

SPring-8データセンター構想

城地 保昌

理化学研究所 放射光科学研究センター

SPring-8サイトのDXにかかわる開発・整備

- リモート実験
- ビームライン制御の高度化
 - スマート化(AI等の導入)
 - 自動化
 - サイト内ネットワーク高度化(10 Gbps → 100 Gbps)
- SPring-8データセンター構想
 - 大規模データのデータ解析能力の強化(第1回WS、参加登録163名)
 - 研究データの共有・公開・再利用(第2回WS)
 - データマネジメント

Contents

- SPring-8におけるデータの状況とデータセンター構想
- SPring-8データセンターのシステム構成(検討中)
- SPring-8データセンター上のサービス(検討中)

SPring-8におけるデータの状況

- 各種実験技術の進歩に伴い、良質な実験データが大量にとれるようになりつつある
 - ✓ 光源・光学系・検出器の高度化により、爆発的に増大する見込み (>~10PB/年)
- SPring-8では、これまでBL毎にデータシステムを整備、個別運用
 - ✓ SACLAは、データセンターを設置済
- ユーザーはビームタイム後に実験データを持ち帰り、成果発表まで時間をかけてデータを解析している
 - ✓ データ解析が律速となって成果創出が遅延(解析律速)

→ 大規模なデータは移動に時間がかかりコストも高いので、近在した場所に置き、利活用するデータセンターを整備



実験データの分類とその課題

小 試料あたりのデータ量 大

少

TB

引き続き個別ビームラインの
データインフラで対応?

解析律速の実験が増大
企業ユーザーのニーズが拡大
→ データ解析リソース拡充&
解析フローの最適化

第1回ワークショップのテーマ

MDX等、データの再利用ニ
ーズへの対応が課題

→ 研究データ管理基盤の標
準化

第2回ワークショップのテーマ

ないと考えているが
データセンター構想で
対応するには工夫が必要

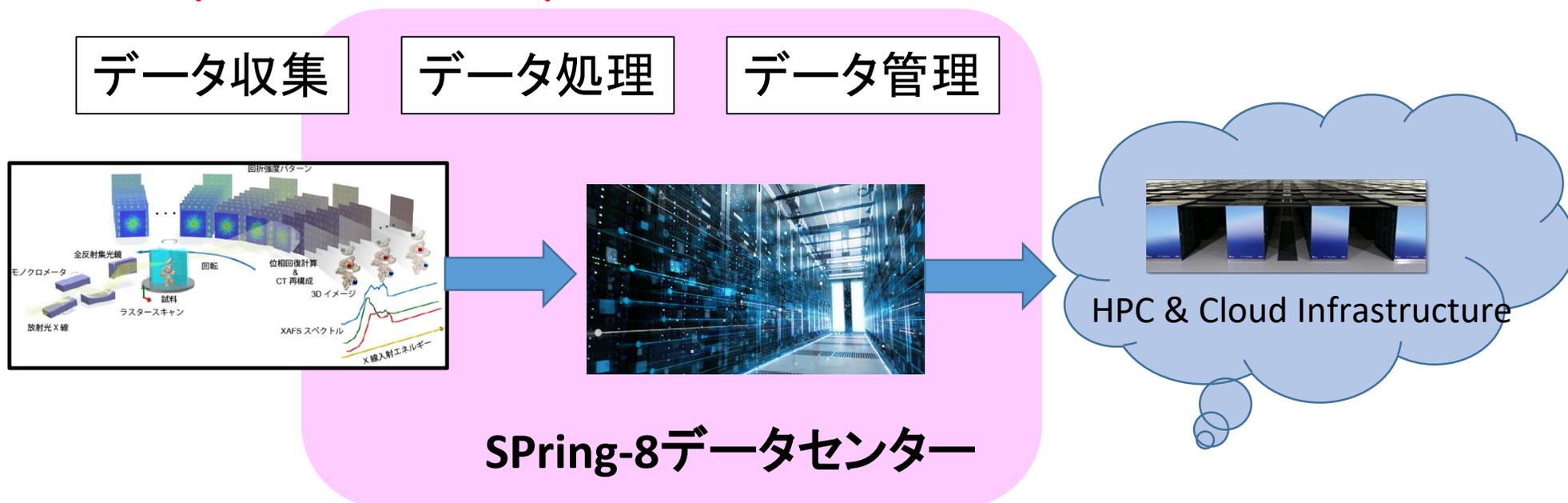
ビームタイムあたりの試料数

100

多

SPring-8データセンター構想(2023年夏~)

