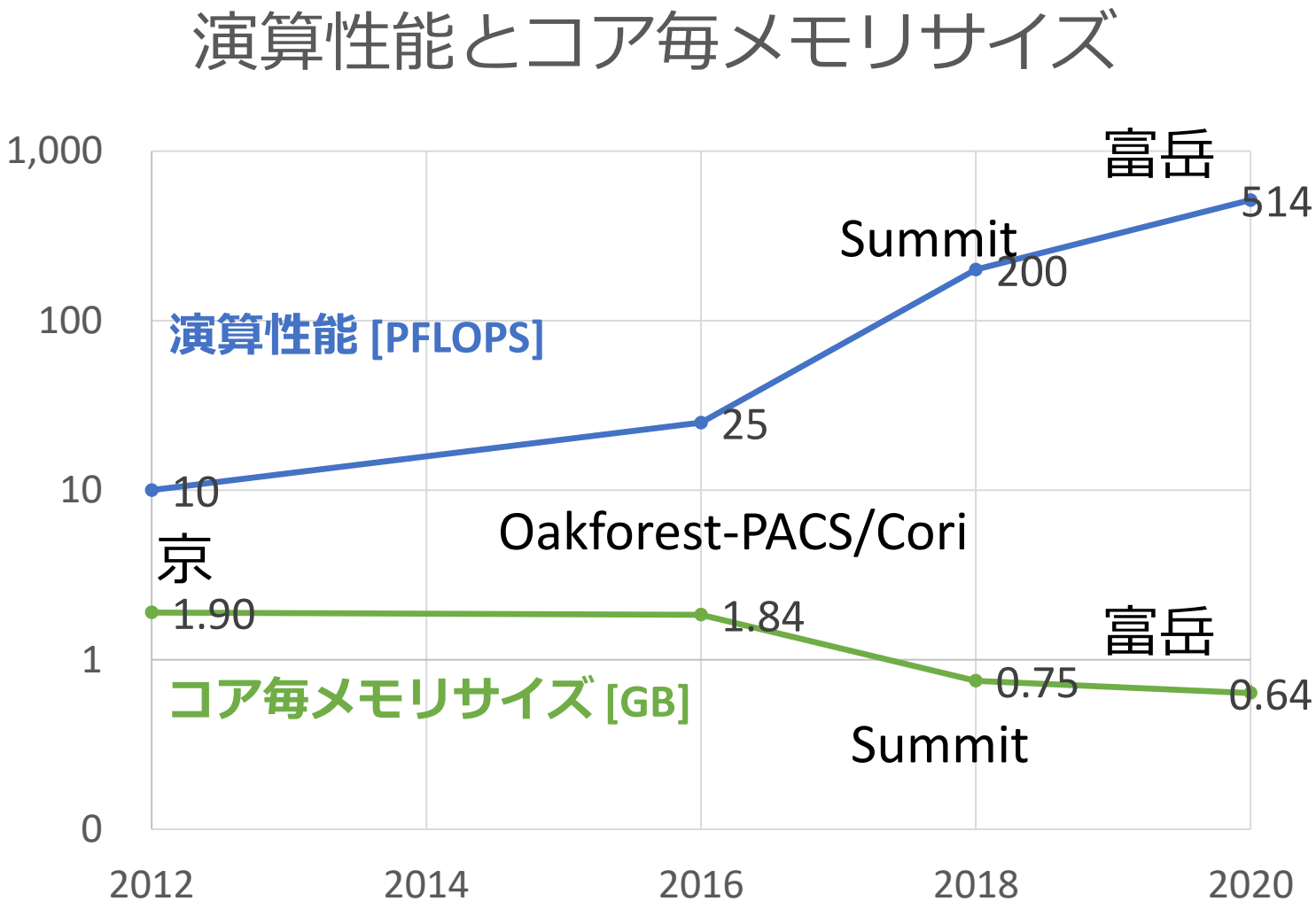
Pegasusビッグメモリスーパコンピュータではじまるこれからのデータ科学・ビッグデータAI

筑波大学計算科学研究センター
建部修見

背景

- ８年間で演算性能は50倍、メモリ容量は3.8倍
- データ駆動科学、AI駆動科学では大問題
  - メモリサイズとストレージ性能が重要

- 不揮発性メモリの導入
  - 低消費電力、コストエフェクティブ

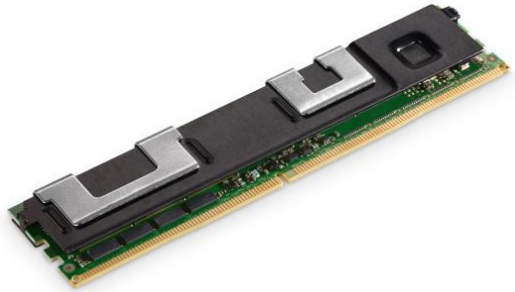


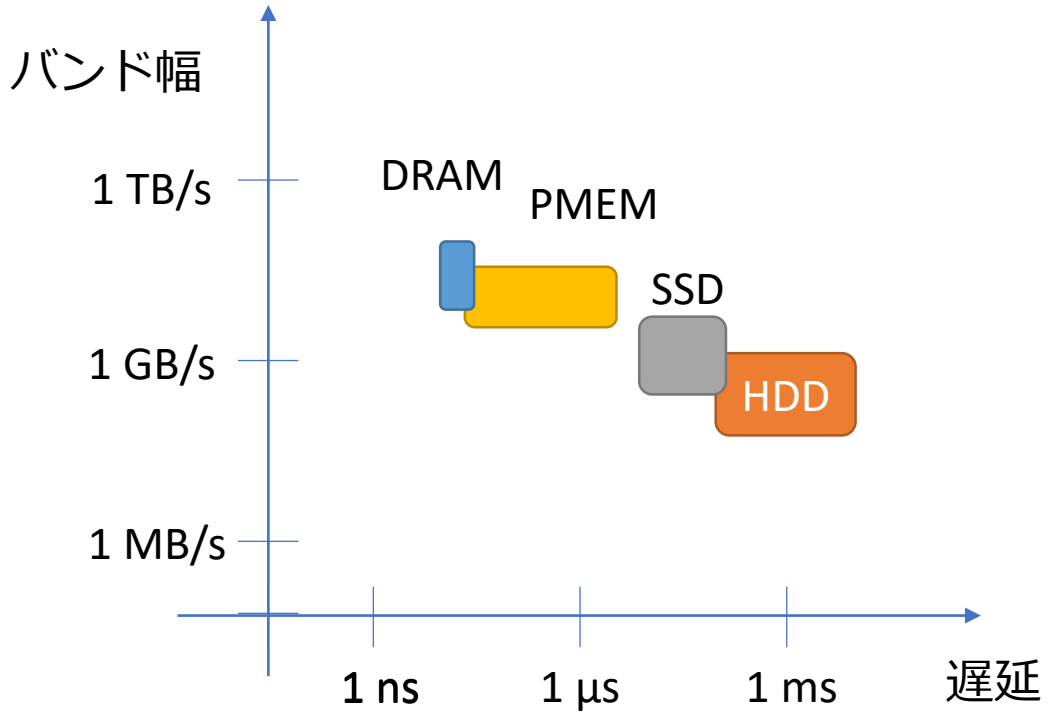
Pegasusの設計目標

- 高帯域大容量メモリと超高速ストレージによるHPC、ビッグデータ解析、超大規模AIの促進
- 不揮発性メモリを活用した大規模データ解析の新分野、ビッグデータAIの新用途、システムソフトウェア研究などの育成

