Gfarm Symposium 2019
2019年10月25日＠市ヶ谷

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/
  - 22,030 downloads since March, 2007
- 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

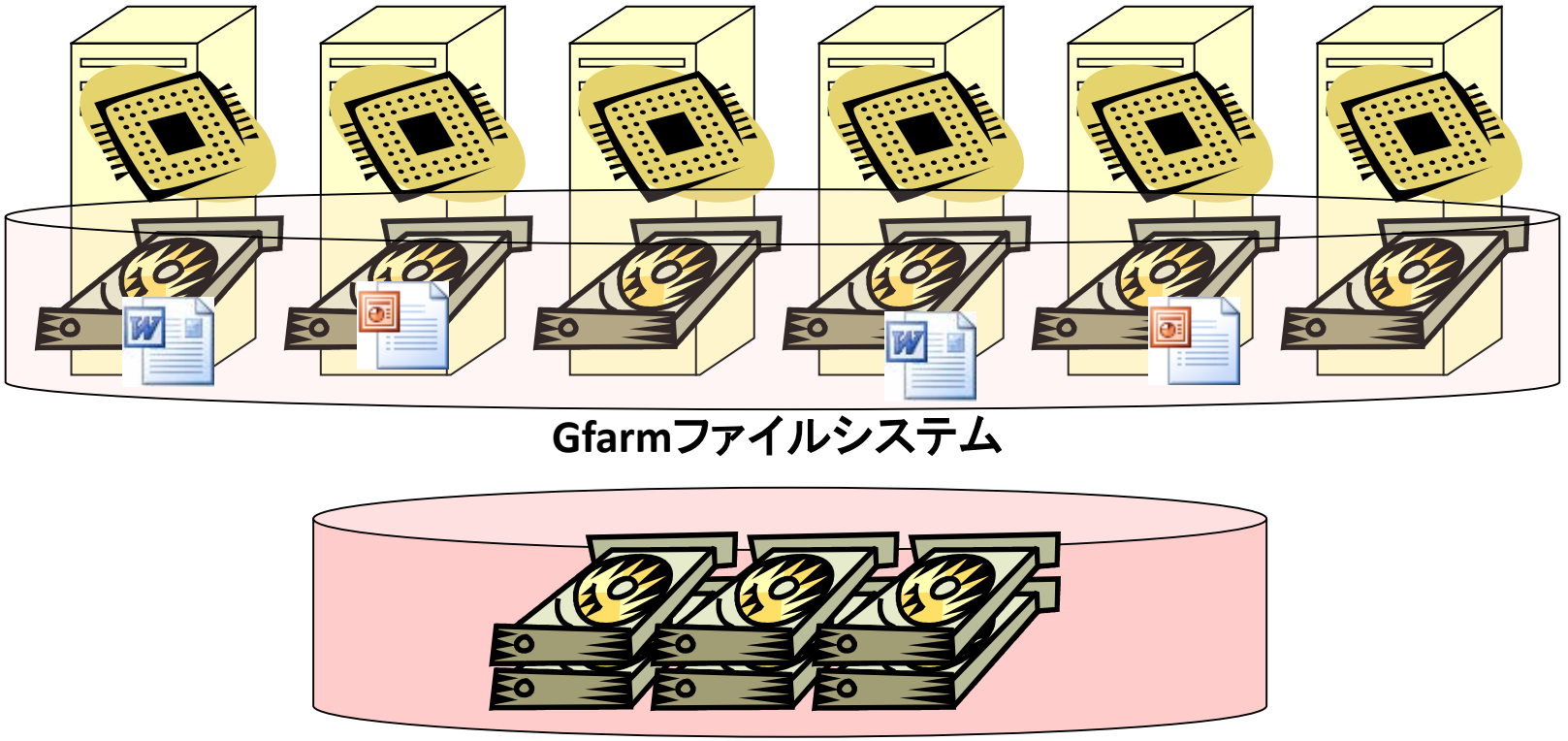
oss-tsukuba.org

Gfarmファイルシステム（2）

- JLDG（10.7PB、8拠点）、HPCI共用ストレージ（~100PB、2拠点）、NICTサイエンスクラウド、（株）クオリティアActive! world等で実運用
- 計算ノードのローカルディスクによるデータ解析
  - すばる望遠鏡データ解析、メタゲノム解析
- Pwrakeワークフローシステム、MapReduce、MPI-IO、バッチキューイングシステム
  - データ局所性を高めるプロセススケジューリング
  - ディスクキャッシュを有効利用するプロセススケジューリング
  - データ局所性を高めるファイル複製作成

