CRIANN

Comité Technique HPC

CRIANN - 20 juin 2024

Comité technique HPC

Ordre du jour

- Bilan d'exploitation
- Focus technique : le stockage (quotas, performances)
- Actualités MesoNET
- Agenda
- Questions / Réponses

Bilan des attributions et des consommations

Au 18/06/2024

- Bilan des appels à projets scientifiques 2024 (et fil de l'eau)
  - 1er AAP 2024 : 58 M.h.c attribuées / 92 projets scientifiques
  - 2ème AAP 2024 : 13,7 M.h.c demandées / 21 projets scientifiques

- Bilan de la production d'heures 2024 au 18/06
  - Industriels : 6,8 M.h.c
  - Académiques : 49,3 M.h.c

Valorisation des heures de calcul académiques

Pour mémoire

- Penser à mentionner l'utilisation des moyens de calcul Criann dans les publications
  - *Ce travail a bénéficié des moyens de calcul du mésocentre CRIANN (Centre Régional Informatique et d'Applications Numériques de Normandie).*
  - *Part of this work / The present work / was performed using computing resources of CRIANN (Normandy, France)*

- Contributions au financement des coûts de calcul
  - Rappel : sur la base du volontariat !!!
  - Via une ligne budgétaire dans les demandes de financement (ANR, Cifre, etc.)
  - Nous informer du retour de vos demandes 2024 (mail à MSC)
  - Pour information : travail en cours entre les directions des établissements et celle du Criann, sur ce sujet, et sur une évolution des statuts du Criann

Austral

Volet exploitation

Arrêts de production / incidents

Depuis le dernier CT (9 avril 2024)

- Aucun arrêt de production
- Incidents
  - 15 & 16 mai 2024 (durée 1,5 jour)
    - maintenance non programmée du système de refroidissement sur une armoire GPU
    - impact sur la tranche c23gpu[1-6]
  - 10 juin 2024 (durée ~8 heures)
    - maintenance du système de refroidissement sur une armoire GPU
    - impact sur la tranche c23hpda[1-5]

Charge d'occupation d'Austral

Pourcentage d'occupation des ressources CPU (partitions parallèles et smp)
et GPU (partitions GPU) (hors jours de maintenance)

CRIANN

Actions mises en place suite au dernier CT

Suivi des consommations

- Suivi des consommations
  - Envoi de mail mensuel aux responsables de projet avec le détail par login
  - Comptabilité dans les fichiers .o des calculs
  - Documentation : https://services.criann.fr/services/hpc/cluster-austral/guide/conso-sacct/
  - Conseil : exploiter ces informations pour optimiser vos soumissions de travaux (en particulier la mémoire)

Actions mises en place suite au dernier CT

Slurm - partage des ressources

- Nouvelles règles de partage des ressources
- Maximum autorisé en simultané par utilisateur
  - Sur GPU A100 : 16 GPU
  - Sur CPU : 22 nœuds / 4224c en semaine, 44 nœuds / 8448c le week-end

- Rappel des préconisations au niveau utilisateur :
  - Consommer régulièrement
  - Utiliser les heures de week-end : soumettre le vendredi soir

Proposition d'évolution

Slurm - partage des ressources

- Organisation actuelle des partitions sur les GPU
  - Partitions gpu, hpda, gpu_all, hpda_mig : durée max de 72 h (3 j)
- Proposition d'évolution
  - Partitions hpda, hpda_mig : durée max de 72h (3j)
    - Ressource max par calcul : 1 serveur (8 GPU)
    - Ressources associées aux partitions : 6 serveurs + 1 serveur MIG
  - Partition gpu, gpu_all : durée max de 24h (1j) ou 48h (2 j)
    - Ressource max par calcul : 2 serveurs (16 GPU)
    - Ressources associées aux partitions : 10 serveurs
  - Conseil : utiliser les fichiers de reprise

Sécurité

Fermeture des comptes inutilisés

- Certains comptes utilisateurs n'ont pas eu de connexion depuis la migration de Myria vers Austral
  - Les accès à ces comptes seront désactivés dans les prochains jours
  - Demande de réactivation possible par le responsable de projet
- Dans un deuxième temps
  - Envoi d'un mail à chaque responsable de projet avec la liste des comptes, pour vérification des comptes à conserver

Point d'attention

Nœuds de calcul diskless

- Les nœuds de calcul d'Austral n'ont pas de stockage local
  - Tout est chargé en mémoire
- Un programme effectuant des écritures dans /tmp écrit dans la mémoire
  - Peut générer un plantage du nœud

