大規模データ処理システム最適化 WG
成果報告書

（WG 活動期間：2018 年 9 月～2021 年 2 月）

2021 年 3 月

サイエンティフィック・システム研究会

大規模データ処理システム最適化 WG

商標について

記載されている製品名などの固有名詞は，各社/各機関の商標または登録商標です.

目次

1. はじめに

1.1. 活動方針

近年の基礎科学プロジェクトの大規模化に伴い，実験・観測装置から産出されるデータ量は加速度的に増えている．国際研究の枠組みにおいては，PB クラスに及ぶ大量データを結果の品質保証を含め短期間に処理しなければ，競争力を持つ科学的成果を上げることは難しい．しかし各研究機関では，限られた予算と時間の中で構築した必ずしも最適ではない計算機システムにおいて苦心して解析処理を行っている実情がある．また，手作業によるリソース割り当てや結果確認など，計算機の効率の問題に加えて人が介在する作業のオーバーヘッドも無視できない．

大規模データ処理を伴う研究が発展を遂げる中で，高速なデータ処理を実現するためには，必要とされる処理とそれに伴うデータ入出力に応じた計算機システムの構成を最適化するとともに，与えられたシステムのリソースを効率的に使い切るためのプロセス割り当ての最適化の双方が必要である．また，機械学習などの技術を用い，プロセス実行や解析処理の異常検知，品質評価などの科学的スキルを要する作業を，トレーサビリティを確保した上で自動化・効率化することも有効と考えらえる．

こうした現状を鑑みて，本 WG ではいくつかの分野において，大規模データ処理のシステムおよび処理手順の最適化に対する要求や技術上の課題を整理し，最新の HPC・ファイルシステム技術，また機械学習等の作業支援技術を調査した上で，それぞれの要請に適した技術の提案を行い，既存技術の課題や今後の方向性について議論を行う．

1.2. 活動内容

WG 前半では，様々な分野・サイトについて，大規模なデータ処理を扱うシステム設計や運用の具体的な必要事例と最適なシステムの構築・運用に関する課題・要求を調査し整理する．また，それらの課題・要求に対応した現在のシステム構築・運用技術に関する技術や将来動向を調査する．調査においては特に，ユースケースに応じた計算機・ストレージ構成，深層強化学習などを活用した計算機資源を効率的に使うためのプロセスとデータ入出力の管理，データの利便性とその長期保存および活用戦略（メディア，圧縮技術，保存すべきデータと方法を含む）やセキュリティを整合的に実現するシステム設計，等の観点に着目する．

WG 後半では，調査した技術と課題・要求との隔たりを整理し，必要とされる技術の検討，ならびにシステム設計・運用のあり方についてまとめる．また，解決できる可能性のある課題・要求については，技術の応用方法も検討し提言する．

1.3. 活動期間

2018 年 9 月～2021 年 2 月（会合　計 9 回）

1.4. WG メンバ

Table 1　WG メンバ

区分	名前	機関	役割
会員	髙田　唯史	国立天文台	担当幹事
	古澤　久徳	国立天文台	推進委員(まとめ役)
	堤　　誠司	宇宙航空研究開発機構	推進委員
	伊達　　進	大阪大学	推進委員　[2019 年 4 月~]
	神武　直彦	慶応義塾大学	推進委員
	真鍋　　篤	高エネルギー加速器研究機構	推進委員
	田村　義保	統計数理研究所	推進委員
賛助会員 (富士通)	山田　久仁	富士通(株)	推進委員(まとめ役)
	堤　　純平	富士通(株)	推進委員
	齋藤　　悠	富士通(株)	推進委員
	齊藤　　哲	富士通(株)	推進委員　[~2020 年 3 月]
	甲斐　俊彦	富士通(株)	推進委員
	岡本　拓也	富士通(株)	推進委員
	秋間　学尚	(株)富士通研究所	推進委員
	大辻　弘貴	(株)富士通研究所	推進委員
	田村　雅寿	(株)富士通研究所	推進委員
事務局	松本　孝之		
	青木　伸子		

1.5. 活動実績

第 1 回会合：2018 年 11 月 28 日（水）　14:00-17:20　富士通（株）本社
出席者：会員 6 名／富士通 9 名／事務局 2 名
・メンバ確認，WG 名称の決定など
・活動方針，活動内容の確認
・メンバ機関(天文台，慶応大，高エネ研)からの現状・問題点などの報告
・今後の進め方の検討

第 2 回会合：2019 年 2 月 8 日（金）　14:00-17:10　富士通（株）本社
出席者：会員 4 名（情報提供者 1 名）／富士通 9 名（情報提供者 1 名）／事務局 2 名
・メンバ機関(JAXA，統数研)からの現状・問題点などの報告
・富士通からの情報提供

第 3 回会合：2019 年 4 月 19 日（金）　13:00-17:00　富士通（株）DTC 東京
出席者：会員 6 名／富士通 6 名／事務局 2 名
・メンバ機関から提供された情報の整理
・ワークショップ形式にて，検討すべき主要な要素の洗い出し・現状・動向調査

第 4 回会合：2019 年 6 月 27 日（木）　14:00-17:30　東京大学工学部 12 号館
出席者：会員 6 名（情報提供者 1 名）／富士通 8 名／事務局 3 名
・ゲストスピーカからの情報提供
・成果目標の設定，重点テーマの決定とケーススタディのインプット情報提供者決め

第 5 回会合：2019 年 10 月 4 日（金）　14:00-17:30　富士通（株）本社
出席者：会員 6 名／富士通 6 名（情報提供者 1 名）／事務局 2 名
・天文台からのデータ量推移情報の提供およびケーススタディの実施
・富士通からの情報提供

第 6 回会合：2020 年 1 月 22 日（水）　14:00-17:00　国立天文台三鷹
出席者：会員 7 名／富士通 5 名／事務局 2 名
・天文台からのシステム構成の情報提供とそれに基づいた課題の討議
・報告書の目次と担当の方針決め

第 7 回会合：2020 年 7 月 2 日（木）　15:00-17:15　オンライン会議
出席者：会員 5 名／富士通 7 名／事務局 2 名

・JAXA・天文台のケースに基づいたシステムモデル化の議論，各機関のケーススタディの実施
・報告書作成内容の方針決め

第 8 回会合：2020 年 10 月 1 日（木）　　14:00-17:55　　オンライン会議
出席者：会員 6 名／富士通 8 名／事務局 2 名
・KEK のケースに基づいたシステムモデル化の議論，各機関のケーススタディ内容の確認
・報告書の構成，記載内容の認識合わせと合意

第 9 回会合：2020 年 12 月 17 日（木）　　14:00-17:30　　オンライン会議
出席者：会員 5 名／富士通 7 名／事務局 2 名
・報告書の各担当執筆済部分の確認，議論，本文とりまとめの実施
・報告書本文のブラッシュアップ方針，イントロ・まとめ部分の検討，完成に向けた進め方の決定

1.6. 本報告書の構成について

本 WG では，大規模なデータ解析を行うためのシステムを設計・構築・運用する際に，最適なシステムの実現のために留意すべき事項について議論した．
データ解析システムの最適化という言葉が示す範囲は非常に広い技術項目を含む．実際，本 WG の各メンバの関心は，各機関で運用するデータ処理システムが扱うデータの性質やそれに応じた処理内容，データの保存や情報保護に対する要請に応じて異なっている．データ処理システムの最適化においてメンバが重要と考える観点は以下のような項目に分類される．
[a] システム（計算機・プロセス）の最適化（クラスタ，ジョブ効率化，継続・復旧）
　　計算機・動的/静的，プロセス・ジョブ割り当て，異常検知・復旧
[b] システムの利用方法（データ配置と移動）
　　ジョブ管理とデータ配置の連動，処理速度とバランスしたデータ転送
[c] データ保持の最適化（データ配置と階層化，圧縮，情報保護）

これらはいずれもデータ処理システムを特徴づける重要な要素であり，最適なシステム設計において総合的に考慮されるべきである．しかしその中でも特に多くのメンバが現在も直面している課題を鑑みて，本 WG では主に[b]データ配置・移動，[c]データ保持，特にデータ量やアクセスの頻度に着目して議論を行うこととした．
この観点で各機関が扱うシステムの特性をワーク形式で分類したところ，Table 2 のような結果を得た．次章で行う各機関のシステム事例紹介では，各システムがどのような特性を持つものか容易に判別できるよう，各節の先頭にこの分類を付記した．この分類を参考にすることで，読者がご自身の関心のあるシステムに類似する他機関システムの運用状況を参照したり，本報告書の議論を今後のシステム設計の参考情報として利用するなどして役立てていただけるのならば幸いである．

データ運用の特性 / 程度	データサイズ	ファイル数	データ量（サイズとファイル数を合わせた場合）	保存期間	I/O の割合
大	数値計算 天文観測 加速器実験 大学計算機センタ	数値計算 天文観測 加速器実験 IoT-POS SNS 医療カルテ 公的データ 大学計算機センタ	数値計算 天文観測 加速器実験 SNS 大学計算機センタ	天文観測 IoT-POS 医療カルテ 公的データ 国民性調査 大学計算機センタ	数値計算 天文観測 SNS 公的データ 加速器 大学計算機センタ
中	SNS 医療カルテ		IoT-POS 医療カルテ 公的データ	数値計算 加速器実験	IoT-POS
小	IoT-POS 国民性調査	国民性調査	国民性調査		医療カルテ

Table 2　データ量とアクセス頻度による各データ解析システムの特性

本報告書では，WG で実際のシステム事例に即して行ったシステム最適化に関する議論とそこから得られた知見を以下のような構成に沿って述べる．まず，2 章で各メンバ機関の扱うデータ処理システム事例の概要と課題を俯瞰する．3 章ではそのうち 3 例を取り上げ，データ量に着目することでシステムの特徴を簡素化してシステム設計の考え方を議論する． 4 章ではシステム設計に関連する最近の技術動向を紹介し，最後に 5 章で議論のまとめを行う.

2. 諸大学研究機関における大規模なデータ解析システムの現状と課題

本章では，数理・工学・基礎科学分野の大学・諸研究機関における大規模なデータ解析またはデータの入出力を伴うシステムを取り上げ，それらの運用に関して現状と課題を示す.

2.1. 国立天文台すばる望遠鏡広視野撮像データ処理システム（東京都三鷹市）

2.1.1. システムの分類

種類：天文観測
データサイズ：大
ファイル数：大
保存期間：大
I/O 割合：大

2.1.2. システムの目的

自然科学研究機構国立天文台（三鷹）（以下,「国立天文台」）では，すばる望遠鏡を含む観測データを国内外の天文学コミュニティに対して提供するためのシステムを運用している．以下では，その中でも国際共同研究の下，すばる望遠鏡の可視光広視野カメラ（HSC）の大きな観測時間を投じて行われる掃天観測データを扱うデータ解析・アーカイブシステムについて概観する．このシステムは，掃天観測の大規模なデータを集約的に解析処理して科学利用できる状態に較正し，さらにその処理結果を共同研究者および一般の天文データユーザに提供するために用いられる．

2.1.3. システム構成の概要

HSC データ解析・アーカイブシステムの概要を以下に示す（Figure 1）．本システムは国立天文台のデータアーカイブ室に設置されている．

システムはデータ解析を行う系（図左側）とデータアーカイブ・公開を行う系（図右側）に大別される．解析系は観測装置が取得した未処理のデータ（生データという）を一時貯蔵場所である生データアーカイブから取り出し集約的に処理し，科学利用できる形に較正する．アーカイブ系は，解析系による処理結果をストレージに転送・格納し，画像データとそれらから得られた天体の測定情報などを含むメタデータを格納したデータベースと合わせてユーザに検索取得可能な状態で公開する．その目的に応じて解析系は合計 1500 コアからなる PC 群と高速な分散仮想化ファイルシステム（Lustre および IBS Spectrum Scale）からなり，アーカイブ系は NFS を用いたファイルサーバとウェブおよびデータベースサーバで構成される．

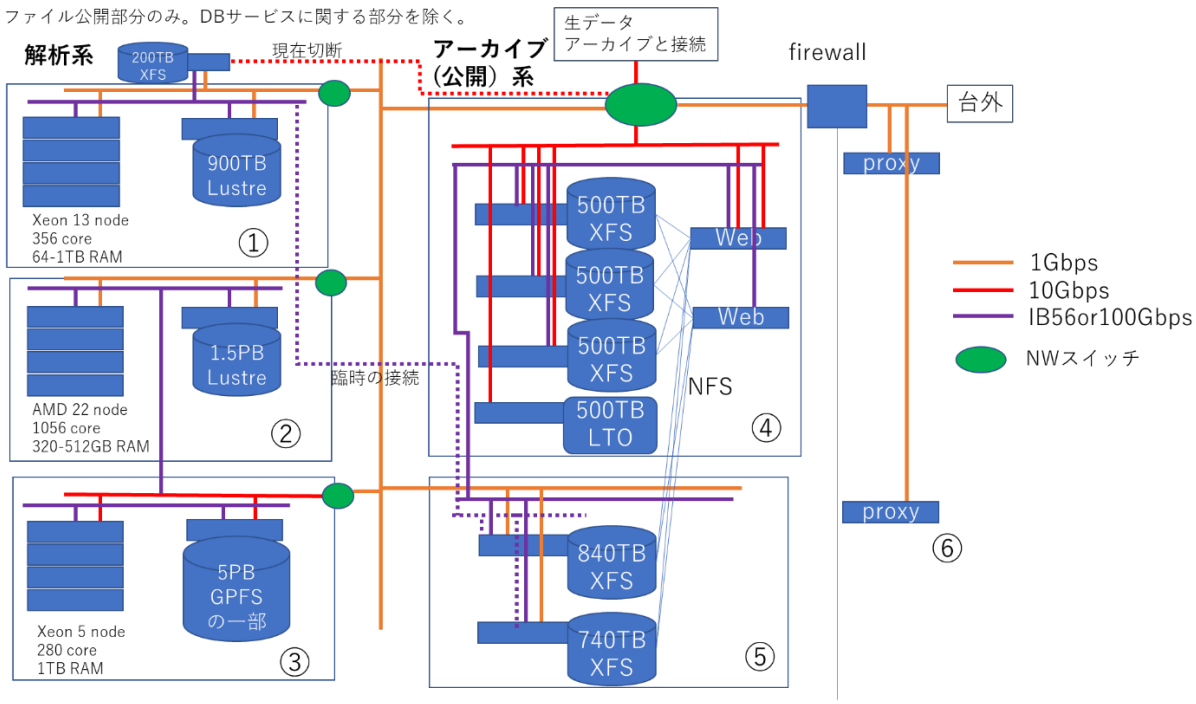


Figure 1　HSC データ解析・アーカイブシステム概要

2.1.4. システムが扱うデータ（種類・性質）と解析処理の概要

本システムが扱うデータを理解するために，まず本システムが扱う解析処理の概要を述べる（Table 3）．本システムは，解析系において広域撮像観測から得られる大量の天文画像データを較正処理し，そこに写っている天体を検出・測定検出する一連の処理を行う．その結果出力をアーカイブ系のストレージおよびデータベースに格納し，国内外のデータ利用者にウェブサーバを通じて公開する．Table 3 では，処理の種類とそれらが必要とするシステムリソースに関係する特性をまとめた．

データ処理の入力データは広視野カメラの個々の検出器(CCD)から得られる画像データであり，観測の一積分あたり 104 個のデータ（計 2GB），一晩あたり 300GB 程度が生成される．ファイル形式はデータの特性を表記したヘッダ部と画素値を格納するデータ部で構成される FITS 形式である．データ処理の大規模なキャンペーンは年 1.5-2 回の頻度で繰り返されるが，ほぼ毎回それまでに取りためられた全データを処理し直すため，入力は数十 TB から 100TB 規模になる．較正処理された CCD 画像データはやはり FITS ファイルとして保存される．それらはさらに決められた天域ごとに合成され結果の画像も同様に保存される．合成画像上で天体が f 検出測定され，天体カタログと呼ばれる個々の天体ごとの測定結果をリストした FITS(BIN-TABLE)形式のファイルが作られる．

以上のような解析処理からは，その結果ファイルおよび中間生成ファイルを合わせて合計で 500TB から 1PB のファイル（入力サイズの約 10 倍）が出力される．個々のファイルサイズは画像で 10～100MB，天体カタログは数十～200MB くらいまでのばらつきがあり，ファイル数は入力約 100 万個に対し出力が 1000 万個のオーダである．

処理ステージ	処理内容	技術	CPU cores	メモリ	ディスクI/O	作業期間
画像処理１(CCDごと)	検出器の感度補正	加減乗除，補間，回帰，主成分解析	100—1000	2-4GB/pcs	入出力が多い	2-3wk
較正	検出器の座標・信号強度決め	最小二乗法・行列演算	>>16 SMP	128GB – 1TB/pcs	入力は高負荷	1-2wk
画像処理２(CCD合成)	足し合わせ	画像変換	100-1000	4-8GB/pcs	入出力が比較的多い	2-3wk
天体測定	天体の検出と測定	畳み込み，測定アルゴリズム，当てはめ	100—1000	8GB -1TB/pcs	入力と出力時にやや負荷	5-8wk

Table 3　HSC データ解析システム内のデータ処理の種類と特性

2.1.5. システム規模の変遷

上で述べたシステム内の各系は 2013 年頃から 5 年以上をかけてデータ量の増加や予算の配分に応じて部分的な更新を続けてきた．そのためシステム全体は異なる構成の計算機やストレージが混在するヘテロジーニアスな系となっており，これらを一体の大型システムとして連動させて運用している．
本格的に稼働を始めた 2014 年度当初は Figure 1 における解析系の①，アーカイブ系の④に相当するサブシステム部分のみが存在した．その後プロジェクトの進行によるデータ量の増加に応じてシステムの増強が必要となり，②および⑤のサブシステムの追加が行われた．また 2019 年には共同利用のデータ解析環境需要の増加を受け，本システムのデータ解析に加えて個々の一般ユーザによる活動もサポートする計算機設備として③が導入された．当初システムは初期 2 年程度のデータを保持できるように設計された．実際は，天候などの影響による観測期間の延長やデータ解析ソフトウェアの変更，データサービスの拡充などもあり，当初想定された以上のデータが生成されている．必要な各サブシステムを一括で増強することは予算措置やデータ量の見積もりの点で難しいため，約 5 年ごとの機関全体の計算機更新に加えて年度ごとに小中規模の部分的な増強を繰り返してきている．

2.1.6. システム運用・設計上の課題

本システムは運用開始から順調にデータ解析・公開の作業を行い一定の成果を挙げてきたが，以下の点において改善の余地があると考えている．

1. 最適な計算機システム構成が確立されていない
   本来はデータ処理手順・データ移動を十分に考慮して最適化したシステム設計を行いたいが，実際は既設の CPU の配置やファイルシステムの空き容量などの制約の中で，必ずしも最適ではないデータ解析や移動を行っているのが現状である．
2. 処理手続きのオーバーヘッドが大きく計算資源を使い切れていない
   解析プロセスの同期・非同期の箇所の，データに依存しメモリなどの必要リソースが変わるアプリケー

ションプロセスを含む解析パイプラインを走らせる際の最適なジョブの割り当て方法・技術が求められる．しかし，解析プロセスの実行は手作業によるデータセットの分割や解析者の経験による定型のリソースの割り当てによるところが大きく，アプリケーションも必ずしも存在するリソースを有効に使い切るようには設計されていない．

3. 処理手続きの自動化・連続実行・異常検知と復旧が不十分である
   処理手続きステージ間の依存関係の処理，結果の品質評価，解析処理の異常検知，自動継続といった解析実施を効率化・自動化する仕組みが整っていない．同期箇所での結果確認や次プロセスの投入にかかる手間と時間，また問題発見までの時間とその修正にかかる時間等のオーバーヘッドは小さくない．
4. システム拡張がシームレスではなく十分に効果的ではない
   既設のシステム内の個々の CPU やディスクなどの物理的な距離の制約，また既存の作業を止められないことなどから，追加するシステム部分が必ずしも既設部分と一体となって動作するようには設定できず，別クラスタや分離したファイルシステムのようになることが多く，増設の効果が最大限に生かせていない．理想的にはファイルシステムは一領域として拡張されると使い勝手が良いが，実際には分断され，本来一領域に格納したいデータセットを分割せざるを得ないなどの非効率が発生する．また，Lustre などの場合，既存 OST と新規 OST の間のファイルの平準化も事実上できておらず性能面でも問題である．

2.2. 高エネルギー加速器研究機構　中央計算機システム（茨城県つくば市）

2.2.1. システムの分類

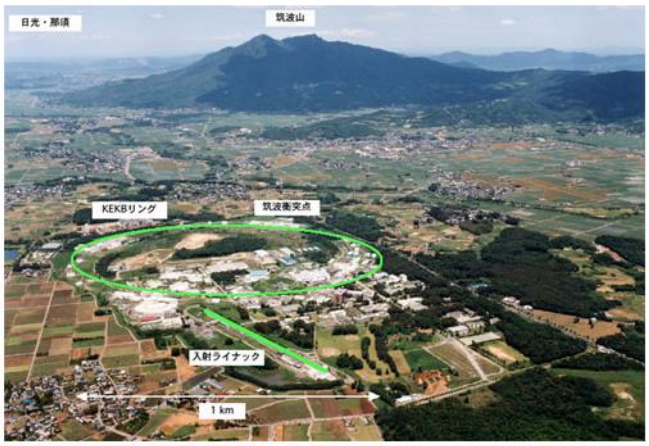
種類：加速器実験
データサイズ：大
ファイル数：多
保存期間：中
I/O 割合：アプリケーションの種類により 少～大

2.2.2. システムの目的

大学共同利用機関法人　高エネルギー加速器研究機構（略称 KEK）では，KEKB 加速器，J-PARC 加速器施設，PF/PF-AR といった大型加速器を利用して，高エネルギー物理学，物質科学，生命科学等の幅広い分野の研究が行われている．計算科学センターでは，これらの研究活動に必要な計算機環境の導入，運用を行っている．KEK 中央計算機システムは(KEKCC)，素粒子・原子核実験，放射光実験，中性子実験，加速器開発，理論計算等の様々な研究ニーズに応じたデータ解析システムを提供する．また，このシステムには，電子メールシステム，Web システム等の情報基盤環境も含まれる．データ解析システムは主に大規模な計算サーバ群とストレージシステムから構成される．

計算サーバは，高速・高集積なサーバ群を高速ネットワークで相互接続した Linux サーバで，インタラクティブ解析やバッチ処理による解析が行われる．ストレージシステムには，大量の実験データ等が保存される．約 20 PB の既存の実験データに加え，本システム稼働中には，数 10PB のデータの蓄積が見積られている．これらの大量のデータへの高速アクセス，高スケーラビリティを実現し，システムを安定的に動作させることは，本システムの最重要課題である．また，本システムは，国内外の大学，研究機関から共同利用されるため，セキュアなネットワークシステムの構築，運用も求められる．すでに，J-PARC 施設では様々な実験が行われ，Belle II 実験も 2019 年春よりデータ収集を開始している．実験データの解析は，国内外の大学や研究機関と連携して行う．Grid 技術を用いた広域分散処理システムを構築し，その運用を行っていくことが要求される．本システムは多数のユーザが 24 時間 365 日利用するシステムである．高負荷状態においてもシステムを安定的に連続運転する必要がある．(2020 年度更新時の仕様書より)

解説：



Super KEKB 加速器：

KEKB 加速器は 電子を 7GeV まで加速し，陽電子を 4GeV まで加速してそれぞれの粒子のビームをぶつけることによって素粒子の実験をするための衝突型加速器である．二つのビームはそれぞれ別々の蓄積型加速器（Storage ring）の中に収められ，リング内一箇所で二つのリングが交差し，ビームが衝突するように作られている．その交叉点を囲むようにして，Belle 測定器が設置されている．また，電子と陽電子は，電子・陽電子線形加速器から供給される．（GeV はエネルギの単位で，1GeV とは，例えば電子を 1GV(10 億 V)の電圧差で加速する場合にあたる．）1999 年より運転していたが，ビーム衝突性能（単位時間に衝突する回数）を 40 倍に増強するため改造を行い 2018 年に初衝突を開始した．

Belle II 実験

世界 26 の国と地域からの 113 の大学・研究機関に所属する約 900 人の研究者が参加（2018 年 11 月時点）する． SuperKEKB 加速器で衝突させた電子・陽電子の反応（衝突の結果，そこから発生した粒子や，その性質，頻度）を詳しく観測することにより，粒子・反粒子の対称性の破れや宇宙初期に起こったはずの極めてまれな現象を再現し，未知の粒子や力の性質を明らかにし，新しい物理法則の解明を図り，宇宙から反物質が消えた謎に迫ることを目的にする．

