MARCHES PUBLICS DE FOURNITURES COURANTES ET SERVICES

Université de Strasbourg
DIRECTION DES FINANCES
Département des Achats et des Marchés
Institut LE BEL
Bureau 336 H - 3ème étage
4 rue Blaise PASCAL
67081 STRASBOURG Cedex
Tél: 03 68 85 12 01

EQUIPEMENT D'EXCELLENCE DE CALCUL HAUTE PERFORMANCE (HPC)

Université de STRASBOURG

Cahier des Clauses Techniques Particulières

Table des matières

Méso-centre de l'Université de Strasbourg

I. Contexte

L'Université de Strasbourg (Unistra) est lauréate d'un équipement d'excellence visant à renforcer son centre de calcul haute performance (HPC). Le but de l'équipement est de faciliter le passage à l'exascale pour les chercheurs qui en seront bénéficiaires.

L'opération se déroule en 3 phases :

1. Mise à niveau du cœur de réseau Infiniband et de l'espace de stockage parallèle *(réalisé)*
2. Urbanisation de la salle machine accueillant les équipements *(en cours de finalisation)*
3. Le présent appel d'offre : « Équipement d'excellence de calcul haute performance (HPC) »

Le présent appel d'offres concerne la 3ème phase de l'opération.

Il s'agit dans le cadre du présent appel d'offres d'acquérir une configuration de serveurs de calcul à des fins d'expérimentation et de production :

- des serveurs de calcul parallèle, incluant des GPU ;
- un réseau haut-débit Infiniband ;
- un socle de logiciels orientés HPC (compilateurs optimiseurs, débogueurs).

La solution proposée devra obligatoirement s'intégrer aux ressources de calcul déjà en place. La solution en place est présentée dans la suite.

Pour cet appel d'offres, le budget prévisionnel est de 700 000 HT, hormis la partie à bons de commande.

II. Configuration existante

2.1) Applications scientifiques

Le méso-centre de l'Université de Strasbourg fait tourner des applications scientifiques utilisant :

- le parallélisme avec MPI, OpenMP. Des applications développées par les chercheurs utilisant OpenMPI cohabitent avec des codes commerciaux utilisant Platform MPI : ADF, Fluent, Turbomole ;
- l'accélération GPU ;
- des tâches séquentielles répétitives.

2.2) Déploiement des systèmes d'exploitation

Actuellement, nous intégrons une solution basée sur Systemimager (http://wiki.systemimager.org), qui déploie des arborescences en mode fichier. Cette solution pose quelques problèmes au vu de l'hétérogénéité des ressources. Nous sommes en phase de migration vers la solution xcat.

2.3) Logiciels de gestion de ressources et de batch

Nous mettons à disposition l'environnement Torque (http://www.clusterresources.com/), piloté par les politiques d'exploitation décrites dans Maui (http://www.clusterresources.com/). Nous assurons le support de ces logiciels auprès de nos utilisateurs. Le paramétrage du gestionnaire de batch est rendu très compliqué par le fait que les ressources de calcul sont financées directement par les utilisateurs, conduisant à une répartition non-équitable des cycles CPU. Nous sommes en phase d'étude de la solution slurm.

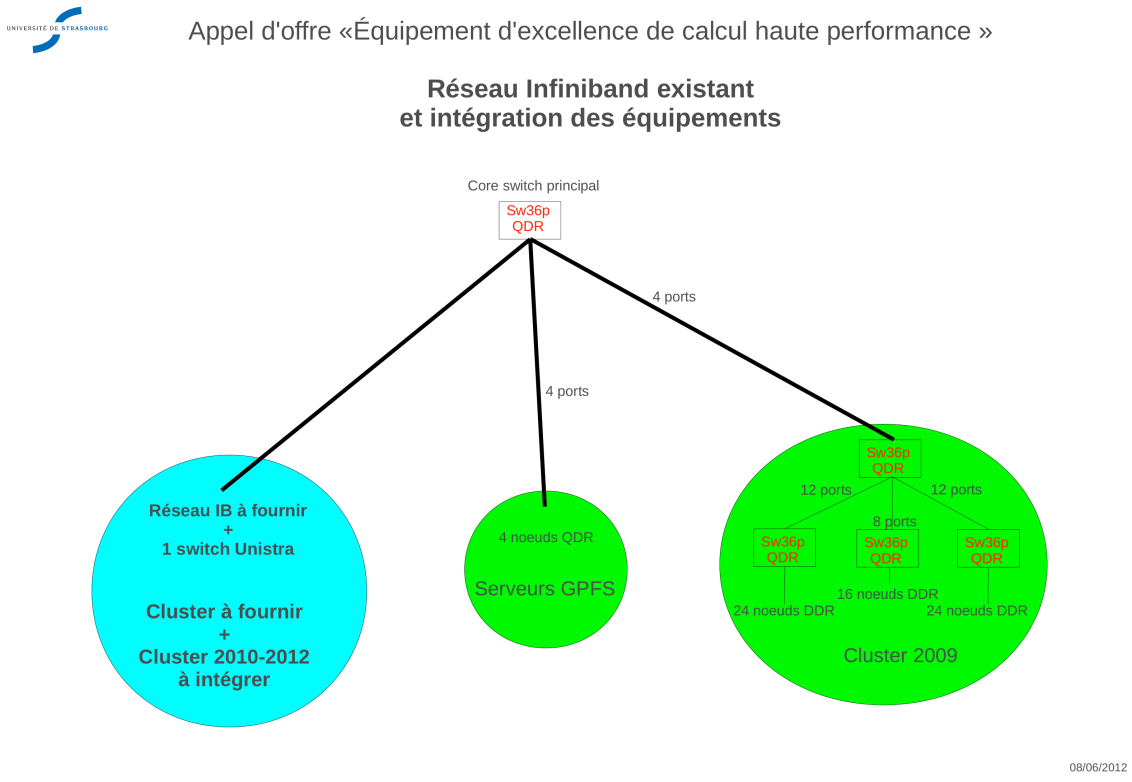
2.4) Socle de logiciels orientés HPC

Actuellement, nous utilisons les logiciels suivants :

