Gfarmシンポジウム＠秋葉原
2023年9月8日

R4文部科学省次世代計算基盤に係る調査研究事業
システム研究調査チーム（代表機関：理研）

次期ストレージシステムの調査

建部修見、平賀弘平、前田宗則（筑波大）
大辻弘貴、岡本拓也、二宮温、宮本巧輝（富士通）
橋爪信明、井原修一（DDN Japan）
野村昴太郎、矢澤克巳、Michael Hennecke、Johann Lombardi （Intel）

次世代計算基盤に係る調査研究事業

- いわゆるポスト富岳のフィージビリティスタディ
  - システム研究調査チーム（理研、神戸大）
  - 新計算原理調査研究チーム（慶応）
  - 運用技術調査研究チーム（東大）

https://www.mext.go.jp/b_menu/boshu/detail/mext_00229.html

理研チーム 研究体制（2023年7月）

GL: グループリーダ
AD: アドバイザー
SGL: サブグループリーダ

システム調査研究チーム
（代表機関）理化学研究所
【代表者：近藤， AD: 松岡(理研)】

アーキテクチャ調査研究グループ

アーキテクチャ調査研究グループ
取りまとめ担当:（代表機関）理化学研究所◎
【GL: 佐野，副GL: 三輪(電通大), AD: 天野(慶大)】

- アーキテクチャ調査研究サブG１
  （理化学研究所）
  【SGL: 泰地】
- アーキテクチャ調査研究サブG２
  （分担機関）富士通株式会社
  【SGL: 新庄】
- アーキテクチャ調査研究サブG３
  （分担機関）インテル株式会社
  【SGL:矢澤】
- アーキテクチャ調査研究サブG４
  （分担機関）日本AMD株式会社
  【SGL:吉田】
- アーキテクチャ調査研究サブG５
  （協力機関）NVIDIA Corporation
  【SGL:Wells】
- アーキテクチャ調査研究サブG６
  （協力機関）Hewlett Packard Enterprise
  【SGL:根岸】
- アーキテクチャ調査研究サブG７
  （協力機関） Arm Ltd.
  【SGL: Lecomber】

アーキテクチャ調査研究グループ

システムソフト・ライブラリ調査研究グループ

システムソフト・ライブラリ調査研究グループ
取りまとめ担当:（代表機関）理化学研究所◎
【GL: 佐藤, 副GL: 片桐(理研),佐藤(豊橋技大) AD : 佐藤(理研)】

システムソフト取りまとめ補助
（協力機関）名古屋大学
【機関代表: 片桐】

- スケジューラ・ランタイムサブG
  （分担機関）東北大学
  【SGL: 滝沢】
- IO・ストレージ・ファイルシステムサブG
  （分担機関）筑波大学
  【SGL: 建部】
  - ストレージアーキ・パターン調査
    （協力機関）株式会社DDNジャパン
    【機関代表: 橋爪】
- OS・仮想化・クラウド連携サブG
  （協力機関）国立情報学研究所
  【SGL: 竹房】
- HPC利用環境サブG
  （分担機関）大阪大学
  【SGL: 伊達】
- 通信ライブラリサブG
  （分担機関）九州大学
  【SGL: 南里】
- コンパイラ・プログラミングモデルサブG
  （理化学研究所）
  【SGL: 辻】
- 数値ライブラリサブG
  （理化学研究所）
  【SGL: 今村】
- AIフレームワークサブG
  （理化学研究所）
  【SGL: Mohamed】

システムソフト・ライブラリ調査研究グループ

アプリケーション調査研究グループ

アプリケーション調査研究グループ
取りまとめ担当:（分担機関）北海道大学◎
【GL: 岩下, 副GL : 高橋(筑波大),深沢(京大), AD: 中島・富田(理研)】

アプリ調査研究取りまとめ補助
（協力機関）京都大学
【機関代表: 深沢】

- 生命科学分野サブG
  （分担機関）横浜市立大学
  【SGL: 寺山】
- 新物質・エネルギーサブG
  （分担機関）物質・材料研究機構
  【SGL: 山地, 副SGL福島(東大)】
- 気象・気候サブG
  （分担機関）海洋開発研究機構
  【SGL: 小玉】
- 地震・津波防災
  （分担機関）東京大学
  【SGL: 藤田】
- ものづくり分野サブG
  （理化学研究所）
  【SGL: 大西】
  - ものづくり分野協力
    （協力機関）宇宙航空研究開発機構
    【機関代表: 高木】
- 基礎科学分野サブG
  （理化学研究所）
  【SGL: 青木】
  - 宇宙・惑星アプリ分野協力
    （協力機関）国立天文台
    【機関代表: 瀧脇】
- 社会科学分野サブG
  （理化学研究所）
  【SGL: 梶本】
- デジタルツイン・Society5.0分野サブG
  （分担機関）東京大学
  【SGL: 下川辺】
  - デジタルツイン・Society5.0分野補助
    （協力機関）日本原子力研究開発機構
    【機関代表: 小野寺】
- 気象庁モデル評価・解析サブG
  （理化学研究所）
  【SGL: 児玉】
  - 気象庁モデル評価・解析サブG補助
    （協力機関）気象庁
    【機関代表: 石田】
- 科学技術計算アルゴリズムサブG
  （分担機関）筑波大学
  【SGL: 高橋】
- 機械学習アルゴリズムサブG
  （分担機関）東京工業大学
  【SGL: 横田】
- ベンチマーク構築サブG
  （理化学研究所）
  【SGL: 村井】
- 性能モデリングサブG
  （理化学研究所）
  【SGL: Domke】

