HPCI 共用ストレージ
R-CCS 第3期システムと今後について

― Gfarm Workshop 2024 ―

理化学研究所 R-CCS
金山 秀智、原田浩
2024/12/18

HPCI共用ストレージおよび運用状況

HPCI共用ストレージ

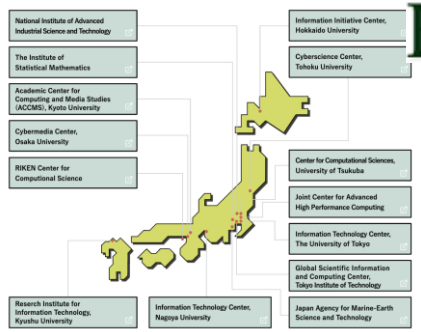
■HPCI

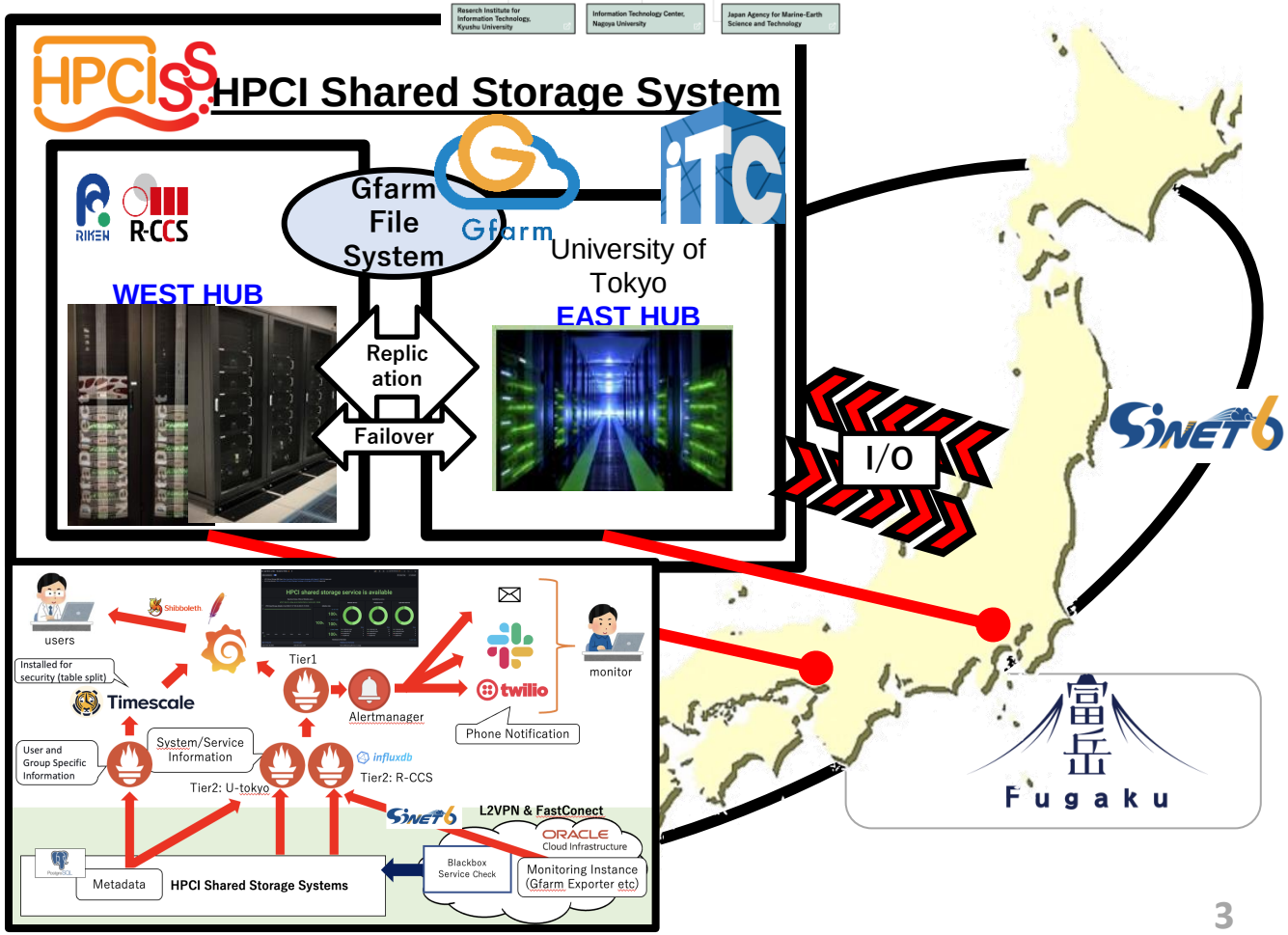
- 国内の大学や研究機関の計算機システムやストレージを高速ネットワークで結んだ共用計算環境基(High Performance Computing Infrastructure)
- 日本のHPCIリソースを効率的に利用と研究環境の提供が目的。
- HPCI provide to service
  - ・スパコン利用者の申請対応
  - ・利用ソフトウェアの提供
  - ・CMSやチケット/ヘルプデスクの提供
- HPCI provide to Infrastructure
  - ・Single sign-on authentication
  - ・Network(SINET - ScIentific NETwork)との連携
  - ・Network Storage Service → HPCI共用ストレージ

