スーパーコンピュータ「富岳」での3年間のファイルシステム運用の振り返り

国立研究開発法人 理化学研究所 計算科学研究センター（R-CCS）
運用技術部門 システム運転技術ユニット 技師
末安 史親（Fumichika Sueyasu）

発表内容

- はじめに
- 「富岳」概要
- ファイルシステム運用で経験した困りごと3選
- おわりに

はじめに

- スーパーコンピュータ「富岳」は2021年3月より共用運用を開始
- その前の2020年4月から一部計算ノードの試行利用を開始し、約3年が経過

- 「富岳」では、Lustreをベースに富士通株式会社が独自に機能拡張したファイルシステム（FEFS）を採用

- 国内最大の超大規模スパコンである「富岳」のファイルシステム運用で経験した困りごとを事例として3つご紹介

「富岳」概要

富岳のこれまでの歩み

2014 | 2019 | 2020 | 2021 | 2022 | 2023

この約3年間を振り返る

▶ フラグシップ2020プロジェクト開始

▶ 詳細設計完了、製造へ

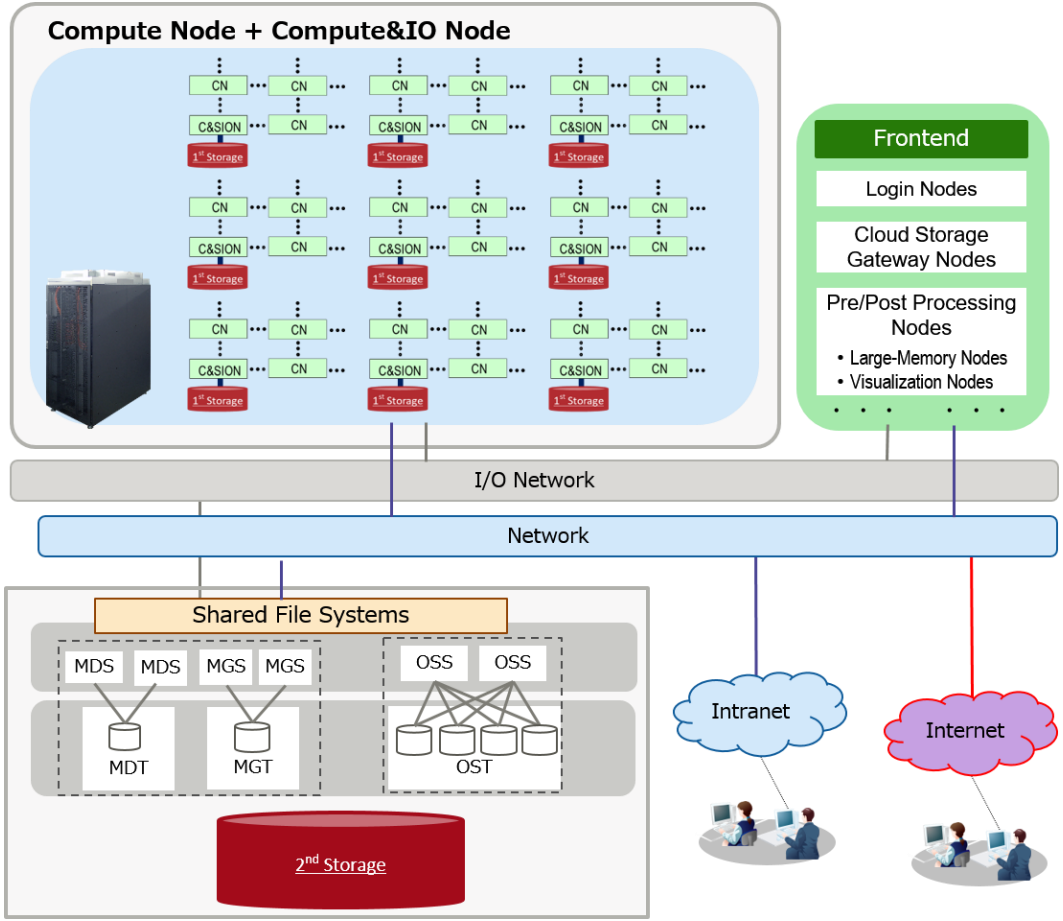
▶ すべての計算ノードのR-CCSへの搬入が完了

▶ 一部の計算ノードを使った**試行利用開始**
