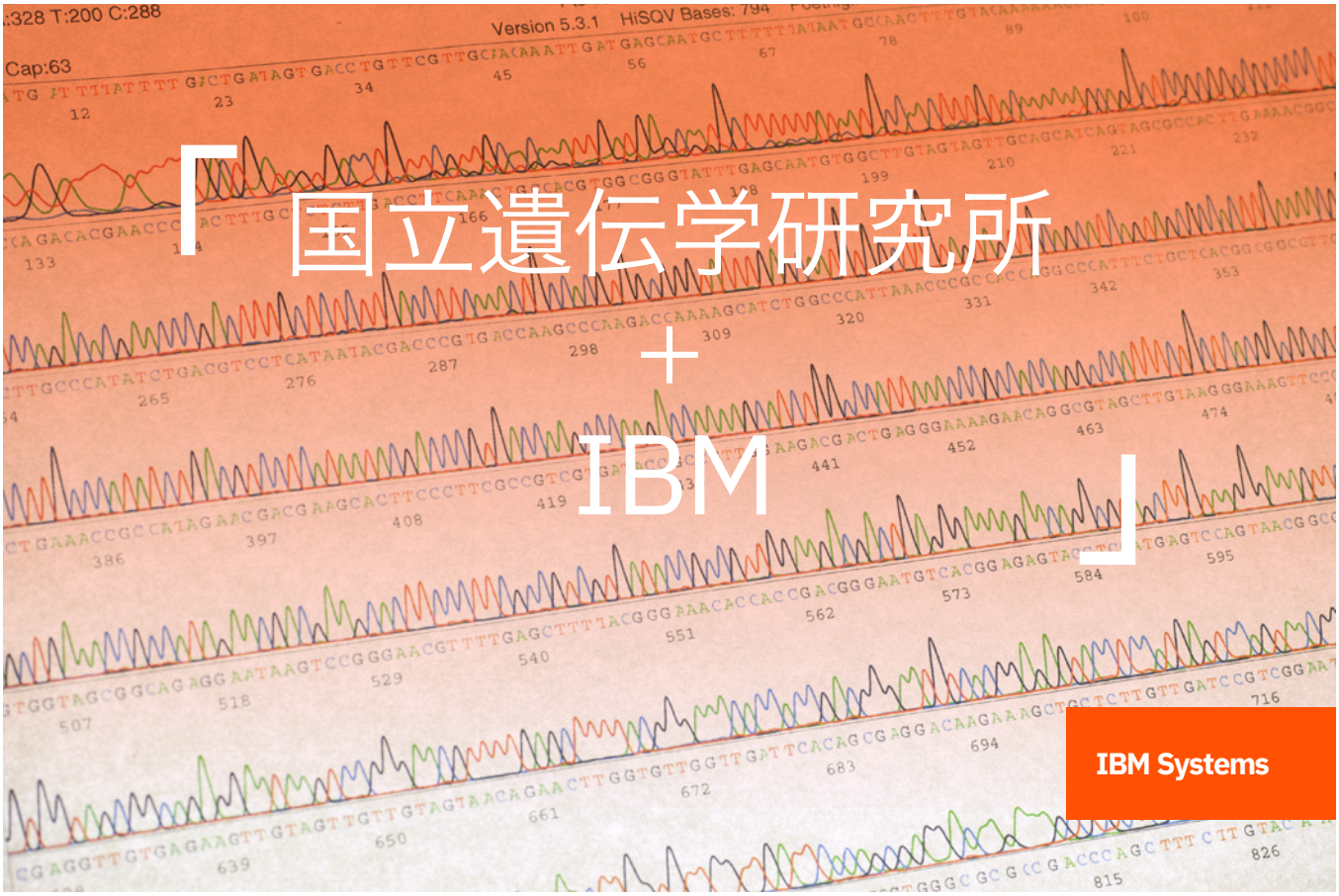
お客様事例



生命科学の解明を目指した先端研究を支える日本最大級となる30PBのDNAデータベース基盤を構築

国立遺伝学研究所DDBJセンターは、日本、米国、欧州の3極連携体制のもと世界中の研究者が自由にデータを利用できる国際塩基配列データベースの日本の拠点として1986年からデータベースの構築・公開を行ってきました。2005年ごろから普及し始めた次世代シーケンサーの影響で、同研究所に保存されるDNAデータ量が近年急速に膨れ上がり、そのデータを安全に保管し、迅速に提供するために、同研究所では2018年春に30PBという日本最大級の階層型ストレージ基盤を構築しました。このストレージ基盤に採用されたのが、高速並列ファイルシステムIBM Spectrum Scaleを搭載した「IBM Elastic Storage Server（ESS）」とテープ・ライブラリー「IBM TS4500」です。安全性と高速性を両立した階層型ストレージが日本と世界の生命科学の発展を支えています。