不揮発性メモリ（PMEM）

- DRAMより一桁上の容量・コストパフォーマンス
- 最低遅延60ns程度（DRAMとほぼ変わらない）
- バンド幅はDRAMの半分程度
- メモリモード
  - アプリの性能をほぼ低下させずメモリ利用量を増加可能
- ダイレクトモード
  - PMEMを直接利用可能。バイトアドレス可能な不揮発性メモリ、超高速ストレージとしての利用





Pegasusハイライト

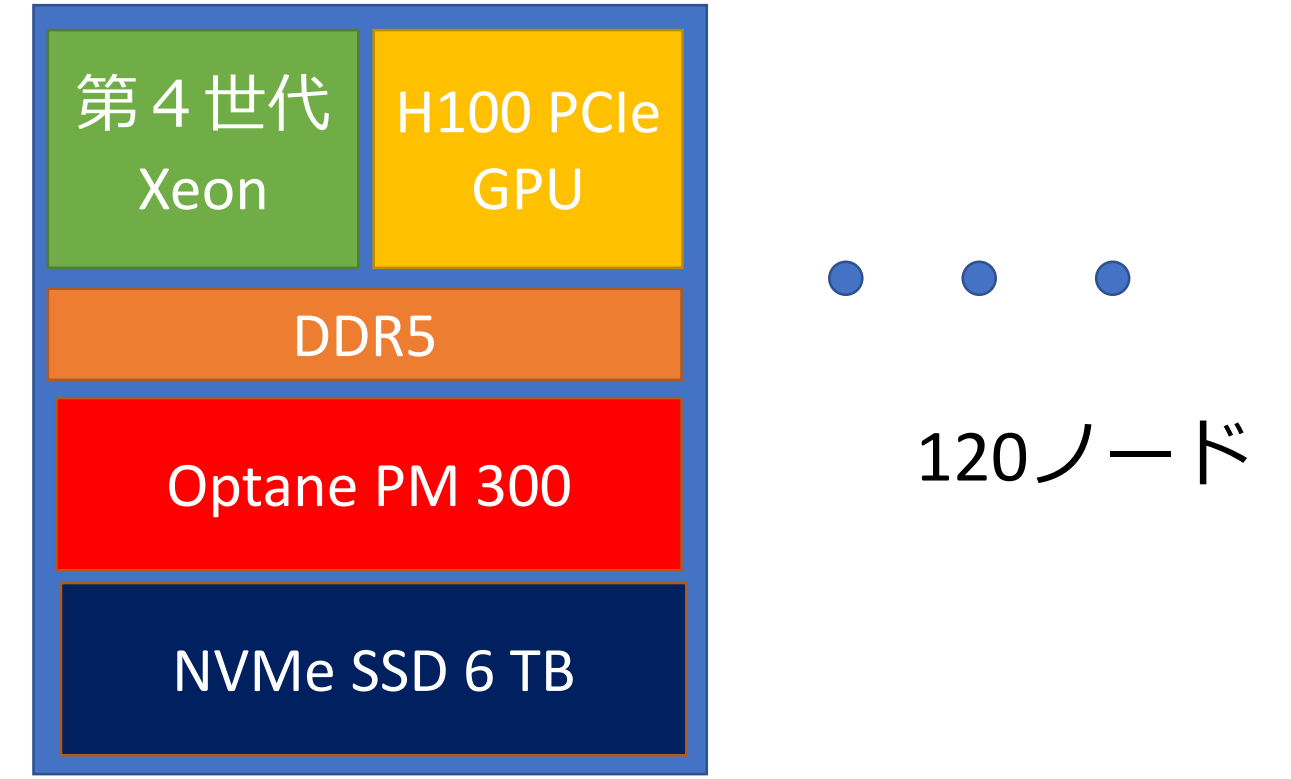
- 世界初NVIDIA H100 PCIe GPUとIntel不揮発性メモリを搭載
- HPC、ビッグデータ解析、超大規模AIを強力に推進

Pegasus仕様

- 2022年Q4に導入予定
- 全体性能
  - 120ノード, > 6.1 PFlops
- 計算ノード仕様
  - Intel第４世代Xeon
  - 51 TFlops NVIDIA H100 PCIe GPU
  - DDR5 DRAM
  - Optane PM 300シリーズ
  - 6 TB NVMe SSD (7 GB/s)
- 相互結合網
  - NVIDIA Quantum-2 InfiniBand (200 Gbps)フルバイセクション
- 並列ファイルシステム
  - 7.1 PByte DDN EXAScaler (40 GB/s)

