資料4-1

運用FS検討報告

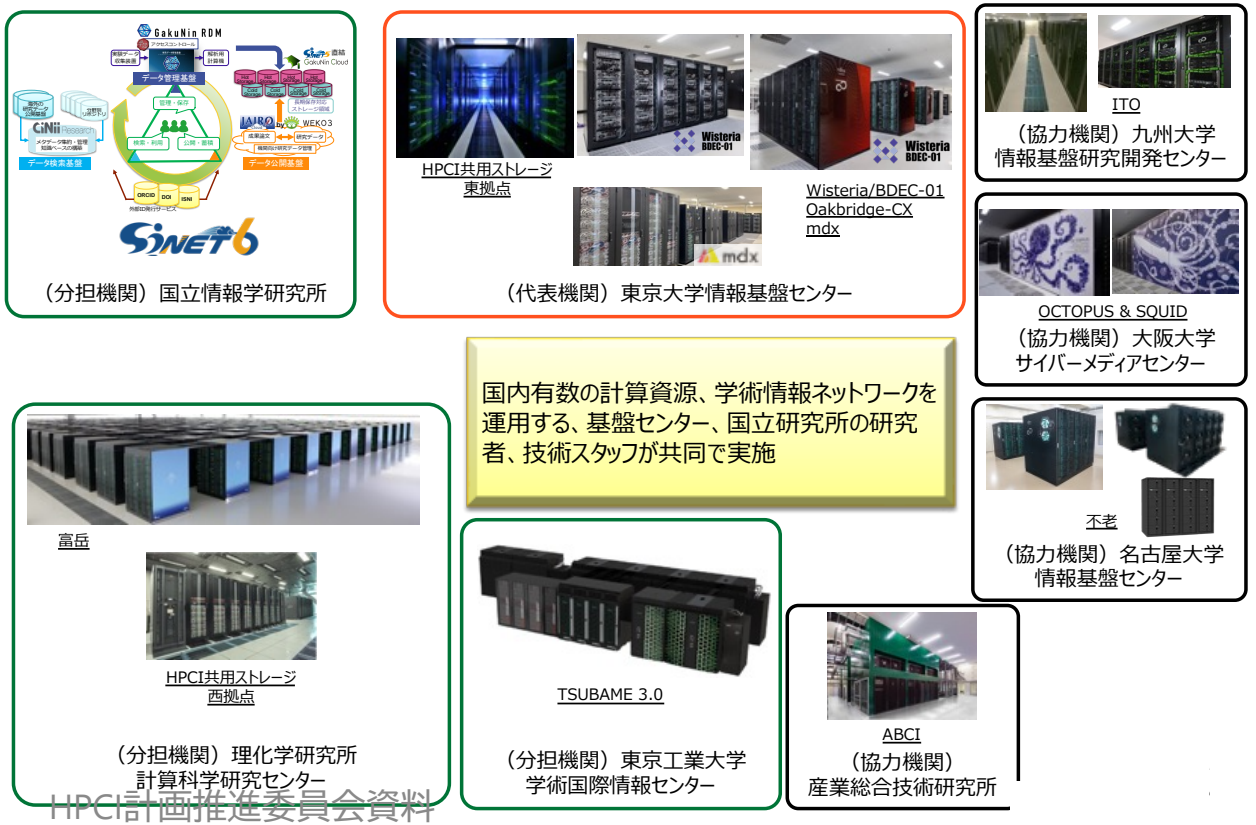
代表： 塙　敏博
東京大学　情報基盤センター

「運用技術調査研究」目的

- 「富岳」ならびにHPCI第2階層システム群、mdxをはじめとする各種データプラットフォーム、GakuNin RDMなどの研究データ基盤、それらを統合する学術情報ネットワークといった多種多様なシステムの設計、運用の知見を集約し、これらがより有機的に結合した、持続可能な次世代計算基盤の実現に向けた検討を実施する。
  - 様々な研究者のニーズに対し、求める適切な資源を提供し、平易で柔軟かつシームレスな利用の実現
    - システム調査研究チームと連携してそれを支えるツールの研究開発について検討する。
  - Society5.0の推進, SDGsの達成に貢献するプラットフォームとして、ひいては国内研究者全般の研究DXに資する共通インフラとして提供することを目指す。

1.	Society5.0運用	4.	カーボンニュートラル
2.	資源管理	5.	データ利活用
3.	施設・設備	6.	HPCI運営

- 以下の3つの層に分けて議論
  - 設備・電力、カーボンニュートラル関連のインフラ的側面
  - 資源管理やSociety5.0、データ利活用などのシステム設計
  - HPCIなどの運用ポリシー、ユーザ視点
- 各メンバーは参加グループで主体的に活動するが、他のグループにも積極的に参加し連携して調査研究を行う。



HPCIの概要

（2023年4月時点）

資料提供：RIST

High Performance Computing Infrastructure
（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）の略

