Gfarmファイルシステムの概要

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

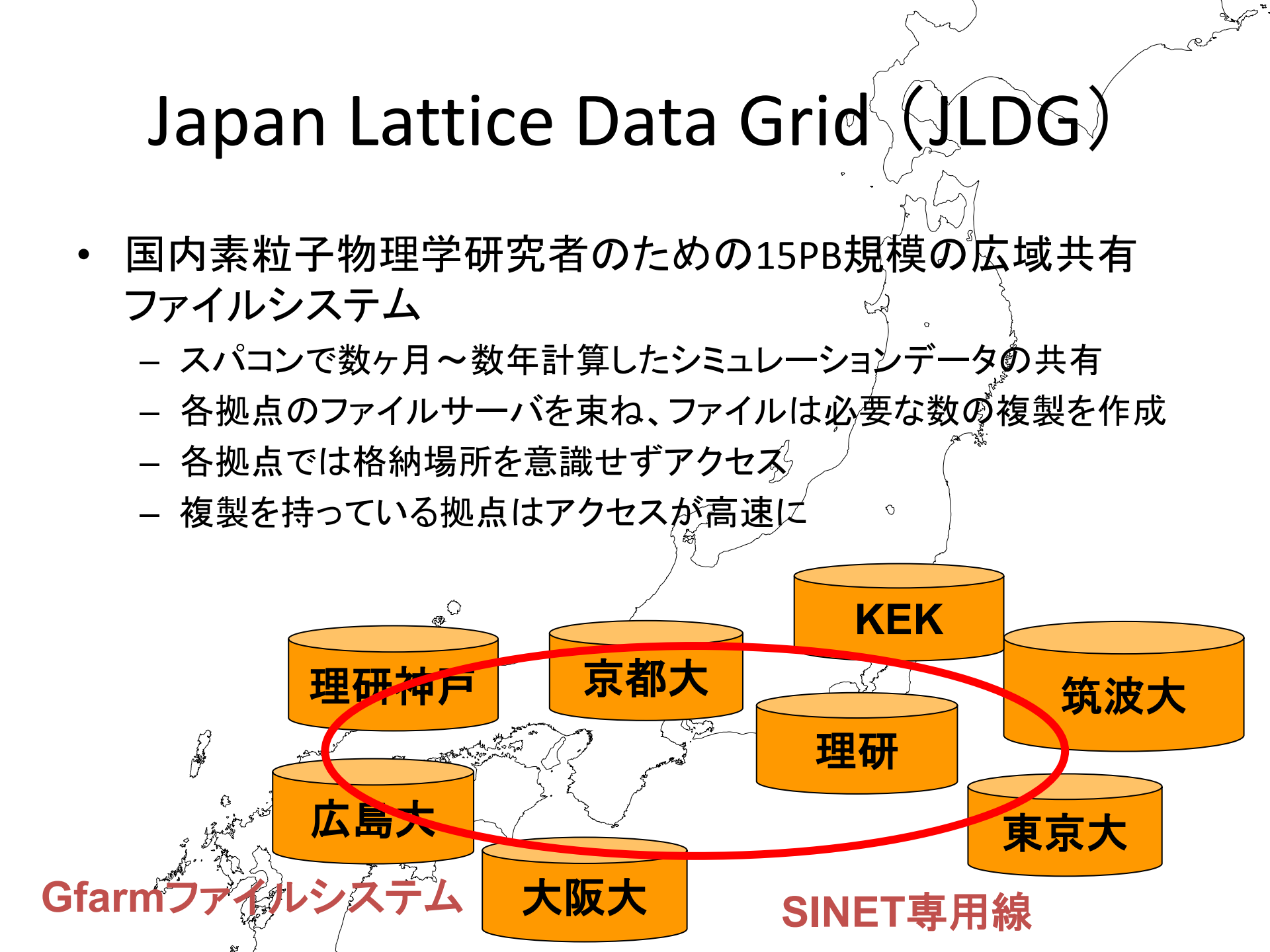
oss-tsukuba.org

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



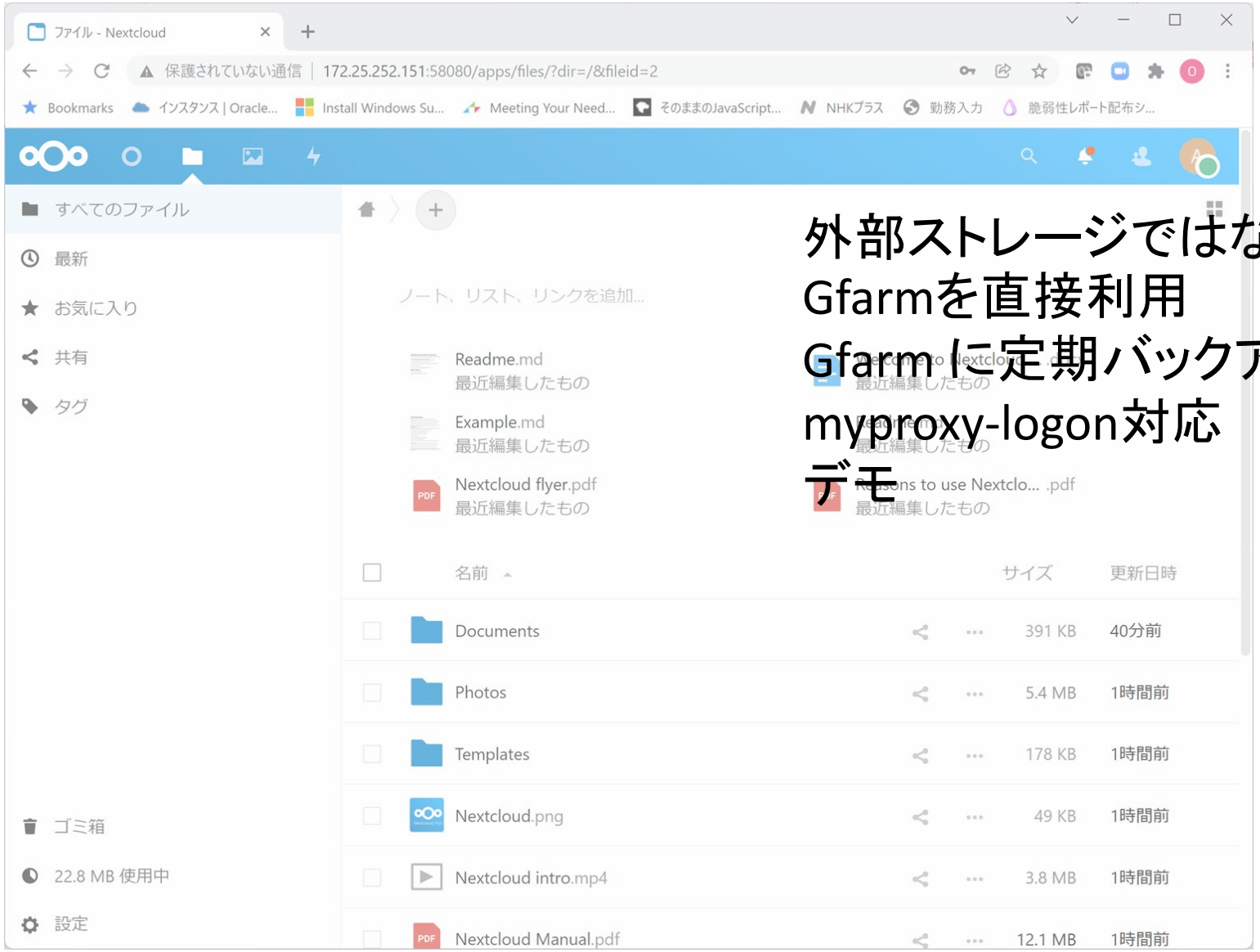
HPCI共用ストレージ

- 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

西拠点（R-CCS）

東拠点（東京大）

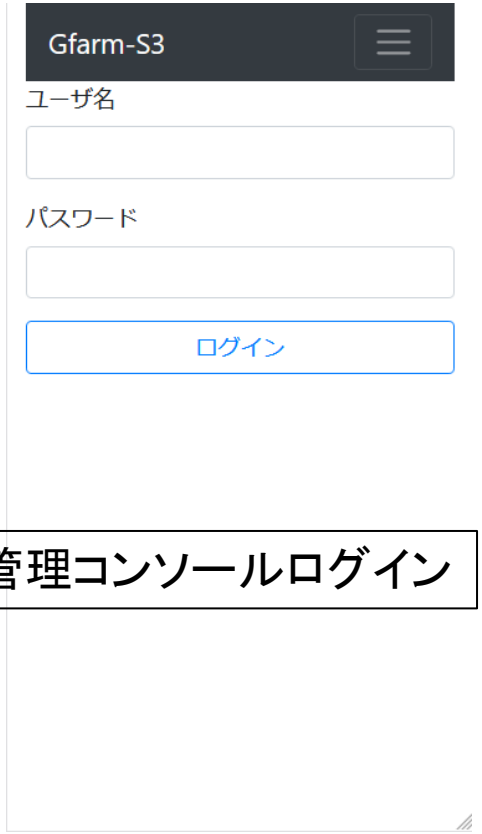
[NEW] NextCloudコンテナ



外部ストレージではなく
Gfarmを直接利用
Gfarm に定期バックアップ
myproxy-logon対応
デモ

[NEW] ARM Linux対応

[UPDATED] Gfarm-S3-MinIO: GfarmのS3互換インタフェイス



管理コンソールログイン

S3サーバ起動

