

Gfarm Workshop 2020
2020年2月7日 @宮崎

Gfarmファイルシステムの 概要と最新機能

建部修見
筑波大学

Gfarmファイルシステム

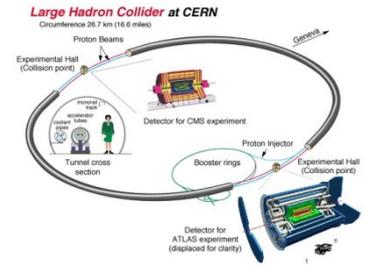


- オープンソース広域分散ファイルシステム
 - <http://oss-tsukuba.org/software/gfarm/>
- サポート
 - NPO法人つくばOSS技術支援センター(日本他)
 - Libre Solutions Pty Ltd(オーストラリア)
- 特徴
 - 性能・容量がスケールアウト
 - データアクセス局所性、ファイル複製
 - 無停止で拡張、更新可能
 - 単一障害点なし
 - 複製数維持機能、ホットスタンバイMDSサーバ
 - データ完全性を保証しサイレントデータ損傷も対応可

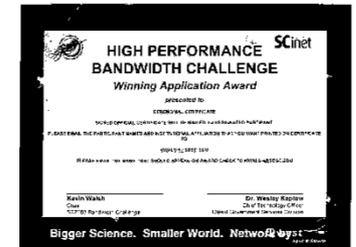
ossTsukuba
oss-tsukuba.org



Congratulation 20th anniversary!



- 2000年6月 電総研と高エネ研のGRIDミーティング
– LHC実験の要求に応えるため日本独自研究開発開始
- 2002年11月 日米間ファイル複製実験で世界最高の741Mbpsを達成！
- 2003年11月 SC03において「分散インフラストラクチャ賞」を受賞！ Gfarm 1.0を公開
- 2005年9月 Gfarm Workshop 2005開催！ Gfarm 1.2公開
- 2005年11月 SC05において「Most Innovative Use of Storage In Support of Science賞」を受賞！
- 2006年11月 SC06におけるHPC Storage Challengeで優勝！ Gfarm 1.4を公開
- 2007年11月 Gfarm 2.0.0公開！
- 2019年11月30日 Gfarm 2.7.16公開！！！！



HPCI共用ストレージ

- 大学情報基盤センターをはじめ全国からマウント可能な共有ファイルシステム(～100PB)
- スパコン間のデータ共有、共有データ格納



西拠点 (R-CCS)



東拠点 (東京大)

最新機能・状況紹介

主なリリース

日付	version	新機能、更新機能
2019/11/30	2.7.16	• Githubへの移行！
2019/10/24	2.7.15	• Gfarmbb status, IB GRH対応
2019/9/10	2.7.14	• Gfarm/BBノバーストバッファ
2016/12/8	2.7.0	• InfiniBand RDMAサポート • ディレクトリクオータ
2016/1/16	2.6.8	• 書込後ベリファイ

