

Projet de Logiciel Cluster

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



Binôme :

Matthias LAPU
Louis AUFFRET

Tables des matières

0	Introduction	3
1	Conformité de la réponse technique au cahier des charges	4
1.1	Partition AmdCpu (AC)	4
1.2	Partition AmdLarge (AL)	4
1.3	Partition AmdGpu (AX)	5
1.4	Partition frontale (F)	6
1.5	Débit vers stockage Lustre	6
1.6	Débit vers backbone	7
1.7	Partition Admin/Service	7
1.8	Offre logicielle & Outils de développement et de profiling	7
2	Nommage du cluster et des composants matériels	8
2.1	Nom du cluster et Organisation logique des nœuds	8
2.2	Définition des groupes clustershell	8
3	Architecture du réseau de contrôle électrique 1Gb du cluster	9
4	Architecture du réseau d'administration 10 Gb du cluster	11
5	Connectivité externe au Backbone de l'université	13
6	Architecture du réseau d'interconnexion Infiniband	14
7	Extensibilité de la solution	15
7.1	Sans rajouter de châssis	15
7.2	Sans rajouter de racks	15
7.3	facteurs limitant et extension de la partie AX	15
7.4	Limitation de l'extensibilité	15

0 Introduction

Lors de ce projet, il nous était demandé d'analyser un appel d'offres émis par une université afin de déterminer si la société répondant à l'appel d'offres a réellement pris en compte tout ce que l'université a demandé. Nous avons remarqué que l'appel d'offre n'a pas fixé de budget, et dans la réponse d'offre, nous avons trouvé de nombreuses incohérences.

Par exemple, dans la page racking situé à la page 6 du sujet, il est question de 2 ports externe 1Gb dans les châssis AZ-B30, cependant, 1 seul port est évoqué dans le paragraphe sur les connexions externes, on a alors supposé la présence de 2 ports. Dans la synthèse de la réponse, située page 5 du sujet, il est noté pour la partition AX, "1 disque de 1 To", cependant, à la fin de la phrase, il est marqué "Aucun disque", on a supposé ces serveurs diskless. Ou encore, laissant place à nos suppositions, nous pouvons prendre l'exemple de la partie racking, ou il n'est pas précisé la taille physique des nœuds de login. On supposera ici qu'ils mesurent 2U. De même, il est question dans le sujet de relier les noeuds de login par 1 port 10 Gb au réseau backbone et 2 câbles 10 Gb au réseau d'administration, alors qu'ils n'ont que 2 ports 10 Gb.

Ces quelques incohérences, nous ont fait questionner si les erreurs que l'on trouvait dans le sujet étaient involontaires, ou si elles étaient là pour nous faire remarquer qu'AzkarHPC a fait quelques erreurs dans sa réponse.

1 Conformité de la réponse technique au cahier des charges

1.1 Partition AmdCpu (AC)

Dans le cahier des charges, il est demandé d'avoir environ 40 000 coeurs, la solution apportée par AzkarHPC comporte 38 656 coeurs, le nombre de coeurs est en dessous de la demande originelle mais est dans la marge des $\pm 5\%$.

N'étant pas certains de l'avantage fourni par les processeurs hyperthreadés, nous nous sommes inspirés de clusters existants : en prenant un grand centre de calcul tel que le TGCC, nous pouvons voir que celui-ci utilise des processeurs hyperthreadés (ex : AMD Irene ROME qui possède des processeurs avec 64 coeurs, 128 threads et 256 MB de cache). C'est également le cas pour le choix effectué par le fournisseur. L'utilisateur n'est pas obligé d'utiliser l'hyperthreading si son code n'en profiterait pas, il semble donc cohérent d'en fournir le support.

Le fournisseur semble avoir fait un bon choix, en effet, dans la gamme de processeurs proposés, le seul ayant plus de cache L3 que les autres est le 9634 et celui-ci ne possède que 3/4 des coeurs du 9754 pour un prix proche. De plus, le 9754 est le processeur fournissant le plus grand nombre de coeurs, et il est hyperthreadé, ce qui permet d'améliorer les performances des codes conçus pour en profiter, il est donc un choix judicieux dans le cadre des codes CPU-bound, pour lesquels cette partition se doit d'être adaptée.

Pour ce qui est de la mémoire, AzkarHPC propose 256Go de mémoire par noeud, un noeud possède 128 coeurs, ce qui correspond donc à 2Go par coeur. Cela respecte la demande de l'université. Il n'y a pas de disques sur cette partition, car ils ne sont ni demandés ni nécessaires : AzkarHPC a fait le choix de boot les noeuds en PXE pour faire des économies.

L'un de leurs points forts réside dans leurs choix de processeurs, un processeur avec beaucoup de coeurs permet de réduire le nombre de noeuds nécessaire pour atteindre le nombre de coeurs demandé. En effet, s'ils avaient décidé d'opter pour le processeur 9634 qui possède le plus de cache L3, il aurait fallu plus de 475 noeuds pour avoir le même nombre de coeurs qu'avec leur proposition, et un code utilisant un plus grand nombre de noeuds est sujet à des effets NUMA inter-noeuds plus forts.

