Septembre 2022

Stratégie de calcul de haute performance de l'Alliance de recherche numérique du Canada

Groupe de travail sur la stratégie de calcul de haute performance

Scott Northrup (président), Bruno Blais, Roy Chartier, Rebecca Davis, Catherine Lovekin, Patrick Mann, John Morton, Florent Parent et Seppo Sahrakorpi

Alliance de recherche numérique du Canada | Digital Research Alliance of Canada

Table des matières

1 Résumé

1.1 Objectif

L'objectif de ce rapport est de présenter une stratégie de préservation et de développement de la plateforme nationale de calcul de haute performance (CHP) à la disposition des chercheuses et chercheurs du Canada. Le présent rapport donne également un aperçu des capacités et de l'état de la plateforme de CHP actuelle, propose une analyse de la demande et de la croissance prévues, et présente des scénarios de préservation, d'amélioration et de soutien de la plateforme.

1.2 Définition

Aux fins du présent document, le calcul de haute performance (CHP) correspond à l'utilisation de systèmes en grappes auxquels un ordonnanceur envoie des tâches de calcul. Sont incluses des ressources de stockage actives directes et, parfois, une gamme de ressources de calcul, notamment divers types d'accélérateurs. L'infonuagique et l'archivage sont toutefois exclus.

1.3 Besoins et vision stratégiques des chercheuses et chercheurs

Le rapport du Conseil de recherche sur les besoins prioritaires[1], les consultations et le rapport sur l'évaluation des besoins des chercheuses et chercheurs[2], et les rapports sur l'état actuel du calcul informatique de pointe (CIP)[3], de la gestion des données de recherche[4] et des logiciels de

[1] Conseil des chercheurs de l'Alliance. *Répondre aux besoins de la communauté de recherche du Canada en matière d'infrastructure de recherche numérique*. https://www.alliancecan.ca/sites/default/files/2022-03/Priorite%CC%81s-du-Conseil-des-chercheurs-28-septembre-2021.pdf (septembre 2021).

[2] Alliance de recherche numérique du Canada. *Évaluation des besoins de la communauté de recherche : résumé des commentaires reçus.* https://alliancecan.ca/sites/default/files/2022-03/EvaluationBesoins_Alliance_20220126.pdf (septembre 2021).

[3] Groupe de travail de l'Alliance sur le calcul informatique de pointe. *État actuel du calcul informatique de pointe au Canada.* https://alliancecan.ca/sites/default/files/2022-03/CIP_Rapport_Etat_Actuel.pdf (mai 2021).

[4] Groupe de travail sur la gestion des données de l'Alliance. *État actuel de la gestion des données de recherche au Canada*. https://alliancecan.ca/sites/default/files/2022-04/2020-11_GDR_Rapport_Etat_Actuel_1.pdf (novembre 2020).

recherche[5] ont permis à l'Alliance de recherche numérique du Canada (l'Alliance) de comprendre pleinement ce dont les chercheuses et chercheurs auront besoin dans l'écosystème canadien de l'infrastructure de recherche numérique (IRN). Ces rapports et les consultations supplémentaires avec les parties prenantes ont jeté les bases du Plan stratégique 2022-2025[6] de l'Alliance, document publié en février 2022 exposant la vision et les priorités stratégiques de l'Alliance.

Si les besoins des chercheuses et chercheurs concernent l'ensemble de l'écosystème, du calcul informatique de pointe aux logiciels de recherche en passant par la gestion des données de recherche, et de l'infrastructure aux services en passant par les opérations, la formation, le financement et des questions personnelles, ce rapport traite principalement des besoins liés aux infrastructures de CIP et de CHP, mais exclut l'infonuagique et les solutions de traitement des données sensibles, qui feront l'objet de stratégies distinctes. Dans ce cadre, le Plan stratégique et la communauté des chercheuses et chercheurs ont clairement établi la nécessité d'accroître considérablement les capacités de calcul informatique de pointe au pays en étoffant l'infrastructure et les services de CIP et de CHP de la Fédération Calcul Canada, notamment en améliorant l'accessibilité et la cybersécurité, et en bonifiant les solutions de stockage, par exemple le stockage à long terme et les solutions de préservation. Le Conseil des chercheurs a d'ailleurs encouragé l'Alliance à doubler les capacités de CIP et de CHP au cours de son premier mandat afin d'atteindre la capacité de calcul moyenne pondérée des pays du G7 par rapport au produit intérieur brut (PIB). En plus d'une hausse de la capacité, il faut trouver des solutions à l'utilisation inefficace des systèmes pour ainsi accroître la disponibilité et maximiser la valeur des investissements. Il faut travailler à rendre l'utilisation des systèmes de CIP et de CHP la plus efficace possible en offrant de la formation et en créant des logiciels et des flux de travail personnalisés et intégrés, adaptés aux besoins de chaque discipline (ex. : sciences humaines numériques). Les investissements en infrastructures de CIP, de CHP et de stockage doivent suivre un plan financièrement viable tenant compte des exigences de maintenance et de mise à jour des infrastructures. Il faudra également élaborer et mettre en place des mesures d'atténuation des risques pour prévenir les interruptions de service.

1.4 Recommandations concernant l'architecture de CHP

La demande des chercheuses et chercheurs en ressources de CHP augmente chaque année; il est plus que prioritaire de maintenir, d'accroître et de soutenir le développement des capacités.

[5] Groupe de travail de l'Alliance responsable des logiciels de recherche de pointe. *Évaluation de l'état actuel des logiciels de recherche*. https://alliancecan.ca/sites/default/files/2022-03/LR_Rapport_Etat_Actuel_0.pdf (septembre 2021).

[6] Alliance. *Plan stratégique 2022-2025*. https://stratplan.alliancecan.ca/wp-content/uploads/2022/03/Alliance_Plan_Strategique_2022_2025.pdf (février 2022).

L'analyse en profondeur exposée dans le présent rapport a mené aux recommandations principales suivantes (sommaire) :

Remplacer l'infrastructure essentielle

- Financer immédiatement le remplacement de l'infrastructure vieillissante pour éviter la diminution du service et conserver les capacités de base actuelles. D'ici la fin de 2024, 234 000 cœurs de processeurs centraux (CPU) et 2 200 processeurs graphiques (GPU) dans l'ensemble des systèmes de la Fédération auront plus de 5 ans et devront être remplacés pour maintenir les capacités de 2021. (Voir le scénario I à la section 4.2 pour en savoir plus.)

Augmenter les capacités de CHP

- Adopter une approche pluriannuelle mixte d'investissement dans les immobilisations et l'exploitation pour faire croître continuellement les capacités de CHP de base. Faire l'évaluation continue des plans de déploiement pour réagir aux changements dans la charge de travail et les exigences des projets de recherche, évaluer la demande et envisager l'utilisation de technologies de pointe.
- Augmenter la compétitivité de la recherche au Canada et soutenir la demande croissante.
    - Doubler la capacité actuelle pour faire passer le Canada de la dernière position à une position centrale parmi les pays du G7 en termes d'opérations en virgule flottante par seconde (flops)/$ de PIB. Pour ce faire, cibler l'ajout de 100 000 cœurs de CPU et 1 000 GPU de plus par année de 2023 à 2025 afin de doubler la capacité de CHP de 2021 d'ici 2025. (Voir le scénario IV à la section 4.2 pour en savoir plus.)
    - Investir pour qu'au moins l'un des systèmes se classe parmi les 50 premiers du palmarès TOP500[7], soit un système conçu pour les chercheuses et chercheurs qui ont des besoins en traitement massivement parallèle. S'assurer qu'au moins trois systèmes demeurent parmi les 250 premiers du palmarès TOP500.

Investir dans le stockage et la gestion des données

- Augmenter au même rythme la mémoire de CHP active et la capacité de CHP, qui elle, devrait être doublée d'ici 2025.
- Élaborer un plan de stockage à long terme intégrant gestion des données de recherche et archivage.
- Déployer une solution de stockage (dans tous les systèmes) pour les jeux de données courants et les données sensibles des utilisatrices et utilisateurs.

[7] TOP500. https://www.top500.org/lists/top500/ (novembre 2021).

Viser un CHP plus vert

- Mettre l'emphase sur l'efficacité du calcul (poids/flop) dans la conception de systèmes.
- Viser les investissements dans les infrastructures et les centres de données qui maximisent leur efficacité énergétique et de refroidissement tout en diminuant leurs coûts opérationnels et leur empreinte carbone.
- Développer des plans de cycle de vie pour l'équipement de CHP prévoyant, lorsque possible, la conversion de l'équipement afin de prolonger sa durée de vie utile.

Augmenter la résilience de la plateforme

- Investir dans la géoréplication des sauvegardes de données sensibles.
- Élaborer des plans de reprise après sinistre et d'atténuation qui concordent avec les objectifs de niveaux de services convenus.

Améliorer le soutien, l'efficacité et la convivialité

- Augmenter l'effectif de soutien à la recherche en fonction de la multiplication des ressources de CHP.
- Promouvoir des initiatives qui augmentent l'efficacité des systèmes et de leur utilisation.
- Augmenter le nombre d'options d'accès au-delà des interfaces de ligne de commande traditionnelles afin de permettre à un bassin plus diversifié de chercheuses et chercheurs d'utiliser les ressources de CHP.

2. État actuel

2.1 Ressources actuelles de CHP

L'actualisation de l'écosystème national de CHP a commencé en 2016, financée par les investissements de la Fondation canadienne pour l'innovation (FCI) dans la cyberinfrastructure canadienne, puis, en 2019, par Innovation, Sciences et Développement économique Canada (ISDE). Les ressources consolidées sont hébergées sur cinq sites principaux de la Fédération Calcul Canada (Calcul Canada). Les systèmes nationaux et les organisations membres de la Fédération affiliées sont les suivants (d'ouest en est) :

- Université de Victoria, Arbutus (Groupe de l'IRN de la C.-B., anciennement WestGrid)
- Université Simon-Fraser, Cedar (Groupe de l'IRN de la C.-B., anciennement WestGrid)
- Université de Waterloo, Graham (Compute Ontario)
- Université de Toronto, Niagara (Compute Ontario)
- Université McGill/Calcul Québec, Béluga et Narval (Calcul Québec)

Cedar, Graham, Béluga et Narval sont des grappes de CHP hétérogènes à usage général pouvant supporter une large gamme de tâches de CHP, communément appelées des grappes de calcul générique. Niagara est une grappe homogène massivement parallèle servant principalement pour les tâches de CHP à grande échelle et évolutives. C'est ce qu'on appelle généralement une grappe parallèle de grande taille. Le tableau 1 répertorie chaque grappe, leur date d'installation, leur capacité de calcul et le stockage allouable. Les ajouts aux systèmes de base ne sont pas indiqués.

Système	Date de mise en service	Cœurs de CPU	GPU	Stockage /project (To)
Béluga	Septembre 2019	32 080	688	31 000
Cedar	Mars 2017	94 528	1 352	42 000
Graham	Juin 2017	34 784	498	20 100
Narval	Septembre 2021	61 760	524	29 000
Niagara	Mars 2018	80 960	216	12 300
Total		304 112	3 278	134 400

Tableau 1 – Ressources nationales de CHP actuellement à la disposition des chercheuses et chercheurs du Canada (mars 2022)

S'ajoute à ces systèmes Arbutus, un système infonuagique pour l'hébergement (principalement sur Linux) de machines virtuelles et d'autres tâches infonuagiques. Des parties des grappes de calcul générique Cedar, Graham et Béluga servent également aux tâches infonuagiques.

Système	Date de mise en service	Cœurs de CPU	GPU	Stockage (To)
Arbutus	Septembre 2016	16 008	108	17 000
Béluga	Septembre 2019	3 072		2 000
Cedar	Mars 2017	1 216		1 300
Graham	Juin 2017	1 368		84
Total		21 664	108	20 384

Tableau 2 – Ressources nationales infonuagiques actuellement à la disposition des chercheuses et chercheurs du Canada (mars 2022)

2.2 Défis des systèmes de CHP

L'exposé de position de 2021 de l'Alliance sur le CIP présente un portrait précis des principaux défis de l'écosystème canadien de CIP. La présente section traite des défis propres au CHP.

2.2.1 Investissement continu

L'investissement continu dans les systèmes de CHP est essentiel pour conserver les capacités actuelles et doter les chercheuses et chercheurs d'une infrastructure robuste et moderne. Le financement soutenu et prévisible des coûts d'immobilisations et d'exploitation permettra de planifier adéquatement le cycle de vie de chaque système et de l'architecture de CHP dans son ensemble.

Les systèmes de CHP ont généralement une durée de vie utile de 3 à 5 ans, 4,2 ans[8] étant la moyenne selon Hyperion (2021). Les améliorations technologiques apportées aux systèmes pour assurer leur performance après cinq ans, surtout pour l'amélioration d'éléments de calcul essentiels des processeurs centraux et graphiques, leur font perdre leur compétitivité et leur rentabilité dans le cadre du CHP. Les composants peuvent à ce moment être réutilisés dans des tâches moins exigeantes. Cela illustre la nécessité de planifier une durée de vie plus longue et démontre que l'investissement continu est nécessaire pour conserver la capacité de CHP et proposer une ressource moderne et fiable.

Vu cette obsolescence programmée, il est essentiel de mettre à jour continuellement l'écosystème de CHP de multiples systèmes pour éviter les situations comme celle que nous vivons présentement, où une portion importante de l'équipement a ou aura bientôt cinq ans. Comme l'indiquent les dates des systèmes dans le tableau 1 ci-dessus, la grande majorité des systèmes de CHP au Canada ont été installés en 2017 et 2018, et pour simplement conserver leurs capacités, doivent être remplacés en 2022 ou 2023. Au-delà de la question de capacité, si le remplacement des systèmes spécialisés, comme le Niagara pour le calcul parallèle à grande échelle, n'est pas considéré, les chercheuses et chercheurs du pays perdront complètement les capacités actuelles.

2.2.2 Manque de ressources de CHP et compétitivité de la recherche

D'après l'exposé de position de 2017 du Conseil du leadership sur la gestion des données (CLIRN) sur le CIP, le rapport *État actuel du calcul informatique de pointe au Canada* (2021) et le rapport *Évaluation des besoins de la communauté de recherche* (2021), le nombre de ressources de CHP ne répond pas à la demande actuellement. Chaque année, le nombre

[8] HPCWire. *Hyperion SC21 Market Update: 2021 Looks Strong (Surprise!); Big Systems, Cloud and AI Are Drivers*. https://www.hpcwire.com/2021/11/15/hyperion-sc21-market-update-2021 (consulté en mai 2022).

d'utilisatrices, d'utilisateurs et de groupes de recherche augmente, et il en va de même pour la demande en ressources de CHP.

Le Canada est le pays du G7 au rang le plus bas dans le palmarès TOP500 en matière de puissance totale de calcul. En ce qui a trait à la puissance de calcul par rapport au produit intérieur brut (Tflops/PIB), le Canada est encore une fois le moins bien classé du G7. « Les faibles capacités de calcul font de faibles concurrents. »[9] Il faut au minimum doubler la capacité de CHP pour fournir aux chercheuses et chercheurs du Canada les outils nécessaires à la réalisation de recherche de pointe (Stratégie pour l'infrastructure de recherche numérique de l'ISDE)[10] et pour amener le Canada au centre du rang du G7.

2.2.3 Gestion des données et résilience

En ce moment, les systèmes et les sites d'hébergement fonctionnent essentiellement indépendamment les uns des autres, à l'exception de quelques services partagés, comme l'authentification et le logiciel centralisé, qui sont conçus pour avoir une haute disponibilité en cas de panne. Cela signifie qu'en cas de panne majeure sur un site, les autres peuvent continuer de fonctionner, et les chercheuses et chercheurs sont en mesure de transférer toute tâche essentielle dans l'un des autres systèmes. Cependant, toutes les données sont présentement propres à un système donné; si ce dernier se déconnecte complètement, les données deviennent inaccessibles. Cet enjeu majeur empêche les chercheuses et chercheurs de passer d'un système à l'autre.

Voici ce qui aiderait grandement les chercheuses et chercheurs à passer d'une ressource à l'autre plus facilement, et améliorerait par le fait même la résilience générale : une stratégie nationale concernant les données ne tenant pas seulement compte du stockage haute performance traditionnel prévu pour le stockage de données à court et moyen terme, mais également de l'archivage à long terme, de la sauvegarde de données sensibles hors site et intersite, et de l'intégration de la gestion des données de recherche. À noter cependant que les systèmes de CHP traitent souvent des jeux de données très volumineux qui, pour être performants, doivent être locaux, et qu'il n'est peut-être pas possible ou même souhaitable de considérer leur reproduction. Pour ces raisons, les modèles et les politiques de stockage doivent être élaborés judicieusement en tenant compte des flux de données de CHP.

2.2.4 Utilisation par les chercheuses et chercheurs et convivialité

Le Canada compte environ 33 000 professeures et professeurs universitaires (titulaires et agrégés). La Fédération Calcul Canada évalue à environ 5 500 le nombre de comptes de

[9] *1st Annual High Performance Computing Users Conference*, Washington, D.C., 13 juillet 2004, rapport de conférence. https ://compete.org/2004/03/16/supercharging-u-s-innovation-competition/

[10] ISDE, *Infrastructure de recherche numérique*. https://ised-isde.canada.ca/site/infrastructure-recherche-numerique/fr (consulté en mai 2022).

chercheuses principales et chercheurs principaux, ce qui signifie qu'approximativement 17 % de ces professeures et professeurs ont créé un compte pour utiliser l'infrastructure de la Fédération. Qui plus est, le corps professoral en sciences humaines et sociales, en commerce et en psychologie représente environ 10 % de ces utilisatrices et utilisateurs, alors qu'il constitue environ 46 % du corps professoral à temps plein dans les universités canadiennes.

Comme la communauté de recherche utilisant le CHP s'élargit, la demande de systèmes compatibles avec de nouvelles interfaces et de nouveaux flux de travail augmente. L'accès aux ressources de CHP se fait principalement par lignes de commande. Or, pour des raisons de convivialité, les solutions comme les infrastructures de bureaux virtuels ou l'accès par navigateur Web, comme les calepins Jupyter Notebook ou les portails Web, gagnent en popularité.

3 Analyse de la demande en ressources de CHP

3.1 Demande de capacité de calcul

L'un des grands problèmes de l'écosystème actuel est que malgré les investissements considérables en ressources de CHP faits récemment, la demande continue de dépasser l'offre.

Qui plus est, quantifier la demande n'est pas chose facile vu le grand nombre de chercheuses et chercheurs sollicitant des ressources et la pluralité de leurs besoins de calcul.

3.1.1 Demande d'allocation de ressources de calcul via le concours

En l'état, l'allocation de 80 % des ressources de la plateforme nationale se fait par l'intermédiaire d'un concours annuel. Les statistiques du concours 2022 pour l'allocation de ressources sont présentées dans le tableau 3. On y voit que la demande dépassait de beaucoup l'offre, particulièrement pour le calcul à l'aide de processeurs centraux et de processeurs graphiques.

Ressource¤	Capacité allouée	Demande	Ratio
Cœurs de processeurs centraux (CPU)	238 950	435 672	**1,8×**
Processeurs graphiques (GPU)	2 450	9 622	**3,9×**

Tableau 3 – Offre et demande dans le concours 2022 pour l'allocation de ressources

Les ressources de CPU sont allouées en cœurs-années (CPU), soit une mesure équivalant à exécuter un programme sur un seul cœur d'un processeur central pendant une année

complète. Les ressources GPU sont allouées en GPU-années, soit une mesure équivalant à exécuter un programme sur un seul processeur graphique pendant une année complète. À noter qu'afin de simplifier le processus d'allocation, l'une et l'autre mesure ne tiennent pas compte des différences dans l'architecture des processeurs centraux et des processeurs graphiques, architecture qui peut varier selon les systèmes. Et si la puissance de calcul tend à rester assez stable chez les processeurs centraux de différentes générations, elle se voit souvent décuplée d'une génération de processeurs graphiques à la suivante. Cette variabilité de la puissance graphique rend particulièrement ardues l'estimation de leurs besoins réels par les chercheuses et chercheurs ainsi que la prédiction de la demande future.

Si l'on regarde les chiffres d'allocation de ressources via le concours dans la décennie de 2012 à 2022, on voit que le nombre de demandes a augmenté de manière constante, et que le nombre moyen de cœurs-années (CPU) sollicités par demande est resté relativement stable. Il est possible de partir de cette tendance et de la moyenne de cœurs sollicités par demande pour extrapoler ce que sera généralement la demande future. Ces projections des besoins sont décrites à l'annexe A – plus précisément à la figure A.4 pour les processeurs centraux, et à la figure A.5 pour les processeurs graphiques. Par exemple, la demande estimée dans le cadre du concours 2024 pour l'allocation de ressources serait de 520 000 cœurs-années et de 14 400 GPU-années.

3.1.2 Analyse de la charge de travail en CHP – Demande en attente

Comme on l'a vu à la section précédente, les données de l'allocation de ressources dans le cadre du concours peuvent faire office de mesure de la demande… mais elles ne rendent compte que des demandes présentées, et non de l'usage réel des systèmes. De plus, les données du concours ne tiennent pas compte de la demande où un grand nombre d'utilisatrices et utilisateurs sollicitent les ressources de CHP, mais ne présentent pas une demande d'allocation dédiée chaque année. Plusieurs utilisatrices et utilisateurs ne soumettent même pas de demande au concours pour l'allocation de ressources en raison de toutes les contraintes, et vont par exemple plutôt passer par leurs collaborations de recherche pour obtenir des ressources accessibles internationalement. Cela dit, comme les systèmes de CHP sont tous programmés pour un ordonnancement en lots, il est possible d'étudier dans le temps les paramètres des tâches qu'ils abattent – taille, temps d'exécution, délai d'attente, mémoire, etc. – pour faire ressortir les tendances et évaluer la demande.