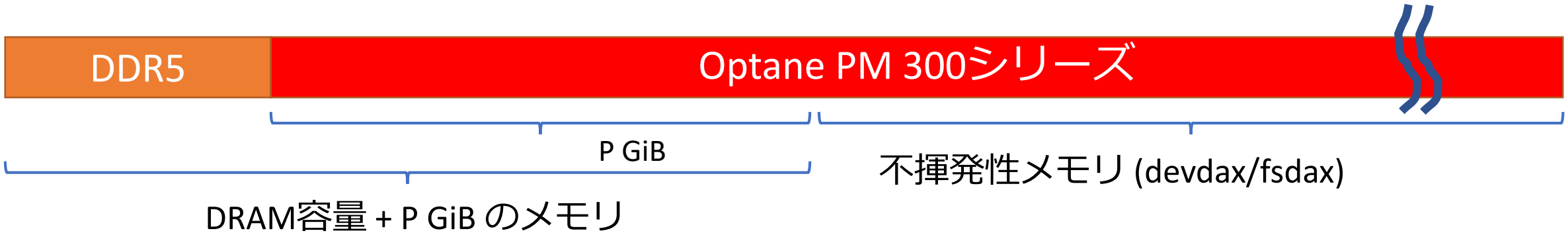
NEC LX B1000E Blade Enclosure

NEC LX 102Bk-6



不揮発性メモリの利用法

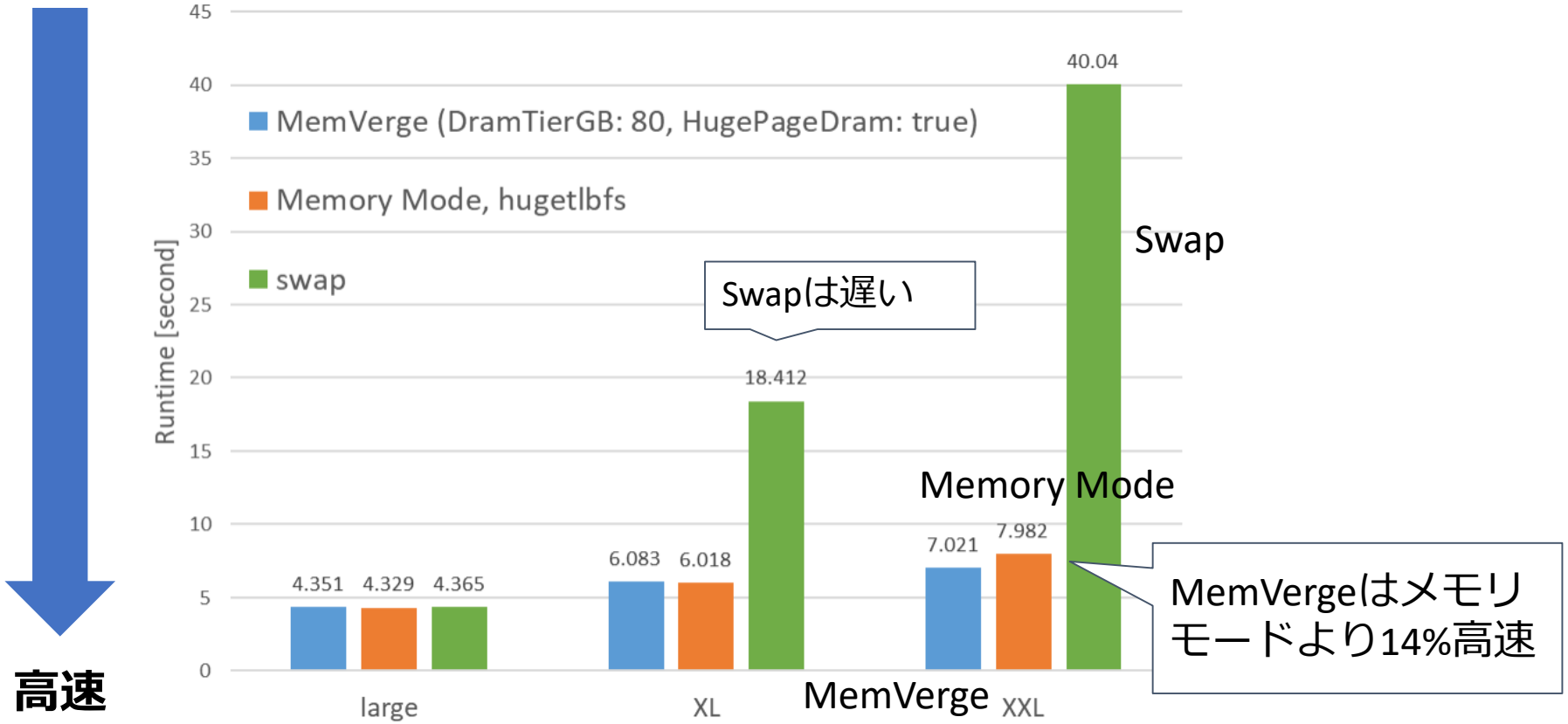
- 不揮発性メモリはダイレクトモードで設定
- 利用者は拡張メモリサイズ（P GiB）を指定



- MemVergeメモリマシンを用いたメモリ拡張

XSBenchベンチマーク (Optane PM 100)

- モンテカルロニュートリノ輸送アルゴリズムの代理アプリ
- メモリ使用量：Large 5.6 GB, XL 120GB, XXL 252GB
- メモリアクセスパターン：ランダム読込

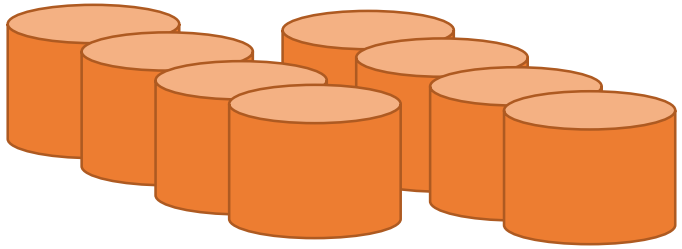


CygnusとPegasusの比較

	Cygnus (2019)	Pegasus (2022)
PFLOPS (DP)	2.3	> 6.1 (2.7x)
CPU	0.16	?.?? (?.?x)
GPU	2.18	6.12 (2.8x)
FPGA (SP)	0.64	0
Storage (PB)	2.4	7.1 (3.0x)

アドホック並列ファイルシステムの研究

- 計算ノードのストレージで構成する一時的な並列ファイルシステム
- 並列ファイルシステムと計算ノードの性能ギャップを埋める



- 不揮発性メモリを活用したCHFSアドホックファイルシステムを開発
  - 性能とスケーラビリティのためメタデータサーバ・逐次処理なし

CHFSの設計目標 [HPC Asia 2022]

- 不揮発性メモリを活用
  - インメモリ永続KVSを利用
- メタデータアクセスのオーバヘッドを極力軽減、スケーラブルに性能を向上させる
  - メタデータサーバなし
  - 逐次処理・集中データ構造なし
- データアクセス性能をスケーラブルに向上させる
  - （バイト単位）チャンクに分割

