

Gfarm Symposium 2021

2021年10月29日 @市ヶ谷、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

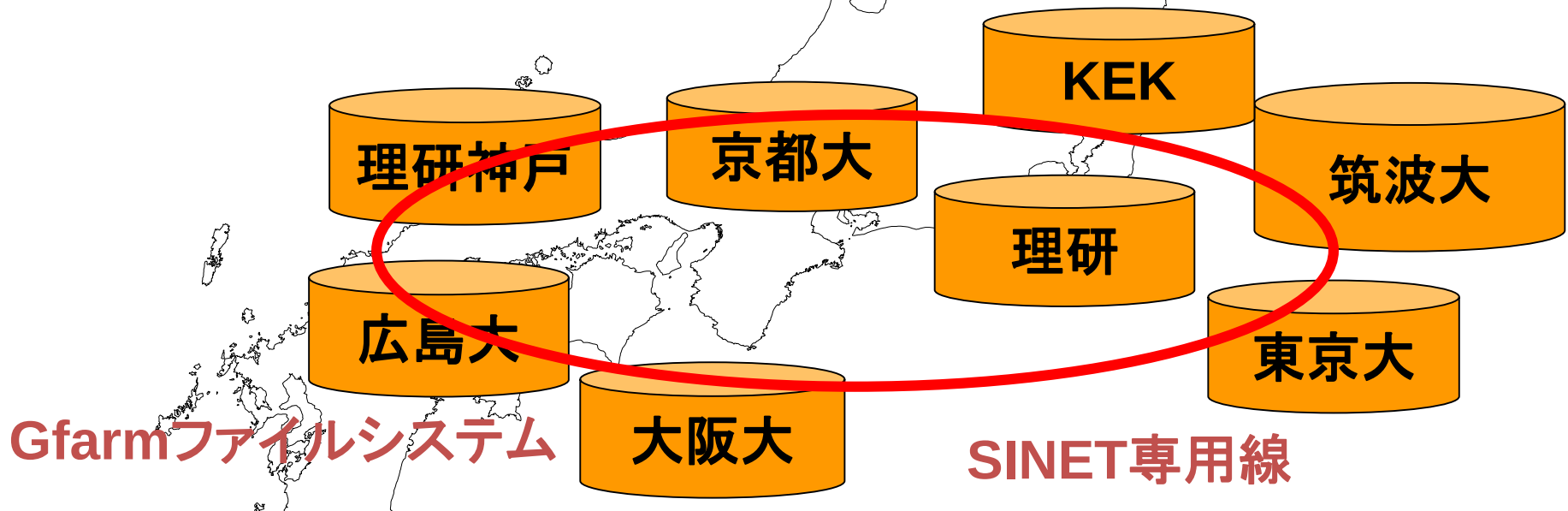


Gfarmファイルシステム(2)