Gfarmファイルシステムの構成

- ローカルディスクを束ねる
- ユーザには、共有ファイルシステムとしてみえる
- 複数のディスクに分散してデータを保持



オブジェクトストレージがダメな理由

- POSIXアクセスできない
  - ディレクトリがない
  - Read/writeができない
- ファイルをコピーしてからデータ解析
- スケールアウトはファイルシステムもできる
- IO-500が動かない

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）

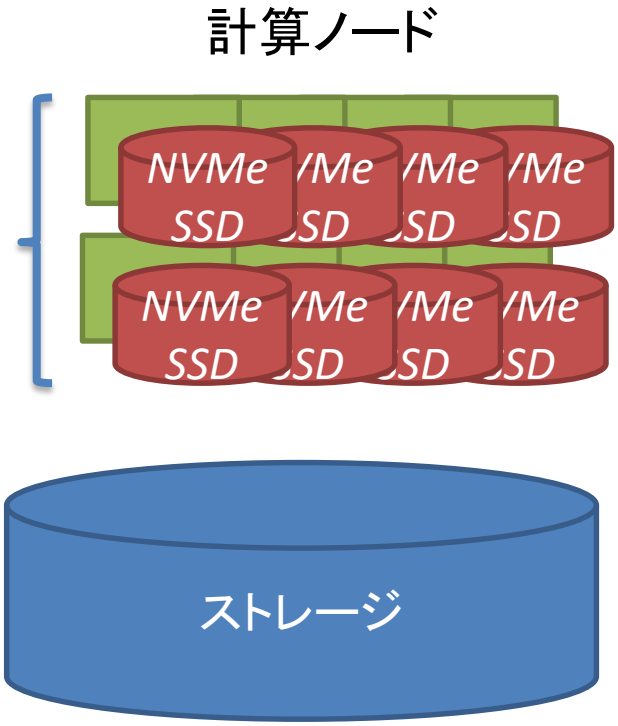
最新機能・状況紹介

主なリリース

日付	version	新機能、更新機能
2019/10/24	2.7.15	• Gfarmbb status, IB GRH対応
2019/9/10	2.7.14	• Gfarm/BBバーストバッファ
2019/3/22	2.7.13	• 読込オンリーモード
2017/10/1	2.7.6	• 書込キャッシュストレージ支援 • データ移行支援
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