Un des angles d'analyse intéressants est de regarder la demande en tant que ratio entre la capacité du système et la charge de travail en attente.

Ce type d'analyse a été appliqué à des systèmes de calcul générique (Graham, Cedar, Béluga); les observations sont rapportées en détail à l'annexe B.1. Il en ressort que les ressources sont toujours sursollicitées de l'ordre de quatre à cinq fois leur capacité, et ce malgré d'importantes expansions qui ont grandement accru l'offre. On peut tirer ici le constat qu'il y a une accumulation considérable de la demande qui est loin d'avoir pu être traitée. On peut aussi voir la constance de cette sursollicitation, même avec les gains de capacité qui ont été faits, principalement comme un indicateur de la patience des chercheuses et chercheurs, la limite de leur tolérance

correspondant au délai d'attente. Cette donnée pourra potentiellement servir d'indicateur pour confirmer la satisfaction des besoins des chercheuses et chercheurs : on la verrait reculer avec l'introduction de nouvelles capacités de calcul. Enfin bref, le fait qu'il y ait constamment une importante charge de travail en attente trahit le besoin d'une expansion substantielle des ressources de CHP, car tout cela est signe que l'on est encore loin de subvenir à la demande.

3.2 Compétitivité de la recherche

Actuellement au Canada, quatre des systèmes de la Fédération figurent dans le palmarès le plus récent (novembre 2021) des super-ordinateurs les plus puissants au monde selon TOP500 : il s'agit de Narval (83^{e}), Niagara (127^{e}), Cedar (137^{e}) et Béluga (288^{e}). Le champion actuel – Fugaku, du Japon – est 76 fois plus rapide que le meilleur des systèmes canadiens. Sont en train d'être mis sur pied trois super-ordinateurs aux États-Unis qui aspirent à dépasser 1 exaflops (10^{18} flops) cette année, et deux autres ont déjà été mis en marche en Chine[11]. L'ère de l'exa-informatique est arrivée… mais le Canada est à peine entré dans celle de la péta-informatique.

Certes, les systèmes de cette magnitude sont des ressources extrêmement dispendieuses à acquérir et à exploiter; la dépense n'est pas réaliste pour le Canada actuellement. Cependant, si le pays ne persiste pas à investir dans la mise à niveau de ses ressources, il se verra bientôt éjecté du palmarès des super-ordinateurs, ce qui viendra plomber la compétitivité des chercheuses et chercheurs canadiens sur la scène internationale. L'ISDE a même fixé explicitement parmi les indicateurs de rendement pour son programme d'IRN le maintien du nombre de machines de CIP dans les 250 premières au palmarès TOP500[12]. Considérant que l'augmentation de la puissance de calcul a été assez constante dans les 20 dernières années (voir la figure 1), on peut s'attendre à ce que l'ordinateur qui arrivera 500^{e} dans la liste en 2025 ait une puissance d'environ 5 pétaflops. S'il ne s'ajoute aucun nouveau système au Canada d'ici là, le pays ne comptera plus qu'un système, Narval, qui arrivera à se qualifier –, et ce en queue de peloton – avec ses 5,88 pétaflops.

[11] The Next Platform. *China Has Already Reached Exascale – On Two Separate Systems*. https://www.nextplatform.com/2021/10/26/china-has-already-reached-exascale-on-two-separate-systems/ (consulté en juillet 2022).

[12] ISDE. *Le Programme de contributions pour l'infrastructure de recherche numérique : guide du programme.* https://ised-isde.canada.ca/site/infrastructure-recherche-numerique/sites/default/files/attachments/ProgrammeContributionsIRN_Guide-du-programme.pdf (avril 2019).

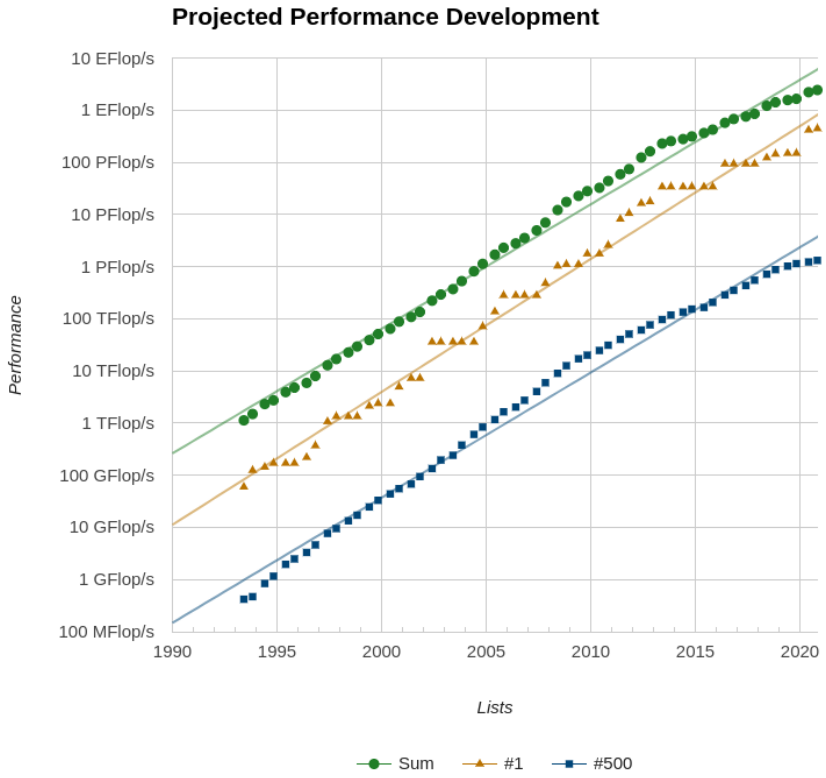

Figure 1 – Puissance des ordinateurs au palmarès TOP500 dans le temps. L'augmentation de la puissance de calcul au fil du temps selon TOP500. Si l'on ne change rien aux systèmes de CHP de l'Alliance, en 2025, le Canada ne comptera plus qu'un système, Narval, au palmarès TOP500, et celui-ci se retrouvera en queue de peloton.

Sinon, plutôt que la position et le nombre de systèmes au palmarès TOP500, une mesure qui serait potentiellement plus utile serait celle de la puissance de calcul totale proposée par pays, normalisée en fonction du PIB. C'est cette information que montre le tableau 4 ci-dessous : il compare entre eux les pays figurant au palmarès TOP500 de novembre 2021 (ceux du G7 sont tramés en vert), sur la base de leurs totaux. Il est à noter que ces statistiques sont produites à partir des chiffres présentés tels quels par TOP500, sans séparation des entrées entre systèmes destinés à la recherche ou réservés à l'usage des établissements d'enseignement. Le tableau est classé selon la puissance de calcul des pays en fonction du PIB. Le Canada arrive en dernier parmi le G7 et en 17^{e} position en général, avec 18,0 téraflops R_{max} agrégés sur le PIB. Pour suivre le rythme de ses pairs du G7 et fournir la même quantité de ressources à ses chercheuses et chercheurs, le Canada doit viser aux alentours de 40 téraflops sur le PIB. Pour y arriver, il faudrait au minimum doubler la capacité de CHP disponible actuellement au pays. C'est un autre signe que les ressources sont insuffisantes pour les chercheuses et chercheurs du Canada, ce qui désavantage ceux-ci par rapport à la concurrence en recherche dans le reste du monde.

Rang – flops/PIB	Rang – TOP500	Pays	Entrées dans le palmarès TOP500	Pétaflops R_{max} agrégés[13]	% du PIB allant à la R-D[14]	Téraflops par milliard USD de PIB
1er	2e	Japon	32	628,2	3,19	128,9
2e	9e	Arabie saoudite	6	55,3	-	80,5
3e	6e	Corée du Sud	7	82,2	4,64	53,7
4e	15e	Finlande	3	13,4	2,795	53,1
5e	1er	États-Unis	149	986,5	3,067	50,6
6e	4e	Allemagne	26	181,4	3,19	49,1
7e	5e	France	19	117,0	2,196	45,3
8e	21e	Tchéquie	2	9,6	1,942	44,4
9e	3e	Chine	173	530,1	2,235	43,3
10e	11e	Pays-Bas	11	35,9	2,184	43,2
11e	7e	Italie	6	78,5	1,466	40,4
12e	13e	Suisse	3	26,2	2,476	38,6
13e	8e	Russie	7	73,7	1,039	28,5
14e	23e	Émirats arabes unis	2	9,0	-	23,6
15e	18e	Suède	4	12,3	3,388	23,0

[13] TOP500. https://www.top500.org/statistics/list/ (novembre 2021).

[14] Organisation de coopération et de développement économiques (OCDE). *Dépenses intérieures brutes de R-D – données 2019*. https://data.oecd.org/fr/rd/depenses-interieures-brutes-de-r-d.htm (consulté en juillet 2022).

16e	10e	Royaume-Uni	11	54,9	1,756	20,8
17e	**12e**	**Canada**	**11**	**29,6**	**1,592**	**18,0**

Tableau 4 – Comparaison des pays figurant au palmarès TOP500 en novembre 2021, classés en fonction de leurs téraflops R_{max} agrégés sur le PIB; les membres du G7 sont en vert

3.3 Demande de stockage

En ce qui concerne le CHP, le stockage se fait typiquement en plusieurs niveaux qui appliquent différentes politiques et technologies. En règle générale, plus rapide et plus performant est ce stockage, plus cher il est par téraoctet (To), mais meilleures sont sa bande passante et sa capacité de traitement d'opérations d'entrée-sortie par seconde (IOPS). Les supports les plus économiques par téraoctet (généralement le ruban magnétique) ne sont vraiment utiles que pour la conservation à long terme nécessitant très peu d'accès en lecture/écriture.

Les systèmes actuels de la Fédération emploient la hiérarchisation que voici pour le stockage :

- **/scratch** – Système de fichiers de haute performance qui s'exécute en parallèle pour le stockage temporaire durant les calculs. Un quota élevé est offert par système. Il n'y a pas de sauvegarde, et la politique en place veut que les fichiers soient purgés.
- **/home** – Système de fichiers modérément performant s'exécutant en parallèle et visant principalement/seulement les données sensibles. Un petit quota est fixé par système, et une sauvegarde s'effectue localement.
- **/project** – Système de fichiers modérément performant s'exécutant en parallèle et se prêtant au stockage des données de recherche de groupe. L'allocation se fait dans le cadre des concours, et une sauvegarde s'effectue localement.
- **dCache** – Système de stockage initialement conçu pour les données préliminaires en physique des hautes énergies. L'allocation se fait dans le cadre des concours.
- **/nearline** – Système de basse performance (principalement le ruban magnétique) pour le stockage à long terme sans usage actif. L'allocation se fait dans le cadre des concours.

3.3.1 Allocation de ressources de stockage de CHP via le concours

À l'instar des ressources de calcul, c'est dans le cadre du processus de concours que sont alloués aux demandeuses et demandeurs trois des types d'espaces de stockage : /project, dCache et /nearline. Les données sur le stockage sollicité dans le concours 2022 pour l'allocation de ressources figurent dans le tableau 5. Contrairement à la demande de ressources de calcul, celle visant le stockage a généralement pu être satisfaite, quoiqu'elle grandisse d'environ 20 % par année; voir l'annexe A pour l'analyse dans le temps de la demande en stockage via le concours.

En l'état actuel des choses, les allocations d'espace de stockage comme celles de ressources de calcul sont locales à chaque système.

Ressources de stockage	Capacité allouée	Demande	Ratio
/project	52,1 Po	62,3 Po	**1,2×**
dCache	13,1 Po	13,1 Po	**1,0×**
/nearline	74,3 Po	74,3 Po	**1,0×**

Tableau 5 – Allocation d'espace de stockage de CHP via le concours 2022

3.3.2 Stockage à long terme et archivage

Selon le processus de concours pour l'allocation de ressources actuel, l'allocation de l'espace de stockage se fait sur une base annuelle la majeure partie du temps, mais elle peut aussi être bis- ou trisannuelle pour certains projets. Si une équipe de recherche ne renouvelle pas son allocation, elle risque de voir ses données être purgées, et c'est pourquoi la communauté réclame une solution de stockage à long terme et d'archivage qui s'intégrerait dans la GDR. C'est toutefois une question qui dépasse le mandat du groupe de travail sur le calcul de haute performance (CHP), et qui sera plutôt traitée par le futur groupe de travail sur le stockage.

3.3.3 Stockage infonuagique

Le stockage infonuagique n'entre pas explicitement dans le mandat du groupe de travail sur le CHP, mais on l'examine tout de même ici puisque le stockage objet se montre prometteur comme option de stockage supplémentaire ou complémentaire directement intégrée aux futurs systèmes de CHP.

L'offre venait égaler la demande pour plusieurs des options de stockage infonuagique dans le cadre du concours 2022 pour l'allocation de ressources (voir le tableau 6), mais si l'on n'augmente pas la capacité, cela ne continuera pas d'être le cas. Aussi, la majorité du matériel infonuagique, y compris les supports de stockage (voir le tableau 2), date de plus de cinq ans et devra bientôt être remplacée. On s'attend à ce que le stockage objet gagne en popularité, surtout s'il vient à être offert directement à même les grappes de CHP, ce qui en ferait une option viable pour le stockage de données partagées.

Ressources de stockage infonuagique	Capacité allouée	Demande	Ratio
Instantanés de volumes	3,0 Po	3,6 Po	**1,2×**
Stockage objet	7,3 Po	7,3 Po	**1,0×**
Nuage partagé	1,6 Po	1,6 Po	**1,0×**

Tableau 6 – Allocation d'espace de stockage infonuagique via le concours 2022

3.4 Demandes particulières en ressources de CHP auxquelles répond l'Alliance

Le rapport du Conseil de recherche sur les besoins prioritaires, les consultations et le rapport sur l'évaluation des besoins des chercheuses et chercheurs, et les rapports sur l'état actuel ont permis à l'Alliance de comprendre pleinement ce dont les chercheuses et chercheurs auront besoin dans l'écosystème canadien de l'infrastructure de recherche numérique (IRN). La communauté des chercheuses et chercheurs a clairement établi la nécessité d'accroître les capacités de calcul informatique de pointe au pays en étoffant l'infrastructure de CHP de la Fédération Calcul Canada.

Dans le cadre du processus de consultation sur l'évaluation des besoins en recherche, on a demandé aux chercheuses et chercheurs de produire des exposés de position. Certains de ces exposés présentaient leurs projections concernant les ressources de CHP ainsi que les exigences opérationnelles; en voici un résumé à titre informatif.

- L'exposé *Large-Parallel Supercomputer Simulations – Frontiers in Canadian Research*[15] fait état du besoin d'un super-ordinateur composé d'environ 10 000 nœuds homogènes pour effectuer des simulations à grande échelle. Autrement dit, il faudrait ici un système

[15] R. Fernandez et coll. *Large-Parallel Supercomputer Simulations – Frontiers in Canadian Research,* exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/ndrio_wp_on_niagara_scale_systems.pdf (décembre 2020).

de nature homogène dont la capacité dépasse de 5 à 8 fois celle de Niagara pour 2023, à savoir un système de l'ordre de 30 à 50 pétaflops (390 à 650 000 cœurs de CPU).

- Dans *DRI Needs Assessment for the Computational Fluid Dynamics (CFD) Research Community*[16], la communauté de recherche en mécanique des fluides numérique, qui s'est vue allouer 28 777 cœurs-années (CPU) en 2020, insiste sur l'importance pour son travail actuel et futur que soient renouvelées les ressources du côté des grandes grappes homogènes parallèles ainsi que des serveurs haute performance d'entrées et sorties de fichiers et des installations réseau connexes.

- L'exposé *Digital Research Infrastructure for Canadian Astronomy*[17] fournit l'estimation générale que la communauté de recherche en astronomie aura besoin d'un temps de traitement de 100 pétaflops-années par processeurs graphiques (5 000 GPU-années) et de 100 pétaflops-années par processeurs centraux (1,3 M de cœurs-années [CPU] – à savoir plus de quatre fois la capacité totale des processeurs centraux des systèmes de la Fédération) ainsi que 75 Po d'espace de stockage en ligne à l'horizon 2025, sans compter les ressources nécessaires à des projets distincts comme le centre régional du SKA.

- Dans *White Paper on Canada's Future DRI Ecosystem – Subatomic Physics in Canada*[18], l'Institut de physique des particules et l'Institut canadien de physique nucléaire décrivent leur besoin d'un fonctionnement et d'un soutien assurés 24 h sur 24, 7 jours sur 7 pour les systèmes de CHP.

- L'exposé *Canada's Future DRI Ecosystem: AI Research Needs*[19] met de l'avant les besoins particuliers de la communauté de recherche en intelligence artificielle, notamment le besoin que s'ajoute une capacité spécialisée pour le travail avec l'IA.

[16] Société canadienne de CFD. *DRI Needs Assessment for the Computational Fluid Dynamics (CFD) Research Community*, exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/dri-needs-assessment-of-the-cfd-community.pdf (décembre 2020).

[17] C. Lovekin et coll. *Digital Research Infrastructure for Canadian Astronomy*, exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/white-paper-on-dri-for-astronomy.pdf (décembre 2020).

[18] Institut de physique des particules (IPP) et Institut canadien de physique nucléaire (ICPN). *White Paper on Canada's Future DRI Ecosystem – Subatomic Physics in Canada*, exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/sap-white-paper.pdf (décembre 2020).

[19] Vector Institute. *Canada's Future DRI Ecosystem: AI Research Needs*, exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/canadas-future-dri-ecosystem_-ai-research-needs.pdf (décembre 2020).

3.5 Projets externes associés à l'Alliance

On compte une poignée de projets de recherche majeurs dont les importants besoins en ressources de calcul sont financés autrement que par l'Alliance. Ces projets ont aussi certains volets qui sollicitent ou vont probablement solliciter les ressources de l'Alliance; par conséquent, il est bon de les garder à l'œil pour garantir que les attentes et la planification concordent. Certains de ces projets présentent également d'importantes exigences de réseautique, ce qui rend particulièrement importante la coordination avec le RNRE et CANARIE.

3.5.1 Physique des hautes énergies

Le Centre canadien d'accélération des particules, TRIUMF, est à l'avant-garde mondiale en physique nucléaire, en physique des particules et dans le domaine des accélérateurs. Le laboratoire exploite plusieurs grappes dédiées et ressources de stockage afin de mener de grandes initiatives de recherche comme le centre de données de premier niveau ATLAS. Une bonne partie des analyses se fait au moyen des ressources de l'Alliance : les systèmes servant aux travaux requièrent différents services, des connexions réseau et un stockage spécialisé (dCache). À l'heure actuelle, la recherche se fait avec l'aide des grappes Cedar, Graham et Arbutus, en plus d'une allocation dCache d'environ 13 pétaoctets. Il est attendu que les besoins en stockage dCache augmentent de 20 % par année. Voir l'exposé de position *Role of TRIUMF within the Digital Research Infrastructure Ecosystem* pour en savoir plus.

D'autres ressources de calcul importantes ont été promises pour les expériences Belle-II, T2K, IceCube, SNOLAB et GlueX, lesquelles comptent toutes une solide participation canadienne. La majorité des ressources de calcul actuelles sont celles fournies par l'Alliance : environ 10 000 cœurs-années (CPU) et plus de 10 pétaoctets sur différentes plateformes de stockage.

3.5.2 Square Kilometer Array (SKA1)

L'une des grandes priorités en astrophysique au Canada relevée dans le Plan à long terme 2020 (Barmby, Gaensler et coll., 2020) est l'instauration d'un centre régional du SKA. Il s'agit d'un des centres de données qui vont collectivement traiter les 5 téraoctets de données que produit le télescope chaque seconde. Le Canada a pris le pari de fournir 6 % de la capacité requise planétairement. Si les choses se poursuivent en l'état, il est attendu qu'à compter de 2029, le Canada fournira :

- 9,7 pétaflops-années de capacité de traitement (126 000 cœurs-années [CPU] ou 485 GPU-années);
- 238 pétaoctets-années de capacité de stockage en ligne (soit pour une année complète de données);
- 654 pétaoctets-années de stockage /nearline (quantité croissante chaque année).

Comme le projet débute à peine, on ne sait pas encore comment ces ressources seront fournies ni comment le tout s'intégrera aux plateformes nationales de l'Alliance.

3.5.3 Environnement informatique pancanadien de l'IA (EIPIA)

Les trois instituts canadiens d'intelligence artificielle – Vecteur, Mila et Amii –, en concertation avec l'Alliance et le CIFAR, sont en train d'investir dans trois nouveaux systèmes spécialement conçus pour la recherche en IA. Ceux-ci représenteraient l'ajout d'un parc substantiel de près de 7 250 processeurs graphiques dans les cinq prochaines années pour traiter des calculs spécialisés. C'est principalement les chercheuses et chercheurs affiliés aux trois instituts qui l'utiliseraient, mais une partie de la capacité serait mise à la disposition du reste du monde de la recherche. Il sera important que la coordination se fasse avec l'Alliance pour assurer la bonne intégration des systèmes à l'écosystème national.