■HPCI共用ストレージ

- 日本国内の大学/研究所の各スパコンのログインノードからアクセス可能なネットワークストレージシステム
- スパコンで生成された研究データを保存し、各スパコン間で共有が可能。
- 並列I/Oを可能とし、50GB/sec以上でのI/Oが可能。
- 東京大学とR-CCSの2拠点で運用
- 課題数: 256件/利用者数: 2278名 (2024/12/10時点)
- 運用期間: 2012年度 ～ (今年で13年目)

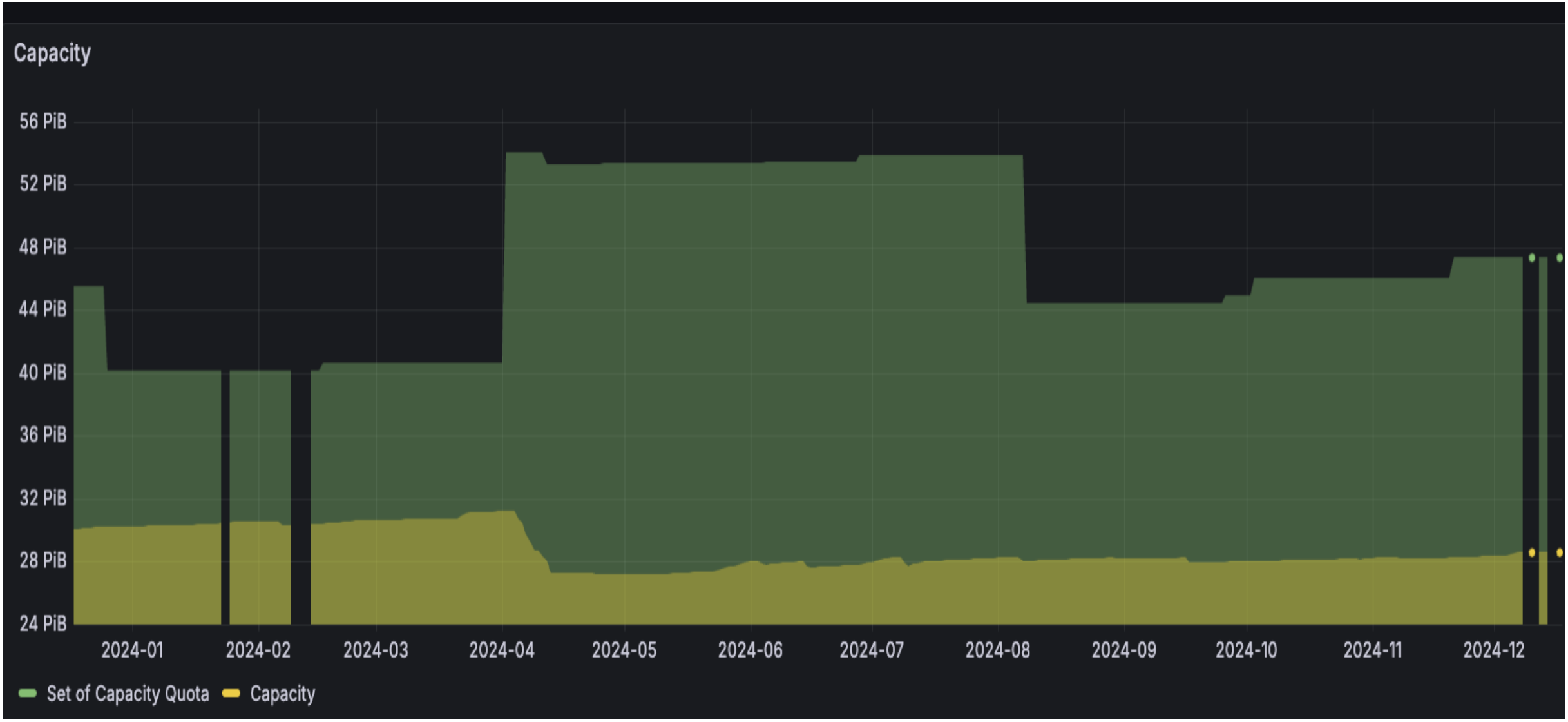
世代	期間(年度)	容量
第1世代	2012 – 2018	10PB → 15PB
第2世代	2018 - 2024	45PB → 50PB
第3世代	2025 -	95PB+(予定)





