Gfarm Workshop 2018
2018年3月2日沖縄

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/
  - 19,000 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

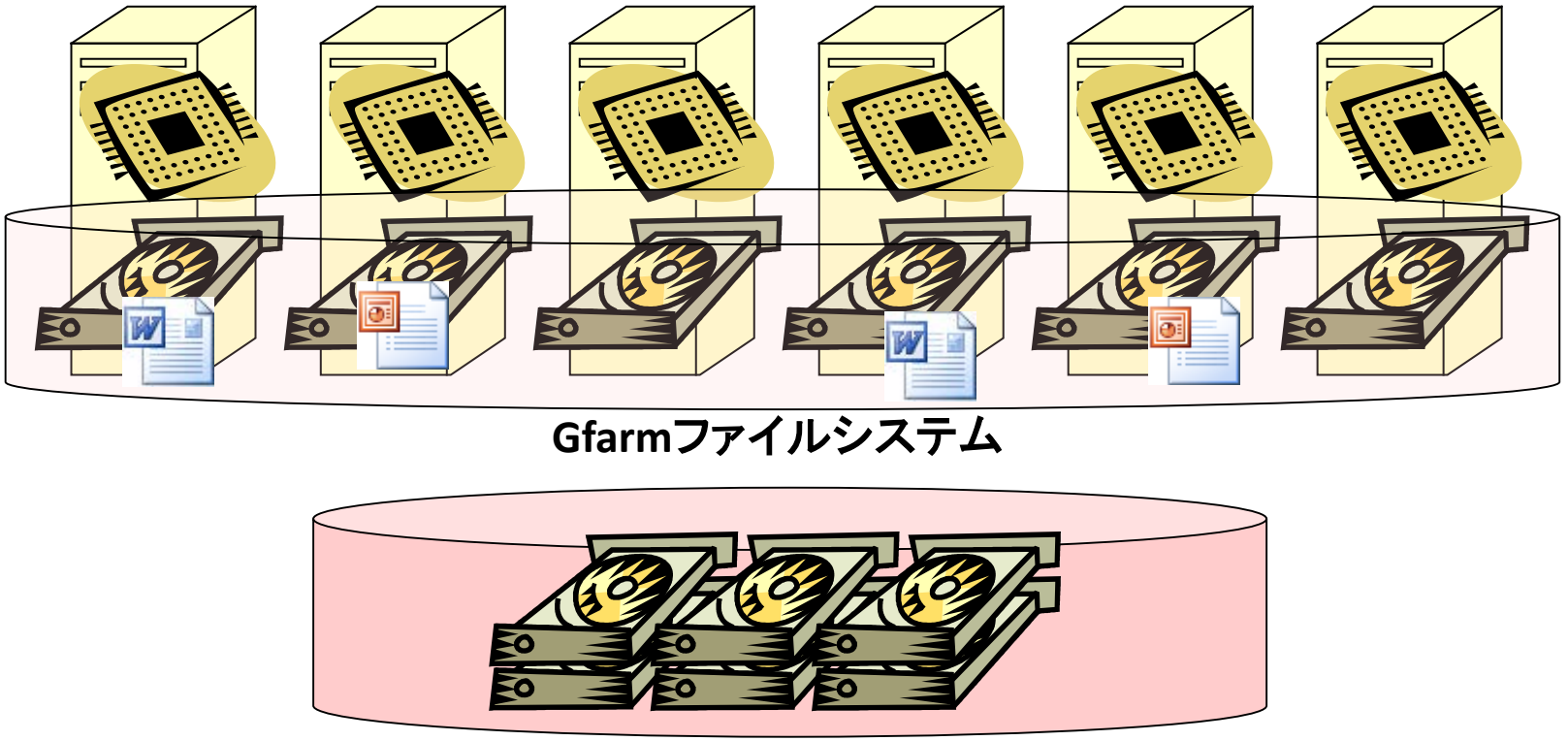
oss-tsukuba.org

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



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（AICS）

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
2018/2/22	2.7.10	• gfmdのレスポンス改善
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
2017/2/4	2.7.2	• Gfarm/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
2015/6/25	2.6.5	• 過剰複製削除
2014/12/17	2.6.0	• End-to-endデーター貫性サポート • クライアント透過なフェイルオーバ