https://github.com/oss-tsukuba/gfarm

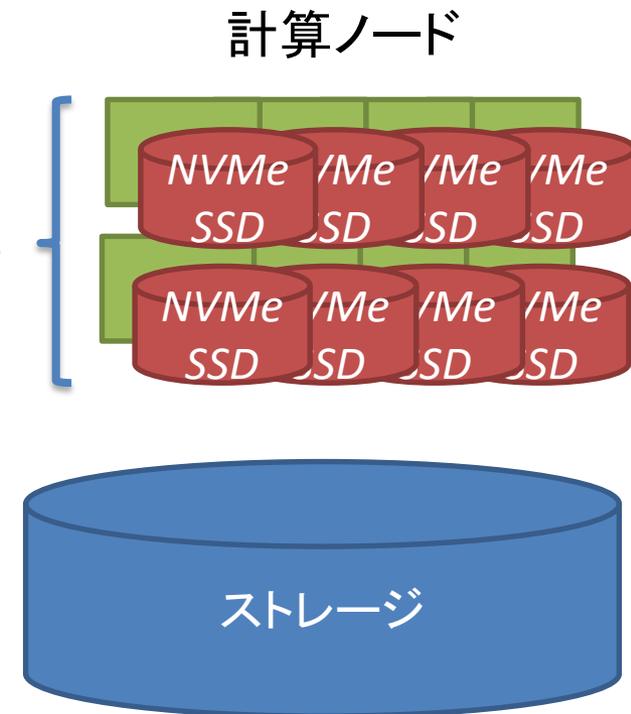
The screenshot shows the GitHub repository page for `oss-tsukuba/gfarm`. The browser address bar displays `github.com/oss-tsukuba/gfarm`. The repository name `oss-tsukuba / gfarm` is shown at the top, along with interaction buttons: `Unwatch` (4), `Star` (4), and `Fork`. Below this is a navigation bar with tabs for `Code`, `Issues` (138), `Pull requests` (0), `Actions`, `Projects` (0), `Wiki`, `Security`, `Insights`, and `Settings`. The repository description reads: "distributed file system for large-scale cluster computing and wide-area data sharing. provides fine-grained replica location control." Below the description is a "Manage topics" link. A statistics bar shows: 5,259 commits, 30 branches, 0 packages, 191 releases, and 10 contributors. A progress bar is visible below the statistics. Below the statistics bar are buttons for `Branch: master`, `New pull request`, `Create new file`, `Upload files`, `Find file`, and `Clone or download`. The commit history table is as follows:

Commit	Message	Time
otatebe	libpq-devel for RHEL8	Latest commit cf6140d 6 days ago
bench	merge r10991 from the replica_check_2 branch (via r10993 from the 2.7...	12 months ago
debian	remove gftool	4 years ago
doc	regen for #1081 - gfmv(1) between different directory quota settings	16 days ago
gftool	fix data race	6 days ago
include/gfarm	assign log message numbers	2 months ago
lib	fix out-of-bound access	6 days ago
linux	assign log message numbers in linux	3 years ago

Gfarm/BBノーストバッファ [建部 2020]

- ノードローカルNVMe SSD等高速ストレージによる一時的な分散ファイルシステム
- アクセス性能の向上
 - ファイルディスクリプタパッシングによるgfsdを経由しない直接アクセス
 - RDMAアクセス
- メタデータ性能の向上
 - メタデータの永続性、冗長性なし
 - ジャーナル書込み、バックエンドDB、スレーブgfsdのオーバヘッドの削減
- 冗長性オーバヘッド削減
 - ファイル複製によるデータの冗長性なし
- ファイルシステム構築、撤去の高速化