国内の大学や研究機関 の 最先端のスパコンやストレージを
高速ネットワークSINET6で接続し 一体的な利用を可能し
産業界や学術界の方に広く提供

フラグシップシステム

理化学研究所
「富岳」
（CPU: Arm）

第２階層計算機システム

11機関＋2機関（2023.10～）
Arm（「富岳」と同じ）、x86、GPU、ベクトルで
多様なニーズに応える

北海道大学
Grand Chariot（CPU:x86）
Polaire（CPU:x86）

産業技術総合研究所
ABCI 2.0（GPU）

東北大学
AOBA-A（ベクトル）
AOBA-B（CPU:x86）
AOBA-S（ベクトル） 2023/8～

筑波大学
Cygnus（GPU）
Pegasus（GPU）

大阪大学
OCTOPUS（CPU:x86/GPU）
SQUID（CPU:x86/GPU/ベクトル）

最先端共同HPC基盤施設
（JCAHPC）・東京大学
Wisteria/Odyssey（CPU:Arm）

理化学研究所
共用ストレージ西拠点

東京大学
Oakbridge-CX（CPU:x86）
Wisteria/Aquarius（GPU）
/ 共用ストレージ東拠点

統計数理研究所
データ同化ｽｰﾊﾟｰｺﾝﾋﾟｭｰﾀ
（CPU:x86）
2023.10～

九州大学
ITO-A（CPU:x86）
ITO-B（GPU）

京都大学
Camphor3
（CPU:x86）
2023.10～

名古屋大学
スーパーコンピュータ「不老」
Type Ⅰ（CPU:Arm）
TypeⅡ（GPU）

東京工業大学
TSUBAME3.0（GPU）

海洋研究開発機構
地球シミュレータ（ES4）
（ベクトル／CPU:x86）

HPCIの運営

- 「富岳」(特定先端大型研究施設):
  - 理研:「特定先端大型研究施設の共用の促進に関する法律(共用法)」に基づく設置者
    - 開発、維持管理等
  - 高度情報科学技術研究機構(RIST): 登録施設利用促進機関(登録機関)
    - 利用促進業務(利用者選定および利用支援など)
- 文部科学省委託事業「HPCIの運営」
  - 委託事業代表機関:RIST、分担機関:理研、東大、筑波大、NII、FOCUS
    - 「富岳」以外のHPCI運営
- **HPCIコンソーシアム**: 整備・運用方針や我が国の計算科学技術の振興策並びに将来のスーパーコンピューティング等について検討し、国や関係機関に提言
  - ユーザコミュニティ、資源提供機関等
- 学際大規模情報基盤共同利用・共同研究拠点(JHPCN): 8国立大学のネットワーク型拠点、HPCIで資源割り当て
  - 北大、東北大、東大、東工大、名大、京大、阪大、九大

