

Gfarm Workshop 2021
2021年3月5日@オンライン

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



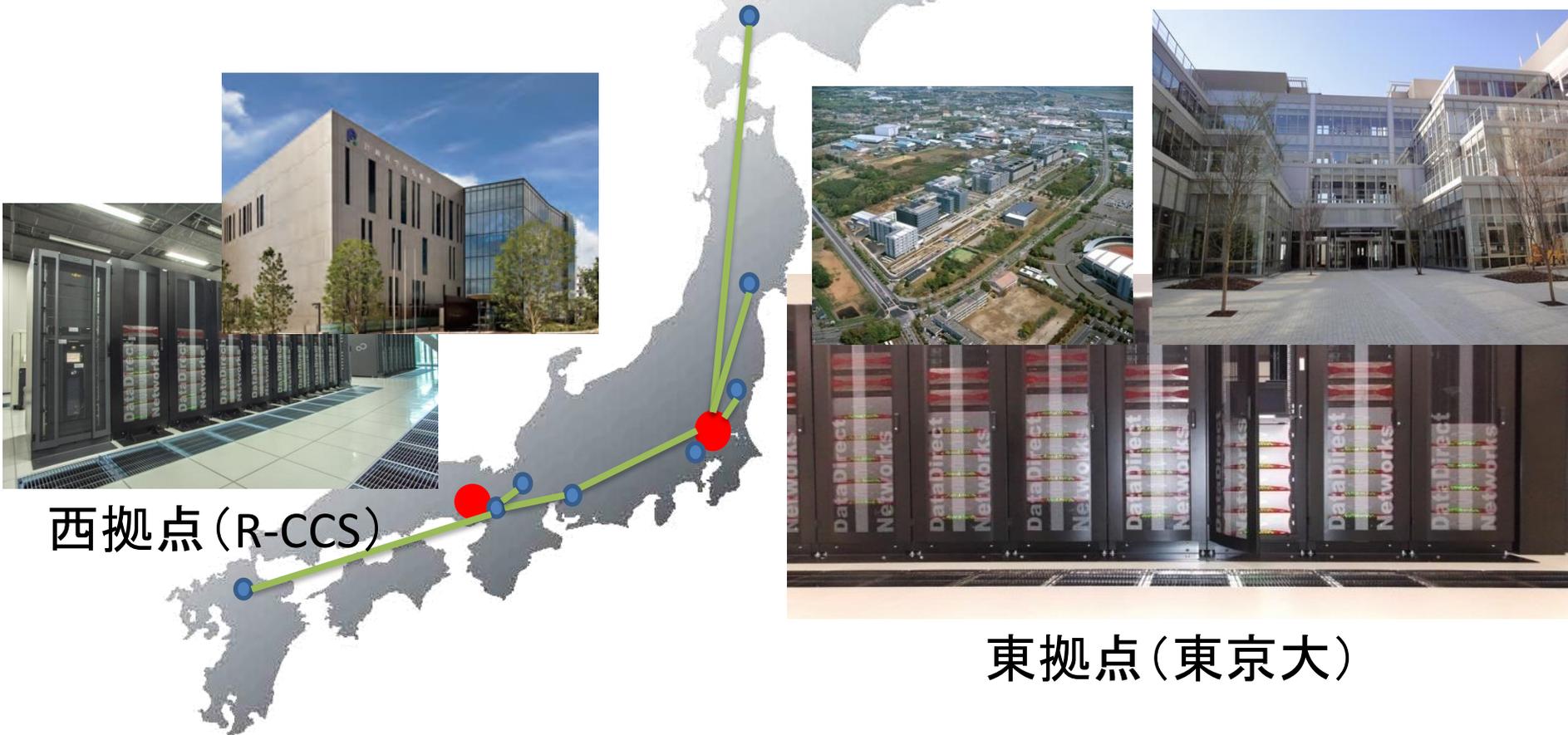
- オープンソース広域分散ファイルシステム
 - <http://oss-tsukuba.org/software/gfarm/>
- サポート
 - NPO法人つくばOSS技術支援センター(日本他)
 - Libre Solutions Pty Ltd(オーストラリア)
- 特徴
 - 性能・容量がスケールアウト
 - データアクセス局所性、ファイル複製
 - 無停止で拡張、縮小可能
 - 単一障害点なし
 - 複製数維持機能、ホットスタンバイMDSサーバ
 - ローリングアップデート
 - データ完全性を保証しサイレントデータ損傷も対応可

ossTsukuba
oss-tsukuba.org



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
2021/?/?		• S3互換IF、TLS通信
2020/9/17	2.7.17	• MTセーフ、ROFS機能、FO強化
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