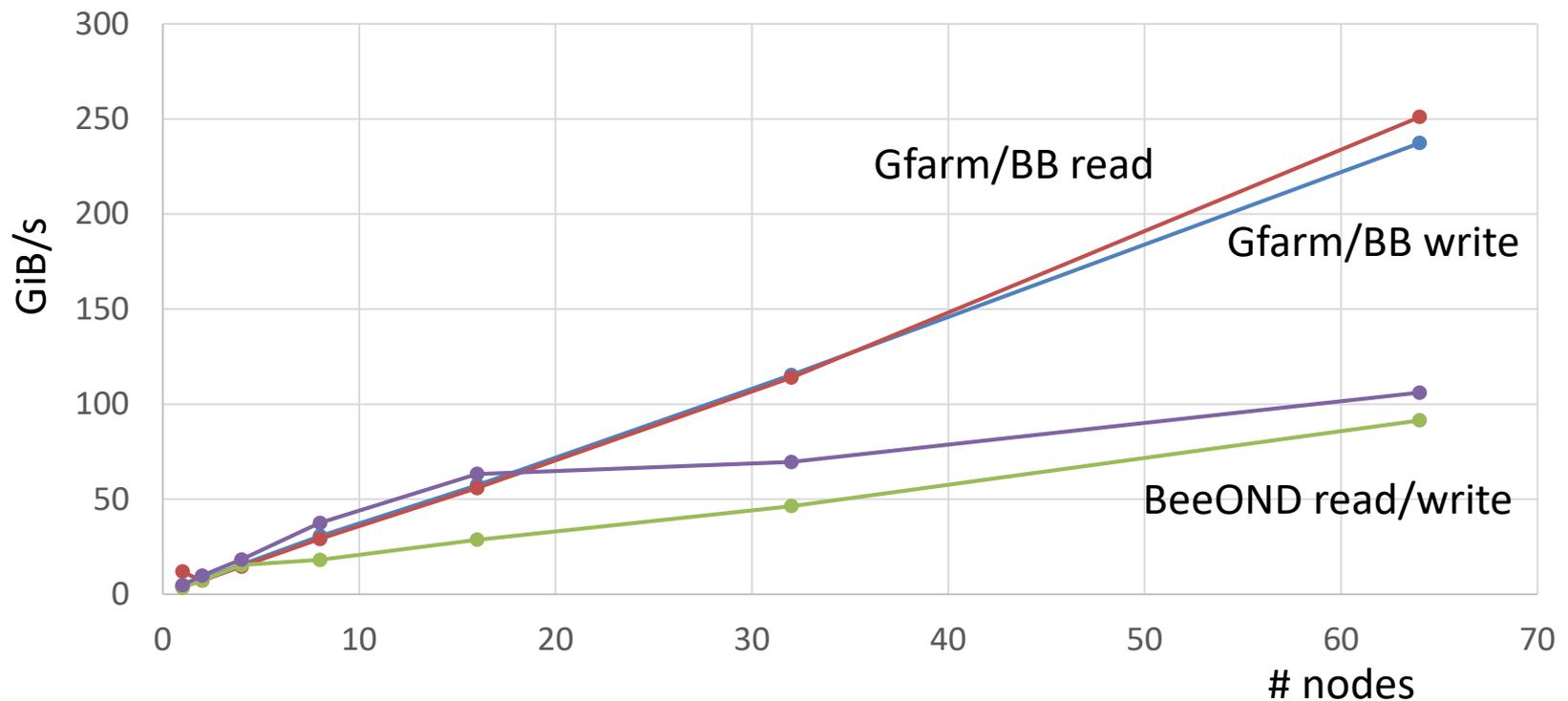
Gfarm/BB



Gfarm/BBノバーストバッファ(2)