Cependant, avoir des coeurs trop nombreux peut ne pas être profitable, car il faudrait des utilisateurs capables d'utiliser 100% d'un processeur avec 128 coeurs et 256 threads, or d'après la loi d'Amdahl, pour pleinement utiliser cette parallélisation il faudrait avoir un code proche de 95% parallèle pour avoir un speedup théorique de 18, cependant en pratique cela est extrêmement dur à atteindre et les utilisateurs ne feront pas tourner de code utilisant autant de threads. De plus, augmenter le nombre de coeurs alloués peut impliquer d'augmenter le nombre de contrôleurs mémoire alloués, ce qui engendre des effets NUMA intra-noeud et peut dégrader les performances.

Les 302 lames peuvent être hébergées dans 11 châssis, un rack peut héberger 3 châssis, donc ces noeuds peuvent bien être hébergés dans 3 racks.

1.2 Partition AmdLarge (AL)

Le nombre de coeurs demandé dans l'appel d'offre est de 4000, AzkarHPC propose 4200 coeurs, ils sont au dessus de ce qui était demandé, mais dans la marge des $\pm 5\%$.

AzkarHPC propose 50 noeuds, avec 512 Go de mémoire par noeud, pour 84 coeurs. Ce qui revient

à environ 6 Go par coeur de mémoire, ce qui est conforme à ce qui est demandé, et même au-delà. Notons que cette partition a pour but de faire tourner des codes memory-bound (à priori, vu la demande en mémoire), et le choix du processeur (9634) est particulièrement judicieux car c'est celui qui possède la plus grande quantité de cache L3, ce qui permet d'accélérer les accès mémoire plus souvent.

La demande de l'université n'est pas assez claire, elle ne précise pas les quantités de cache des CPUs, une réponse avec très peu de cache serait conforme à la demande, mais aurait des performances moindres sur des codes memory-bound. AzkarHPC n'a pas pris de risque et a proposé un cache L3 le plus important possible. Cependant, il aurait été possible de prendre un autre CPU, tel que le 9754 qui possède plus de coeurs, mais celui-ci possède moins de cache, donc le ratio cache/coeur est beaucoup plus faible, ce qui explique le choix du 9634.

En effectuant un changement de CPU, pour des 9754, avec 32 noeuds, il y aurait 4096 coeurs, et 134 400 \$ économisés sur les CPUs uniquement.

Au total, avec moins de cache L3 (car nous ne savons pas si c'est que l'université désire réellement), nous pouvons faire une économie de plus de 134 000\$ tout en ayant un nombre de coeurs élevé et 4 Go de mémoire par coeur, mais des performances mémoire moindres, en respectant malgré tout le cahier des charges. L'université n'ayant pas spécifié de budget, nous gardons ici la réponse proposée par AzkarHPC, mais en cas de réponse défavorable de l'université pour cause de budget, nous gardons en tête qu'il est ici possible de faire des économies.

Les 50 lames peuvent être hébergées dans 2 châssis, ce qui peut bien rentrer dans 1 rack. Cependant, nous avons fait le choix de mutualiser les emplacements (châssis et racks) pour les partitions AC et AL pour économiser de la place et des ports sur les switches évoqués dans les parties ci-dessus. Ainsi, avec 302 noeuds AC et 50 noeuds AL, on remplit :

- 10 châssis avec 30 noeuds AC chacun
- 1 châssis avec 30 noeuds AL
- 1 châssis avec 20 noeuds AL et 2 noeuds AC

Ainsi les partitions AC et AL peuvent rentrer dans 12 châssis au lieu de 13, donc dans 4 racks au lieu de 5, ce qui nous fait économiser de la place et permet de garder un rack libre si on voulait étendre les partitions.

1.3 Partition AmdGpu (AX)

AzkarHPC propose 50 noeuds avec des processeurs 9224, leur proposition offre au total 1200 coeurs, ce qui est loin d'être suffisant. De plus, avec une carte par noeud, il y aurait donc 50 cartes GPU Nvidia, ce qui ne correspond également pas à la demande émise par l'université. La quantité de mémoire par coeur excède 10 Go, ce qui est bien au-delà de la demande et n'est à priori pas nécessaire. De plus, il est marqué un disque de 1 To (incohérence car ils sont dits diskless juste après), ce qui n'est pas demandé. Remarquons également qu'il est demandé 60 cartes Nvidia A100, seulement 50 sont proposées, ce qui est en-dessous de la marge des $\pm 5\%$.