- ビームタイム中に実験データの品質を確認できるデータ収集・処理基盤(高品質データの効率的創出)
- 詳細データ解析等のために、スパコン・クラウドなどハイエンドな計算資源との有機的な結合を可能とするデータ基盤(データ利活用)



実験からデータ解析までの効率的なWorkflowの実現を支えるデータ基盤

解析律速の原因と対策

- 計算が遅い
 - ✓ 計算リソースの確保
 - ✓ 並列化など計算科学技術の適用
 - ✓ Workflow (pipeline)ツールの導入(SACLAではKMR活用の実績あり)

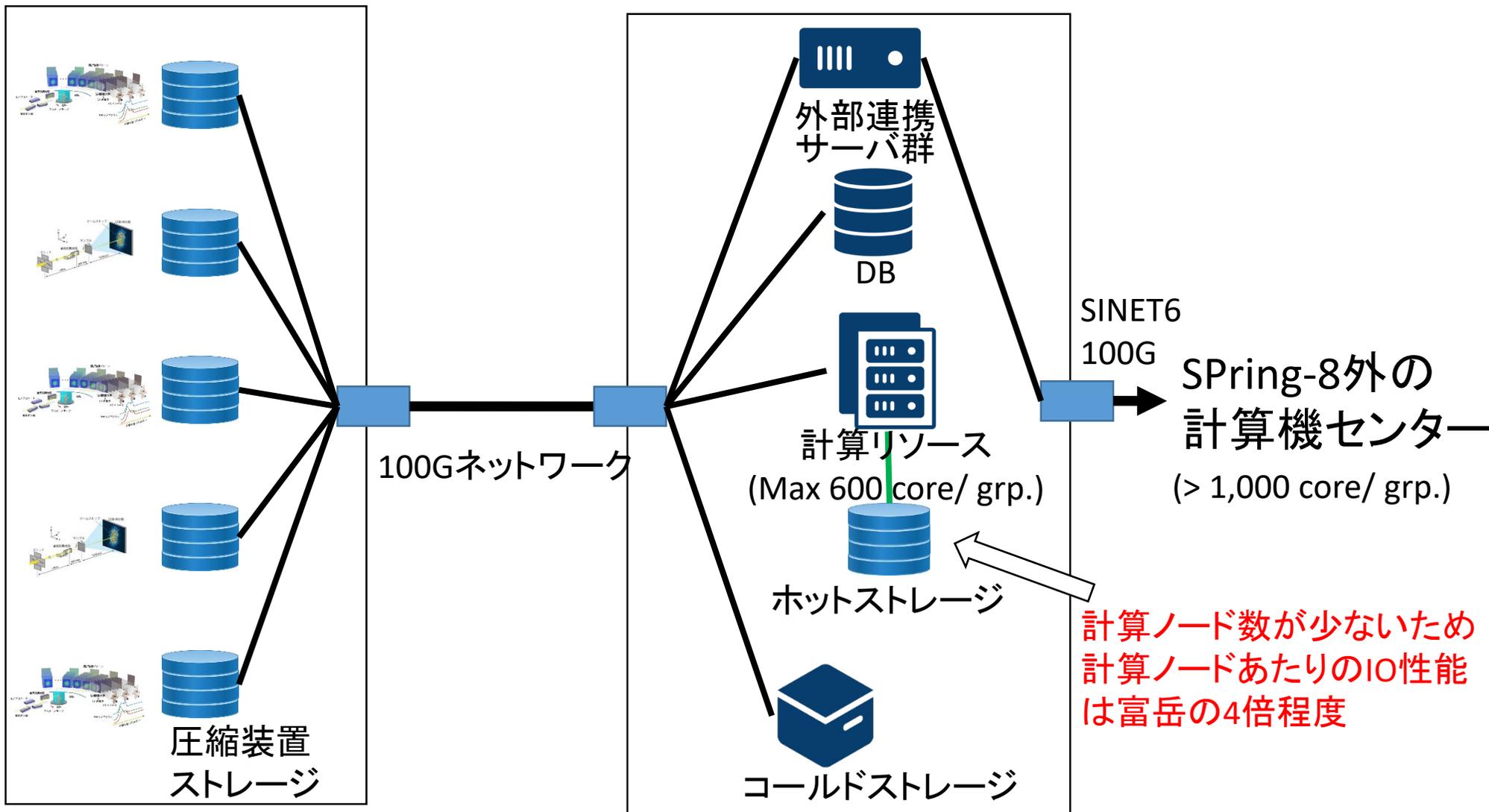


- 良い解析方法がない(試行錯誤に時間がかかる)
 - ✓ 新しい解析方法の探索
 - ✓ 統計・情報科学(狭義のデータサイエンス)との連携
 - ← 研究者間の共同研究促進のため、データの効率的なやりとりが必要

SPring-8データセンターのシステム構成(案)

ビームライン

電算機室



解析律速を計算科学により解消

SPring-8データセンターの性能(検討中)

[計算性能]

- CPU 5,120 core, 180 TFLOPS
