

今後のHPCIに向けて

東京大学 塙 敏博

運用FSで目指したこと

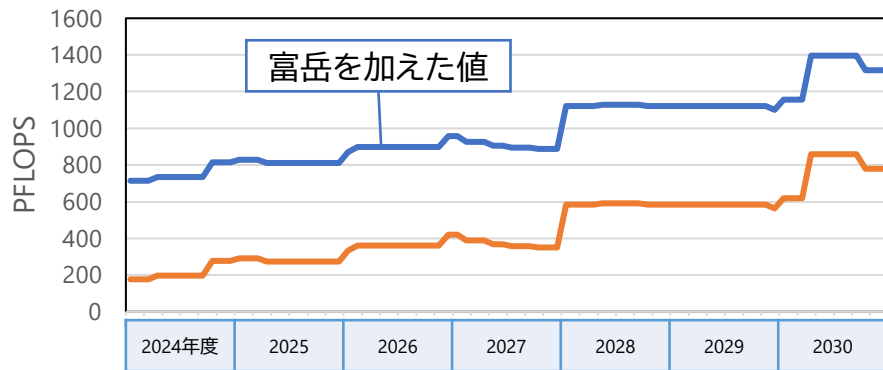
- サステイナブルな計算資源の提供
- 研究データの提供・共有
- データセントリックで複雑なワークロードを実現
- 実験・観測設備等との連携
- 既存の環境から移行を容易に

→柔軟な利用形態の実現を目指す

運用FSの成果・提言→今後のHPCIの理想形

次期(+その先の)フラグシップシステム整備に向けて

- これまでのフラグシップシステム開発の問題
 - 計画立案からシステムの運用終了までのサイクルが10年程度(京の場合)
 - 継続的な計画がなかったため、次期システムが決まるまで運用期間の方針も立たない
 - 富岳もいつまで運用するか現時点での見通しが無い
 - システム移行のタイミング:いわゆる**端境期**→ ユーザの研究計画に多大な影響
- 理想的な形 → そもそも**端境期を作ってはいけない**
 - 更新時期をずらし一体的な並行運用について検討
 - **研究開発の継続性も維持、技術動向の変化に即時対応できる**
 - 計算機システムの寿命は導入からせいぜい**6年**: コンポーネントの保証(End-Of-Life)の観点、これまでの運用経験から見て妥当な線
 - 製品のライフサイクルは、技術革新のサイクルの影響で延びてきているものの、6年目以降は厳しい
 - 7年目から保守費が急激に増加(IT機器は5年程度が多い)、機材によってはリプレースしないと継続運用できない
 - 利用者にとっては陳腐化し魅力が低下



2030年度までの9大学基盤センターの導入計画について調査
(2023~4年度時点での合計値)

- **HPCIにはこの一部を提供: 2024年度(平均): 55.3 PF / 186 PF**
→ 約30%
- 2028年度以降はあくまでも希望的観測
→ **NIS(大学基盤センターなど)だけでは端境期は乗り切れない**

富岳

導入・運用(案)

導入

システム A1

入れ
替え

システム A2

引退?

導入

システムB1

入れ
替え

システム
B2

次世代計算基盤の整備に向けて:基盤センター群

- 基盤センター群を「第2階層」と呼ぶのはやめよう!! → **NIS: National Infrastructure Systems**

- 多様なアーキテクチャの選択、運用方針がセンター独自の特色、それらが尊重されてきた
- システム導入予算の減少
 - 大学予算の一部、独法化後、運営費交付金の削減→センター予算にも波及
 - システム整備、保守等の出費増(WiFi設備、クラウド等のライセンス整備など)
 - 電気代高騰、物価上昇、円安など外的要因
- システム価格の高騰
 - 物価上昇、製造コストの上昇、ムーアの法則の終焉



システム更新しても、性能が下がる可能性が高い
近年の多くのシステムはCPU→GPUシフトで
性能は向上したが、しかしその先は厳しい

- 基盤センター群に対しても**導入・運用のための予算配分**
 - **[導入]** システム**価格の上昇**、高効率な冷却設備のための初期導入コスト（運用コスト減少分は導入費用に組入れできない）
 - **[運用]** **電力料金の高騰**+今後求められる可能性が高い**再エネ利用のコスト増**
 - センター間連携の後押し
- フラグシップだけでなく**基盤センター群+アカデミッククラウド+ストレージ**の充実・継続性も重要
 - 先進的・特色あるシステム、**量子**との連携など、**導入・運用**を後押しする**仕組みがあってもいい**
 - 基盤センターには、ユーザー層拡大、共同研究、手厚いユーザー支援、などの役割の充実が期待される
 - Cf. GPU移行に向けたセンター間の協力体制

