Mise en place d'un supercalculateur dans un centre informatique.

1 Modalités pratiques

Ce projet est destiné à vous faire réfléchir sur tout ce que vous avez vu pendant les cours et les travaux pratiques de l'UE Logiciel Cluster.

Merci de rendre un dossier bien rédigé sans fautes d'orthographe (avec les correcteurs orthographiques, c'est possible !) avec des schémas si nécessaire et des explications détaillées de vos choix. Vous pouvez poser des hypothèses supplémentaires pour appuyer vos solutions.

Attention aux unités :

- Gb = Gigabit
- Go = Gigaoctet = GB = Gigabyte

Reprenez la structure du document en répondant aux questions qui vous sont posées.

Si vous avez des questions, des doutes , vous pouvez nous contacter par mail. Si vous pensez avoir trouvé une erreur, merci de nous la signaler.

Prenez le soin de lire deux fois le document avant de vous lancer. Les réponses attendues ne nécessitent pas de recherche sur Internet, tout a été vu en cours et en TP. Inspirez vous en. Vous pouvez néanmoins laisser libre cours à la réflexion et l'imagination. Nous apprécions les propositions originales.

Une fois toutes les parties traitées, effectuer plusieurs relectures pour vérifier la cohérence de l'ensemble.

Toutes les parties sont notées. Un partie non rendue est évidemment notée 0. C'est un travail personnel, ne recopiez pas le travail d'un autre. Une partie recopiée d'un autre projet est aussi notée 0.
La note finale est une moyenne de toutes les notes des parties.

N'attendez pas la dernière semaine pour travailler. Un travail régulier et quotidien est la garantie d'un travail de qualité et d'une bonne note finale.

Je vous rappelle que ce projet compte pour coefficient 2 dans la note de l'UE et qu'il n'y a pas de rattrapage.

Vous devez nous retourner la première partie [4.1-4.7] **avant le lundi 22 avril midi**, puis nous envoyer le document complet avant le **vendredi 10 mai 2024, midi**. Envoyer les documents nommées TP-00-AppelOffre-2024-<Noms>.pdf au format PDF à lc.ensiie@gmail.com, copie à philippe.gregoire@cea.fr, lise.jolicoeur@cea.fr, martial.lameth@cea.fr, emmanuel.penot@cea.fr.

2 Cahier des charges de l'appel d'offres

2.1 Cahier des charges

L'université 3 de Toulouse a obtenu le financement d'un supercalculateur pour répondre à ses nouvelles missions :

- Accueillir des collaborations universitaires,
- Accueillir des collaborations avec des startups et des PMEs locales.

Après avoir rencontré ses futurs utilisateurs et discuté de leurs besoins, l'université a rédigé un cahier des charges dont les principaux points sont :

- une partition AmdCpu (AC) de nœuds de calcul d'environ 40000 cœurs avec au moins 2 Go de mémoire par cœur,
- une partition AmdLarge (AL) de nœuds de calcul d'environ 4000 cœurs avec au moins 4 Go de mémoire par cœur,
- une partition AmdGpu (AX) d'au moins 3000 cœurs avec au moins 2 Go de mémoire par cœur, permettant d'exécuter des codes CUDA sur cartes NVIDIA et/ou d'utiliser des solutions de visualisation à distance. Cette partition doit offrir l'accès à au moins 60 cartes NVIDIA.
- une partition frontale (F) contenant un nombre suffisant de nœuds de login pour séparer les différentes communautés d'utilisateurs et leur offrir suffisamment de puissance de calcul interactive pour développer, compiler et déverminer leurs applications, exécuter des outils de post-traitement. Chaque nœud doit disposer de 24 cœurs, 2 Go de mémoire par cœur, 2 disques sécurisés pour le système avec espace temporaire (/tmp) d'au moins 3 To au total. Les administrateurs prévoient de les relier directement connectés au backbone 10Gb Ethernet de l'université.
- un débit I/O global d'au moins 140 Go/s vers un cluster de stockage Lustre (qui a fait l'objet d'un autre appel d'offres). Le réseau de stockage Lustre est un réseau Infiniband HDR 200Gbs.
- un débit I/O global d'au moins 60 Gbits/s vers le backbone pour les nœuds des partitions AC, AL, AX,
- une partition Admin/Service (Srv) contenant au moins 10 nœuds de services intégrés au cluster pour installer des services nécessaires au fonctionnement du cluster et à son intégration dans le centre informatique de l'université et/ou de futurs services. Ces nœuds doivent disposer de
  - 2 disques sécurisés de 1 To pour le système ,
  - 2 disques sécurisés de 4To pour les données système,
  - au moins 16 cœurs et 256 Go de mémoire par nœud.
- une offre logicielle à base de RHEL9 dans un souci d'homogénéité avec le matériel déjà en place,
- des outils de développement et de profiling adaptés aux processeurs sélectionnés.

Les différentes communautés d'utilisateurs ont été organisées en groupes Unix nommés ugpXX. Certains groupes peuvent cohabiter ensemble, d'autres non pour des raisons de sécurité et de confidentialité. Les noms des groupes ont été choisis de façon à ce que des rangs consécutifs puissent cohabiter sur des nœuds de login. En fonction du nombre de personnes dans les différents groupes et de leurs besoins, l'université a prévu de regrouper comme suit les différentes communautés d'utilisateurs sur les nœuds de login :

Groupes	Nombre de nœuds de login	Nombre de personnes en tout
ugp[0-2]	2	40
ugp[3-4]	3	80
ugp[5-7]	2	30
ugp[8-9]	3	70

L'université acceptera des offres divergeant de 5% par rapport aux nombres demandés concernant le nombre de cœurs des partitions AmdCpu, AmdLarge et AmdGpu, la capacité mémoire pour toutes les partitions, les volumes des disques et les performances réseau BB et Lustre.

