Calcul intensif sur GPU : retour d'expérience de l'UMR d'(Epi)génomique Fonctionnelle et Physiologie Moléculaire Du Diabète et Maladies Associées du CNRS (1283/8199) à Lille

Anne-Sophie Ledoux
UMR1283-8199 EGID – PreciDIAB / Équipe Informatique
1 place de Verdun
59 000 Lille

Mehdi Derhourhi
UMR1283-8199 EGID – PreciDIAB / Équipe Bio-informatique
1 place de Verdun
59 000 Lille

Résumé

Les progrès de la recherche en génétique de ces dernières années ont été rendus possibles par l'arrivée du séquençage dit « haut débit », qui permet de séquencer le génome complet de plusieurs dizaines de patients en quelques jours seulement. Cette technologie ouvre la porte de la médecine de précision, où chacun pourra bénéficier d'un traitement sur mesure, mais au prix de volumes de données de plus en plus conséquents.
Les traitements bio-informatiques qui visent à analyser ces données sont devenus un facteur limitant, et un enjeu crucial pour poursuivre la démocratisation du séquençage.
Pour répondre à ce challenge, la société Nvidia a développé une solution d'analyse reposant sur l'utilisation de GPU, associés à des versions adaptées des principaux logiciels d'analyses bio-informatique de données de séquençage (GATK, BWA, …).
Dans ce retour d'expérience, nous proposons une comparaison entre les solutions d'analyses habituelles utilisant des CPU, et l'approche Nvidia utilisant des GPU, et nous aborderons également une solution concurrente basée sur une carte FPGA (Dragen / Illumina).

Mots-clefs

CPU, GPU, GATK, BWA, Parabricks, Nvidia, DELL, FPGA, DRAGEN, calculs parallélisés, haut débit, génétique

1 Contexte scientifique

Les thématiques de recherche des laboratoires de génétique utilisant le séquençage à haut débit sont caractérisées par la production d'un très grande quantité de données en un temps limité.

Lors de chaque *run*, jusqu'à 1,2 To de données brutes sont générées par l'appareil Illumina Novaseq, ce qui représente 30 génomes humains, et ce en 40 heures seulement. L'analyse de ces données génère jusqu'à 10 To supplémentaires, et doit être réalisée dans des délais

compatibles avec les impératifs cliniques et le débit de production de données des séquenceurs.

Le projet PreciDiab mené par notre laboratoire, qui vise à accélérer le développement de la médecine de précision dans le cadre du diabète, de l'obésité et des complications qui y sont associées, nécessite l'acquisition d'équipements de séquençage haut débit et d'équipements informatiques.

Ce projet qui va générer une volumétrie importante de données génétiques de l'ordre de 1 Po (10 000 échantillons) concerne la mise en place d'une plateforme de séquençage haut débit pouvant générer jusqu'à 400 Go de données par jour et étudier plus de 1500 génomes humains entiers par an.

2 Contexte technologique

Actuellement, la majorité des outils utilisés pour analyser les données de séquençage à haut débit sont des outils libres, fonctionnant sous environnement Linux et basés sur l'utilisation de CPU.

Ces outils sont de plus en plus difficilement adaptables à la quantité de données croissante à analyser et sont souvent difficiles à paramétrer efficacement si l'on cherche à améliorer leur vitesse d'exécution, par exemple via l'usage du calcul parallélisé : plusieurs tâches simultanément et/ou une tâche sur plusieurs CPU.

Face à cet état de fait, la société NVIDIA a récemment développé des versions « GPU accelerated » d'outils phares de la bio-informatique, tels que GATK (*Genome Analysis Tool Kit*) ou BWA (*Burrow-Wheeler Aligner*). Ces développements visent à répondre à la problématique des temps de calcul nécessaires à l'analyse bio-informatique de données de génétique, qui constitue l'un des grands challenges actuels en particulier face au développement de la médecine de précision.

3 Infrastructure informatique

Avec l'acquisition de plateformes de séquençage et de génotypage à très haut débit, la plus grande partie de l'activité de l'équipe informatique consiste à améliorer et à augmenter la capacité de stockage et de calcul du système informatique du laboratoire et à optimiser la bande passante du réseau.

Nous avons multiplié les contacts avec les fournisseurs aussi bien sur les technologies de séquençages que les matériels liés au calcul et au stockage.