アプリケーション調査研究グループ

中間報告

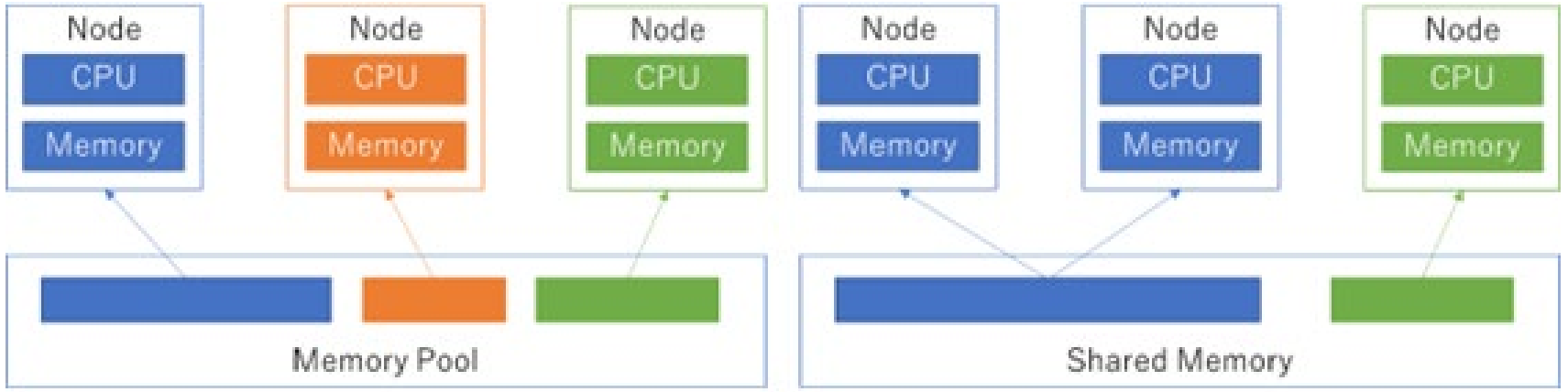
- ストレージデバイスの動向
- アクセスパターンの調査
- ストレージアーキテクチャの調査
- 次期ストレージシステムについての考察

ストレージデバイスの動向

- HDD
  - 2022年は22TBまで出荷されており18TBがボリュームゾーンである。今後も容量は増加していき、2026年頃には50TB超の出荷も予想される。容量当たりのコストはSSDに比べ一桁ほど優位となり続けるであろう
- SSD
  - 2022年は30TBまで出荷されている。今後も容量は増加していき2024年頃には120TBの出荷も予想される。バンド幅はPCIeの世代とともに増加する。消費電力も増加する。
- PCIe
  - 2025年頃にPCIe Gen7の標準化が行われ、対応する製品の出荷はその3〜5年後が予想される