gfmdのレスポンス改善

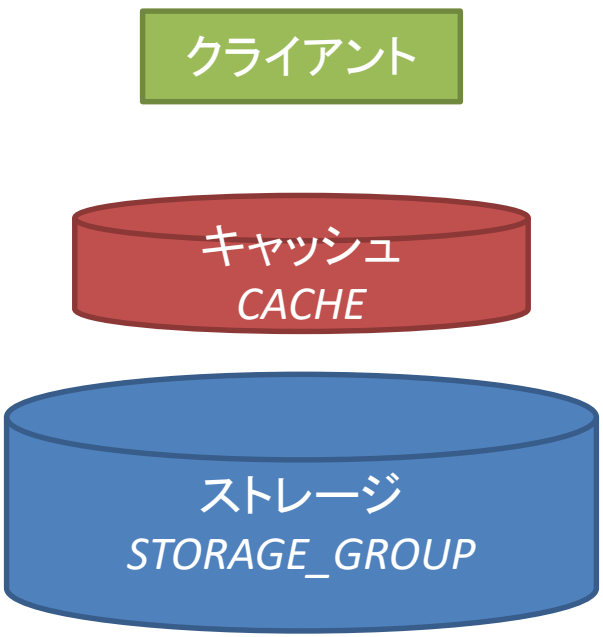
- 大量ファイル削除、大量ファイル複製作成、ファイル更新、replicainfo有効時の複製チェック動作時など
- たまにレスポンス時間が数秒～10数秒となる

- gfmdのファイル削除処理の効率化、mutexロックの公平化により改善

書込キャッシュストレージ支援

- 一時的な書込キャッシュとしての利用を支援
- STORAGE_GROUPに複製を保持

  % gfncopy -S *STORAGE_GROUP*:2 /
- キャッシュに書込むように
  - STORAGE_GROUPの仮想負荷を上げる or
  - write_target_domainでCACHEを指定
- キャッシュに保持する期間、キャッシュ容量の設定
  - replica_check_remove_grace_time, replica_check_remove_grace_used_space_ratio