covid-19対策など

▶ **共用運用開始**

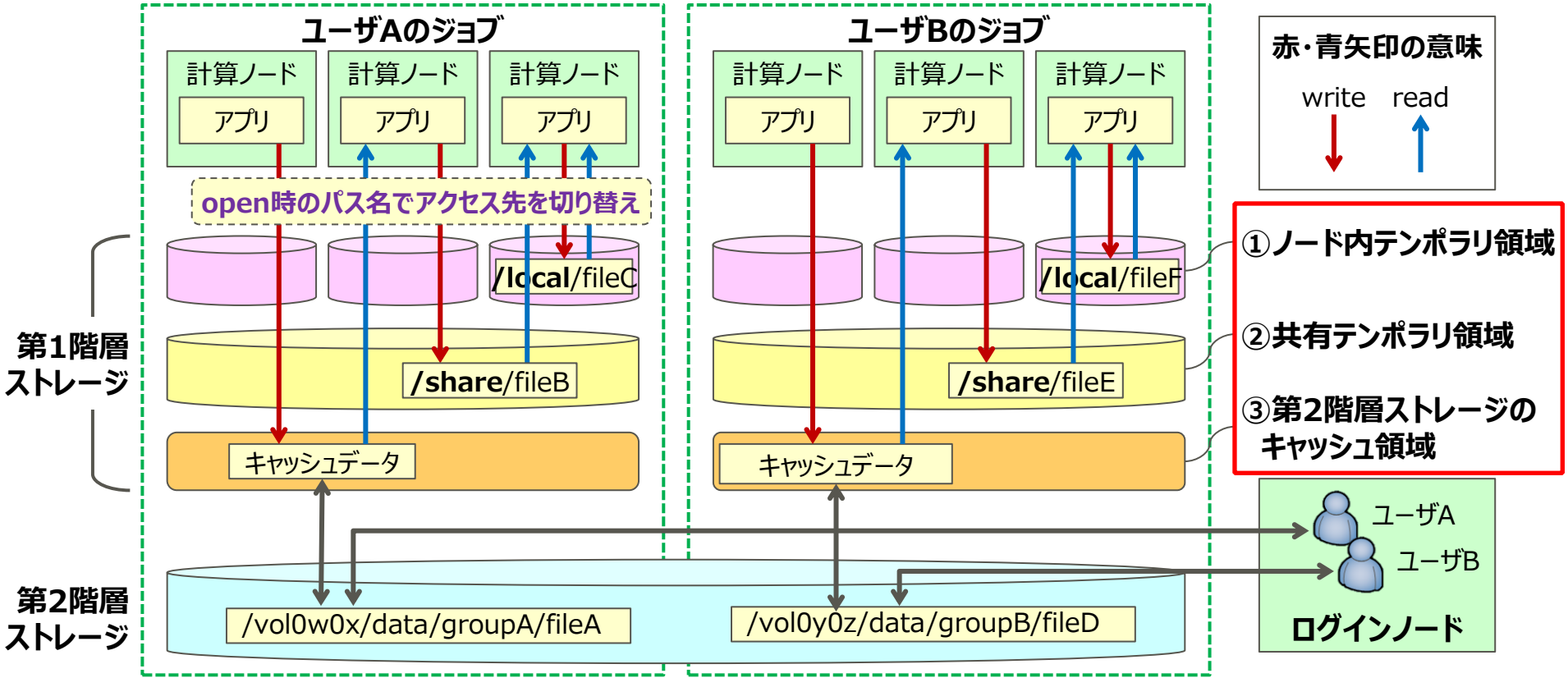
4期連続四冠達成
（TOP500、HPCG、HPL-AI、Graph500）

「富岳」システム構成

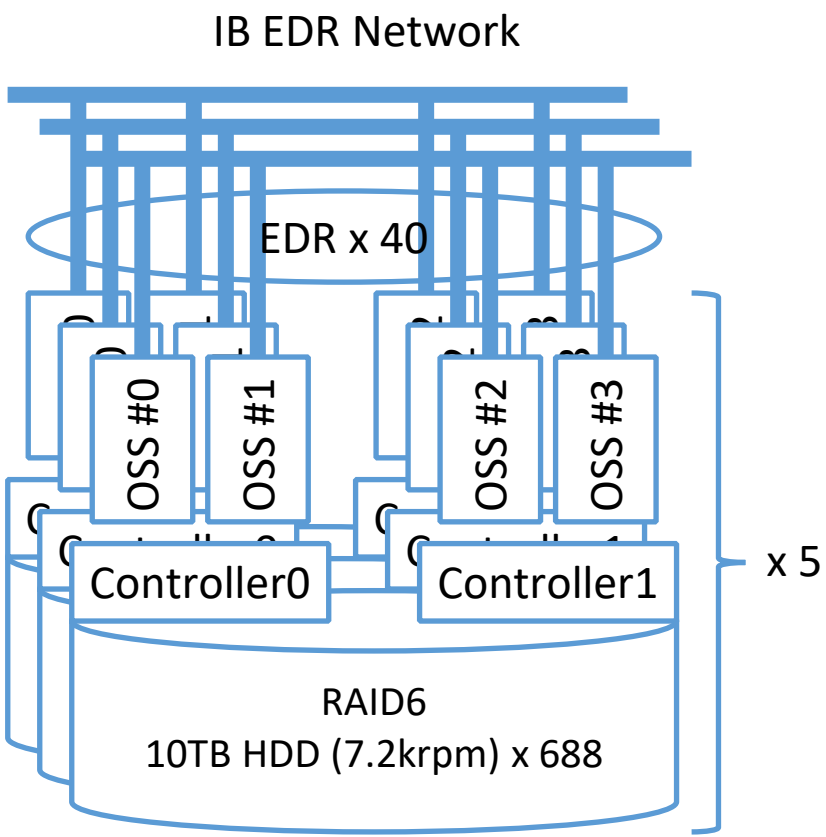


- **計算ノード**
  - 432 ラック (**158,976 ノード**)
    - 384ノード/ラック（一部ラックは 192ノード/ラック）
- **インターコネクト**
  - TofuD (28 Gbps x 2 lanes x 10 ports)
    - 6D mesh/torus (物理構成)
    - 3D torus (論理構成)
- **ストレージシステム**
  - 第1階層（LLIO）
    - 第2階層のファイルキャッシュ
    - テンポラリ領域
      - ローカルファイルシステム
      - 共有ファイルシステム
  - 第2階層
    - Fujitsu FEFS: Lustre-based file system 約125 PB
  - 第3階層
    - 商用クラウドストレージ / HPCI ストレージ

第1階層（LLIO）構成


ユーザAのジョブ
ユーザBのジョブ
計算ノード
アプリ
open時のパス名でアクセス先を切り替え
/local/fileC
/local/fileF
/share/fileB
/share/fileE
キャッシュデータ
第1階層ストレージ
第2階層ストレージ
/vol0w0x/data/groupA/fileA
/vol0y0z/data/groupB/fileD
赤・青矢印の意味
write
read
①ノード内テンポラリ領域
②共有テンポラリ領域
③第2階層ストレージのキャッシュ領域
ユーザA
ユーザB
ログインノード

第2階層ストレージ構成



- **FEFS (Fujitsu Exa-scale File System)**
  - 富士通が開発した Lustreベース(Ver.2.10)のファイルシステム
- **合計容量 150PB → 125PB** (2023年4月以降)
  - 6ファイルシステムに分割
    - 1ファイルシステムの構成例
      - DDN装置5台で構成
      - 容量：25PB
      - 20 OSSs, 60 OSTs
- **QoSを設定**
  - 計算ノード と ログインノード間
  - ユーザ間