2.2 Description du centre informatique de l'université

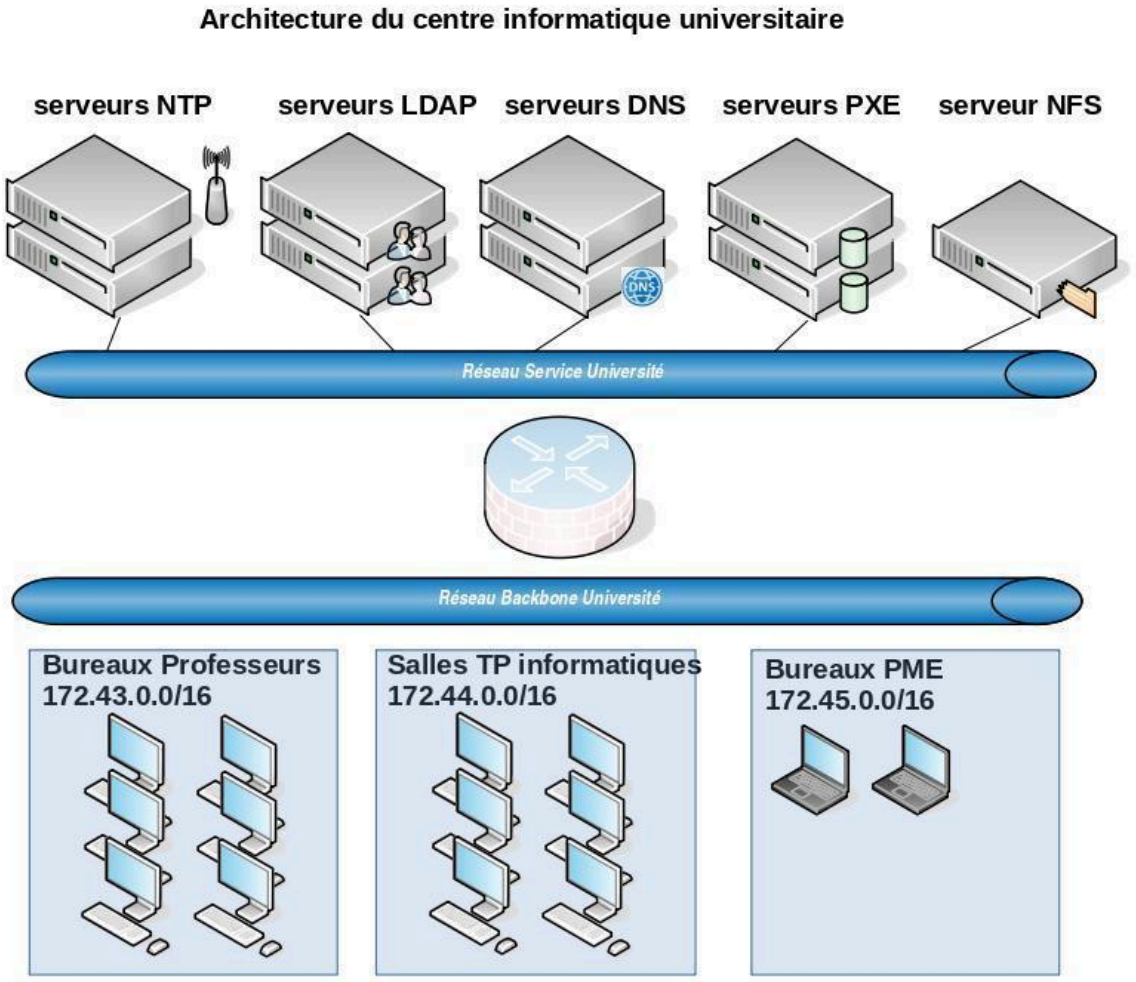
L'université dispose déjà d'un centre informatique hébergeant des services opérationnels :

- Deux serveurs LDAP : montcalm1.univ-tlse3.fr montcalm2.univ-tlse3.fr
- Deux serveurs DNS : marbore1.univ-tlse3.fr marbore2.univ-tlse3.fr
- Deux serveurs NTP reliés à un boitier GPS-NTP NTS-6002-GPS-WWVB: ossau1.univ-tlse3.fr ossau2.univ-tlse3.fr
- Un serveur NFS maladeta.univ-tlse3.fr qui contient les répertoire HOME des utilisateurs.
- Deux serveurs d'installation PXE pour installer automatiquement les stations Linux des salles d'études, de travaux pratiques et les bureaux des professeurs : perdu1.univ-tlse3.fr perdu2.univ-tlse3.fr. Ces serveurs ont déjà un système Cobbler opérationnel avec des dépôts RHEL9,
- Des stations Linux présentes dans les bureaux des professeurs, dans les salles de TPs pour les élèves, et des salles réservées à l'accueil de start-ups

Liste des adresses utilisées :

montcalm[1-2].univ-tlse3.fr	172.4.24.[11-12]
marbore[1-2].univ-tlse3.fr	172.4.24.[13-14]
ossau [1-2].univ-tlse3.fr	172.4.24.[15-16]
perdu[1-2].univ-tlse3.fr	172.4.24.[17-18]
maladeta.univ-tlse3.fr	172.4.24.19
Bureaux professeurs	172.43.0.0/16
Salles TPs	172.44.0.0/16
Bureaux PME	172.45.0.0/16

Toutes ces machines sont connectées à un switch Ethernet 1/10Gb configurés avec plusieurs VLANS pour bien compartimenter les différentes populations.



L'université souhaite que les utilisateurs du cluster puissent accéder à leur répertoire HOME à partir des nœuds de logins du cluster .

3 Réponse à l'appel d'offres

Après sélection des candidatures, remise des réponses techniques et financières, la société AzkarHPC a remporté l'appel d'offres en s'appuyant sur la dernière génération des processeurs AMD Genoa EPYC™ 9004 Series qu'elle a intégrée dans sa propre gamme de serveurs HPC:

Extrait de https://www.amd.com/en/products/specifications/processors?s_family%5B%5D=23336&s_series%5B%5D=25666

	Nb de cœurs	Nb de threads	Fréquence	Nb max de sockets	Cache L3	Débit mémoire par socket	Prix
9754	128	256	2,25GHz	2	256MB	460.8 GB/s	11900$
9754S	128	128	2,25GHz	2	256MB	460.8 GB/s	10200$
9734	112	224	2,20GHz	2	256MB	460.8 Go/s	9600$
9634	84	168	2.25GHz	2	384MB	460.8 Go/s	10304$
9554	64	128	3.1GHz	2	256MB	460.8 Go/s	9087$
9454	48	96	2,75GHz	2	256MB	460.8 Go/s	5225$
9334	32	64	2,7GHz	2	128MB	460.8 Go/s	2990$
9224	24	48	2,5GHz	2	64MB	460.8 Go/s	1825$

La réponse technique de la société AzkarHPC est la suivante :

3.1 Synthèse de la réponse

La partition AmdCpu (AC) compte 302 nœuds , avec par nœud : 1 processeurs AMD 9754, 256 Go mémoire DDR5, 1 port Infiniband HDR200, 1 port Ethernet 1Gbit, 1 port Ethernet BMC. aucun disque,

La partition AmdLarge (AL) compte 50 nœuds , avec par nœud : 1 processeurs AMD 9634, 512 Go mémoire DDR5, 1 port Infiniband HDR200, 1 port Ethernet 1Gbit, 1 port Ethernet BMC. aucun disque,

La partition AmdGpu (AX) compte 50 nœuds , avec par nœud: 1 processeurs AMD 9224, , 256 Go mémoire DDR5, 1 disque de 1 To, 1 carte GPU Nvidia A100 , 1 port Infiniband HDR200, 1 port Ethernet 10Gbit, 1 port Ethernet BMC, Aucun disque.

Les nœuds des partitions AC, AL et AX sont diskless.