- Compilateurs Intel Version 12, 5 jetons flottants comprenant :
  - C/C++
  - Fortran 90/95
  - Bibliothèque Mathématique MKL
- Compilateurs PGI Cluster Development Kit, version 11.2, incluant PGI Accelerator : 2 jetons flottants
- Débogueur parallèle Totalview : licence *Team* pour 4 processus

2.5) Réseau haut-débit

Nous utilisons un réseau haut-débit Infiniband (cartes hôte et commutateurs) QDR. Nous disposons de commutateurs Mellanox MTS 3600. Le réseau est représenté ci-dessous :



2.6) Réseau Gigabit Ethernet

Notre réseau Gigabit ethernet actuel, utilisé pour le déploiement de fichiers, est structuré par 2 switchs HP Procurve 2900-48G et un switch Dell Powerconnect, relié par un backbone 10 Gbit de connectique CX4.

2.7) Système de fichiers parallèle

Nous disposons d'une solution GPFS s'appuyant sur le réseau Infiniband présenté ci-dessus. Dans la configuration GPFS, nous disposons de 4 serveurs d'entrée/sortie reliés à 2 baies DS3512. La version du logiciel GPFS est 3.4

2.8) Cluster de calcul

Le cluster **actuellement** en place est hétérogène et composé comme suit :

Nombre et usage	Type de processeur	Nb de processeurs	Constructeur	Dénomination
64 nœuds de calcul	AMD 2384	2	HP	Cluster 2009
4 nœuds de calcul	Intel X5550	2	Dell	Cluster 2010
23 nœuds de calcul	Intel X5650	2	Asus	Cluster 2010
5 nœuds de calcul	Intel X5650	2 + 2 Nvidia M2050	Supermicro	Cluster 2010
16 nœuds de calcul	Intel X5650	2	Supermicro	Cluster 2010
4 nœuds de calcul	Intel E5-2630	4	Supermicro	Cluster 2010
2 serveurs frontaux	AMD 8350	4	Supermicro	
1 serveur d'infra	Intel E5345	2	HP	

III. Prestations attendues

3.1) Descriptif général de la solution

Le présent appel d'offres comprend deux parties une partie forfaitaire et une partie à bons de commande.

3.1.1) Partie forfaitaire

La partie forfaitaire concerne l'ensemble de la solution (serveurs de calcul, serveurs de calcul multi-GPU, réseau haut débit, logiciels HPC).

La solution comprend un volet cluster de calcul, à base de nœuds d'architecture x86_64 à des fins de compatibilité avec l'existant. Dans ce même but, le **réseau haut-débit à fournir** doit être de technologie Infiniband.

Nous souhaitons des nœuds multi-cœurs bi-socket, avec 4 GO de mémoire par cœur. Les nœuds GPU doivent être équipés d'au moins 2 GPU chacun.

Sur l'ensemble de la solution, nous souhaitons disposer de 10% de GPU compatibles avec les codes déjà en production au méso-centre de calcul (développés en CUDA), et souhaitons pouvoir rajouter 10% de GPUs dans des emplacements libres. Par exemple, si la solution proposée comporte 400

nœuds, elle devra intégrer initialement 40 GPUs, et des emplacements libres pour 40 GPUs supplémentaires. La répartition des GPUs sera réalisée selon l'exemple suivant

Pour 40 GPUs achetés initialement :

- Si les nœuds proposés peuvent contenir exactement 2 GPUs, 20 nœuds seront installés avec 2 GPUs chacun, 20 seront dépourvus de GPUs
- Si les nœuds proposés peuvent contenir exactement 4 GPUs, 20 nœuds seront installés avec 2 GPUS chacun.
- Si les nœuds proposés peuvent contenir exactement 8 GPUs, 10 nœuds seront installés avec 4 GPUs chacun.

Le débit dédié à un GPU sur le bus PCI Express sera pris en compte dans la notation (cf « ANNEXE 2 à l'AE - résultats des tests »).

Nous souhaitons également disposer d'un serveur de *login* permettant la compilation et les tests d'applications : un serveur quadri-socket multi-cœurs, disposant de 8 GO de mémoire par cœur.

Un serveur dédié au déploiement système et au gestionnaire de batch et de ressources doit également être proposé : un serveur bi-socket multi-cœurs, disposant de 1 GO de mémoire par cœur.

Spécifications minimales communes aux nœuds de calcul :

- Processeurs d'architecture x86_64 performants ;
- Rackables dans des armoires 19 pouces standard. Une solution de type Blade est possible ;
- 1 Port Infiniband ;
- Disque dur hot-swap de 500 GO minimum. Technologie de disque dur au choix ;
- Alimentation **non redondante certifiée 80+ platinium ;**
- Carte d'administration à distance compatible IPMI utilisant un port Ethernet **partagé** avec le système.