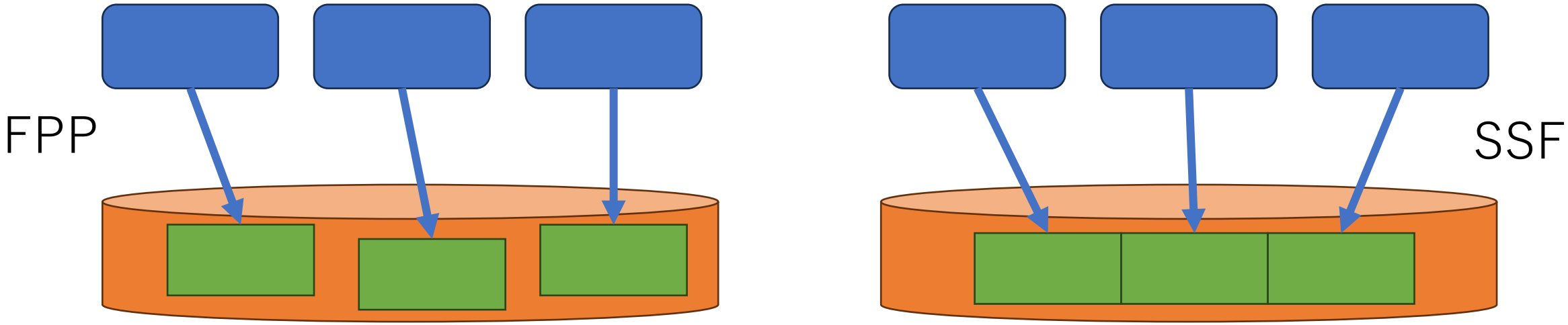
CXL – Compute Express Link

- CPUとデバイス（メモリ、GPU）間のキャッシュコヒーレントなリンク
- メモリプーリングと共有メモリ（CXL 3.0）



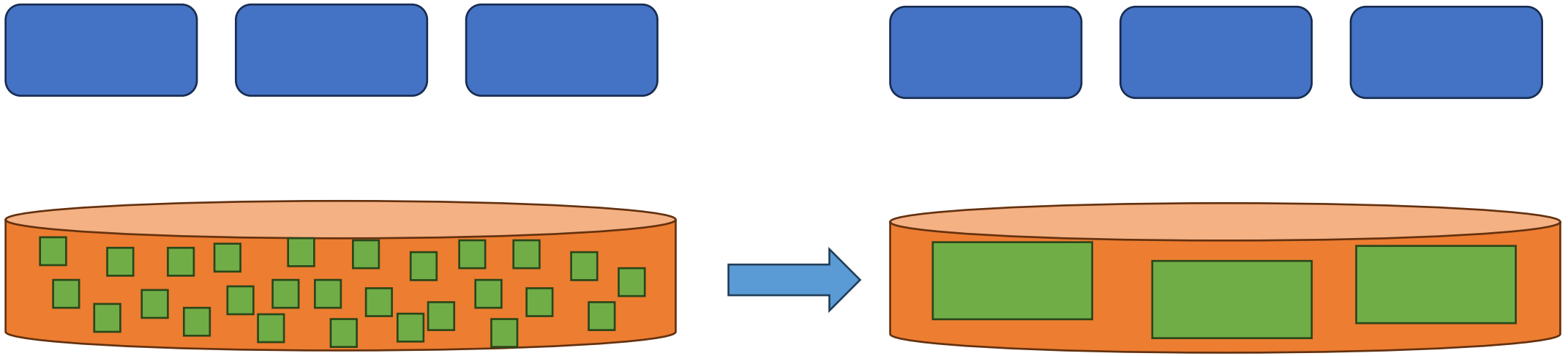
アクセスパターンの調査

- 科学技術アプリケーション
  - プロセス毎にファイルをアクセスするFPPと、単一ファイルをアクセスするSSFがある。大規模アプリケーションにおいてはSSFが望まれるが、ストレージ性能に問題がある場合はFPPが使われる。
  - 富岳ユーザには、POSIXによるFPPとNetCDF、PnetCDF、MPI-IO、HDF5によるSSFの両方が存在する



アクセスパターンの調査

- 深層学習
  - サイズの小さい多くのファイルのランダムアクセスが求められているが、ストレージアクセス性能向上のため複数のファイルをまとめる最適化が用いられる傾向にある
  - 深層学習のモデルが大きくなり、計算コストに対するファイル入出力コストが下がる傾向にある。チェックポイントのコストは増大
  - データ整備のため探索的データ解析、キュレーション等の前処理が必要となるが、これらの処理では高いストレージ性能が求められる

