Mai 2021

État actuel du calcul informatique de pointe au Canada

Mise à jour concernant le rapport du CLIRN sur le calcul informatique de pointe (2017)

Rédaction : Groupe de travail de l'Alliance sur le calcul informatique de pointe

Seppo Sahrakorpi, Maxime Boissonneault, Emmanuel Château-Dutier, Chris Loken, Catherine Lovekin, Carolyn McGregor, Felipe Pérez-Jvostov, Ghilaine Roquet et Lisa Strug.

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

Financé par le gouvernement du Canada | Canada

Table des matières

Graphiques

Tableaux

1 Résumé analytique

Cette mise à jour concerne le rapport sur le calcul informatique de pointe, que le Conseil du leadership sur l'infrastructure de recherche numérique (CLIRN) a soumis à Innovation, Sciences et Développement économique Canada (ISDE) en 2017. Il s'agit d'un aperçu de l'environnement du calcul informatique de pointe (CIP) au Canada. On y présente les forces, les défis et les possibilités du CIP à l'heure actuelle, ainsi que l'Alliance de recherche numérique du Canada (l'Alliance). Le Groupe de travail de l'Alliance sur le CIP a rédigé ce rapport pour illustrer la vision et les priorités de l'équipe, mais il ne s'agit pas d'une déclaration officielle de l'Alliance sur l'état actuel du CIP au Canada.

Grâce à ce travail, l'Alliance pourra tracer une voie, à partir de l'état actuel, pour faire progresser le CIP conjointement avec les autres éléments de l'infrastructure de recherche numérique (IRN), pour promouvoir l'excellence en recherche au Canada. Les conclusions et les observations de ce document, ainsi que les publications sur l'évaluation de l'état actuel de la gestion des données de recherche (GDR) et des logiciels de recherche (LR), visent à fournir des conseils et des renseignements contextuels aux analystes et à la direction de l'Alliance, au conseil d'administration de l'Alliance et au Conseil des chercheurs de l'Alliance, afin d'appuyer l'évaluation des besoins et l'élaboration du nouveau modèle de prestation de services, du plan stratégique et du financement de l'Alliance.

Le CIP est essentiel à la recherche moderne et la demande connaît une croissance phénoménale (pour les données massives et de l'intelligence artificielle par exemple). Il s'agit d'un domaine très concurrentiel ; la technologie et les outils évoluent continuellement et rapidement, notamment avec les nouvelles disciplines qui émergent et se transforment. Tout cela exige un écosystème agile et très réactif, ainsi que du personnel hautement qualifié (PHQ). Le financement doit être durable et prévisible si le Canada veut relever les défis du 21^{e} siècle et demeurer concurrentiel à l'échelle internationale.

Le Canada est un très grand pays diversifié avec une longue histoire et la communauté de recherche canadienne reflète cette diversité, autant sociale que géographique. Les écosystèmes canadiens de CIP et l'IRN doivent desservir tous les membres de cette communauté diversifiée pour mieux faire avancer la recherche nationale. Ces éléments profitent à la société et contribuent au progrès. Le CIP peut non seulement aider à résoudre des problèmes et des questions scientifiques, qui seraient autrement difficiles à résoudre, mais aussi contribuer à des solutions qu'on ne pourrait pas trouver avec les moyens habituels (calcul analytique ou expérimental sur des postes de travail).

Le CIP est un ensemble de technologies et de services numériques permettant aux chercheurs de résoudre des problèmes de recherche qui sont trop importants ou trop complexes pour les entreprendre seuls. Outre les grappes de serveurs traditionnelles, le CIP moderne repose également sur des infrastructures virtuelles et dématérialisées. Les chercheurs bénéficient d'un environnement numérique de pointe, tout comme l'enseignement supérieur s'équipe physiquement d'infrastructures et d'installations traditionnelles pour mener des activités de recherche, dans un environnement où les ressources sont très fortement utilisées et mises en commun. Voici les principaux éléments du CIP :

- Calcul (unités centrales de traitement et processeurs graphiques),

- Stockage actif et sauvegarde (stockage en cours d'exécution, de proximité et temporaire),
- Soutien, formation et consultation par du personnel hautement qualifié (PHQ),
- Gestion et soutien logiciel (logiciels d'exploitation, protocoles de communication et de bibliothèque d'usage courant),
- Confidentialité, sécurité et authentification,
- Connexions haute vitesse aux réseaux de recherche, ainsi qu'aux équipement nationaux et internationaux, ainsi qu'entre les sites,
- Soutien et coordination avec les autres composants de GDR (mise en réseau, gestion des données, stockage à moyen et long terme, logiciels de recherche).

Pour la Fédération Calcul Canada (FCC), ces composants n'incluent pas le matériel, les systèmes et les services de stockage à moyen terme (dépôt) et à long terme (archivage) pour des raisons de financement. Ils font néanmoins partie intégrante du cycle informatique de la recherche moderne et de l'IRN dans son ensemble. Ces composants et considérations de stockage sont un élément essentiel du prochain mandat de l'Alliance.

Résumé de l'état actuel

Les chercheurs ont divers degrés d'accès aux ressources de CIP, à l'échelle collective, ministérielle, institutionnelle, régionale, nationale et internationale. Notons que plusieurs chercheurs ne connaissent pas les ressources de la FCC, pensent que ces ressources ne leur sont pas adressées ou utilisent des systèmes autres que ceux de la fédération. À l'heure actuelle, nous sommes uniquement en mesure de quantifier et d'analyser la dimension nationale de la FCC, même si les chercheurs canadiens ont accès à plusieurs autres ressources. Ce rapport inclut presque exclusivement les données d'utilisation de la FCC.

Le nombre d'**utilisateurs inscrits de la FCC** a augmenté considérablement dans la dernière décennie, pour atteindre 16 000 en 2020. Depuis 2014, le taux de croissance annuel composé (TCAC) est d'environ 12 %. Au 1er janvier 2020, le corps enseignant constituait le plus important groupe d'utilisateurs avec une proportion de 27 %. Les quatre plus importants groupes d'utilisateurs (corps enseignant, étudiants à la maîtrise, au doctorat et au postdoctorat) représentaient les trois quarts de tous les utilisateurs de la FCC.

Le corps enseignant en génie, biologie et sciences de la vie est le plus important groupe d'utilisateurs, avec une proportion de 19 % chacun. En sciences humaines, sciences sociales, commerce et psychologie, le corps enseignant représente environ 10 % de la base d'utilisateurs, alors que ces disciplines comptent environ 46 % de l'ensemble du corps enseignant universitaire canadien, ce qui montre clairement que ces disciplines sont sous-représentées pour le nombre d'utilisateurs du CIP au Canada (même s'il est difficile de déterminer si cet écart s'explique par les divers besoins de CIP entre les disciplines) et le virage numérique de la recherche universitaire. Le nombre d'utilisateurs dans ces disciplines a augmenté d'environ 40 % sur trois ans depuis 2017. Par conséquent, il y a un grand potentiel de croissance en ce qui concerne le nombre absolu d'utilisateurs en sciences humaines et autres, ce qui est une belle occasion de promouvoir le soutien et la formation à la GDR. Il y a un réel besoin d'accès aux ressources

informatiques de pointe en sciences humaines et sociales. Ces disciplines se transforment considérablement, ce qui nécessite des ressources particulières, dont le Canada ne dispose pas suffisamment et qui explique dans une certaine mesure la sous-représentation de ces disciplines.

Dernièrement, la **Fondation canadienne pour l'innovation (FCI) a largement investi dans la cyberinfrastructure canadienne,** ce qui a permis de consolider les ressources, de mettre au point de nouveaux systèmes, en plus d'augmenter considérablement la capacité et le potentiel du CIP au Canada. Les cinq principaux sites de la FCC hébergent cinq nouveaux systèmes nationaux, qui sont affiliés à ces organisations régionales et membres de la FCC :

- University of Victoria, Arbutus (Calcul Canada/WestGrid),
- Simon Fraser University, Cedar (WestGrid),
- University of Waterloo, Graham (Calcul Ontario),
- University of Toronto, Niagara (Calcul Ontario),
- McGill University /École de technologie supérieure, Béluga (Calcul Québec).

Cedar, Graham et Béluga sont des systèmes de CIP hétérogènes qui répondent à diverses charges de travail dans ce domaine. Arbutus est un système de CIP en infonuagique à usage général pour l'hébergement de systèmes virtuels (principalement basées sur Linux) et pour répondre à d'autres charges de travail dématérialisé. Niagara est un système de CIP homogène massivement parallèle pour les charges de travail évolutives dans ce domaine.

La FCC a utilisé environ 200 000 CPU en 2019, soit quatre fois plus depuis 2010, ce qui représente un taux de croissance annuel composé (TCAC) d'environ 16 %. En revanche, l'utilisation est toujours limités aux ressources disponibles. En d'autres termes, la communauté a recours aux ressources dont dispose la FCC. Trois domaines de recherche (ingénierie, physique et astronomie, chimie et biochimie) ont consommé environ trois quarts des ressources de CPU. Les sciences sociales, la psychologie, le commerce et les sciences humaines utilisent huit fois plus de ressources depuis dix ans. En termes absolus, l'utilisation demeure modeste, avec environ 1250 CPU en 2019. La croissance des CPU par année pour le CIP montre un fort intérêt et le potentiel de l'IRN à l'avenir dans ces disciplines. Par contre, l'indicateur de CPU par année n'est pas nécessairement le plus pertinent pour évaluer l'utilisation des ressources dans ces disciplines.

Si l'on considère l'**utilisation des GPU** dans l'ensemble de l'écosystème de la FCC, l'utilisation totale des ressources GPU était d'environ 1300 GPU en 2019, ce qui correspond à un TCAC d'environ 56 % depuis 2012. Cette croissance a été sévèrement limitée par l'offre disponible et elle n'est donc pas indicative de l'augmentation du taux de croissance réel des besoins en GPU à usage général (GPU). Quatre disciplines (informatique et sciences de l'information ; chimie et biochimie ; sciences biologiques et sciences de la vie ; physique et astronomie) ont utilisé environ 87 % des ressources GPU en 2019.

La FCC ne suit pas actuellement de façon détaillée l'**utilisation de l'infonuagique** au sein de la fédération. Au mois de janvier 2021, la FCC offrait des ressources d'informatique dématérialisée sur plusieurs systèmes et dans diverses régions, soit environ 7 % de la capacité totale de calcul CPU de la FCC. Le système d'infonuagique Arbutus contribue à 87 % de toutes les ressources

dématérialisées de l'infrastructure de la FCC. La recherche sur la COVID-19 et la physique sont les domaines qui ont le plus utilisé le système Arbutus l'année dernière. Cet indicateur intéressant montre comment les ressources en nuage peuvent être déployées de manière flexible pour de nouvelles recherches et de nouveaux besoins, comme en témoigne la recherche sur la COVID-19, qui a utilisé le plus grand nombre de ces ressources.

Pour les **logiciels de recherche**, l'Alliance publiera, au cours de l'été 2021, un document détaillé sur l'état actuel de la RS, qui traitera en détail de l'utilisation des logiciels de recherche et des tendances. En ce qui concerne le soutien et l'informatique de recherche, on peut conclure que pour chaque progiciel, un très petit nombre d'utilisateurs utilise de nombreux produits logiciels (notamment depuis que la FCI a mandaté la FCC pour soutenir l'ensemble de l'informatique de recherche au-delà des postes de travail). Ceci présente manifestement ses propres défis dans l'écosystème de la maintenance et du soutien logiciel.

La FCC a produit environ **12 000 billets de soutien pour la recherche** en 2019, avec une forte croissance entre 2016 et 2019, avec la migration des utilisateurs et des anciens systèmes depuis les anciens systèmes régionaux/institutionnels vers des systèmes nationaux. La très grande majorité des billets de soutien de la FCC sont liés à l'utilisation générale des systèmes et de l'infrastructure de CIP. En d'autres termes, la plupart des billets de soutien ne se rattache pas à des besoins scientifiques propres à un domaine (programmation, etc.).

La **formation** est une activité très importante de la FCC, notamment les séminaires, les ateliers et les cours d'été, entre autres. Elle est essentielle pour l'adoption, la sensibilisation, la formation de la main-d'œuvre numérique, ainsi que la mise à jour des compétences des chercheurs. En 2019-20, le consortium de la FCC a dispensé 46 000 heures de formation au total (heures d'événement multipliées par le nombre de participants) à environ 14 000 participants lors de 460 événements en personne.

Principaux atouts de l'écosystème canadien du CIP

Important approvisionnement en services de CIP

Dans le cadre du processus annuel de renouvellement des comptes, la FCC mène un sondage auprès des utilisateurs pour avoir leur avis sur les ressources et les services de la fédération. En 2020, 85 % des utilisateurs des plateformes de CIP étaient « satisfaits » ou « très satisfaits » des services de la FCC, tandis que seulement 3 % des répondants étaient « insatisfaits » ou « très insatisfaits ». Les utilisateurs de toutes les disciplines de recherche semblent être à peu près également satisfaits des ressources et services de la FCC.

Conception et actualisation de l'infrastructure du CIP

Depuis plus de 20 ans, la FCC et ses prédécesseurs offrent aux chercheurs canadiens des systèmes haut de gamme de CIP extrêmement productifs, par le biais de consortiums en CIP. La FCC a mis en place l'ensemble actuel des cinq systèmes nationaux entre l'automne 2016 et le printemps 2019, grâce à l'Initiative pour la cyberinfrastructure de la FCI. Avec des fonds supplémentaires d'ISDE, la FCC a étendu et mis à niveau quatre des systèmes au début de 2020. De plus, un nouveau système sera installé d'ici l'automne 2021. Conçus à l'origine comme étant semblables, les trois systèmes à usage général varient considérablement en taille à cause d'un

écart de financement qui se multiplie par trois. Au total, ces systèmes représentent un peu moins de 170 millions de dollars d'investissements fédéraux et provinciaux.

Le Canada compte actuellement 5 systèmes de recherche qui figurent dans le plus récent classement (novembre 2020) du Top500, soit les superordinateurs les plus puissants du monde : Beluga, Cedar, Niagara, ainsi que deux systèmes du gouvernement fédéral qui servent principalement au climat et à la météorologie. Aujourd'hui, le plus rapide superordinateur du monde, le Fugaku du Japon, est environ 123 fois plus rapide que le système canadien qui se classe parmi les premiers du pays.

Prestation de services centralisée

Dans le cadre de la modernisation des services de la FCC, l'organisation est passée à un modèle d'exploitation et de soutien plus national, notamment avec un environnement informatique et de données plus homogène et cohérent, ainsi que plusieurs équipes nationales. Le modèle de service a été amélioré, notamment avec un accès uniforme par le biais d'identifiants centralisés, une meilleure qualité de documentation (centralisée et bilingue), des services de transfert de données améliorés, une approche centralisée pour des services de stockage plus uniformes (par le biais de politiques et d'un système de fichiers normalisés), une restructuration de la sécurité, une pile logicielle centralisée, ainsi qu'un processus de demande centralisé pour les comptes et l'allocation de ressources. Les utilisateurs finaux disposent désormais d'un point de contact unique pour le soutien informatique à la recherche, tandis que le personnel de soutien de l'établissement est accessible au besoin. La FCC et ses affiliés ont également amélioré la portabilité des charges de travail entre les plateformes. Avec une programmation unique par lots, les utilisateurs finaux peuvent utiliser des scripts de soumission semblables sur différents systèmes avec des modifications mineures. La FCC a également amélioré les rapports sur l'état des systèmes grâce à son service centralisé de publication des ressources, qui fournit des informations actualisées sur les ressources disponibles.

Une solide communauté de personnel hautement qualifié (PHQ)

Le réseau de la FCC est composé d'environ 200 équivalents temps plein (ETP) qui sont membres du personnel hautement qualifié. Ils gèrent les activités et les sites de la fédération à travers le Canada. Ces personnes fournissent divers services essentiels liés à l'administration des systèmes de CIP, l'approvisionnement, la maintenance, la mise en réseau, l'exploitation, la gestion, la planification, le financement, le soutien, le développement de logiciels de recherche, la gestion des données, la formation, la gestion des comptes et des allocations, les communications, la mobilisation, etc. De nombreux analystes et administrateurs de systèmes sont titulaires de diplômes d'études supérieures. Ils ont également de l'expérience au niveau de la recherche et comme utilisateurs du CIP. Les systèmes de CIP sont en fait hautement sophistiqués et complexes en ce qui concerne la configuration, les piles logicielles et matérielles, l'exploitation et l'utilisation, ce qui exige une expertise de niveau supérieur et plusieurs années de spécialisation. Par conséquent, l'écosystème canadien de l'IRN dépend de ces compétences et de la rétention du personnel hautement qualifié.

La FCC mène régulièrement un sondage sur la satisfaction après l'émission de billets pour évaluer le service à la clientèle. Les réponses témoignent de la grande qualité du service que fournit le PHQ de la FCC. En ce qui concerne la rapidité de réponse, 94 % des personnes

sondées l'ont jugée bonne ou excellente. Pour ce qui est de la « solution fournie », 91 % des personnes étaient satisfaites de la solution.

Engagement financier renouvelé

Le gouvernement canadien, par l'intermédiaire d'Innovation, des Sciences et du Développement économique (ISDE), voit clairement la valeur de l'IRN pour la société canadienne, comme le prouve l'engagement budgétaire de 572,5 millions de dollars pour 2018. Du côté de l'Alliance, cela se traduit par un financement fédéral total de 375 millions de dollars jusqu'en mars 2024, ce qui assure une continuité importante (relativement) à long terme pour le financement de l'IRN. De plus, cette restructuration équilibre et centralise également le financement, qui rassemble les trois éléments clés d'une IRN moderne.

Principaux défis et possibilités de l'écosystème canadien de CIP

Coordination de la planification stratégique et opérationnelle nationale

L'IRN canadien dépend d'une approche plus coordonnée et centralisée dans la planification stratégique et opérationnelle nationale, notamment par l'entremise d'un financement plus durable et prévisible. Il est important d'avoir une approche nationale pour accroître les synergies et les efficacités (p. ex. en optimisant l'utilisation des ressources), améliorer l'interopérabilité et la convivialité, en plus de maximiser l'expertise du PHQ à l'échelle canadienne. En cas d'incendie, d'inondation ou de toute autre catastrophe majeure sur un site hôte, la FCC pourrait potentiellement perdre entièrement ce site, ainsi que toutes les données qui y sont stockées. De plus, la planification doit tenir compte explicitement du CIP traditionnel, du stockage à court, moyen et long terme, des sauvegardes hors site ou intersites, de la gestion des données de recherche et des besoins en logiciels de recherche, sous une seule enveloppe, tout en mettant l'accent sur les disciplines, les communautés et les publics mal desservis. Par ailleurs, il faut garder à l'esprit que chacun des systèmes de CIP est différent et que les disciplines et communautés ont divers besoins en ce qui concerne les infrastructures et services de CIP.

Étant donné les contraintes financières, les fournisseurs passés et actuels de systèmes et de services de CIP n'ont pas eu l'occasion de concentrer sur la gestion des données de recherche, l'accessibilité et la convivialité des logiciels de recherche, le stockage à long terme pour la prestation de services, ou encore les besoins de publics et de disciplines plus vastes comme les sciences humaines et sociales. Cette situation n'est pas nécessairement due à un manque de vision ou de reconnaissance de la part des fournisseurs de CIP, mais plutôt aux limitations des mandats de financement.

Approvisionnement en CIP insuffisant

En ce qui concerne l'**offre et la demande pour les ressources informatiques CPU** de la FCC entre 2012 à 2020, la capacité totale disponible a fluctué d'environ 155 000 à 230 000 années CPU, tandis que les demandes présentées au Concours pour l'allocation de ressources (CAR) de Calcul Canada (CC) ont augmenté d'environ 100 000 années CPU à 450 000 années CPU. En termes de TCAC, la croissance de la demande de cycles informatiques CPU était d'environ 21 %. En 2020, environ 40 % de la demande était satisfaite. Les 270 000 années CPU de demande non satisfaite correspondent à environ 3,4 fois la valeur des ressources informatiques

du superordinateur Niagara. Les systèmes sont en général très sollicités, soit environ 90 % de tous les cycles théoriquement disponibles.

Parmi ses homologues du G7, le Canada se classe au dernier rang en ce qui concerne la puissance de calcul totale cumulative du Top500. Le rapport entre la puissance de calcul et le produit intérieur brut (tFlops/PIB) place le Canada à l'avant-dernier rang au sein du G7. Il faudra au moins doubler notre capacité de CIP pour rattraper les pays du G7 qui se situent dans la moyenne.

La demande de **ressources informatiques GPU** a augmenté rapidement aux cours des 10 dernières années, pour atteindre près de 13 000 années GPU dans le cadre du CAR 2020. Les ressources GPU ont connu une souscription qui excédait environ 5 fois la capacité. Il faut garder à l'esprit que de nombreux flux de travail et applications ne peuvent pas exploiter pleinement les GPU, ce qui mène à une sous-utilisation et un manque d'optimisation. En termes absolus, la demande non satisfaite de GPU en 2020 était d'environ 11 000 années GPU, ce qui équivaut à environ huit cycles GPU du superordinateur Cedar actuel. Pour illustrer cette échelle autrement, avec des cartes NVIDIA V100 Volta GPU modernes, le coût total uniquement pour les cartes d'accélération serait d'environ 100 millions de dollars. Le coût réel pour répondre aux besoins non satisfaits de 2020 en matière de GPU serait encore plus élevé si l'on inclut le coût des milliers de serveurs de CIP et des autres infrastructures de soutien.

Par rapport au déficit de capacité en matière de CPU et GPU mentionné ci-dessus, la FCC a beaucoup mieux répondu à la demande ces dernières années en ce qui concerne la **capacité de stockage** actif. Il s'agit d'une tendance positive si l'on garde à l'esprit que les besoins de stockage sont souvent non transitoires et très différents de l'utilisation ponctuelle des ressources CPU et GPU. Les chercheurs ne s'attendent pas à ce que leurs allocations de stockage disparaissent dans les prochaines années. Par ailleurs, on ne peut pas octroyer à autrui un stockage déjà utilisé. En 2020, la capacité de stockage totale de 140 Po était supérieure d'environ 30 Po à la demande totale, sachant qu'une grande partie de cette marge de manœuvre supplémentaire dans l'infrastructure de stockage est nécessaire au bon fonctionnement du système. La demande a augmenté d'environ 5 fois et elle est relativement linéaire depuis cinq ans, ce qui correspond à un TCAC approximatif de 39 %.

Par ailleurs, l'analyse du stockage actif ci-dessus ne tient pas compte de l'offre ou des besoins en matière de données d'archivage ou de dépôt à long terme. Dans le cadre de son mandat, la FCC ne fournit pas ce type de stockage, même si les systèmes de bandes qui sont utilisés pour le stockage nearline pourraient, d'un point de vue technique, répondre à ces besoins. La FCC dispose également de l'expertise et du PHQ nécessaires. Il faudrait un investissement substantiel fédéral pour stockage nearline et d'archivage à long terme, conjointement avec la capacité de stockage de sauvegarde correspondante, pour soutenir les initiatives de GDR de manière cohérente et durable.

Évolution au même rythme que la diversité technologique et culturelle

L'écosystème canadien actuel du CIP et les fournisseurs de services peinent à évoluer au même rythme que la diversité technologique et culturelle (notamment en ce qui concerne les facteurs de géographie et d'âge). En plus des progrès méthodologiques et de la chaîne de compilation de l'infrastructure de recherche numérique (IRN), cet écosystème est un nouvel outil précieux pour les utilisateurs et les disciplines non traditionnelles du CIP, notamment les sciences sociales,

sciences humaines, sciences de la santé, études autochtones, etc. Avec l'accroissement des données et du contenu disponibles, ces disciplines reconnaissent l'énorme potentiel de l'IRN pour leur recherche. En revanche, ces initiatives créent aussi des préoccupations concernant la sensibilité, confidentialité, propriété et sécurité des données. Dans le contexte des sciences humaines, ce ne sont pas seulement les données, mais aussi le contenu qui sont numérisés. La manipulation de contenus numérisés et de documents nés numériques (issus des réseaux sociaux, internet, etc.) renforce le mouvement vers l'utilisation de méthodes informatiques qui existent déjà dans ces disciplines. Par exemple, plusieurs projets font appel au traitement informatique pour l'analyse automatisée de textes, voix, sons, images ou vidéos, que ce soit pour l'exploration ou la classification de données.

La société reconnaît désormais un peu plus les besoins de divers groupes sous-représentés, notamment les communautés racialisées, LGBTQ+ et autochtones. Elle tente de mieux y répondre. Si les chercheurs issus de ces disciplines et de ces communautés ne connaissent pas les systèmes et les outils modernes du CIP, ils peuvent bénéficier d'innovations en matière d'IRN, de formation, de documentation, de PHQ ou d'outils leur permettant d'accéder aux systèmes de l'IRN. Par ailleurs, les nouvelles technologies de CIP émergent rapidement et se diversifient dans une certaine mesure. Notons par exemple l'adoption croissante de l'informatique GPU, les nouvelles architectures de puces d'IA, l'informatique quantique, l'infonuagique (dont les infrastructures en tant que service IaaS, les plateformes en tant que service PaaS, les logiciels en tant que service SaaS), etc.

L'écosystème canadien d'IRN actuel n'est pas bien équipé ou financé pour répondre aux besoins susmentionnés. L'accent a été mis davantage sur les besoins des communautés d'utilisateurs traditionnels du CIP, avec une certaine croissance dans l'essai de nouvelles technologies (p. ex. de nouveaux paradigmes d'utilisation au-delà de l'accès en ligne de commande, en tirant parti des passerelles scientifiques, etc.), mais sans financement ou effort coordonné à l'échelle nationale.

Sensibilisation des chercheurs et adoption du CIP

Le manque de connaissance et d'adoption du CIP et de l'IRN chez les chercheurs est un problème majeur au Canada et dans le monde. Il y a environ 33 000 professeurs d'université titulaires et associés au Canada, alors que la FCC compte actuellement environ 5 500 comptes de chercheurs principaux (CP). En d'autres termes, environ 17 % des professeurs titulaires et associés se sont inscrits pour utiliser l'infrastructure de la FCC.

Entre autres, les sciences sociales, les sciences humaines, la psychologie, le commerce et les sciences de la santé sont actuellement sous-représentés, mais pourraient énormément bénéficier du CIP et de l'infrastructure de recherche numérique pour faire progresser leurs domaines. À l'échelle mondiale, plusieurs organisations intéressantes participent au financement et à l'exploitation de l'IRN pour ces disciplines qui deviennent de plus en plus informatisées. La Fédération EGI est une infrastructure électronique internationale qui fournit des services informatiques et d'analyse de données avancés ; Parthenos Virtual Research Environment (VRE) est un environnement en ligne pour les sciences humaines intégrant le stockage en nuage avec des services et des outils pour le cycle de vie des données de recherche ; ARIADNEplus offre des environnements virtuels en nuage pour la recherche archéologique basée sur les données ; DARIAH est une infrastructure paneuropéenne pour les chercheurs en arts et en sciences humaines ; IPERION HS est une plateforme européenne d'infrastructure de recherche intégrée

pour les sciences du patrimoine. En France, l'infrastructure Huma-Num fournit aux chercheurs en sciences humaines et sociales des services de CIP, mais aussi une gamme complète de services d'IRN tout au long du cycle de vie de la recherche. Cette infrastructure est considérée comme une « Très Grande Infrastructure de Recherche (TGIR) » et bénéficie d'un financement public. Elle fournit des plateformes et des outils pour le traitement, la conservation, la diffusion et la préservation à long terme des données numériques de recherche. En ne desservant pas ces disciplines à leur plein niveau de représentativité, le Canada risque fort de perdre des occasions et de se faire dépasser par la concurrence mondiale.

Outre les disciplines et les communautés qui ne profitent pas de l'IRN, certains chercheurs dans les disciplines « traditionnelles » n'accèdent pas aux systèmes de la FCC pour diverses raisons. Parfois ils ne connaissent pas la gamme de services, ils considèrent que les interfaces utilisateur et l'utilisation sont trop compliquées, ils laissent tomber parce que leur demande a été rejetée ou ils ont eu une mauvaise expérience. Par ailleurs, certains chercheurs répondent à leurs besoins en matière de CIP par d'autres moyens et n'ont donc pas besoin d'un compte ou des ressources de la FCC. Ces chercheurs ont dans certains cas l'expérience nécessaire pour utiliser leurs propres systèmes qui sont adaptés à leurs besoins, en plus travailler sur des problèmes qui ne nécessitent pas de ressources informatiques et de stockage puissant et haut de gamme.

Il est donc important de sensibiliser les chercheurs en déployant des efforts particuliers et concentrés pour atteindre toutes les communautés qui ne profitent pas encore de l'IRN. Ces efforts doivent s'accompagner de ressources d'IRN, au niveau de l'infrastructure, des services, de la formation et du personnel de soutien, afin que tous les chercheurs intéressés puissent bénéficier de l'IRN. Tout en sensibilisant ces parties prenantes, il faut améliorer les technologies d'accès et d'utilisation pour répondre aux besoins des chercheurs qui ne sont pas nécessairement intéressés ou à l'aise avec les technologies de l'information et l'IRN. Par conséquent, il faudrait plus de ressources pour aider les chercheurs à exploiter de manière efficace le CIP et l'IRN dans leurs travaux, notamment avec des formations ciblées, du soutien, une documentation, ainsi que des nouvelles passerelles et des intergiciels innovants pour accéder à l'IRN.

Égalité, diversité, inclusion (EDI) et représentation des minorités

On ne reconnaît pas suffisamment l'importance de l'égalité, diversité, inclusion (EDI) et représentation des minorités dans le domaine du CIP, ce qui pose d'importants problèmes, en plus du manque de solutions d'EDI. Pour la prestation de services, il faut tenir compte des besoins des communautés autochtones, des immigrants au Canada, des chercheurs dans les régions éloignées et rurales, des chercheurs en début de carrière et chercheurs chevronnés, ainsi que des chercheurs handicapés ayant des besoins particuliers en matière d'accessibilité. À titre d'exemple, l'infonuagique a un énorme potentiel pour réduire les inégalités d'accès aux ressources. La FCC ne recueille pas actuellement de données sur l'EDI, donc la fédération ne connaît pas précisément l'état de cette situation. L'EDI ne devrait pas être considéré comme un élément isolé, mais doit plutôt faire partie de toutes les discussions et prises de décision.

L'écosystème de l'IRN doit aussi répondre aux besoins des personnes dont l'anglais n'est pas la langue maternelle, notamment les communautés francophones et les utilisateurs allophones (avec par exemple de la documentation clairement rédigée). Tous les documents et services devraient être accessibles dans les deux langues officielles. La qualité de la traduction doit être équivalente à celle du document d'origine et non à celle d'un logiciel automatique. Par ailleurs, la documentation et les services importants doivent être accessibles dans certaines langues

autochtones. Pour les principaux événements et les conférences, il faut prévoir des services d'interprétation en langue des signes et d'interprétation bidirectionnelle (voire multidirectionnelle).

2 Méthodologie

Ce rapport sur l'état actuel du CIP a été produit entre l'hiver 2020 et le printemps 2021 dans le cadre d'un processus en plusieurs étapes, incluant des consultations avec des représentants de la communauté canadienne du CIP. Il a été rédigé par le groupe de travail de l'Alliance sur le CIP, qui est composé des membres suivants :

- Seppo Sahrakorpi (analyste principal en CIP, Alliance, président)
- Ghilaine Roquet (vice-présidente, stratégie et planification, Alliance)
- Felipe Pérez-Jvostov (analyste principal, communications et mobilisation, Alliance)
- Maxime Boissonneault (chef d'équipe, soutien à la recherche, Calcul Canada/Calcul Québec)
- Chris Loken (directeur de la technologie, Calcul Ontario)
- Emmanuel Château-Dutier (muséologie numérique, Université de Montréal)
- Catherine Lovekin (astronomie, Université Mount Allison)
- Carolyn McGregor (informatique de santé, Institut universitaire de technologie de l'Ontario)
- Lisa Strug (biostatistique, Université de Toronto).

Les analystes principaux de l'Alliance, Shahira Khair (analyste principale en gestion des données de recherche) et Qian Zhang (analyste principale en logiciels de recherche) ont également aidé le groupe de travail sur le CIP dans ce processus. Les membres se sont réunis une fois par semaine, en plus de fournir des conseils à l'Alliance dans le cadre du processus d'évaluation des besoins.

Les résultats historiques détaillés du Concours pour l'allocation de ressources (CAR) et d'autres données historiques sur l'utilisation des ressources internes, que Calcul Canada a généreusement fournies, constituent les principales sources de données qui ont servi au rapport. Comme le temps et les ressources étaient limités, le groupe de travail sur le CIP n'a pas pu mener de nouvelles recherches ou sondages pour appuyer ses conclusions. Certains éléments de cette recherche seront menés dans le cadre des projets d'évaluation des besoins et d'analyse de l'environnement de l'Alliance durant le premier semestre de 2021.

Les conclusions et les observations de ce document, ainsi que les publications sur la gestion des données de recherche et l'évaluation de l'état actuel des logiciels de recherche, ont pour but d'aider les analystes, la direction, le conseil d'administration et le Conseil des chercheurs de l'Alliance l'évaluation des besoins, afin de définir un nouveau modèle de prestation de services,

un plan stratégique et des modèles de financement pour l'écosystème canadien de l'IRN qui s'étendra jusqu'en 2024.

3 Introduction

3.1 Qu'est-ce que le calcul informatique de pointe ?

Dans un rapport en 2017, le Conseil du leadership sur l'infrastructure de recherche numérique (CLIRN) définit le CIP comme suit : « Le CIP offre aux chercheurs une technologie et une expertise numériques pour les aider à résoudre des problèmes de recherche qui sont trop importants ou complexes pour qu'ils puissent les entreprendre seuls. Il s'agit des services, des conseils, du matériel et des logiciels que gère le personnel hautement qualifié (PHQ) pour les activités de recherche qui ont d'importants besoins en matière de calcul, d'acquisition de données, de simulation, d'expérimentation, d'analyse et d'exploration ».[1] En plus des grappes de serveurs traditionnels, le CIP moderne comprend également des infrastructures virtuelles. Le CIP doit s'adapter à la recherche dans un environnement numérique, tout comme l'enseignement supérieur s'équipe physiquement d'infrastructures et d'installations traditionnelles pour ses activités de recherche.

Westgrid, qui est affilié à la Fédération Calcul Canada (FCC), définit le CIP comme suit : « Le calcul informatique de pointe comprend tout ce qui va au-delà d'un poste de travail traditionnel, notamment l'informatique dématérialisée, les superordinateurs, le calcul de haute performance (HCP), la gestion et le stockage des données pour la recherche ».[2] En 2019, un autre affilié de la FCC, Calcul Ontario, a défini le CIP comme un prolongement du HCP et de la recherche effectuée sur les superordinateurs. « La recherche moderne, dans pratiquement tous les domaines, implique souvent un travail de calcul important qui ne nécessite pas forcément des superordinateurs et des codes massivement parallèles. Les décideurs politiques du Canada ont adopté d'expression « calcul informatique de pointe » (CIP) pour désigner la gamme complète des besoins informatiques des chercheurs, tout en utilisant le terme « calcul de haute performance » (HCP) pour désigner le sous-ensemble de ces besoins informatiques auxquels un superordinateur ne peut pas répondre ».[3] Dans le présent rapport, nous utilisons cette définition canadienne qui repose sur la distinction entre le CIP et le HCP.

Comparativement à l'informatique d'entreprise, le CIP accorde une grande importance au rendement et à l'agilité. La recherche (méthodes, techniques, logiciels) et la technologie (y compris la capacité et la rentabilité) évoluent toutes deux sur des échelles de temps très rapides (deux ans peuvent être considérés comme une longue période dans certains cas). Pour que le CIP soit concurrentiel et pertinent pour la recherche de pointe, il doit rester à jour sur tous les fronts. L'informatique d'entreprise est quant à elle axée sur la stabilité et la fiabilité, ce qui est

[1] Rapport sur le calcul informatique pointe (CIP), CLIRN (août 2017).
[2] Westgrid: What We Do https://www.westgrid.ca/about_westgrid/what_we_do (consulté en novembre 2020).
[3] Calcul Ontario : Thinking Forward Through the Past:A Brief History of Supercomputing in Canada and its Emerging Future https://computeontario.ca/wp-content/uploads/2019/07/A-Brief-History-of-Supercomputing-in-Canada-and-its-Emerging-Future.pdf (juin 2019).

essentiel pour les systèmes de paie, de courrier électronique et d'inscription des étudiants. En effet, ces systèmes et technologies évoluent sur des échelles de temps plus longues. Généralement, on ne voit pas une très grande différence si un système d'entreprise traite soudainement les salaires 50 % plus rapidement, mais si les systèmes de CIP fonctionnent tout à coup 50 % plus vite, la recherche s'accélère considérablement, pour ainsi répondre à la demande énorme et croissante.

Par ailleurs, il y a une importante différence entre le CIP et les services informatiques ordinaires d'un établissement. Ces derniers sont prévus pour un environnement de production nécessitant des ententes de niveau de service (ENS) strictes, tandis que l'infrastructure informatique de recherche vise généralement à obtenir le meilleur rendement possible avec l'argent disponible. Elle repose parfois sur des technologies de pointe qui ne sont pas toujours aussi fiables que les services informatiques d'entreprise. Cette infrastructure met l'accent sur la flexibilité dans la prestation et la configuration de services, ce qui n'exige donc pas une ENS aussi contraignante. Le CIP se caractérise également par l'interconnexion et l'interaction de plusieurs technologies de pointe qui permettent ensemble de produire les résultats attendus, dans un environnement partagé entre plusieurs utilisateurs qui fonctionne à très haute capacité.

En fonction du milieu et de l'administration, le CIP est connu sous divers noms à consonance légèrement différente : la cyberinfrastructure est aux États-Unis un synonyme de CIP, notamment dans le contexte de l'infrastructure en réseau, du calcul de haute performance pour les systèmes de CIP (excluant souvent les systèmes constitués de postes de travail ou de serveurs) et les opérations de plus haut niveau, ainsi que le calcul intensif pour les systèmes de CIP très haut de gamme (dont les superordinateurs Bluegene d'IBM et Cray qui sont équipés matériel, E/S et systèmes de communication sur mesure). Sauf au Canada où le HCP est considéré comme un sous-ensemble du CIP, le calcul informatique de pointe se distingue à l'échelle internationale du calcul de haute performance en mettant l'accent sur des applications scientifiques et la recherche plutôt que sur le rendement. En revanche, le HCP concerne davantage la performance et inclut parfois des environnements de production, notamment dans le secteur commercial et des agences de sécurité. En ce qui concerne les entreprises et le marché, Intersect360 Research (société de conseil et d'études de marché spécialisée dans le calcul de haute performance) définit le HCP comme ayant un champ d'application plus vaste que le CIP. Au-delà de l'informatique de recherche, cette définition inclut également l'informatique de production à grande échelle. Elle caractérise le calcul de haute performance comme étant « l'utilisation de serveurs, de grappes et de superordinateurs, ainsi que de logiciels, d'outils, de composants, de stockage et de services associés, pour des tâches scientifiques, techniques ou analytiques particulièrement intensives pour le calcul, l'utilisation de la mémoire ou la gestion des données. Le calcul intensif sert aux scientifiques et aux ingénieurs, autant pour la recherche, la production, l'industrie, l'administration et le monde universitaire ». [4]

En plus des considérations générales précédentes, le CLIRN caractérise le CIP en six fonctions essentielles, [1] que nous vous présentons ci-dessous avec des commentaires additionnels (*en italique*) : [1]

- Calcul : utilisation de ressources de calcul comme des cœurs, des processeurs graphiques *ou d'autres accélérateurs*, ainsi que de la mémoire.

[4] Recherche Intersect360 - https://www.intersect360.com/what-is-hpc (consulté en septembre 2020).

- Stockage actif : données utilisées régulièrement pendant la durée du projet (contrairement à l'archivage pour la conservation à long terme).

- Conseils, soutien et formation : leadership de Calcul Canada et des consortiums régionaux au nom de la communauté du CIP ; expertise du personnel pour aider les chercheurs, *en plus de renforcer et créer de nouvelles approches numériques dans la recherche.*

- Gestion et soutien des logiciels : entretien, maintenance, évolution et développement de logiciels utiles à de nombreux chercheurs et projets au niveau des systèmes et des applications. *En particulier, les logiciels de recherche utilisés et développés par les chercheurs sont considérés comme relevant du pilier/cadre des logiciels de recherche*.

- Confidentialité, sécurité et authentification : exigences strictes, détaillées ou inhabituelles, au-delà de celles fournies par défaut.

- Soutien et coordination avec d'autres composants de l'IRN (réseau, gestion des données, stockage, logiciels) : les fournisseurs de CIP doivent coordonner efficacement les autres composants de l'écosystème de l'IRN, *y compris les éventuelles initiatives de science ouverte*.

Ce rapport inclut donc la catégorisation fonctionnelle légèrement mise à jour ci-dessus dans la définition du CIP, tout en reconnaissant que le mandat futur de l'Alliance ne sera pas aussi limité. Autrement dit, cette définition fonctionnelle n'inclut pas le matériel, les systèmes et les services de stockage à moyen terme (dépôt) et à long terme (archives), qui font partie intégrante du cycle informatique de la recherche moderne.

Il est intéressant de noter que dans certaines disciplines, le CIP est davantage axé sur la fonctionnalité que sur la technologie comme ci-dessus. Par exemple, le projet de l'Union européenne CLARIN (*Common Language Resources and Technology Infrastructure*) en sciences humaines et sociales se concentre sur les outils et le processus de recherche : « CLARIN ERIC est né en 2012 dans le but de créer et de maintenir une infrastructure pour le partage, l'utilisation et la durabilité des données et des outils linguistiques pour la recherche en sciences humaines et sociales. À l'heure actuelle, CLARIN offre un accès simple et durable aux données linguistiques numériques (sous forme écrite, orale ou multimodale) pour les chercheurs en sciences sociales, humaines et autres. CLARIN offre également des outils avancés pour découvrir, explorer, exploiter, annoter, analyser ou combiner ces ensembles de données, où qu'ils se trouvent ».[5] Alors que CLARIN est axé sur les ressources linguistiques, DARIAH est une infrastructure paneuropéenne pour les arts et les sciences humaines. Les opérations de cette infrastructure reposent sur quatre centres de compétences virtuels (CCV) et des groupes de travail. Chaque CCV est doté d'un mandat et de priorités. Par exemple, le CCV1 se concentre sur les fondations technologiques et l'infrastructure électronique de CLARIN. [6]

[5] CLARIN: CLARIN in a nutshell https://www.clarin.eu/content/clarin-in-a-nutshell (consulté en décembre 2020).
[6] DARIAH-EU: DARIAH in a nutshell https://www.dariah.eu/about/dariah-in-nutshell/ (consulté en février 2021).

3.2 Qui utilise les systèmes de HPC et CIP à l'échelle mondiale ?

Selon l'étude de marché d'Intersect360 Research, le calcul de haute performance représentait mondialement un marché d'environ 39 milliards de dollars en 2019, avec une croissance d'environ 8,2 % par an par rapport à 2018. Les principaux marchés verticaux étaient le milieu universitaire (17,1 %), le gouvernement (25,4 %) et l'industrie (57,5 %). [7] Dans la catégorie « gouvernement », les principaux utilisateurs de HPC étaient la sécurité nationale (12 %), les laboratoires de recherche nationaux (10 %) et les agences nationales (3 %). Du côté de l'industrie, les principaux groupes d'utilisateurs étaient répartis de manière assez égale, les secteurs les plus importants étant les services financiers (13 %), la fabrication de gros produits (8 %), les biosciences (8 %), l'énergie (5 %), la fabrication de produits de consommation (5 %) et le commerce de détail (5 %). D'une part, la grande majorité des investissements dans le calcul intensif est donc motivée par une consommation non universitaire, de sorte que les solutions proposées par le marché ne servent pas principalement à répondre aux besoins de la recherche, mais plutôt à l'utilisation commerciale et la sécurité nationale. Par exemple, les laboratoires de recherche nationaux (10 %), les biosciences (8 %), l'énergie (5 %) et l'ingénierie chimique (4 %) représentent 27 % du marché, ce qui, ajouté au secteur universitaire (17 %), donne un total de 44 % du marché axé sur les besoins « traditionnels » en matière d'informatique de recherche et de CIP.

Les deux nouveaux changements principaux en 2019 par rapport aux années précédentes concernent la croissance du gouvernement par rapport à l'industrie, ainsi qu'une croissance majeure dans les déploiements d'infonuagique et de systèmes semblables. Les quatre principales catégories de dépenses étaient les serveurs (environ 14 milliards de dollars), les logiciels (environ 9 milliards de dollars), le stockage (environ 6 milliards de dollars) et les services (environ 4 milliards de dollars), dans cet ordre. Soulignons que les dépenses liées aux solutions d'infonuagique ont connu une forte croissance, mais sont restées inférieures à environ 2 milliards de dollars dans l'ensemble. [8]

3.3 Qui participe à la mise en œuvre du CIP pour les chercheurs canadiens ?

Prestation de services de CIP à l'échelle locale, régionale et nationale au Canada

Au Canada, les services de CIP reposent sur un réseau non centralisé d'organisations locales, régionales et nationales, ce qui n'a pas beaucoup changé depuis le rapport du CLIRN en 2017 sur le CIP.

Au niveau local, l'écosystème canadien de l'IRN est dynamique, les universités offrant divers services et soutiens informatiques pour la recherche, que ce soit dans le cadre de leurs opérations informatiques centrales, de leurs bibliothèques ou d'opérations informatiques de recherche indépendante au niveau universitaire. La croissance locale bénéficie d'un financement national,

[7] Intersect360 Research - Worldwide HPC Market 2019 Actuals, 2020-24 Forecast, Including Effects of COVID-19 https://www.intersect360.com/presentations/Intersect360%20WW%20HPC%202019%20market%20and%202020-24%20forecast.pdf (septembre 2020).

[8] Intersect360 Research - Worldwide HPC Market 2019 Actuals, 2020-24 Forecast, Including Effects of COVID-19 https://www.intersect360.com/presentations/Intersect360%20WW%20HPC%202019%20market%20and%202020-24%20forecast.pdf (septembre 2020).

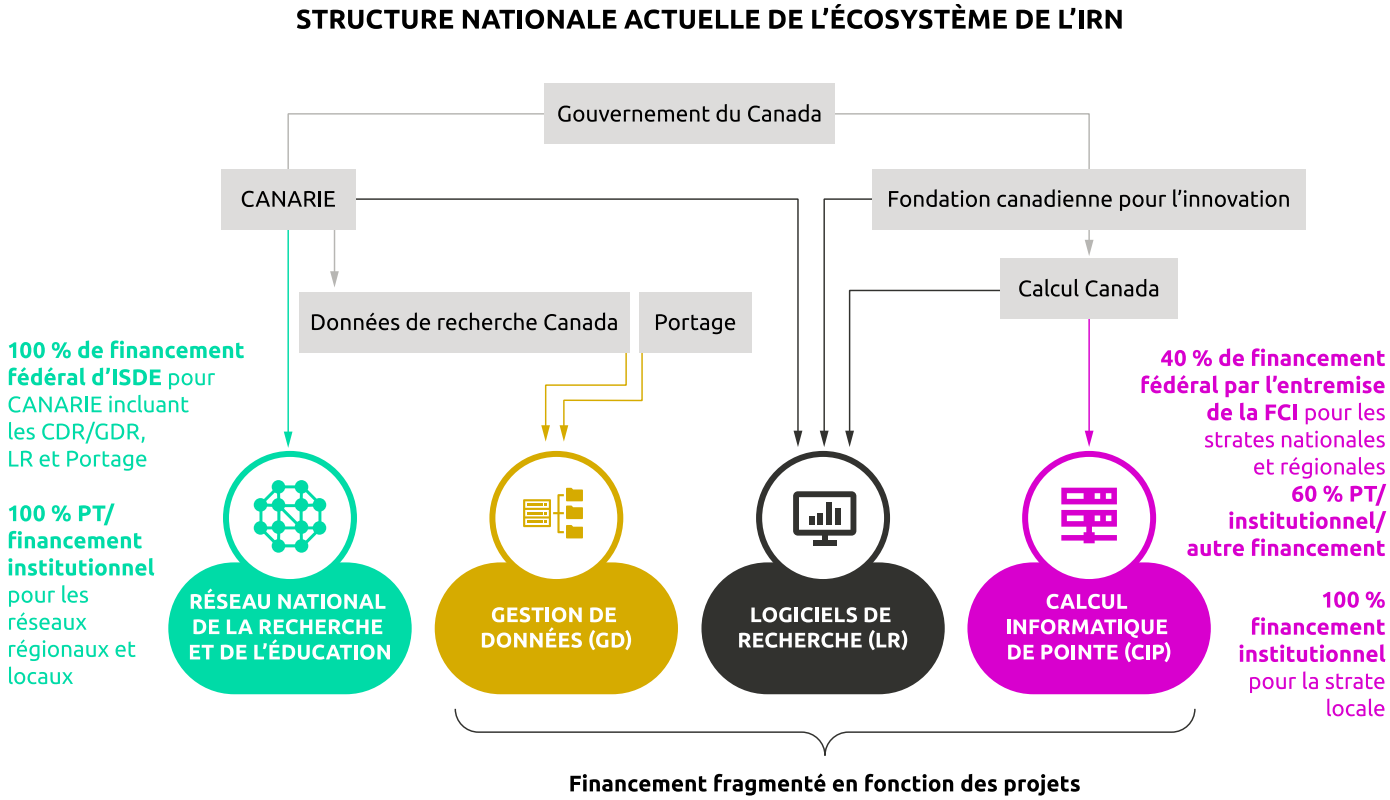
par exemple les initiatives de CANARIE en matière de logiciels de recherche et de gestion des données. En revanche, les services sont toujours fragmentés et variés en ce qui concerne l'infrastructure de recherche numérique (INR) dans le milieu universitaire canadien. Certains établissements mènent d'importantes opérations et offrent beaucoup de soutien pour l'IRN, tandis que d'autres en offrent très peu dans ce domaine.

