HITACHI
Inspire the Next

2011 年 9 月 27 日
株式会社日立製作所
国立遺伝学研究所

日立が国立遺伝学研究所の協力のもと ゲノムデータの大量処理への Hadoop 技術適用を検証

約 5 分の 1 のコストで従来システムと同等の処理性能を実現

株式会社日立製作所(執行役社長:中西 宏明/以下、日立)は、大学共同利用機関法人 情報・システム研究機構 国立遺伝学研究所(所長:小原 雄治/以下、遺伝研)五條堀孝教授グループの協力のもと、このたび、Hadoop(*1)技術を用いた大量ゲノムデータの分散処理環境を試作し、その有用性の検証を行いました。その結果、従来の分散処理システムと比較し、約5分の1のコストで同等のデータ処理能力が実現できました。

*1 Hadoop:オープンソースソフトウェアコミュニティ Apache Software Foundation にて開発・公開されているソフトウェアで、大規模データを効率的に分散処理・管理することに優れている。

現在、ゲノム研究分野では、次世代 DNA シーケンサー(*2)の発展が著しく、ヒトゲノムをはじめとした各種生物のゲノム情報の網羅的解析が加速しています。最新の次世代DNAシーケンサーでは、断片配列データと呼ばれる DNA の配列を解析単位に断片化したデータが、一回の計測で約 60 億個(約 1.8TB)も生成されます。この次世代 DNA シーケンサーの登場により、世界中の研究機関で生成される断片配列データ量は爆発的に増加しており、2010 年に生成された塩基配列(*3)の断片データ量は PB(ペタバイト)(*4)オーダーに及びます。一方で、データ量の増大に比例してデータ解析にかかる時間が増大しており、研究のスピード化を妨げる要因の一つとなっています。このため、データ解析にかけるコストを大きく増やすことなく、解析スピードを向上させる大量データ処理システムの開発が求められています。

*2 DNA シーケンサー:化学処理した DNA サンプルに対して様々な分析処理を行うことで、DNA の塩基配列を自動的に読み取るための装置。
*3 塩基配列:各染色体(1~22、X、Y)を構成する分子(塩基)の配列。
*4 PB(ペタバイト):約 1000 兆バイト。GB(ギガバイト)の 1,000,000 倍。

遺伝研は、現在、文部科学省「革新的細胞解析研究プログラム(セルイノベーション)」に参画し、次世代シーケンサーから産出される大量のゲノム関連情報を扱う解析拠点の整備を進めています。今回、日立は、ゲノムデータ解析の国内中核拠点である遺伝研の協力のもと、現在と同等のコストで飛躍的に解析スピードを向上させるための方策として Hadoop に着目し、検証を行いました。検証の内容としては、現在遺伝研が使用しているゲノム解析フローを、日立が構築した Hadoop 検証環境に移植し、さまざまな条件設定のもとゲノム解析を実行し、遺伝研でのゲノムデータ解析環境との処理性能の比較を行いました。この結果、従来システムと比較した場合、約 5 分の 1 のコストで同等のデータ処理性能が実現できました。これにより、今回の検証で構築したシステムを実用化した場合、従来システムと同等のコストで約5倍の処理性能が実現できることになり、研究機関はゲノム解析のスピードを大きく向上させることが可能となります。