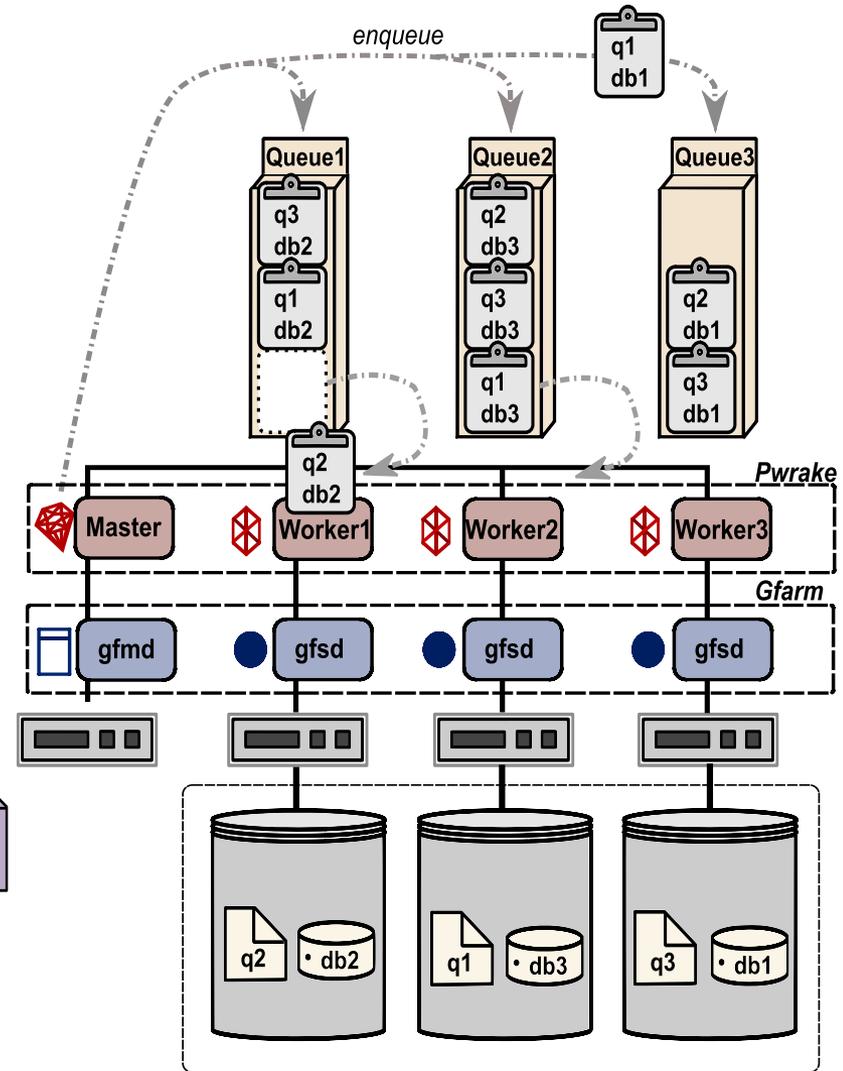
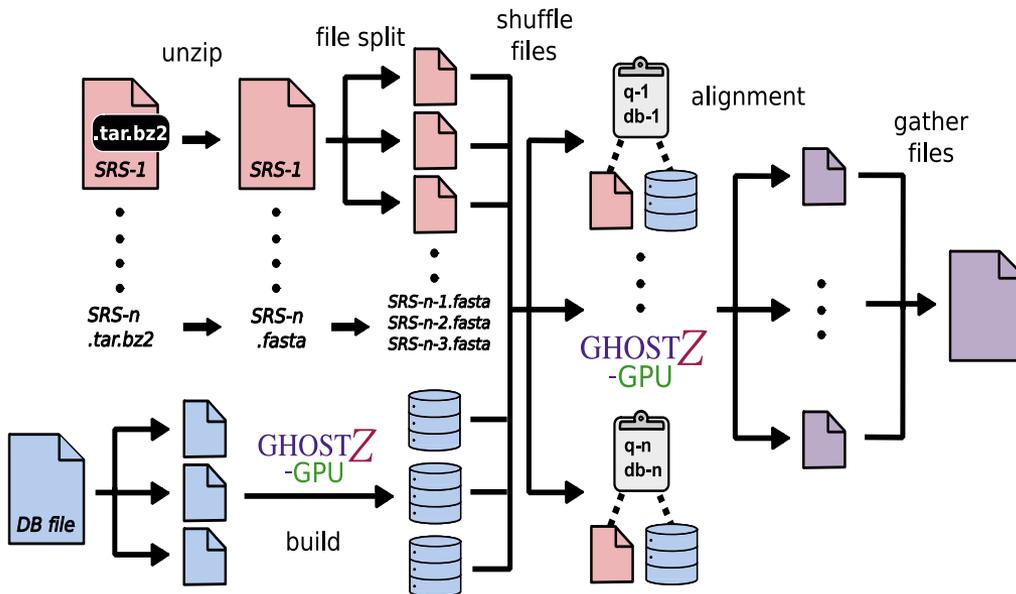
```
gfarmbb -h hostfile -m mount_point start  
...  
gfarmbb -h hostfile stop
```