Au niveau régional, WestGrid, Calcul Ontario, Calcul Québec et ACENET coordonnent la prestation de services d'IRN. L'important financement de la FCI pour actualiser l'infrastructure de CIP au Canada a consolidé davantage le rôle des cinq principaux sites de la Fédération de Calcul Canada (FCC) : Université de Victoria, Université Simon Fraser, Université de Waterloo, Université de Toronto et Université McGill. Il est important de noter que trois des cinq sites (SFU, UW et McGill) sont gérés par des équipes distribuées qui incluent des membres externes à l'établissement. Par ailleurs, plusieurs tâches nécessaires au fonctionnement de ces infrastructures (soutien aux utilisateurs, documentation, installation de logiciels, surveillance, programmation) sont prises en charge par des équipes nationales composées de personnes issues d'établissements qui embauchent du personnel de la FCC. Le consortium régional ACENET n'a pas de site principal d'hébergement de la FCC pour CIP, mais il exploite la grappe à haute performance Siku à l'Université Memorial.

Au niveau national, le lancement de l'Alliance de recherche numérique du Canada (l'Alliance) en 2019 constitue le principal changement. ISDE a confié à l'Alliance le mandat de coordonner et de financer l'écosystème de l'IRN émergent du Canada, incluant non seulement le CIP, mais aussi les logiciels de recherche (LR) et la gestion des données (GD).[9] Grâce à cette approche, le financement sera beaucoup plus centralisé, cohérent et prévisible pour l'ensemble de l'écosystème de l'IRN. Dans le cadre de ce nouveau mandat, l'Alliance prendra en charge les activités de Calcul Canada à compter du 1er avril 2022. En ce qui concerne la gestion des données, l'Alliance est déjà responsable des activités et du financement de l'ABRC Portage. Elle deviendra responsable de la gestion des données de CANARIE et des activités de Données de recherche Canada (DRC) d'ici avril 2022. Pour la gestion des données et les logiciels de recherche, l'Alliance intégrera les équipes, les politiques, les procédures et les principales initiatives de CANARIE et de Calcul Canada.

[9] Alliance : Historique https://engagedri.ca/about-engage-dri/background (consulté en décembre 2020).

Plan général des mécanismes actuels financés par le gouvernement au Canada pour l'IRN



Innovation, Sciences, et Développement économique Canada, 2020. Ajustements de l'Alliance de recherche numérique du Canada (l'Alliance).

Graphique 1 Structure nationale actuelle de l'écosystème de l'IRN

La structure fondamentale du financement de l'IRN au Canada n'a pas changé depuis 2017. ISDE, ministère du gouvernement canadien, est toujours le principal bailleur de fonds de l'écosystème et des activités de l'IRN Canada, par l'entremise du Fonds d'innovation (FI), du Fonds des initiatives scientifiques majeures (FISM) de la FCI et de CANARIE. Le financement des systèmes de CIP est relativement bien établi, puisque la FCI verse directement le financement du FI et du FISM aux sites d'hébergement individuels et aux universités sous la direction de Calcul Canada. La formule de financement en contrepartie est de 40/60 : 40 % du financement provient de la FCI et 60 % proviennent de sources comme les provinces, universités et dons en nature. Habituellement, les infrastructures locales de CIP ne sont pas admissibles au financement de la FCI, mais certaines exceptions peuvent être faites pour des besoins en temps réel, de données sensibles ou de calcul de pointe, par exemple. En général, l'infrastructure de CIP doit se trouver sur les principaux sites de la FCC et contribuer à l'infrastructure existante. Les divers modèles de financement compliquent en quelque sorte les choses, notamment en ce qui concerne la différence entre le financement en capital (FI) et le financement des coûts d'exploitation (FISM). De plus, cette distinction est moins pertinente avec l'émergence de l'informatique dématérialisée, qui déplace les dépenses en capital vers des dépenses de fonctionnement.

Parmi ses futurs mandats, l'Alliance devra regrouper les trois principales composantes de l'IRN (CIP, LR et GD) sous un même cadre de financement et de planification, en intégrant possiblement des modèles de financement nouveaux ou améliorés. À cette fin, l'Alliance a obtenu un nouveau financement quinquennal important pour l'IRN, qui totalise 375 millions de dollars jusqu'en mars 2024.

Organisations affiliées à la Fédération Calcul Canada

La Fédération Calcul Canada est composée de quatre organisations partenaires régionales : ACENET, Calcul Québec, Calcul Ontario et WestGrid. Voici les principales caractéristiques de ces organisations :

ACENET

Créé en 2003, ACENET est l'organisation chargée de l'infrastructure de recherche numérique du Canada atlantique. Il s'agit d'un partenariat entre 14 universités et collèges de la région représentant presque tous les établissements d'enseignement postsecondaire. Cette organisation entrevoit l'infrastructure de calcul informatique de pointe (CIP), le soutien technique et le développement des compétences numériques pour plus de 1000 chercheurs universitaires, étudiants postsecondaires et chercheurs industriels des Maritimes.

L'infrastructure régionale du CIP se trouve à l'Université Memorial, qui détient une désignation spéciale de la FCI lui permettant d'intégrer des systèmes fournis par les chercheurs pour des groupes de recherche individuels. En échange de leur exploitation, de leur maintenance et de leur gestion par ACENET, les cycles de calcul excédentaires sont reversés dans l'ensemble de ressources partagées, ce qui permet de maximiser les ressources et le financement du CIP. L'infrastructure d'ACENET compte actuellement quatre systèmes contribués par des chercheurs, tandis que cinq autres sont en cours d'acquisition.

ACENET est une organisation distribuée qui compte 20 employés. Elle est dirigée par un conseil d'administration composé de vice-présidents à la recherche (ou de leurs représentants), qui sont issus des six établissements hôtes : Université Dalhousie, Université Memorial, Université Saint Francis Xavier, Université Saint Mary's, Université du Nouveau-Brunswick et Université de l'Île-du-Prince-Édouard. L'équipe dirigeante d'ACENET s'appuie sur les conseils d'une direction de la recherche qui inclut 10 chercheurs actifs et interdisciplinaires provenant de plusieurs établissements membres du Canada atlantique. La FCI, l'APECA et les quatre provinces de l'Atlantique assurent le financement d'ACENET. [10]

Calcul Québec

Calcul Québec est le partenaire régional de Calcul Canada pour cette province. Ce consortium est une société constituée sans but lucratif composé de onze universités québécoises. Les universités membres ont mis en commun leurs investissements et leurs ressources locales en CIP pour former la coalition. Plus de 550 groupes de recherche et environ 1925 utilisateurs profitent de ces ressources.[11] Le financement de Calcul Québec provient de la province et de la FCI. La connectivité du réseau est assurée par le Réseau d'informations scientifiques du Québec

[10] Ines Hessler, directrice technique d'Acenet, communication privée (juillet 2021).
[11] Calcul Québec : Qui sommes-nous ? https://www.calculquebec.ca/en/about-us/who-are-we/ (consulté en novembre 2020).

(RISQ) et de CANARIE.[12] Calcul Québec emploie plus de 40 personnes hautement qualifiées et son modèle de gouvernance bien établi repose sur un conseil d'administration de 13 membres, un conseil scientifique de 10 membres, un conseil d'exploitation, ainsi qu'un conseil de développement technologique, d'exploitation et de soutien à la recherche. [13]

Calcul Québec héberge Béluga, qui est l'un des principaux systèmes nationaux de la FCC. Il s'agit d'une grappe de CIP polyvalente qui appartient à l'Université McGill. Le système se trouve à l'École des technologies supérieures (ETS) et il est exploité par une équipe composée de membres du consortium. Depuis janvier 2021, Calcul Québec exploite également Helios[14], un superordinateur de génération antérieure, qui est à l'Université Laval. L'organisation exploite aussi MP2, qui se trouve à l'Université de Sherbrooke.

Calcul Ontario

Calcul Ontario a été constitué en 2014 en tant que société à but non lucratif, avec pour mandat de coordonner le CIP en Ontario. [15] Cette organisation repose sur deux décennies de travail dans la province et les consortiums de calcul de haute performance en Ontario. Calcul Ontario est financé par le ministère des Collèges et Universités (MCU) de l'Ontario et la connectivité du réseau régional est assurée par ORION en collaboration avec CANARIE.

Les consortiums partenaires de Calcul Ontario sont les suivants : Centre for Advanced Computing, SHARCNET, SciNet et HPC4Health.[16] Calcul Ontario travaille avec les consortiums pour centraliser la stratégie et la planification des actifs d'informatique de pointe, notamment le matériel, les logiciels, la gestion des données, le stockage, la sécurité, la connectivité et le personnel hautement qualifié.

SHARCNET est un consortium composé de 18 collèges, universités (dont Université Western est l'établissement principal) et instituts de recherche qui exploitent un réseau de groupes informatiques à haute performance dans le sud-ouest, le centre et le nord de l'Ontario. SHARCNET héberge Graham, un superordinateur hétérogène polyvalent situé à l'Université de Waterloo, l'un des principaux sites nationaux de la FCC.

L'Université de Toronto dirige le consortium SciNet, qui héberge Niagara, un superordinateur homogène massivement parallèle qui appartient à cet établissement universitaire et se trouve sur l'un des principaux sites nationaux de la FCC.

Westgrid

WestGrid est une coalition de sept établissements membres de la Colombie-Britannique, de l'Alberta, de la Saskatchewan et du Manitoba. Ces derniers reçoivent un financement pour répondre aux besoins de WestGrid en matière d'exploitation et d'entretien par l'entremise du

[12] Calcul Québec : Partenaires https://www.calculquebec.ca/en/about-us/partenaires/ (consulté en novembre 2020).
[13] Calcul Québec : Gouvernance https://www.calculquebec.ca/a-propos/gouvernance/ (consulté en novembre 2020).
[14] Calcul Canada : Hélios https://docs.computecanada.ca/wiki/H%C3%A9lios/en (consulté en novembre 2020).
[15] Calcul Ontario : À propos de Calcul Ontario https://computeontario.ca/about-compute-ontario/ (consulté en novembre 2020).
[16] Calcul Ontario : Partenaires https://computeontario.ca/partners/ (consulté en novembre 2020).

FISM de la FCI. Les partenaires provinciaux et les établissements universitaires versent un financement en contrepartie cette subvention. [17]

Les installations de WestGrid sont reliées par un réseau central de WestGrid, qui utilise l'infrastructure du réseau national CANARIE et du Réseau national de la recherche et de l'éducation (Colombie-Britannique/BCNET/Alberta/Cybera/Saskatchewan/SRNET et Manitoba/MRnet).[18] L'Université Simon Fraser, affiliée à WestGrid, héberge et gère un superordinateur hétérogène à usage général, Cedar, qui est situé à la SFU, l'un des principaux sites nationaux de la FCC. En plus des importants systèmes régionaux que gère WestGrid, l'Université de Victoria héberge Arbutus, un système de CIP en infonuagique polyvalent qui fait partie des ressources de la FCC. [170]

Établissements universitaires non membres de la FCC

L'écosystème canadien de l'IRN est très dynamique, avec des organisations qui évoluent et innovent constamment. Elles lancent régulièrement de nouveaux projets avec divers champs d'application et offres de services. La section suivante présente un **échantillon non exhaustif** d'établissements canadiens qui ne sont pas membres de la FCC, mais classés dans les catégories « infrastructure », « services et plateformes » et « recherche et formation ». Ces catégories ne s'excluent pas forcément mutuellement, puisque certains grands établissements offrent des services dans toutes les catégories. De plus, certaines activités ci-dessous sont offertes par des établissements affiliés à la FCC, mais qui ne s'inscrivent pas dans l'offre nationale principale de la FCC, accessible au public et située sur l'un des cinq principaux sites d'hébergement.

Fournisseurs d'infrastructures

Les fournisseurs exploitent des infrastructures physiques d'IRN et offrent souvent en plus des services, plateformes ou formations en matière d'IRN. D'autres mènent de la recherche.

ARC UBC

Advanced Research Computing (ARC) de l'Université de la Colombie-Britannique (UBC) est une importante infrastructure de CIP qui ne se rattache pas à la FCC, mais appartient à UBC. Elle comprend la grappe de calcul Sockeye et le système de stockage d'objet Chinook. Le financement initial de Sockeye de 7,9 millions de dollars s'est matérialisé en 2018. Au début de l'année 2020, un investissement supplémentaire de 10,1 millions de dollars a permis d'améliorer le système.19 À l'heure actuelle, il compte plus de 16 000 cœurs de CPU, 200 GPU, une interconnexion EDR InfiniBand 20et jusqu'à 20 Po de stockage, distribués entre les campus de Vancouver et d'Okanogan. Le système dispose également de 192 To de stockage flash pour les besoins d'E/S temporaires rapides. Si l'on considère uniquement le nombre de cœurs d'unités

[17] WestGrid: What We Do https://www.westgrid.ca/about_westgrid/what_we_do (consulté en novembre 2020).
[18] WestGrid: Partners https://www.westgrid.ca/about_westgrid/members-partners (consulté en novembre 2020).
[19] UBC ARC: Enhancing Support for Advanced Research Computing https://arc.ubc.ca/enhancing-support-advanced-research-computing (consulté en janvier 2021).
[20] UBC ARC: Sockeye – Detailed Technical Specification https://arc.ubc.ca/sockeye-techspecs (consulté en janvier 2021).

centrales, le système fait environ la moitié de la taille de la grappe de calcul polyvalente Beluga de la FCC, qui se trouve à Montréal. Les services de CIP d'UBC sont accessibles aux chercheurs qui sont membres du corps enseignant de l'établissement et aux chercheurs principaux d'UBC, notamment pour aider les nouveaux professeurs et les chercheurs doivent héberger leurs données localement et non sur les serveurs de stockage partagés de la FCC. [21]

CAC

L'Université Queen's dirige le Centre for Advanced Computing (CAC), qui compte aussi parmi ses membres l'Université Carleton, l'Université d'Ottawa et le Collège militaire royal du Canada. Le CAC se spécialise dans les ressources informatiques de pointe sécurisées, en plus de répondre aux besoins chercheurs universitaires et médicaux. Le CAC soutient plus de 400 groupes de recherche canadiens totalisant quelque 2100 chercheurs qui travaillent dans divers domaines.[22] Le CAC n'est pas l'un des principaux sites nationaux de la FCC, mais les plateformes Frontenac et Katarokwi du centre sont accessibles aux chercheurs affiliés à la fédération. En revanche, l'accès n'est pas gratuit comme pour les principaux systèmes de la FCC.

RCCDR

Le Réseau canadien des centres de données de recherche (RCCDR) est un partenariat entre un consortium d'universités canadiennes et Statistique Canada, dont le siège est à l'Université McMaster. Par le biais de son Centre de données de recherche (CDR), le RCCDR a pour mission de mettre à la disposition des chercheurs des microdonnées confidentielles sociales, économiques et de santé. À l'heure actuelle, on y accède dans les bureaux sécurisés avec des postes de travail et des serveurs strictement contrôlés, situés sur des campus universitaires canadiens. Le financement de base du RCCDR provient d'une combinaison de subventions dirigées par le CRSH et les IRSC et du Fonds des innovations scientifiques majeures (FISM) de la FCI. Le RCCDR bénéficie également d'un soutien en nature et financier des universités hôtes et de Statistique Canada.[23] En 2017, le réseau a obtenu un financement important par le biais du Fonds d'innovation de la FCI pour créer une infrastructure nationale de HPC centralisée afin de répondre aux besoins croissants prévus en matière de traitement et de stockage des données dans le réseau. [24]Plus récemment, les spécifications de conception de la plateforme HPC/centre de données de recherche virtuel (vRDC) se sont élargies pour inclure la capacité d'accès à distance, c'est-à-dire l'accès aux ressources centrales depuis l'extérieur des espaces de bureaux sécurisés du CDR. [25]Cela implique une conception et un examen détaillés de la sécurité du système afin de répondre aux exigences strictes en matière de cybersécurité qui s'appliquent aux fichiers de microdonnées Protégé B [26](telles que définies par le Conseil du Trésor du

[21] UBC ARC Sockeye https://arc.ubc.ca/ubc-arc-sockeye (consulté en janvier 2021).
[22] Centre for Advanced Computing: What is CAC ? https://cac.queensu.ca/about_us/ (consulté en novembre 2020).
[23] CRDCN: About the CRDCN https://crdcn.org/about-crdcn (consulté en novembre 2020).
[24] McMaster University : McMaster Data Center receives 2.7 millions dollars to support research infrastructure development for economic, social & health data https://www.economics.mcmaster.ca/news/mcmaster-data-center-recieves-2-7m-for-study-of-economic-social-and-health-data (octobre 2017).
[25] CRDCN 2019-24 Strategic Plan https://crdcn.org/sites/default/files/strategic_plan_0.pdf (juin 2019).
[26] CRDCN September 2020 Newsletter https://us4.campaign-archive.com/?u=c3b811df1cf083f6ae6fb612b&id=5a4556d80e (consulté en novembre 2020).

gouvernement du Canada, le Centre de la sécurité des télécommunications et les exigences en matière de sécurité des TI et de microdonnées de Statistique Canada).

CYBERA

Cybera est une filiale sans but lucratif de CANARIE qui se trouve en Alberta. Elle est chargée d'exploiter le CYBERANET, qui fait partie du Réseau national de recherche et d'éducation (RNRE) du Canada.[27] Elle fournit également une infrastructure et des services de CIP au programme DAIR (Digital Accelerator for Innovation and Research) de CANARIE, qui permet aux petites et moyennes entreprises (PME) d'accéder aux bancs d'essai et à la technologie en nuage. Elle héberge et fournit également l'accès à Rapid Access Cloud, un service dématérialisé gratuit pour les universitaires et les PME d'Alberta qui ne sont pas admissibles au programme DAIR. CYBERA collabore également avec le Pacific Institute for Mathematical Sciences (PIMS) pour fournir aux chercheurs canadiens la très populaire passerelle scientifique Syzygy Jupyter. Le chiffre d'affaires de Cybera pour l'exercice 2018-19 était d'environ 5,9 millions de dollars et son personnel comptait 39 personnes. [28]

HPC4Health

HPC4Health est un consortium regroupant SickKids et University Health Network dans la région de Toronto, qui construit un moteur de calcul en infonuagique sécurisé pour la recherche clinique. Les services sont principalement accessibles aux membres de ces deux établissements, mais les organisations extérieures peuvent accéder aux services en payant des frais pour le recouvrement de coûts.[29] HPC4Health héberge une infrastructure dématérialisée sur l'ensemble de logiciels OpenStack de 7000 CPU afin que chaque établissement de soins de santé puisse accéder à son propre nuage entièrement privé, tout en bénéficiant des avantages de la mise en commun de ressources. Chaque établissement participant se voit garantir une quantité minimale de cœurs de CPU lorsqu'il en a besoin (équivalent à 80 % de sa contribution), ce qui permet de répondre aux besoins informatiques essentiels et urgents. Les 20 % restants sont partagés, ce qui permet à tous les utilisateurs de tirer parti de la capacité sous-utilisée. [30]

CNRC

Le Conseil national de recherches du Canada (CNRC) est le plus grand organisme fédéral de recherche et de développement du Canada. Il gère à la fois ses propres activités de recherche, en plus de collaborer avec des universités et des établissements de recherche canadiens. Le CNRC finance aussi les activités de recherche, les petites et moyennes entreprises (PME) et les industries au Canada. Au cours de l'exercice 2019, le CNRC comptait 4109 ETP et affichait des revenus de 184 millions de dollars pour des dépenses totales de 1214 millions de dollars.[31] Les équipes du CNRC exploitent et tirent parti de nombreuses ressources de CIP, à l'interne et en collaboration avec Services partagés Canada. En revanche, même une organisation de l'envergure du CNRC n'est pas toujours en mesure de répondre à la demande pour des services

---

[27] CYBERA: Services https://www.cybera.ca/services/ (consulté en janvier 2021).
[28] CYBERA: 2018-2019 Annual Report https://www.cybera.ca/wp-content/uploads/2020/03/Cybera_Annual_Report_2018-19.pdf (octobre 2019).
[29] HPC4Health: Accessing Our Services http://www.hpc4health.ca/services.html (consulté en novembre 2020).
[30] HPC4Health: Overview http://www.hpc4health.ca/overview.html (consulté en novembre 2020).
[31] Conseil national de recherches du Canada : Rapport annuel 2019-20 https://nrc.canada.ca/sites/default/files/2020-08/annual-report-2019-2020.pdf (août 2020).

de CIP. En 2018 par exemple, la capacité de CIP du Centre de recherche sur les technologies de sécurité et de rupture du CNRC était insuffisante et désuète. Il prévoit donc de travailler éventuellement avec Calcul Canada pour des améliorations.[32] Malheureusement, l'information publique sur les activités de CIP du CNRC n'est pas facilement accessible.

Ouranos

Ouranos est un organisme à but non lucratif qui se trouve à Montréal, Québec. Il se spécialise dans les changements climatiques et leurs impacts, ainsi que les vulnérabilités socio-économiques et environnementales pertinentes, afin d'orienter les politiques et les stratégies d'adaptation. Ouranos emploie plus de 50 personnes, participe à 13 programmes scientifiques et plus de 100 projets. Les principaux membres d'Ouranos sont la province du Québec, Hydro-Québec, l'UQAM, l'Université McGill, l'Université Laval et Environnement et Changement climatique Canada. Les revenus annuels se situent entre 8 et 12 millions de dollars et le financement provient de diverses sources, dont la province du Québec. [33]En 2015, Ouranos disposait de trois superordinateurs Cray et collaborait avec Calcul Québec pour les ressources de CIP. [34]

SciNet4Health

En septembre 2020, l'Université de Toronto et SciNet ont annoncé[35] une nouvelle initiative, SciNet4Health, qui « permettra aux chercheurs et aux cliniciens scientifiques de l'Université de Toronto et de ses hôpitaux partenaires d'accéder à des bases de données massives d'informations sur la santé des patients et de les analyser (d'une manière sécurisée qui protège la vie privée des patients) en utilisant des technologies telles que l'apprentissage machine ». Le système central a un rendement maximal théorique d'un pétaFlop et il comprend 20 nœuds de calcul, chacun équipé de processeurs AMD EPYC et de huit accélérateurs GPU AMD Radeon Instinct, offerts gratuitement par AMD. Le système sera hébergé dans le centre de données principal de SciNet, à côté du superordinateur Niagara. L'initiative repose sur l'expérience de HPC4Health en matière de procédures et de protocoles. Les deux organisations prévoient de travailler ensemble offrir des services de CIP pour la prestation de soins de santé dans la région de Toronto.

Siku

Acenet héberge Siku, une grappe informatique de haute performance qui compte 2300 cœurs d'unité centrale. Le système est fonctionnel depuis 2019 et il se trouve à l'Université Memorial à St. John's, Terre-Neuve. Il ne s'inscrit pas dans les systèmes nationaux de la FCC et ne se trouve pas sur l'un des sites principaux de la fédération. Siku est financé en grande partie par l'APECA dans le but de générer des avantages économiques régionaux en faisant participer les industries locales, tout en soutenant la recherche universitaire dans la région de l'Atlantique. Le système est uniquement accessible à certains clients, notamment des chercheurs industriels et des

[32] Bureau de la vérification et de l'évaluation du CRNC : Évaluation du Centre de recherche sur les technologies de sécurité et de rupture du CNRC - Rapport final https://nrc.canada.ca/sites/default/files/2019-03/sdtech_report_2018_f.pdf (janvier 2018).
[33] Ouranos https://www.ouranos.ca/en/ouranos/ (consulté en novembre 2020).
[34] Ouranos: PLAN STRATÉGIQUE 2014-2020 (décembre 2014).
[35] University of Toronto: U of T and AMD launch supercomputing program dedicated to big-data health research https://www.utoronto.ca/news/u-t-and-amd-launch-supercomputing-program-dedicated-big-data-health-research (consulté en septembre 2020).

groupes de recherche universitaires. La priorité est donnée aux utilisateurs industriels, ce qui permet aux utilisateurs universitaires d'utiliser gratuitement les ressources restantes. Grâce à cet aspect industriel, le modèle de financement est autosuffisant. Il dispose à la fois d'un accès traditionnel au système de CIP par lots et d'une interface d'infonuagique. [36]

SOSCIP

La Southern Ontario Smart Computing Innovation Platform, dont le siège est dans le MaRS Discovery District de Toronto, a été créée en 2012. Il s'agit d'une coalition entre 15 établissements postsecondaires de l'Ontario, IBM Canada et diverses petites et moyennes entreprises de l'Ontario. Elle offre aux projets admissibles un accès payant à des ressources de GPU, ainsi que des ressources d'infonuagique massivement parallèle.[37]Les projets admissibles doivent provenir d'une collaboration entre l'industrie et le milieu universitaire, répondre à des besoins informatiques avancés et avoir des « objectifs de commercialisation clairs et réalisables ». Le partenaire universitaire doit être un chercheur principal issu d'un établissement membre de SOSCIP, tandis que le partenaire commercial doit se situer dans le sud de l'Ontario. [38]En 2020, SOSCIP aura travaillé avec plus de 120 PME ontariennes et aura soutenu plus de 195 projets de R&D.[39]La plateforme de GPU actuelle de SOSCIP, Mist, a été lancée en 2020 et comprend 54 nœuds IBM Power9, chacun doté de quatre cartes GPU Nvidia V100, connectés via Mellanox InfiniBand EDR. La plateforme se trouve à côté du superordinateur massivement parallèle Niagara de SciNet et partage son système de fichiers utilisateur. [40]La ressource massivement parallèle est une allocation équivalente à 2880 cœurs du superordinateur Niagara de SciNet. Cette plateforme d'analyse dématérialisée est basée sur OpenStack et elle est conçue comme système d'analyse de données massives en libre-service. Elle se compose de plus de 4600 cœurs de CPU dans une grappe mixte x86 et PowerPC, en plus de 70 GPU Nvidia. [41]

SPC/ECCC

Services partagés du Canada (SPC) héberge les principaux systèmes de CIP d'Environnement et Changement climatique Canada (ECCC). À l'heure actuelle, SPC exploite deux superordinateurs Cray pour ECCC, Banting et Daley. Le premier a été mis en service en 2017, tandis que le second a été mis en ligne en 2020. [42]Les systèmes soutiennent les services de modélisation et de prévision météorologique d'ECCC. [43] Ils ne semblent pas être disponibles aux chercheurs universitaires canadiens en général.

---

[36] ACENET: Siku https://www.ace-net.ca/wiki/Siku (consulté en novembre 2020).
[37] SOSCIP https://www.soscip.org/ (consulté en novembre 2020).
[38] SOSCIP: Project Requirements https://www.soscip.org/project-guide/ (consulté en novembre 2020).
[39] SOSCIP: SOSCIP By the Numbers https://www.soscip.org/soscip-by-the-numbers/ (consulté en novembre 2020).
[40] SciNet: Mist GPU cluster https://www.scinethpc.ca/mist/ (consulté en novembre 2020).
[41] SOSCIP: SOSCIP's Advanced Computing Platforms https://www.soscip.org/platforms/ (consulté en novembre 2020).
[42] Services partagés Canada : Calcul de haute performance https://www.canada.ca/en/shared-services/corporate/data-centre-consolidation/high-performance-computing.html ; et Mise à niveau de l'environnement de calcul de haute performance à l'appui du gouvernement numérique https://www.canada.ca/en/shared-services/campaigns/stories/hpc-upgrade.html (consulté en septembre 2020).
[43] ECCC : Analyses et Modélisation : https://meteo.gc.ca/mainmenu/modelling_menu_f.html (consulté en novembre 2020).

Fournisseurs de services et de plateformes

Les fournisseurs de services et de plateformes d'IRN proposent aussi souvent des formations ou entreprennent de la recherche, mais contrairement aux fournisseurs d'infrastructures, ils ne possèdent ou n'exploitent pas principalement leur propre infrastructure de CIP.

CCDA

Le Centre canadien de données astronomiques (CCDA) du CNRC a été créé en 1986 et il se trouve au Herzberg Astronomy and Physics (HAA) Research Centre à Victoria, en Colombie-Britannique.[44] Il a pour mandat d'héberger les données de télescopes canadiens et d'exploiter sa plateforme scientifique pour l'astronomie à forte intensité de données. Le CCDA offre des services d'infonuagique, de stockage géré par l'utilisateur, de gestion de groupe, de publication de données et de stockage permanent pour les principales collections de données. Le CCDA ne dispose pas de sa propre infrastructure de CIP, donc il fait plutôt appel aux services fournis en collaboration entre Services partagés Canada, Calcul Canada, CANARIE et des universités par l'entremise du financement de la FCI. Une enquête de 2016[45] sur le HAA montre que l'infrastructure informatique du CCDA est désuète, ce qui nuit à ses activités. Les capacités réseau sont également limitées. De plus, « le transfert de l'infrastructure informatique du HAA vers Services partagés Canada (SPC) a eu des impacts majeurs sur la possibilité de planifier, de mettre en œuvre et d'acquérir de nouveaux équipements informatiques et une capacité réseau ». Le CCDA héberge le Réseau de pointe canadien pour la recherche en astronomie (CANFAR). L'étude citée plus haute indique également que « la solution d'infonuagique de CANFAR est une évolution importante pour que le CCDA réponde aux besoins de la communauté astronomique canadienne ».

CANFAR

Le Réseau de pointe canadien pour la recherche en astronomie (CANFAR) a été créé afin de fournir aux astronomes canadiens des services de CIP pour soutenir leur recherche intensive en données. La suite intégrée de services comprend la gestion de données de recherche, le stockage géré par l'utilisateur, le traitement en nuage et des services spécialisés de visualisation et d'analyse.[46] CANFAR ne possède pas sa propre infrastructure de CIP, donc le réseau fait plutôt appel aux serveurs Openstack de Calcul Canada, notamment la grappe Arbutus. [47]

CMC

CMC Microsystems est un organisme sans but lucratif qui gère le Réseau national de conception du Canada (CNDN). Il s'agit d'un réseau national de 10 000 participants universitaires et de 1 000 entreprises axées sur la recherche et l'innovation dans le domaine des micronanotechnologies.[48] Le réseau a reçu une subvention du Fonds des initiatives scientifiques majeures (FISM) de la FCI. CMC offre plus de 50 outils logiciels de conception assistée par

[44] Centre canadien de données astronomiques : À propos du CCDA https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/fr/about.html (consulté en janvier 2021).
[45] CNRC: Évaluation du portefeuille CNRC Herzberg, Astronomie et astrophysique https://nrc.canada.ca/sites/default/files/2019-03/haa_evaluation_report_2016_f.pdf (novembre 2016).
[46] CANFAR Portfolio https://www.canfar.net/assets/CANFAR_portfolio.pdf (mai 2016).
[47] CANFAR: OpenStack Cloud https://www.canfar.net/en/docs/openstack_cloud_portal/ (consulté en janvier 2021).
[48] CMC Microsystems: About Us https://www.cmc.ca/about-us/ (consulté en février 2021).

ordinateur (CAO) par l'entremise de sa plateforme d'octroi de licences CADpass, ainsi que des plateformes supplémentaires de grappes de calcul et d'infonuagique pour les universitaires canadiens, le tout sous la marque CAD.[49] CMC offre aux utilisateurs de la FCC des licences de logiciels commerciaux qui sont utilisés dans l'infrastructure de la FCC. En plus de CAD, CMC fournit également aux partenaires universitaires des plaquettes multiprojets et des services de fabrication connexes, ainsi que des outils connexes pour les besoins d'essai et de démonstration.

GenAP

La plateforme d'analyse en génétique et génomique GenAP est une infrastructure informatique et un environnement logiciel destinés aux chercheurs en sciences de la vie. Elle a été mise en place en 2015 grâce au financement de CANARIE, Génome Québec, la FCI et le CRSNG. GenAP offre des applications web clés en main fonctionnant sur l'infrastructure du Service infonuagique de Calcul Canada (Arbutus) et de HPC. [50]

LINCS

Le projet Linked Infrastructure for Networked Cultural Scholarship (LINCS) a été mis sur pied en avril 2020 dans le but de créer une infrastructure web sémantique pour convertir de grands ensembles de données en un ensemble de ressources organisées, interconnectées et traitables par machine pour la recherche culturelle canadienne.[51] Il s'agit de sous-systèmes pour 1) convertir et interconnecter les diverses sources de données, 2) stocker les données et 3) accéder aux données avec la possibilité de filtrer, d'analyser, d'annoter et de modifier les résultats sémantiques créés automatiquement.[52] Le projet compte des dizaines de partenaires universitaires, du secteur privé et institutionnel du Canada et des États-Unis. La principale source de financement est la subvention de 2 millions de dollars de l'Initiative sur la cyberinfrastructure de la FCI en 2019, [53]tandis que l'infrastructure dorsale sera fournie par les sites d'hébergement de la Fédération Calcul Canada.

OHDP

Au cours de l'été et de l'automne 2020, l'Ontario a lancé une plateforme de données sur la santé (OHDP), qui offre aux chercheurs un accès intégré aux données liées à la COVID-19. Cette plateforme répond aux exigences de sécurité et confidentialité des données. Les principaux objectifs sont de fournir des informations sur le dépistage de la COVID-19 au sein des populations, déterminer les facteurs de risque, prédire les éclosions, évaluer l'efficacité des mesures de traitement et optimiser l'allocation de ressources.[54] Sur le plan méthodologique, l'OHDP repose sur l'intelligence artificielle et l'apprentissage machine, notamment par l'entremise

[49] CMC Microsystems: CAD https://www.cmc.ca/cad/ (consulté en février 2021).
[50] GenAP: Genetics & Genomics Analysis Platform: Introduction to GenAP https://genap.ca/p/help/introduction (consulté en janvier 2021).
[51] LINCS https://lincsproject.ca/ (consulté en janvier 2021).
[52] LINCS: Research Data Infrastructure https://lincsproject.ca/development/ (consulté en janvier 2021).
[53] University of Guelph: U of G-Led Network Gets $2 Million to Link Cultural Researchers https://news.uoguelph.ca/2019/02/university-of-guelph-led-network-gets-2-million-to-link-cultural-researchers/ (février 2019, consulté en janvier 2021).
[54] Ontario News Release: Province Developing New Health Data Platform to Help Defeat COVID-19 https://news.ontario.ca/en/release/56659/province-developing-new-health-data-platform-to-help-defeat-covid-19 (avril 2020).

du CIP.[55] La plateforme est le fruit d'une collaboration entre le ministère de la Santé de l'Ontario, Calcul Ontario, l'Université Queen's, Vector Institute et d'autres établissements de recherche ontariens.[56] Le financement provient du ministère de la Santé de l'Ontario.

Syzygy

Syzygy est un exemple canadien récent majeur de plateforme en tant que service (PaaS) en infonuagique. Il s'agit d'une collaboration entre le Pacific Institute for the Mathematical Sciences (PIMS), Calcul Canada et Cybera, lancée en 2017, qui fournit gratuitement aux chercheurs canadiens des ressources informatiques basées sur Jupyter Notebook.[57] Les utilisateurs finaux peuvent se connecter en utilisant les identifiants de leur université si celle-ci collabore avec Syzygy ou des comptes Google, pour faire du développement de code, en plus de lancer des productions légères et des essais. Le fondateur de cette plateforme indique qu'elle a connu un « succès catastrophique »[58], avec plus de 34 000 utilisateurs à la fin de 2020.

Prestataires de recherche et de formation

Ces prestataires offrent des services de recherche et de formation dans leurs disciplines respectives, mais ils exploitent aussi parfois des plateformes ou des infrastructures dans ces domaines.

Amii

L'Alberta Machine Intelligence Institute (Amii) a été créé en 2002 et se spécialise dans l'intelligence artificielle et l'apprentissage machine. Il s'agit de l'un des trois principaux instituts d'IA au Canada. L'institut Amii se trouve à l'Université d'Alberta et il a été constitué en société à but non lucratif en 2017. Il compte actuellement plus de 25 chercheurs boursiers et une vingtaine d'employés [59] et il est financé par Alberta Innovates, le CIFAR, la province de l'Alberta et l'Université d'Alberta. [60]

IQC

L'Institute for Quantum Computing (IQC) de l'Université de Waterloo est un institut de recherche créé en 2002 qui compte 32 professeurs affiliés et plus de 300 chercheurs qui mettent au point de nouvelles technologies quantiques.[61] Il est financé par Mike et Ophelia Lazaridis, le gouvernement du Canada, le gouvernement de l'Ontario et l'Université de Waterloo. L'institut a

---

[55] OHDP: About https://computeontario.ca/covid-19-health/about-ohdp/overview/ (consulté en novembre 2020).
[56] OHDP: Project Team https://computeontario.ca/covid-19-health/about-ohdp/project-team/ (consulté en novembre 2020).
[57] Syzygy.ca https://syzygy.ca/# (consulté en décembre 2020).
[58] James Colliander at Berkeley Computing, Data Science, and Society's 2020 National Workshop on Data Science Education - National Scale Interactive Computing https://data.berkeley.edu/academics/resources/data-science-education-resources/2020-national-workshop-data-science-education (consulté en décembre 2020).
[59] Amii: Our People https://www.amii.ca/about/our-people/ (consulté en novembre 2020).
[60] Amii: Our Story https://www.amii.ca/about/our-story/ (consulté en novembre 2020).
[61] University of Waterloo: About Institute for Quantum Computing https://uwaterloo.ca/institute-for-quantum-computing/about (consulté en janvier 2021).

obtenu plus de 30 millions de dollars de fonds de recherche au cours de l'exercice 2019-2020.[62] L'IQC n'héberge pas ou n'exploite pas de système informatique quantique pour les clients, mais se spécialise plutôt dans la recherche et l'avancement des technologies d'informatique quantique.

IVADO

La mission de l'Institut de Valorisation des Données (IVADO) est d'accroître les talents scientifiques et industriels en intelligence numérique (en science des données, intelligence artificielle et recherche opérationnelle), en plus d'accélérer son adoption.[63] Il a été créé en 2016 grâce à une importante subvention de 93,6 M$ du Fonds d'excellence en recherche « Apogée Canada ».[64] En 2019, l'institut a accordé 4,5M$ à plus de 40 projets de recherche impliquant plus de 350 personnes.[65] L'IVADO compte 95 membres industriels actifs et plus de 1400 membres dans sa communauté scientifique. Il s'agit d'un important prestataire de formation à l'utilisation des outils d'intelligence numérique, dont plus de 750 personnes ont bénéficié en 2019.

MILA

L'Institut des algorithmes d'apprentissage de Montréal (MILA) est un partenariat à but non lucratif entre l'Université de Montréal et l'Université McGill. Polytechnique et HEC Montréal sont également étroitement liées au MILA, qui a été créé en 1993 et constitué en société en 2017. Le MILA se spécialise dans l'intelligence artificielle, l'apprentissage machine et l'apprentissage profond. Il compte plus de 500 chercheurs.[66] Il s'agit de l'un des trois principaux instituts de recherche en IA au Canada. Au cours de l'exercice 2018-19, les revenus du MILA étaient d'environ 7 millions de dollars, dont la plupart provenaient de subventions gouvernementales (6 millions de dollars), financées par le CIFAR, ainsi que le ministère de l'Économie et de l'Innovation du Québec. [67]

IORC

L'Institut ontarien de recherche sur le cancer (IORC) est un organisme de recherche collaborative à but non lucratif qui se trouve dans le centre MaRS à Toronto, en Ontario. Il se spécialise en recherche interdisciplinaire sur le cancer dans des domaines tels que la génomique, l'immuno-oncologie, l'informatique, la découverte de médicaments et la pathologie moléculaire. Les partenaires de l'IORC proviennent du domaine de la santé, de la recherche, du gouvernement et du secteur privé.[68] Il emploie plus de 300 personnes et soutient près de 2000 chercheurs hautement qualifiés en Ontario. L'institut ne dispose pas d'important système de CIP sur place, mais il entreprend plusieurs initiatives de CIP, dont un programme de biologie computationnelle axé sur le développement de nouveaux algorithmes, logiciels et outils de visualisation liés à la

---

[62] IQC: Annual Report April 1, 2019-March 31, 2020 https://uwaterloo.ca/institute-for-quantum-computing/sites/ca.institute-for-quantum-computing/files/uploads/files/iqc_report_to_ised_2019-2020-eng_aug_2020.pdf (juillet 2020).

[63] IVADO https://ivado.ca/en/ivado/ (consulté en janvier 2021).

[64] HEC Montréal : IVADO obtient une subvention de 93,6 M$ d'Apogée Canada https://www.hec.ca/en/news/2016/IVADO-receives-93-6-M-grant-from-Canada-First.html (septembre 2016, consulté en janvier 2021),

[65] IVADO : Rapport d'activité 2019 https://ivado.ca/rapport-activites-2019-EN/IVADO_RapportActivites2019_ENG_v3_web.pdf (février 2020).

[66] MILA : À propos de MILA https://mila.quebec/mila/ (consulté en novembre 2020).

[67] MILA : Rapport annuel du 1er avril 2018 au 31 mars 2019 https://mila.quebec/wp-content/uploads/2020/01/Mila-Annual-Report-2018-2019.pdf (janvier 2020).

[68] IORC : À propos de nous https://oicr.on.ca/about-us/ (consulté en novembre 2020).

génomique pour les grands ensembles de données.[69] L'IORC est le deuxième plus important organisme de financement de la recherche sur le cancer au Canada. Ses revenus pour l'exercice 2019-20 étaient d'environ 85 millions de dollars, financés principalement par le ministère des Collèges et Universités de l'Ontario. [70]

Vector Institute

Vector Institute se spécialise dans l'intelligence artificielle, l'apprentissage machine et apprentissage profond. Il a été constitué en 2017 comme société à but non lucratif avec l'aide de l'Université de Toronto. L'institut représente une communauté de plus de 500 chercheurs et 1000 étudiants à la maîtrise, qui sont issus d'une vingtaine d'établissements universitaires ontariens.[71] Au cours de l'exercice 2019-20, les revenus de Vector se sont élevés à près de 27 millions de dollars, dont environ 10 millions proviennent de la province de l'Ontario et environ 5,5 millions du gouvernement fédéral. Par ailleurs, quarante-sept entreprises ont financé les activités de Vector à près de 9 M$. Soulignons que le financement de Vector connaît d'importantes fluctuations annuelles parce que la province de l'Ontario fournit son financement en début de période. Les dépenses totales de Vector au cours de l'exercice 2019-20 étaient d'environ 18 millions de dollars, ce qui reflète plus précisément le volume général des activités annuelles.[72] Vector Institute exploite sa propre infrastructure d'intelligence artificielle, d'une valeur de 6 M$. Elle est composée de 1163 GPU répartis sur près de 200 serveurs, dont 11 nœuds de calcul à grande mémoire basés uniquement sur le CPU. [73]

Organisations commerciales

Plusieurs organisations commerciales fournissent des services de CIP aux chercheurs canadiens. L'infonuagique s'est imposée comme le principal mécanisme d'accès aux ressources externes de CIP grâce à sa facilité relative d'utilisation, à son faible coût d'entrée et à son modèle économique par répartition. En ce qui concerne les risques, les incertitudes liées à la budgétisation et la concentration actuelle du secteur entre les mains de quelques acteurs exposent les utilisateurs au risque de changements de prix incontrôlés. Les trois grands fournisseurs d'infonuagique sont AWS d'Amazon, Azure de Microsoft et GCP de Google, qui représentaient environ 88 % des déploiements de HPC à l'échelle en 2019, selon InsideHPC. [74]Ces trois fournisseurs d'infonuagique sont également certifiés par le gouvernement du Canada pour héberger des données classifiées Protégé B dans certaines offres de services. [75]

---

[69] IORC : Biologie computationnelle https://oicr.on.ca/research-portfolio/computational-biology/ (consulté en novembre 2020).

[70] IORC : Institut ontarien de recherche sur le cancer - États financiers https://oicr.on.ca/wp-content/uploads/2020/09/OICR-Financials-1920.pdf (septembre 2020).

[71] Vector Institute: April 2019-March 2020 Annual Report https://vectorinstitute.ai/wp-content/uploads/2020/11/vector_annual-reportv8.pdf (novembre 2020).

[72] Vector Institute: Financial Statements for the Year Ended March 31, 2020 https://vectorinstitute.ai/wp-content/uploads/2020/11/2020-fs-final-vector-institute.pdf (novembre 2020).

[73] Vector Institute Position Paper submission to NDRIO: Canada's Future DRI Ecosystem: AI Research Needs https://engagedri.ca/wp-content/uploads/2020/12/Canada%E2%80%99s-Future-DRI-Ecosystem_-AI-Research-Needs.pdf (décembre 2020).

[74] InsideHPC White Paper: Could Adoption for HPC: Trends and Opportunities https://insidehpc.com/white-paper/cloud-adoption-for-hpc-trends-and-opportunities/ (novembre 2019).

[75] Services infonuagiques GC (Protégé B) https://cloud-broker.canada.ca/s/central-provider-page-v2?language=en_CA (consulté en novembre 2020).

Amazon AWS

Amazon est le premier fournisseur d'infonuagique et un exemple typique de ce genre de service, en général et plus précisément pour le CIP. L'entreprise a lancé Amazon Web Service (AWS) en 2002, puis Elastic Compute Cloud (EC2) en 2006. Depuis, le portefeuille de services a connu une croissance astronomique et comprend désormais une suite complète de services [76]de mise en réseau, de routage, de stockage et de CIP, notamment des interconnexions à haut débit, nœuds de mémoire de grande taille, calcul GPU, circuits logiques programmables et calcul quantique (Amazon Braket)[77]. Amazon a enregistré environ 46 % des déploiements mondiaux de HPC dématérialisé en 2019.[72] L'ampleur d'utilisation au sein de la communauté universitaire canadienne n'est pas connue pour le moment, mais on peut supposer qu'elle est substantielle grâce à la facilité relative d'utilisation et à la nature dynamique de ce service.

Google Cloud Platform

Comme AWS, Google Cloud Platform (GCP) est un service d'infonuagique bien ancré qui offre des infrastructures en tant que service (IaaS), plateformes en tant que service (PaaS) et services sans serveur.[78] Google Compute Engine (GCE) est l'équivalent d'EC2 d'Amazon. Il fournit des nœuds de calcul CPU, GPU et de grande mémoire pour le CIP, mais ne fournit pas de circuits logiques programmables. En 2019, GCP a capté environ 18 % des déploiements mondiaux de HPC dématérialisé. [72] Google offre des services d'intelligence artificielle et d'apprentissage machine par l'entremise de sa plateforme[79], mais l'entreprise met également au point le processeur interne Google TPU ML, qui est accessible au public par le biais du service d'infonuagique Google TPU. [144]

Microsoft Azure

Microsoft Azure est le troisième grand fournisseur de CIP dématérialisé avec environ 24 % des déploiements mondiaux en 2019.[72] Comme AWS, le service d'infrastructure en tant que service (IaaS) de CIP d'Azure comprend des instances de calcul optimisées pour les CPU, les circuits logiques programmables, les GPU et l'interconnexion à haut débit.[80] L'environnement de production et de test d'informatique quantique est également accessible par l'entremise d'Azure Quantum depuis février 2021.[81] En ce qui concerne le CIP hautement spécialisé, Azure propose également les superordinateurs Cray XC et CS dans le cadre de son offre d'infonuagique. En revanche, ils sont accessibles par location à court terme (taux horaire). Le service inclut

---

[76] Amazon White Paper: Overview of Amazon Web Services https://d0.awsstatic.com/whitepapers/aws-overview.pdf (août 2020).

[77] Amazon: Quantum computing is now available on AWS through Amazon Braket https://aws.amazon.com/about-aws/whats-new/2020/08/quantum-computing-available-aws-through-amazon-braket/ (consulté en novembre 2020).

[78] Google Cloud Platform Services Summary https://cloud.google.com/terms/services (consulté en novembre 2020).

[79] Google AI Platform https://cloud.google.com/ai-platform/ (consulté en novembre 2020).

[80] Microsoft Azure High-Performance Computing https://azure.microsoft.com/en-us/solutions/high-performance-computing/ (consulté en novembre 2020).

[81] InsideHPC: Azure Quantum Now in Public Preview https://insidehpc.com/2021/02/azure-quantum-now-in-public-preview/ (consulté en février 2021).

l'infrastructure de stockage standard basée sur Azure, mais il ne s'inscrit pas dans l'infrastructure d'infonuagique en tant que service (IaaS). [82]

OVHcloud

Même si les trois grands fournisseurs d'infonuagique représentent près de 90 % des déploiements de CIP dématérialisé, plusieurs membres de la communauté font appel à de nombreux fournisseurs spécialisés ou plus ciblés. Par exemple, OVH est une société privée créée en 1999, qui est aujourd'hui le plus grand fournisseur d'hébergement en nuage d'Europe.[83] Elle se présente comme une alternative transparente et sécurisée aux trois grands fournisseurs, notamment en offrant des solutions de stockage et d'infonuagique pour des données sensibles.[84] La société a déclaré un chiffre d'affaires d'environ 500 millions d'euros en 2018, avec des investissements agressifs de plusieurs milliards d'euros en 2021-26 pour rattraper ses rivaux. [85] Elle a annoncé en janvier 2021 une initiative majeure de stockage en tant que service, conjointement avec IBM et Atempo. Il s'agit d'un service de stockage sécurisé, souverain et résilient pour les entreprises et les institutions publiques européennes. [86]

IBM Canada Watson et Bluegene

Depuis 2012, IBM Canada collabore avec des chercheurs canadiens et leur fournit d'importantes ressources informatiques, notamment en collaboration avec SOSCIP et CAC de l'Université Queen's. Auparavant, les chercheurs de l'Ontario avaient accès à une plateforme analytique IBM Watson (au CAC), au superordinateur IBM Bluegene/Q (à SciNet pour le compte de SOSCIP), à une plateforme de circuit logique programmable, à une plateforme accélérée par GPU et aux services d'infonuagique d'IBM. [87] Les trois premiers services semblent avoir été abandonnés, tandis que les deux derniers sont accessibles via l'offre payante de SOSCIP. [88, 89]

Organisations internationales accessibles aux Canadiens

Les chercheurs canadiens ont accès à diverses ressources internationales pour la recherche informatique, généralement dans le cadre de collaborations avec des chercheurs étrangers. Le CP étranger soumet la demande principale d'accès. Le chercheur canadien y obtient ensuite accès, en tant que collaborateur étranger de cette équipe de recherche. Dans certains cas,

---

[82] Microsoft Azure: Cray in Azure https://azure.microsoft.com/en-us/solutions/high-performance-computing/CIPy/ (consulté en novembre 2020).

