

Gfarm Workshop 2019
2019年2月1日@鹿児島

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ワークショップ2019

- ライブ配信
 - つくばOSS YouTube で検索
- プログラム
 - HPCI共用ストレージ
 - Lustre
 - Excelero
 - IBM Storage Software
 - 産総研ABCIのストレージシステム

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/>
- 21,200 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

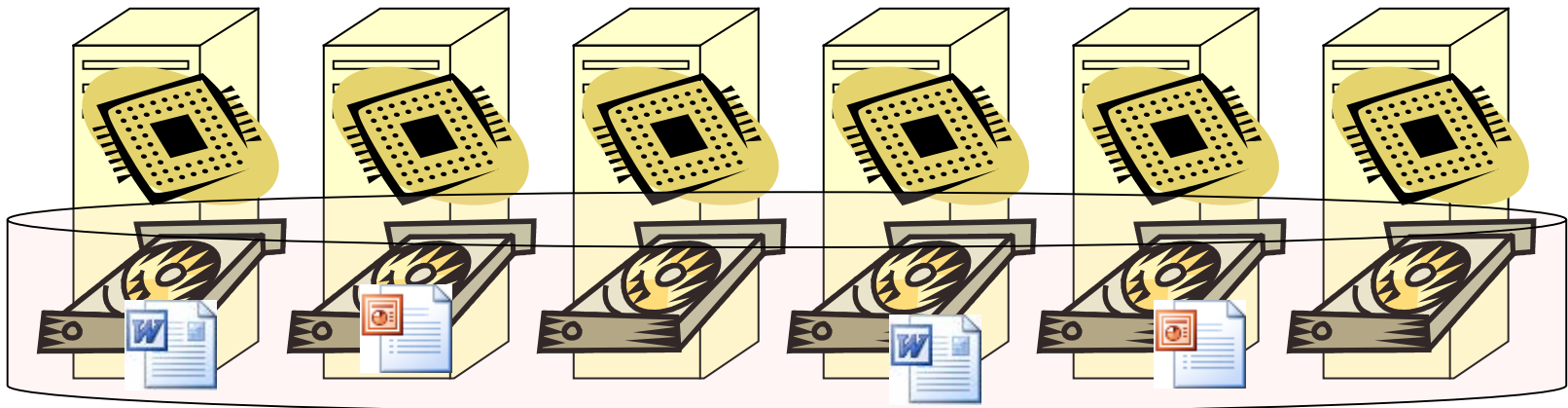


Gfarmファイルシステム(2)

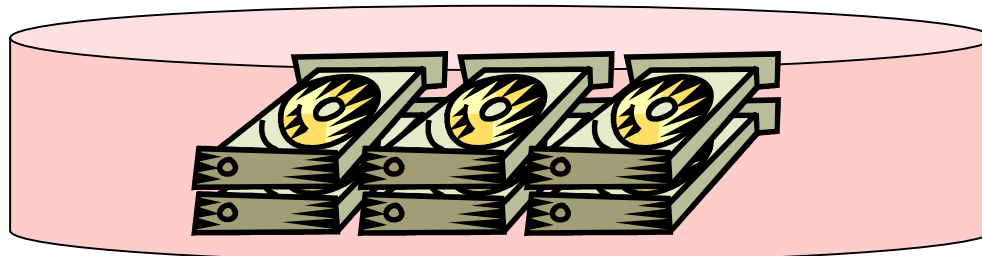
- JLDG(10.7PB、8拠点)、HPCI共用ストレージ(～100PB、2拠点)、NICTサイエンスクラウド、(株)クオリティアActive! world等で実運用
- 計算ノードのローカルディスクによるデータ解析
 - すばる望遠鏡データ解析、メタゲノム解析
- Pwrakeワークフローシステム、MapReduce、MPI-IO、バッチキューイングシステム
 - データ局所性を高めるプロセススケジューリング
 - ディスクキャッシュを有効利用するプロセススケジューリング
 - データ局所性を高めるファイル複製作成

