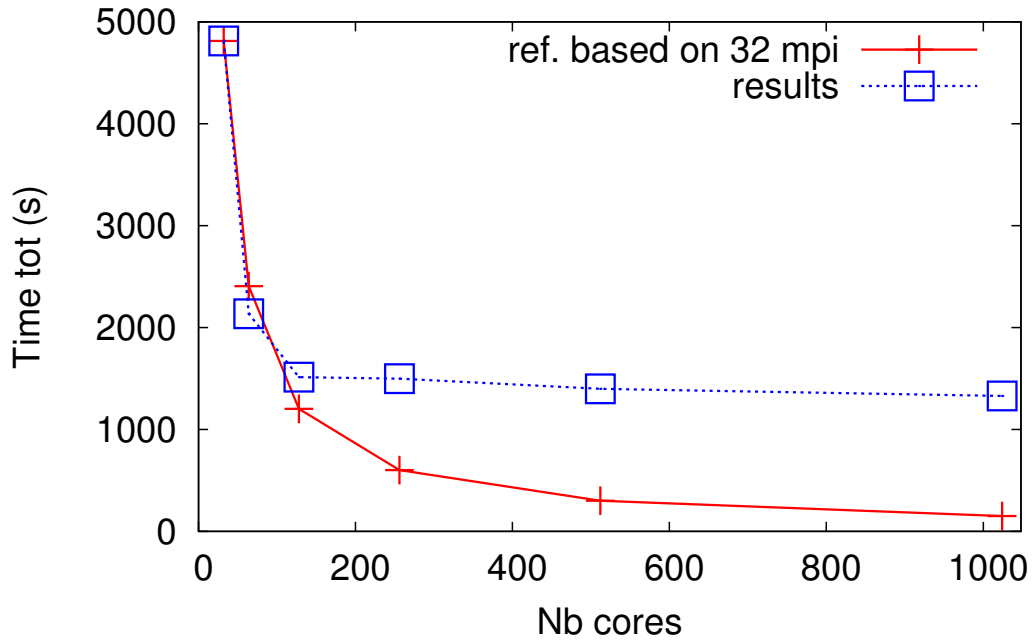
- Flops, CPU : compteurs PAPI, kCachegrind
- Mémoire : valgrind, collectl
- I/O : collectl, darshan
- Communication : Vampir, ITAC
- Temps : score-p, TAU, gprof (-pg)



Occigen

Les outils d'analyse de performance, d'optimisation

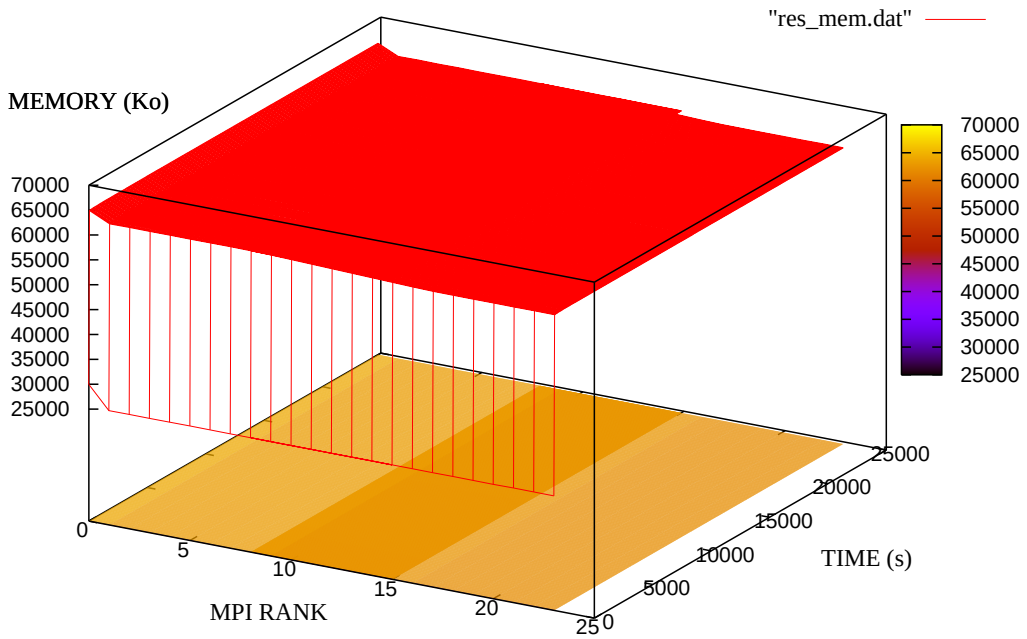
- Options de compilation : -xCORE-AVX2
- Courbes de scaling