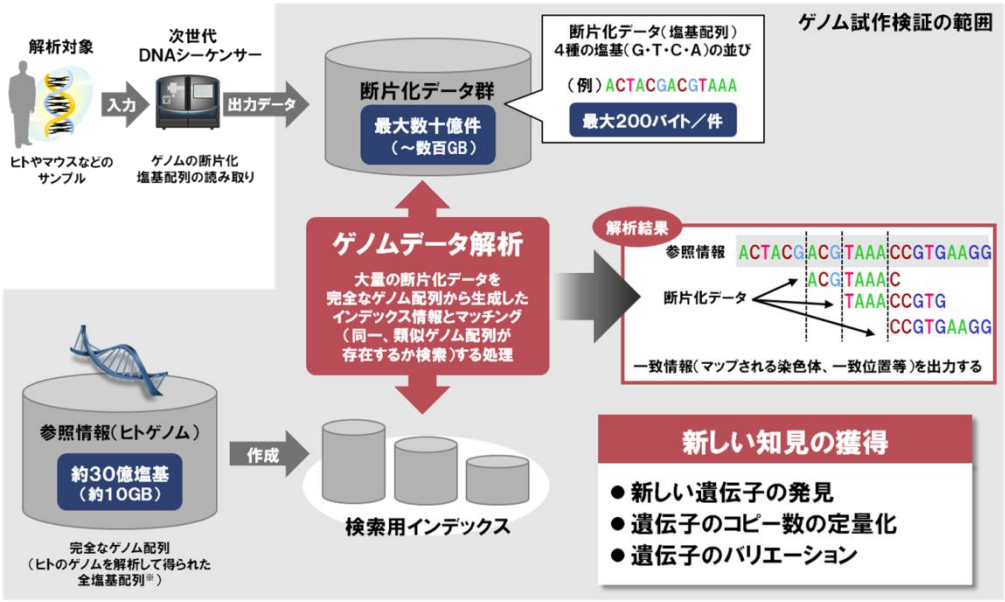
なお、今回の検証環境の構築にあたっては、Hadoop の特長であるスケールアウト(*5)に適した、コストパフォーマンスに優れ、省スペースかつ省電力なサーバブレードを多数集約可能な日立のエントリーブレードサーバ「HA8000-bd/BD10」を用いました。また、ゲノム解析プログラムは Sanger 研究所開発の「Burrows-Wheeler Aligner(BWA)」を用いています。

*5 スケールアウト:サーバの台数を増やすことで、システム全体の処理性能を向上させる手法。

今後、日立は、ゲノム解析プログラムの選定をはじめ、それらを Hadoop 上で効率的に動かすための最適なシステム構成を実現するなど、今回の検証で得た知見と、ゲノム研究分野へ最先端の解析環境整備を行ってきたこれまでのノウハウをもとに、Hadoop を適用した大規模データ分散処理ソリューションの事業化をすすめ、ゲノム研究に注力する大学、研究所、製薬業界、食品業界などに提供していきます。さらに、そのほかの分野においても、大量データ活用に関してコンサルテーションからシステム構築までシステムライフサイクル全般におけるサービスを、株式会社日立コンサルティングをはじめ、日立グループ全体で充実させていきます。

また遺伝研は、より高速で効率的に収集・解析可能な大規模データ解析基盤の構築に取り組んでいきます。

■ゲノムデータ解析のイメージ



※塩基配列：各染色体（1～22、X、Y）を構成する分子（塩基）の配列

■大学共同利用機関法人 情報・システム研究機構 国立遺伝学研究所について

国立遺伝学研究所（静岡県 三島市）は生命科学分野における遺伝学の中核拠点として、細胞機能・発生・分化・進化・生物多様性・ゲノム情報などについて、国際水準の先端的研究を行っています。また、知的基盤整備として、生命科学を先導するデータベース、バイオリソース事業を進めています。

URL: http://www.nig.ac.jp/

■日立 大量データ分散処理に関する取り組みについて

日立では、大量データを活用したお客さまの新たなビジネスの開拓に向けて、Hadoopやバッチジョブ分散処理、ストリームデータ処理などの並列処理技術の適用性を判断するアセスメントサービスを提供しているほか、システム構築に必要な製品・サービス群を提供し、大量データ分散処理システムの構築を総合的に支援しています。

URL: http://www.hitachi.co.jp/soft/big_data/

■本件についてのお問い合わせ先

株式会社日立製作所 情報・通信システム社 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：佐々木、米山］
〒136-8632 東京都江東区新砂一丁目 6 番 27 号 新砂プラザ
URL: http://www.hitachi.co.jp/Div/jkk/inquiry/inquiry.html

以上

補足資料

生命科学研究におけるバイオインフォマティクスの重要性（別紙 1）

■「シーケンシング」と「バイオインフォマティクス」

生命科学研究の分野では、細胞の中に存在する DNA の塩基配列を専用の機器により物理的に読み取ることを「シーケンシング」と呼んでいます。この「シーケンシング」を行う装置であるシーケンサーの性能が飛躍的に向上しており、2011 年現在では 2000 年時の 1 万倍以上の解析スピードを実現した次世代シーケンサーが実用化されています。現在も解析スピードは 1 年で 50 倍のペースで向上していると言われており、解析にかかるコストも大幅に低下しています。このシーケンサーの進化により、数年前までは考えられなかった膨大な量の塩基配列データが世界中で生み出されており、そのデータ量は年間 PB（ペタバイト）オーダーと言われています。