J-PARC 加速器施設

J-PARC (Japan Proton Accelerator Research Complex) は，素粒子物理，原子核物理，物質科学，生命科学，原子力など幅広い分野の最先端研究を行うための陽子加速器群と実験施設群の呼称．世界に開かれた多目的利用施設である J-PARC の最大の特徴は，世界最高クラスの強度(1MW) の陽子ビームで生成する中性子，ミュオン，K 中間子，ニュートリノなどの多彩な 2 次粒子（大型加速器により加速された粒子を炭素などの静止標的に当てた結果そこから発生する多種の粒子）ビーム利用にある．J-PARC では，初段が 400MeV まで加速するリニアック，次に 3GeV まで加速する RCS (Rapid Cycling Synchrotron) ，さらに一部を 30GeV まで加速する MR (Main Ring) の陽子加速器構成で世界最高クラスの大強度陽子ビーム生成を目指している．

2.2.3. システム構成の概要

KEK 中央計算機システム(KEKCC)の概要と主要仕様を以下に示す．本システムは 2016 年から 4 年間運用され，2020 年 9 月に更新された．2020 年 9 月に更新された新システムも 4 年間のリース契約であり 2024 年に次の更新が予定されている．前述したように，このシステムは Belle II 実験，J-PARC 実験という高エネルギー加速器研究機構の主要な二つの大型加速器実験のデータ解析が主要な利用割合を占めるが，CERN LHC 実験や，宇宙背景輻射測定などの他の実験，理論研究もふくめて高エネルギー加速器研究機構でおこなわれている多くの研究で利用されている．主要な 2 実験が必要とする計算機需要は加速器の性能向上や検出器の性能向上という予定はあるが不確定な要素により決まるため，4 年ごとに 2024 年までのできる限り正確な予測をもとに計算機システムの調達を行っている．利用者はインタラクティブな計算やプログラム開発を 12 台のワークグループサーバで，主な計算をバッチ処理により372 台の計算サーバで実行する．ストレージは主に共有ファイルシステムによるホーム・グループディスク領域（17PB）と，GPFS のインターフェースをもつ階層型ファイルシステム（HSM）HPSS(100PB)からなる．HSM は，キャッシュとして用いられる GPFS-HPSS Interface (GHI) 8.5PB の

GPFS サーバを通してアクセスされる．また，計算サーバは，それぞれ 1 TB の SSD を持ち，ノードでの一時記憶領域として利用される．そのほか，Grid 関係のサーバ，メールサーバ，Web サーバ，各種機能サーバを付帯しているが，それらの説明については割愛する．

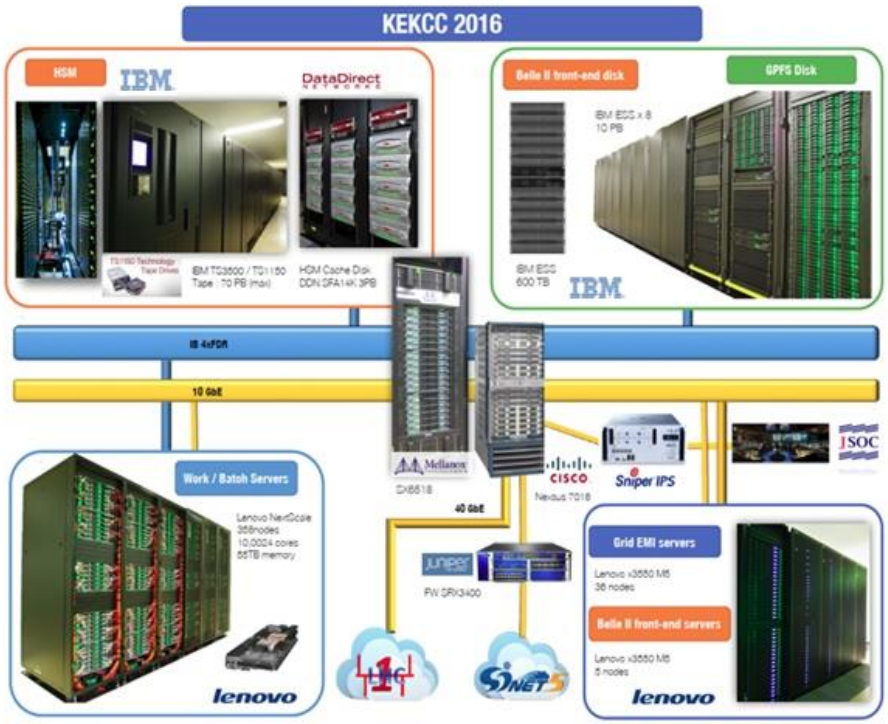


Figure 2　KEKCC 2016-2020　システム概要

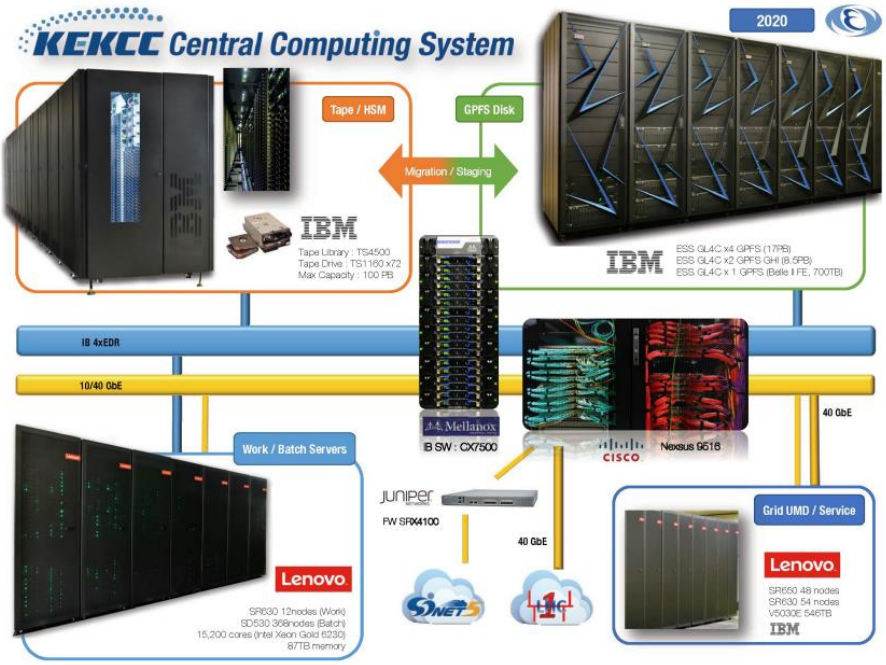


Figure 3　KEKCC　2020-2024　システム概要

	2016 システム	2020 システム	増強比
CPU	Xeon E5-2697 (2.6GHz 14 cores)	Xeon Gold 62300 (2.1GHz 20 cores)	
# of CPU cores	10,024	15,200	× 1.5
SpecInt2000	60k	120k(est.)	× 2
OS	Scientific Linux 6.10	CentOS 7.X	
Disk	10+3(HSM cache) PB	17+8.5(HSM) PB	× 2
Tape Drive	IBM TS1150	IBM TS1160	
# of tape driv.	54	72	
Media	7TB/巻(JC), 10TB/巻(JD) 360MB/s	7TB/巻(JC), 15TB/巻(JD6), 20TB/巻(JE) 400MB/s	
Capacity	70PB	100PB	× 1.4

Table 4 新旧 KEKCC 性能比較

2.2.4. システムが扱うデータの概要（種類･性質）

大型高エネルギー実験の多くにおいては実験そのものが 10 年以上続く．天文のデータのようにその時に取ったデータが，後日取ったデータと異なる情報（その時刻に収集したデータが 2 度と取れないデータ）

であるのにくらべ，高エネルギー実験で収集されるデータは，いつ取ったデータも同じ条件であれば同じ内容をふくむ（はず）という，本質的に時間によらない普遍性がある．その一方，実験データの精度は，データの量による統計性による（たくさんのデータがあればあるほどより正確になる）ため，少しでも多くのデータを活用する必要がある．結局収集した実験データは，その実験が最終的に終了し，完全にデータ解析が終了するまでの 20~30 年程度のデータ保存期間を有する（実験の前半にとられたデータは長い期間保存され，終盤にとられたデータの保存期間は短い）．また，収集された生データは後述する種々の解析を経過して物理量となるが，その最終的な物理量は収集された生データから比べると非常に少ない量となる．生データ自体は，巨大な加速器や検出器を膨大な予算を投じて収集したという観点から非常に貴重なものであるが，将来 新しい加速器や検出器がつくられて 10 年がかりで貯めたデータが新しい実験では数カ月で同じ統計量をとることができるようになったりすることもある．また，生データを解析するには実験装置を熟知した実験グループの多大なノウハウが必要であり，実験をおこなったグループ以外が解析することがなかなか困難なことから，実験生データを長期保存し世界中で実験グループ外も含めて公開し再利用とする動きはあるものの，あまり進んでいない．

2.2.5. システム規模の変遷

KEK が創設されて以来，メインフレーム，SMP，クラスタシステムと形を変えながら KEK 計算科学センターで運用する大型計算機システムは 12GeV 陽子シンクロトロン，トリスタン加速器による実験，J-PARC 実験施設における実験，KEKB 加速器，Super KEKB 加速器における Belle 実験など KEK における主要な高エネルギー実験で使用されてきた．Figure 4 に過去の CPU 性能，ストレージ性能について示す．

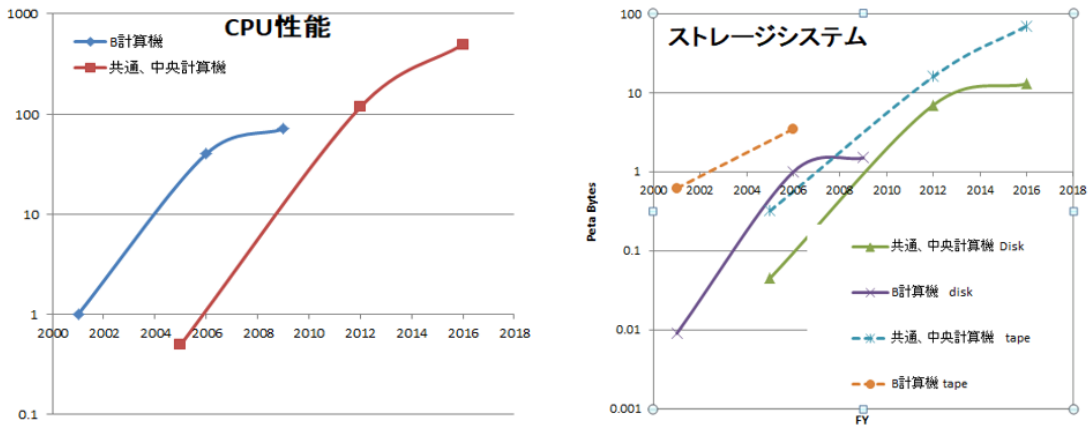


Figure 4　KEK における中央計算機システムの性能の変遷

2.3. JAXA JSS2 システム (東京都調布市)

2.3.1. システムの分類

種類：数値計算
データサイズ：大
ファイル数：大
保存期間：中
I/O 割合：大

2.3.2. システムの目的

スーパーコンピュータと共に発展してきた数値流体力学(CFD)は，理論・実験に並ぶ第三の手法としてさまざまな分野で利用されている．また，スーパーコンピュータの利用分野も地球観測データの処理や AI など広がっている．2014 年 10 月から段階的に稼働開始した JAXA スーパーコンピュータシステム(JSS2)では，このようなニーズに対し，プリポスト処理を含めた計算機環境を提供することを目的とする．

- 航空宇宙分野の国際競争力を強化する数値シミュレーション実行基盤
- 大規模データ解析基盤
- 新たなニーズを受け入れる研究開発基盤

2.3.3. システム構成の概要

JSS2 の構成を Figure 5 に，各システムの主要諸元を Table 5 に示す．JSS2 はヘテロジーニアスな構成であり，数値計算を実行する SORA-MA (富士通 FX100 3240 ノード，理論演算性能 3.49 PFlops，FX100 単体では 32 コアで理論演算性能 1 TFlops，主記憶 32 GB)，ISV を含めたプリポスト処理を行うための x86 アーキテクチャをベースとした GPU を持つ SORA-PP，大きなメモリ空間が必要となるアプリを実行するための大メモリシステム(SORA-LM)，そしてログインシステム(SORA-LI)から構成される．JSS2 は Lustre ベースの FEFS(Fujitsu Exabyte File System)ファイルシステムを採用し(以下，Lustre)，SORA-MA，SORA-PP，SORA-LM，SORA-LI は 7 PB から成るディスク領域を共有する．JSS2 より以前は，計算結果をポスト処理するためにデータを移動する必要があったが，増大する計算規模を考えるとデータの移動を極力少なくし，かつ遠隔地からも ISV や OSS を使って遠隔並列に処理する必要性が増してきた．そのため，JSS2 では SORA-MA で計算して出力したデータを移動させることなく，SORA-PP から処理が可能となるようなシステム構成である．SORA-MA 内は Tofu インターコネクトを採用しおり，各システム間は InfiniBand (FDR)で結合されている．SORA-MA は 408 GB/s (=60 ノード×6.8 GB/s)，SORA-PP はノード辺り 6.8 GB/sec，SORA-LI はノード辺り 13.6 GB/sec でファイルサーバ(SORA-FS)と物理的に結合されている．Lustre OSS と OST 間は FC-HBA で結合され，最大 144 GB/sec の帯域を持つ．数値計算ジョブから見ると，SORA-MA のノードメモリ帯域は 1.5 PB/sec と非常に広いのに対して，Tofu インターコネクトは 108 TB/sec と 1/14 倍に狭くなり，Lustre-OST は

144 GB/sec と更に 1/750 倍(ノードメモリと比較すると 1/10^4 倍)も帯域が狭くなる．JSS2 は 2020 年 1 月時点で 70 PB のアーカイバ(J-SPACE)を持ち，SORA-LI と 10GbE イーサネットで結合されている．SORA-LI の 1 ノード辺り 2 本，J-SPACE は 8 本の 10GbE NIC で接続されている．以上のシステムはすべて JAXA 調布航空宇宙センターに設置されているが，筑波宇宙センター，相模原キャンパス，角田宇宙センターそれぞれに遠隔システムがあり，例えば筑波宇宙センターには SORA-TPP という小規模なプリポスト処理専用システムが設置されている．

本成果報告書執筆時点の 2020 年 12 月から次期スーパーコンピュータ JSS3(富士通 FX1000 5760 ノード，理論演算性能 19.4 PFlops)が稼働を開始した．本成果報告書で述べる内容は JSS2 にて取得したデータに基づくことを断っておく．

2.3.4. システムが扱うデータの概要(種類・性質)

JSS2 では，CFD に代表される数値シミュレーションでは定常解析(Reynolds averaged Navier-Stokes(RANS) 解析)と非定常解析(Large-eddy simulation(LES)，Direct numerical simulation(DNS))に分類される．取り扱うデータ規模が大きいのは後者であり，1 つの計算ケース辺り概ね 10~100TB のデータがバイナリ形式で Lustre ファイルに出力される．プログラムによっては領域分割されたデータを 1 つにまとめることもあるが，規模の大きな計算ほど分割された領域ごとに並列出力する．ポスト処理後は，暫く使わないデータはアーカイバに転送される．領域分割されたデータを 1 つのファイルにまとめない場合，1 プロセス辺り 10 領域(10 ファイル)で 10^3 プロセス，データを出力する総ステップ数が例えば 10^3 ステップとすると，(10 ファイル/プロセス)×(10^3 ステップ)×(10^3 プロセス)=10^7 ファイルとなる．後述するように，ファイル数が大きくなるとオーバーヘッドが非常に大きくなり取扱いが大変になることから，領域分割されたデータを 1 つまとめる，出力間隔を広げるなど現実的な対応をとっている場合が多い．計算実行時間の中に占める IO 時間はそれほど気にならないが，定量的にあまり評価したことがないため，気にしていないだけなのかもしれない．数値シミュレーションの場合，ロードモジュール，入力ファイル(計算格子，初期条件，計算設定ファイル)があれば問題になることはあまりない．ただし，システムが変わるとコンパイラの違いにより，計算結果がマシン誤差オーダで変わることがあるが，収束解が得られれば問題がないレベルである．データの保存期間は，いわゆる賞味期限と同じである．JAXA で進められる数値計算は，航空宇宙機の開発のために行われるエンジニアリングと流体科学といったサイエンスに分類され，前者は航空宇宙機の開発期間やスパコンのリプレースと同じぐらいの概ね 5 年程度であろう．スパコンがリプレースされればより精度のよい計算が可能となるからである．一方，後者はもう少し長く，10～15 年程度である．

2.3.5. システム規模の変遷

Table 6 に直近 10 年のシステムの変遷を示す．2009 年 4 月に稼働した 120 TFlops の JSS は 1 PB の共有ディスクと 10 PB のアーカイバから構成されていた．2014 年 10 月から稼働した次の JSS2(3.49

PFlops)では，計算機の性能向上に伴って非定常解析が増えた．そのため，導入当初は共有ファイルシステムが 5 PB であったが，空きスペースが枯渇したため，2017 年 4 月に 7 PB へ増設した．同様に，アーカイバも 2017 年 7 月には 40 PB へ増強し，2020 年 1 月には JSS3 の導入に先立って 70 PB のシステムシステムに移行した．非定常解は益々増加しディスクスペースへの要求が増えることは必至であったため，2020 年 12 月から稼働した JSS3 は理論演算性能が 5.6 倍になった一方で，共有ファイルシステムは HDD が 8 倍の 40 PB，NVMe の SSD が 10 PB と合計 10 倍と増加した．

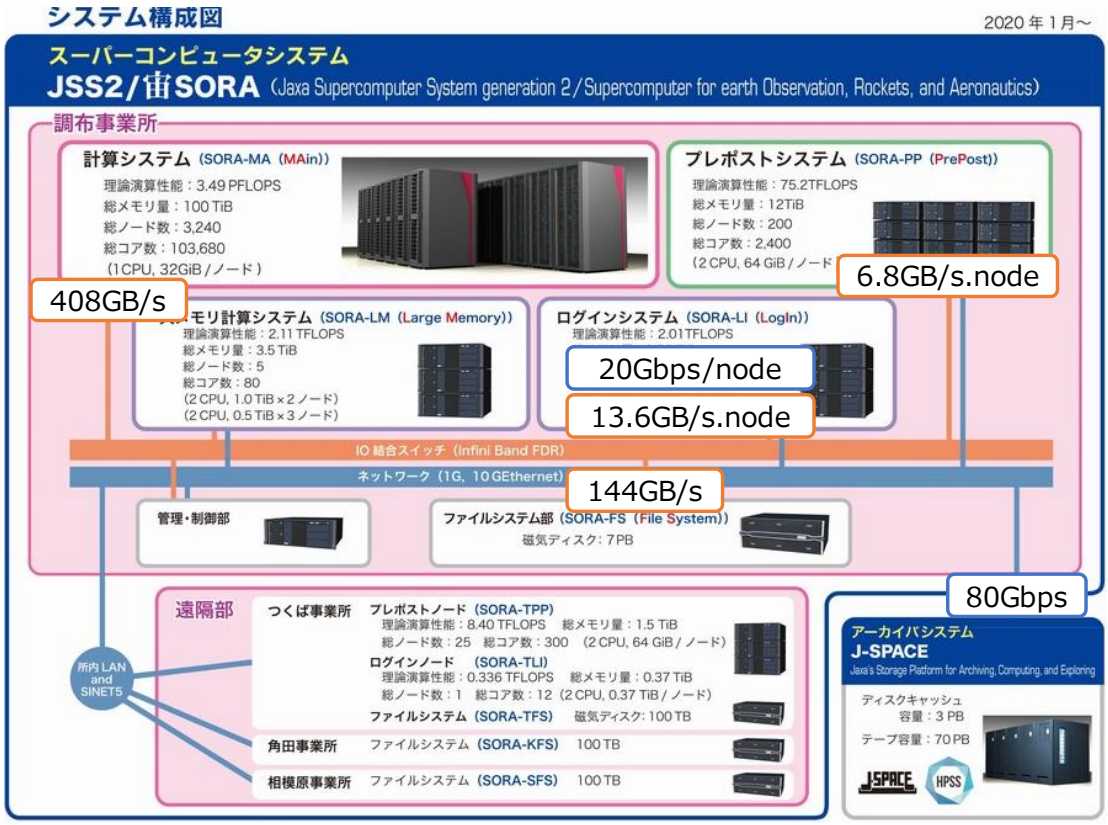


Figure 5　JSS2 システム構成図

Table 5 JSS2 主要諸元

	メイン計算システム (SORA-MA)	プリポストシステム (SORA-PP)	大メモリ計算システム (SORA-LM)	筑波プリポストシステム (SORA-TPP)
機種名	富士通 FX100	富士通 RX350 S8	富士通 RX350 S8	富士通 RX350 S8
ノード数	3,240	200	5	25
総理論演算性能	3.49 PFlops	75.2 TFlops	2.11 TFlops	8.40 TFlops
コア数/CPU	32	6	8	6
CPU 数/ノード	1	2	2	2
メモリ量/ノード	32 GB	64 GB	1024/512 GB	64 GB

Table 6 JAXA スーパーコンピュータシステムの変遷

	ピーク性能	ファイルシステム	アーカイバ
JSS (2009 年 4 月~)	120 TFlops (富士通 FX1 3,008 ノード)	1 PB	10 PB
JSS2 (2014 年 10 月~)	3.49 PFlops (富士通 FX100 3,240 ノード)	5 PB 7 PB (2017 年 4 月~)	20 PB 40 PB (2017 年 7 月~) 70 PB (2020 年 1 月~)
JSS3 (2020 年 12 月~)	19.4 PFlops (富士通 FX1000 5,760 ノード)	50 PB (HDD: 40 PB, SSD: 10 PB)	70PB

2.3.6. システム運用・設計上の課題

JSS2 で実行されるジョブは数値計算が大部分を占めるため，本報告書では数値計算に着目し，地球観測衛星データ処理については言及しない．数値計算におけるデータ処理のユースケースを Figure 6 に示す．SORA-MA で実行される計算ジョブは Luster OST に計算結果を書き込む．ポスト処理の際は，Lustre OST から SOR-PP へとデータを読み込む．Lustre ファイルシステムの容量には制約があることから，定期的に(もしくは計算中に)J-SPACE へと計算結果をアーカイブする．なお，Lustre OST から J-SPACE へは SORA-LI を経由してデータ転送する．

JSS2 に登録したユーザ数の年度履歴を Figure 7 に示す．概ね 700 人前後であるが，実際になにがしかの計算処理を行っているユーザ数は 500 ぐらいである．ユーザが利用したファイルシステムの利用量推移を Figure 8 に示す．ユーザは/home，/data，/ltmp の 3 つのディレクトリを利用する．/home は数

値計算のソースプログラムなどの小規模，少量ファイルを保存するために用いる．/home は J-SPACE に定期的にバックアップされ，UPS にも接続されていることから不意の電源障害にも対応可能である．/data は大規模，大量な計算結果が対象である．後に述べるように，JSS2 稼働当初は J-SPACE に定期的にバックアップを行っていたが，近年はデータ容量，ファイル数の増大により一部の領域しかバックアップできていない．/ltmp は 60 日で自動削除される 1 次保存データ用であり，バックアップは行っていない．Figure 8 から観察されるように，2015 年 5 月以降利用量は増加し始めるが，2017 年で当時の空き容量が枯渇したため，ユーザには不要なデータの削除をお願いしたため減少に転じている．その後また増加し始め，2019 年 7 月には空き容量が再度枯渇して不要データの削除を行い，現在に至る．ファイル容量の推移は線形近似で概ね近似することができ，増加量は 1 日辺り 2.4 TB である．数値計算のユーザからするとディスクはあればあるだけ使うため，2.4 TB/日という増加量は 7 PB というシステムの上限値に強く影響されていると考えられる．次に，J-SPACE(アーカイバ)の使用量履歴(データ容量，ファイル数)を Figure 9 に示す．J-SPACE では二重コピーをしている．OST と同様に，JSS2 運用期間中に容量が逼迫したため 20PB から 40PB へと増設し，さらに 2020 年 1 月には次期システム(JSS3)稼働より少し早めに 70 PB へと換装した．2020 年現在，利用量の合計は 20 PB である．(コピーを考えると 40 PB) 20 PB の内訳は，約 12 PB が数値計算，約 2 PB が地球観測衛星データ，約 6 PB がファイルしシステムのバックアップである．データ容量はほぼ線形で，一日辺り 10 TB(年間 3.7 PB)で増加している．一方，ファイル数は時々突発的に増加する傾向が見られるが，敢えて線形で近似するならば 1 日辺り 0.1M 個で増加している．Lustre ファイルシステムと同様に，J-SPACE もシステム容量に影響を受けるのか，計算機の演算性能に影響を受けるのか，それとも別の因果があるのかは定かではないが，JAXA の次期システム(JSS3)や他拠点での利用量の推移と比較してみたいところである．