THÉMATIQUE SCIENTIFIQUE	NOM DU LOGICIEL
SIMULATION ATOMISTIQUE ET OUTILS CONNEXES	CHARMM
	GROMACS
	NAMD
	MOLPRO
	VASP
	PSI4
	DALTON
	AMF
	LAMMPS
	QCORE
	ASE
	ICMR-GAUSSIAN
	Quantum Espresso
	VMD
BIOLOGIE	Augustus
	Guppy
	Dorado
MÉCANIQUE DES FLUIDES	Star CCM+
	SWASH
	MODULEF
	DUALPHYSICS
	YADE
	TELEMAC-MASCARET
	OPENFOAM
	FOAM-EXTEND
	CODE_SATURNE
MODÉLISATION ATMOSPHÉRIQUE, CLIMATOLOGIE, OCÉANOGRAPHIE	WRF - WPS
	NCL
	WGRIB
	GEOS
	GDAL
	CDO
	R_TERRA
	NCO
	SIRANE

THÉMATIQUE SCIENTIFIQUE	NOM DU LOGICIEL
MÉCANIQUE, ACOUSTIQUE	Code ASTER
	CAST3M
	HYPERWORKS
	LS-DYNA
	NASTRAN
	SALOME-MECA
MATHÉMATIQUES, STATISTIQUES	FREEFEM ++
	OCTAVE
	SCILAB
	R
	Python/dask
	Python/pandas
MACHINE LEARNING, DEEP LEARNING	PyTorch
	TensorFlow/Keras
	Horovod
	Scikit-learn
	OpenCV
MAILLAGES	GMSH
COUPLEURS	Oasis
	Precice
VISUALISATION	Paraview
	Ferret
	Xmgrace
	Molden
	Ncview

Logiciels disponibles sur Austral (mars 2024)

Austral

Logithèque

- Applications scientifiques
  - Installations à la demande, optimisées pour l'architecture
  - Possibilité d'effectuer sa propre installation

Focus technique : stockage

Stockage

Pourquoi ce focus ?

- La migration de Myria vers Austral s'est accompagnée de quelques changements au niveau technique
  - Un stockage de moindre capacité => 2 Po sur Austral (vs 2,5 Po sur Myria)
  - 2 types de disques pour le stockage des données
    - 1 Po NVme et 1 Po disques rotatifs => performance crête ~150 Go/s (vs ~25 Go/s)
  - Un nouveau système de fichiers parallèle => Lustre (vs GPFS)
- Bilan après ~8 mois d'exploitation, il est impératif
  - de limiter l'occupation du système de stockage
    - en volumétrie
    - en nombre d'inodes (lié au nombre de fichiers)
  - d'adapter certains workflows pour obtenir de bonnes performances

Stockage

Limitation des ressources

- Mise en place des quotas sur la volumétrie et les inodes
  - par identifiant sur la totalité de la baie de disques
    - volumétrie : **quota strict de 30To**
    - nombre d'inodes : **quota strict de 5M**
  - **Objectif : éviter un incident causé par un job**
    - Ce quota n'est pas un droit à consommation !

Stockage

Limitation des ressources

- Un quota strict sera appliqué sur les dossiers d'accueil
  - À la création des nouveaux comptes et des nouveaux projets
    - /home/projet_id/identifiant : mise en place d'un quota de **50 Go**
    - /home/projet_id/PARTAGE : mise en place d'un quota de **200 Go**, augmentable sur demande justifiée
  - Sur les comptes et projets déjà existants
    - Mise en place de quotas, adaptés dans un premier temps, qui seront revus à la baisse dans un 2ème temps (~septembre)
    - Faire du ménage, supprimer les données inutiles ou redondantes
    - Rapatrier vos résultats dans vos laboratoires
    - Revoir vos workflows, en particulier sur les bases de données d'IA

Stockage

Limitation des ressources

- Surveiller votre utilisation avec la commande ***cri_quota***
- Si votre utilisation atteint 90% de l'une des limites, votre consommation est affichée lors de la connexion
- L'effacement des données des anciens jobs sur /dlocal/run est effectué
  - durée de rétention actuelle 60 jours
  - évolution en cours : 40 jours après la fin de job

- Quelques préconisations
  - Tourner dans /dlocal/run et rapatrier uniquement les fichiers à conserver
  - Utilisateurs d'OpenFoam : penser à utiliser l'option "-fileHandler collated" (cf documentation https://services.criann.fr/services/hpc/cluster-austral/guide/data-management/)