La partition Admin/Service (Srv) compte :

- 2 nœuds maîtres pour déployer et administrer le cluster, avec par nœud : 2 processeurs AMD 9224, 256 Go mémoire DDR5, 1 port Infiniband HDR200, 4 ports Ethernet 1Gb/s, 4 ports Ethernet 10Gb, 1 port Ethernet BMC séparé, 1 contrôleur RAID, 2 disques de 3To,
- 9 nœuds de service cluster pour héberger les services internes nécessaires à son administration, avec par nœud : 2 processeurs AMD 9224, 256 Go mémoire DDR5, 1 port Infiniband HDR200, 2 ports Ethernet 1Gb/s, 4 ports Ethernet 10 Gb, 1 port Ethernet BMC séparé, 1 contrôleur RAID, 2 disques de 3To,
- 9 nœuds routeurs I/O pour la connexion au système de stockage Lustre et flux Backbone,

avec par nœud : 2 processeurs AMD 9224, 256 Go mémoire DDR5, 2 ports Infiniband HDR200, 2 ports Ethernet 1Gb/s, 6 ports Ethernet 10 Gb, 1 port Ethernet BMC séparé, 1 contrôleur RAID, 2 disques de 1To.

La partition frontale (F) contient 10 nœuds. Chaque nœud comporte: 1 processeurs AMD 9634, 96 Go mémoire DDR5, 1 port Infiniband HDR200, 2 ports Ethernet 1Gb/s, 2 ports Ethernet 10 Gb/s, 1 port Ethernet BMC séparé, 1 contrôleur RAID, 2 disques de 4 To.

La société AzkarHPC a intégré aussi un serveur de stockage NFS hautes performances de 300 To pour vous permettre de stocker les données système nécessaires à la gestion du cluster. Ce serveur intègre 4 ports 10 Gb Ethernet, 1 port ethernet BMC séparé, et 30 disques de 18 To. Il supporte une grande variété de protocoles dont iscsi.

L'offre logicielle comprend une licence RHEL9, les rpms Lustre pour les nœuds clients et routeurs, des compilateurs Intel et AMD, un debugger Totalview et des logiciels OpenSource.

3.2 Racking

Les nœuds de la partition AC et AL sont constitués par des lames AZ-L30 intégrées dans des châssis AZ-B30.

Un châssis AZ-B30 accueille jusqu'à 30 lames AZ-L30 et intègre un fond de panier qui fournit l'interconnexion IPMI, Ethernet et Infiniband HDR200. En effet chaque châssis AZ-B30 intègre trois switchs dans le fond de panier :

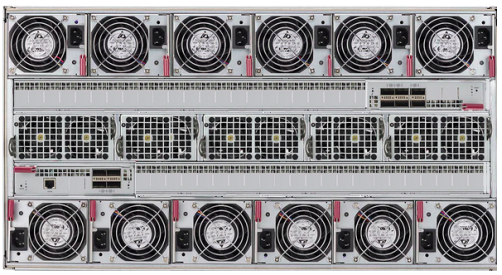
- un switch Infiniband HDR200 de 40 ports,
- un switch Ethernet 100Mb de 36 ports internes + 2 ports 1Gb externes
- un switch Ethernet 1Gb de 36 ports internes + 4 ports externes 10Gb.

Chaque lame se connecte donc au fond de panier via 3 ports qu'elle intègre:

- 1 port ethernet 100Mb IPMI,
- un port ethernet 1Gb (administration),
- et un port Infiniband HDR200.

Pour les connexions externes, le châssis AZ-B30 utilise lui-même un port 1Gb ethernet (IPMI), 4 ports Ethernet 10 Gb (administration) et au maximum 10 ports Infiniband HDR200. La BMC du châssis permet de contrôler individuellement chaque lame/nœud.

Châssis AZ-B30 - vue avant



Châssis AZ-B30 - vue arrière

Un châssis AZ-B30 occupe 16U dans un rack. Les nœuds de la partition AX et Srv sont des serveurs classiques 2U. Le serveur NFS occupe aussi 2U.

Les switchs Ethernet de contrôle AZ-W40-1g et d'administration AZ-W40-10g ainsi que les switchs Infiniband AZ-W40-IB occupent aussi 2U.

La société AzkarHPC prévoit de livrer des racks 48 U:

- 4 racks 48U pour la partition AC . Ces rack nommés rackAC[1-4] contiennent les châssis de la partition AC ,
- 1 racks 48U pour la partition AL. Ces rack nommés rackAL[1- 1] contiennent les châssis de la partition AL,
- 3 racks 48U pour la partition AX. Ces rack nommés rackAX[1-3] contiennent les serveurs de la partition AX,
- 2 racks pour les nœuds de service . Ces racks nommés rackSrv[1-2] contiennent tous les nœuds de service (administration, routeurs, login, serveur NFS),
- 1 rack(s) pour les switchs ethernet et infiniband non intégrés dans les châssis AZ-B30, nommé(s) rackW[1-1]. (rack sWitch).

Chaque rack intègre deux alimentations redondantes placées verticalement de chaque côté du rack ce qui permet de fiabiliser l’alimentation électriques des équipements (PDU)

Les alimentations électriques (PDU) des rack sWitch rackW[1-1] intègrent en plus un contrôleur intelligent et pilotable par Ethernet (via le protocole SNMP) ce qui permet de couper indépendamment l'alimentation de chaque prise électrique. Ceci permet de contrôler électriquement les équipements n'intégrant pas un contrôle électrique par Ethernet/IPMI comme les switchs Infiniband et les switchs Ethernet.

Les alimentations sont nommées rackW[1-1]pdu[1-2].

3.3 Architecture réseau du cluster

3.3.1 Réseau Ethernet de contrôle (IPMI et SNMP)

Ce réseau Ethernet 1Gb/s permet de contrôler électriquement :

- tous les serveurs Srv via leurs ports BMC sauf les nœuds maîtres,
- le serveur NFS via son port BMC,
- tous les nœuds de login F via leurs ports BMC,
- les nœuds des partitions AC, AL via leurs châssis AZ-B30,
- les nœuds des partitions AX via leurs ports BMC,
- tous les switchs IB et Ethernet externes via leurs alimentations reliées à des PDU contrôlables par Ethernet.

Le port BMC de chaque nœud maître est branché sur un port 1Gb de l'autre nœud maître. Ainsi il est possible de contrôler un nœud maître à partir de l'autre.

Les switchs IB externes aux châssis et les switchs Ethernet n'ont pas de port BMC et sont contrôlés électriquement via les PDUs sur lesquelles ils sont branchés.

La société AzkarHPC a choisi de vous fournir 3 switchs AZ-W40-1g de 40 ports 1Gb pour ce réseau. Ces switchs Ethernet sont empilés et interconnectés de façon à former un switch virtuel. La perte d'un switch entraîne la perte de ses ports mais laisse les autres switchs fonctionnels.