上記で紹介した JSS2 における実績を基に，10 年後の想定される課題について議論する．

■ データ容量

ファイルシステムに関しては，JSS2 の 2.4 TB/日という増加量から試算すると，10 年後はたかだか 14 PB である．しかし，上述したように，7 PB というシステム上限値に強く影響された値であることは容易に理解できる．次期システム JSS3 では 50 PB(HDD: 40 PB, NVMe: 10 PB)に増加した．NVMe は一時的な入出力で利用されると考えると，データ保存用の HDD 領域は理論演算性能(5.6 倍)と同等の割合で増やした．JSS3 のさらに後継(JSS4)では，同様に理論演算性能に比例して増加すると考え，例えば理論演算性能が 6 倍になると仮定すると 240 PB という容量になる．

一方，アーカイバについては JSS2 の利用実績を参考にすると 10 年後に 60 PB(二重コピーを考慮して 120 PB)となる．ファイル数に関しては 500M 個になると想定され，ファイル数の増大に関わるオーバーヘッドが大きくなることは必至である．

このように増大すると考えられるデータを保存するためのシステムが，電力，設置場所，予算を含めて考えた場合に導入可能なのかは課題である．また，下記のような課題も挙げられる．

■ バックアップ

JSS2 稼働当初，/home, /data は数週間～数か月以上かけて定期的に J-SPACE へバックアッ

プをしていた．バックアップは find で更新されたファイルのリストを採取してコピーし，ファイルサイズに応じて tar と hsi，もしくは hsi のみを利用してコピーする．しかし，Figure 8 に示すように/data は線形的に増加し続け，近年はバックアップが追いつけない状況であり，JSS2 では概ね 1 PB が定期的なバックアップの限界のようであった．JSS3 では 50 PB のディスク領域を持つため，既存のやり方を踏襲するだけでは不可能な量である．現時点でも運用ポリシの見直しが求められる．

■ アーカイバ換装に伴うデータ移行作業

上述の通り，2019 年 9 月に新 J-SPACE を導入し，2020 年 1 月から運用を開始した．この間，旧 J-SPACE(LTO4 テープ)から新 J-SPACE(3592 テープ)に 1.3 PB のデータを移行した．移行作業は JSS2 運用中に実施し，約 80 日(平均転送速度 190 MB/sec)を要した．Figure 9 に示すように一日辺り 10 TB で増加してことから，次のアーカイバの換装時点ではかなりの量であることは必至で，現実的な日数で移行できるかどうかは不明である．

次に，JSS2 での実績から見えている運用・利用上の課題を紹介する．

■ 不要データの削除

スパコンの性能が増大してより高精度な数値計算が可能となれば，古い結果は陳腐化する．そのため，数値計算結果には賞味期限がある．JAXA で進められる数値計算は，航空宇宙機の開発のために行われるエンジニアリングから，流体科学といったサイエンスに分けられる．前者は航空宇宙機の開発期間やスパコンのリプレースと同じぐらいの賞味期限であり，概ね 5 年程度であろう．一方，後者はもう少し長く，10～15 年程度である．一旦アーカイブしたデータは忘れられがちなので，データの賞味期限を適切に管理する機能が求められる．

■ ファイルシステム関連

本来，Lustre は分散ファイルシステムであるため，分散ファイル形式の方が高い IO 性能が得られるはずだが，JSS2 ではそうではなかった．また，大規模な非定常計算を行うとファイル数が膨大になり，ls といったファイル操作にかかるオーバーヘッドが大きい．膨大なファイル数を操作する仕組みが求められる．

■ アーカイバの UI

アーカイバの UI は htar，his のみであり，必ずしも使いやすいとは言えない．また，アーカイブ時はキャッシュの容量を意識する必要があり，ファイル数が増えるとより重要になる．rsync や lsyncd のような UI で，キャッシュリソースをジョブスケジューラのように管理する仕組みが欲しいところである．

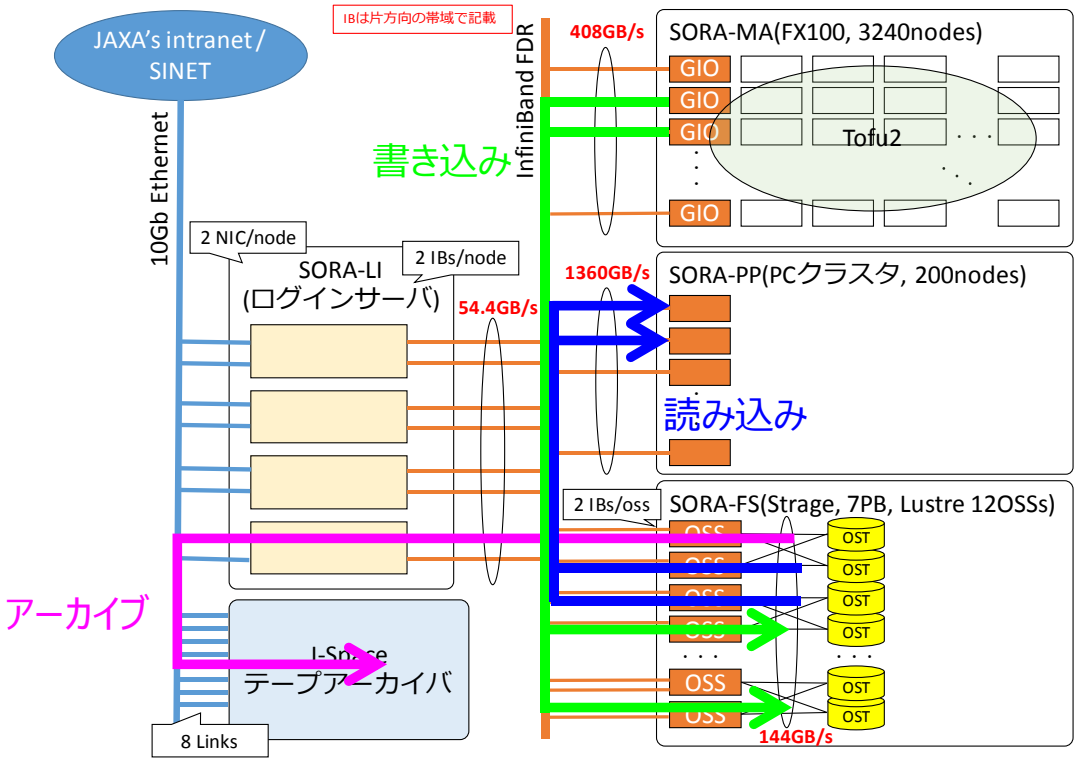


Figure 6 数値解析におけるデータ処理ユースケース

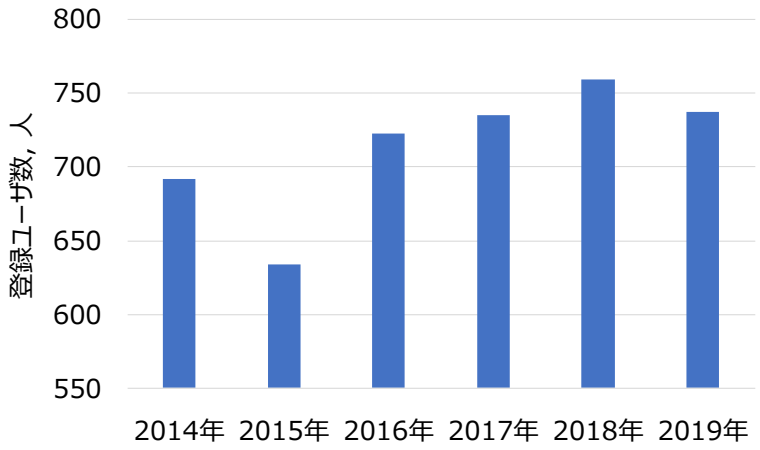


Figure 7 ユーザ数の推移

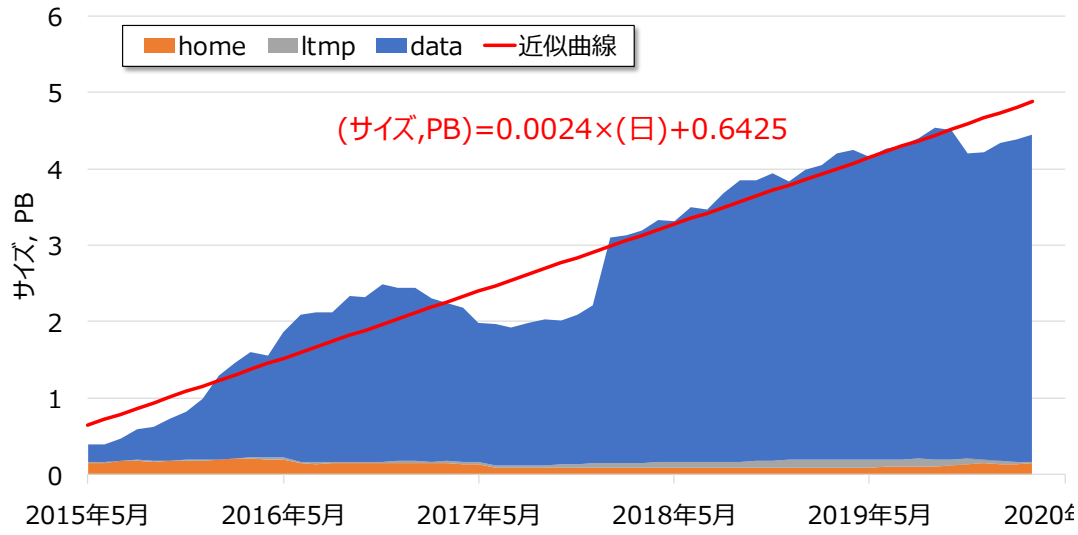


Figure 8 Lustre ファイルシステム利用量推移.

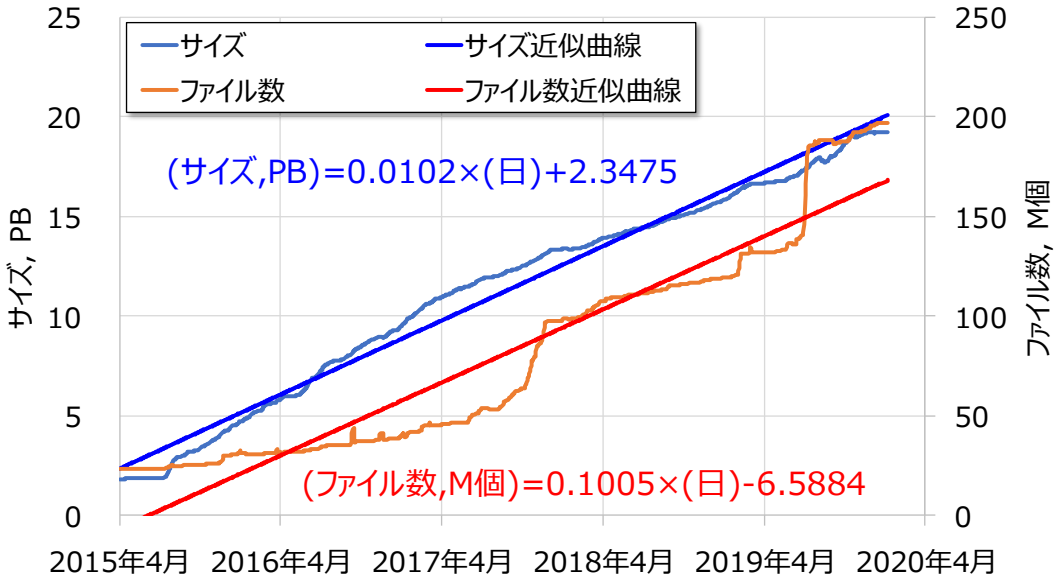


Figure 9 アーカイバ使用量推移.

2.4. 統計数理研究所・統計科学技術センタースーパーコンピュータシステム（東京都立川市）

2.4.1. システムの分類

種類：国民性調査，公的データ，医療カルテ，IoT-POS，SNS
データサイズ：小～中
ファイル数：小～大
保存期間：大
I/O 割合：小～大

2.4.2. ISM 統計科学スーパーコンピュータシステムの概要

少し冗長になるが，統計数理研究所の計算機システムの導入ポリシを知らせるために，仕様書（平成 25 年 10 月 21 日の前システムのものであるが，基本的方針は現システムでも同様である．）の前書きの一部を示す．

統計的データ解析の中には，大規模な線形演算の占める割合が非常に多いものがあり，ベクトル演算機構を付加することにより計算速度の大幅な向上が期待できる．このため，平成元年の初代統計科学計算機システムでは，内蔵アレイプロセッサ付きの M682H を，平成 6 年の第 2 代統計科学計算機システムではベクトル演算機構を有する S-3600/120 を高速・大規模計算を行うために導入した．平成 11 年の統計科学スーパーコンピュータシステムでは擬似ベクトル機構を有する SR8000 をこの目的のために導入した．

平成 16 年の統計科学スーパーコンピュータシステムでは，統計科学用ベクトル計算機として，SX-6(12CPU)を導入している．同時に導入した分散共有記憶型計算機 Altix3700 は，ベクトル化は困難であるが並列に計算することにより高速化が図れる手法が統計科学の研究の中心となっているために導入している．

ギッブス・サンプラー，非線形フィルタリング，ブートストラップ法等，計算統計学の分野の手法を研究し，計算可能な形にするためには，いわゆる並列計算機と呼ばれる計算機が不可欠である．このために，計算統計学支援システムとして平成 8 年 1 月に分散記憶型並列計算機（IBM RS/6000 SP）を，平成 12 年 1 月には共有記憶型並列計算機（SGI Origin2000）を，平成 18 年 1 月には分散記憶型並列計算機（HP XC4000）を導入した．さらに，平成 13 年 1 月にはパソコンクラスタであるブートストラップシステム（SGI1200（Pentium III（800MHz），512MB メモリー，物理乱数ボード）100 台）を導入した．

昨今，ビッグデータやデータ中心科学の重要性が指摘されている．研究所がこの潮流の中心的な研究機関として研究を推進していく必要がある．このような計算を行うためには，共有記憶型と分散記憶型の計算機が共に必要である．大規模主記憶の共有が可能であることにより初めて計算可能な課題

も多い．しかしながら，計算量が莫大となるために，個々の CPU に付属する主記憶が比較的少なくても計算は可能であるが，高い並列度が必要な課題も存在するためには分散記憶型並列計算機が適していると考える．

統計科学のソサエティだけでなく統計解析を必要とする研究者等に，クラウド環境で研究所の成果を利用できるようにし，成果を社会に還元できるようなシステムであることも重要である．平成 24 年度補正予算で措置された経費を用いて共有記憶型計算モデルを計算するためのシステムとしてデータ同化スーパーコンピュータシステムを，クラウド利用に適したシステムとして共用クラウド計算システムの導入手続きをすでに開始している．

仕様書前書きに書いた「データ同化スーパーコンピュータシステム」，「共用クラウド計算システム」をあわせて，統計数理研究所のスーパーコンピュータ環境は，平成 26 年度から平成 29 年度までは 3 システム体制であった．データ同化スーパーコンピュータシステム（愛称「A」），統計科学スーパーコンピュータシステム（愛称「I」），　共用クラウド計算システム（愛称「C」）という愛称をつけていた．赤池弘次先生の業績である赤池情報量規準（AIC）を想起する名称になっている．しかし，統数研がつけたものはなく，公募の結果である．詳細は，下記を参照して欲しい．

https://www.ism.ac.jp/computer_system/jpn/outline/outline.html

データ同化スーパーコンピュータシステムは 2018 年 3 月で運用を停止したが，それまでは，京コンピュータを中心として HPCI 事業にリソースを提供していた．共有クラウド計算システムは部分的なバージョンアップで継続利用してきているが，現在，新システムの導入準備中である．

統計科学スーパーコンピュータシステムの現有システムは 2018 年 10 月 1 日から稼働している．システムの概要は次の通りである．

統計科学スーパーコンピュータシステム

HPE SGI 8600 ・計算ノード（376 ノード） CPU：Intel Xeon Gold 6154 x 2 (18core 3.0GHz) Mem：384GB/Node ・GPU 計算ノード（8 ノード）CPU：Intel Xeon Gold 6154 x 2 / node (18core 3.0GHz) GPU ： NVIDIA P100 x 4 /node Mem：384GB/Node システム合計理論性能:：1.49PFlops 総 Core 数： 13824 core 総主記憶容量: 144TB ノード間接続: EDR InfiniBand

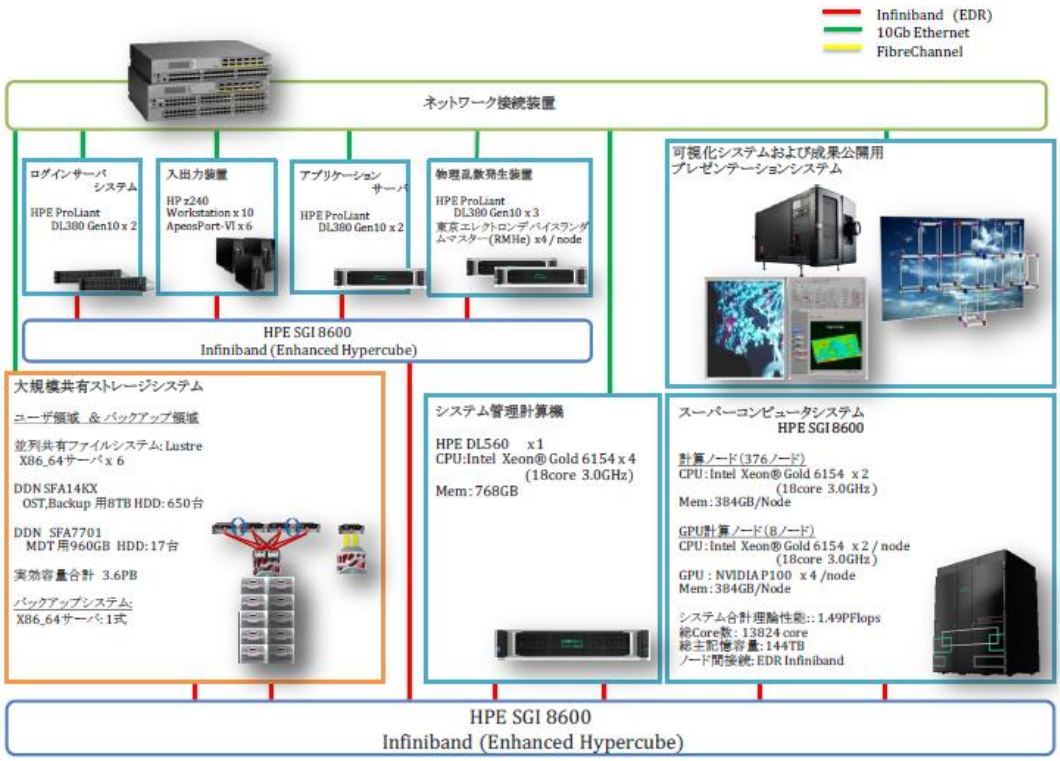


Figure 10 統計科学スーパーコンピュータシステム構成図

Table 7 統計科学スーパーコンピュータシステム　主要諸元

	CPU	主記憶	GPU	インターコネクト	ノード数
計算ノード1	Intel Xeon-Gold 6154 (3.0GHz/18-core/200 W)×2	DDR4、RDIMM 384GB 2666MHz	無し	4x EDR InfiniBand ×1	376
計算ノード2	Intel Xeon-Gold 6154 (3.0GHz/18-core/200 W) ×2	DDR4、RDIMM 384GB 2666MHz	Nvidia TeslaP100 x 4	4x EDR InfiniBand ×2	8

	製品	数量	備考
他システムとの接続	4x EDR InfiniBand	1	
所内との接続	システム管理計算機経由	-	
開発環境	Intel 開発環境 PGI 開発環境 R 言語 リモート乱数ライブラリ	-	
システムソフトウェア	Red Hat Linux PBS バッチ・システム Lustre ファイルシステム	-	

計算ノードは「5 次元ハイパーキューブ」トポロジーにより接続されています。

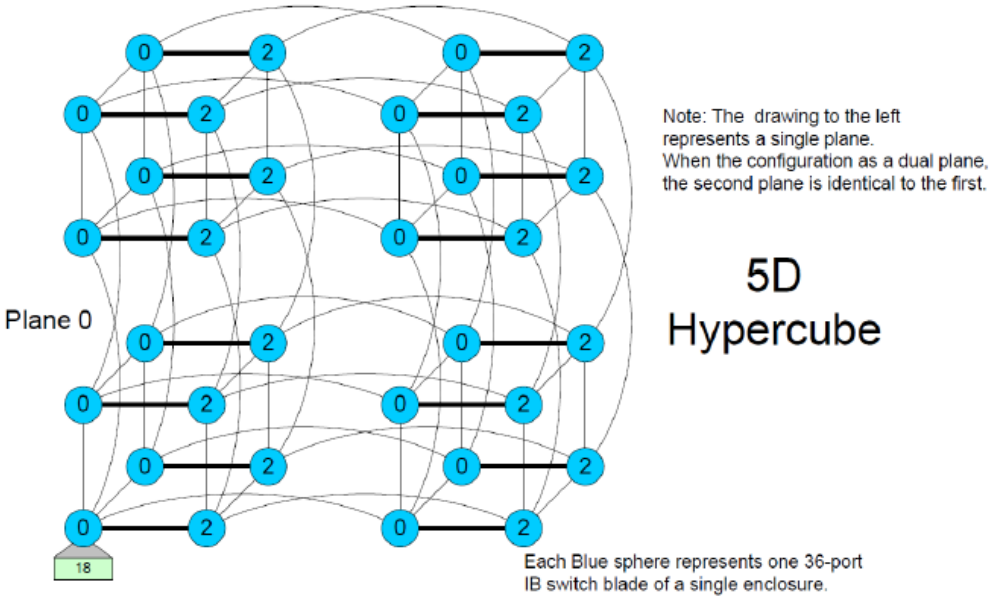


Figure 11 計算ノード間の接続の概念図

この他に特記すべきこととして物理乱数発生ボードを 4 枚装着したノードを 3 ノード用意していることがある．東京エレクトロンデバイス製で 1 枚の実効発生速度は約 600MB/秒であり，並列発生が可能となっている．シミュレーション（データ同化を含む）が統計科学スーパーコンピュータシステムの中心タスクであると考える．疑似乱数も利用されているが，田村は物理乱数を用いるべきだと考えている．

2.4.3. 統計科学スーパーコンピュータのディスクに関する情報

少し古いが 2019 年 11 月 15 日に取得した情報を示す．

実容量
/home 領域 1.8PB + /backup 領域 1.8PB
/home はユーザ様が実際に利用する領域
/backup は/home のバックアップ領域

形式
ラスター(Lustre)
/home と /backup で 2 つのラスター領域

現在利用されている容量
/home で現在利用されている容量は 460TB

バックアップの方法
/home から /backup へ rsync を実行．

ユーザあたりの制限
ユーザあたり 100TB

ディスクサーバのシステム構成
ディスクサーバは MDS 2 台, OSS 4 台 (いずれも DDN 1U サーバ).
各サーバのスペック
MDS Intel Xeon-Platinum 8168 (2.7GHz/24-core)CPUx2 基,192GBMem,
InfiniBand 4xEDR(100Gbps)
OSS Intel Xeon-Gold 6142(2.6GHz/16-core)CPUx2 基,192GBMem,
InfiniBand 4xEDR(100Gbps)
計算機本体とは InfiniBand 4xEDR(100Gbps) ネットワークで接続

システム領域

各サーバが 300GB の内蔵ディスクを 2 つ持っており RAID1 構成

その他(ディスクに関する情報)

8TB の SAS(ニアライン SAS)ディスクを　計 650 本用いて 2 つのラスター領域を構築

2.4.4. 統統計科学スーパーコンピュータの利用と統計分野での計算機利用

統計科学スーパーコンピュータシステムは最近は 5 年のリース契約で更新している．ピーク性能や主記憶容量は更新時に 5 倍から 10 倍になることを目指している．前システムまでは，ほぼ満たしているように思うが，今後は難しいように思う．磁気ディスク容量は更新時に 2 倍なるようにはしている．この容量は，利用最終時点で，磁気ディスクの利用率が 50%程度に納まることを目指した容量である．