En supposant qu'un noeud AX ne peut accueillir qu'une seule Nvidia A100 (les spécifications du noeud ne sont pas précisées), on optera pour 60 noeuds comportant chacun une carte Nvidia A100, un processeur 9454 (48 coeurs, 96 threads), et 128 Go de RAM DDR5. On aurait alors 2880 coeurs,

2.67 Go de mémoire par coeur, et 60 cartes Nvidia : ce qui est conforme à la demande. Ces 60 noeuds peuvent remplir exactement 2.5 racks, on n'a pas besoin de plus que les 3 racks fournis.

Si un noeud pouvait accueillir 2 cartes, on opterait pour 30 noeuds ayant chacun 2 cartes Nvidia A100 (on remplirait à peine plus d'un rack), un processeur 9734 et 256 Go de DDR5 : on aurait alors 60 cartes Nvidia, 3360 coeurs et 2.29 Go de RAM par coeur, ce qui serait conforme à la demande (la demande spécifiant au moins 3000 coeurs, et non $3000 \pm 5\%$).

Si le budget proposé par l'université est dépassé, les modifications que nous avons proposé sur la partie AL seraient alors effectuées. En effet, Le budget originel proposé par AzkarHPC pour les CPU est de 91 250 € pour cette partie, si nous ajoutons la somme théorique dû au économies faites lors de la partie précédente, nous avons alors un nouveau budget de 230 450 \$. Cependant nous ne connaissons pas les prix des cartes Nvidia ni des cartes mères, seulement ceux des processeurs.

1.4 Partition frontale (F)

La société propose 10 noeuds de login, ce qui est le nombre demandé par l'université. Chaque noeud possède un 9634, ce processeur possède 84 coeurs, il en était demandé 24, c'est donc bien au-delà de la demande. De plus, cela ne laisse pas assez de mémoire pour chaque coeur (1,14 Go).

Il faudrait par conséquent changer de processeur, ce qui ferait gagner en budget. Le coût pour 10 noeuds est de 103 040 \$. Or, il est demandé 24 coeurs par noeud, nous pouvons donc utiliser le processeur 9224. Avec 10 noeuds 9224, cela reviendrait à 18 250 \$, cela ferait économiser une somme importante (84 790 \$), qui serait mieux utilisée ailleurs (partie AX). Notons par ailleurs que parmi les processeurs proposés, le 9224 a le meilleur rapport théorique GFLOPS/\$ (en supposant que les FLOPS théoriques se calculent par $nb_coeurs \times frequence$, sous les hypothèses qu'une opération sur des flottants dure un cycle, et qu'un code utilisant le processeur 100% du temps ne profite pas de l'hyperthreading, d'où le choix du calcul avec le nombre de coeurs physiques, bien que le 9224 supporte bel et bien l'hyperthreading).

Avec cette modification, la quantité mémoire originelle (96Go de mémoire DDR5) est trop élevée. En effet, cela reviendrait à avoir 4 Go par coeur, l'université n'en demande que 2. On optera ici pour 64 Go de RAM par noeud, ce qui fait 2.67 Go de RAM par coeur.

Le stockage requis ici est pour une partition /tmp et une partition bootable qui contient le système, les compilateurs, les bibliothèques de calcul, les debuggers, etc. La partition système peut être montée en read-only. La partition /tmp est montée en read-write. On a besoin de 3 To au total, on peut donc choisir 3 disques de 2 To et utiliser le contrôleur RAID pour avoir 2 disques de stockage (donc 4 To) et un de redondance (basé sur la parité, disque 3 = disque 1 XOR disque 2).

Le point fort de leur offre est le nombre important de coeurs, le point faible est le prix, qui est plus important que le nombre nécessaire pour des noeuds de login, et le contrôleur RAID qui peut être mieux utilisé. De plus, la RAM disponible est également moins importante que la demande de l'université. Changer le processeur utilisé ainsi que la taille de la RAM permettrait de faire des économies et de respecter la marge de 5% de l'université.

1.5 Débit vers stockage Lustre

La société propose de mutualiser les 2 fonctions Lustre et Backbone sur 9 routeurs I/O. Pour la fonction lustre, ils seront connectés à l'aide de port IB HDR200 de part et d'autre. Il y a 9 routeurs

qui sont connectés.

Chaque routeur à un débit réel de 24 Go/s, en parallèle cela fait donc un débit théorique total de 216 Go/s. Cela dépasse de loin le débit demandé originellement, cependant, l'université indique "au moins", nous garderons donc la proposition d'AzkharHPC. L'un des points forts de cette offre repose dans le nombre de routeurs conséquent, permettant de proposer beaucoup plus de débit que demandé vers le stockage Lustre, ce qui peut s'avérer nécessaire en cas de panne ou de besoin d'extensibilité. De plus, la société propose les processeurs les moins chers, ce qui est un bon choix, car tous les processeurs possèdent le même débit mémoire, qui est amplement suffisant. Le fait d'avoir 2 processeurs par noeud permet également de répartir correctement la charge de travail, par exemple entre un processeur s'occupant uniquement du backbone et l'autre uniquement de Lustre.

