Gfarm Workshop 2016
2016年10月21日神戸

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

- 2000年より研究開発
  - 国際会議SC03、SC05、SC06で受賞
- 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/
  - 18,000 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）
  - 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
  - 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
  - 無停止で拡張、更新可能
  - データ完全性を保証しサイレントデータ損傷も対応可

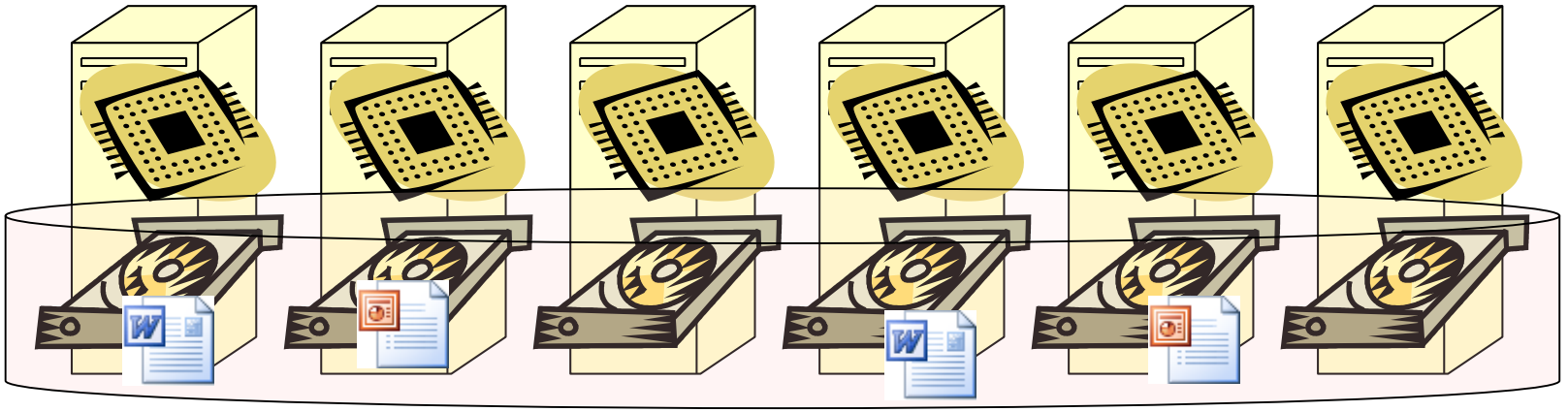
oss-tsukuba.org

Gfarmファイルシステム（2）

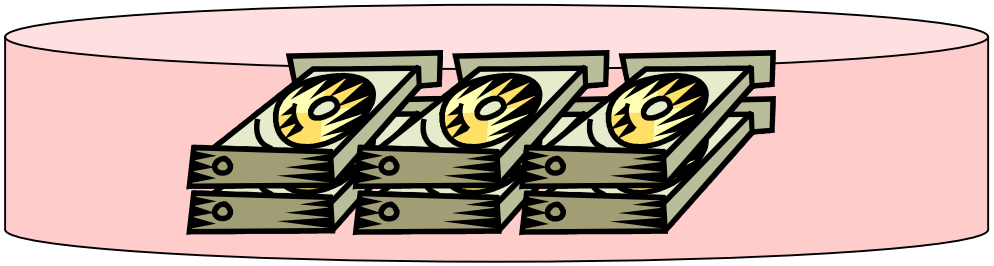
- JLDG（7PB、8拠点）、HPCI共用ストレージ（22.5PB、3拠点）、NICTサイエンスクラウド、（株）クオリティアActive! worldで実運用
- 計算ノードのローカルディスクを利用するデータ解析
- Pwrakeワークフローシステム、MapReduce、MPI-IO、バッチキューイングシステム
  - データ局所性awareなプロセススケジューリング
  - ディスクキャッシュawareなプロセススケジューリング
  - データ局所性を高めるファイル複製作成

Gfarmファイルシステムの構成

- 一般的なPCの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

- 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（22.5PB）
- スパコン間のデータ共有、共有データ格納