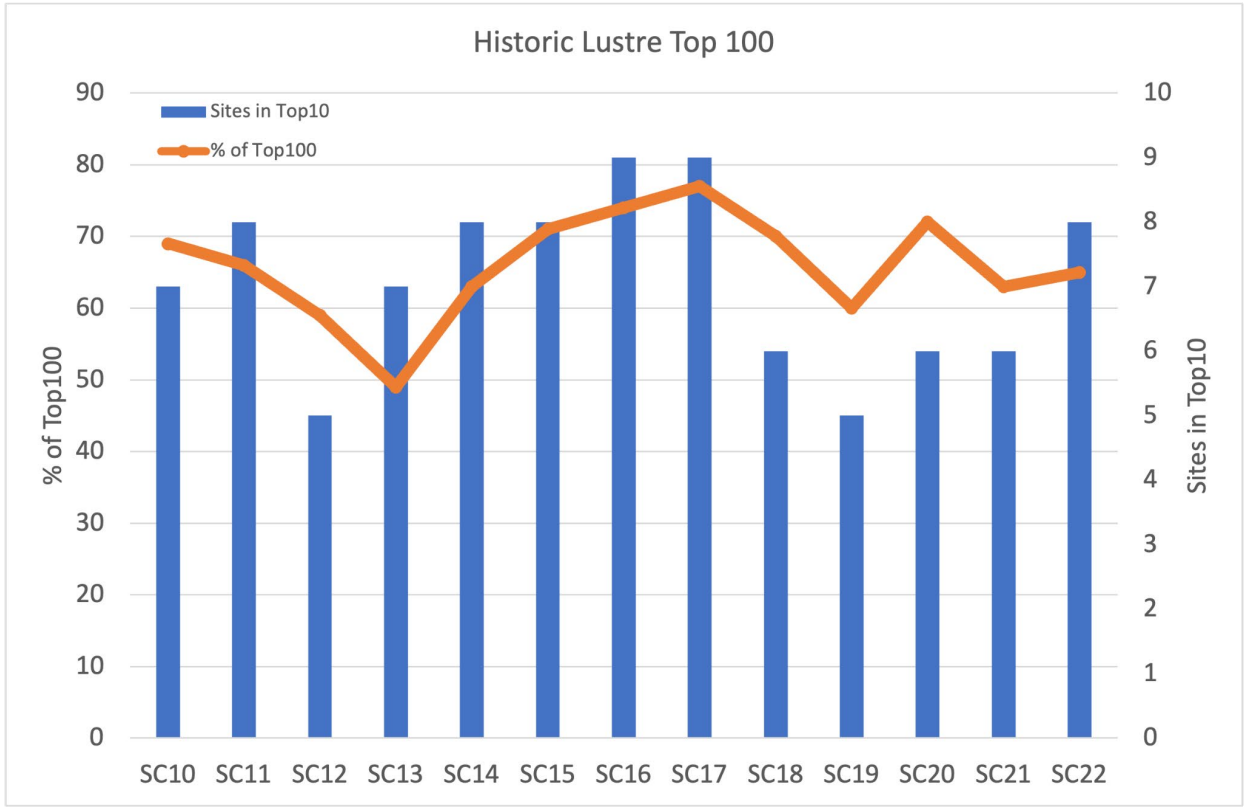


ストレージアーキテクチャの調査

- Lustre
- DAOS
- LLIO

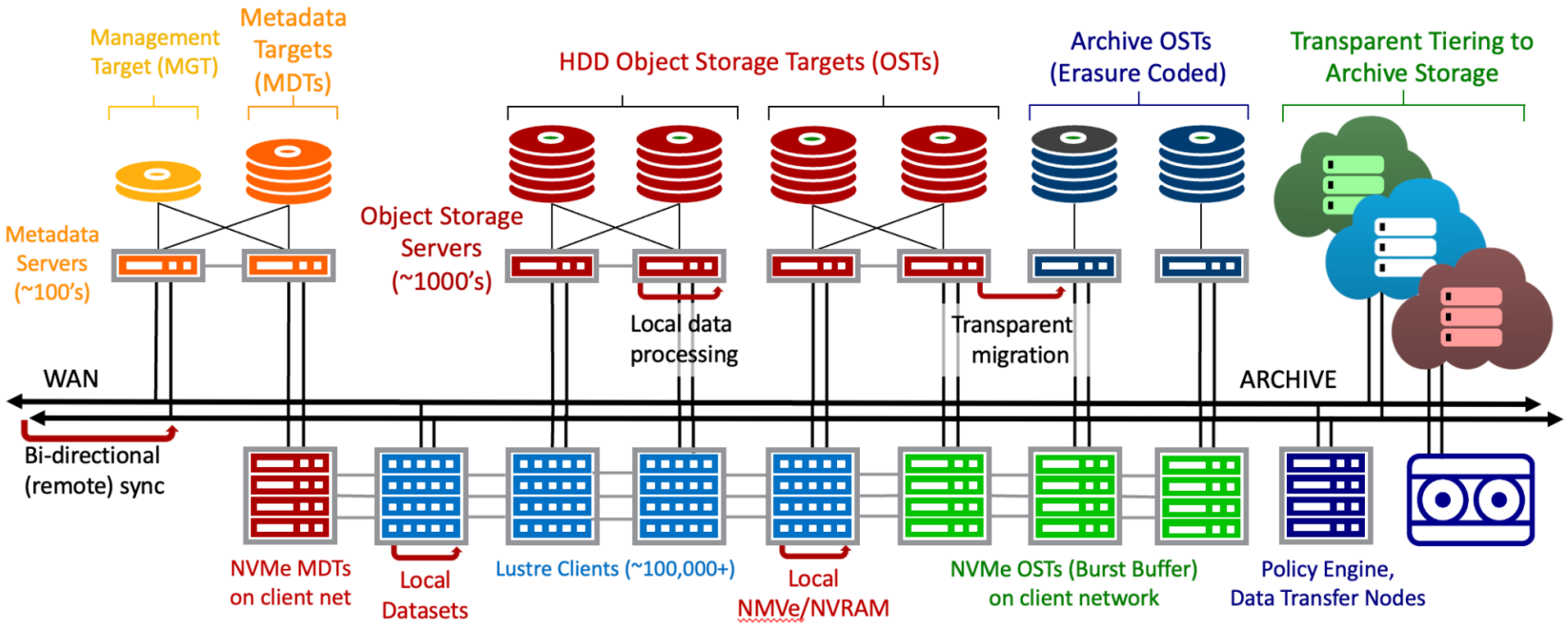
Lustre

- オープンソース。20年以上開発が継続。富岳第２階層のベース
- 2022年Top10の8システム、Top100の7割強
- 最新は Lustre 2.15.1
  （2023/3時点）



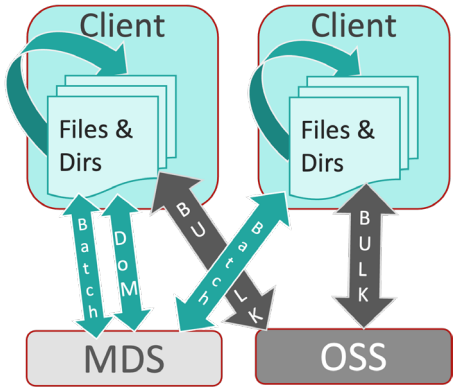
最新のLustreアーキテクチャ

- ハイブリッドストレージ
  - 計算ノードローカル、NVMe、HDD、テープ、クラウド
  - NVMe-HDD間自動ティアリング、永続クライアントキャッシュ