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



Gfarmファイルシステム



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

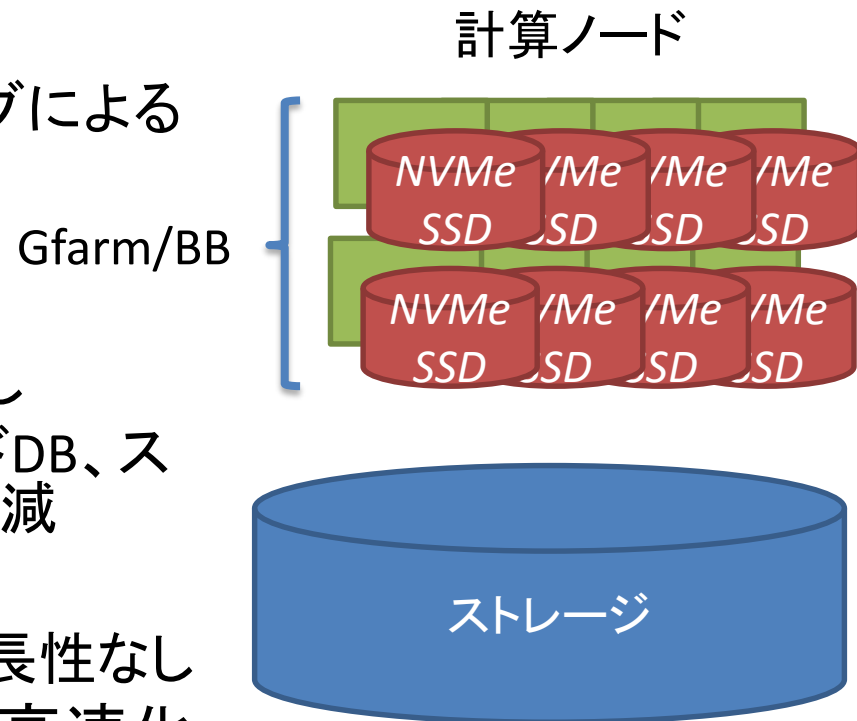
日付	version	新機能、更新機能
soon	2.7.12	• Replicainfo高速化 • gfdf -aの修正 (Gfarm 2.7.6以降)
2018/2/22	2.7.10	• gfmdのレスポンス改善
2017/10/1	2.7.6	• 書込キャッシュストレージ支援 • データ移行支援
2017/2/4	2.7.2	• Gfarm/バーストバッファ
2016/12/8	2.7.0	• InfiniBand RDMAサポート • ディレクトリクォータ
2016/1/16	2.6.8	• 書込後ベリファイ

