Gfarm Symposium 2020
2020年10月9日＠秋葉原

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

GRID DATA FARM

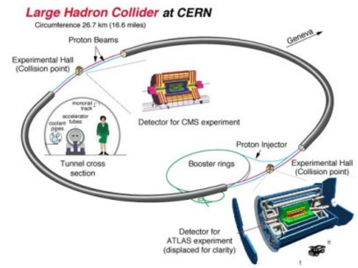
- オープンソース広域分散ファイルシステム
  - http://oss-tsukuba.org/software/gfarm/
- サポート
  - NPO法人つくばOSS技術支援センター（日本他）
  - Libre Solutions Pty Ltd（オーストラリア）
- 特徴
  - 性能・容量がスケールアウト
    - データアクセス局所性、ファイル複製
    - 無停止で拡張、更新可能
  - 単一障害点なし
    - 複製数維持機能、ホットスタンバイMDSサーバ
  - データ完全性を保証しサイレントデータ損傷も対応可

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
-
-
-
- 2.0.0公開！
- m 2.7.17公開！！！

日刊工業新聞　2002年（平成14年）11月22日（金曜日）

日米間、毎秒707メガバイト達成

グリッドコンピューティングシステム　データ処理、パソコン利用

産総研

産業技術総合研究所は21日、パソコンクラスターを利用した日米間を結んだグリッドコンピューティングシステムで、これまでの記録を大幅に上回る毎秒707メガビット（707Mbps）の高速で、超大規模データを処理することに成功したと発表した。

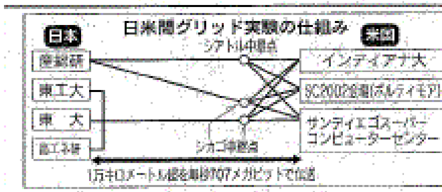
目標の800Mbpsには達しなかったが、1万キロメートル離れた日米間で、両国の7拠点に分散しているテラバイト級のデー

コンピューターに関する国際会議「SC2002」会場に産総研が開発したグリッドコンピューティング用ソフト「データファーム」と機材を持ち込んで行った。産総研のほか高エネルギー加速器研究機構、東京工業大学など日米6機関が参加し、米サンディエゴ・スーパーコンピュータ・センターなど7拠点を結んだ。

で実施した。

ネットで結ばれたパソコンは7システム1900台。つくば広域情報通信網（WAN）、アジア太平洋高速ネットワーク（APAN）など日米高速回線を利用、パソコンだけで18テラバイト（1テラは1兆文字）の大容量データを高速処理する環境構築に世界で初めて成功した。

2002年（平成14年）11月22日（金曜日）　日経産業新聞



複数結ぶ「グリッドコンピューティング」

日米間通信で最速記録

毎秒707メガビット達成　産総研など両国チーム

産業技術総合研究所（茨城県つくば市）など日米の共同研究チームは二十一日、複数のコンピューターが協調して高速処理する「グリッドコンピューティング」で過去最速の日米間通信実験に成功したと発表した。多数の並列パソコンを利用し、スーパーコンピューターを超える演算をこなすグリッド技術の実用化に近付く成果としている。

米ボルティモアで開催したスーパーコンピューティングの国際学会「SC2002」会場や米インディアナ大学などと、一万キロ以上離れた日本の産総研、東京大学、東京工業大学など計七カ所を結んだ。素粒子物理学の実験を想定した十八テラ（テラは一兆）バイトのデータをやり取りした。通信速度は毎秒七百七メガ（メガは百万）ビット。

天文学や衛星観測、生命情報学などの分野では、一回の実験データが数ペタ（ペタは千兆）バイトを超えるため、ネット経由では実験施設から研究所までデータを送りきれず、磁気テープなどに記録して運ぶ例が多い。高速化技術が実用化されれば、こうした手間が省け、研究の効率化につながる。

またデータを各地に分散して利用できることから、企業の重要情報を遠隔保存するデータセンターなどへの応用も見込める。産総研はグリッド実験に使ったソフトウエアを無償で公開し、民間などでの普及を図る。（つくば）

産経新聞　2002年（平成14年）12月1日（日曜日）

科学

産総研　超大規模データ転送に成功

グリッドコンピューティング　世界初の実証

独立行政法人「産業技術総合研究所」（茨城県つくば市）などは、日米七カ所を拠点として百九十台のパソコンを国際回線で結び、通常は大型コンピューターで行う一八テラバイト（CD－ROM約三万枚分相当）もの超大規模データを高速処理できるグリッドコンピューティング技術の実証実験に、世界で初めて成功した。

実験では、米サンディエゴ・スーパーコンピューターセンターや東京工業大学など日米の拠点に、それぞれ数十台のパソコンをつないだ「パソコンクラスター」を構築。七つのクラスターを産総研製のソフトで統合し、陽子衝突のシミュレーションデータ（一八テラバイト）を使ってデータ転送、解析を実証した。