[83] IT World Canada: Canadian customers' heads are still in the clouds, and so is VMware's https://www.itworldcanada.com/article/canadian-customers-heads-are-still-in-the-clouds-and-so-is-vmwares/421294 (consulté en janvier 2021).

[84] Reuters: France's OVH partners with Google for European cloud computing push https://www.reuters.com/article/ctech-us-ovh-google-cloud-idCAKBN27Q0OP-OCATC (10 novembre 2020, consulté en janvier 2021.)

[85] Reuters: France's OVH to triple spending to take on Google, Amazon in cloud computing https://www.reuters.com/article/us-ovh-strategy-idUSKCN1MS17L (18 octobre 2018, consulté en janvier 2021).

[86] InsideHPC: OVHcloud Teams with IBM and for Cloud Storage https://insidehpc.com/2021/01/ovhcloud-teams-with-ibm-and-atempo-for-cloud-storage/ (consulté en janvier 2021).

[87] IBM Canada: Why investing in Canadian R&D matters https://www.ibm.com/ibm/ca/en/ibmcanada100/investing-in-canadian-rd.html (consulté en novembre 2020).

[88] SOSCIP https://www.soscip.org/ (consulté en novembre 2020).

[89] CAC: CAC Services https://cac.queensu.ca/services/ (consulté en novembre 2020).

l'accès aux systèmes de calcul peut être limité. Par exemple, l'accès pour les étrangers aux superordinateurs Summit et Frontier d'Oak Ridge National Lab (ORNL) est contrôlé par un mécanisme PAS (système de gestion des accès) du laboratoire [90]et n'est parfois pas du tout accessible aux étrangers, ou seulement dans des circonstances limitées. Si l'accès aux systèmes internationaux est limité aux collaborateurs de chercheurs principaux étrangers, cela pourrait priver de leurs droits certains chercheurs canadiens, parce qu'ils n'ont pas établi de telles collaborations internationales. Dans ce qui suit, nous examinons quelques exemples représentatifs des États-Unis, de l'Union européenne et de l'Australie.

Aux États-Unis, le Department of Energy (DOE) et la National Science Foundation (NSF) sont les deux principaux bailleurs de fonds de la cyberinfrastructure, qui reposent sur plusieurs mécanismes. Les principaux programmes du DOE pour l'hébergement de systèmes de calcul intensif [91]incluent le National Nuclear Security Administration (NNSA, qui héberge le superordinateur Trinity aux Los Alamos National Labs et le superordinateur Sierra aux Lawrence Livermore National Labs)[92], Oak Ridge Leadership Computing Facility (OLCF, qui héberge les systèmes Summit et Titan)[93], Argonne Leadership Computing Facility (qui héberge les nouveaux systèmes Aurora et Theta)[94] et National Energy Research Scientific Computing Center (NERSC, qui héberge les nouveaux systèmes Perlmutter et les superordinateurs Cori actuels)[95]. Les systèmes hébergés par la NNSA ne sont pas accessibles pour une utilisation scientifique générale (même pour les citoyens américains), car ils exécutent des simulations très sensibles de l'arsenal nucléaire. L'accès public et l'allocation de ressources aux systèmes de l'ORNL et de l'ALCF se font principalement par le biais du programme INCITE (Innovative and Novel Computational Impact on Theory and Experiment) du DOE Office of Science. [96] Ces laboratoires ont également des programmes discrétionnaires et d'autres mécanismes pour attribuer des ressources à plus petite échelle et à court terme. Par exemple, le programme ALCC (Leadership Computing Challenge) de l'Advanced Scientific Computing Research (ASCR) permet d'attribuer des ressources importantes à l'OLCF, ALCF et NERSC pour des simulations à haut risque et à haut rendement. [97] NERSC a son propre processus d'allocation de ressources informatiques pour la recherche énergétique (ERCAP). [98]

L'OAC (Office of Advanced Cyberinfrastructure) est le principal mécanisme de la NSF pour l'octroi de ressources de CIP. Par exemple, la NSF verse un important financement au National Center

---

[90] Oak Ridge National Laboratory: Applying for a user account https://docs.olcf.ornl.gov/accounts/accounts_and_projects.html#applying-for-a-user-account (consulté en novembre 2020).

[91] US Department of Energy: Supercomputing and Exascale https://www.energy.gov/science-innovation/science-technology/computing (consulté en novembre 2020).

[92] National Nuclear Security Administration: Maintaining the Stockpile https://www.energy.gov/nnsa/missions/maintaining-stockpile (consulté en novembre 2020).

[93] Oak Ridge Leadership Computing Facility: Compute Systems https://www.olcf.ornl.gov/olcf-resources/compute-systems/ (consulté en novembre 2020).

[94] Argonne Leadership Computing Facility: ALCF Resources https://www.alcf.anl.gov/alcf-resources (consulté en novembre 2020).

[95] NERSC: Systems https://www.nersc.gov/systems/ (consulté en novembre 2020).

[96] INCITE Leadership Computing: INCITE Program https://www.doeleadershipcomputing.org/about/ (consulté en novembre 2020).

[97] Argonne Leadership Computing Facility: ALCC Allocation Program https://www.alcf.anl.gov/science/alcc-allocation-program (consulté en novembre 2020).

[98] NERSC: Allocations of computer time and storage https://www.nersc.gov/users/accounts/allocations/ (consulté en novembre 2020).

for Supercomputing Applications (NCSA),[99] qui se trouve à l'Université de l'Illinois, Urbana-Champaign, et qui héberge le superordinateur Blue Waters, en plus de diriger le projet XSEDE (Extreme Science and Engineering Discovery Environment). Ce programme fournit aux chercheurs américains un accès centralisé à de multiples ressources de calcul intensif, notamment celles des autres centres de calcul intensif financés par la NSF, par exemple l'Université de l'Indiana (IU/TACC, Jetstream), le Pittsburgh Supercomputing Center (PSC, Bridges, Bridges-2, Anton 2), le San Diego Supercomputer Center (SDSC, Comet, Expanse) et le Texas Advanced Computing Center (TACC, Ranch, Stampede2).[100]L'accès aux diverses ressources financées par la NSF n'est pas simple, par exemple le superordinateur Frontera du TACC n'est pas accessible par le mécanisme d'allocation XSEDE. En effet, le TACC a plutôt mis en place un cadre pour les demandes d'allocation[101]. De plus, le superordinateur Cheyenne du National Center for Atmospheric Research (NCAR) ne fait pas partie du programme XSEDE.[102]

En plus des exemples ci-haut de superordinateurs de « classe dirigeante » financés par le DOE et la NSF, l'infrastructure de CIP aux États-Unis se caractérise par plusieurs systèmes à plus petite échelle, gérés par les universités, les coalitions ou les États où ils se trouvent. Un grand nombre de ces systèmes sont potentiellement accessibles aux chercheurs canadiens par le biais de collaborations internationales. Par exemple, l'Université Harvard a mis en commun ses ressources en 2011 avec le Massachusetts Institute of Technology et d'autres universités locales, l'État et le secteur privé pour créer le Massachusetts Green High Performance Computing Center, une installation de 168 millions de dollars américains. [103] Le centre fournit des installations et des infrastructures aux universités membres pour l'hébergement de leur infrastructure de CIP. Par exemple, l'Université de Harvard exploite les principaux composants de sa grappe hybride Cannon de 100 000 CPU. [104]

En Europe, le Partnership for Advanced Computing (PRACE) est l'organisation principale qui coordonne les ressources de CIP. Elle a été créée en 2010 avec un financement total de 125 millions d'euros jusqu'en 2019.[105] L'écosystème européen de CIP que gère PRACE est réparti en trois niveaux : le niveau 0 concerne les superordinateurs (systèmes en pétaFlops), le niveau 1 concerne les systèmes nationaux et le niveau 2 concerne les systèmes régionaux et universitaires.[106] Les pays membres hôtes (Allemagne, France, Italie, Espagne et Suisse) se sont engagés à financer et à fournir des services d'infrastructure de recherche à la coalition de 26 pays membres de PRACE. Actuellement, l'infrastructure de CIP de PRACE se compose de sept systèmes de niveau 0, dont certains sont hébergés en plusieurs endroits ou possèdent plusieurs segments fonctionnellement différents. Le plus récent est HAWK, situé au High-Performance

---

[99] NCSA: About NCSA http://www.ncsa.illinois.edu/about (consulté en décembre 2020).
[100] XSEDE : XSEDE Resource Information https://portal.xsede.org/allocations/resource-info (récupéré en décembre 2020).
[101] TACC: FRONTERA ALLOCATION SUBMISSION GUIDELINES https://frontera-portal.tacc.utexas.edu/allocations/policy/ (consulté en décembre 2020).
[102] NCAR: Allocations https://www2.cisl.ucar.edu/user-support/allocations (consulté en décembre 2020).
[103] The Harvard Crimson: Harvard Helps Build $168M Supercomputing Facility https://www.thecrimson.com/article/2011/10/31/supercomputers-research-facility-holyoke/ (consulté en décembre 2020).
[104] Harvard University Faculty of Arts and Sciences Research Computing: Cluster Architecture https://www.rc.fas.harvard.edu/about/cluster-architecture/ (consulté en décembre 2020).
[105] PRACE: Introduction https://prace-ri.eu/about/introduction/ (consulté en décembre 2020).
[106] ARCHER Training: HPC in Europe https://www.archer.ac.uk/training/course-material/2017/11/intro-epcc/slides/L12-PRACE.pdf (consulté en décembre 2020).

Computing Centre Stuttgart (HLRS), qui a été mis en ligne en 2020 avec un rendement maximal de 26 pétaFlops. [107]Il existe dix-neuf systèmes de niveau 1, pour une puissance totale de calcul de 16 pétaFlops. Les systèmes PRACE sont en principe accessibles aux chercheurs canadiens, avec d'éventuelles limitations imposées par les établissements hôtes. De plus, le cadre de référence du dernier appel à ressources indique que la collaboration avec un PI d'un pays européen qui contribue à PRACE améliorera les chances d'approbation des ressources.[108] PRACE encourage également la collaboration internationale via des appels à collaboration avec XSEDE et RIST.[109] L'écosystème européen de CIP étant assez complexe, PRACE a récemment lancé le portail HPC-in-Europe afin de coordonner les services européens de CIP par une approche ascendante, en fournissant un guichet unique aux utilisateurs de CIP. Ce portail, qui se trouve à l'adresse hpc-portal.eu, est encore en cours de développement. [110]

En 2018, l'Union européenne crée le European High-Performance Computing Joint Undertaking (EuroHPC JU) dans le but de coordonner les efforts et de financer les ordinateurs à l'échelle exa, en plus de gérer un financement de 1,1 milliard d'euros pour l'exercice 2019-20. Les trois principaux sites d'hébergement des systèmes antérieurs à l'échelle exa seront le Barcelona Supercomputing Centre (Espagne), le CSC (Finlande) et le CINECA (Italie). Ce dernier compte actuellement cinq superordinateurs en construction, dont le plus grand (LUMI) aurait une performance maximale de 552 pétaFlops et devrait être mis en service en 2021. [111]Selon Oriol Pineda de PRACE, environ 5 % des ressources de PRACE ont été utilisées par des étrangers, tandis que les politiques d'admissibilités des chercheurs étrangers à l'utilisation des ressources EuroHPC sont encore en cours d'élaboration. [107]

L'Australie a deux centres de calcul intensif de niveau 1, qui sont financés par la National Collaborative Research Infrastructure Strategy (NCRIS) du ministère de l'Éducation du gouvernement australien. [112] L'organisation National Computational Infrastructure (NCI) héberge le superordinateur le plus rapide d'Australie, Gadi, qui a une performance maximale de 9 pétaFlops, [113] tandis que le Pawsey Supercomputing Centre (PSC) exploite le superordinateur Magnus, dont le rendement se situe à l'échelle du pétaFlop. [114] En octobre 2020, le PSC a annoncé que son prochain superordinateur serait un système HP/Cray de 50 pétaFlops. [115] Les

[107] PRACE: HPC Systems https://prace-ri.eu/hpc-access/hpc-systems/ (consulté en décembre 2020).
[108] PRACE: PRACE Project Access Terms of Reference – 22nd Call for Proposals https://prace-ri.eu/wp-content/uploads/Terms_of_Reference_Call22.pdf (consulté en décembre 2020).
[109] PRACE: Collaborative Calls https://prace-ri.eu/hpc-access/collaborative-calls/ (consulté en décembre 2020).
[110] SC20: European HPC Ecosystem Summit Presentation by Oriol Pineda https://cdmcd.co/Qqz7Eq (voir Q&R à 37:55, novembre 2020).
[111] EuroHPC: Discover EuroHPC https://eurohpc-ju.europa.eu/discover-eurohpc (consulté en décembre 2020).
[112] NCRIS Network: Infrastructure Projects Funded by NCRIS https://www.ncris-network.org.au/capabilities (consulté en décembre 2020).
[113] NCI Australia: HPC Systems https://nci.org.au/our-systems/hpc-systems (consulté en décembre 2020).
[114] PSC: Magnus https://pawsey.org.au/systems/magnus/ (consulté en décembre 2020).
[115] PSC: Powering the next generation of Australian research with HPE https://pawsey.org.au/powering-the-next-generation-of-australian-research-with-hpe/ (consulté en décembre 2020).

systèmes actuels de Pawsey sont en général accessibles aux chercheurs étrangers s'ils collaborent avec un chercheur principal admissible. [116]

3.4 Comment le CIP est-il mis en œuvre et financé dans d'autres pays ?

Le rapport du CLIRN de 2017 explore la prestation de services de CIP au niveau mondial au chapitre 4.3 et à l'annexe C. Depuis, le paysage général du financement international n'a pas beaucoup changé. Il y a des modèles de prestation et de financement supranationaux (PRACE et EuroHPC dans l'UE), des modèles nationaux complexes (aux États-Unis le DOE et la NSF gèrent plusieurs sources de financement au niveau régional et local), des modèles nationaux centralisés (au Japon par exemple) et des modèles nationaux collaboratifs (en Australie par exemple). On tente toujours de caractériser le financement du CIP à l'échelle mondiale. Comme l'Alliance doit proposer à ISDE un plan stratégique et un nouveau modèle de financement de l'IRN à l'automne 2021, une analyse approfondie du paysage international de la prestation et du financement de l'IRN est nécessaire pour fournir des idées potentielles pour la prestation de futurs services au Canada. Vu l'ampleur des ressources requises pour une telle étude, un examen et une analyse détaillés des modèles de prestation et de financement de l'IRN à l'échelle mondiale (incluant le CIP, les LR et la GDR) seront effectués dans le cadre d'un projet d'analyse environnementale au début ou au milieu de 2021.

Comme nous l'avons vu dans la section précédente, la création d'EuroHPC est un développement majeur dans le paysage international du CIP au cours des dernières années. Cette initiative a pour but de fournir des ressources de superordinateurs de classe mondiale dans l'Union européenne, notamment avec des augmentations de financement importantes au-delà et indépendamment de l'enveloppe budgétaire de PRACE. En ce qui concerne les avancées techniques, le chapitre précédent aborde également divers nouveaux superordinateurs étrangers qui sont devenus accessibles ou le seront bientôt aux chercheurs canadiens par le biais de collaborations internationales potentielles au niveau individuel.

Comme nous l'avons vu au chapitre 4.1, le CIP au Canada est encore largement utilisé par les disciplines traditionnelles des sciences exactes comme la physique, l'ingénierie et l'informatique. La diversité de l'utilisation du CIP s'est accrue très rapidement au Canada, non seulement en matière de demande, mais de services d'IRN plus complets. À l'échelle mondiale, certaines disciplines de recherche mal desservies bénéficient déjà d'un soutien important pour l'IRN. En France par exemple, l'infrastructure Huma-Num fournit aux chercheurs en sciences humaines et sociales des services de CIP, mais aussi une gamme complète de services d'IRN tout au long du cycle de vie de la recherche, tandis que Prodego concerne les données des sciences sociales. Cette infrastructure est considérée comme une « Très Grande Infrastructure de Recherche (TGIR) » au niveau du financement gouvernemental. Elle fournit des plateformes et des outils pour le traitement, la conservation, la diffusion et la préservation à long terme des données de recherche numériques. [117]

[116] Pawsey: Application Process https://support.pawsey.org.au/documentation/display/US/Application+Process (récupéré en décembre 2020).

[117] Huma-Num: About us https://www.huma-num.fr/about-us (consulté en décembre 2020).

3.5 Tendances d'architectures, marchés, besoins de CIP et d'IA

La section suivante concerne les tendances et les besoins en matière de CIP et d'IA. Au cours de l'hiver 2020 et du printemps 2021, l'Alliance a mené trois évaluations détaillées et ciblées sur les tendances futures : 1) enquête sur l'évaluation des besoins ; analyse des besoins et tendances dans la communauté canadienne de l'IRN, 2) analyse de l'environnement international et 3) examen des progrès technologiques.

Malgré la COVID-19 et les circonstances économiques très difficiles en 2020 (et probablement dans les années suivantes), la croissance des marchés du CIP et de l'IA devrait se poursuivre considérablement dans les cinq prochaines années. Le taux de croissance annuel composé (TCAC) est estimé à 7,1 %, pour atteindre 55 milliards de dollars américains en 2024, selon Intersect360 Research. Cette croissance risque de fluctuer au fil du temps, donc les dépenses diminueront en 2020, mais la demande laissée sans réponse augmentera considérablement les dépenses en 2021. En revanche, ces fluctuations ne devraient pas avoir d'impact sur le TCAC à long terme (comparativement aux estimations avant la COVID-19).[118] Hyperion Research prévoyait un TCAC d'environ 8,7 % avant la crise. De plus, les estimations de ce groupe sont plus prudentes que celles d'Intersect Research en ce qui concerne les dépenses à long terme après la crise. Il cite un probable resserrement budgétaire dans le secteur gouvernemental à partir de 2023.[119] Lors de sa conférence SC20 en novembre 2020, Hyperion Research estimait que l'effet de la COVID-19 serait court, mais prononcé, avec une chute de 11,5 % des revenus du marché de CIP au cours du premier semestre de 2020. [120]

En ce qui concerne les clients, les dépenses publiques devraient augmenter au cours des cinq prochaines années, en termes absolus et relatifs, à cause d'une diminution des dépenses dans le secteur commercial (énergie, commerce de détail et fabrication de gros produits). [115] Les biosciences, la défense et les dépenses de laboratoires publics devraient augmenter au cours des prochaines années, notamment avec un appui additionnel pour les biosciences en 2020-22 grâce à la recherche sur la COVID-19. [116]

Infonuagique

Le secteur du CIP en infonuagique est très dynamique, avec des innovations continuelles et une croissance rapide, ce qui pourrait faire l'objet d'une tout autre recherche. On définit l'infonuagique comme étant « la prestation de services informatiques (serveurs, stockage, bases de données, mise en réseau, logiciels, analyse et renseignements) sur internet (« le nuage »). [121]

---

[118] Addison Snell of Intersect360 Research at HPC-AI Advisory Council 2020 Australia Conference: Supercomputing to the Rescue : HPC/AI Market Update http://www.hpcadvisorycouncil.com/events/2020/australia-conference/pdf/HPCAIMarketUpdate_020920_ASnell.pdf (septembre 2020).

[119] InsideHPC: Hyperion Research Forecasts Widespread COVID-19 Disruption to HPC Market https://insidehpc.com/2020/04/hyperion-research-forecasts-widespread-covid-19-disruption-to-hpc-market/ (consulté en septembre 2020).

[120] InsideHPC : At SC20 : Hyperion Sees COVID HPC Impact Sharp but Short ; HPE in Server Lead ; Aurora 12+/- Months Late ; Cloud HPC Heating Up https://insidehpc.com/2020/11/at-sc20-hyperion-sees-covid-hpc-impact-sharp-but-short-hpe-in-server-lead-aurora-12-months-late-cloud-hpc-heating-up/ (consulté en novembre 2020).

[121] Microsoft Azure: What is cloud computing? https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/ (consulté en mai 2021).

La section suivante porte sur quelques nouvelles offres de services intéressantes et les taux de croissance prévus. Pour ce qui est de la première question, il y a plusieurs plateformes en tant que service (PaaS), qui se positionnent entre l'infrastructure en tant que service (IaaS, Amazon EC2) et le logiciel en tant que service (SaaS, plateforme de collaboration dématérialisée Microsoft Office 365) plus traditionnel. Syzygy est un important exemple canadien de PaaS. Il s'agit d'une collaboration entre le Pacific Institute for the Mathematical Sciences (PIMS), Calcul Canada et Cybera, lancée en 2017, qui fournit gratuitement aux chercheurs canadiens des ressources informatiques basées sur Jupyter Notebook. [122]

La plateforme communautaire PanGeo.io est un autre projet intéressant, qui développe des outils libres python pour les données massives en géosciences. Ceci inclut une modularité pour les pétaoctets de données dans des environnements de CIP en infonuagique et sur place.[123] La plateforme bénéficie d'un financement et d'un soutien à l'international de la part d'agences gouvernementales (NSF, NASA, UK Met Office), d'établissements universitaires (Université de Washington) et du secteur privé (Anaconda). Les développeurs de Pangeo aident par exemple la NASA à traiter des ensembles de données massives dans Earthdata Cloud en exploitant les métadonnées pour obtenir et traiter uniquement ce qui est pertinent dans ces ensembles. [124] La plateforme n'est pas limitée aux géosciences. Par exemple, l'institut Neuro de l'Université McGill mène le projet pilote PanNeuro, qui fait appel à la plateforme PanGeo, dont le volet d'infonuagique est un environnement de science des données basé sur AWS ou Google Cloud. Cette dernière est en phase de développement expérimental sans financement direct à long terme. [125]

L'utilisation des services d'infonuagique est relativement simple quand on dispose d'une carte de crédit, mais il est très difficile de comprendre le coût réel de ces services, qui dépend de la nature exacte des calculs effectués. Pour résoudre ces problèmes, l'Université de Californie à San Diego, avec le soutien de la NSF, a collaboré avec plusieurs institutions californiennes pour créer CloudBank. Il s'agit d' « une entité d'accès à des nuages publics, dont les services sont simplifiés et gérés pour la communauté informatique dans le domaine de la recherche et de l'éducation ».[126] De façon pratique, le service fournit aux chercheurs financés par la NSF un portail utilisateur et une formation connexe pour accéder aux ressources d'infonuagique de plusieurs fournisseurs à des tarifs avantageux grâce à l'évitement des coûts indirects et à la mise en commun du pouvoir d'achat.

Si l'on considère la croissance future et les changements prévus les catégories de produits et de services, on estime que l'infonuagique connaîtra une croissance beaucoup plus forte que le reste du marché, avec un TCAC de plus de 20 % selon Intersect360 Research. [115]. Hyperion Research prévoit un TCAC de près de 25 % jusqu'en 2023, qui s'explique par le virage des groupes de travail (systèmes dont le prix est inférieur à 100 000 USD) vers l'infonuagique. [127] Le volume total

[122] Syzygy.ca https://syzygy.ca/# (consulté en décembre 2020).
[123] Pangeo: About Pangeo http://pangeo.io/about.html (consulté en décembre 2020).
[124] NASA Earthdata: The Pangeo Project: Developing Community Tools For a New Era of Data Analysis https://earthdata.nasa.gov/learn/articles/pangeo-project (consulté en décembre 2020).
[125] Pangeo: Pangeo Cloud http://pangeo.io/cloud.html (consulté en décembre 2020).
[126] CloudBank: About CloudBank https://www.cloudbank.org/about (consulté en décembre 2020).
[127] Hyperion Research white paper : Bringing HPC Expertise to Cloud computing https://www.dellemc.com/resources/en-us/asset/analyst-reports/products/ready-solutions/hyperion-dell-cloud-hpc-ai.pdf (avril 2020).

du marché de CIP en infonuagique était estimé à environ 2,2 à 2,8 milliards de dollars américains en 2019. [128]

Selon InsideHPC en 2019, environ 30 % de la communauté de CIP utilise déjà l'infonuagique dans le domaine de la production, tandis que plus de 90 % songent à l'utiliser. Parmi les organisations qui font déjà appel à ces services, plus de 60 % constatent des résultats positifs ou considèrent l'infonuagique comme un « réel avantage ». [116] Hyperion Research indique que volume de travail dans le domaine du CIP en infonuagique est passé de 10 % à 20 % entre 2018 et 2019. Cette importante augmentation en tout juste un an découle de l'amélioration des applications logicielles, des modèles d'accès simplifié à l'infonuagique et des capacités de calcul accrues. Plus tôt dans la décennie, l'adoption du CIP en infonuagique était motivée par la possibilité de tester et de développer des technologies de pointe en infonuagique et d'exécuter des charges de travail très parallèles. [124]

Solutions sur place

En ce qui concerne les serveurs traditionnels sur place, les achats devraient se concentrer de plus en plus sur les systèmes haut de gamme et les superordinateurs, car ce segment n'est pas encore bien desservi par les fournisseurs commerciaux d'infonuagique à cause du niveau de technologie avancée et sur mesure de ces systèmes. Un vaste mouvement demande de financer les efforts de calcul à l'échelle exa, besoin auquel les systèmes d'infonuagique ne sont pas (encore) en mesure de répondre. [115,116] Pour le CIP moins haut de gamme, notamment les postes de travail ou les serveurs de moins de 100 000 USD, les achats ont baissé de 25 % entre T2 2019 et T2 2020, ce qui indique une pression sur ce segment (au-delà des problèmes à court terme liés à la COVID-19). Par ailleurs, ces charges de travail (continuent) de se déplacer vers les offres d'infonuagique plus flexibles et variables. [129]

Changements dans les profils des fournisseurs et les flux de travail du CIP

En ce qui concerne les fournisseurs et les prestataires de services, le nombre de grands fournisseurs devrait diminuer, tandis que les solutions spécialisées et les jeunes pousses vont augmenter. [130] Il y aura également une concurrence accrue entre les fabricants de puces (par exemple x86, ARM et le marché croissant des puces sur mesure) et une demande accrue pour les accélérateurs à base de GPU. [127] Les Par exemple, les puces sur mesure pour les solutions d'IA sont potentiellement 10 à 100 fois plus rapides que les solutions standard x86 ou à base de GPU. Le financement par capital-risque aux États-Unis pour l'accélération de l'IA est estimé à plus de 4 milliards de dollars US.[131] Selon Hyperion Research, le marché de l'informatique quantique se chiffrait à environ 320 millions de dollars US en 2020, dont une part importante a

---

[128] Addison Snell of Intersect360 Research: Pre-SC20 Market Update http://www.intersect360.com/LiteratureRetrieve.aspx?ID=158848 (consulté en novembre 2020).

[129] InsideHPC: Hyperion: COVID-19 Driving Down HPC Server Revenues, But Impact May Be Moderating https://insidehpc.com/2020/09/hyperion-research-covid-19-driving-decline-in-hpc-server-revenues-but-may-be-moderating/ (consulté en septembre 2020).

[130] Allan Williams of NCI Australia at HPC-AI Advisory Council 2020 Australia: HPC Impact: Future of Scientific Computing http://www.hpcadvisorycouncil.com/events/2020/australia-conference/pdf/FutureofSciComp_020920_AWilliams.pdf (septembre 2020).

[131] Kathy Yelick of Lawrence Berkley National Laboratories at HPC-AI Advisory Council 2020 Australia: AI for Science https://www.youtube.com/watch?v=sLjI9p3u7Mo&list=PLafs-cr09EuW71NepWOlQ98Ht8K40VnJx&index=7 (septembre 2020).

été investie dans l'informatique quantique dématérialisée. Le groupe indique que ce marché pourrait connaître un taux de croissance annuel moyen d'environ 27 %. [132]

Les principales tendances techniques et opérationnelles futures concernent les flux de travail et de l'efficacité améliorés dans l'écosystème de l'IRN. Il faudra donc mettre en place des flux de travail intégrés avec des composants en réseau et en infonuagique. [128] L'intégration de l'informatique périphérique à diverses étapes des flux de travail permettra de réduire les volumes massifs de données et d'améliorer l'efficacité du transfert et du stockage des données. Ces technologies incluent notamment le traitement sur capteur, le traitement déployable sur le terrain, le traitement en temps réel et à proximité du capteur, les interconnexions de CIP intelligentes et les accélérateurs spécialisés, comme indiqué ci-dessus. [128] « L' « IA pour l'IA » est un élément important pour rassembler les composants individuels en les « installations intégrées en continu » qui exécutent des expériences et des analyses automatisées. Le principal avantage étant que les installations intégrées amélioreront la productivité et la reproductibilité dans le domaine scientifique. [128]

Dans le domaine de la recherche, la National Computational Infrastructure, une organisation centrale en Australie, cite comme tendances majeures la demande accrue pour des services de CIP, la hausse du nombre de travaux de CIP, l'augmentation des disciplines, ainsi que le besoin de flux de travail intégrés. [127]

Stockage et gestion de données

Le stockage et la gestion des données vont prendre plus d'ampleur dans les domaines établis (CIP traditionnel) et dans les domaines émergents (données massives, intelligence artificielle, apprentissage machine et apprentissage profond) du CIP. La croissance annuelle de ces marchés sera considérable. Rien que dans le domaine des sciences de la vie, le marché de l'analyse, du stockage et de la gestion des données devrait connaître une croissance annuelle moyenne de 17,1 % entre 2018 et 2024, pour atteindre 41,1 milliards de dollars américains à l'échelle mondiale.[133] Hyperion Research prévoit que le marché international du stockage du CIP pourrait connaître un TCAC de 7 %, passant de 5,5 milliards de dollars américains en 2018 à 7,0 milliards de dollars américains en 2023. [134] Les méthodes de simulation itératives et la croissance des charges de travail en données massives et IA, qui génèrent des quantités massives de fichiers et de volumes de données, exigent une meilleure gestion des données pour une utilisation plus efficace de la capacité de stockage et de meilleures performances. La plupart des utilisateurs du stockage de CIP estiment que le taux de croissance annuel de leur capacité est inférieur à 50 %, dont la majorité se situe entre 25 et 50 %. Bien que les paragraphes suivants décrivent les défis techniques, le principal obstacle de stockage est le recrutement et l'embauche de personnel hautement qualifié. [131]

---

[132] Hyperion Research at High Performance Computing AWS Conference: HPC in the cloud with Hyperion Research https://hpcaws.splashthat.com/ (novembre 2020).

[133] MarketsandMarkets : HPC, Data Analysis, Storage & Management Market in Life Sciences By Products & Services (Data Analysis, Cloud Computing), Applications (NGS, Microscopy, Chromatography), End User (Pharmaceutical & Biotechnology, Hospitals) - Global Forecast to 2024 https://www.marketsandmarkets.com/Market-Reports/hpc-data-analysis-storage-management-market-47829739.html (avril 2019, sommaire public consulté en 2020).

[134] Hyperion Research white paper (sponsored by Panasas): New Study Details Importance of TCO for HPC Storage Buyers https://www.panasas.com/wp-content/uploads/2020/04/Hyperion_Importance-of-TCO-for-HPC-Storage-Buyers_Q1-20_FINAL_2020-04-22.pdf (avril 2020).

L'intelligence artificielle, l'apprentissage machine et l'apprentissage profond sont axés sur les caractéristiques de stockage de CIP existantes, tout en établissant de nouveaux flux de travail de stockage qui sont différents du CIP traditionnel. Il est important de considérer les éléments essentiels suivants 1) réseaux haute performance, 2) stockage partagé, 3) stockage à plusieurs niveaux, 4) accès parallèle, 5) prise en charge de plusieurs protocoles et 6) gestion avancée des métadonnées. [135]

Les réseaux haute performance sont bien sûr une source de revenus importante pour les systèmes de CIP traditionnels, mais pour le stockage, ces réseaux peuvent être conçus pour des solutions mémoires de stockage en réseau à faible latence, comme NVMe Over Fabrics (NVMe-oF). Le stockage partagé, où tous les chargements ont accès à l'ensemble du stockage, est une caractéristique commune du stockage de CIP. Pour l'IA en revanche, ces réseaux sont très importants pour répartir les chargements plus uniformément et pour maximiser l'utilisation des unités GPU coûteuses. Le stockage à plusieurs niveaux à l'ère de l'intelligence artificielle et de l'apprentissage machine (IA/AM) doit tenir compte des caractéristiques (contradictoires) suivantes : les ensembles de données d'apprentissage sont souvent énormes, il faut y accéder avec des E/S très rapides quand on les utilise et il faut les conserver longtemps, car il est coûteux, voire impossible, de recueillir de nouveau ces données. D'une part, le coût des solutions entièrement Flash est parfois prohibitif, ce qui nécessite une solution hybride à plusieurs niveaux avec de la mémoire continue, Flash, ainsi qu'un ensemble de disques.[136] D'autre part, ces systèmes de stockage à plusieurs niveaux peuvent être déroutants et peu pratiques pour les utilisateurs finaux. Pour remédier au problème, le futur superordinateur Perlmutter de NERSC prévoit uniquement deux niveaux de stockage pour ces utilisateurs : la hiérarchisation sera gérée automatiquement en arrière-plan par des solutions logicielles.[137] L'accès parallèle est une autre caractéristique commune au stockage du CIP, mais ce besoin est aussi accentué par l'ampleur des processus de calcul dans les charges de travail typiques de l'IA/AM, qui exigent un accès simultané aux (mêmes) données. La prise en charge de plusieurs protocoles, par exemple FS NFS, SMB et S3 en parallèle, est nécessaire, car les données d'IA/AM sont parfois recueillies à partir de dispositifs ayant leurs propres protocoles, notamment des dispositifs connectés au réseau de l'internet des objets, alors que d'autre part les données sont consommées dans le système de stockage massivement parallèle du CIP. La gestion avancée des métadonnées n'est pas propre à l'IA/AM, mais les milliards de fichiers auxquels sont attachées des métadonnées exercent une pression supplémentaire sur les E/S du système de stockage. [132]

Échelle exa

Au niveau international, la communauté du CIP vise des performances qui se situent dans l'échelle exa, même s'il ne s'agit pas d'un objectif fondamental ou réaliste à moyen terme au Canada. Actuellement, les travaux à très grande échelle et massivement parallèles ne sont pas si fréquents dans les systèmes de CIP. 96 % des charges de travail exécutées sur les systèmes

[135] Storage Switzerland LLC White Paper (financé par Panasas): Is Tour Storage Infrastructure Ready for the coming AI Wave? https://insidehpc.com/white-paper/is-your-storage-infrastructure-ready-for-the-coming-ai-wave/ (janvier 2020).

[136] The Next Platform: Divide Deepens Between HPC and Enterprise Storage https://www.nextplatform.com/2020/10/20/divide-deepens-between-hpc-and-enterprise-storage/?mc_cid=61a88daab6&mc_eid=79a1266800 (consulté en octobre 2020).

[137] HPCWire: SC20 Panel – OK, You Hate Storage Tiering. What's Next Then? https://www.hpcwire.com/2020/11/25/sc20-panel-ok-you-hate-storage-tiering-whats-next-then/ (consulté en novembre 2020).

de CIP sont inférieures à 2048 cœurs et 72 % sont inférieures à 1024 cœurs. Il faut garder à l'esprit que la plupart des grappes de CIP peuvent en théorie exécuter des tâches de plus de 30 000 cœurs. En revanche, il faut approfondir la question pour déterminer si cette situation est simplement due à un manque d'accès aux ressources à plus grande échelle ou à un manque de codes massivement parallèles évolutifs pouvant tirer parti d'une telle échelle. Pour que les chercheurs canadiens puissent profiter de l'informatique à l'échelle exa massivement parallèle, il pourrait être avantageux que le Canada fasse partenariat avec des fournisseurs aux États-Unis afin d'exécuter ces travaux d'une telle ampleur.

Selon Nic Dube, technologue en chef de HPE/Cray, « rien n'est simple à l'échelle exa ».[138] Les principaux défis découlent des exigences en matière d'énergie, de logiciels et de résilience du système. Les besoins en énergie d'un système à l'échelle exa sont estimés à 30-40 MW, voire plus. Il faudra des mises à jour majeures pour le centre de données de l'infrastructure. Sur le plan logiciel, l'ampleur du parallélisme exigera des efforts de développement de code ciblés pour exécuter des applications réelles (en utilisant, par exemple, les bibliothèques CUDA qui font abstraction du parallélisme dans une certaine mesure) et non uniquement des codes Linpack de type « trophée » qui s'adressent aux systèmes du classement Top500. En ce qui concerne la résilience du système, le travail sur les logiciels système effectué dans les centres de superordinateurs américains offre une « apparence de résilience » aux utilisateurs finaux, même si les composants sous-jacents ne sont pas résilients (ou n'ont même pas besoin de l'être). [135]

Pour atteindre l'échelle exa, les deux approches architecturales dominantes reposent sur l'utilisation exclusive de cœurs de CPU ou la combinaison de cœurs de CPU et d'accélérateurs GPU.[139] Le plus important système du palmarès Top500, le Fugaku du Japon basé sur l'architecture ARM, est un exemple de la première approche. En revanche, le processeur sur mesure du Fugaku est à la fois un processeur ARM à plusieurs cœurs et un processeur accéléré de type GPU, qui maintient à la fois un appui massif pour les applications grâce aux principes de conception « primauté de l'application » et « positionnement comme leader du HPC ». [140]

Les récents systèmes américains de premier plan (Summit et Frontera) sont des exemples de la seconde approche (CPU+GPU). Par exemple, le système « Leonardo » d'EuroHPC JU, doté de plus de 200 Pf de pointe et installé à CINECA, reposera sur une architecture CPU Intel et de GPU Nvidia A100 dans un rapport d'un à quatre.[141] Il est intéressant de noter que le nouveau système Leonardo provient d'Atos/Bull, ce qui témoigne de la croissance rapide de sa présence sur le marché du CIP et fait concurrence aux grands acteurs habituels (HPE/Cray, Dell, Lenovo et IBM).[142] Par rapport à l'approche « cœurs + GPU », l'approche « cœurs exclusifs » se caractérise

---

[138] InsideHPC: Getting to Exascale: Nothing Is Easy https://insidehpc.com/2020/10/getting-to-exascale-nothing-is-easy/ (consulté en octobre 2020).

[139] Prof. Mark Parsons of EPCC at HPC-AI Advisory Council 2020 UK Conference: Exascaling AI http://www.hpcadvisorycouncil.com/events/2020/uk-conference/pdf/day-one/M_Parsons_ExascalingAI_131020.pdf (octobre 2020).

[140] SC20 conference presentation by Satoshi Matsuoka: Fugaku: 'Exascale' and 'Applications First' https://cdmcd.co/Kra6dr (novembre 2020).

[141] HPCWire : Nvidia and EuroHPC Team for Four Supercomputers, Including Massive 'Leonardo' System https://www.hpcwire.com/2020/10/15/nvidia-and-eurohpc-team-for-four-supercomputers-including-massive-leonardo-system/ (consulté en octobre 2020).

[142] The Next Platform : With Another Key Supercomputer Win, Atos Looks Stronger Than Ever https://www.nextplatform.com/2020/10/15/with-another-key-supercomputer-win-atos-looks-stronger-than-ever/?mc_cid=ee478a79f8&mc_eid=79a1266800 (consulté en octobre 2020).

par une durée de vie plus longue des systèmes, une mise en œuvre plus facile des logiciels pour les codes de simulation traditionnels, des performances inférieures en matière d'intelligence artificielle et des besoins plus importants en termes d'espace physique et de puissance. [136] Afin d'équilibrer les besoins des différents utilisateurs, EuroHPC adopte une approche hybride avec son futur superordinateur pré -exa LUMI. Ce dernier comportera une grappe de nœuds CPU exclusivement avec diverses allocations de mémoire, puis une très grande grappe de nœuds CPU+GPU, tous connectés via une interconnexion Ethernet à haute vitesse partagée. Le système devrait offrir une capacité de pointe de 552 PF et devrait être construit en 2021. [143] La famille de processeurs Epyc d'AMD et ses GPU Instinct gagnent en popularité, principalement grâce à la densité de calcul et à la rentabilité d'Epyc par rapport à Xeon d'Intel. Par exemple, les systèmes Frontier et El Capitan aux États-Unis, ainsi que le nouveau superordinateur du Pawsey Supercomputing Center à Perth, Australie, utiliseront ces puces. [144]

Si l'on considère les tendances technologiques à plus long terme qui sont nécessaires pour rendre possible l'informatique « post-exa », l'opinion dominante (surtout aux États-Unis) estime que les améliorations à l'échelle doivent se transformer en améliorations de développement : le ralentissement la loi de Moore, avec la diminution l'échelle d'améliorations de la densité des transistors, unités d'exploitation, fréquence d'horloge, efficacité énergétique et nombre de cœurs par connecteur logiciel, nécessitera des innovations au-delà de ces approches « traditionnelles ». [145]

Les efforts de développement de nouvelles architectures peuvent être organisés en trois catégories (qui se chevauchent) : 1) construction de systèmes à usage spécial, 2) conception de systèmes de puces avec intégration hétérogène de plusieurs accélérateurs sur mesure pour des « micro tâches » précises et 3) meilleure adaptabilité de la charge de travail du système de CIP par la désagrégation de ressources. La première catégorie comprend des systèmes sur mesure semblables à la série Anton de D. E. Shaw Research pour la dynamique moléculaire[146], à l'ASIC TPU de Google pour l'apprentissage machine[147], aux puces neuromorphiques personnalisées pour l'IA basée sur les réseaux neuronaux à impulsions, ou potentiellement à un superordinateur sur mesure pour les calculs basés sur la théorie de la fonctionnelle de la densité (DFT) qui représentent actuellement 25 % de la charge de travail du NERSC. [142]

La deuxième catégorie (intégration hétérogène d'accélérateurs sur mesure) implique l'agrégation de puces d'accélération hautement sur mesure située immédiatement à côté de la puce du

[143] The Next Platform : The Resurrection Of Cray And AMD In A Trifurcating HPC Space https://www.nextplatform.com/2020/10/22/the-resurrection-of-CIPy-and-amd-in-a-trifurcating-hpc-space/?mc_cid=61a88daab6&mc_eid=79a1266800 (consulté en octobre 2020).

[144] The Next Platform: HPE and AMD Bag The Big Supercomputer Deal Down Under https://www.nextplatform.com/2020/10/19/hpe-and-amd-bag-the-big-supercomputer-deal-down-under/?mc_cid=ee478a79f8&mc_eid=79a1266800 (consulté en octobre 2020).

[145] John Shalf of Lawrence Berkeley National Laboratories at Oklahoma Supercomputing Symposium 2020: Pathfinding for Post-Exascale HPC http://www.oscer.ou.edu/Symposium2020/oksupercompsymp2020_talk_shalf_20200930.pdf (septembre 2020).

[146] Pittsburgh Supercomputing Centre: Anton https://psc.edu/resources/computing/anton (consulté en octobre 2020).

[147] Google Cloud: Cloud TPU https://cloud.google.com/tpu/ (consulté en octobre 2020).

processeur principal (au-delà des accélérateurs GPU à usage général intégrés habituels).[148] Parmi les exemples commerciaux existants, citons les puces Bionic d'Apple, hautement sur mesure, pour l'apprentissage automatique et l'analyse du mouvement dans les téléphones intelligents,[149] [150]ou la gamme de puces ARM sur mesure AWS Graviton d'Amazon pour les charges de travail en nuage.[151] Le programme Networking and Information Technology Research and Development (NITRD) aux États-Unis étudie les possibilités d'applications du CIP dans le cadre de son projet 38.[152] Une technologie intéressante connexe est l'unité de traitement des données (DPU), une unité de silicium sur puce (SoC) qui se trouve à côté des CPU et GPU traditionnels et qui combine des cœurs de processeur ARM avec des capacités réseau et GPU intégrées pour décharger certaines des tâches de calcul et de gestion réseau de la CPU vers la DPU. [153]

La troisième catégorie (désagrégation de ressources) vise à créer des nœuds de calcul reconfigurables avec une interconnexion à très haut débit entre les composants clés. Une interconnexion interne à la vitesse de la bande passante photonique permettrait une allocation flexible et une (re)configuration de bas niveau des ressources mémoire, CPU, GPU, E/S et réseau, permettant à un seul système d'exécuter efficacement diverses charges de travail (par exemple, l'apprentissage de l'intelligence artificielle, l'inférence de l'intelligence artificielle, l'exploration de données ou l'analyse de graphes) avec des besoins très différents en matière de mémoire, de calcul, de réseau et d'E/S. Une telle conception pourrait être basée, par exemple, sur des modules multipuces photoniques (MCM), comme le propose le projet PINE (Photonic Integrated Networked Energy Efficient Datacenter) de l'Université Columbia. [154] Un autre avantage de la désagrégation matérielle est la souplesse d'adaptation aux différentes charges de travail, par exemple entre les charges de travail traditionnelles du CIP, les charges de travail à haut débit (HTP) et l'IA. Aux États-Unis, le ministère de la Défense (DoD) a choisi d'acquérir des

[148] Prof. Simon McIntosh-Smith of University of Bristol at HPC-AI Advisory Council 2020 UK Conference: Exascale Research and Development Opportunities http://www.hpcadvisorycouncil.com/events/2020/uk-conference/pdf/day-one/S_McIntoshSmith_ExaRandDOpps_131020.pdf (octobre 2020).

[149] Apple : Apple unveils all-new iPad Air with A14 Bionic, Apple's most advanced chip https://www.apple.com/newsroom/2020/09/apple-unveils-all-new-ipad-air-with-a14-bionic-apples-most-advanced-chip/ (consulté en octobre 2020).

[150] Wired: An Exclusive Look Inside Apple's A13 Bionic Chip https://www.wired.com/story/apple-a13-bionic-chip-iphone/ (consulté en octobre 2020).

[151] ZDNet : AWS Graviton2 : What it means for Arm in the data center, cloud, enterprise, AWS https://www.zdnet.com/article/aws-graviton2-what-it-means-for-arm-in-the-data-center-cloud-enterprise-aws/ (consulté en octobre 2020).

[152] NITRD Project 38 Technical Report: HPC Performance Improvements Through Innovative Architecture https://www.nitrd.gov/Presentations/files/HPC-Performance-Improvements-Project-38.pdf (octobre 2019).

[153] EnterpriseAI : Nvidia Expands Its DPU Family, Unveils New Datacenter on Chip Architecture https://www.enterpriseai.news/2020/10/05/nvidia-expands-its-dpu-family-unveils-new-datacenter-on-chip-architecture/ (consulté en octobre 2020).

[154] Columbia University in the City of New York: Photonic Integrated Networked Energy efficient datacenter (PINE) https://lightwave.ee.columbia.edu/research-projects/photonic-integrated-networked-energy-efficient-datacenter-pine (consulté en octobre 2020).

superordinateurs et des solutions désagrégées mettant l'accent sur la flexibilité plutôt que sur la mesure (plus traditionnelle) du nombre de flops par dollar. [155]

3.6 COVID-19

La COVID-19 a considérablement accru l'intérêt pour l'infonuagique, sur le marché général et sur celui du CIP. Les clients qui avaient déjà recours à ces services (notamment pour des raisons expérimentales) les utilisent davantage parce que la demande d'informatique générale a augmenté, ce qui déplace cette demande vers le nuage et accélère le taux d'adoption par rapport aux périodes antérieures. On adopte plus rapidement l'infonuagique aussi parce que cette technologie permet de gérer la volatilité accrue des besoins informatiques. La communauté scientifique s'est beaucoup investie dans la recherche sur la COVID-19, grâce aux ressources spécialisées de CIP et d'IA dématérialisé, ainsi qu'aux subventions spéciales des fournisseurs commerciaux. En ce qui concerne l'infrastructure, de nombreuses sociétés et institutions qui envisageaient auparavant l'infonuagique pour le CIP comme un projet pluriannuel ont accéléré leurs tests et l'adoption à cause de l'évolution des exigences en milieu de travail, ainsi que des difficultés et des dépendances liées à la gestion des centres de données de CIP sur place. [156]

En ce qui concerne la fédération Calcul Canada, les principaux sites d'hébergement n'ont presque pas connu d'interruption à cause de la COVID-19 et ils ont été en mesure de fournir des services tout au long de la pandémie. La CIP a également beaucoup aidé les chercheurs canadiens dans ce contexte, notamment en « fournissant un accès aux ressources dématérialisées, aux grappes haute performance, au stockage ou en favorisant la priorité des tâches. Les chercheurs ont bénéficié de conseils en matière de calcul haute performance (HCP), de gestion de données, d'analyse de données, d'apprentissage machine et de visualisation. Ils ont aussi créé des liens avec des scientifiques canadiens dans d'autres domaines et avec des établissements de recherche afin de poursuivre leurs collaborations ». [157]Calcul Canada a également contribué aux ressources du projet international de calcul distribué folding@home en optimisant le code pour GPUS et par des simulations de la structure des protéines du SRAS-CoV-2 sur le nuage Arbutus et le superordinateur Cedar.[158] À la fin du mois de septembre 2020, le CIP avait fourni 27 allocations à des projets sur la COVID-19 qui ont nécessité environ 4 600 années-cœurs de CPU d'accès prioritaire aux cycles de calcul. À cette date, les projets de recherche avaient utilisé 1300 années CPU de ces ressources. Cette utilisation représente environ 0,7 % du total des ressources en années CPU accessibles en 2020 pour le CIP. [159] En revanche, ce chiffre ne tient pas compte de toutes les recherches sur la COVID-19 effectuées sur les systèmes de la FCC, mais uniquement des groupes de recherche qui ont demandé des

[155] TheNextPlatform : For HPC And AI, Composability Might Trump Cheap Flops https://www.nextplatform.com/2020/10/27/for-hpc-and-ai-composability-might-trump-cheap-flops/?mc_cid=95e0f6bf8a&mc_eid=79a1266800 (consulté en novembre 2020).