Spécifications minimales communes au serveur de login et au serveur de déploiement et de batch :

- Alimentations **redondantes certifiées 80+ platinium ;**
- Contrôleur RAID ;
- Rackables dans des armoires 19 pouces standard ;
- Port Infiniband ;
- Carte d'administration à distance compatible IPMI utilisant un port Ethernet **dédié**.

3.1.2) Partie à bons de commande

La partie à bons de commande permettra d'étendre la configuration par l'achat de :

- nœuds de calcul par 2 ou par 4 ;
- switchs Infiniband ;
- câbles Infiniband ;
- cartes GPU.

3.2) Réseau Haut-Débit et accès aux fichiers

Nous rappelons que l'accès aux serveurs de fichiers GPFS se fait en utilisant le réseau Infiniband déjà en place. Cette possibilité devra impérativement être conservée.

En cohérence avec le réseau Infiniband déjà en place, un réseau haut-débit dédié au HPC (commutateurs, câbles) sera mis en place par le titulaire du marché.

Le facteur de blocage du réseau proposé sera compris entre 1:1 (réseau non bloquant) et 2:1 (2 ports descendants pour 1 port remontant).

Le réseau proposé devra être composé de switchs « stackés ».

Le niveau de stacking supérieur (core) devra disposer de suffisamment de **ports en attente** pour relier ultérieurement des switchs au niveau de stacking inférieur « côté nœuds » (edge), pour une capacité supplémentaire de 60 nœuds qui seront installés dans l'armoire 4..

Chaque switch core doit être relié au core switch principal de notre réseau existant pour permettre l'accès au serveur de fichiers GPFS. Dans notre configuration, le core switch principal est un Mellanox MTS 3600.

Chaque switch du niveau de stacking inférieur (edge) devra disposer de suffisamment de ports en attente pour les connexions aux switchs core de l'armoire 5 (à venir).

Il est possible au soumissionnaire de proposer des switchs ou des cartes hôte de marque différentes. Dans tous les cas, l'intégration au réseau existant est à la charge du titulaire. Cette intégration sera vérifiée dans la VA (vérification d'aptitude au bon fonctionnement) et dans la VSR (vérification de service régulier), en particulier par l'interrogation de la topologie Infiniband.

Pour la phase de câblage, il faudra répartir les core switchs dans 4 armoires et les switchs edge dans le nombre d'armoires occupés par la solution. La liaison à notre core switch devra être réalisée par le titulaire.

Le soumissionnaire fournira un schéma détaillé dans le mémoire technique, qui mettra en évidence le nombre de câbles entre les switchs et le nombre de ports en attente (cf. 5.1 du CCTP)

Si la bande passante du réseau proposé est supérieure à la bande passante de notre réseau actuel QDR, le titulaire devra :

- remplacer notre core switch principal par un switch à la bande passante proposée ;
- remplacer les cartes hôtes et les liens des quatre serveurs GPFS à la bande passante proposée (4 interfaces Infiniband en tout).

3.3) Réseau Ethernet

Tous les nœuds de calcul, les nœuds GPU, le serveur de login et le serveur de batch seront reliés à un réseau Ethernet, pour permettre entre autres l'accès aux cartes d'administration à distance.

Le réseau Ethernet sera constitué de 6 switchs Gigabit administrables de 48 ports en « top of rack », à raison de 2 switchs par armoire (l'Unistra fournit 2 switchs 48 ports).

Le soumissionnaire a le choix de relier les 8 switchs Gigabit via un switch administrable de 24 ports ou via des uplink 10Gbits.

Les serveurs de login et de batch seront reliés au réseau Ethernet par 2 ports :

- un port pour l'interface réseau du système ;
- un port pour la carte d'administration à distance.

Les nœuds de calcul et GPU seront reliés au réseau Ethernet par 1 port :

- port partagé pour l'interface réseau du système et la carte d'administration à distance.

Ce réseau Ethernet sera relié au réseau de la salle informatique via un câble fourni par le titulaire.

3.4) Cartes d'administration à distance compatible IPMI

Tous les nœuds de calcul, les nœuds GPU ainsi que les serveurs de login et de batch seront munis d'une carte d'administration à distance compatible IPMI de caractéristiques suivantes :

- Adressage IP automatique par DHCP ;
- possibilité de redirection de la console texte du serveur ;
- accès ligne de commande depuis le réseau ;
- possibilité d'allumage / d'extinction / de redémarrage du serveur ;
- possibilité de modification des paramètres du Bios.

L'accès aux cartes d'administration à distance se fera :

- pour les serveurs de login et de batch, via un port Ethernet dédié ;
- pour les nœuds de calcul et les nœuds GPU, via le port Ethernet système.

3.5) Installation et mise en œuvre

L'installation et la mise en œuvre de la solution est soumise aux délais définis au chapitre 7.1 du CCTP. Se référer à l'annexe 1 au CCTP pour la numérotation des armoires informatiques.

En cas de dépassement de ces délais, des pénalités de retard seront appliquées conformément aux conditions décrites dans le CCAP.

3.5.1) Installation et mise en œuvre : Infrastructure

Le titulaire devra intégrer matériellement les serveurs dans nos **4 armoires informatiques réfrigérées**, situées dans une salle au premier étage du bâtiment I.U.F.M., 141 avenue de Colmar, 67100 Strasbourg.

Les armoires sont au format 19'', 42U, référence RITTAL TS8 (7831.812 baie High Performance Cooling), dimensions 600 mm x 2.000 mm x 1.200 mm.

