Renouvellement des ressources

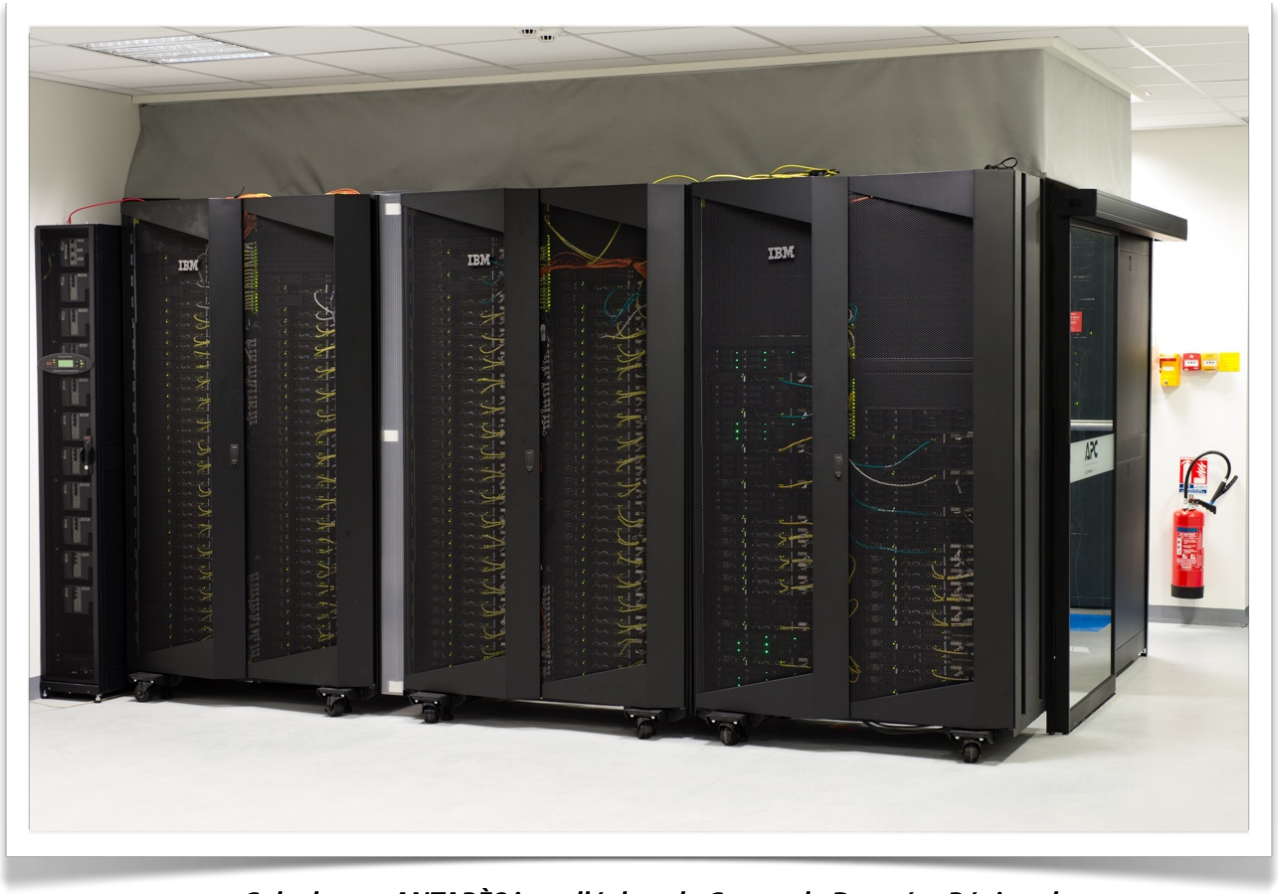
Patrick BOUSQUET-MÉLOU (pbm@crihan.fr)

16 Juin 2015

Ressources actuelles

Ressources actuelles

IBM iDataPlex «ANTARÈS»



Calculateur ANTARÈS installé dans le Centre de Données Régional

Ressources actuelles (2)

IBM iDataPlex «ANTARÈS»

ANTARÈS
45 TFlops CPU
13 TFlops GPU - 2 TFlops Phi
3692 cœurs CPU et 13,4 To de RAM