- 分散キーバリューストアを基盤とし集中データ構造をもたない設計

ファイルシステムの設計

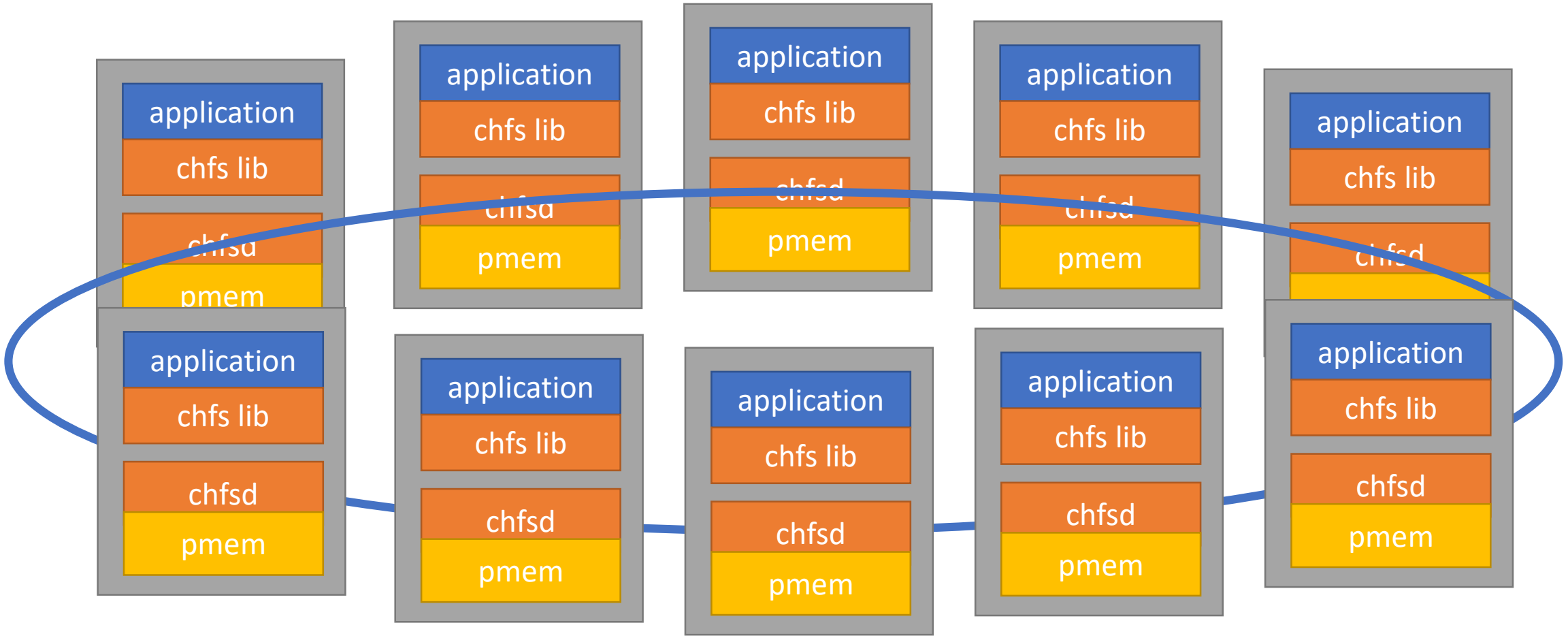
- 分散キーバリューストアにメタデータ、ファイルデータを保持

Key	Value	
フルパス名＋チャンク番号	メタデータ	ファイルデータ

メタデータ（64 byte）
mode, uid, gid, size chunk_size mtime, ctime

- チャンク数、全体のファイルサイズは保持しない
- ディレクトリも同形式

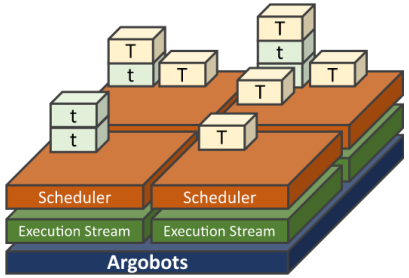
CHFSのシステムアーキテクチャ



CHFSの実装

- Mochi-Margo [JCST 2020]
  - https://mochi.readthedocs.io/en/latest/
  - Mercury, Argobots等を用いた通信ライブラリ
- Mercury [Cluster 2013]
  - 非同期RPC, RDMA通信ライブラリ
  - libfabric, UCX, 共有メモリプラグイン
- Arogbots [IEEE TPDS 2018]
  - 軽量スレッドライブラリ
- pmemkv
  - インメモリ永続KVS

