課題番号 jh150028-MD01

計算資源の連携を目指した 将来のサイエンスビッグデータ共有機構の開発

松岡 聡 （東京工業大学）

概要　近年の各種センシング技術の発展により、科学技術分野で生成されるデータ量は飛躍的に増加し続けている。例えば、遺伝学研究所が管理する国際塩基配列データベースは現状で 1.5 PB 程度ではあるが、数年のうちに 5 PB に達すると試算されている。このように急激に成長続けるサイエンスビッグデータを、効率的に保存・活用する技術が求められている。加えて、生成データ量の増加により、それを解析するための計算量も増加し、既存のスーパコンピュータの計算資源が不足することが懸念されている。このことから、外部機関のスーパコンピュータや、クラウド DC などの計算資源と連携し、必要な計算資源量を確保する必要がある。本研究では増加し続けるサイエンスビッグデータを、保存・共有するためのシステム環境を構築し、遠隔の計算資源より高効率にアクセス可能にする技術を開発する。本年度は、実証実験基盤の整備と遠距離計算資源へのアクセス技術であるクラウドバーストバッファの拡張等を行った。

1. 共同研究に関する情報

(1) 共同研究を実施した拠点名

- 北海道大学
- 東京大学
- 東京工業大学
- 九州大学

(2) 共同研究分野

- 超大規模データ処理系応用分野
- 超大容量ネットワーク技術分野
- 超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

- 松岡 聡: 全体統括
- 小笠原 原理: 遺伝学研究所のデータ提供
- 三浦 信一: システムの設計・構築とデータ保存技術の開発・評価、およびシステムの結合
- 大田 達郎: データベース連携
- 徐 天棋: 高速データアクセス手法の開発・評価
- 佐藤 賢斗: 高速データアクセス手法の検討

2. 研究の目的と意義

近年の次世代シーケンサの普及に伴い生命系データベースのデータ量は飛躍的に増大している。International nucleotide sequence database collaboration で管理されている国際塩基配列データベースを例にすると、現状のデータベースサイズは 1.5PB 程度であるが、数年のうちに 5PB にも達する見込みである。このような巨大なサイエンスビッグデータをより効率的に保存する技術が必要である。国立遺伝学研究所（以下、「遺伝研」という）では、国際塩基配列データベースや生命系データベースを管理するとともに、このサイエンスビッグデータの解析用にスーパコンピュータを提供している。しかしながら、近年の生命系計算科学の重要性から生命系研究者の必要計算量が増加しており、スーパコンピュータの混雑が発生し、ジョブの実行までの待ち時間が長くなりつつあり、場合によっては外部のスーパコンピュータやクラウドデータセンターなどとの連携も考慮する必要がでてきている。

一方で、遺伝研が管理する国際塩基配列データベースなどの生命系データベースは、遺伝研が運営するスーパコンピュータだけではなく、計算手法や目的に応じて理化学研究所計算科学研究機構が運用する「京コンピュータ」や東京工業大学学術国際情報センター（以下、「東工大」という）が運用する「TSUBAME2.5」のような大型のスーパコンピュータを用いた利用がなされている。しかし

ながら、これらの生命系データベース等は、手動のファイルコピーによってなされており、データベース等の追加・更新がなされた場合は、再度データのコピーなどが必要であるなど、一貫的なデータ管理という意味では、望ましくない運用となっている。可能であるならば、全国の計算資源より直接アクセス可能な分散ファイルシステム（データレポジトリ）を用意し、一貫した生命系データの管理が望まれている。しかしながら、生命系データベースの一部、特にヒトゲノムデータ等は、個人情報などが含まれているため、そのデータの公開には厳しい公開条件が課せられている。そのため、サイエンスビッグデータを安心・安全に共有可能にするためのシステム基盤が必要になっている。

このような中で、革新的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ー（以下、「HPCI」という）では、HPCI 資源提供機関間のデータ共有システムとして、約 20 PB の HPCI 共用ストレージが構築され運用されている。この HPCI 共用ストレージによって、京コンピュータをはじめとする全国のスーパコンピュータ間で容易に計算データの連携が可能になり、有効に活用されている。しかしながら、限りある資源の有効利用の観点から、利用には HPCI 課題として採択される必要があり、本研究提案が想定する恒常的なサイエンスビッグデータの共有・公開用のデータストレージとしての利用が難しい。