西拠点（AICS 9.5PB）

東拠点（東京大、東工大 13.0PB）

最新機能・状況紹介

gflsof - オープン中のファイル表示

```
% gflsof
USER   GPID FD TYPE INODE GEN SIZE/OFF CLIENT      GFSD
tatebe    6  1w REG   929  12        0 host1:38034 host1:38037
```

- ユーザ名、プロセスID、デスクリプタ番号、オープン種別（rwu）、ファイルタイプ、iノード番号、生成番号、サイズ、クライアント、GFSDを表示

ディレクトリクオータ機能（１）

- ディレクトリ単位でファイル数、利用サイズの制限をユーザ権限で行う
- gfdirquota – ディレクトリセット作成、設定

```
% gfdirquota –c projectX
% gfdirquota –a projectX /work/projectX
```

- gfedquota, gfquota – クオータ設定、表示

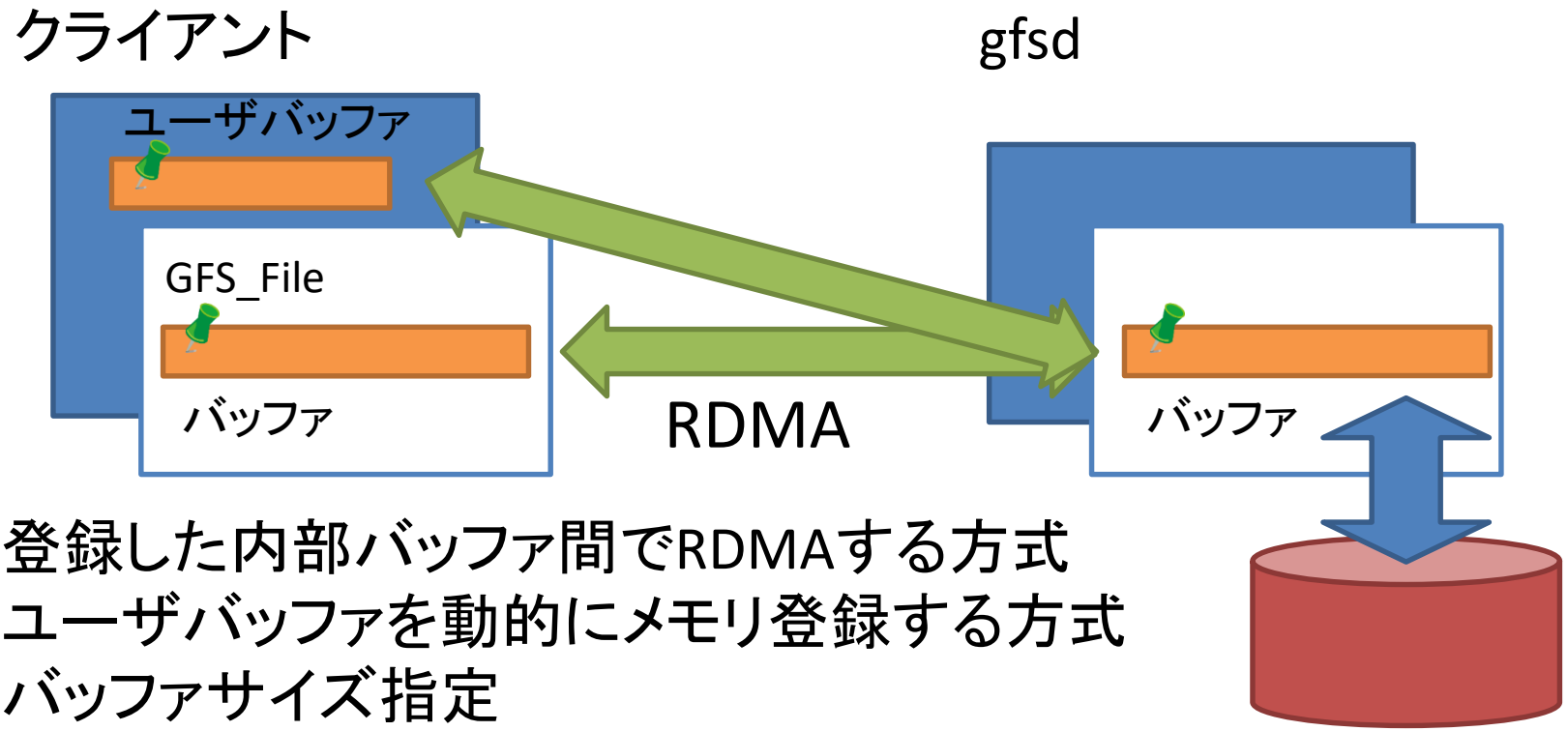
```
% gfedquota –D projectX –H 1000000
                                  # ハードリミット設定
% gfquota –D projectX
```