On suppose que la raison pour laquelle il y a 2 processeurs par noeud est de fournir 1 par fonction (Lustre et Backbone) : cela pourrait permettre un meilleur usage du cache pour chaque processeur, car une interaction sur la mémoire par un processeur n'affectera pas le cache de l'autre tant que les zones mémoire utilisées sont distinctes.

1.6 Débit vers backbone

L'université demande 7.5 Go/s vers le backbone pour chacune des 3 partitions de calcul, AzkharHPC connectera 2 ports au backbone, si cela est bien configuré il est possible que cela double le débit. Au total 18 connexions 10 Gb seront établies, le débit sera donc entre 9 Go/s et 18 Go/s. Le point positif de leurs solutions repose dans la rapidité donnée par la double connexion des routeurs au backbone et le fait que le pire des cas apporte un meilleur résultat que ce que l'université demande.

1.7 Partition Admin/Service

La solution apportée par Azkhar nécessite 20 nœuds, cela est plus de nœuds que demander par l'université. Leurs solutions impliquent 2 nœuds maîtres permettant l'administration et le déploiement du cluster, ces 2 nœuds possèdent 2 disques de 3 To, nous supposons que ces disques contiendront les données nécessaires au déploiement. Ces 2 nœuds maîtres utiliseront également les données contenues dans les 2 disques de 3To des nœuds services pour stocker les données liées aux services internes. L'université demandait au total, au moins 80 To pour les données systèmes. AzkharHPC propose de séparer ces données dans différents nœuds, en ne suivant pas le format système/données systèmes demandés par l'université. Cependant, la partie maîtres et service seraient équivalents à la partie données systèmes, et la partie routeur, à la partie système. La solution proposée par Azkhar ne possède donc pas suffisamment d'espace par noeud pour correspondre à la demande de l'université. Cependant, leurs approches, utilisant des nœuds maîtres et des nœuds de services, permettraient, selon moi, d'avoir le résultat escompté avec moins d'espace, car un espace conséquent serait économisé avec une approche nœuds de services/nœuds maîtres.

1.8 Offre logicielle & Outils de développement et de profiling

La société propose l'offre logicielle demandée par l'université, c'est-à-dire RHEL9. De plus, il y a également des rpm lustre, des compilateurs Intel et Amd. Uniquement, des processeurs Amd sont utilisés, nous supposons que le compilateur Intel sert uniquement afin d'avoir de la compatibilité si certains nécessitent des instructions assembleurs Intel. Sinon, ce compilateur est inutile. Il propose également des debugger ainsi que des logiciels Opensource. Ce qui est demandé par l'université.

2 Nommage du cluster et des composants matériels

2.1 Nom du cluster et Organisation logique des nœuds

Un nom doit être simple, et facile à retenir. L'université se nomme "Université Toulouse III - Paul Sabatier". **Nous appellerons donc le cluster "sabatier"**, c'est un nom simple et évocateur pour les personnes utilisant le supercalculateur de l'université.

Pour l'organisation logiques des noeuds, ceux-ci seront notés sabatierXXX. Dans la synthèse de la réponse, il y a déjà proche de 440 noeuds au total, dont 302, uniquement pour la partie AmdCPU. Afin de prendre en compte de futurs extension, nous utiliserons des plages assez grandes. Chaque groupe possédera son nombre afin de faciliter l'administration.

2.2 Définition des groupes clustershell

En s'inspirant du clustershell du cluster de l'ENSIE. Nous appliquons la configuration suivante sur le fichier `/etc/clustershell/groups.d/cluster.yaml`.

```
1 login: sabatier[000-009]
2 master: sabatier[020-021]
3 services: sabatier[040-049]
4 io: sabatier[060-069]
5 ax: sabatier[100-149]
6 al: sabatier[200-249]
7 ac: sabatier[300-601]
8 all: sabatier[000-601]
```

Pour la partie 4.3 et 4.4, nous avons compris qu'il était demandé de faire avec ce qu'AzkarHPC proposait, malgré le fait que lors de la 4.1 nous évoquions d'autres solutions, ou que des choses différentes étaient marquées dans l'énoncé (2 ports au lieu d'1 par exemple pour la partie AX 10 Gb etc.). Même si cela ne semblait pas toujours pertinent, nous avons préféré rester avec ce qu'il y avait dans l'énoncé afin de ne pas être pénalisés. Nous avons cependant pris le temps d'expliquer comment nos modifications auraient affecté ces questions.