また、一般に遠隔に構築されたデータストレージを有効に利用するためには、如何に広帯域なネットワークを有効活用するかが重要になる。実際に HPCI 共有ストレージでは、SINET5 の広帯域なネットワークを有効に利用し、大きなデータセットの共有を効果的に行うことができる機構を有する。遠隔からのファイルアクセスでは、物理距離に合わせて増加する通信遅延時間が影響し、個々のファイル操作の応答時間が長くなるが、HPCI 共用ストレージでは、多数のファイルを同時に扱うことで通信遅延の影響を隠蔽し、広帯域ネットワークを有効に活用することもできる。しかしながら、通信遅延時間の増大とともに個々のファイルアクセスに対する応答時間は長くなることは避けられない。ファイルキャッシュ機構を利用することで解決することも可能であるが、単純なファイルキャッシュ機構では複数のノード間でファイルの一貫性を維持することが難しい。本問題の解決のためには、複数のノード間でファイルの一貫性を維持可能な何らかの新たな機構を追加する必要がある。

以上を踏まえ、本研究では生命系データベースのような巨大なサイエンスビッグデータの保存・共有するためのシステム環境を構築し、サイエンスビッグデータを遠隔の計算資源より高効率でアクセス可能にする技術を開発することを目標とする。現時点で本研究がターゲットとするサイエンスビッグデータは、遺伝研が管理する生命系データベースではあるが、開発する技術は気象データや天文データなど、他分野のサイエンスビッグデータも想定し、成果を共有可能なものとする。

3. 当拠点公募型共同研究として実施した意義

本共同研究は遺伝研が運用管理する国際塩基配列データベースなどの実データの提供をうけ、大容量のサイエンスビッグデータを複数の拠点で効率的に共有する方法を開発するものである。そのため、実際の評価などには、比較的距離が離れた複数の計算資源を必要とし、拠点公募型共同研究により複数の資源を確保できることが必要である。同様の公募型研究として、HPCI による研究公募が存在するが、本研究は異なる 2 分野の共同研究によるフレームワークの開発であることに加え、HPCI では計算することを主目的とすることから、JHPCN を選択した。