3.6 Systèmes contributifs intégrés (FCI)

Les chercheuses et chercheurs présentent des demandes afin d'obtenir du financement pour leurs systèmes, le plus souvent adressées à la FCI (dans le cadre du FLJE ou du FI), mais parfois aussi à d'autres bailleurs de fonds. L'Alliance a publié des lignes directrices[20] sur la manière dont on devrait considérer ces systèmes. En effet, ceux-ci peuvent dans certains cas comporter des composants matériels inédits ou encore se prêter à des usages particuliers qui sont incompatibles avec la plateforme nationale, et donc être vus non pas comme des systèmes contributifs, mais bien comme des systèmes exploités par un groupe de recherche ou d'autres équipes d'appoint locales externes à la plateforme nationale.

Les systèmes contributifs qui sont semblables ou identiques à d'autres systèmes de la plateforme nationale, s'ils sont intégrés correctement, ne demandent aucun temps de personnel supplémentaire pour leur exploitation. Ajouter quelques nœuds de plus à une grappe ne change pas grand-chose non plus au travail global d'administration du système si tout est pleinement intégré au système hôte et que rien n'est trop spécialisé. L'expansion des ressources disponibles bénéficie non seulement au groupe contributeur qui en a besoin, mais aussi à l'ensemble des utilisatrices et utilisateurs.

Des coûts vont parfois s'ajouter pour certains logiciels, comme les ordonnanceurs de tâches qui font l'objet d'une licence par nœuds, mais la hausse reste très minime. La facture d'énergie va aussi grimper vu le besoin d'alimenter et de refroidir le matériel informatique, mais cela reste vrai qu'il s'agisse ou pas de systèmes contributifs. S'ils sont contributifs, les systèmes pourront tirer avantage de l'infrastructure en place dans les installations nationales, par exemple les nœuds en

[20] Alliance. *Nouvelles directives politiques sur le matériel informatique intégré (systèmes contribués)*. https://alliancecan.ca/sites/default/files/2022-03/Nouvelles-directives-politiques-sur-le-mate%CC%81riel-informatique-inte%CC%81gre%CC%81-syste%CC%80mes-contribue%CC%81s.pdf (octobre 2021).

charge du réseau, de la connexion et de l'administration, ce qui permettra au groupe contributeur d'acquérir une infrastructure informatique plus vaste que ce que ses moyens lui permettraient normalement.

Certains organismes de financement jugent durement les demandes visant des ressources de calcul, s'ils ne les refusent pas carrément, sous prétexte que la recherche pourrait être menée sur la plateforme nationale. Or, cela n'est vrai que si ladite plateforme est suffisamment équipée pour prendre en charge le travail exigé.

Il est important de tenir compte de la dimension temporelle dans l'ajout de matériel contributif. Greffer de nouveaux appareils à un système qui date de plus de trois ans va créer un déséquilibre dans la capacité et la durée de vie utile de l'ensemble du matériel. Idéalement donc, les contributions devraient se faire tôt dans le cycle opérationnel du système principal; celles qui arrivent plus tard devraient plutôt s'ajouter à un autre système plus récent.

Le financement des systèmes contributifs s'accompagne souvent d'une participation provinciale de contrepartie. Il est donc difficile, voire impossible, de se prévaloir de cette contribution dans le cas d'un système national hors de la province en question, même si c'est là où ladite contribution serait la mieux venue.

3.7 Collaborations internationales

L'Alliance devrait examiner les possibilités de collaboration pouvant donner aux chercheuses et chercheurs du Canada un accès à certains systèmes spécialisés ou encore très vastes comme on en trouve aux États-Unis (ex. : systèmes ACCESS de la National Science Foundation, super-ordinateurs du département de l'Énergie) ou en Europe (ex. : systèmes PRACE et EuroHPC de l'Union européenne).

3.8 Concordance avec les appels de financement, les trois conseils et les autres initiatives

L'Alliance devrait s'assurer que ses investissements dans le CHP et les autres ressources d'IRN cadrent bien avec les initiatives 2022-2027 des trois conseils qui pourraient nécessiter beaucoup de ressources.

3.9 Ressources infonuagiques

Le présent document ne s'étend pas sur la stratégie infonuagique ni la charge de travail qui doit se faire sur le nuage, comme la question est plutôt du ressort du groupe de travail

spécial de l'Alliance qui planche sur la stratégie infonuagique. C'est dans le cadre de cette stratégie que l'on intégrera les considérations concernant les différentes tâches, y compris l'emploi des ressources infonuagiques pour le CHP. Cela vaudra probablement la peine d'explorer certaines options, par exemple le recours à une infrastructure comme Magic Castle afin de proposer un environnement de CHP familier aux utilisatrices et utilisateurs pour les charges intensives ou le traitement des urgences. Les composants matériels des ressources infonuagiques actuelles de la Fédération sont présentés dans ce rapport par souci d'exhaustivité.

4 L’architecture de CHP du futur

4.1 Maintien des capacités de base – besoins immédiats

Comme on l’a vu à la section 2.1, l’écosystème actuel compte environ 300 000 cœurs de processeurs centraux (CPU) et 3 000 processeurs graphiques (GPU) au total, pour une capacité de calcul combinée de 40 pétaflops. Cela dit, l’âge de toutes ces machines varie; la mise en service de certaines est toute récente, mais pour d’autres, elle remonte à plus de 5 ans (voir le tableau 1). Échelonner les acquisitions dans le temps est généralement une bonne stratégie, car le remplacement de l’équipement se fait petit à petit, au profit des toutes dernières technologies. Cependant, elle part aussi du principe que la durée de vie typique d’un système sera de 5 ans, après quoi on le remplacera.

Nous en sommes là aujourd’hui avec les systèmes Cedar et Graham, dont une bonne partie des ressources a dépassé l’âge de remplacement. Plus précisément, c’est 60 000 de leurs cœurs (CPU) et 900 de leurs GPU ainsi que la majorité du stockage principal qui ont passé le cap des 5 ans en 2022. D’ici la fin de 2024, toutes les ressources déployées dans la Fédération, à l’exception du système Narval qui vient d’entrer en ligne, auront plus de 5 ans et devront être remises à niveau. Si l’on n’investit pas immédiatement pour remplacer ces équipements vieillissants, la capacité disponible va bientôt décliner massivement.

Il est certes possible de continuer d’exploiter les machines passé leur 5^{e} et dernière année de vie prévue, mais cela peut s’avérer très coûteux : celles-ci requièrent des garanties prolongées et une maintenance accrue, les pannes et défectuosités se font plus fréquentes, et le rendement par Watt est beaucoup plus faible que pour les nouveaux équipements.

C’est d’autant plus vrai pour les technologies qui évoluent rapidement, comme celle des processeurs graphiques. Comparons par exemple un GPU NVIDIA P100 de 5 ans à un GPU A100 de dernière génération : les deux consomment la même énergie – 300 W –, mais l’un a une puissance de 4,3 téraflops et l’autre, de 19,5 téraflops. C’est une différence de 4,5 fois le rendement en flops/W. Le second GPU présente aussi le double de mémoire, une plus grande bande de mémoire, et toutes sortes de fonctionnalités qui n’existent pas chez le premier.

4.2 Projections de la capacité de CHP

Considérant les prévisions du côté de la demande en ressources de CHP que l’on a vues à la section 3, on projette ici quatre scénarios différents (résumés dans le tableau 7 et chiffrés dans le tableau 8) :

I. Remplacement suivant les besoins critiques pour maintenir la capacité.

II. Ajout de la capacité requise pour satisfaire 100 % de la demande en CPU et en GPU attendue dans le cadre du concours 2024 pour l'allocation de ressources.
III. Ajout de la capacité requise pour satisfaire 70 % de la demande en CPU et 50 % de la demande en GPU attendue dans le cadre du concours 2024.
IV. Ajout de capacité supplémentaire pour amener le Canada en milieu de peloton du G7 au palmarès TOP500 pour ce qui est des flops agrégés sur le PIB (implique de doubler la capacité actuelle).

4.2.1 Scénarios de remplacement et d'expansion

	Capacité totale actuelle (mars 2022)	**I. Remplacement** selon les besoins critiques en 2022-2024	**II. Ajout** de la capacité requise pour satisfaire complètement la demande projetée dans le concours 2024	**III. Ajout** de la capacité requise pour satisfaire partiellement la demande projetée dans le concours 2024	**IV. Ajout ponctuel** de capacité pour placer le Canada vers le milieu des pays du G7 au palmarès TOP500 (flops agrégés sur le PIB)
CPU (cœurs)	304 112	234 000	224 000 (total de 520 000 en 2024)	68 000 (satisfaction de 70 % de la demande)	Doublement de la puissance de calcul (gain grossièrement équivalent à 300 000 cœurs)
GPU (unités)	3 278	2 200	14 400 (total de 17 500 en 2024)	5 650 (satisfaction de 50 % de la demande)	Doublement de la puissance de calcul (gain grossièrement équivalent à 3 000 unités)
Stockage actif (To) (/project, dCache)	100 000	80 000	48 000 (selon une croissance annuelle de 20 %, pour un total de 148 000 en 2024)	48 000	48 000

/nearline (To)	88 000				
Réseautique	100 Gb/s (chaque site)		200 Gb/s[21, 22]		

Tableau 7 – Résumé des scénarios de remplacement et d'expansion de la capacité de CHP

Le tableau 7 ci-dessus présente l'état actuel des choses ainsi que l'estimation des besoins selon le type de matériel (première colonne) en fonction des différents scénarios projetés (première rangée). Les besoins de calcul par CPU (processeurs, serveurs, mémoire, connexions réseau haute vitesse internes, stockage /scratch) sont mesurés en cœurs individuels; ceux de calcul par GPU sont mesurés en unités (processeurs graphiques) individuelles. Le stockage actif comprend les catégories /project et dCache, mais pas la capacité de type /scratch ou /nearline. Le stockage /nearline à des fins de dépôt et d'archivage, lequel se fait sur disque et sur bande magnétique, fait l'objet de sa propre rangée. Le dernier composant architectural considéré est celui des connexions réseautiques externes nécessaires pour le raccordement aux réseaux nationaux et internationaux.

La deuxième colonne donne, à titre informatif, la capacité de calcul agrégée pour les principaux systèmes de la Fédération de l'Alliance (anciennement la Fédération Calcul Canada) en date de mars 2022. Ces données sont fondées sur le tableau 1, vu plus tôt, qui présente le détail des ressources de calcul par système, y compris le tout dernier à s'ajouter, Narval de Calcul Québec.

La troisième colonne présente les projections pour le scénario I : le remplacement à court et à moyen terme des équipements actuels à la première colonne selon les besoins les plus critiques. On parle ici du remplacement de toute l'infrastructure qui aura dépassé ses cinq ans de vie utile en date de décembre 2024. Au vu des dates de mise en service présentées dans le tableau 1, cela signifie de renouveler la capacité de tous les systèmes sauf Narval dans les deux années et demie à venir. Dans ce scénario visant uniquement les besoins critiques pour le maintien des capacités de base, aucune nouvelle capacité de calcul ne serait ajoutée en tant que telle, outre les gains d'efficacité inhérents qui viennent avec les avancées technologiques.

La quatrième colonne présente les projections pour le scénario II : l'ajout de la nouvelle capacité nécessaire pour satisfaire à 100 % les besoins attendus dans le cadre du concours 2024 pour l'allocation de ressources. Lesdits besoins attendus sont calculés à l'aide des estimations que

[21] Selon les requis prévus pour le LHC et Belle-II, il faudra atteindre 100 Gb/s d'ici 2025, et 200 Gb/s d'ici 2027. Le tout **en supplément** de la réseautique actuelle du site. Une capacité supplémentaire dépassant la barre des 100 Gb/s est déjà nécessaire pour Cedar (connexion au centre de données de premier niveau).

[22] Toute capacité réseautique supplémentaire doit être négociée auprès du RNRE concerné et de CANARIE.

l'on verra à l'annexe A; la demande en CPU est rapportée dans la figure A.4, et la demande en GPU dans la figure A.5. À noter que les besoins infrastructurels dans le cadre du scénario II s'ajoutent à ceux du scénario I, qui ne visaient qu'à maintenir les capacités de base actuelles.

La cinquième colonne présente les projections pour le scénario III : l'ajout de la nouvelle capacité nécessaire pour satisfaire *partiellement* les besoins attendus dans le cadre du concours 2024 pour l'allocation de ressources. Cela signifie, du côté de la demande en CPU, de subvenir à 70 % des besoins totaux pour le concours 2024 (à titre comparatif, les besoins pour le concours 2022 avaient été satisfaits à 54 %; voir la figure A.4), et du côté de la demande en GPU, de subvenir à 50 % des besoins (à titre comparatif, les besoins pour le concours 2022 avaient été satisfaits à 24 %; voir la figure A.5). Le scénario III est une version plus modeste du scénario II, et une fois encore, ces besoins infrastructurels s'ajoutent aux besoins critiques du scénario I. Fait intéressant, le besoin en nouveaux CPU dans le scénario III sera « uniquement » d'environ 68 000 cœurs grâce à l'arrivée de Narval, qui fait passer à lui seul le taux de satisfaction de la demande pour 2022 de 54 % à 68 % environ (chiffres non inclus dans la figure A.4). À savoir qu'il ne manque qu'une hausse assez modeste du nombre de cœurs (CPU) dans les deux prochaines années pour poursuivre sur cette lancée et atteindre les 70 % de satisfaction des besoins.

La sixième et dernière colonne présente les projections pour le scénario IV : l'ajout de la nouvelle capacité de CIP et de CHP nécessaire pour faire passer le Canada de la fin au milieu du peloton des pays du G7 pour ce qui est des flops agrégés sur le PIB (voir le détail dans le tableau 4). En pratique, cela nécessiterait au strict minimum d'au moins doubler la capacité de l'infrastructure de CIP (par rapport à 2022). Par exemple, du côté des CPU, il faudrait effectuer des investissements comme priorité immédiate afin de passer d'environ 300 00 à 600 000 cœurs. Tout comme pour les scénarios II et III, les besoins infrastructurels à combler dans le scénario IV viennent d'ajouter aux besoins critiques du scénario I.

4.2.1 Estimation des coûts pour chaque scénario

	Capacité totale actuelle (mars 2022)	**I. Remplacement** selon les besoins **critiques**	**II. Ajout** de la capacité requise pour satisfaire complètement la demande projetée dans le concours 2024	**III. Ajout** de la capacité requise pour satisfaire partiellement la demande projetée dans le concours 2024	**IV. Ajout ponctuel** de capacité pour placer le Canada au milieu des pays du G7 au palmarès TOP500 (flops agrégés sur le PIB)
Coûts d'immobilisations (CPU)	-	69 M$	65 M$	20 M$	87 M$
Coûts d'immobilisations	-	46 M$	303 M$	119 M$	63 M$

(GPU)					
Coûts d'immo-bilisations (stockage)	-	15 M$	9 M$	9 M$	9 M$
Coûts d'immo-bilisations totaux[23]	**-**	**130 M$**	**378 M$**	**148 M$**	**159 M$**
Pétaflops (R_{peak})	40	58	298	115	82
Espace (n^{bre} de bâtis)		162	500	192	220
Énergie (MW)	5,0	5,0	14,3	5,4	6,0
Services publics	5,0 M$	5,0 M$	14,3 M$	5,4 M$	6,0 M$
Coûts d'exploitation par année	29 M$	29 M$	7,2 M$	7,2 M$	7,2 M$
Coûts d'exploitation totaux[24]	**34 M$**	**34 M$**	**21,5 M$**	**12,6 M$**	**13,2 M$**

Tableau 8 – Estimation des coûts pour les scénarios du tableau 7

Le tableau 8 résume dans leurs grandes lignes les coûts d'acquisition du matériel dans les quatre scénarios figurant dans le tableau 7. Les coûts d'immobilisations pour les CPU et GPU sont estimés à partir d'un coût normalisé par nœud qui prévoit une portion allant au processeur central, à la mémoire, au réseau et à l'espace de stockage de type /scratch, le tout multiplié par le nombre de cœurs (CPU) et de GPU. Il faut toutefois savoir que le coût des technologies est hautement

[23] Annexe E – Voir le coût pour les configurations de référence.

[24] Frais d'exploitation totaux de la Fédération d'après le budget 2022-2023 (personnel compris).

changeant; les chiffres présentés sont donc offerts plus à titre informatif que prescriptif. On se donnera donc comme principe directeur de maximiser la capacité pour les chercheuses et chercheurs en fonction du budget prédéterminé au moment de la demande de propositions. La méthodologie et les prix appliqués pour en arriver à ces estimations de coûts seront expliqués à l'annexe E.

Les coûts d'exploitation sont estimés à partir du budget de fonctionnement 2022-2023 pour les cinq sites d'hébergement et leurs quelque 200 membres de personnel, à savoir environ 34 M$. À partir de la ventilation des coûts d'environ 5 M$ en services publics pour alimenter les systèmes en place en environ 5 MW de puissance, il est possible d'extrapoler ces coûts pour les quatre scénarios. La même approche peut être appliquée pour en arriver aux coûts en dotation additionnelle pour chaque scénario. L'effectif actuel à l'échelle de la Fédération est de 114 aides à la recherche et de 69 membres du personnel technique. Les expansions envisagées ne nécessiteraient qu'une modeste hausse d'une à deux personnes de plus au soutien technique par système, vu la nature dupliquée du matériel. Cependant, si l'on augmente considérablement les ressources disponibles, le nombre de demandes de soutien risque de suivre la même tendance; par conséquent, il faudrait ajouter un nombre adéquat d'effectifs pour le soutien à la recherche. Ainsi, pour les expansions envisagées dans les scénarios II à IV, on estime à la lumière du budget de 2022 qu'il faudrait voir s'ajouter environ 10 personnes au soutien technique et 50 au soutien à la recherche. Moyennant une charge salariale de 120 000 $ par personne en moyenne, cela revient à 7,2 M$ en coûts supplémentaires chaque année.

Résumons le tout : les coûts de remplacement selon les besoins critiques pour tenir les systèmes à jour afin de maintenir la capacité de base existante d'ici la fin de 2024 seraient de 136 M$ en immobilisations, plus des coûts d'exploitation de 34 M$ par année. Pour les expansions envisagées dans les scénarios II à IV, les coûts additionnels en immobilisations seraient respectivement de 378 M$, 148 M$ ou 159 M$, auxquels s'ajouteraient des coûts d'exploitation de 21,5 M$, 12,6 M$ ou 13,2 M$ respectivement.

Par exemple, la facture totale estimée pour une modernisation complète qui doublerait la capacité actuelle des systèmes d'ici la fin de 2024 reviendrait à 295 M$ en coûts d'immobilisations et à 47,2 M$ par année en coûts d'exploitation.

4.2.3 Coûts de remplacement des ressources infonuagiques

	Remplacement selon les besoins **critiques** en 2022-2024
Coûts	5 M$

d'immobilisations (CPU)	
Coûts d'immobilisations (GPU)	2 M$
Coûts d'immobilisations (stockage)	2 M$
Coûts d'immobilisations totaux[25]	**9 M$**

Tableau 9 – Estimation des coûts pour le remplacement des ressources infonuagiques de la Fédération

Le tableau 9 résume dans leurs grandes lignes les coûts d'acquisition de l'équipement de remplacement pour l'infrastructure infonuagique de la Fédération (voir le tableau 2 à la section 2.1). Tout cet équipement infonuagique aura plus de 5 ans à la fin de décembre 2024 et devra absolument être remplacé. Les coûts d'immobilisations sont ici estimés par application de la même méthodologie que pour les systèmes de CHP plus tôt, à savoir en partant de coût par nœud normalisé qui inclut une portion pour les processeurs, la mémoire et le réseau et que l'on multiplie ensuite en fonction du nombre de cœurs (CPU) et de GPU. La méthodologie et les prix appliqués pour en arriver à ces estimations seront expliqués à l'annexe E. Pour ce qui est de la demande et de l'expansion des ressources infonuagiques, la question est laissée à la considération du groupe de travail sur l'infonuagique.

4.3 Architecture de stockage de CHP

Le stockage et la gestion de données influent sur tous les aspects de l'IRN. Puisqu'un groupe de travail sur la stratégie de l'architecture de données se penchera sur la question de façon plus poussée et globale, la présente section ne s'attarde qu'aux exigences de stockage pour le CHP, en tenant toutefois compte de leur intégration dans une stratégie plus globale.

[25] Annexe E – Voir le coût pour les configurations de référence.

4.3.1 Hiérarchisation du stockage

La demande de stockage dédié (actif et de proximité) est détaillée dans le scénario de capacité à la section 4.2. Cela dit, la question du stockage ne se limite absolument pas à la capacité totale; la performance, les permissions, les politiques, les sauvegardes et la sécurité des données sont autant d'autres facteurs à prendre en compte. Comme nous l'avons brièvement mentionné dans la section 3.3, les systèmes actuels de la Fédération utilisent une approche de hiérarchisation plutôt courante dans les systèmes de CHP. La hiérarchie est décrite ci-dessous, avec quelques ajouts pour la prochaine génération de systèmes.