Les switchs de ce réseau sont nommés ewc[1-3]. Les ports 1 Gb sont numérotés de 1 à 40. On désignera le port P du switch N de contrôle par ewcNpP. Par exemple , le port 3 du second switch de contrôle est nommé ewc2p3. Les ports 10 à 17 de ce switch sont désignés par ewc2p[10-17].

Tous les switchs ewc[1-4] sont intégrés dans le rack rackW1.

La société AzkarHPC livre tous les câbles Ethernet 1Gb nécessaires au réseau de contrôle.

3.3.2 Réseau Ethernet d'administration

Tous les nœuds du cluster sont interconnectés par un réseau de services d'administration 10 Gb Ethernet, soit indirectement par leur châssis pour les nœuds AZ-L30, soit directement pour les autres nœuds. Ce réseau est utilisé pour installer les nœuds et/ou faire ensuite passer tous les services d'administration. Le serveur NFS fourni par la société AzkarHPC est aussi connecté sur ce réseau par tous ses ports 10Gb.

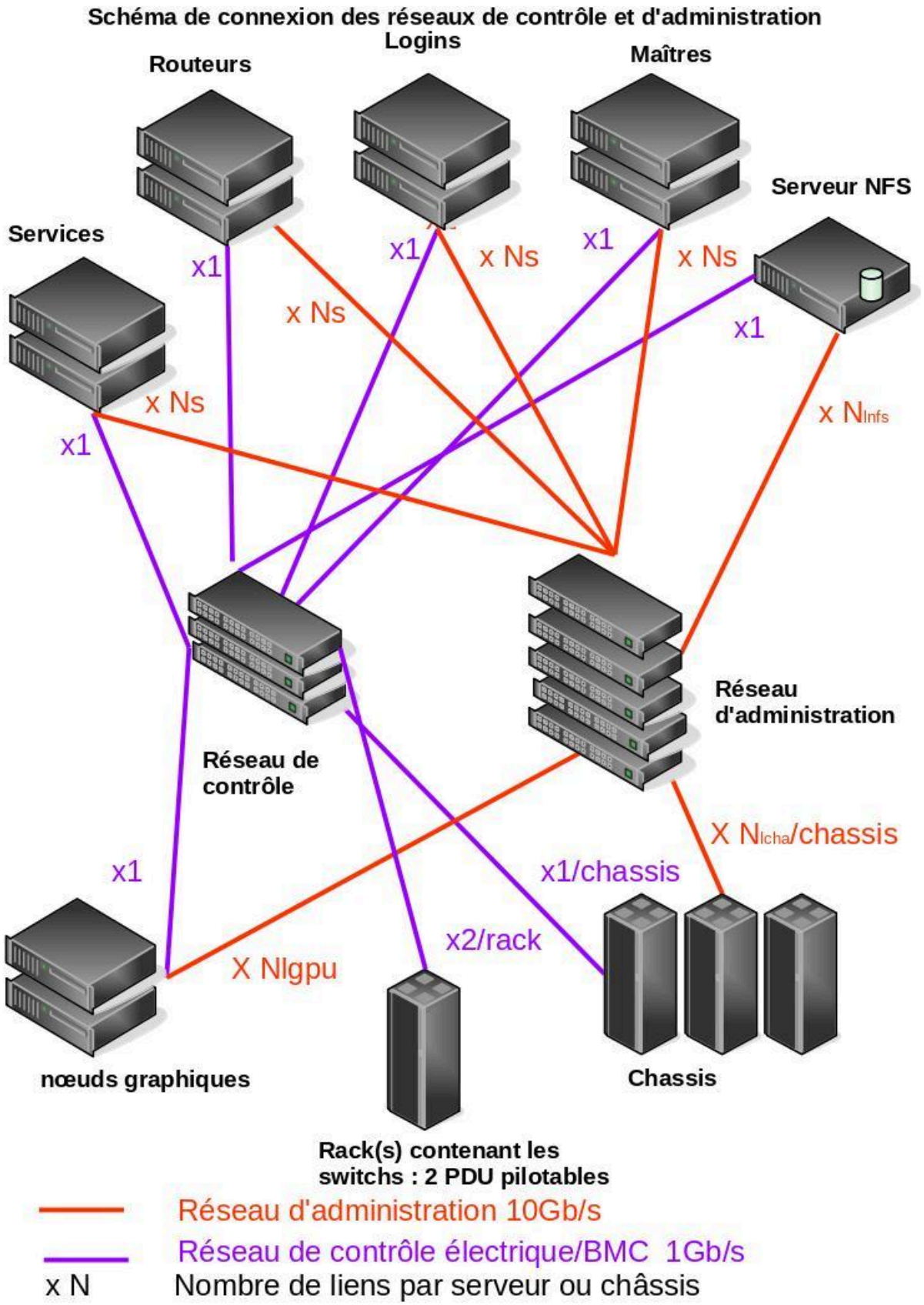
La société AzkarHPC a choisi de vous fournir 6 switchs AZ-W40-10g de 40 ports 10Gb pour ce réseau. Ces switchs Ethernet sont empilés et de façon à former un switch virtuel. La perte d'un switch entraîne la perte de ses ports mais laisse les autres switchs fonctionnels.

Les switchs de ce réseau sont nommés ewa[1- 6]. Les ports 10Gb sont numérotés de 1 à 40. On désignera le port P du switch N de contrôle par ewaNpP. Par exemple , le port 4 du premier switch de contrôle est nommé ewa1p4.

Tous les switchs ewa[1-5] sont intégrés dans le rack rackW1

Le serveur NFS est relié par ses 4 ports 10Gb à ce réseau.
Chaque nœud de la partition Admin/Service (Srv) est relié par 2 câbles 10Gb à ce réseau.
Chaque nœud de la partition frontale (F) est relié par 2 câbles 10Gb à ce réseau.
Chaque nœud de la partition AmdGpu (AX) est relié par 2 câbles 10Gb à ce réseau.
Chaque châssis de la partition AmdCpu (AC) est relié par 4 câbles 10Gb à ce réseau.
Chaque châssis de la partition AmdLarge (AL) est relié par 4 câbles 10Gb à ce réseau.

La société AzkarHPC livre tous les câbles Ethernet 10Gb nécessaires au réseau d'administration.



3.3.3 Réseau d'interconnexion Infiniband

Le réseau d'interconnexion est constitué de switchs AZ-W40-IB HDR200 avec 40 ports.

Tous les nœuds du cluster sont donc interconnectés par un réseau Infiniband HDR200 de topologie Fat Tree constitués par les 17 switchs de niveau L1 (Leaf switch) et 6 switchs de niveau L2 AZ-W40-IB (Top switchs). Chaque switch de niveau L1 est raccordé à jusqu'à 30 nœuds et aux 6 switchs de niveau L2 (Top switchs)

Les switchs de niveau L1 (Leaf switch) pour les nœuds des partitions AC et AL sont intégrés dans les châssis AZ-B30. Il reste donc 10 ports dans chaque châssis. AzkarHPC a choisi de câbler 6 ports sur chacun des 6 switchs de niveau L2 (Top switchs).