ディレクトリクオータ機能（２）

- 異なるディレクトリセット間でのディレクトリ移動はできない
- 登録するディレクトリは空ディレクトリ
  - いずれも管理者は可能だがその間Gfarmはロックされる（メンテナンスなどに行うことを想定）
- 異なるディレクトリセット間でハードリンクはできない
- ディレクトリセットからディレクトリを除くためにはそのディレクトリを削除する

RDMAアクセス

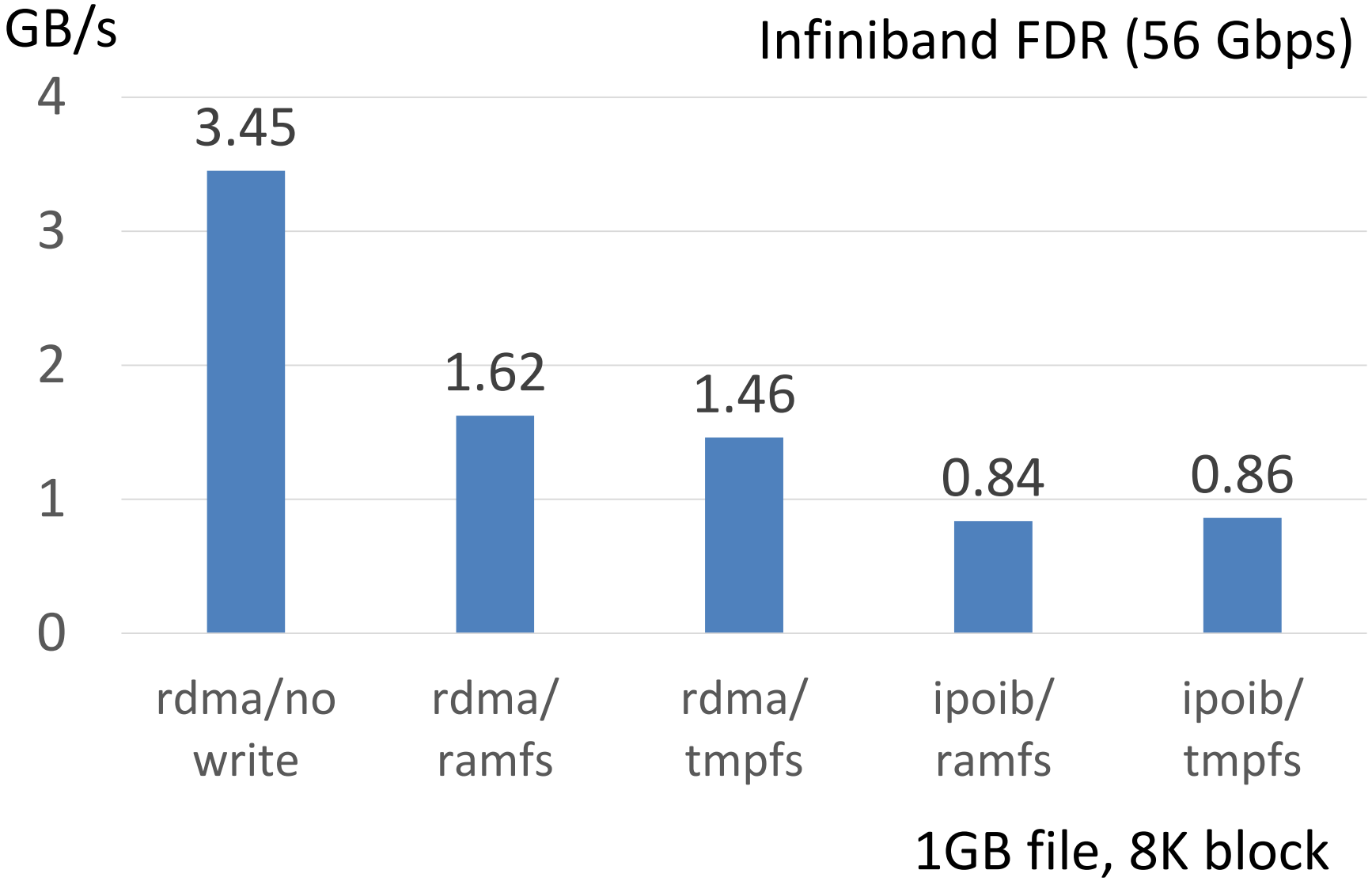
- Gfarm Version 2.7で導入（年内リリース予定）



登録した内部バッファ間でRDMAする方式
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

- RDMA通信に失敗したとき、IP通信へのフォールバック

RDMAファイル書込み性能



シンボリックリンクのフェイルオーバ対応 [2.6.12]

- Gfarm URL形式のシンボリックリンク
  - gfarm://serverA:port/full/path/name

  % gfln -s gfarm://serverA:port/full/path/name name

- serverAがserverBにフェイルオーバしてもシンボリックリンクがたどれる

  metadb_server_list serverA:port serverB:port

データ完全性（1）

- サイレントデータ障害の検知
- digestを書込時に計算しメタデータに保存
- 逐次書込（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
- ファイル複製作成時のdigest検査
- 書込み後ベリファイによるdigest検査 [2.6.8]
- クライアントからのEnd-to-endのデータ完全性