Lustre開発中の機能

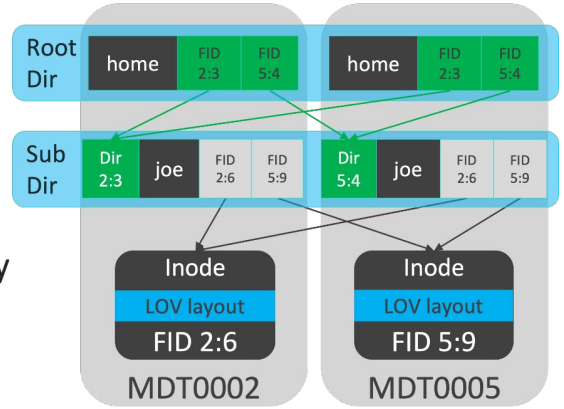
- メタデータサーバのスケーラビリティ
  - DNE3 – 複数のMDTを自動ロードバランシング
- メタデータのローカリティ
  - MDTプール機能 – MDTをグループ化してディレクトリに割当
- バッチメタデータRPCとメタデータWrite Backキャッシュ
  - RPC発行回数の削減
- メタデータ冗長化
  - MDS/MDTのHA（自動縮退）
  - メタデータの冗長化
    を段階的に実装
    （2025 Lustre 2.18以降）



メタデータWrite Backキャッシュ

In early discussion and architecture stages

- **LMR1a: Replicate services to other MDTs**
  - Mirror FLDB, Quota, `flock()` across MDTs
- **LMR1b: DNE transaction performance**
  - Transactions have excessive ordering/sync
  - Improves **all** DNE operation performance
- **LMR1c: Replicate top-level dirs for availability**
  - **ROOT/** dir (rarely changed) mirrored over MDTs
  - No per-file metadata replication initially
- **LMR2/3 phases needed for full redundancy**
  - Full tree replication, inode replication, configurable per directory
  - Recovery, LFSCK, rebuild replicated directories after MDT loss



DAOS

Research ISC23 List

Customize | Download

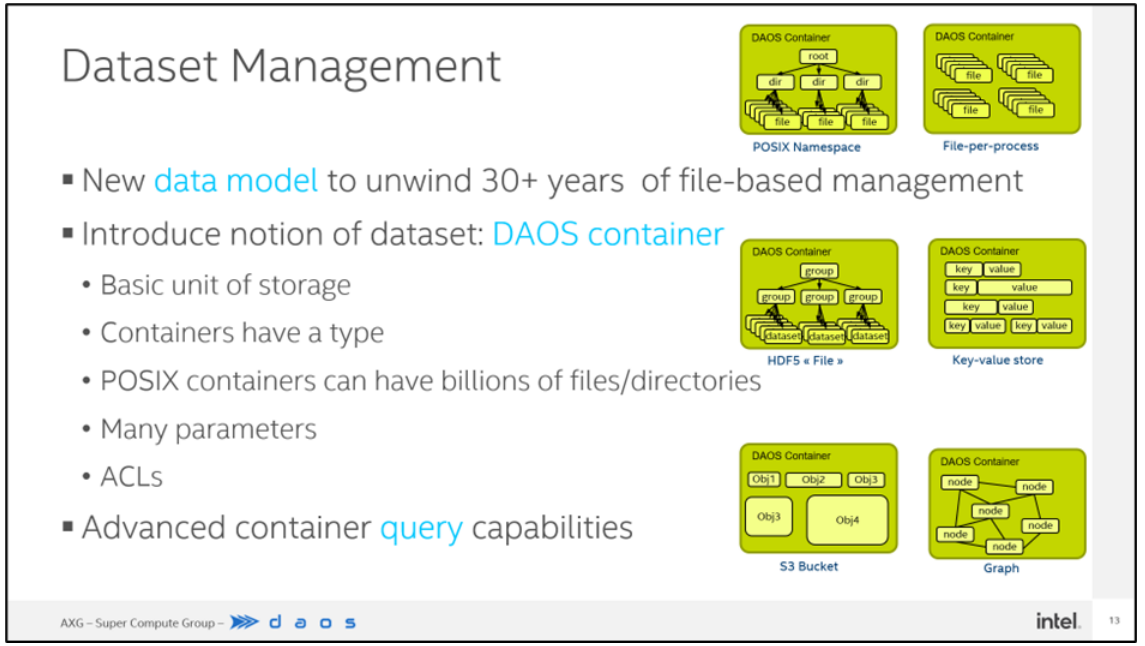
Production | 10 Node Production | **Research** | 10 Node Research | Full | Historical

Ranking of the research system submissions. This is a subset of the Full List of submissions, showing only one highest-scoring result per storage system. This list also contains all valid IO500 submissions prior to the creation of the Research List.