4. 前年度までに得られた研究成果の概要

今年度初採択につき該当事項無し。

5. 今年度の研究成果の詳細

5.1. 実証実験基盤の構築と効率的なデータ保存方法の開発・評価

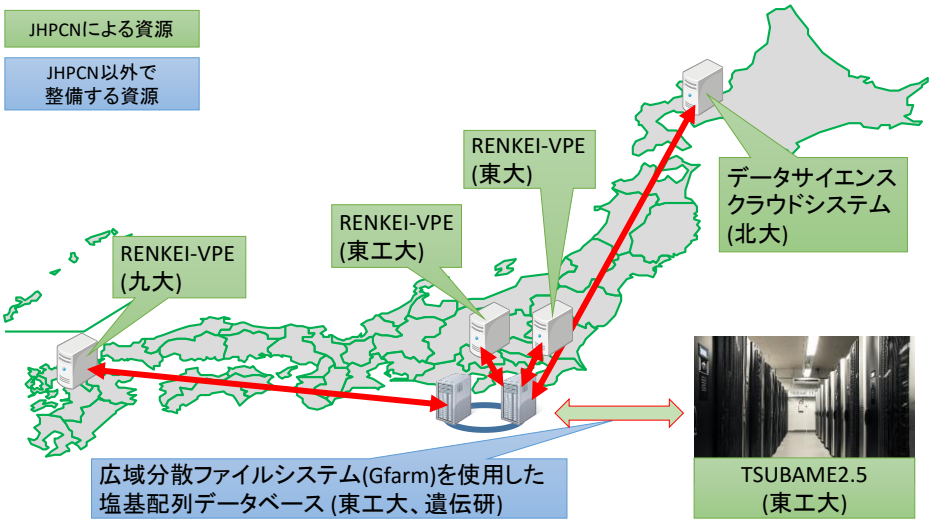

図 1 システム基盤の構築

本研究の評価の基盤として、広域分散ファイルシステムを構築し、このファイルシステム上に実際の生命系データベースを構築する。具体的には図 1 のように、Gfarm ファイルシステムを用い、東工大と遺伝研に広域分散ファイルシステムを構築し、遺伝研が有する国際塩基配列データベース等を展開する。本環境を用いて、I/O スループットなどの基本性能を計測するとともに、本データを使用するアプリケーションを東工大設置の TSUBAME や、各拠点に整備された RENKEI-VPE で評価する。最も遠隔にある北大のデータサイエンスクラウド環境を用いて、アクセス特性などを評価する。これらの評価を通じて、既存の広域分散ファイルシステムの問題を把握し、サイエンスビッグデータ向けの広域分散ファイルシステムのあり方を検討する。

今年度は、本実験基盤の中核となる東工大、遺伝研の 2 拠点にまたがる分散ファイルシステムを構築した。分散ファイルシステムとして、現在 HPCI で利用されている Gfarm ファイルシステムを用いた。

今回 Gfarm ファイルシステムを構成するファイルサーバでは SMR: Shingled Magnetic Recording 技術を用いた HDD を用いている。SMR 技術は今後の HDD の高密度化を実現する核となる技術である。安価でありながら大容量を実現することが可能であるため、今後のファイルアーカイブに用いられる HDD に利用されていくことが期待されている。一方で SMR 技術を用いた HDD では、細かいデータの read/write において、性能が得られないという特性をもっており、本 HDD を用いて RAID を構成した場合、性能が得られないという問題がある。実際、今回構築した Gfarm ファイルシステムでは、1 台のファイルサーバあたり 24 台の SMR 型 HDD で成り立ち、RAID6 を用いて 1 つのデータスプールを構築した。この結果として、データスプールあたり 50MB/sec 程度の性能しか得ることができず、システムの評価環境として全く適さないものとなってしまった。このことより、SMR 型の HDD を使う場合は、RAID 構成ではなく、HDD 単体をデータスプールとして扱わなければならない。残念ながら、現在の Gfarm ファイルシステムの実装では、データスプール 1 つに対して 1 つの IP アドレスが必要になり、今回構築する数百台規模の HDD を用いたデータスプール構成では、IP アドレスが枯渇しかねない。また、数千台規模の HDD 構成によるシステム構築も検討しており、現状の Gfarm ファイルシステムを用いたシステムでは HDD を RAID 構成にせざるを得ず、性能か得られない可能性がある。

以上の理由により、評価環境として、Gfarm ファイルシステムを構築したが、基本となるデータスプール自体の問題で必要となる性能が得られなかったことから、本ファイルシステムを用いた詳細