Gfarm-S3-MinIO: GfarmのS3互換インタフェース

The image displays three sequential screenshots of the Gfarm-S3 management interface:

- Management Console Login:** The first screenshot shows a login form with fields for 'ユーザ名' (Username) and 'パスワード' (Password), and a 'ログイン' (Login) button.
- S3 Server Start:** The second screenshot shows the server status as '起動中' (Running). It includes a '停止' (Stop) button, access key information (K4XcKzocrUhrnCAKrx2Z), a '表示' (Show) button for the secret key, and a '変更' (Change) button. The authentication method is 'GSI (grid-proxy-init)' and the expiration time is 'Tue Feb 16 09:50:33 2021'. A '延長' (Extend) button is also present.
- Sharing Settings:** The third screenshot shows the '共有設定' (Sharing Settings) page. It lists two buckets: 'test1' and 'test2'. The 'test1' bucket is selected, showing its creation time as 'Mon Feb 15 23:22:08 2021'. A table of permissions is shown below:

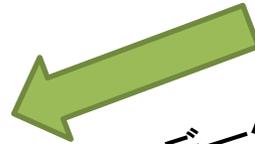
	user/group	read	write
×	MY GROUP	<input type="checkbox"/>	<input type="checkbox"/>
×	OTHER	<input type="checkbox"/>	<input type="checkbox"/>
×	user:user2(user2)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
×	[Dropdown]	<input type="checkbox"/>	<input type="checkbox"/>

Below the table, there is an '工:' (Work) dropdown menu with options: 'group:gfarmadm', 'group:gfarmroot', 'user:Gfarm administrator(gfarmad', and 'user:Gfarm'. An '更新を適用' (Apply Update) button is also visible.