統計数理研究所ではスパコンを設置し，利用している．また，共同利用で外部の研究者にも利用を許可している．本報告書の読者が「統計」での計算機利用をどのように想像されているかは不明であるが，統計数理研究所の研究者の計算機利用は統計科学の研究者の中ではかなり特異であると考える．国内外を問わず，多くの統計科学の研究者はパソコンで計算できる程度のデータ解析しか行っていないように考える．統数研の研究者が対象としているデータも，天文学，気象学，地震学などに関係しているものである．これらのデータに関係した，データ解析，データ同化，シミュレーションを行っている．このため，生データよりも，計算結果のデータの方が多いように考える．

MPI を用いた並列計算を行っているが，1 CPU あたりの計算規模を大きくした方が良いために，主記憶を大きくするような調達となっている．UV2000 のような共有記憶型計算機の方が望ましいが，製造されていないために，分散型で 1 ノードあたりの主記憶容量を大きくしている．

ユーザ数の最近の数字は把握していないが，所内外をあわせて登録ユーザは 100 程度である．実際の利用者は 50 ユーザ程度ではあるが，常にキューの待ちがあることから考えると稼働率は 100%である．CPU の稼働率で見ると 60 から 70%である．所内のユーザと所外のユーザで利用時間はほぼ同程度あるが，ジョブ数は所内の方が少ない．所内ユーザの方が大規模ジョブを実行している．

2.4.5. 統計分野での計算機利用

ビッグデータあるいは DX というワードを聞くことが多いように，データ集積，データ活用が注目されている．POS データを集めている機関は多いと思うが，そのシステムは公開されていない．また，データの保存期間も公表されていない．本 WG で講演いただいた筑波大佐藤教授の話では，2 年程度というものであったと考える．公的機関が行っている統計調査においては，おそらく法律で保存期間，保存方法も決められているのだろうが，計算機システムの詳細は公開されていない．しかし，集計された統計データは e-stat から利用可能である．

このようなあやふやなことを書いても意味が無いよう思う，しかし，日本における現状はこの程度である．デジタル庁に期待したいものである．e-stat に未だに，テキストリーダブルでない pdf の報告書が公開されているが，速やかに，やめて欲しい．一部の地方自治体が，オープンデータとして，かなりの情報を，テキストリーダブルで公開していることも付け加えておく．

2.5. 大阪大学サイバーメディアセンター 大規模計算機システム （大阪府茨木市）

2.5.1. システムの分類

種類：大学計算機センタ
データサイズ：大
ファイル数：大
保存期間：大
I/O 割合：大

2.5.2. システムの目的

大阪大学サイバーメディアセンター(大阪府茨木市)(以下，「CMC」)では，本 WG 活動時点において，学術研究・教育に伴う計算および情報処理をおこなえる全国共同利用施設として，全国の大学・研究機関の研究者，および，最近では産業利用枠として一部民間企業の研究者に対して，ベクトル型スーパーコンピュータ SX-ACE，スカラ型スーパーコンピュータシステム OCTOPUS (Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer)，大規模可視化対応 PC クラスタ VCC(PC Cluster for large-scale Visualization)を提供している．これらのスーパーコンピュータシステムは，文部科学大臣の認定を受け北海道大学，東北大学，東京大学，東京工業大学，名古屋大学，京都大学，九州大学と連携して推進している，学際大規模情報基盤利用・共同研究拠点(JHPCN: Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructure)，および，「京」と全国の大学や研究機関に設置されたスーパーコンピュータやストレージを高速ネットワーク(SINET-5)で結び，多様なユーザニースに応える革新的な協働計算環境基盤 HPCI（High Performance Computing Infrastructure）の一拠点を形成している．

以下では，本 WG 活動時点におけるシステム概要ついて概説する．

2.5.3. システム構成の概要

CMC の概要を以下に示す（Figure 12）．本システムは大阪府茨木市に在する CMC の IT コア棟サーバ室に設置されている．

CMC の提供するスーパーコンピューティング環境は，上述した通り， ベクトル型スーパーコンピュータ SX-ACE (423TFlops)，スカラ型スーパーコンピュータシステム OCTOPUS(1.463PFlops)および VCC を中核とし，利用者のデータを収容するストレージ，利用者からのログインを受け付けるログインノード，利用者がコンパイル，プログラム開発を行うフロントエンドノードから構成される．その他，利用者の可視化要求に応えるべく，複数の計算機と複数のモニタ群を連動させ，あたかも一台の表示装置のように扱うことができる大型立体表示システム(Figure 13)もある．これら CMC のスーパーコンピューティング環境を構成するコンポーネントは，10Gbps ネットワークをベースとする Supercomputer Network を経由して相互に接続されている．ここで Supercomputer Network へのスーパーコンピュータシステム群からの接続は，

複数のリンクから構成されることを補足しておく.

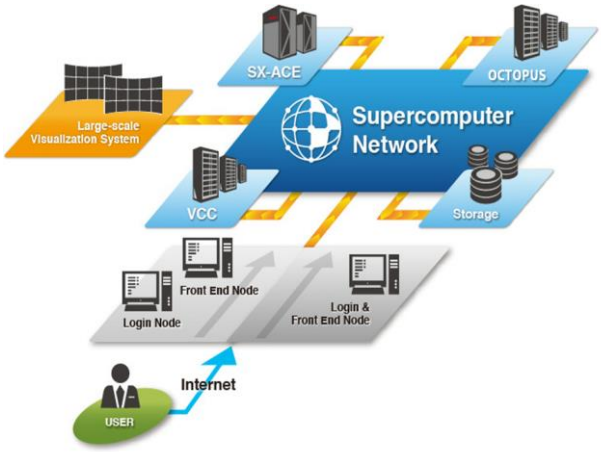


Figure 12 CMC のスーパーコンピューティング環境の概要.

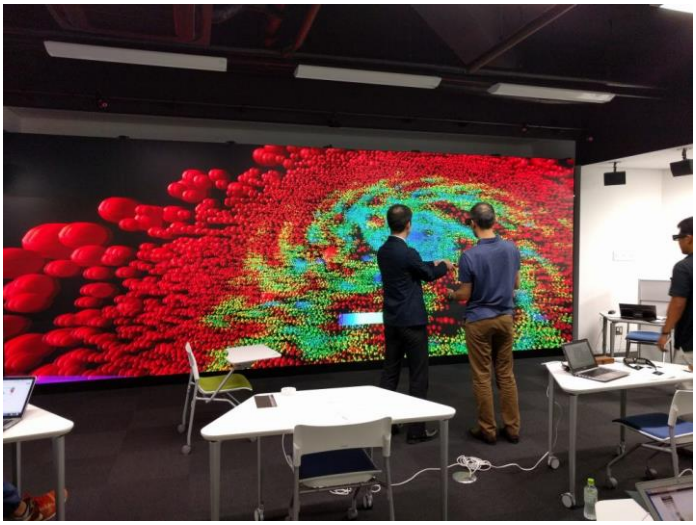


Figure 13 24 面立体表示システムの概要.

ここで，本 WG での関心事項となるストレージ環境についてすこし詳細に説明しておく．CMC のスーパーコンピューティングシステムの調達は，それぞれ異なる調達で行われる．すなわち，CMC では基本的にスーパーコンピューティングシステムの一部としてストレージを導入しており，スーパーコンピューティングシステムに独立した形でストレージを調達していない．これは大阪大学が国の機関（正確には，指定国立大学法人）であるため，調達活動は公平性のもと実施されなければならないことに一部起因している．例えば，それぞれの調達活動の結果の各スーパーコンピュータシステムの導入業者が異なるという状況も発生しえる．その際に起こりうるデータ移行時，運用時，障害時の業者間連携などの想定外の状況への対応可否・困難性を考慮しなければならない．CMC では，このような問題点だけでなく，調達活動

の負荷の大きさ，調達規模によるスケールなどを考慮して，ストレージを独立して調達するということは行なっていないのが現状である．

Table 8 に CMC のスーパーコンピューティングシステムが保有するストレージシステムの概要をまとめる．Table 8 に示すように，CMC では現状 5PB のストレージ容量を有している．SX-ACE のファイルシステムは，NEC の開発する並列ファイルシステム NEC Scalable Technology File System (ScateFS)，OCTOPUS には Lustre ファイルシステムが導入されている．なお，VCC はいわゆる補正予算で導入されたスーパーコンピュータシステムであり，計算ノード数を優先した調達活動の結果，既設の SX-ACE システムの NEC ScateFS をマウントして利用できるように構築されている．さらに，SX-ACE の ScateFS でマウントされたデータを OCTOPUS からアクセスできるように，ScateFS プロトコルを NFS プロトコルに変換してデータを export する NFS サーバを設置している．同様に，OCTOPUS 上にも Lustre プロトコルを NFS プロトコルに変換して export する NFS サーバを設置し，SX-ACE システムから OCTOPUS 上のデータにアクセスできるよう相互連携させている．これにより，ベクトル型スーパーコンピュータ SX-ACE を利用する利用者，およびスカラ型スーパーコンピュータ OCTOPUS あるいは VCC を利用する利用者，およびベクトルおよびスカラ型スーパーコンピュータをともに利用する利用者のデータ移動要求に応える形を取っている．

Table 8　CMC のストレージシステム．

システム	ファイルシステム	容量
SX-ACE	NEC Scalable Technology File System (ScateFS)	2 PB
OCTOPUS	DDN Lustre	3 PB
VCC	N/A (SX-ACE の ScateFS をマウント)	N/A

2.5.4. システムが扱うデータの概要（種類・性質）

CMC が扱うデータの種類を一言で種類・性質を分類・整理することは非常に難しい．これは，上述したように，CMC で提供するシステムは，全国の大学・研究機関（一部民間企業にも開放）の研究者による学術研究を対象としている．また，CMC としてもより多様な計算ニーズを収容できるスーパーコンピュータシステムの構築・提供・運用を通じて，我が国の学術研究に貢献することをミッションとしており，CMC のスーパーコンピュータシステムを利用する研究者の扱うデータの種類・性質は多様である．そのため，ここでは，CMC のスーパーコンピュータシステムを利用する利用者の視点から，システムが扱うデータの概要をまとめたい．

CMC では，全国共同利用施設として，全国の研究者を対象とした計算機サービスを提供してきた．とりわけ，CMC では長らくベクトル型スーパーコンピュータを中核とした計算機サービスの提供をおこなってきている．ベクトル型スーパーコンピュータは，その特性上，同じような計算を繰り返し行う計算，すなわち，SIMD 型の計算，さらにいえば，気象，地震，水理などの大規模科学シミュレーションに高性能を発揮する．それゆえ，CMC での利用者の多くは，こういった科学シミュレーションであったり，数値計算

を行っている傾向がある．こういった利用者による計算処理では，詳細な分析が必要ではあるが，入力データよりも出力データが大規模化・大容量化する傾向がみられる．CMC では，利用者に無制限にストレージを提供するのではなく，基本となる容量を超えてストレージ容量が必要な場合には 1TB 単位での利用者負担金（SX-ACE 10000 円/年　OCTOPUS 3000 円/年）が必要となるが，ベクトル型スーパーコンピュータ SX-ACE 利用者のなかには「計算を行いたいが，ストレージ容量がないので計算をできない．ストレージを購入するよりは，利用負担金を計算に回したい．」というような利用者の声がある．このことは，CMC の SX-ACE システムでは，利用者の計算を妨げる要因が計算資源容量ではなく，ストレージ資源容量にあることを意味している．なお，CMC では上述の通り，5PB の容量がある．ディスクストレージの提供は，各利用者に公平性の観点から，利用者に対して一律のディスク課金を行なわざるを得ない．そのため上述したような利用者のディスク容量を満たすためには，運用上の負荷なども考慮すると，当該利用者に追加で購入いただく，あるいは，すべての利用者に公平に一律ディスク容量の上限を増加させるの 2 択となるが，利用者の中にはデフォルトで割り当てられるディスク容量で十分な方もおり，判断が難しい現状がある．

次に，OCTOPUS 導入後の利用者のデータ特性について考える．OCTOPUS は 2017 年 12 月より稼働する，汎用 CPU(Xeon Skylake)，GPU(Skylake + P100)，メニーコア(Xeon Phi)，大容量主記憶搭載マシン(6TB)から構成されるスーパーコンピュータである．当該システムは，プロセッサ，アクセラレータ技術の進展による計算アーキテクチャの多様化，および，それに伴う利用者の計算ニーズの多様化を考慮して導入されたシステムである．今日では，様々な科学技術分野において，計測技術の発展，プロセッサ・アクセラレータ技術の発展による計算性能の向上によって取り扱うべきデータ容量がますます増大化傾向にある．加えて，AI(artificial intelligence)，ML (machine learning)，DL (deep learning)に代表される高性能データ分析 (HPDA: High Performance Data Analysis) 分野での計算ニーズが急拡大している．そうしたことから，OCTOPUS の利用者が取り扱うデータの種類と特性は，上述の SX-ACE とは若干異なる特性が見られつつある．例えば，運用中に 1 ディレクトリで作成可能なファイル，ディレクトリ数が数百万を超過した，とのアラートが見られたり，利用者から「ディレクトリ内にファイルが増えるとアクセス速度がさがるのか？」などといった質問がよせられることが度々ある．

加えて，近年では，医療情報などのプライバシなどを考慮しないといけない機密性・秘匿性を考慮しなければないデータをスーパーコンピュータ上で利活用したいという利用者ニーズが急速に拡大している．例えば，大阪大学歯学部附属病院の推進する S2DH(Social Smart Dental Hospital)では，医療情報と AI の利活用により，サイバ空間での新たな医療のあり方を模索しているが，大量の医療情報を AI で処理する計算基盤・データ基盤が必要不可欠となっている．今後のスーパーコンピュータシステムの整備には，こうしたデータセキュリティの要求されるデータの取り扱いについても考慮しなければならない．

また，近年では，OCTOPUS，SX-ACE を利用する研究者らの研究形態もまた急速に変容しつつある．とりわけ，研究者の研究活動の広域化・グローバル化は顕著にあらわれつつある．各国立大学や研究機関において，産学共創，国際連携が研究評価基準になっており，今日までのように，大学設置の各計算機センタの提供するスーパーコンピューティングシステムやストレージ単独だけで研究開発が行われるケースだけでなく，他計算機センタやクラウド計算資源やストレージとの相互連動性を考慮しなけ

ればならない状況がうまれつつある．例えば，米国研究機関で生成されたデータを，国内の研究機関・計算機センタで処理し，欧州の研究所の研究者とそのデータと計算結果を共有したい，というケースがあり，そうした場合にどのようにすればいいのか？あるいはそういったケースを考慮したシステム整備を期待する，といた声が寄せられている．

2.5.5. システム規模の変遷

CMC では，スーパーコンピューティングシステムの調達は，本報告書執筆時点において，大きくは 2 系統の調達活動によっておこなわれる．現行システムでは，SX-ACE とOCTOPUS の調達の 2 系統となる．それぞれの調達は，いずれも 5 年レンタル，あるいはリースを想定した契約に基づき行われる前提となっている．規模としては，前者は後者の約 10 倍弱の規模となっている．本報告書執筆時点では，これらの調達はおよそ 2.5 年ズレており，新システムが導入されると，次期システムの調達が行われる状況となっており，利用者に常に最新のシステムをご利用いただけるようになっている．しかし，5 年のリースあるいはレンタル契約は，後継機の調達にともなう市場調査の結果に基づき，しばしば延長されることがおおく，今後もプロセッサ，アクセラレータなどの状況に影響をうけることが予想される．

システム規模としては，計算性能側面とストレージ側面があるので，一概に数字で示すことはむずかしいが，現行の SX-ACE（423TFlops）に対して，現在 SX-ACE の後継として進められている調達では，10PFlops-20PFlops のピーク演算性能を想定して進められている．ストレージ容量については，SX-ACE では 2PB であるが，後継機では，15PB-25PB を想定した調達が本報告書執筆時点において行われている．また，これまで HDD ベースでの並列ファイルシステムの提供を行なっていたが，後継機では一部高速データアクセス領域として 1 PB 程度の SSD ベースのファイルシステムを提供することも考慮されている．さらに，クラウドや学内外の研究機関とのデータ移動性を考慮して，S３対応ストレージの導入も検討されている．

上述したように，CMC での調達は，大学法人という特性上，公平に行われなければならない．そのため，既存システムの導入業者が，その後に行われる調達の導入業者となることは定まっていない．それゆえに，インクリメンタルなシステム導入は CMC のような大学法人では難しい要素がある．例えば，ある調達の契約機関が終了し，他の調達に影響を及ぼしてしまうような状況を回避しなければならない事情があり，調達システムの機能拡張という形で調達をおこなうことは契約期間後の運用保守の視点からは困難な現状がある．それゆえに，大学計算機センタの調達では，契約期間の間に変容しうる計算ニーズ，データニーズを先読みしつつ検討しておかなければならない事情がある．

2.5.6. システム運用・設計上の課題

今後のシステム運用・設計上の課題としては，

1. システムおよぶストレージへの費用投資割合

   スーパーコンピュータシステムの調達においては，計算ノード，ストレージ，インターコネクトにどの程度予算を投資するかが重要な課題となる．理論演算性能，計算ノード数を追求すれば，ストレージやインターコネクトの予算を縮小しなければならないし，ストレージをリッチにすれば計

算ノード，インターコネクトにしわ寄せがいく．こうしたことから，スーパーコンピュータシステムの契約期間における利用者の計算ニーズやデータニーズにあわせて適切な設計を行うことが重要であるが，現状では，利用者アンケートによる利用者の声に基づいた設計を行っているものの，ベストな設計となっているかどうかは評価が難しい．

2. 研究活動の広域化・グローバル化に伴うデータ移動機能

上述したように，今日では研究者が行う研究活動が広域化・グローバル化傾向にある．こうした際に，高速ネットワークを活用して，自由自在にデータを移動できなければならない．とりわけ，高性能データ分析分野 HPDA のニーズ拡大により，とりわけ各種センサや AI 処理したいデータをスーパーコンピュータに集約するといった機能性がもとめられている．また，海外の共同研究者とクラウドを通じたデータ共有をしたいなどの機能性もまた求められている．したがって，スーパーコンピュータ上だけでデータを扱うのではなく，国際的な研究基盤を担う一拠点としてスーパーコンピュータ，ストレージを設計していくことは重要となるであろう．

3. セキュアなデータを取り扱える設計

医療情報などの AI 利活用など，機密性・秘匿性が要求されるデータをスーパーコンピュータシステムで取り扱いという要望が急拡大している．どのように研究者のデータセキュリティ要望を満たし，高性能な計算サービスを提供できるかについて考慮したシステム・ストレージ設計は，ライフサイエンス分野等のセキュアなデータを取り扱う科学の発展にとって必要不可欠である．

4. 研究データ保存，reproducibility を確保したデータ管理

今日では，研究データの保存・管理に関する議論が急速に高まっている．スーパーコンピュータが生成する計算結果も例外ではなく，科学計測機器が生成するデータとともに，研究成果が再現性(reproducibility)を保つ形で保存・管理されなければならない．今後のシステムでは，こうした視点でも設計を考慮していく必要があるだろう．

2.6. 慶應大学センシング・デザインラボによるセンシングデータ解析（神奈川県 横浜市）

2.6.1. システムの分類

種類：多様なセンシングデータ，衛星データ
データサイズ：小―大
ファイル数：小―大
保存期間：小―大
I/O 割合：小―大

2.6.2. システムの目的

慶應義塾大学大学院システムデザイン・マネジメント研究科センシング・デザインラボでは，様々なセンサから得られるデータを活用し，多方面の社会の課題の解決を目指している．用いられるセンサは人工

衛星から身に着けられるサイズのものまで多岐に渡る．また，観察やインタビュー，アンケートといった人を介在してセンシングするデータも課題の解決には重要な要素として捉え，扱っている．近年は高機能で廉価な IoT デバイスが利用できるようになり，これらから得られる客観的なデータと実地のフィールドワークによる対話的なデータを組み合わせることで，分野・地域を超えて社会課題に取り組み新しい価値を生み出している．
以下ではその応用例を紹介する．

2.6.3. マレーシアにおけるアブラヤシ伐採・植樹の ICT による作業支援

従来目視と手作業で行われていたアブラヤシの伐採・植樹の作業をリモートセンシングによる地形データに基づいて行うことで効率化し，パーム油の収穫増を目指すプロジェクトである．リモートセンシングデータにより植樹計画を行い，作業時にはマルチ GNSS 信号を受信できる受信機を作業者が身に着けることで，計画に基づいて正確な位置で伐採・植樹することが出来る．

Figure 14　従来の伐採・植樹計画と作業の流れ

Figure 15　リモートセンシングと GNSS 位置情報による計画・作業支援の概念

2.6.4. インド・カンボジアにおける農家信用評価の最適化

インドやカンボジア郊外の貧困な農家では金融機関が融資を行うための十分な情報が揃わないため，正当な評価を受けづらく，結果，生産性が上がらないという状況があった．このプロジェクトでは，モバイル端末により農家の人員構成や耕作の状況を正確に集約するとともに，衛星による撮像・レーダー測定を用いて農地や農家の地理的な情報などを組み合わせることで，金融機関が公正なデータに基づいて評価し，融資を行うことが出来るようにすることを目指している．

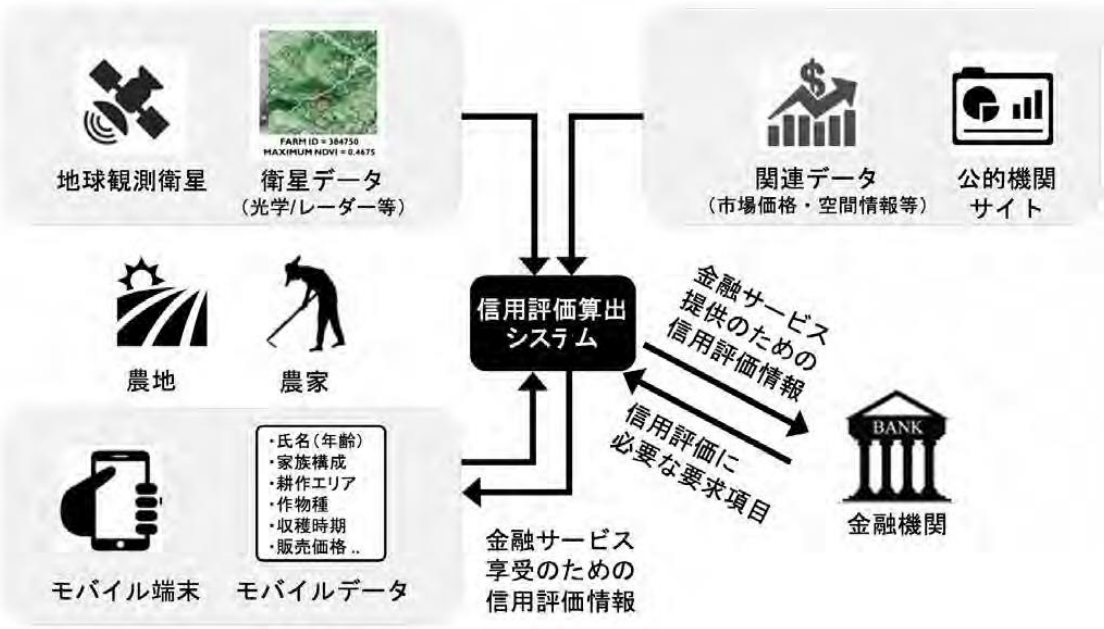


Figure 16　農家情報収集と信用評価計算の最適化の概念

2.6.5. 身体情報のモニタによるスポーツ選手のコンディション管理と子供のスポーツ振興

心拍をモニタし GNSS 信号を受信して位置情報をトラックできるセンサをトップアスリートが装着することで，運動量や速度，加速度といった練習中や試合中の選手の動きをデータ化することが出来る．このデータを解析して，選手のコンディションを把握し怪我を防止したり，最適なトレーニング方法やチーム戦略の立案などに活用し，怪我の低減や効果的なトレーニングへの効果を確認している．

Figure 17　選手が装着するセンサの例

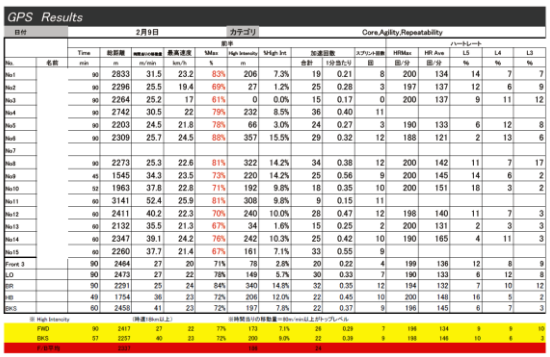


Figure 18　選手から取得されるデータの例

また，小学生や中学生といった育成年代の子供たちへのスポーツ振興にも同様の技術が生かされている．子供たちにセンサを装着してもらうことで，自分たちの動きを可視化して伝えられるため，効果的に指導を行うことができる．