Occigen

Les outils d'analyse de performance, d'optimisation

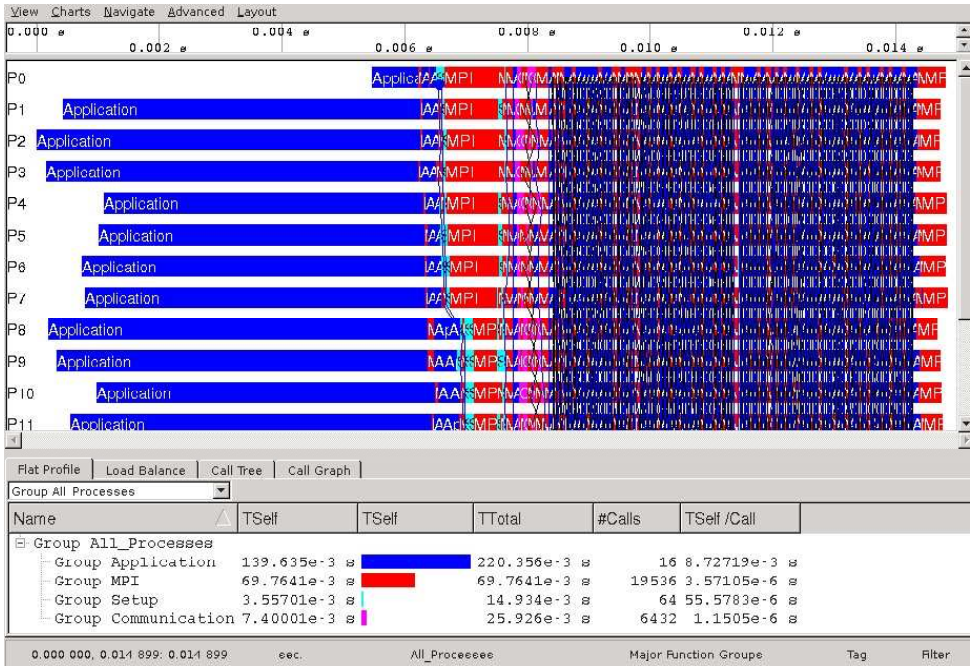
- Flops, CPU : compteurs PAPI, bullxprof
- Mémoire : valgrind

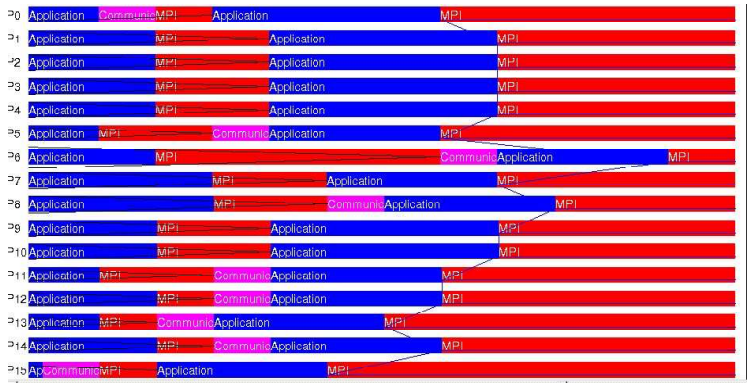


Occigen

Les outils d'analyse de performance, d'optimisation

- I/O : bullxprof, IOtop, Darshan
- Communication : ITAC, bullxProf, ipm

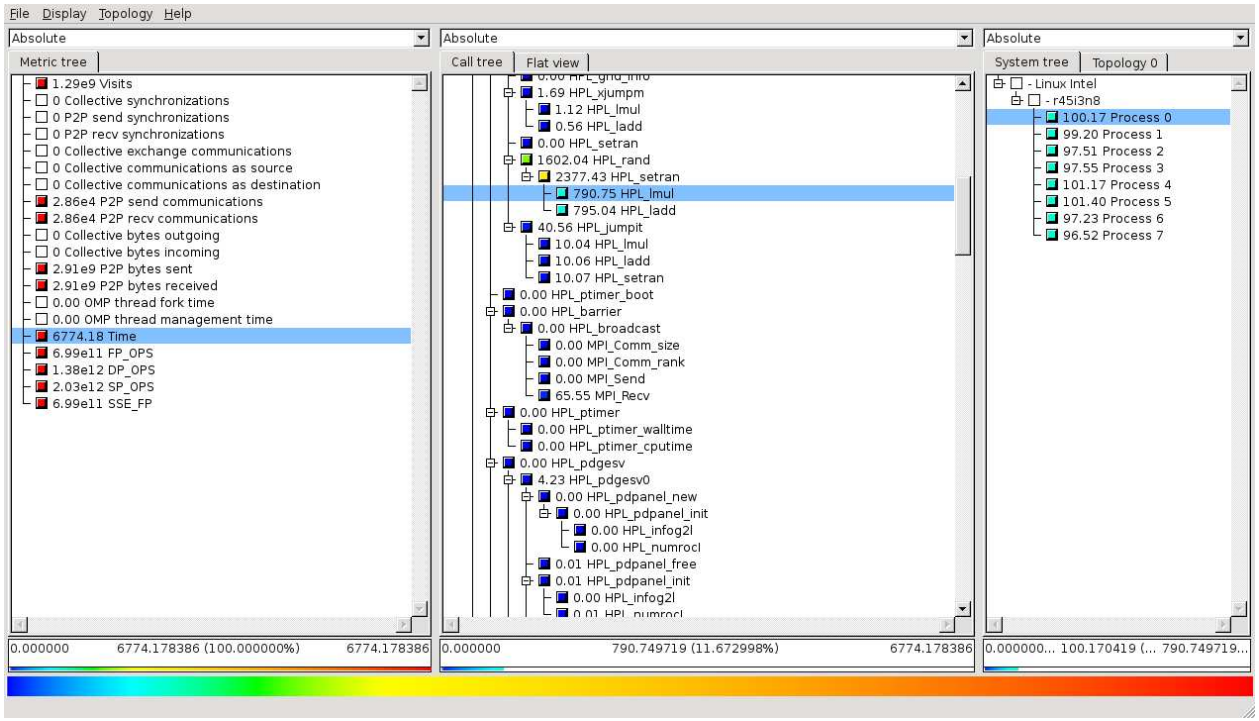




Occigen

Les outils d'analyse de performance, d'optimisation

- Temps : score-p, bullxprof, gprof (-pg)



Occigen

Sources

- Introduction à Jade (2013)
- cines.fr
- edari.fr
- ark.intel.com
- bullxDEUser'sGuide
- computing.llnl.gov/linux/slurm
- software.intel.com
- valgrind.org
- vi-hps.org
- workshop Bull Nov. 2014
  P. Girard, C. Mazauric
- icl.cs.utk.edu/papi