[156] Altair HPC Virtual Summit 2020 – Cloud Roundtable: Is Cloud Officially Inevitable? Experts from Azure, Oracle, Advania and Google get candid about 2020's biggest cloud computing trends and challenge: https://player.vimeo.com/video/455986574 (septembre 2020).

[157] Compute Canada: Support for COVID-19 Projects https://www.computecanada.ca/featured/support-for-covid-19-research-projects/ (consulté en septembre 2020).

[158] Compute Canada: Harnessing Power of Scientific Cloud Computing to Fight COVID-19 https://www.computecanada.ca/featured/harnessing-the-power-of-scientific-cloud-computing-to-fight-covid-19/ (consulté en septembre 2020).

[159] Source: Base de données de Calcul Canada, fournie par Maxime Boissonneault (octobre 2020).

ressources supplémentaires ou une priorisation au-delà de leur allocation habituelle. Pour illustrer l'ampleur de ces activités de recherche, la communauté et les chercheurs dans le domaine du CIP ont été mentionnés dans au moins 70 articles et messages sur les médias sociaux liés à COVID-19 entre mars et octobre 2020. [160]Deux cas d'utilisation du CIP ont eu un impact direct sur la prise de décision et ces projets n'étaient pas de nature exploratoire : toutes les modélisations effectuées pour l'INSPQ (Institut national de santé publique du Québec), qui oriente les décisions en matière de santé publique, ont été réalisées par un groupe de l'Université Laval, avec le soutien de Charles Coulombe, un analyste de CQ. Ce groupe de recherche a utilisé Graham [161]. Un groupe McGill effectue un séquençage pour repérer des variants de la COVID-19 sur Béluga, avec l'aide d'un analyste de CQ. [162]

À l'échelle mondiale, divers instituts et sociétés ont également augmenté leurs investissements dans le CIP à cause de la COVID-19. Par exemple, AMD a donné 5 pétaFlops de puissance de calcul pour la recherche universitaire sur la COVID-19 en septembre 2020, [163]en plus d'un important don de matériel pour la nouvelle initiative SciNet4Health de SciNet.[164] En novembre 2020, le DOE américain a acquis un superordinateur grâce au financement en vertu de la loi CARES (Coronavirus Aid, Relief and Economic Security). Le système à grande mémoire Mammoth sera situé aux Lawrence Livermore National Labs (LLNL) et il compte 64 nœuds de serveur jumeaux basés sur AMD Epyc, dotés chacun de 2 To de RAM et de 4 To de mémoire non volatile. Ils sont reliés par une interconnexion à haut débit Omnipath. Le système est conçu pour la recherche sur la COVID-19, dont l'analyse génomique, les simulations de CIP non traditionnelles et l'analyse de graphes. [165] Le superordinateur le plus rapide du monde, le Fugaku du RIKEN, a été mis en ligne près d'un an avant la date prévue afin d'entreprendre des recherches sur la COVID-19. Le superordinateur est utilisé à grande échelle pour étudier, par exemple, la transmission par les gouttelettes, la ventilation, le niveau de filtrage des masques et les effets de l'humidité sur la viabilité du coronavirus, ce qui a eu un impact direct sur l'élaboration des politiques japonaises. [166]Aux États-Unis, la capacité du neuvième superordinateur le plus

[160] Analyse de l'équipe de communication de Calcul Canada fournie par Maxime Boissonneault (novembre 2020).

[161] Québec INSPQ : Épidémiologie et modélisation de l'évolution de la COVID-19 9 avril 2021 - Mise à jour des projections du 18 mars https://www.inspq.qc.ca/covid-19/donnees/projections/9-avril-2021 (récupéré en avril 2021).

[162] Calcul Québec : COVID-19 : Transparence en matière de données https://www.calculquebec.ca/recherche/covid-19-transparence-en-matiere-de-donnees-publiques/ (consulté en avril 2021).

[163] AMD : AMD COVID-19 HPC Fund Adds 18 Institutions and Five Petaflops of Supercomputer Power Processing Power to Assist Researchers Fighting COVID-19 Pandemic https://www.amd.com/en/press-releases/2020-09-14-amd-covid-19-hpc-fund-adds-18-institutions-and-five-petaflops (consulté en septembre 2020).

[164] University of Toronto: U of T and AMD launch supercomputing program dedicated to big-data health research https://www.utoronto.ca/news/u-t-and-amd-launch-supercomputing-program-dedicated-big-data-health-research (consulté en septembre 2020).

[165] HPCWire: Lawrence Livermore Announces Mammoth Cluster to Fight COVID-19 https://www.hpcwire.com/2020/11/04/lawrence-livermore-announces-mammoth-cluster-to-fight-covid-19/ (consulté en novembre 2020).

[166] HPCWire: It's Fugaku vs. COVID-19: How the World's Top Supercomputer Is Shaping Our New Normal https://www.hpcwire.com/2020/11/09/its-fugaku-vs-covid-19-how-the-worlds-top-supercomputer-is-shaping-our-new-normal/ (consulté en novembre 2020).

rapide du monde, Frontera, augmentera d'environ 5 % en janvier 2021 grâce à une subvention spéciale de la National Science Foundation (NSF) et un don de Dell. [167]

3.7 Retour sur investissement en milieu universitaire du CIP

Le CIP et l'IRN ont de nombreux avantages pour les chercheurs, la science, les sociétés et les industries. Le calcul informatique de pointe permet de résoudre des problèmes et des questions scientifiques qui seraient autrement difficiles à résoudre, mais il contribue aussi à des solutions qui ne seraient pas possibles par des moyens ordinaires (calcul analytique, expérimental ou au niveau des postes de travail par exemple). Les problèmes résolus grâce au CIP vont de l'échelle nanométrique, comme la découverte de médicaments, à l'échelle macroscopique, comme les simulations de phénomènes météorologiques violents ou des changements climatiques. Il a souvent un impact direct sur les individus et la société. De plus, le CIP devient actuellement une composante essentielle des sciences sociales et des humanités numériques. La concurrence de l'industrie canadienne dépend de l'efficacité et de la valeur ajoutée de ses produits, dont la conception est souvent facilitée par les systèmes de CIP. Une entreprise peut par exemple tester diverses versions de produits à l'aide du CIP, puis construire uniquement les produits candidats les plus intéressants, ce qui réduit les coûts et raccourcit les délais de mise en marché.

Dans le cadre d'une étude plus approfondie sur les avantages du CIP, il faudrait également tenir compte des coûts afférents, en examinant le retour sur investissement du CIP, notamment en milieu universitaire. Il s'agit là d'un exercice très complexe, car le rendement monétaire et non monétaire est très difficile à quantifier. Concrètement, il est aussi dur d'estimer avec précision la dimension d'investissement.

La Coalition for Academic Scientific Computation (CASC), basée aux États-Unis, a récemment étudié le retour sur investissement pour les universités en termes de coût total de possession (CTP) et d'avantages financiers et non financiers. Pour le dénominateur (coûts), il faut prendre en compte non seulement les coûts d'investissement immédiats (par exemple, l'achat de matériel et de logiciels, les garanties, les licences et l'amortissement), mais aussi les coûts d'exploitation permanents (formation, personnel, licences, alimentation, refroidissement, réseau, maintenance, sécurité, surveillance et facturation), les coûts de construction et d'infrastructure des installations et l'amortissement correspondant.[168] Au-delà du seul coût explicite, il faut également prendre en compte et définir le champ d'application, par exemple si les installations sont exploitées ou financées conjointement.

Du côté du numérateur, il faut considérer au moins les éléments suivants en ce qui concerne les avantages financiers potentiels du CIP :

[167] InsideHPC: TACC's Frontera HPC System Expansion for 'Urgent Computing' – COVID-19, Hurricanes, Earthquakes https://insidehpc.com/2020/11/taccs-frontera-hpc-system-expansion-for-urgent-computing-covid-19-hurricanes-earthquakes/ (consulté en novembre 2020).

[168] Craig E. Stewart et al : Assessment of financial returns on investments in cyberinfrastructure facilities: A survey of current methods - PEARC '19 : Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) July 2019 Article No. : 33 Pages 1-8 https://doi.org/10.1145/3332186.3332228 (juillet 2019).

1. Avantages pour l'utilisateur final des installations du CIP dans la recherche (temps gagné)
2. Ressources du système de CIP (économies par rapport aux solutions alternatives)
3. Ressources humaines (valeur du soutien du fournisseur de CIP)
4. Valeur de la formation
5. Revenus de subventions (valeur monétaire des revenus de subventions reçus par rapport aux occasions manquées)
6. Produits et brevets (revenu monétaire)
7. Impact économique (avantages financiers régionaux indirects, emplois et rentrées fiscales). [162]

On pourrait cependant faire valoir que les véritables avantages du CIP et de l'IRN en général se situent du côté non financier, c'est-à-dire au niveau des impacts et des résultats, où les avantages indirects et à long terme peuvent potentiellement être presque incalculables, si l'on considère par exemple le développement de nouveaux vaccins qui sauvent des vies, ou l'utilisation mentionnée plus haut de la modélisation pour informer les décideurs dans le contexte de la COVID-19. Voici plus concrètement une liste de ces avantages :

1. Nouvelles découvertes rapportées dans les médias (amélioration de la qualité de vie)
2. Personnes formées dans de nouveaux domaines (main-d'œuvre mieux formée pour l'économie)
3. Prix, communiqués de presse (réputation des individus et organisations concernés)
4. Brevets (produits améliorant la qualité de vie ou la protection de la vie terrestre). [169]

Du côté des entreprises, le US Council on Competitiveness a mené une étude sur les avantages du CIP auprès de ses membres. Le rapport Solve indique que les meilleures mesures pour justifier l'investissement de l'industrie dans le CIP étaient le « temps nécessaire pour trouver une solution », « l'impossibilité de résoudre le problème par d'autres moyens », le « retour sur investissement » et la « réduction des coûts par rapport aux méthodes physiques ».[170]Au-delà des avantages du CIP, les estimations du retour sur investissement du CIP dans le contexte commercial varient considérablement. Par exemple, Hyperion Research rapporte un retour moyen de 44 USD sur chaque dollar investi (c'est-à-dire un retour sur investissement irréaliste

[169] Craig E. Stewart et al: Assessment of non-financial returns on cyberinfrastructure: A survey of current methods - HARC '19 : Proceedings of the Humans in the Loop: Enabling and Facilitating Research on Cloud Computing July 2019 Article n° : 2 Pages 1-10 https://doi.org/10.1145/3355738.3355749 (juillet 2019).

[170] US Council on Competitiveness: Solve. The Exascale Effect: The Benefits of Supercomputing Investment of U.S. Industry https://www.compete.org/reports/all/2695 (octobre 2014).

d'environ 4300 %),[171] tandis qu'Intersect360 Research affirme que le véritable retour sur investissement doit être beaucoup plus faible et qu'il est très complexe à estimer. [172]

4 État actuel

Le CIP et l'IRN sont essentiels pour un nombre important et sans cesse croissant de chercheurs en raison des avancées technologiques et des nouveaux paradigmes de recherche. Ces tendances ne se manifestent pas uniquement dans les sciences exactes traditionnelles, mais aussi dans d'autres disciplines, telles que l'intelligence artificielle, le traitement du langage naturel, l'analyse des médias sociaux, l'analyse d'enquêtes qualitatives et quantitatives à grande échelle et le séquençage génétique.

Il est difficile de mesurer l'ampleur et la portée des activités de CIP au Canada parce que ces entreprises sont très largement répandues. Les chercheurs canadiens utilisent des ressources de CIP au sein de groupes de recherche, de départements, d'établissements (universités, collèges, hôpitaux de recherche), d'instituts de recherche (IORC, OBI, etc.), en plus d'y avoir recours à l'échelle provinciale, nationale et internationale. Certaines de ces ressources sont réservées à un groupe, un institut ou une discipline, tandis que beaucoup sont partagées à un certain niveau. Parfois elles sont principalement destinées à des organismes gouvernementaux (par exemple, les ressources hébergées par SPC et utilisées par ECCC), mais elles sont également utilisées par des universitaires. Cette diversité doit être considérée comme une force de l'écosystème canadien des CIP, car elle permet de répondre à des besoins géographiques, techniques et propres à des domaines diversifiés et complexes.

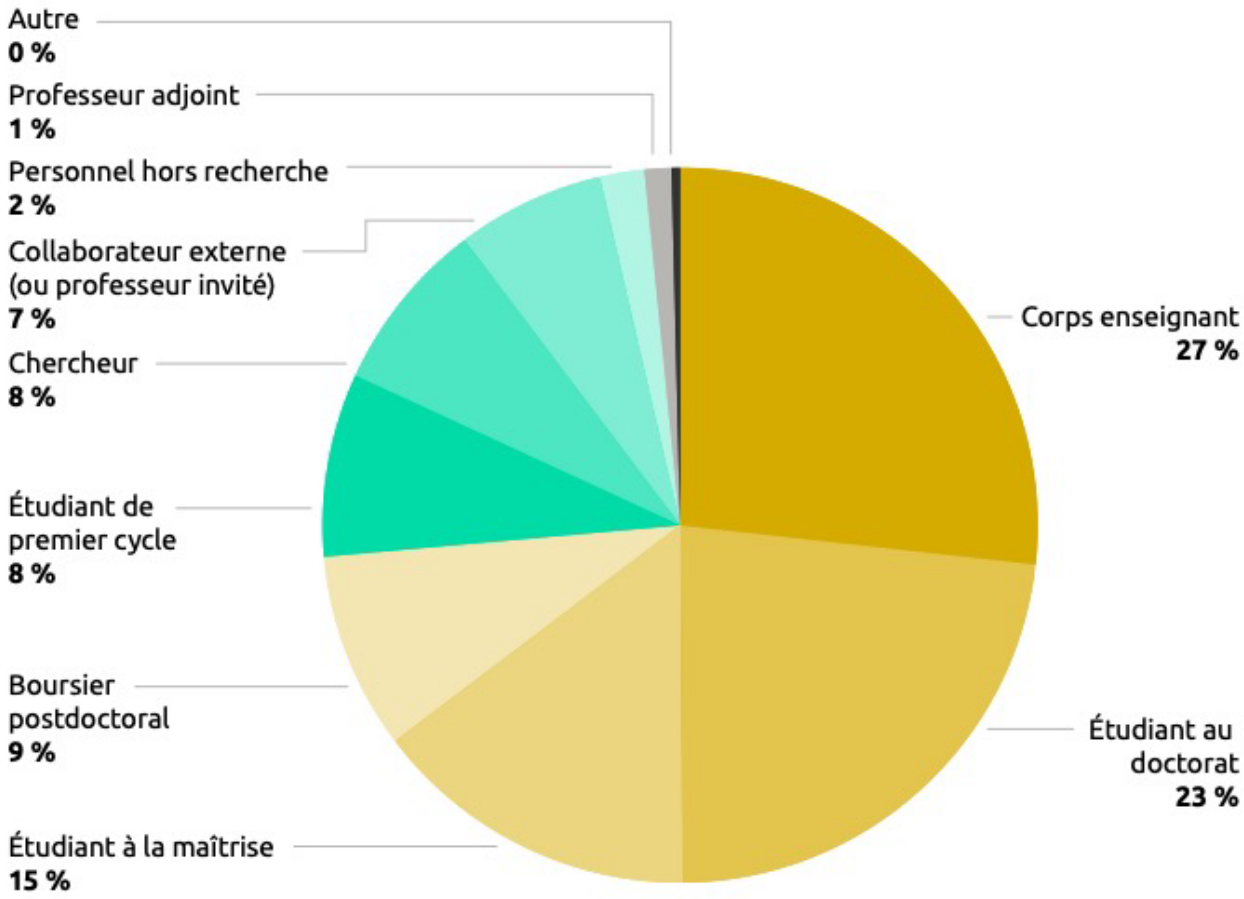
Les ressources de la FCC sont les seules ressources de CIP qui sont accessibles à tous les chercheurs universitaires canadiens. L'utilisation du CIP en milieu universitaire au Canada est relativement bien connue au sein de l'écosystème principal de la FCC. L'analyse ci-dessous est fondée sur l'utilisation des installations de la FCC uniquement. Il est important de souligner que plusieurs chercheurs ne connaissent pas l'existence des ressources de la FCC, ou ils pensent que ces ressources ne leur sont pas destinées ou ils utilisent les autres systèmes décrits plus haut. Les données suivantes ne tiennent pas compte de ces utilisations.

[171] Hyperion Research : HPC Investments Bring High Returns https://www.dellemc.com/resources/en-us/asset/analyst-reports/products/ready-solutions/hyperion-hpc-investment-brings-high-returns.pdf (juillet 2020).

[172] HPCWire: ROI: Is HPC Worth It? What Can We Actually Measure? By Addison Snell, Intersect360 Research https://www.hpcwire.com/2020/10/15/roi-is-hpc-worth-it-what-can-we-actually-measure/ (consulté en octobre 2020).

4.1 Utilisateurs inscrits aux systèmes de la FCC

Répartition par occupation



Graphique 2 : Occupation des utilisateurs inscrits de la FCC

Le graphique 2 ci-dessus montre les diverses occupations des utilisateurs inscrits à la FCC depuis le 1er janvier 2020. Les utilisateurs finaux précisent leur occupation sur une liste de la FCC dans sa base de données centrale des comptes utilisateurs lors de leur renouvellement annuel. Ils sont également validés par un membre de l'équipe de la FCC. La plupart des occupations sont explicites, mais il n'y a pas de distinction entre les différents postes de professeur (par exemple, assistant, titulaire, titulaire d'une chaire de recherche du Canada, etc.). De plus, le terme « chercheur » peut s'appliquer de manière générale et il n'est pas strictement défini dans ce contexte.

Au total, près de 16 000 utilisateurs « inscrits » ont été répertoriés dans la base de données de la FCC. Ces derniers sont définis comme étant des utilisateurs qui, au cours du processus de renouvellement annuel de leur compte, se sont inscrits à la base de données de Calcul Canada. Ils précisent ainsi qu'ils souhaitent que leur compte reste actif, qu'ils soient connectés ou non à un système (la FCC appelle parfois ces utilisateurs des « utilisateurs actifs »). Les pourcentages

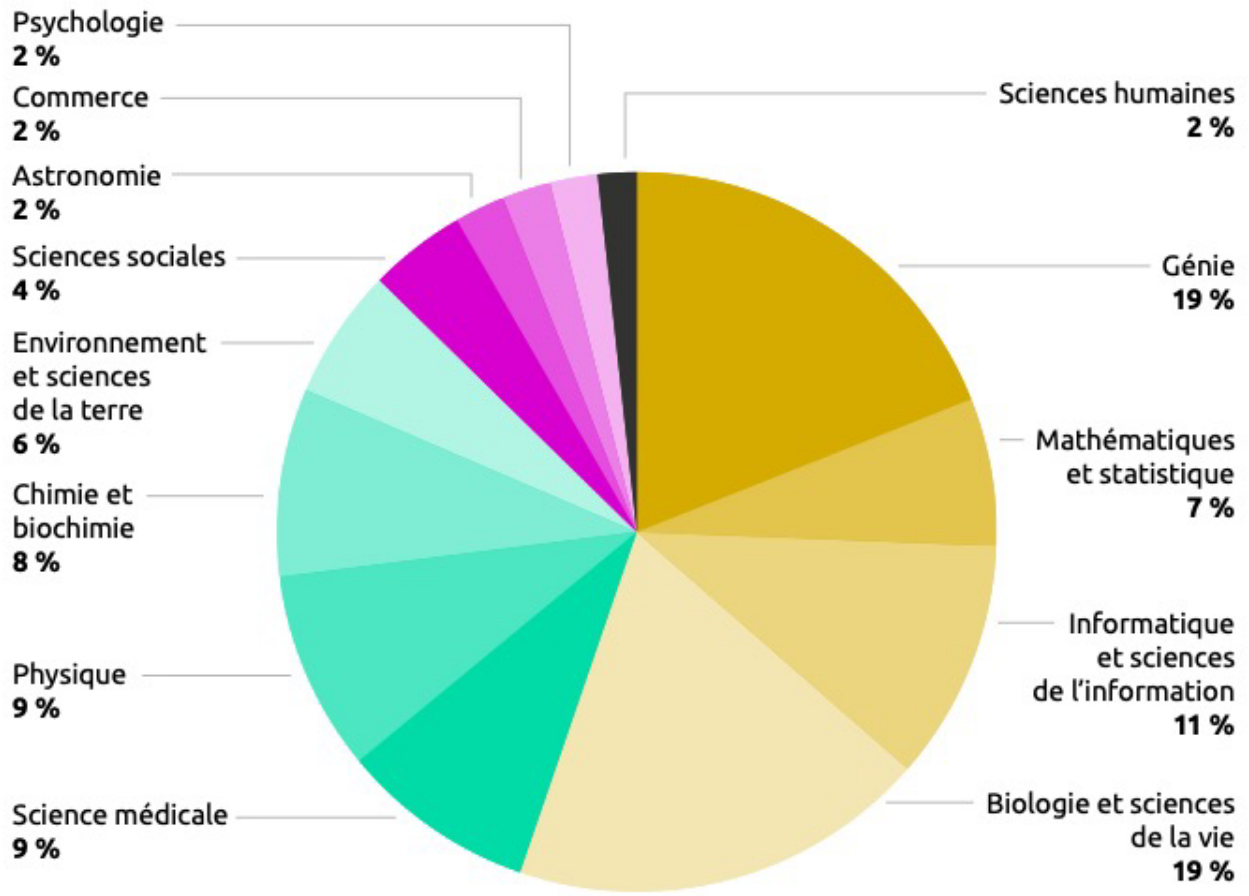
représentent le nombre d'occupations dans la base d'utilisateurs et n'indiquent pas dans quelle mesure ces utilisateurs ont utilisé les ressources.

Le bassin d'utilisateurs de la banque de données de Calcul Canada (BDCC) connaît un roulement important d'une année à l'autre, ce qui indique que plusieurs personnes évoluent dans le milieu du CIP même si elles quittent le milieu universitaire. Le taux de roulement élevé nécessite une formation continue sur une base annuelle ou même plus souvent. Cette proximité au développement d'une main-d'œuvre qui connaît les outils d'IRN et sait comment les utiliser. Le taux de roulement est estimé à 20-50 % par an, ce qui indique que le total cumulatif de personnes uniques qui travaillent à proximité du CIP est important.

Le groupe d'utilisateurs le plus important est celui des professeurs (27 %), suivi des doctorants et des étudiants à la maîtrise (23 % et 15 % respectivement). Soulignons que pour la plupart des occupations, il faut être parrainé par le chercheur principal, ce qui explique en partie le pourcentage élevé de postes de professeurs. Si l'on inclut le quatrième plus important groupe d'utilisateurs, les étudiants postdoctoraux (9 %), ces quatre principaux groupes d'utilisateurs représentent 74 % de tous les utilisateurs de la FCC. Si l'on considère les chercheurs plus chevronnés, incluant les professeurs, professeurs adjoints, chercheurs et collaborateurs externes, ils représentent environ 43 % de tous les utilisateurs de la FCC. Si l'on incluait les étudiants postdoctoraux dans ce groupe, les chercheurs chevronnés représenteraient plus de la moitié de tous les utilisateurs. Les chercheurs en début de carrière, incluant les étudiants postdoctoraux et les étudiants de premier cycle, représentent 55 % de l'ensemble des utilisateurs. Quelle que soit la façon dont on considère l'ancienneté, environ la moitié des utilisateurs des systèmes de la FCC sont des chercheurs chevronnés et l'autre moitié des chercheurs en début de carrière.

Le niveau d'aisance d'une personne dans le domaine du CIP dépend beaucoup du stade de sa carrière. Il y a un besoin continu de formation au sein de l'IRN, qui s'explique par l'utilisation accrue l'IA dans de nombreux domaines de recherche ou simplement par une croissance exponentielle des données ou des besoins informatiques. Le CIP évolue constamment parce que la technologie, les logiciels et les méthodes de recherche s'améliorent et changent continuellement, de sorte que les chercheurs doivent continuellement mettre à jour leurs compétences. Par ailleurs, les besoins et les intérêts scientifiques individuels des chercheurs évolueront et nécessiteront des outils et un soutien plus ambitieux, plus évolués et plus ciblés en matière d'IRN.

Faculté par domaine de recherche



Graphique 3: Faculté par domaine de recherche

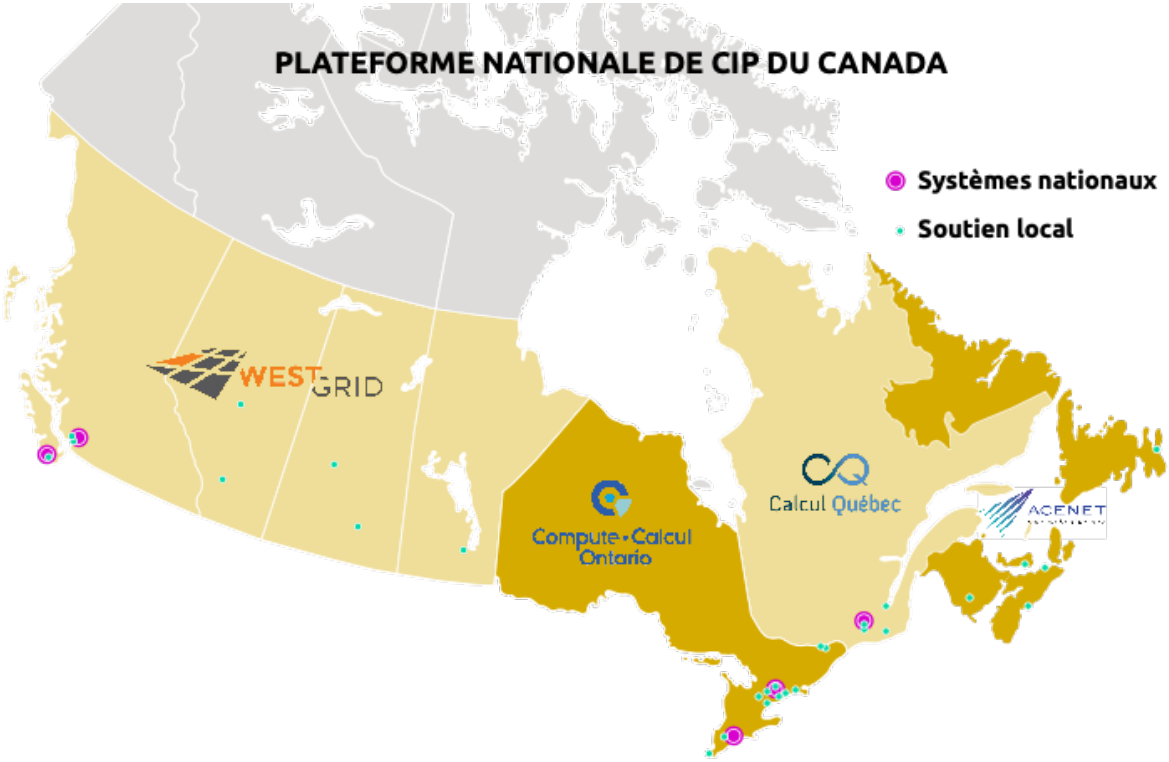
Le graphique 3 ci-dessus montre la répartition des postes de professeur parmi les utilisateurs de la FCC dans différentes disciplines de recherche. Au total, environ 4 200 utilisateurs se sont déclarés comme membres du corps enseignant en janvier 2020 (sur un total d'environ 16 000 utilisateurs, comme indiqué ci-dessus). Encore une fois, ces statistiques n'indiquent pas l'utilisation réelle (qui sera abordée à la section 4.3 ci-dessous), mais uniquement la répartition des disciplines de recherche avec les utilisateurs de la FCC. Les groupes d'utilisateurs les plus importants sont l'ingénierie, les sciences biologiques et de la vie, qui représentent 19 % respectivement. Environ la moitié des professeurs utilisateurs de la FCC se trouvent dans plusieurs disciplines comme les mathématiques, la statistique, l'informatique, les sciences de l'information, les sciences médicales, la physique, la chimie et la biochimie, qui représentent toutes entre 7 % et 11 % de la base d'utilisateurs. En ce qui concerne les groupes de disciplines, l'ingénierie, les mathématiques, la statistique et l'informatique totalisent 37 % de la base d'utilisateurs du corps professoral (c'est-à-dire un total de 1530 professeurs). Si l'on compare ceci avec le rapport du CLIRN en 2017, cette cohorte représentait 35 %[1]. Le pourcentage relatif des utilisateurs du corps professoral est donc resté à peu près constant entre 2017 et 2020. Les sciences biologiques, les sciences de la vie et les sciences médicales totalisent 28 % dans le cas précis et 26 % dans le rapport de 2017. Les domaines traditionnels du CIP, soit la physique,

l'astronomie, la chimie, la biochimie, les sciences de l'environnement et de la terre, représentent 25 % dans le cas précis et environ 30 % dans le rapport de 2017. La baisse de 5 % en trois ans est due à une diminution de quelques points de pourcentage du nombre relatif d'enseignants en chimie, en biochimie et en physique. Il convient de noter que certaines disciplines, par exemple les sciences médicales et les sciences sociales, nécessitent souvent l'utilisation de systèmes alternatifs tels que des infrastructures hautement sécurisées ou conformes à l'information personnelle sur la santé.

Le corps professoral en sciences humaines, sciences sociales, affaires et psychologie représente environ 10 % de la base d'utilisateurs de la FCC, alors que le corps professoral de ces disciplines (sciences humaines, sciences sociales et comportementales, droit, affaires, gestion et administration publique) représente environ 46 % de l'ensemble du corps professoral universitaire à temps plein au Canada[173], ce qui indique clairement que le nombre d'utilisateurs du CIP dans ces disciplines n'illustre pas leur représentation relative générale dans le milieu universitaire canadien. En revanche, il est important de garder à l'esprit que ces disciplines n'ont pas toutes accès également aux ressources de CIP. Elles ont augmenté leur utilisation du CIP depuis 2017, où elles représentaient environ 7 % de la base d'utilisateurs. La croissance de 3 points de pourcentage correspond à une croissance d'environ 40 % pour ce groupe d'utilisateurs sur trois ans, alors que le nombre absolu d'utilisateurs dans ces disciplines a encore une marge et un potentiel de croissance importants. Outre le nombre de professeurs, le nombre d'utilisateurs parrainés par un membre du corps professoral est plus faible en sciences sociales et dans les autres domaines sous-représentés. Il y a une différence entre la sous-représentation et les besoins non satisfaits, ces derniers n'étant qu'une des raisons potentielles de la première. Comme toutes les autres disciplines, les sciences sociales et humaines sont en général de plus en plus touchées par la technologie numérique. L'augmentation de la quantité de données accessibles et de contenus numérisés s'accompagne d'une utilisation croissante des méthodes informatiques. À titre d'exemple, le Canada se distingue au niveau international par ses recherches en humanités numériques. Dans ce domaine, l'offre actuelle de CIP/IRN ne couvre pas les besoins particuliers des utilisateurs. L'évolution des services pourrait certainement faciliter l'adoption chez de nombreux autres utilisateurs en sciences humaines et sociales dont les besoins ne sont pas suffisamment pris en compte.

[173] Canadian Association of University Teachers (CAUT) : Age Distribution of Full-Time University Teachers by Age, Sex and Major Discipline, 2017-2018 https://www.caut.ca/sites/default/files/3.16_age_distribution_of_full-time_university_teachers_by_sex_and_major_discipline_2017-2018.xlsx (consulté en avril 2021).

4.2 Quels sont les principaux systèmes de la FCC pour la prestation de CIP ?



Graphique 4 : Plateforme nationale de CIP du Canada

Le graphique 4 ci-dessus montre les cinq sites d'hébergement nationaux au Canada. Les récents investissements substantiels de la FCI dans la cyberinfrastructure canadienne ont permis de déployer de nouveaux systèmes et d'augmenter considérablement la capacité et le potentiel du CIP au Canada. Pour ce processus, il a également fallu consolider les installations de la FCC dans cinq sites principaux de centres de données.[174] Ces derniers hébergent cinq systèmes nationaux et sont affiliés aux organisations régionales membres de la FCC suivantes, d'ouest en est :

- Université de Victoria, Arbutus (WestGrid) ;
- Université Simon Fraser, Cedar (WestGrid),
- Université de Waterloo, Graham (Calcul Ontario),
- Université de Toronto, Niagara (Calcul Ontario) ; et
- Université McGill/École de technologie supérieure, Béluga (Calcul Québec). [175]

[174] Compute Canada: Renewing Canada's Advanced Research Computing Platform https://www.computecanada.ca/techrenewal/ (consulté en septembre 2020).
[175] Compute Canada: Available Resources https://www.computecanada.ca/research-portal/accessing-resources/available-resources/ (consulté en septembre 2020).

Les caractéristiques générales de ces systèmes nationaux sont les suivantes :

- **Arbutus est un système de CIP en infonuagique polyvalent** destiné à l'hébergement de systèmes virtuels (principalement basés sur Linux) et d'autres charges de travail dématérialisé.[176]Il est basé sur une infrastructure de nuage OpenStack à source ouverte et dispose de plus de 16 000 cœurs de processeurs Intel, répartis sur 450 nœuds avec un total de 140 To de mémoire (soit 300 Go par nœud ou 10 Go par cœur). La communication d'arrière-plan est basée sur Ethernet, de 10 à 25 gigabits. La capacité de stockage principale est de 17 Po. La capacité de l'accélérateur GPU du système est minimale. [177]

- **Cedar est une grappe de CIP hétérogène polyvalente** destinée à une variété de charges de travail de CIP. Elle compte près de 95 000 cœurs de processeurs Intel, répartis sur 2 470 nœuds, avec une mémoire disponible par nœud allant de 125 Go à 3 To (c'est-à-dire de 4 Go à 90 Go par cœur). La communication d'arrière-plan est une matrice Intel Omni-Path à 100 Gbit/s à faible latence. Le système de stockage à plusieurs niveaux va de la petite capacité persistante « maison » (d'une capacité totale de plus de 500 To), à la capacité persistante « projet » (23 000 To), en passant par la capacité non persistante haute vitesse « complètement nouvelle » (5400 To). Le système dispose d'un total de 1350 cartes GPU Nvidia comme accélérateur. [178]

- **Graham est une grappe de CIP hétérogène polyvalente** destinée à une variété de charges de travail de CIP. Elle compte près de 42 000 cœurs de processeurs Intel, répartis sur 1185 nœuds, avec une mémoire disponible par nœud allant de 124 Go à 3 To (c'est-à-dire de 4 Go à 50 Go par cœur). La communication d'arrière-plan est une matrice InfiniBand Mellanox FDR (EDR) à 56 Gbit/s (100 Gbit/s) et à faible latence. Le système de stockage hiérarchisé va de la petite capacité persistante « maison » (d'une capacité totale de plus de 130 To), à la capacité persistante « projet » (16 000 To), en passant par la capacité non persistante haute vitesse « complètement nouvelle » (3600 To). Le système dispose d'un total de 520 cartes GPU Nvidia comme accélérateur. Environ un quart des cartes GPU sont des cartes Turing T4 de dernière génération, conçues pour les charges de travail d'apprentissage profond. [179]

- **Niagara est une grappe de CIP homogène massivement parallèle** pour le CIP évolutif. Elle compte près de 81 000 cœurs de processeurs Intel, répartis sur 2016 nœuds, avec une mémoire accessible par nœud fixée à 200 Go (soit 5 Go par cœur). La communication d'arrière-plan est une matrice InfiniBand Mellanox EDR à 100 Gbit/s, à faible latence et à haut débit, qui exploite la topologie Dragonfly+ de pointe. Le système de stockage à plusieurs niveaux va de la petite capacité persistante « maison » (200 To de capacité totale), à la capacité persistante « projet » (2000 To), en passant par la capacité non persistante haute vitesse « complètement nouvelle » (7000 To) et la capacité non

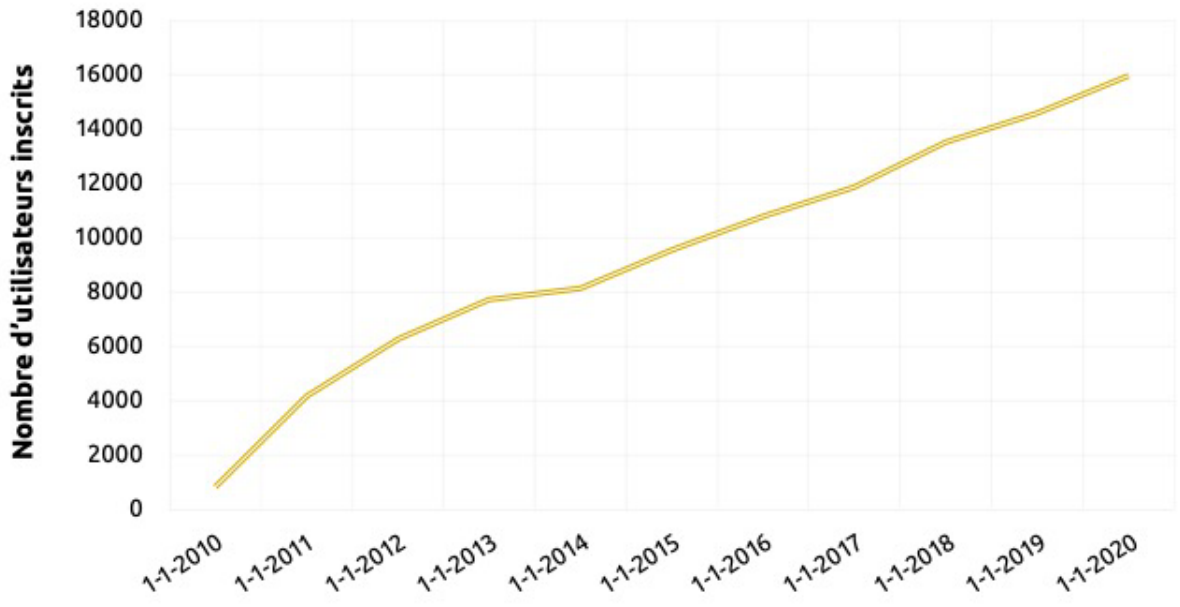
[176] Compute Canada: National Systems https://www.computecanada.ca/techrenewal/national-systems/ (consulté en septembre 2020).
[177] Compute Canada: Arbutus cloud https://docs.computecanada.ca/wiki/Cloud_resources#Arbutus_cloud_.28arbutus.cloud.computecanada.ca.29 (consulté en septembre 2020).
[178] Compute Canada: Cedar https://docs.computecanada.ca/wiki/Cedar (consulté en septembre 2020).
[179] Compute Canada: Graham https://docs.computecanada.ca/wiki/Graham (consulté en septembre 2020).

persistante haute vitesse « éclat » (230 To). Auparavant, le système n'avait pas de capacité d'accélération GPU, mais grâce au récent financement de l'expansion, il dispose actuellement de 64 GPU. [180]

- **Béluga est une grappe de CIP hétérogène polyvalente** destinée à une variété de charges de travail en CIP. Elle compte près de 35 000 cœurs de processeurs Intel, répartis sur 872 nœuds, avec une mémoire accessible par nœud allant de 90 Go à 750 Go (c'est-à-dire de 2 Go à environ 20 Go par cœur). La communication d'arrière-plan est une matrice InfiniBand Mellanox EDR de 100 Gbit/s à faible latence. Le système de stockage à plusieurs niveaux va de la petite capacité persistante « maison » (d'une capacité totale de plus de 100 To), à la capacité persistante « projet » (25 000 To), en passant par la capacité non persistante haute vitesse « complètement nouvelle » (2600 To). Le système dispose d'un total d'environ 688 cartes GPU Nvidia comme accélérateur. [181]

4.3 Quelle est le niveau d'utilisation actuel et antérieur du CIP dans les installations de la FCC ?

Utilisateurs inscrits :



Graphique 5: Utilisateurs inscrits de la FCC

[180] Compute Canada: Niagara https://docs.computecanada.ca/wiki/Niagara (consulté en septembre 2020).
[181] Compute Canada: Beluga https://docs.computecanada.ca/wiki/B%C3%A9luga/en (consulté en septembre 2020).

Dans le graphique 5 ci-dessus, le nombre d'utilisateurs inscrits de la FCC est représenté de façon temporelle, au 1er janvier de chaque année. En d'autres termes, les « utilisateurs inscrits » ont un compte actif lors du renouvellement annuel de la base de données de la FCC (qui désigne souvent ces derniers comme des « utilisateurs actifs »). Plusieurs de ces personnes n'exécutent pas nécessairement des tâches tout au long de l'année et parfois elles n'accèdent même pas aux systèmes. Certains utilisateurs ont tout simplement un compte actif en cas de besoins futurs, ou si un projet de recherche est retardé ou réorienté. Tous les chercheurs principaux (CP) doivent avoir un compte actif pour parrainer leurs étudiants, mais ils n'ont pas forcément accès aux systèmes de CIP ou n'exécutent pas de travaux eux-mêmes. Ces professeurs et CP sont essentiels en tant que coordonnateurs et bailleurs de fonds de la recherche, même s'ils n'utilisent pas au quotidien les systèmes de CIP. Certaines données, dont les téléchargements de logiciels, comme indiqué plus loin dans ce document, indiquent qu'environ 10 000 utilisateurs ont eu recours activement aux systèmes principaux de janvier à octobre 2020, par rapport à environ 18 000 utilisateurs inscrits en octobre 2020.

Le nombre d'utilisateurs inscrits a augmenté considérablement au cours de la dernière décennie, passant de 829 en 2010 à 15 994 en 2020, ce qui correspond à une croissance de plus de 1700 % en dix ans. Par ailleurs, la croissance a été relativement linéaire au fil des ans, mais elle est divisée en deux segments. Entre 2010 et 2013, la croissance a été très rapide, puis presque statique de 2013 à 2014. Par la suite, elle a été régulière et plus lente de 2014 à aujourd'hui. Le taux de croissance annuel composé (TCAC) était d'environ 34 % tous les ans au cours de la dernière décennie. Jusqu'en 2013, le TCAC était encore plus élevé, à environ 110 %, alors que depuis 2014, il est d'environ 12 %.

À la fin du mois d'octobre 2020, il y avait environ 1800 d'utilisateurs avec un compte et inscrits aux systèmes de la FCC, ce qui indique une augmentation d'environ 2 000 utilisateurs au cours des dix premiers mois de l'année. Cette augmentation est légèrement supérieure au TCAC annuel. Voici la répartition de ces utilisateurs par région :

- Westgrid : 5100 utilisateurs affiliés inscrits
- Calcul Ontario : 4800 (SharcNet : 2765, CAC : 1019, SciNet : 1865)
- Calcul Québec : 4600
- ACENET : 1000

Au total, 16 400 utilisateurs inscrits étaient affiliés à des universités qui, à leur tour, étaient membres de leurs organisations régionales de la FCC. À l'inverse, environ 900 utilisateurs n'étaient pas membres des principales organisations affiliées à la FCC (environ 700 utilisateurs n'ont indiqué aucune affiliation dans la base de données).

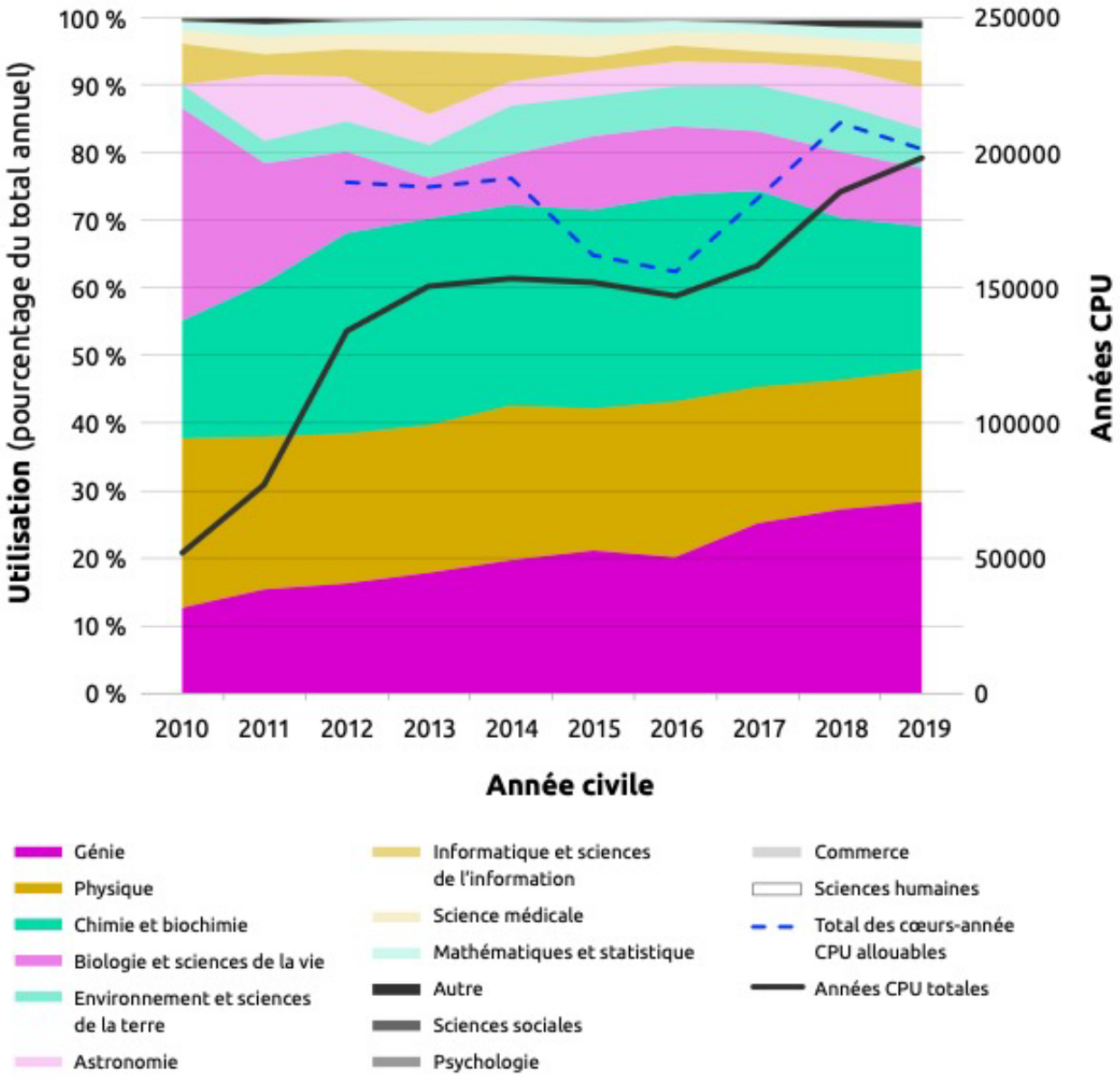
Le nombre d'établissements utilisant les ressources de la FCC à la fin octobre 2020 était supérieur à 600, dépassant largement le nombre d'établissements canadiens. Ceci reflète l'importance de la collaboration internationale. Les établissements ayant plus de 500 utilisateurs de la FCC sont les suivants, par ordre décroissant :

- Université McGill (1700)
- Université de Toronto (1500)

- Université de la Colombie-Britannique (1500)
- Université d’Alberta (1100)
- Université de Waterloo (800),
- Université de Montréal (800)
- Université Simon Fraser (600)
- Université de Western Ontario (600)

Ces établissements représentaient au total environ 8500 utilisateurs inscrits, soit environ 47 % de tous les utilisateurs inscrits de la FCC. D’autre part, plus de 600 établissements comptaient dix utilisateurs, ce qui reflète la diversité des utilisateurs et des établissements.

Utilisation du CPU :



Graphique 6: Utilisation historique de CPU pour le CIP (regroupée et par discipline de recherche)

Le graphique 6 ci-dessus montre l'utilisation totale et par discipline des ressources de CPU dans les systèmes de la FCC. Elle est mesurée en années-cœur de CPU, c'est-à-dire la ressource de calcul utilisée en exécutant un programme sur un seul cœur de CPU pendant une année complète. En revanche, cette mesure ne tient pas compte de l'évolution de la puissance de calcul CPU, ou les progrès d'architecture et de technologie au fur et à mesure des nouvelles générations de CPU. La durée est indiquée en années civiles complètes.