RIST: ポータルサイト運営、ヘルプデスク等の窓口業務、利用者選定、利用支援

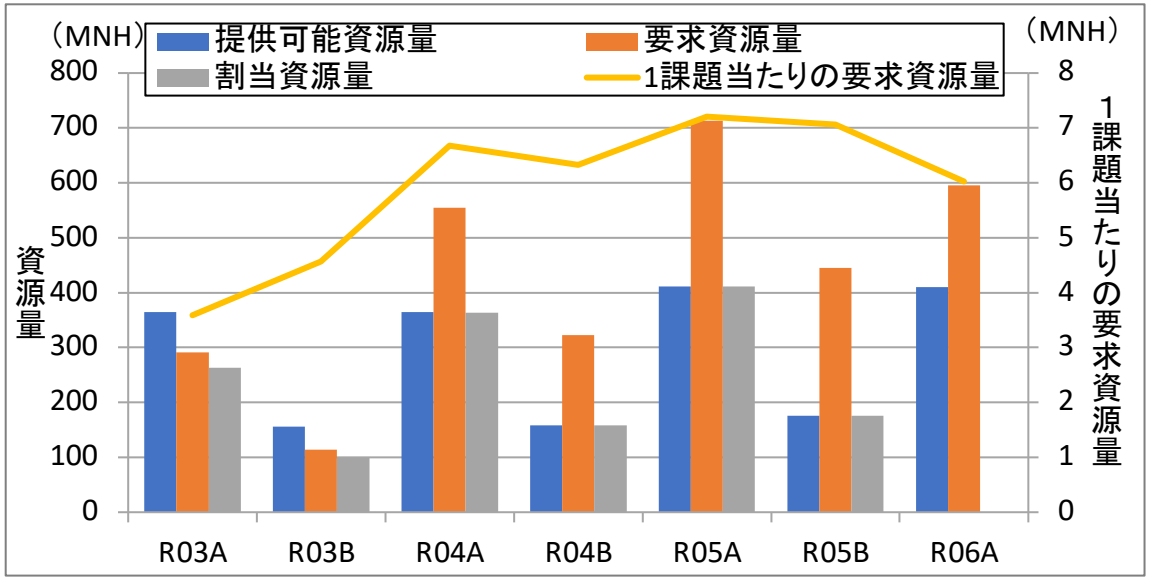
HPCIの構成

- 文部科学省委託事業「HPCIの運営」の委託事業代表機関 ➔ RIST
- HPCIシステムを構成する計算機資源
  - HPCI全体で共通運用され、かつ、文部科学省委託事業「HPCIの運営」の委託事業代表機関として実施する一括した課題選定の対象とする、計算機資源(共用計算資源)
    ➔ R-CCS, 9大学+JCAHPC, JAMSTEC, 統数研
  - HPCI全体で共通運用されるが、文部科学省委託事業「HPCIの運営」の委託事業代表機関として実施する一括した課題選定の対象とせず、各機関独自のルールで利用に供する計算機資源 ➔ AIST
- 認証基盤及び高速ネットワーク ➔ NII
- 共用ストレージ ➔ 理研R-CCS, 東大
- プライマリセンター：HPCIアカウント発行管理業務
  ➔ 上記の「計算機資源を提供するシステム構成機関」提供機関
- 最寄りセンター：対面認証を実施 ➔ ほぼプライマリセンターと同じ+RIST

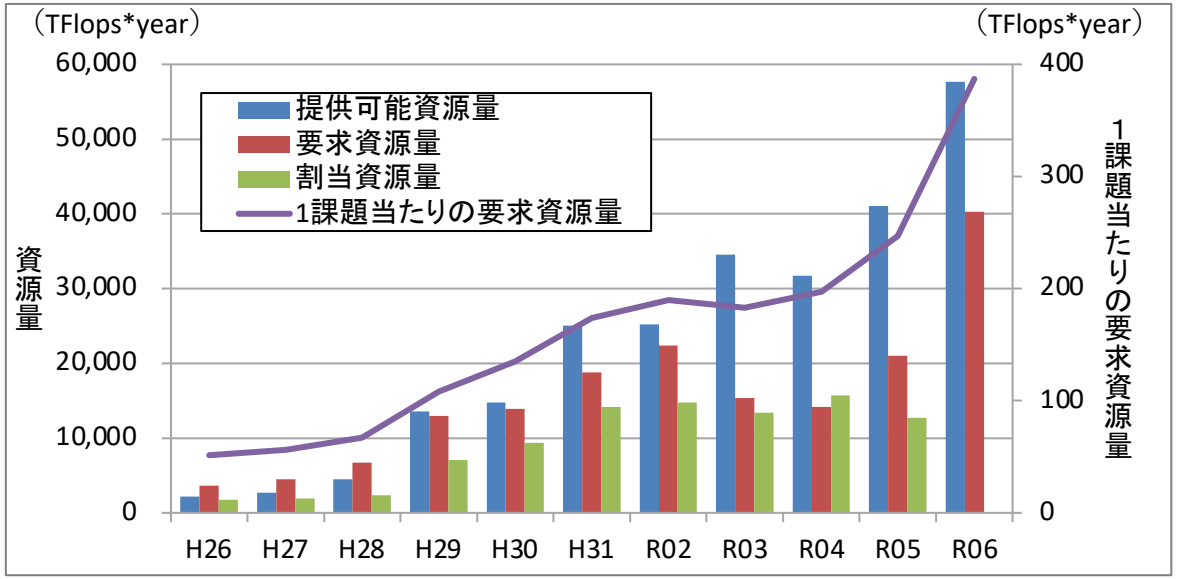
定期募集における要求・割当資源量の推移

資料提供:
RIST+補足

「富岳」(1MNH=約386TFlops*year)