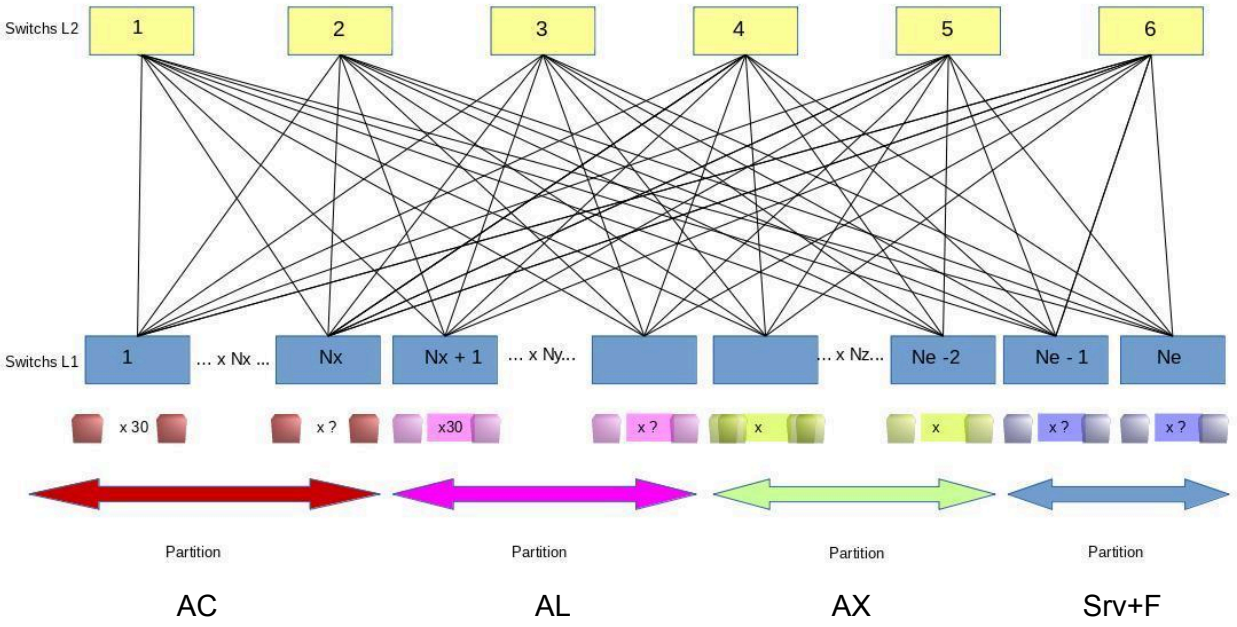
AzkarHPC a fait le choix de connecter les nœuds de la partition AX sur des switchs L1 dédiés. Les nœuds de service et les nœuds de login sont connectés sur deux switchs L1 dédiés.

Les switchs AZ-W40-IB L1 des partitions AX et Srv ainsi que les switchs AZ-W40-IB L2 (Top switchs) sont intégrés dans le rack rackW1. Le contrôle électrique (power on/off) de ces switchs HDR200 Infiniband se fait via les alimentations électriques (PDU) sur lesquelles ils sont branchés dans les racks et qui sont pilotables par Ethernet.

Il y a 17 switchs L1, nommés iwh1n[1-17]. Les switchs L2 sont nommés iwh2n[1-6]. Les ports HDR200 sont numérotés de 1 à 40. On désignera le port P du switch Infiniband N de niveau L par iwhLnNpP. Par exemple , le port 14 du cinquième switch de niveau 1 est nommé iwh1n5p14.

La société AzkarHPC livre tous les câbles Infiniband nécessaires au réseau Infiniband HDR200.

Schéma du réseau d'interconnexion Fat Tree Infiniband HDR du cluster



3.3.4 Connexion au réseau de stockage Lustre et au réseau Backbone

La société AzkarHPC a prévu de mutualiser les fonctions de routeurs Lustre et routeurs Backbone Ethernet sur les 9 nœuds routeurs I/O.

Pour la fonction Lustre, ces routeurs seront donc connectés par un port IB HDR200 sur le futur réseau de stockage Infiniband HDR200 et par l'autre port IB HDR200 sur le réseau d'interconnexion du cluster de calcul.

Pour la fonction Backbone, ces routeurs seront donc connectés par 2 port(s) 10Gb au backbone Ethernet de l'université. Le routage Backbone pour les nœuds des partitions AC, AL, AX se fera à travers leur port IB (via un protocole IP over IB) . Pour être plus explicite, un flux backbone sera routé depuis le nœud de calcul via son port IB vers un port IB d'un nœud routeur sur le réseau d'interconnexion du calculateur puis du port ethernet 10Gb du routeur vers le backbone de l'université.

Les cartes Infiniband fournies sont des cartes ConnectX6 Mellanox HDR200 200 Gbits. Dû à un encodage 64/66bits, le débit réel est de (200*64/66)/8 = 24 Go/s.

Les cartes 10Gb Ethernet ont un débit réel de 1 Go/s.

