2021 年 2 月 25 日
国立大学法人東京大学医科学研究所ヒトゲノム解析センター
株式会社日立製作所
エヌビディア合同会社

東大医科研ヒトゲノム解析センターが、がんゲノム医療推進や新型コロナウイルス対策など全ゲノム解析の高速化に向け解析基盤を強化

日立の技術支援のもと、ヒトゲノム解析用スーパーコンピュータ SHIROKANE に NVIDIA Clara Parabricks を全面導入

■本発表のポイント

- 国立大学法人東京大学医科学研究所(所長:山梨　裕司／以下、東大医科研)ヒトゲノム解析センター(センター長:井元　清哉)は、全ゲノムシークエンスデータ解析の大幅な高速化のため、株式会社日立製作所(執行役社長兼 CEO:　東原　敏昭／以下、日立)とエヌビディア合同会社(日本代表兼 米国本社副社長:大崎　真孝／以下、NVIDIA)の協力のもと、最新型のヒトゲノム解析用スーパーコンピュータシステム SHIROKANE (以下、SHIROKANE)に、従来の約 40 倍(注 1)の高速化を可能とするゲノムデータ解析ソフトウェア NVIDIA Clara™ Parabricks(以下、Parabricks)を全面導入します。これにより、2021 年 3 月 1 日の稼働開始後は SHIROKANE 環境下において、Parabricks による処理容量が約 6 倍(注 2)となり、さらなる全ゲノム解析の高速化が期待できます。
- SHIROKANE の学術機関・民間機関の利活用を大きく推進し、全ゲノムシークエンスに基づく、がんゲノム医療や新型コロナウイルス研究など、産官学民の英知を結集し推進するべき喫緊の課題への取り組みを強力に後押しします。複数のユーザーで同時に解析可能な基盤設計のため、個々のユーザーの利用環境に合わせたサービスの提供を実現します。

■概要

東大医科研ヒトゲノム解析センターは、このたび、日立の技術支援のもと、最新型の生命科学データサイエンス用スーパーコンピュータシステム SHIROKANE の解析基盤の強化に向け、Parabricks を全面的に導入しました。また、Parabricks は GPU(注 3)の並列演算性能を活用して実行されるため、GPU 環境の強化として、5 ペタフロップス(注 4)の AI 性能を備えた世界最先端の GPU サーバ、NVIDIA® DGX™ A100 システム(以下、DGX A100)を増設するとともに、全国の研究機関など全ての SHIROKANE ユーザーが利用できる環境を構築しました。あわせて、既存システムを含む SHIROKANE 全体の最適化も実施し、複数のユーザーが同時に全ゲノムデータ解析を行う場合のボトルネックを解消することで、システム性能を最大限に発揮できるように構成しました。本システムは、3 月 1 日から運用を開始し、4 月 1 日にユーザーへの提供を開始します。

東大医科研ヒトゲノム解析センターは、日立と共同で、SHIROKANE を用いて、がんゲノム(注 5)医療における全ゲノムデータ解析の高速化や解析時間の短縮化に取り組んできました。今回、Parabricks を SHIROKANE の GPU サーバにて評価した結果、今後の大規模全ゲノム解析時代に不可欠な全ゲノムシークエンスデータ解析の高速性と機能を持つことを認めたことから、全 GPU サーバへの導入に至りました。従来の CPU サーバ数百ノード分に相当する全ゲノムデータ解析能力を GPU サーバに実装し、SHIROKANE ユーザー向けに、ゲノム研究を加速する最新鋭の高速全ゲノムデータ

解析環境を実現します。今回の SHIROKANE の強化によって、日本の生命科学分野における研究開発の進化に寄与し、医科学の発展と社会へ貢献することをめざします。

■背景・課題

個別化医療とは一人ひとりの体質や病態にあった適切な医療を提供することであり、そのためには全ゲノム解析(注 6)により取得したパーソナルゲノム情報に基づいた予防・診断・治療法の検討が必要です。厚生労働省においても国家戦略として、2019 年 12 月にがんや難病領域の「全ゲノム解析等実行計画」を策定し、がんと難病の患者計約 92,000 人分の検体を対象に最大 3 年間かけて解析することを発表しました。一方、ゲノム研究において、全ゲノムシークエンスは情報の網羅性が高いことから研究面での有用性が広く認識され一般化してきたと言えます。近年では従来の 5 倍以上のシークエンス深度(注 7)で、がん全ゲノムを解析する研究も発表されています。また、がん研究以外の感染症などさまざまな研究領域においても全ゲノムシークエンスデータ解析のニーズが高まっています。