Figure 19 育成年代への位置センサを用いた競技指導の例

2.6.6. システム運用・設計上の課題

センサ技術の有効活用においては，良質のデータを収集することが非常に重要である．その社会課題への有効な利用のために，データの適切な管理とポータビリティが確保されなければならない．
取得されたデータ（特に取得しづらいデータ）は可能な限りは保存されてきているが，プロジェクトの種類によって保存・破棄の扱いは異なる．例えばスポーツに関する取得データは人に帰属するため一定の解析利用の後は捨てられることが多い．これまでは実社会の課題に根差したデータ活用という観点を重視し，データの取得と解析利用に主眼が置かれてきたと言える．しかし，データの増大や価値のあるデータの蓄積の観点から，今後は統計的にデータを代表する部分のみを保存する，圧縮するなどのデータを保管するための技術を検討していく必要もあるだろう．

2.7. メンバ機関のシステム事例に学ぶ課題

いずれの機関のケースでも，増大するデータの保管と移動については懸案が示された．データ解析後すぐに削除して良いデータは少ないようである．新規に生成されるデータを扱う一方で，一度解析された過去のデータについても一定の期間（場合によっては 20－30 年以上の長期に渡って）保管し続けなければならないようなシステムが多く，システム移行への対応を含む長期運用は大きな問題である．また，特に実験系の研究機関では，過去のデータは保管するだけでなく研究グループ自身による再解析や他者による解析の可能性を作ることが求められており，利用頻度の異なる複数のデータに対して長期保管と解析での利用効率の両立が課題となっている．

また，システムの調達計画の難しさも共通の課題であることが分かった．リース契約を含む 4-5 年間の計算機計画を行う上で，システム設計に必要な情報であるユーザのニーズやプロジェクトのデータ生成率の予測が難しいことが議論された．予算の制約から必要量の計算資源を一度に調達できないことなどの現場の悩みも見られた．また，大学・研究機関等の公的法人において複数の調達により構成される

システムが必ずしもインクリメンタルにならないことから，効率的で安定的な運用の難しさも共通の問題であることが分かった．

3. 事例に基づくデータ解析システム構成のモデル化

本章では，前章で紹介した WG メンバ機関の現行システムのうち，国立天文台，JAXA，KEK の 3 機関のデータ解析・運用システムの事例を取り上げ，それらのシステムに必要とされるデータ処理に応じたシステム設計の考え方を議論する．この議論では，特にそれぞれのシステムが処理・保管・移動するデータ量に着目して各事例のシステムを単純化（以下モデル化と呼ぶ）し，運用実績あるいは実績から推測されるデータ量に基づいて必要とされる計算機リソースを議論する．

3.1. 国立天文台事例のシステムモデル

3.1.1. データ解析処理とデータの動き

本システムの大枠の構造を理解するために，システム内で行われる作業とそれに応じたデータの動きを 2 章で示したシステム外観図（Figure 1）に即して説明する．大規模なデータ解析の作業は 1 年に 1 ないし 2 回程度の頻度で行われる．

まずデータ解析の準備として，解析すべき未処理の生データ(～数十 TB)を本システム内のストレージに配置する．生データは本システムとは独立の生データアーカイブシステムで管理されているので，そこからダウンロードを行う．データは一旦，図中④または⑤に運ばれ保存され，かつデータのメタ情報がデータ解析専用のデータベースに登録される．ここで直接解析系の作業場所にダウンロードしないのは，アーカイブ系のストレージの方が解析系よりも生データアーカイブシステムとのネットワーク的な距離が近いためである．アーカイブ系システムはデータ保存と配布に特化されたシステムである．そのストレージは容量重視で比較的低速な XFS・NFS からなり，ウェブやデータベースをホストするための最小限の CPU しか持ち合わせないため，そこで直接解析作業を行うことはできない．次に，アーカイブ系ストレージから実際に解析を行う各解析系クラスタの作業用ファイルシステム（①，②，③のいずれか）に生データをコピーする．解析系クラスタが複数のシステムに分離しているのは導入時期や管理体制等の内部的な事情によるところが大きい．解析系とアーカイブ系の間のネットワーク帯域(1Gbps)があまり広くないこともあり，この作業にはおよそ 2-3 週間を要する．このデータ移動作業において現システム構成は最適とは言えない．

データ配置が整った後，①，②，③のいずれかの解析系クラスタによって 2.1 節で述べたデータ処理が施される．ほとんどの解析処理はバッチジョブを通して分散処理されるが，解析はいくつかの段階に分かれており，それぞれの間で同期が必要である．同期ポイントでの結果確認や次段階の解析の実施の多くは手作業に頼っており，この作業の負担，操作ミス，オーバーヘッドの軽減は課題のひとつである．

解析結果は，検出器ごとの較正・測定結果，検出器・複数回の撮像を横断した合成画像から得られる測定結果など，多種類のファイルとして出力されるが，一般に出力は生データの 10 倍程度のサイズ

（～数百 TB）になる．これらは同じく①，②，③の比較的高速なファイルシステム（Lustre または Spectrum Scale）に書き込まれる．この解析作業の主要部分には順調な場合で約 2 か月間を要する．

同環境下で解析結果の品質確認を行い問題がないことが確認されれば，全結果出力ファイルは④または⑤のアーカイブ系ストレージに再び転送され，ユーザの利用に付すため比較的長期(1～数年)に保存される．

結果出力のうちで，天体カタログと呼ばれる天体の測定結果については，ユーザが検索しやすいようにデータベースに展開され，SQL を介した検索サービスに付される．また，天体カタログのデータベース投入前の元ファイル（フラットファイル）や画像ファイルについてはウェブサーバ（HTTP）およびプロキシサーバを経由して外部のユーザへ配布される．上述の通り解析は定期的に実施されるが，多くの場合データ解析のためのソフトウェアや較正用の参照カタログが更新されるため，その都度，それまでに取得された全データを解析し直すことが多い．そのため解析により作成されたデータプロダクトには世代が発生し，本システム内では複数の世代を管理し続ける必要がある．現在のシステムは全世代のプロダクトをアクティブにサービスに付すために十分なディスク容量とサーバ数を持ち合わせないため，２世代以上前のプロダクトは LTO テープに保存するか圧縮してサービスから外した場所で保管するなどしている．それでも増大するデータに対応するために定期的にストレージを追加調達することが必須となっている．

解析結果がアーカイブ系に保存されれば原理的には解析系に置かれたデータは削除が可能となる．ただし，実際には世代を超えたデータを横断して追解析を施したり品質評価を行ったりするケースが少なくなく，解析系にも 1-2 年程度の期間は出力データを保持しなければならない．古い解析結果は本来生データから再現できるはずだが，実際にはソフトウェア動作のアーキテクチャや各種ライブラリへの依存性などの再現性の問題があり，加えて 2 か月以上の大規模な処理の再現は現実的に難しいため，一度科学活動に使われた全世代のプロダクトを保全しなければならない点に留意が必要である．

3.1.2. データ処理とデータ量の関係

Table 9 は上で述べたデータ解析処理の前後でデータ量（サイズと個数）がどのように変動するかを示している．表中の「1 回あたり処理済みデータサイズ」とは，1 年に 1-2 回の解析作業を行う度に生成される 1 回あたりの出力データのサイズであり，「処理済みデータ総サイズ」はそれらを積算した本システムが保管しなければならない全世代を通した処理済みデータサイズの総量を表している．また，2015 年から 2020 年までの数値は実際のシステム運用による実績値で，2021 年と 2024 年はそれらに基づいた予想値である．

この表より， 2024 年度中には 1 回のデータ解析作業で入力する未処理データは 100TB 程度，出力データは 1PB 規模になり，システムが保持しなければならないプロダクトのデータ量は 90 億個以上のファイルからなる 4PB 以上のサイズに及ぶことが分かる．なお，2016 年から 2017 年にかけて処理済みデータサイズが減少しているのは，この時点で解析ソフトウェアの一部の出力データが自動的に可逆圧縮されるようになったためである．以下では，このデータ量の推移をもとに，システムに求められる資源の規模を議論する．

年度	未処理データサイズ(TB)	未処理データ数(個)	1 回あたり処理済みデータサイズ(TB)	処理済みデータ総サイズ(TB)	処理済みデータ総数(億個)
2015	6	3,440	66	66	記録なし
2016	16	8,465	430	430	記録なし
2017	28	15,243	278	650	記録なし
2018	32	17,379	420	990	記録なし
2019	39	21,347	550	1,435	0.5
2020	46	25,254	700	2,000	1.1
2021(*)	60(*)		850(*)	2,700(*)	1.9(*)
2024(*)	~100(*)		1,000(*)	4,300(*)	92.8(*)

Table 9　HSC データ解析システム内のデータ解析処理前後のデータ量とその年次変化.
(*)は予想値

3.1.3. データ量に対して要求される CPU の規模

約 6 年間のシステム運用の実績に基づき，入力データサイズに対して解析作業で実際に使われた CPU コア数の関係を調べた．ここで解析に使われた CPU コア数とは，あくまでも各回の解析作業の時点で利用可能な状態にあったリソースであり，実際のデータ量や計算時間に応じて動的に割り当てられたものではない．しかし，各解析作業に際しては，データの配置や結果配布の準備を除くメインの解析処理をほぼ 2 か月程度で完了できるように計算機リソースを拡張してきており，実際に毎回おおむね 2 か月強の作業で解析の大部分を完了させてきた経緯がある．したがって，使われた CPU コア数は，おおむねデータ量に対して必要とされる CPU コア数に近いと考えて良い．

Figure 20 には解析に入力される未処理データサイズに対して処理に使われた CPU コア数の関係を示した．図から，必要とした CPU コア数には未処理データのサイズと良い相関があることが分かる．

Figure 21 では入力データサイズに加えて，出力データサイズと使われた CPU コア数の関係も併せて示した．出力データサイズの系列が 300-400TB 付近で逆行しているのは上で述べた 2016 年から 2017 年での出力データ量の逆転に対応している．入力データに対する相関と同様に，出力データに対しても大きな傾向としては使われた CPU コア数と正の相関が見られる．ただし，新規の解析作業に際してはソ

フトウェア・測定方法等のアルゴリズムの更新，データ格納形式の変更などがあるため，出力サイズは入力サイズに比べて単純な相関にはならない．

3.1.4. データ量に対して要求されるストレージサイズの規模

次に前節と同様，解析で扱ったデータサイズに対して実際に用いられているシステムのストレージサイズの関係を本システムの運用実績に基づいて調べた（Figure 22）．ここでは，横軸を入力データサイズとした場合（●）と各解析の出力（処理済み）データサイズとした場合（■）に対応するストレージサイズという２つの系列の情報を表示している．処理済みデータサイズの系列で 300-400TB 付近で折れ線が逆行しているのは，前節同様に可逆圧縮の導入による．

この図より，入力データサイズ，各解析出力サイズの双方ともに，必要とされるストレージサイズと良い相関があることが分かる．本システムの全世代のプロダクトを保持しようとした場合，すでに 5PB 程度のストレージでは安定的な運用が難しくなっていることが理解できる．本システムのデータ解析では，入力データサイズより出力データサイズが有意に卓越しているため，出力データの推移がより重要である．システムの設計・調達においては数年後のデータ規模を予測することでシステムのストレージサイズを検討する必要がある．

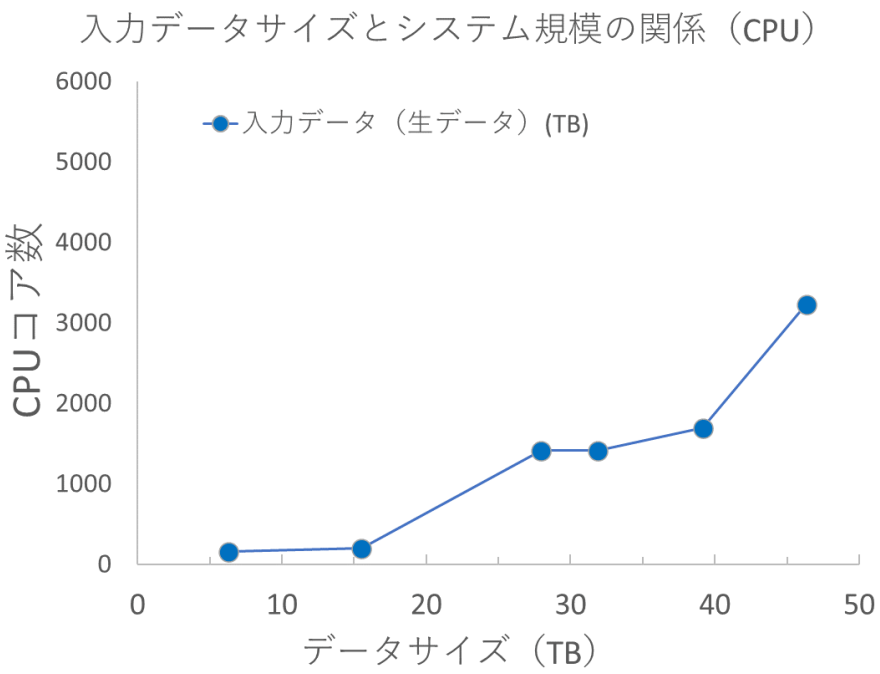


Figure 20　処理を行う入力未処理データサイズと必要 CPU コア数

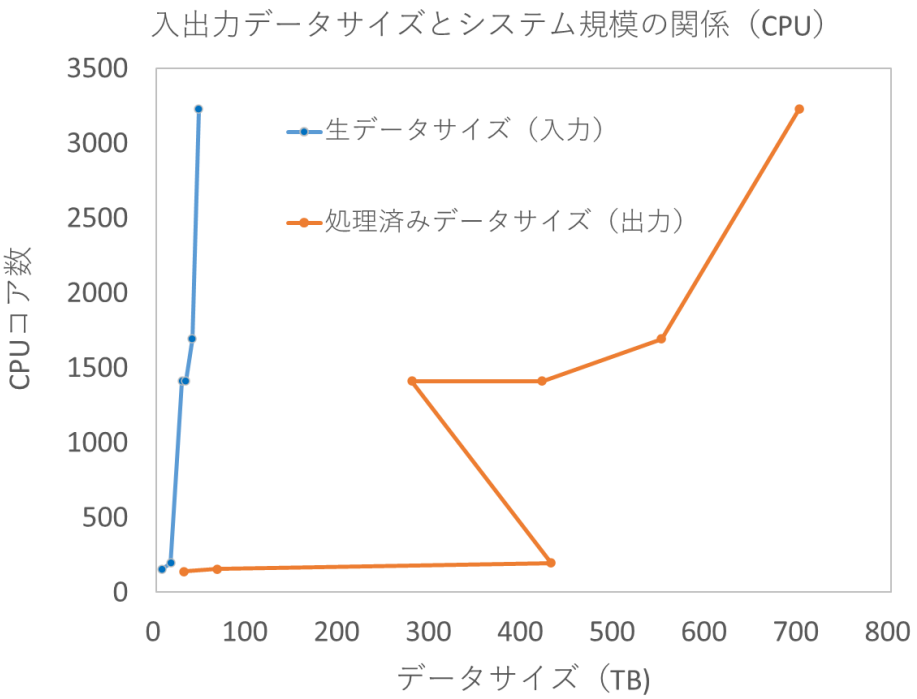


Figure 21　処理を行うデータサイズ（入力と処理済み出力）と必要 CPU コア数

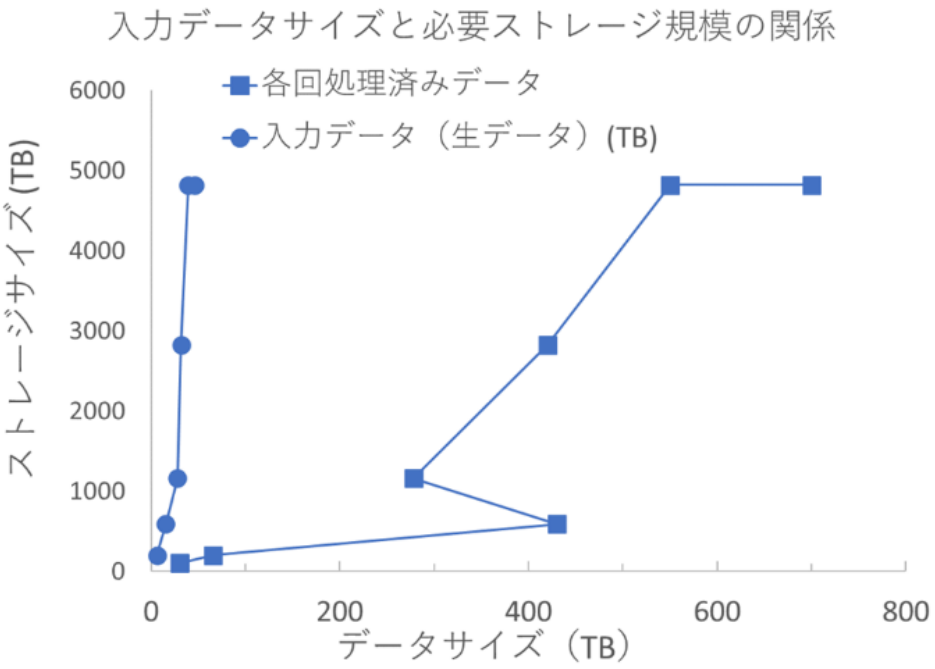


Figure 22　処理を行うデータサイズと必要ストレージサイズ

3.1.5. システムモデルの議論から導かれる考察

以上のように，本システムで各解析作業時に用いられた CPU コア数とストレージサイズは入力のデータサイズおよび出力データサイズと良い相関を示すことが分かった．CPU コア数については，入力データサイズと良い相関があり，出力データ量に対しても大きな傾向としては正の相関がある．システム設計においてはこれらを参考にし，データ解析の内容に応じて必要とされる計算資源の規模を検討すべきである．ストレージサイズも入力・出力データサイズともに良い相関がある．本システムにおいては特に出力データサイズの推移に基づいてストレージサイズを検討することが求められる．

本システムが扱うソフトウェアは比較的新しい天文学のデータ解析アルゴリズムを含んでおり，今後 5 年程度でソフトウェア面の進展が計算機資源に与える影響はさほど大きくないと予想されるので，今回の議論は 5 年程度の計算機調達計画には応用することができるだろう（もちろん今後の進展次第で想定以上の要求が出てくる可能性もある）．

具体的には以下のことに留意してシステム設計を行うべきと考えられる．

1) 5 年程度の先を見越したデータ量の経年予測を行うことが重要であり，それに基づいてシステムの計算資源の規模を検討し，調達計画を立案できると良い．
2) 運用方法を具体的に考慮し，大きなデータの移動を極力避ける方法を考える必要がある．
   特に解析システム，アーカイブシステムそれぞれのストレージは可能な部分を一体とし，データの移動にかかる負荷を最小限に出来ると良い．可能ならば高速にアクセスすべきデータを高速なディスク上に置くといった階層化が実現できると良い．
3) 大きなデータ移動がやむを得ず発生する部分を見出し，そのような箇所には転送のボトルネックが生じないよう広帯域のネットワークを導入すべきである．

一方，以下のような課題も見えた．これらはメンバ機関を通じて共通の課題であることも分かった．

1) 現代科学活動においてはデータ解析の手法の変更やプロジェクトの進捗等により計算負荷やデータレートの予想が外れることも多く，想定以上のデータ増大が発生することがあり得る．与えられた予算の範囲で一度の調達でまかなえるシステムの規模には限界があり，そのようなデータ増大に対応するためには運用中の断続的なシステム拡張を行わざるを得ない．このような拡張に対し，CPU・ストレージ共に可能な限り初期導入部分と一体で運用できるような拡張計画が望ましい．
2) 次期システム設計に十分に生かせるようなシステム運用記録が必ずしも十分には取られていない．
   特に個々のユーザ活動やデータの利用頻度等をシステム動作の最適化に利用できるレベルで追うことが望ましいが実際は難しい．

本システムと類似の処理を行うデータ解析システムにおいても，データ量の増加予測に基づいて必要とされるシステムの計算機資源の規模の同様の考察が可能であろう．

3.2. 高エネルギー加速器研究機構のシステムのモデル

KEKCC は，高エネルギー加速器研究機構の素粒子原子核研究所，物質構造学研究 所，加速器研究施設，共通基盤研究施設で共用して使われているために，一概にどのような計算やデータが標準であるとは言えない．しかしながら，KEKCC のデータ解析システムの大きな割合が Belle 実験及び J-PARC のハドロン実験で利用されているため，主な割合としては下記のようなモデルにより利用されているといえる．

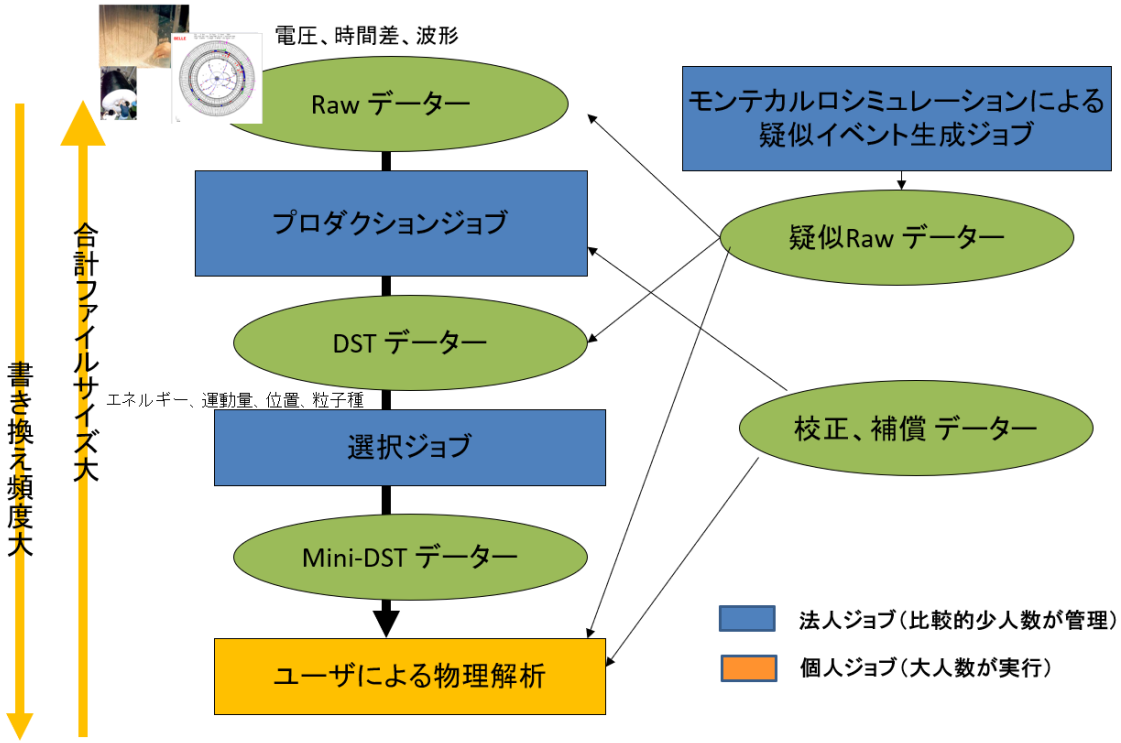


Figure 23 KEKCC で主に行われている高エネルギー実験の解析モデル

加速器で加速された粒子同士を衝突させた結果生じる各種の粒子は検出器により様々な情報が取得される．収集されたデータは時間情報，波形情報，波高情報などで，最終的にデジタル化されて，興味ある現象に特徴的なデータを含む衝突事象（イベント）から生じたデータのみを荒く選別されて蓄積される．収集されたデータは raw(生)データと呼ばれ，記録された時間情報，波形情報等から，生成された粒子の軌跡の位置情報，粒子種情報，エネルギー，運動量情報などの値に　多量の計算機パワーをつかって再現・再構成する．このプロセスを「プロダクション」と呼ぶ．プロダクションには，実験データを補正するために検出器や既知の物理現象をシミュレーションした結果が利用される．このシミュレーションは，入力データは少ないが，非常に時間がかかり出力データも多量である．プロダクションによって解析されイベントから発生した現象を再構成したデータを DST データ（Data Summary Tape：歴史的事情によりテープという名前がついている）と呼ぶ．　DST データには生データを含む場合と含まない場合

があるが，含む場合は，プロダクションのジョブの入出力データはほぼ同じで，非常に多量となる．生データへのアクセス（すなわち　プロダクションの回数自体は　解析がうまくいけば　データ当たり数回とそれほど多くないが，多量のデータがあるため，非常に時間がかかる．また，このプロダクションは実験グループ全体で管理して行われている．DST データから　興味の対象によってデータを選別し　その後，多くの研究者によって種々の物理的事象の解析が行われる．