約一万キロ離れた日米間で、CD－ROM一枚分のデータを五・七秒で送れる秒速七〇七メガビット（一[illegible]）の転送速度を達成。多数のパソコン間で並列転送することで、これまでに例のない高速転送が可能になったという。

その結果、CD－ROM一枚を〇・一秒で読み込める秒速六六〇〇メガビットの高速処理性能も実現。産総研では「さらに大規模・高速化も可能。今後は欧州へも拡大して実証実験を進めていく」としている。

（伊藤壽一郎）

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
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 • ディレクトリクオータ
2016/1/16	2.6.8	• 書込後ベリファイ

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

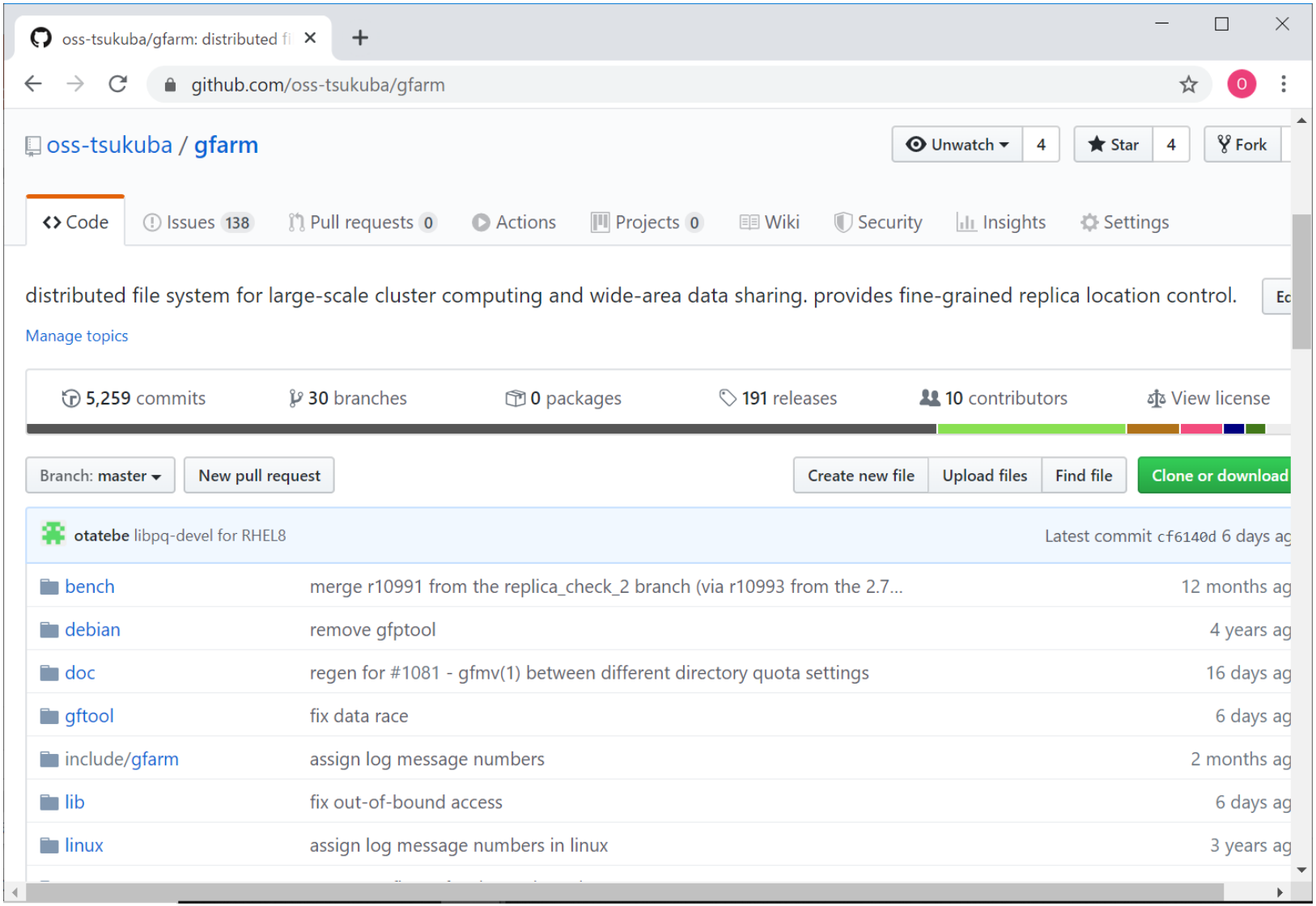
ROFS – Gfarm Read only FS

- ファイルシステムを完全にread onlyに

```
% gfstatus -Mm ‘read_only enable’
```

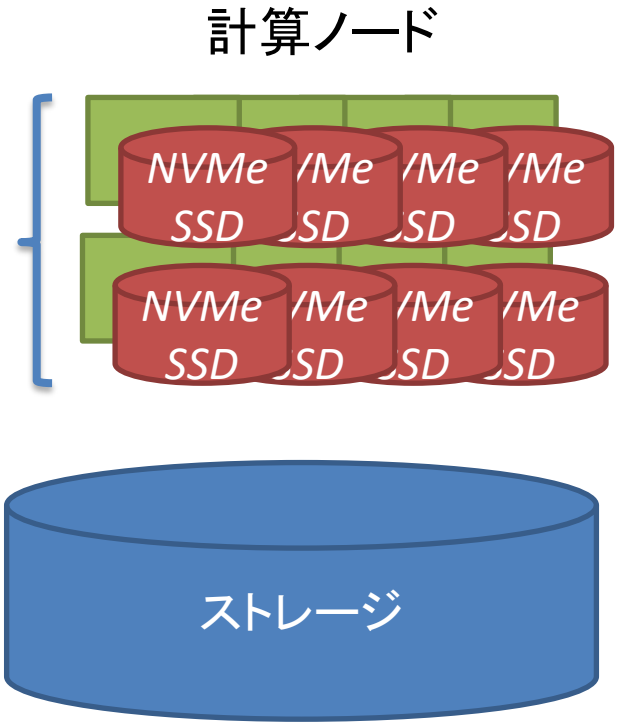
- Zabbixフェイルオーバスクリプトでは、split brainの可能性が残る場合にROでフェイルオーバ
  - 確認が取れ次第read onlyを解除

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



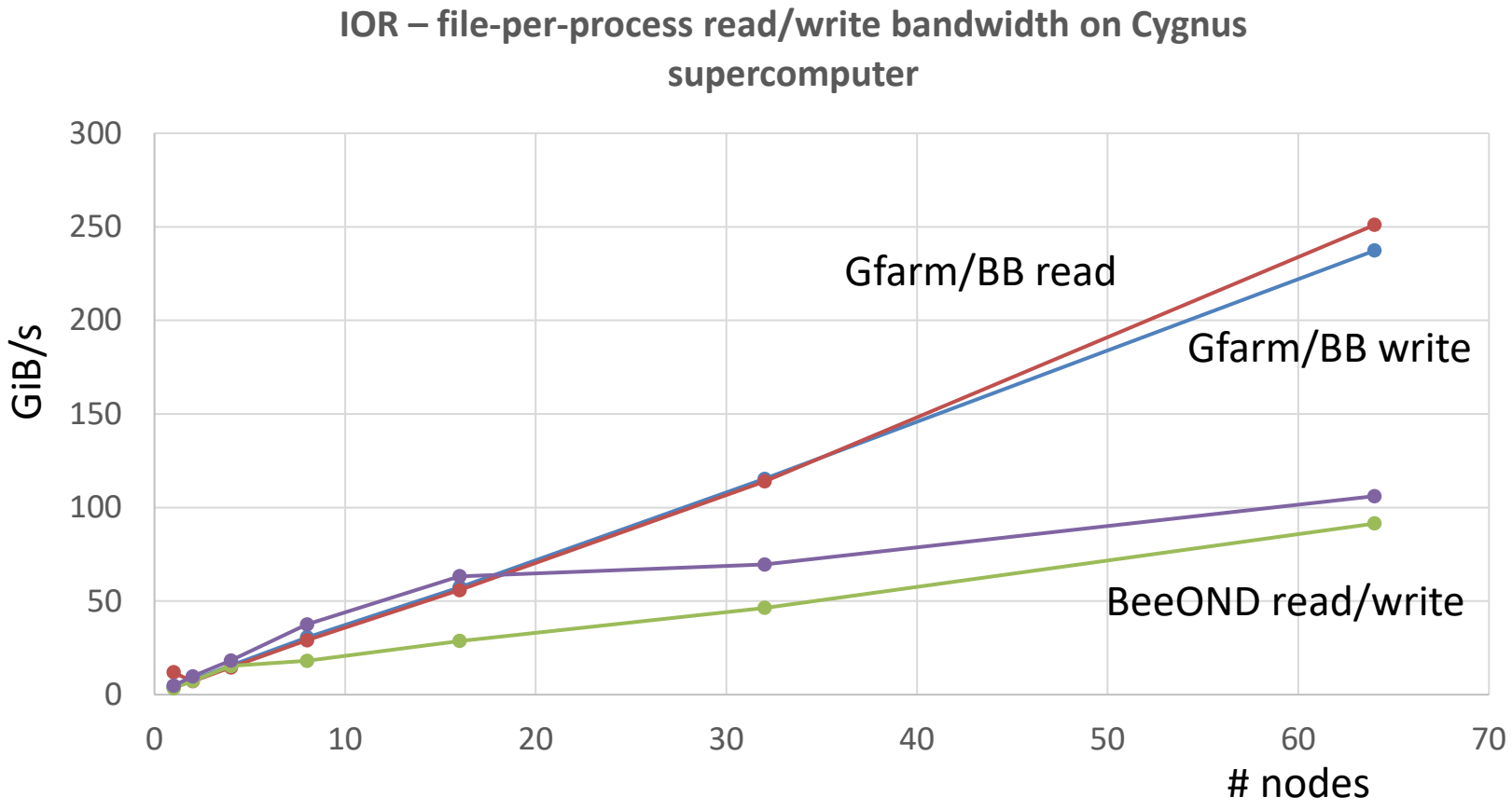
Gfarm/BBバーストバッファ [建部 2020]