[導入製品]
- IBM Elastic Storage Server
- IBM Spectrum Scale（旧GPFS：General Parallel File System）
- IBM Spectrum Protect（旧TSM：Tivoli Storage Manager）
- IBM TS4500 テープ・ライブラリー

課題
- 年率1.4倍から2倍で増える高速シーケンサーの出力データを含み、十数億件規模になる国際塩基配列データを安全に保管し、迅速に提供するストレージ基盤が必要とされていた

ソリューション
- データ保全と高速性を両立するIBM ESSによって30PBの階層型ストレージ・システムを構築した

効果
- パフォーマンスが向上し、データ保全が担保できたことで、データ量が急速に増大するなかでの多数の利用者からのデータ登録・利用の要求への対応が強化された。また解析サービスの拡充といった新たな分野での活動も強化できる

[お客様課題]

予想されるゲノム解析の広がりに
対応できるデータベース環境が必要に

国立遺伝学研究所は、日本の生命科学の中核拠点として、1949年に設立されました。現在は、国立極地研究所、国立情報学研究所、統計数理研究所と共に大学共同利用機関法人情報・システム研究機構を構成し、遺伝学の先端研究、教育、そして遺伝学の共同利用・共同研究の場を研究者に提供しています。

生命科学の進歩を支えるゲノム解析は、コンピューターの処理能力に大きく影響されます。同研究所でも1996年にスーパーコンピューターを導入し、以来数年ごとにリプレースし、能力を増強してきました。このスーパーコンピューターを利用して、生命科学研究のデータの共有や、解析サービスを提供しているのが、研究支援事業の一環として設置されているDDBJ（DNA Data Bank of Japan）センターです。

DDBJセンターの役割は大きく2つあります。DNAのデータベースを構築して、研究のために提供することと、解析のためのコンピューター・リソースなどデータ解析環境を提供することです。

DNAの研究者は、研究成果発表の際には塩基配列データを世界の3カ所にあるデータベースのいずれかに登録することが義務付けられていますが、その1つがDDBJです。日米欧の3極で毎日データを交換し合うことにより共通のデータベースとして運用され、論文から参照できる世界中全てのDNA塩基配列データを収録した世界の共有財産として活用されています。このデータ公開・データ共有の仕組みは1980年代前半に整備され、データ駆動型科学の先駆的取り組みとして世界的にも高く評価されています。

また、DDBJセンターの計算機システムがDDBJが構築・維持するDNA塩基配列データベースに高速にアクセスできる利点を生かし、DDBJセンターは研究者にデータ解析のためのコンピューター・リソースの提供を行っています。主なデータは次世代シーケンサーによる高速出力データおよび個人ゲノムです。近年急速に増え続ける個人ゲノムのデータにどう対応するかは、大きな課題になってきています。

2008年に日米欧の3極体制で次世代シーケンサーのデータ・アーカイブを開始したことを受け2012年に大規模データに対応した現在のスーパーコンピューター・システムを導入しました。552ノード相当の分散メモリー型スーパーコンピューターおよび10TBのメインメモリーを利用可能な計算ノードを含む11台の共有メモリー型スーパーコンピューターを導入して大量データの解析に対応すると共に、ストレージ環境を増強してきました。しかし、次世代シーケンサーによるDNAデータの量の拡大、政府が進める生命科学事業へのコンピューター・リソースの提供、DDBJセンターの解析環境の強化などもあって、さらなる増強が必要とされていたのです。

DDBJセンターのシステム管理部門長である特任准教授の小笠原 理氏は「これまで扱ってきたオープンアクセスデータについて、多数のデータ登録・データ利用の利便性を高めることが重要であることに加え、近年の課題は個人ゲノムデータにどう対応するかということでした。個人ゲノムのデータは年率2倍のペースで増えていて、今後保管データ全体の半分を占めると予想されています。まだデータ量が少ないうちに、容量や性能を容易に拡張可能なストレージ基盤を整備し、運用ルールを確立しておく必要がありました」と語ります。

[ソリューション]