大きな特徴として　Belle 等の大型実験においては機構内部の計算資源だけでデータ解析がおこなわれるのではなく，国際共同により　広く世界中の関係機関の計算機資源を利用してデータ解析がおこなわれることである．どの国のどの拠点がどの程度の計算機資源を出しあうかは，博士論文の数などから算出され，各機関に割り当てられる．従って，高エネルギー加速器研究機構としての実験解析は KEKCC だけで完結せず，世界的な実験解析のための分散計算機環境の一部として位置づけられる．これについては，本研究会の「クラウド時代のデータ共有技術 WB 成果報告書」の高エネルギー加速器研究機構事例に詳しく書かれているのでそちらを参照いただきたい．

3.2.1. データの流れ

KEKCC の実験データや，解析結果は，国際的な Grid 分散計算機環境により解析，分散保存される．このために KEKCC はファイルサーバ・計算サーバ毎に　10GbE　2　本が接続され，100GbE で SINET に接続されて外部とデータ流通を行うが，その部分を除いて，内部でのデータの流れを示したのが Figure 24 である．　前述のように KEKCC の主要なストレージは主に　GPFS 共有ファイルシステムによるホーム・グループディスク領域（17PB）と，GPFS のインターフェースをもつ階層型ファイルシステム（HSM）HPSS(100PB)からなる．　HSM はキャッシュとして用いられる GPFS-HPSS Interface (GHI) 8.5PB の GPFS サーバを通してアクセスされる．また，計算サーバには，それぞれ 1 TB の SSD を持ち，ノードでの一時記憶領域として利用される．そのほか，Grid 関係のサーバ，メールサーバ，Web サーバ，各種機能サーバを付帯しているが，それらの説明については割愛する．システム内の通信やデータ読み書きの速度については以下の仕様が指定されている（Figure 25）．

1) ホーム・グループディスク領域と計算サーバの間で合計 100GB/秒以上
2) ホーム・グループディスク領域，GHI ディスク領域 それぞれと計算サーバの 1 コネクションあたり 400MB/秒以上
3) HSM（HPSS）テープライブラリのドライブへの装填時間（ドライブが空いているとして）は最大 30 秒以内，平均 15 秒以内，装填後のロード時間は 15 秒以内，テープの読み出し速度は 150MB/s 以上．
4) HSM(HPSS)テープライブラリと，GHI ディスク領域への合計転送速度は，テープの読み出し・書き出し速度も含め 25GB/秒以上
5) GHI ディスク領域と計算サーバの合計転送速度は　60GB/秒以上

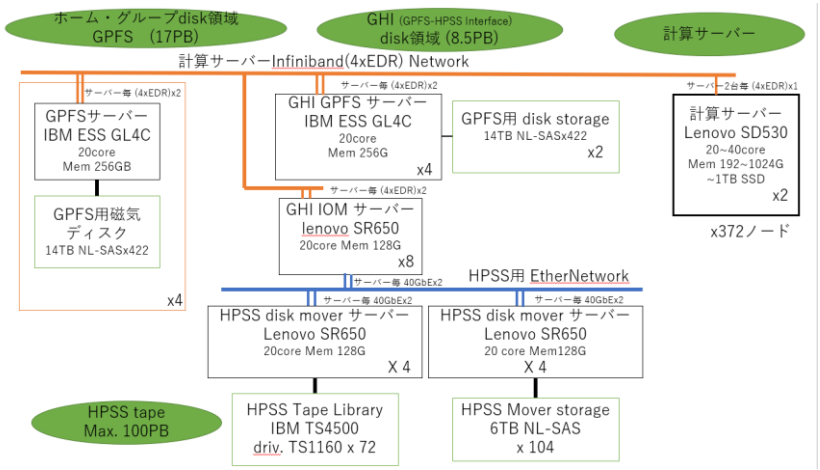


Figure 24　KEKCC システム内部でのデータの流れ

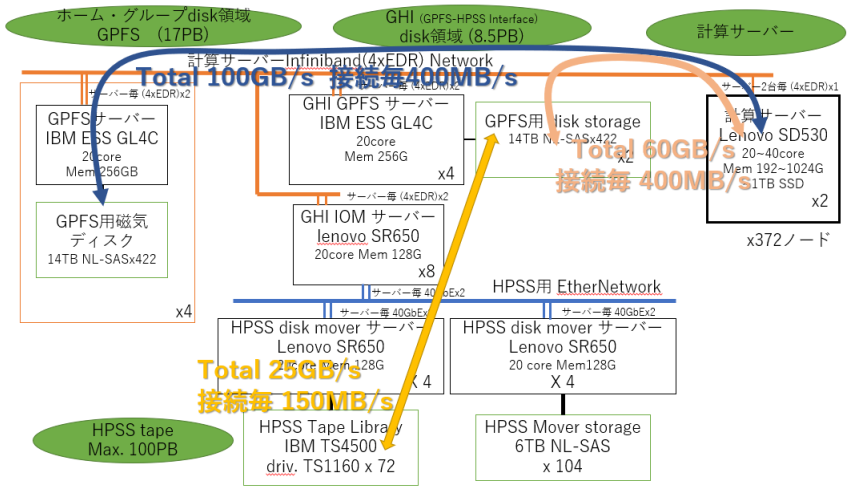


Figure 25　KEKCC システム内データ転送性能の要求仕様

これらの要求は，各実験の需要の合計値から算出されている．

Figure 26 は，2016 年システム（Disk 10+3(HSM)PB, Tape 70PB）における，毎年のストレージ使用量を示している．（Disk についてはグループ領域のみを加算している） HSM 以外の Disk 領域については 2 年目でほぼハードウエア容量上限に達し，Disk に書き込めないため HSM を利用してテープの利用が急激に伸びている．テープは後で消耗品として買い足すことができ，初期投資が少なくて済み，またアクセスしない限り電力消費が 0 であることから，KEK での調達を含めたモデルには適しているといえる．

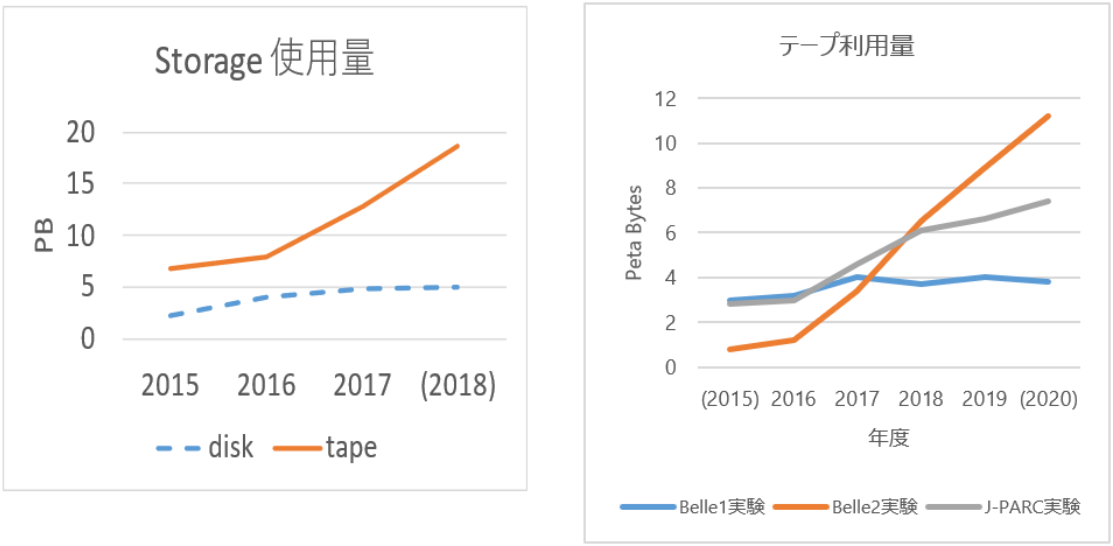


Figure 26　毎年のストレージ使用量

KEKCC では，JOB の実行状況はもちろん，各ネットワークの利用状況，HSM においては，テープライブラリ-ディスク領域間のデータ移動の頻度やテープライブラリにおけるドライブの利用量などの 現状や統計情報をモニタしている． Figure 27－Figure 32 は，画面のハードコピーで見にくいが過去 1 か月の状況を示している．これらモニタ結果から，概ね 繁忙期は 仕様の上限の 転送能力，テープドライブなどの性能を使い切っていることがわかるが，これは，それぞれの指標単体をとりだしたものであり，計 算サーバノードがもつ SSD，JOB のコントロールの仕方，ネットワーク接続の方法，ディスクサーバの配置，テープドライブなどコストがかかるどの部分を増強すれば全体としての性能・コストを最適化できるかについては，指標がない．最も典型的には，階層型ファイルシステムにおける GPFS ディスク階層の容量があり，もちろん容量があればあるほど使いやすくなるわけだが，どの容量がコストを含めてこのシステム全体を最適化できるのか不明であるし，各種相関をとったモニタがとれると何らかの最適化がなされるのではないかと考える．

Figure 27　過去 1 か月ジョブ実行量　（縦軸一番上目盛 14,000JOB 色はジョブの種類）

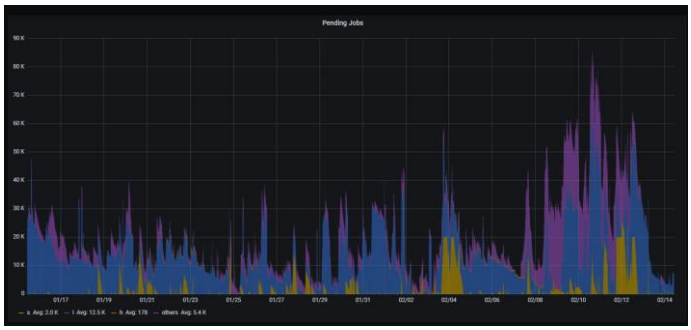


Figure 28　過去一カ月の JOB 待機数（縦軸一番上目盛 90,000JOB　色はジョブの種類）

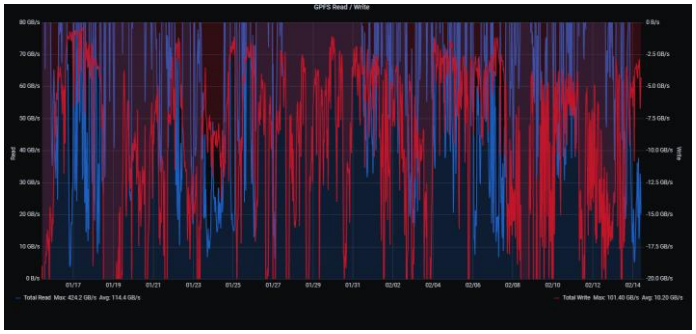


Figure 29　過去一カ月 GPFS 領域データ転送量（縦軸一番上目盛 80GB/s　赤：write　青：read）

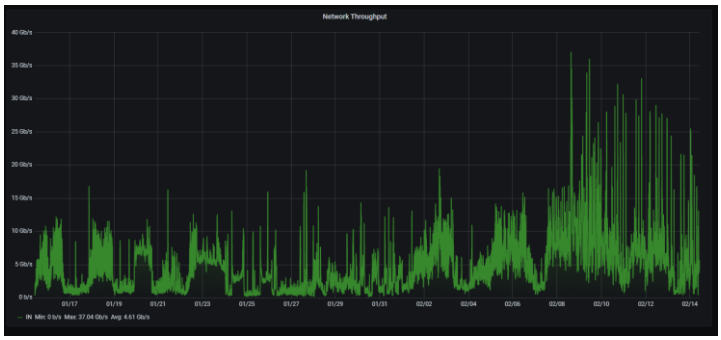


Figure 30　過去 1 か月外部との接続ネットワーク転送量（縦軸一番上目盛　40GB/s 5min 平均）

Figure 31　過去一カ月　InfiniBand データ転送量（縦軸一番上目盛　500GB/s）

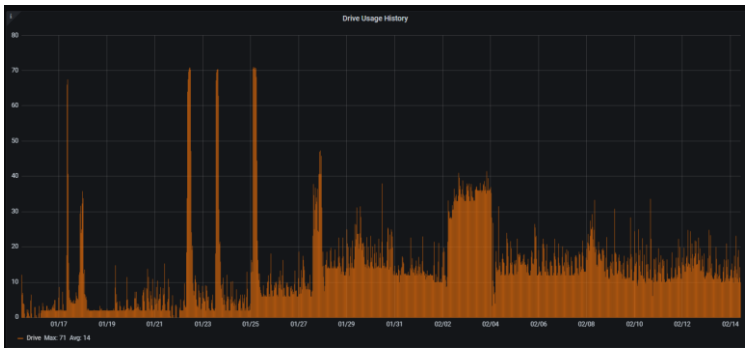


Figure 32 過去 1 か月　テープライブラリ　ドライブ同時占有数（縦軸一番上 80）

3.2.2. データ移行事例紹介

中央計算機においては，4 年～6 年のシステム更新に合わせて，そのすべてのデータを新システムに移行する必要がある．　2020 年のシステム移行においては，そのデータ量はテープ量において 30PB，ディスク量において約 10PB になり，現在および将来にわたりこのデータ移行は大きな課題となっている．事例として 2020 年に行われたシステム更新の際の　データ移行の実際を紹介する．

3.2.2.1.　テープデータ

今回の更新においては，テープライブラリシステムが更新前システムの上位互換となったため，テープ上のデータ自体は，テープ媒体を　更新前システムのライブラリ棚から，新システムのライブラリ棚へ移動するだけでよかった．（テープの総本数は約 4000 巻）　上位互換のシステムであればテープ媒体を移動することで移行の大部分ができることは，テープシステムを使う際の大きなメリットである．もちろん，階層型ファイルシステム（HSM）として新しいシステムとなるため，テープ上のデータのカタログ DB は新システムに移行する必要があったが，HSM システム自体は，同じ HPSS であることから，複雑な変換は不要で，量自体も全体のデータ量にくらべれば微々たるものであるため，特に問題にならない．階層型ファイルシステムとしての更新前ディスクキャッシュの容量は 3PB あった．必要に応じてテープからディスクへ on demand でデータがコピーされるため，本来的にはディスクキャッシュのデータは移行の必要がないが，システムサービスイン直後より効率的に利用を継続する観点からは，このキャシュについてもデータをコピーして移行することが望まれ，それを予定していた．　しかしながら，スケジュールの遅れ等から，これは結局断念した．

前述のように今回は，テープ全体システムが　上位互換であったため，スムーズにいったが，ベンダーロックの観点や，またこの製品自体が EOL になった場合には大きな問題となり，本システムの大きな課題である．

3.2.2.2.　磁気ディスクデータ

KEKCC の磁気ディスクは，前述したように利用者ごとに用意されたホームディスク，実験グループごとに領域を確保したグループディスク，階層型ファイルシステムの磁気ディスク領域（HPSS-GHI），Belle2

実験用に特別に用意された，実験データが最終的に蓄積されるまで一時的にバッファとして記録される Belle 実験一時領域がある．今回の移行において対象になった容量，ファイル数，移行期間，移行方法を表に示す．　データが利用できる期間を最小限にするため，運用を継続しながら一次移行をおこない，最終的に運用を止めて確定した内容により行う最終移行により一次移行期間に変更，追加されたファイルを移行した．また，ホーム＋グループ領域においては，6 台のサーバ(移行元と移行先合わせて 12 台)と，それぞれのサーバに 10GbE ネットワークを接続した移行専用のネットワークにより行った．移行専用のネットワークを利用し，移行量をコントロールすることにより，1 次移行においては，通常の運用をある程度続けながら完遂することが可能であった．Table 10 にそれぞれの領域の data 量，ファイル数，移行期間，移行方法を，Figure 33 にデータ移行のために占有した移行用ネットワークの構成を示した．

システム	移行対象 データ容量 (フィアル数)	移行期間	移行方法
磁気ディスク (ホーム)	約 80TB (約2.8億個)	(一次移行) 2020/7/11 – 2020/8/21 (最終移行) 2020/8/26 – 2020/8/28	rsyncを複数回実行 (checksumオプションも使用) グループと合計で最大120並列
磁気ディスク (グループ)	約 8.2 PB (約10億個)	(一次移行) 2020/7/11 – 2020/8/27 (最終移行) 2020/8/19 – 2020/8/31	rsyncを複数回実行(checksumオプションも使用) グループと合計で最大120並列
磁気ディスク (Belle実験一次領域)	約500TB (約86万個)	(一次移行) 2020/7/29 – 2020/7/30 (最終移行) 2020/8/30 – 2020/8/30	rsyncを複数回実行(checksumオプションも使用) 最大40並列
HPSS-GHI	約29PB (約6.1億個)	(disk-metadata) 2020/8/29 – 2020/9/2 (tape-data) 2020/8/27 – 2020/8/27	(disk) メタデータはGPFSの機能(AFM)にて移行 データはテープ移行後に順次読み出し処理 (tape) メタデータはデータベースを移行 データはテープを物理移動

※ファイル数にはハードリンク等のリンクファイルも含まれる。
※GHIディスクのメタデータ容量は、約4.1TB。

Table 10　各システムの data 量，ファイル数，移行期間，移行方法

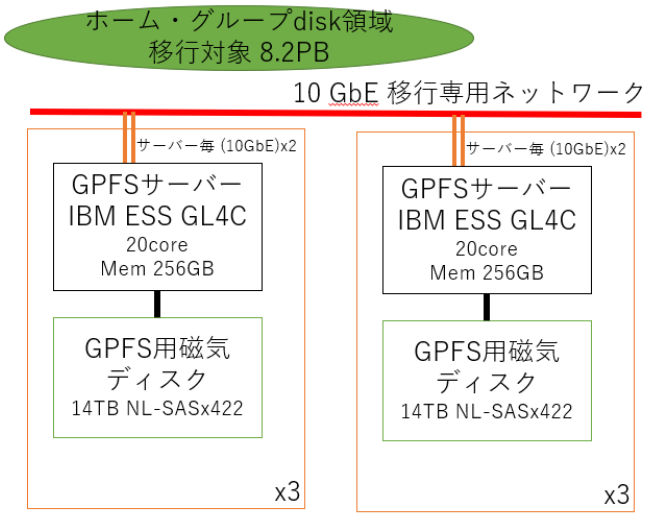


Figure 33　データ移行用ネットワークの構成

3.2.3. システム構築における課題

いままで述べてきたも含め，本システムの課題をまとめる．

課題 1: システム更新時における移行データの巨大化と，移行時間の増大．
課題 2: コストが限定されている状態でのシステムの仕様の最適化

- 階層化されたファイルシステムのどこに，どれだけの容量を用意するか
- 上記とネットワークを含めた転送路の性能や，各デバイスのアクセス速度性能
- ジョブの性格によるディスパッチ順やディスパッチ先の最適化
- どのようなデータを　どこに配置するか？　自動的に行う観点では，単に最近使ったデータは近くにあればよいというアルゴリズムしかない．特に転送コストが高い．全世界的な広域分散システムの場合には，どこにどの様にコピーを置けばよいかは非常に重要．

上記に関して

- スナップショット的な使用量のログはとれるが，それぞれの相関的なデータが取れない．
- それぞれの仕様のパラメータを振ってのシミュレーションができない．
- proactive なデータ配置/preemptive な資源確保が限定的にしかできない．

課題 3: 調達の都合上，使っただけ資源を用意するということが，テープや，限られた範囲のクラウド利用でしかできない．リースの初期には需要に対して供給が過大となり，終了期にはかなり不足となる．

課題 4: 需要を満たすだけの資源を購入するだけの予算の確保が困難であると同時に，実際に，数年後の計算資源需要を予測すること自体も難しくなっている（実験装置建設の予定，予算に強く影響を

受ける）
課題 5: KEK では保存時の運用コスト，省エネルギーの観点からテープ利用を今後も継続していきたいが，大規模なテープライブラリ，高性能なテープドライブ，それを扱うためのソフトウェアすべてについて，供給ベンダーが少なくなってきており選択肢が限られるとともに，ベンダーが開発や製造を中止した場合の先行きに不安がある．

3.3. JAXA JSS2 システムの運用実績にもとづくシステムモデル化

3.3.1. データ処理とデータの動き

JSS2 におけるデータの動き Figure 6 に示す．1.5 章で述べたように，SORA-MA から Luster OST に書き込まれたデータは，再度 Luster OST から SORA-PP へと読み込まれ，SORA-PP 上で可視化解析や統計処理といったポスト処理が行われる．Lustre ファイルシステムの容量には制約があることから，定期的に(もしくは計算中に)J-SPACE へと計算結果をアーカイブする．なお，Lustre OST から J-SPACE へは SORA-LI を経由してデータ転送する．

3.3.2. 計算機リソースとデータ量，及び要求されるストレージの規模

数値シミュレーションで利用されるスーパーコンピュータは，予算，電力，設置面積などの境界条件を考慮しながら，目標演算性能を設定して設計される．ファイルシステム，インターコネクト，アーカイバなどは目標演算性能とバランスするだけの性能が必要であるが，ある程度はこれまでの実績を基に決定される．例えば，JSS2 では想定以上に非定常解析が実行されたことにより，当初の予測を超えた容量が必要になったことから，運用途中にファイルシステムを 5 PB から 7 PB へ，アーカイバは 20 PB から 40 PB へ増設することになった．

1.2 章で述べたように，JSS2 ではファイルシステム，及びアーカーブされるデータ量 X_d，ファイル数 X_f の増加量は，日数 d に対して下記のような線形近似でモデル化される．

ファイルシステム

$$X_d = 0.0024d + 0.645, \mathrm{PB}$$

アーカイバ

$$X_d = 0.0102d + 2.3475, \mathrm{PB}$$

$$X_f = 0.1005d - 6.5884, \mathrm{M\ 個}$$

ただし，いずれも JSS2 の Luster-OSS の上限値(7 PB)，また演算性能(3.49 PFlops)などに影響を受けていると考えられ，次のシステムである JSS3 や他サイトでもこの線形近似モデルを利用可能かは定かではない．

3.3.3. モデル化と考察

JSS2 の SORA-MA，SORA-PP，SORA-LM，SORA-TPP においてジョブスケジューラからバッチ/インタラクティブ実行された数値計算ジョブ，及びプリポストジョブに関する 6 か月間の統計データを使い，ジョブやノード単位の IO 量について議論する．まず，使用する記号を以下にまとめる．

1 か月間の平均合計 IO 量: $\mathcal{D} = \sum_{u=1}^{U} \sum_{j=1}^{J_u} d(u,j)$
1 か月間のユーザ毎の平均ジョブ数: J_u
ジョブの IO 量: $d(u,j)$
ユーザ数: U
ユーザインデックス: u
ジョブインデックス: j
ジョブ毎の使用ノード数: $\mathcal{N}(u,j)$

Figure 34 に 1 か月間にユーザが利用した合計 IO 量 $\sum_{j=1}^{J_u} d(u,j)$ に関して，6 か月間の平均値を示す．なお，以下では合計 IO 量が 1 GB 未満のユーザは無視する．最も IO 量が大きいノード区分は 129～256 であるが，1 ノードで実行されたジョブの IO 量は比較的大きく，ほとんどがプリポスト処理の IO 量だと考えらえる．各ノード区分で実行された 1 か月辺りの平均ジョブ数 $\sum_{u=1}^{U} J_u$ を Figure 35 に示す． 1 ノードのジョブ数が最も多いが，プリポスト処理の需要が大きいといえる．9～16 ノードの比較的小規模なジョブ数がもっとも多く，これよりノード数が増えるとジョブ数は徐々に減る傾向にある．次に，これらの取得されたデータを使い，ジョブ辺りの平均 IO 量 $\left(\frac{\mathcal{D}}{\sum_{u=1}^{U} J_u} = \frac{\sum_{u=1}^{U} \sum_{j=1}^{J_u} d(u,j)}{\sum_{u=1}^{U} J_u} \right)$ を求めた．その結果を Figure 36 に示す．JSS2 ではジョブ辺りの Elapse time は 10 時間であるため，本結果を 10 時間で割ることでジョブが 1 秒間に出力する平均 IO 量が算出でき，その結果を Figure 37 に示す．ただし，1 ノード辺りに集中しているプリポスト処理ジョブは必ずしも 10 時間であるわけではないため，この評価法は必ずしも正確ではないということを断っておく．1 ノードで実行されたジョブはジョブ数が多いためジョブ辺りの 1 秒間の平均 IO 量でみるとそれほど多くはないが，大規模なジョブほど非定常解析であることから平均 IO 量は増え，257～1024 ノードで最大 180 MB/sec の頻度で IO が発生している．最後に，1 つのジョブが 1 ノード辺りに発生させる平均の IO 量 $\left(\frac{\sum_{u=1}^{U} \sum_{j=1}^{J_u} \left(\frac{d(u,j)}{\mathcal{N}(u,j)} \right)}{\sum_{u=1}^{U} J_u} \right)$ を Figure 38 に示す．主に 1 ノードで実行されるプリポストジョブはノード辺りの IO 量が大きく，平均で 15 GB ほどの IO が発生している．(意外と小さい) 数～数十ノードは定常計算が多いと考えられるためノード辺りの IO 量は比較的少ないが，それ以上の並列度の数値計算はほぼ非定常解析であると考えらえるため IO 量は増加する．ノード数が増えても weak scaling 的に計算規模を増やすと考えられることから，ノード辺りの平均 IO 量は一定値をとるであろうと予想していた．しかし，Figure 38 の結果を見る限り 4～15 GB とバラつきが大きく，予想とは異なる結果となった．