3.1 Infrastructure existante

La refonte complète de l'infrastructure du laboratoire s'inscrit dans le cadre du projet de création du centre national de médecine de précision sur le diabète PreciDIAB.

En effet ce projet de recherche génère une volumétrie importante de données génétiques : 1Po au total, 400Go par jour.

Après état des lieux, les équipements de stockage existant ne permettaient pas d'accueillir cette production de très grandes quantités de données et les performances de calculs étaient insuffisantes pour l'étude de génomes humains.

Le principal objectif était de faire évoluer les systèmes informatiques du laboratoire pour permettre à la fois l'augmentation du volume de données générées et garantir un maximum de performances pour la bio-informatique.

Nous avons renouvelé et avons obtenu les budgets conséquents pour cela, le stockage capacitif NetApp, les puissances de calcul, mais aussi renforcé la sécurisation des données et du réseau.

En effet le projet complet d'infrastructure comprend le renouvellement du stockage de NetApp à Isilon, l'upgrade du réseau de 10 à 25Gb/s, les puissances de calculs, la sauvegarde sur disque en remplacement de la sauvegarde sur bandes lto, la réplication sur distant distant avec la mise en place d'un site de PRA, et enfin l'intégration d'un pare-feu de site PaloAlto.

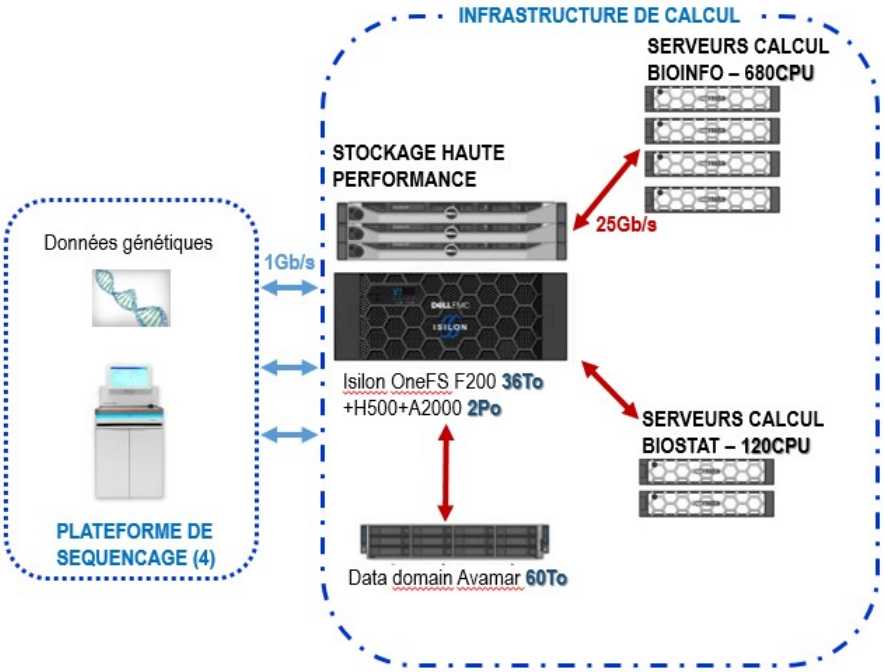

Figure 1 – Nouvelle infrastructure de calcul

3.2 Infrastructure projetée

Dans l'objectif d'améliorer la puissance de calcul du laboratoire, nous avons orienté nos recherches aux dernières avancées informatiques utilisées dans le domaine de la génétique.

Nous avons étudié les *white-papers* publiés par DELL portant sur les innovations technologiques de cartes GPU dans plusieurs domaines, dont celui de la génétique.

Nous avons découvert que l'utilisation des cartes dédiées à l'analyse génétique commençait à se démocratiser et qu'elles diminuaient considérablement les temps de calcul. Les *white-papers* étudiés sont ci-dessous:

Technologie GPU Whitepaper DELL/NVIDIA mars 2020

https://www.delltechnologies.com/en-gb/collaterals/unauth/white-papers/solutions/parabricks-isilon-nvidia-wp.pdf

Technologie FPGA Whitepaper DELL/DRAGEN (Illumina) avril 2020 :

https://www.delltechnologies.com/en-gb/collaterals/unauth/white-papers/partner/dell-emc-isilon-and-illumina-dragen-bioit-platform-wp.pdf

Dans un premier temps, nous avons testé laechnologie GPU DELL avec NVIDIA.