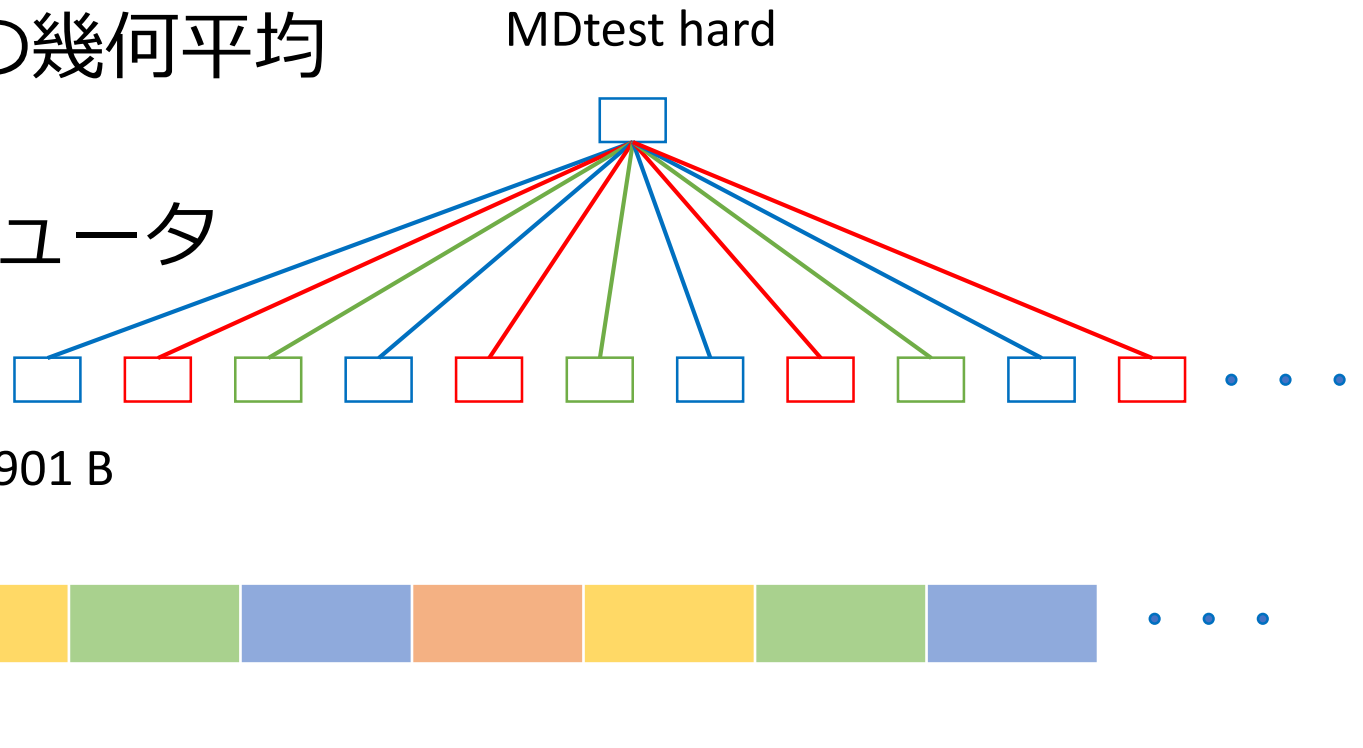
Mochi



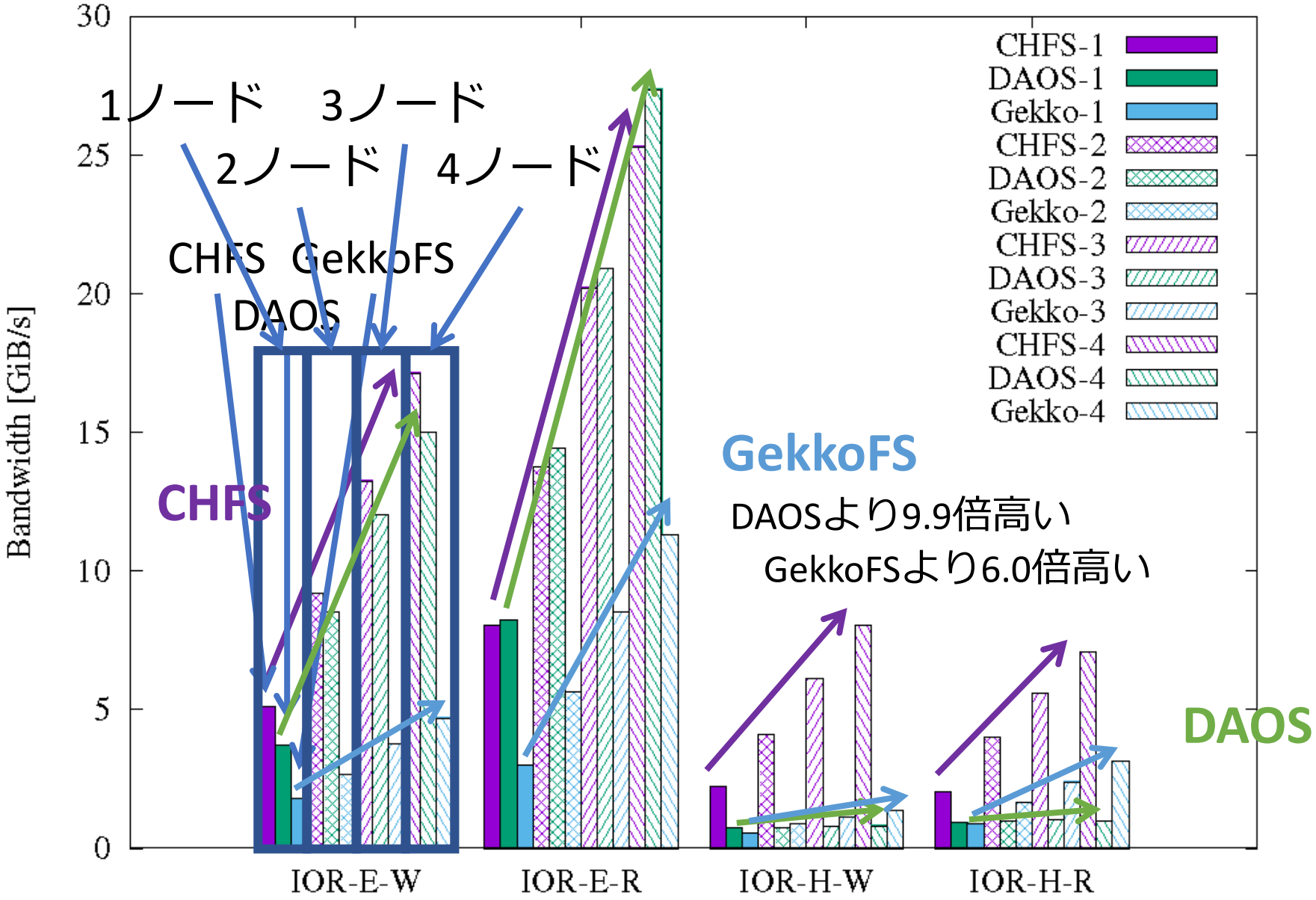
pmem.io

性能評価

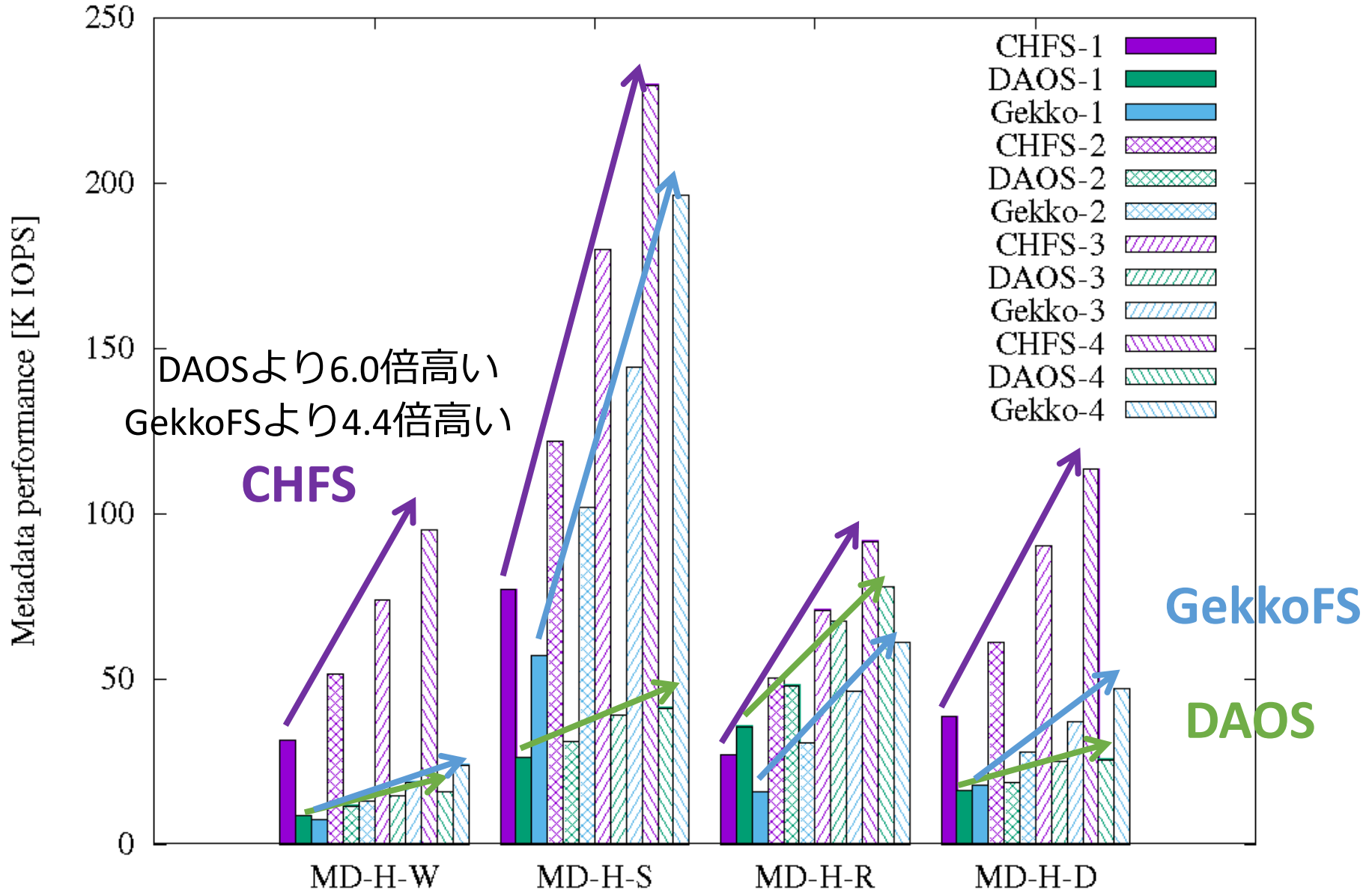
- IO500ベンチマーク
  - BW: IOR easy/hard – 4 benchmarks
  - MD: MDtest easy/hard, Find – 8 benchmarks
- スコアはベンチマーク結果の幾何平均
- 4ノードPMEMクラスタと
  78-node Cygnusスーパコンピュータ



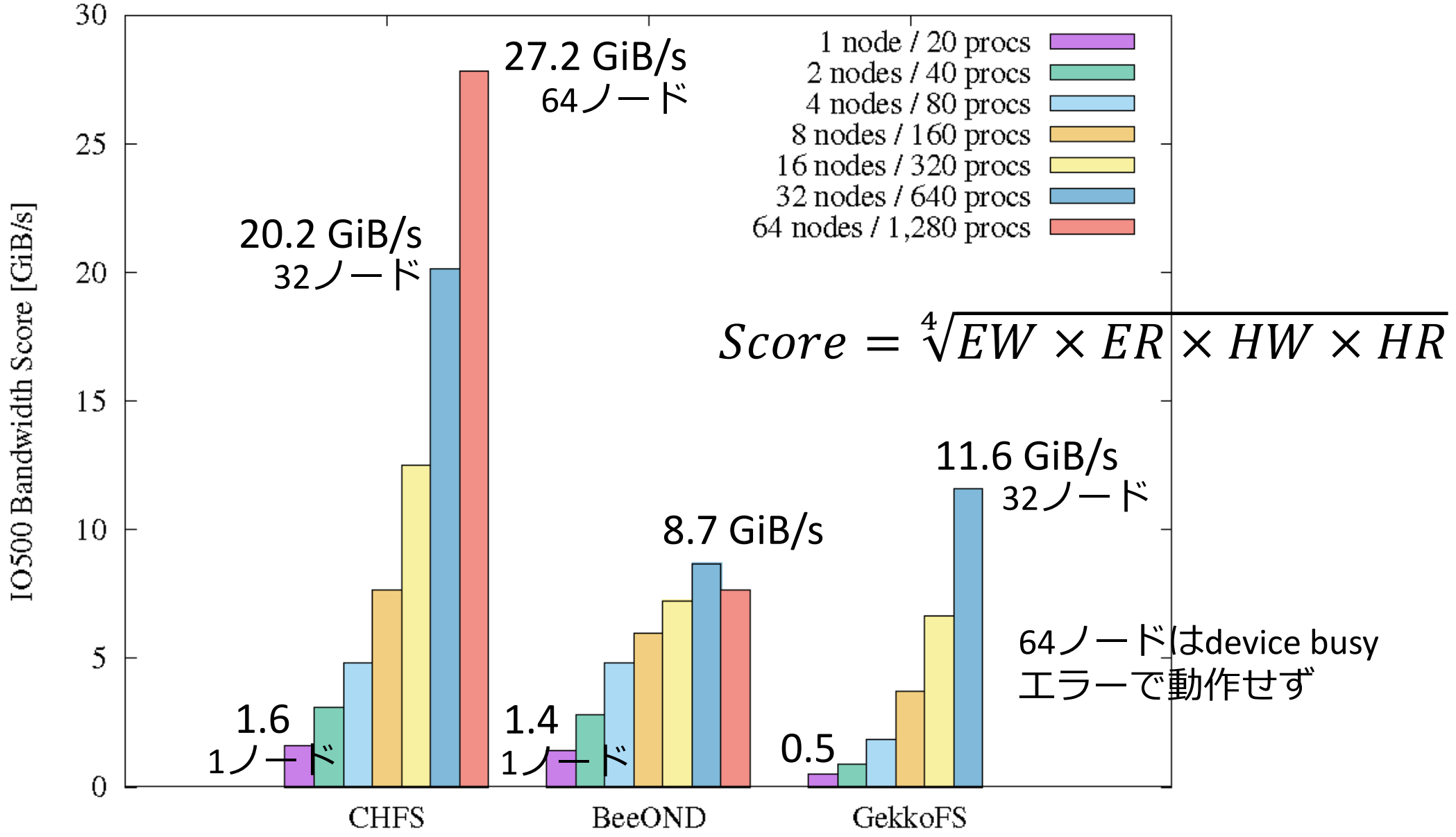
IO500バンド幅 (CHFS/DAOS/GekkoFS)