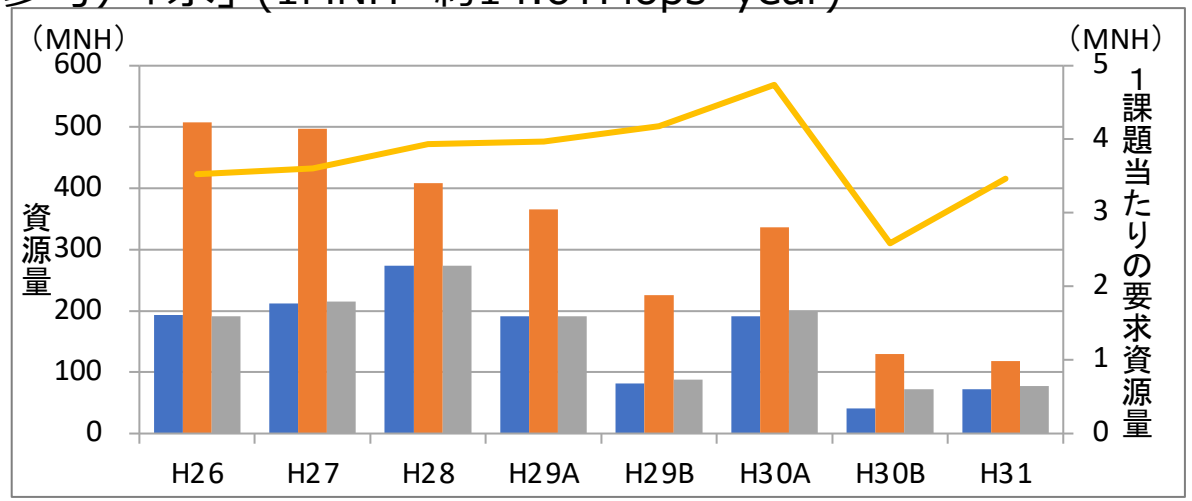


HPCI共用計算資源（第二階層）



※H31、R02年度の提供可能資源量には成果創出加速プログラム課題（ポスト京課題）を含む

（参考）「京」(1MNH=約14.6TFlops*year)



※計算ノード単体における演算性能は、「富岳」3.3792TFlops（ブーストモード：倍精度）、「京」128GFlops

（凡例）

○「富岳（京）」（単位はノード＊時間（百万））
提供可能資源量：HPCI計画推進委員会の審議による「富岳（京）」の資源配分比率に基づき、運用機関より定期募集毎に課題（一般、産業）に提供される資源量の総和
要求資源量　：「富岳（京）」の利用を希望する課題の要求資源量の総和
割当資源量　：「富岳（京）」の利用に採択された課題の割当資源量の総和

○HPCI共用計算資源（第二階層）（単位はTFlops＊年（演算量換算））
提供可能資源量：第二階層資源提供機関から提供可能とされた資源量の総和
要求資源量　：第二階層資源を第一希望とする課題及び「富岳」との同時利用として第二階層資源（第一希望）を希望する課題の要求資源量の総和
割当資源量　：第二階層資源の利用に採択された課題の割当資源量の総和

HPCI資源提供機関としての役割

例：東大(+JCAHPC)の場合, 2023年度

- HPCIシステムを構成する計算機資源提供
  - Oakbridge-CX: 200ノードx半年相当
  - Wisteria/BDEC-01 Aquarius: 4ノード年相当
  - Wisteria/BDEC-01 Odyssey ➔ JCAHPCとしての資源提供 2,304ノード年相当
- HPCI共用ストレージ
  - 約45PB
- **プライマリセンター：IdPサーバ管理等**
  - 東大情報基盤センター
  - (HPCI共用ストレージ)
  - (JCAHPC)
- 最寄りセンター
  - 東大情報基盤センター
- SINET6接続 (HPCIのための専用回線)
  - 100G x2

一般利用課題と同じ利用負担金

一般利用課題と同じ利用負担金（筑波大+東大で応分に充当）

補正予算(設備)/「HPCIの運営」委託費（再委託）

赤字：HPCIの運営に(持ち出しで)協力

次期HPCIの運営に向けて：認証基盤

- 現状：プロトコル変更作業が進行中
  - これまで： GSI (Grid Security Infrastructure)認証、代理証明書

GSIを提供するGlobus Toolkitのサポート終了

  - 国際標準の認証：OAuth2.0、認可: OpenID Connectへ、2024年度に本格移行予定
- 認証基盤の学認フェデレーション(GakuNin)との共通化
  - 上記移行により技術的には可能：（学術）ユーザーにとってアカウント管理が容易に、企業ユーザにも対応可能 (gBizIDなど)
  - 各HPCI資源提供機関が別途IdPやプライマリセンターを運用するコストを削減
  - 一方、各参加組織の認証レベル(IAL, AAL)がHPCI運用のセキュリティレベルと合うかどうか、は課題、利用時に確認する仕組みが要るはず