3 Architecture du réseau de contrôle électrique 1Gb du cluster

Noeuds	Port Switch AZ-W40-1g	Commentaire
nœud master1 ports Gb [1-2]	ewc[1-2]p1	On élimine les SPOFs en le connectant à 2 switches.
nœud master2 ports Gb [1-2]	ewc1p2 ,ewc3p1	idem
1 serveur NFS admin, port BMC	ewc1p3	C'est un SPOF
9 nœuds de service, port BMC	ewc1p[4-6], ewc[2-3]p[2-4]	Réparti équitablement parmi les 3 switches
10 nœuds de login, port BMC	ewc1p[7-8], ewc[2-3]p[5-8],	2,4,4 . ewc1 possède déjà les 2 masters, on met plus de nœuds sur les 2 autres.
9 nœuds routeur, port BMC	ewc[1-3]	on répartit également équitablement
50 nœuds graphiques, port BMC	ewc1p[12-27], ewc[2-3]	on répartit à peu près équitablement (16,17,17)
Châssis AZ-B30, port BMC	ewc1p[28-35] ,ewc[2-3]p[29-36]	il y a 12 châssis, soit 24 port BMC, qu'on répartit équitablement
Les alimentations électriques (PDU)	ewc[2-3]p37	les 2 masters sont déjà sur ewc1, on répartit les PDU sur les 2 autres

Utilisation des ports :

- Switch 1 : 35 ports utilisés, 5 ports restants
- Switch 2 : 37 ports utilisés, 3 ports restants
- Switch 3 : 37 ports utilisés, 3 ports restants

Finalement, lorsque tout est connecté, 109 ports sont utilisés et il reste 11 ports.

Ce nouveau branchement, rend le réseau de contrôle électrique 1Gb plus résilient, car tout est réparti de manière à peu près uniforme. S'il y a une panne de port, il reste suffisamment de port pour le déplacer et le rendre de nouveau fonctionnel en attendant de réparer le problème de port. S'il y a une panne de port et que celui-ci n'est pas déplacé, les fonctionnalités permettant d'allumer et d'éteindre la dite machine est donc perdue (nœuds de calcul, nœuds de service, etc.) . Sur le réseau de contrôle, les flux utilisés sont en lien avec la gestion de nœuds.

Il n'est pas précisé comment les switches sont interconnectés entre eux pour former un switch virtuel. S'il est nécessaire de les relier entre eux en utilisant l'un des ports. 2 ports en plus, par switch seront utilisés pour faire un graphe complet. Il resterait donc : 3 ports pour le switch 1, et 1 port pour le switch 2 et 3.

Notons que la partition AX utilise à elle seule presque la moitié des ports, et y ajouter 10 nœuds pour se conformer à la demande de l'université implique qu'il n'y a plus assez de ports et qu'il faudrait rajouter un switch et revoir les attributions des ports pour mieux distribuer la charge.

Les 2 nœuds maîtres possèdent chacun 1 port 1 Gb relié au port BMC de l'autre, ainsi que 3 autres ports 1 Gb. Ici, en utiliser seulement 2 est suffisant pour éviter que la panne d'un switch empêche complètement l'accès au réseau électrique par l'un des nœuds maîtres, c'est donc la solution retenue.

Ceci dit, nous pensons qu'1 seul port BMC pour le serveur NFS Admin représente un SPOF conséquent, en effet, si le switch tombe en panne, les administrateurs perdent l'accès BMC au NFS

permettant de stocker les données du système nécessaires à la gestion du cluster. Celui-ci est censé rester allumé, mais par précaution nous avons pensé à une alternative. S'il était possible d'effectuer des modifications à la proposition d'AzkarHPC, nous aurions rendu la partie NFS résiliente ainsi :

Noeuds	Port Switch AZ-W40-1g	Commentaire
nœud master1 ports Gb [1-2]	ewc[1-2]p1	On élimine les SPOFs en le connectant à 2 switches.
nœud master2 ports Gb [1-2]	ewc1p2 ,ewc3p1	idem
1 serveur NFS maître proposé par Azkhar, port BMC	ewc1p3	On ajoute de la résilience avec un NFS par switch
1 serveur NFS maître supplémentaire , port BMC	ewc2p2	
1 serveur NFS esclave , port BMC	ewc3p2	
9 nœuds de service, port BMC	ewc1p[4-6] , ewc[2-3]p[3-5]	Réparti équitablement parmi les 3 switches
10 nœuds de login, port BMC	ewc1p[7-8] , ewc[2-3]p[6-9]	2,4,4 . ewc1 possède déjà les 2 masters, on met plus de noeuds sur les 2 autres.
9 nœuds routeur, port BMC	ewc1p[9-11] , ewc[2-3]p[10-12]	on réparti également équitablement
50 nœuds graphiques, port BMC	ewc1p[12-27] , ewc[2-3]p[13-29]	on réparti à peu près équitablement (16,17,17)
Châssis AZ-B30, port BMC	ewc1p[28-35] ,ewc[2-3]p[30-37]	il y a 12 châssis, soit 24 port BMC
Les alimentations électriques (PDU)	ewc[2-3]p38	les 2 masters sont déjà sur ewc1, on réparti les PDU sur les 2 autres

4 Architecture du réseau d'administration 10 Gb du cluster

Nous rappelons qu'il y a 12 châssis.