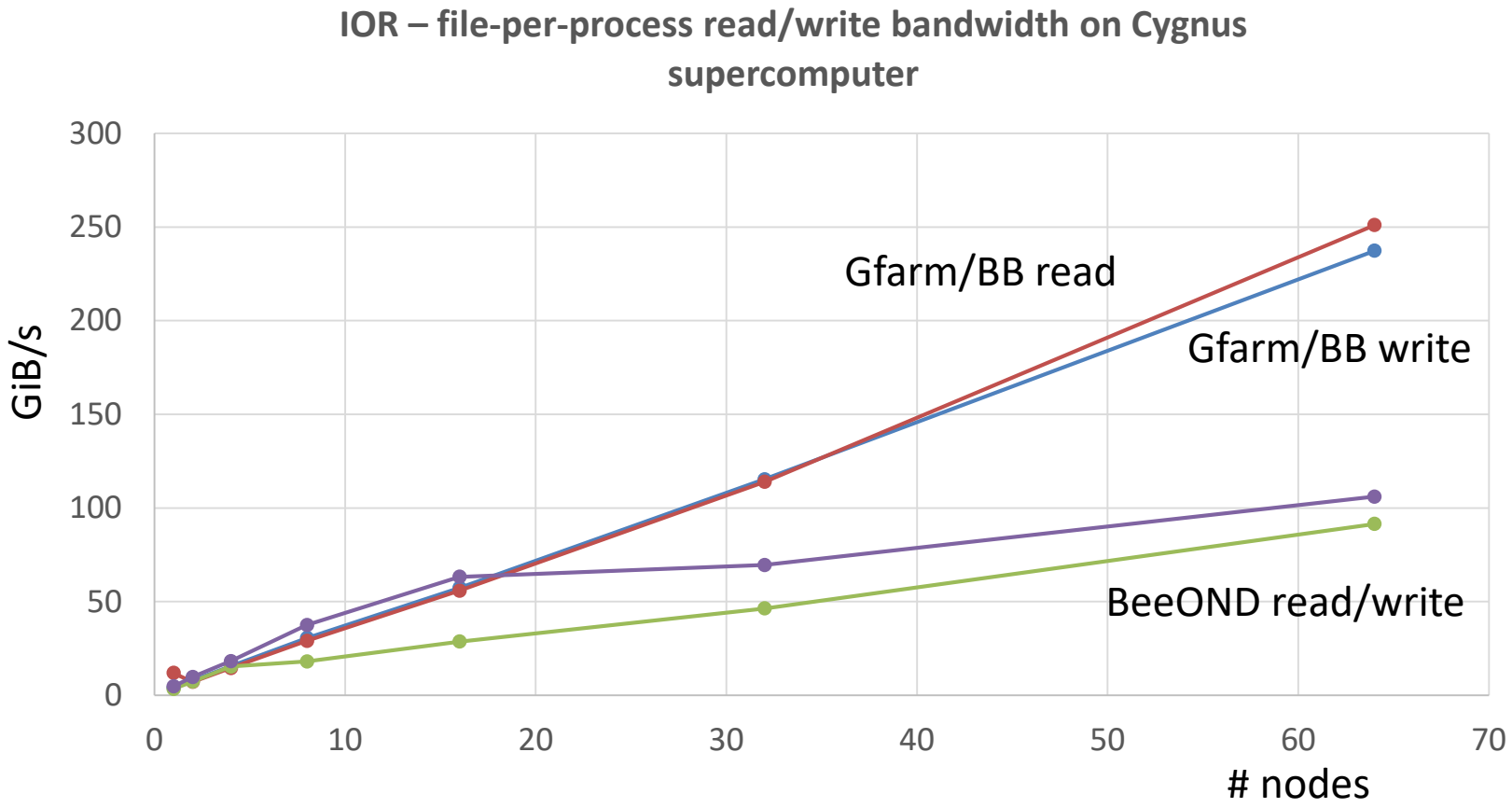
Gfarm/BBバーストバッファ

- 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gfmd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

