CENTRE D'OCÉANOLOGIE DE MARSEILLE
UNIVERSITÉ DE LA MÉDITERRANÉE AIX-MARSEILLE II

ANNEXE TECHNIQUE

Acquisition, livraison, installation et mise en service d'une grappe de PC de calcul

Contexte et matériel existant :

Le Centre d'Océanologie de Marseille (COM) possède depuis 2003 un cluster de calcul (grappe de PC dédiés au calcul intensif) composé de :

- 1 nœud frontal : supportant 3.5 To de disque (2To en DAS LVM et 1.5To sur une baie iSCSI). Le noeud frontal est utilisé pour les connexions des utilisateurs,

- 6 nœuds de calcul hétérogènes (3 Dell Precision 670 et 3 Dell Poweredge 1950 dotés de 2 à 4Go de RAM), sont utilisés pour le traitement des modèles de calcul

Ce cluster de calcul est actuellement géré par le système RocksCluster (**http://www.rocksclusters.org**) qui donne satisfaction dans le sens où il fournit une installation aisée et automatique de tous les nœuds de calcul en Linux Centos, ainsi que des outils connus de clustering (Ganglia pour le monitoring, et SGE pour la soumission des processus de calcul sur les nœuds)

Les codes de calcul qui tournent sur cette plateforme de calcul sont des codes de modélisation biogéochimiques, pour certains séquentiels, pour d'autres parallélisés avec MPI (Intel)**.** Les codes actuels sont compilés avec le compilateur F90 de Intel et utilisent les librairies de parallélisation MPI de Intel

Objectifs :

Les besoins de calcul évoluant au sein de notre laboratoire, nous souhaitons désormais améliorer notre cluster actuel, et mettre en place une solution plus puissante, mieux intégrée et plus évolutive dans le temps. Nous souhaitons des nœuds de calculs plus puissants et homogènes, un réseau de meilleure qualité, et donc une solution plus pérenne et professionnelle qui améliore la puissance de calcul tout en optimisant également les paramètres physiques : consommation électrique, dégagement de chaleur, et niveau sonore.

L'architecture demandée

- Le nœud maître frontal sera destiné :

- à accueillir les connexions et un espace disque centralisé qui contiendra «les home Directory » d'une vingtaine d'utilisateurs qui s'y connecteront en ssh...
- à compiler les codes de calcul (compilateur fortran90 et MPI de Intel)
- à soumettre l'exécution des codes au moyen d'un ordonnanceur comme SGE (Sun Grid Engine) ou openPBS (torque/MAUI)

- Les autres nœuds de la grappe seront destinés uniquement au calcul et recevront les processus distribués par l'ordonnanceur SGE et/ou les librairies de parallélisation MPI
  - Ces PC de calcul auront seulement un disque de base de faible capacité destiné uniquement à installer un système d'exploitation. Ces PC feront un montage de l'espace disque global partagé du nœud frontal.

- L'ensemble nœud maître + nœuds de calcul devra être intégrable dans un châssis (~38U) *« à fournir » dans l'offre*

- *Nota: pour ne pas perdre la totalité de notre cluster actuel, nous demandons à pouvoir intégrer et réutiliser nos 3 nœuds de calcul existants Dell Poweredge 1950 (1U), dans le nouveau cluster proposé... (ceux ci seront bien entendu hors garantie)*

Nous souhaitons que ce cluster soit évolutif dans le temps et que nous puissions au terme de la garantie et de la durée de vie des PC, remplacer les PC par des modèles de caractéristiques techniques les meilleures du moment.

Les Processeurs :

Tous les PC seront équipés au minimum de Bi-processeurs Dual Core les plus puissants du moment pour le calcul.

Le prestataire fournira dans son offre les caractéristiques exactes des processeurs proposés (nombre de processeurs, nombre de cœur par processeur, vitesse, fréquence, cache...) *(par exemple: Dual Core Intel® Xeon® 5110, 4MB Cache, 1.60GHz, 1066MHz FSB)*

L'Espace disque :

Le nœud frontal possédera un espace disque qui devra être partagé (/home) sur tous les nœuds de calcul.

- Un espace disque minimal de 4 To utile est demandé, géré par un RAID matériel (Raid 5 minimum) plus un disque de Spare
- disques hotswap si possible

Le volume disque *devra pouvoir être augmenté simplement* et aisément par ajout de disques hotswap. Cette capacité d'extension devra être prévue dans la configuration proposée.