IO500メタデータハード (CHFS/DAOS/GekkoFS)



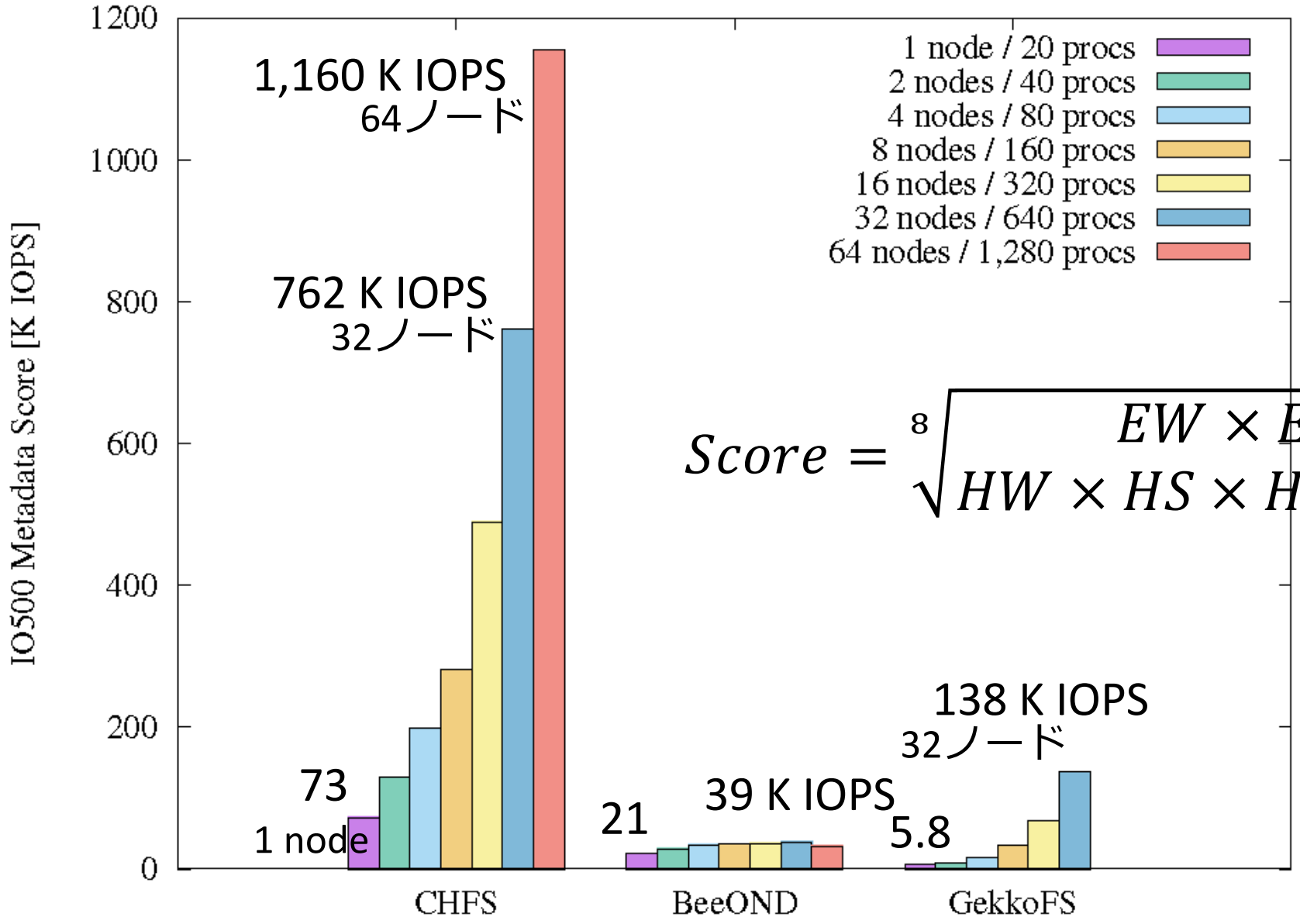
IO500バンド幅スコア on Cygnus (CHFS/BeeOND/GekkoFS)



$$Score = \sqrt[4]{EW \times ER \times HW \times HR}$$

64ノードはdevice busy
エラーで動作せず

IO500メタデータスコア on Cygnus (CHFS/BeeOND/GekkoFS)



$$Score = \sqrt[8]{\frac{EW \times ES \times ED \times}{HW \times HS \times HR \times HD \times FIND}}$$

3	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
5	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120	415.04	112.17	1,535.63
9	SC19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	249.50	86.97	715.76
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800	239.37	45.79	1,251.32
11	ISC21	Olympus Storage Technology Innovation Lab	OceanStor	Huawei	OceanFS	10	960	220.10	69.49	697.15
12	SC20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS)& BSC (NEXTGenIO)	GekkoFS	10	240	167.64	22.97	1,223.59
13	SC20	DDN	DIME	DDN	IME	10	110	161.53	101.60	256.78
14	SC19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2,610	156.51	56.22	435.76
15	ISC21	University of Tsukuba	Cygnus	OSS	CHFS	10	240	148.69	30.39	727.61
16	ISC21	Joint Institute of Nuclear Research	Govorun	RSC	DAOS	10	160	132.06	20.19	863.69
17	SC20	TACC	Frontera	DDN	IME	10	280	109.91	176.23	68.55

#15 in 10 node list

#23 in full list

CHFS/Cacheの設計目標 [ESSA 2022]

- ノードローカル不揮発性メモリを利用
- 高いメタデータ性能、バンド幅、スケーラブルな性能
  - CHFSをベースにキャッシュ機能を追加
- メタデータ性能を犠牲とせずバックエンド並列ファイルシステムと同期
  - 分かりやすく一貫性を緩和