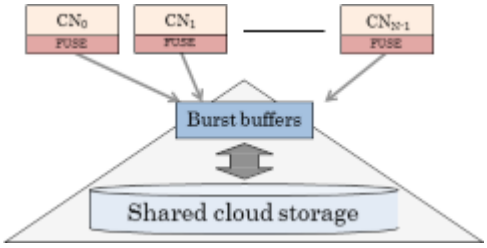

図 2 Two-level storage hierarchy in CloudBB

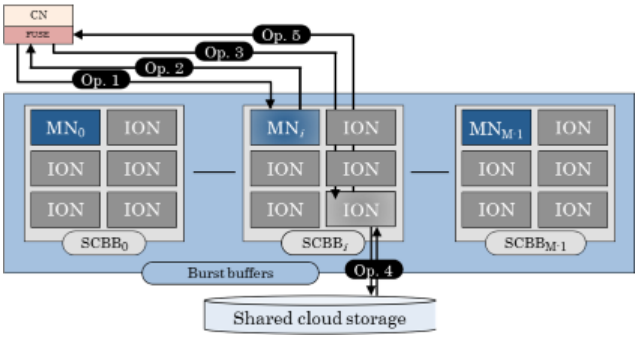

図 3 Architecture of CloudBB

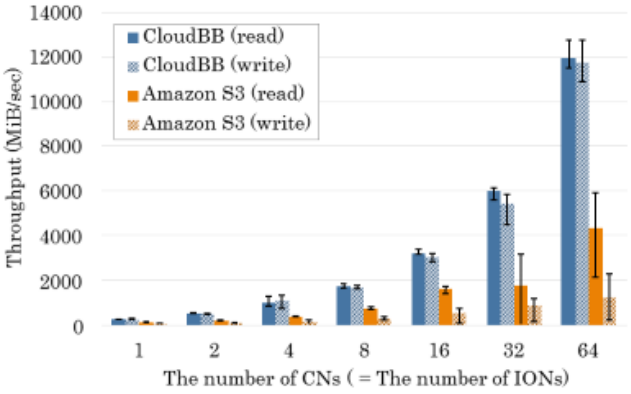

図 4 Sequential I/O Performance

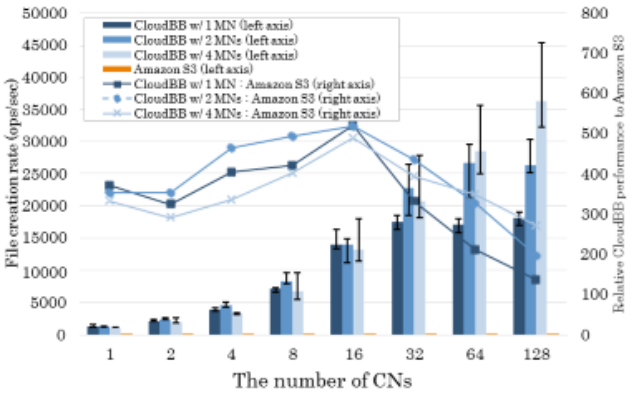

図 5 File Create Performance

な性能評価を断念した。これらを踏まえて、IP アドレスを最小限にしつつ数千のデータスプールを扱うことが可能な Swift や Ceph などの分散オブジェクトストレージでシステムを再度構築することとした。

5.2. 遠隔計算資源より広域ファイルシステムに高速にアクセスする技術の開発

現在のクラウド環境で用いられている共有ファイルシステムは、HPC アプリケーションに求められる I/O スループットに十分対応することができていない。これは、クラウド上では実際にファイルを使用する計算ノード（以下「CN」という）と共有ファイルシステム間の物理的なネットワーク性能が、HPC 計算の主力であるスーパコンピュータのそれよりも、遅延時間で 10 倍以上 転送帯域では 10～100 分の 1 以下であることに起因する。一般的には、データの移動コストは非常に高いため、この問題を解決する最も単純かつ最善な方法は、共有ファイルシステムの位置を固定したうえで、使用する物理的な CN を実際の共有ファイルシステムに近傍に配置することである。しかしながら、このような CN の配置制限は、本来クラウド環境が持つ柔軟な計算リソース確保に制約を与えるため望ましいものではない。一般的なスーパコンピュータと同様、クラウド環境下においても CN 上のローカルファイルシステムにデータステージングする手法は有効ではあるが、複数の CN 間でデータの一貫性確保が難しく、複数の CN からある特定のファイルに read/write を繰り返すような、データインテンシブアプリケーションには有効に機能しない。

本問題を解決するために、我々は遠隔のファイルシステムに各々の計算機資源から効率的に直接アクセス可能にする技術であるクラウドバーストバッファ（以下、「CloudBB」という）を提案・実装している。CloudBB の概念を図 2 に示す。CloudBB はオンデマンドのキャッシュ制御機構を有するファイルバッファ（第 1 階層）と既存ストレージ（第 2 階層）から成り立つ。クラウド上の各 CN が、CloudBB が提供する第 1 階層にあるファイルバッファを経由し、第 2 階層の実ファイルシステムが