Performances

Pool flash - pool disk

- Stockage NVme (pool flash)
  - Max mesurés (IOR) : écriture 167 GiB/s, lecture : 304 GiB/s
- Stockage disques rotatifs (pool disk)
  - Max mesurés (IOR) : écriture : 44,5 GiB/s, lecture : 49 GiB/s
- En cas de taux de remplissage élevé dans le pool flash
  - Migration des données anciennes et volumineuses sur le pool disk
  - -> performances réduites sur les accès à ces fichiers
- Voir la documentation : https://services.criann.fr/services/hpc/cluster-austral/guide/data-management/

Performance Lustre et IA

Performance Lustre et IA

Système de fichiers Lustre

- Système de fichiers parallèle optimisé pour les stockages de grande capacité (plusieurs Po)
  - I/O performantes sur les gros fichiers
  - I/O peu performantes sur les petits fichiers (< 1Mo)
  - I/O peu performantes sur les répertoires contenant beaucoup de fichiers

Performance Lustre et IA

ImageNet sur Lustre

- Base de données d'images utilisée pour l'entraînement en Deep Learning
  - De l'ordre de **10 000 000** petits fichiers images (entre **~10Ko et ~1.5Mo**)
  - Organisée en un millier de répertoires
  - => usage non optimal sur Lustre
- Tests de performances effectués
  - Entrainement d'un modèle resnet50 avec Pytorch 2.0 sur 1 GPU nvidia A100 et 8 cpus pour le chargement des données
    - Une passe sur 65 536 fichiers de la base, par paquet de 512
  - Cas pour comparaison : même modèle resnet50 ; images extraites à la volée de fichiers vidéos - 64 images extraites par vidéo (x 8)
  - Résultats attendus : durées d'entrainement similaires entre les 2 cas - de l'ordre de **85s** (référence Jean-Zay- IDRIS)

Performance Lustre et IA

ImageNet sur Lustre

- Performances observés
  - Cas vidéos : entre **~ 72s et ~85s**
  - Cas ImageNet : variation entre **~ 75s et ~180s**
- Explication des performances (analyse avec HPE)
  - Lustre non performant sur les petits fichiers
  - Configuration actuelle de Lustre inadaptée (stripping)
  - CPUs fortement sollicités par les processus Lustre
  - Performances parfois correctes si les données sont dans le cache Lustre

Performance Lustre et IA

ImageNet sur Lustre

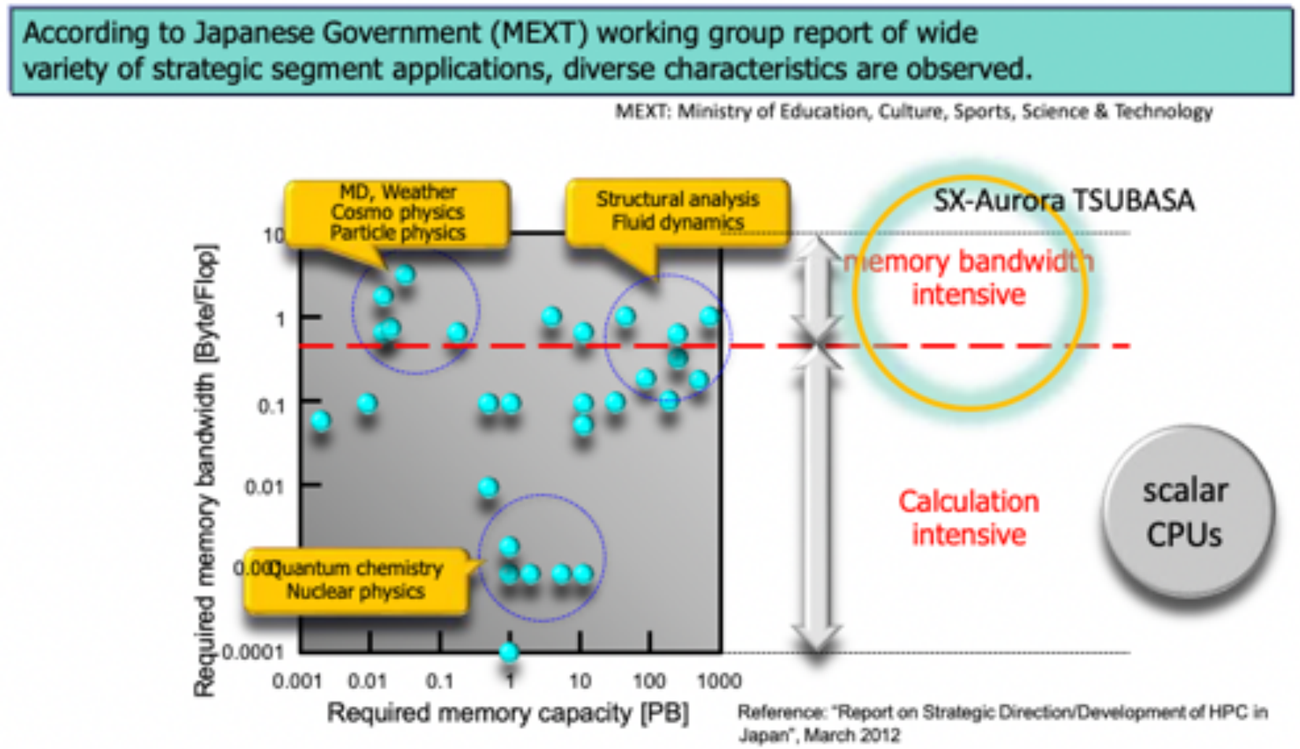
- Solutions (**travaux en cours avec HPE**)
  - Augmenter la « capacité CPUs » des nœuds **gpu** et **hpda** en activant l'hyperthreading (test à venir)
  - Modifier le stripping des répertoires contenant les bases
  - Taille de lecture/écriture des blocs de données
  - Nombre de serveurs d'IO utilisé
- **Créer une archive à partir de la base (=> un seul gros fichier) et monter cette archive à la volée lors des jobs d'entrainement**
  - archive squashfs créée une seule fois
  - temps de restitution entre ~85s et ~105s
    - variation des performances explicables par la charge des CPUs