このような今までの数倍のシークエンス深度、かつ膨大なサンプル数が必要となる全ゲノムシークエンスデータを遅延なく迅速に網羅的に解析することは、従来の大型計算機を使っても膨大な時間を要するものでした。世界で全ゲノム情報を医療に活用する取り組みが加速するほか、日本においても、全ゲノム解析の実現性が議論されており、全ゲノム情報に基づくゲノム医療を多くの患者に提供するためには、そのデータ解析基盤の構築が喫緊の課題となっています。

■今回の取り組み

東大医科研ヒトゲノム解析センターは、日立の協力のもと、全ゲノム解析環境を強化すべく、2020 年 2 月に SHIROKANE に搭載されたデータセンター向けの GPU サーバ 80 基のうち 16 基に Parabricks を導入し(注 8)、2020 年 6 月から、研究機関やライフサイエンス関連企業など SHIROKANE ユーザーに開放しています。従来の想定を大きく上回る解析速度が評価されユーザー数が増加したことから、解析のジョブ待ちが多数発生するなど、基盤強化が求められていました。

東大医科研ヒトゲノム解析センターでは、今回、GPU サーバ(DGX A100)を新たに増設するとともに、さらに、全 88 基の GPU サーバに Parabricks を搭載し、一般的な CPU 環境で 1 サンプル当たり 20 時間以上を要する計算処理を 30 分以内で完結できる、解析基盤の強化を実現しました(注 1)。この全面導入にあたり、日立は、既存システムとの連携を考慮し、SHIROKANE の一部として最大性能が発揮できるよう構成の最適化を行いました。SHIROKANE ユーザーが利用できる Parabricks 導入ノードが増えることで、日本のさまざまなゲノム研究に対する支援を強化するとともに、ユーザーの利用環境に合わせたサービスのより一層の向上をめざします。

なお、東大医科研ヒトゲノム解析センターは、新型コロナウイルス感染症の研究を加速するため、必要とする研究機関に対して、2020 年 4 月から SHIROKANE の無償提供を行うほか、ヒトゲノム解析センターの研究者自身も、7 大学・研究機関の異分野の専門家からなる共同研究グループ「コロナ制圧タスクフォース」(注 9)をはじめさまざまな新型コロナウイルス感染症の研究に参画しています。今回の SHIROKANE の基盤強化は、新型コロナウイルス感染症の研究に係る研究者を強力に支援するものです。

■今後の取り組み

東大医科研ヒトゲノム解析センターは、SHIROKANE を最先端のゲノム研究の礎とし、超高速に全ゲ

ノムシークエンスデータの解析が可能な最新の全ゲノム解析環境と質の高いサービスを SHIROKANE ユーザーに提供することにより、日本のゲノム研究を大きく加速させ、ゲノム医療の実現を通して医学の発展と社会に貢献します。

日立は、「誰もが快適に、安心して、健やかに暮らせる社会」の実現をめざし、社会イノベーション事業を推進しています。新型コロナウイルス対策などの社会課題にも対応するため、ゲノム解析基盤やオープンソースソフトウエアの構築技術と最新技術を組み合わせ、お客さまとの協創により Society 5.0(注 10)時代のゲノム情報を活用した個別化医療の実現に寄与します。

NVIDIA は、世界中の医療機関が未来を切り拓くための支援をしています。個別化医療、ケアの質の向上、そしてゲノム解析をはじめとした医学生物学研究におけるブレークスルーなど、次世代のヘルスケアには新しいコンピューティングパラダイムが求められています。NVIDIA は人工知能(AI)およびハイパフォーマンス コンピューティング(HPC)のテクノロジの提供により、これらのニーズに応えます。