- GPU 16 基, 320TFLOPS(FP32), 160TFLOPS(FP64)

(参考)富岳 7.6 M core, 488 PFLOPS

[ストレージ容量]

- ホットストレージ >10 PB (データ保持期間3年?)
- コールドストレージ 小規模から始めて段階的に増強

(参考)

- 富岳データ解析用ストレージ 150 PB
- HPCI共用ストレージ 45 PB (東西拠点間二重化)

SPring-8データセンター上のサービス(検討中)

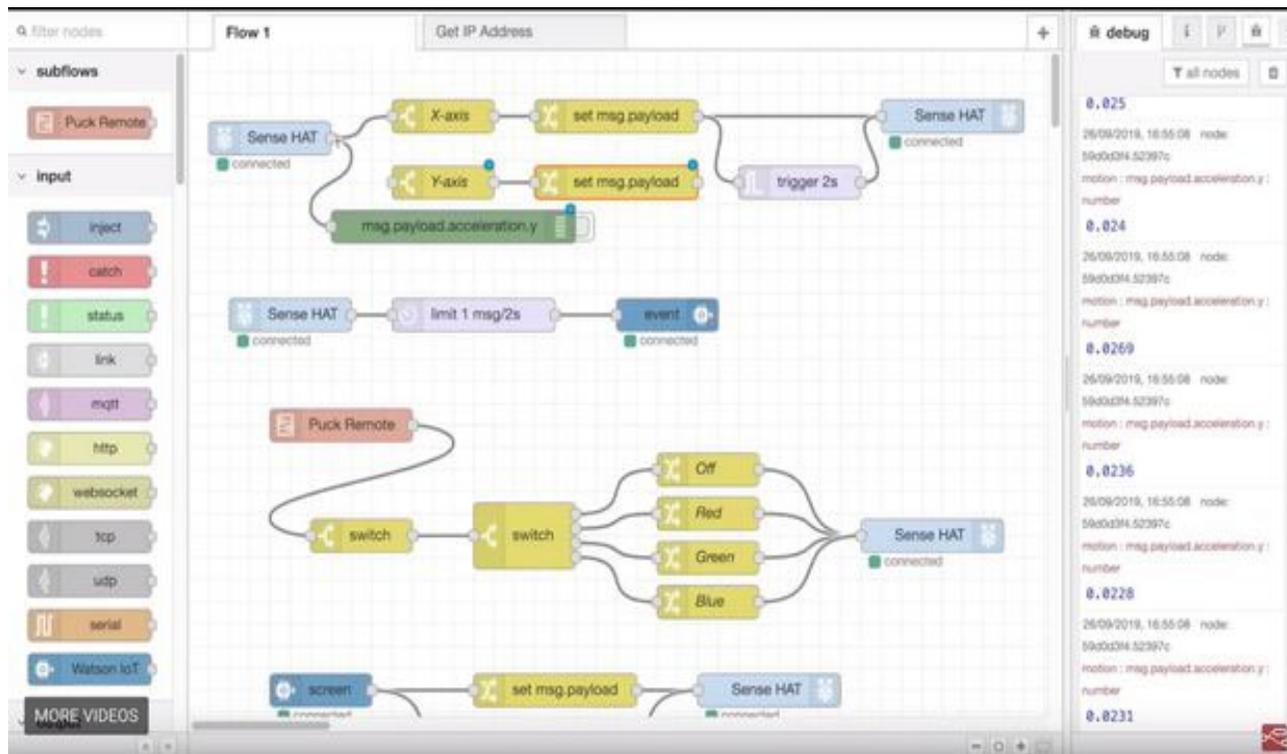
- ビームラインからデータセンターへの高速自動データ転送
- ソフトウェア開発の負担軽減。
- データセンターへの効率的・簡便なアクセス
 - ✓ 外部からのGUI操作の高速化
 - ✓ Linuxに不慣れなユーザーへの対応
- データ共有・持ち帰り(所属機関、クラウド) 機能
 - ✓ オンラインストレージサービス
- 外部計算資源(富岳等HPCI)とのシームレスな接続による大規模・高速データ解析
- データサイエンスとの連携を念頭においたデータの共有・公開機能(第2回WSのテーマ)
 - ✓ 国立情報学研究所が提供する研究データ管理サービスとの連携を検討中