- **/scratch** – La taille de ce stockage équivaut à environ 10 fois la mémoire du système, et sa performance doit être suffisante sur les plans de la bande passante et des IOPS, avec un système de fichiers s'exécutant en parallèle accessible par toutes les ressources de calcul. L'emploi de disques entièrement à semi-conducteurs (SSD ou NVMe) pour l'espace /scratch est de plus en plus populaire et abordable. Il est aussi important de s'assurer que les utilisatrices et utilisateurs suivent les politiques de purgeage et les quotas. L'espace /scratch est temporaire et n'est donc pas sauvegardé.

- **/home** – Les quotas pour ce type de stockage sont petits, si bien que la capacité n'est généralement pas un enjeu majeur. Néanmoins, le nombre de fichiers peut être très élevé. Les données sont sauvegardées, et il pourrait être pertinent de les préserver dans plusieurs sites malgré leur petite taille, en raison de leur importance.

- **/project** – Devant présenter un équilibre de performance et de capacité, ce stockage prend la forme d'un système de fichiers s'exécutant en parallèle. Toutefois, comme le coût par téraoctet est généralement le facteur décisif, il se compose principalement de disques mécaniques rotatifs. Puisqu'il sert de stockage actif principal dédié, des quotas et des politiques de gestion doivent être en place pour en prévenir les utilisations malveillantes et le stockage à trop long terme de données qui devraient être migrées (préférablement de manière automatique) vers un stockage de proximité /nearline. Ce stockage est distribué aux projets sur une base annuelle. Ses sauvegardes sont locales, fait dont la planification budgétaire de la capacité doit tenir compte. Actuellement conçu comme un espace unique de plusieurs pétaoctets, il ne pourrait probablement pas faire l'objet d'une sauvegarde complète dans d'autres sites.

- **/nearline** – Conçu dans une optique de préservation à long terme, ce stockage dédié vise une capacité optimale. À l'heure actuelle, les utilisatrices et utilisateurs doivent procéder manuellement à la migration des données vers ce stockage de proximité. Or, il serait préférable que le tout se fasse automatiquement, selon des politiques sur la durée de statisme des données. Cela permettrait de réduire considérablement les données laissées sur des disques rotatifs, mais n'obligerait pas les utilisatrices et utilisateurs à trier leurs données, ce qui pourrait mener à une plus lourde charge de stockage. La Fédération

utilise déjà plusieurs solutions de ce type : IBM HPSS, Spectra BlackPearl et une solution maison basée sur Robinhood.

4.3.2 Stockage spécialisé

Outre la hiérarchie de stockage décrite plus haut, certains éléments plus spécialisés du stockage devront être examinés ou remplacés dans la prochaine génération de systèmes.

- **Mémoire tampon pour les utilisations intensives** – Le système Niagara est actuellement doté d'un espace /scratch encore plus performant juste pour les opérations IOPS intensives. Utilisant des disques NVMe et la technologie NVMesh, ce stockage partagé offre une très haute performance. Cette approche était assez courante pour les grands systèmes de CHP puisque les coûts des disques à semi-conducteurs étaient astronomiques. Aujourd'hui, cet espace pourrait probablement être remplacé par une expansion du stockage /scratch répondant à la demande.

- **dCache** – Ce stockage est utilisé par ATLAS, T2K, SNO+, DEEP et quelques autres petits projets de physique des hautes énergies.

- **CVMFS** – Ce stockage est employé dans la fourniture de logiciels au sein de la Fédération et est mis à l'essai pour les jeux de données partagées courants.

- **Stockage d'objets infonuagique** – Cette technologie de stockage infonuagique est dotée d'interfaces normalisées non POSIX, comme Swift/S3. Elle pourrait servir de réceptacle commun aux jeux de données et possiblement aux données utilisateur partagées entre les systèmes.

4.3.3 Autres considérations pour la conception du stockage

Les systèmes de stockage déployés ayant une durée de vie supérieure aux cinq années précédemment mentionnées, ils peuvent servir plusieurs générations de systèmes de calcul. C'est d'autant plus vrai pour les systèmes axés sur la capacité et à bande magnétique. Contrairement aux systèmes de calcul, les systèmes de stockage, surtout ceux à grande capacité comme /project, sont parfois longs à remplacer, et le processus peut causer beaucoup d'inconvénients aux utilisatrices et utilisateurs. Leur conception et leur cycle de vie doivent donc être examinés de près. En construisant un système qui peut être élargi et mis à niveau, on permet de diminuer les coûts, d'éviter des migrations de données et d'augmenter la capacité au gré des besoins pour réduire l'achat initial.

4.3.4 Autres considérations pour la gestion des données

Les considérations globales ci-dessous ne sont que brièvement abordées. Elles devraient être explorées plus en détail dans le cadre d'une stratégie nationale de gestion des données.

- Le fait que le stockage des données (actif, de proximité et de sauvegarde) soit entièrement local est l'un des plus gros obstacles au transfert de tâches et de fonctions de calcul d'un système à l'autre. Il existe des outils pour faciliter les mouvements de données entre les sites, mais aucun n'est automatique ou à haute disponibilité. De plus, les utilisatrices et utilisateurs sont laissés à eux-mêmes pour le transfert ou la duplication de leurs données.
- Une grande quantité de données actuellement stockées sur des disques rotatifs locaux n'ont pas été utilisées depuis au moins six mois. En outre, il n'existe actuellement aucune façon normalisée de centraliser les données couramment utilisées, comme les grands jeux de données, ce qui peut entraîner une duplication importante même à l'échelle locale.
- Le CHP requiert un stockage local de haute performance, mais une stratégie coordonnée visant à assurer un certain stockage de projet (non local) à haute disponibilité demeurerait pertinente, entre autres pour augmenter la résilience de l'écosystème de CHP.
- La planification et l'allocation du stockage doivent se faire à long terme (sur plus d'un an). La GDR et le CIP devraient également être coordonnés. Des politiques, une planification et des investissements à long terme amélioreraient considérablement les options de stockage pour les chercheuses et chercheurs.
- Le stockage devrait être sécurisé, et son utilisation et ses accès, encadrés par des politiques de gouvernance.
- Les sauvegardes devraient être planifiées (à l'échelle locale et hors sites).

4.4 Considérations pour la prochaine génération de systèmes de CHP

La section 4.2 balise les exigences de capacité avec des termes généraux, par exemple les cœurs-années (CPU), les GPU-années et le stockage générique. Toutefois, les systèmes de CHP sont généralement conçus en fonction de tâches précises; ils peuvent aussi bien prendre la forme de machines aux classes de capacités hautement spécialisées (comme le système Frontier Exascale de l'ORNL) pour des tâches massivement parallèles que de systèmes au matériel plus modeste principalement utilisés pour calculer un large éventail de petites tâches mixtes.

Dans l'optique d'appuyer l'ensemble des flux de travail et des chercheuses et chercheurs, l'écosystème actuel du Canada se divise en trois parties : quatre grappes hétérogènes de calcul générique, qui répondent à la plupart des besoins de CHP à petite et à moyenne échelle et fournissent les ressources de GPU; le grand système homogène parallèle Niagara, conçu pour les tâches massivement parallèles à grande échelle; et le système infonuagique Arbutus.

4.4.1 Tendances actuelles des tâches de CHP

Pour orienter la conception des systèmes de CHP et déterminer l'équilibre approprié de ressources générales et spécialisées, on peut se pencher sur les caractéristiques des tâches confiées à l'infrastructure actuelle et en dégager des tendances. L'équipe nationale d'analyse de données conserve et analyse les données de l'ordonnanceur pour dresser le portrait des différentes tâches. En classant les charges de travail selon l'ampleur de l'utilisation des processeurs centraux et graphiques ainsi que selon le type de tâches confiées à chaque système de CHP, on peut dégager des tendances assez régulières. Les principales observations sont résumées ci-dessous; les données exactes sont fournies à l'annexe B.

Petites tâches

En examinant la taille des tâches confiées aux systèmes de calcul générique, on constate qu'environ 50 % de l'utilisation des processeurs centraux se concentre dans les tâches d'un nœud ou moins, un nœud équivalant à 32 à 48 cœurs. En outre, le nombre de cœurs de CPU par système continue d'augmenter, et l'afflux de petites tâches devrait donc se maintenir. Ainsi, le besoin de matrices/nuages de réseautique internes à haute vitesse pour les communications dans les systèmes de CHP risque de diminuer, mais pas nécessairement de disparaître, car ces technologies pourraient demeurer pertinentes en raison des besoins de stockage à haute vitesse. En effet, plus il y aura de tâches par nœud, plus la demande de stockage sera grande, d'où l'importance de réseaux solides. En outre, la moins grande proportion de nœuds majeurs vient considérablement diminuer le coût de la connexion à haute vitesse (lié au nombre de cartes réseau et de montages contacteurs) par rapport au coût du nœud.

Exigences de mémoire

Dans l'ensemble des systèmes, environ 85 % des tâches n'utilisaient que 4 Go ou moins de mémoire par cœur. Il existe certes une demande pour des cœurs offrant une mémoire supérieure, mais elle demeure relativement petite; les ressources en ce sens devraient donc être développées en conséquence. De plus, la tendance étant à augmenter le nombre de cœurs de CPU par nœud, si le seuil de 4 Go par cœur est maintenu, la mémoire totale par nœud augmentera naturellement. Dès lors, il existera des nœuds à plus grande capacité de mémoire pour les tâches plus lourdes. Néanmoins, l'augmentation de la mémoire totale par nœud avec la multiplication des cœurs devra tenir compte des éléments architecturaux nécessaires pour porter une largeur de bande de la mémoire accrue.

Charge des processeurs graphiques

Bien que la demande de ressources de GPU soit élevée, la charge demeure largement dominée par de petites tâches utilisant entre un et quatre GPU; plus de 50 % des tâches n'en utilisaient qu'un, et 95 % en utilisaient quatre ou moins (soit un seul nœud). Et avec l'amélioration rapide de la performance des GPU et le nombre encore limité de tâches et de codes bases capables de les exploiter efficacement (même dans les scénarios à un seul GPU), cette tendance devrait se

maintenir au moins à court terme. Qui plus est, puisque les données préliminaires indiquent que les GPU ne sont pas toujours utilisés au maximum de leur capacité, il y aurait lieu d'envisager de nouveaux systèmes permettant de les subdiviser, afin que plusieurs petites tâches puissent utiliser le même. Ce faisant, les utilisatrices finales et utilisateurs finaux disposeraient de ressources moindres, mais qu'ils pourraient exploiter de façon optimale, sans devoir déployer d'efforts majeurs de programmation. Les technologies comme le processeur graphique à instances multiples de NVIDIA, compatible avec ses GPU de dernières générations, seraient une option.

Tâches massivement parallèles

Contrairement aux systèmes de calcul générique, avec Niagara, c'est 60 % de l'utilisation qui passe par des tâches requérant 512 cœurs ou plus. Ce n'est là rien de surprenant puisque ce système a été conçu pour les tâches massivement parallèles, mais cela confirme tout de même la nécessité d'offrir au moins un tel système. Ce besoin est aussi nommé dans des exposés de position de domaines comme l'astrophysique et la dynamique des fluides numériques, qui dépendent encore largement de superordinateurs massivement parallèles homogènes alimentés par des CPU.

> « Nous recommandons un renouvellement et un développement ambitieux de la capacité de simulation massivement parallèle de l'écosystème canadien de l'IRN, afin de tirer parti des progrès en cours, de même qu'une expansion des ressources humaines spécialisées qui y sont associées et sont essentielles aux découvertes scientifiques[26]. »

La communauté de recherche a de toute évidence besoin d'une capacité de CPU homogène de prochaine génération. Cela dit, il faudrait aussi qu'un futur système massivement parallèle comprenne une portion dotée d'accélérateurs. Cette approche hybride se voit déjà dans des configurations comme le nouveau système LUMI d'EuroHPC visant la compatibilité avec un plus grand nombre de codes bases existants et émergents[27].

Concours pour l'allocation de ressources

Le concours distribue généralement 80 % des ressources disponibles dans une année. À l'exception du système Niagara, seuls 60 % de ces ressources sont utilisés par les comptes du concours. Ce n'est pas qu'elles sont gaspillées, mais plutôt qu'elles sont principalement

[26] R. Fernandez et coll. *Large-Parallel Supercomputer Simulations – Frontiers in Canadian Research*, exposé de position soumis à l'Alliance. https://alliancecan.ca/sites/default/files/2022-03/ndrio_wp_on_niagara_scale_systems.pdf (décembre 2020).

[27] LUMI. *LUMI's full system architecture revealed*. https://www.lumi-supercomputer.eu/lumis-full-system-architecture-revealed/ (consulté en mai 2022).

exploitées par des utilisatrices et utilisateurs par défaut qui n'ont pas d'allocations dédiées. Il y aurait donc une potentielle demande de ressources flexibles à petite échelle qui devrait orienter les prochaines politiques et stratégies d'allocation.

4.4.2 Équilibre des accélérateurs et des processeurs centraux

L'utilisation et la demande d'accélérateurs (qui sont le plus souvent des GPU) dans les systèmes de CHP augmentent, car la performance et l'efficacité des accélérateurs en flops/$ ou en flops/poids avancent plus vite que celles des CPU traditionnels. Les GPU sont en effet hautement efficaces pour certaines tâches; les nouveaux concepts et environnements de programmation optent d'ailleurs pour des systèmes compatibles avec un plus grand nombre d'utilisations. Cependant, ces processeurs peuvent aussi être utilisés de façon très peu optimale; ils ne devraient ni ne pourraient donc remplacer complètement les CPU. En effet, de nombreux problèmes et usages ne peuvent pas être aisément transposés ou transférés vers des GPU (voire ne peuvent pas l'être du tout). C'est pourquoi le système Frontera du TACC, financé par la NSF à des fins de recherche universitaire générale, repose principalement sur une base pure de processeurs centraux, avec des processeurs graphiques en complément.

On notera par ailleurs que les trois instituts canadiens de l'IA collaboreront avec l'Alliance pour déployer l'environnement informatique pancanadien de l'IA, qui comprendra trois nouveaux systèmes spécialisés conçus expressément pour les tâches d'IA, donc probablement axés sur les GPU. Ces systèmes pourront aussi être utilisés, dans une certaine mesure, pour des usages de recherche plus généraux, ce qui ne peut qu'alléger la demande croissante de ressources de calcul par GPU.

4.4.3 Configuration de la prochaine génération de systèmes

L'utilisation actuelle confirme un besoin continu de systèmes de calcul générique pour les petites tâches de traitement par processeurs centraux ou graphiques, ainsi que d'au moins un grand système parallèle adaptable. En outre, bien que la demande de GPU continue de croître, les systèmes les plus récents (Béluga et Narval) offrent d'importantes ressources graphiques. Ainsi, il suffirait d'en mesurer l'usage pour déterminer l'équilibre idéal de CPU et de GPU. Le ratio actuel d'utilisation des ressources de calcul générique et massivement parallèles semble pour sa part convenir à la demande. Avec les multiples déploiements de systèmes de calcul générique, il serait possible de planifier les remplacements et les améliorations de sorte que les dernières technologies soient toujours disponibles dans la plateforme nationale, tout en maintenant – et préférablement en développant – la capacité. Pour ce qui est du grand système parallèle, il devrait être suffisamment important pour durer plusieurs années, puisque l'écosystème ne verra un nouveau système du genre que tous les quatre à cinq ans.

4.5 Réseautique

La réseautique du CHP se divise généralement en deux parties : un réseau interne à haute vitesse et faible latence utilisé pour les communications entre les tâches et les fichiers entrants et sortants, et un réseau externe utilisé pour le mouvement des données et la connectivité Internet au site.

4.5.1 Réseau à haute vitesse

Le réseau interne à haute performance (ex. : InfiniBand) prend souvent la forme d'une matrice non Ethernet à faible latence offrant un large éventail de topologies de réseaux; c'est généralement la principale distinction entre un système de CHP et un système d'entreprise ou d'infonuagique classique.

Au départ, la réseautique à haute vitesse servait principalement à la communication entre les nœuds dans les tâches parallèles. Aujourd'hui, puisque les nœuds simples comptent beaucoup plus de cœurs qu'avant, elle sert aussi à améliorer la performance des entrées et sorties de fichiers. La tendance à l'élargissement des nœuds a aussi permis de réduire le coût de la réseautique proportionnellement au coût total du système de CHP, en réduisant le nombre de ports réseau requis. Ainsi, même les systèmes de calcul générique qui traitent des petites tâches devraient être déployés avec une forme ou une autre de réseautique à haute vitesse. L'annexe C explore ce sujet plus en détail.

4.5.2 Réseau externe

Les réseaux étendus externes à protocole IP ont généralement une bande passante suffisamment élevée pour traiter les mouvements de données. Tous les sites d'hébergement actuels sont reliés par une connexion de 100 Gb/s, mais la capacité de la connectivité commerciale (le lien Internet) varie d'un site à l'autre. Les discussions sur la réseautique externe ont aussi des répercussions directes sur la planification de la cybersécurité, car ce sont avec ces connexions que les utilisatrices et utilisateurs se connectent aux systèmes pour travailler à distance.

- CANARIE a besoin d'une connexion de 200 Gb/s pour les sites qui hébergent des tâches à forte intensité de données (Atlas).
- Il faudrait prévoir une mise à niveau à 200 Gb/s pour les sites qui instaureront le plan de gestion des données et de résilience.
- La mise à niveau des réseaux étendus devrait être coordonnée avec le RNRE et CANARIE.

Voir l'annexe C pour des explications plus détaillées des considérations de réseautique externe.

4.6 Acquisition et déploiement de matériel de prochaine génération

4.6.1 Planification pluriannuelle des investissements en immobilisations et en exploitation

L'infrastructure de CHP moderne vient avec des coûts d'immobilisations et d'exploitation considérables. De plus, les ressources humaines nécessaires devraient être assurées avant tout investissement dans l'infrastructure. Le financement des immobilisations de l'IRN a souvent été sporadique, tous les quatre ou cinq ans pour répondre à un besoin immédiat, si bien que les achats majeurs sont espacés de plusieurs années. En outre, aucun financement soutenu n'a été consenti par les organismes de financement, de sorte qu'aucune planification systémique et plurigénérationnelle à long terme n'est possible. Les systèmes ont donc été construits en réponse à des besoins ponctuels, sans s'inscrire dans un plan d'écosystème durable tenant compte du cycle de vie et des calendriers d'actualisation. Les coûts d'exploitation ne faisaient pas non plus partie de l'appel de financement et étaient plutôt couverts par un programme distinct. Résultat : l'acquisition de nouveaux systèmes et d'infrastructures de CHP s'est faite sans égard aux coûts et aux enjeux d'exploitation correspondants, malgré le lien fort entre ces éléments.

L'échelonnement des dates de mise en service et de mise hors service des systèmes de CHP, bien qu'avantageux côté continuité des services et améliorations techniques, peut causer des problèmes dans l'harmonisation des budgets d'immobilisations et d'exploitation. Le gouvernement du Canada et ISDE reconnaissent ce problème et ont donc donné comme mandat à l'Alliance de gérer l'enveloppe d'exploitation des systèmes de CHP (anciennement gérée par le Fonds des initiatives scientifiques majeures de la FCI), afin que toutes les décisions majeures de financement des immobilisations et de l'exploitation de l'IRN soient prises dans une optique de globalité. La planification du financement d'exploitation devra d'ailleurs prendre en compte les coûts liés aux systèmes contributifs.

Le processus de planification et d'acquisition devrait être révisé régulièrement afin de confirmer que les investissements répondent aux besoins de la communauté de recherche, de cerner les éventuelles lacunes et d'adapter le plan aux changements technologiques fréquents anticipés dans le domaine du CHP.

Une telle approche d'investissement continu pour les multiples sites assurait la continuité des services dans l'écosystème entier tout en permettant d'actualiser les ressources et d'y ajouter de nouvelles technologies et capacités.

4.6.2 Considérations pour le déploiement et le cycle de vie du matériel de CHP

4.6.2.1 Déploiement unique à grande échelle

Par le passé, les nouveaux systèmes de CHP étaient généralement achetés tous en même temps dans un déploiement unique à grande échelle, en raison du modèle de financement des immobilisations. Cette façon de faire a ses avantages, notamment d'offrir un grand système homogène aux ressources initiales supérieures et d'obtenir des prix d'achat en gros auprès des fournisseurs. Par conséquent, cette méthode demeurait pertinente pour les grappes de grandes charges parallèles. Or, elle comporte aussi un risque, à savoir que le système restera sensiblement le même sur tout son cycle de vie (généralement cinq ans) même si la conception initiale ne répond pas bien aux besoins des chercheuses et chercheurs ou s'adapte mal aux changements et aux améliorations technologiques qui surviennent. Enfin, les systèmes pour lesquels aucune amélioration n'avait été prévue initialement pourraient être plus difficiles à mettre à niveau.

4.6.2.2 Déploiement par étapes

Plutôt que de tout déployer d'un coup, on peut aussi acquérir un système par étapes, selon un plan d'expansion ou d'amélioration continue. Un système ainsi déployé serait conçu de façon à ce que sa puissance et son stockage puissent être développés au fil du temps, avec des composantes essentielles comme la réseautique centrale adaptables à la croissance future. Il pourrait alors mieux répondre aux besoins des chercheuses et chercheurs et accueillir de nouvelles technologies plus rapidement, les étapes étant généralement assez rapprochées. Cette méthode a fait ses preuves dans des systèmes comme le superordinateur Pleiades de la NASA. Toutefois, elle n'est pas sans défaut, car elle peut produire des systèmes hétérogènes composés de plusieurs générations d'appareils, ce qui peut compliquer la coordination et l'exploitation et réduire la convivialité pour les utilisatrices finales et utilisateurs finaux. Il s'agit déjà d'un enjeu avec les systèmes de calcul générique, particulièrement avec les composants contributifs, mais le problème pourrait être partiellement atténué par la planification d'étapes d'une taille respectable pour limiter l'hétérogénéité. Autre problème potentiel : il se pourrait que les dernières étapes soient restreintes par la conception physique initiale du système et de son centre de données, par exemple un budget d'électricité ou de refroidissement maximal contraignant.