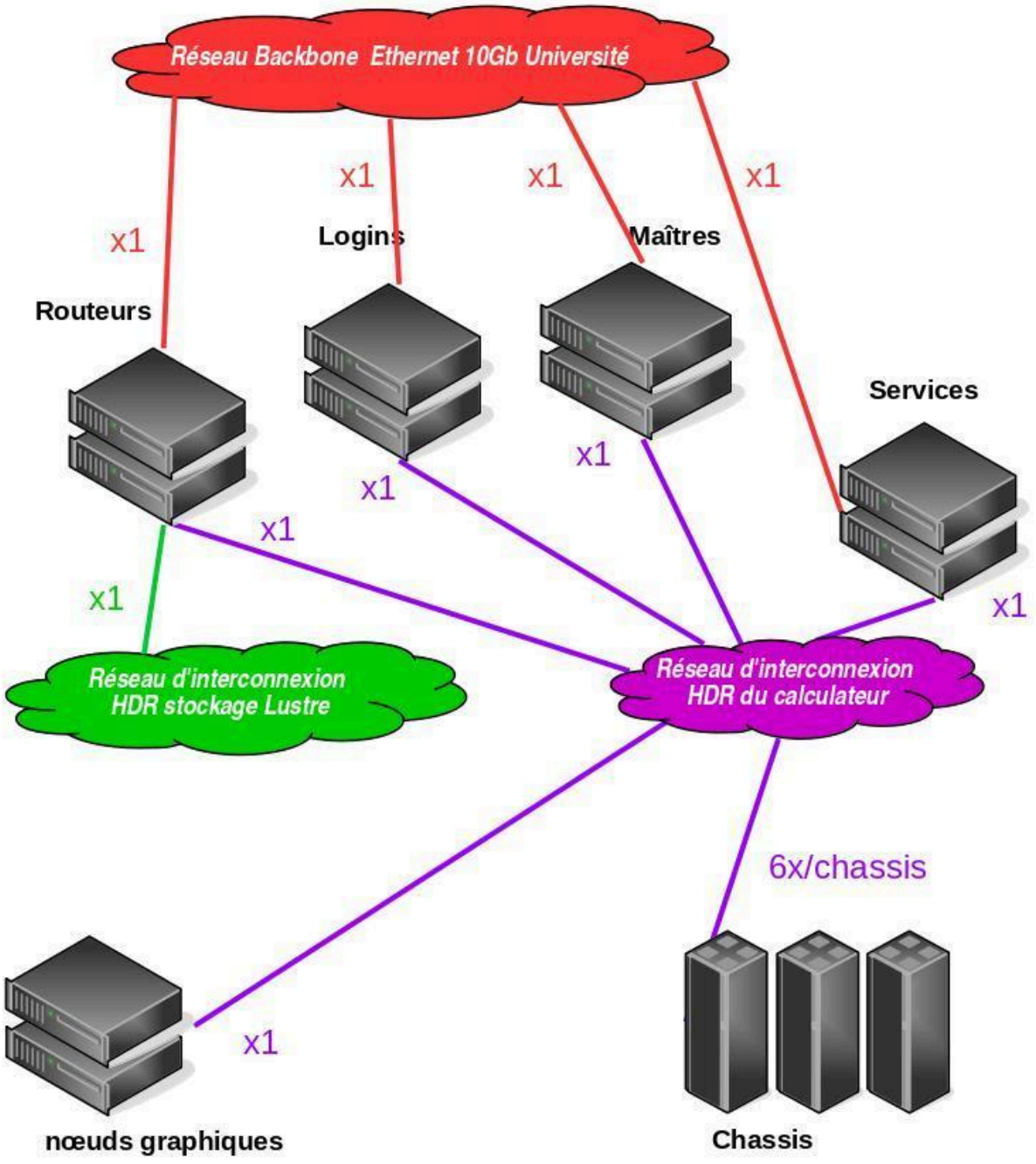
Les nœuds de login (les nœuds sur lesquels les utilisateurs du cluster se connectent) sont directement reliés au backbone de l'université par 1 port 10Gb

Les nœuds maîtres (les nœuds sur lesquels les administrateurs du cluster se connectent) sont directement reliés au backbone de l'université.

Les nœuds routeurs sont directement reliés au backbone de l'université.

Les nœuds de service peuvent être directement reliés au backbone de l'université si ils ont besoin d'accéder à un des services fournis par les serveurs de l'université.

Schéma de connexion des réseaux Infiniband et Backbone



Réseau backbone 10Gb/s

Réseau Infiniband HDR calculateur

Réseau Infiniband HDR stockage

x N Nombre de liens par serveur ou châssis

4 Analyse de la réponse et préparation de l'installation

Rédigez un rapport reprenant les entêtes des paragraphes suivants en répondant aux questions qui vous sont posées. N'hésitez pas à ajouter des schémas à vos explications. Vous devrez utiliser les nœuds de service à votre convenance pour implémenter les services.

Si des fichiers de configurations sont demandés, validez la syntaxe de ces fichiers comme vu en TP et donner les paths complets de ces fichiers avant d'en donner le contenu. Vous pouvez utiliser la commande tree pour donner une représentation d'un répertoire si nécessaire.
Exemple :
serveur X "/etc/foo/foo.conf"
…. contenu du fichier "/etc/foo/foo.conf"
tree /etc/dhcp/
/etc/dhcp/
├── dhclient.d
│ ├── chrony.sh
│ └── ntp.sh
├── dhcpd6.conf
└── dhcpd.conf

serveur X "/etc/dhcp/dchpd.conf"
… contenu du fichier "/etc/dhcp/dchpd.conf"

4.1 Conformité de la réponse technique au cahier des charges

Comparer la réponse technique au cahier des charges. Faire la liste des points forts, des points faibles. Relever les points qui ne répondent pas parfaitement au cahier des charges. Vérifier la correction et la cohérence de la solution.
Proposer des améliorations, soit en restant dans le cadre de la proposition technique proposée par AzkarHPC, soit en demandant des modifications légères : attention aux surcoûts, le budget doit être respecté.

4.2 Nommage du cluster et des composants matériels

4.2.1 Nom du cluster et Organisation logique des nœuds

Choisir un nom pour le cluster. Quels sont les critères de choix pour un nom de cluster - penser aux critères techniques et pratiques.
Proposer une numérotation des nœuds du cluster (en respectant la notation nodeset de clustershell) qui permet de distinguer leur appartenance aux différents types de nœuds et d'étendre plus tard le cluster c'est-à-dire de rajouter des nœuds de différents types.

4.2.2 Définition des groupes clustershell

En réfléchissant aux différents types et groupes de matériels présents dans le cluster et aux différents rôles, définir une liste de groupes clustershell - sous la forme @nom du groupe : nodeset - utiles pour l'administration quotidienne du cluster. Soyez exhaustifs, il ne faut négliger aucun aspect de l'administration. Remplissez le tableau ci dessous en y ajoutant des lignes :

Nom du groupe @xxxx	Nodeset	Commentaire

Vous pourrez vous référer plus tard à ce système de nommage dans vos réponses. Vous pourrez rajouter des groupes à cette liste au fur et à mesure de l'étude.

4.3 Architecture du réseau de contrôle électrique 1Gb du cluster

AzkarHPC propose le plan de câblage suivant pour le réseau de contrôle :

Noeuds	Port Switch AZ-W40-1g	Commentaire
nœud master1 ports Gb [1-2]	ewc1p[1-2]	
nœud master2 ports Gb [1-2]	ewc2p[1-2]	
1 serveur NFS admin, port BMC	ewc1p3	
9 nœuds de service, port BMC	ewc1p[4-...]	
10 nœuds de login, port BMC	ewc2p[3-...]	
9 nœuds routeur, port BMC		En terminant de remplir ewc1 puis ewc2, puis ewc3 etc ...
50 nœuds graphiques, port BMC		idem
Châssis AZ-B30, port BMC		idem

Modifier ce tableau pour amener plus de résilience et éliminer les SPOFs (Single point of failure/point unique de défaillance) en donnant vos arguments.
Il manque une ligne dans ce tableau. Trouvez l'élément oublié et ajoutez le.

Évaluer le nombre de ports libres sur le réseau de contrôle . En tirer des conclusions sur la résilience et la maintenabilité (exemple panne d'un port).
Définir les types de flux/protocoles/services qui passent sur le réseau de contrôle.