ユーザーがデータセンター内のシステム構成図を意識することを出来る限り少なくできるサービスを実現したい

標準的なデータ解析の効率化に向けて

- 標準的な処理・解析モジュール(要素)は施設側で開発予定
- ユーザーはWorkflowツール(下図参照、検討中)で、各要素を接続し、データ解析を実行
- ユーザー開発のモジュールも追加できるようにしたい
- 2024年度中のサービス開始を目指す(2023年夏~は計算科学者の協力を得て対応)

実現のためには皆様からのインプットが重要です
ご協力よろしくお願ひいたします

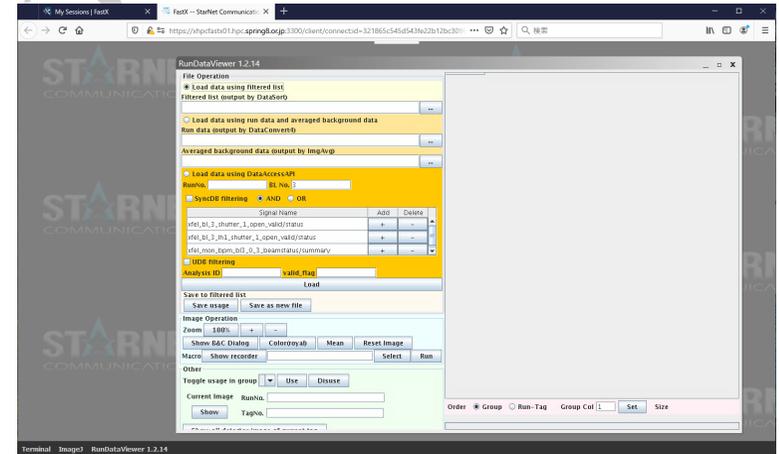


(Workflowツールによるノーコード開発の例)

SPring-8外からのデータセンターへのアクセス機能

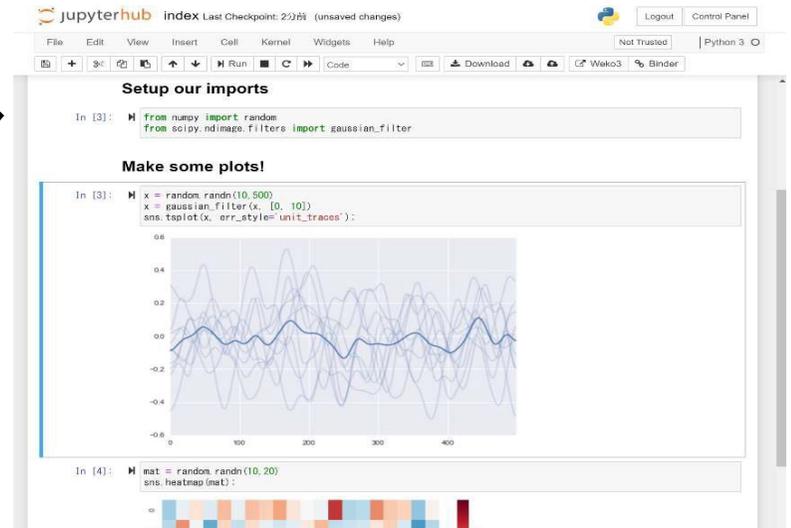


- GUI動作の高速化のためにFastXを導入予定 (SACLAで運用中)
 - ✓ ユーザーは特別なソフトウェアをインストールすることなしにWebブラウザからデータセンターにログイン可能