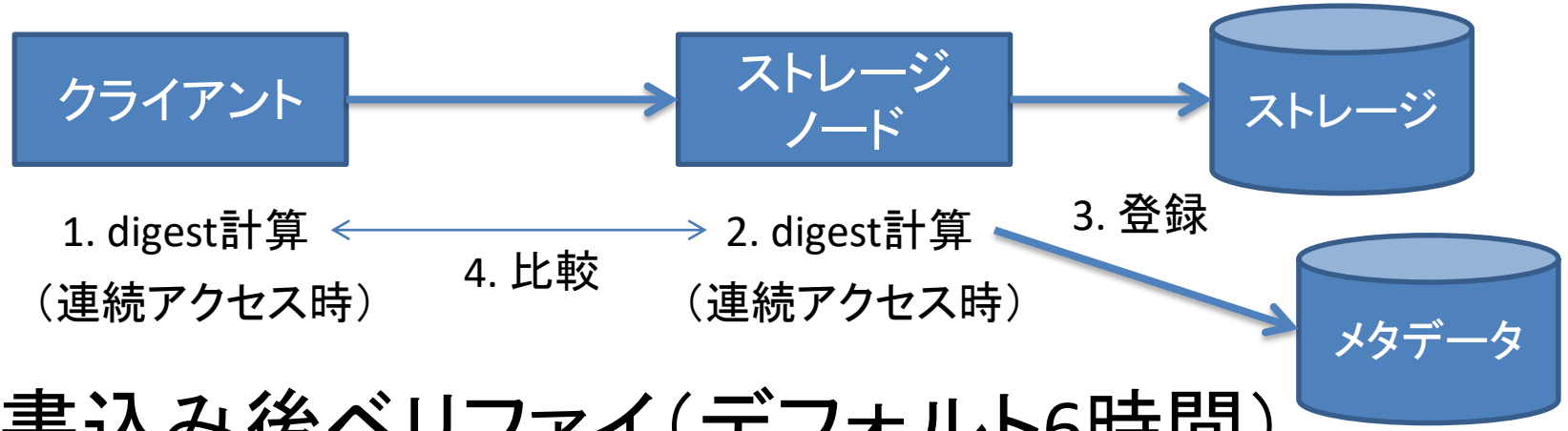
データ完全性（2）

- 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

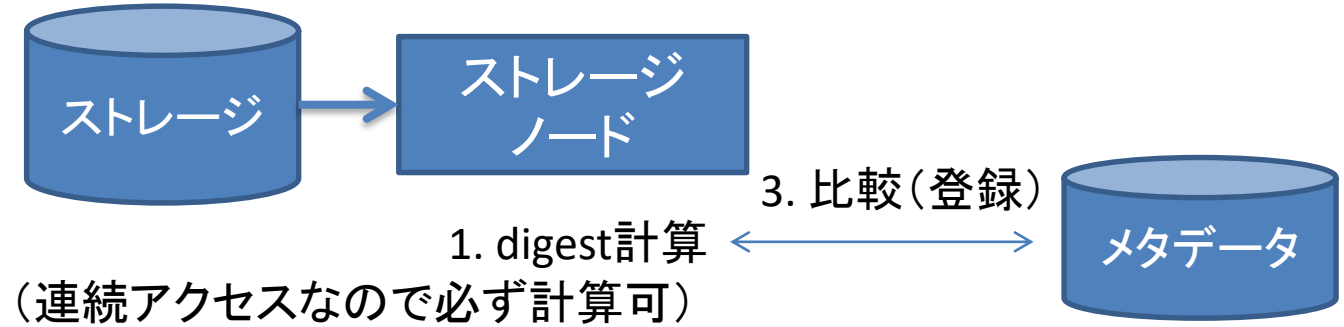
% **gfcksum** *file*
  – ファイルのdigestを表示
% **gfcksum** *-c [-h host] file*
  – （hostに格納されている）ファイルのdigestを計算し確認