WebUI



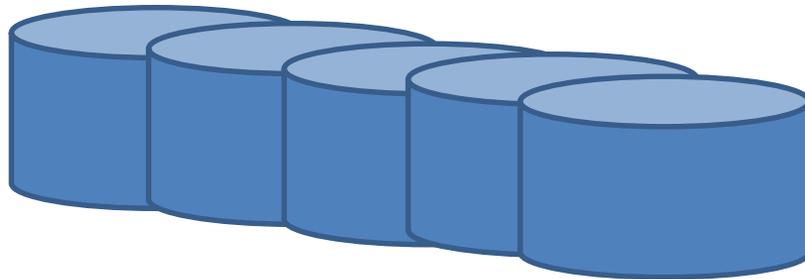
サーバ制御



データ格納



S3クライアント



Gfarmファイルシステム

暗号化ファイルシステム

- <https://github.com/oss-tsukuba/gfarm/blob/master/doc/encfs.ja.md>

Gfarm暗号化ファイルシステム

EncFS(*)を用いることにより、Gfarmファイルシステムに暗号化されたデータを格納することができます。

(*) <https://github.com/vgough/encfs/blob/master/encfs/encfs.pod>

インストール

EncFSをインストールします

```
# yum install encfs
```

使い方

1. Gfarmファイルシステムをマウントする

```
$ gfarm2fs /tmp/gfarm
```

この例では、/tmp/gfarmにGfarmファイルシステムをマウントしています。

2. 暗号化ファイルシステムを作成しマウントする

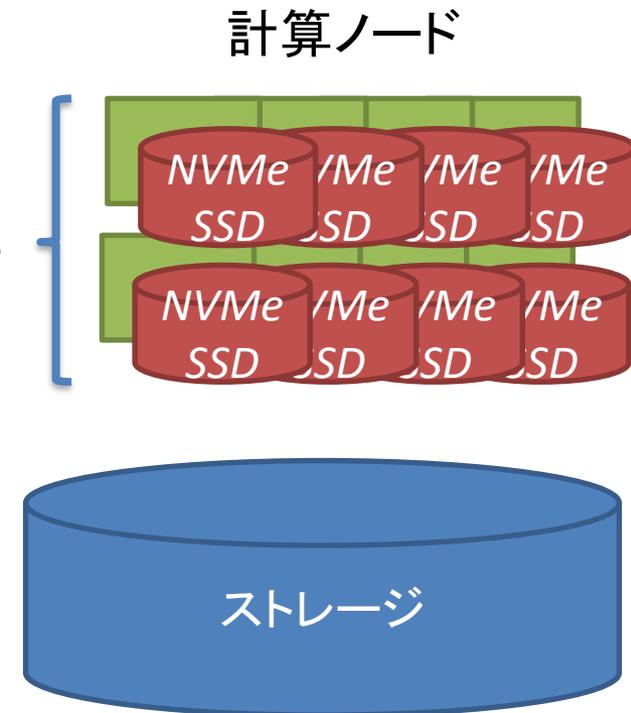
ROFS – Gfarm Read only FS

- ファイルシステムを完全にread onlyに
% `gfstatus -Mm 'read_only enable'`
- Zabbixフェイルオーバースクリプトでは、split brainの可能性が残る場合にROでフェイルオーバー
 - 確認が取れ次第read onlyを解除