(FastXによるSACLA GUIへの接続例)

- Jupyter等の環境(R&D中)
 - ✓ バッチジョブ等に不慣れなユーザーが、自分のPCのWebブラウザから計算ノードのJupyter Notebook等を起動可能
 - ✓ PythonスクリプトをWebブラウザ上で記述・実行し、データセンター内のデータを解析



(Jupyterによる解析の例) ¹²

データ共有・持ち帰り(所属機関、クラウド) 機能の強化

- 実験データを外部の研究者と共有するためのオンラインストレージの導入: Amazon S3互換サービス (R&D中)
 - ✓ APIによるデータアクセス
 - ✓ S3接続に対応したソフトウェア(S3 Browser、WinSCP等)によるデータ転送

計算機(施設外)

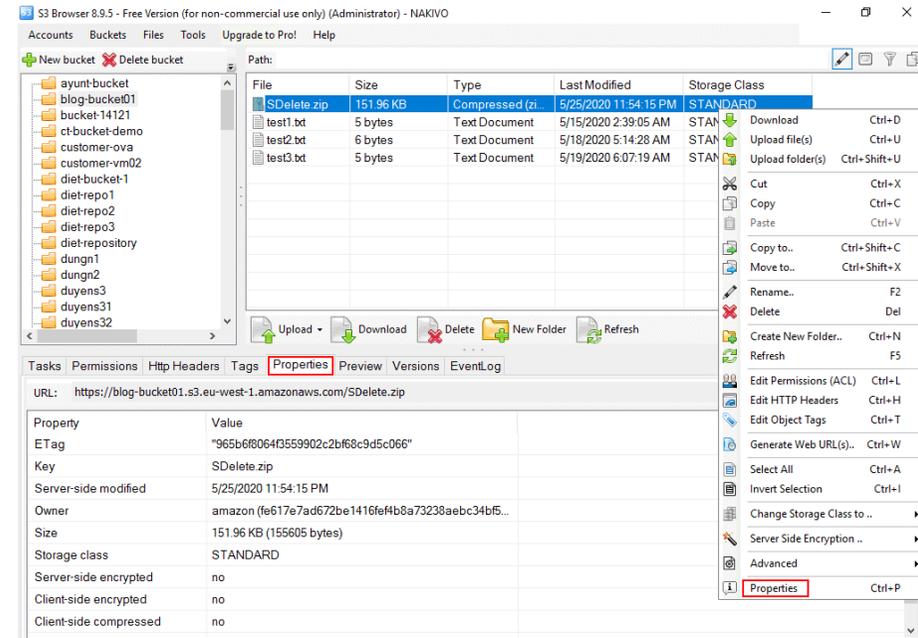
アプリケーション

API

HTTP/REST

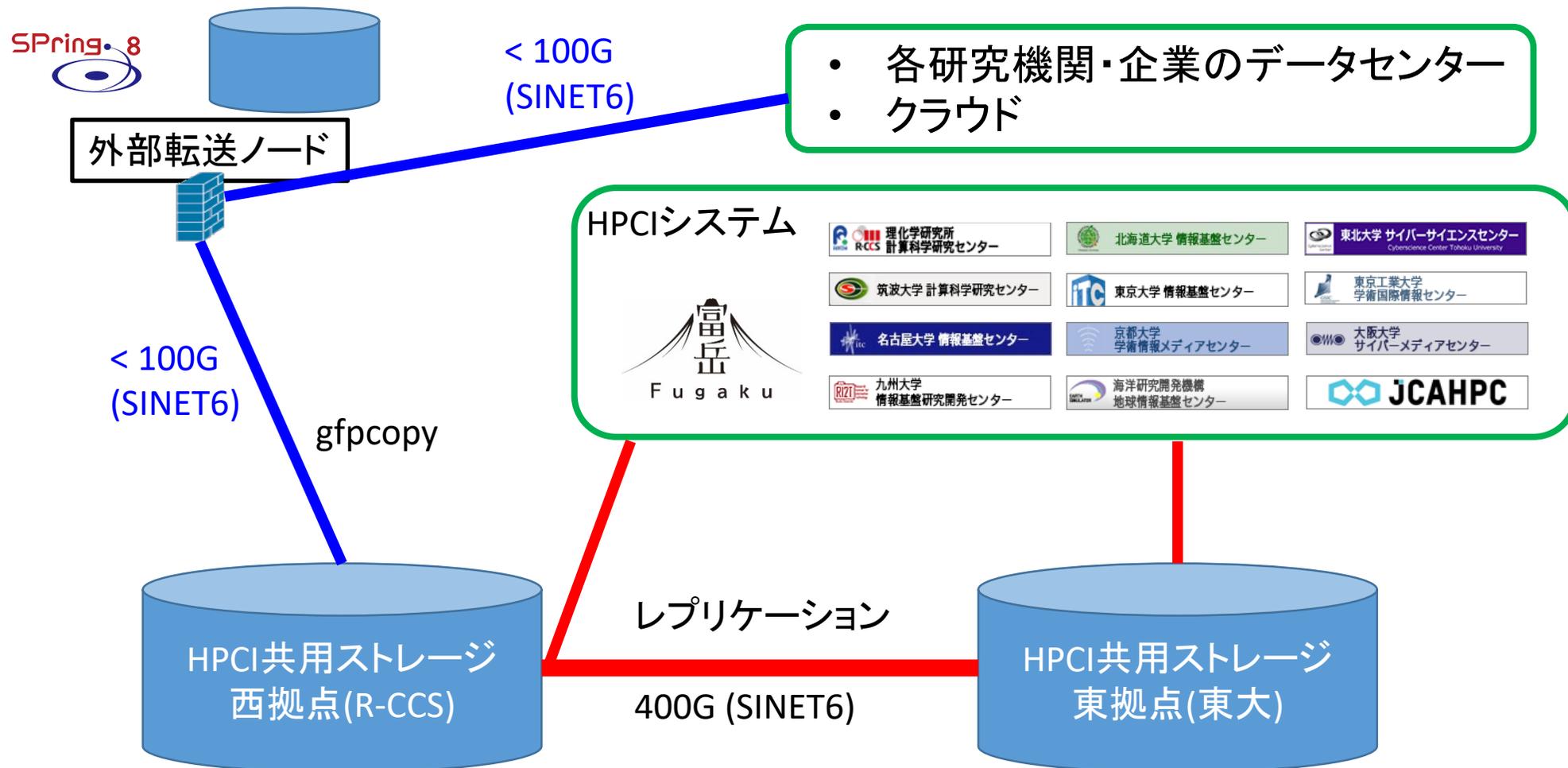
オブジェクトストレージサービス

SPRING-8データセンター