IOR – file-per-process read/write bandwidth on Cygnus
supercomputer



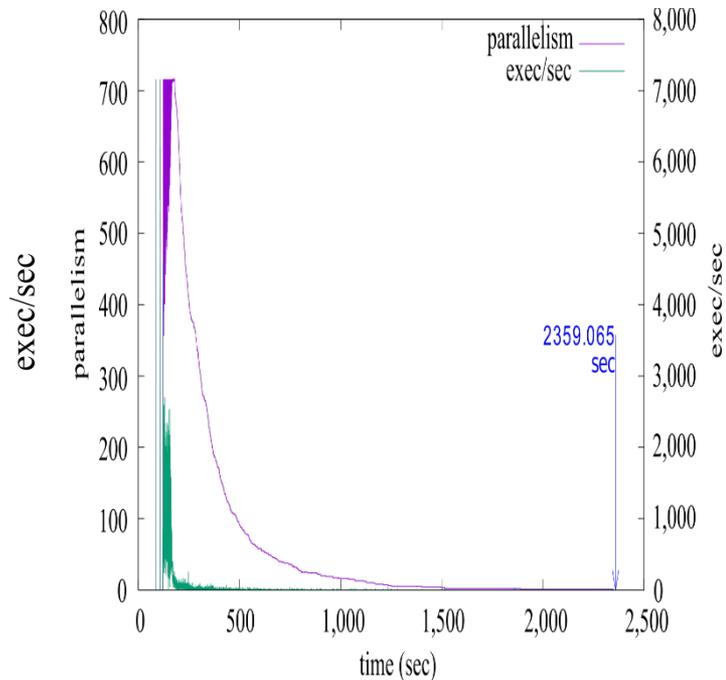
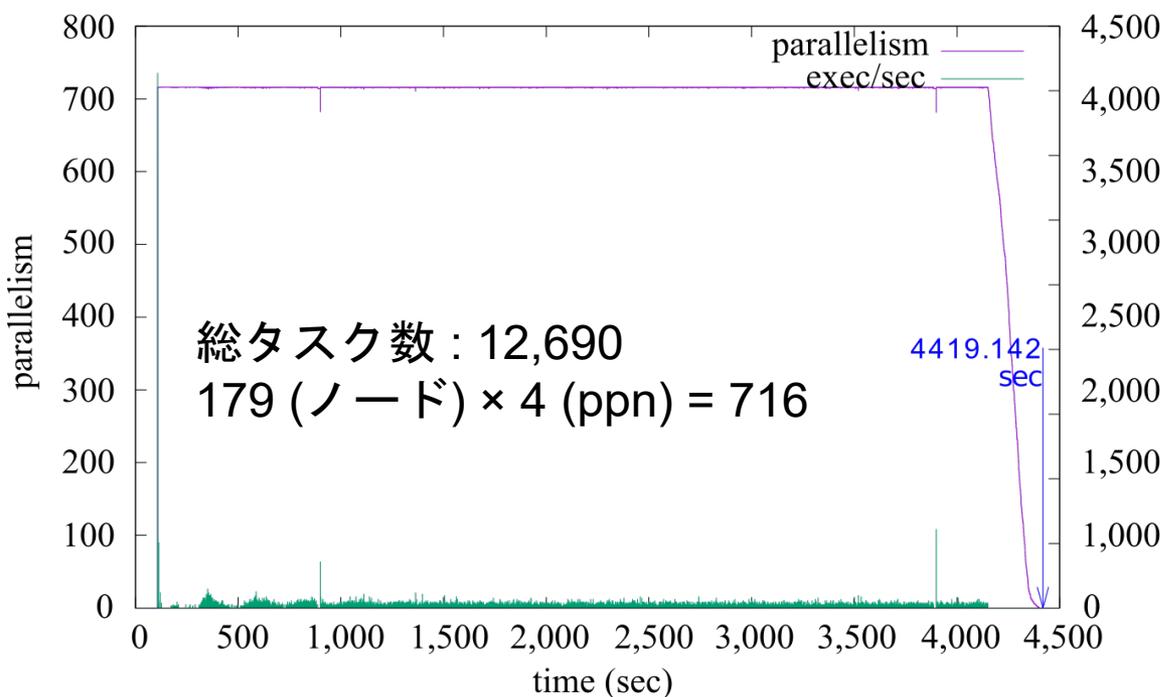
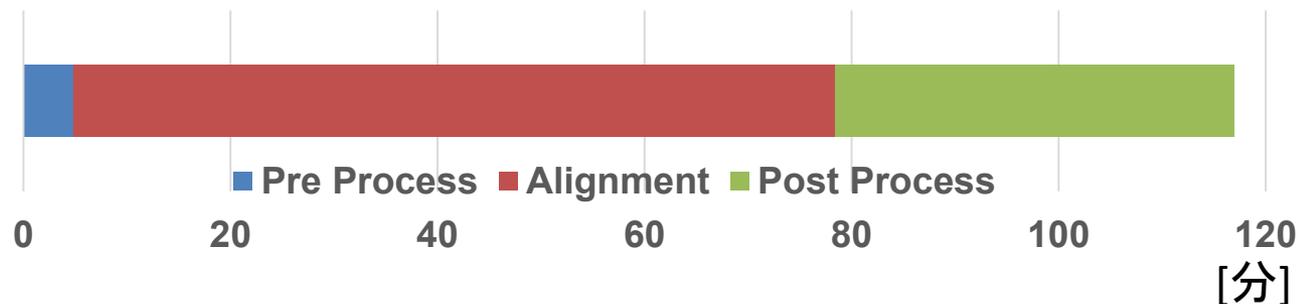
大規模メタゲノム解析 [町田 2019]