量子・スパコン連携プラットフォーム(JHPC-quantum)

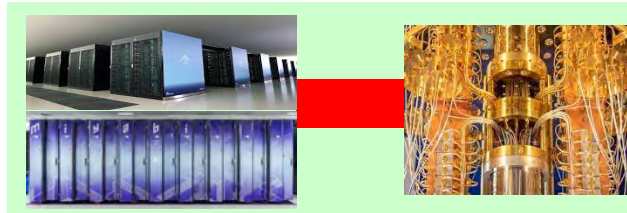
<http://jhpc-quantum.org/>



- JHPC-quantum (FY.2023-FY.2028)
 - 理研R-CCS, ソフトバンク, 東大, 阪大
 - 経産省・NEDOポスト5G関連事業, 産業応用への期待
- 2基の量子コンピュータ(QC)を導入(2025年)
 - IBM(超伝導型), 156+Qubit, 理研神戸(6月)
 - Quantinuum(Ion-Trap型), 20+Qubit, 理研和光(2月)
- 量子・HPCハイブリッド連携
 - QCを加速装置として扱う
 - 量子科学, 量子機械学習



CPU		GPU		Others	
A64FX Arm	X86	NVIDIA Intel AMD Arm	AMD Intel	Sambanova Cerebras Graphcore etc.	Quantum
h3-Open-SYS/WaitIO					



- 東大の役割: 連携環境向けシステムソフトウェア
 - h3-Open-BDECの拡張
- 2025年秋以降: 富岳・Miyabiと量子コンピュータの連携開始
 - 連携のためにmdxも使用予定

スパコン調達スキームの改善

- **問題点: 日本では開発途上のシステムでの調達がほぼ不可能**
 - 「スーパーコンピューター導入手続」の制約
 - 仕様を満たせない事態が起きた場合、リスクを見通せない
 - 調達やり直し、ベンダに過大なペナルティ、調達者にとっても負担
 - 「落札した当該供給者は、機種を指定する納期までに納入しなければならない。もし、当該供給者が納入を行うことができない場合には、調達全体は再度入札公告に付されるものとする。」
 - 「落札したシステムは、納入前にベンチマーク・テストを行い、予測性能値と同等又はそれ以上の結果を示すと共に仕様を満たさなくてはならない。」
 - 製品仕様が確定するまでは入札公告ができない → **欧米より半年～1年以上の遅れ**
- 米国では contingency planを契約に含めることができる(EUでもおそらく可能)
 - 提案に **Non-Recurring Engineering (NRE)** を含めることが可能(次世代テクノロジー実現のための開発支援)
 - 進捗に応じた段階的な契約が可能 (例: 1st phaseが成功→2nd phaseへ、1st phaseの進捗によっては、2nd phaseは契約内容を変更、または契約しない)
 - NREの進捗も条件に物品契約 (Plan Aがうまくいかなければ Plan B、など)
 - 結果的に、契約金額に幅があることがほとんど
 - 技術審査: 定量的ではない加点要素も加味
 - 日本の制度で言う「企画競争」に近い: mdxで実施

複数システム間・異種基盤間の連携

- 計算基盤の一体運用 & オンプレシステム、実験機器、研究分野DB等とのワークフロー連携など

- 全ての資源やサービスで共通の認証・認可、認証基盤
- メタスケジューラ、オーケストレータ
- ワークフローとKubernetesによるコンテナ連携
- ワークロードのマイグレーション
- 電力需給バランス次第で実行するシステムを変更