Grâce à un partenariat que nous avons mis en place avec DELL et NVIDIA, nous avons obtenu en prêt et configuré un serveur Dell PowerEdge R740 Bi Pro Intel Xeon Gold 6248R 3GHZ (2 CPU - 48 cores - 96 threads) et 4 cartes Nvidia GPU Tesla T4. Les cartes nouvelles générations de Nvidia basées sur leur architecture Ampère comme la carte NVIDIA A40 n'étaient pas encore disponibles sur le marché.

Nous avons réalisé un POC (*Proof Of Concept*) basé essentiellement sur le calcul. Ces tests ont été réalisés sur des échantillons préparés au laboratoire.

Puis nous avons testé la technologie FPGA DRAGEN d'Illumina.

Illumina étant le fournisseur principal des équipements de séquençage NGS (*Next Generation Sequencing*) du laboratoire, afin de n'écarter aucune autre alternative, nous avons demandé et obtenu en prêt un serveur Illumina muni d'une carte DRAGEN capable d'analyser les génomes entiers.

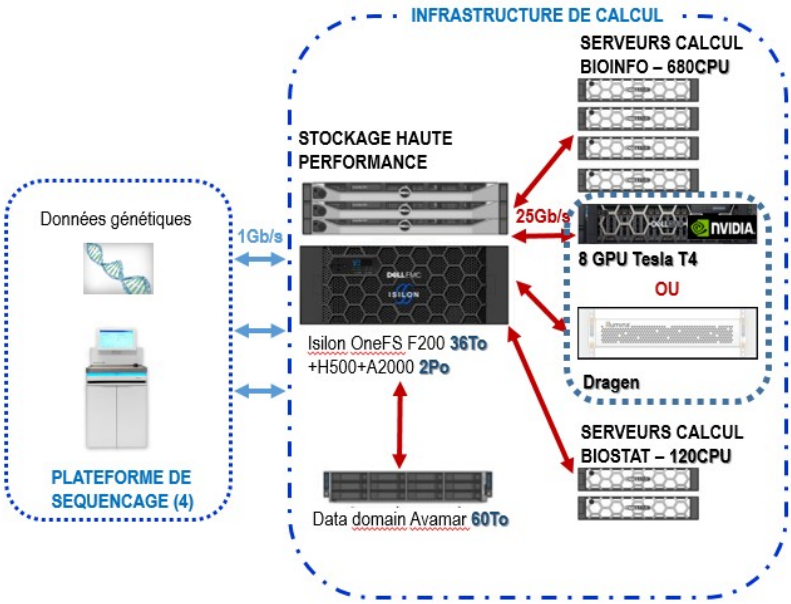

Figure 2 – Infrastructure de calcul projetée

Tests de comparaison

Les tests réalisés ont comparé 3 solutions différentes :

- une solution purement CPU basée sur un serveur de calcul Dell R740XD. Étant donné les limitations des logiciels utilisés, une partie des étapes a pu être réalisée en utilisant jusqu'à 15 CPU simultanément (BWA), et le reste sur un seul CPU (GATK) ;

- une solution utilisant 4 GPU Nvidia T4 associés à la suite logiciel Nvidia Parabricks, qui reprend les mêmes logiciels (BWA/GATK) et étapes que pour la solution CPU, mais adaptés pour tirer parti de l'architecture GPU ;

- une solution utilisant une carte FPGA Dragen de la société Illumina, associée à leur solution logiciel (Dragen) qui reprend les mêmes étapes que les solutions CPU ou GPU.

Pour comparer les différentes solutions testées, nous avons procédé à l'analyse de deux échantillons humains représentant deux grandes méthodes de séquençage : un échantillon séquencé par la méthode génome complet (WGS) et un échantillon séquencé par la méthode exome (WES). L'exome représente uniquement les parties du génome codant des protéines, soit environ 1 à 2% du génome total.

La profondeur de séquençage, qui représente le nombre de fois où chaque zone à séquencer (l'ensemble du génome pour le WGS ou uniquement les parties codantes pour le WES) est effectivement séquencée, est de 30X pour l'échantillon WGS et 50X pour l'échantillon WES, soit des valeurs standards pour ces deux méthodes. Le séquençage répété de chaque zone à séquencer permet de compenser les erreurs de séquençage (en moyenne 1/10000 bases séquencés avec les séquenceurs de la société Ilumina), et de détecter des variations qui ne sont pas présentent dans l'ensemble des fragments d'ADN séquencés.