Nœuds de calcul
Intel NEHALEM, WESTMERE et IVY BRIDGE

158 nœuds 8 cœurs / 24 Go RAM
141 nœuds 12 cœurs / 48 Go RAM
3 nœuds 12 cœurs / 96 Go RAM
12 nœuds spécialisés I/O
18 nœuds 20 cœurs / 64 Go RAM
13 nœuds hybrides CPU / GPU NVIDIA M2050
1 nœud hybride CPU / co-processeur Xeon Phi

Interconnexion InfiniBand
280 To de disques partagés

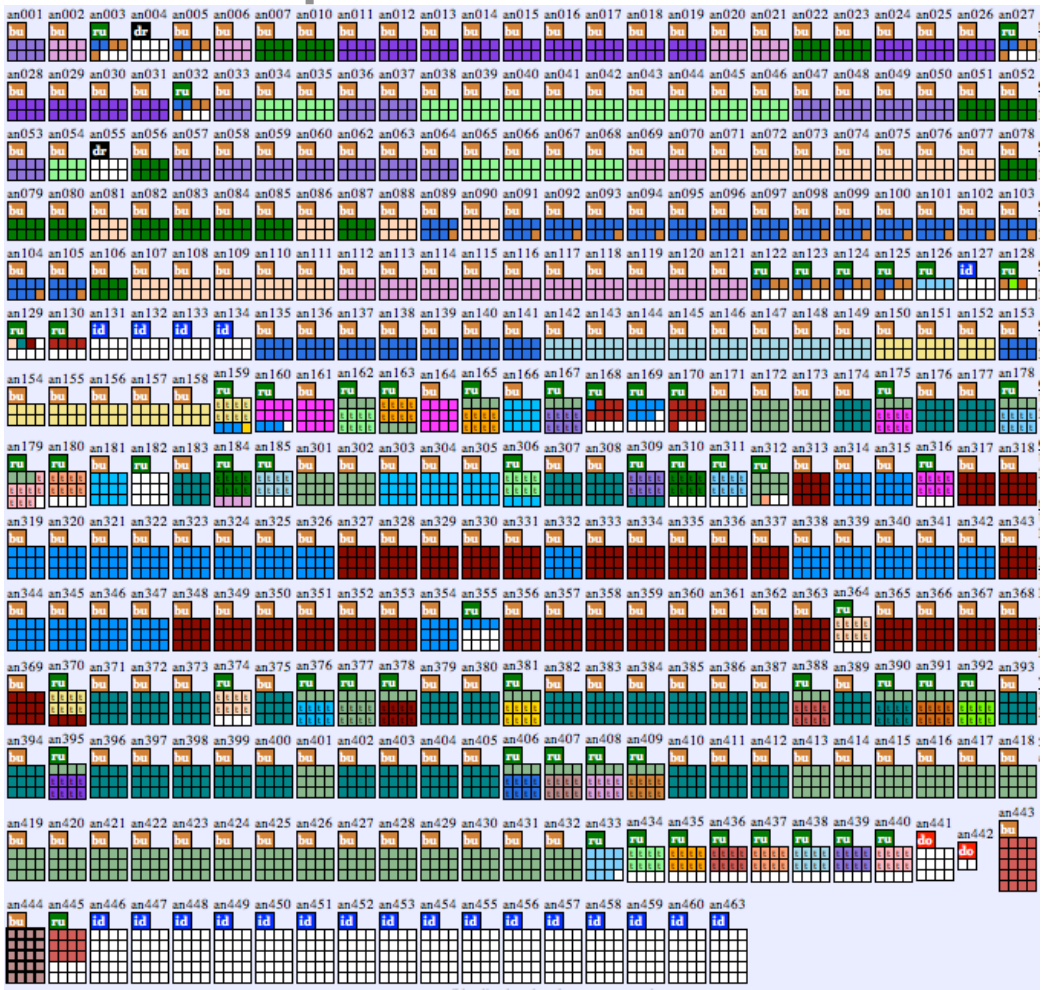
Connexion 2 x 10 Gbit/s
sur SYRHANO

- 3692 cœurs
  - 3048 CRIHAN, 624 ECN
- 347 nœuds de calcul
  - dont 1 serveur de visualisation
- 13,4 To de mémoire (DDR3)
- 280 To de disques partagés
- Système de fichiers GPFS
- Batch LoadLeveler
- Connexion sur SYRHANO
  - 2 x 10 Gbit/s

Plates-formes envisagées en 2016

Grappe de calcul actuelle (ANTARÈS)

Variété de profils des travaux / de nœuds de calcul



125 nœuds «Nehalem» (8 cœurs, 24 Go RAM)

Calculs MPI «standards» (≤ 3 Go de mémoire / cœur) **de taille petite ou moyenne**
≤ 128 cœurs sur 24 heures
ou 256 cœurs sur 12 heures

CFD, dynamique moléculaire, physique des matériaux, etc.