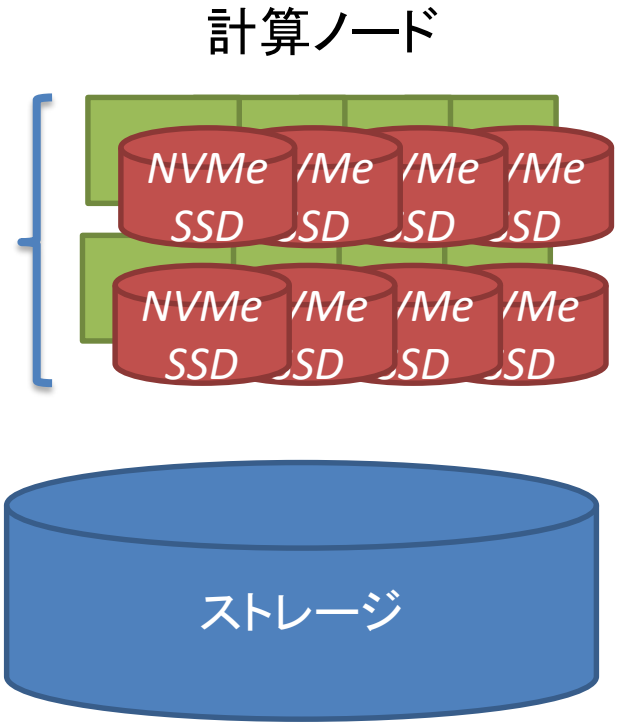


データ移行支援

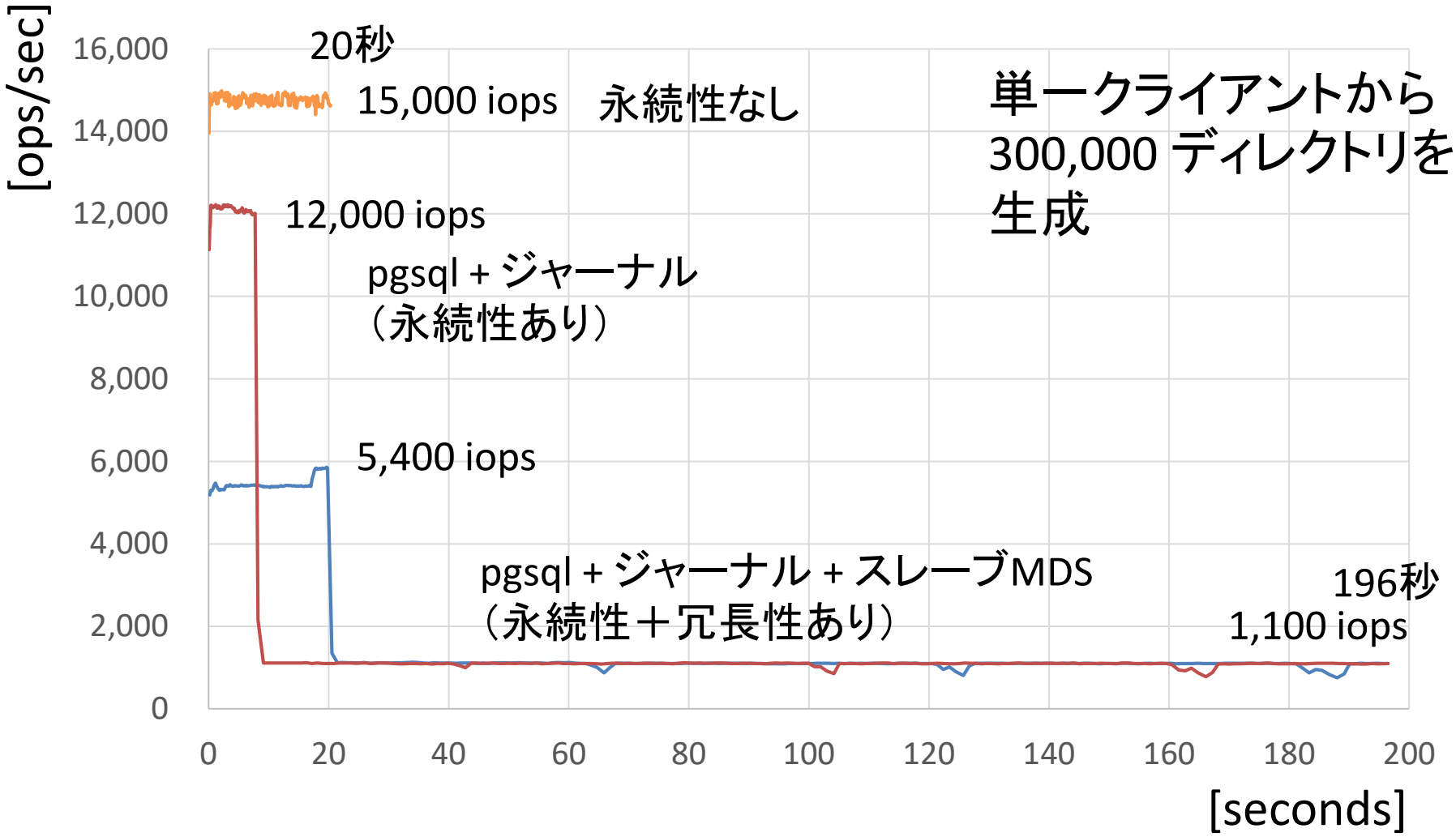
- 書込禁止ノードの設定（移動対象ノード）
  - 読込のみ可とする（readonly設定）
  - 仮想負荷を上げる（spool_base_load設定）
- 書込禁止ノードへのファイル複製作成
  - replication_busy_host enable/disable
  - gfpcopyの-Bオプション
- replicainfo機能の有効・無効化
  - 無効の場合はncopyに従う
- ノードへの最大並列複製作成数の実行時変更