- JLDG(15PB、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

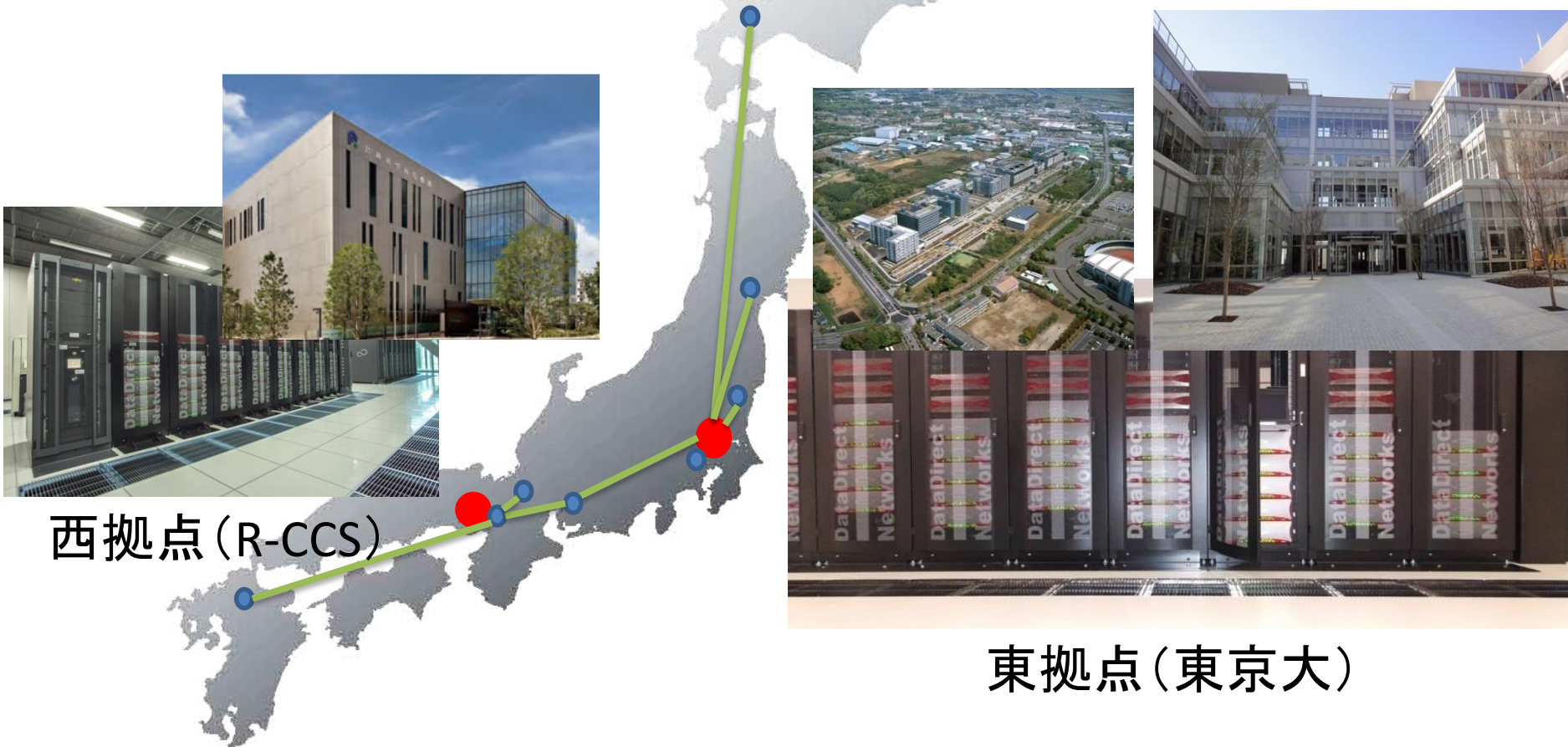
Japan Lattice Data Grid (JLDG)

- 国内素粒子物理学研究者のための15PB規模の広域共有ファイルシステム
 - スパコンで数ヶ月～数年計算したシミュレーションデータの共有
 - 各拠点のファイルサーバを束ね、ファイルは必要な数の複製を作成
 - 各拠点では格納場所を意識せずアクセス
 - 複製を持っている拠点はアクセスが高速に