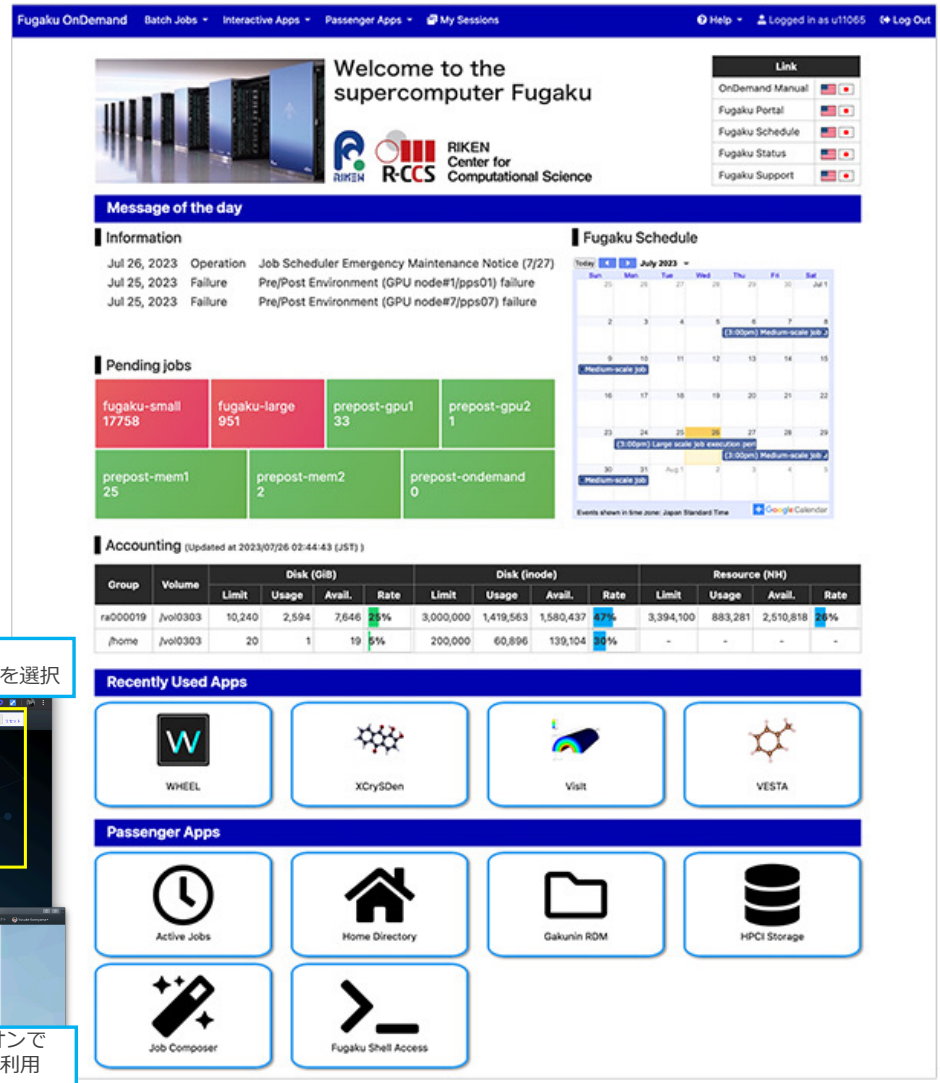
➔HPCIの資源だけでなく、他の研究管理基盤、ストレージ等のサービスとの連携には必要不可欠、整備が重要

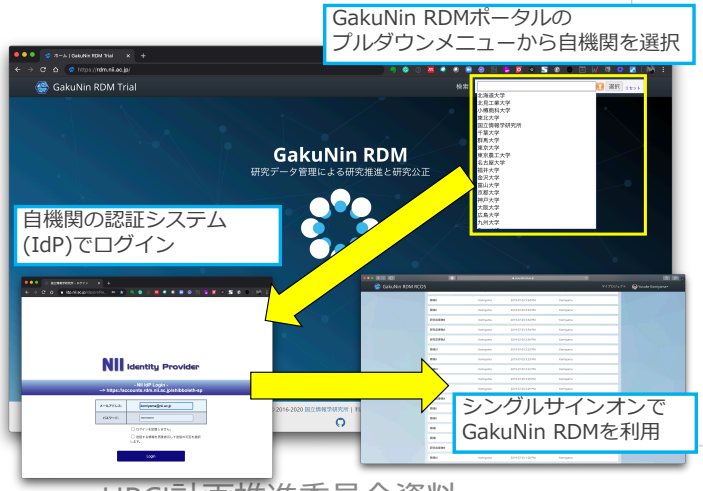
次期HPCIの運営に向けて：セキュリティ

- 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」
  - 必須項目 / 推奨項目
- 資源提供機関のポリシーに依存する部分
  - 共通化するためにレベルの低いポリシーに合わせなければいけなかった（仕方なく推奨にしている項目もあり）
- セキュリティガイドライン：「セキュリティ要件」のアップデート
  - NII を中心に策定中： 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
    - 昨年度 NIST Cyber Security Framework (CSF)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」に実施
    - 侵入や内部不正を前提として被害を最小限に抑える「サイバーレジリエンス」も考慮すべき
  - ソフトウェアの安全性チェックも必要
    - 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

次期HPCIの運営に向けて：ユーザビリティ

- 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
- 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で導入、「富岳」で利用可能
- 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