La ligne noire pleine et épaisse indique l'utilisation totale absolue des années-cœurs de CPU au cours de la dernière décennie (voir l'axe vertical à droite pour les unités), tandis que la ligne bleue en pointillés indique l'offre de CPU en années-cœurs par processus demande au CAR (ce qui montre que la demande est limitée par l'offre). En 2010, l'utilisation était d'environ 50 000 années CPU, tandis qu'en 2019, l'utilisation était d'environ 200 000 années CPU, ce qui indique que l'utilisation des ressources de CPU s'est multipliée par quatre. Cela correspond à un TCAC d'environ 16 %. La croissance a notamment été rapide entre 2010 et 2013, puis elle s'est stabilisée à environ 150 000 années CPU entre 2013 et 2017 à cause d'un déficit de financement de 9 ans pour les systèmes de CIP, notamment NPF (2006) et CI (2015). Cette stagnation s'explique aussi par des finalisations budgétaires et l'espacement des dépenses qui en découlent. Depuis 2017, l'utilisation augmente de nouveau rapidement et de façon presque linéaire avec la mise en ligne des nouveaux systèmes financés par la modernisation de la FCI. Historiquement et à l'heure actuelle, l'utilisation du CIP (au Canada et ailleurs dans le monde) est limitée par l'offre. Ce n'est donc pas que les ressources sont plus sollicitées, mais que la demande en soi augmente. Par ailleurs, l'utilisation augmente chaque année parce que les chercheurs ont accès à plus de ressources.

Les lignes pleines de couleur fine et les zones ombrées correspondantes montrent la répartition relative de l'utilisation des années processeur entre les différentes disciplines (voir l'axe vertical à gauche pour les unités). En 2010, le groupe d'utilisateurs le plus important était celui des sciences biologiques et de la vie avec environ 30 %, suivi de la physique et de l'astronomie avec environ 25 % (l'astronomie n'a pas été suivie séparément en 2010), de la chimie et de la biochimie avec environ 18 % et de l'ingénierie avec environ 12 % de l'utilisation totale. Au total, ces quatre disciplines ont utilisé environ 85 % des ressources.

En 2019, ces groupes d'utilisateurs étaient les plus importants :

- Ingénierie (environ 28 %)
- Physique (environ 20 %) et astronomie (environ 5 %) (total combiné de 25 %)
- Chimie et biochimie (environ 20 %)

Ces trois domaines de recherche ont consommé environ trois quarts des ressources, tandis que les domaines suivants ont consommé les 25 % restants :

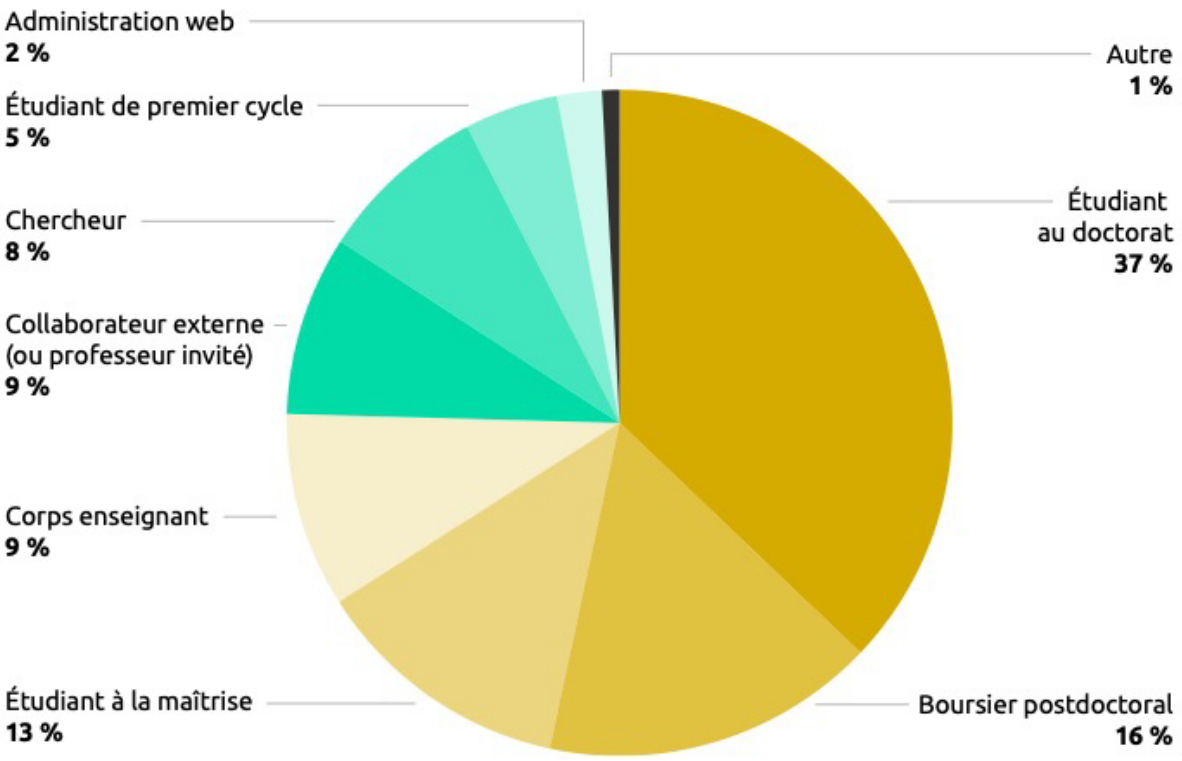
- Sciences biologiques et de la vie (environ 8 %)
- Sciences de l'environnement et de la terre (environ 6 %)
- Informatique et sciences de l'information (environ 4 %)
- Mathématiques et statistique (environ 2 %)
- Sciences médicales (environ 2 %)

Ces cinq disciplines ont consommé un total d'environ 22 % des ressources, ne laissant que 3 % aux « autres » (environ 1 %), soit les sciences sociales, la psychologie, les affaires et les sciences humaines.

Au cours de la dernière décennie, les principales tendances ont été la forte croissance de l'ingénierie, la baisse de la position relative de la consommation de ressources des sciences biologiques et de la vie, ainsi que l'émergence et la forte croissance de nouvelles disciplines exploitant le CIP. Parmi les trois premières disciplines, l'ingénierie est passée d'une allocation de 12 % à 28 % au cours de la dernière décennie, tandis qu'en physique, en astronomie, en chimie et en biochimie elle est restée à peu près identique. Les gains réalisés en ingénierie sont, dans une certaine mesure, proportionnels à la baisse des sciences biologiques et du vivant, qui n'ont consommé « que » 8 % environ de toutes les ressources en 2019, alors qu'en 2010, la consommation était d'environ 30 %. En termes absolus, l'utilisation des ressources par les sciences biologiques et de la vie est restée constante à environ 16 000 années CPU, tandis que l'allocation relative s'est divisée par quatre au cours de la dernière décennie. Dans la fourchette moyenne de consommation des ressources (entre 2 et 6 % par discipline), les mathématiques et la statistique (croissance absolue multipliée par plus de sept) et les sciences de l'environnement et de la terre (croissance absolue multipliée par près de sept) ont augmenté leur utilisation des ressources par rapport aux autres au cours de la dernière décennie, c'est-à-dire à un rythme plus rapide que la croissance générale multipliée par quatre. Le taux de croissance de l'utilisation en sciences médicales était d'environ 4,5. En informatique et sciences de l'information, il n'a augmenté que de 2,5 au cours de la décennie.

Il est intéressant de noter que les sciences sociales, la psychologie, les affaires et les sciences humaines ont considérablement augmenté leur utilisation des ressources CPU au cours de la dernière décennie. En 2010, ce groupe consommait environ 0,3 % des ressources de CPU, tandis qu'en 2019, il consommait environ 0,6 % des ressources de CIP, ce qui correspond à une multiplication par deux de l'utilisation par rapport aux autres disciplines. Cette croissance est encore plus impressionnante en termes absolus, puisqu'elle s'est multipliée par huit au cours de la décennie, passant de 157 années CPU en 2010 à environ 1250 années CPU en 2019. La croissance de l'utilisation de CPU en CIP dans ces disciplines indique un fort intérêt potentiel pour l'IRN à l'avenir dans ces disciplines. Étant donné le type de tâches effectuées dans ces disciplines, la mesure brute des cycles CPU n'est pas nécessairement pertinente pour refléter l'utilisation des ressources informatiques.

Utilisation de CPU par occupation



Graphique 7 : Utilisation de CPU par occupation

Le graphique 7 ci-dessus montre l'utilisation relative de CPU par occupation dans les systèmes de la FCC, de mars 2020 à février 2021. Les données ne sont pas normalisées en fonction du nombre d'utilisateurs dans chaque groupe. Elles montrent le temps total de CPU pour les utilisateurs dans chaque type d'occupation. Les étudiants au doctorat constituent le groupe d'utilisateurs le plus important (37 %), suivis des étudiants au postdoctorat (16 %), étudiants à la maîtrise (13 %), ce qui représente les deux tiers de l'utilisation totale des ressources de CPU. Les professeurs utilisent directement 9 % des ressources de CPU pour le CIP. Le quart restant des ressources est utilisé par une variété de groupes externes et internes, avec des étudiants de premier cycle (5 % du total). L'utilisation du GPU en général est semblable à l'utilisation de CPU par occupation, sauf que les étudiants à la maîtrise utilisent 22 % des ressources de GPU, c'est-à-dire près du double de leur utilisation relative de CPU. Le corps enseignant utilise seulement 2 % des ressources de GPU par rapport à 9 % pour le CPU.

La comparaison entre l'utilisation des ressources de CPU et les occupations des utilisateurs (comme indiqué précédemment dans ce document) fournit des indications intéressantes : les professeurs représentent 27 % de la base d'utilisateurs du CIP, avec 9 % des ressources de CPU. Les étudiants au postdoctorat et les étudiants au doctorat représentent 32 % des utilisateurs, mais consomment 53 % des ressources de CPU. Quant aux étudiants à la maîtrise, ils utilisent les ressources à peu près au même niveau que leur nombre le laisserait supposer (13 % d'utilisation par rapport à 15 % de la base d'utilisateurs). En général, ce type de sous-représentation et de surreprésentation dans l'utilisation des ressources n'est pas surprenant étant

donné que les professeurs et les chercheurs chevronnés doivent souvent se concentrer sur la direction de leurs équipes de recherche, la rédaction des demandes de subventions et la mobilisation, ce qui laisse moins de temps pour exécuter des charges de travail sur les systèmes de CIP.

Utilisation de CPU par site hôte comparativement à la région du chercheur principal (CP)

		PI's regional affiliation				
		ACENET	Calcul Ontario	Calcul Québec	Westgrid	Total (années CPU)
Ressources régionales	Beluga	3,68%	11,14%	**68,02%**	17,16%	23 052
	Cèdre	9,29 %	22,11 %	17,03 %	**51,56 %**	74 780
	Graham	14,34 %	**47,49 %**	14,17 %	24,00 %	30 573
	mp2	3,77 %	21,51 %	**65,72 %**	9,00 %	15 723
	Niagara	0,82 %	**73,07 %**	7,77 %	18,33 %	71 774
	Total	6,19%	41,43%	22,54%	29,84%	215,903

Tableau 1: Répartition régionale de l'utilisation de CPU par affiliation régionale du CP

Le Tableau 1 ci-dessus montre l'allocation régionale des ressources de CPU par région de la FCC en fonction de l'établissement d'origine du chercheur principal. Les données portent sur une période de 12 mois, à partir d'avril 2020. Chaque ligne correspond à l'un des 4 sites nationaux, en plus de Mammouth Parallel 2 (mp2), une grappe d'ancienne génération toujours allouée en 2020. Les colonnes trois à six indiquent ensuite les régions d'où proviennent les CP. Les valeurs de chaque ligne/ressource s'additionnent à cent pour cent (elles ne sont pas indiquées en pourcentage, mais la valeur totale absolue figure dans la dernière colonne) et montrent l'allocation régionale et l'utilisation de la ressource. Par exemple, les utilisateurs affiliés à Calcul Québec utilisent 68 % des cycles de CPU sur Beluga, mais seulement 8 % de tous les cycles de CPU sur Niagara. L'utilisation locale de l'infrastructure, chez les usagers de cette région, est mise en évidence en gras pour plus de clarté. Pour donner un contexte absolu à la distribution relative sur chaque ligne, la dernière colonne indique le nombre total d'années CPU pour chaque ressource. Les ressources primaires sont Niagara avec 72 000 années CPU et Cedar avec 75 000 années CPU, tandis que Beluga et Graham fournissent respectivement 23 000 et 31 000 années CPU.

La dernière ligne indique la distribution relative de l'allocation totale de 216 000 années CPU entre les utilisateurs des différentes régions. Les pourcentages relatifs de chaque colonne (rangées trois à six) ne sont pas censés s'additionner à la valeur en pourcentage du « total » de la dernière rangée.

En ce qui concerne la dernière ligne et la dernière colonne, les utilisateurs de la région de Calcul Ontario (4 800 utilisateurs inscrits) consomment environ 41 % de tous les cycles de CPU, tandis que Calcul Ontario fournit environ 47 % des cycles de calcul (c'est-à-dire 102 000 sur 215 000). Le deuxième groupe d'utilisateurs le plus important provient de la région Westgrid (5100 utilisateurs inscrits), qui consomme environ 30 % des cycles CPU, tandis que Westgrid fournit environ 35 % des ressources de calcul. Les utilisateurs provenant des régions affiliées à Calcul Québec (4600 utilisateurs inscrits) utilisent environ 23 % du total, tandis que la grappe Beluga de Calcul Québec fournit environ 11 % des ressources informatiques. Les utilisateurs régionaux d'Acenet (100 utilisateurs inscrits) consomment environ 6 % du total des cycles de CPU. Le reste de la capacité de calcul est fourni par l'ancienne grappe Mammouth Parallel 2 (mp2), exploité par Calcul Québec.

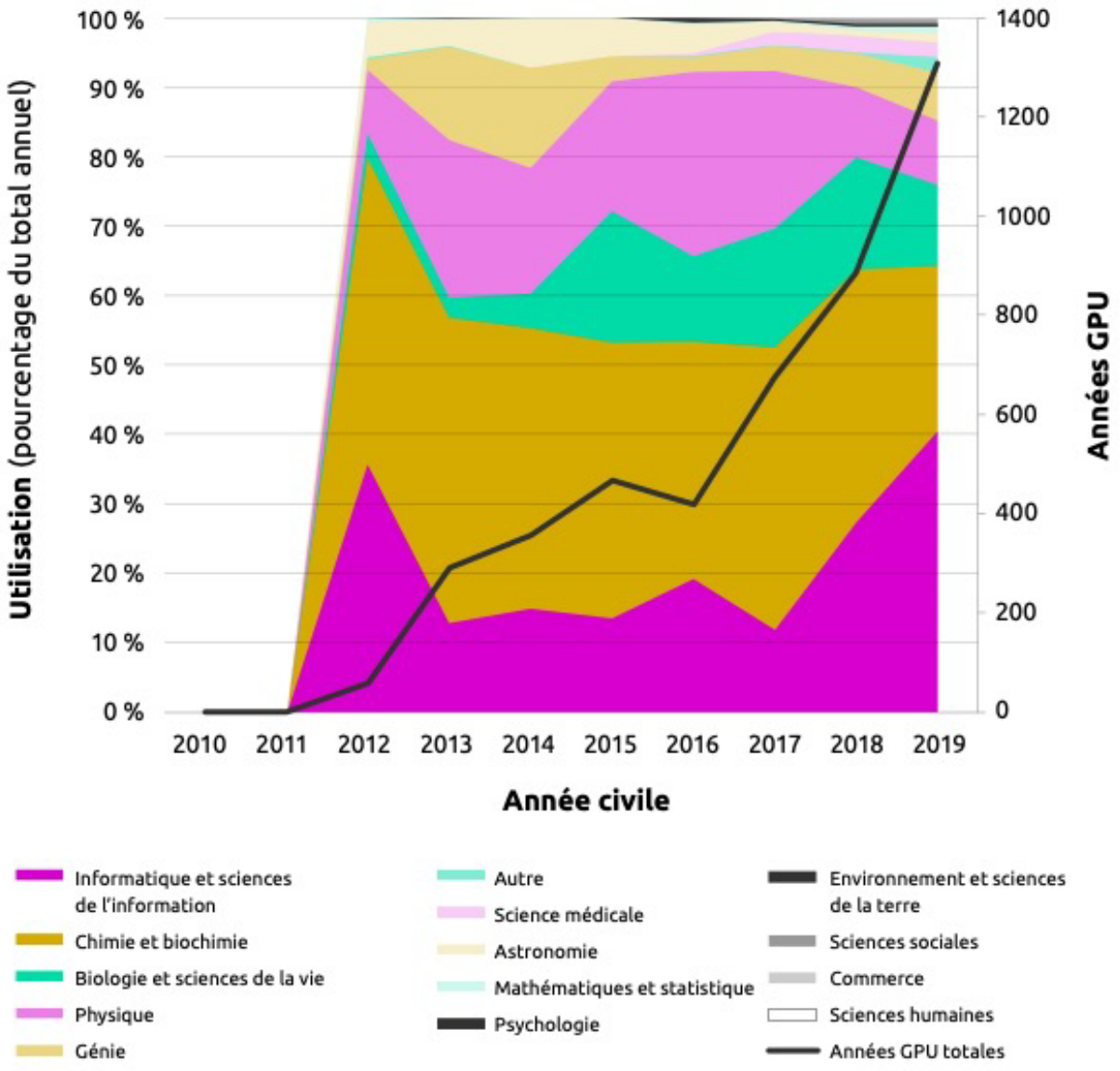
Pour les lignes trois à six, soit la distribution régionale des utilisateurs pour chaque file d'attente de services, ce sont en général les usagers de la région qui font appel aux ressources « locales ». À cet égard, les usagers affiliés à Calcul Ontario utilisent 73 % des ressources CPU de Niagara. De même, les usagers de calcul Québec utilisent environ 68 % des ressources sur Beluga. Les deux autres grappes principales, Graham et Cedar, affichent quant à elles une utilisation d'origine locale d'environ 50 %. La moitié restante sur Cedar de Westgrid est répartie à peu près également entre les utilisateurs de Calcul Ontario et de Calcul Québec. Dans le cas de Graham, les usagers de la région Westgrid utilisent environ 25 % des ressources, tandis que le quart restant est réparti à peu près également entre les affiliés d'Acenet et de Calcul Québec.

Calcul Ontario a effectué une analyse indépendante de l'utilisation régionale en fonction de l'affiliation régionale du CP, ce qui ajoute un peu de couleur aux données ci-dessus. Les résultats indiquent que la grande majorité des CP utilisent leur système régional, mais aussi qu'au moins 50 % des CP d'une région donnée calculent également en dehors de leur région (18 % des CP utilisent 3 systèmes sur 5). L'analyse a également porté sur le rôle de l'allocation et du CAR, montrant que, par exemple, les CP basés en Ontario ont utilisé 40 % de tous les cycles tout en recevant 40 % des allocations du CAR. L'utilisation régionale repose également sur les données démographiques, c'est-à-dire que les utilisateurs de différentes régions utilisent les ressources de CIP en fonction de la taille de la population dans cette région.

La corrélation ci-dessus entre le site d'une infrastructure et celle des chercheurs est surtout historique et compliquée. Tous les chercheurs inscrits à la BDCC peuvent accéder à l'ensemble des systèmes (qu'ils aient ou non une allocation). Le choix du ou des systèmes qu'ils utilisent pour leurs calculs peut être motivé par de nombreux facteurs, y compris des raisons techniques et d'autres plus individuelles. Certains facteurs évidents incluent : l'allocation du CAR (des bourses sont allouées pour des systèmes précis), l'accessibilité du matériel (par exemple, des nœuds à grande mémoire, un type particulier de GPU), la politique d'ordonnancement (par exemple, autoriser des travaux avec de longs temps d'horloge murale ou de grands nombres de cœurs), la formation, les recommandations de collègues, la perception personnelle du rendement d'un système et d'autres préférences individuelles. La nécessité d'avoir des données accessibles localement pour un calcul efficace (ainsi que la gestion des données en général) peut tendre à

minimiser le nombre de systèmes qu'un CP utilise. D'un autre côté, l'utilisation de plus de systèmes augmente la quantité de cycles « par défaut » accessibles pour un CP et peut contribuer à réduire les temps d'attente. L'allocation des CP et leur utilisation peuvent également être liées au moment où les systèmes sont entrés en production. De façon générale, nous pouvons calculer qu'environ 60 % des cœurs de CPU sont utilisés par des usagers de la région qui héberge l'infrastructure, tandis que 40 % des cœurs de CPU sont utilisés par des usagers d'autres régions.

Utilisation de GPU



Graphique 8: Utilisation historique de GPU pour le CIP (regroupée et par discipline de recherche)

Le graphique 8 ci-dessus montre l'historique d'utilisation des accélérateurs GPU dans les systèmes de CIP au cours de la dernière décennie. Les unités absolues sur l'axe vertical de droite sont les années GPU, soit l'équivalent de l'exécution d'un programme sur une unité de traitement

GPU pendant une année civile complète. L'efficacité de calcul de cette unité n'est pas constante en raison des avancées architecturales, donc le nombre de « micro » cœurs de calcul au sein d'un GPU massivement parallèle peut changer au fil du temps. Par exemple, la dernière génération de carte GPU Nvidia Ampere peut avoir plus de 8000 cœurs de calcul CUDA par GPU, alors que l'unité de traitement GPU Nvidia Turing de la génération précédente peut avoir environ 3000 cœurs CUDA par GPU.[182]L'augmentation de l'efficacité du calcul est également vraie pour les nœuds de calcul (le nombre de cœurs augmente) et les cœurs de calcul (en raison des améliorations apportées aux microprocesseurs, mais malheureusement plus au même degré).

La ligne noire épaisse et continue indique la croissance d'utilisation des GPU en termes absolus. En 2010, l'utilisation des GPU en CIP était pratiquement nulle, mais le calcul par GPU émergeait comme une tendance importante du CIP à ce moment-là. 2012 marque la première année où les GPU ont été production pour le CIP et qu'ils ont fait l'objet d'un suivi et de rapports. Au total, 58 années GPU ont été utilisées. En 2019, l'utilisation totale des ressources de GPU était d'environ 1 300 années GPU, ce qui correspond à un TCAC d'environ 56 % depuis 2012. Comme indiqué ailleurs dans ce document, cette croissance a été sévèrement limitée par l'offre et elle n'est donc pas indicative du taux de croissance réel du calcul par GPU en général.

Les lignes continues de couleur fine et les zones ombrées correspondantes montrent l'allocation relative de l'utilisation de GPU-année entre les différentes disciplines de recherche (voir l'axe vertical à gauche pour les pourcentages). En 2019, le plus grand groupe d'utilisateurs était l'informatique et les sciences de l'information (40 %), suivi de la chimie et de la biochimie (24 %), des sciences biologiques et de la vie (12 %), puis de la physique et de l'astronomie (11 %). Au total, ces quatre disciplines ont utilisé environ 87 % des ressources de GPU en 2019. Historiquement, l'utilisation des GPU était d'abord et avant tout associée à la dynamique moléculaire (chimie et biochimie), GROMACS et NAMD étant les principales applications. Comparativement à leur utilisation respective des CPU, l'informatique et les sciences de l'information sont devenues beaucoup plus importantes en tant qu'utilisateur de GPU en CIP.

[182] NVIDIA White Paper: NVIDIA AMPERE GA102 GPU ARCHITECTURE https://www.nvidia.com/content/dam/en-zz/Solutions/geforce/ampere/pdf/NVIDIA-ampere-GA102-GPU-Architecture-Whitepaper-V1.pdf (septembre 2020).

Utilisation de GPU par site hôte par région de chercheur principal (CP)

		Affiliation régionale du CP				
		ACENET	Calcul Ontario	Calcul Québec	Westgrid	Total (années GPU)
Ressource régionale	Beluga	3,37%	17,46%	69,65%	9,52%	547,47
	Cèdre	2,21 %	29,99 %	27,02 %	40,78 %	989,13
	Graham	8,29 %	50,08 %	24,92 %	16,71 %	306,49
	Helios	0,00 %	52,43 %	46,12 %	1,45 %	87,40
	Total	3,41 %	30,64 %	39,64 %	26,31 %	1930,49

Tableau 2: Répartition régionale de l'utilisation des ressources de GPU par affiliation régionale du CP

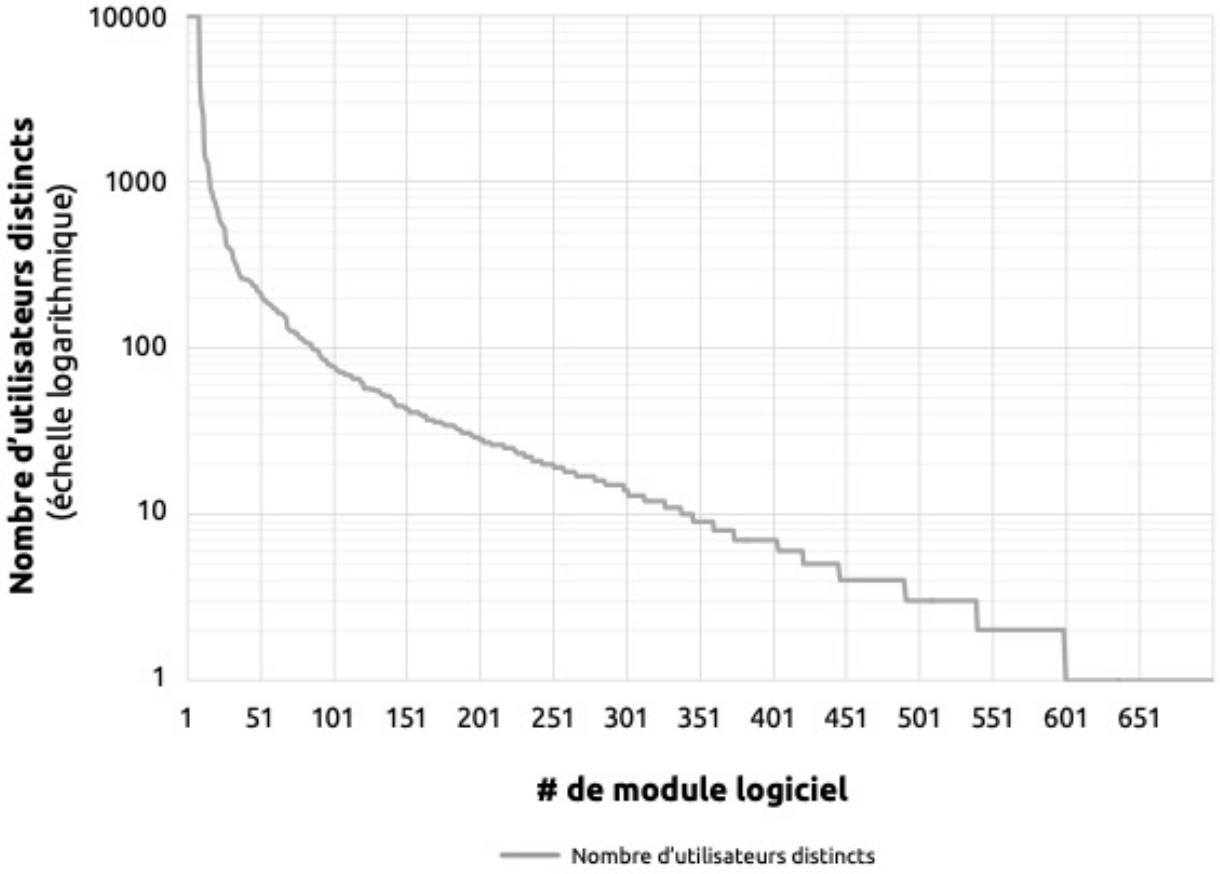
Le tableau 2 ci-dessus montre la répartition régionale d'utilisation des ressources de GPU en fonction de l'établissement d'origine du chercheur principal. Les données portent sur une période de 12 mois, à partir d'avril 2020. Chaque ligne correspond à l'un des 3 sites nationaux qui ont des GPU, en plus d'Helios, une grappe d'ancienne génération qui était toujours allouée en 2020 et 2021. Les colonnes trois à six montrent ensuite les différentes régions d'où proviennent les CP. Les valeurs de chaque ligne/ressource s'additionnent à cent pour cent (elles ne sont pas indiquées en pourcentage, mais la valeur totale absolue figure dans la dernière colonne) et montrent la répartition régionale de l'utilisation de la ressource. Par exemple, les utilisateurs affiliés à la région de Calcul Québec utilisent 70 % du total des cycles de GPU accessibles sur Beluga et 27 % de tous les cycles GPU accessibles sur Cedar. Pour donner un contexte absolu à la distribution relative de chaque ligne, la dernière colonne indique le nombre total d'années GPU pour chaque ressource respectivement. Les ressources primaires sont Cèdre avec 990 années GPU, Beluga avec 550 années GPU, tandis que Graham fournit 310 années GPU. Le système d'ancienne génération Helios, exploité par Calcul Québec, a fourni 90 années GPU depuis avril 2020. La dernière ligne indique la répartition relative totale des 1930 années GPU entre les utilisateurs des différentes régions. Les pourcentages relatifs dans chaque colonne (rangées trois à six) ne doivent pas s'aditionner à la valeur en pourcentage du « grand total » de la dernière rangée.

En ce qui concerne la dernière ligne et la dernière colonne, les utilisateurs de la région de Calcul Québec consomment environ 40 % de tous les cycles de GPU, tandis que Calcul Québec fournit environ 33 % des cycles de calcul (c'est-à-dire 634 années GPU sur 1930). Le deuxième groupe d'utilisateurs le plus important provient de la région de Calcul Ontario, qui consomme environ 31 % des cycles GPU, tandis que Calcul Ontario fournit environ 16 % des ressources de calcul. Les usagers provenant des régions affiliées à Westgrid utilisent environ 26 % du total, tandis que la grappe Cedar de Westgrid fournit environ 51 % des ressources de GPU pour le CIP. Les utilisateurs régionaux d'Acenet consomment environ 3 % du total des cycles GPU.

Si l'on examine les lignes trois à six, c'est-à-dire la distribution régionale des utilisateurs pour chaque file d'attente de services, la tendance générale est que chaque ressource « locale » est principalement utilisée par les usagers de cette même région (à l'exception de la faible utilisation d'Helios). À l'extrémité la plus locale, 70 % des ressources de GPU sur Beluga sont consommées par les utilisateurs affiliés à Calcul Québec. Les utilisateurs affiliés à Calcul Ontario consomment ensuite 17 % des ressources de GPU sur Beluga, tandis que les usagers affiliés à Westgrid utilisent environ 10 % des ressources. Les deux autres grappes principales, Graham et Cedar, ont un taux d'utilisation locale d'environ 50 % et 41 % respectivement. Dans le cas de Graham, les usagers de la région de Calcul Québec utilisent environ 25 % des ressources de GPU, tandis que le quart restant des ressources de GPU est réparti entre les utilisateurs affiliés à Westgrid (17 %) et Acenet (8 %). Les 61 % restants de la capacité en GPU de Westgrid Cedar sont répartis à peu près équitablement entre les utilisateurs de Calcul Ontario et de Calcul Québec, avec respectivement 30 % et 27 %. Soulignons que cette corrélation entre le site d'une infrastructure et celui des chercheurs est surtout historique et accidentelle. En effet, les utilisateurs ont tendance à migrer vers une nouvelle grappe lorsque les anciennes sont mises hors service. Les considérations techniques sont prises en compte lorsque le personnel la FCC recommande un site précis à un groupe de recherche. De façon générale, environ 50 % des cœurs de processeur sont utilisés par des usagers de la région qui héberge l'infrastructure, tandis que 50 % des cœurs de processeur sont utilisés par des usagers d'autres régions.

Utilisation de logiciels

L'infrastructure de CIP de la FCC prend en charge et fournit plusieurs logiciels de recherche à la communauté des utilisateurs. Les mécanismes de distribution et l'accessibilité des logiciels varient même entre les principaux systèmes, dont les détails sont abordés au chapitre 4.5 du présent document. L'un des principaux mécanismes de distribution de logiciels gérés de manière centralisée par le personnel de la FCC repose sur la technologie de conditionnement des modules, qui est ouverte aux utilisateurs de tous les sites.



Graphique 9 : Nombre d'utilisateurs distincts de modules logiciels

Le graphique 9 ci-dessus montre le nombre d'utilisateurs distincts (axe des y, sur une échelle logarithmique) qui ont chargé des modules logiciels précis par l'entremise du système des progiciels de modules de la FCC. Chaque progiciel de module est indiqué par une étiquette numérotée (axe des x). Au cours des dix premiers mois de 2020, plus de 700 progiciels différents ont été chargés à l'aide du mécanisme central de « modules » de la FCC. Plus de 1000 utilisateurs distincts ont chargé 15 modules SW, tandis qu'au moins 100 utilisateurs distincts ont chargé plus de 600 modules SW. De plus, sur ces 600 modules SW, moins de 20 utilisateurs distincts ont chargé 450 modules SW. En d'autres termes, un nombre relativement petit d'utilisateurs consomme plusieurs modules SW, ce qui met clairement sous pression l'écosystème de maintenance et d'assistance pour ces modules. De plus, l'équipe de la FCC conçoit la plupart de ces progiciels pour plusieurs versions de ceux-ci et pour plusieurs générations d'architecture de CPU, afin de garantir un rendement maximal sur une infrastructure donnée.

Près de 10 000 utilisateurs ont chargé les plus populaires progiciels par défaut, ce qui correspond au nombre d'utilisateurs actifs ayant exécuté des travaux dans les systèmes de la FCC au cours des dix premiers mois de 2020 (excluant de nombreux travaux sur Niagara). Étant donné que de nombreux modules logiciels sont chargés pour des progiciels interconnectés et interdépendants, il est très difficile d'interpréter l'utilisation explicite réelle des logiciels et leur popularité à partir des

chargements de modules individuels. On peut néanmoins déduire certaines informations pour les progiciels qui sont relativement indépendants et doivent être explicitement chargés par l'utilisateur final. Si l'on considère de tels cas, les compilateurs GCC ont été chargés environ 3000 fois, ce qui indique qu'environ 30 % des utilisateurs voulaient passer des compilateurs Intel par défaut aux compilateurs GNU. Les chargements de modules indiquent également qu'environ 14 % des utilisateurs actifs avaient recours ou étaient intéressés par le calcul GPU (puisqu'environ 1400 utilisateurs ont chargé les bibliothèques CUDA). Le progiciel scientifique R est relativement indépendant. Par conséquent, les quelque 1100 téléchargements de ce progiciel indiquent qu'environ un dixième des utilisateurs actifs y ont eu recours. Comme Python est utilisé par de nombreux progiciels, on ne peut pas tirer de conclusions semblables à partir des quelques 4100 chargements correspondants du progiciel des modules Python. Pour la pile logicielle de la FCC par exemple, 38 modules dépendent explicitement de Python, alors que seulement 3 modules dépendent explicitement du progiciel R.

Utilisation de l'infonuagique

En janvier 2021, la FCC proposait des ressources d'infonuagiques sur plusieurs systèmes et dans diverses régions, notamment Arbutus (environ 16 000 cœurs de CPU), East (environ 600 cœurs), Cedar (environ 1 000 cœurs) et Graham (environ 800 cœurs), soit un total d'environ 18 400 cœurs de CPU. Comme l'indiquent les chiffres, le système d'infonuagique Arbutus contribue à 87 % de toutes les ressources dématérialisées de l'infrastructure de la FCC. L'infrastructure de calcul intensif de la FCC dispose d'environ 268 000 cœurs de CPU, ce qui signifie que l'offre d'infonuagique représente environ 7 % de la capacité totale de calcul par CPU. En ce qui concerne les ressources de GPU pour l'infonuagique, Arbutus offre environ 100 GPU, alors que les superordinateurs traditionnels de la FCC comptent ensemble environ 2500 GPU. En termes relatifs, les systèmes dématérialisés de la FCC ont environ la moitié de la capacité de GPU (environ 4 %) par rapport à la capacité de CPU disponible. Il y a deux principaux types d'utilisateurs : ceux qui lancent des systèmes virtuels (SV) en tant qu'instances persistantes et ceux qui lancent des instances de calcul/plateforme temporelles en tant que service ou de logiciels en tant que service par l'entremise de plateformes d'intergiciels, dont CANFAR et Syzygy entre autres. Certains utilisateurs de la FCC n'appartiennent pas à ces catégories, comme les usagers aguerris exploitent tous les types de ressources tant qu'elles sont accessibles.

Domaine de recherche	Total d'années-cœurs de CPU	Total des projets
Anthropologie	9,45	1
Astronomie	2 990,68	51
Sciences biologiques et de la vie	2 785,39	52
Commerce	97,55	12
Personnel de la FCC	279,25	36
Chimie et biochimie	887,56	8
Informatique et sciences de l'information	2 093,78	69
COVID-19	7 815,38	4
Ingénierie	767,37	35
Sciences de l'environnement et de la terre	525,54	20
FRDR	143,26	4
Histoire	9,02	1
Sciences humaines	573,52	37
Mathématiques et statistique	359,19	11
Sciences médicales	428,37	23
Équipe ou service national	350,46	24
Physique	7 069,72	18
Psychologie	160,79	13
Gestion des données de recherche	0,18	1
Sciences sociales	348,84	35
Formation	232, 44	11
Total	**27 927,76**	**466**

Tableau 3: Utilisation de l'infonuagique par domaine de recherche en 2020 sur Arbutus

À l'heure actuelle, la FCC n'entreprend pas de suivi détaillé sur l'utilisation de l'infonuagique dans les systèmes précédents. Le tableau 3 ci-dessus montre l'utilisation en 2020 de l'infonuagique par domaine de recherche sur le principal système fournisseur d'infonuagique de la FCC, la grappe Arbutus. L'utilisation qui apparaît dans la deuxième colonne est rapportée en années-cœur de CPU avec l'hyperthreading activé. Chaque cœur de CPU physique est présenté sous forme de deux cœurs virtuels par le système d'exploitation, notamment pour l'allocation de ressources, ce qui double en effet la ressource de calcul accessible aux utilisateurs finaux. L'utilisation totale en 2020 était d'environ 28 000 années CPU, soit à peu près 13 000 années-

CPU de plus que la capacité stricte de 15 000 années CPU sans hyperthreading de cette grappe. Avec la sursouscription basée sur l'hyperthreading, Arbutus est présenté comme une ressource de 30 000 cœurs de CPU. La sursouscription est une caractéristique standard des systèmes d'infonuagique bien gérés, notamment avec des charges de travail de serveurs qui n'utilisent pas toujours les ressources à 100 % comme le font souvent les charges de travail traditionnelles du CIP. Le ratio de sursouscription dans Arbutus est d'environ 1,9. Un taux optimal à cet égard dépend fortement du type de charges de travail et peut aller de 1,0 (c'est-à-dire pas de sursouscription, par exemple pour les charges de travail de CIP traditionnel à forte intensité de calcul CPU) à plus de 5 (par exemple pour les applications bureautiques légères ou les portails web peu utilisés).

Les principales disciplines qui ont utilisé Arbutus l'année dernière sont la recherche sur la COVID-19 (environ 7,8 k années CPU), la physique (7,1 k), l'astronomie (3,0 k), les sciences biologiques et de la vie (2,8 k) et les sciences informatiques et de l'information (2,1 k), totalisant environ 22,8 k années CPU, soit 81 % de l'utilisation totale. La recherche sur la COVID-19 figure en tête d'utilisation des ressources, ce qui indique de façon intéressante que les ressources dématérialisées peuvent être déployées de manière flexible pour de nouvelles recherches et de nouveaux besoins. Soulignons aussi que la recherche sur la COVID-19 s'inscrit uniquement dans quatre projets, alors qu'en physique, la même quantité de cycles de calcul est répartie sur 18 projets.

Il y a une autre comparaison intéressante entre la physique et l'astronomie (environ 10 000 années CPU, 69 projets) et les sciences sociales et humaines (0,9 000 années CPU, 72 projets), ce qui indique que les disciplines sous-représentées adoptent plus l'infonuagique comme l'indique le nombre de projets (avec des exigences de puissance de calcul beaucoup plus faibles que le groupe de comparaison).

Domaine de recherche	Total des utilisateurs
Astronomie	41
Sciences biologiques et de la vie	189
Commerce	20
Personnel de la FCC	236
Chimie et biochimie	20
Informatique et sciences de l'information	221
Ingénierie	87
Sciences de l'environnement et de la terre	76
Sciences humaines	57
Mathématiques et statistique	28
Sciences médicales	70
Physique	37
Psychologie	29
Sciences sociales	63
Total général	**1174**

Tableau 4: Utilisateurs d'Arbutus par domaine de recherche en 2020

Le tableau 4 ci-dessus illustre le nombre d'utilisateurs d'infonuagique par discipline de recherche (en utilisant une nomenclature légèrement différente de celle du tableau précédent de la FCC) en 2020 sur Arbutus. Le plus grand groupe d'utilisateurs est celui des sciences informatiques et de l'information (221 utilisateurs), correspondant au plus grand nombre de projets en nuage (69 projets). Il est intéressant de noter qu'il y a moins d'utilisateurs d'astronomie et de physique (78) que d'utilisateurs de sciences sociales et humaines (120).

Les statistiques ci-dessus sont néanmoins difficiles à interpréter et peuvent être trompeuses. En ce qui concerne les statistiques sur le nombre d'utilisateurs d'infonuagique, il convient de noter que plusieurs groupes gèrent d'importants portails qui desservent de grandes communautés. Par exemple, le projet Syzygy de James Colliander apparaît comme un utilisateur unique dans les statistiques ci-dessus, alors qu'il accueille en moyenne 1000 utilisateurs par jour et compte une base d'utilisateurs en pleine croissance de plusieurs dizaines de milliers d'individus. Ces

statistiques ne reflètent pas non plus tout à fait les plateformes comme CANFAR[183], GenAp[184], Magic Castle[185], ou iReceptor [186]qui desservent de grandes communautés.

Soulignons que les données ci-dessus n'incluent pas les autres ressources dématérialisées de la FCC, ni l'utilisation commerciale du nuage, cette dernière pouvant être substantielle. D'autres recherches ciblées sont manifestement nécessaires pour évaluer l'étendue de l'utilisation de l'IRN en infonuagique au Canada.

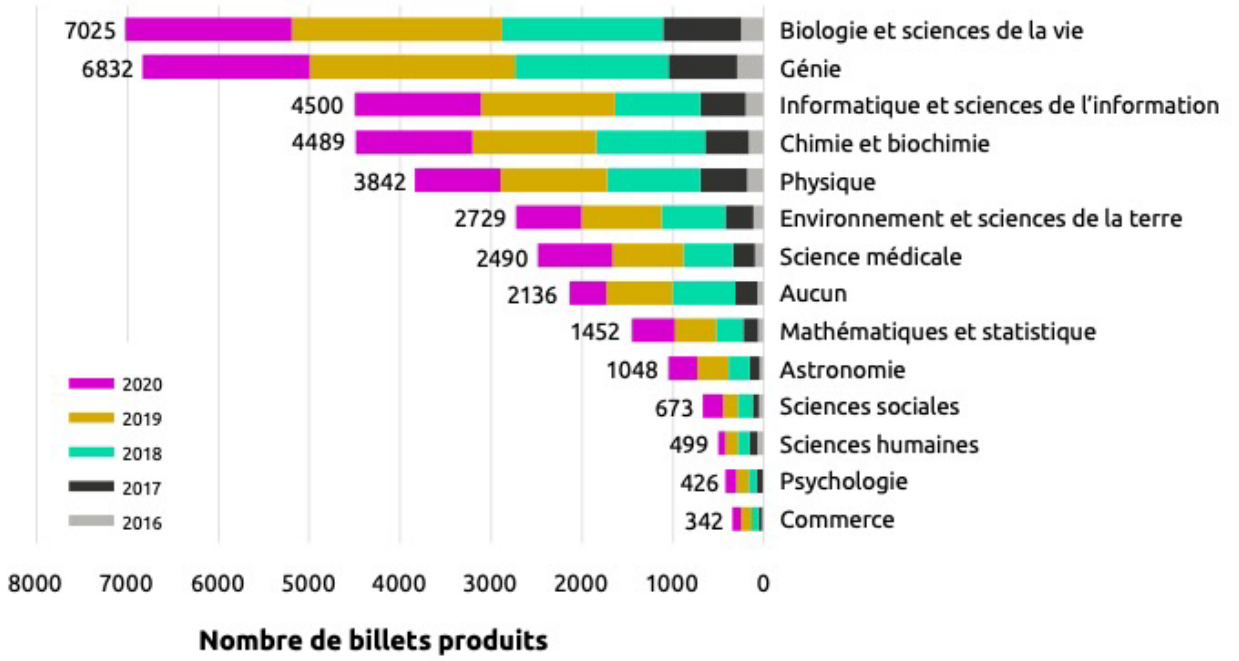
Utilisation du stockage

Comme les ressources sont limitées, le groupe de travail sur le CIP n'a pas eu l'occasion d'analyser les modèles d'utilisation réelle du stockage sur les sites d'hébergement et les systèmes de stockage de la FCC. En revanche, l'Alliance a formé un groupe de travail distinct sur le stockage qui examinera à la fois l'utilisation actuelle du stockage et les besoins futurs de stockage dans l'infrastructure de la FCC. Ce travail ne prendra pas seulement en compte le stockage actif dans le cadre du mandat actuel de la FCC auprès de la FCI, mais aussi les besoins et les politiques de stockage nearline, de dépôt et d'archivage. Les conclusions du groupe de travail sur le stockage seront accessibles au printemps-été 2021 et contribueront au plan stratégique, ainsi qu'aux nouveaux modèles de service de l'Alliance à l'automne 2021.

Plutôt que d'examiner l'utilisation réelle du stockage, le CAR de la FCC montre les demandes et les allocations de stockage actif actuelles et historiques. Ces données agrégées sur le stockage actif sont abordées au chapitre 4.5. Il convient de noter que les données du CAR qui s'adressent à l'utilisateur final ne tiennent pas compte des besoins en matière de sauvegarde, de dépôt et d'archivage des données.

[183] CANFAR: CADC https://www.canfar.net/en/nodes/cadc/ (consulté en avril 2021).
[184] Calcul Canada: Canadians lead in transforming genomic data into knowledge to drive medical innovations https://www.computecanada.ca/news/canadians-lead-in-transforming-genomic-data-into-knowledge-to-drive-medical-innovations/ (consulté en avril 2021).
[185] InsideHPC: Compute Canada's Magic Castle: Terraforming the Cloud for HPC https://insidehpc.com/2020/02/compute-canadas-magic-castle-terraforming-the-cloud-for-hpc/ (consulté en avril 2021).
[186] SFU: iReceptor Architecture http://ireceptor.irmacs.sfu.ca/architecture (consulté en avril 2021).

Utilisation du soutien informatique



Graphique 10: Répartition des billets de soutien par domaine de recherche

Le graphique 10 ci-dessus montre le nombre de billets de soutien soumis à l'équipe d'assistance de la FCC entre 2016 et fin octobre 2020. Chaque barre horizontale correspond à différentes disciplines de recherche, tandis que chaque couleur correspond à différentes années. La longueur totale de chaque barre correspond au nombre cumulé de billets de soutien pour cette discipline. Le nombre total de billets est d'environ 38 500 sur les cinq années. Historiquement, le nombre de billets était plus faible en 2016 et 2017, avec environ 1600 et 4400 billets respectivement, mais depuis, la croissance s'est ralentie avec environ 9600 billets en 2018 et environ 12 300 billets en 2019. Une grande partie de l'augmentation initiale était due à la migration des utilisateurs d'anciens systèmes avec une assistance régionale/institutionnelle vers des systèmes nationaux avec une assistance nationale. Le total de fin octobre 2020 est d'environ 10 500 billets, ce qui indique qu'en 2020, le nombre total de billets sera à peu près le même qu'en 2019. La grande majorité des billets de soutien transmis à la FCC est liée à l'utilisation générale des grappes et de l'infrastructure de CIP de la FCC. Une grande majorité des billets de soutien n'est pas liée aux besoins scientifiques propres à un domaine.

La croissance du nombre de billets est nettement supérieure depuis 2016 à l'augmentation correspondante du nombre d'utilisateurs (comme indiqué plus haut, le TCAC pour la croissance des utilisateurs de la FCC est d'environ 12 % depuis 2014). Ceci montre comment l'équipe centrale de la FCC a créé et mis à niveau le service d'assistance afin de répondre aux besoins de la communauté. La stabilisation en 2019 et 2020 du nombre de billets traités s'explique par quelques facteurs. Comme indiqué plus haut, une grande partie de la croissance initiale (2016, 2017, 2018) était due à la transition depuis les anciens systèmes vers les systèmes nationaux, où les utilisateurs sont passés d'un service d'assistance régionale, à un service national. La croissance du service d'assistance nationale correspond probablement à une diminution des services d'assistance régionale (même si le groupe de travail sur le CIP ne dispose pas de

données régionales pour étayer cette hypothèse). De plus, le nombre de billets est plus étroitement lié au nombre de *nouveaux* utilisateurs, plutôt qu'au nombre total d'utilisateurs. Les utilisateurs aguerris ont tendance à poser moins de questions que les débutants. Le nombre de billets devrait donc suivre le taux de croissance plutôt que le nombre absolu d'utilisateurs.

Le tableau 5 ci-dessous indique le pourcentage de billets de soutien par discipline de recherche en 2019 (deuxième colonne). À des fins de comparaison, la troisième colonne montre la répartition des utilisateurs, tandis que les quatrième et cinquième colonnes indiquent l'utilisation relative des ressources de CPU et GPU (comme indiqué précédemment dans ce chapitre).

