

Instrument de Recherche Numérique

Nuwa du Laero-OMP

Didier Gazen

didier.gazen@aero.obs-mip.fr

10/02/2023, HPC@OMP2023

Cluster de calcul hétérogène Nuwa (5 baies) :

Piloté sous Linux openSUSE Leap 15.2 par :

- 1 nœud Dell R540 (24c, 192To) : utilisateurs (\$HOME:33To) + outils de dev. + Slurm
- 1 Intel NUC (2c, 8Go) : outils admin SluBK + base accounting Slurm

- **143** nœuds de calcul (\simeq 2700 cœurs) :
 - **17%** HPE **DL360** (20c|32c,192|256|384Go), HPE **DL380** (56c, 384Go)
 - **31%** Dell **R440** (20c, 96|192Go), **R740** (36c|40c, 192Go)
 - **37%** Dell **R430**, **R620** (16c|20c, 32|64|128Go), **R720** (16c/32Go, 20c/128Go)+GPUs
 - **15%** Dell **R410** (8c, 24Go)

- dont **6** nœuds avec **bi-gpus NVidia** (Titan(V), **Tesla V100**)

- 1 cœur réseau ethernet : 168 ports 1Gb/s + 4 ports 10Gb/s

- 4 îlots Infiniband (dédié MPI) : 3x36 ports 40Gb/s, 1x36 ports 56Gb/s

Cluster de stockage Ceph CNUwa (2 baies) :

- mise en production en 2020 (remplacement stockage 900To SAS)
- équipements :
 - 1 nœud admin Intel NUC (2c, 8Go)
 - **3 nœuds MON/MDS** Dell R410 (8c, 32Go)
 - **31 nœuds OSDs** Dell R740xd/HPE Apollo 4200 (20c, 192|256Go), **16x8To** sas/sata, 1x480Go ssd
 - réseau public ethernet : 1x48 ports 10Gb/s
 - réseau privé Infiniband : 1x36 ports 56Gb/s
- caractéristiques Ceph :
 - Version 14.2.22 (Nautilus)
 - **522 OSDs**, 11 pools(EC8.3), 1 cephfs
 - 4000 pgs, 500M objets, **100M fichiers** cephfs

Volume total **3.5Po brut** délivrant **2Po** via **cephfs**

- Projets de recherche **utilisateurs** (soutiens Laero, OMP, région, ANR, Europe) pour **nœuds de calcul** et **nœuds de stockage**
- **Soutien de base Laero** pour équipements communs (jouvence clim, cœur réseau, armoire électrique, matériel IB, **nœud frontal, stockage Ceph**)
- **OMP** pour 4 nœuds avec accélérateurs et facture électricité
- **Hébergement de ressources** :
 - **Calcul** (% total nœuds) : Laero (71%), Legos (17%), Get (4.2%), OMP (2.8%), Irap (5%)
→ **3k€ HT** (matinfo5) pour HPE DL360 : **20c, 384Go**
 - **Stockage Ceph** (% espace total) : Laero (63%), Legos (20%), Sedoo (10%), Get (7%)
→ **8k€ HT** (matinfo5) pour HPE Apollo 4200 : **16x8To** (64To utile, 125€/To)

- Equipe technique **1 ETP** depuis fin 2019 (2 auparavant) collabore avec les ingénieurs/chercheurs pour dimensionner les besoins, installe, surveille et maintient matériels/logiciels, forme les utilisateurs (wiki Nuwa) et assure le support technique
- Installation d'équipements contrainte par place/puissance imposée par le datacenter Laero (7 baies, 84kW, saturé aujourd'hui)
- **Utilisateurs autorisés** : financeurs de la ressource de calcul + invités + demandes ponctuelles. **Adhésion à une politique de partage** mise en œuvre via le gestionnaire de ressource Slurm
- **Aucun comité de programme** : chaque équipe décide de l'utilisation de son matériel (partitions Slurm, création de comptes et partage inter-équipe)

→ utilisateurs accèdent au matériel financé, voire plus !

- **183 utilisateurs** enregistrés : Laero (52%), Legos (17%), Get (15%), Irap (8%), OMP (5%), Extérieurs (3%, Cnrm, UPS, Laas).

- Outil de **développement, prototypage** et **débogage** des codes numériques (MésóNH, Sirocco, Croco. . .) + **portage** notamment sur gpus
- Outil pour le **traitement de données** (Iagos, CAMS, Iasi, Saetta. . .) ou la **prévision** (campagnes) : tâches difficiles ou impossibles à réaliser sur des centres régionaux/nationaux
- **Plateforme de tests** (1^{ère} marche) pour accès aux centres régionaux, nationaux et internationaux avec architecture et environnement de travail équivalents
- **Services particuliers** : visualisation déportée (**VituaGL** sur GPUs), travail distant via **x2go**, environnements Python gérés par **Conda** (éco-système **Pangeo**, IA **Keras**), **JupyterLab** sur nœuds de calcul, conteneurs **Singularity**

	2019	2020	2021	2022
Nbre cpus	1962	2033	2180	2313
Taux Occupation % Total (Mois max)	22%(30%)	19%(28%)	27%(37%)	25%(46%)
Heures CPU allouées en millions	3.7	3.4	5.0	5.0
Nbre Utilisateurs allocation \geq 5000 heures/an (total)	40 (92)	28 (80)	36 (93)	35 (87)

- Augmentation ressource cpus à volume/enveloppe énergétique constante
- Baisse occupation en 2020 : effet COVID (ni campagne de prévision, ni réservation)
- Nombre utilisateurs soumettant des jobs $>$ 80/an, réguliers \simeq 35/an

- Outil maîtrisé favorisant la **veille technologique** (GPUs, Xeon Phi, conteneurs)
- **Jobs longs** (>15 jours) et **réservations** Slurm à la carte autorisés
- Calcul “**écoresponsable**” : gestion de l'énergie avec **arrêt des nœuds inactifs**
- **Important volume de données (2Po) au plus prêt du calcul.** Pas de sauvegarde (snapshots cephfs à explorer)
- Mutualisation de la ressource de calcul et centralisation de l'administration système a fait ses preuves. Solution en voie de déploiement sur d'autres unités (Legos, Irap) à l'OMP (sous réserve soutien RH de la DSI-OMP)
- Positionné comme **Tier-3**, l'IRN Nuwa :
 - participe à la formation des étudiants en HPC
 - s'avère très flexible pour s'adapter aux besoins des utilisateurs
 - fournit des **services complémentaires** à ceux délivrés par les centres Tier-2 et Tier-1