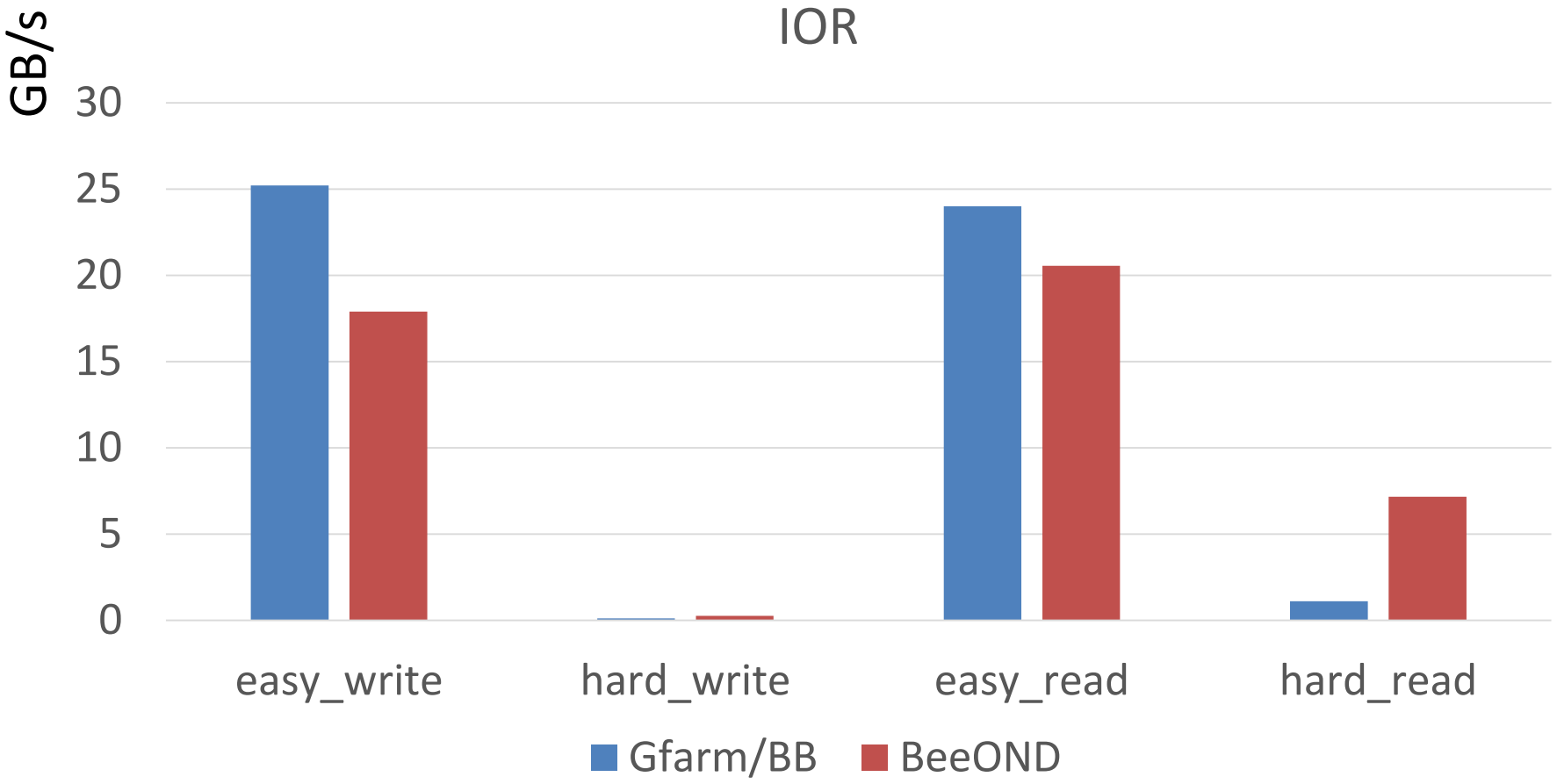


Gfarm/BBバーストバッファ（2）

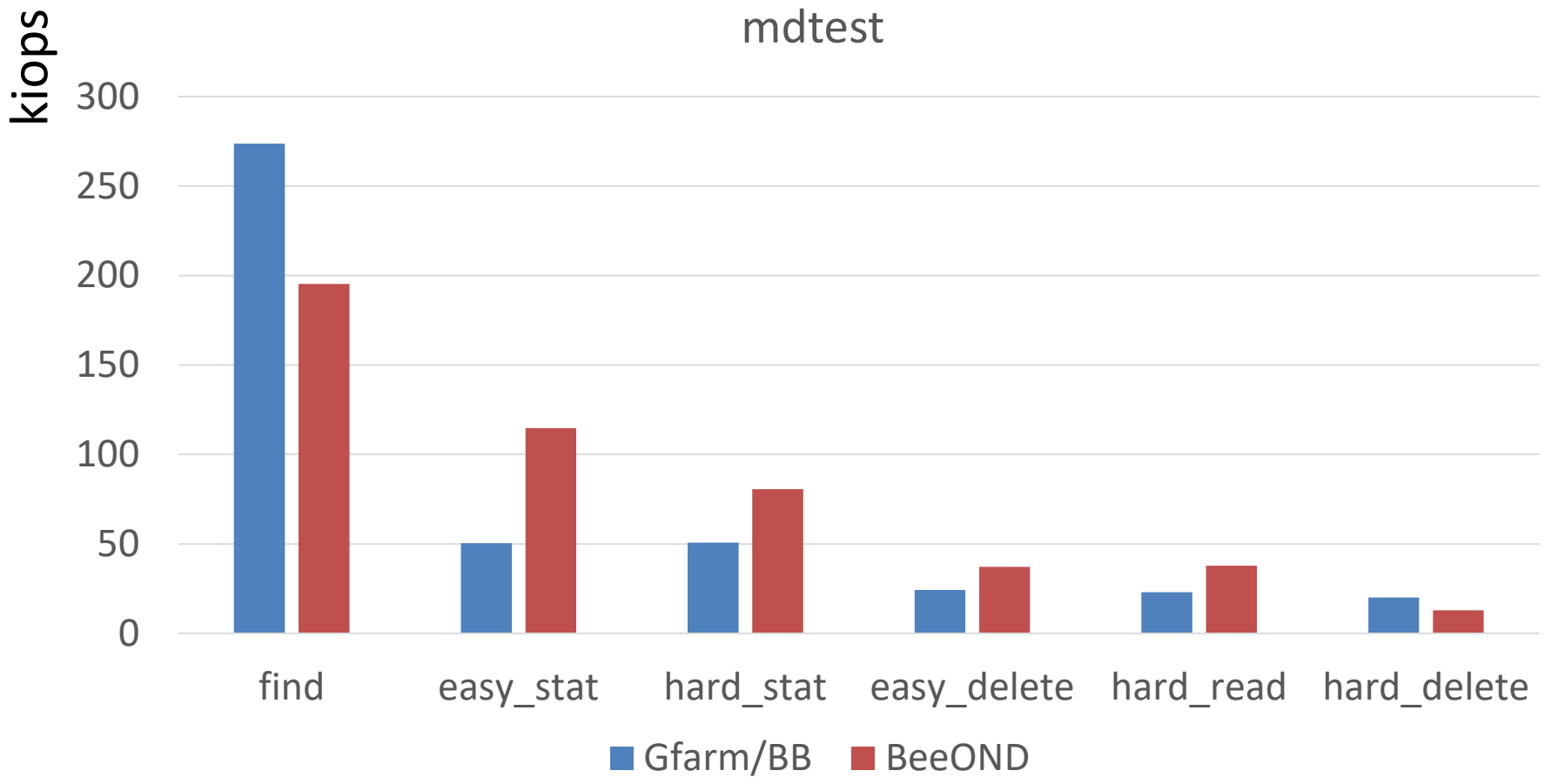
```
gfarmbb –h hostfile –m mount_point start
…
gfarmbb –h hostfile stop
```



IO500 10 node challenge (1)



IO500 10 node challenge (2)



ディレクトリクオータ機能

- ディレクトリ単位のファイル数、利用サイズの制限
  - XFSと異なりグループクオータと併用可、またユーザ権限で設定可能

- gfdirquota – ディレクトリセット作成、設定
  - 複数ディレクトリでクオータ制限可能
- gfedquota, gfquota – クオータ設定、表示

データ完全性

- サイレントデータ障害の検知
- 書込時にdigestを計算しメタデータに保存
- 読込時にgfsdでdigest検査
  - 破損ファイルは読込時にEIO (checksum error)を返し、読込失敗。lost+foundへ移動させ自動修復
- 複製作成時のdigest検査
- 書込後ベリファイによるdigest検査
- クライアントからのEnd-to-endのデータ完全性

JLDGにおける運用例

- 10.7 PB, 8拠点, 46ファイルサーバ
  - 物理学研究者による全国規模のストレージ
  - 9.9 PB利用, 111 Mファイル
- md5によるEnd-to-end一貫性チェックと書込み後ベリファイ利用
- 2016年8月19~22日
  - 書込み後ベリファイで5ファイル、複製作成で1ファイルの損傷ファイルを検出
  - I/Oエラーは起こっていない

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 2.7.15を10/24にリリース
- Gfarm/BB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
  - クラウドストレージ連携