確実にデータを守る仕組みがあるから
大容量のファイルでも安心して扱える

DDBJセンターは、次世代のストレージ基盤について2015年ごろから検討を開始し、2016年には仕様を固め、2017年夏頃に入札を行いました。求める条件は3つありました。1つ目はアクセス頻度が低いデータもオンラインであること、2つ目はデータが確実に保全されていること、3つ目は構築時の容量が30PBで拡張性に富み、大きなデータを高速に扱えることです。

こうした条件のもとで選ばれたのが、IBMが提案したIBM ESSによる階層型ストレージ・システムです。大容量のファイルのやりとりに対応できる分散ファイル管理ソフトウェアIBM Spectrum Scaleを搭載した15PBのディスク・ストレージと、IBM TS1155テープ・ドライブを6台搭載したIBM TS4500テープ・ライブラリー（当初の容量は15PB）から構成されています。

IBM ESSの導入によって処理速度が大幅に向上することがわかりました。そして何よりもデータの安全性を担保する独自の仕組み、End-to-end checksumが魅力でした。

国立遺伝学研究所
DDBJセンター　システム管理部門長
特任准教授 博士（理学）
小笠原 理氏

はじめに、すべてのデータを効率よくオンライン状態で保管する必要がありました。「登録されているDNAデータは共有財産として活用されるものです。データの中にはアクセス頻度が少ないデータもありますが、国際塩基配列データベースの性格上そのようなデータも保管されている場所からFTPで取り出せるようにしておかなければなりません。これに加えて全データを用いてデータ解析を行うといった要望にも対応する必要があります。このため、どのようにストレージを構成するかが重要なポイントでした」と小笠原氏は語ります。

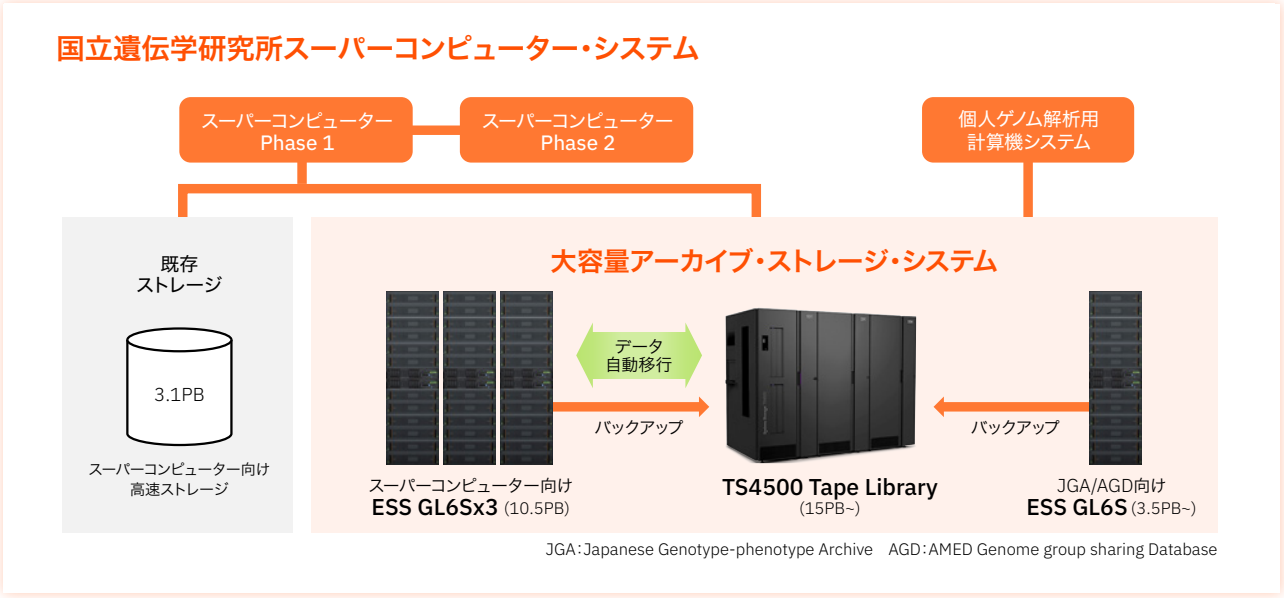
今回採用された階層型ストレージは、ユーザーやアプリケーションが意識する必要なく、シングル・ネーム・スペースで管理されます。アクセス頻度が低いデータは自動的にディスクからテープ・メディアへと移動し、データの保管コストを抑制することができます。また、アクセス頻度が低いデータへのアクセスが発生した場合には、テープ・アーカイブの中の必要なファイルのみを素早く取り出すことができます。