4.6.2.3 Déploiements échelonnés

Dans le contexte d'une plateforme nationale comprenant plusieurs sites et systèmes, le scénario optimal serait probablement un équilibre des deux approches : des déploiements de systèmes complets échelonnés dans le temps et entre les sites. Ainsi, la plateforme serait continuellement actualisée en fonction des nouvelles technologies et des besoins de la recherche sans causer d'hétérogénéité majeure qui serait difficile à gérer et complexifierait l'utilisation des systèmes. D'un point de vue financier, cette approche est aussi avantageuse puisque les dépenses annuelles en immobilisations seraient plus régulières, sans compter que la répartition des dépenses dans le temps pourrait faciliter l'obtention de financement de contrepartie.

4.6.2.4 Cycle de vie

Peu importe le scénario choisi, la conception devrait tenir compte du cycle de vie complet de chaque composant du système. La durée de vie typique d'un composant de système de CHP est de quatre à cinq ans, après quoi les avantages des nouvelles technologies supplantent les avantages d'un système établi. Cela dit, le matériel ne perd pas toute sa valeur; de nombreuses fonctions à petite échelle, d'informatique en grille ou même d'infonuagique ne sont pas limitées par la performance et peuvent donc s'exécuter sur des appareils légèrement plus vieux et moins puissants. En gardant le cycle de vie des systèmes en tête, on peut exploiter les mêmes ressources plus longtemps, pour réduire les répercussions environnementales et économiser sur les dépenses en immobilisations.

4.6.3 Conception flexible

Compte tenu des besoins variés de la communauté de recherche, présentés dans l'analyse de la charge de travail, et des multiples acquisitions et déploiements de technologies en évolution rapide, il est crucial que la conception des systèmes soit flexible. La cohérence dans les interfaces et les fonctions est importante pour les utilisatrices et utilisateurs, mais elle ne devrait jamais être si rigide qu'elle empêche l'intégration d'une solution technologique optimale ou tente d'appliquer une approche unique à tous les flux de travail. L'un des avantages d'un écosystème national de CHP, c'est que les composants de la plateforme peuvent différer en fonction des spécialisations requises tout en assurant un accès à l'ensemble de la communauté de recherche du pays.

4.7 Considérations en lien avec le centre de données et l'environnement

Les centres de données de CHP sont généralement plus gourmands en électricité, en systèmes de refroidissement et en superficie que les centres de données de TI des entreprises. C'est pourquoi tous les systèmes actuels de la Fédération se trouvent dans des centres de données de CHP spécialisés, dans les sites d'hébergement. Les sites ont tous des capacités d'expansion différentes, indiquées à l'annexe D. L'espace physique n'est cependant pas le facteur limitant; ce sont l'électricité et le refroidissement qui déterminent la puissance de calcul possible.

4.7.1 Soutien aux centres de données pour les systèmes nouveaux et élargis

Selon la date de mise hors service des systèmes et la taille des nouveaux systèmes, par exemple ceux des scénarios d'expansion aux tableaux 7 et 8, il risque d'être nécessaire de modifier les centres de données, voire d'aménager de nouveaux sites d'hébergement. Même sans expansion, les nouvelles technologies de CHP, comme les GPU en développement et les CPU à cœurs nombreux, demandent un apport en ressources intensif pouvant dépasser les capacités

d'alimentation électrique et de refroidissement des centres existants, imposant ainsi des modifications ou une expansion.

Les coûts engendrés par ces facteurs n'ont jusqu'ici pas été admissibles au financement fédéral, ce qui peut créer des inefficacités à l'échelle nationale et des inégalités entre les établissements d'hébergement, qui assument des coûts importants auxquels la majorité des autres établissements ne contribuent pas. Le rapport *État actuel du calcul informatique de pointe au Canada* se penche plus amplement sur le financement de la construction, de la maintenance et de l'exploitation des centres de données de CHP, aux pages 121 et 122.

4.7.2 Impact environnemental

Les modifications et les expansions de centres de données sont de bonnes occasions de moderniser les installations de CHP et d'investir dans une infrastructure plus efficace et écologique. Des options comme le refroidissement direct à l'eau chaude, qui conviendraient à la charge thermique plus élevée des nouveaux serveurs de CHP, pourraient aussi réduire considérablement les coûts d'exploitation et diminuer l'empreinte carbone. Selon l'emplacement et la configuration des centres de données, il y aurait même une possibilité de réutiliser le surplus de chaleur, ce qui viendrait diminuer de façon importante l'empreinte carbone des centres.

Quel que soit le cas, les acquisitions futures devraient prioriser les conceptions écoénergétiques et les pratiques exemplaires. Le groupe de travail pour le CHP écoénergétique a d'ailleurs publié un rapport sur le sujet, *Energy Efficient Considerations for HPC Procurement Document* (2021)[28], qui fait état des exigences pour la mesure fiable et exacte de l'efficacité énergétique des fonctionnalités et capacités des systèmes (données de référence, conception et mesure de l'alimentation électrique et du refroidissement, interface avec les installations).

4.7.3 Considérations pour la redondance, la résilience et le taux de disponibilité

Les centres de données de CHP ont souvent des points de défaillance unique connus; ils choisissent de dépenser moins sur l'infrastructure physique (ex. : aucune redondance pour le refroidissement ni système complet d'alimentation sans coupure [ASC]) et plus sur la capacité de l'infrastructure de CHP essentielle. Les coûts d'une pleine redondance pour l'alimentation, le refroidissement, le stockage et la réseautique à l'échelle des systèmes de CHP seraient immensément élevés, ce qui réduirait de beaucoup la quantité de ressources offertes aux chercheuses et chercheurs.

Ainsi, en pratique, la résilience pourrait être l'option plus pertinente à considérer. On pensera aux conceptions qui assurent un fonctionnement minimal de l'infrastructure et des éléments critiques comme le stockage en cas de panne, au moyen d'un système d'ASC et d'un générateur, une

[28] Lawrence Livermore National Labs' (LLNL) Energy Efficient HPC Working Group. *Energy Efficiency Considerations for HPC Procurement Document: 2021*. https://drive.google.com/file/d/1aB7uv47anaHUcHzw140tJLUSe9xVkQfQ/view (octobre 2021).

option bien moins coûteuse que la pleine redondance. La redondance de la réseautique peut aussi prendre la forme d'un lien à plus faible capacité pour maintenir une connectivité même lorsque les gros transferts de données doivent être suspendus.

On notera aussi que, même si les systèmes ne sont pas vraiment individuellement redondants à l'échelle locale, le fait qu'ils soient répartis dans plusieurs centres de données dans différentes régions du pays est aussi une forme de résilience. À l'heure actuelle, les systèmes fonctionnent presque indépendamment les uns des autres, à l'exception de quelques services centraux assurés par une infrastructure à haute disponibilité. La capacité totale demeure toutefois problématique, car les systèmes actuels ont atteint leur capacité maximale et n'ont donc plus les moyens d'absorber une hausse de la demande. Le fait que les données sont localisées et non redondantes oppose un autre obstacle majeur à la redondance entre les systèmes; toutes les données sont locales à leur système et ne sont reproduites dans aucun autre site, les copies de sauvegarde étant aussi stockées localement. Seul le système Niagara fait exception puisque ses copies de sauvegarde et son stockage de proximité sont dupliqués au Center for Advanced Computing (CAC).

Les systèmes futurs devraient au minimum stocker des copies des données essentielles dans d'autres sites à des fins de redondance partielle. L'accès à un stockage délocalisé par grappes (ex. : un stockage d'objets délocalisés ou une autre forme de stockage à haute disponibilité) contribuerait aussi de beaucoup à la résilience du service et réduirait considérablement le risque en cas de défaillance d'un site.

Enfin, il y aurait lieu d'envisager des options comme un soutien au fonctionnement disponible 24 heures sur 24 (actuellement limité aux heures de bureau et aux autres moments lorsque possible) dans l'établissement des objectifs de niveau de service (ONS) pour chaque site d'hébergement.

4.8 Technologies futures et environnements de test

Les architectures de CHP évoluent constamment en fonction des dernières technologies pour devenir plus performantes. C'est pourquoi il est important de continuellement repenser l'offre et de fournir au personnel et aux chercheuses et chercheurs des ressources pour explorer les technologies émergentes et différentes avant leur déploiement à grande échelle. Il en va de même pour la mise à l'essai de fonctionnalités et de services potentiels, comme des logiciels ou des options de stockage. Ainsi, l'Alliance devrait envisager le financement de petits systèmes de test et de développement, par exemple les suivants :

- Un carrefour centralisé pour l'évaluation et la mesure de la performance du matériel
- Des systèmes de test et de développement pour les tests bêta avant le déploiement en production

- Des ressources d'intégration et de déploiement continus en soutien au développement de logiciels de recherche et de plateformes de CHP
- Des petits environnements de test flexibles pour étudier les architectures de CHP émergentes
    - Orchestration (Kubernetes, Fuzzball, etc.)
    - Infrastructures composables

4.9 Soutien, efficacité et convivialité

4.9.1 Soutien et formation

Le financement de ressources élargies est un morceau important de la réponse aux besoins des chercheuses et chercheurs, mais il est tout aussi crucial de financer la dotation en personnel pour voir au fonctionnement de ces ressources et au soutien des utilisatrices et utilisateurs. Le personnel de soutien à la recherche joue un rôle essentiel dans la maintenance et l'exploitation des systèmes. Tout comme les ressources n'arrivent pas à fournir à la demande, l'effectif de la Fédération est trop petit pour encadrer correctement la base croissante d'utilisatrices et utilisateurs. La Fédération emploie actuellement environ 200 spécialistes de l'IRN de l'Alliance équivalents temps plein (ETP), soit à peu près 6 ETP par établissement contributeur ou 1 ETP pour 80 utilisatrices et utilisateurs enregistrés. Comparativement, le Texas Advanced Computing Center (TACC), qui héberge le système Frontera, compte environ 190 spécialistes, lesquels sont répartis entre les milliers d'utilisatrices et utilisateurs selon un ratio deux à cinq fois plus favorable, entre 1:16 et 1:55.

Une hausse du personnel permettrait de former davantage les utilisatrices et utilisateurs et de fournir un soutien propre à chaque discipline. Avec le nombre croissant de personnes travaillant avec le CHP et le roulement rapide aux cycles supérieurs, il est essentiel d'offrir des formations régulières pour assurer l'utilisation efficace des ressources et réduire le nombre de problèmes causés par des erreurs ou une mauvaise utilisation. En embauchant du personnel pour accompagner de près les utilisatrices et utilisateurs dans des tâches comme l'analyse des flux de travail et des programmes, on irait chercher d'importants gains d'efficacité, qui augmenteraient à leur tour la productivité de la recherche et la bonne utilisation des ressources de CHP. Ce point est d'autant plus vrai pour les flux de travail qui sont difficiles à exécuter de façon parallèle, qui font une utilisation inefficace des GPU ou qui suivent des modèles d'accès aux fichiers entrants et sortants sous-optimaux.

4.9.2 Efficacité des systèmes et du travail

Les systèmes de prochaine génération devraient accorder une importance particulière aux outils et aux fonctionnalités axés sur l'efficacité afin que le travail effectué tire pleinement parti des ressources offertes. L'Alliance devrait donc encourager des initiatives de développement d'outils d'instrumentation, d'automatisation et de récolte de données sur l'utilisation pour les utilisatrices

et utilisateurs et le personnel de soutien. Ces outils pourraient aussi servir à surveiller et à mesurer la performance des systèmes, dans une optique d'utilisation efficace.

4.9.3 Expérience utilisateur

Comme la communauté de recherche utilisant le CHP s'élargit, la demande de systèmes compatibles avec de nouvelles interfaces et de nouveaux flux de travail augmente. Afin de rendre ses services plus conviviaux, l'Alliance devrait s'intéresser à des solutions plus diversifiées, comme les infrastructures de bureau virtuel ou les flux de travail Web. Les ressources consacrées à des services interactifs comme les calepins Jupyter Notebook ou des portails Web devraient également être élargies.

Annexe A : Données d'offre et demande du concours pour l'allocation de ressources

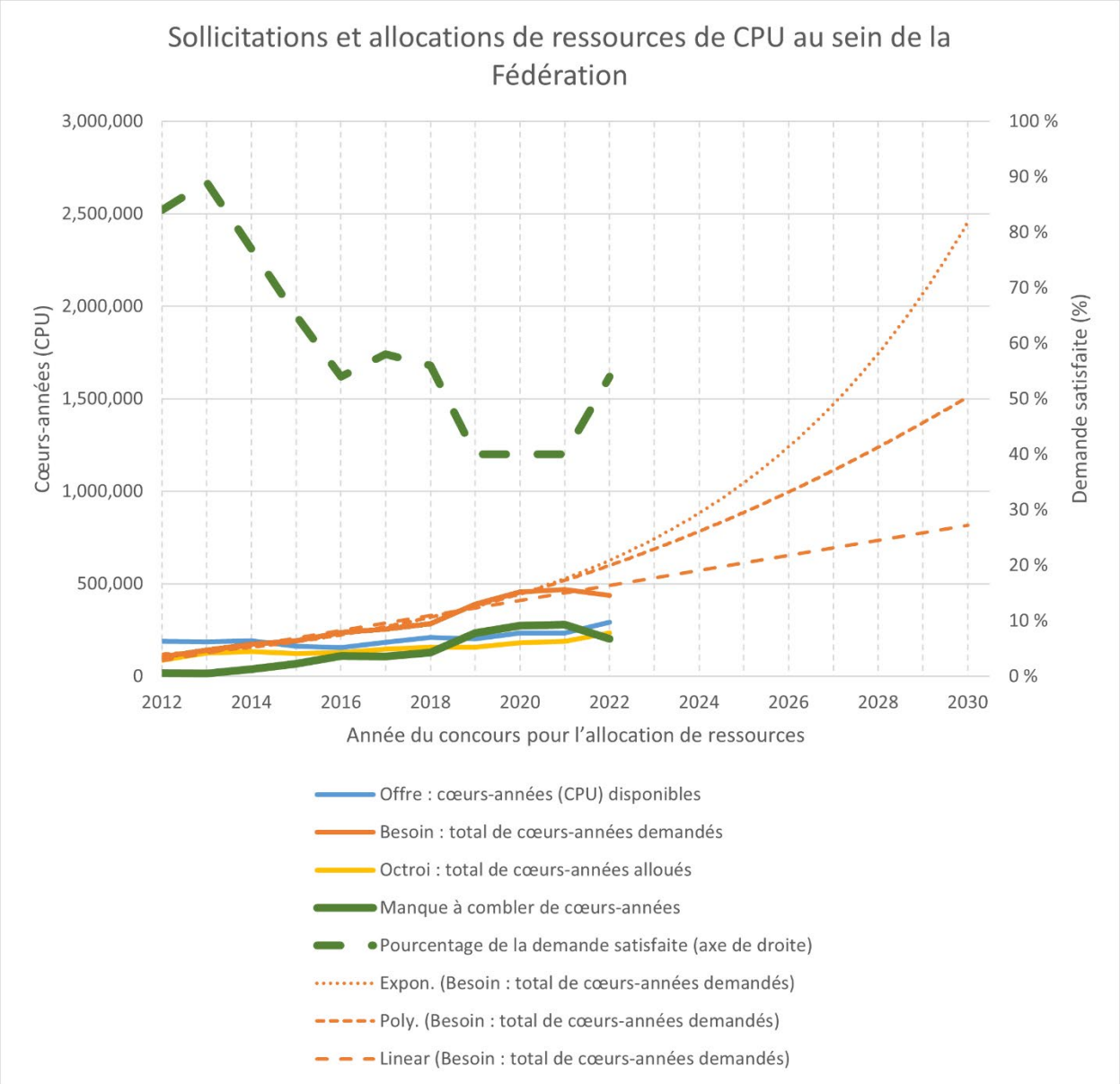

Figure A.1 – Demande de ressources de CPU et tendances projetées

La figure A.1 ci-dessus compare l'offre et la demande des ressources de calcul par CPU dans le consortium de la Fédération Calcul Canada de 2012 à 2022. L'offre équivaut à la capacité de CPU réelle disponible dans les systèmes de la Fédération, et la demande, aux sollicitations de ressources reçues dans le cadre du concours annuel pour l'allocation de ressources de la Fédération. L'axe horizontal représente les années d'allocation, et l'axe vertical, les cœurs-années (CPU).

La ligne bleue représente la capacité de calcul annuelle brute totale disponible dans les principaux systèmes de la Fédération. Cette capacité fluctuait initialement dans une fourchette assez étroite, avant d'augmenter plus récemment pour atteindre 293 000 cœurs-années au concours de 2022, soit près du double de l'offre d'il y a 10 ans. Notons que la mesure des cœurs-années ne tient pas compte de la puissance de calcul réelle de chaque cycle, c'est-à-dire l'augmentation de la capacité de calcul par processeurs grâce au développement de l'architecture.

La ligne orange pleine illustre le total de capacité sollicitée dans le cadre du concours. Dans les 10 dernières années, la demande est passée d'environ 100 000 cœurs-années à un sommet de 470 000 en 2021, avant de redescendre aux alentours de 440 000 en 2022. Les raisons du ralentissement et de la baisse de la demande ces deux dernières années ne sont pas entièrement claires, mais on peut supposer qu'il s'agit principalement d'un mélange de fluctuations naturelles, et de perturbations (temporaires) des activités et priorités de recherche universitaire dues à la COVID-19. Le groupe de travail anticipe que la croissance de la demande reprendra rapidement avec le retour à la normale de la communauté de recherche. Cette conclusion est appuyée par les projections internationales des besoins de CHP, qui augmenteront considérablement dans un avenir proche. La croissance totale de la demande a été fulgurante et semi-linéaire, mais ne semble pas exponentielle. Le taux de croissance annuel composé (TCAC) de la demande de cycles de calcul par CPU entre 2012 et 2020 s'élève à environ 21 % par an.

Les lignes orange minces projettent trois tendances de la demande. Les projections utilisent les données des concours de 2012 à 2020; elles excluent les données de 2021 et de 2022 en raison de la baisse anormale et temporaire. Les trois lignes correspondent à trois hypothèses fonctionnelles de croissance future : la ligne pointillée montre une croissance exponentielle; la ligne de petits tirets, une croissance polynomiale de deuxième degré; et la ligne de gros tirets, une croissance linéaire. Ensemble, ces projections illustrent la fourchette de croissance anticipée de la demande. Au concours de 2030, la demande se situerait donc entre 800 000 et 2 400 000 cœurs-années.

La ligne jaune indique la capacité réelle allouée par le concours. Une partie de la capacité totale (délibérément plafonnée à environ 80 %) est accordée aux utilisatrices finales et utilisateurs finaux dans le cadre du concours, tandis que la capacité restante (environ 20 %) est mise à la disposition de quiconque en a besoin sans processus de demande formelle. En 2022, les ressources totales disponibles se chiffraient à environ 294 000 cœurs-années. Le concours en a distribué 230 000, les 64 000 restants ayant été utilisés de façon libre, ou « opportuniste », par l'entremise du service d'accès rapide (SAR).

L'évolution de la capacité allouée (ligne jaune) suit de près celle de la capacité disponible (ligne bleue), avec la marge susmentionnée d'environ 20 % pour le service d'accès rapide. En comparant l'offre (ligne bleue) et la demande (ligne orange), on constate que, jusqu'au concours

de 2020, la capacité de CPU tirait de l'arrière par rapport à la croissance rapide des besoins, et que le développement de l'infrastructure de CIP ne suivait pas non plus la hausse de la demande. La situation s'est toutefois améliorée dans les deux dernières années, en raison de la diminution temporaire de la demande (expliquée plus haut) et de la hausse importante de l'offre (grâce à la mise en service de Narval à l'automne 2021).

La ligne verte pleine épaisse montre la demande non répondue en termes absolus. Au concours de 2022, celle-ci équivalait à environ 200 000 cœurs-années. La ligne verte pointillée épaisse illustre l'ampleur de la situation en montrant le pourcentage de la demande de calcul comblée par des ressources allouées. On peut voir qu'environ 80 % de la demande était satisfaite en 2012, pourcentage tombé à 40 % en 2020 avant de remonter à 54 % en 2022. Ainsi, malgré le répit temporaire récent, l'incapacité à répondre à la demande de calcul par CPU a augmenté considérablement dans la dernière décennie, tant en termes absolus qu'en termes relatifs.

Il faut néanmoins se rappeler que les ressources de CIP sont constamment prisées par leur nature même, en raison du nombre toujours plus grand de disciplines utilisant l'IRN, de la hausse de la résolution des instruments expérimentaux et observatoires, et du besoin de simulations plus précises (résolution et exactitude), souvent ajusté en fonction de la puissance de calcul allouée. Dès que les ressources augmentent, les chercheuses et chercheurs trouvent à les utiliser de nouvelles façons.