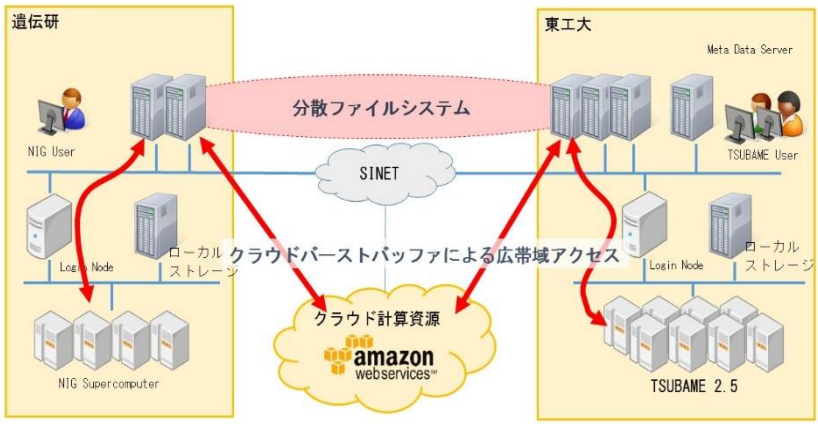

図 2 スーパコンピュータとクラウドリソースの連携

有するファイルにアクセスすることで、データの一貫性を確保したまま、ファイル I/O を効率的に行うことを可能にする。CloudBB は第 2 階層のストレージとして、Amazon S3 といったクラウドストレージを主なターゲットとして開発したものであるが、一般的に Linux がマウント可能なファイルシステム（NFS、Lustre、GPFS、Gfarm 等）であれば使用できるとともに、TSUBAME2.5 のような一般的なスーパコンピュータでも利用できる。

既存の CloudBB の実装では、第 1 階層でのファイルの一貫性確保のために、一般的な並列ファイルシステムと同様に、メタデータサーバを用いるマスタワーカ型の実装となっていた。しかしながら、マスタワーカ型の実装はシステムの拡張性に乏しく、複数の CN が多量のファイル I/O を行う場合に、メタデータの性能がシステム全体の性能のネックとなる。このような拡張性の制限を取り除くためには、オブジェクトストレージなどで用いられている、Key-Value モデルを用いることが一般的であるが、Key-Value モデルはアプリケーションによってはデータ一貫性の維持が難しいという問題がある。今年度の開発では、データの一貫性を維持しつつ、性能向上を得るために、マスタワーカモデルと Key-Value モデルのハイブリットであるマルチマスタワーカ型へ CloudBB を拡張した。

マルチマスタワーカを実現する、CloudBB の構成を図 3 に示す。マルチマスタワーカ型の CloudBB は、複数のサブクラウドバーストバッファ（以下、「SCBB」という）で構成される。各 SCBB は、メタデータを管理するマスターノード(以下、「MN」という) 1 台と、その管理下で実際のバッファとして機能する IO ノード(以下、「ION」という) 複数台から成り立つ。CN は、ファイルパスのハッシュ値を用いて SCBB を選択し、その SCCB に紐づけられた MN と ION を用い、ファイル I/O を行う。このように CloudBB を拡張することで、データの一貫性を維持しつつ、大量のメタデータ変更が必要なファイル I/O が発生した場合においても、システムを追従させることが可能になる。

評価には実際のクラウド環境として Amazon EC2/S3 の Asia Pacific(Tokyo)リージョンを用いた。CN、MN および ION の各仮想マシンのインスタンスは M3.xlarge (spot instance)を選択した。各仮想マシンの構成は vCPU が 4 つとし、メモリーサイズは 15GB とした。また、Network 性能として High を選択している。各 CN は FUSE 経由で CloudBB をマウントし、各 ION が Amazon S3 を s3fs 経由でマウントした各 ION 経由でファイル I/O を行った。図 4 に MN を 1 台に設定した場合のシーケンシャル I/O の評価結果を示す。s3fs 単体で用いる場合では、read 4.3GiB/sec、write 1.2GiB/sec の性能が得られることに対して、CloudBB を用いる場合では、それぞれ 12GiB/sec、11.7GiB/sec となり、CloudBB により高い I/O スループットが得ることが確認できた。メタデータ処理の性能評価として

MN を 1,2,4 台にした場合のファイル作成の性能評価結果を図 5 に示す。評価の結果では、16 台の CN と 2 台の MN を用いた場合、CloudBB は s3fs を直接使う場合に対して約 518 倍高速にメタデータ操作が可能であることが分かった。