注 1: 一般的な CPU 環境で 1 サンプル当たり 20 時間以上を要する計算処理を 30 分以内で完結できるため、40 倍の高速化を実現。データは一般公開されている NA12878 (https://precision.fda.gov/challenges/truth) から深度 x30 に生成。CPU による所要時間は GATK4.1 を用い、32 vCPU (3.1Ghz Intel Xeon® Platinum 8175M) 128GB RAM 環境で計測。GPU による所要時間は Parabricks 3.2 を用い、DGX A100 環境で計測。

注 2: 2020 年 2 月の SHIROKANE 搭載 GPU 80 基のうち 16 基に Parabricks を導入、2021 年 3 月には新規導入 GPU の DGX A100 を含め GPU 88 基へ Parabricks の導入が完了するため、処理容量が約 6 倍へと増加。

注 3: GPU (Graphics Processing Unit):高度な画像処理を行うためのプロセッサ。1999 年に NVIDIA が世界ではじめて開発。高度な並列演算性能を備えており、AI (ディープラーニング) や科学技術計算などに活用される。

注 4: NVIDIA DGX A100 システムの性能: AI 処理を中心とした FP16 Tensor 演算では最大 5 ペタフロップス(毎秒 5,000 兆回の浮動小数点演算)、Parabricks の大部分やその他様々なアプリケーションで利用される FP32 演算では最大 156 テラフロップス(毎秒 156 兆回の浮動小数点演算)の性能を発揮。

注 5: ゲノムとは、遺伝子をはじめとした遺伝情報の全体を意味する。また、がんゲノム医療は、遺伝子情報に基づくがんの個別化治療の 1 つ

注 6: 全ゲノム解析とは:ヒトの全ゲノムは約 30 億塩基対で構成されているが、一般的な次世代シークエンサはその機構上、巨大なゲノムを 100~150 塩基対程度の断片に切断しなければ塩基配列情報を読むことができない。そのため、次世代シークエンサからは、数億個の断片に分割された塩基配列情報が出力されることになる。これを意味のある情報に変換するためには、膨大な断片を破綻のない形で本来の姿である 30 億塩基対の繋がりに復元する必要がある。さらにその後リファレンス配列と呼ばれる塩基配列に対し「30 億塩基対のどこに変異があるのか」を検出することで、はじめて有用な情報となる。

注 7: シークエンス深度とは、対象のゲノム領域に対して何回シークエンスを行ったかを意味する。次世代シークエンサでは配列読み取りエラーが発生するため、ゲノム上の同じ位置を繰り返しシークエンスを行うことで確度を高める。

注 8: 2020 年 3 月 10 日付 国立大学法人東京大学／株式会社日立製作所による共同プレスリリース (https://www.hitachi.co.jp/New/cnews/month/2020/03/0310.html)

注 9:「コロナ制圧タスクフォース」:慶應義塾大学、東京医科歯科大学、大阪大学、東京大学医科学研究所、国立研究開発法人国立国際医療研究センター、北里大学、東京工業大学、京都大学の感染症学、ウイルス学、分子遺伝学、ゲノム医学、計算科学を含む、異分野の専門家が共同で立ち上げた、研究グループ

注 10:日本政府が掲げる新たな社会像であり、その実現に向けた取り組みのこと。AI や IoT、ロボットなどの革新的な科学技術を用いて、社会の様々なデータを活用することで、経済の発展と社会課題の解決を両立し、人間中心の豊かな社会をめざす。狩猟社会、農耕社会、工業社会、情報社会に続く 5 番目の新たな社会として位置づけられている。

■本件に関するお問い合わせ先

国立大学法人東京大学医科学研究所ヒトゲノム解析センター
教授・センター長 井元 清哉 (いもと せいや)
〒108-8639 東京都港区白金台 4-6-1
TEL: 03-5449-5611
URL: https://www.at.hgc.jp/
E-Mail: imoto@ims.u-tokyo.ac.jp

株式会社日立製作所 公共システム営業統括本部
カスタマ・リレーションズセンタ［担当: 森下］
〒140-8512 東京都品川区南大井六丁目 23 番 1 号 日立大森ビル
URL: https://www.hitachi.co.jp/public-it-inq/

エヌビディア合同会社 広報部
Japan-PR@nvidia.com
URL：https://www.nvidia.com/ja-jp/

以 上