Gfarm/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



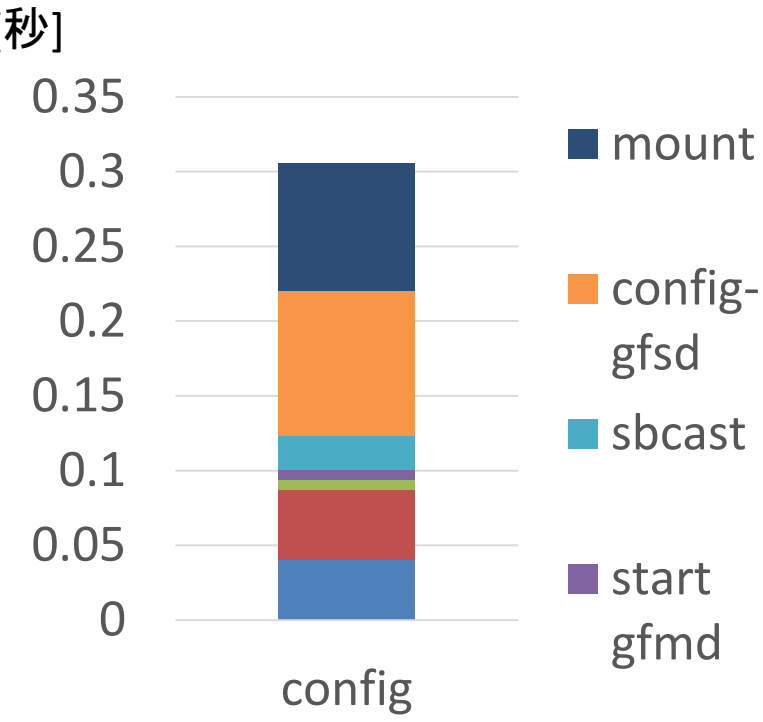
ディレクトリ生成のIOPS（メタデータ性能向上）



Gfarmファイルシステム構築・撤去

各計算ノードのTDIR以下にファイルシステムを構築

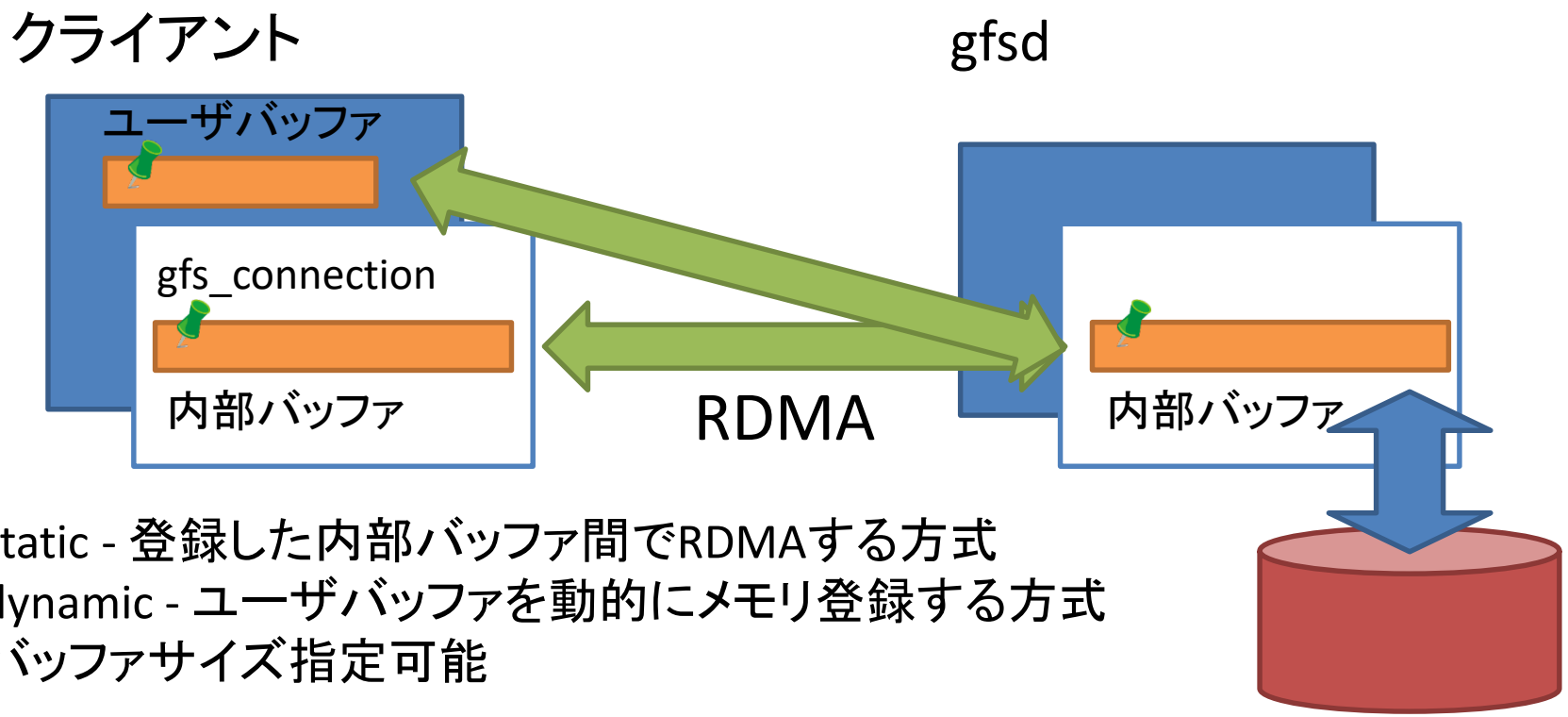
```
GFARM_CONFIG_FILE=$TDIR/etc/gfarm2.conf
srun mkdir -p $TDIR/etc
config-gfarm --prefix $TDIR -S -A $USER -b none
sbcast -p $GFARM_CONFIG_FILE $GFARM_CONFIG_FILE
sbcast -p $TDIR/etc/gfsd.conf $TDIR/etc/gfsd.conf
sbcast -p $TDIR/etc/usermap $TDIR/etc/usermap
srun $BINDIR/config-gfsd --prefix $TDIR -S
srun $BINDIR/mount.gfarm2fs $GFARM_CONFIG_FILE $MDIR
```