ファイルシステム運用で経験した困りごと3選

困りごと その1（ユーザへの継続的な周知・指導）（1/2）

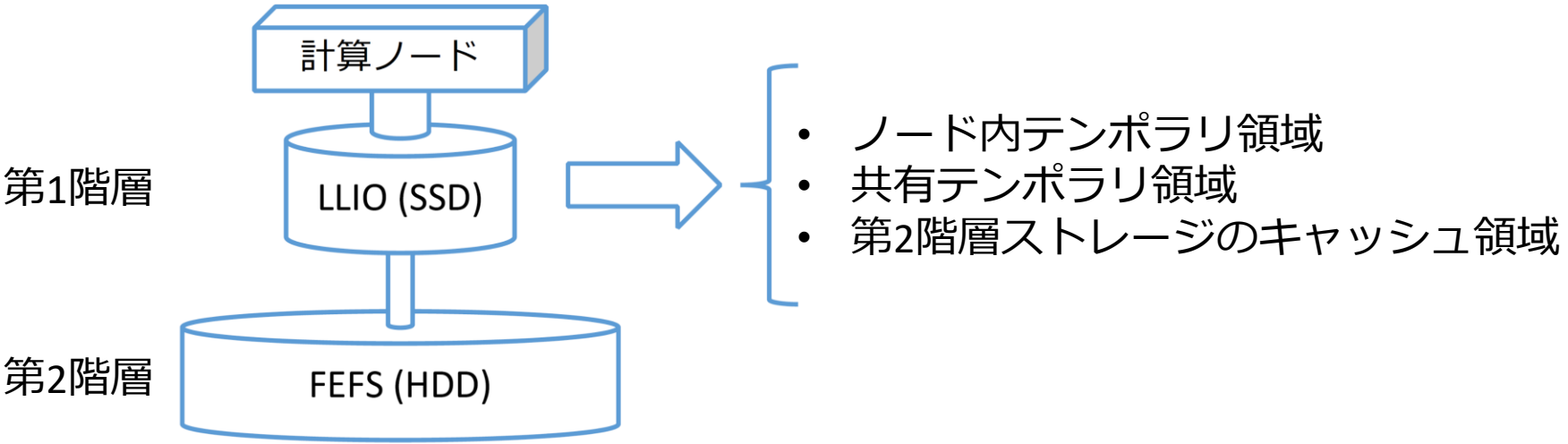
- **「富岳」は利用者の裾野が広く、多分野にわたって多くのユーザが利用**
  - HPC利用への習熟度に差がある
- **HPCに不慣れなユーザによる、「富岳」での大量ジョブ実行**
  - 同一ファイルを多数の計算ノードから同時に読み込み
  - 多数の計算ノードから同一ディレクトリ配下への大量のファイル出力
- **負荷集中・過負荷が発生**
- **IOノードダウンやファイルシステム全体のスローダウンにつながる**

困りごと その1（ユーザへの継続的な周知・指導）（2/2）

- **「富岳」では第2階層（FEFS）への負荷軽減のためLLIOを導入**
- **しかし、利用方法を理解して使える人がいなかった**
  - 並列数に関する制限を超過する
  - ジョブ実行に必要なファイルの事前読み込み指示が漏れる（llio_transfer）
- **そのためユーザへの継続的な周知・指導が必要**
  - 年度ごとの課題入れ替わりでユーザが去り、新規ユーザが利用を開始
    → 同じような状況が繰り返されてしまう
  - 大規模ジョブ実行には一定規模以上での動作実績が必須
- **今後のファイルシステムへの期待**
  - ユーザが過負荷を気にせず使っても安定的に動作する

参考：LLIOとは

- **LLIO（ Lightweight Layered IO-Accelerator）とは**
  - 並列分散ファイルシステムであるFEFSと計算ノードの間に位置する、SSDを使用した高性能なファイルシステム、またはそれを実現する技術
  - ジョブ用の一時ファイルをFEFSに書き出さない、またはLLIOからFEFSへの書出しを計算処理中に非同期に行うことで高速化を実現