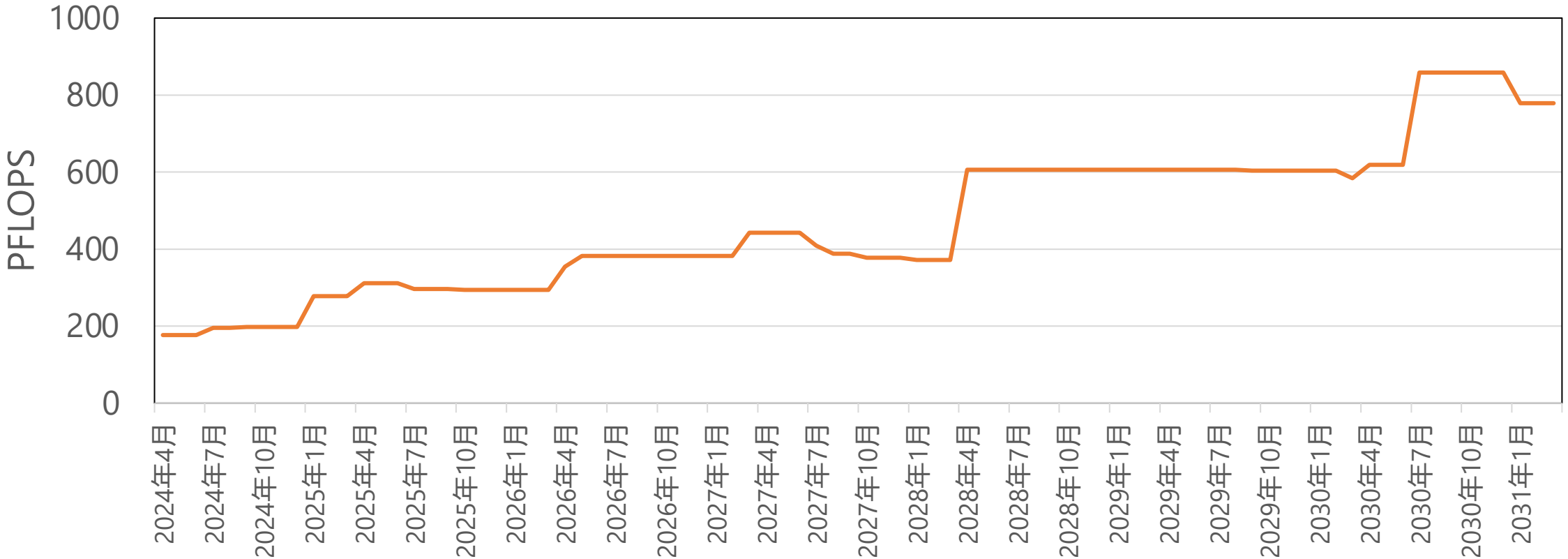




- 「富岳」Open OnDemandにおけるGakuNin RDMとのデータ転送アプリケーションの開発 (2023/7/26)
- 「富岳」Open OnDemandにおけるHPCI共用ストレージとのデータ転送アプリケーションの開発 (2023/9/7)

計算基盤整備に向けた調査

- HPCI第2階層: ここでは9大学基盤センターの導入計画について調査(合計値)
  - 2027年頃までの運用計画は決まりつつある、2028年以降は不透明



- EUの動向の調査(委託業務)
  - 基盤整備のポリシー, 登録機関業務についての調査

次世代計算基盤の整備に向けて（データ基盤・利活用）

- データ利活用とアカデミッククラウドの重要性
  - 様々な計算資源・エッジデバイス・ストレージの間をつなぐ上で、VMの柔軟性は重要
    - データ利活用の自由度の観点から、商用クラウドに頼らずアカデミアでストレージを担保すべき
  - セキュリティ・プライバシに配慮したストレージも必要、TEE (Trust Execution Environment)
  - mdxでは運用費の担保はない、利用負担金(＋運営機関の持ち出し)
    - 今後の継続に向けて各センターからのendorseはある

次世代計算基盤の整備に向けて（基盤センター群）

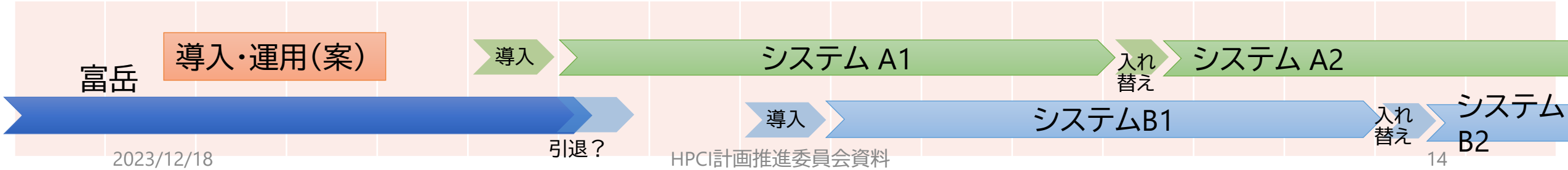
- 基盤センターのシステム導入予算
  - 大学予算の一部、そもそもセンターに独立して予算が担保されるわけではない
  - 独法化後、運営費交付金の削減➔センター予算にも波及
  - 新たな整備、保守等の出費増(WiFi設備、クラウド等整備のライセンスなど)
  - 電気代高騰、物価上昇、円安など外的要因

- フラグシップだけでなく**第2階層システム＋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、量子との連携も
  - ユーザー層拡大、手厚いユーザー支援、共同研究
- 基盤センター群の運営費負担増に対して**導入・運用のための予算配分が重要**
  - **技術革新の鈍化、物価高騰、為替レート悪化**によるシステム**価格の上昇**
  - **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**
  - 効率の良い冷却設備のための初期導入コスト（ランニングコストの圧縮効果とはリンクしない）