169 nœuds «Westmere» (12 cœurs, 48 ou 96 Go RAM)

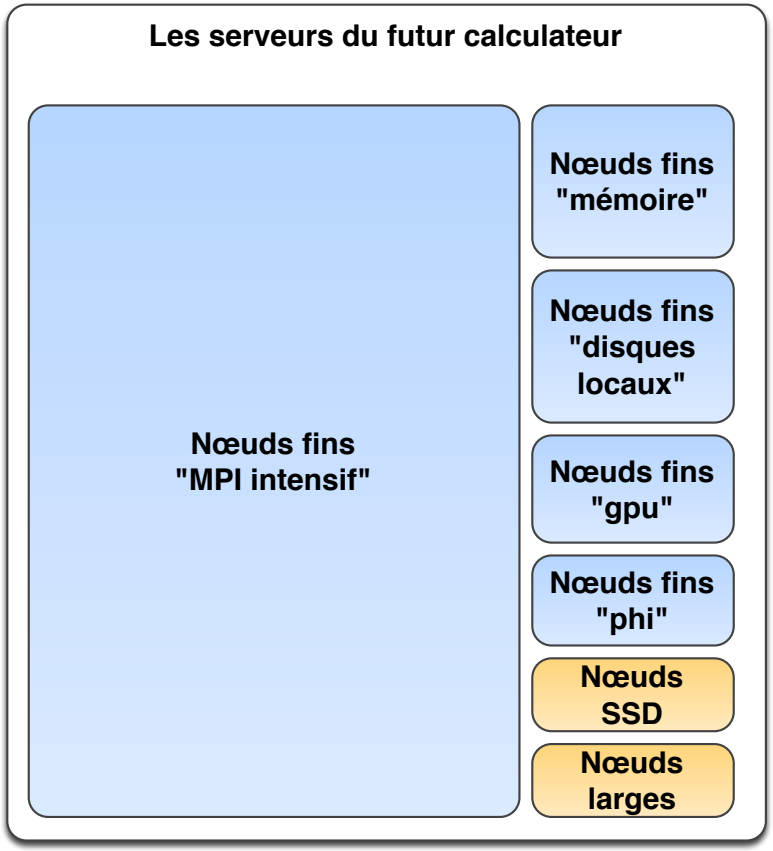
12 nœuds «disque» : travaux intancœud de chimie quantique (*GAUSSIAN*, *GAMESS*, *JAGUAR*)

3 nœuds «mémoire» (96 Go RAM / nœud) : travaux de chimie du solide (*ABINIT*) : MPI 20 cœurs avec 8 Go / cœur

13 nœuds GPGPU

141 nœuds standards (4 Go RAM / cœur) : travaux MPI «standards» **(4 Go RAM / cœur) de grande taille (≤ 1032 cœurs) sur 24 heures, ou de taille petite ou moyenne (≤ 144 cœurs) sur 48 heures**, de CFD, dynamique moléculaire, physique des matériaux, etc.

CRIHAN



Serveurs envisagés

Serveurs thématiques

- Sous-ensemble de grande taille pour des codes MPI fortement parallèles
- Sous-ensembles spécifiques de plus petite taille
  - Codes nécessitant beaucoup de mémoire par cœur
  - Codes générant beaucoup d'I/O
  - Codes exploitants des GPU
  - Plate-forme d'expérimentation Xeon Phi (KNL)

Grappe de calcul

Architecture distribuée

- Nœuds de calcul fins
  - Standards pour besoins «MPI» (2 à 4 Go de mémoire / cœur) : CFD, dynamique moléculaire, physique des matériaux, etc.
  - Certains gonflés en mémoire (512 Go par nœud) : chimie du solide, mécanique des structures, pré - / post - traitements séquentiels, etc.
  - D'autres à disques internes capacitifs et performants : chimie quantique (*GAUSSIAN 03*, *GAMESS*)