Gfarm/BBノバーストバッファ [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

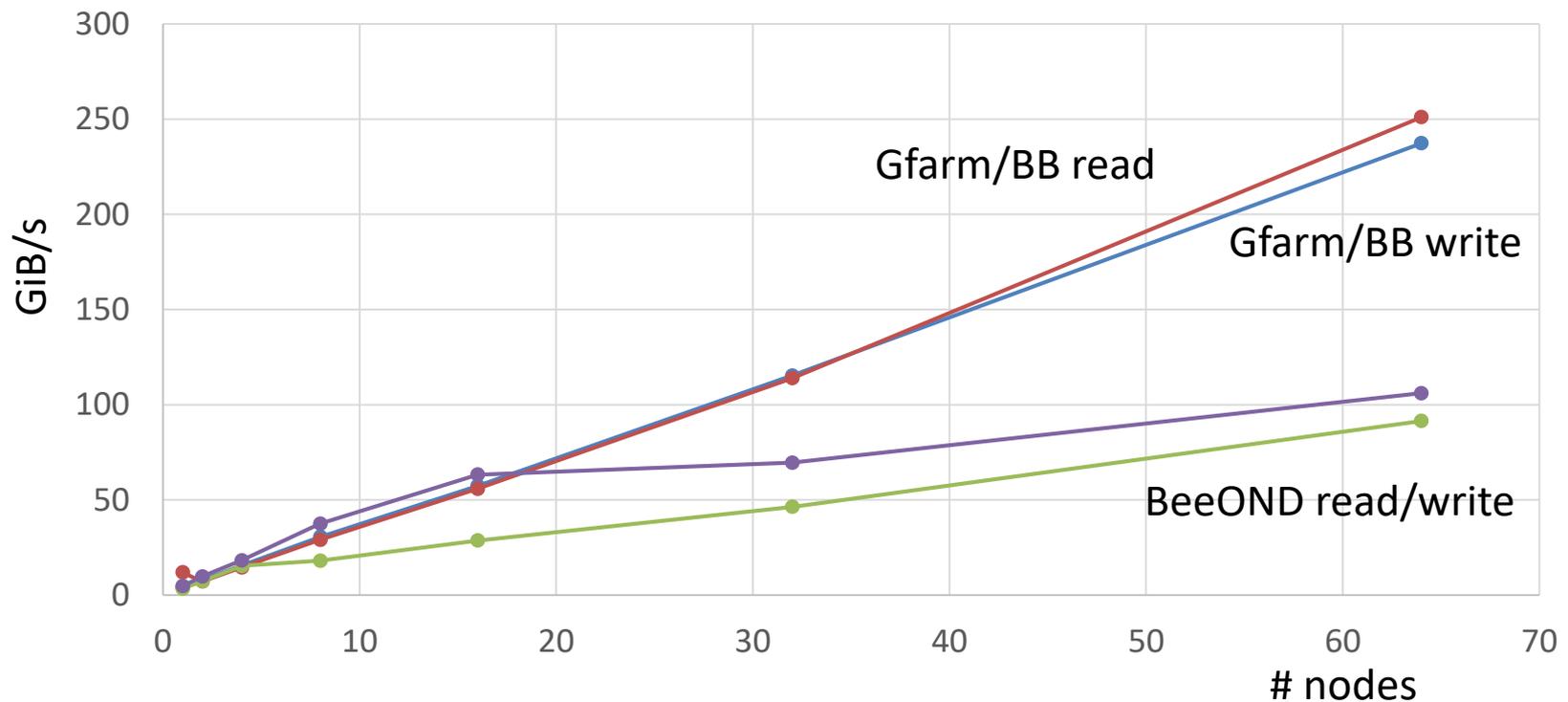
Gfarm/BB



Gfarm/BBノバーストバッファ(2)

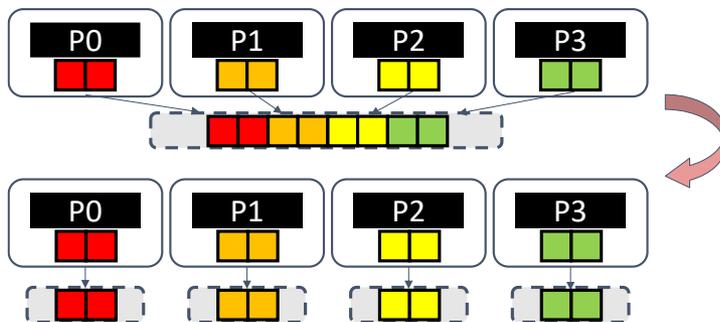
```
gfarmbb -h hostfile -m mount_point start  
...  
gfarmbb -h hostfile stop
```

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

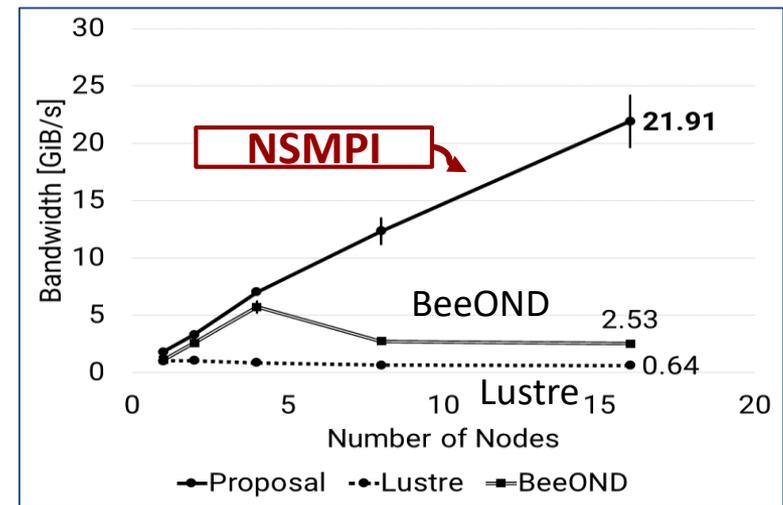


NSMPI - MPI-IO for Node-local Storage [杉原 2020]