Gfarmによるデータ完全性保証（1）

- データ書き込み時

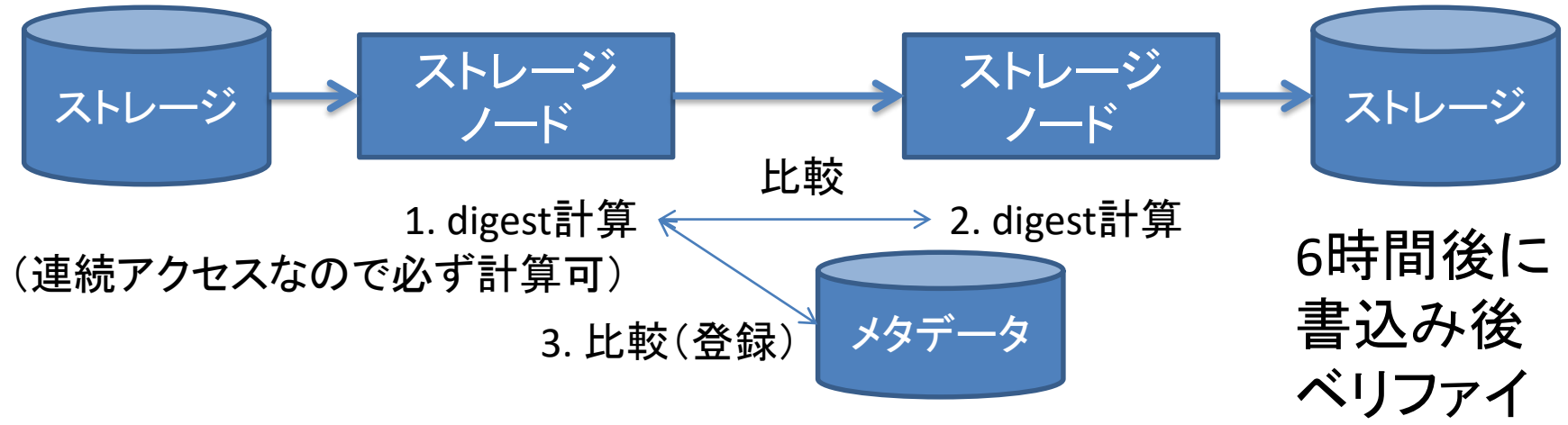


- 書込み後ベリファイ（デフォルト6時間）



Gfarmによるデータ完全性保証（2）

- ファイル複製作成時（書込み後すぐ）

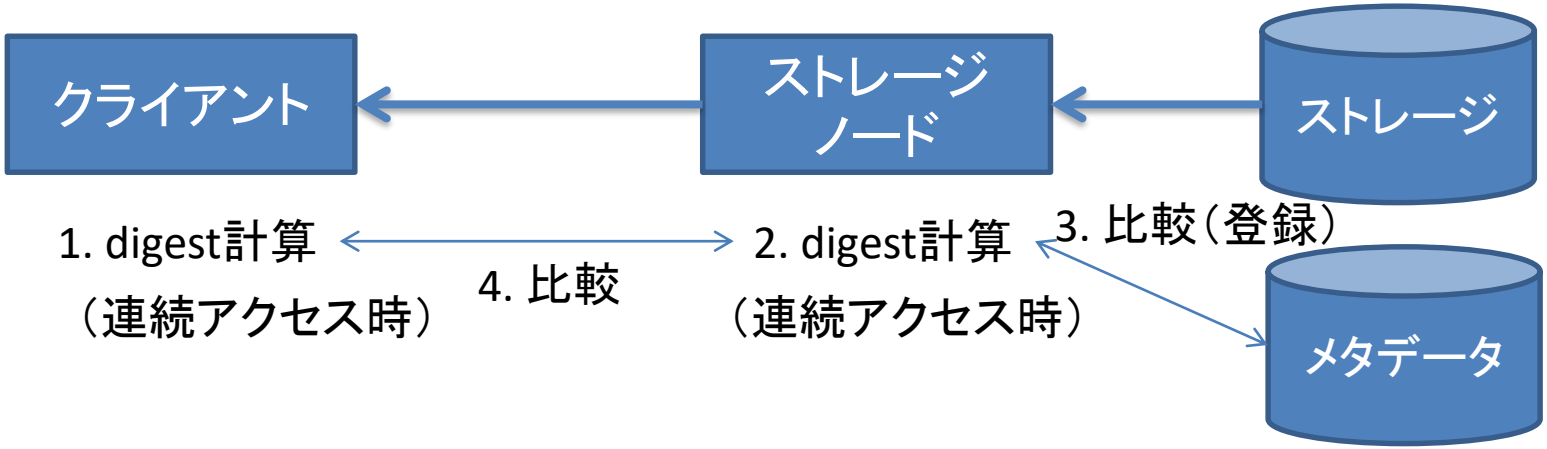


チェックサムが異なる場合、lost+foundへ移動

⇨ ファイル複製による自動修復

Gfarmによるデータ完全性保証（2）

- データ読み込み時



チェックサムが異なる場合、全体ファイル読み込み時にreadがI/Oエラーを返し、lost+foundへ移動