Le poids maximum net admissible par une armoire informatique est de 735 Kg.

La puissance froid de chaque armoire est de 30kW.

Chaque armoire dispose de 2 boîtiers de dérivation au sol contenant chacun une arrivée électrique tri-phasée 32A.

Les bandeaux de prises électriques sont à fournir par le titulaire, et devront être reliés aux boîtiers au sol par le titulaire.
Ces bandeaux devront être équipés de protections par différentiels 30mA.
Dans la mesure du possible, les bandeaux devront être installés verticalement (zéro U).

Une installation dense sera privilégiée. Par exemple les switchs ethernet seront intégrés tête-bêche, deux par deux dans un seul emplacement 1U (dans ce cas, les préconisations du fabriquant des armoires réfrigérées relatives à la circulation des flux d'air seront respectées).

L'ensemble des câbles nécessaires à la solution devra être fourni. Tout le câblage sera réalisé par le titulaire.

Les machines et les câbles devront être étiquetés suivant la convention qui sera fournie par l'Unistra.

Les emballages devront être éliminés par le titulaire.

3.5.2) Installation et mise en œuvre : Pré-configuration

Le titulaire livrera dans un délai de 6 semaines maximum à partir de la date de notification du marché un nœud type pour permettre à l'Unistra de définir les éléments de pré-configuration. Ce délai est compris dans le délai de 8 semaines correspondant au délai de livraison de la solution.

Le titulaire devra :

- Pré-configurer chaque nœud ou serveur pour que le boot réseau soit effectué en priorité ;
- Fournir la liste des adresse MAC des machines sous forme d'un fichier texte ;
- Activer l'ensemble des redirections (Bios en particulier) possibles sur la carte de management de la machine. Cette configuration devra être validée par l'Unistra ;
- paramétrer l'utilisation du port Ethernet partagé avec le système pour la carte IPMI des noeuds de calcul ;
- L'Unistra se réserve le droit d'ajouter des étapes de pré-configuration, dont la liste exhaustive sera définie sur mise à disposition d'une machine type par le titulaire.

3.5.3) Installation et mise en œuvre : Gestionnaire de ressources et de batch

Dans le cas où l'option obligatoire n°1 «Gestionnaire de ressources et de batch» serait notifiée, le titulaire mettra en œuvre sa solution pour l'ensemble des nœuds après déploiement de l'OS par l'Unistra.

3.5.4) Déménagement du cluster 2010

Le titulaire assurera le déménagement du Cluster 2010. L'opération consiste en :

- Déplacement des 56 nœuds du cluster 2010 (33 U) dans les armoires 3 et 4
- Le câblage de ces nœuds sur :
    - En partie sur un des switchs edge du réseau Infiniband fourni par le titulaire.
    - En partie sur un Switch MTS 3600 que nous fournissons, qui sera intégré au niveau edge.
- Le câblage de ces nœuds sur les switchs Gigabit
- Le raccordement de ces nœuds au réseau électrique

3.6) Compilateurs

Nous souhaitons disposer de compilateurs/optimiseurs de dernière génération : C, Fortran, bibliothèques mathématiques, outils de développement multi-thread, parfaitement compatibles avec les processeurs qui seront proposés. Il n'est pas nécessaire de fournir les bibliothèques MPI associées.

La maintenance pour 5 ans doit être incluse.

3.7) Débogueurs parallèles

Nous souhaitons disposer de licences flottantes pour des débogueurs parallèles existants, afin de déboguer des applications MPI/OpenMP jusqu'à 32 processus ainsi que des applications sur GPU.

La maintenance pour 5 ans doit être incluse.

3.8) Option obligatoire n°1 «Gestionnaire de ressources et de batch»

La solution comportera sous forme d'**option obligatoire** un gestionnaire de ressources et de batch. Ce gestionnaire, destiné à remplacer notre solution actuelle Torque/Maui est décrit dans la suite.

Il n'est pas nécessaire d'assurer la compatibilité avec les scripts Torque déjà en place.

Les références d'autres sites du milieu HPC utilisant le logiciel proposé seront fournies, ainsi que les contacts techniques correspondants.

L'installation et la configuration du gestionnaire de ressource et de batch seront assurées par le titulaire du marché.

3.8.1) Fonctionnalités du gestionnaire de ressources

- Intégration fine avec les implémentations MPI *OpenMPI* et *Platform-MPI*. Par intégration fine nous entendons :
    - démarrage des processus sur les nœuds et cœurs choisis par le gestionnaire de batch
    - suppression de l'ensemble des processus à la fin d'un job, ou sur demande de l'utilisateur ou de l'administrateur
    - possibilité d'endormir (suspend) et de réveiller (resume) l'ensemble des processus d'une application parallèle
- Partage des nœuds entre plusieurs jobs :
    - Actuellement les nœuds sont partagés entre plusieurs jobs par la création de *cpusets*. Les mécanismes mis en œuvre par le gestionnaire de ressources devront être décrits.
- Placement des processus
    - A la soumission des jobs, le placement sur les différents sockets et cœurs des nœuds de calcul doit être possible

3.8.2) Fonctionnalités du gestionnaire de batch

Notre configuration actuelle fait un usage intensif du fairshare, défini comme suit. Nous disposons de sous-clusters appelés *Cluster 2009* et *Cluster 2010*. Différents groupes de chercheurs ont financé des parts différentes de ces clusters. À chaque groupe nous restituons, sur une fenêtre de temps donné et pour un sous-cluser donné, un pourcentage de cycles CPU correspondant au pourcentage de machines achetées. Quand les machines ne sont pas utilisées, d'autres utilisateurs n'ayant pas financé de machines peuvent y accéder. Leurs jobs seront *suspendus* si un job provenant d'un chercheur ayant financé les machines a besoin des ressources. À l'heure actuelle, les jobs suspendus sont endormis puis réveillés quand le job prioritaire est terminé. D'autres politiques sont possibles, comme la re-soumission par exemple.