- クエリ、DBの増大
→ クエリ、DB双方を分割して分散計算
- Gfarm/BB, Pwrake, GhostZ-GPU

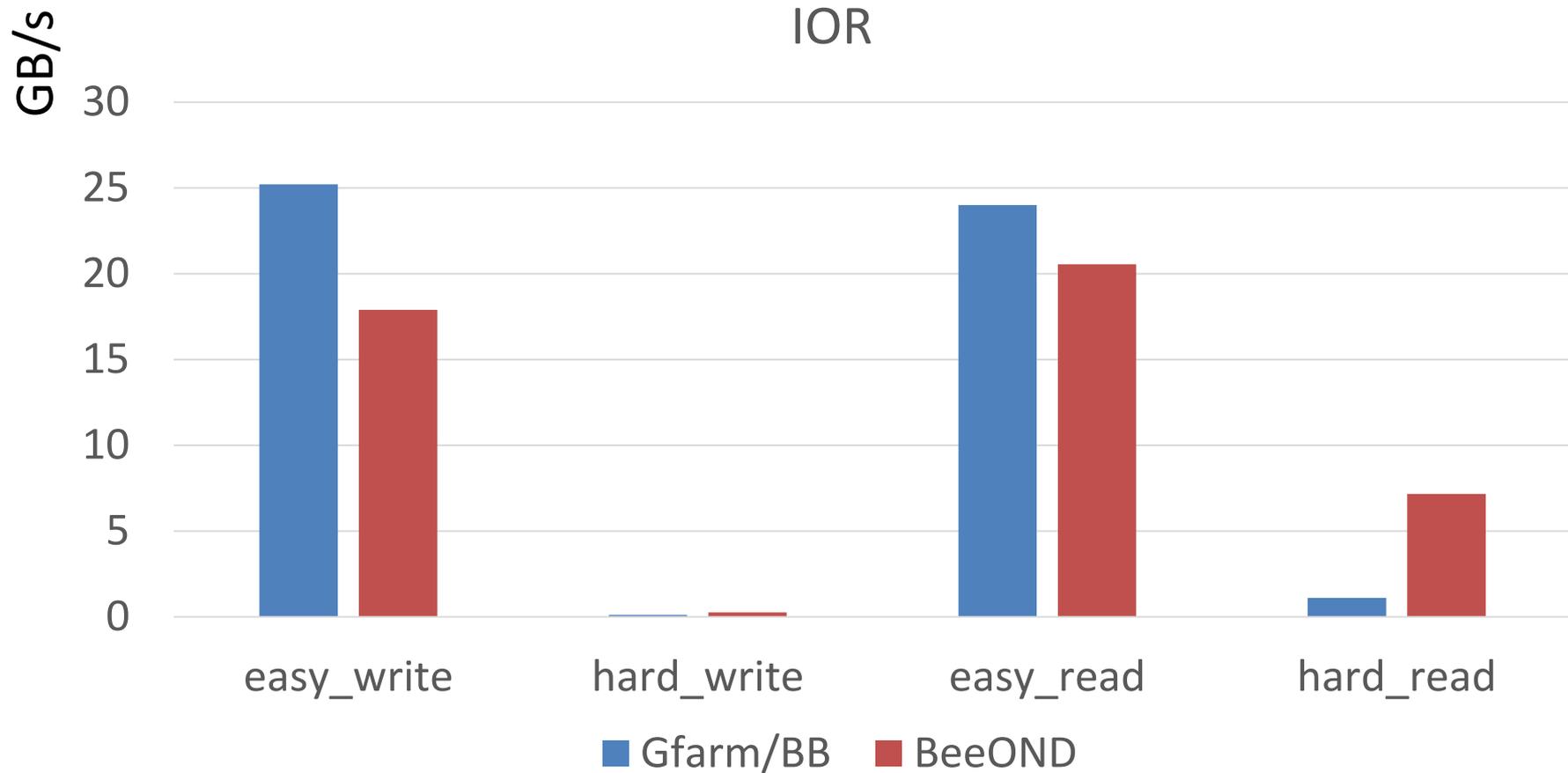