- 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

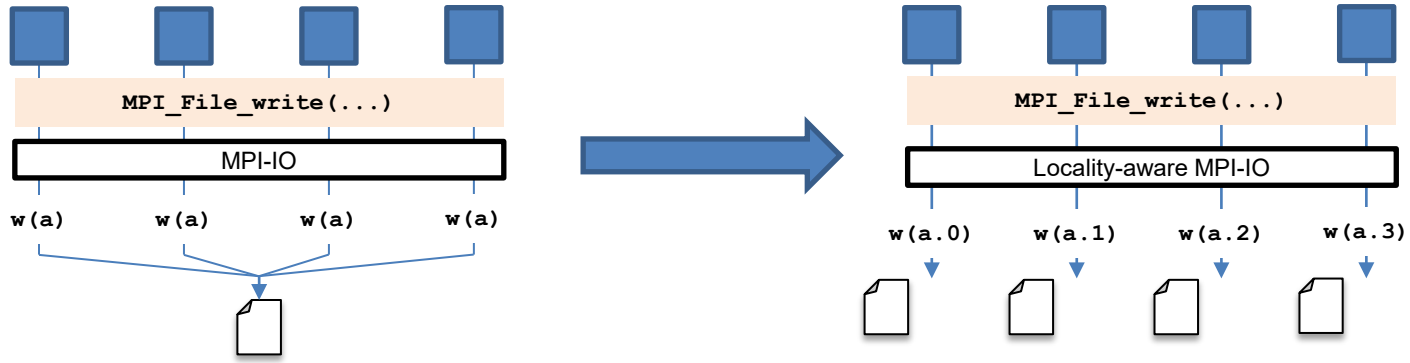


Gfarm/BBバーストバッファ（2）

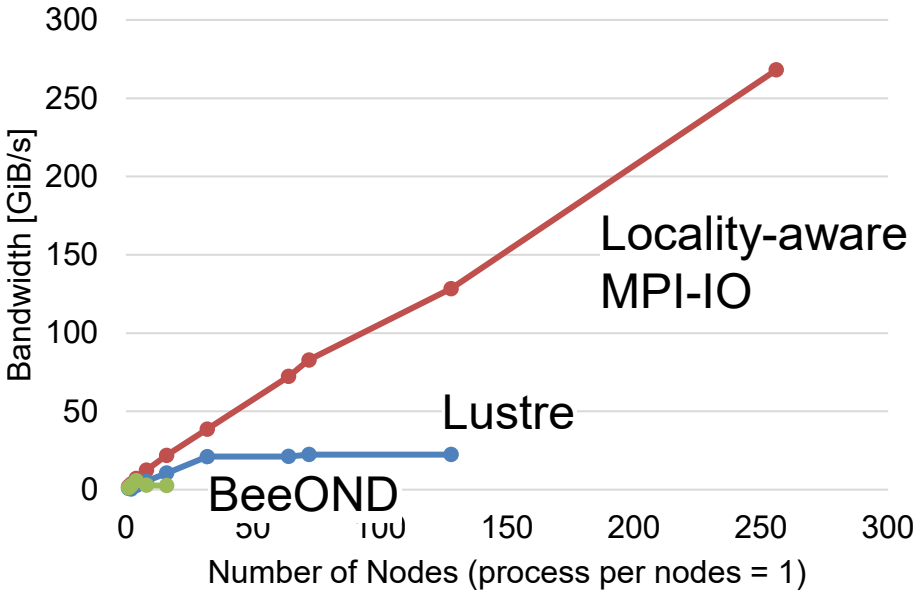
```
gfarmbb –h hostfile –m mount_point start
…
gfarmbb –h hostfile stop
```



Locality-aware MPI-IO [杉原 2020]



- 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 2.7.17を9/17にリリース
  - https://github.com/oss-tsukuba/
- 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）