2ノードで構成、マウントに0.31秒、撤去に0.19秒

InfiniBand RDMAによるアクセス

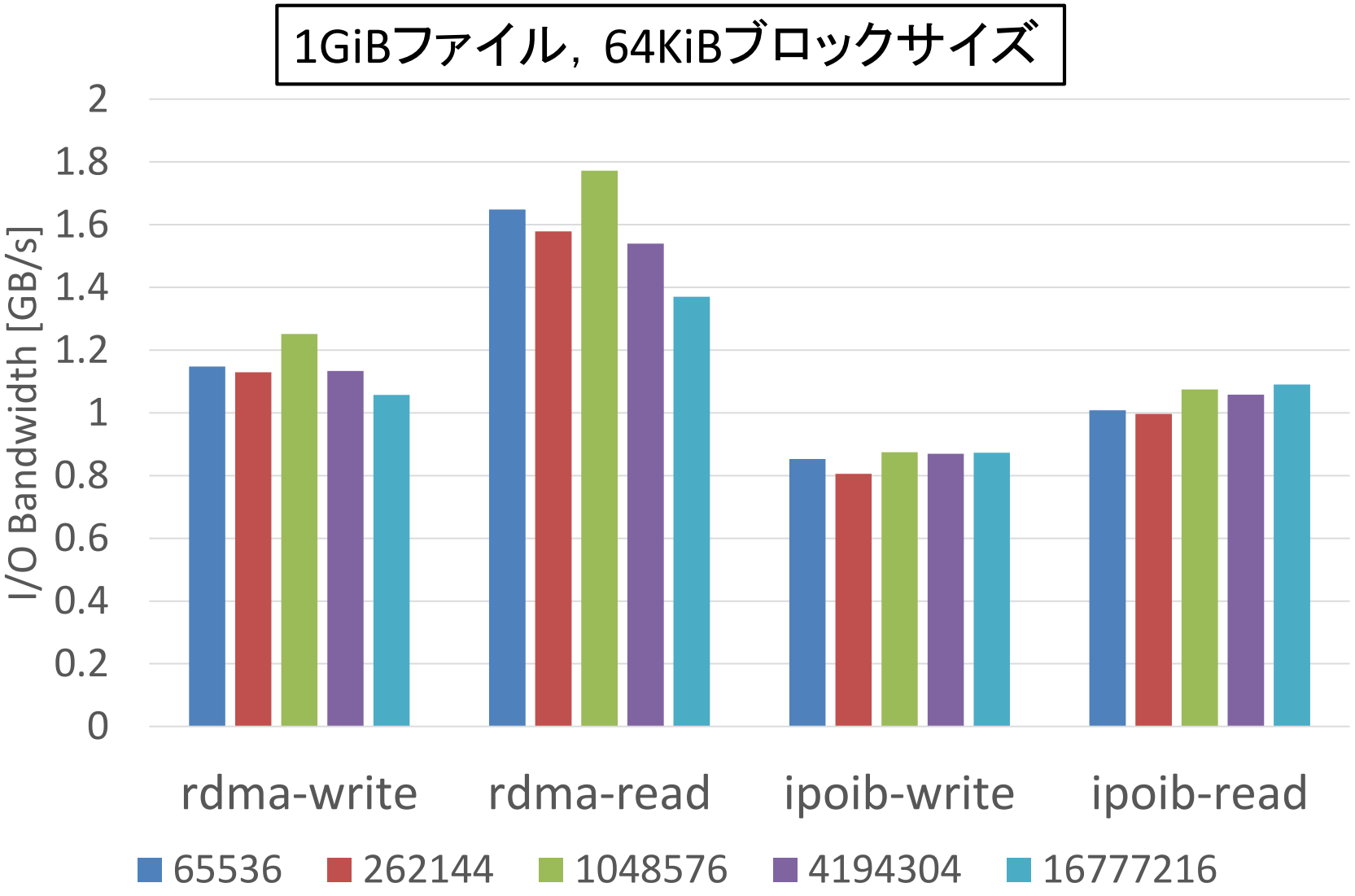
- ファイルアクセスをRDMAで行う
  - 読込はRemote Write, 書込みはRemote Read