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



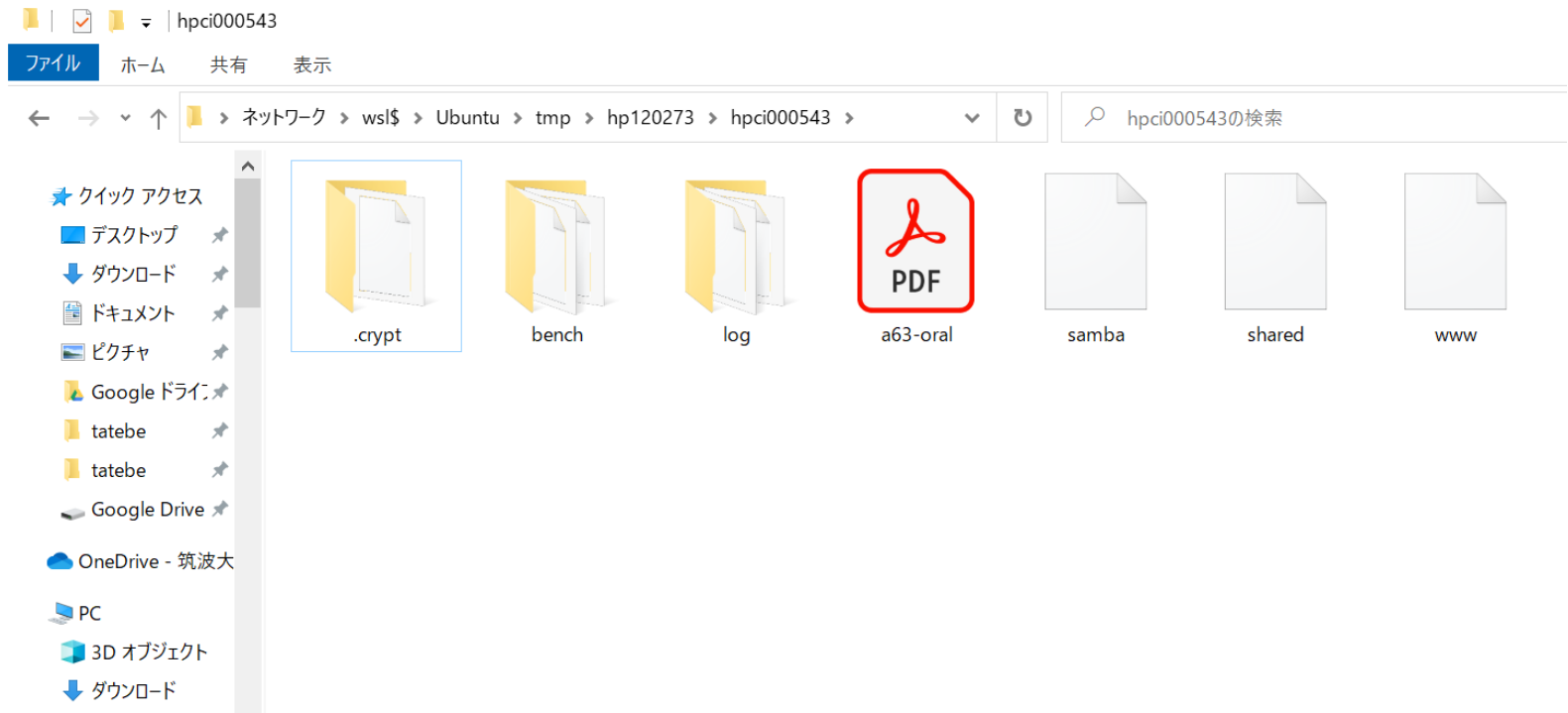
最新機能・状況紹介

主なリリース

日付	version	新機能、更新機能
2021/6/2	2.7.19	• 軽微修正
2021/4/23	2.7.18	• network_send_timeout
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！
2019/9/10	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

Windowsからのアクセス

- WSLでマウント
\$ mount.hpci allow_other
- Explorerでアクセス



Gfarm-S3-MinIO: GfarmのS3互換インタフェース

Gfarm-S3

ユーザ名

パスワード

ログイン

管理コンソールログイン

Gfarm-S3

起動中

停止

アクセス K4XcKzocrUhrnCAKrx2Z
キー
シークレットキー

表示

変更

認証方式 GSI (grid-proxy-init)

有効期限 Tue Feb 16 09:50:33 2021

延長

S3サーバ起動

Gfarm-S3

共有設定

test1

test2

共有設定

Gfarm-S3

test1

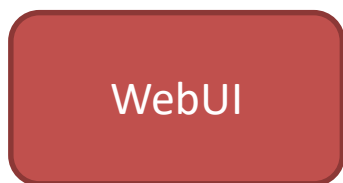
Mon Feb 15 23:22:08 2021

user/group	read	write
× MY GROUP	☐	☐
× OTHER	☐	☐
× user:user2(user2)	☒	☒
×	☐	☐

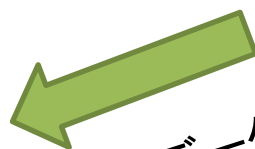
工

group:gfarmadm
group:gfarmroot
user:Gfarm
administrator(gfarmad
user:Gfarm

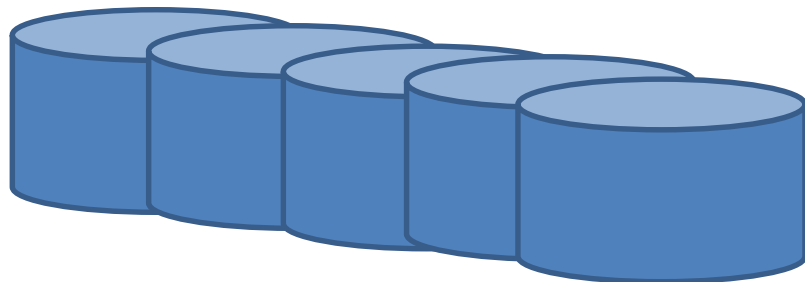
更を適用



サーバ制御



データ格納



Gfarmファイルシステム

並列コピー

- gfpconcat – 複数のファイルを結合して並列コピー

```
$ gfpconcat [ -j 4 ] -o out file1 file2 ...
```