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(提案)
  - 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
  - 引き続き実現可能性を検討中

端境期をどう解消するのか？



次期(＋その先の)フラグシップシステム整備:施設

一方で

- 次期システム向けに既存の富岳の設備の大規模な改修が必要
  (理研R-CCS施設運転技術ユニットによる)
  - 冷却システムの主となっているコジェネレーションシステム(CGS)設備の寿命
  - 冷却配管の摩耗: 全交換が必要、並行配管には経路が不足
  - これらのリプレイス等には富岳**停止後2年ほどの工事期間**が必要
- 富岳と新システムを並行設置するには、現行の計算機棟では、追加設置場所・電力設備容量の点で不足

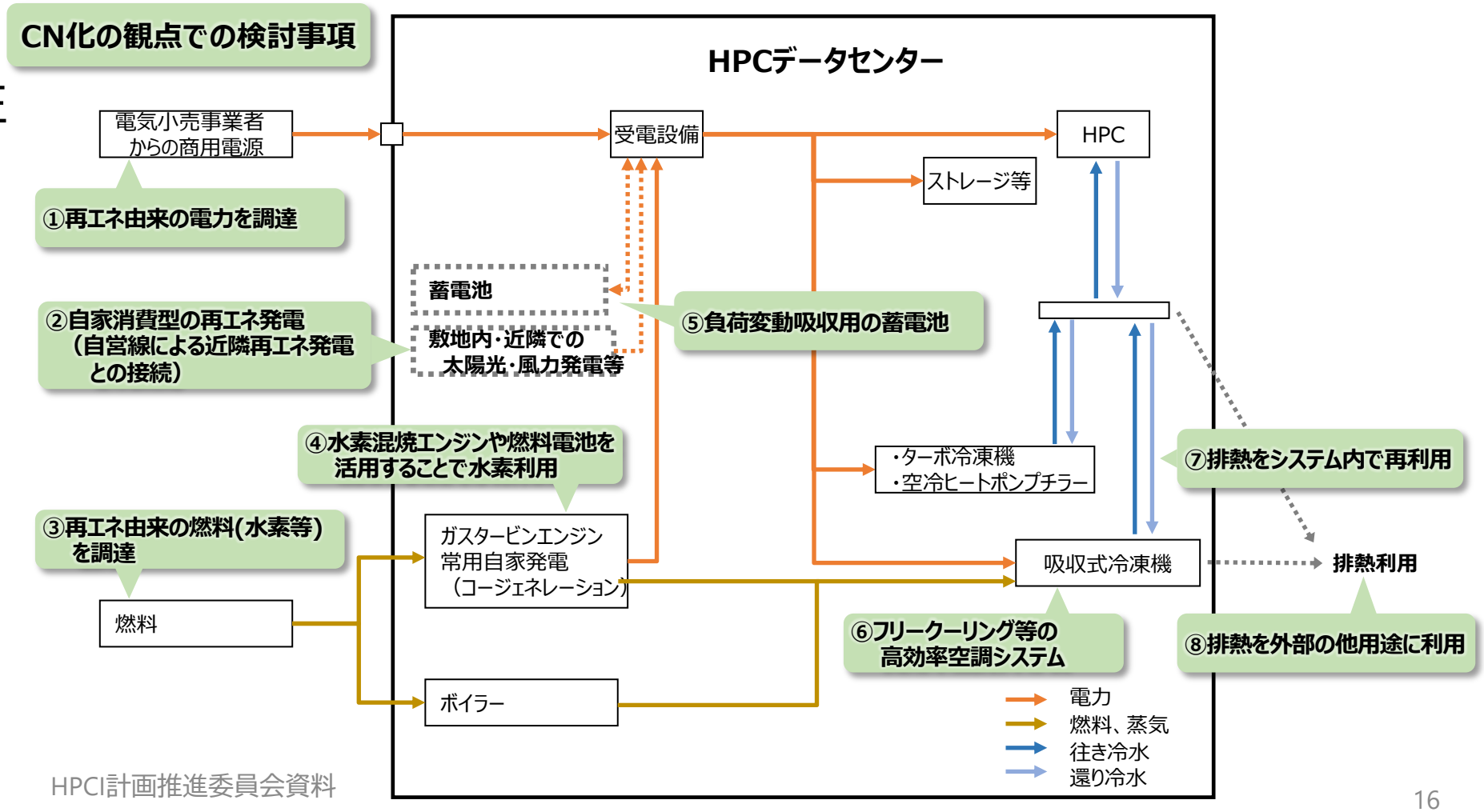
- 次期システムについて、計算ノードの検討だけでなく、建屋等の施設や冷却等の設備の拡大や増強についても、早急に検討を進めることが必要

次期(＋その先の)フラグシップシステム整備:カーボンニュートラル

- 政府目標: 2030年度において、温室効果ガス46%削減(2013年度比)を目指す(50%とも)、2050年度カーボンニュートラル(CN)化達成(ネットゼロ)
  - 政府の調達電力、再エネ35% (2023年度目標) ~ 60% (2030年度目標)