	INFORMATION							IO500	
# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)
1	ISC23	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory and Tsinghua University	SuperFS	300	36,000	210,254.98	4,847.48	9,119,612.35
2	ISC23	JNIST and HUST PDSL	Cheeloo-1 with OceanStor Pacific	Huawei	OceanFS2	10	9,600	137,100.02	2,439.37	7,705,448.04
3	SC22	Argonne National Laboratory	Aurora Storage	Intel	DAOS	260	27,040	20,694.50	6,048.69	70,802.51
4	SC22	Sugon Cloud Storage Laboratory	ParaStor	Sugon	ParaStor	10	2,560	8,726.42	718.11	106,042.93
5	SC22	SuPro Storteck	StarStor	SuPro Storteck	StarStor	10	2,560	6,751.75	515.15	88,491.65

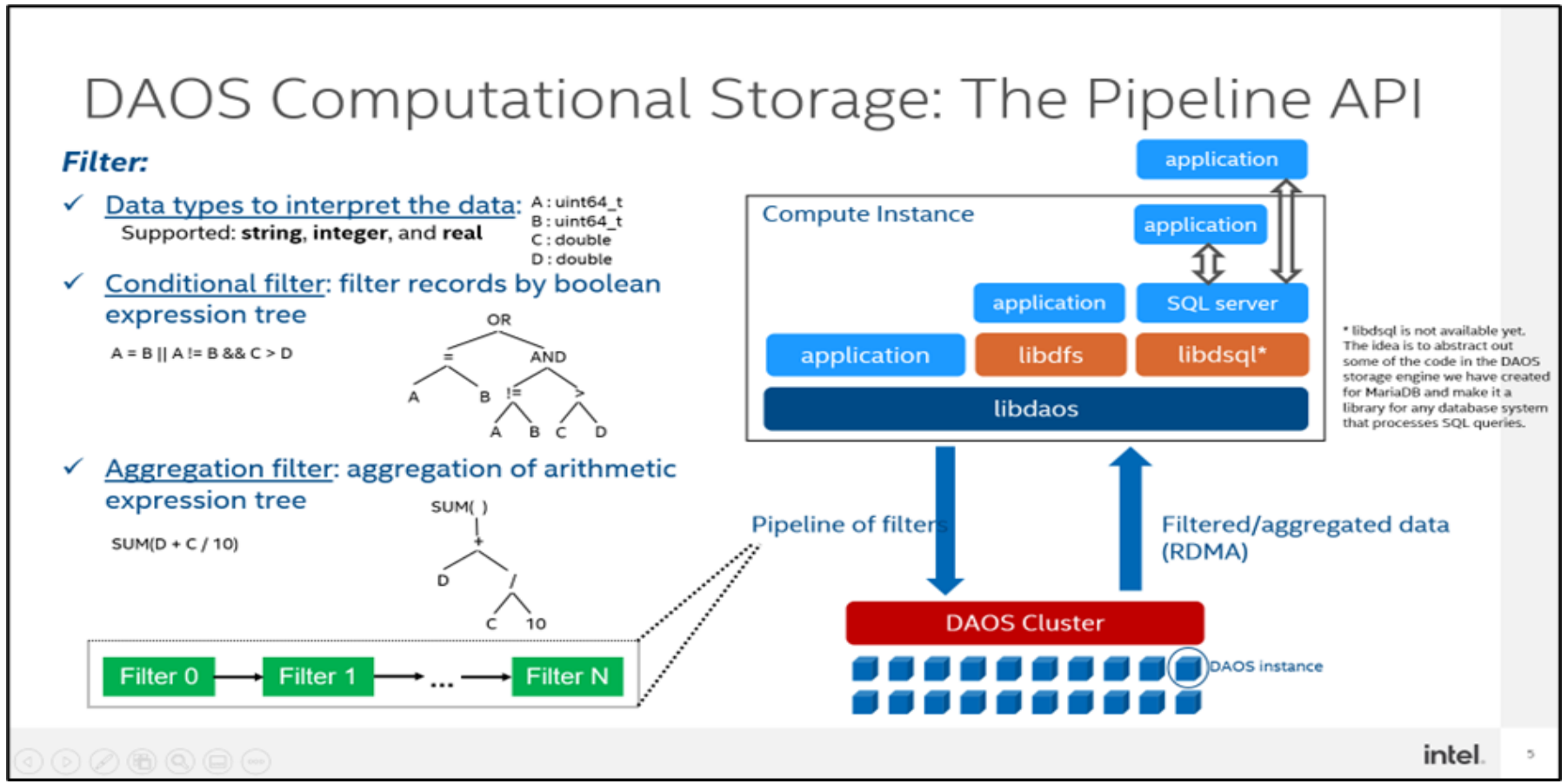
DAOSの特徴

- オープンソース
  - 多くの機能はANLのAurora(230PB+, 20TB/s+)のため設計
- 完全分散型メタデータ（メタデータ専用サーバなし）
- バージョニングオブジェクトストア（ログ構造マージ木なし）
- ユーザ空間で動作
- 今後階層ストレージ
  に対する開発計画あり