- 計算ノードローカル NVMe SSD、PMを用いた MPI-IOの設計
- 共有ファイルをSparse Segmentsで表現することでN-Nアクセスに変換

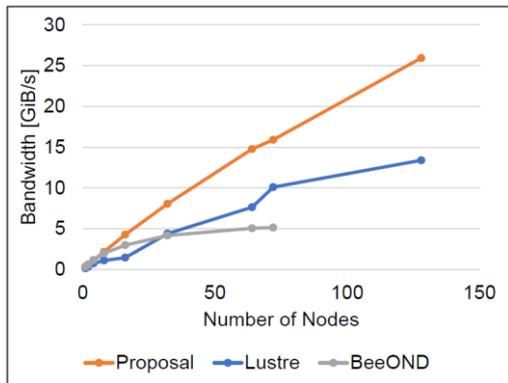


- HPC IOR Benchmark
 - N-1アクセスでスケール
ブルな性能を実現

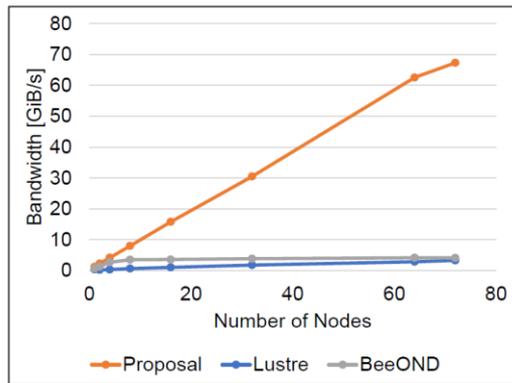


Write

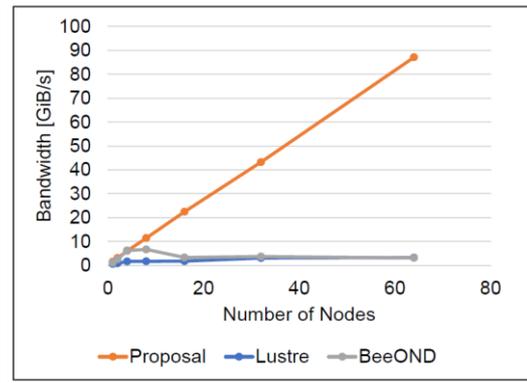
NSMPI (2) アプリケーション ベンチマーク



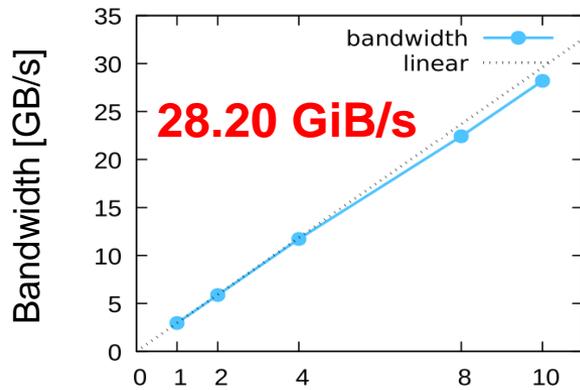
S3D-IO



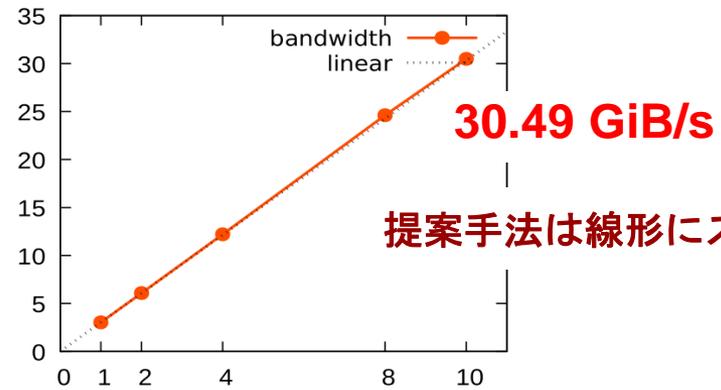
LES-IO



VPIC-IO



IOR SSF (a) Write ノード数



IOR SSF (b) Read ノード数

提案手法は線形にスケール

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 2.7.17を9/17にリリース
 - <https://github.com/oss-tsukuba/>
- Gfarm/BBバーストバッファ
- HPCI共用ストレージ、JLDGなど実運用実績
- まもなくリリース
 - IPv6対応 (Gfarm 2.8)
 - S3互換IF
 - TLS対応