Les groupes de chercheurs sont identifiés par des groupes Unix dans le système d'exploitation. Dans certains cas, ayant financé deux groupes de machines, ils appartiennent à deux groupes et spécifient alors leur groupe secondaire à la soumission d'un job si besoin.

Les clusters 2009 et 2010 sont identifiés par des files d'attente spécifique.

Le gestionnaire de batch proposé doit pouvoir faire cohabiter deux politiques d'exploitations :

- sur les machines déjà en place, la politique présentée ci-dessus ;
- sur les machines fournies dans le cadre de cet appel d'offres, une répartition équitable entre tous les utilisateurs, y compris ceux ayant financé des machines *Cluster 2009* et *Cluster 2010*. La priorité d'un utilisateur dépend donc non seulement de son groupe d'appartenance mais également de la file d'attente utilisée.

3.8.3) Option obligatoire n°2 «Support logiciel et maintenance du Gestionnaire de ressource et de batch»

Le soumissionnaire proposera, pour une durée de 5 ans à partir de l'admission des matériels :

- un support logiciel, sous forme d'une assistance téléphonique ;
- une maintenance incluant les mises à jour mineures et majeures des logiciels.

Le support et la maintenance seront d'un niveau et d'une qualité telle que décrits dans le chapitre 4 du présent CCTP.

La résolution des incidents liés à des bugs logiciels du gestionnaire de ressources et de batch doit se faire dans un délai maximum de 2 semaines après ouverture d'un incident par l'Unistra.

En cas de dépassement de ces délais, des pénalités de retard seront appliquées conformément aux conditions décrites dans le CCAP.

IV. Support et maintenance

4.1) Option obligatoire n°3 « Guichet Unique HPC »

Pour l'ouverture et le suivi des incidents, et au vu de la spécificité d'utilisation des machines, nous souhaitons avoir, sous la forme d'une **option obligatoire**, un accès de type guichet unique avec une **équipe spécialisée dans le domaine HPC**.

En particulier, la configuration logicielle plus matérielle devra être référencée auprès de ce guichet comme un tout et non comme un ensemble de serveurs indépendants.

Nous demandons un accès direct au guichet unique, dont les modalités seront précisées dans la réponse à cet appel d'offre (cf.5.1 du CCTP).

4.2) Maintenance – Garantie

4.2.1) Conditions

Le soumissionnaire fera une proposition de contrat de maintenance matérielle sur les équipements proposés pour une **durée de cinq ans**. Cette période de cinq ans doit englober la période de garantie initiale des matériels, et prend effet à compter de l'admission de la VSR.

La maintenance devra couvrir l'ensemble des pannes matérielles de la solution, avec le remplacement de toute pièce défectueuse, main d'œuvre et déplacements inclus.

Toutes les interventions dans le cadre de la maintenance devront être effectuées **sur site par le titulaire, en particulier les changements de disques**.

La notification d'un incident par l'Unistra au titulaire du marché devra pouvoir se faire soit par téléphone, soit par mail, soit par l'intermédiaire d'un site web. Dans le cas où l'ouverture d'un incident s'accompagne d'un diagnostic de panne réalisé par l'Unistra, le soumissionnaire doit s'engager sur la mise à disposition d'un **circuit court** pour le traitement de l'incident. Les moyens mis en œuvre doivent être décrits dans le mémoire technique.

Par exemple, si l'Unistra appelle pour signaler une panne de disque dur, le titulaire du marché devra immédiatement planifier l'intervention d'un technicien sur site.

Lorsque l'Unistra ouvre un appel sur des éléments de la configuration, son interlocuteur doit immédiatement disposer d'une vue de l'ensemble de la configuration et non du seul élément impliqué. Le soumissionnaire s'engagera sur ce point et décrira les moyens mis en œuvre.

En particulier, le système d'exploitation Linux utilisé sur cette configuration devra être connu des services de support du titulaire.

Pour les serveurs de login et de batch, le contrat sera avec garantie de temps de rétablissement sous 8 heures après ouverture d'un incident.

Pour les équipements réseau et les nœuds CPU et GPU, le contrat sera avec délai maximum d'intervention J+1 après ouverture d'un incident.

Les interventions devront être effectuées de 08h00 à 18h00 les jours ouvrés.

Si la panne concerne un nœud de calcul, celui-ci doit être remplacé intégralement si la remise en service du nœud ne peut se faire dans un délai inférieur ou égal à 15 jours calendaires après ouverture de l'incident.

En cas de dépassement de ces délais, des pénalités de retard seront appliquées conformément aux conditions décrites dans le CCAP.

4.2.2) Partie à bons de commande

Le contrat de maintenance qui couvre la partie forfaitaire doit couvrir également les équipements acquis via la partie à bons de commande.

La facturation de la partie du contrat de maintenance correspondant aux équipements acquis via la partie à bons de commande sera faite au **prorata du temps restant sur les cinq ans** du contrat.