Discipline	Billets de soutien en 2019	Utilisateurs par discipline	Utilisation de CPU en 2019	Utilisation de GPU en 2020
Sciences biologiques et de la vie	19 %	18 %	8 %	12 %
Ingénierie	18 %	19 %	28 %	7 %
Informatique et sciences de l'information	12 %	14 %	4 %	41 %
Chimie et biochimie	11 %	9 %	20 %	24 %
Physique	10 %	10 %	20 %	9 %
Sciences de l'environnement et de la terre	7 %	6 %	6 %	0,4 %
Sciences médicales	6 %	8 %	2 %	2,0 %
Mathématiques et statistique	4 %	5 %	2 %	0,8 %
Astronomie	3 %	3 %	6 %	1,4 %
Sciences sociales	**1,3 %**	**2,2 %**	**0,3 %**	0,2 %
Sciences humaines	**1,1 %**	**1,2 %**	**0,03 %**	0,0 %
Psychologie	**1,1 %**	**1,5 %**	**0,2 %**	0,5 %
Commerce	**0,9 %**	**1,1 %**	**0,2 %**	0,2 %

Tableau 5 : Billets de soutien par discipline de recherche

Si l'on considère les disciplines pour lesquelles plus de 3 % des billets ont été soumis, le nombre de billets semble être en corrélation avec le nombre d'utilisateurs par discipline de la FCC. L'utilisation de CPU varie selon la discipline, de sorte que la biologie et les sciences de la vie font appel à des services d'assistance proportionnellement à leur représentation, mais n'exploitent que la moitié des années-CPU en termes relatifs. D'autre part, la physique et l'astronomie

soumettent un nombre de billets tout aussi représentatif, mais (sans surprise) consomment le double des ressources de CPU par rapport à la base d'utilisateurs. En ce qui concerne l'informatique et les sciences de l'information, il convient de noter qu'elles consomment relativement peu de ressources de CPU (4 % du total) et beaucoup plus de ressources de GPU (40 % du total), il est donc probable que leurs demandes de soutien soient principalement liées à cette ressource.

Les disciplines dans les quatre catégories qui ont soumis le moins de billets (sciences sociales, sciences humaines, psychologie et commerce) contribuent à environ 4,4 % de tous les billets de soutien, tout en représentant 6 % des utilisateurs par discipline. Si l'on regarde l'utilisation des ressources de CPU correspondantes, ces disciplines consomment environ 1 % des ressources de CPU et GPU. Les chercheurs dans ces domaines soumettent des billets de soutien à un rythme correspondant à leur représentation en nombre d'utilisateurs, mais ne consomment qu'un quart des ressources de CPU en termes relatifs par rapport aux gros utilisateurs traditionnels. Cet écart pourrait indiquer qu'il est nécessaire de cibler ces disciplines, non seulement en obtenant des comptes pour les chercheurs, mais aussi en permettant à ces derniers d'exploiter efficacement les ressources de CIP et de l'IRN pour leur travail, notamment par le biais de formation, d'assistance et de documentation ciblées, ainsi que de nouveaux intergiciels et passerelles innovants pour accéder à l'IRN.

Formation

La formation est une activité très importante au sein de la Fédération Calcul Canada (FCC), notamment les séminaires, les ateliers et les écoles d'été. Elle est essentielle pour l'adoption, la diffusion, la formation de la main-d'œuvre numérique, la mise à jour des compétences des chercheurs (dont plusieurs apprenaient des techniques d'apprentissage machine pour la première fois dans le cadre de ces activités). La FCC doit également produire des rapports annuels à ce sujet pour la FCI.

Région	Nombre total d'événements en personne organisés dans la région	Nombre total de sites où des événements en personne ont été organisés	Nombre total de participants à tous les événements en personne	Nombre total d'heures de formation lors d'événements en personne	Nombre total d'événements en ligne organisés dans la région
ACENET	75	7	1 298	3 522	2
Calcul Québec	44	11	905	4 453	2
Calcul Ontario	254	18	10 132	30 114	85
WestGrid	85	7	1,883	7,898	14
TOTAL	**458**	**43**	**14 218**	**45 987**	**103**

Tableau 6 : Formation au sein de la FCC

Le tableau 6 ci-dessus répertorie les formations offertes par la FCC entre avril 2019 et mars 2020. Les lignes de la partie supérieure présentent les formations dispensées par les régions, tandis que la partie inférieure (en italique) énumère les formations de Calcul Ontario que des affiliés locaux ont offertes. Les heures de formation dans la cinquième colonne sont calculées comme le total des heures de formation reçues par personne, c'est-à-dire qu'une séance de formation d'une heure avec dix participants est comptée comme dix heures de formation dispensée.

En 2019-20, la FCC a offert un total de 46 000 heures de formation (soit 5,2 années) à environ 14 000 participants dans le cadre de quelque 460 événements en personne. À l'échelle régionale, les deux tiers des heures ont été dispensés par Calcul Ontario, suivi de Westgrid avec environ 8000 heures, puis d'Acenet et de Calcul Québec avec environ 4000 heures. Si l'on considère le nombre d'événements organisés en personne, Calcul Ontario représente environ la moitié des 458 événements en 2019-20, suivi de Westgrid avec 19 % des événements, d'ACENET (16 %) et de Calcul Québec (10 %). Le nombre moyen de participants par événement était d'environ 31, ce qui indique la demande et la popularité de ces formations. En examinant le nombre moyen de participants par région, Calcul Ontario est en tête de liste avec 40 participants par événement, tandis que dans les autres régions les taux est à peu près égal, avec environ 20 personnes par événement.

Le contenu de la formation a évolué au cours de la dernière décennie. Chez SciNet, les heures de formation sur le CIP « traditionnel » (MPI et OpenMP) dispensée sont restées à peu près les mêmes depuis 2012, soit environ 1000 heures. Les sciences des données (python, R et ML) sont devenues la plus grande catégorie de formation, de sorte qu'en 2018, SciNet a offert environ

5 000 heures de formation en sciences des données. Le calcul scientifique est un autre domaine émergent, pour lequel environ 2000 heures de formation ont été offertes en 2018.

En ce qui concerne la formation sur le CIP à l'échelle internationale, la FCC et SciNet ont participé à l'International HPC Summer School. Cette école d'été a eu lieu à Toronto en 2015 et elle sera offerte par SciNet en juillet 2021.[187] Tous les ans, 10 étudiants canadiens diplômés participent à cet événement. En plus de la formation, ils bénéficient de mentorat avec des instructeurs experts et du personnel canadien hautement qualifié.

4.4 Quelles sont les forces actuelles de la plateforme de CIP du Canada ?

Dans un rapport sur l'état actuel du CIP en 2017, le CLIR a dressé une liste des nombreuses forces de l'écosystème du CIP au Canada :

- Très bonne prestation de services de CIP
- Évolution et mise à niveau de l'infrastructure de CIP
- Amélioration des services offerts aux chercheurs grâce à une meilleure coordination
- Solide communauté composée de personnel hautement qualifié qui s'engage à fournir des infrastructures et des services de premier ordre aux chercheurs du Canada
- Culture forte et centres d'innovation bien structurés
- Excellent bilan en matière d'adaptabilité et de diversité
- Mise au point d'un environnement réglementaire stable

Depuis 2017, les points forts de l'écosystème de CIP se sont maintenus et même améliorés. Par ailleurs, le financement à long terme renouvelé et centralisé d'ISDE pour l'IRN, qui sera géré par l'Alliance, sera une force additionnelle pour cet écosystème à l'avenir. La section suivante porte sur certains des éléments ci-dessus, notamment en ce qui concerne les changements importants ou le besoin d'accroître les efforts dans ces domaines. Pour un examen détaillé de ces forces historiques, veuillez consulter le chapitre 4.1 du rapport du CLIRN sur le CIP (2017).

Très bonne prestation de services de CIP

Dans le cadre de son processus annuel de renouvellement des demandes de compte, la FCC mène un sondage auprès de ses utilisateurs pour connaitre leurs impressions des ressources et services de CIP. En 2020, 10 900 personnes ont répondu au sondage : 85 % des utilisateurs des plateformes de CIP étaient « satisfaits » ou « très satisfaits » de l'offre de la FCC en général. À l'inverse, seuls 3 % des répondants se sont déclarés « insatisfaits » ou « très insatisfaits ». Le niveau de satisfaction est resté élevé et stable au cours des dernières années. En 2020 par exemple, le pourcentage d'utilisateurs satisfaits était le même qu'en 2017, soit 85 %.

Les utilisateurs de toutes les disciplines de recherche semblent être également satisfaits des ressources et services de la FCC. Il y a peu de variations entre les disciplines et en 2020, la

[187] International HPC Summer School https://www.ihpcss.org/index.html (consulté en mai 2021).

moyenne totale était de 4,32 sur l'échelle de 1 à 5. En ce qui concerne les résultats du sondage sur la satisfaction des utilisateurs en fonction de leur occupation, région ou CAR, il n'y a pas de grandes variations statistiques claires non plus.

Évolution et mise à niveau de l'infrastructure de CIP

Depuis plus de 20 ans, la FCC et ses prédécesseurs mettent au point, soutiennent et proposent des systèmes de CIP haut de gamme à la communauté de recherche canadienne par le biais de ses consortiums. Dans un rapport récent intitulé « Thinking Forward Through the Past : A Brief History of Supercomputing in Canada and its Emerging Future », Calcul Ontario résume bien tous les progrès depuis le début des années 1950. [188] Le taux de satisfaction élevé des utilisateurs de CIP témoigne des excellents principes fondamentaux sous-jacents.

Dans son budget 2018, le gouvernement du Canada a investi l'importante somme de 575,5 M$ dans l'infrastructure canadienne de recherche numérique. La majeure partie de cet investissement avait pour but de financer à plus long terme l'Alliance (375 M$) et CANARIE (145 M$) pour soutenir les réseaux universitaires, mais le budget prévoyait également 50 M$ pour l'évolution et la mise à niveau de l'infrastructure de CIP dans l'immédiat.[189] Si l'on tient compte de la contrepartie des provinces et des partenaires, l'investissement total pour l'amélioration du CIP canadien était de 94 M$, répartis entre les principaux sites d'hébergement : Université McGill (28,1 M$), Université Simon Fraser (39,7 M$), Université de Victoria (9,6 millions M$), Université de Toronto (11 M$) et Université de Waterloo (5,6 M$)[190]. Cet investissement a contribué aux principales mises à niveau du système, comme indiqué au chapitre 4.2. Entre l'hiver 2020 et l'hiver 2021, les quatre sites auront chacun une capacité en ligne.

En plus du financement pour mettre à niveau le CIP, ISDE voudrait également répondre aux besoins potentiels immédiats de stockage à court terme pour le CIP (exercice 2021-22). L'Alliance a formé un groupe de travail sur le stockage qui présentera à la FCI une proposition de mise à niveau du stockage fondée des données probantes au début de 2021.

Top500

À l'heure actuelle, 12 systèmes canadiens (dont 8 pour le CIP) figurent dans le classement Top500 des plus puissants superordinateurs du monde. Parmi ceux-ci, deux sont d'anciennes versions de Cedar et deux sont des versions CPU et GPU du Cedar actuel, donc à proprement parler le Canada a 9 systèmes distincts dans le Top500. Tous les systèmes qui ne sont pas en nuage récemment mis en service par la FCC, à l'exception de Graham, figurent sur la liste : Niagara au 82e rang, Cedar (GPU) au 89e rang et Beluga au 188e rang. En juin 2020, Graham figurait dans le Top500, mais il a disparu de la liste en novembre 2020. Le système

[188] Calcul Ontario: Thinking Forward Through the Past: A Brief History of Supercomputing in Canada and its Emerging Future https://computeontario.ca/wp-content/uploads/2019/07/A-Brief-History-of-Supercomputing-in-Canada-and-its-Emerging-Future.pdf (juin 2019).

[189] Gouvernement du Canada, Innovation, Sciences et Développement économique Canada - Infrastructure de recherche numérique https://ised-isde.canada.ca/site/infrastructure-recherche-numerique/fr (consulté en janvier 2021).

[190] Gouvernement du Canada, Innovation, Sciences et Développement économique Canada - Infrastructure de recherche numérique - Questions et Réponses : https://ised-isde.canada.ca/site/infrastructure-recherche-numerique/fr/questions-reponses (consulté en janvier 2021).

canadien le plus performant (Niagara) atteint environ 3,6 pétaFlops de performance de calcul mesurée, tandis que le plus lent, Beluga, atteint environ 2,3 pétaFlops. Les systèmes les plus lents de la liste de novembre atteignent 1,3 pétaFlops. Par conséquent, Graham n'est pas tout à fait assez puissant (1,2 pétaFlops) pour se classer dans le palmarès. Le Fugaku (Japon), qui est le superordinateur le plus rapide du monde, atteint environ 442 pétaFlops. Il est donc environ 123 fois plus rapide que le meilleur système canadien. [191]

Au-delà des systèmes affiliés à la FCC, le Top500 compte plusieurs autres entrées canadiennes : Banting (128e) et Daley (139e) de Services partagés Canada (SPC). Ces superordinateurs Cray XC50 sont hébergés par SPC et utilisés principalement par Environnement et Changement climatique Canada (ECCC).[192] Dans le palmarès, il y a également quatre systèmes canadiens qui sont basés sur Lenovo et proviennent d'un fournisseur de services d'infonuagiques. Ils sont classés du 333e au 336e rang. Ces systèmes proviennent fort probablement du même propriétaire/exploitant, mais il n'est pas connu publiquement.

Le tableau 7 ci-dessous compare les entrées du Canada dans le Top500 de novembre 2020 aux autres pays du G7 de façon regroupée, ce qui inclut les systèmes de la FCC et deux des autres fournisseurs. Ces statistiques reposent sur les résultats du Top500 tels quels, sans analyse des entrées multiples pour des systèmes identiques ou similaires. Pour des raisons de cohérence et de comparabilité, le nombre d'entrées pour le Canada est de 12 et non de 9 comme indiqué ci-dessus.

Si l'on considère le nombre d'entrées, le Canada se classe au 5e rang, devant l'Italie et à égalité avec le Royaume-Uni. Si l'on considère la puissance de calcul totale agrégée (pétaFlops Rmax), le Canada est dernier, tandis que les États-Unis sont clairement en tête pour la puissance de calcul absolue. Le Japon suit de près, avec environ trois fois la puissance du Canada. L'Italie, qui n'a que la moitié des entrées par rapport au Canada, a trois fois plus de puissance de calcul agrégée. La comparaison entre des pays individuels comme le Canada et les membres de l'UE n'est cependant pas simple en raison de la mise en commun des ressources au sein de l'UE. L'avant-dernière colonne tente de fournir une mesure plus représentative des investissements dans le CIP, en comparant la puissance de calcul agrégée au PIB national.[193] Le Japon a une nette avance sur cette mesure, tandis que l'Italie, l'Allemagne, la France et les États-Unis se situent en milieu de classement. Le Canada se situe à l'avant-dernier rang parmi les pays du G7 (avec un total de 16 téraFlops Rmax sur le PIB en dollars US). Compte tenu de notre richesse nationale, notre capacité en matière de CIP devrait être au moins deux fois plus élevée pour suivre le rythme de nos pairs (qui se situent entre 31 et 39 téraFlops Rmax sur le PIB en dollars US).

---

[191] Top500: Novembre 2020 https://www.top500.org/lists/top500/2020/11/download/TOP500_202011.xlsx (consulté en janvier 2021).

[192] Services partagés Canada : Calcul de haute performance https://www.canada.ca/fr/services-partages/organisation/regroupement-centres-donnees/calcul-haute-performance.html ; et Mise à niveau de l'environnement de calcul à haute performance à l'appui du gouvernement numérique https://www.canada.ca/fr/services-partages/campagnes/histoires/mise-a-niveau-chp.html (consulté en septembre 2020).

[193] Banque mondiale : PIB (dollars US actuels) https://donnees.banquemondiale.org/indicateur/NY.GDP.MKTP.CD?view=chart (consulté en janvier 2021).

Pays	Entrées dans le Top500	PétaFlops Rmax agrégés	PIB (2019, milliards de dollars US)	TéraFlops Rmax agrégés/PIB (téraFlops par milliard USD)	Classement dans le G7 : puissance de calcul du Top500 sur le PIB
Canada	12	27	1736	16	6
France	18	89	2715	33	3 (ex aequo)
Allemagne	17	131	3861	34	3 (ex aequo)
Italie	6	79	2003	39	2
Japon	34	594	5081	117	1
Royaume-Uni	12	34	2829	12	7
États-Unis	113	669	21 433	31	3 (ex aequo)

Tableau 7: Comparaison des classements au Top500 et du PIB des pays du G7

Dans le classement Top500 en novembre 2013, la puissance de calcul agrégée du Canada sur le PIB était d'environ 1 369 gigaFlops par milliard du PIB, soit environ 1,4 téraFlops sur le PIB.[194] En d'autres termes, au cours des sept dernières années, le Canada (ainsi que tous les autres pays en général grâce aux progrès et à la valeur du HPC) a atteint onze fois plus de puissance de calcul par dollar du PIB, tout en avançant dans le classement du G7, passant du 7e au 6e rang.

Amélioration des services offerts aux chercheurs grâce à une meilleure coordination

Dans le cadre de la modernisation de ses services, la FCC a adopté un modèle opérationnel et de soutien plus national, ce qui inclut un environnement informatique et de données plus homogène et cohérent. Les améliorations apportées au modèle de service comprennent notamment un accès uniforme par le biais d'identifiants centralisés (basés sur le protocole LDAP), une meilleure qualité de documentation (centralisée et bilingue), des services de transfert de données améliorés, une approche centralisée pour du stockage plus uniforme (avec une organisation de systèmes de fichiers et des politiques normalisées), ainsi qu'un processus de demande centralisé pour les comptes et l'allocation de ressources. Les utilisateurs finaux disposent désormais d'un point de contact unique pour le soutien informatique à la recherche, tandis que le personnel de soutien de l'établissement est accessible au besoin. Le soutien centralisé offre plusieurs avantages, par exemple l'accès à une expertise plus approfondie à travers le Canada, à un soutien bilingue à travers le Canada, en plus d'une meilleure répartition

[194] Top500: Who are the Top500 list countries? https://www.top500.org/news/who-are-the-top500-list-countries/ (13 juin 2014, consulté en janvier 2021).

du personnel et des ressources accessibles sur plusieurs fuseaux horaires. [195] Outre le soutien centralisé, au moins un tiers des billets de soutien pour le système Niagara de SciNet sont traités en dehors du système central de soutien de la FCC pour le traitement des billets.

La FCC et ses affiliés ont amélioré la capacité de transférer la charge de travail entre les plateformes : « Les nouveaux systèmes permettront aux utilisateurs de Calcul Canada de déplacer plus facilement leur charge de travail entre les différents systèmes, afin d'optimiser l'utilisation des ressources accessibles. Cela sera possible par le déploiement d'un seul système de calcul de haute performance (SLURM), par l'utilisation d'un schéma de dénomination commun pour les logiciels, les modules et les points de montage des systèmes de fichiers, ainsi que l'intégration de mécanismes de déplacement des données avec Workload Manager. » [196] Avec un seul système d'ordonnancement par lots, les utilisateurs finaux peuvent avoir recours aux mêmes scripts de soumission de tâches sur différents systèmes avec des modifications mineures. Malheureusement, les utilisateurs doivent toujours se connecter à des systèmes individuels pour soumettre des tâches, c'est-à-dire que les systèmes d'ordonnancement individuels ne sont pas centralisés.

Les piles logicielles sur les grappes polyvalentes Graham, Cedar et Béluga sont identiques, tandis que sur la grappe massivement parallèle Niagara, le système logiciel est divisé pour que la pile logicielle principale de la FCC et les piles de Niagara soient accessibles, ces dernières étant les piles par défaut.[197] L'environnement de la pile SW principale de la FCC est documenté, suivi et publié sur Github.[198] La pile SW est transférable, évolutive et optimisée pour l'architecture CPU. Elle est accessible aux affiliés de la FCC et à la population générale du monde entier via la technologie CVMFS (Virtual Machine File System du CERN).[199] Au Canada, le CNRC, divers établissements et groupes de logiciels l'utilisent.

En plus des améliorations susmentionnées du service d'identification, d'autorisation et du service de distribution de logiciels, la FCC a également amélioré les rapports système grâce à son service centralisé qui fournit des informations actualisées sur les ressources accessibles.

Solide communauté composée de personnel hautement qualifié qui s'engage à fournir des infrastructures et des services de premier ordre aux chercheurs du Canada

Le réseau de la FCC inclut environ 250 membres du personnel hautement qualifié (PHQ), qui dirigent les activités et les sites de la fédération au Canada. Ces personnes offrent divers services essentiels pour l'administration des systèmes de CIP, l'approvisionnement, la maintenance, le réseautage, les opérations, la gestion, la planification, le financement, le soutien, le développement de logiciels de recherche, la gestion des données, de la formation et des

[195] Exposé de position de l'Alliance soumis par Maxime Boissonneault : https://engagedri.ca/successes-and-shortfalls-of-the-current-canadian-arc-platform-and-ideas-to-improve-it-further (consulté en janvier 2021).
[196] Calcul Canada : Utilisation et capacités https://www.computecanada.ca/la-plateforme-canadienne-de-cip-pour-la-recherche/utilisation-et-capacites/?lang=fr (consulté en septembre 2020).
[197] Calcul Canada : Logiciels disponibles https://docs.computecanada.ca/wiki/Available_software/fr (consulté en septembre 2020).
[198] Github: Compute Canada Software Management https://github.com/ComputeCanada/software-stack/blob/master/doc/INDEX.md (consulté en janvier 2021).
[199] Calcul Canada : Accès à CVMFS https://docs.computecanada.ca/wiki/Accessing_CVMFS (consulté en janvier 2021).

allocations, les communications et la sensibilisation. Les systèmes de CIP sont hautement sophistiqués et complexes à presque tous les niveaux de leur configuration, de leurs piles logicielles et matérielles, de leur exploitation et de leur utilisation. Ceci exige une très grande expertise et plusieurs années de spécialisation. Il est donc très important de maintenir ces compétences et le PHQ de la FCC afin de gérer l'IRN au Canada.

Certains membres du personnel doivent se trouver sur les sites d'hébergement pour accéder physiquement et rapidement au matériel et aux réseaux dans les centres de données. En revanche, la plupart des autres fonctions se font nécessairement à distance. De nombreux systèmes de CIP sont conçus pour être accessibles à distance, de sorte que la seule limite pour de nombreuses opérations est la bande passante et la latence de la connexion réseau, ainsi que d'autres facteurs externes, dont les possibles exigences additionnelles en matière de sécurité, les besoins, les ressources humaines ou les préférences en matière de constitution d'équipes. Dans tous les cas, le CIP se prête au travail à distance et permet aux fournisseurs canadiens de l'IRN de puiser leurs ressources humaines dans plusieurs bassins, peu importe leur situation géographique. Dans certains cas, il est même préférable que le personnel travaille à distance, notamment pour être en mesure d'offrir des services de soutien aux universités et communautés d'utilisateurs à travers les régions.

	Westgrid	Calcul Ontario	Calcul Québec	Acenet	Total
ETP prévus au budget	62,1	67	53	19	201,1
Nombre d'employés	98	76	55	20	249,0
Nombre d'établissements	7	14	8	6	35

Employés du siège social de Calcul Canada non inclus

14 établissements en Ontario incluant 2 hôpitaux de recherche

Tableau 8: Répartition du personnel — Budget 2020-2021 du CIP

Le tableau 8 présente les niveaux d'effectifs PHQ qui sont prévus au budget 2020-21 du CIP. On y retrouve la répartition régionale du personnel, à l'exclusion des employés centraux de CC. La communauté de CIP au Canada est composée d'environ 250 membres du personnel hautement

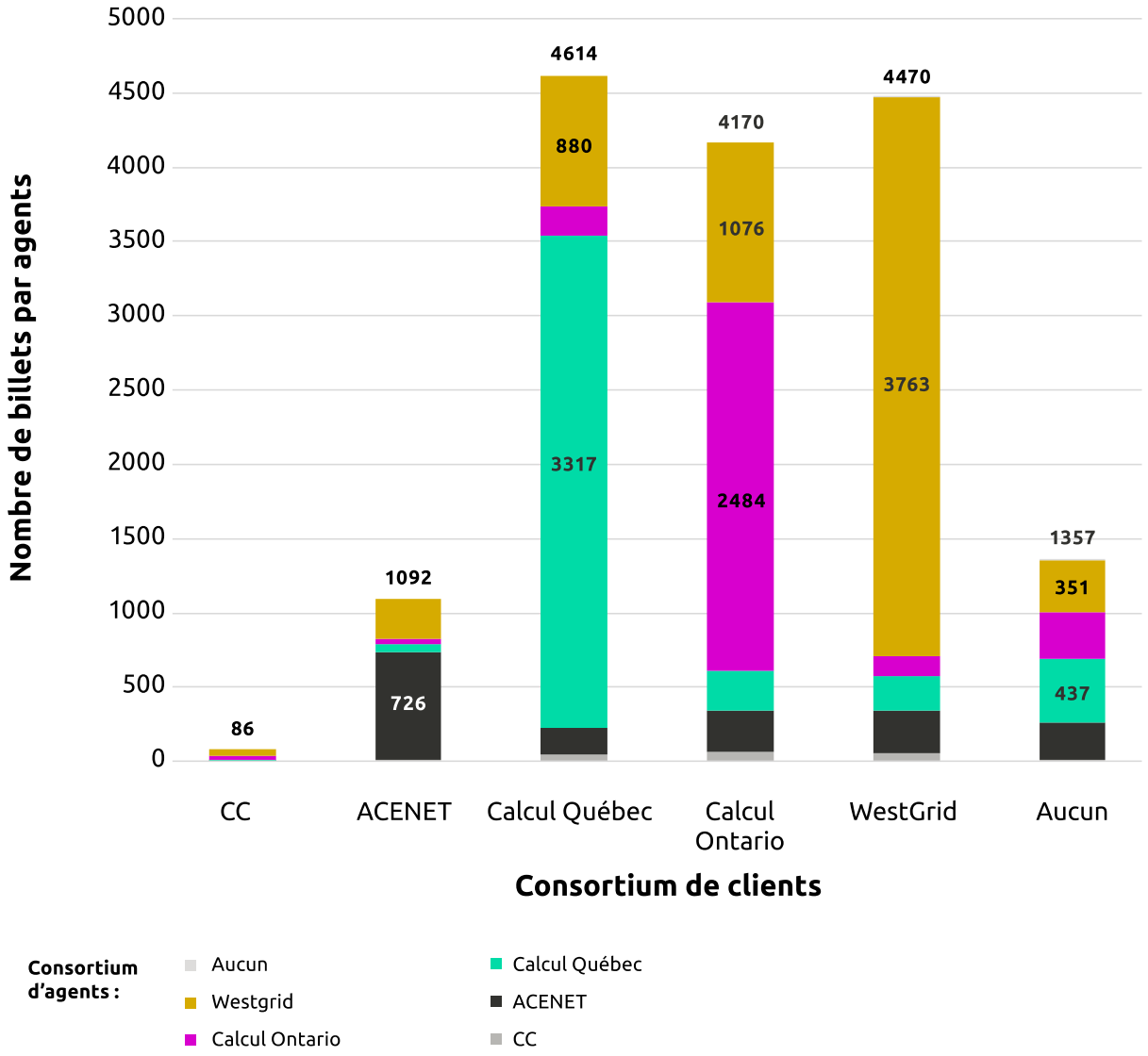
qualifié qui se trouvent dans diverses régions du pays. Les ressources officielles en équivalent temps plein (ETP) pour l'exercice 2020-21 étaient d'environ 200.

L'effectif de 200 PHQ ETP représente environ 6 ETP par établissement membre, un rapport d'environ 1:80 entre le nombre d'ETP du PHQ et les utilisateurs inscrits de la FCC. Il est intéressant de noter que ce rapport est plus favorable, par exemple, au Texas Advanced Computing Center (TACC), le site hôte de Frontera, l'un des superordinateurs les plus rapides au

monde, qui dessert « plusieurs milliers d'utilisateurs »[200] avec un effectif d'environ 190 personnes[201]. Ceci correspond à un rapport personnel-utilisateur entre 1:16 et 1:55, qui est cinq fois plus favorable à presque deux fois plus favorable au TACC. D'un autre côté, il est remarquable de constater que l'équipe de PHQ de la FCC obtient régulièrement un taux de satisfaction de 85 % avec beaucoup moins de ressources par utilisateur qu'au TACC, par exemple.

À l'échelle régionale, Westgrid et Calcul Ontario contribuent le plus grand nombre d'ETP, soit 62 et 67 respectivement, tandis que Calcul Québec contribue 53 et Acenet 19 ETP. Si l'on compare ces ressources ETP au nombre d'utilisateurs des mêmes régions, le ratio est de 82 utilisateurs inscrits pour un PHQ ETP chez Westgid, de 84 chez Calcul Ontario, de 87 chez Calcul Québec et d'environ 55 chez Acenet. En d'autres termes, le ratio de ressources PHQ fournies par utilisateur inscrit est à peu près le même dans les trois plus grandes régions, alors que la situation est plus favorable chez Acenet. Cette comparaison ne tient pas compte du fait que toutes les ressources en personnel hautement qualifié ne sont pas consacrées aux services pour les utilisateurs, tels que l'assistance et la formation. Par exemple, les trois plus grandes filiales régionales du CIP ont un besoin important de personnel administratif afin de gérer leurs principaux sites d'hébergement (ce dont Acenet ne dispose pas actuellement). De plus, le soutien et la formation au sein de la FCC ne se limitent pas au niveau régional, donc les utilisateurs peuvent obtenir un soutien d'une ressource de PHQ qui n'est pas locale ou située dans leur région d'origine.

[200] TACC: Texas boosts U.S. science with the fastest academic supercomputer in the world https://www.tacc.utexas.edu/-/texas-boosts-u-s-science-with-fastest-academic-supercomputer-in-the-world (consulté en décembre 2020).
[201] TACC: Staff Directory https://www.tacc.utexas.edu/about/directory (consulté en décembre 2020).



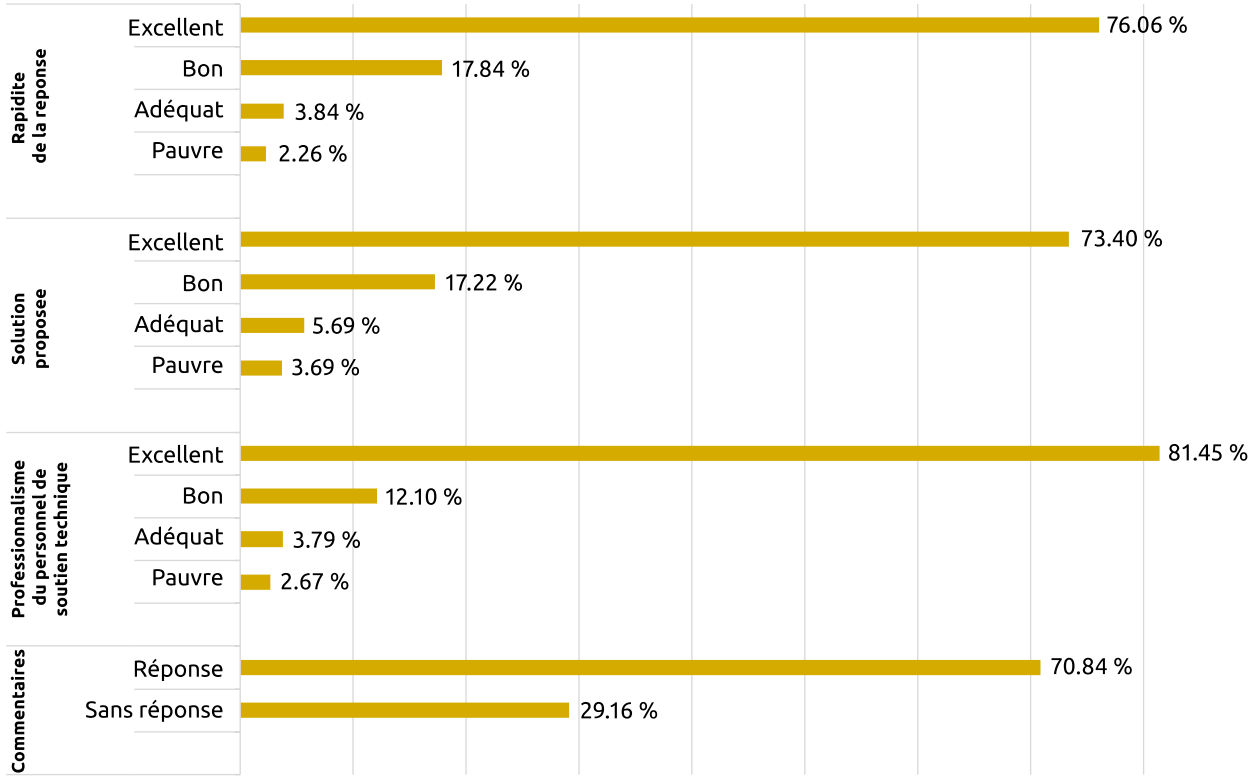
Graphique 11: Nombre de billets de soutien par client et par consortium d'agents

Le graphique 11 ci-dessus montre le nombre de billets de soutien qui proviennent de clients situés dans différents consortiums régionaux de la FCC. Les données concernent l'année civile 2020. Le nombre total de billets provenant des différentes régions est indiqué en haut de chaque barre. Les utilisateurs de trois régions, Calcul Québec, Westgrid et Calcul Ontario, ont soumis un nombre similaire de billets, soit environ 4600, 4500 et 4200 billets, respectivement. Les utilisateurs affiliés à la région Acenet ont soumis environ 1100 billets. Les barres de couleur à l'intérieur de chaque barre principale indiquent les consortiums qui ont traité le billet. Par exemple, le personnel de soutien de Calcul Québec a traité environ 3 300 billets provenant d'utilisateurs affiliés à Calcul Québec et Calcul Ontario a traité environ 2 500 billets de soutien des utilisateurs qui lui sont affiliés. La plupart des billets de soutien soumis par les utilisateurs dans toutes les régions sont traités principalement par l'affilié régional correspondant de la FCC. Par ailleurs, il peut être trompeur de considérer uniquement le nombre de billets. Ce type d'analyse ne tient pas

compte de toutes les personnes impliquées dans la résolution d'un billet donné, ni de l'effort requis pour les billets (qui peut varier d'une minute à plusieurs heures) ou du fait que les experts pour des questions précises peuvent ne pas être répartis de manière égale entre les régions. L'analyse actuelle ne tient pas compte non plus du sujet abordé dans le billet.

Un examen plus approfondi des données révèle des tendances intéressantes. Si l'on garde à l'esprit que la Colombie-Britannique, l'Ontario et le Québec hébergent les principaux systèmes de la CIP et doivent avoir le personnel administratif correspondant, ces régions ont, en termes relatifs, moins de personnel disponible pour le traitement des billets de soutien informatique pour la recherche. Le volume total de billets provenant d'utilisateurs des régions de Calcul Québec et de Calcul Ontario est supérieur au nombre de billets traités par ces régions respectivement. Les quelque 4600 utilisateurs inscrits de Calcul Québec ont soumis environ 4600 billets, tandis que le personnel de Calcul Québec (environ 53 ETP) a traité environ 4100 billets (comme l'indique l'addition de toutes les zones vertes). Les quelque 5600 utilisateurs inscrits de Calcul Ontario ont soumis environ 4200 billets, tandis que le personnel de Calcul Ontario (environ 67 ETP) a traité environ 3100 billets (comme l'indique l'addition de toutes les zones mauves). En d'autres termes, si le personnel de soutien de Calcul Québec et de Calcul Ontario traitait tous les billets provenant de leurs régions, il resterait des billets locaux supplémentaires à traiter par les autres régions, soit 500 et 1100 billets, respectivement.

Les billets supplémentaires de Calcul Québec et Calcul Ontario doivent être pris en charge par d'autres régions, à savoir Acenet et Westgrid. Dans le cas d'Acenet, le personnel local (environ 19 ETP) traite plus de la moitié des billets d'origine locale (environ 700 sur 1 100 billets) soumis par les quelque 1 050 utilisateurs locaux inscrits, puis il prend en charge les billets soumis par des utilisateurs de différentes régions dans une proportion à peu près égale par région, pour un total d'environ 1 400 billets traités par le personnel d'Acenet (comme l'indique l'addition de toutes les zones noires). En d'autres termes, le personnel de soutien d'Acenet a traité environ 300 billets de plus que ceux soumis par tous les utilisateurs de la région d'Acenet. Dans le cas de Westgrid, le personnel local (environ 62 ETP) traite plus de 80 % des billets d'origine locale (environ 3800 sur un total d'environ 4500 billets) soumis par les quelque 5050 utilisateurs locaux inscrits et prend ensuite en charge de manière substantielle les billets soumis par les utilisateurs d'autres régions, pour un total d'environ 6300 billets traités par le personnel de Westgrid (comme indiqué en additionnant toutes les zones brunes). En d'autres termes, le personnel de soutien de Westgrid a traité environ 1800 billets de plus que ceux soumis par tous les utilisateurs de la région Westgrid. Le personnel de Westgrid a traité environ 1100 billets provenant de la région de Calcul Ontario et environ 900 billets provenant de la région de Calcul Québec.



Graphique 12: Réponses au sondage sur la satisfaction post-traitement des billets de soutien de la FCC

Les réponses du récent sondage sur la satisfaction post-traitement des billets de soutien de la FCC reflètent la grande qualité des services offerts par le PHQ de la fédération. Le graphique 12 ci-dessus montre la répartition des réponses recueillies entre septembre 2019 (date à laquelle la FCC a commencé à recueillir ces données) et début octobre 2020. 94 % des participants au sondage ont jugé que la vitesse de réponse est bonne, voire excellente. En ce qui concerne la « solution fournie », 91 % se sont déclarés satisfaits. On note la même tendance pour la question sur la serviabilité du personnel, où 94 % des répondants ont jugé le personnel bon ou excellent. Ces résultats indiquent un niveau élevé de satisfaction à l'égard de l'équipe de soutien de la FCC. Le sondage n'a pas un échantillon suffisant d'utilisateurs pour le superordinateur Niagara de SciNet, car de nombreux billets sont encore envoyés au service de soutien local de SciNet. Pour une analyse et des conclusions plus fiables, il faudrait recueillir plus de données et mener un autre sondage.

Culture forte et centres d'innovation bien structurés

Les principaux sites d'hébergement de la FCC sont gérés par du personnel hautement qualifié, qui fournit des solutions et des infrastructures à l'échelle nationale adaptées aux besoins des universités locales et des régions. Ces centres attirent du talent et servent de terrains de formation pour la nouvelle génération de PHQ, tout en attirant également les universitaires et les chercheurs qui apprécient l'accès local au soutien et aux ressources. La collaboration entre les sites d'hébergement permet d'offrir des services au niveau national (dans la plupart des cas), en tirant

parti de l'innovation individuelle pour améliorer l'ensemble. Comme l'indiquent les classements du Top500, les principaux systèmes de calcul de la FCC sont véritablement de classe mondiale.

Excellent bilan en matière d'adaptabilité et de diversité

Les principaux sites d'hébergement de la FCC doivent satisfaire les nombreuses exigences et besoins d'utilisateurs, tout en maintenant une base stable de cycles de CIP et de services de stockage pour les gros utilisateurs. La coalition de la FCC continue de relever ce défi en fournissant des services de calcul CPU et GPU, ainsi que divers services d'infonuagique pour plus d'adaptabilité. Ce désir de servir la communauté des utilisateurs finaux doit être équilibré avec les limitations et les possibilités de financement actuelles liées, par exemple, aux systèmes contribués, au stockage sur bande, à la capacité de fournir un stockage sécurisé séparé physiquement, aux programmes ciblés pour les communautés EDI et au manque de flexibilité concernant la séparation entre le financement du capital et le financement opérationnel.

Mise au point d'un environnement réglementaire stable

Le Canada est une démocratie stable, progressive et sûre, dotée d'échelons gouvernementaux qui fonctionnent bien, d'un excellent système d'éducation publique et d'un système de soins de santé universel, le tout offrant le cadre et la stabilité nécessaires, directement ou indirectement, aux activités de recherche numérique avancée. Les investisseurs et les bailleurs de fonds, les entreprises, les universités, le personnel hautement qualifié et les chercheurs peuvent compter sur des politiques et des investissements fondés sur des données et la science. Un tel environnement sécuritaire favorise l'innovation, car les gens peuvent être sûrs que les services et les ressources ne disparaîtront pas du jour au lendemain. Par exemple, la réorganisation du financement et des opérations de l'écosystème canadien de l'IRN a été annoncée dans le cadre du budget fédéral de 2018, de sorte que l'Alliance entreverra les activités quotidiennes en avril 2022, laissant à toutes les parties suffisamment de temps pour planifier.

Il est important d'avoir des cadres réglementaires pour protéger la vie privée et la sécurité des personnes, tandis que le libre accès à la science et aux données favoriser les connaissances, la recherche et l'innovation. La communauté de l'IRN canadien devra relever le défi d'équilibrer ces deux dimensions concurrentielles au profit de la société.

Engagement de financement renouvelé du gouvernement l'IRN

Le gouvernement canadien, par l'intermédiaire d'Innovation, des Sciences et du Développement économique du Canada (ISDE), voit clairement la valeur de l'IRN pour la société canadienne, comme le prouve l'engagement budgétaire de 572,5 millions de dollars pour 2018, tel que mentionné ci-dessus. Pour l'Alliance, cela se traduit par un financement fédéral total de 375 millions de dollars jusqu'en mars 2024, ce qui assure une continuité importante (relativement) à long terme au financement de l'IRN. De plus, cette restructuration du financement permet d'équilibrer le CIP, les logiciels de recherche et la gestion des données de recherche, ce qui rassemble les trois piliers principaux d'un écosystème moderne d'IRN.

4.5 Quels sont les défis et les possibilités actuels de la plateforme de CIP du Canada ?

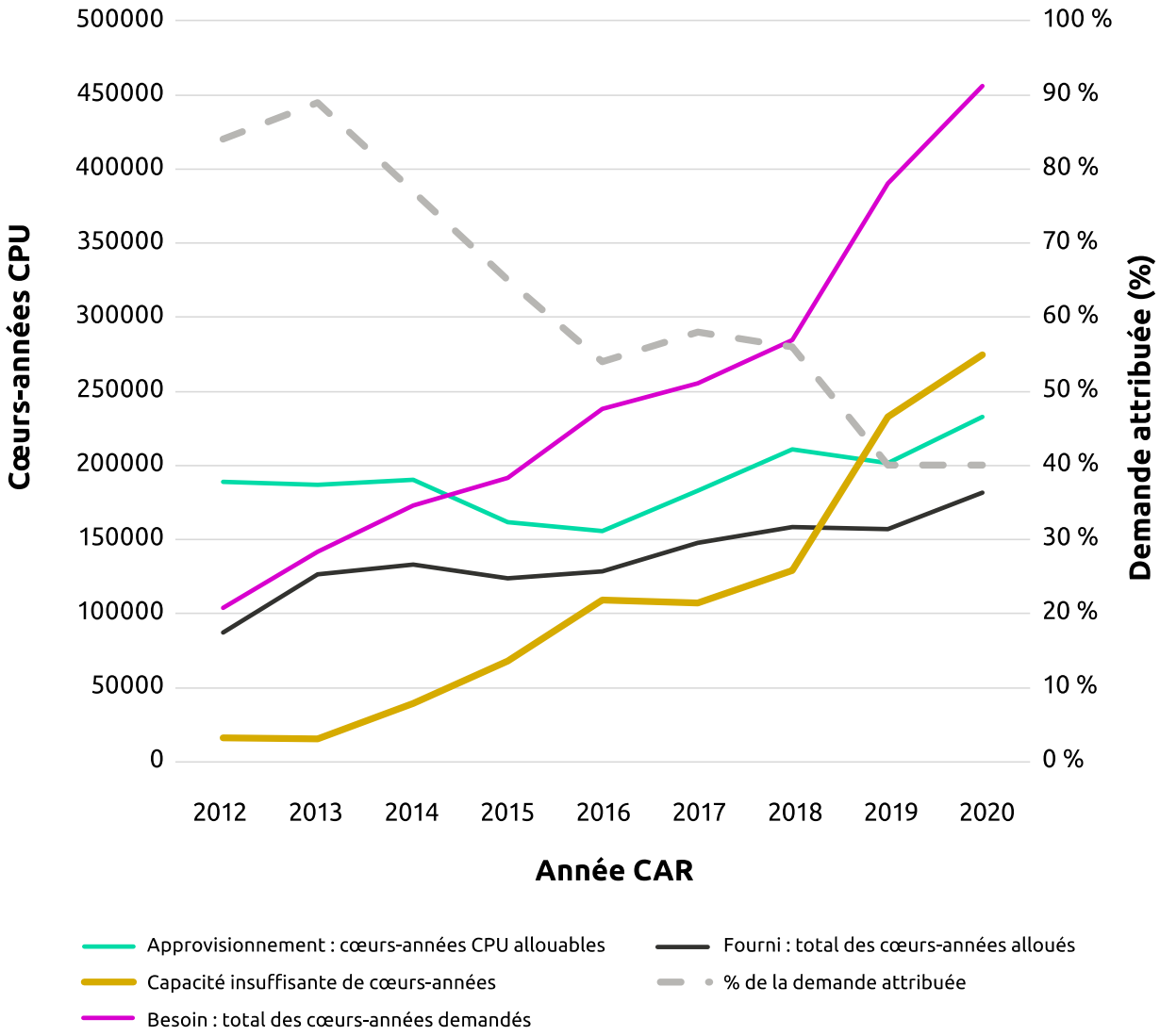
Le rapport du CLIRN sur le CIP en 2017 relève de nombreux défis pour l'écosystème du CIP canadien :

- Offre de CIP insuffisante pour répondre à la demande actuelle et future
- Manque de financement durable et prévisible
- Développement d'une plateforme nationale et modèle de financement actuel
- Coordination de la planification stratégique et opérationnelle nationale
- Attraction et rétention du personnel hautement qualifié
- Collaboration internationale et compétitivité
- Coordination des investissements scientifiques fédéraux et des services de CIP
- Suivi de l'évolution des technologies et du marché
- Exploitation des ressources intersectorielles du CIP
- Sensibilisation des chercheurs et adoption du CIP
- Impact sur l'environnement.
- Sécurisation de la plateforme nationale

Depuis 2017, une grande partie de ces défis n'est toujours pas résolue, même si d'importants investissements ont été réalisés pour contrer le problème des ressources limitées. De plus, ISDE a récemment demandé à l'Alliance de gérer et d'atténuer plusieurs de ces problèmes. La section suivante concerne certains éléments énumérés ci-dessus, notamment en ce qui concerne les changements importants ou le besoin d'accroître les efforts dans ces domaines. Pour un examen détaillé de ces défis historiques, veuillez consulter le chapitre 4.2 du rapport sur le CIP du CLIRN (2017).

Offre de CIP insuffisante pour répondre à la demande actuelle et future

Offre et demande de ressources pour le calcul CPU



Graphique 13: Allocation historique et demande relative aux ressources de CPU de la FCC

Le graphique 13 ci-dessus compare l'offre et la demande de ressources de calcul CPU de la FCC entre 2012 et 2020. L'offre est basée sur la capacité de CPU réellement disponible dans les systèmes de la FCC. Ensuite, la demande est basée sur les demandes d'allocation de ressources soumises lors du Concours annuel pour l'allocation de ressources (CAR) de la FCC. L'axe horizontal correspond aux années d'allocation du CAR, tandis que les unités sur l'axe vertical correspondent à l'équivalent de l'utilisation d'un seul cœur de CPU à 100 % de sa capacité pendant un an (c'est-à-dire les années de cœur de CPU). La ligne verte représente la capacité de calcul brute annuelle totale disponible dans les principaux systèmes de la FCC. Au cours des

huit dernières années, la capacité totale disponible a fluctué dans une fourchette relativement étroite comprise entre environ 155 000 et 232 000 années-cœurs de CPU. En 2014, une capacité importante a été mise hors ligne en raison de son âge et des décisions de la FCC quant à la possibilité de maintenir les équipements plus anciens. Cet élément se reflète dans la baisse de capacité (ligne verte) après 2014, qui n'a pas dépassé de manière significative les niveaux de 2012 avant 2020, date à laquelle l'expansion d'ISDE s'est déployée. La mesure de l'année du cœur du processeur ne tient pas compte de la puissance de calcul réelle de chaque cycle, c'est-à-dire de l'augmentation de la capacité de calcul du processeur grâce aux développements d'architecture.

La ligne mauve représente la capacité totale demandée dans le Concours pour l'allocation de ressources (CAR). Au cours des huit dernières années, la demande est passée d'environ 100 000 années processeurs à 450 000 années processeurs. La croissance de la demande est très rapide et semi-linéaire, mais ne semble pas être exponentielle. En termes de TCAC, la croissance de la demande de cycles de calcul CPU était d'environ 21 % par an.

La ligne noire indique la capacité réelle allouée par CAR. De plus, la capacité totale est fournie aux utilisateurs finaux par le biais du processus de demande au CAR (délibérément plafonné à environ 80 % de la capacité), tandis que la capacité restante (environ 20 %) est accessible aux utilisateurs en fonction de leurs besoins, sans qu'il soit nécessaire de présenter une demande formelle. En 2020, la ressource totale disponible d'environ 232 000 années CPU était répartie entre le CAR (environ 182 000 années CPU) et l'utilisation non allouée/non concurrentielle. La capacité non allouée de 50 000 années-processeurs sert aux utilisateurs qui doivent accéder au système d'allocation de ressources de manière rapide et opportune. Pour illustrer l'échelle, cette capacité totale non allouée sur l'ensemble des sites nationaux en 2020 correspond à environ 60 % de la capacité de calcul annuelle totale de Niagara. Par ailleurs, l'utilisation globale des systèmes est élevée, soit environ 90 % de tous les cycles théoriquement disponibles (le reste s'explique par les temps d'arrêt des composants individuels de systèmes, les arrêts planifiés et non prévus et le fait que la planification des tâches sur les systèmes partagés n'est jamais parfaite).

La capacité allouée (ligne noire) suit de près la capacité disponible (ligne verte), laissant une marge d'environ 20 % pour les services d'accès rapide. Si l'on compare l'offre (ligne verte) et la demande (ligne mauve), on constate que la capacité de calcul de l'unité centrale est en retard par rapport à la croissance rapide des besoins et que le développement de l'infrastructure de CIP n'a pas suivi la demande.