Grappe de calcul (2)

Architecture distribuée (2)

- Nœuds de calcul fins (suite)
  - D'autres dotés de GPUs
    - GPU NIVIDIA Maxwell (2014) et/ou Pascal (2016)
      - Au moins 12 Go de mémoire par carte GPU (vs. 3 Go dans les cartes Fermi d'Antares actuellement)
      - Mêmes GPUs adaptés au calcul (GPGPU) et à la visualisation
  - D'autres dotés de Xeon Phi (KNL)

  - Dimensionnement ciblé : ~ 12 000 cœurs, ~ 500 nœuds, ~ 500 TFlop/s
  - Processeur Haswell ou Broadwell
    - 12 à 16 cœurs @ 2,3 à 2,6 GHz

Serveurs spécifiques

Autres architectures

- Nœuds de calcul larges (≥ quadri-processeurs)
  - ≥ 1 To de mémoire et > 32 cœurs
  - Usage
    - MPI avec beaucoup de mémoire par processus
      - Mécanique des structures
    - Codes OpenMP efficaces sur un nombre de threads > 30
    - Pré - / post - traitements séquentiels ou peu parallèles demandant beaucoup de mémoire par processus

Soumission des travaux

Evolution

- Batch
  - Quel logiciel ?
    - LoadLeveler (IBM) est en fin de vie
    - Slurm, LSF ou PBS Pro font partie des logiciels candidats
  - Quelles améliorations par rapport au batch actuel du CRIHAN ?
    - Notion de travail multi-étapes (1 seul script pour lancer des étapes de profils différents et ayant des dépendances ; exemple : pré - traitement séquentiel avec beaucoup de mémoire, puis solveur MPI prenant peu de mémoire par cœur)

Interface Web

Portail unifié

- Interface Web pour
  - la soumission des travaux
  - le transfert de données
  - la visualisation graphique à distance

Stockage

Actullement

- Stockage «rapide» (/home + scratch, GPFS)
  - 180 To
  - 4,5 Go/s de débit agrégé d'entrées-sorties

- Stockage «moyen terme» non intégré au service HPC
  - Données non accessibles par le service de visualisation des données
    - Service stockage.syrhano.net (http://www.crihan.fr/services/stockage/)
      - environ 500 To (environ 160 To utilisés par des données «calcul»)

Stockage

Evolution

- Stockage attaché au calculateur : 1 Po envisagé
  - /home
  - Scratch
    - /work avec espaces personnels
    - /scratch pour répertoires temporaires de calcul
      - ≥ 15 Go/s de débit agrégé d'entrées-sorties
- Stockage «moyen terme» intégré au service HPC : 1 Po envisagé
  - données visibles à partir des frontales et du service de visualisation
- Service de migration des données (gestion du cycle de vie)

Consultation

Mode de consultation

Dialogue compétitif (i.e. appel d'offres sur performances)

- Expression de besoins fonctionnels
- Dialogue avec les candidats retenus
  - Lors des réunions d'étape, présence souhaitée :
    - des développeurs des codes de benchmark
    - de représentants des utilisateurs dans la commission technique
      - Suivi puis évaluation des offres
- Convergence pour l'élaboration d'un cahier des charges
- Évaluation des offres et choix

Benchmarks

Codes prévus

- Jeu de codes-tests représentatifs

Code	Origine / utilisateurs	Domaine	Type de parallélisme	Caractéristiques
2 codes	CORIA	CFD : diphasique / combustion	MPI (2048 - 4096 cœurs)	Réseau, Bande passante mémoire
1 code	GPM	Physique des matériaux	MPI (512 cœurs)	FFT
1 code	Libre, open source / IRCOF (Rouen), Paris et Lille	Chimie quantique	MPI (intranœud pour usage de disques internes)	Intensif en I/O
1 code	Libre, open source / Lille 1, Rouen (IRCOF), Caen (CERMN)	Dynamique moléculaire	MPI / GPU	GPU

Planning

Renouvellement des ressources

- Mars 2016
  - Clotûre de consultation, choix du constructeur et de sa solution

- Eté 2016
  - Ouverture de la machine aux utilisateurs
    - Vérification de Service Régulier (VSR)
    - Mise en production officielle

Centre de Ressources Informatique de Haute-Normandie