- L'offre proposera les prix pour une configuration avec des disques SATA de base, avec une possibilité d’extension ou de ***mixité*** avec des disques SAS

Les nœuds de calcul auront juste un disque système (~80G) non redondé

RAM:

Tous les nœuds de calcul seront équipés d'un minimum de 8Go de RAM. L'entreprise fournira les caractéristiques physiques des mémoires proposées

Caractéristiques physiques et environnementales

L'offre précisera tous les paramètres physiques du cluster ainsi que les critères environnementaux suivants :

- les côtes d'encombrement pour l'installation au sol,
- le poids de l'ensemble,
- les dimensions dans toutes les directions (y compris lors de la livraison);
- la puissance crête par nœud de calcul (GFLOPS)
- type de raccord au réseau électrique du laboratoire (référence des éléments de connectique)
- la consommation totale de courant et les raccords électriques nécessaires
- la puissance électrique totale consommée (VA / kW) ,
- la chaleur dégagée (BTU/h)
- le niveau sonore en Db pour le système complet ;

Le Réseau d'interconnexion des nœuds

Dans la configuration de base du cluster, par défaut chaque nœud sera équipé d'une carte Ethernet Gb/s

Cependant, l'entreprise proposera en option 1 le coût pour une mise à niveau vers un *réseau à faible latence de type Infiniband ou 10GbE*

Logiciels, système d'exploitation, évolutivité du cluster

Les PC devront tourner sous le système Linux (Debian bienvenue, ou RedHat like (centos ou autre)...)

L'offre décrira précisément :

- le Système et les outils d'administration système utilisés
- le Gestionnaire de batch embarqué : de type SGE, ou open PBS
- Les procédures d'installation ou de réinstallation des nœuds de calcul
  - *l'ajout ou le retrait de nœuds devra être une manipulation simple et aisée.*
  - *L'évolution du cluster dans le temps devra être une chose aisée.*
    - Nous voulons pouvoir au fil des années faire évoluer le cluster en ôtant les nœuds aux performances obsolètes et en les remplaçant par

les meilleures machines du moment. Cette opération devra être possible et rendue simple par le système d'installation.

Le système devra pouvoir permettre une installation et diffusion aisée de logiciels additionnels sur tous les nœuds de calcul comme « R », MATLAB, le Fortran90, MPI de Intel , ou tout autre logiciel ou librairie de calcul rendu nécessaire par les calculs etc...

Tests de performances - benchmark

Si nécessaire les équipes concernées du centre d'océanologie de Marseille tiennent à disposition des codes de calculs séquentiels et parallélisés en MPI, avec leurs conditions d'utilisation, codes qui pourront être utilisés par les entreprises pour déterminer les meilleures configurations matérielles (cpu, réseau, disque) du cluster demandé. **Les entreprises fourniront les résultats des tests et l'argumentaire pour confirmer leur choix de matériel.**

Services et maintenance demandés

1- garantie sur trois ans comprenant au minimum :

a-garantie des pièces envoi à J+1et mise à jour mineure et majeure du système.

b-support par hot-line, accéder à une base de connaissances et obtention la documentation adéquate.

2- Formation et transfert de compétences pour 2 personnes.

Détail de l'offre

L'entreprise devra remplir le CCP valant acte d'engagement (page 12) sur lequel apparaissent les éléments suivants :

- le prix de l'ensemble du cluster de 8 PC. L'entreprise proposera également dans son offre le prix ***unitaire par PC ajouté au cluster.***

- En option 1, la proposition de mise à niveau sur réseau « infiniband » ou 10GbE.
- En option 2 : la fourniture de l'armoire 19 pouces pour contenir les PC du cluster.

Rappel : Le prix du cluster contenant les 8 PC devra inclure le transport, la livraison, l'installation, la mise en service du matériel, la configuration du système du cluster et la garantie de 3 ans.

Le bordereau de prix que vous établirez comportera séparément tous les éléments : cluster de 8 pc, baie de disques, élément réseau de base, armoire de 19 pouces, mise à niveau en réseau faible latence infiniband).

L'entreprise devra fournir un mémoire technique comportant les renseignements cités ci-dessous.

- décrire très précisément la solution matérielle proposée :
  - configuration matérielle de chaque nœud : processeurs, RAM, Disque, taille mémoire cache, mémoire par cœur et débits proc/mémoire
  - proposition pour 4 To utiles en disques SATA et SAS
  - type de réseau fourni (argumentaire pour trancher entre 10GbE et infini band bienvenu)

- décrire très précisément la solution logicielle proposée :
  - procédures d'administration du cluster
    - procédures d'installation, ajout, retrait des noeuds et de leur système
    - procédures d'installation de logiciels supplémentaires devant être partagés sur les nœuds de calcul (R, Matlab, MPI..)
  - système de supervision de l'ensemble du cluster (style ganglia)
  - système de soumission de batch (style SGE)

- décrire précisément les caractéristiques physiques et environnementales
  - les côtes d'encombrement pour l'installation au sol,
  - le poids de l'ensemble,
  - les dimensions dans toutes les directions (y compris lors de la livraison)
  - la puissance crête par nœud de calcul (GFLOPS)
  - type de raccord au réseau électrique du laboratoire (référence des éléments de connectique)
  - la consommation totale de courant
  - la puissance électrique totale consommée (VA / kW) ,
  - la chaleur dégagée (BTU/h)
  - le niveau sonore en Db pour le système complet ;

- indiquer précisément le contenu de la prestation de service comme demandé ci-dessus (paragraphe « services et maintenance demandés page 4)