DAOS計算ストレージ機能

- ストレージへの計算のオフロード。Find、データ前処理のユースケース



LLIO

- 富岳向けに開発された第一階層ストレージ。当時、要件を満たすストレージが存在しなかったため独自開発
- キャッシュ、共有テンポラル、ノード内テンポラルを提供
- 富岳においては、SSF性能のスケーラビリティ、同時アクセスクライアント数、メタデータ性能、llio_transfer指定漏れによる高負荷等問題があり、次期システムにおいては改善が望まれる

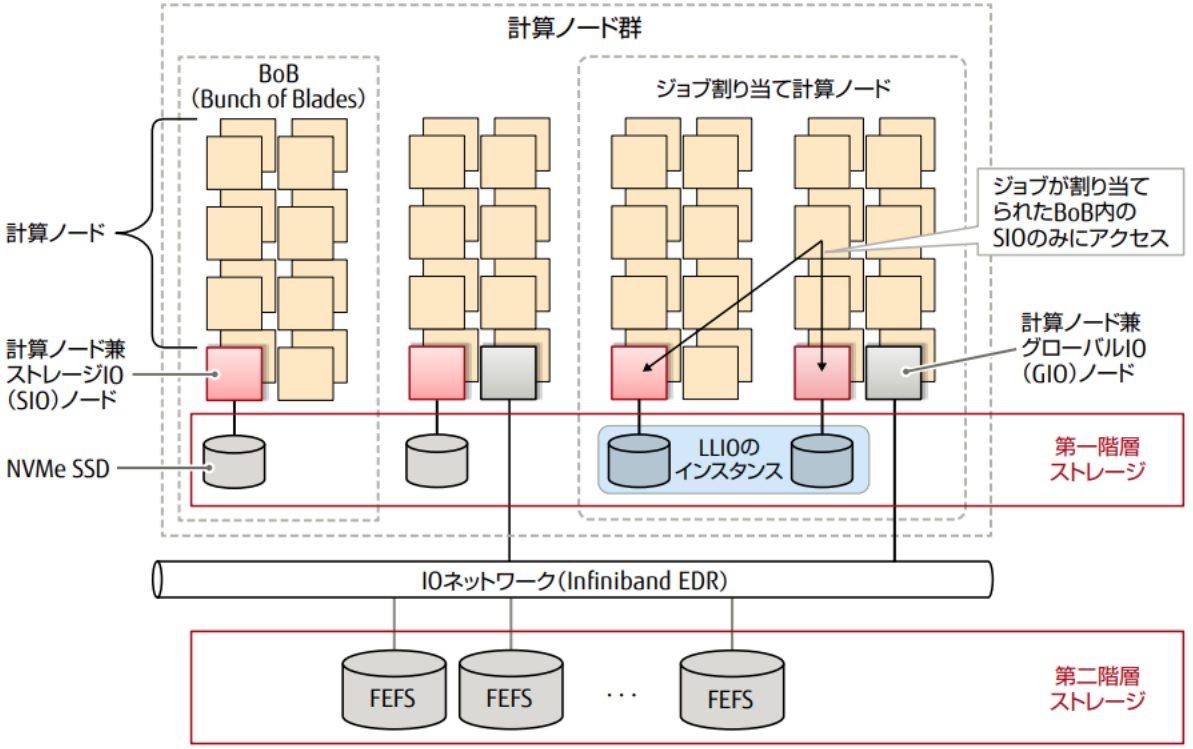


図-2 「富岳」ストレージシステムの概要

富士通テクニカルレビューより

次期ストレージシステムについての考察

- ストレージ性能について
  - 演算性能、メモリサイズ、ノード数などとのバランスが重要である
  - 全システムを用いた全プロセスからのアクセスにも耐えられるメタデータ性能、アクセス性能が求められる
  - SSFについても十分なバンド幅が求められる
- ストレージシステムの構成要素について
  - フラグシップシステムにおいては最先端の技術をいち早く導入することが求められる