共有設定

WebUI

データ格納

S3クライアント

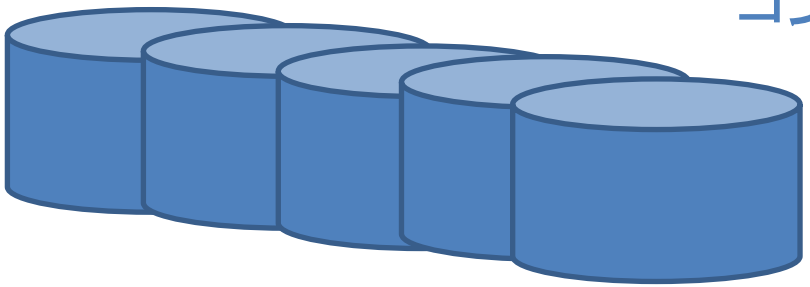
サーバ制御

MinIO

Gfarm gateway

アップロード高速化

NGINX対応
一般ユーザ権限での起動
コンテナ化



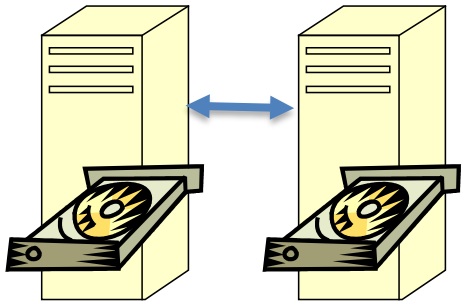
Gfarmファイルシステム

認証・セキュリティについて

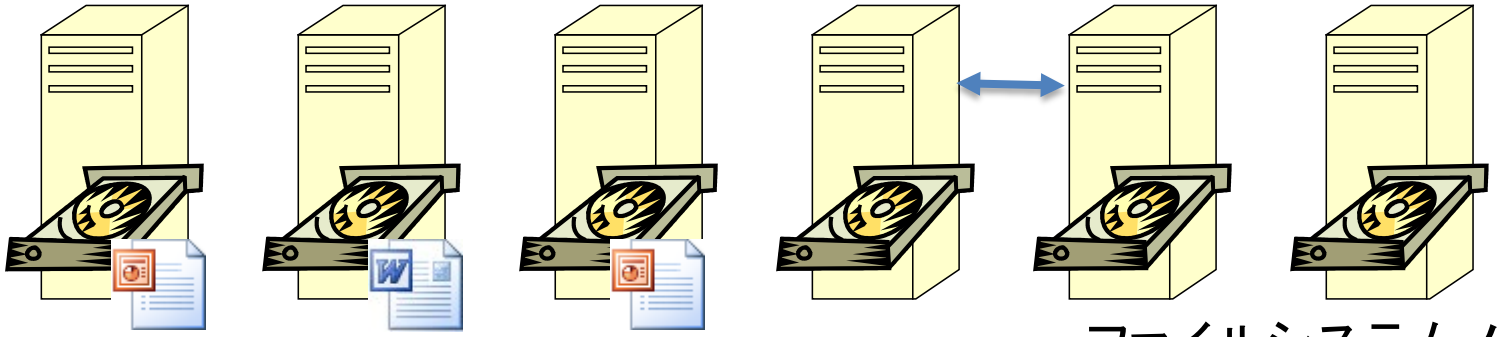
共有鍵（Gfarm）
(代理)証明書（GSI/TLS*）
チケット*（Kerberos）
トークン*（OAuth/OIDC）

Plain/GSI/TLS*

Plain/GSI/TLS*



メタデータサーバ
共有鍵（Gfarm）
証明書（GSI/TLS*）
チケット*（Kerberos）



ファイルシステムノード群
共有鍵（Gfarm） 証明書（GSI/TLS*）チケット*（Kerberos）

CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory

- スパコン用計算ノードの不揮発性メモリ・SSDを用いたad hoc並列ファイルシステム
- Gfarmとは両極端
  - No metadata server, no central data structure
  - No sequential operation
  - Based on highly parallel KVS
- 論文 https://dl.acm.org/doi/10.1145/3492805.3492807
- 発表 https://youtu.be/K6sO66kAkJ8
- キャッシュ機能については3月のHPC研究会で発表予定

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 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
  - https://github.com/oss-tsukuba/
- HPCI共用ストレージ、JLDGなど実運用実績
- 暗号化ファイルシステム、完全性
- まもなくリリース
  - TLS対応, OAuth2, Kerberos認証
  - S3コンテナ、nextcloudコンテナ