V. Données techniques à fournir

Les documents suivants sont à **compléter** et à retourner avec la proposition :

- « ANNEXE 2 à l'AE - résultats des tests.xls » (voir CCTP 5.3) ;
- « ANNEXE 3 à l'AE - données techniques à fournir.xls » (voir CCTP 5.2) ;

ATTENTION : les documents ci-dessus sont à fournir au format électronique.

Des **archives au format tgz** contenant les résultats de chaque test de performance est également à retourner avec la proposition (voir CCTP 5.3 ci-après).

Un **mémoire technique** au format libre (PDF de préférence) est également à fournir.

Suit dans le détail la description des données à fournir.

5.1) Mémoire technique

Le mémoire technique comprendra entre-autres :

- La description globale de la solution ;
- La description détaillée de chaque composant matériel et logiciel, avec les fiches techniques correspondantes ;
- Un schéma d'implantation physique des matériels dans les armoires informatiques ;
- Le schéma de câblage du réseau Infiniband ;
- La description du guichet unique HPC.

5.2) Synthèse des données techniques

Une synthèse technique sera à fournir dans le document « **ANNEXE 3 à l'AE - données techniques à fournir.xls** » Cette synthèse contiendra notamment :

- Les consommations électriques et dissipations calorifiques armoire par armoire (voir 3.5 du CCTP), pour les deux scénarii suivants :

- **un taux de charge CPU et GPU de 100% ;**
- **un taux de charge CPU et GPU de 80%.**
- Les spécifications techniques et caractéristiques de chaque composant de la solution.

5.3) Tests de performance pour recette de la solution

Il est demandé au soumissionnaire de réaliser des tests sur des machines identiques à celles proposées dans l'offre et sous le système d'exploitation Linux. Le soumissionnaire s'engagera sur les temps d'exécution obtenus, qui seront vérifiés à titre de recette des machines selon le même mode opératoire que celui décrit dans ce document : nous tolérons un écart de +/- 5% sur les temps d'exécution.

En cas de non reproductibilité des résultats, le titulaire aura 15 jours pour appliquer des mesures correctives (matérielles – ajout de serveur ou logicielles). En cas d'impossibilité de reproductibilité des résultats, l'Unistra se réserve le droit de résilier le marché.

Les scripts utilisés pour les tests sont configurés pour regrouper les processus de manière compacte sur les nœuds (tous les cœurs d'un nœud seront utilisés). Ce mode de fonctionnement ne pourra pas être modifié.

Tout problème dans l'exécution des tests devra être signalé au bureau des marchés de l'Unistra.

Résultats consolidés :

Les résultats des tests sont à consolider dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** ». A cet effet, chaque test (Code 1 à Code 5) produit un fichier « *avx.time » et/ou « *noavx.time ». Il faut reporter le meilleur des deux résultats entre les deux versions avx et noavx et indiquer quelle version a été retenue (ligne « Version »).

L'annexe 2 devra être remplie en intégralité (27 zones vertes).

Par ailleurs, le soumissionnaire fournira, pour chaque application, une courbe d'accélération représentant :

- en abscisse, le nombre de cœurs utilisés ;
- en ordonnée, l'accélération par rapport à 8 cœurs.

Résultats détaillés :

Les répertoires contenant les tests pourront être fournis dans des archives, au format tgz, nommées « Resultats_Code_<numéro du test>.tgz ».

Les résultats des tests seront utilisés pour calculer l'accélération CPU globale de la solution : équivalent en nombre de cœurs Intel Xeon 5650 que représentent l'ensemble des nœuds CPU fournis.

Les formules de calcul se trouvent dans l'annexe précitée. Le calcul de l'accélération se base sur des temps de calculs obtenus sur le cluster actuel de l'Unistra.

Paramétrage des nœuds de calculs utilisés pour les tests :

Les fonctionnalités suivantes doivent être **désactivées** sur les nœuds de calcul utilisés pour les tests :

- Turbo boost
- Hyperthreading

MPI :

En cas de besoin, la version de OpenMPI utilisée sur notre site pour faire tourner cette application peut être fournie.

5.3.1) Code 1 – Chimie

- ***Description***

Il s'agit de lancer une application parallèle sur 8 à 64 cœurs. Cette application commerciale, ADF, fonctionnera quel que soit le type de processeur et nous permettra d'évaluer la performance «brute» du processeur. Ce code engendrera un grand nombre de communications entre processeurs.

Pour ce run, il est indispensable de demander une licence de test à l'éditeur à l'adresse mail support@scm.com, en précisant qu'ils s'agit de benchmarks pour l'Université de Strasbourg. L'éditeur est averti de la démarche.