La ligne brune épaisse et continue montre la demande non satisfaite en termes absolus. En 2020, cela représentait environ 274 000 années-processeurs. Cette demande non satisfaite correspond à environ 3,4 fois la capacité du superordinateur Niagara. La ligne grise en pointillés souligne l'ampleur du problème avec le pourcentage de la demande de calcul qui a été effectivement alloué. Historiquement, ce pourcentage diminue, passant d'environ 80 % de la demande satisfaite en 2012 à seulement 40 % en 2020.

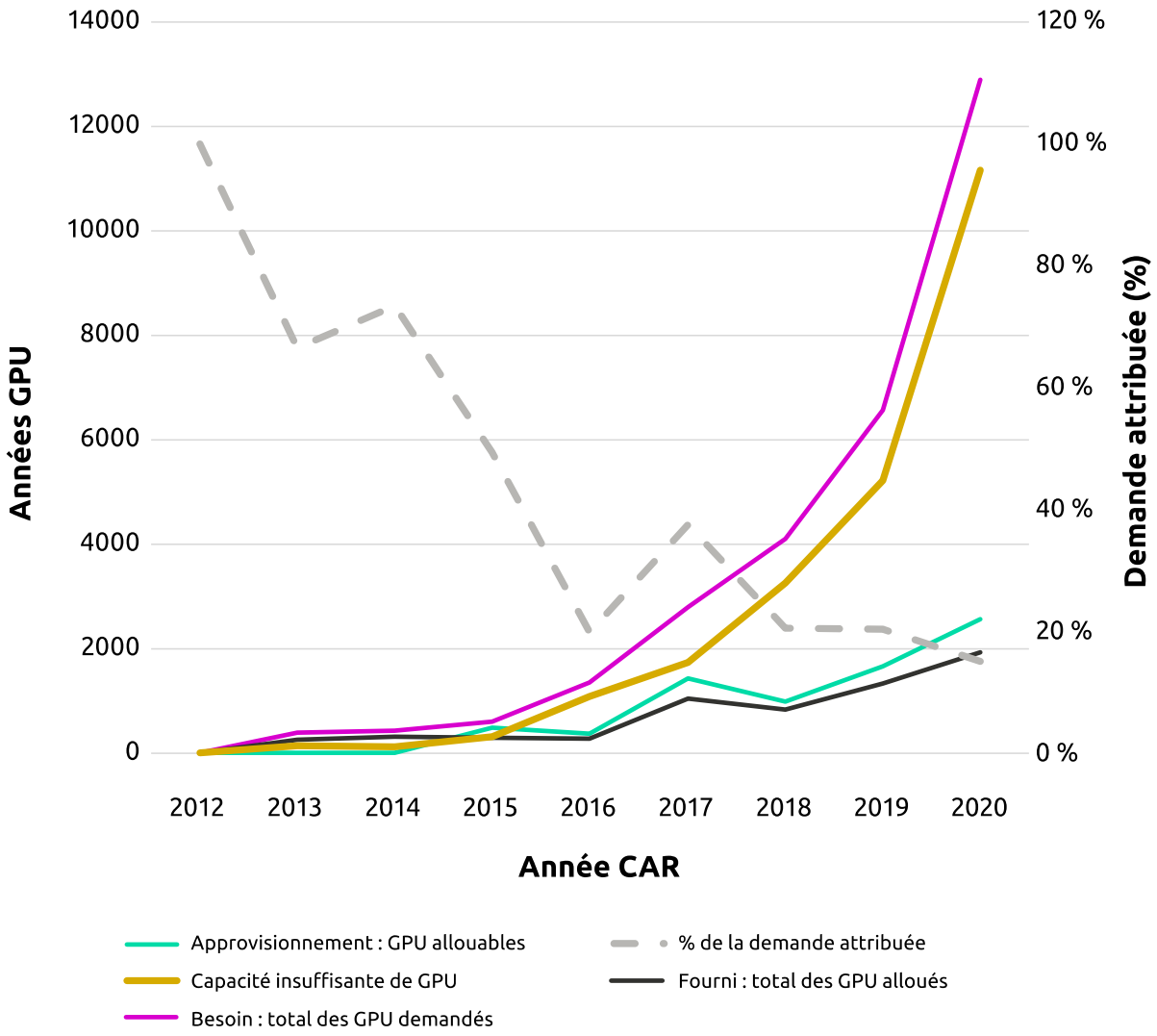
La modernisation de l'infrastructure en 2016-18 a permis de stabiliser temporairement le déficit (voir la ligne brune épaisse), mais n'a pas été en mesure de le réduire en termes absolus ou relatifs. En 2019, la mise hors service de MP2 a de nouveau réduit la capacité de calcul disponible, tandis que les ajouts plus récents ont permis d'augmenter modérément la capacité totale disponible de quelque 30 000 années CPU en 2020. Par ailleurs, l'augmentation rapide de

la demande de ressources a encore aggravé l'écart absolu entre l'offre et la demande. En résumé, au cours de la dernière décennie, le déficit de capacité de calcul CPU s'est considérablement accru, en termes absolus et relatifs.

L'évaluation de la situation de l'offre et de la demande du CIP basée sur l'étude des demandes d'utilisateurs et de l'allocation par le biais du CAR ci-dessus peut donner une image incomplète ou trompeuse des conditions et de la situation sous-jacentes. Les administrateurs du CIP sont conscients des situations où les chercheurs peuvent demander trop de ressources et n'utilisent pas une portion de ces allocations. Cependant, la FCC n'a pas suffisamment de personnel à l'heure actuelle pour aider les chercheurs dans ces situations, notamment pour évaluer correctement l'efficacité des codes et les optimiser pour une meilleure utilisation des ressources. L'Alliance pourrait éventuellement investir à ce niveau pour s'assurer que l'IRN soit optimisé.

Par ailleurs, les ressources de CIP, de par leur nature, sont toujours insuffisantes à cause de l'augmentation constante du nombre de disciplines exploitant l'IRN, de la résolution croissante des instruments expérimentaux ou d'observation et du besoin de simulations à plus haute résolution et précision, souvent ajustées à l'allocation de calcul disponible. Si la ressource informatique disponible augmente, les chercheurs passent rapidement à l'exploitation de cette capacité pour de nouvelles activités scientifiques.

Offre et demande d'accélérateurs GPU



Graphique 14: Allocation historique et demande relative aux ressources de GPU de la FCC

Le graphique 14 ci-dessus compare l'offre et la demande d'accélérateurs GPU de la FCC entre 2012 et 2020. L'offre est basée sur la capacité GPU réellement disponible dans les systèmes de la FCC. Ensuite, la demande est basée sur les demandes d'allocation de ressources soumises lors du concours annuel pour l'allocation de ressources (CAR) de CC.[202] L'axe horizontal correspond aux années d'allocation du CAR, tandis que les unités sur l'axe vertical correspondent à l'équivalent du fonctionnement d'un seul accélérateur GPU à 100 % de sa capacité pendant un an (c'est-à-dire des années GPU). La ligne verte représente la capacité annuelle totale allouable,

[202] Calcul Canada : Résultats des concours d'allocation de ressources pour 2020 https://www.computecanada.ca/page-daccueil-du-portail-de-recherche/acces-aux-ressources/concours-dallocation-des-ressources/resultats-du-concours-dallocation-de-ressources-pour-2020/?lang=fr (consulté en août 2020).

tandis que la ligne mauve reflète la capacité demandée dans le cadre du CAR. La ligne noire est la capacité allouée aux utilisateurs finaux par le CAR.

La demande pour les ressources de GPU a fortement augmenté entre 2012 et 2020. En 2012, les demandes étaient minimes, avec 10 années GPU, alors qu'en 2020, la demande totale du CAR s'élevait à près de 13 000 années GPU. La croissance n'a pas été linéaire, mais exponentielle d'année en année. En termes de TCAC, la croissance était d'environ 67 % depuis 2017 (lorsque la demande était d'environ 2800 années GPU), ce qui indique une augmentation très rapide de la demande pour cette ressource. Cette demande correspond aux tendances mondiales, par exemple, la liste actuelle du Top500 compte un nombre record de 149 systèmes avec accélérateurs (dont 146 utilisent des GPU Nvidia) et 6 des 10 premiers utilisent des GPU (dont le Summit #2 avec plus de 24 000 années GPU).

La capacité de calcul GPU prend du retard par rapport à la croissance exponentielle des besoins. D'un côté cela est positif, montrant un intérêt croissant significatif pour les technologies d'accélération, mais d'un autre côté, l'infrastructure du CIP n'a pas suivi la demande. La modernisation de l'infrastructure en 2016-17 a permis de rattraper et de suivre temporairement la demande, mais depuis, la demande pour ces accélérateurs a clairement dépassé l'offre.

En termes absolus, la capacité GPU non satisfaite en 2020 était d'environ 11 100 années GPU, comme le montre la ligne brune épaisse du graphique ci-dessus. Cet écart entre l'offre et la demande de GPU équivaut à environ huit superordinateurs Cedar actuels. Pour illustrer l'échelle autrement, le coût total de l'achat d'environ 11 000 cartes accélératrices serait d'environ 100 millions de dollars pour les cartes GPU modernes NVIDIA V100 Volta (prix catalogue d'environ 9500 dollars CAD)[203]. Le coût réel du rattrapage des besoins en GPU non satisfaits en 2020 serait encore plus élevé si l'on inclut le coût des milliers de serveurs de CIP et des autres infrastructures de soutien.

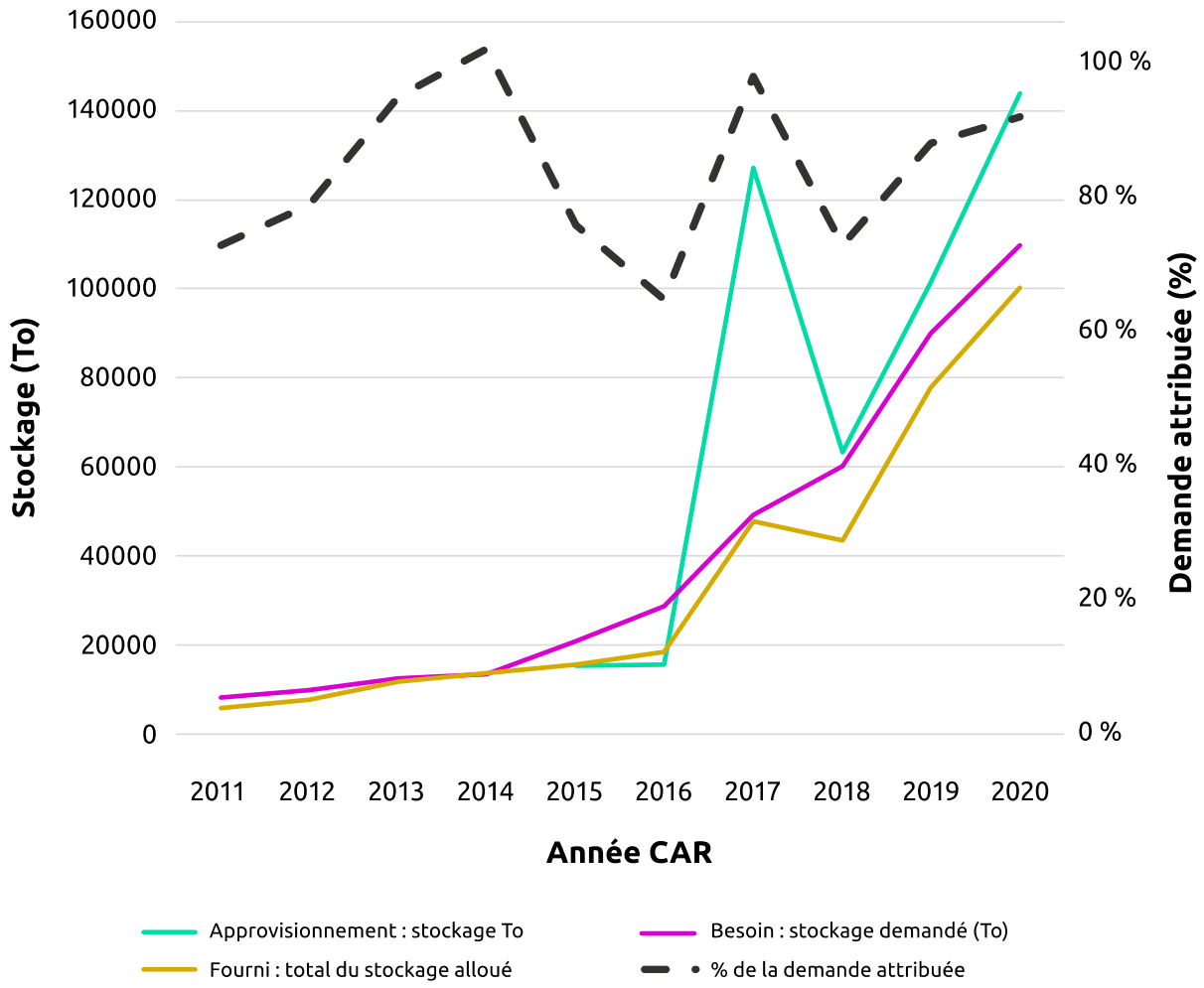
En termes relatifs, la demande non satisfaite (ligne grise épaisse en pointillés) est passée d'environ 100 % en 2012 à environ 20 % en 2020. En revanche, le déficit relatif s'est stabilisé à ce niveau au cours des trois dernières années grâce aux mises à jour des infrastructures, ce qui indique que l'offre et la demande suivent à peu près les mêmes trajectoires de croissance, mais à des niveaux absolus différents, comme indiqué ci-dessus.

Par ailleurs, des tendances semblables ont été observées dans les centres de HPC mondialement lors de la montée en puissance du calcul de pointe CPU. L'utilisation de GPU pour le HPC s'est nettement accrue en 2010, mais elle est difficile en raison de la complexité de la programmation, même avec les premiers paradigmes de programmation GPU orientés vers la programmation scientifique comme CUDA. Les codes de dynamique moléculaire (DM) ont été parmi les premiers à être portés et l'adoption dans cette communauté a été particulièrement forte. L'émergence de l'IA au cours de la dernière décennie a plus récemment fait exploser la demande de ressources GPU. Avec de telles complexités (nouvelle technologie, nouvelles méthodes et nouveaux paradigmes de calcul), il est possible que la demande mesurée par les propositions de CAR soit gonflée encore plus que pour les CPU. Avec l'émergence et l'adoption rapides des GPU, les chercheurs n'ont pas de bonnes bases de référence pour la performance du code, ou pour le nombre de cycles d'entraînement requis ou le temps humain nécessaire pour gérer et interpréter

[203] CDW online store: NVIDIA Tesla V100 – GPU computing processor - Tesla V100 - 16 Go https://www.cdw.ca/product/NVIDIA-Tesla-V100-GPU-computing-processor-Tesla-V100-16-GB/4939179 (consulté en septembre 2020).

les résultats. Il en résulte une surestimation des besoins en calcul GPU. Par ailleurs, le portage de l'ensemble du problème scientifique sur les GPU peut être très difficile, voire impossible. Par conséquent, il y a potentiellement une inefficacité intégrée dans l'utilisation des ressources GPU selon la loi d'Amdahl, lorsque des parties de l'exécution sont liées au CPU uniquement tout en gardant des GPU réservés.

Offre et demande d'allocation de stockage



Graphique 15: Allocation et demande de stockage historique auprès de la FCC

Le graphique 15 ci-dessus montre l'offre et la demande historique de stockage auprès de la FCC. L'offre disponible agrégée des différents types de stockage disponibles est représentée par la ligne verte continue. En 2015, la capacité totale était d'environ 15 Po et elle a presque décuplé pour atteindre environ 143 Po en 2020, avec des variations annuelles substantielles dues au retrait des anciens systèmes et à la mise en service de nouveau. Ce stockage total est réparti entre des types de stockage fonctionnellement différents, notamment les systèmes de stockage

Project, dCache, Cloud et Nearline. Project (57 Po en 2020) [204]est le stockage principal pour les données et les fichiers de recherche active, dCache (15 Po) est un système de stockage de fichiers objet pour les grands ensembles de données (en particulier dans le domaine de la physique des hautes énergies), Cloud (4 Po) est destiné aux instances en nuage et Nearline (68 Po) est un système de fichiers hybride disque-bande pour les données moins actives.

Du côté de la demande, la ligne mauve indique la demande historique totale de stockage, tandis que la ligne brune indique le stockage fourni. La demande s'est multipliée par cinq, passant d'environ 21 Po en 2015 à environ 110 Po en 2020. En 2020 notamment, la capacité totale de stockage (ligne verte) était supérieure d'environ 34 Po à la demande totale (ligne mauve). Une grande partie de la marge de manœuvre additionnelle dans l'infrastructure de stockage est nécessaire pour que le système fonctionne bien. Comme la capacité du système de stockage a encore augmenté au cours de cette période, l'offre a pu suivre la demande, de sorte que le stockage fourni a augmenté au même rythme que la demande. En termes absolus, la demande non satisfaite en 2020 était d'environ 9 Po. La croissance de la demande, multipliée par environ cinq, était plus ou moins linéaire au cours des cinq dernières années, ce qui correspond à un TCAC d'environ 39 %.

La ligne pointillée noire indique la demande allouée en pourcentage des demandes de stockage annuelles. Cette demande satisfaite a varié de 72 % en 2011 à 91 % en 2020, tout en chutant à 64 % en 2016. Grâce à l'augmentation de l'offre, le système de stockage dans son ensemble a été en mesure de répondre à la demande.

Conformément aux politiques de conservation des données de la FCC, le stockage nearline n'est pas un système d'archivage ou de sauvegarde de fichiers. Il est donc uniquement accessible pour des projets actifs.[205] L'archivage ou le stockage à long terme est habituellement considéré comme un stockage pluriannuel (5, 10, 20 ans), qui exige un financement prévisible à long terme. En vertu de son mandat, la FCC ne fournit pas ce type de stockage, même si les systèmes de bandes d'entreprise utilisés pour le stockage nearline à la FCC peuvent, d'un point de vue technique, répondre à de tels besoins. La fédération dispose également de l'expertise et du PHQ nécessaires pour fournir ce type de stockage et pourrait exploiter des systèmes à bandes pendant des années si elle avait le financement et les politiques pour ce faire. L'Alliance a mis en place un groupe de travail sur le stockage à la fin de l'année 2020, afin d'étudier en profondeur les besoins de stockage à court et à long terme dans l'écosystème canadien de l'IRN.

Besoins prévisionnels en astronomie et astrophysique

Comme nous l'avons vu plus haut, l'astronomie et l'astrophysique utilisent considérablement de ressources d'IRN, tant au Canada qu'à l'échelle mondiale. Ces disciplines ont également des écosystèmes d'IRN internationaux éprouvés et bien établis, par exemple pour le traitement de données massives d'instruments d'observation et pour l'exécution de simulations astronomiques. Avec le lancement de nouveaux instruments et l'augmentation des résolutions, les besoins de

[204] Calcul Canada : Résultats du concours d'allocation de ressources pour 2020 https://www.computecanada.ca/page-daccueil-du-portail-de-recherche/acces-aux-ressources/concours-dallocation-des-ressources/resultats-du-concours-dallocation-de-ressources-pour-2020/?lang=fr (consulté en novembre 2020).

[205] Documentation de Calcul Canada : Terminologie des concours pour l'allocation de ressources https://docs.computecanada.ca/wiki/Technical_glossary_for_the_resource_allocation_competitions/fr (consulté en novembre 2020).

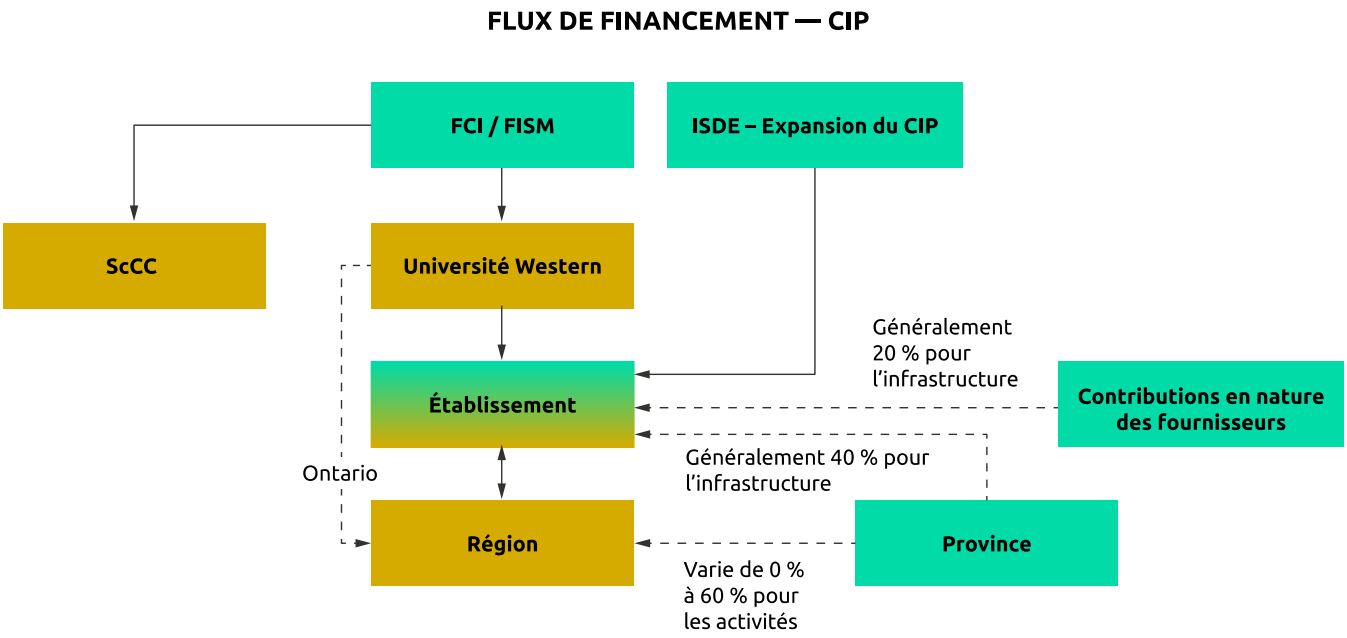
ces « disciplines traditionnelles du CIP » augmentent également rapidement. En décembre 2020, la Société canadienne d'astronomie (CASCA) a publié le plan à long terme PLT2020 de l'astronomie et l'astrophysique canadiennes pour la période de 2020 à 2030.[206] Le besoin général estimé en 2025 est de 100 Pf années de capacité CPU, 100 Pf années de capacité GPU, et 75 Po de stockage en ligne. Si l'on compare ces besoins à la capacité actuelle de toutes les installations de la FCC, ils correspondent, selon la CASCA, à environ dix fois la capacité totale actuelle de CPU et à 25 fois la capacité actuelle de GPU. Les besoins de stockage en 2025 correspondent au volume total de tous les projets de stockage actuels dans les systèmes de la FCC. Outre les besoins généraux en superordinateurs, le projet Square Kilometer Array (SKA1) nécessitera 10 années Pf supplémentaires de capacité de calcul CPU et près de 900 Po de stockage. Les besoins généraux en matière de calcul seront probablement financés par l'Alliance, tandis que les besoins en matière de calcul de SKA1 seront financés par le projet SKA. Il est clair que ces besoins prévus en matière de superordinateurs et d'IRN nécessiteront une attention et un financement particuliers au cours de la prochaine décennie.

Manque de financement durable et prévisible

Le manque de financement durable et prévisible continue d'être un problème dans l'IRN au Canada. Cela a eu des effets majeurs sur la planification à long terme de cet écosystème. Le manque de clarté quant au moment ou à l'ampleur du prochain cycle de financement empêche de planifier efficacement la croissance, les nouvelles technologies et les capacités, ce qui provoque un cycle d'expansion et de ralentissement. L'engagement de financement substantiel dans le budget 2018 pour la création de l'Alliance offre plus de prévisibilité et de continuité jusqu'en mars 2024. Comme indiqué dans le rapport 2017 du CLIRN sur le CIP, les problèmes de continuité du financement peuvent nuire à l'efficacité du système, entraver la capacité des chercheurs à planifier des projets de recherche pluriannuels à long terme et avoir un effet négatif sur la rétention du PHQ en raison des contrats à court terme. [1] Le décalage entre le financement en capital et des opérations à cause des règles et politiques à cet égard est également un problème majeur. Les mécanismes de financement distincts (et les périodes de temps) pour les dépenses d'investissement et les dépenses opérationnelles n'affectent pas seulement les activités traditionnelles de CIP dans des centres de données, mais aussi l'adoption de technologies modernes d'infonuagique à cause des possibilités de financement « artificiellement » limitées. La nécessité d'établir un financement prévisible et durable pour le CIP est la principale recommandation du rapport d'Hyperion Research pour Calcul Ontario en novembre 2019. [207]

[206] Société canadienne d'astronomie - Rapport PLT 2020 https://casca.ca/?page_id=11501&lang=fr (décembre 2020).

[207] Hyperion Research: Summary Report of Phase 1 Study to Support Compute Ontario ARC Planning https://computeontario.ca/wp-content/uploads/2019/11/Hyperion_Summary.pdf (novembre 2019).



Graphique 16: Flux de financement du CIP

Le graphique 16 illustre la circulation très complexe des fonds dans le modèle de financement actuel du CIP. À cause d'un système hérité, le financement de la FCI pour l'infrastructure et les activités de CIP est acheminé par une seule université (Western) et non par Calcul Canada, l'autorité centrale. De plus, le financement du site principal est acheminé ultimement aux universités individuelles qui hébergent le centre de données et non par l'intermédiaire de l'organisation affiliée régionale de la FCI. Les récents fonds additionnels de 50 millions de dollars d'ISDE pour l'expansion du CIP ne sont pas passés par l'Université Western, mais ils ont plutôt été versés directement aux sites d'hébergement principaux. Sans entrer dans les détails et sans discuter du financement opérationnel, il est clair que le financement et le flux des fonds pour le CIP sont inutilement compliqués au Canada. La création de l'Alliance et le fait qu'ISDE et la FCI reconnaissent ces complexités devraient améliorer la situation à l'avenir.

Puisque le financement est souvent sporadique, notamment pour des besoins immédiats tous les 4 à 5 ans, on note des achats importants à plusieurs années d'intervalle. En revanche, les technologies informatiques progressent continuellement, donc il serait plus rentable de pouvoir faire des achats différés. Comme nous l'avons vu dans la section sur le classement Top500, un investissement relatif à peu près identique (en termes de PIB) au cours des huit dernières années a permis d'obtenir plus de dix fois plus de puissance de calcul brute au Canada.

Développement d'une plateforme nationale et modèle de financement actuel

De façon générale, le financement du CIP au Canada se classe en trois volets, le premier étant le financement en capital pour l'achat de matériel, de logiciels de soutien et de garanties. Le deuxième concerne le financement opérationnel pour l'exploitation du matériel, y compris les coûts des services publics, de location, d'assurance et la dotation en personnel de l'administrateur des systèmes. Le troisième porte sur le financement opérationnel pour le soutien

des utilisateurs, y compris les coûts de dotation en personnel connexes. Dans le premier volet, le financement pour l'infrastructure est versé par vagues, ce qui mène potentiellement à des lacunes pluriannuelles et l'incertitude du financement à long terme. Pour le deuxième, le financement est accordé annuellement par le biais du Fonds pour les initiatives scientifiques majeures de la FCI. Il se rattache à l'exploitation des infrastructures sur les sites d'hébergement du premier volet. Le troisième volet est indépendant du deuxième et il concerne les établissements d'attache des chercheurs qui utilisent les ressources de CIP au Canada. La formule de financement de contrepartie est actuellement de 40/60, c'est-à-dire que pour 40 % de financement fédéral, il faut 60 % de contrepartie des universités, provinces, escomptes de fournisseurs, etc. Dans le volet de financement en capital, le rapport 40/40/20 se traduit par un rapport d'environ 50/50 grâce aux généreux escomptes que les fournisseurs accordent à la FCI (ils accordent presque automatiquement un escompte de 20 %, ce qui équilibre les contributions fédérales et les autres contreparties). Dans le deuxième et le troisième volet, le rapport de 40/60 est plus systématique. Le récent financement de 50 millions de dollars pour le programme d'expansion du CIP a injecté des fonds dans le premier volet du financement en capital, mais ne prévoit pas d'augmentation correspondante dans les volets de financement opérationnel.

Comme indiqué dans le rapport 2017 du CLIRN, le modèle de financement de 40 % de la FCI et de 60 % en contrepartie pose problème aux universités et aux régions quand il s'agit de répartir équitablement les coûts en capital et les coûts d'exploitation. Dans certains cas, les établissements qui hébergent l'infrastructure principale de la FCC se sont engagés à financer des activités et des coûts permanents au-delà de leur utilisation locale ou régionale. Par ailleurs, certaines universités et régions utilisent les mêmes ressources gratuitement, ou ne contribuent qu'aux coûts d'infrastructure, mais pas aux coûts opérationnels, ou vice versa. Seule une petite partie des universités et collèges (35) contribue au financement des coûts du CIP. Certaines régions fournissent moins de financement ou de soutien pour le CIP, alors que dans d'autres cas, le financement de contrepartie est accessible. Le financement peut également dépendre d'autres priorités budgétaires émergentes. Par conséquent, on ne peut pas compter (entièrement) sur ce financement à cause de sa nature sporadique. Il faudra notamment garder à l'esprit les conséquences budgétaires de la pandémie actuelle de la COVID-19 à l'échelle provinciale.

Le modèle de financement actuel et les processus d'allocation ne favorisent pas non plus, par leur nature, la collaboration entre les régions et les sites d'hébergement, car il est difficile et peu probable (mais pas impossible) que les provinces financent des projets en dehors de leurs frontières. Les principaux sites d'hébergement possèdent et exploitent les systèmes qui se trouvent dans leur établissement, donc plusieurs administrateurs de systèmes locaux exploitent ces systèmes en fonction de leurs besoins et expertise. En ce qui concerne le soutien aux utilisateurs, les activités se rattachent davantage aux chercheurs des établissements locaux. Par conséquent, le niveau de contribution aux initiatives nationales varie considérablement. À cause d'un manque inhérent d'incitation à la collaboration, il est plus difficile pour le gouvernement fédéral et les agences de financement de coordonner, d'homogénéiser et d'optimiser l'exploitation de plateforme au niveau national et de créer un modèle de service pour une infrastructure d'IRN fédérale cohérente.

À cause de la grande complexité des flux de financement du modèle actuel, les rôles et les responsabilités des diverses entités locales, régionales et nationales ne sont pas clairs, ni faciles à encadrer. Avec un modèle opérationnel plus simple, il y aurait un lien plus étroit entre le financement, les responsabilités opérationnelles et la gestion.

L'Alliance devrait se tourner vers l'UE pour savoir comment l'IRN est financé et exploité dans un modèle fédéré pour les domaines sous-représentés (sciences sociales et humaines par exemple). De nombreuses organisations intéressantes participent au financement et à l'exploitation de l'IRN pour ces disciplines, dont la Fédération EGI, qui propose une infrastructure numérique pour les services de calcul et d'analyse de données avancés[208] ; Parthenos Virtual Research Environment (VRE), un environnement virtuel pour les sciences humaines intégrant le stockage en nuage avec des services et des outils pour le cycle de vie des données de recherche [209]; ARIADNEplus, une infrastructure de données pour la recherche archéologique[210] ; DARIAH, une infrastructure paneuropéenne pour les chercheurs en arts et en sciences humaines [211]; et IPERION HS, une infrastructure européenne de recherche intégrée en sciences du patrimoine[212]. L'ESFRI, un forum stratégique européen sur les infrastructures de recherche, est une importante agence de financement de la recherche en Europe, qui subventionne par exemple le projet DARIAH mentionné ci-dessus et le partenariat PRACE pour le CIP. [213]

Dans son annonce de financement majeur de 2018 pour l'écosystème canadien de l'IRN, le gouvernement du Canada a reconnu certains des problèmes susmentionnés. Par conséquent, le nouveau financement de l'IRN est à plus long terme et il s'étend sur 5 ans jusqu'en mars 2024. L'Alliance a été créée pour planifier et attribuer de manière centralisée la plupart des fonds. Cette continuité permet à l'Alliance de prévoir les achats plutôt que d'être soumise à des cycles. Les achats d'infrastructure peuvent plutôt être répartis sur plusieurs années, en particulier sur les trois derniers exercices de la période de financement : 2021-22, 2022-23 et 2023-24. Une partie du mandat de l'Alliance consiste également à concevoir les plateformes nationales d'IRN de manière holistique, en tenant compte des besoins en matière de CIP, de logiciels de recherche et de gestion des données de recherche, ainsi que des efficiences au niveau national. Un autre changement majeur concerne les exigences en matière de contrepartie. À l'avenir, ISDE/l'Alliance sera le principal bailleur de fonds avec (en général) 60 % du financement en capital et opérationnel, tandis que les 40 % restants proviendront de contreparties.[214] Les principales parties évaluent et négocient actuellement les détails du ratio.

Coordination de la planification stratégique et opérationnelle nationale

L'infrastructure actuelle du CIP, ainsi que le mandat précédent relatif aux activités et le financement de la FCC, repose sur les besoins, les modèles de fonctionnement, les disciplines des utilisateurs et usagers traditionnels du CIP. En pratique, la prestation des services de CIP dépend de systèmes et de grappes à haute performance et massivement parallèles, ainsi que de systèmes de stockage actifs d'exécution ou de projet, avec quelques systèmes en nuage mis en ligne plus récemment. Dans une certaine mesure, cette approche ne répond pas aux besoins d'utilisateurs et de disciplines plus larges, dont les sciences humaines et sociales. Elle ne répond

[208] EGI: About https://www.egi.eu/about/ (consulté en février 2021).
[209] Parthenos: About VRE https://parthenos.d4science.org/web/parthenos_vre/about-parthenos-vre (consulté en février 2021).
[210] ARIADNEplus: About https://ariadne-infrastructure.eu/about-ariadne/ (consulté en février 2021).
[211] DARIAH-UE: DARIAH in a nutshell https://www.dariah.eu/about/dariah-in-nutshell/ (consulté en février 2021).
[212] IPERION HS: About http://www.iperionhs.eu/about/ (consulté en février 2021).
[213] ESFRI Roadmap 2018: ESFRI Projects http://roadmap2018.esfri.eu/media/1044/part1-project-landmarks-list.pdf (consulté en février 2021).
[214] ISDE : PROGRAMME DE CONTRIBUTION POUR L'IRN - GUIDE DU PROGRAMME https://ised-isde.canada.ca/site/infrastructure-recherche-numerique/fr (récupéré en mars 2021).

pas non plus aux besoins de publics plus diversifiés. De plus, l'utilisation et l'accès à ces systèmes reposent sur des technologies traditionnelles, par exemple en utilisant des lignes de commande au lieu de solutions basées sur une interface utilisateur graphique, ou en utilisant des ordonnanceurs par lots traditionnels au lieu de solutions infonuagiques qui cachent souvent l'ordonnanceur à l'utilisateur final. Si l'on considère les tâches interactives (les interfaces graphiques étant par définition interactives) par rapport aux tâches par lots, le principal compromis est de savoir s'il faut préserver les ressources inactives (gaspillées) ou les utiliser immédiatement (interactives). En général, on tente de ne pas laisser les ressources inactives. Par ailleurs, les systèmes et les fournisseurs précédents et actuels dans une certaine mesure de services de CIP n'offrent pas de services en matière de GDR, de LR et de stockage à long terme. Dans plusieurs cas, cette situation n'est pas due à un manque de vision ou de reconnaissance de la part des fournisseurs de CIP, mais plutôt à un manque de ressources (surtout en ce qui concerne les ETP) et à des limitations des mandats de financement.

Il faut une approche plus coordonnée et centralisée de la planification stratégique et opérationnelle de l'IRN au Canada. Cette approche doit avoir une portée nationale pour accroître les synergies et les efficacités (p. ex. en optimisant l'utilisation des ressources), améliorer l'interopérabilité, améliorer la convivialité et mieux utiliser l'expertise du PHQ au Canada. Le manque de haute disponibilité et de haute redondance de l'infrastructure du CIP est également préoccupant. En cas d'incendie majeur, un site d'hébergement pourrait potentiellement perdre totalement et complètement toutes les données stockées sur ce site. Actuellement, il n'y a pas de reproduction hors site et les données sur bande sont généralement hébergées à côté de la grappe. Avec approche coordonnée à l'échelle nationale, ainsi que le financement de systèmes de sauvegarde et de stockage d'archives, on pourrait prévoir la reproduction hors site pour prévenir la perte de données.

Une planification plus centrale et coordonnée doit également tenir compte explicitement non seulement du CIP traditionnel, mais aussi des besoins de stockage à court, moyen et long terme, de GDR et LR de manière générale sous une seule enveloppe, tout en tenant davantage compte des disciplines, publics et communautés mal desservis. Depuis 2018, le Canada et ISDE reconnaissent ce besoin crucial. Par conséquent, l'Alliance devra, en vertu de son mandat, adopter une telle approche plus coordonnée pour financer et exploiter l'écosystème canadien de l'IRN. L'Alliance procède actuellement à une évaluation des besoins, notamment en ce qui concerne l'état actuel du CIP, de la GDR et des LR. L'organisation a aussi lancé un appel à communications et documentation sur l'IRN, en plus d'un sondage sur les besoins des chercheurs et d'assemblées générales. Ce processus aboutira au plan stratégique de l'Alliance pour l'IRN canadienne, au nouveau modèle de prestation de services et à la proposition complète de financement de l'IRN, qui sera soumise à ISDE à la fin de 2021.

Pour harmoniser la prestation des services de CIP au niveau fédéral, l'Alliance doit créer des politiques cohérentes en matière d'accord de participation et d'accord de niveau de service (ANS) pour tous les fournisseurs, les régions et les sites d'hébergement. À l'heure actuelle, les services sont fournis sans aucun accord de niveau de service ou par l'entremise d'accords de niveau de service avec peu d'indicateurs clés de rendement (ICR) mesurables. Afin d'optimiser l'investissement fédéral dans l'IRN et de gérer les attentes en matière de prestation de services, les ANS doivent dorénavant inclure des ICR clairs et un cadre d'application correspondant. Dans le cadre de ce processus, il faudra également consulter les chercheurs pour cerner l'évolution

des attentes. Par exemple, ceci permettra de déterminer si la communauté des utilisateurs a besoin de soutien 24/7 ou si la perte de données est nulle pour le stockage.

Attraction et rétention du personnel hautement qualifié

Les systèmes de CIP sont presque par définition des systèmes de pointe très complexes, combinant diverses technologies avancées qu'exigent de nombreux types d'utilisation distincts. Il faut donc du personnel de haut niveau hautement qualifié, dont la formation prend des années, pour exploiter ces systèmes. Ce personnel doit également s'investir dans la formation continue et à l'apprentissage en cours d'emploi pour suivre l'évolution constante des tendances. De nombreux exploitants recherchent les compétences de ce personnel hautement qualifié, surtout depuis l'émergence ou l'adoption au niveau commercial de l'infonuagique, de l'IA et de l'informatique quantique.

La FCC dispose d'un bassin d'environ 250 employés très qualifiés et motivés. L'Alliance doit donc se donner comme priorité de maintenir ce personnel pour fournir des ressources et des services de haute qualité dans les années à venir. La rémunération est importante, peu importe le degré de motivation de la personne. Malheureusement, les taux de rémunération des universités et des gouvernements ne sont pas concurrentiels par rapport au secteur privé. En revanche, la FCC et la communauté du CIP de l'Alliance bénéficient de facteurs non tangibles avec les avantages sociaux, la sécurité d'emploi, l'environnement de travail, l'équilibre entre vie professionnelle et vie privée qu'offrent les universités. L'Alliance doit donc collaborer avec tous les fournisseurs de l'IRN pour renforcer ces facteurs afin d'être concurrentielle et préserver le PHQ dans l'IRN du secteur public.

La sécurité d'emploi à long terme du personnel hautement qualifié est problématique, notamment parce que plusieurs postes de la FCC pour le PHQ sont temporaires. En Ontario, le personnel dans le domaine du CIP est embauché explicitement pour les systèmes nationaux, alors que le financement est quinquennal et que la FCI (dans une certaine mesure CC également) peut le revoir et modifier annuellement. Il est peu probable que les universités créent des postes permanents et il est assez courant qu'un poste dépende d'une subvention. Cette situation pourrait potentiellement poser problème, même si plusieurs membres du PHQ ont dix ans d'ancienneté malgré ces circonstances. Fait intéressant, le personnel de la FCC est plus préoccupé par l'incertitude que cause la transition vers l'Alliance et son statut d'emploi après 2022.

Le groupe de travail sur le CIP n'a pas assez de données sur le nombre de contrats à durée limitée pour le PHQ, en dehors de l'observation générale selon laquelle « de nombreuses personnes ont des contrats annuels ». Les mandats à court terme (surtout de moins de 12 mois) sont décourageants pour les employés à cause de l'incertitude à plus long terme et peuvent facilement pousser les employés à chercher un poste permanent ailleurs. Si les contrats à durée déterminée sont nécessaires pour des raisons budgétaires, il est important de l'expliquer aux employés en soulignant la continuité à long terme. Les bailleurs de fonds devraient plutôt s'engager à financer la dotation en personnel au-delà de la durée explicite des subventions du projet. De plus, le principal bailleur de fonds, par exemple ISDE ou le gouvernement du Canada, devrait s'engager à financer les activités et la dotation en personnel de base de l'IRN de façon permanente et non temporaire.

En 2018, Calcul Ontario a demandé un rapport sur le personnel hautement qualifié dans la province. Le rapport Malatest[215] propose douze domaines d'intervention pour attirer, retenir et assurer la formation continue du PHQ. Les recommandations suivantes concernent le recrutement de talents :

1. Promouvoir l'image du CIP en tant qu'important parcours professionnel, à l'instar « Research Software Engineer » (RSE) qui a vu le jour au Royaume-Uni et qui est devenu une initiative internationale[216] ;
2. Soutenir les femmes qui travaillent dans le domaine du CIP ;
3. Accroître la reconnaissance du PHQ qui travaille dans le CIP ;
4. Promouvoir le CIP en sciences sociales et humaines.

Pour retenir le PHQ au-delà de l'Ontario et au Canada en général, le rapport recommande ce qui suit :

1. Favoriser l'emploi à long terme dans le monde universitaire ;
2. Soutenir la communauté du PHQ au Canada ;
3. Établir le Canada comme référence en matière de CIP ;
4. Promouvoir le Canada comme étant un pays où il fait bon vivre.

Enfin, le rapport souligne l'importance d'approfondir l'expertise et les compétences du PHQ :

1. Offrir plus de formations continues pour les compétences techniques et les aptitudes en calcul informatique ;
2. Désigner des champions au sein d'établissements ;
3. Favoriser le développement de programmes d'enseignement du CIP ;
4. Promouvoir le travail d'équipe et la communication.

Le groupe de travail sur le CIP souligne aussi l'importance du financement à long terme, de la diversité et des femmes qui travaillent dans l'IRN, en plus de promouvoir l'IRN en sciences sociales et humaines.

Collaboration internationale et compétitivité

La collaboration internationale est primordiale dans de nombreux domaines. Pour que le Canada reste concurrentiel et pertinent en sciences à cette échelle, il est important d'avoir un écosystème national d'IRN compétitif. Le manque de financement durable et prévisible, ainsi que l'insuffisance des ressources informatiques du CIP sont problématiques. De plus, la situation ne s'est pas

[215] Malatest, rapport commandé par Calcul Ontario: Highly Qualified Personnel Study https://computeontario.ca/publications/reports/highly-qualified-personnel-study/ (avril 2018).
[216] Society of Research Software Engineering: https://society-rse.org/about/history/ (consulté en mars 2021).

foncièrement améliorée depuis le rapport 2017 du CLIRN sur le CIP, qui soulève ce problème. D'autre part, la collaboration internationale est désormais plus cruciale pour le partage de connaissances, de financement et de données sur par exemple les changements climatiques et la COVID-19. Les Canadiens collaborent déjà à l'échelle internationale dans de nombreux domaines à forte intensité de CIP. Parmi ceux-ci s'inscrivent le projet ATLAS-TRIUMF du CERN en physique des particules, les plateformes d'analyse intégrées correspondantes de niveau 2 et 3[217] et le Centre canadien de données astronomiques (CCDA)[218], qui exploite une grande partie des ressources de CIP au Canada. L'écosystème de l'IRN doit donc s'étendre au-delà des actuelles solutions axées sur des disciplines et des projets pour inclure de nouveaux projets et collaborations internationaux. Il faudrait notamment fournir une infrastructure et des installations de base pour soutenir les nouveaux systèmes de CIP comme celui du CCDA et offrir des services d'IRN plus généraux pour attirer la recherche et les collaborateurs internationaux dans des disciplines nouvelles et mal desservies.

Les chercheurs canadiens risquent de ne pas être considérés comme des collaborateurs internationaux attrayants s'ils ne sont pas en mesure de pleinement exploiter des ressources nationales de CIP pour leurs projets. Dans certains cas, les chercheurs doivent utiliser les allocations de CIP de leurs collaborateurs internationaux, aux États-Unis, en Australie, dans l'UE ou en Chine par exemple. On ne mesure pas encore pleinement le problème des universitaires qui doivent obtenir des ressources de CIP à l'étranger. L'Alliance mène actuellement un sondage à ce sujet dans le cadre de son processus d'évaluation des besoins. Ceci permettra également d'évaluer l'utilisation des ressources internationales de CIP. Les résultats donneront un aperçu de l'ampleur du problème même s'ils ne sont pas entièrement quantitatifs.

De plus, il pourrait être difficile d'attirer les meilleurs talents universitaires à l'échelle mondiale. Avec le marché hautement concurrentiel des professeurs universitaires de haut niveau, les établissements canadiens risquent d'avoir plus de mal à recruter ces professionnels si leurs homologues étrangers offrent de meilleurs services et ressources en matière de CIP aux niveaux local, régional et national. Comme nous l'avons vu dans ce rapport, le Canada se classe à l'avant-dernier rang parmi les pays du G7 en ce qui concerne la puissance de calcul par PIB dans le Top500, ce qui est un indicateur du manque de concurrence à l'échelle internationale.

En plus des initiatives internationales, la collaboration régionale et interprovinciale s'est aussi accrue au cours des dernières années. Tout comme à l'échelle internationale, ces initiatives connaissent d'importants défis en matière de droit, de politique, d'interopérabilité et de financement, qui exigent des ressources et un soutien à la recherche substantiels pour être résolus et atténués. L'Alliance devrait à l'avenir contribuer au réseautage et à la collaboration internationale, provinciale et régionale afin de réduire et d'éliminer les obstacles aux initiatives de recherche collaborative.

Coordination des investissements scientifiques fédéraux et des services de CIP

Dans son rapport de 2017 sur le CIP, le CLIRN aborde la question du financement scientifique fédéral et de l'allocation des ressources du CIP comme suit : « La recherche financée par le

[217] TRIUMF: ATLAS Tier-1 Data Centre https://www.triumf.ca/atlas-tier-1-data-centre (consulté en février 2021).

[218] FCI : Centre canadien de données astronomiques https://navigator.innovation.ca/fr/facility/conseil-national-de-recherches-canada/centre-canadien-de-donnees-astronomiques (consulté en février 2021).

gouvernement fédéral dépend de plus en plus de l'accès aux ressources de CIP. Cependant, Calcul Canada et les organismes subventionnaires ne coordonnent pas leurs efforts, que ce soit par des programmes à fort impact comme les chaires de recherche du Canada (CRC), les centres d'excellence nationaux (CEN) ou le Fonds d'excellence en recherche Apogée Canada. Par conséquent, les chercheurs et les administrateurs du CIP sont souvent pris par surprise en ce qui concerne l'accès et les exigences en matière de ressources. Avec une coordination mieux établie et plus formelle, ces parties pourraient réaliser des économies, exploiter au maximum les investissements et offrir aux chercheurs des processus plus rationalisés, ce qui réduirait le risque qu'ils se retrouvent dans une situation où ils n'ont pas accès aux ressources de CIP dont ils ont besoin, même s'ils ont des subventions ». [1]

Actuellement, Calcul Canada n'a pas accès aux demandes de subvention des trois organismes, ce qui l'empêche de prévoir les besoins futurs en ressources, d'établir la viabilité technique de la demande ou d'évaluer les tendances générales futures en informatique. À cause de ce décalage, les chercheurs principaux doivent parfois faire une demande de subvention à l'un des trois organismes par le biais d'un processus d'examen des pairs, pour ensuite devoir présenter une demande dans le cadre d'un processus indépendant distinct, également examiné par les pairs, pour obtenir des cycles de CIP. Ce processus est inefficace et lourd. Par ailleurs, le rejet potentiel ou la réduction des ressources informatiques lors du second volet peut faire dérailler une recherche autrement financée.

Dans certains domaines, dont les sciences sociales et humaines, les organismes subventionnaires n'autorisent pas le financement des coûts administratifs et opérationnels des systèmes de TI. De telles politiques exacerbent les problèmes liés à l'adoption de l'IRN dans ces domaines, qui sont déjà sous-représentés et qui n'ont pas nécessairement les ressources informatiques internes pour un éventuel système de CIP, comparativement aux domaines traditionnels du CIP. Pour le financement des sciences sociales dans l'Union européenne par exemple, les aspects techniques et les besoins de soutien opérationnel pour l'IRN sont évalués dans le cadre du processus de demande.