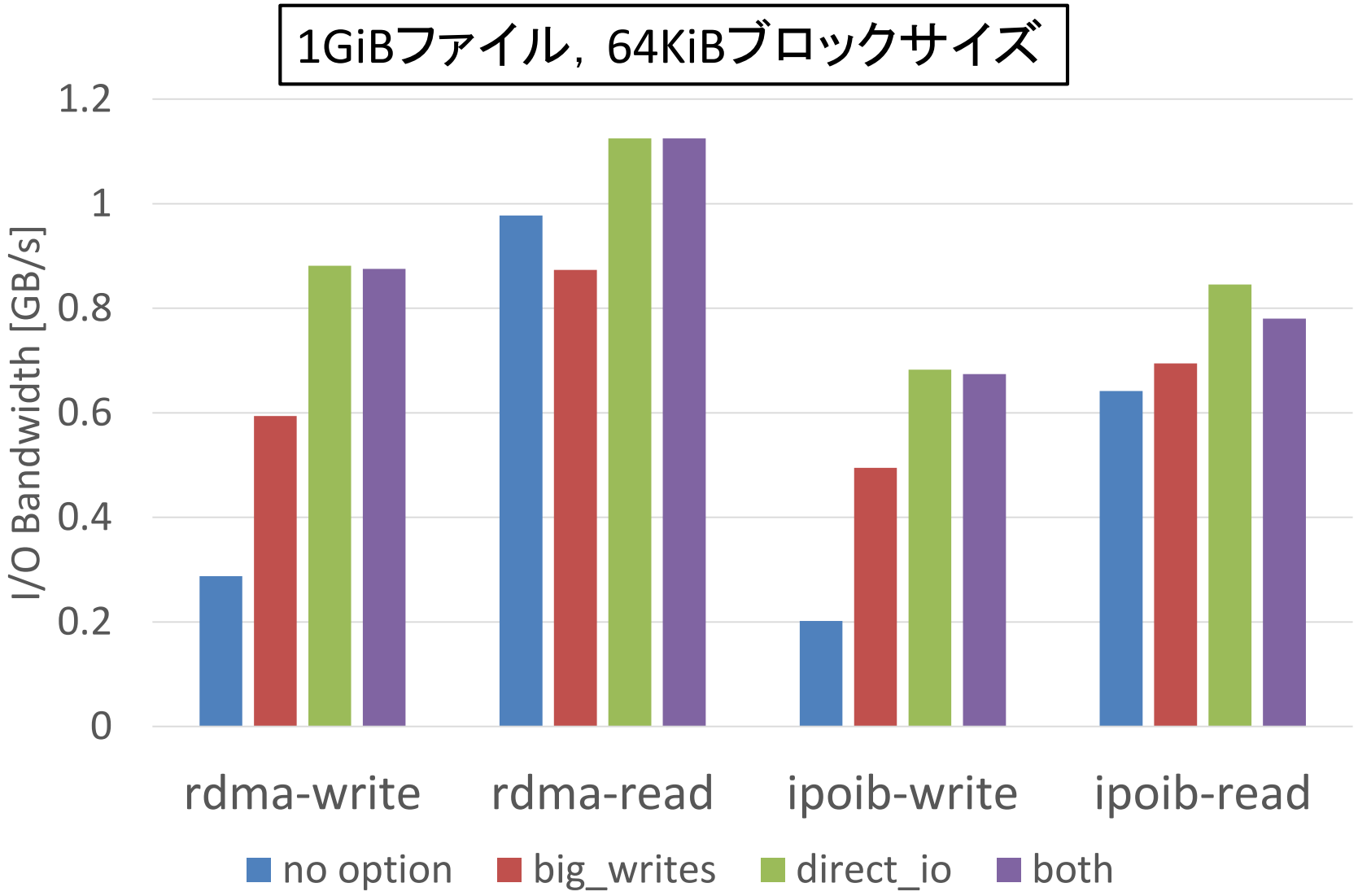
static - 登録した内部バッファ間でRDMAする方式
dynamic - ユーザバッファを動的にメモリ登録する方式
バッファサイズ指定可能

- RDMA通信に失敗したとき、IP通信へフォールバック

Gfarm APIによるファイルアクセス性能



POSIX APIによるファイルアクセス性能



ディレクトリクオータ機能（１）

- 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（1）

- 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

データ完全性（2）

- データ完全性の保証
  digest *digest_type*　　// gfmd.conf
  md5, sha1, sha256, …をサポート
- End-to-endデータ完全性の保証
  client_digest_check enable　　// gfarm2.conf

% **gfcksum** *file*
  – ファイルのdigestを表示
% **gfcksum** *-c [-h host] file*
  – （hostに格納されている）ファイルのdigestを計算し確認

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 2.7.10を2018/2/22にリリース
- gfmdのレスポンス改善
- 書込キャッシュ支援、データ移行支援
- 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
- IPv6対応進行中
- クラウドストレージ連携も進行中