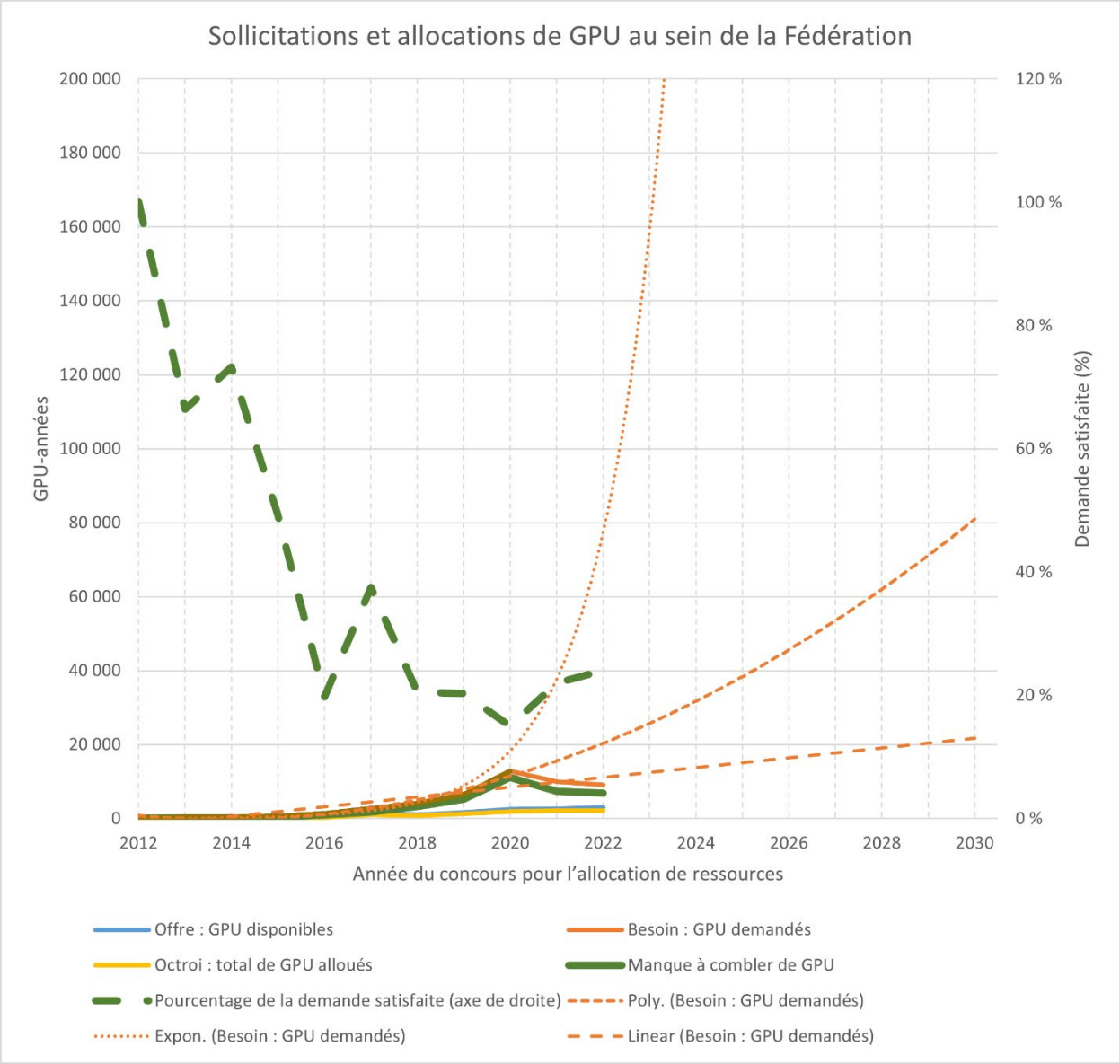

Figure A.2 – Demande de ressources de GPU et tendances projetées

La figure A.2 ci-dessus compare l'offre et la demande des ressources d'accélération par GPU dans le consortium de la Fédération Calcul Canada de 2012 à 2022. L'offre équivaut à la capacité de GPU réelle disponible dans les systèmes de la Fédération, et la demande, aux sollicitations de ressources reçues dans le cadre du concours annuel pour l'allocation de ressources de la Fédération. L'axe horizontal représente les années d'allocation, et l'axe vertical, les GPU-années. La ligne bleue représente la capacité totale pouvant être distribuée chaque année, et la ligne orange, la capacité demandée lors du concours. La ligne jaune illustre la capacité allouée aux utilisatrices finales et utilisateurs finaux dans le cadre du concours.

La demande de ressources de calcul par GPU (ligne orange pleine) a augmenté considérablement entre 2012 et 2020; en 2012, la demande était minimale, se chiffrant à 10 GPU-années, tandis qu'au concours de 2020, elle atteignait un total de près de 13 000 GPU-années. Il s'agit d'une croissance non linéaire et exponentielle d'une année à l'autre. Le TCAC de la croissance se chiffre à environ 67 % depuis 2017 (alors que la demande était d'environ 2 800 GPU-années), indiquant un essor fulgurant. La demande a plus récemment diminué, aux concours de 2021 et 2022, probablement pour deux raisons : 1) d'abord, comme pour la demande de processeurs centraux, en raison d'une baisse et d'une restructuration de la recherche universitaire pendant la COVID-19; et 2) parce que la demande de processeurs graphiques était gonflée par les complexités des nouvelles technologies et méthodes de calcul, et par les paradigmes et les difficultés liés à l'estimation du besoin réel. Avec l'émergence et l'adoption rapides des GPU, les chercheuses et chercheurs n'ont pas toujours les bases pour écrire un code performant, ou manquent de formation et de temps pour gérer et interpréter les résultats. Même avec la venue d'une capacité d'IA adaptée aux flux de travail dans les principaux établissements d'IA du Canada, le groupe de travail anticipe que la croissance de la demande reprendra rapidement avec le retour à la normale de la communauté de recherche. Cette conclusion est appuyée par les projections internationales des besoins de GPU, qui augmenteront considérablement dans un avenir proche.

En faisant fi de la diminution temporaire récente, la capacité de calcul par GPU prend du retard sur la croissance rapide de la demande. On peut voir du positif dans l'intérêt fortement accru pour les technologies accélératrices, mais on constate aussi que l'infrastructure de CIP ne suffit pas. Dans l'absolu, ce sont environ 7 000 GPU-années qui manquaient en 2022, comme l'indique la ligne verte pleine épaisse dans la figure plus haut. Relativement, la demande satisfaite (ligne verte pointillée épaisse) est passée de près de 100 % en 2012 à environ 20 % en 2020, avec un regain à 24 % au concours de 2022.

Les lignes orange minces projettent trois tendances de la demande. Les projections utilisent les données des concours de 2012 à 2020; elles excluent les données de 2021 et de 2022 en raison de la baisse anormale (et fort probablement temporaire). Les trois lignes correspondent à trois hypothèses fonctionnelles de croissance future : la ligne pointillée montre une croissance exponentielle; la ligne de petits tirets, une croissance polynomiale de deuxième degré; et la ligne de gros tirets, une croissance linéaire. L'estimation de la croissance exponentielle dépasse les extrêmes du graphique. Toute croissance exponentielle importante à plus long terme ralentira considérablement en raison de la maturation des technologies, comme expliqué plus haut. Les projections linéaire et polynomiale illustrent la fourchette de croissance anticipée de la demande : au concours de 2030, la demande se situerait entre 20 000 et 80 000 GPU-années.

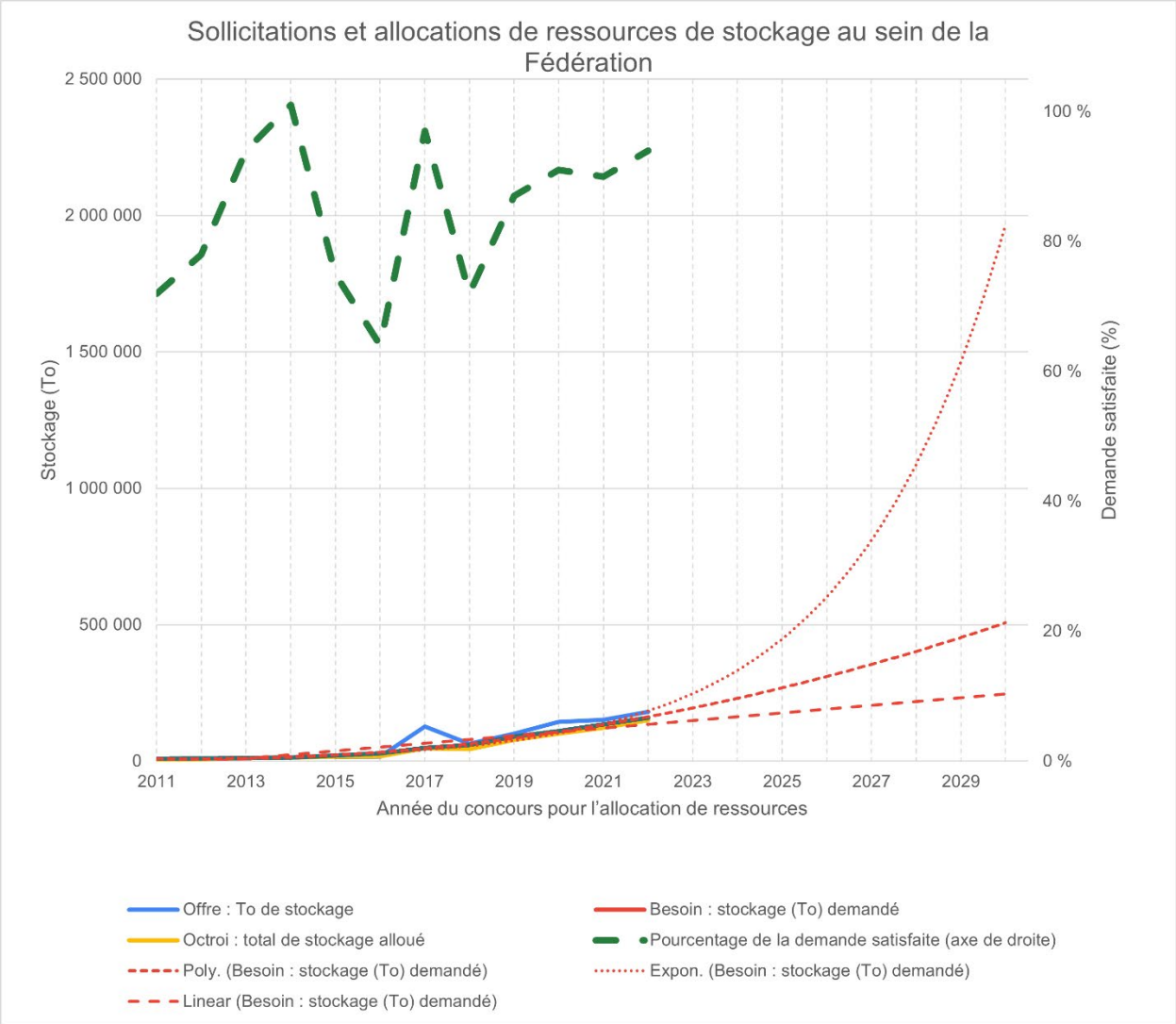

Figure A.3 – Demande de stockage et tendances projetées

La figure A.3 ci-dessus compare l'offre et la demande de stockage au sein de la Fédération Calcul Canada au fil des ans. La ligne bleue pleine représente le total de stockage disponible, tous types confondus. En 2015, la capacité totale s'élevait à environ 15 Po. En 2022, elle avait plus que décuplé, atteignant environ 180 Po, avec des variations annuelles importantes au rythme des mises hors service et des déploiements de systèmes. La capacité totale est un amalgame de plusieurs types de stockage : /project, dCache, /nearline, infonuagique, etc.

Du côté de la demande, la ligne rouge indique le total sollicité pour tous les types de stockage, et la ligne jaune, le total fourni. La demande se chiffrait environ à 21 Po en 2015 et 160 Po en 2022, presque huit fois plus. Cette hausse fulgurante sur sept ans équivaut à un TCAC d'environ 34 %. Point intéressant : en 2022, la capacité totale de stockage (ligne bleue) dépassait la demande totale (ligne rouge) d'environ 20 Po. La ligne verte pointillée représente le pourcentage de la

demande annuelle pour laquelle des ressources ont été attribuées. Ce pourcentage est passé de 72 % en 2011 à 94 % en 2022, avec une chute à 64 % en 2016. Grâce à l'augmentation de la capacité, le système de stockage dans son ensemble arrive à répondre à la demande.

Les lignes orange minces projettent trois tendances de la demande. Les trois lignes correspondent à trois hypothèses fonctionnelles de croissance future : la ligne pointillée montre une croissance exponentielle; la ligne de petits tirets, une croissance polynomiale de deuxième degré; et la ligne de gros tirets, une croissance linéaire. Ensemble, ces projections fonctionnelles illustrent la fourchette de croissance anticipée de la demande. Au concours de 2030, la demande de stockage actif se situerait donc entre 250 Po et 2 000 Po.

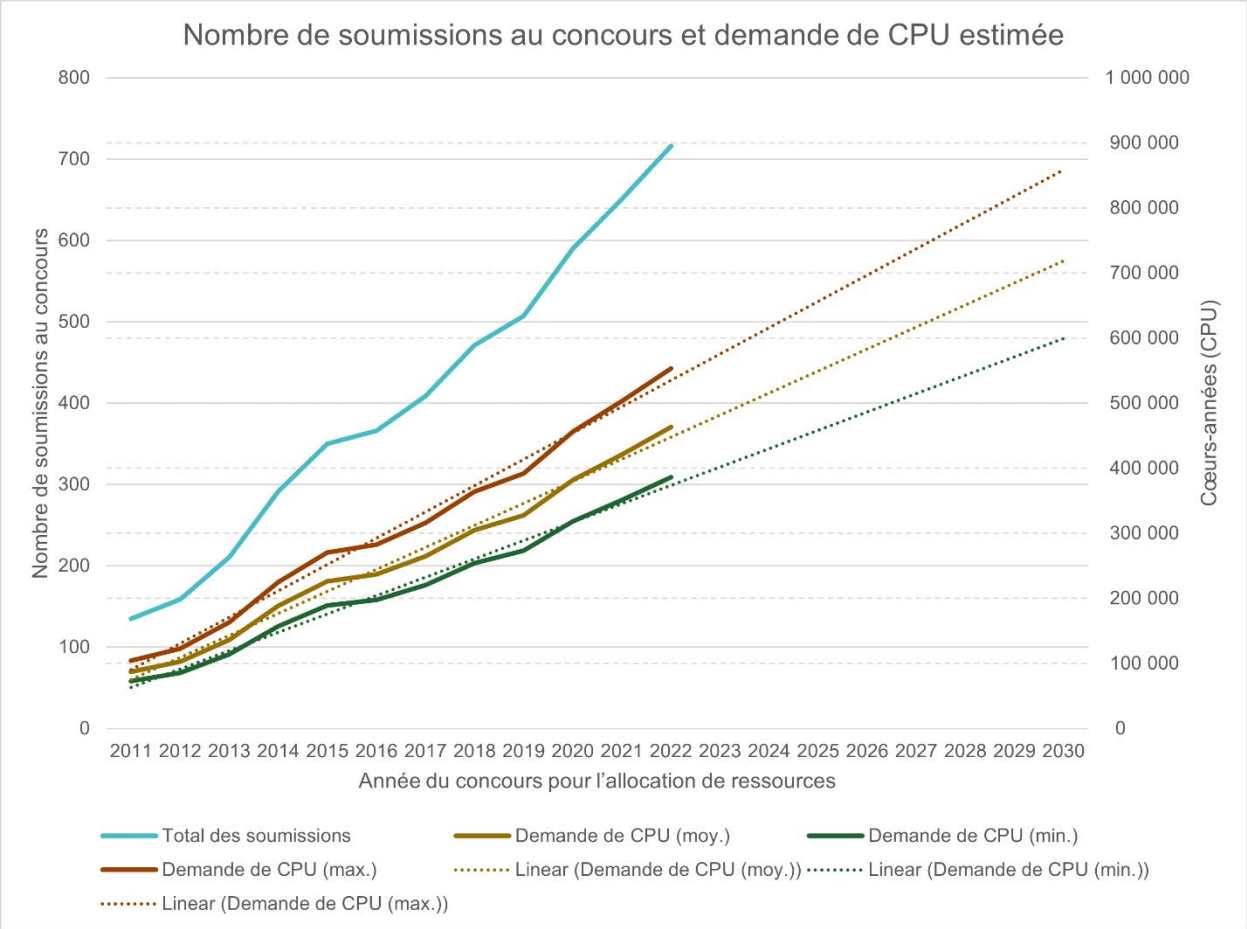

Figure A.4 – Nombre de soumissions au fil des ans et estimation de la demande de cœurs-années (CPU)

La figure A.4 ci-dessus montre le nombre de soumissions au concours pour l'allocation de ressources au fil des ans (ligne bleue pleine, axe vertical de gauche), et l'estimation de la demande de cœurs-années (CPU) selon ces soumissions (lignes pointillées, axe vertical de droite). Le nombre de soumissions croît chaque année, étant passé de 135 en 2011 à 716 en 2022. Sa croissance quasi linéaire devrait aussi se poursuivre, du moins du côté des utilisatrices traditionnelles et utilisateurs traditionnels des systèmes de CHP.

La moyenne de cœurs-années demandés dans chaque soumission (non illustrée) est demeurée relativement stable dans la décennie, avec une certaine variation d'une année à l'autre. Dans l'ensemble, la demande moyenne de 2011 à 2022 était de 647 cœurs-années par soumission, avec une fourchette de variation de 540 à 773 cœurs-années.

Les trois lignes pleines et les extrapolations en pointillé (en rouge, en brun et en vert; axe vertical de droite) montrent l'estimation de la demande de cœurs-années, selon une croissance linéaire du nombre de soumissions et une moyenne assez constante de demande par soumission.

Plus précisément, les lignes représentent la moyenne, le minimum et le maximum demandés par soumission, multipliés par le nombre annuel de soumissions et extrapolés de façon linéaire jusqu'en 2030. Selon la **moyenne** à long terme, la demande totale est estimée à **700 000 cœurs-années en 2030** (ligne brune pointillée), tandis que le minimum (ligne verte pointillée) et le maximum (ligne rouge pointillée) balisent une **fourchette potentielle de 600 000 à 870 000 cœurs-années en 2030**. Ces trois projections peuvent être utilisées pour approximer dans un premier temps le besoin de base de cœurs de CPU en 2030, si l'on suppose que les tendances et les profils d'utilisation demeureront relativement stables. À noter que la fourchette estimée pour 2030 équivaut environ à 2,5 fois la capacité actuelle (soit à peu près 296 000 cœurs).

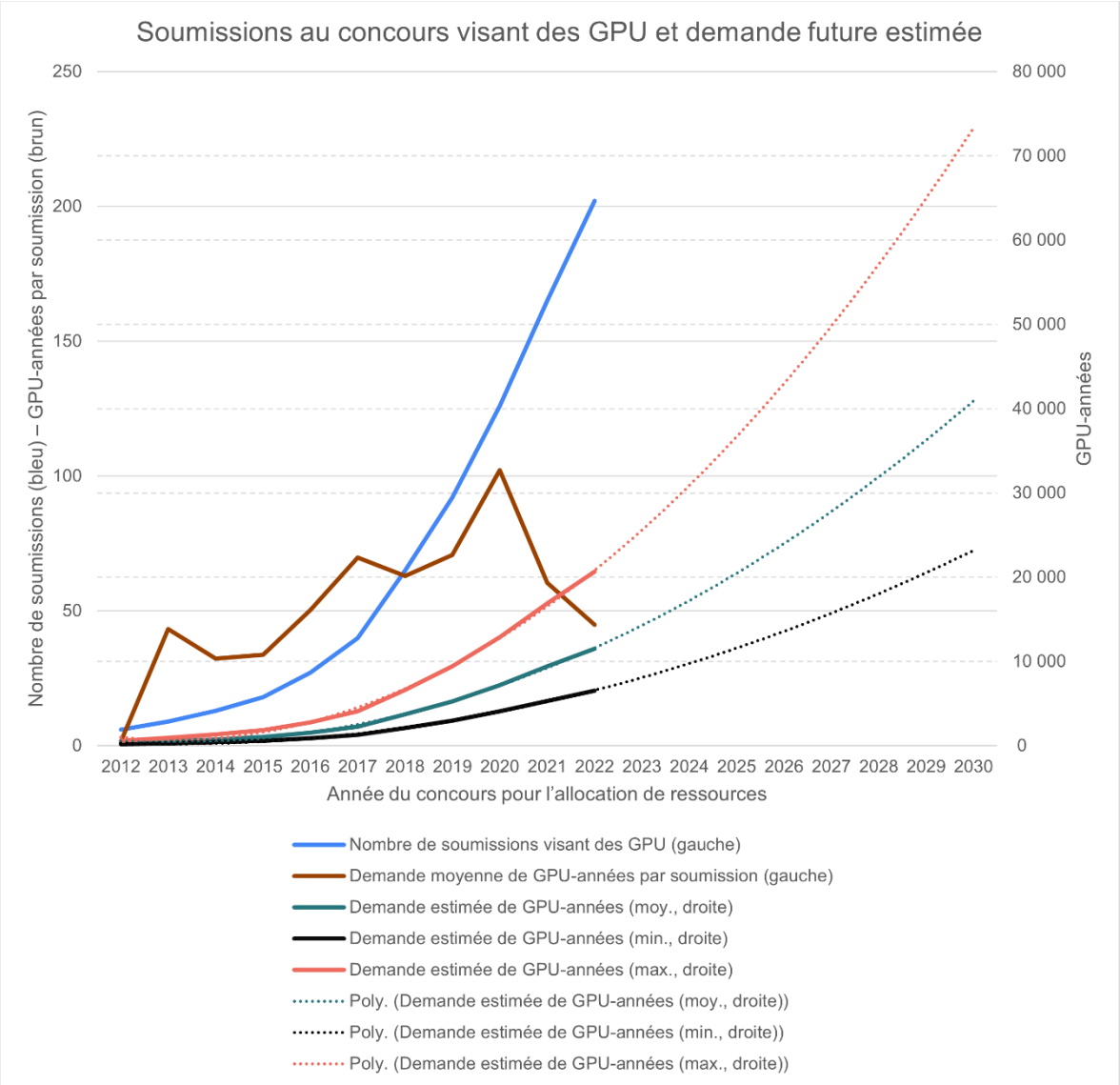

Figure A.5 – Nombre de soumissions visant des GPU au fil des ans et estimation de la demande associée

La figure A.5 ci-dessus montre le nombre de soumissions au concours pour l'allocation de ressources visant des GPU au fil des ans (ligne bleue pleine, axe vertical de gauche), et l'estimation de la demande de GPU-années selon ces soumissions (lignes pointillées, axe vertical de droite). La moyenne annuelle de ressources de GPU demandées par soumission est aussi indiquée (ligne brune pleine, axe vertical de gauche). En observant la ligne bleue pleine, on constate que le nombre de soumissions visant des processeurs graphiques croît chaque année depuis le concours de 2012, étant passé de 6 en 2012 à 202 en 2022. La croissance rapide initiale non linéaire indique qu'une proportion grandissante des soumissions comprennent une demande de GPU. Cette proportion se chiffrait à 4 % en 2012 et à 28 % en 2022. Ainsi, cette croissance est peu susceptible de se prolonger. On anticipe plutôt une croissance quasi linéaire mélangée à une croissance linéaire stable du nombre total de soumissions au concours (figure A.4).