Les exécutables dépendent des bibliothèques des compilateurs Intel. Nous les avons testé avec les compilateurs Intel version 11 et 12. Sur nos systèmes, les dépendances notables sont les suivantes :

libimf.so => /opt/intel/composerxe/lib/intel64/libimf.so

libsvml.so => /opt/intel/composerxe/lib/intel64/libsvml.so

libintlc.so.5 => /opt/intel/composerxe/lib/intel64/libintlc.so.5

- ***Procédure de test***

1. Récupérer l'archive : http://hpc-web.u-strasbg.fr/Equipex/Code_1.tgz.
2. Décompacter l'archive, cela crée un répertoire Code_1
3. cd Code_1
4. Lire le fichier README, qui décrit la procédure de lancement. Créer le ou les fichiers demandés. Si indiqué, modifier les scripts.
5. ./run.adf
6. Reporter les données du fichier « *.time » dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** »
7. Archiver l'ensemble du répertoire Code_1 dans un fichier Resultats_Code_1.tgz

5.3.2) Code 2 – Mécanique des fluides

Description

Il s'agit de lancer une application parallèle sur 8 à 64 cœurs. Nous fournissons deux versions de l'exécutable :

- nsmb.noavx compilée sans avx

- nsmb.avx, compilée avec avx

Les deux versions de l'application ont été compilées avec les compilateurs intel version 12.0, et la MKL correspondante. Sur nos systèmes, les dépendances notables sont les suivantes :

libgfortran.so.1 => /usr/lib64/libgfortran.so.1

libmkl_intel_lp64.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_lp64.so

libmkl_sequential.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_sequential.so

libmkl_core.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_core.so

libmpi_f90.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f90.so.0

libmpi_f77.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f77.so.0

libmpi.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi.so.0

libopen-rte.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-rte.so.0

libopen-pal.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-pal.so.0

libm.so.6 => /lib64/libm.so.6

libpthread.so.0 => /lib64/libpthread.so.0

libifport.so.5 => /opt/intel/composerxe/lib/intel64/libifport.so.5

libifcoremt.so.5 => /opt/intel/composerxe/lib/intel64/libifcoremt.so.5

libimf.so => /opt/intel/composerxe/lib/intel64/libimf.so

libsvml.so => /opt/intel/composerxe/lib/intel64/libsvml.so

libintlc.so.5 => /opt/intel/composerxe/lib/intel64/libintlc.so.5

- ***Procédure de test***

1. Récupérer l'archive : http://hpc-web.u-strasbg.fr/Equipex/Code_2.tgz.
2. Décompacter l'archive, cela crée un répertoire Code_2
3. cd Code_2
4. Lire le fichier README, qui décrit la procédure de lancement. Créer le ou les fichiers demandés. Si indiqué, modifier les scripts.
5. ./run.nsmb
6. Reporter les données du fichier « *.time » dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** »
7. Archiver l'ensemble du répertoire Code_2 dans un fichier Resultats_Code_2.tgz

5.3.3) Code 3 – Simulation moléculaire

- ***Description***

Il s'agit de lancer l'application CHARMM en parallèle sur 8 à 64 cœurs.

Les deux versions de l'application ont été compilées avec les compilateurs intel version 12.0, et la MKL correspondante.

Sur nos systèmes, les dépendances notables sont les suivantes :

libmpi_f90.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f90.so.0

libmpi_f77.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f77.so.0

libmpi.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi.so.0

libopen-rte.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-rte.so.0

libopen-pal.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-pal.so.0

libifport.so.5 => /opt/intel/composerxe/lib/intel64/libifport.so.5

libifcoremt.so.5 => /opt/intel/composerxe/lib/intel64/libifcoremt.so.5

libimf.so => /opt/intel/composerxe/lib/intel64/libimf.so

libsvml.so => /opt/intel/composerxe/lib/intel64/libsvml.so

libintlc.so.5 => /opt/intel/composerxe/lib/intel64/libintlc.so.5

libiomp5.so => /opt/intel/composerxe/lib/intel64/libiomp5.so

libmkl_intel_lp64.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_lp64.so

libmkl_intel_thread.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_thread.so

libmkl_core.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_core.so

- ***Procédure de test***

1. Récupérer l'archive : http://hpc-web.u-strasbg.fr/Equipex/Code_3.tgz
2. Décompacter l'archive, cela crée un répertoire ***Cod***e_3
3. cd Code_3
4. Lire le fichier README, qui décrit la procédure de lancement. Modifier le script run.charmm
5. ./run.charmm
6. Reporter les données du fichier « *.time » dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** »
7. Archiver l'ensemble du répertoire Code_3 dans un fichier Resultats_Code_3.tgz

5.3.4) Code 4 – Simulation moléculaire

- ***Description***

Il s'agit de lancer l'application CPMD en parallèle sur 8 à 64 cœurs.

Les deux applications ont été compilées avec les compilateurs intel version 12.0, et la MKL correspondante.

Sur nos systèmes, les dépendances notables sont les suivantes :

libmpi_f90.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f90.so.0

libmpi_f77.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f77.so.0

libmpi.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi.so.0

libopen-rte.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-rte.so.0

libopen-pal.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-pal.so.0

libmkl_intel_lp64.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_lp64.so

libmkl_intel_thread.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_thread.so

libmkl_core.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_core.so

libiomp5.so => /opt/intel/composerxe/lib/intel64/libiomp5.so

libifport.so.5 => /opt/intel/composerxe/lib/intel64/libifport.so.5

libifcoremt.so.5 => /opt/intel/composerxe/lib/intel64/libifcoremt.so.5

libimf.so => /opt/intel/composerxe/lib/intel64/libimf.so

libsvml.so => /opt/intel/composerxe/lib/intel64/libsvml.so

- ***Procédure de test***

1. Récupérer l'archive : http://hpc-web.u-strasbg.fr/Equipex/Code_4.tgz
2. Décompacter l'archive, cela crée un répertoire **Cod**e_4
3. cd Code_4
4. Lire le fichier README, qui décrit la procédure de lancement. Modifier le script run.cpmd
5. ./run.cpmd
6. Reporter les données du fichier « *.time » dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** »
7. Archiver l'ensemble du répertoire Code_4 dans un fichier Resultats_Code_4.tgz

5.3.5) <u>Code 5 – Dynamique moléculaire</u>

- ***Description***

Il s'agit de lancer l'application PMEMD en parallèle sur 8 à 64 cœurs.