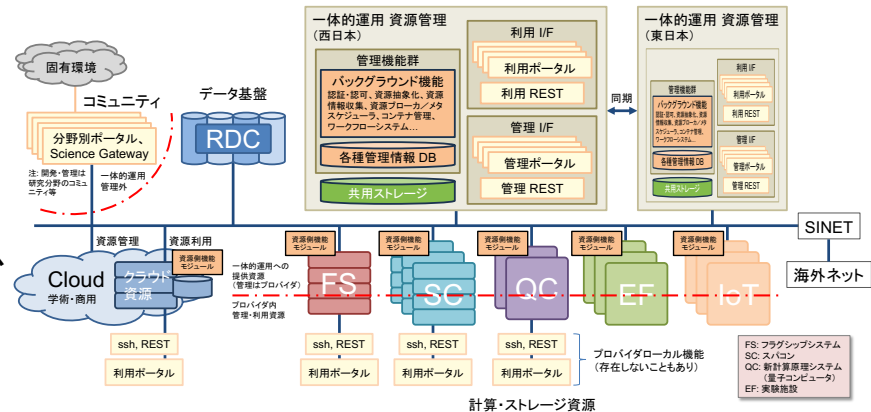
- イベントが起こると短時間での起動・再開が求められるジョブ

- 緊急ジョブ、実験施設連携ジョブ、ストリーミングデータ、LLM推論など
- Gang Scheduling+高速な同期切替機構など

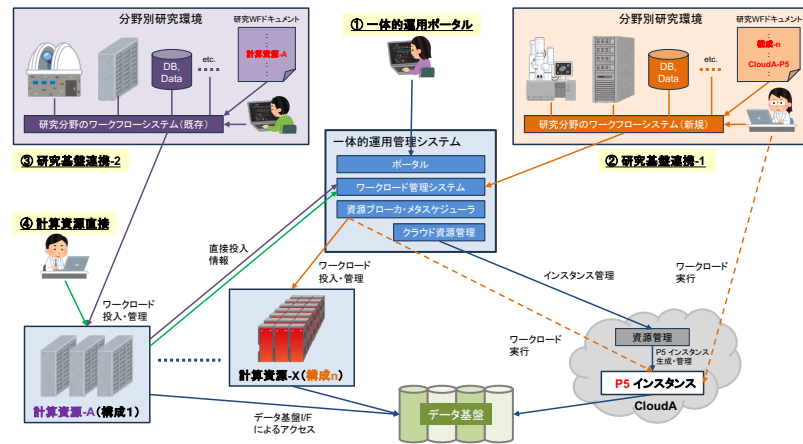
- 欧米の計算基盤における異種基盤連携調査

- 米国では Integrated Research Infrastructure (IRI)を推進: ANL Auroraと放射光施設 (APS) など

➔ 得られた知見を踏まえて連携のあり方を検討中、
推進・支援する人材も極めて重要



一体的運用計算基盤の構成例



一体的運用計算基盤と研究基盤の連携例

HPCシステムのセキュリティ

- 現状:「HPCI共通セキュリティ要件」を元に運用

- 目的:不正アクセス・サービス妨害の防止や、インシデントへの緊急対応・事後調査・恒久対応
 - 認証局運用機関向けには別途「HPCI認証局運用規定」
- 資源提供機関のポリシーに依存する部分: 共通化するためレベルの低いポリシーに合わせたもの



- HPCシステム用セキュリティガイドラインを策定

- NISTのCyber Security Frameworkを基にした セキュリティガイドライン
 - NII ストラテジックサイバーレジリエンス研究開発センターの協力
 - 境界防御型から侵入やミスを前提とするゼロトラスト型のセキュリティ
- CSFが組織間連携を考慮した2.0に更新されたため、ガイドラインも更新
 - 2022年度CSF 1.1 に基づくサイバーセキュリティアセスメントを「富岳」, 「Wisteria/BDEC-01」に対して実施
- 10の学術機関からCSF2.0対応ガイドラインドラフトについて意見を収集、反映

- ソフトウェアの安全性も必要

- 例: github、spackのようなautomatic check
- SBOM : Software Bill of Materials, 使用したコンポーネント等のリストを一覧化
- セキュリティ・プライバシーに配慮したストレージ、TEE (Trust Execution Environment)によるセキュアな計算環境も必要

HPCI共通セキュリティガイドラインチェックシート

ユーザビリティ向上

- Webベースのユーザインタフェース → 新規スパコンユーザの開拓
- システム間で統一されたユーザインタフェース → システムの相互利用を容易に
- ユーザポータル: Open OnDemand
 - オハイオ州立大学で開発、日本ではR-CCSでいち早く導入し「富岳」で利用可能に
 - TSUBAME4.0@東京科学大、玄界@九州大、AOBA@東北大、Miyabi@JCAHPC、ABCI@産総研など、ボトムアップ的に各所で利用可能に
- Jupyter Lab
 - データ解析、機械学習ではde facto
 - スパコンジョブスケジューラとの連携、プロトタイプ開発中 → 簡便なワークフローの実現
- Webベース研究データ管理基盤: GakuNin RDM
 - NIIで開発
 - 各計算資源やストレージを繋ぐためのプラグイン → データを中心とした計算資源との連携に向けて、ストレージ群と一体の検討が必要
 - 既存のワークフローエンジンに変更を加え「富岳」などからGRDMを用いたワークフロー処理を可能に
- コンテナ技術を組み合わせ複数システムを連携させるワークフローについても検討
- 認証基盤も重要、すべての連携対象が対応する必要