- gfcop – 単一ファイルの並列コピー

```
$ gfcop [ -j 4 ] src dest
```

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

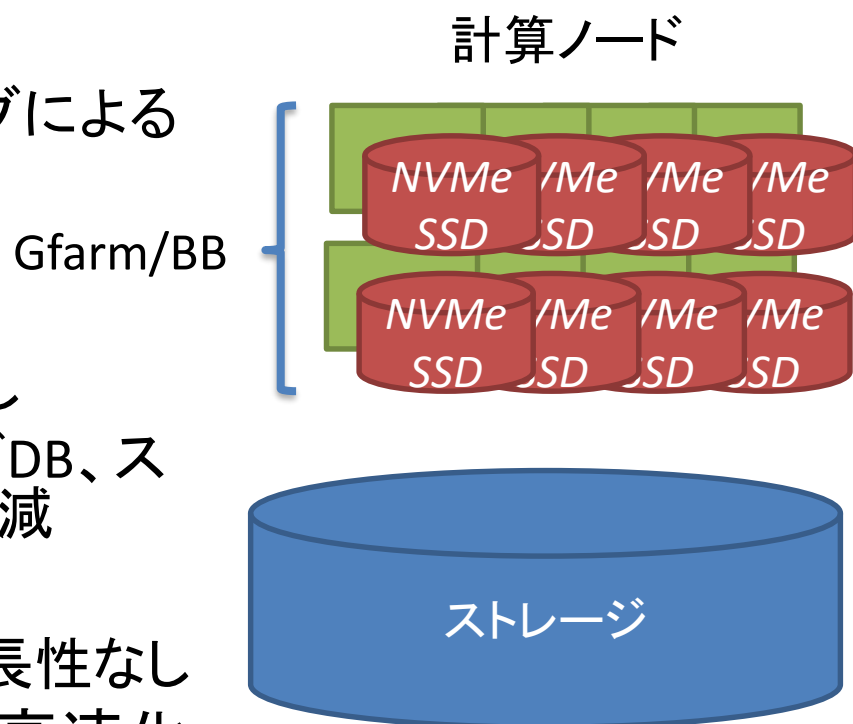
```
$ gfarm2fs /tmp/gfarm
```

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

Gfarm/BBノーストバッファ [建部 2020]