以上の評価結果により、CloudBB を用いることで、遠隔のファイルシステムにクラウド上の計算リソースから、効率的なアクセスが可能になることが示せた。

5.3. 外部のスーパコンピュータやクラウド DC との連携技術の開発

解析するべきデータはセンサーの性能向上と共に増大し続けており、必要十分な計算資源を一拠点のみが提供することが難しくなっている。また、Amazon web service 等パブリッククラウドの発展により、クラウド資源を利用したデータ解析も有効な手段になりつつある。そこで、複数のスーパコンピュータ資源間での連携やスーパコンピュータとクラウド資源の連携について模索する。これらの研究は過去においてグリッド関係でなされてきているが、VM やコンテナ技術の発展で使用可能な技術が成熟しており、現在の技術を用いた最適なシステムを構築する。まずはじめとして、スーパコンピュータからクラウド資源へのジョブマイグレーションなどを検討する。その際に各資源間でファイルをシームレスに共有する手段としてクラウドバーストバッファのようなシステムの利用を検討する（図 2)。

今年度は、5.2 節に示したクラウドバーストバッファをスーパコンピュータである TSUBAME2.5 に適用し動作することを確認した。

6. 今年度の進捗状況と今後の展望

実証実験基盤の構築では、分散ファイルシステムの一部を設置する遺伝研内部において、想定のネットワーク速度が得られないという問題が発生した。調査の結果、接続したネットワーク装置が想定した性能を提供していないことが原因であることが判明した。また、遠隔の計算リソースとして、北大のデータサイエンスクラウドを用いて評価を行う予定であったが、データサイエンスクラウド上の仮想マシンから、外部ネットワークへの接続に対してデータ転送性能が得られないという問題も見られ、最終的に今年度は遠隔地からファイルアクセスなどの評価は見送った。また 5.1 に示したように、当初評価に用いるために構築していた大容量分散ファイルシステムが、性能評価に耐えうる性能が得られず、システムに合わせて再構築を必要とするなど、今年度は研究のインフラ面の整備で大きな課題を残した。

他方ソフトウェア部の開発では、遠隔計算資源より広域ファイルシステムに高速にアクセスするための基本システムであるクラウドバーストバッファの性能向上と安定化を進めることができた。これにより、今後の外部のスーパコンピュータとクラウド DC との連携に目処をつけることができた。しかしながら、現状のクラウドバーストバッファの評価では、Amazon EC2 上での評価にとどまり、本来の計画した、TSUBAME2.5 等のスーパコンピュータと外部ストレージとの連携評価が行えておらず、TSUBAME2.5 上でのクラウドバーストバッファの基本動作の確認にとどまった。

このように、今年度は個別の技術要素の整備と技術開発に終始し、実システムへの適用、特に整備したストレージ環境とクラウドバースバッファを用いた性能評価などは、今後の課題として残った。

今後は個別の技術要素の一体化を進め、本課題の最終的な目標となる、外部のスーパコンピュータやクラウド DC との連携技術の確立を進めていく。

7. 研究成果リスト

(1) 学術論文

該当事項なし

(2) 国際会議プロシーディングス

該当事項なし

(3) 国際会議発表

- Tianqi Xu, Kento Sato and Satoshi Matsuoka,

"Towards Cloud-based Burst Buffers for I/O Intensive Computing in Cloud", In HPC in Asia Workshop in conjunction with the International Supercomputing Conference (ISC'15), Frankfurt, Germany, July, 2015. (Refereed Poster)

- Tianqi Xu, Kento Sato and Satoshi Matsuoka, "Design and Modelling of Cloud-based Burst Buffers", In Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis 2015 (SC15), Austin, USA, Nov, 2015. (Refereed Poster)

(4) 国内会議発表

- Tianqi Xu, Kento Sato and Satoshi Matsuoka, "Cloud-based Burst Buffers for I/O Acceleration", IPSJ SIG Technical Reports 2015-HPC-150, Beppu, Japan, Aug, 2015.

(5) その他

該当事項なし