L'analyse a consisté à démarrer des données non alignées des échantillons (fichiers Fastq), qui représentent l'ensemble des données séquencées sous forme brute. Ces fichiers Fastq contiennent plusieurs dizaines ou centaines de millions de fragments génétiques d'environ 100 bases (ATCG) de long chacun.

La première étape à réaliser est l'alignement de ces millions de fragments génétiques sur un génome de référence pour générer un fichier d'alignement, ou *Sequence Alignment Map file* (fichier .sam). Cette étape nécessite de comparer chaque fragment séquencé avec l'ensemble du génome de référence, pour déterminer les coordonnées du meilleur emplacement possible de chaque fragment. Elle fait appel au logiciel BWA pour les solutions CPU et GPU, et au logiciel Dragen pour la solution FPGA.

Une fois l'alignement réalisé, la seconde étape consiste à rechercher les variations génétiques, c'est-à-dire les différences entre les fragments alignés d'un échantillon et le génome de référence, pour générer la liste de ces variations sous la forme d'un fichier VCF (*Variant Calling File*). Cette étape fait appel au logiciel GATK pour les solutions CPU et GPU, et au

logiciel Dragen pour la solution FPGA. Le génome de référence utilisé est le génome humain dans sa dernière version (hg38).

Les deux étapes décrites précédemment représentent l'analyse primaire effectuée classiquement dans le cadre du séquençage d'ADN humain.

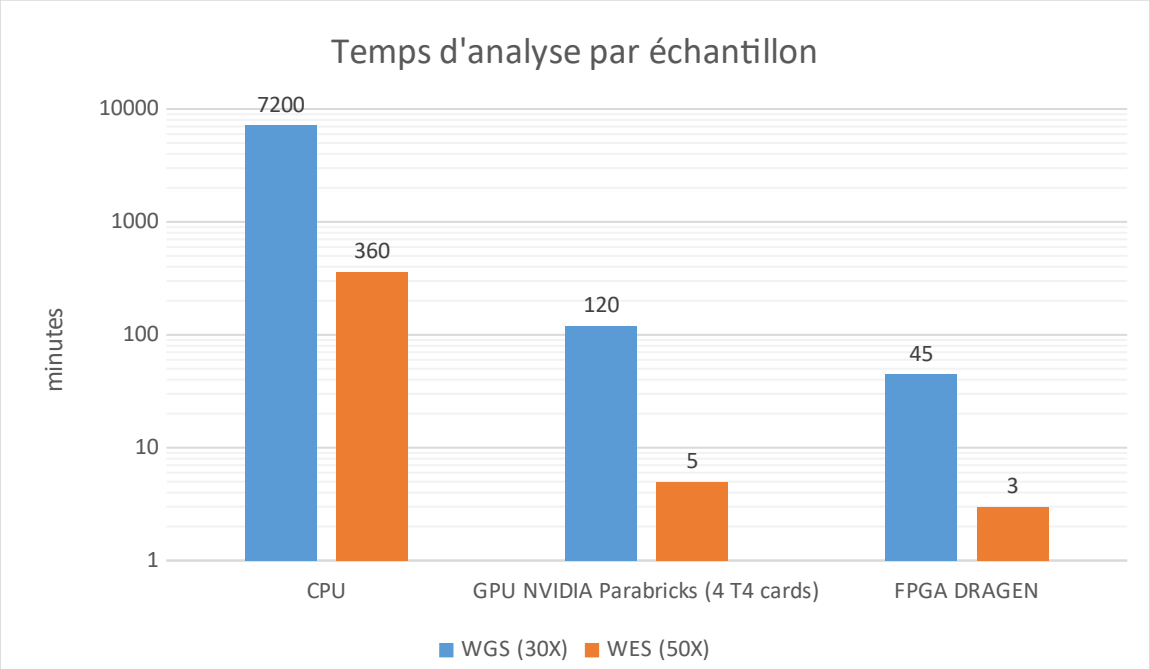

Figure 3 - Comparaison des temps d'analyse primaires de deux échantillons séquencés par WGS et WES, à l'aide de 3 méthodes utilisant : le calcul purement CPU, le calcul GPU a l'aide de la solution Parabricks de la société Nvidia, le calcul sur carte FPGA Dragen de la société Illumina.