En bref, il faudrait un processus d'examen plus approfondi des subventions au Canada pour les projets qui nécessitent des ressources de l'IRN. Ce processus doit inclure les organismes subventionnaires et les fournisseurs de services concernés, en plus d'évaluer la faisabilité technique et d'envisager le financement des coûts opérationnels. L'Alliance et les trois organismes subventionnaires s'efforcent de résoudre ces problèmes majeurs, avec par exemple la possibilité d'une évaluation préalable à l'allocation où les besoins en matière d'IRN d'une demande seraient examinés avant l'approbation. Il faut donc une planification approfondie et coordonnée, ainsi qu'une coopération entre les principaux bailleurs de fonds fédéraux pour construire un écosystème d'IRN viable, qui intègre les besoins de CIP, GDR et LR pour les disciplines actuelles et futures des utilisateurs de l'IRN.

En plus de ce qui précède, les coûts d'exploitation permanents des systèmes dits « contribués » sont problématiques. À l'heure actuelle, la FCI exige que les systèmes financés par ses programmes d'innovation soient installés sur les sites d'hébergement principaux de la FCC pour contribuer aux ressources du bassin informatique commun lorsqu'ils ne sont pas utilisés par le principal bénéficiaire de la subvention. S'il estime que le système de CIP demandé ne peut pas faire partie des systèmes principaux de la FCC, la demande doit être approuvée par Calcul Canada, puis par l'Alliance à partir d'avril 2022. Si on autorise qu'un système ne soit pas mis à contribution et qu'il ne soit pas installé dans un centre de données de la FCC, les coûts

d'exploitation de ces systèmes incombent naturellement au demandeur ou à son établissement d'attache. Dans le cas de systèmes contribués, la situation est actuellement compliquée, car il n'y a pas de lien entre l'installation du système contribué et le financement opérationnel. Pour un système qui est partiellement utilisé par un groupe de recherche individuel et partiellement utilisé par la communauté du CIP en général, le financement opérationnel n'est pas abordé par les organismes subventionnaires au moment de l'approbation. Il incombe à l'établissement hôte de résoudre la question avec le demandeur, souvent par la suite. ISDE, conjointement avec la FCI et l'Alliance, reconnaît ce problème. Le ministère a donc créé un groupe de travail dont le mandat est de définir des conditions et des politiques pour le financement des opérations de systèmes contribués. Il serait potentiellement plus facile pour les chercheurs d'acheter des services informatiques directement auprès de l'Alliance, plutôt que d'avoir à acheter eux-mêmes le matériel.

Suivi de l'évolution en matière de diversité technologique et culturelle

Historiquement, les progrès technologiques de l'IRN concernent sur le CIP, mais on voit aussi depuis peu la pertinence et les avantages des LR et de la GDR pour la recherche moderne. L'IRN doit inclure ces trois aspects dans sa conception. En plus des avancées méthodologiques et de la chaîne de compilation, l'écosystème de l'IRN est devenu un outil émergent précieux pour les utilisateurs et les disciplines non traditionnelles du CIP, dont les sciences sociales, les sciences de la santé et les études autochtones. Avec l'accès accru aux données, ces disciplines reconnaissent l'énorme potentiel de l'IRN pour leurs recherches. En revanche, ces initiatives donnent aussi lieu à des préoccupations en matière de sensibilité, confidentialité, propriété et sécurité des données. La société reconnaît aujourd'hui davantage les besoins et les revendications de divers groupes sous-représentés, par exemple les communautés racialisées, LGBTQ+ et autochtones. Par contre, les chercheurs de ces disciplines et de ces communautés ne connaissent pas toujours les systèmes et les outils modernes de l'IRN. Ils ont donc parfois des besoins relatifs aux innovations de l'IRN, à la formation, à la documentation, au PHQ ou aux outils dédiés pour accéder aux systèmes de l'IRN. La section « Équité, diversité, inclusion, accès en milieu autochtone et représentation des minorités » ci-dessous aborde cette question plus en profondeur.

En ce qui concerne la technologie traditionnelle du CIP, le paysage évolue rapidement, notamment avec l'émergence et l'adoption croissante de l'informatique GPU, de nouvelles architectures de puces d'IA, de l'informatique quantique ou de l'infonuagique. Par conséquent, les experts des établissements hôtes et régionaux de l'Alliance, de CC et du CIP suivent les récents progrès technologiques dans le domaine du CIP.

Tous les groupes et les tendances susmentionnés requièrent plus d'attention et de nouvelles ressources pour répondre à leurs besoins et aux progrès technologiques connexes. L'écosystème actuel de l'IRN au Canada n'est pas bien équipé ou financé pour répondre à ces besoins. De plus, l'importance était accordée aux utilisateurs traditionnels du CIP, avec une certaine avancée pour l'essai de nouvelles technologies (par exemple de nouveaux paradigmes d'utilisation au-delà de l'accès en ligne de commande par l'entremise de passerelles scientifiques), sans pour autant coordonner le financement à l'échelle nationale.

Comme nous l'avons vu, le gouvernement canadien a reconnu dans son annonce budgétaire en 2018 pour l'IRN et la création de l'Alliance certains problèmes de diversité technologique et culturelle. À l'avenir, l'Alliance tiendra compte de ces préoccupations dans la planification et le

financement de ses activités pour l'IRN au niveau central et national. Par ailleurs, l'équipe responsable de l'IRN de l'Alliance suivra attentivement les tendances technologiques, par exemple les publications techniques internationales, en plus de participer aux conférences internationales (actuellement virtuelles) sur l'IRN. L'Alliance évaluera les besoins actuels au cours de l'hiver et du printemps 2021 afin de déterminer les futures technologies de l'IRN dont les chercheurs canadiens auraient besoin. Ces données permettront ensuite d'analyser les lacunes et de planifier stratégiquement, pour ensuite présenter une proposition de financement à ISDE à la fin de 2021.

Exploitation des ressources intersectorielles du CIP

On ne connait pas les ressources de CIP accessibles aux chercheurs canadiens en général de SPC (ECCC), du Centre de la sécurité des télécommunications (CST) et du Service canadien du renseignement de sécurité (SCRS). Les activités de CIP du CST et du SCRS se limitent naturellement aux services (classifiés et non classifiés) uniquement pour les utilisateurs de base et elles ne sont pas intégrées à l'infrastructure du CIP. En ce qui concerne les ressources de CIP d'ECCC, les scientifiques dans le domaine climatique et météorologique ont toujours des mécanismes pour accéder aux systèmes qu'ECCC héberge pour SPC. Les services météorologiques achètent généralement deux systèmes à la fois, un pour la production et un système qui sert à la recherche (à moins que les autres ne tombent en panne).

Aux États-Unis, l'écosystème diversifié des ressources de CIP est plus largement accessible aux chercheurs universitaires. En guise d'exemple, les ressources qui sont principalement financées par la défense sont également (partiellement) accessibles aux initiatives scientifiques générales, conformément à leurs mandats. Le troisième superordinateur le plus rapide du monde, le Sierra qui est hébergé par le Lawrence Livermore National Laboratory (LLNL) Computing Centre et principalement financé par le DOE et la NNSA, sert à la fois pour des simulations informatiques classifiées relatives à la gestion de la réserve d'armes nucléaires des États-Unis et pour des simulations non classifiées dans les domaines de la physique numérique, des changements climatiques et de la sécurité mondiale.[219]À titre d'exemple de collaboration intersectorielle en matière de CIP aux États-Unis, la Maison-Blanche a récemment annoncé en mars 2020 un partenariat pour la lutte contre la COVID-19, qui permet aux chercheurs internationaux d'accéder aux ressources des superordinateurs du DOE en plus des fournisseurs commerciaux d'infonuagique. [220]

Aucune information publique n'est disponible concernant les capacités de CIP du Centre de la sécurité des télécommunications (CST) et du Service canadien du renseignement de sécurité (SCRS). Aucun de ces systèmes ne figure non plus sur le palmarès actuel du Top500[221], ce qui laisse croire que le CST et le SCRS n'ont pas soumis leurs systèmes au classement, probablement pour des raisons de sécurité nationale. En raison des travaux à forte intensité de calcul que mènent ces institutions, elles sont susceptibles de disposer de superordinateurs très puissants. Comme le montre l'exemple américain ci-dessus, de tels systèmes pourraient

[219] Lawrence Livermore National Laboratories Livermore Computing Centre: Mission Support https://hpc.llnl.gov/about-us/mission-support (consulté en septembre 2020).

[220] Lawrence Livermore National Laboratories: New partnership to unleash U.S. supercomputing resources in the fight against COVID-19 https://www.llnl.gov/news/new-partnership-unleash-us-supercomputing-resources-fight-against-covid-19 (consulté en septembre 2020).

[221] Top500: November 2020 https://www.top500.org/lists/top500/2020/11/download/TOP500_202011.xlsx (consulté en janvier 2021).

potentiellement intéresser les chercheurs canadiens, même si, naturellement, une telle collaboration exigerait de modifier considérablement les mandats (comme dans l'exemple de la double fonction au sein du DOE).

L'accès aux ressources de CIP classifiées (CST, SCRS) ou non classifiées (SPC/ECCC), même de façon limitée, que le gouvernement détient et finance profiterait aux universitaires canadiens et au pays en général. On pourrait imaginer que certains projets de recherche d'importance nationale (p. ex. le développement d'un vaccin contre le virus de la COVID-19) qui utilisent des codes hautement évolutifs et nécessitent une capacité de calcul informatique de pointe soient réalisés à partir des ressources de CIP non traditionnelles financées par le gouvernement canadien et fournies par le SPC/ECCC. Un organisme central de l'IRN, comme l'Alliance, pourrait jouer un rôle dans le lancement, la négociation et la gestion des collaborations intersectorielles de l'IRN et des cadres connexes.

Sensibilisation des chercheurs et adoption du CIP

Les chercheurs connaissent peu et n'ont pas beaucoup adopté le CIP et l'IRN, ce qui constitue un problème majeur au Canada et ailleurs dans le monde. Il y a environ 33 000 professeurs universitaires titulaires et associés au Canada, [222]alors que la FCC dénombre actuellement environ 5500 comptes de chercheurs principaux (CP) dans le cadre du CAR. En d'autres termes, environ 17 % des professeurs titulaires et associés utilisent potentiellement les ressources de CIP (directement). Parmi ces 5500 comptes CP et groupes de recherche correspondants, 3177 ont utilisé le CPU au cours des 12 derniers mois, c'est-à-dire qu'environ 10 % des groupes de recherche potentiels dirigés par des CP au Canada ont utilisé les ressources de la FCC. En fixant le seuil d'utilisation du CPU à plus de 4 années-cœur (l'équivalent approximatif d'un ordinateur portable semi-moderne fonctionnant 365/24/7), ce chiffre tombe à 1545 groupes dirigés par des CP et à 578 groupes de recherche dirigés par des CP avec un seuil de 50 années-cœur pour le CAR (l'équivalent approximatif de 5 à 10 postes de travail fonctionnant 365/24/7). Ce dernier groupe, les « gros » utilisateurs des installations de la FCC, correspond à environ 2 % des groupes de recherche dirigés par des chercheurs principaux admissibles au Canada.

L'adoption de l'IRN peut également être considérée du point de vue de l'utilisation des cycles de calcul dans différentes disciplines. Les chercheurs dans les disciplines sous-représentées (sciences sociales, sciences humaines, psychologie et commerce) totalisent 17 % de tous les utilisateurs du corps enseignant dans la base de données des utilisateurs de la FCC, mais n'utilisent que 1 % de toutes les ressources de CPU. Une partie de ce déséquilibre s'explique par la taille plus réduite des groupes de recherche de ces domaines qui sont dans la base de données de la FCC (ratio utilisateur/CP). Le déséquilibre provient aussi des simulations lourdes en CPU que les chercheurs en physique et en astrophysique entreprennent dans leurs recherches, contrairement à d'autres disciplines qui ne sont pas nécessairement aussi intensives en calcul. En principe, rien ne devrait empêcher les disciplines sous-représentées d'utiliser beaucoup plus le CIP si l'on considère la taille et la diversité des ressources potentiellement accessibles pour les données lourdes, par exemple les bases de données et les données longitudinales et transactionnelles du secteur privé, des gouvernements provinciaux et fédéral, de Statistique Canada, des autorités sanitaires régionales, etc. La combinaison et le recoupement de ces

[222] Statistique Canada : Nombre du personnel à temps plein dans les universités canadiennes, selon le rang et le sexe https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=3710007601&request_locale=fr (consulté en avril 2021).

sources d'information nécessiteraient des ressources de stockage et de calcul considérables. L'analyse potentielle de ces ressources avec l'intelligence artificielle et l'apprentissage machine permettrait d'approfondir les connaissances dans les domaines sous-représentés. En d'autres termes, les sciences sociales, les sciences humaines, la psychologie, le commerce pourraient utiliser davantage le CIP et l'IRN pour faire avancer leurs domaines. En ne desservant pas ces domaines à leur niveau représentatif complet, le Canada risque de rater des occasions et de se faire dépasser par la concurrence internationale.

En plus des disciplines et des communautés qui n'exploitent pas l'IRN, certains chercheurs dans les disciplines « traditionnelles » n'accèdent pas aux systèmes de Calcul Canada pour diverses raisons. Parfois, ils ne connaissent pas la gamme de services, ils considèrent que les interfaces utilisateur et l'utilisation sont trop compliquées ou ils ont abandonné parce que leur demande a été rejetée ou qu'ils ont eu une mauvaise expérience. Il faut donc plus de sensibilisation pour attirer ces communautés et individus. Les efforts de sensibilisation ne devraient pas seulement porter sur les solutions déjà accessibles dans ces domaines, mais aussi inclure des séances de remue-méninges tournées vers l'avenir, au cours desquelles des spécialistes de l'IRN rencontreraient des experts de ces disciplines afin d'élaborer de nouvelles approches et technologies potentielles pour faire progresser la science. Il faudrait également souligner les cas d'utilisation dans les domaines concernés pour montrer comment les outils de l'IRN peuvent leur profiter. Tout effort de sensibilisation doit s'accompagner des ressources de l'IRN en matière d'infrastructure, de services, de formation, de personnel de soutien, ainsi que de nouvelles technologies d'accès et d'utilisation améliorées. Ceci permettrait à tous les chercheurs intéressés de tirer avantage de l'IRN. Sans avoir les ressources correspondantes, la sensibilisation risque de déboucher sur des hypothèses, sans action réelle. Parmi les approches intéressantes en matière de convivialité figurent les solutions d'infrastructure de bureau virtuel (VDI), permettant aux utilisateurs finaux d'accéder aux ressources de CIP dans un environnement de bureau familier. L'accès par un navigateur web (portails ou Jupyter Notebook) est aussi une solution intéressante. Pour favoriser l'adoption du CIP et de l'IRN, il est aussi important d'améliorer les bibliothèques et piles de logiciels, ainsi que l'intégration correspondante. Pour tout ce qui précède, il faut également prévoir des formations ciblées.

Dans le cadre de son mandat futur, l'Alliance tentera d'aborder les questions susmentionnées et de fournir des solutions liées pour sensibiliser les chercheurs et favoriser l'adoption dans les disciplines sous-représentées. En intégrant les LR et la GRD, on pourra concevoir des solutions d'IRN plus complètes qui devraient être plus accessibles et plus pertinentes pour des publics plus larges. L'Alliance s'efforce déjà d'inclure les communautés de recherche sous-représentées dans son processus d'évaluation des besoins, au-delà des utilisateurs traditionnels du CIP (qui figurent dans les bases de données de la FCC). L'organisation veut aussi sensibiliser les organisations qui se rattachent aux disciplines concernées, ainsi que la direction de toutes les universités canadiennes. Le nouveau plan stratégique, le modèle de prestation de services et le financement ont pour but d'aborder et de résoudre les problèmes de sous-représentation.

Impact sur l'environnement

Les centres de données du CIP ont d'importants serveurs et infrastructures de soutien qui fonctionnent presque au maximum de la capacité de calcul, ce qui consomme beaucoup d'énergie (environ 1 % de la consommation d'électricité mondiale). Pour les centres de données de Calcul Canada, cette consommation est estimée de 3 à 5 MW. L'approvisionnement en

électricité du Canada provient à 67 % de sources renouvelables et à 82 % de sources n'émettant pas de gaz à effet de serre, ce qui réduit l'empreinte des émissions de gaz à effet.

L'efficacité énergétique du calcul (flops/W) peut servir de mesure indicative pour déterminer la viabilité environnementale du CIP. Cette mesure ne tient pas compte de l'énergie consommée par le système de CIP lui-même ou de facteurs additionnels, par exemple toute l'énergie consommée doit être refroidie, ce qui augmente l'impact énergétique et environnemental total. On estime que la surcharge de refroidissement actuelle est de l'ordre de 10 à 30 % pour les sites nationaux de Calcul Canada. Au-delà de l'efficacité, la réutilisation de la chaleur générée par une grappe est un important élément qui manque à de nombreux sites CC. Il s'agit notamment d'une fonction de l'infrastructure du centre de données et non des grappes en soi. L'amélioration de l'efficacité énergétique du calcul et la réutilisation de la chaleur sont donc essentielles pour réduire l'impact environnemental des superordinateurs.

Les palmarès Top500 et Green500 sont des références bien connues et bien établies pour le CIP. Malgré certaines limitations considérables, ils peuvent servir d'indicateurs de la performance et de l'efficacité du CIP. Ces palmarès représentent la performance semi-théorique du matériel à partir de microréférences qui ne sont souvent pas représentatives des charges de travail réelles exécutées par la communauté de recherche. De plus, les GPU sont théoriquement efficaces, mais ils peuvent aussi être très mal utilisés ou ne pas être utilisables du tout par diverses applications. Il faudrait plutôt se concentrer sur les performances du code de la recherche réelle et sur l'efficacité des centres de données (réutilisation de l'énergie, etc.). En revanche, il serait difficile d'étendre ces mesures à l'ensemble de l'écosystème international pour faire des comparaisons.

Malgré ces réserves, les systèmes de la FCC récemment installés sont relativement bien positionnés au niveau international dans le classement Green500. Le Canada a cinq entrées parmi les 100 premiers du classement Green500 avec Cedar (deux entrées avec des combinaisons CPU+GPU différentes : # 16, environ 11 gigaFlops/W et # 33, environ 8 gigaFlops/W), Beluga (# 21, environ 9,5 gigaFlops/W), Cèdre (CPU pur, # 33, environ 8 gigaFlops/W) et Niagara (# 67, environ 3,9 gigaFlops/W).[223] Il est encourageant de constater que les systèmes canadiens récemment acquis sont bien positionnés pour l'efficacité énergétique du calcul. D'autre part, si l'on considère la performance brute en gigaFlops, le Canada n'a que deux systèmes parmi les 100 premiers du Top 500, Niagara (70e) étant le système de CIP canadien le mieux classé et Cedar (accéléré par GPU, 74e) qui le suit de près. [224]En d'autres termes, les systèmes de CIP canadiens sont mieux positionnés à l'international pour leur efficacité énergétique que pour leur puissance de calcul brute.

Pour certaines applications et certains cas d'utilisation, l'efficacité des accélérateurs est impressionnante, en ce qui concerne la puissance de calcul brute et l'efficacité énergétique du calcul. Le système le mieux classé dans le palmarès Green500 est le système japonais MN-3, avec environ 21 gigaFlops/W. Il fonctionne avec des puces accélératrices ASIC MN-Core conçues sur mesure pour une tâche spécifique, soit la phase d'entraînement des charges de

[223] Green500: June 2020 listing https://www.top500.org/files/green500/green500_top_202006.xls (consulté en septembre 2020).

[224] Top500: November 2020 https://www.top500.org/lists/top500/2020/11/download/TOP500_202011.xlsx (consulté en janvier 2021).

travail d'apprentissage profond.[225] Parmi les 10 premiers du Top 500, six systèmes utilisent des accélérateurs. L'efficacité énergétique pour le calcul des accélérateurs GPU est démontrée en comparant Niagara, basé uniquement sur un CPU (# 70 du Top500) et Cedar, accéléré par le GPU Nvidia V100 (# 74 du Top500). Niagara atteint à peu près la même puissance de calcul que Cedar, tout en consommant près de trois fois plus d'électricité (920 kW contre 310 kW). En revanche, de nombreux problèmes et applications ne peuvent pas être facilement (ou pas du tout) redistribués ou portés sur les GPU. Pour cette raison, Frontera de TACC, un système financé par la NSF pour la recherche universitaire générale, est principalement un système de CPU pur avec une capacité GPU supplémentaire.

Il est important de noter que si on considère uniquement l'efficacité énergétique du calcul, on ne tient pas compte d'autres facteurs environnementaux importants, comme la viabilité environnementale de la production d'électricité, l'impact environnemental total des systèmes et des matériaux utilisés pendant leur durée de vie, l'efficacité énergétique des systèmes de climatisation et de refroidissement (si nécessaire) ou l'utilisation secondaire potentielle de la chaleur résiduelle. La réutilisation efficace de toute chaleur excédentaire est une application pratique essentielle, notamment pour chauffer les bâtiments adjacents ou même les communautés. Par exemple, la chaleur produite par Mammouth à l'Université de Sherbrooke et Colosse à l'Université Laval est utilisée pour chauffer une partie de leurs bâtiments respectifs. [226] [227]

La réutilisation des équipements usagés permet de prolonger leur durée de vie et d'éviter de les jeter inutilement. Par contre, l'efficacité énergétique de calcul du vieil équipement est inférieure à celle du matériel informatique plus récent. Le vieil équipement tombe en panne plus souvent, ce qui augmente les coûts d'électricité, réduit l'efficacité de l'utilisation de l'espace et augmente les coûts de personnel.

Sécurisation de la plateforme nationale

Selon le Service canadien du renseignement de sécurité (SCRS), la fréquence et la sophistication des cybermenaces contre les intérêts de la recherche canadienne ont augmenté ces dernières années, notamment dans les secteurs de la biopharmacie et de la santé, de l'intelligence artificielle, de l'informatique quantique, ainsi que des technologies océaniques aérospatiales.[228]Les atteintes à la sécurité sont souvent liées à des facteurs humains (mots de passe inadéquats ou ingénierie sociale par exemple), mais elles peuvent aussi inclure des solutions technologiques. Par exemple, les États-Unis s'inquiètent de la sécurité des équipements de réseautique 5G de Huawei et encouragent leurs proches alliés à ne pas utiliser ces

[225] InsideHPC White Paper - Supermicro Contributes to the MN-3 Supercomputer Earning #1 on Green500 list https://insidehpc.com/white-paper/supermicro-contributes-to-the-mn-3-supercomputer-earning-1-on-green500-list/ (consulté en septembre 2020).

[226] U. Sherbrooke: L'ordinateur Mammouth au premier rang au Canada https://www.usherbrooke.ca/sciences/accueil/nouvelles/nouvelles-details/article/17844/ (consulté en mai 2021).

[227] CBC Radio-Canada : Des serveurs informatiques pour chauffer l'Université Laval https://ici.radio-canada.ca/nouvelle/1143499/serveurs-informatiques-chauffage-universite-laval-recuperation-energie (consulté en mai 2021).

[228] Global News: China and Russia 'aggressively' targeting Canadians, CSIS director warns https://globalnews.ca/news/7629494/china-and-russia-targeting-canadians-csis-director/ (consulté en février 2021).

équipements. [229]Le gouvernement canadien n'a pas encore pris de décision à ce sujet, alors que ces questions sont d'actualité et préoccupent l'écosystème de l'IRN canadien, puisque Huawei a fabriqué certains équipements de la FCC (Graham par exemple). Il faut donc trouver un équilibre entre les préoccupations et pressions publiques concernant certains fournisseurs et les politiques et règles officielles d'approvisionnement, par exemple les demandes de propositions (DP).

Comme l'infrastructure de l'IRN devient plus nationale et centralisée, toutes les mesures pour sécuriser l'infrastructure doivent être coordonnées entre les participants, au niveau des sites d'hébergement locaux, des régions et du gouvernement fédéral. Ces efforts doivent protéger les données sensibles, les renseignements personnels, la propriété intellectuelle, ainsi que les actifs de recherche numériques et stratégiques.

Les solutions de CIP sont traditionnellement conçues et exploitées dans des réseaux isolés, donc les menaces externes ne constituaient pas une préoccupation majeure. Durant la conception, l'accent portait plutôt sur les caractéristiques de performance traditionnelles du CIP, dont la puissance de calcul brute, la latence, la bande passante ou l'efficacité énergétique. En tant que tels, les systèmes de CIP n'ont pas été explicitement conçus pour la prévention des attaques ou des tentatives malveillantes. À mesure que de nouvelles possibilités, des ensembles de données sensibles, de nouveaux paradigmes comme l'informatique périphérique et de nouvelles disciplines de recherche exploitent le CIP, les systèmes deviennent plus accessibles et plus répandus à l'échelle nationale. Ils sont donc plus exposés aux tentatives d'accès externe non autorisé et aux menaces. Ces dernières ont beaucoup évolué au cours des deux dernières années, avec des attaques généralisées et très médiatisées sur de nombreux systèmes de HPC, comme le piratage de l'infrastructure européenne de superordinateurs lié au minage de cryptomonnaies en mai 2020.[230] Les sites de CIP de la FCC ont également subi divers types d'attaques au fil des ans. En réponse aux menaces émergentes, ces sites ont augmenté les ressources et les activités pour résoudre les problèmes de sécurité, notamment en améliorant la culture de la sécurité, le financement et les réunions de groupe. Calcul Canada emploie un responsable de la sécurité depuis plusieurs années et tous les sites ont embauché du personnel de sécurité à temps plein à partir de 2020. Les sites d'hébergement ont également augmenté la surveillance du réseau et travaillent beaucoup plus étroitement qu'auparavant ensemble et avec les responsables de la sécurité de l'information (CISO) institutionnels.

À l'avenir, l'équilibre entre la performance, l'accessibilité et la sécurité du CIP sera un défi, car les systèmes de l'IRN doivent être renforcés contre les cyberattaques, au niveau technologique, opérationnel et politique. L'Alliance a pour mandat de renforcer la cybersécurité et devra donc faire appel aux ressources et à l'expertise du Centre canadien pour la cybersécurité (CCCS), en collaborant avec ce centre pour planifier et sécuriser les activités et les services de l'IRN dans le cadre de son enveloppe financière. L'Alliance formera également un groupe de travail sur la cybersécurité afin d'intégrer les considérations de sécurité de l'IRN dans sa planification, ses opérations et ses décisions de financement.

[229] CBC: Biden team sees Huawei as a threat and wants to talk to allies https://www.cbc.ca/news/world/biden-huawei-canada-1.5900991 (consulté en février 2021).

[230] BBC News: Europe's supercomputers hijacked by attacks for crypto mining https://www.bbc.com/news/technology-52709660 (consulté en mars 2021).

Prestation de services hétérogènes dans les systèmes nationaux de CIP

Les services et exigences en matière de CIP et d'IRN sont complexes. De plus, les besoins et technologies des différentes disciplines et groupes d'utilisateurs ne sont pas homogènes. Par exemple, des systèmes individuels ayant des objectifs fonctionnels similaires ne sont pas identiques (par exemple, les systèmes Cedar et Graham comportent chacun plusieurs types de CPU et de GPU à cause des cycles de financement et d'acquisition différents). Niagara est conçu pour répondre aux besoins de performance du calcul parallèle à grande échelle. Le nuage Arbutus est naturellement très différent, en ce qui concerne sa conception, des systèmes de calcul à usage général. Les systèmes pour les données sensibles devront être adaptés pour se conformer aux exigences locales, régionales et nationales, en plus de répondre aux besoins de discipline et de la fonction. Du côté commercial, AWS d'Amazon offre plusieurs instances (plus de 40) et services (plus de 300) aux clients à l'échelle internationale. Pour l'IRN, l'homogénéité réduirait l'innovation et de la position concurrentielle des chercheurs. Il est donc important de se doter de normes communes si possible, mais pas uniquement pour des raisons d'uniformité.

D'un autre côté, le manque d'homogénéité dans les services et la configuration peut potentiellement augmenter les coûts pour le personnel de soutien et l'exploitation. Ces coûts peuvent aussi s'accroître si les efforts se chevauchent au sein de la FCC, ce qui provoquerait également une confusion chez les utilisateurs finaux. Si plusieurs groupes offrent les mêmes services, ces derniers risquent de varier d'un utilisateur à l'autre, en plus de mettre en péril la sécurité, comme le niveau d'ancienneté et d'expertise n'est pas le même dans toutes les équipes.

Compte tenu de ce qui précède, les systèmes actuels de la FCC sont très différents en ce qui concerne la configuration, l'accès au soutien et à la documentation et la gamme de services. Voici les plus importantes différences entre les principaux systèmes de la FCC : possibilité d'accéder à l'internet par des nœuds de calcul (bloqué dans certains systèmes et autorisé dans d'autres) ; nœuds d'entrée/connexion qui ne sont pas configurés de manière uniforme (impossibilité de configurer sur certains systèmes l'exécution programmée de tâches routinières avec « crontab », différences de limites de mémoire système et de CPU, politiques d'accès différentes pour les transferts de données) ; politiques de programmation (différences dans les politiques sur la durée d'exécution maximale et le nombre de tâches) ; système de transfert de fichiers Globus (différences fondamentales dans l'authentification, par Calcul Canada ou par Globus dans certains cas). Les différences au niveau du système donnent potentiellement lieu à des problèmes et des inconvénients pour les utilisateurs finaux lorsqu'ils migrent leurs charges de travail entre les systèmes. Bien que la localisation des données soit une préoccupation majeure pour la portabilité des charges de travail, le déplacement d'ensembles de données volumineux ou complexes d'un système à l'autre peut s'avérer impossible.

Différents systèmes offrent également une gamme de services sensiblement différente, par exemple en ce qui concerne l'infrastructure de bureau virtuel (VDI) pour le travail à distance, NextCloud pour le stockage en nuage et la passerelle scientifique Jupyter Notebook pour la convivialité et la collaboration qui est accessible sur certains sites de la FCC. Même si le même service est offert sur plusieurs sites, il est souvent conçu et mis en œuvre par des équipes locales diverses, ce qui peut créer des différences dans la prestation de service et la configuration entre les sites. La documentation et le service de soutien sont également très différents dans l'un des sites par rapport aux autres, avec un wiki unilingue anglais comme principale source de documentation et un service de soutien autre que l'assistance technique nationale de la FCC.

Dans la mesure du possible, il faut donc privilégier une offre de services uniforme. Si ce n'est pas possible, un catalogue centralisé de tous les services de la FCC, par exemple un portail de services, permettrait aux chercheurs de trouver au même endroit ce qui est disponible. Ce portail pourrait inclure des listes et des descriptions de services, mais aussi des liens vers les principaux documents, ainsi que des renseignements sur les fournisseurs de services et les personnes à contacter pour plus d'information. Un catalogue central constituerait un point d'entrée unique de haut niveau pour les utilisateurs nouveaux et existants de la FCC.

Équité, diversité, inclusion, accès en milieu autochtones et représentation des minorités

L'équité, la diversité et l'inclusion (EDI) sont d'importants impératifs moraux, que reconnaissent les démocraties libérales modernes. Historiquement, l'EDI n'était pas considéré comme un facteur important dans la prestation de CIP. Par exemple, le rapport 2017 du CLIRN sur le CIP ne soulève pas de préoccupations à cet égard. De plus, la FCC ne recueille pas actuellement de données sur l'EDI, donc on n'en connait pas l'état actuel. En revanche, la FCC collecte des données sur les affiliations institutionnelles, ce qui reflète la diversité régionale, mais elles ne sont pas suffisantes pour étudier correctement l'EDI au sein de la fédération.

Les principales ressources de CIP au Canada et ailleurs dans le monde sont centralisées sur des sites d'hébergement spécialisés, auxquels on accède à distance (avec un accès terminal SSH ou des portails web Jupyter). En principe, l'accès à distance de ces systèmes devrait favoriser l'équité pour les régions plus éloignées. Par contre, les régions rurales et nordiques du Canada ont des problèmes d'accès à un internet fiable et à haute vitesse. En août 2020, CANARIE a annoncé une étape importante pour l'équité dans le réseau internet universitaire avec l'adhésion du Nunavut Arctic College au Réseau national de recherche et d'éducation (RNRE) du Canada,[231] ce qui donne aux treize provinces et territoires un accès à internet haute vitesse pour l'enseignement et la recherche. Par ailleurs, la FCC et les fournisseurs d'IRN et de réseaux doivent collaborer étroitement pour favoriser un accès efficace et équitable aux ressources de l'IRN.

L'écosystème de l'IRN doit également s'adapter aux personnes dont l'anglais n'est pas la langue maternelle, notamment les communautés francophones et les utilisateurs allophones (avec une documentation utilisateur clairement rédigée par exemple). De plus, la documentation et les services doivent être accessibles dans les deux langues officielles. La qualité de la traduction doit être équivalente à celle du texte d'origine et non à celle des logiciels de traduction automatique. La documentation et les principaux services clés doivent en outre être offerts dans certaines langues autochtones. Les événements et principales conférences devraient prévoir un service d'interprétation en langue des signes. En tant qu'organisation fédérale, l'Alliance devrait aussi fournir un service d'interprétation bidirectionnelle (voire multidirectionnelle) lors d'événements majeurs.

À l'avenir, l'Alliance devrait systématiquement recueillir des données sur l'EDI concernant les activités et la prestation de services afin d'évaluer l'état actuel de ces impératifs. Ceci permettrait de relever les problèmes et de trouver des solutions, au-delà des simples « belles paroles »

[231] CANARIE Press Release: Nunavut Joins Canada's National Research and Education Network to Enhance Nunavummiut Access to Colleagues, Data, and Scientific Facilities https://www.canarie.ca/nunavut-joins-canadas-nren/ (consulté en février 2021).

concernant l'EDI. Les chercheurs qui ont répondu au sondage de l'Alliance sur l'évaluation des besoins en février 2021 ont indiqué qu'ils étaient très impressionnés par les questions audacieuses et rigoureuses sur l'EDI, ce qui reflète l'engagement de l'Alliance à cet égard. Ces données fourniront des informations essentielles pour comprendre et résoudre les problèmes d'EDI dans l'écosystème de l'IRN canadien.

L'EDI ne doit pas être considéré comme un élément isolé, mais doit plutôt s'inscrire dans toutes les discussions et les prises de décision. À titre d'exemple, lors de création du Conseil des chercheurs, l'Alliance a accordé une attention particulière à l'EDI. Par conséquent, la composition actuelle du conseil reflète presque toute la société canadienne, à l'exception des communautés autochtones (il était difficile de trouver des universitaires autochtones qui avaient le temps nécessaire pour siéger au conseil). En revanche, le groupe de travail sur la GDR de l'Alliance compte actuellement deux membres d'origine autochtone, qui apportent une expertise précieuse en matière de droits et de propriété des données aux discussions et à la planification de la gestion des données de recherche.

Manque d'installations et de services pour les données sensibles

Le potentiel d'analyse des sources de données sensibles est devenu une tendance importante au cours des dernières années au Canada et à l'échelle internationale. Les discussions avec la communauté de recherche canadienne, le rapport sur l'évaluation des besoins et le sondage auprès des chercheurs indiquent un besoin urgent au sein de l'IRN en ce qui concerne les données sensibles au Canada. Ces dernières peuvent inclure des données personnelles sur la santé, les autochtones, les recensements statistiques, les finances, les impôts, les médias sociaux et les transactions commerciales, municipales, provinciales et fédérales. D'un côté, l'analyse et le recoupement potentiels de ces données peuvent apporter d'énormes avantages pour l'avancement des sciences sociales, de l'économie, des sciences humaines et du secteur de la santé humaine, ce qui profite à la société canadienne, aux politiques et à la prise de décision. D'un autre côté, ces ensembles de données impliquent souvent des données individuelles et commerciales très sensibles, qui comportent des exigences strictes en matière de confidentialité et de protection. Trouver un équilibre entre les avantages et les risques liés à l'accès et à la recherche de ces données pose continuellement plusieurs types de défis :

- Perception du public,
- Communications,
- Gestion des attentes,
- Lois et exigences municipales, provinciales, fédérales et internationales ; politiques et procédures mandatées connexes,
- Questions de propriété des données et d'accès ouvert

Solutions technologiques de soutien :

- Cryptage au repos et en cours de transfert,
- Cybersécurité,

- Sécurité au niveau des applications,
- Technologies pour la mise en œuvre et l'exécution de politiques

En plus de ces solutions, la facilité d'utilisation, ainsi que l'accessibilité pour les utilisateurs finaux et les chercheurs sont préservées. Tous les facteurs précédents doivent être considérés de manière holistique quand les plateformes de données de recherche sécurisées traitant des données sensibles sont conçues.

À l'heure actuelle, la FCC n'a pas de plateforme de CIP entièrement nationale, gérée et fournie de manière centralisée et accessible au public pour les données sensibles au Canada. Le Réseau canadien des centres de données de recherche (RCCDR) est probablement l'entité la plus nationale. En collaboration avec Statistique Canada (SC), Services partagés Canada, CANARIE et les universités locales, le RCCDR donne accès aux chercheurs en sciences sociales à des données appartenant principalement à SC, par l'intermédiaire des bureaux sécurisés des Centres de données de recherche (CDR), qui se trouvent dans plus de 30 établissements universitaires du pays. L'accès à ces ressources est offert à tous les chercheurs canadiens admissibles dans les universités participantes. Ce système n'est pas encore entièrement centralisé, mais le RCCDR développe une grappe centralisée et accessible à distance (« CDRv » ou CDR virtuel) pour remplacer l'infrastructure actuelle des postes de travail et des serveurs de chaque CDR. Ce système respectera des exigences très strictes en matière de protection des données protégées B du gouvernement du Canada, ainsi que les exigences additionnelles de Statistique Canada, puis éventuellement les exigences locales et régionales (notamment pour recouper les données appartenant à SC avec les données transactionnelles sur les soins de santé provinciaux, dont le Régime d'assurance-maladie de l'Ontario ou le RAMO). Aux niveaux régional et local, il existe de nombreuses installations pour les données sensibles, par exemple l'initiative HPC4Health, dont il est question plus haut dans le rapport et qui a connu un grand succès.

Le défi pour l'Alliance et l'écosystème de l'IRN canadien consiste à fournir des solutions pour le traitement des données sensibles au niveau national et à grande échelle, tout en respectant les exigences locales, régionales et fédérales. Étant donné que les exigences juridiques et autres pour divers ensembles de données (par exemple les données sur la santé par rapport aux données provenant des médias sociaux) sont souvent très différentes, la conception de systèmes d'IRN en mesure de traiter tous les types d'utilisation au niveau national constitue un défi. Pour y parvenir, l'exploitation, la coordination et la coopération nationales sont très importantes. D'un point de vue juridique, la « chaîne de commandement » doit être clairement définie, ce qui est complexe dans un modèle fédéré, quand il faut tenir compte de toutes les exigences locales, régionales et fédérales. L'Alliance prévoit de créer des groupes de travail sur les données sensibles afin de trouver des solutions pour les données sensibles et sécurisées.

Financement de la construction, de la maintenance et de l'exploitation des centres de données

La construction d'un centre de données et de son infrastructure pour les systèmes de CIP exige beaucoup de capital, notamment pour l'acquisition de biens immobiliers et d'installations. Il faut également de bons systèmes d'électricité, de chauffage, de ventilation, de climatisation et de réseau. En général, les installations de CIP comportent aussi des exigences plus rigoureuses pour l'alimentation, le refroidissement, l'espace au sol et même la charge au sol, comparativement aux centres de données d'entreprise. Ces investissements doivent être prévus pour des besoins à long terme, ce qui implique une planification et une estimation importantes

des tendances et des besoins futurs. Il faut de préférence des engagements à long terme de la part de toutes les parties prenantes pour justifier les coûts et l'amortissement correspondant. Pour l'instant, la FCI et les autres organismes de financement fédéraux ne couvrent pas les coûts d'infrastructure des centres de données. La construction de ces centres de données est plutôt souvent couverte par les sites d'hébergement locaux dans le cadre d'un financement provincial, ce qui entraîne des inefficacités potentielles au niveau national et des inégalités entre les établissements d'enseignement supérieur canadiens, comme certains établissements assument des coûts substantiels, alors que de nombreux établissements ne partagent pas les coûts d'infrastructure.

Les coûts d'exploitation du CIP ne s'inscrivent pas dans l'enveloppe du Fonds d'innovation de la FCI. Ils sont plutôt financés par le Fonds des initiatives scientifiques majeures de la FCI. Par conséquent, l'acquisition de nouveaux systèmes et d'infrastructure pour le CIP ne s'inscrit pas dans le financement des coûts opérationnels et dans les préoccupations correspondantes, même si ces deux éléments sont étroitement liés. Les frais pour l'exploitation de ces centres de données, au-delà de l'électricité et du personnel, devraient être inclus dans toutes les prises de décision au niveau national. Actuellement, les coûts d'exploitation de la FCI ne reflètent pas la réalité, puisque l'espace du centre de données, la sécurité physique (alarmes, surveillance, etc.) et la maintenance des infrastructures (refroidisseurs, pompes, etc.) ne sont généralement pas des coûts éligibles.

En d'autres termes, l'infrastructure moderne du CIP implique d'importants coûts en capital, coûts opérationnels et besoins en personnel, qui devraient tous être considérés et garantis au moment de l'investissement dans l'infrastructure. Il faudrait aussi résoudre le problème du manque d'harmonisation entre les échéanciers. L'échelonnement des délais pour la mise en service et hors service pour les systèmes de CIP est avantageux pour la continuité des services et des mises à jour techniques, mais pose problème quand il s'agit de concilier le capital et les budgets opérationnels. Puisque le gouvernement canadien et ISDE reconnaissent ce problème, l'Alliance a été créée pour « remplacer » le Fonds d'innovation de la FCI en ce qui concerne les coûts d'exploitation du CIP. À l'avenir, les principales décisions sur les investissements et l'exploitation de l'IRN seront prises de manière holistique. La planification du financement opérationnel doit également tenir compte des coûts liés aux systèmes contribués.

Absence de planification à long terme en raison des ressources limitées

Comme indiqué plus haut dans ce rapport, le PHQ de la FCC exploite l'infrastructure du CIP et les services connexes avec un ratio utilisateurs/personnel relativement élevé par rapport au TACC, par exemple. En revanche, il manque de temps et de ressources pour se concentrer sur les tendances et les progrès à long terme. Il doit donc se limiter aux besoins opérationnels à court terme. Avec les ressources nécessaires pour suivre les tendances futures, tester et développer de nouvelles technologies, on améliorerait tout l'écosystème de l'IRN, grâce à une meilleure prestation de services, une prise de décision plus éclairée à tous les niveaux et une meilleure rétention du personnel hautement qualifié. Il faudrait donc potentiellement augmenter le financement opérationnel et embaucher plus de PHQ. À l'avenir, l'Alliance devrait tenter de réduire la charge de travail quotidienne du personnel et investir plus de ressources dans le développement de nouveaux services et produits. Selon certaines données anecdotiques, les ressources de CPU et GPU ne sont pas exploitées efficacement pour de nombreuses tâches. En revanche, le personnel de la FCC n'a pas le temps de se mobiliser et de travailler avec les

utilisateurs pour optimiser ces inefficacités. La réduction de la charge de travail quotidienne du personnel de la FCC optimiserait donc les ressources du système et offrirait plus d'efficacité.

5 Liens entre le CIP, la GD et les LR

Les sites de CIP ont connu certaines difficultés pour la gestion des données et les logiciels de recherche. Calcul Canada a des systèmes de stockage à plusieurs niveaux depuis au moins 10 à 15 ans à cause de ces problèmes, en plus d'avoir déployé des LR et des intergiciels (notamment pour la physique des hautes énergies ou PHE). L'histoire du CIP et des superordinateurs est marquée par les avancées, l'amélioration continue, les progrès technologiques et l'évolution des besoins de chercheurs. Par ailleurs, les disciplines qui utilisent depuis longtemps le CIP, comme le climat, l'astronomie et la physique des hautes énergies, développent des LR et améliorent la GDR depuis des décennies (même si elles les nommaient autrement). Les logiciels de recherche et la gestion des données de recherche sont depuis environ dix ans plus sophistiqués, organisés, bien établis et importants grâce aux progrès de la technologie et des logiciels, ainsi qu'à l'évolution des demandes dans les disciplines de recherche et l'explosion des données massives. Comme de nouvelles disciplines ont d'importants besoins en matière de données, les solutions de CIP, LR et GDR et RDM plus polyvalentes et faciles à utiliser sont très recherchées et mises au point.

Traditionnellement, le lien entre le CIP et le développement de logiciels reposait sur la migration des codes de logiciels de recherche existants vers de nouvelles classes de superordinateurs, de plus en plus parallèles, voire dotés de nouveaux types d'accélérateurs. Dans le cadre de ces efforts, les États-Unis ont financé le développement de débogueurs commerciaux massivement parallèles afin d'améliorer la convivialité et de faciliter le portage des LR vers de nouvelles classes de superordinateurs. Depuis quelque temps, les communautés du CIP se concentrent davantage sur la maturation des écosystèmes et du soutien pour les LR et la GDR, car ils sont reconnus étant essentiels pour la viabilité de l'IRN.

Les écosystèmes bien établis de LR incluent des pratiques de codage professionnelles, des contrôles de version par des dépôts de code source, une variété d'offres de compilateurs, de pilotes et de bibliothèques, une assurance de la qualité dans l'ensemble du système, la possibilité de réutiliser, des environnements conteneurisés pour la duplication, la documentation et la formation, des passerelles scientifiques accessibles, etc. Les écosystèmes bien établis de GDR respectent les principes FAIR (faciles à trouver, accessibles, interopérables et réutilisables) pour les données de recherche, ainsi que les principes TRUST (transparence, responsabilité, orientation vers l'utilisateur, durabilité et technologie) pour les dépôts de données. L'offre et l'activation de LR et de GDR nécessitent des systèmes de CIP qui sont conçus à cette fin, notamment des intergiciels, des portails web et d'hébergement, des réseaux haute vitesse et des solutions de stockage sur mesure à moyen et à long terme pour les données en ligne et les données d'archives, en plus du stockage de sauvegarde correspondant. Un bon nombre de ces composants existent déjà sous une forme ou une autre. Le CIP a toujours été axé sur l'amélioration continue, vu les progrès de la technologie, ainsi que l'évolution des besoins d'utilisateurs et des cadres d'un financement concernés. Par conséquent, les besoins en matière de LR et de GDR ne sont pas nouveaux en soi, mais ils relèvent historiquement du financement.

À l'avenir, l'Alliance devra aller au-delà de l'approche traditionnelle pour intégrer, conformément à son mandat, le CIP, les LR la GDR dans sa planification, ses activités et ses décisions de

financement. Une telle approche offrira aux chercheurs canadiens un écosystème d'IRN mieux ancré et plus cohérent, ainsi que des gains d'efficacité à long terme grâce à la science ouverte. Ceci donnera des avantages directs à tous les Canadiens grâce aux nouvelles recherches, à l'innovation et aux connaissances qui contribueront à l'économie, la compétitivité et l'élaboration de politiques au Canada.

Annexe A : Commentaire de la communauté (été 2021)

Le rapport sur l'état actuel du CIP a été présenté aux groupes d'intervenants et à la communauté au cours de l'été 2021 afin de recueillir leurs réactions et leurs commentaires généraux. Ceci inclut le PHQ et les PDG régionaux de la FCC, les groupes de travail sur les logiciels de recherche et la gestion des données de recherche, ainsi que le Conseil des chercheurs de l'Alliance. La plupart des commentaires étaient généraux et n'ont pas révélé de problèmes factuels majeurs, donc seuls quelques changements de formulation ont été apportés au rapport.

Les commentaires de la communauté du CIP, des groupes de travail sur les logiciels de recherche et la gestion des données de recherche et du Conseil des chercheurs de l'Alliance portent sur les questions suivantes :

- Il serait important d'avoir des données supplémentaires afin de différencier les charges de travail infonuagiques, haut débit et HPC (tâches massivement parallèles à grande échelle avec d'importants besoins de communication interprocessus) pour la prise de décision future sur l'allocation de ressources. Par exemple, il serait intéressant d'établir une corrélation entre les disciplines de recherche et l'ampleur actuelle des soumissions de tâches.

- Ce serait utile de répartir les postes du PHQ ETP pour avoir plus d'information pertinente, notamment pour savoir quelle proportion du PHQ assure d'administration de systèmes, la gestion de projet, le soutien à la recherche informatique, le soutien à l'infonuagique, etc.

- Les comparaisons internationales avec des pays semblables, par exemple l'Australie, constitueraient de précieuses informations.

- Il serait intéressant d'effectuer une comparaison internationale pour voir comment le taux d'adoption de 17 % des systèmes de CIP de la FCC par les universitaires canadiens, dont 10 % sont en SSH, se compare au niveau mondial.

- Les taux de satisfaction de l'offre et de la demande doivent être compris dans un contexte plus large. À première vue, le taux de satisfaction de 40 % pour les ressources de CPU ne semble pas disproportionnel au taux d'approbation général des demandes de subvention. Les données du rapport sur le CIP ne mesurent pas explicitement les besoins en ressources, qui n'ont jamais été formulés à la FCC par l'entremise du CAR, notamment parce que les chercheurs estimaient qu'il était inutile ou difficile d'obtenir des ressources suffisantes, ou parce que les chercheurs utilisaient déjà des ressources de CIP autres que celles de la FCC et provenant de l'international. En d'autres termes, la demande réelle est beaucoup plus élevée que celle rapportée par les seules demandes du CAR.

- Il faut être prudent quant à la portée et à la validité des projections lorsqu'on utilise des données passées pour prédire l'avenir.
- L'auteur, l'objectif, le public visé et l'organisation de la publication du document doivent être clairement indiqués.
- Les liens entre les données et les déclarations pourraient être plus clairs.
- La valeur et l'importance du PHQ de la FCC devraient être davantage mises en évidence.
- À l'avenir, l'Alliance devrait recueillir des informations sur les billets de soutien concernant les grappes problématiques pour les utilisateurs.
- La diversité des disciplines de recherche devrait être abordée séparément des questions d'égalité des genres et d'équité entre les régions.