4.4 Architecture du réseau d'administration 10 Gb du cluster

AzkarHPC propose le plan de câblage suivant pour le réseau d'administration:

Noeuds	Port Switch AZ-W40-10g	Commentaire
nœud master1 ports 10Gb [1-2]	ewa1p[1-2]	
nœud master2 ports 10Gb [1-2]	ewa2p[1-2]	
1 serveur NFS adm, 4 port 10Gb	ewa[1-4]p3,	
9nœuds de service, 1 port 10Gb		En terminant de remplir ewa1 puis ewa2, puis ewa3 etc ...
10 nœuds de login , 1 port 10Gb		idem
9 nœuds routeur, 1 port 10Gb		idem
50nœuds 50 nœuds graphiques, 1 port 10Gb		idem
Châssis AZ-B30, 4 ports 10Gb		

Modifier ce tableau pour amener plus de résilience et éliminer les SPOFs (Single point of failure/point unique de défaillance) en donnant vos arguments.

Évaluer le nombre de ports libres sur le réseau d'administration . En tirer des conclusions sur la résilience (exemple panne d'un port). Quel est l'impact de la perte d'un switch AZ-W40-10g sur le fonctionnement du cluster ?
Définir les types de flux/protocoles/services qui passent sur le réseau d'administration.

4.5 Connectivité externe au Backbone de l'université

Vous prévoyez de connecter tous les nœuds de service au backbone de l'université. Au total, combien de connexions 10Gb au backbone demanderez-vous à l'équipe réseau ?
Pour chaque type de nœud du cluster connecté directement au backbone, précisez les types de flux/protocoles qui passent sur ses liens backbone et vers quels serveurs de l'université.

Nodeset	Serveurs université	Flux/protocoles
@login		
@admin etc ...		

Vous aurez certainement à revenir sur ce tableau une fois les autres parties réalisées.

4.6 Architecture du réseau d'interconnexion Infiniband

La société AzkarHPC ne propose pas de solution pour les connexions des nœuds des partitions Srv sur le réseau Infiniband.
Remplissez le tableau ci-dessous en utilisant des nodeset en proposant une solution résiliente et redondante.
Expliquez comment vous répartissez les communautés d'utilisateurs sur les nœuds de login.

Noeuds	**Port IB**	**Commentaire**
2 nœuds master		
nœuds de service		
nœuds de routeurs		
nœuds de login		

4.7 Extensibilité de la solution

Revoir le schéma du réseau d'interconnexion du réseau Infiniband.
Sans rajouter de châssis, combien de nœuds peuvent être ajoutés dans les partitions AC et AL ?

Sans rajouter de racks, combien de châssis et donc de nœuds peut-on rajouter dans les les partitions AC et AL ?
Quels sont les facteurs limitant pour étendre la partition AX ? Comment pourrait t-on l'étendre ?
En dehors de la place physique dans les racks, quels sont dans l'architecture du cluster les autres facteurs ou composants limitant son extensibilité.

4.8 Récupération des logs et des consoles

Pour faciliter l'administration du cluster, vous voulez récupérer tous les messages des nœuds des partitions C, M, A, logins et services et centraliser ces messages dans un répertoire /var/cluster accessible depuis les 2 nœuds maîtres.

Les fichiers de journaux devront être organisés sous la forme d'une arborescence : /var/cluster/logs/<Hostname>/messages.

En utilisant un ou plusieurs nœuds de service, proposez une architecture hiérarchique et scalable pour le service RSYSLOG. Faites attention à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) . Expliquer le choix des serveurs du cluster assignés à ce service. Illustrez l'architecture par un schéma et détailler le rôle des serveurs.

Écrire un fichier de configurations rsyslog par type de serveurs et un fichier de configuration pour un nœud de calcul.

De même, vous voulez centraliser les consoles des nœuds sous l'arborescence :/var/cluster/consoles/ . Quel est le logiciel qui permet l'accès interactif aux consoles des nœuds et l'enregistrement de ces consoles dans des fichiers ?

Relire la proposition technique de AzkarHPC et proposer un espace de stockage adéquat pour le répertoire /var/cluster/ .

Proposez une politique de gestion des journaux du cluster.

4.9 Synchronisation temporelle du cluster

Choisissez 4 nœuds de service pour implémenter une architecture NTP interne au cluster qui s'appuie sur l'infrastructure NTP existante de l'université. Vous utiliserez les modes client/serveur et symétrique de NTP.

Votre architecture NTP doit être résiliente, le service NTP du cluster doit survivre à une panne des deux serveurs NTP de l'université ou une coupure du backbone.

Faites attention à la connectivité Backbone (4.5) et à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) . Expliquer le choix des serveurs du cluster assignés à ce service.

Illustrez l'architecture NTP par un schéma en détaillant les modes NTP.

Écrire un fichier de configuration pour chaque type de serveurs NTP et un fichier de configuration pour un nœud client.

Est-il nécessaire de modifier les configurations des serveurs NTP de l'université ?

4.10 Architecture LDAP

Proposer une architecture interne au cluster pour le service LDAP qui s'appuie sur l'infrastructure existante LDAP du centre informatique. Décrire les mécanismes mis en œuvre.

Faites attention à la connectivité Backbone (4.5) et à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) . Expliquer le choix des serveurs du cluster assignés à ce service.

Préciser la configuration LDAP des clients.

Écrire les fichiers de configurations pour les différents types de nœuds selon le rôle LDAP.

Indiquer les changements de configuration à apporter sur les serveurs LDAP de l'université de

l'université avec les éventuelles modifications de schéma.

4.11 Architecture DNS

Proposer une architecture interne au cluster pour le service DNS qui s'appuie sur l'infrastructure existante DNS de l'université.

Faites attention à la connectivité Backbone (4.5) et à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) . Expliquer le choix des serveurs du cluster assignés à ce service.

Décrivez vos choix de sous-domaine, ... pour les différentes zones en précisant comment elles sont gérées.
Décrivez les différentes interactions entre les serveurs, en précisant les mécanismes mis en œuvre.

Préciser la configuration DNS des clients.
Si nécessaire, indiquez les changements de configuration à apporter sur les serveurs DNS de l'université.

4.12 Configuration du contrôleur SLURM

Faites attention à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) .
1) Expliquer le choix des serveurs du cluster assignés à ce service.
2) Préciser la configuration slurm qui :
- Permettra aux utilisateurs du cluster de soumettre leurs jobs sur les différents types de nœuds de calcul
- Optimisera l'ordonnancement des jobs et le choix des nœuds lors de leur allocation.
- Mettra en place une comptabilité des ressources consommées .