Les deux applications ont été compilées avec les compilateurs intel version 12.0, et la MKL correspondante.

Sur nos systèmes, les dépendances notables sont les suivantes :

libmpi.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi.so.0

libmkl_intel_lp64.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_lp64.so

libmkl_intel_thread.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_intel_thread.so

libmkl_core.so => /opt/intel/composerxe/mkl/lib/intel64/libmkl_core.so

libmpi_f90.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f90.so.0

libmpi_f77.so.0 => /usr/local/openmpi-1.4.i11/lib/libmpi_f77.so.0

libopen-rte.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-rte.so.0

libopen-pal.so.0 => /usr/local/openmpi-1.4.i11/lib/libopen-pal.so.0

libiomp5.so => /opt/intel/composerxe/lib/intel64/libiomp5.so

libimf.so => /opt/intel/composerxe/lib/intel64/libimf.so

libsvml.so => /opt/intel/composerxe/lib/intel64/libsvml.so

libintlc.so.5 => /opt/intel/composerxe/lib/intel64/libintlc.so.5

libifport.so.5 => /opt/intel/composerxe/lib/intel64/libifport.so.5

libifcoremt.so.5 => /opt/intel/composerxe/lib/intel64/libifcoremt.so.5

- ***Procédure de test***

1. Récupérer l'archive : http://hpc-web.u-strasbg.fr/Equipex/Code_5.tgz
2. Décompacter l'archive, cela crée un répertoire **Cod**e_5
3. cd Code_5
4. Lire le fichier README, qui décrit la procédure de lancement. Modifier le script run.pmemd
5. ./run.sander
6. Reporter les données du fichier « *.time » dans le document « **ANNEXE 2 à l'AE - résultats des tests.xls** »
7. Archiver l'ensemble du répertoire Code_5 dans un fichier Resultats_Code_5.tgz

VI. Organisation du marché

Le marché comporte :

- une partie forfaitaire constituée de l'ensemble des matériels, logiciels et prestations décrites aux chapitres 3 et 4 du présent CCTP. Montant prévisionnel pour la partie forfaitaire : 700 000 €. ;
- une partie à bons de commande. La durée du marché à bons de commande sera de 18 mois à compter de la notification.

VII. Délais et recette

7.1) Délais

Le délai de livraison est fixé à :

- Pour la partie forfaitaire : 8 semaines maximum à partir de la date de la notification du marché. Ce délai recouvre le délai de 6 semaines correspondant au délai de livraison du nœud type. Des pénalités de retard seront appliquées conformément aux conditions décrites dans le CCAP ;
- Pour les équipements acquis via le marché à bons de commande : 4 semaines maximum à partir de la commande.

Le délai pour l'installation et la mise en œuvre est fixé à :

- Pour la partie forfaitaire : 4 semaines maximum à partir de la date de livraison. Des pénalités de retard seront appliquées conformément aux conditions décrites dans le CCAP ;
- Pour les équipements acquis via le marché à bons de commande : 2 semaines maximum à partir de la date de la livraison.

7.2) Recette (VA – VSR)

La recette de la solution se décomposera de la manière suivante :

- Une vérification d'aptitude au bon fonctionnement (VA) qui permettra la validation des fonctionnalités et des équipements conformément au CCTP. Cette vérification se fera dans un délai de 15 jours maximum à compter de la date de fin d'installation et de mise en œuvre des équipements.
  Les vérifications porteront notamment sur :
- les tests de performances, qui seront rejoués par l'Unistra dans les conditions décrites dans le paragraphe 5.3 (en cas de non reproductibilité, le titulaire aura 15 jours pour prendre des mesures correctives).
- le bon fonctionnement des interfaces IPMI ;
- la bonne intégration des nœuds au gestionnaire batch

  Un procès-verbal attestera de la VA.

- Une vérification de service régulier (VSR), durant une période de 3 mois maximum à compter de la VA, pendant laquelle les équipements de la solution devront fonctionner de manière satisfaisante. En particulier, nous exécuterons une application utilisant l'ensemble des nœuds de calcul acquis dans le cadre du présent appel d'offre (grand challenge). Un procès-verbal attestera de la VSR.

VIII. Critères de jugement des offres

- Critères techniques : 70% ;
  - Accélération CPU globale de la solution : équivalent en nombre de cœurs Intel Xeon 5650 que représentent l'ensemble des nœuds CPU fournis (y compris les nœuds équipés de GPU). Ce nombre est calculé sur la base des résultats des tests. La formule de calcul se trouve dans le document « ANNEXE 2 à l'AE - résultats des tests.xls » : 60% ;
  - Débit total sur l'ensemble des bus PCI Express disponible pour les cartes GPU : 10% ;
  - Facteur de blocage du réseau Infiniband : 5% ;
  - Rapport « Débit global théorique du réseau Infiniband (sur le total des ports utilisés par les noeuds) / nombre de cœurs total de l'ensemble des nœuds CPU (y compris les nœuds équipés de GPU) » : 5% ;
  - Consommation électrique de l'ensemble de la solution à 80% de charge CPU et GPU : 5% ;
  - Occupation totale en nombre de U : 5% ;
  - Mémoire technique + annexe 3 à l'AE : 10% ;
- Prix : 30%.