Noeuds	Port Switch AZ-W40-10g	Commentaire
noeud master1 ports 10Gb [1-2]	ewa1p1, ewa4p1	On répartit équitablement
noeud master2 ports 10Gb [1-2]	ewa2p1, ewa5p1	idem
1 serveur NFS adm, 4 port 10Gb	ewa1p2, ewa2p2, ewa3p1, ewa4p2	idem
9 noeuds de service, 1 port 10Gb	ewa1p3, ewa2p3, ewa3p2, ewa4p[3-4], ewa5p[2-3], ewa6p[1-2]	idem
10 noeuds de login, 1 port 10Gb	ewa1p4, ewa2p[4-5], ewa3p[3-4], ewa4p5, ewa5p[4-5], ewa6p[3-4]	idem
9 noeuds routeur, 1 port 10Gb	ewa1p5, ewa2p6, ewa3p5, ewa4p[6-7], ewa5p[5-6], ewa6p[5-6]	idem
50 noeuds graphiques, 1 port 10Gb	ewa1p[6-13], ewa2p[7-14], ewa3p[6-14], ewa4p[8-15], ewa5p[7-14], ewa6p[7-15]	idem
12 châssis AZ-B30, 4 ports 10Gb	ewa1p[14-21], ewa2p[15-22], ewa3p[15-22], ewa4p[16-23], ewa5p[15-23], ewa6p[16-23]	idem

Nous avons réparti tous les noeuds entre les différents switches de manière uniforme en utilisant du Round-robin. Cela nous permet d'éliminer les SPOFs en évitant que des ports équivalents (venant du même noeud ou de noeuds redondants entre eux) se retrouvent attribués aux ports d'un même switch, ce qui ferait du switch en question un SPOF

Utilisation des ports :

- Switch 1 : 21 ports utilisés, 19 ports restants
- Switch 2, 3 : 22 ports utilisés, 18 ports restants
- Switch 4, 5, 6 : 23 ports utilisés, 17 ports restants

Finalement, lorsque tous les noeuds sont connectés, il reste 106 ports libres.

Il faut également établir la topologie nécessaire à la création de ce switch virtuel. En effectuant une topologie cyclique, c'est-à-dire que le switch est relié au switch $n+1$ et $n-1$ modulo le nombre de switches disponible. Avec cette topologie, une panne d'un des switches n'empêche pas les autres de communiquer entre eux, 146 ports seraient utilisés et il resterait 94 ports libres.

Si nous effectuons une topologie full mesh (graphe complet), 164 ports seraient utilisés et il resterait 76 ports libres. Étant donné le nombre de ports restants qui est conséquent, nous pouvons partir sur une topologie full mesh, permettant donc un accès à chaque switch depuis 1 machine externe en au plus 2 switches.

Si un port tombe en panne, il restera assez de place pour le déplacer le temps de résoudre la panne. La perte d'un port entraîne la perte de l'accès au service lié à celui-ci en passant par ce port. Par conséquent, les noeuds n'ayant qu'un seul port ne seront plus administrable. Les types de flux/protocoles utilisés sont en lien avec les services d'administration, il y a également les noeuds qui auront besoin de faire des requêtes (mises à jour, puppet, etc.).

Si un switch tombe en panne, la résilience de notre solution permet aux autres de toujours fonctionner, cependant, cela réduit également l'administration des services du cluster d'1/6.

Remarquons que dans l'énoncé, il est marqué que les noeuds de services, de login, les routeurs et les noeuds AX sont connectés au réseau 10Gb par 2 câbles et non 1 comme indiqué dans le tableau. Il faudrait alors rajouter plus de câbles qu'indiqué dans le tableau en augmentant les tailles des plages allouées, il y a la place pour cela, et cela remplirait presque les switches. Cela permettrait d'ajouter plus de résilience et de permettre l'administration de chaque machine même si un port tombe en panne. Sauf pour les noeuds de login, qui n'ont que 2 ports 10 Gb, l'un d'eux étant réservé au réseau backbone. De plus, si nous prenons en compte la solution énoncée lors de la partie précédente il faudrait également rajouter des ports pour le serveur NFS admin rajouté, ainsi que le serveur NFS

esclave. C'est une solution plus viable que le tableau donnée dans l'énoncé. Et la solution que nous aurions utilisé s'il était possible d'effectuer l'architecture à partir de nos modifications fait à la partie 1.

5 Connectivité externe au Backbone de l'université

Le réseau backbone fournit des services LDAP, DNS, NTP, NFS et PXE, ainsi qu'une communication directe avec les stations de travail de l'université. Le LDAP sert notamment à l'identification et l'authentification. Le DNS sert surtout pour l'accès à Internet, car s'il y a besoin de noms de domaines pour certaines machines au sein du réseau, un simple fichier `/etc/hosts` suffit, donc le serveur DNS n'est accessible qu'aux administrateurs sur les noeuds maîtres. Pour choisir sur quel VLANs placer les noeuds, il est nécessaire de prévoir les services dont ils auront besoin :