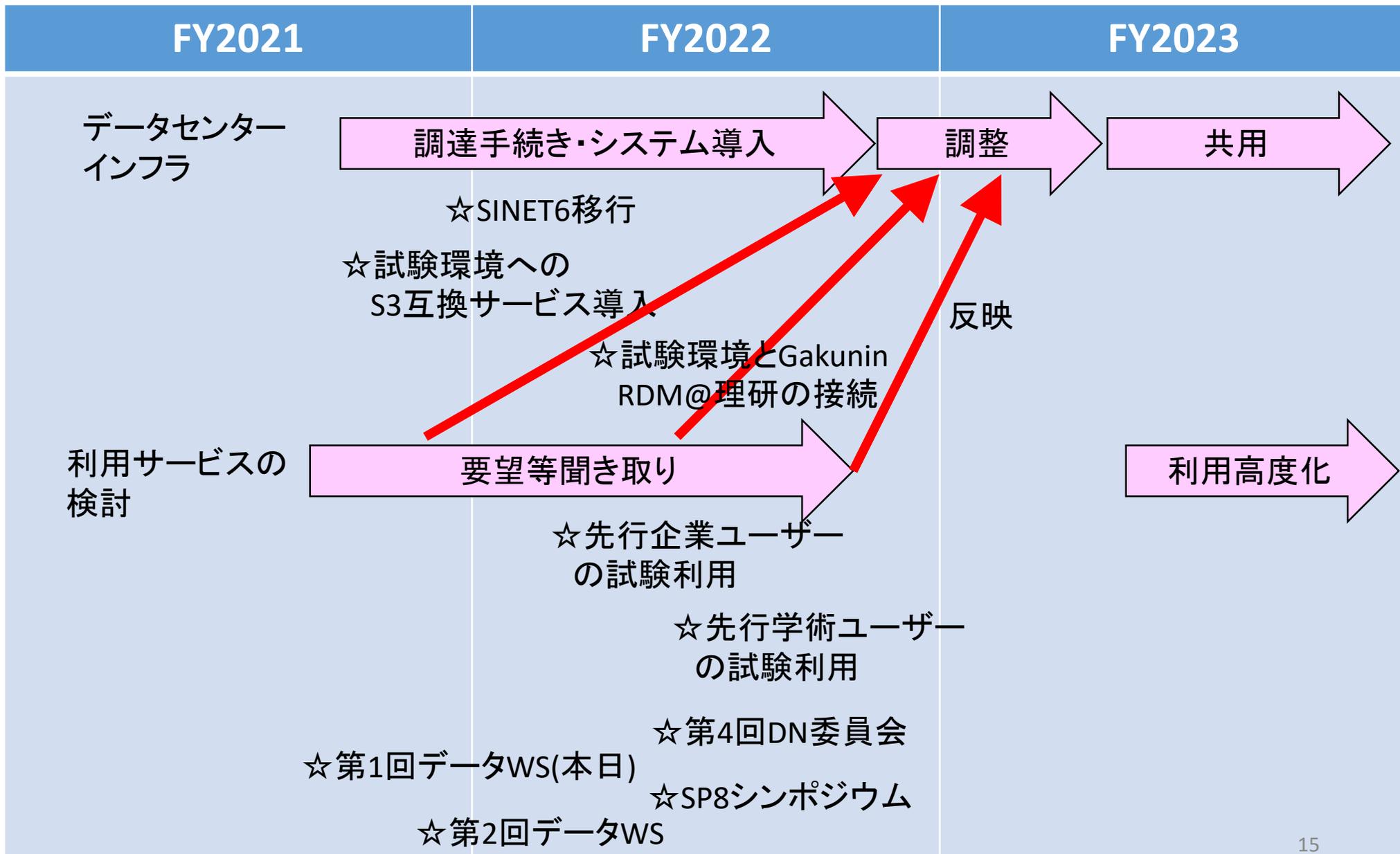
S3 Browserによるファイルアクセス例

外部計算資源への高速データ転送サービス



理研計算科学研究センター(R-CCS)との共同研究によりSACLAに導入済
SPring-8データセンターにも外部転送サービスを導入予定

今後のスケジュール



SPring-8データセンターに関する情報公開

- <https://dncom.spring8.or.jp/>
 - ✓ データ・ネットワーク委員会HPを、SPring-8のデータ・ネットワークの運用全般に関わるポータルサイトとしてリニューアル予定
 - ✓ 第3回データ・ネットワーク委員会(2021年12月1日) データセンター利用規程を議論(12月24日確定予定)
 - ✓ データワークショップ等の情報も掲載予定
- ご意見のある方は、是非ご連絡をお願いします。
dncs@spring8.or.jp

まとめ

- SPring-8データセンター(DC)を設置し、データ解析が成果創出の律速となる解析律速の問題に対応
- 2022年度中のDCインフラ導入に向け、構成を検討中
- 2023年度夏のDC共用開始に向けてサービスを検討中
 - ✓ Workflowツールの導入等によるソフト開発の負担軽減
 - ✓ DCへの効率的・簡便な接続サービス
 - ✓ データ共有機能の強化
 - ✓ 外部計算機資源へのデータ転送の効率化
 - ✓ データ圧縮、認証等の基盤技術

データ解析の効率化により、成果創出を加速していきたい

謝辞

[SPring-8データセンター構想]

中嶋享(JASRI), 本村幸治(理研), 平木俊幸(理研), 杉本崇(JASRI),
山鹿光裕(JASRI), 初井宇記(理研), 矢橋牧名(理研),
中町将貴(理研)、渡邊一輝(理研)

[HPCI連携、データ圧縮、Workflowツール、認証]

理研計算科学研究センター(R-CCS, 富岳)

松田元彦, 原田浩、金山秀智, 山本啓二, 佐藤賢斗,
佐野健太郎, 庄司文由, 松岡聡

[研究データ管理]

理研情報統合本部(R-IH)

實本英之, 富士健太郎, 美濃導彦

国立情報学研究所(NII)

込山悠介, 山地一禎, 喜連川優