➔ 次期システムにおけるCNに向けた可能性を探る

- シナリオ: 富岳の現有設備をCN対応にするには?
  - ①~⑥までの実現可能性、⑦⑧の利用可能性を検討

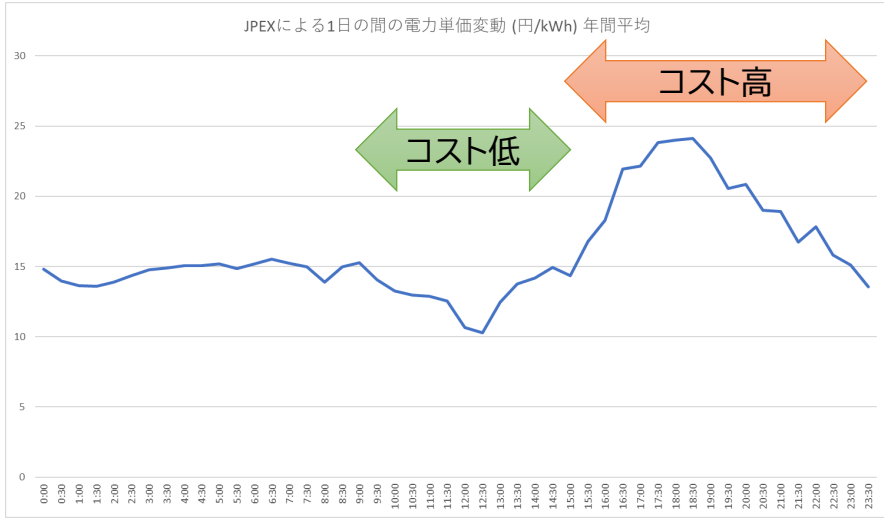


次期(＋その先の)フラグシップシステム整備：CN（続き）

- 定量的な検討については継続中
- 炭素排出量の削減候補
  1. 再エネ由来の電力購入… 可能性あり、ポートアイランドまで十分託送できるかどうか
  2. 自家消費型再エネ… 期待薄、コスト効率が悪い

  3,4.再エネ由来燃料(水素等)… コスト高で非現実的
  5. 負荷変動吸収用蓄電池… コストは高いが部分的には効果あり？(1,2と連携しないとあまり意味はない)
  6. 高効率冷却… 検討が必要、次ページへ

  7,8.排熱再利用、転用… R-CCS内での活用は困難



↑安価な電力の活用
(電力コスト削減)
- 蓄電池の利用
- 電力余剰時：稼働計算機資源を多く稼働、ひっ迫時：少なく
  9-15時 コスト低、
  15-24時 コスト高（年平均）

次期(＋その先の)フラグシップシステム整備:冷却システム

冷却条件が達成可能な水冷方式に限って検討、CGSは効率・導入コスト・運用コストの観点から検討から除外した

指標	チラーのみ	冷却塔+チラー/ チルドタワー	冷却塔	備考
達成可能PUE	× 1.3程度	△ 1.1強	○ 1.1以下	• 電力料金に直結 • 冷却塔はチラー比で**全消費電力が2割減**
温度追従性	× 「富岳」以上の負荷変動には追従不可	△ チラーが2割までなら冷却塔で吸収可	○	• 追従できない分は冷却電力が増加
生成可能な冷却水温度(夏季)	○ 20℃	△ 30℃	△ 32℃	• TSUBAME3やABCIは32℃で運用実績あり
保守コスト	× 費用大・人力監視	△ 人力監視/監視は不要	○	• チラーはある程度人力の監視が必要(追従性の問題)
導入コスト	×× 高価	×/×× 二重投資/高価	○	• チラーの容量あたり導入費用は冷却塔の数倍 • 冷却塔＋チラーは設置面積の点で困難

- 今後の低消費電力化を踏まえると、負荷変動幅は「富岳」よりも大 →冷却システムの追従性が重要
  - 圧縮機を用いるチラーは起動・停止に時間を要する(操作してから温度に反映されるまで数分程度)
    - ジョブの開始抑制のみの運用対応は既に限界: ジョブ実行中の電力変動には対処できない
  - 蓄熱槽外付けも考えられるが、設置要件やPUEがさらに悪化する

  →「富岳」以上の電力効率達成には、冷却塔のみで冷却可能な**32℃以上の冷却水温度に対応**
- 季節毎の冷却水の効率的な生産のため、**幅広い冷却温度(20～32℃程度)に耐える**システムが望ましい
- 冷却効率を高めるためには、**入出力温度差(Δt)を高温にする**ことも有効
  - 廃熱再利用には**排出水温度が高温**であることが望ましい (例: 温室利用なら40℃, Δt=8K程度)

スパコン調達スキームの改善(特にHPCI第二階層)

- **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(次世代テクノロジ実現のための開発支援)
    - NREは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」が近い: mdxで実施