MesoNET, vectoriel

Architecture spécialisée

Machine vectorielle du CRIANN, Boréale : cibles

- Architecture : https://www.mesonet.fr/documentation/user-documentation/arch_exp/boreal/description



Langages
- C/C++/FORTRAN, OpenMP, MPI
- Python : Numpy porté par NEC

Bibliothèques optimisées
- BLAS, LAPACK, ScaLAPACK, FFTW

Parmi les applications les plus connues dotées d'une version portée efficacement sur l'architecture NEC Aurora :
- VASP 6
- Quantum Espresso 6.4.1 et 7.1
- NEMO (circulation océanique)

SDK : https://sxauroratsubasa.sakura.ne.jp/Documentation#SDK

Architecture spécialisée

Machine vectorielle du CRIANN, Boréale : accès

- Accès à MesoNET :
  - https://www.mesonet.fr/documentation/user-documentation/acces/portail
    - Compte à demander sur https://iam.mesonet.fr/login
    - Puis dépôt de projet sur le portail https://acces.mesonet.fr
      - e.g. 1000 heures x VE (Vector Engine) à demander pour Boreale

- Ou écrire à support-boreale@criann.fr

Agenda

Arrêt de production d’Austral

Mise à jour système

- Planifié à partir du lundi 26 août
- Retour en production prévu mercredi 4 septembre

Formations

À venir

- Criann site satellite de la **Gray Scott School**, du 1er au 11 juillet
  - Programmation et optimisation sur architectures hétérogènes
  - Semaine 1 : CPU, semaine 2 : GPU
  - Inscriptions jusqu'au 24 juin, merci de préciser les jours suivis
- En cours d'organisation sur le 2ème semestre
  - **CPE** (Cray Programming Environment), par HPE
    - Outils de compilation, débugage, profilage
  - **MLDE** - Machine Learning Development Environment, par HPE
    - cf. présentation au CT d'avril
  - Conférence MesoNET « **État de l'art du calcul quantique** » par Eviden
    - 3h - fin septembre/début octobre, co-organisation avec l'Insa Rouen
  - **Python pour le HPDA**
  - Formations de prise en main à planifier

Agenda

- JCAD 2024, Bordeaux, du 4 au 6 novembre
  - https://jcad2024.sciencesconf.org/?lang=fr

- Prochain Comité Technique
  - 10 octobre 2024 envisagé

Le plateau de calcul intensif du Criann est cofinancé par la Région Normandie, l'État français et l'Union européenne (Fonds Feder).
MesoNET bénéficie d'un financement de l'Agence nationale de la recherche au titre des Investissements d'avenir.
Le Centre de Compétence EuroCC français est cofinancé par l'Union européenne et par l'État français.
Le réseau régional Syvik est cofinancé par la Région Normandie et par l'Union européenne (fonds Feder).
Le fonctionnement du Criann bénéficie du soutien de la Région Normandie.

RÉGION NORMANDIE

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

UNION EUROPÉENNE
Fonds européen de développement régional

CRIANN

Centre Régional Informatique et d'Applications Numériques de Normandie
www.criann.fr