- les noeuds de service auront accès au service NTP (strate n) et serviront eux-mêmes de service NTP (strate n+1) pour les noeuds au sein du réseau IB, et ils auront accès au service PXE : les noeuds diskless du réseau IB vont boot en PXE sur les noeuds de service, les noeuds de service vont soit aller chercher les images de boot sur le serveur PXE de l'université, soit les envoyer directement s'ils les ont en cache (le cache permet d'éviter de surcharger le PXE de l'université)
- les noeuds maîtres sont réservés aux administrateurs, ils ont accès à tous les services, et les stations de travail peuvent y avoir accès en SSH
- les noeuds de login ont besoin de LDAP pour l'identification et l'authentification, et de NFS pour monter les répertoires home des utilisateurs, un serveur SSH devra écouter sur le réseau des stations de travail
- les routeurs n'ont à priori besoin de rien sur le réseau backbone, car récupérer des contenus provenant du cluster Lustre peut se faire en passant par les noeuds de login, mais il serait intéressant d'héberger un serveur NFS sur ces routeurs pour avoir un accès direct aux contenus du cluster Lustre depuis les stations de travail.

Nodeset	Serveurs université	Flux/protocoles
@login	montcalm[1-2], maladeta, stations	LDAP, NFS, SSH
@master	montcalm[1-2], maladeta, ossau[1-2], marbore[1-2], perdu[1-2], stations	LDAP, NFS, NTP, DNS, PXE, SSH
@service	ossau[1-2], perdu[1-2]	NTP , PXE
@io	stations	NFS

Pour ce qui est du nombre de ports :

- les 9 noeuds de service (4 ports 10 Gb) ont déjà 1 port 10 Gb sur le réseau d'administration, il est envisageable d'en ajouter un 2e sur ce même réseau, donc on utilise les 2 restants pour le réseau backbone (18 ports à demander)
- les 2 noeuds maîtres (4 ports 10 Gb) ont déjà 2 ports 10 Gb sur le réseau d'administration, on utilise les 2 restants pour le réseau backbone (4 ports à demander)
- les 10 noeuds de login (2 ports 10 Gb) ont déjà 1 port 10 Gb sur le réseau d'administration, on utilise celui qui reste pour le réseau backbone (10 ports à demander)
- les 9 noeuds routeurs (6 ports 10 Gb) ont déjà 1 port 10 Gb sur le réseau d'administration, il est envisageable d'en ajouter un 2e sur ce même réseau, donc on utilise les 4 restants pour le réseau backbone (36 ports à demander)

Il faudra demander au total 68 ports sur le réseau backbone à l'université.

6 Architecture du réseau d'interconnexion Infiniband

Tous les équipements du cluster (sauf le NFS, qui n'est que sur le réseau d'administration) possèdent au moins un port InfiniBand. Les routeurs en ont deux, mais l'un d'eux est réservé au cluster Lustre. Le seul port InfiniBand disponible sur chacun des noeuds sera donc relié au réseau InfiniBand du cluster. La topologie proposée par AzkarHPC est en étoile, avec 6 switches L2 redondants au centre de l'étoile et 17 switches L1 comme branches, chacun connecté aux 6 switches L2. On a alors 17 ports de chaque switch L2 utilisés. Pour les noeuds des partitions AC et AL, les switches InfiniBand sont ceux des châssis (30 noeuds/L1), or on a 12 châssis AC/AL donc il faudrait répartir le reste sur 5 switches L1. AzkarHPC propose d'avoir des switches L1 dédiés pour la partition AX (il en faudrait alors 2 au vu du nombre de noeuds, avec 25-30 noeuds/L1), un switch L1 dédié aux noeuds de service (9 noeuds/L1) et un dédié aux noeuds de login (10 noeuds/L1). D'après la solution proposée, il ne reste qu'un seul switch L1 pour les routeurs et les noeuds maîtres (soit 11 noeuds/L1).

Cependant, même si cette solution est conforme en termes de nombre de ports utilisés, elle manque de redondance et de résilience : les 9 routeurs sont sur le même switch, or les routeurs font les I/O les plus intensives. De plus, si le switch en question ne fonctionnait plus, les noeuds maîtres n'ont plus accès au réseau IB, et le réseau IB n'a plus accès au cluster Lustre. Si le switch qui héberge les noeuds de login ne fonctionnait plus, aucun utilisateur n'aurait accès au réseau IB. Si le switch qui héberge les noeuds de service ne fonctionnait plus, le cluster entier se met à dysfonctionner.

Étant donné que chaque équipement ne possède qu'un seul port InfiniBand pour ce réseau, on ne peut pas introduire de la redondance en connectant un équipement à plusieurs switches L1 : les L1 seront des SPOFs quoi qu'il arrive. Cependant, on peut répartir sur plusieurs switches L1 les noeuds dont les I/O sont les plus intensives (les routeurs) et les noeuds redondants (noeuds de service, noeuds maîtres, noeuds de login utilisés par les mêmes groupes).

