２０１０年４月１日

ライフサイエンスにおける革新的 HPCI への期待

東京大学医科学研究所　宮野 悟

大規模データストレージと直結した 10 ペタフロップス級のコンピュータと全国の大学等の機関に導入される数十～数ペタフロップスのコンピュータを高速ネットワークで結んだ HPCI により、世界初の次世代クラウドコンピューティング環境を整備し、ベンチマーク世界一を超えてライフサイエンスにおけるユーザ満足度世界一を目指すこと。

【背景・世界の動向】

ヒトゲノム計画(1980 年代終わりごろ～2003 年）によりヒトゲノムが解読された直後、2003 年に米国 NIH はヒトゲノム解読後の"Roadmap for Biomedical Research"を発表した。そこには、"Biology is changing fast into a science of information management"というメッセージが出されており、医療・ヘルスケア開発のパラダイムシフトが起こることが示唆された。その後、個人個人の違いをゲノムのレベルで解明した国際ハプマップ計画(2002 年～2007 年）により、ヒトの病気や薬に対する反応性に関わる遺伝子を発見するための基盤整備が行われた。2007 年末には Science がこのハプマップ計画の成果を"Human Genetic Variations"として"BREAKTHROUGH OF THE YEAR 2007"に認定した。2008 年、中国・英国・米国は、世界各地の 1000 人以上の全ゲノム DNA 配列を解析し、医学的応用価値のある人類のゲノム地図を作成する「国際 1000 人ゲノム計画」が開始され、2010 年に終了することがアナウンスされている。収集されたデータと研究成果は全世界の研究者に提供されることになっている。

また、同年 2008 年からは、日本も含めた国際がんゲノムコンソーシアムが開始された。同じく 2008 年、主要ながんのゲノム異常（変異）カタログを作成するために、がん細胞のゲノムをシークエンスする「国際がんゲノムコンソーシアム」が開始され、日本を含む世界 10 カ国が参加している。日本は最大で 500 人の肝臓がんのゲノムを正常なゲノムと合わせて 1000 のゲノムをシークエンスする。このコンソーシアムでは数千のがん及び正常細胞のゲノムデータがでてくることになる。こうしたプロジェクトから 10000 のがんゲノムがでてくるとすると、データだけで１６ＰＢのストレージが必要となる。我が国にはこれだけのデータを無理なく収納し利用できるところは現在ない。また、我が国には現時点でこの規模のデータを収納し計算による解析をする仕組みはない。さらに、シーケンス技術だけではなく、ＤＮＡ

チップなどの遺伝子発現計測技術やタンパク質計測技術も日々更新されており、未曾有のデータが押し寄せてきている。この膨大なデータを解析するためには大容量ストレージだけでなく、それに直結された大容量メモリと高速ＣＰＵを備えたハイパフォーマンス・コンピューティング・インフラストラクチャの構築が必須である。

【ライフサイエンス・医療開発における大規模計算の特徴】

- 今後の技術革新により、質・量・速さ共にどのようにデータが出てくるか安易に予測することは出来ない。
- **多様なデータが解析対象**。配列データ（DNA、RNA、タンパク質)、量的データ（RNA、タンパク質、代謝産物など)、画像データ（静的、動的)、時間スケール、場所（染色体上、細胞内、生体内、地球環境内）など。
- データの**超高次元性**。遺伝子レベルでは数万次元、個人レベルの SNP を取り扱うと数十万次元、遺伝子の発現をエクソンレベルで解析すると数百万次元になる。また超高次元性という特徴があり、遺伝子レベルでは数万次元、個人レベルの SNP を取り扱うと数十万次元、遺伝子の発現をエクソンレベルで解析すると数百万次元。
- 新しいデータに即対応することが必須。２年かけてソフトウェアを開発することはライフサイエンスではあり得ない。したがって汎用的・オープンソース的なソフトウェアの利用環境が必須（Java や R、Perl、GCC など)。
- 計算資源の利用も短時間・小容量なものから長時間・大容量なものまで多種多様であり様々な解析手法・ソフトウェアを駆使し解析を行うことが必須。

以上のような特徴のあるデータの解析には、データの量と質、そして解析目的にフレキシブルに応じた計算資源が必要である。

【ライフサイエンスに必要な HPCI の必須仕様】

- 大規模ストレージ直結大規模メモリ実装ハイパフォーマンス・コンピュータが必要。
- シングルプロセスによる大量の（データ並列性のある）計算から超並列計算まで様々なレベルの様々なアプリケーションを利用可能な環境が必須。
  - シングルプロセスで数時間～数日程度のジョブを数千から数十万ジョブ実行。
  - 大規模ストレージデータ(数ペタバイト)への高頻度・高負荷アクセス。
  - １週間レベルの長時間計算ジョブを同時に数百ジョブ実行する計算。
  - 大規模メモリ（100TB～） AND/OR 高並列（数万並列）計算。短時間計算から長時間計算までの多様なジョブ。
  - データベース検索など「インタラクティブな計算資源の利用」も必須。
- ゲノムという究極の個人データを扱うためセキュリティには特段の配慮が必須。

- 大容量データを保存する必要性から物理的なデータの受け取り（ＨＤＤ持ち込みなど）対応も必須。
- 世界各地のゲノムデータ・遺伝子発現データなどの公開データのミラーリング。
- ライフサイエンスの分野で使われる一般的なソフトウェアの提供およびその維持管理。
- **これを実現するのはクラウドコンピューティング。**

【次世代クラウドスーパーコンピューティングプラットフォーム】

ライフサイエンスの現場では、データの質・量と解析目的に応じたフレキシビリティのある計算資源が必要である。また、計算の規模やスタイルは、データ並列型の大規模計算、超高並列計算、大規模メモリを使う計算、ストレージへの高頻度アクセス計算、リアルタイム計算と長時間計算など多様である。10 ペタフロップス次世代スパコンというハードウェアとネットワークをつないだだけではベンチマーク世界一という点を通過したに過ぎない。現場のユーザ満足度世界一を目指すためには、10 ペタフロップスの次世代スーパーコンピュータと直結した大規模ストレージセンターを核にして、日本各地の大型コンピュータやデータベースをシームレスに接続し、ブラウザ上でデータのアップロード、スパコン計算の実行、データの比較、データベース参照などを一元的に利用可能とするクラウドアプリの構築が不可欠である。

次世代クラウド スーパーコンピューティングプラットフォーム
ベンチマーク世界一を超えてユーザ満足度世界一へ