La moyenne de ressources de GPU sollicitées par soumission (ligne brune pleine) a fluctué au cours de la dernière décennie, suggérant une difficulté des chercheuses et chercheurs à estimer leurs besoins face à cette technologie émergente. Dans l'ensemble, la demande moyenne de 2013 à 2022 était de 57 GPU-années par soumission (seulement les soumissions visant des processeurs graphiques), avec une fourchette de variation de 32 à 102 GPU-années.

Les trois lignes pleines et les extrapolations en pointillé (en rouge, en vert et en noir; axe vertical de droite) montrent l'estimation de la demande de GPU-années, selon la croissance du nombre de soumissions visant les processeurs graphiques et une moyenne assez constante de demande par soumission (suivant une méthodologie semblable à celle employée pour les processeurs centraux plus haut).

Plus précisément, les lignes représentent la moyenne, le minimum et le maximum demandés par soumission, multipliés par le nombre annuel de soumissions visant des GPU et extrapolés de façon linéaire jusqu'en 2030. Selon la **moyenne** à long terme, la demande totale est estimée à **40 000 GPU-années en 2030** (ligne verte pointillée), tandis que le minimum (ligne noire pointillée) et le maximum (ligne rouge pointillée) balisent une **fourchette potentielle de 23 000 à 63 000 GPU-années en 2030**. Ces projections peuvent être utilisées pour approximer dans un premier temps le besoin de base de GPU en 2030, si l'on suppose que les tendances et les profils d'utilisation demeureront relativement stables. À noter que la fourchette estimée pour 2030 équivaut environ à 10 à 20 fois la capacité actuelle (soit à peu près 3 000 GPU).

Il faut aussi savoir que la performance des processeurs graphiques augmente beaucoup plus rapidement que celle des autres technologies, ce qui rend souvent imprécises les estimations des chercheuses et chercheurs quant à leurs besoins. Un GPU-année n'est au mieux qu'une estimation grossière; un processeur NVIDIA P100 de 2017 atteint un sommet IEEE FP64 de 5 téraflops, tandis qu'un processeur H100 de 2022 atteint 60 téraflops. C'est 12 fois plus en à peine cinq ans. En comparaison, un cœur de CPU n'est devenu que deux à trois fois plus puissant (en virgule flottante à double précision) sur la même période.

Annexe B : Analyse de la charge de CHP

Les données d'ordonnancement des systèmes de CHP de la Fédération ont été analysées et résumées à l'aide d'un ensemble d'outils de l'équipe nationale d'analyse de données. Ce qui suit en est un petit échantillon permettant de dégager des tendances dans les tâches, d'encadrer l'ordonnancement et d'assurer l'utilisation efficace des systèmes.

B.1 Charge de travail

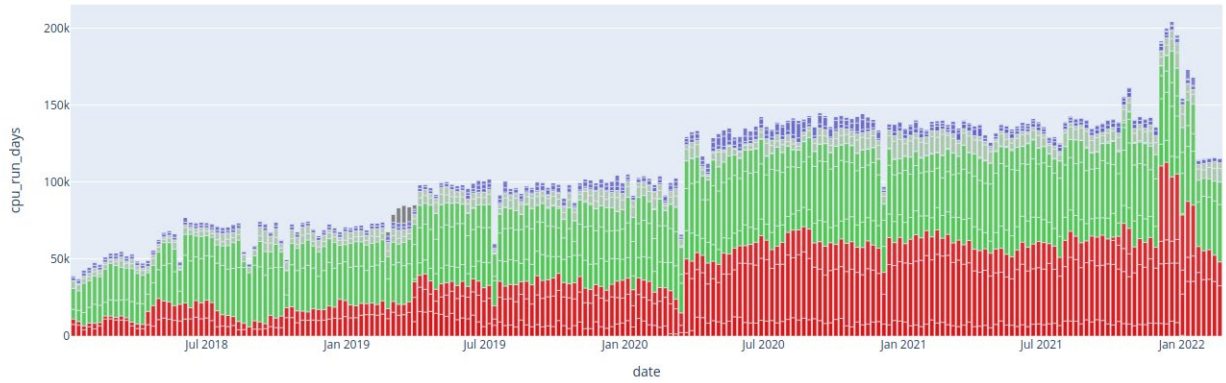

Figure B.1 – Charge de travail des CPU dans les systèmes de calcul générique, 2018 à 2022. La charge de travail des CPU dans les systèmes de CHP génériques au fil du temps. La hausse de l'utilisation au fil du temps est le résultat de deux expansions de Cedar et de l'ajout de Béluga.

La figure B.1 montre la charge de travail totale assumée par les CPU des systèmes de calcul générique Graham, Cedar et Béluga, de 2018 à 2022. L'utilisation est illustrée en journées d'exploitation des CPU, soit l'utilisation d'un cœur pendant 24 heures. Les couleurs correspondent au type de compte : le rouge représente les comptes par défaut, le vert, les comptes du concours pour l'allocation de ressources, le bleu, les comptes contributifs, et le gris, les autres comptes (voir la section B3). La hausse de l'utilisation au fil du temps est le résultat de deux expansions de Cedar et de l'ajout de Béluga. À l'aide des données de l'ordonnanceur, on peut évaluer la demande de ressources en comparant le travail effectué au travail en attente. Plus particulièrement, en examinant la demande pendant les périodes d'augmentation de la capacité, on peut observer l'effet des expansions sur la charge en attente.

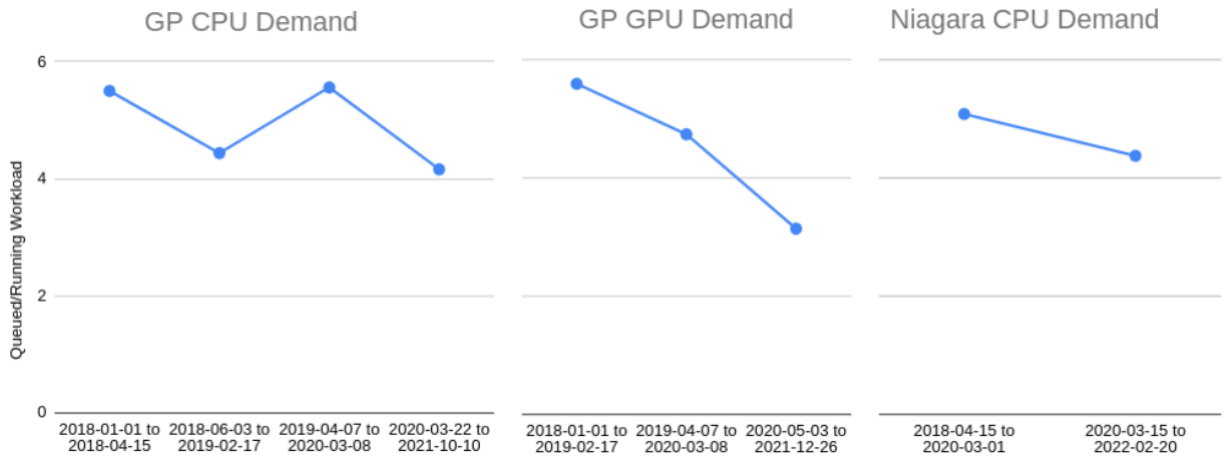

Figure B.2 – Charge de travail pendant les périodes d'expansion de la capacité de CHP. Évolution de la demande relative pendant les périodes d'expansion de la capacité de CHP. La demande par rapport aux ressources disponibles demeure relativement stable, même après d'importantes augmentations de la capacité de calcul.

La figure B.2 montre la demande en attente moyenne, avant et après les expansions des CPU et des GPU de calcul générique et des CPU de Niagara, normalisée en fonction de la capacité de fonctionnement des systèmes. Malgré certaines fluctuations, la demande en attente de ressources de CPU atteint généralement environ quatre à cinq fois la capacité de fonctionnement, même après d'importantes expansions de la capacité. On observe toutefois une réduction minime de la charge en attente du côté des GPU après les déploiements de ressources, par exemple après la mise en service du système Béluga, qui offre une grande capacité de GPU. La familiarité accrue des utilisatrices et utilisateurs avec cette technologie émergente aide aussi. Cela dit, malgré la diminution de la demande de GPU, la charge en attente demeure trois fois plus élevée que la capacité.

B.2 Caractéristiques de la charge sur les ressources

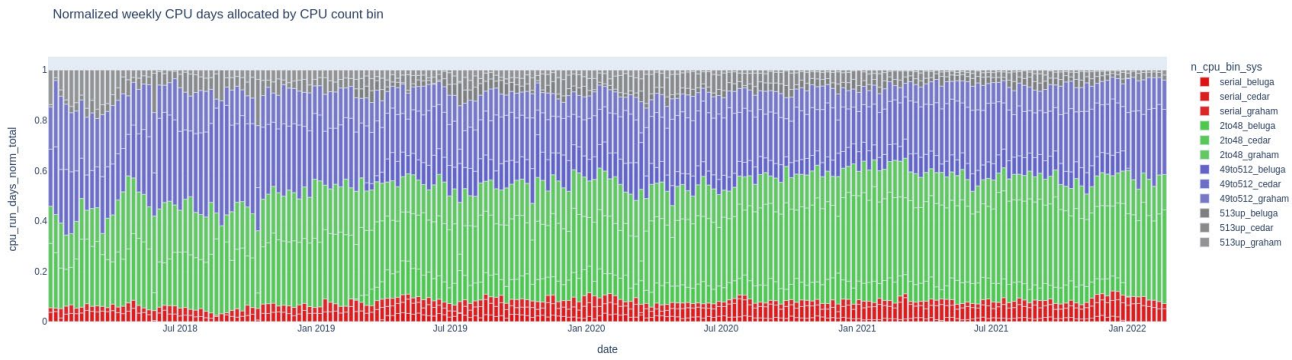

Figure B.3 – Taille des tâches par CPU dans les systèmes de calcul générique. Taille des tâches par CPU dans les systèmes de CHP génériques au fil du temps. Environ 50 % de l'usage des systèmes de CPU génériques ne demandent qu'un seul nœud, voire moins, une tendance qui se maintient sensiblement sur les quatre années de données.

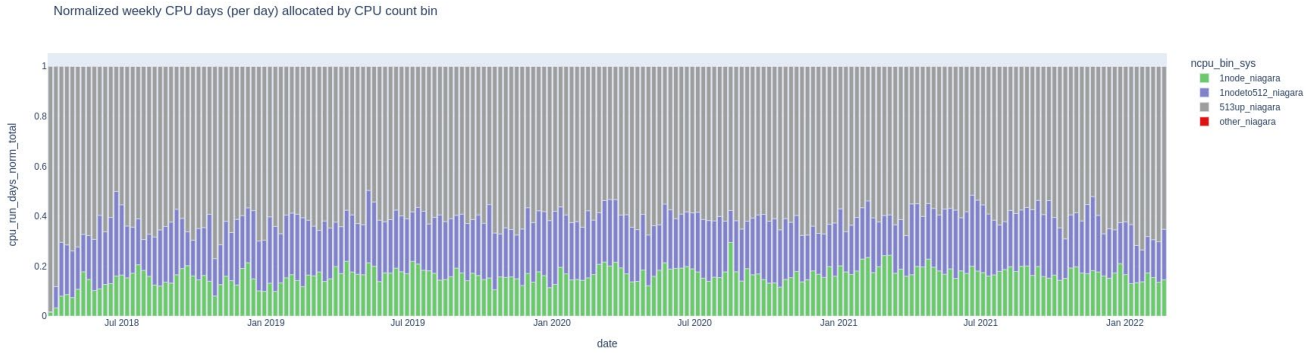

Figure B.4 – Taille des tâches par CPU dans Niagara. Taille des tâches par CPU pour le grand système parallèle Niagara. Environ 60 % de l'usage était consacré à des tâches requérant 512 cœurs ou plus.

La figure B.3 montre l'utilisation normalisée des ressources de CPU au fil du temps, de 2018 à 2022, dans les systèmes de calcul générique Cedar, Graham et Béluga, par tailles des tâches. La figure B.4 montre la même information, mais pour le grand système parallèle Niagara. Les barres rouges représentent les tâches en série, les vertes, les tâches utilisant 2 à 48 cœurs, les bleues, les tâches utilisant 49 à 512 cœurs, et les grises, les tâches utilisant plus de 512 cœurs. On peut voir qu'environ 50 % des fonctions de calcul générique exécutées par CPU ne demandaient qu'un seul nœud, voire moins, une tendance qui se maintient sensiblement sur les quatre années des données. Du côté de Niagara, environ 60 % de l'usage était consacré à des

tâches requérant 512 cœurs ou plus. Cela va de soi, car le système a été conçu et dédié expressément pour les tâches massivement parallèles. On notera aussi que la répartition des tâches selon leur taille ne change pas vraiment dans le temps, ce qui suggère que l'allocation actuelle d'infrastructure de calcul générique vs parallèle de grande taille convient toujours aux besoins.

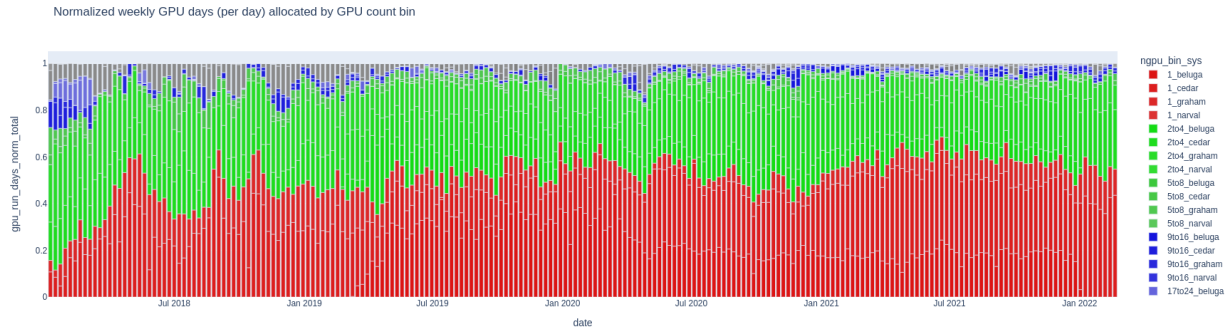

Figure B.5 – Taille des tâches par GPU dans les systèmes de calcul générique. aille des tâches par GPU dans les systèmes de calcul générique au fil du temps. Plus de 50 % des tâches exécutées par GPU ne demandent qu'un seul processeur et 95 % n'en demandaient que quatre ou moins (un seul nœud).

La figure B.5 montre l'utilisation normalisée des ressources de GPU au fil du temps, de 2018 à 2022, dans les systèmes de calcul générique Cedar, Graham et Béluga, par tailles des tâches. Les barres rouges représentent les tâches n'utilisant qu'un seul GPU, les vert pâle, les tâches utilisant 2 à 4 GPU, les vert foncé, les tâches utilisant 5 à 8 GPU, les bleues, les tâches utilisant 9 à 16 GPU, et les grises, les tâches utilisant plus de 17 GPU. On peut voir que plus de 50 % des tâches exécutées par GPU ne demandaient qu'un seul processeur, et 95 %, quatre ou moins (un seul nœud). Ces systèmes ont rarement été utilisés pour des tâches de GPU hétérogènes parallèles de grande taille. Ici aussi, la répartition selon la taille ne change pas vraiment dans le temps.

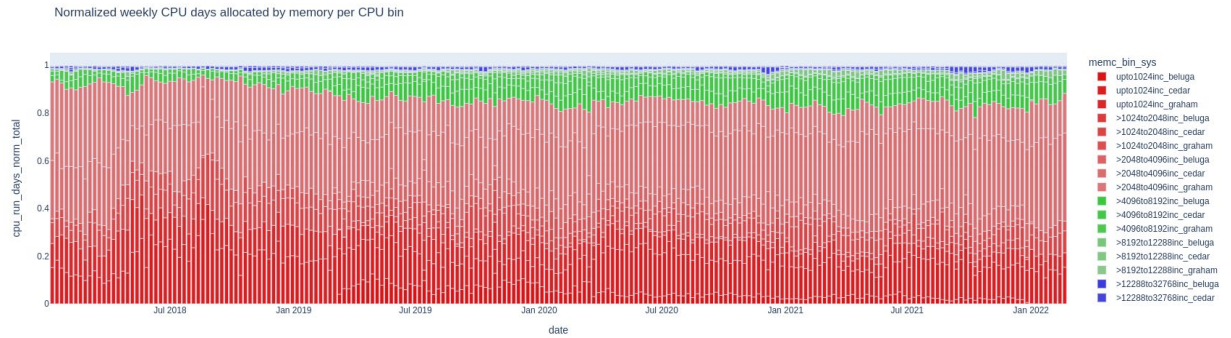

Figure B.6 – Usage de la mémoire par CPU dans les systèmes de calcul générique. Taille de la mémoire par CPU dans les systèmes génériques au fil du temps. Environ 85 % ont utilisé 4

Go/cœur ou moins, ce qui n'a rien de surprenant, car la majorité des systèmes sont conçus en fonction d'une telle utilisation.

La figure B.6 montre l'utilisation normalisée des ressources de CPU au fil du temps, dans les systèmes de calcul générique Cedar, Graham et Béluga, par consommation de mémoire. Les barres rouge foncé représentent les tâches requérant un maximum de 1 Go/cœur, les rouge pâle, les tâches requérant un maximum de 4 Go/cœur, les vertes, les tâches requérant un maximum de 12 Go/cœur, et les bleues, les tâches requérant jusqu'à 32 Go/cœur. Dans l'ensemble des systèmes, ce sont environ 85 % tâches qui ont utilisé 4 Go/cœur ou moins, ce qui n'a rien de surprenant puisque la majorité des systèmes sont conçus en fonction d'une telle utilisation, soit avec un nœud de 32 cœurs à 128 Go de mémoire vive. Encore une fois, la répartition selon la taille ne change pas vraiment dans le temps. Le système Niagara n'est ordonnancé que par nœuds complets, de sorte que toutes les tâches disposent d'environ 4,5 Go/cœur.

B.3 Tâches selon les types d'allocations au concours pour l'allocation de ressources

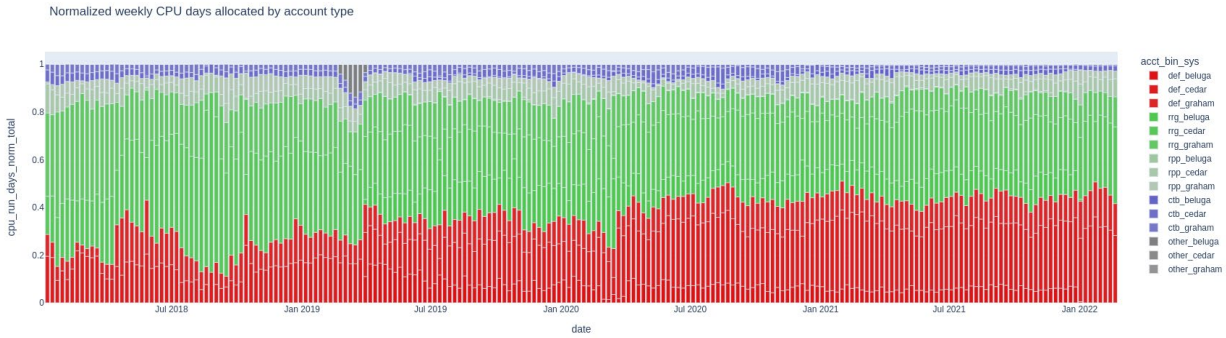

Figure B.7 – Utilisation de CPU dans les systèmes de calcul générique, par types d'allocations. Utilisation de CPU par type d'allocations au concours dans les systèmes de calcul génériques au fil du temps. Dans les systèmes de calcul générique, les allocations par défaut peuvent atteindre 40 % des ressources de CPU.