Nous proposons donc de garder les 12 switches dédiés à AC/AL, et les 2 switches dédiés à AX, mais de répartir le reste des noeuds sur les 3 switches restants :

Noeuds	Ports IB L1	Commentaire
Switch L2 x6	iwh1n[1-17]p[1-6]	chaque L1 relié à tous les L2
Routeur x9	iwh1n[1-3]p[10-13]	répartition de charge, redondance
Service x9	iwh1n[1-3]p[15-17]	redondance
Maître x2	iwh1n[1-2]p20	redondance
Login x2 (groupes ugp[0-2])	iwh1n[2-3]p21	redondance
Login x3 (groupes ugp[3-4])	iwh1n[1-3]p22	redondance
Login x2 (groupes ugp[5-7])	iwh1n{1,3}p23	redondance
Login x3 (groupes ugp[8-9])	iwh1n[1-3]p24	redondance

ici iwh1[4-17] sont dédiés aux partitions AC, AL et AX

Ainsi, si l'un des 3 switches ne fonctionnait plus, seul l'un des deux noeuds maîtres n'aura plus accès au réseau IB, chaque groupe d'utilisateurs aurait accès à au moins un noeud de login ayant accès au réseau IB, et le débit vers Lustre est réduit. Si 2 des 3 switches ne fonctionnait plus, soit un groupe d'utilisateurs n'a plus aucun accès au réseau IB, soit les 2 noeuds maîtres n'y ont plus accès, et le débit vers Lustre est très réduit. De plus, la charge des routeurs est répartie équitablement sur les 3 switches L1.

7 Extensibilité de la solution

Le schéma du réseau IB a déjà été revu lors de la question précédente.

7.1 Sans rajouter de châssis

Dans la situation actuelle, il y a 12 châssis, chaque châssis contient 30 lames. D'après l'appel d'offres, il y a actuellement 352 nœuds, 300 de la partition AC et 50 de la partie AL. Or, chaque lame peut contenir 30 nœuds, nous pouvons donc rajouter 8 nœuds sans rajouter de châssis.

7.2 Sans rajouter de racks

1 châssis fait 16U, 1 racks en fait 48U. Avec les 12 châssis de la partie AC+AL, nous remplissons 4 racks. D'après le nombre de racks indiqué dans l'appel d'offres, il reste donc 1 rack vide pour la partition AC+AL. En rajoutant des châssis, nous pouvons alors le remplir. Cela rajouterait donc 3 châssis, donc 90 nœuds supplémentaires. Par conséquent, sans rajouter de racks, mais uniquement en rajoutant des nœuds et des châssis, ils seraient possibles de rajouter 98 nœuds (90 nœuds de cette partie + les 8 nœuds de la partie précédente).

7.3 facteurs limitant et extension de la partie AX

Le problème de la partie AX repose dans le fait qu'il n'y a qu'un GPU par nœud. Cela fait donc beaucoup de connexions aux différents switches. En effet, en comparant le nombre de nœuds AX et le nombre de nœuds de la partition AC+AL, nous pouvons voir que 50 nœuds AX occupent 50 ports sur chacun des réseaux du cluster (IB, administration, contrôle électrique), alors que 12 châssis avec 352 nœuds et la topologie proposée occupent 72 ports sur les switches (externes aux châssis) du réseau IB, 48 ports sur le réseau d'administration, et 24 ports sur le réseau de contrôle électrique. Il devient envisageable d'introduire des switches supplémentaires dans les racks de la partition AX, mais il y a une meilleure solution.

En optant pour des serveurs permettant de mettre plusieurs GPU dans un nœud tel que le "Serveur De Calcul APY SCG GPGPU 2U 4 GPU AMD EPYC Serie 7003", nous pourrions diminuer de manière conséquente le nombre de nœuds AX. Cela permettrait de mettre 4 cartes Nvidia A100 par nœud, augmentant grandement l'extensibilité sans rajouter de rack et sans augmenter de manière absurde le nombre de ports nécessaires.

Cependant, cela est coûteux et nécessiterait de revoir tous les nœuds AX et de refaire les branchements pour les switches.

7.4 Limitation de l'extensibilité

Cependant, même en appliquant les modifications précédentes, si le budget n'est pas une limite pour l'université, le débit et la latence le seront. En effet, il faut des connexions à chaque nœud lié, notamment, à la topologie, car un full mesh n'est pas faisable si le cluster continue de s'agrandir. Et même des topologies tel que dragonfly poseraient problème passé un certain nombre de nœuds.

De plus, un problème de refroidissement serait à prévoir si nous utilisons les noeuds avec 4 GPU proposés précédemment et que nous remplissons des racks avec ceux-ci. Une autre limitation pourrait être un problème d'espace lié au bâtiment et au poids de chaque rack. Ce qui forcerait une modification de l'infrastructure du bâtiment contenant le calculateur, ou une limitation du nombre de noeuds par rack, amenant à acheter plus de racks. Ce qui représente un budget non négligeable, que ce soit pour une université ou non.