TSUBAMEグランドチャレンジ

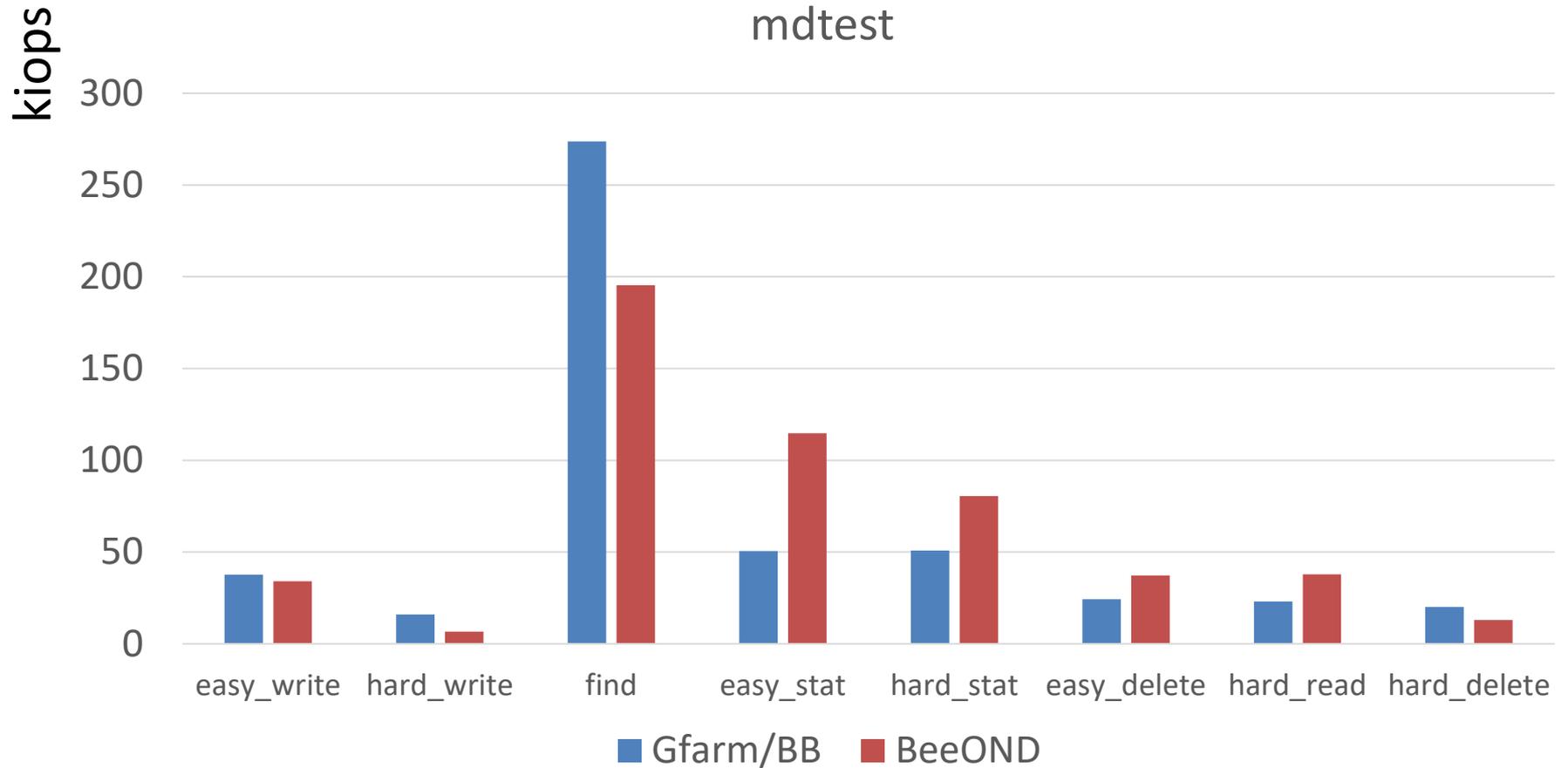
- 180ノード
- 62GB nr DB,
71GB クエリ
(口内歯垢)