以上のように，JSS2 の運用実績をもとに各ジョブの IO 量を整理した．ここで得られた結果に関しては，計算システムや数値計算のアプリが変われば異なった結果となると考えられる．2020 年 12 月から稼働した JSS3 のデータや他システムのデータを集めることによって，様々なシステムやアプリに関する傾向が得ることができ，モデル化が可能になるであろう．

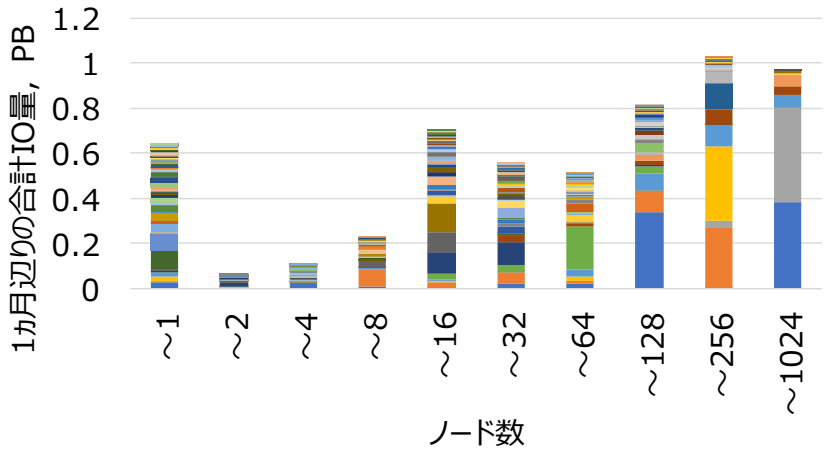


Figure 34　1 か月間で合計した IO 量の平均値　(ユーザ毎に色分け)

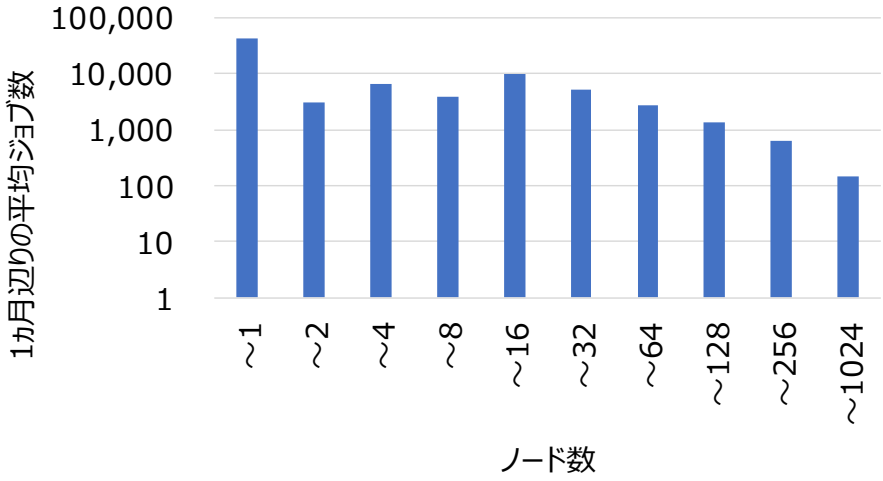


Figure 35 1 か月辺りの平均ジョブ数

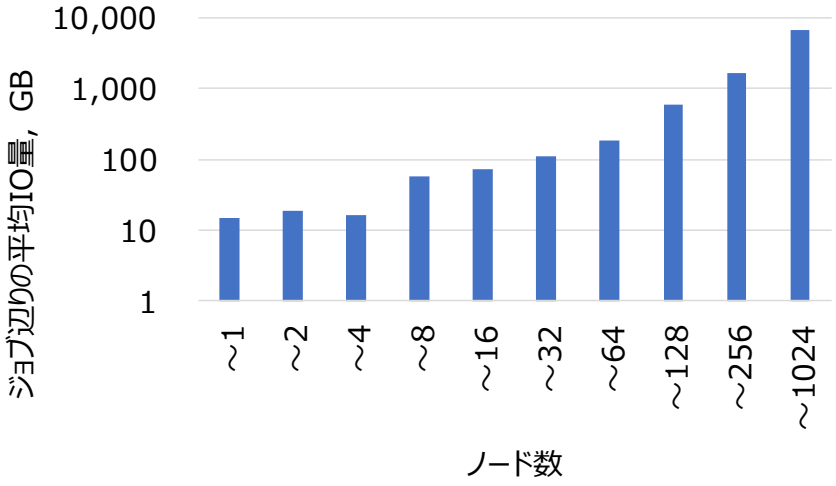


Figure 36 ジョブ辺りの平均 IO 量

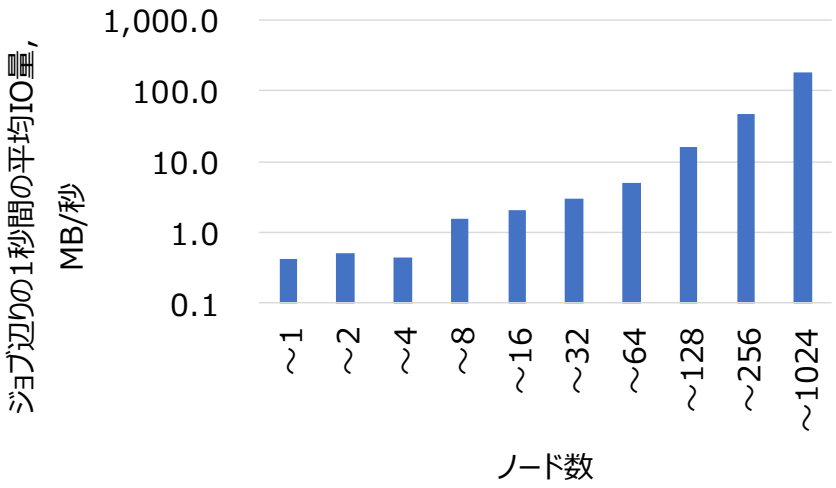


Figure 37 ジョブ辺りの 1 秒間の平均 IO 量

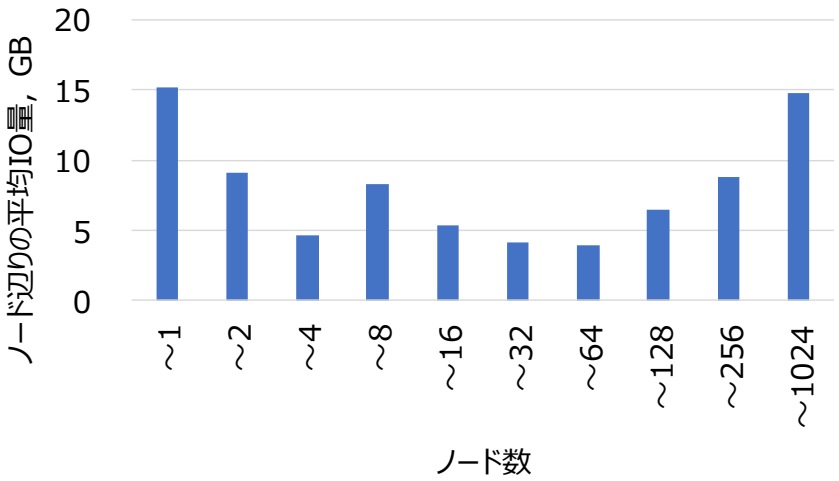


Figure 38 1 つのジョブが 1 ノード辺りに発生させる平均 IO 量

3.3.4. 技術提案とニーズへの答え

3.3.4.1. In situ/in transit 処理

近年のスーパーコンピュータは，計算ノードのメモリ帯域とファイルシステムの帯域の間には大きな差がある．1.1 章で紹介したように，JSS2 ではノードメモリの帯域幅は 1.5 PB/sec であるのに対して IO 性能は 144 GB/sec とその差は 4 桁も開いている．その結果，数値計算のワークフロー全体の中で I/O がホットスポットとなる可視化といったポスト処理に関して，その作業コストが増大し，Capability Computing だけではなく Capacity Computing ですら困難になりつつある．JSS2 では，計算結果を移動することなく ISV を利用したポスト処理までを一気通貫して行うことができるようなシステム構成とした．しかし，増大する計算規模を考えた場合，10 年後も同様のシステム構成でポスト処理ができるとは限らない．そこで，ファイルシステムへデータを書き出すことなく，ノードメモリ上にある計算中のデータを直接ポスト処理する tightly-coupled in situ 処理(以下，in situ)，及び in transit 処理に着目し，JSS2 ではポスト処理の中で時間を要する可視化で実装し，そして JSS2 において試行した．JSS2 における in situ/in transit 処理の例を Figure 39 と Figure 40 に示す．Figure 39 に示すよう，in situ 処理では計算プロセスとポスト処理プロセスを同一の CPU 上に起動し，ノードメモリ上のデータを共有することから，データコピーが全く発生しないという特徴(zero-copy)を持つ．一方，計算プロセスとポスト処理プロセスが CPU やメモリ帯域を共有するため，計算性能への影響が懸念される．また，ポスト処理プロセスが何らかの原因で異常終了した場合，計算プロセスも連動して異常終了してしまうという問題がある．一方，Figure 40 に示す in transit 処理では計算プロセスとポスト処理プロセスを分離し，計算プロセスの主記憶上にあるデータをポスト処理プロセスの主記憶へ Remote Direct Memory Access (RDMA)等で転送する．ノードメモリ帯域よりも 1/14 倍ほど狭いが，ディスク帯域よりは 750 倍速いインターコネクトを利

用しようというのが in transit 処理の狙いである．計算プロセスとポスト処理プロセスを同一の計算機内(JSS2 の場合は SORA-MA)に割当てることもできれば，計算プロセスは SORA-MA，ポスト処理プロセスは GPU を持つ SORA-PP へ割当てるなど，柔軟に計算リソースを割り当てることが可能である．また，計算性能に与える影響を最小限で済む，ポスト処理プロセスの異常終了に計算プロセスが影響を受けることはないといったメリットもある．本報告書では数値流体力学(CFD)の可視化を対象に，in situ/in transit 処理を試行した．In situ 可視化には VisIt/Libsim[1]を使った．in transit 可視化においても可視化には VisIt/Libsim を利用し，ADIOS2[2]を使って計算・ポスト処理プロセス間の通信を行った．In situ/in transit 可視化のいずれも，JAXA のインハウス CFD プログラムである upacs-LES[3]に実装した．CFD 結果の可視化は，可視化アプリ(クライアント)の GUI 機能を使って流体場を解析するインタラクティブ可視化と，ある程度流れ場の現象が捉えられた後に決められた静止画や動画を作成するバッチ可視化の 2 つに分けられ，いずれも利用できる環境を構築する必要がある．バッチ可視化については，in situ/in transit いずれのアプローチでも，一旦プログラムに組み込んでしまえば平易に利用できた．実行性能を比較すると，可視化プロセスのノード数を十分確保すれば in transit 可視化では通信のオーバーヘッドといった影響をほぼ無視することができ，in situ 可視化とそん色ない速度で，かつ堅牢に処理することが可能であった．インタラクティブ可視化ではユーザ端末上で起動する可視化クライアントと SORA-MA で走らせる可視化サーバの間でソケット通信をする必要があるが，JSS2 のような典型的なスパコンシステムの場合，計算ノードはユーザ端末からは直接アクセスはできない．ログインシステム，もしくは JSS2 ではプリポスト専用の SORA-PP と計算システムの IO ノードだけが結合しているのが一般的である．In situ 可視化では，可視化サーバの先頭プロセス(ランク 0)を IO ノードに割当て，ログインシステムの SORA-LI を介してポートフォワードすることで，ユーザ端末に起動した VisIt クライアントと Libsim(可視化サーバ)の間でソケット通信することができた．in transit 可視化では，可視化サーバの処理は SORA-PP にある GPU 資源を有効利用することが望ましい．しかし，SORA-MA は Tofu インターコネクト，SORA-MA と SORA-PP 間は InfiniBand で結合されていることから，異なるネットワークを跨いで直接 RDMA は難しい．そこで，IO ノードにブリッジプロセスを立ち上げ，計算プロセスからブリッジプロセスに Tofu を使ったデータをいったん送信し，次にブリッジプロセスから SORA-PP 上の可視化プロセスに InfiniBand を使ってデータを送信することで，計算プロセスから可視化プロセスにデータを通信することが可能となった．ユーザ端末からは SORA-PP のノードにアクセスすることが可能であるため，ユーザ端末上で起動した可視化クライアントと SORA-PP 上の可視化サーバはソケット通信を行ってインタラクティブ可視化が可能となった．インタラクティブ可視化処理は計算実行中に常時接続して実施するわけではなく，必要な際にサーバとクライアント間の接続が可能となる Attach/Detach の機能が必須である．VisIt/Libsim は Attach/Detach の機能は実装されておらず，in situ 可視化ではいったん切断すると，

[1] http://wci.llnl.gov/simulation/computer-codes/visit.
[2] Godoy et al., SoftwareX, Vol. 12, 100561, 2020. https://doi.org/10.1016/j.softx.2020.100561
[3] S. Tsutsumi, and K. Terashima, Trans. JSASS ATJ, 15 (2017). http://dx.doi.org/10.2322/tastj.15.7.

2 度と再接続することはできなかった．一方，ADIOS2 の通信エンジンに使った in transit 可視化では，Sustainable Staging Transport (SST)を使うことで何度も Attach/Detach が可能であった．以上のように，典型的なスパコンシステムの 1 つである JSS2 において，なんとかインタラクティブ可視化を実現することができたが，ユーザに対して使い勝手が良いわけではない．もし in situ/in transit 処理を使ってインタラクティブ可視化を実施したいのであれば，スパコンシステムの設計段階からこのような利用を加味してシステム設計する必要がある．最後に，詳細な議論は文献[4]が詳しい．

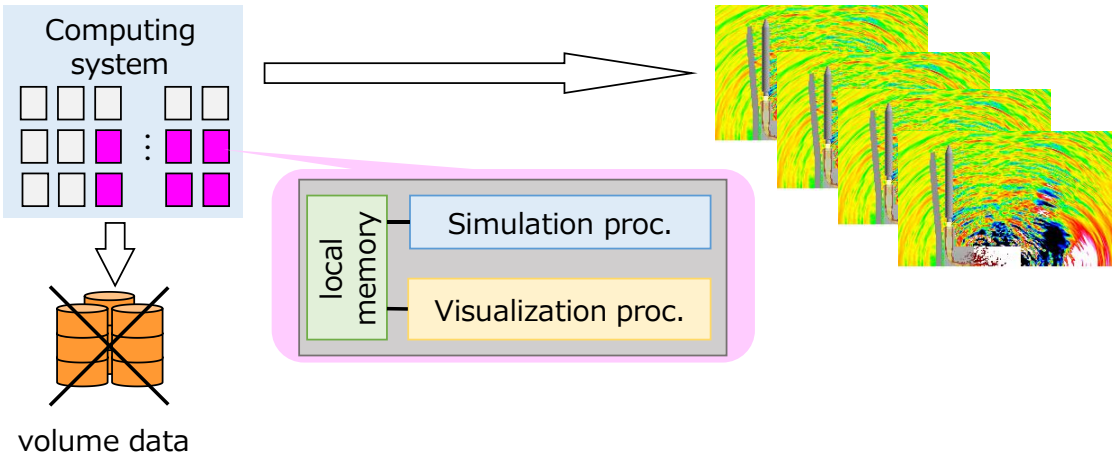


Figure 39 In situ 処理

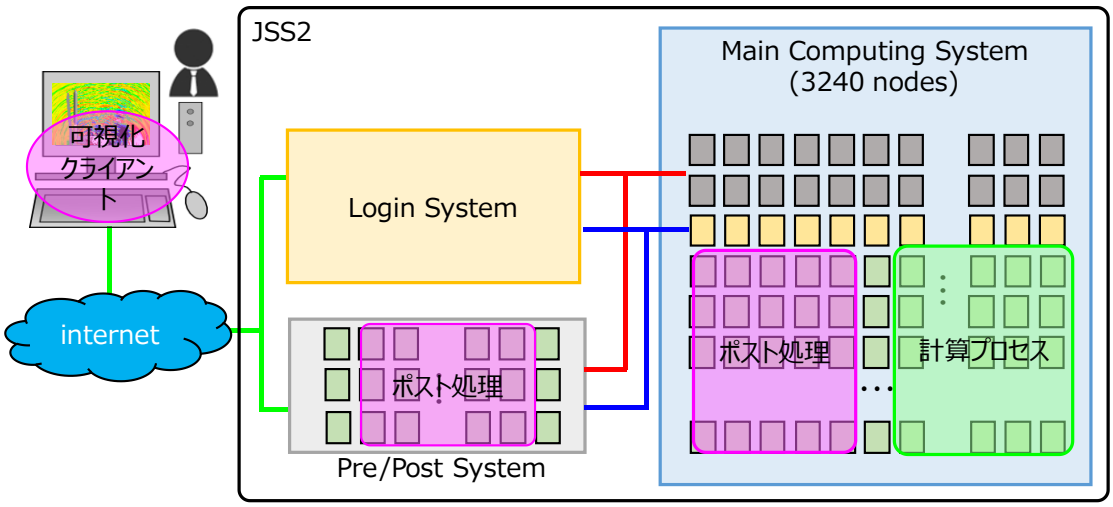


Figure 40 In transit 処理

[4] 堤，他，数値流体力学シンポジウム, B12-1, 2019.

3.4. システムモデル化の議論から見るシステム最適化における課題

以上，３機関のシステムを単純化することでデータ解析システムの最適化における課題を議論した．これらのケーススタディから以下のような共通の課題が見いだされる．既存の技術を用いていくつかの工夫はすでに行われているが（In-situ・In-transit 解析，テープライブラリとハードディスクの間の階層的なデータ保持など），その効果は限定的であり，データ解析作業全体に対してより効果的な解決方法が今後求められる．

1) システム設計は近未来のデータ生成率や利用方法・ユーザ活動[5]を十分考慮して行うが，正確な予測は難しく，必要とされる計算機資源を一度に調達することは現実的には難しい．システムの性能や利用効率を落とさない形で，随時の CPU やストレージの増設に対応していく必要がある．
（キーワード：ストレージ技術と拡張性，ファイルシステム・階層化・平準化，ユーザ活動の記録）
2) 数 PB クラスに及ぶデータ規模に対して，システム更新にともなうデータ移行を効率的に行う必要がある．
（キーワード：ストレージ技術と拡張性，高速なネットワーク）
3) 最新の技術動向に応じて，予算の制約内での CPU，ストレージ，インターコネクト・ネットワーク等の各リソースへの割当の最適化が必要である．
（ストレージ，ネットワーク・インターコネクト，演算装置，メモリなどの各種技術）
4) 大規模データ（サイズ，個数双方）[6]をなるべくストレージ間・拠点間で動かさない最適なハードウェア・ソフトウェアや運用方法の模索が必要である．
（高速なネットワーク，インターコネクト，メモリ，プロセス間通信，ユーザ活動の記録）

次節ではこれらの課題の改善につながりそうな技術動向の調査結果を報告する．

[5] 本研究会「ファイルシステム利用技術 WG」の報告書にも関連の議論がある．

[6] 本研究会「クラウド時代のデータ共有技術 WG」の報告書にも関連の議論がある．

4. データ解析システム構築に関する技術動向

システム調達・システム設計を行う上で，システム規模とその中でのデータの利用方法を明らかにし，システムを実現させるために必要な技術要素のロードマップを十分に理解して生かすことが重要であると言える．前章では以下４つの課題が挙げられた．

課題１．ストレージ技術と拡張性
課題 2．高速なネットワーク
課題 3．ファイルシステム・階層化・平準化
課題 4．ユーザ活動・システムの利用のされ方の記録

本章では，上記課題１．と課題 2．に対し，以下の技術動向を記載しているため，参考にしていただきたい．

1) 記録メディアの技術動向
   SSD/HDD/Tape/Archival Disc/次世代フラッシュメモリ
2) 接続インターフェースの技術動向
   FIBRE CHANNEL/NVMe/NVMe over Fabrics
   Ethernet/InfiniBand/PCI Express(PCIe)
3) 参考情報
   LTO ULTRIUM/Archival Disc

SSD/HDD容量単価比較

FUJITSU

- 大雑把に言うと1桁違う
- NAND Flashは3D/多値化により大容量化は進んでいるが、容量単価の差は埋まっていない

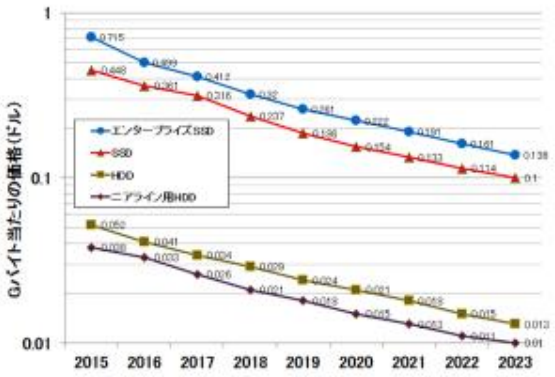


Figure 40　**HDDとSSDのGバイトあたりの価格の推移**
（出所：テクノシステムリサーチのデータをもとに作成）

FUJITSU

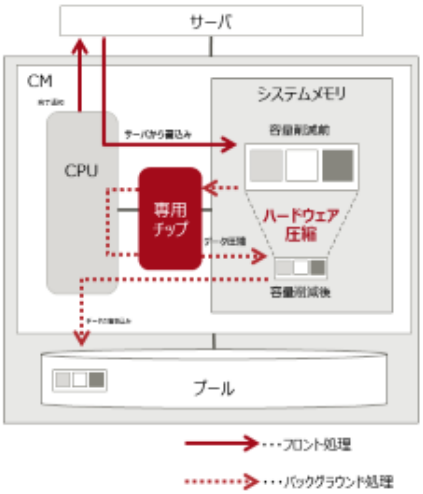
- 米Western Digital社が2017年に公表した資料は、2028年の時点でもHDDとSSDの間には10倍程度の差があると指摘

SLC：Single Level Cell　MLC：Multi-Level Cell　TLC：Triple-Level Cell　QLC：Quad-Level Cell
PMR：Perpendicular Magnetic Recording　MAMR：Microwave Assisted Magnetic Recording
Figure 41　**HDDとSSDのTバイト当たりの価格の推移**
（出所：Western Digital社）

ストレージ装置側でのSSD活用の取組み

FUJITSU

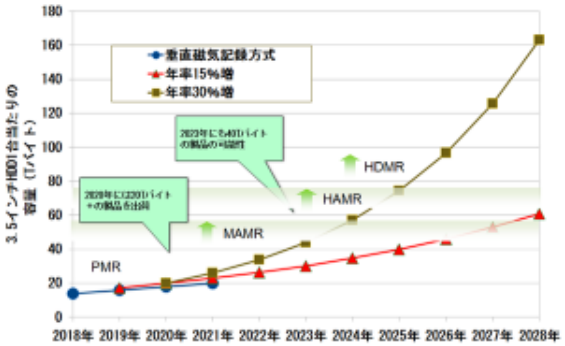
- 重複排除・圧縮による書込みデータ量削減
  - SSDに書き込む物理データ量を削減し、論理容量当たりのコストを改善する
- 課題もある
  - 重複排除は、プール容量を大きくすると、システムメモリー上のハッシュテーブルの容量も増える
    - メモリーの拡張は可能だが、停電時の不揮発デバイスへの退避に時間がかかり、必要となるバッテリー容量が増大
  - 圧縮はCPU負荷が大きい
    - ETERNUS DX上位モデルでは専用チップに圧縮機能をオフロード(右図)