困りごと その2（ストレージ割り当て）（1/2）

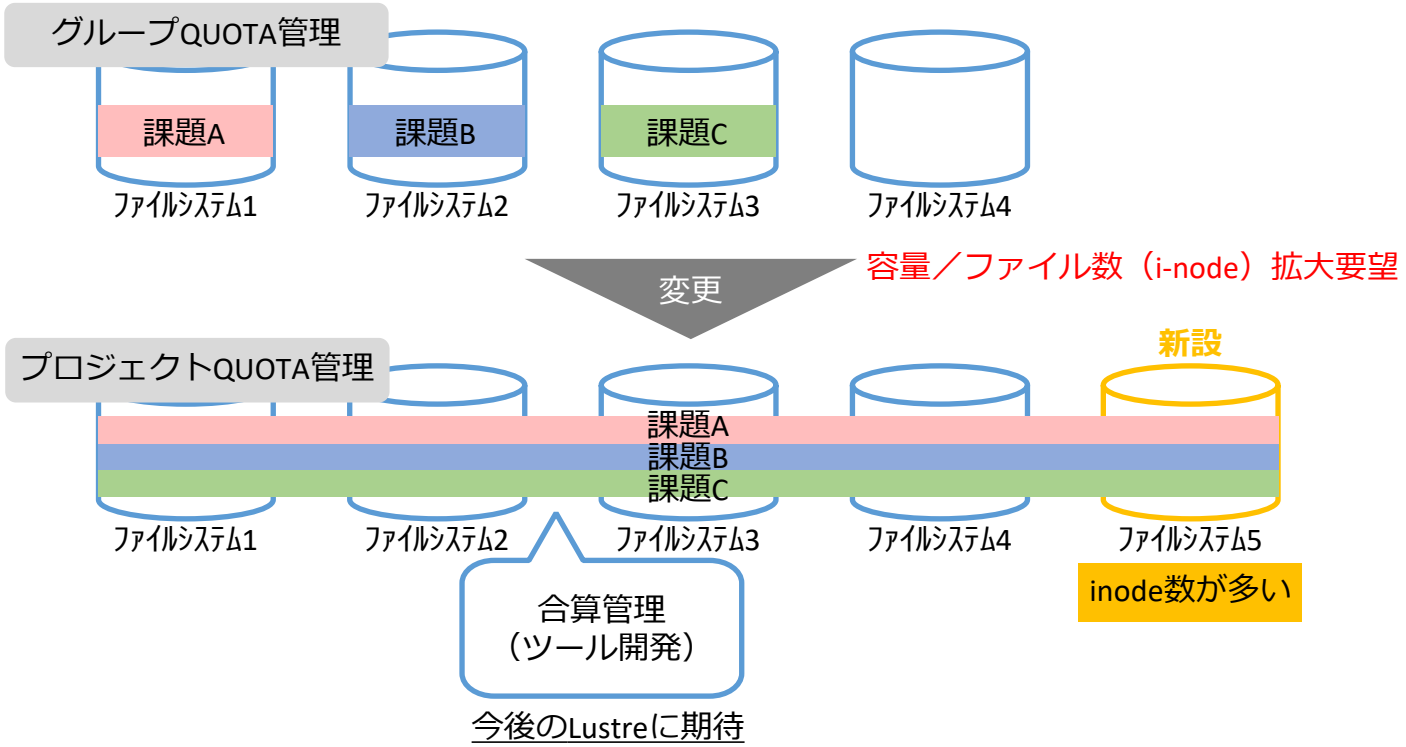
- **運用当初のストレージ管理（スパコン「京」踏襲）**
  - 1課題(*)1ファイルシステムへのデータ領域割り当て　　*「富岳」利用の単位、グループに相当
  - 課題単位のデータ領域をグループQUOTAで管理

- **運用中に発生**
  - 課題単位のデータ領域が1ファイルシステムに収まらなくなってきた
    - 数百TB割り当てている課題もある
  - より大量のファイル数を扱いたいという需要があった

- **ストレージ割り当ての運用を変更した**

困りごと その2（ストレージ割り当て）（2/2）

- **運用変更イメージ**



困りごと その3（データ削除）（1/2）

- **運用開始から3年以上が経ち、ファイルシステムへの領域割り当てがひっ迫**
  - 全体の8~9割割り当て済み
  - 節約のための指導が必要な段階
  - 割り当てが大きい課題から優先的に、データの削除と割り当て縮小を打診

- **しかし、データ削除のための各ユーザの使用量の情報の採取は容易ではない**
  - QUOTAの情報から割り出せない
  - ユーザの全ファイル（数十億ファイル）の所有者情報を取得し集計
  - 1周20日以上かかっている

使用量の単位	ｸﾞﾙｰﾌﾟ QUOTA	ﾕｰｻﾞ QUOTA	ﾌﾟﾛｼﾞｪｸﾄ QUOTA
ﾕｰｻﾞ	×	○	×
ｸﾞﾙｰﾌﾟ	○	×	○
ﾕｰｻﾞ & ｸﾞﾙｰﾌﾟ	×	×	×

困りごと その3（データ削除）（2/2）

- **削除対象データが非常に多い場合**
  - 一気に削除できない（負荷集中・過負荷の問題）
  - ファイルシステムの負荷状況を見ながら徐々に削除

- **今後のファイルシステムへの期待**
  - QUOTAのように、ユーザのファイル情報が容易かつリアルタイムに確認できる
    - オーナー情報
    - サイズ
    - アクセス日付　等
  - データ一括削除など処理量が多い運用オペレーションでも過負荷にならない

おわりに

- 「富岳」は2020年4月の共用前試行利用から3年以上経過し、
  今回振り返りを実施
- 約3年のファイルシステム運用での困りごとを3つご紹介

- アイデア・解決策など、お気づきのことがあれば議論・情報交換させてください


Thank you for your attention !