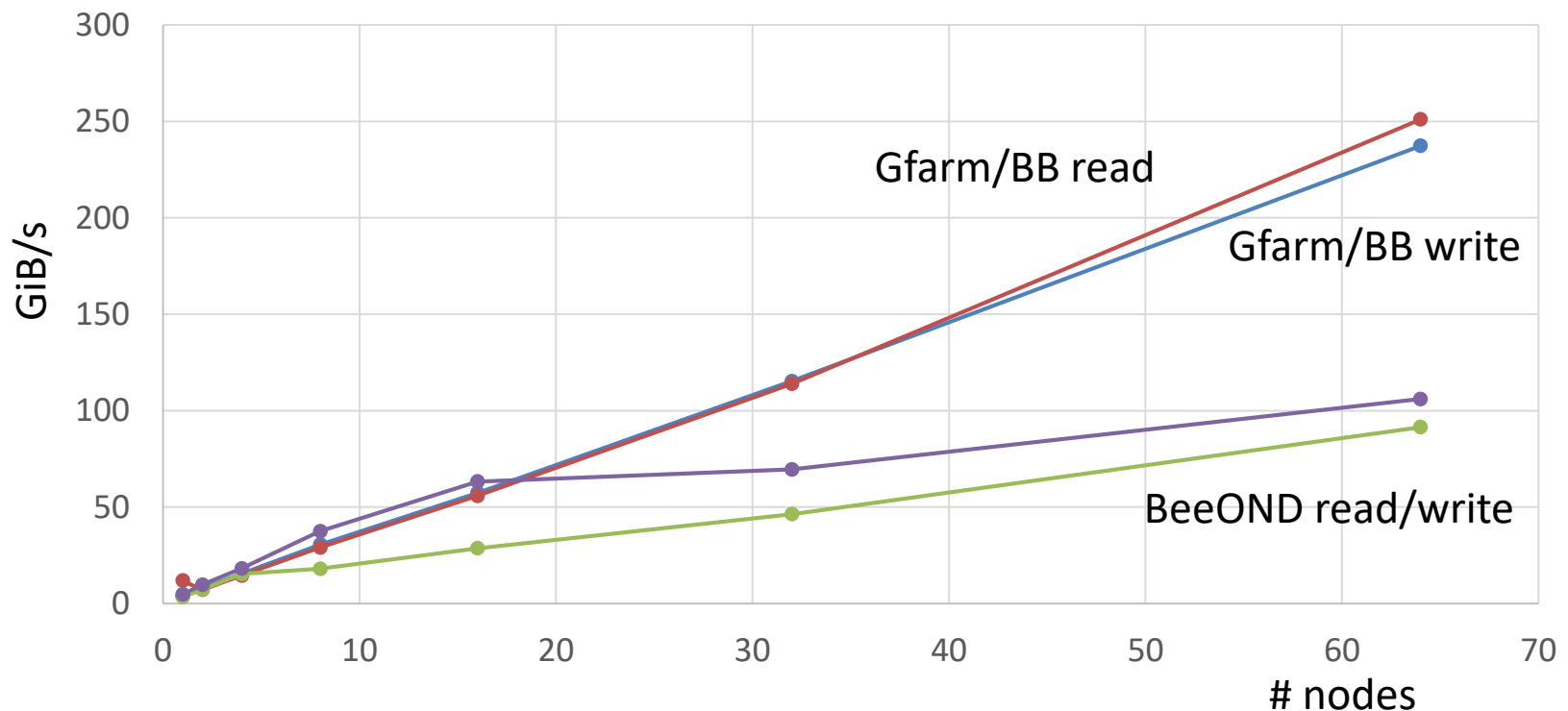
- ノードローカルNVMe SSD等高速ストレージによる一時的な分散ファイルシステム
- アクセス性能の向上
 - ファイルディスクリプタパッシングによるgfsdを経由しない直接アクセス
 - RDMAアクセス
- メタデータ性能の向上
 - メタデータの永続性、冗長性なし
 - ジャーナル書込み、バックエンドDB、スレーブgfmmdのオーバヘッドの削減
- 冗長性オーバヘッド削減
 - ファイル複製によるデータの冗長性なし
- ファイルシステム構築、撤去の高速化