- 中尾、山本@R-CCS:「HPCクラスタにおけるバッチジョブ投入のためのWebアプリケーションOpen Composerの開発」、HPC&CPSY研究会、情報処理学会研究報告、2024年12月、など



システムの利用の現状と問題点

- 現状: ローカルアカウント + HPCIアカウント
 - ローカルアカウントが主
 - HPCIアカウントはストレージ利用が主で、他にはあまり使われていない
 - 運用側としてもローカルアカウントしか意識しない
 - 実質HPCIアカウントだけではユーザとのやりとりが困難
- ユーザ
 - 各システムのアカウントを個別に管理
- 運用
 - 相互乗り入れを考えたときにアカウントの変換が必要

次期認証基盤

- 「学認」をHPCIにも使う
- スパコン、共用ストレージはSP (サービスプロバイダ)として登録
- ユーザ認証情報のセンター内での管理が不要になる
 - IdPは利用ユーザが所属する各組織が提供
 - システム利用の許可情報のみを管理
- 真のシングルサインオンの実現
 - スパコン・ストレージ間の連携を進めるには絶対必要



従来できなかった利用手段の提供

新しい利用形態

- Web UIによる利用システムの柔軟な選択
- 電力逼迫・利用不可能な場合のシステム間マイグレーション
 - 同一アーキテクチャ間であれば、融通は可能なはず
 - 共用ストレージを活用
- 実現のための課題
 - 計算資源を随時利用可能にする仕組み
 - コンテナなどによる利用環境のポータビリティ
 - アカデミア (≡学認が使える組織)以外のユーザはどうする？
 - とりあえずはOrthros (<https://ui.orthros.gakunin.nii.ac.jp>)で救える？

フラグシップシステムとストレージ・データ管理

- ・フラグシップ(に限らず)システム更新に伴うデータ移行は大問題だった

→ストレージのライフサイクルはシステム本体とは(当初から)独立に計画すべき: システム本体より経年劣化は深刻

→ ユーザに意識させずにデータ移行を進める観点で、データ共有のフレームワーク導入を検討

- ・ 海外/国内機関との連携には、研究データの共有および管理が必須、米国/欧州では、Globus を用いたデータ共有がデファクトスタンダードに

→これらのツールを積極的にHPCI/次期フラグシップに取り込み、フラグシップシステムのシームレスな移行を可能にするだけでなく、研究データの共有、連携促進へ繋げる

- 例: Globus

- FaaS (Function as a Service)に近いデータ管理/転送、認証などをサポートしたプラットフォーム

- 昔のGridソフトウェアのGlobusとは別物

- DOIを付与することによって、研究データ、論文への到達性、一意性、永続性を保証 (<https://www.doi.org>)



ストレージサービス

- 冗長性のある分散共有ストレージサービスは必要
 - Gfarmでなければならない理由は無いが、今のところは必要十分(?)
 - シングルストリームのread/write性能の問題もあるが、robustnessを考えるとやむを得ないところもある
- ようやく昨年度からHPCI全ユーザにストレージ容量が自動的に割り当てられるようになった
- 従来の各システムのログインノード、専用クライアントだけでなく、クラウドサービスをもっと前面に出してもよいのでは
 - NextCloudで利用可能、データ共有が容易に
- フラグシップシステムが複数存在する時期には、ストレージへの負荷は多大になる可能性
→ フラグシップシステム専用のストレージ、HPCI共用ストレージ等、計算基盤全体でのストレージ階層の定義と、次期SINETも見据えた上で、必要なバンド幅、設置形態も含めた議論が必要では

日本の研究力を支える計算基盤

- 「日本の研究力向上に資する研究DXを支える計算基盤」に向けて
 - 運用FSとしての当初目標の一つ
- 例えば、科研費のような国プロにおける研究プラットフォームとしてHPCIの利用を推奨する、など認知度を上げられないか
 - 研究データマネジメントのための費用については、プロジェクト予算に計上する義務が今後出てくる？ → 計算資源も同様なのでは
 - データマネジメントと連動して、最初からある程度の計算資源は付いてくる、ようなことも考える
 - プロジェクトごと個々にクラスタが導入されていくのは無駄が多いはず
 - もちろんシステムそのものの研究や、セキュリティの観点では一定のオンプレミス環境は必要だろう
 - カーボンフットプリントに賦課金かけられるような時代がくれば、計算資源の集約が必須になるのでは
 - 集約する代わりに、さまざまなタイプの計算基盤を整備する必要は生じるかも
 - 各種システム実験のためのオープンラボ
 - ユーザ数の爆発的な増加も見込まれ、学認の利用が前提