Les résultats montrent que l'utilisation de la solution GPU Nvidia Parabricks permet un gain de temps d'un facteur d'environ 60 lors de l'utilisation simultanée de 4 cartes T4, en comparaison de la solution purement CPU. Ce gain est globalement proportionnel au nombre de cartes utilisées simultanément (une carte permet un gain d'un facteur 20, deux cartes d'un facteur 35).

La solution FPGA Illumina Dragen permet un gain de temps d'un facteur 120 à 150 selon l'échantillon, en comparaison de la solution purement CPU.

Les résultats obtenus pour le WGS ont été comparés au *set* de variation de référence de l'échantillon (NA12878), de façon à obtenir une mesure de la sensibilité de détection pour les deux types de variations génétiques recherchés par ces méthodes : les variations ponctuelles et les insertions/délétions de petites tailles (moins de 20 bases). Cette comparaison a été effectuée à l'aide du logiciel GATK Concordance.

La solution CPU a permis d'obtenir des sensibilités de détection de 99,4% pour les variations ponctuelles et de 97% pour les insertions/délétions de petites tailles.

La solution GPU Nvidia Parabricks a permis d'obtenir des sensibilités de détection de 99,3% pour les variations ponctuelles et de 97,5% pour les insertions/délétions de petites tailles.

La solution FPGA Illumina Dragen a permis d'obtenir des sensibilités de détection de 99,5% pour les variations ponctuelles et de 98,5% pour les insertions/délétions de petites tailles.

Les sensibilités obtenues permettent de considérer ces 3 solutions de façon équivalente d'un point de vue résultats obtenus.

4 Solution retenue

Ces résultats sont à considérer dans le contexte d'analyse classique d'un ***run*** de séquençage, qui comprend généralement entre plusieurs dizaines (WGS) et plusieurs centaines (WES) d'échantillons.

Dans le cas d'une analyse basée sur l'utilisation de CPU, la durée d'analyse plus élevée par échantillons en comparaison des solutions GPU ou FPGA est en partie compensée par la possibilité d'analyser plusieurs échantillons en simultané.

Néanmoins pour obtenir une durée d'analyse comparable à celles obtenues par les solutions GPU ou FPGA, plusieurs serveurs de calcul CPU seraient nécessaires.

Ces deux solutions non basées sur le CPU sont donc intéressantes dans le cadre d'une activité de séquençage soutenue, car les coûts supplémentaires liés aux licences et aux GPU ou à la carte FPGA sont compensés par la nécessité d'utiliser un plus grand nombre de serveurs CPU.

Ce constat est particulièrement vrai concernant la solution Nvidia Parabricks, qui grâce à des coûts d'acquisition matérielle et de licence très inférieurs à ceux de la solution Illumina Dragen, permet un ratio coût/échantillon plus intéressant.

De plus la possibilité d'ajuster la puissance disponible en faisant varier le nombre et le type de GPU permet de répondre plus efficacement aux différents besoins.

Remplacer à moyen terme ces cartes Tesla par des cartes de nouvelles générations de type Ampere, si le matériel sous-jacent est capable de les prendre en charge (par exemple T4 PCIe 3.0 x16 168mm ; A40 PCIe 4.0 x16 267 mm), aurait du sens du point de vue de la durée de traitement et financièrement. En effet nous pourrions envisager d'utiliser deux cartes A40 à la place de quatre cartes T4. À performances égales, le coût de la licence Nvidia Parabricks ramené à deux cartes au lieu de quatre serait donc divisé par deux. À noter qu'il n'est actuellement pas possible d'utiliser plusieurs modèles de cartes sur un même serveur. Malgré sa flexibilité moindre, la solution Illumina Dragen reste néanmoins intéressante dans le cadre d'une utilisation très haut débit nécessitant une capacité d'analyse supérieure à celle de notre laboratoire.

En conclusion, ces solutions sont une alternative efficace au calcul purement CPU dans le cadre de l'analyse d'un nombre élevé d'échantillons.

Cette technologie nous a permis d’augmenter notre puissance de calcul et ainsi d’analyser des génomes humains entiers à grande échelle.

À terme, il est prévu de multiplier par deux le nombre de serveurs Dell PowerEdge R7xx avec deux cartes NVIDIA A40.

L’évolution des performances des cartes graphiques, ainsi que la possibilité de monter en gamme sur le modèle sélectionné, permet d’envisager une augmentation des performances de calcul permettant de répondre à une grande variété de besoins.