⇨ データ化けファイルの読み込みを防ぐ

ファイル複製による自動修復

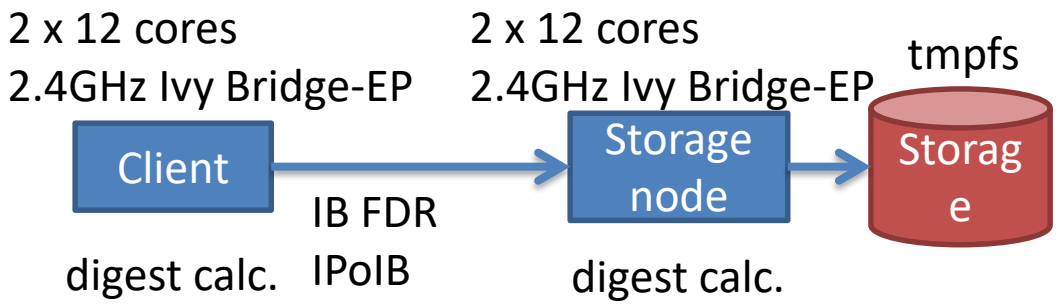
Openssl digest 計算性能

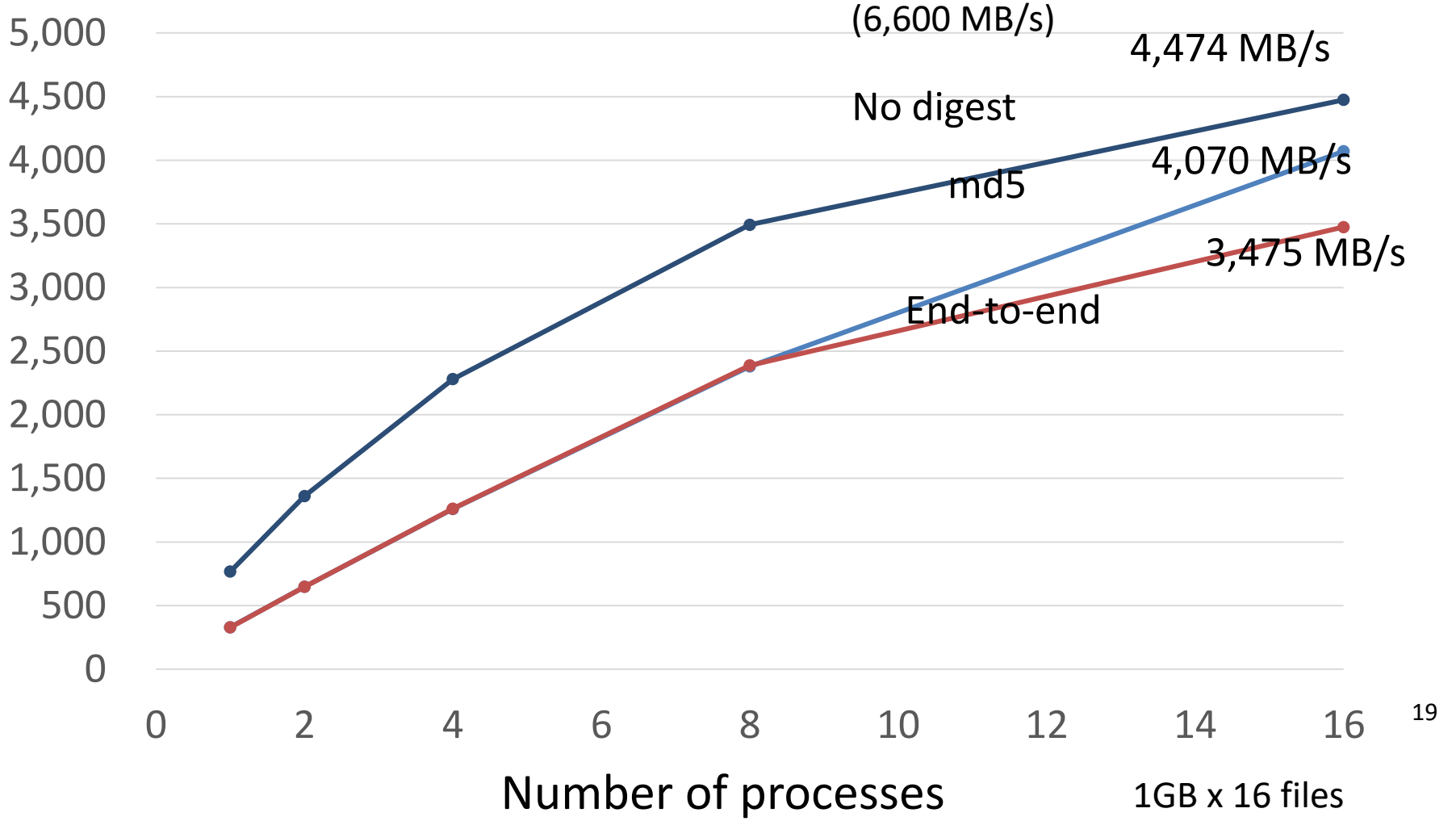
	md5	sha1	sha256	sha512
2.4GHz Xeon E5-2695 v2 (Ivy Bridge-EP)	541	584	218	337
2.4GHz Xeon E5-2665 (Sandy Bridge-EP)	564	585	176	274
2.4GHz Xeon E5620 (Westmere-EP)	483	417	150	238

8KB block size, MB/s

書込み性能

Write Bandwidth [MB/s]





JLDGにおける運用例

- 7.8 PB, 7 拠点, 39 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
  - 7.2 PB利用, 109 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
- 大容量、高速なデータ共有
- 耐障害性に優れる
- データ完全性、サイレントデータ損傷対応
- RDMAデータアクセス
- HPCI共用ストレージ、JLDGなど実運用実績
- ビッグデータ処理