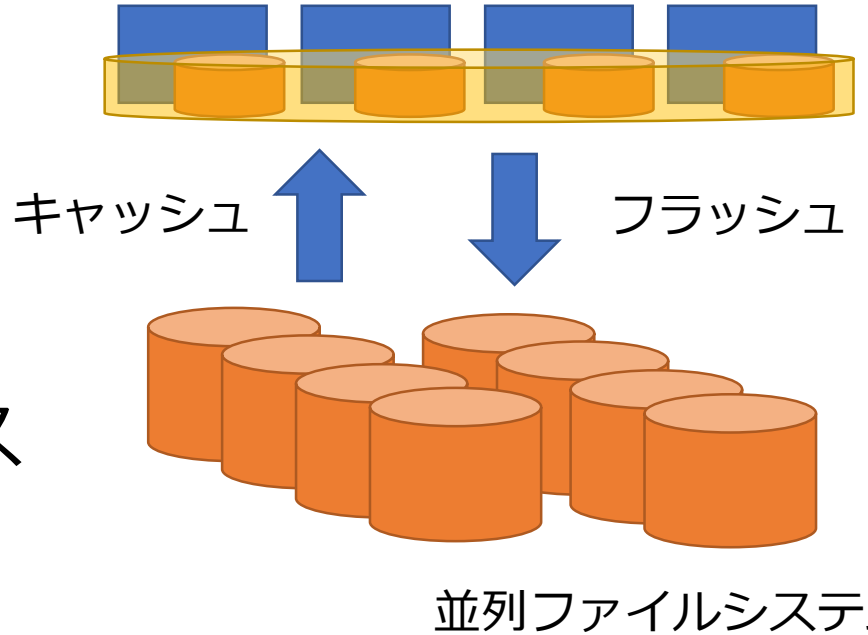
Oakforest-PACSのIO500スコア

	ハンド幅 [GiB/s]	メタデータ [kIOP/s]	合計
Lustre	21.4	88.78*	42.18*
IME	471.25	21.85	101.48

一貫性の緩和

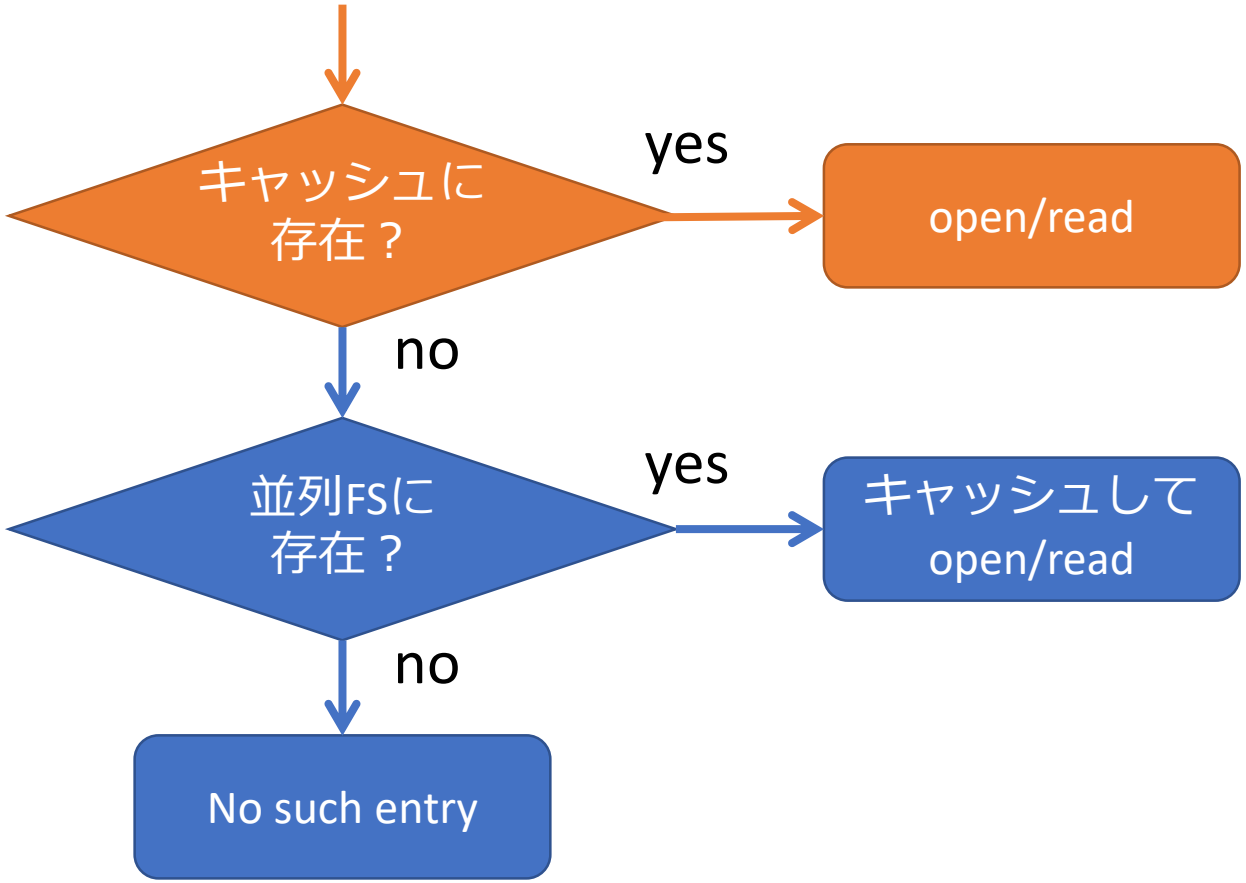
仮定

1. 入力データはジョブ実行中は変更されない
2. ジョブの新規エントリ作成は成功する
3. ジョブが既存ファイルを更新するときは読込んでから
4. 更新はフラッシュするまで反映されない
   （キャッシュファイルシステムをアクセスすれば分かる）
5. フラッシュはジョブ終了までに行われる



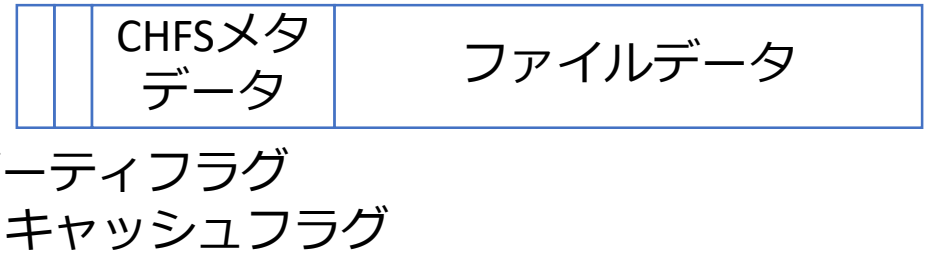
ファイル操作の設計

- ファイルオープン・読込
  - 右図（仮定1より）
  - キャッシュに存在すれば
    並列FSに対し更新の確認
    はしない
- ファイル生成・書込
  - キャッシュに生成・書込
    （仮定2, 3, 4より）
  - 並列FSは参照しない