(参考)アクセラレータを念頭においたユーザビリティ JCAHPC第二世代システム “Miyabi”



導入の経過

- 2022年11月より調達プロセスを開始
 - 2022年6月には事前性能評価によりGPUアーキテクチャを決定
- 2023年11月に開札、富士通が落札
 - 準備期間を1年以上確保

システムの特徴

- システム全体性能の飛躍的な向上のため、演算加速装置としてGPUを主体とするシステムへ
 - 消費電力も削減 → 電力当たり性能の劇的な向上
 - CPU-GPU間が高速リンクで密結合され、既存アプリのGPU化が容易に
- GPU化が困難なアプリケーションのため、汎用CPUのみの計算ノードも導入

GPU移植・移行の計画

- NVIDIA Japanの協力
- 3,000人以上のOFP利用者:2つの形態
- 「自己移植(Self Porting)」:様々なオプション
 - 1週間のハッカソン(ミニキャンプ), 3ヶ月に1回, オンライン・ハイブリッド, Slack併用
 - 毎月開催される「相談会」(Zoom, 非ユーザーも自由に参加できる)
 - 素晴らしく充実した「移行ポータルサイト」, 各種講習会
 - https://jcahpc.github.io/gpu_porting/
- 「サポート移植(Supported Porting)」, 2022年10月開始
 - 多くのユーザーを有するコミュニティコード(19種類, 次頁), OpenFOAM(NVIDIA)
 - 外注のための予算も確保(富士通が担当する予定)
 - 「サポート移植」グループメンバー(主に若手)はハッカソン・相談会にも積極的に参加
- 基本的にOpenACC/StdPar(Standard Parallelism)推奨



Category	Name (Organizations)	Target, Method etc.	Language
Engineering (5)	FrontISTR (U.Tokyo)	Solid Mechanics, FEM	Fortran
	FrontFlow/blue (FFB) (U.Tokyo)	CFD, FEM	Fortran
	FrontFlow/red (AFFr) (Advanced Soft)	CFD, FVM	Fortran
	FFX (U.Tokyo)	CFD, Lattice Boltzmann Method (LBM)	Fortran
	CUBE (Kobe U./RIKEN)	CFD, Hierarchical Cartesian Grid	Fortran
Biophysics (3)	ABINIT-MP (Rikkyo U.)	Drug Discovery etc., FMO	Fortran
	UT-Heart (UT Heart, U.Tokyo)	Heart Simulation, FEM etc.	Fortran, C
	Lynx (Simula, U.Tokyo)	Cardiac Electrophysiology, FVM	C
Physics (3)	MUTSU/iHallMHD3D (NIFS)	Turbulent MHD, FFT	Fortran
	Nucl_TDDFT (Tokyo Tech)	Nuclear Physics, Time Dependent DFT	Fortran
	Athena++ (Tohoku U. etc.)	Astrophysics/MHD, FVM/AMR	C++
Climate/ Weather/ Ocean (4)	SCALE (RIKEN)	Climate/Weather, FVM	Fortran
	NICAM (U.Tokyo, RIKEN, NIES)	Global Climate, FVM	Fortran
	MIROC-GCM (AORI/U.Tokyo)	Atmospheric Science, FFT etc.	Fortran77
	Kinaco (AORI/U.Tokyo)	Ocean Science, FDM	Fortran
Earthquake (4)	OpenSWPC (ERI/U.Tokyo)	Earthquake Wave Propagation, FDM	Fortran
	SPECFEM3D (Kyoto U.)	Earthquake Simulations, Spectral FEM	Fortran
	hbi_hacapk (JAMSTEC, U.Tokyo)	Earthquake Simulations, H-Matrix	Fortran
	sse_3d (NIED)	Earthquake Science, BEM (CUDA Fortran)	Fortran

まとめ

- 次世代計算基盤 → 次世代HPCIへの期待
 - **継続性**: 重要なインフラであり、様々なアーキテクチャの計算資源、ストレージの提供を将来にわたって担保することが求められる
 - 端境期をどのように乗り越えるかが課題
 - 運用コスト削減・カーボンニュートラルに向けて、電力、冷却設備の最適化
 - 将来に向けて継続的に検討する体制
- フラグシップシステムだけでなく**基盤センターシステム+アカデミッククラウド+ストレージ**の充実も、複数システム間・異種環境との接続とワークフロー
- セキュリティ: ガイドラインの整備、セキュリティ・プライバシーに配慮した環境の提供
- ユーザビリティの向上: 共通UI、認証基盤の整備