HDDの技術動向

FUJITSU

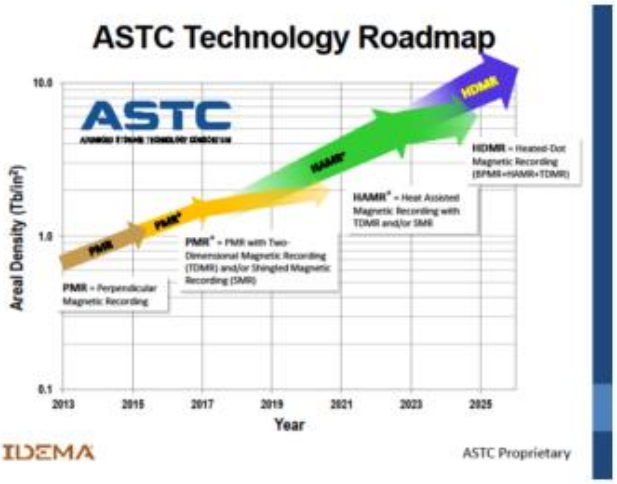
- 近年、容量の伸びが鈍化していたが新たな技術で容量増加が期待される
- 各技術での最大容量
  - MAMR(Microwave Assisted Magnetic Recording)/Western Digital: ～60TB/3.5"
  - HAMR(Heat Assisted Magnetic Recording)/Seagate Technology: 70～80TB/3.5"
  - いずれも今年サンプル出荷か？
- ガラス基板で媒体枚数を8→12枚にすることで1.5倍の容量となり、100TBのドライブが実現か
- ただし、線密度としてはあまり上がらず性能の伸びは少ない(4%程度)



PMR: Perpendicular Magnetic Recording
MAMR: Microwave Assisted Magnetic Recording
HAMR: Heat Assisted Magnetic Recording
HDMR: Heated-Dot Magnetic Recording
Figure 42 **HDD容量の推移**
（出所: 各社の発表を基に推定）

【補足】ASTC Technology Roadmap

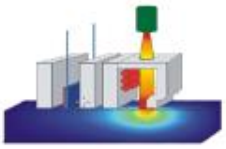
FUJITSU



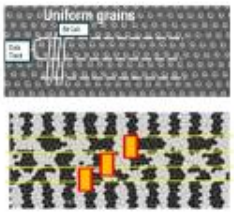
【補足】

FUJITSU

- HAMR／MAMR(Heat Assited Magnetic Recording／Microwave Assited Magnetic Recording)
  - 記録前に媒体を加熱し記録補助
  - 微細化に伴う熱揺らぎによるビット反転を防ぐため、異方性エネルギーの強い媒体材料を使用



- HDMR : Heated-Dot Magnetic Recording
  HAMR + BPMR + TDMR
  - BPMR(Bit-Patterned Magnetic Recording)
    - 隣接干渉の影響を抑制のため、加工した各磁性ドット毎にデータを記録
  - TDMR(Two Dimensional Magnetic Recording)
    - 複数の再生ヘッドで媒体上の信号を読み取り、トラックの信号強度を増幅し隣接干渉の影響を抑制



HDD/TapeのTCO比較

FUJITSU

- 米Enterprise Strategy Group（ESG）社の2018年の調査によれば、業界標準のLTO第8世代（LTO-8）を利用した場合で10年間にかかるTCO（210万ドル）は、全てをHDDに保存した場合（1520万ドル）の14%に留まるという
- 米INSIC（Information Storage Industry Consortium）によれば、2016年のESGによる同様な調査や、2015年に米Clipper Groupが実施した類似の調査から、磁気テープとHDDの間のTCOの比率は年々拡大している

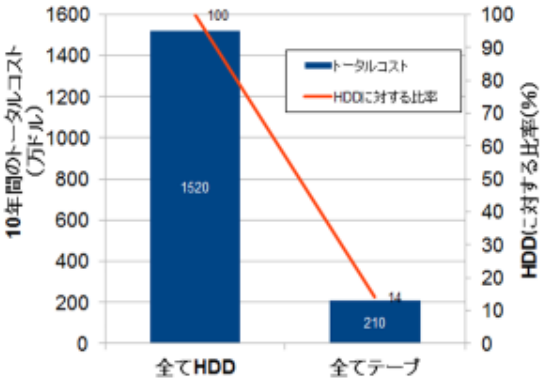


Figure 43　**磁気テープ、HDDのTCOの比較**
（出所：米Enterprise Strategy Group社）

HDD/Tape容量

FUJITSU

- INSICが2019年に公開した技術ロードマップでは、2019年の1巻当たりの非圧縮容量が25Tバイトであるのに対し、年率40%増で成長し、2029年には同723Tバイトに達するという（図-5）。HDDの容量が仮に年率30%増で拡大したとしても差は開く

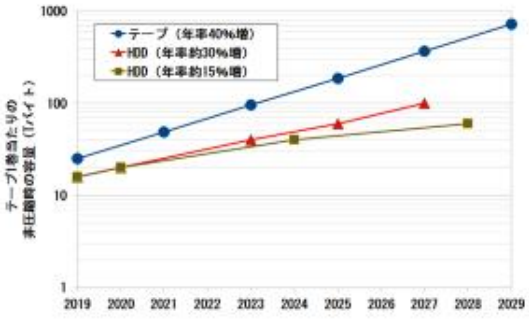


Figure 44　**磁気テープの容量の推移**
（出所：INSICのロードマップのデータを基に加筆）

Archival Discは異なる価値で

FUJITSU

- 耐環境性で優位
  - 媒体のみの保管であれば、世界のほとんどの都市で空調なしで保管可能
  - 水害などで水没しても洗浄(ベンダー作業)後、読出し可能
  - 磁気記録ではないので、電磁波パルスでもデータ消失しない
- 記録容量・性能では対抗できない
  - HDD/Tapeと比べ単位体積当たりの記憶容量が少ない
  - ハードウェア暗号化や圧縮機能なし

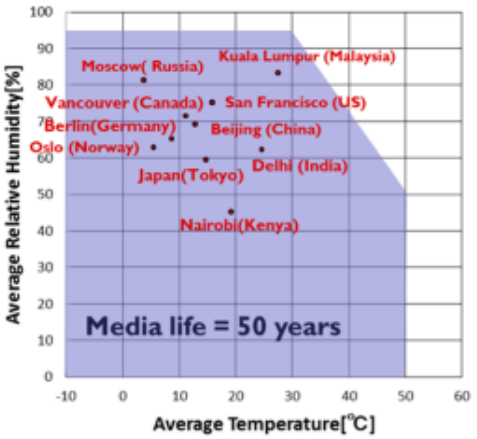


Figure 45　**Archival Discの耐環境性**
（出所：Archival Disc TechnologyのWhite Paper）

次世代フラッシュメモリー

FUJITSU

- PM(Persistent Memory)/SCM(Storage Class Memory)としてPCRAM/PCMが商品化されているが、NAND Flashの置き換えには当面至らない

MRAM(Magnetoresistive Random Access Memory)
PCRAM(Phase Change Random Access Memory), PCM(Phase Change Memory)
ReRAM(Resistive Random Access Memory)

次世代フラッシュ比較

FUJITSU

	DRAM	STT-MRAM	PCRAM/PCM	ReRAM	NAND	NRAM
記録方式		スピン注入 (抵抗値変化)	相変化 (抵抗値変化)	抵抗変化 (抵抗値変化)	電荷	Carbon Nanotube (抵抗値変化)
セルファクター	4-6F^2	6-14F^2	4F^2	4F^2	4F^2	
書込み電圧	1-1.5V	1-1.5V	1.5-3V	~3V	18-20V	-2V
書込み電流	20μA	49μA	100μA	25μA	<1μA	1μA
書込み時間	5ns	<7ns	<100ns	<10ns	10μs	20ns
書換え回数	10^{15}	10^{15}	10^{12}	10^{6}	10^{5}	>10^{12}
データリテンション	<100ms	10Years	10Years	10Years	10Years	>10Years
研究企業		IBM, Micron Technology, SK Hynix, Samsung Electronicsなど	Intel(3D XPoint), Micron Technology, Samsung Electronics	Samsung Electronics, Micron Technology, Cypress Semiconductor		Nantero,富士通セミコンダクター
状況		構造が複雑なため大容量化で苦戦。性能などは非常に有望であり課題解決が望まれる。	Intel, Micronは3D XPointとして商品化。期待性能は出ていない。歩留まりが悪い。	3D実装で高密度化(大容量)が必要。		性能、消費電力の少なさから期待されている。

記録メディアに関するまとめ

FUJITSU

- 各記録メディアは今後も大容量化、高性能化していくが、少なくとも今後10年程度は、置き換えることはなく共存していく
  - フォログラム記録などのアイディアは出ているものの実用化には遠い
  - 次世代フラッシュについても、小型化／大容量化と歩留まり／コスト改善にまだ時間がかかる

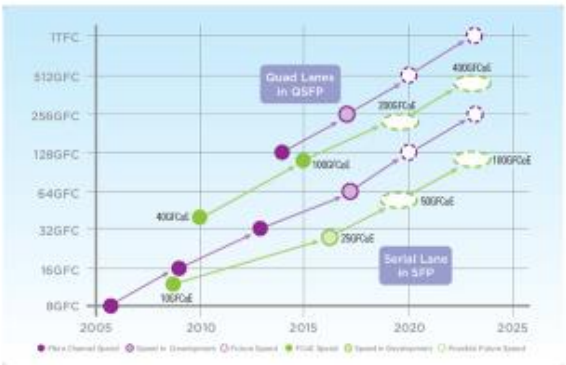
FUJITSU

接続インターフェースの技術動向

FIBRE CHANNEL ROADMAP

FUJITSU

- 次世代Fibre Channel(Gen7)
  - SFP: 64Gbps(50GbEに比べ9%高速)
    - 2018年に規格化済
    - FCIAのレポートでは2019年に製品が出ると書かれているが、光モジュールの開発が遅れており、SWが2020年後半、HBAは2021年と想定
    - BER(ビット誤り率):Ethernet 10^{-12} , FC 10^{-15} 1000倍高品質
  - QSFP: 256Gbps(64Gbps*4)
    - 2019年規格化
    - 製品出荷は2021年以降では？
- 128/512Gbpsも規格化の予定
  - 1Tbpsを2029年に規格化予定



出典：FCIA(Fibre Channel Industry Association)
https://fibrechannel.org/roadmap/

NVMe/NVMe over Fabrics

FUJITSU

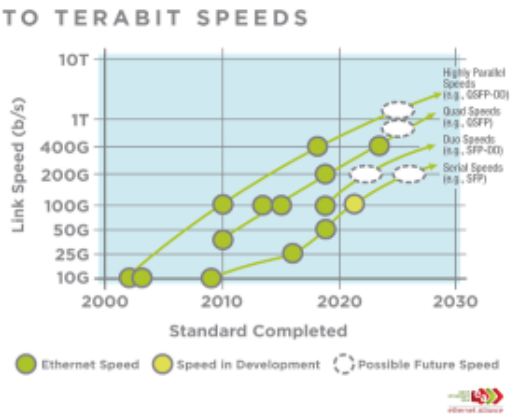
- SSDはNVMe化
  - All FalshのドライブIFはNVMe
    - PCIe Gen3/4/5と高速化
    - 高速化に伴い消費電力が増加
  - ハイブリドはSAS
    - SASはSAS 3.0(12Gbps)からSAS 4.0(24Gbps)
- 外部ストレージはNVMe-oF
  - 高速化に伴いRDMA転送が主流か？
  - 物理層はEthernet/FC/IB

NVMe™ Specification Roadmap

Ethernet Roadmap

FUJITSU

- Ethernet Alliance
  - シングルレーン25GbE製品が一般的に入手可能になった今、開発はより高速のシングルレーン速度と、デュオ、クワッド、8レーン構成を含むいくつかのマルチレーン仕様に集中
  - ファイルプロトコルとオブジェクトストレージプロトコルだけでなく、ブロックプロトコルのストレージアプリケーションにも影響



InfiniBand Roadmap

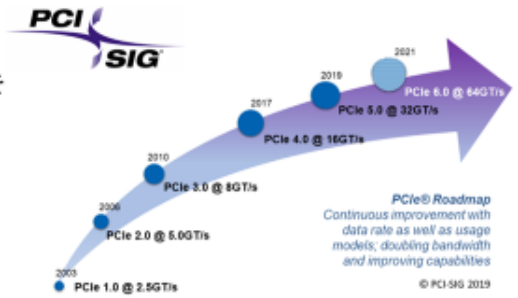
FUJITSU

- Beyond HDR
  - NDRとXDRが発表され、ロードマップ上にある。NDRは2022年以降に利用可能になると推定されているが、XDRの日付は公開されていない。

PCI Express (PCIe)

FUJITSU

- PCIExpress®(Peripheral Component Interconnect Express)
  - 仕様化団体、PCI-SIG® (PCI Special Interest Group)
  - PCIe 4.0
    - 2017年10月、PCIe 4.0、バージョン1.0仕様をリリース
  - PCIe 5.0
    - PCIe 5.0仕様が2019年5月に完成



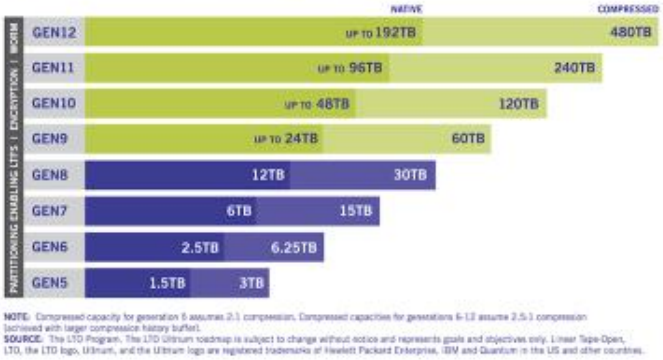
FUJITSU

【参考】LTO ULTRIUM

LTO ULTRIUM ROADMAP

FUJITSU

■ 第12世代まで規格化されており継続的な容量の伸びが期待される



LTO ULTRIUM ROADMAP

FUJITSU

- これまで継続的に容量、転送速度が向上
- LTO-9以降の容量・転送速度については目標値で、未発表の部分もある
- 出荷時期については約2.5年で世代が上がっている(LTO9以降は予想)

規格／世代		LTO-1	LTO-2	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10	LTO-11	LTO-12
出荷年		2000	2003	2005	2007	2010	2013	2015	2017	2021?			
非圧縮	容量[TB]	0.1	0.2	0.4	0.8	1.5	2.5	6	12(9)	24	48	96	192
	転送総度[MB/s]	20	40	80	120	140	150	300	360(300)	708	1,100	TBA	TBA
圧縮	容量[TB]	0.2	0.4	0.8	1.6	3.0	6.25	15	30(22.5)	60	120	240	480
	転送総度[MB/s]	40	80	160	240	280	400	750	900(750)	1,770	2,750	TBA	TBA

ドライブの提供は事実上IBMの1社
テープ媒体はグローバルに見ても、富士フイルムとソニーの2社のみ

LTO仕様

FUJITSU

規格／世代		LTO-1	LTO-2	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10	LTO-11	LTO-12
出荷年		2000	2003	2005	2007	2010	2013	2015	2017	2020	2022	2025	2027
非圧縮	容量[TB]	0.1	0.2	0.4	0.8	1.5	2.5	6	12/9	24	48	96	192
	転送総度[MB/s]	20	40	80	120	140	150	300	360/300	708	1,100	-	-
圧縮	容量[TB]	0.2	0.4	0.8	1.6	3.0	6.25	15	30/22.5	60	120	240	480
	転送総度[MB/s]	40	80	160	240	280	400	750	900/750	1,770	2,750	-	-
Tape length		609 m		680 m	820 m	846 m		960 m					
Tape width		12.650 mm ± 0.006 mm											
Tape thickness		8.9 µm		8 µm	6.6 µm	6.4 µm	6.1 µm	5.6 µm					
Magnetic pigment material		Metal Particulate (MP)					MP or BaFe	BaFe					
Base material		Polyethylene naphthalate (PEN)											
Data bands per tape		4											
Wraps per band		12	16	11	14	20	34	28	52/42				
Tracks per wrap (read/write elements)		8		16				32					
Total tracks		384	512	704	896	1,280	2,176	3,584	6,656/ 5,376				
Linear density (bits/mm)		4,880	7,398	9,638	13,250	15,142	15,143	19,094	20,668/ 19,104				
Encoding		RLL 1,7	RLL 0,13/11; PRML			RLL 32/ 33;PRML	32/33 RLL NPML						

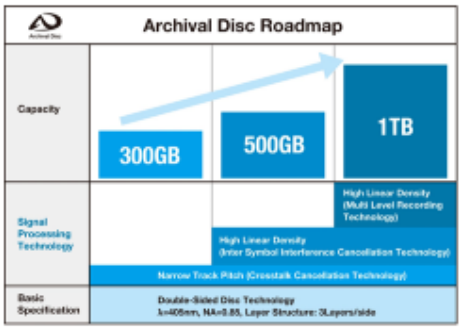
FUJITSU

【参考】Archival Disc

Archival Disc

FUJITSU

- ソニー株式会社とパナソニック株式会社が、デジタルデータを長期保存するアーカイブ事業の拡大に向けて、業務用次世代光ディスク規格Archival Disc(アーカイバルディスク)を策定
- Archival Discは両面ディスク(片面3層)、データ記録容量の大容量化を実現。下位互換性を高く保つことができるため、データの長期運用に適したアーカイブメディア



- 2社の仕様は非互換
- 次期媒体(500GB)の提供時期
- その次(1TB)は未定

【補足】Archival Discの改善点

FUJITSU

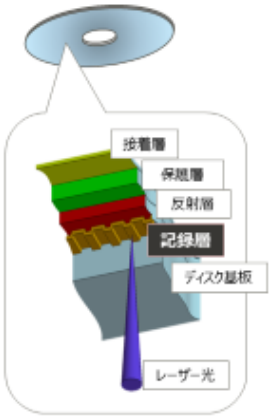


図. 光ディスクの概略構造

Disc種別	耐環境性	媒体寿命	耐久性
DVD-R ・媒体材料 色素材料(有機物) ・記録方式 レーザー照射による **色素変化**	△ ・経年変化により、光(太陽、蛍光灯)、温度、湿度の影響を受ける可能性有り	△ ・10年以上 (メーカによってバラツキ有り)	△  ケース無し ・取扱い時にキズ、指紋、ほこりなどが付きやすい
Archival Disc ・媒体材料 酸化材料(無機物) ・記録方式 レーザー照射による **物理変形**	◎ ・無機物を物理変形させるため、経年変化しない	◎ ・100年以上	◎  ケース有り(マガジン) ・マガジン単位での取扱いになり、キズ、指紋、ほこりなどの影響を受けない

メディア比較

FUJITSU

媒体種別 (記録方式)	寿命	セキュリティ	TCO (トータルコスト)	レイテンシー
HDD (磁気変化)	△ (5年程度) ・HDD内部メカ部品の寿命のため	△ ・磁気変化のため物理的に書換え可能	△ ・寿命の度にデータマイグレーションが必要 ・消費電力が大きい ・ビットコストが高い	◎ ・数msec~数十msec
Tape (磁気変化)	○ (30年以上) ・テープのベースフィルム寿命のため	△ ・磁気変化のため物理的に書換え可能	◎ ・世代間の互換性があるため、10年毎にデータマイグレーションが必要 ・消費電力が小さい ・ビットコストが安い	△ ・数min
Archival Disc (レーザ照射による物理変形)	◎ (100年以上) ・無機物を物理変形させるため経年変化しない	◎ ・物理変形のため書換え不可(改ざん防止)	○ ・全世代でデータ読取り互換があるため、データマイグレーション不要 ・消費電力が小さい ・ビットコストはHDDより安く、Tapeより高い	△ ・数min

5. まとめ

データ解析は科学活動の進展の根幹を成す作業である．科学技術の発展や各種プロジェクトの大型化にともない，データや処理に必要な計算機資源も今後ますます大規模になっていくと予想される．本 WG ではこのようなデータ解析を扱うためのシステムの最適化について議論を行った．各機関プロジェクトにおける既存システムの運用事例を学んだところ，ひとえにデータ解析と言ってもその中で扱われるデータの特性や必要とされるシステムへの要請は多岐に渡ることが分かった．そのため，本 WG ではその中でも主に各機関の関心が高いデータの保存や利用効率に関わるストレージ周りの議論を取り上げた．その上で，データの量とシステム内の利用方法に着目して現在各メンバ機関が直面する，あるいは近い将来に想定される課題について議論を行った．

本 WG の議論の帰結として重要なことは，増大するデータ量への対応が科学分野の大学研究機関のデータ解析において共通の課題として見いだされたことである．また同時に現時点で直接的な解決につながるハードウェア・ソフトウェア技術がただちには存在しないことも重要である．各システムでは経験に基づき既存の技術を用いていくつかの工夫をすでに行っており一定の効率化を実現している．しかし真の適化を行うためにはシステムごとに重視する項目・メトリックの確立が必要であり，今後の課題である．

議論から得られた，データ解析システムの最適な設計を行う際に考慮すべき観点は以下の 4 点である．

1) システムに必要とされるリソースや構成は，データ量（サイズ・個数）・データアクセス頻度などの観点で単純化して考えること

2) 既存システムの運用実績や今後のデータ解析のプロジェクト計画やユーザ活動のニーズ・予定に基づいて，システム規模の推移を予測すること

3) データがどのように利用されるかを考慮し，最適なシステム内のデータ移動やプロセスの配置を設計すること

4) システム設計の課題に応じて，システムを構成する技術要素のロードマップを考慮し設計すること

既存システムでは特にデータ保管や移動の負荷への対応が大きな課題であり，各システム運用で一定の工夫をしているものの，今後直接的な解決に至る技術の登場が求められる．そのような状況でも，既存システムの運用実績やデータ解析プロジェクトの要請を考慮することで，必要な計算機資源を見積もり，技術要素のロードマップを考慮してシステム設計することの重要性が確認された．4 章では関連する技術要素のロードマップについて最近の動向を調査し報告した．大学・公的研究機関等の場合，5 年前後での計算機システムの更新に際し，次の更新期間までのデータ増大を予測し，技術動向を見据えて調達を行うことは重要な命題と言える．今後もこうした WG 等を通してそのような技術情報が関係諸機関と効率的に共有できると良いだろう．

一方で，本 WG はデータ解析に関わる多くの課題を俎上に載せ，その中で比較的共通の課題を見出すところまでの大枠の議論に留まった．我々は本 WG の企画時には，データ解析処理の効率に関わ

る現行の課題に対して，システムをどのように設計・構成・運用すべきか具体的な解決方法を検討することを目指した．しかし，実際には個々の実装の検討は将来への課題となった．これは上述のように「データ解析」という作業が扱う内容・データの特性は非常に幅広く，各メンバの持つ喫緊の課題それぞれが重要で多岐に渡ったためだが，この事実も本 WG の議論の知見として重要であろう．

具体的にシステム最適化を実現し実際に解析効率を改善していくためには各課題について技術的な実装を議論することも重要である．たとえば，本 WG でも議題には上がりながら具体的な検討に至らなかった将来的な検討課題として以下のようなものがある．

システム設計で考慮すべきパラメータ：

- 機能的な事項
  - 処理性能に関する技術
    （CPU・アクセラレータ，主記憶，ストレージ・テープ，InterConnect，ネットワークなど）
  - データ量に対応する技術
  - 利用統計の収集・解析とその自動的なファイルシステムの階層化と平滑化への活用
  - AI の活用
- 設置環境に関する事項
  - 電力，冷却，設置スペース，重量
- 調達方法に関する検討
  - 方針・政治
  - 予算
    - ある予算で 5 年後に保持できる安全なデータサイズといった観点での検討
    - 調達のサイクル，一括・分割調達の考え方（コストパフォーマンス，運用しやすさ）

一般的にシステム設計では機能的事項に注力しがちだが，設置環境的な項目[7]，効率的な調達方法に関する検討も考慮に入れる必要がある．こうした課題に対して，本 WG のような大枠の議論を行う場とは別に興味範囲を絞り技術的なテーマを明らかにした上でメンバを募ったサブ WG あるいは新 WG などで今後議論を重ねることが重要であろう．

以上．

[7] 本研究会「情報システムの効率的エネルギー活用検討 WG」の報告書にも関連の議論がある．

==========

大規模データ処理システム最適化 WG 成果報告書

【発行者】：サイエンティフィック・システム研究会
【編　集】：サイエンティフィック・システム研究会　大規模データ処理システム最適化 WG
【発行日】：2021 年 3 月 29 日
【連絡先】：サイエンティフィック・システム研究会 事務局
〒105-7123 東京都港区東新橋 1-5-2 汐留シティセンター
Email: ssken_office@ml.css.fujitsu.com
http://www.ssken.gr.jp/MAINSITE/

==========