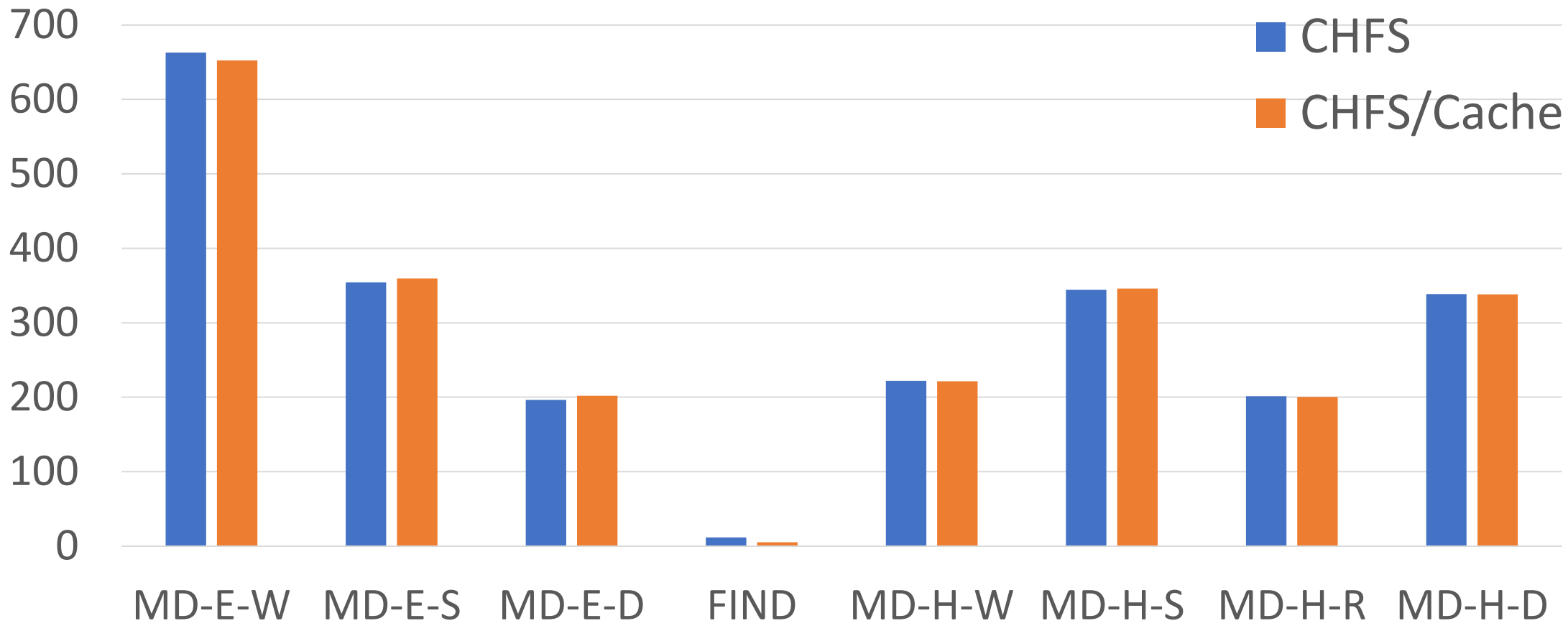
キャッシュ機構の設計

- メタデータフラグの設計
  - ダーティフラグ – 更新を示す
  - キャッシュフラグ – 既存ファイルを示す
    - readdirで重複して返さないため



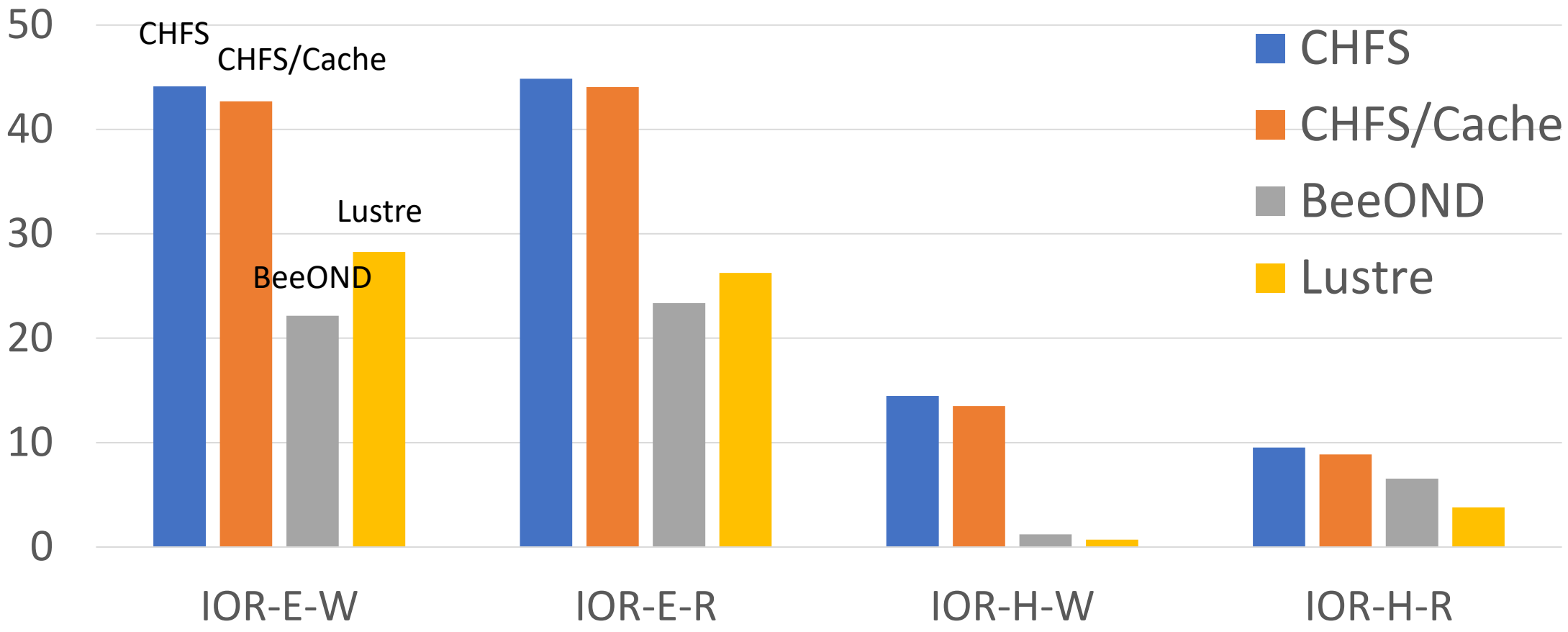
- オープンフラグの設計
  - キャッシュフラグ – ダーティではない書込みを示す
    - キャッシュしたファイルはクリーン

Cygnus Pノードにおけるメタデータ性能（4ノード）



オーバヘッドはFINDを除き3%以下。FINDは54%であったが、大きな問題ではない

Cygnusにおけるバンド幅（48ノード）



Lustreと比べIOR-Eは1.5～1.6倍、IOR-H-Wは19～20倍、IOR-H-Rは2.3～2.5倍高速

まとめ

- Pegasusは2022年Q4に導入予定
  - データ駆動・AI駆動科学のための高い演算性能と高帯域大容量メモリ、超高速ストレージを提供
- CHFSアドホックファイルシステムの開発
  - 不揮発性メモリを活用しスケーラブルな性能を実現
  - 2021 June IO500 10ノードリストで15位、全体リストで23位
  - 2022/10/21 CHFS 2.0.0をリリース
    - https://github.com/otatebe/chfs
- CHFS/Cacheキャッシング並列ファイルシステムの開発
  - 一貫性緩和でメタデータ性能問題を解決
  - バンド幅、メタデータ性能ともにバックエンドPFSを上回る