次期ストレージシステム

- 次期スパコンの想定　　富岳

演算性能 (DP)	20 EFlops	537 PFlops	37x
メモリ容量	20 PB	4.85 PiB	4x
ノード数	数千～数十万	158,976	

- 次期ストレージシステムの性能要件　　富岳

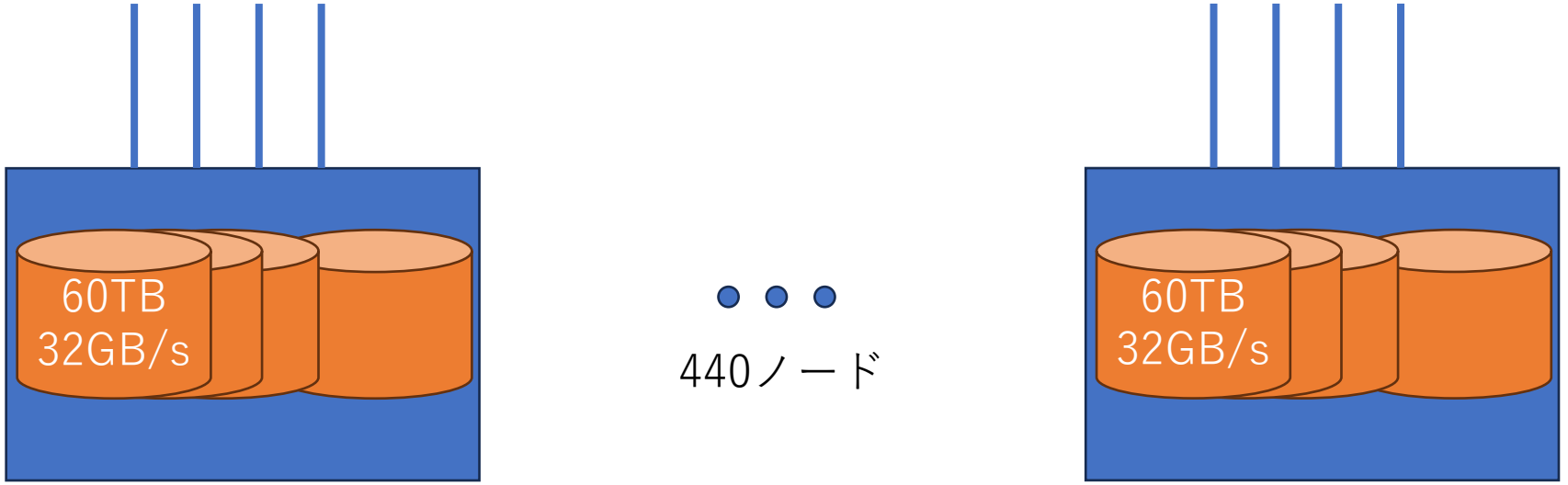
バンド幅（SSF）	350 TB/s（メモリダンプ1分）	20 TB/s, 1.3 TB/s（FFP）	17x
容量	600 PB（メモリの30倍）	15.8 PB, 150 PB	4x
メタデータ性能	100M IOPS（数千万プロセス）	???	
広域ストレージ	6 EB（10倍の容量）	200PB	30x

ストレージデバイス

- 2028～2030頃 PCIe Gen7の製品出荷開始
  - 16レーンで 1.6 Tbps（= 200 GB/s）
  - 4レーンで800 Gbps（＝50 GB/s）
- HDD　　50 TB、600 MB/s、10 W
- SSD　　240 TB、32 GB/s、40 W

ストレージ構成（その１）

200GB/s x 4 = 800GB/s



26 x 60TB = 1.56PB
26 x 32GB/s = 832GB/s

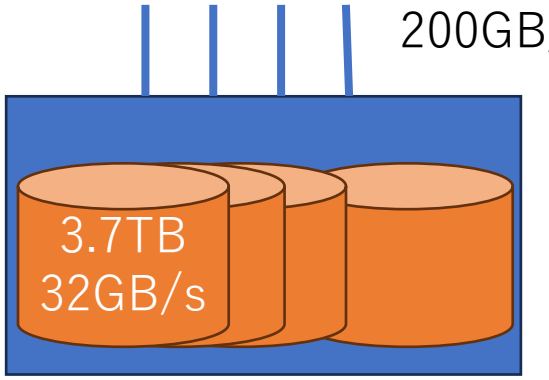
440 x 1.56PB = **686.4PB**
440 x 800GB/s = **352TB/s**

数千万プロセスから100M IOPSの
メタデータ性能
SSFにおける350TB/sの安定した
バンド幅を持つストレージシステム

ストレージ構成（その２）

1次ストレージ

200GB/s x 4 = 800GB/s



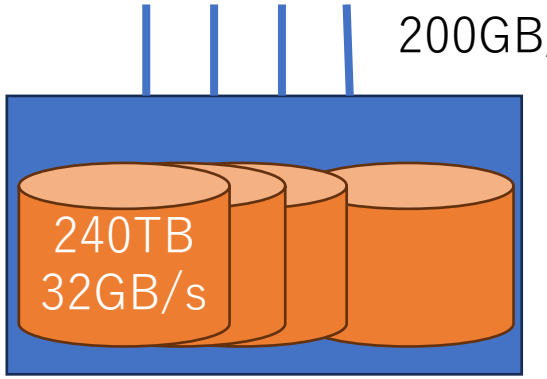
440ノード

440 x 96.2TB = **42.3PB**（メモリの倍）
440 x 800GB/s = **352TB/s**

数千万プロセスから100M IOPSの
メタデータ性能
SSFにおける350TB/sの安定した
バンド幅を持つストレージシステム

2次ストレージ

200GB/s x 4 = 800GB/s



106ノード

26 x 240TB = 6.24PB
26 x 32GB/s = 832GB/s

106 x 6.24PB = **661.4PB**
106 x 800GB/s = **84.8TB/s**（20PBを5分）

数十万プロセスから10M IOPSの
メタデータ性能
SSFにおける70TB/sの安定した
バンド幅を持つストレージシステム

透過的なキャッシュ機構

今後の調査について

- アプリケーションのヒアリングによる要求の明確化
- 次期ストレージシステムの要求を満たすストレージアーキテクチャの提示