Replicainfo高速化

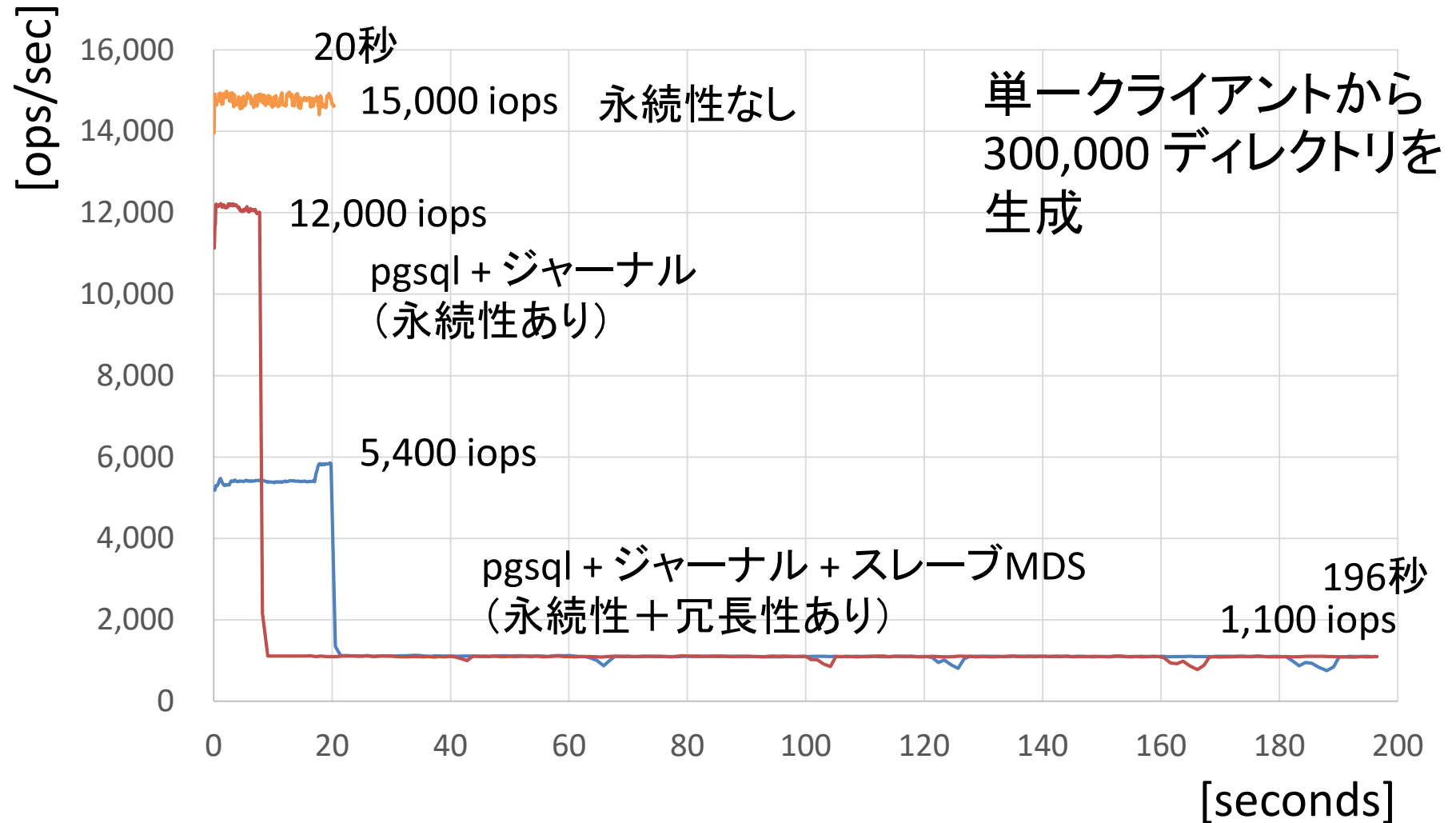
- Replicainfo
 - 複製配置機能
 - A:1,B:1
 - ファイルシステムグループAに一つ、Bに一つ
- パトロール(replica check)の高速化
 - 正しい複製配置がなされているか全ファイルチェック
 - ビットマップの集合演算で書き換え
 - 8Mファイルで3時間が5分半(33倍高速化)

Gfarm/ノーストバッファ

- 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gfmmd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)

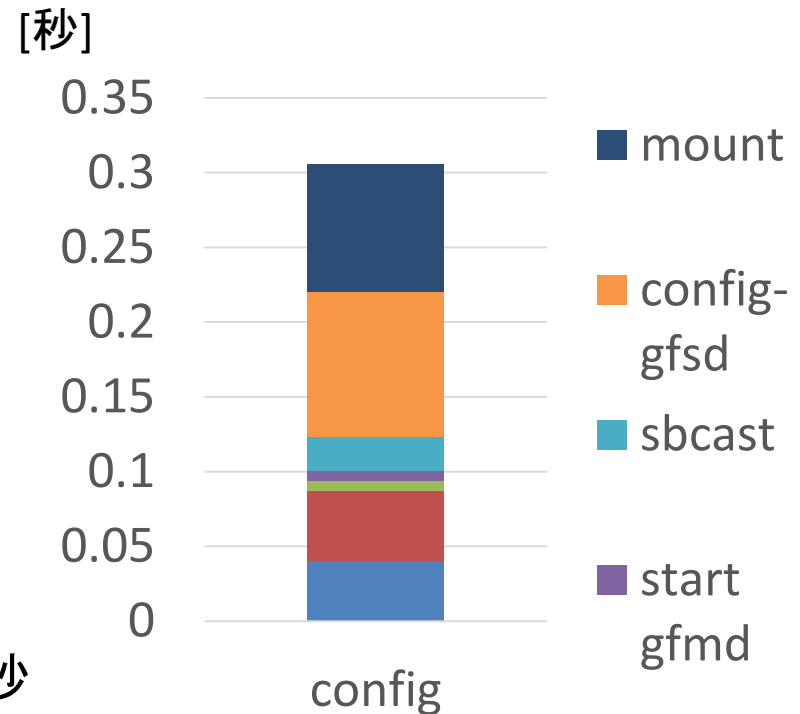


Gfarmファイルシステム構築・撤去

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
srunk 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
srunk $BINDIR/config-gfsd --prefix $TDIR -S
srunk $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