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

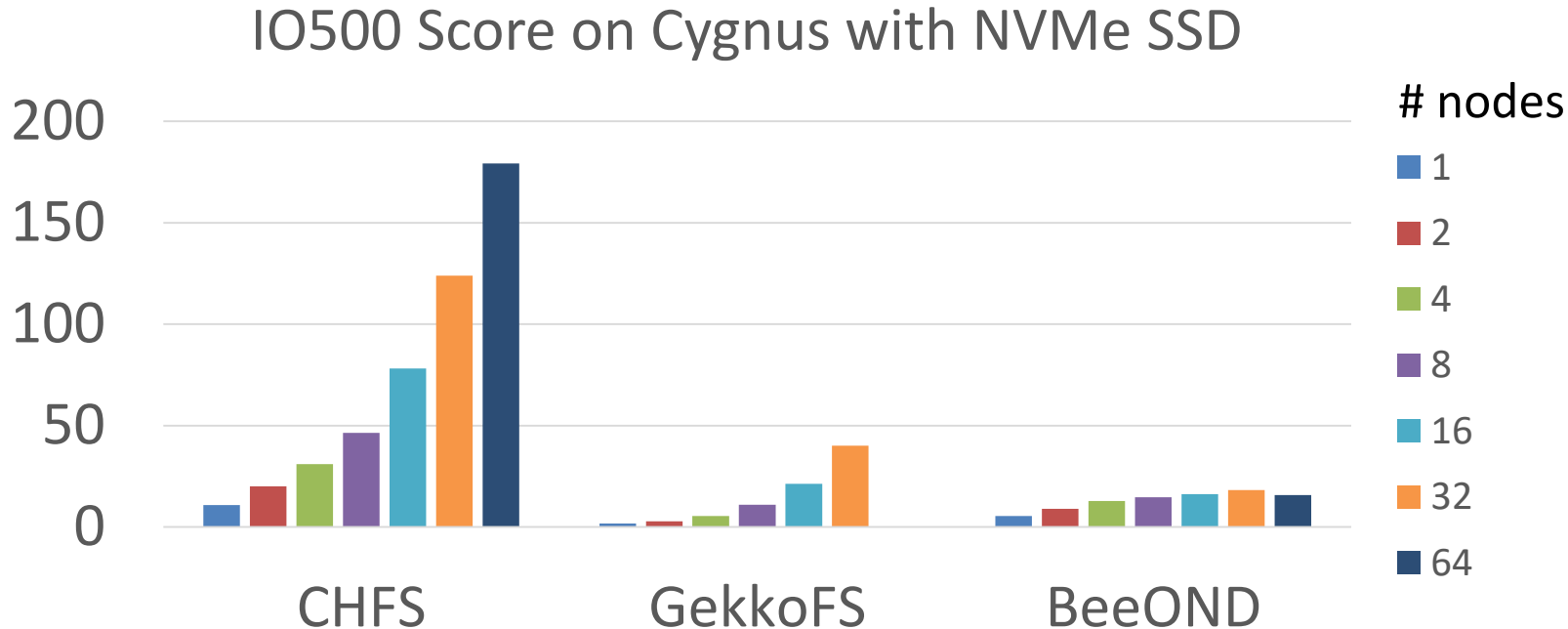


CHFS – 並列分散ファイルシステム [HPC Asia 2022]

- 不揮発性メモリを活用
 - In-memory persistent key-value storeを利用
- メタデータアクセスのオーバヘッドを極力軽減、スケーラブルに性能を向上させる
 - メタデータサーバなし、逐次処理なし
 - ノードローカルストレージで構成した分散キーバリューストアを基盤
- データアクセス性能をスケーラブルに向上させる
 - (バイト単位)チャンクに分割

Cygnusにおける性能評価

- IO500ベンチマーク
 - 計12種のバンド幅とメタデータ性能計測
 - IO500スコアは幾何平均



# ↑	INFORMATION								IO500	
	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW	MD
									(GIB/S)	(KIOP/S)
1	ISC21	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng	MadFS	10	1,800	2,595.89	193.77	34,777.27
2	ISC21	Intel	Endeavour	Intel	DAOS	10	1,440	1,859.56	398.77	8,671.65
3	ISC20	Intel	Wolf	Intel	DAOS	10	420	758.71	164.77	3,493.56
4	ISC21	Lenovo	Lenovo-Lenox	Lenovo	DAOS	10	960	612.87	105.28	3,567.85
5	ISC20	TACC	Frontera	Intel	DAOS	10	420	508.88	79.16	3,271.49
6	ISC21	National Supercomputer Center in GuangZhou	Venus2	National Supercomputer Center in GuangZhou	kapok	10	480	474.10	91.64	2,452.87
7	ISC20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	10	380	440.64	95.80	2,026.80
8	ISC21	Supermicro		Supermicro	DAOS	10	1,120	415.04	112.17	1,535.63
9	SC19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	249.50	86.97	715.76
10	SC20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3,800	239.37	45.79	1,251.32
11	ISC21	Olympus Storage Technology Innovation Lab	OceanStor	Huawei	OceanFS	10	960	220.10	69.49	697.15
12	SC20	Johannes Gutenberg University Mainz	MOGON II	JGU (ADA-FS)& BSC (NEXTGenIO)	GekkoFS	10	240	167.64	22.97	1,223.59
13	SC20	DDN	DIME	DDN	IME	10	110	161.53	101.60	256.78
14	SC19	WekaIO	WekaIO	WekaIO	WekaIO Matrix	10	2,610	156.51	56.22	435.76
15	ISC21	University of Tsukuba	Cygnus	OSS	CHFS	10	240	148.69	30.39	727.61
16	ISC21	Joint Institute of Nuclear Research	Govorun	RSC	DAOS	10	160	132.06	20.19	863.69
17	SC20	TACC	Frontera	DDN	IME	10	280	109.91	176.23	68.55
15	ISC21	University of Tsukuba	Cygnus	OSS	CHFS	10	240	148.69	30.39	727.61
16	ISC21	Joint Institute of Nuclear Research	Govorun	RSC	DAOS	10	160	132.06	20.19	863.69
17	SC20	TACC	Frontera	DDN	IME	10	280	109.91	176.23	68.55

#15 in 10 node list
#23 in full list

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 2.7.19を6/2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)
 - TLS対応, OAuth2, Kerberos認証
- CHFS並列分散ファイルシステム
 - #15 in 10 node list, #23 in full list (June 2021)