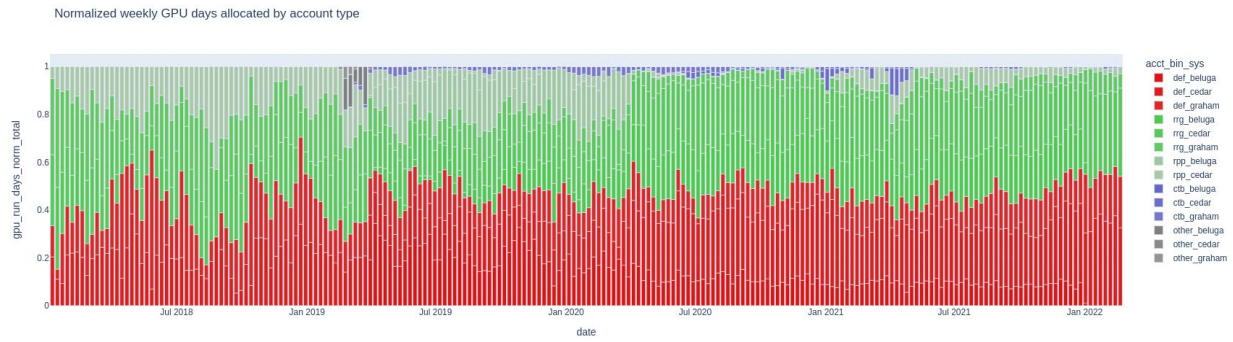

Figure B.8 – Utilisation de GPU dans les systèmes de calcul générique, par types d'allocations. Utilisation de GPU par type d'allocations au concours dans les systèmes de calcul génériques au fil du temps. Dans les systèmes de calcul générique, les allocations par défaut peuvent atteindre 50 % des ressources de GPU.

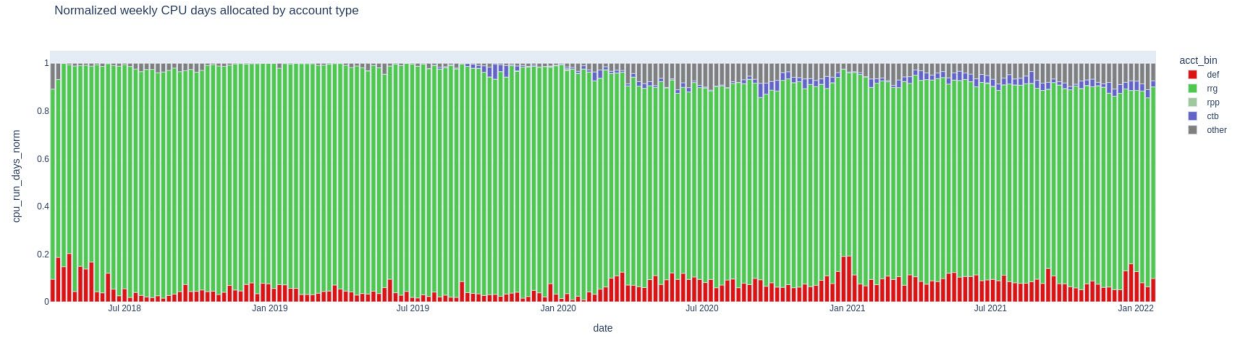

Figure B.9 – Utilisation de CPU dans le système Niagara, par types d'allocations. Utilisation de CPU par type d'allocations au concours pour le grand système parallèle Niagara au fil du temps. La grande majorité des ressources utilisées pour Niagara sont les ressources allouées par concours pour les groupes de recherche.

Les figures B.7, B.8 et B.9 illustrent l'utilisation des ressources de calcul générique par CPU, de calcul générique par GPU et du système Niagara, par types d'allocations. Les barres rouges représentent les tâches exécutées dans le cadre d'allocations par défaut, les vertes, les tâches exécutées dans le cadre de l'allocation de ressources aux groupes de recherche, les vert pâle, les tâches exécutées dans le cadre de l'allocation de ressources aux portails et plateformes de recherche, les bleues, aux tâches contributives, et les grises, aux tâches autres. Le processus du concours pour l'allocation de ressources distribue habituellement 80 % des ressources annuelles disponibles, mais à l'exception du système Niagara, seulement 60 % de ces ressources sont exploitées par des comptes ayant reçu des ressources du concours. Dans les systèmes de calcul générique, les allocations par défaut peuvent atteindre 40 % des ressources de CPU et plus de 50 % des ressources de GPU. Il n'y a pas de changement majeur dans la répartition de l'utilisation selon les types d'allocations, à l'exception d'une hausse de l'allocation de CPU par défaut dans les systèmes de calcul générique, d'environ 20 % en 2018 à environ 40 % en 2022.

Annexe C : Réseautique

C.1 Interconnexions internes à haute vitesse

Les systèmes de CHP ont besoin d'un réseau de haute performance pour travailler comme un tout plutôt que comme une collection de nœuds séparés. Or, la conception d'un réseau et de ses capacités dépend de l'usage prévu.

Les choix de technologies et de conception doivent être soigneusement étudiés pour que le réseau soit capable de soutenir la charge de travail de la grappe sans tomber dans la surconception de la matrice. Une réflexion consciencieuse est d'autant plus importante que les coûts d'interconnexion peuvent engloutir une grande partie du budget total d'une grappe.

La plateforme nationale doit pouvoir répondre à tous les besoins de CHP, petits et grands; elle doit être développée et élargie. Les systèmes déployés devraient pouvoir convenir à tout l'éventail des usages ciblés. Les données plus haut sur la taille des tâches sont un outil précieux dans la conception du réseau, car elles montrent l'utilisation qui est faite de la plateforme nationale et les façons dont cette dernière peut s'y adapter.

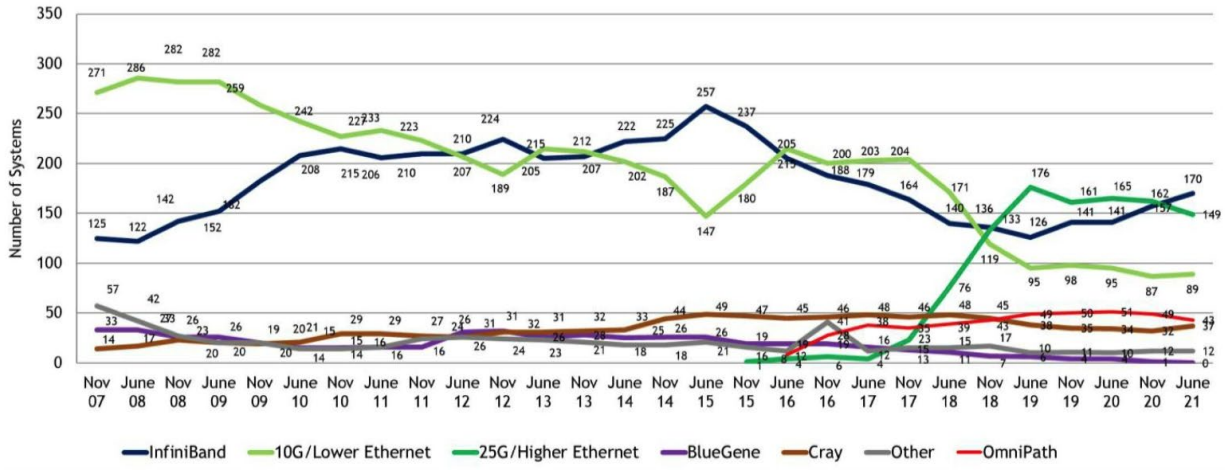

Figure C.1 – Distribution des interconnexions dans les systèmes de TOP500 au fil du temps.[29]
Distribution des interconnexions dans les systèmes au palmarès de TOP500 au fil du temps. InfiniBand prime sur le marché depuis des années et est encore la matrice interconnectée de CPH non Ethernet de prédilection.

[29] TOP500. http://top500.org/statistics/list/ (novembre 2021).

Comme la montre la figure C.1, InfiniBand prime sur le marché depuis des années et est encore la matrice interconnectée de CHP non Ethernet de prédilection. Mais ce n'est pas la seule technologie. L'Ethernet à haute vitesse est aussi une option courante et habituellement moins coûteuse, qui n'a toutefois pas la performance globale d'InfiniBand et offre moins d'options de protocoles de communication et de topologie. Il pourrait néanmoins être un compromis raisonnable selon le type de travail visé.

Il existe également des options moins connues qui pourraient convenir. Cela dit, il est important pour une question de risques de ne pas déployer une nouvelle technologie dans plusieurs sites. Notamment, la solution OPA d'Intel semblait prometteuse, mais n'est aujourd'hui plus très utilisée, ce qui complique le soutien technique et l'expansion. De nouvelles matrices comme HPE Slingshot et Rockport, maintenant disponibles, pourraient être mises à l'essai dans des contextes limités.

Les matrices assurent les communications entourant les tâches, mais aussi l'accès aux fichiers. Avec le nombre croissant de calculs axés sur les données, la matrice doit pouvoir rapidement transporter des fichiers du stockage central aux nœuds de calcul. Par ailleurs, même si les nœuds de grande taille réduisent les communications internœuds en accueillant un plus grand nombre de tâches, ils demandent davantage de bande passante pour acheminer les données à tous leurs cœurs.

Éléments à prendre en compte dans l'acquisition de la matrice :

1. Soutien pour les différents types et tailles de tâches prévus (bande passante et latence des communications intertâches, mouvements des données entrantes et sortantes)
2. Diversité de la matrice dans l'ensemble de la plateforme nationale
3. Coût, extensibilité et disponibilité de chaque option
4. Expérience du personnel avec la technologie

C.2 Réseautique externe

La connectivité des réseaux étendus dans les sites nationaux se résume en trois types de services :

1. Services IP de CANARIE et du Réseau national de la recherche et de l'éducation (RNRE)
2. Internet commercial (grand public)
3. Réseaux de projet de recherche

Tous les sites bénéficient des services IP de CANARIE et du RNRE et de l'Internet commercial (1,2). Ces services sont décrits dans les sous-sections ci-dessous.

C.2.1 Services IP de CANARIE et du RNRE (recherche et éducation)

Chaque site national utilise le réseau dorsal de recherche canadienne administré par CANARIE et ses partenaires régionaux du RNRE, visant à assurer une connectivité à haute vitesse entre les sites et avec les établissements reliés au réseau de CANARIE. Ce dernier veille aussi à la connectivité des réseaux internationaux de recherche et d'éducation.

CANARIE ne fournit pas directement la connexion aux sites; la configuration de chaque connexion dépend du partenaire régional du RNRE qui octroie l'accès au réseau dorsal. Aucun trafic Internet commercial ne passe par le réseau IP de CANARIE.

Le tableau C.1 présente la connectivité des sites nationaux de recherche et d'éducation au RNRE et à CANARIE.

Exemples de trafic utilisant ce service :

- Trafic intersite (ex. : transferts de données entre les sites)
- Trafic des campus de recherche (établissements connectés à leur branche provinciale du RNRE)

C.2.2 Internet commercial (grand public)

Un service IP d'Internet commercial est nécessaire pour que les utilisatrices et utilisateurs puissent se connecter aux systèmes de CHP et aux nuages à partir d'Internet ainsi que pour assurer les activités normales des systèmes (DNS, correctifs, courriels, etc.). Ce service est fourni par le partenaire régional du RNRE ou le campus local, qui détermine la bande passante nécessaire avec le site.

Bien que les transferts de jeux de données de recherche soient censés passer par le service IP à haute vitesse de CANARIE et du RNRE, les nuages commerciaux sont de plus en plus utilisés à cette fin et usent généralement de services Internet commerciaux à plus basse vitesse. Dès qu'un accord d'appairage est signé[30] entre CANARIE ou le RNRE et un fournisseur de nuage commercial, une connexion à plus haute vitesse est mise en place pour ces transferts.

L'utilisation de ressources infonuagiques commerciales augmente (ex. : transferts de jeux de données entre les sites d'hébergement et le nuage), et la capacité de connexion du réseau doit s'y adapter (bande passante supérieure ou appairage avec CANARIE ou le RNRE).

C.2.3 Réseaux de projet de recherche

Bien que l'architecture IP dorsale de CANARIE et du RNRE permette une connectivité à haute vitesse pour la recherche et l'éducation, le réseau doit être puissant de bout en bout (bande

[30] CANARIE. *Diffusion de contenu*. https://www.canarie.ca/fr/reseau/services/sdc/ (consulté en mai 2022).

passante, latence, friction) pour répondre aux besoins des projets scientifiques internationaux et à grande échelle à forte intensité de données.

Parmi les réseaux adaptés, on pensera à LHCONE de l'initiative ATLAS (3.5.1 Physique des hautes énergies). Au Canada, les sites Cedar, Arbutus et Graham sont interconnectés avec LHCONE et participent à l'initiative, Cedar étant le site d'hébergement pour le centre de données de premier niveau ATLAS.

Le projet Square Kilometer Array (SKA) est une autre initiative scientifique à grande échelle dont on anticipe la production et la distribution massives de données de recherche (3.5.2 Square Kilometer Array [SKA1]). L'architecture pour le mouvement des données n'a pas encore été définie, mais devrait venir avec son lot d'exigences de réseautique pour les sites participants du Canada.

Site	**RNRE**	**Lien de recherche et d'éducation**	**Fournisseur du lien grand public**	**Capacité du lien grand public***	**Réseau de projet de recherche (en date de 2022)**
Narval/Béluga	RISQ	100 Gb/s	RISQ	400 Mb/s	
Graham	ORION	100 Gb/s	ORION	2 000 Mb/s	ATLAS/LHCONE
Niagara	GTAnet /ORION	100 Gb/s	Université de Toronto		
Cedar	BCNET	100 Gb/s	BCNET	200 Mb/s	ATLAS/LHCONE (niveau 1)
Arbutus	BCNET	100 Gb/s	BCNET	600 Mb/s	ATLAS/LHCONE

Tableau C.1 – Connectivité Internet des sites d'hébergement. * Capacité des accords d'appairage avec les fournisseurs non comprise.

C.2.4 Estimations et projections du trafic

En utilisant les données des systèmes de surveillance réseau de chaque site, on peut dresser un portrait de l'utilisation des réseaux étendus pour analyser et estimer les volumes de trafic. Puisque les données sont résumées sur une longue période, les pointes temporaires ne sont pas visibles.

- Les sites hébergeant ATLAS/LHCONE (Cedar, Arbutus et Graham) reçoivent plus de trafic sur leur réseau étendu que les autres (Béluga, Niagara).
- Cedar est le site d'hébergement ATLAS de premier niveau, et son réseau étendu est le plus utilisé. En 2021, sa capacité de 100 Gb/s était pleinement utilisée.

Le tableau C.2 ci-dessous indique les exigences de bande passante de niveau 1 estimées par LHC pour la prise de données 3 d'ATLAS. Ces exigences sont additives – et non substitutives – à celles déjà en place pour le reste du trafic.

Année	**Bande passante**
2023	60 Gb/s
2025	100 Gb/s
2027	200 Gb/s

Tableau C.2 – Exigences de bande passante de niveau 1 pour la prise de données 3 d'ATLAS

En plus des exigences de bande passante, la Grille de calcul mondiale pour le LHC (WLCG) demande que tous les sites participants soient compatibles avec le protocole IPv6. Bien que l'équipement réseau actuel réponde à cette condition depuis des années, il sera important que les prochaines acquisitions soient à double pile (IPv4 et IPv6).

Le trafic du projet Belle-II (centre de données hébergé à l'Université de Victoria) devrait être minimal dans les prochaines années, mais pourrait éventuellement rattraper celui de l'initiative ATLAS dans Arbutus (niveau 2).

Le projet Square Kilometer Array (SKA) risque d'apporter son lot d'exigences de réseautique pour les sites participants et CANARIE, qui devront planifier une connexion internationale capable de transporter les données des sites éloignés. Cela dit, comme il en est encore à l'étape de la conception architecturale et des prototypes, il ne s'agit pas d'un besoin immédiat.

Résumé

- La capacité de la bande passante devra être augmentée à court terme pour accueillir le trafic du centre de premier niveau du LHC dans Cedar.
- Toute hausse de la capacité du réseau devra être coordonnée avec CANARIE et le RNRE, et avec le campus local dans certains sites.

- Du matériel, du personnel et de l'équipement doivent être consacrés et déployés pour l'exploitation des réseaux de transport de données et des outils de mesure de la réseautique de prochaine génération (ex. : perfSONAR, utilisation des liens).

Annexe D : Capacité des centres de données des sites d'hébergement

Il existe actuellement cinq sites d'hébergement nationaux. La superficie, l'alimentation électrique et l'infrastructure de refroidissement disponibles pour chacun sont indiquées au tableau D.1. Ce sont généralement ces deux derniers facteurs qui sont limitants, et non la superficie. En effet, comme les systèmes de CHP deviennent plus énergivores, la capacité d'alimentation et de refroidissement doit souvent être augmentée, ou l'architecture, modifiée. À l'exception de l'Université de Victoria, qui n'utilise que le refroidissement à air, tous les sites emploient principalement des échangeurs thermiques à porte arrière alimentés en eau refroidie. Les bâtis standards avec refroidissement à air sont généralement limités à 10 à 15 kW de puissance par bâti, et les bâtis à échangeurs thermiques, eux, cessent le plus souvent d'être efficaces au-delà de 30 à 40 kW pour chacun.

		Simon-Fraser	**Waterloo**	**Toronto**	**McGill**	**Victoria**	**Total**
Bâtis	Utilisés	75	60	72	77	105	389
	Max.	120	200+	120	230	140	810+
Puissance (MW)	Utilisée	1,45	0,6	1,45	1,75	0,496	5,746
	Max.	10	1,4	3,75	2	3	21,15
Refroidisse-ment (T)	Utilisé	400	200	475	400	120	1595
	Max.	950	260	735	515	266	2 726

Tableau D.1 – Capacité actuelle des centres de données des sites d'hébergement

Annexe E : Coût des ressources

Afin d'estimer les coûts de la capacité et de son expansion, deux systèmes de référence ont été créés : un système avec 100 000 cœurs de CPU, et un système avec 1 000 GPU. Les deux utilisent un réseau InfiniBand à débit binaire NDR/HDR bloqué 3:2 doté d'un système de fichiers de travail parallèle de haute performance, /scratch, utilisant un stockage à semi-conducteurs équivalant à environ 10 fois la mémoire du système. La configuration de référence des nœuds de calcul et des systèmes est indiquée aux tableaux E.1 et E.2, respectivement. Les coûts ont été estimés à partir d'acquisitions récentes de systèmes importants et des prix habituels avec garantie de cinq ans pour les établissements d'éducation. Les configurations de référence ont ensuite été multipliées selon les facteurs des scénarios à la section 4.2 pour obtenir les estimations finales.

Nœud de calcul (+ fraction de /scratch et du réseau)	Puissance	Coût
2x Intel à 32 cœurs, 256 Go de mémoire vive, HDR 200G	800 W	19 000 $
2x AMD à 24 cœurs, 512 Go de mémoire vive, 4x NVIDIA A100 80 Go, HDR 200G	3 000 W	84 000 $

Tableau E.1 – Nœuds de calcul de référence

Système	Nombre de nœuds	/scratch	Puissance	Bâtis	Coût
100 000 cœurs de CPU	1 562	4 Po	1 250 kW	40	29,5 M$
1 000 GPU	250	1,3 Po	750 kW	28	21 M$

Tableau E.2 – Systèmes de CHP de référence

À noter que le coût des technologies est hautement changeant. On se donnera donc comme principe directeur de maximiser la capacité pour les chercheuses et chercheurs en fonction du budget prédéterminé au moment de la demande de propositions. Le même principe s'applique

au choix des technologies, qui a été simplifié dans le présent document aux fins d'examen et de comparaison. Il est d'autant plus important avec les GPU, pour lesquels les options se multiplient à tous les niveaux de performance et de prix. Le système de référence utilise le GPU le plus complet et performant, mais puisqu'il est aussi très dispendieux, il est peu probable que cette option haut de gamme soit la seule utilisée si l'on souhaite optimiser la valeur offerte.

Une approche semblable a été utilisée pour estimer le coût de différentes technologies de stockage à trois niveaux (tableau E.3). La conception d'un système de fichiers et les choix de logiciels et de technologies peuvent grandement influer sur le prix. Néanmoins, ces estimations généralisées supposent une configuration standard de la Fédération. L'espace /project estimé, avec la double copie de sauvegarde sur bande magnétique, s'élève à 150 000 $/Po. Le stockage /nearline, qui utilise aussi une double copie sur bande magnétique, est estimé à 50 000 $/Po.

Type de stockage	Technologie	Niveau de stockage	Coût/Po
Haute performance	Semi-conducteurs	/scratch	500 000 $/Po
Milieu de gamme	Disque mécanique	/project, dCache, nuage	100 000 $/Po
Sur bande magnétique	Sur bande magnétique	Sauvegarde, /nearline, archivage	25 000 $/Po

Tableau E.3 – Systèmes de stockage de référence