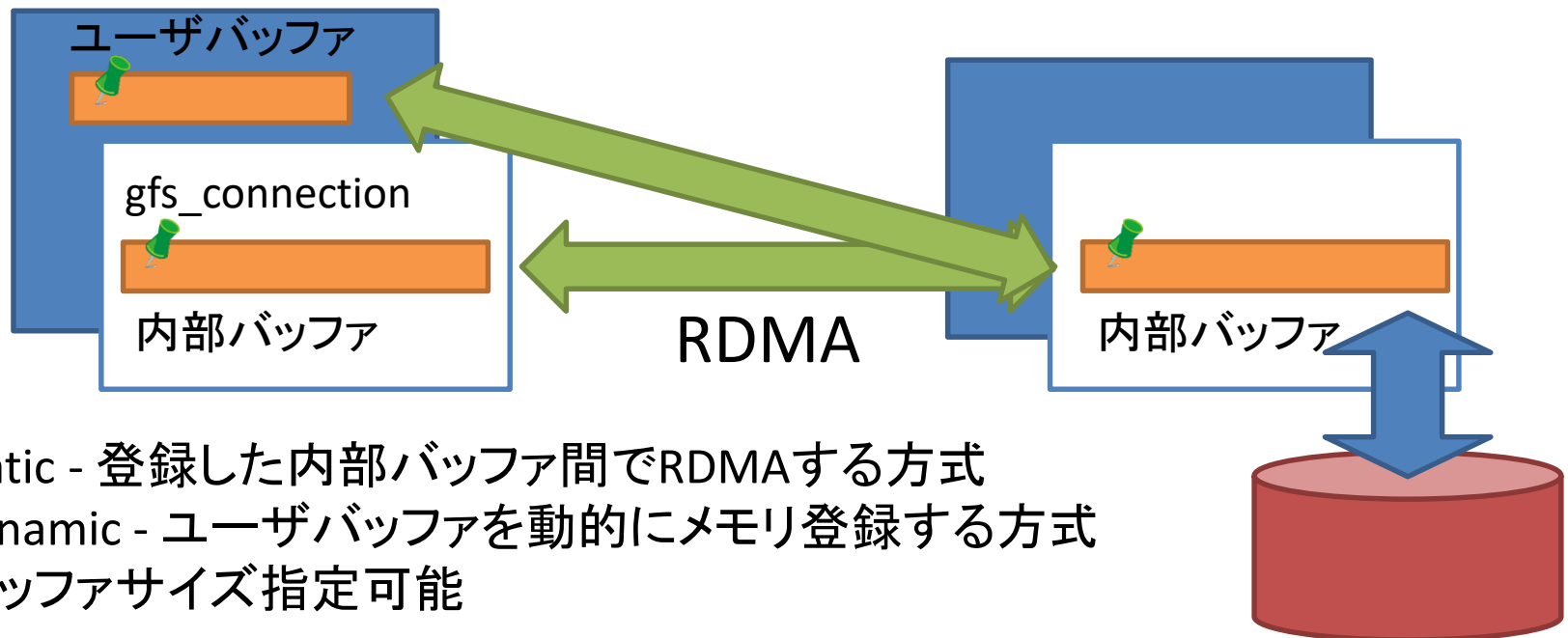


InfiniBand RDMAによるアクセス

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

クライアント

gfsd



static - 登録した内部バッファ間でRDMAする方式

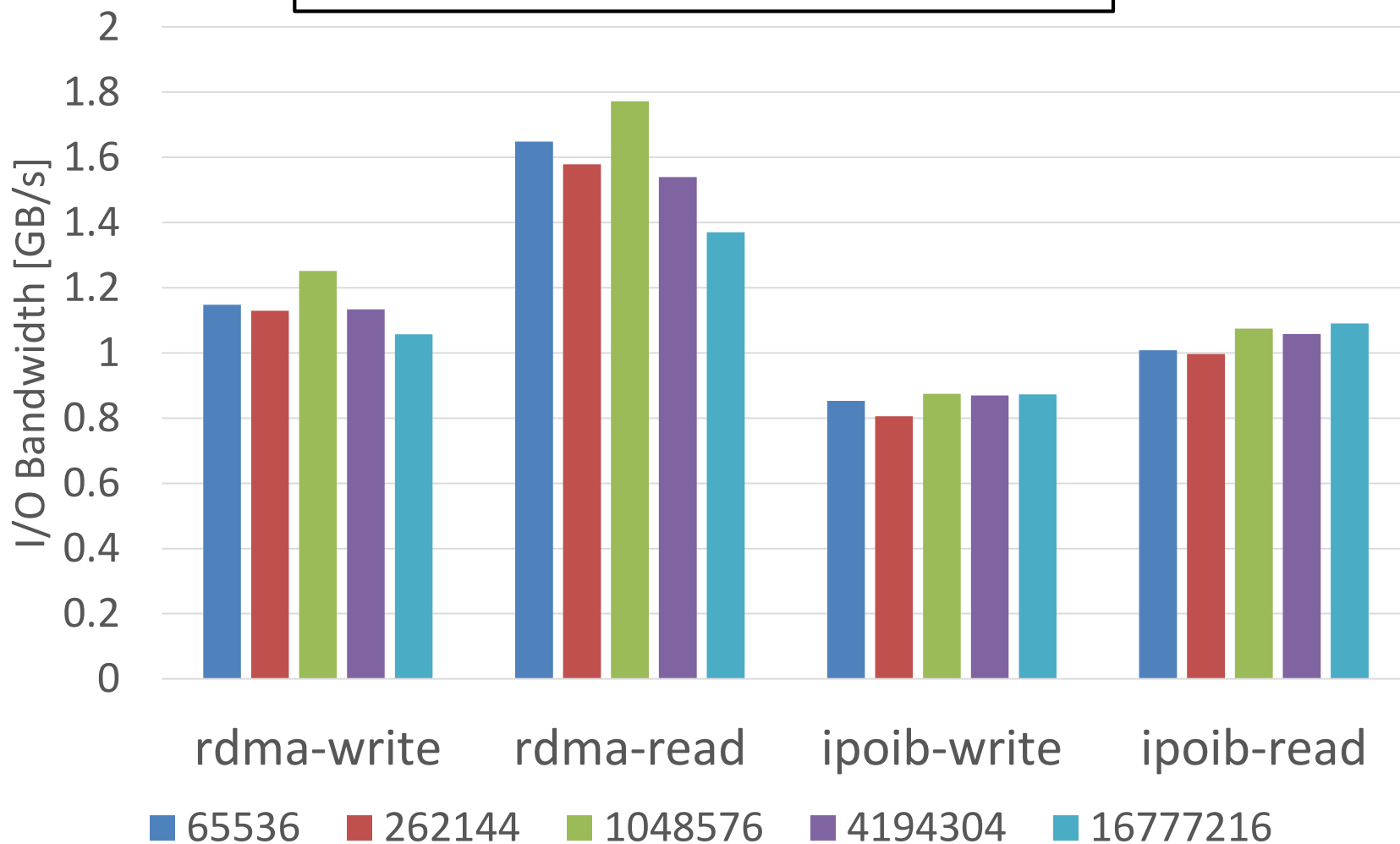
dynamic - ユーザバッファを動的にメモリ登録する方式

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

1GiBファイル, 64KiBブロックサイズ



ディレクトリクォータ機能

- 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ォータと併用可、またユーザ権限で設定可能
- `gfdircquota` – ディレクトリセット作成、設定
 - 複数ディレクトリでクォータ制限可能
- `gfredquota`, `gfquota` – クォータ設定、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 2.7.12をまもなくリリース
- Replicainfo高速化
- InfiniBand RDMA、ディレクトリクォータ機能
- 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)
 - クラウドストレージ連携