一方で、この読み取られた大量の塩基配列データを、IT を活用して整理、管理、解析する技術は「バイオインフォマティクス」と呼ばれています。例えば、「シーケンシング」によって次世代シーケンサーから短い DNA 断片（リードといいます）が得られた際、「バイオインフォマティクス」により、その DNA 断片が既知のゲノム配列(参照ゲノムといいます)のどの部分に該当するかを決定（マッピングといいます）したり、ゲノム配列が解析されておらず参照ゲノムが存在しない生物が対象の場合には複数の DNA 断片から元の長い塩基配列を再構築（アセンブリといいます）するなどといった様々な解析を行います。これにより生み出された塩基配列のデータはようやく意味を持ったデータとなります。いわば「シーケンシング」と「バイオインフォマティクス」は生命科学研究を進める上での両輪のようなものであり、「シーケンシング」で生み出されたデータを「バイオインフォマティクス」によって解析することで研究が進んでいます。

■「バイオインフォマティクス」がボトルネックに

「シーケンシング」と「バイオインフォマティクス」の関係性を踏まえて生命科学研究の現状を見てみると、「シーケンシング」の急速な発展と比較し「バイオインフォマティクス」の発達が十分ではないことが世界中の研究者の共通認識となっています。「バイオインフォマティクス」に関わる人材の不足、コストの高どまりなどにより、世界中で生み出される大量のデータを処理しきれずにボトルネックとなる事態がここ数年続いており、これは今後数年でより顕著な状況になると見られており、世界中の研究機関で解消に向けた議論が続けられています。

今回の検証結果は、この課題の解決に寄与するものであり、「バイオインフォマティクス」の高速化、低コスト化に向けた流れを加速するものになると考えています。

■生命科学研究における国際競争の現状

「シーケンシング」の分野においては、シーケンサーの開発力においてアメリカが突出しており、技術をほぼ独占している状況です。また近年では中国が低コストでのシーケンシングサービスの提供に力を入れており、その存在感が増しつつあります。一方で日本は「シーケンシング」に関わる技術開発で突出した成果をあげられていません。こういった中、「バイオインフォマティクス」の分野においては各国が横並びで試行錯誤している状況であり、IT 技術を得意とする日本が、「バイオインフォマティクス」の分野で世界をリードすることも可能であると考えられます。

■生命科学研究の応用分野について

「シーケンシング」の高速化・低コスト化により、膨大なゲノムデータが必要となる応用研究も現実の物となっています。たとえば、11 カ国が協力して 50 種のがん患者 2 万 5 千人のゲノムを解析し正常な細胞と比較しがんの原因を探る「国際がんゲノムコンソーシアム」などが進められています。このままシーケンシングの進化が進めば、個々人が自分のゲノム情報を丸ごと解析し、病気の治療などに活かすことも可能となると思われます。ただ、そうした研究を進めるためには、解析されたゲノムデータを整理・管理・分析できることが大前提であり、まさに「バイオインフォマティクス」の発展が待たれる状況となっています。

適用技術「Hadoop(ハドゥープ)」の概要(別紙2)

■Hadoop とは

「Hadoop」とは、プログラム開発者に対して、複数のコンピュータを並列につなぎ、大量のデータを分散処理するために必要な基本的な機能を提供する目的で作られたフリーソフトウェアで、Apach 財団がその管理を行っています。世界的に見ると大量データの処理を進めるためのソフトウェアの基盤として採用されるケースが出始めており、IT ベンダ各社が研究を進めています。

■Hadoop の特長

・ソフトウェア面

複数のコンピュータを並列につなぎ分散処理するためのソフトウェアを開発するためには、高度な技術とノウハウ、資金が必要となります。Hadoop はこの仕組みをフリーソフトとして提供するため、開発者は分散処理部分については Hadoop にまかせ、従来通り一台のマシン上で動くと同様の形でプログラムを開発し、Hadoop が実現する分散処理環境上で実行することができます。ただし、Hadoop 上で実行するプログラムを動かすためのチューニングは必要となり、最適なチューニングを施すためのノウハウが必要となります。

・ハードウェア面

比較的単純なアルゴリズムのプログラムを実行する場合、Hadoop を利用した分散処理は非常に効率が良く、従来起こりがちだった並列処理による性能劣化(10 の能力のマシンを 10 台つないでも 50 の能力しか出ないなど)がかなり抑えられます。このため、低コストで性能の低いマシンを複数台つなぎ並列処理を行わせることで、高性能なコンピューティング環境を実現することが可能となりました。

以上