次に、データ保全も重要でした。貴重な共有財産であるアーカイブ・データは確実に保管されていることが求められます。小笠原氏は「データは無くしてはいけないものであり、システムの信頼性が重要となります。これまではトラブルに備えて2重、3重の安全策を講じてきましたが、30PBという容量で同じことはできません。だからこそ確実にデータを守れる仕組みが必要でした」と語ります。

小笠原氏が注目したのはIBM ESSが提供する「End-to-end checksum」という機能です。大容量のファイルを転送する際に、小さな単位に分割してブロック単位ごとに整合性をチェックするものです。「扱うデータ量が大きいだけにトラブルが起きる可能性はあります。しかし、全体の容量が30PBもあるだけに冗長化は困難です。だからこそ、ファイル転送をしながら確実にデータを守ってくれるEnd-to-end checksumという機能に魅力を感じました。あったらいいなと思っていた機能が提供されていたのです」と小笠原氏はIBM ESSを選定した理由を語っています。通常では、こうした手法を採用すると転送速度が遅くなりますが、高速なIBM Power Systemsサーバーと IBM Spectrum Scaleソフトウェアが緊密に統合されたIBM ESSではそうした影響はありませんでした。

最後に、30PBという容量と拡張性、大きなデータを高速に扱えることです。「他のセンターでのIBM Spectrum Scale導入実績については知っていましたが、性能を評価するために、ベンチマークテストでスループットを計測してみました。すると、これまで使っていたZFSより処理速度が大幅に向上することがわかりました」（小笠原氏）。

IBM Spectrum Scaleは、世界中の数々のスーパーコンピューター・プロジェクトで実績があり、20年以上にわたりユーザーからの高速化のニーズに応えるべく進化を続けてきました。さらに、今回採用されたエンタープライズ・テープ・ドライブIBM TS1155は、非圧縮データの転送速度360MB／秒に達しており、高速なハードディスクをさらに超えるスピードを安定して発揮することができます。容量に関しても、IBM TS4500の拡張フレームを追加することで、消費電力を最小限に抑えつつ、ストレージ容量を大幅に拡大することができます。



【効果/将来の展望】

ストレージのボトルネックが改善されることで解析サービスの拡充にも貢献できる

IBM ESSとテープ・ライブラリーは2018年1月に搬入され、セットアップと性能確認が行われた後、3月から稼働を開始しました。「半年くらいかけてデータを移行する予定です。従来システムでは、オープンデータ用に5.5PB、計算リソース用に7PBのストレージ容量がありますが、どちらも9割程度使用しているので、パフォーマンスに影響が出ています。新しい階層型ストレージ環境へ移行することで容量に余裕ができ、ストレージがボトルネックになっている現状が解消されると期待しています」と小笠原氏は話します。

DDBJセンターが取り組むゲノム研究に関するデータは現在も増え続けており、ストレージ環境には拡張性があることも重要な要素です。「5年後にはオープンアクセス用が30PBになり、個人ゲノムと合わせて50PBになると予想しています。さらに、機能ゲノミクスやエピゲノム、メタボロームなど、研究領域がゲノムから発展しています。今後は、そうした多様なデータにも対応していかなければなりません」(小笠原氏)。さらに、顕微鏡写真のような画像データもアーカイブして、海外の計算拠点とデータ転送したいという研究者コミュニティーからの要望への対応も検討されています。

「基本的にはデータベースなので、その隣で解析するサービスを提供する必要性も感じています。2013年頃からMinIONが話題になっているように、例えばベッドサイドで小型なシーケンサーを用いてDNAが分析できるようになっていくと、それに応じた解析サービスも必要になるでしょう。今回ストレージ基盤を強化したことで、できることが広がりました」と小笠原氏は話します。

ゲノム解析がより身近になることは、生命科学の解明に繋がります。DDBJセンターの新たなストレージ基盤には、国内外の遺伝学研究者向けのクラウドセンターとしてサービス拡充の契機となり、その解明に貢献することが期待されています。

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

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

〒411-8540 静岡県三島市谷田1111
https://www.nig.ac.jp/

国立遺伝学研究所は日本で唯一の遺伝学の専門研究機関として1949年に設立され、理論から実験まで、幅広い分野で研究に取り組み、生命科学分野で大きな成果を生んできました。近年ではスーパーコンピューターによる研究支援に力を入れています。

〒103-8510 東京都中央区日本橋箱崎町19-21