大量に作成されるエピゲノムデータと宇宙プラズマデータへの対応

九州大学情報基盤研究開発センター
深沢 圭一郎

要旨：

ビックデータとして近年注目されているデータの一つにエピゲノムがある。細胞がもつゲノム修飾の総体をエピゲノムと呼ぶが、次世代シークエンサーによりこの情報を効率的に取り出すことが可能になり、現在爆発的にそのデータ量が増えている。九州大学では今までシークエンサーのデータを箱崎地区の九州大学情報基盤研究開発センター（九大センター）内にある高性能演算サーバで解析・加工し、数倍にふくれあがったデータを箱崎から離れた病院地区にあるストレージに送り、保存・データ公開を行っていた。この際、データが増えることと、ストレージが設置されている研究室の内部ネットワークの帯域不足のために効率的な解析システムとなっていなかった。そこで、シークエンサー以外をすべて、九大センター内に設置することで、ボトルネックを解消し、効率的な運用が行えるようになった。さらに、爆発的に増え続けるエピゲノムデータのために、専用の解析サーバシステムとストレージを導入し、今後の容量不足、計算資源不足に対応を行っている。

一方、従来から大規模なストレージを利用する数値シミュレーションも近年の計算機の発達により、シミュレーションデータが巨大になっており、その保存、解析が難しくなっている。今回取り上げる宇宙プラズマは Vlasov 方程式でその振る舞いを計算することができるが、Vlasov 方程式は位置だけで無く、速度空間を扱う方程式であり、惑星磁気圏といった大規模構造を計算することは現在の計算機システムではできない。そこで、Vlasov 方程式の近似式である MHD 方程式を用いて計算を行っているが、それでもまだまだ解像度が粗い。そのため計算機性能が上がるにつれて、解像度を上げるため、計算サイズが大きくなり、出力データサイズも増えている。今回紹介する計算例では、その計算サイズは利用する計算機システムに依り異なるが、九大センターの計算機では 768GB～6TB の出力がある。これらのデータが時間発展分貯まっていくため、各システムで 100TB 以上のストレージ領域を使っている。計算機システムは共同利用計算機のため、あまりにストレージを占有することはできず、現在外部の大規模ストレージも利用している。HPCI 共有ストレージは 20PB の容量が有り、HPCI システムにマウントして、利用可能で有り利用が容易である。また NICT サイエンスクラウドのサービスの 1 つであるクラウドストレージでは全国各地にノードを置いてある分散共有ストレージであり、4PB 程度の容量を持つ。こちらは接続地点が限られているが、地域分散による災害時に強いシステムで有り、比較的安心して利用ができる。このように複数のシステムを利用することで、宇宙プラズマデータの保存に対応をしている。