3) Restreindre l'allocation des ressources en fonction de 4 projets projet[0-3] de la façon suivante :
Chaque regroupement de communauté d'utilisateurs, est rattaché à un « projet » ie
projet0 : utilisateurs de ugp[0-2], … ,projet3 : utilisateurs de ugp[8-9].
- Limiter le nombre de jobs et de ressources utilisées par défaut par utilisateur ,
- Offrir la possibilité aux utilisateurs des projets 1 et 2 de tourner rapidement des jobs pour debugger mais en restreignant encore plus le nombre de ces jobs ainsi que leur durée,
- Ne pas donner l'accès aux nœuds de la partition AmdGpu des utilisateurs du projet0,
- Attribuer respectivement aux 4 projets 15 %, 20 %,40 %, 25 % d'utilisation de la machine.

Donner les commandes nécessaires à la création de cette politique en y associant au minimum un utilisateur par projet. Le nom des utilisateurs contiendra son numéro de ugp et son projet.: exemple userXugpYprojetZ.

Donner les commandes permettant de vérifier la politique mise en place.

4.13 Configuration de la surveillance

Le contrat de disponibilité spécifie 5 critères de disponibilité.

1) Au moins un nœud de login doit être opérationnel pour chaque communauté,
2) Un débit minimal de 32 GB/s vers le système de stockage Lustre doit etre assuré,
3) 90% des nœuds de la partition AC doivent être disponibles,
4) 80% des nœuds de la partition AL,
5) 60 % des nœuds de la partition AX.

La sociéte AzkarHPC utilise le logiciel Shinken comme logiciel de surveillance et a écrit des sondes (scripts) pour vérifier les critères de disponibilités :

- la sonde verif-noeud vérifie qu'une liste de nœuds est up
- la sonde verif-lustre-routeur vérifie que le service Lustre est opérationnel sur une liste de nœuds
- la sonde verif-nodeset-ssh vérifie que le service ssh est opérationnel sur une liste de nœuds

Spécifier comment vous pourriez utiliser ces sondes pour vérifier les critères de disponibilités (paramètres d'appel de ces sondes)
Proposer l'implémentation d'une sonde qui surveillera les seuils de disponibilité pour une partition slurm.
Donner 3 points spécifiques à surveiller sur les nœuds de login pour qu'ils puissent assurer un service interactif confortable pour les utilisateurs.
Au niveau du centre de calcul, quels sont les services critiques à surveiller ?
Pour faire une analyse plus fine des performances du cluster, que faut-il mettre en place ?

4.14 Configuration Puppet

Proposer une architecture des serveurs Puppet permettant la configuration (en un temps acceptable) des nœuds et serveurs du cluster.
Faites attention à l'absence de SPOF (Reportez vous au tableau que vous avez écrit en 4.4) .
Expliquer le choix des serveurs du cluster assignés à ce service.
En vous appuyant sur le travail de configuration effectué auparavant, faire une liste de tous les services utilisés dans le cluster, puis partitionner tous les nœuds et serveurs en groupes partageant les mêmes rôles (un nœud de calcul, un nœud de login, etc...). En dériver la liste des profils (client DNS, nœuds en IB, nœud "utilisateur", etc...) qui composent chacun de ces rôles.
Présentez le résultat de ces réflexions ment sous forme de tableau.

Groupe/Rôle	Profiles	Nodeset	Commentaire
Ex : nœud bleu	Liste de profiles	...	...
Ex : nœud rouge	Liste de profiles	...	...

Proposer une liste de modules communautaires pouvant remplir certains de ces rôles, donner également les profils dont le module puppet sera à développer car trop spécifique ou inexistant.
Donner ainsi l'organisation des modules et source de données 'Hiera'.

4.15 Environnement Utilisateur

Grâce aux réunions avec les futurs développeurs et utilisateurs de la machine, vous avez identifié les besoins suivants :

- un environnement de développement Intel complet
- un environnement de développement GNU de base
- un environnement de développement pour des codes accélérés par GPU/NVIDIA .
- Une version récente de MATLAB et MATLAB-Engine

Vous anticipez d'autres demandes et décidez d'utiliser EasyBuild pour générer et offrir les

environnements logiciels voulus. Précisez les toolchains que vous utiliserez et les produits que vous offrirez avec leurs versions. Vous trouverez les informations adéquates sous : https://docs.easybuild.io/en/latest/version-specific/Supported_software.html

Dans quels types de systèmes de fichiers pouvez-vous installer ces produits ? Préciser les avantages et inconvénients.

4.16 Déploiement

Pour satisfaire l'université, vous avez décidé d'installer la version RHEL9 sur tous les nœuds du cluster. Il faut déterminer quelles versions RHEL9, Lustre et MOFED (driver infiniband) sont nécessaires.
Vous avez sélectionné la version MOFED -2310213 LTS. Donner la signification de cet acronyme LTS et expliquer quelles sont les caractéristiques d'une version LTS.
Consulter la page
https://docs.nvidia.com/networking/display/mlnxofedv23102131lts/general+support
Vérifier la liste des versions RHEL9 supportées par cette version MOFED. Vérifier la liste des cartes ConnectX supportées par cette version. Vérifier aussi quelles versions de Lustre sont supportées par cette version des drivers Infiniband.
Déterminer la version de firmware pour les cartes ConnectX-6 compatible avec cette version MOFED.
Consulter la page https://wiki.whamcloud.com/display/PUB/Lustre+Support+Matrix et déterminer quelle version Lustre est compatible avec la version RHEL9.
Remplir la matrice de compatibilité ci dessous :

RedHat	Lustre	MOFED	Firmware ConnectX-6

Proposer une méthode d'installation pour les nœuds maîtres en utilisant l'infrastructure du centre. Décrire d'une façon générale l'ordre et la méthode de déploiement du cluster, une fois les nœuds maîtres installés.

Que faut-il préparer pour pouvoir démarrer les nœuds des partitions AC, AL, et AX ?

4.17 Planification de l'installation du cluster

A cause de problèmes de logistique et d'approvisionnement de composants, la société AzkarHPC ne peut pas vous livrer tout le matériel comme prévu dans le contrat en semaine 17/2024. Elle ne pourra livrer que 2 racks de calcul par semaine à partir de la semaine 18/2024. De plus, le personnel de la société ne travaille pas le week-end et les jours fériés. Enfin, l'agenda des sociétés de transport ne permet que des livraisons le mercredi matin à 8h00. L'installation matérielle d'un rack prend une demi journée : installation du rack en salle, câblage interne au rack, câblage électrique, Son intégration avec le reste du cluster et son raccordement au centre de calcul prend une demi journée.
Identifier les éléments indispensables pour commencer l'installation au plus tôt, et préparer un ordre et un planning de livraison et d'installation matérielle et logiciel qui permette de minimiser le temps total d'installation et d'ouvrir à temps la machine aux utilisateurs. Dans ce planning, vous intégrerez toutes les étapes qui amènent à la mise en production du cluster et à l'ouverture aux utilisateurs que vous avez vues en cours.
Pour optimiser ce temps, vous pouvez imposer un ordre de livraison à la société AzkarHPC. Vous présenterez ce planning sous forme de tableau excell ou un diagramme de Gant.