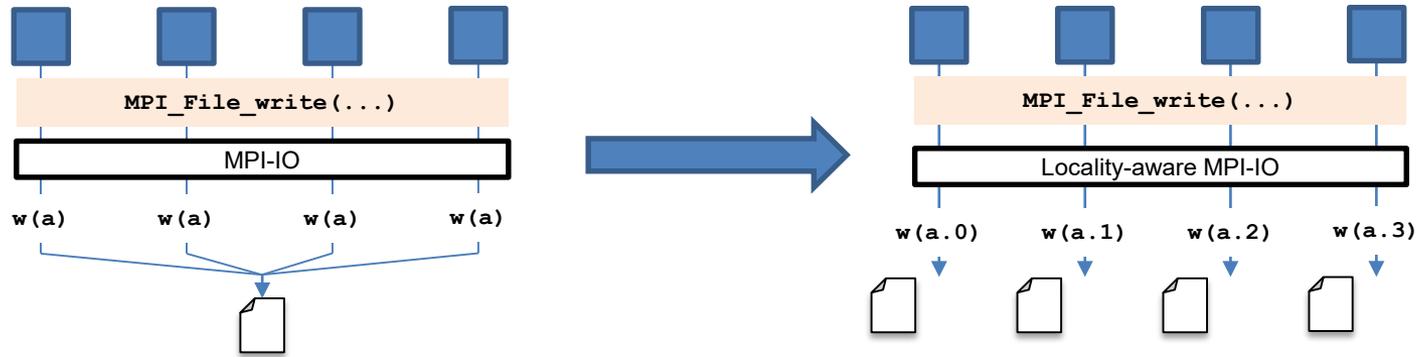
IO500 10 node challenge (1)



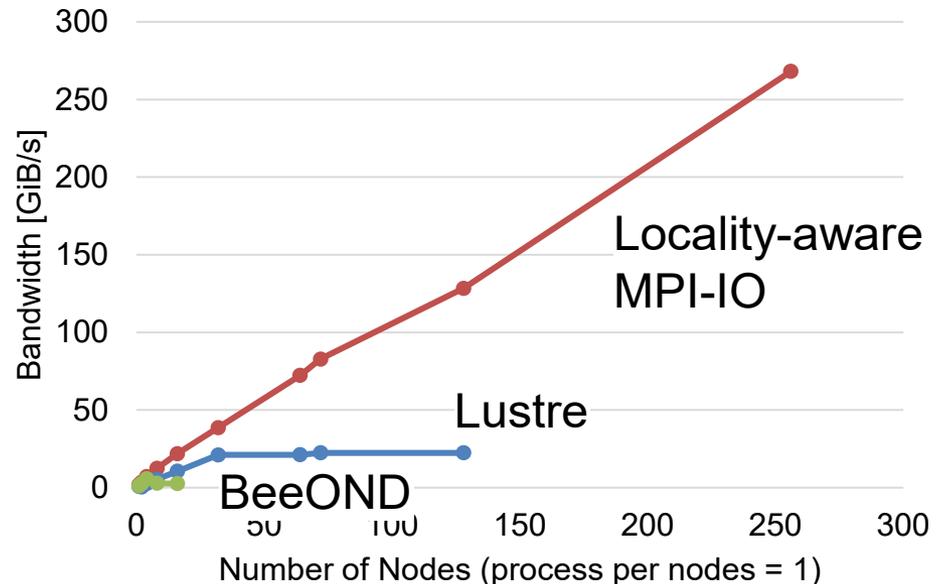
IO500 10 node challenge (2)



Locality-aware MPI-IO [M1 杉原]



- MPI-IOで分割セグメントの内部形式で格納して、N-1アクセス→N-Nアクセスに変換
- IOR N-1アクセス



まとめ

- Gfarmファイルシステム
 - NPO法人つくばOSS技術支援センターによるサポート
 - Gfarm 2.7.16を11/30にリリース
 - <https://github.com/oss-tsukuba/>
- Gfarm/BBノバーストバッファ
- InfiniBand RDMA、ディレクトリクオータ機能
- データ完全性、サイレントデータ損傷対応
- HPCI共用ストレージ、JLDGなど実運用実績
- 進行中
 - IPv6対応 (Gfarm 2.8)
 - 暗号化対応