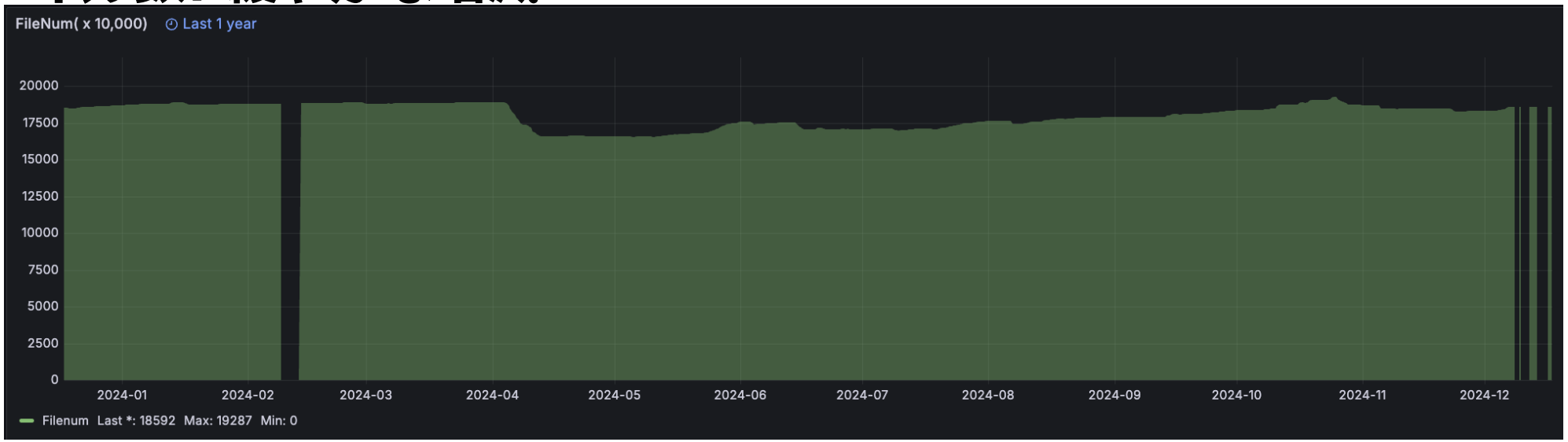
HPCI共用ストレージの利用状況: 容量

- 容量は安定して増加/割当量は既に頭打ち気味



HPCI共用ストレージの利用状況: ファイル数

- ファイル数: 緩やかな増減

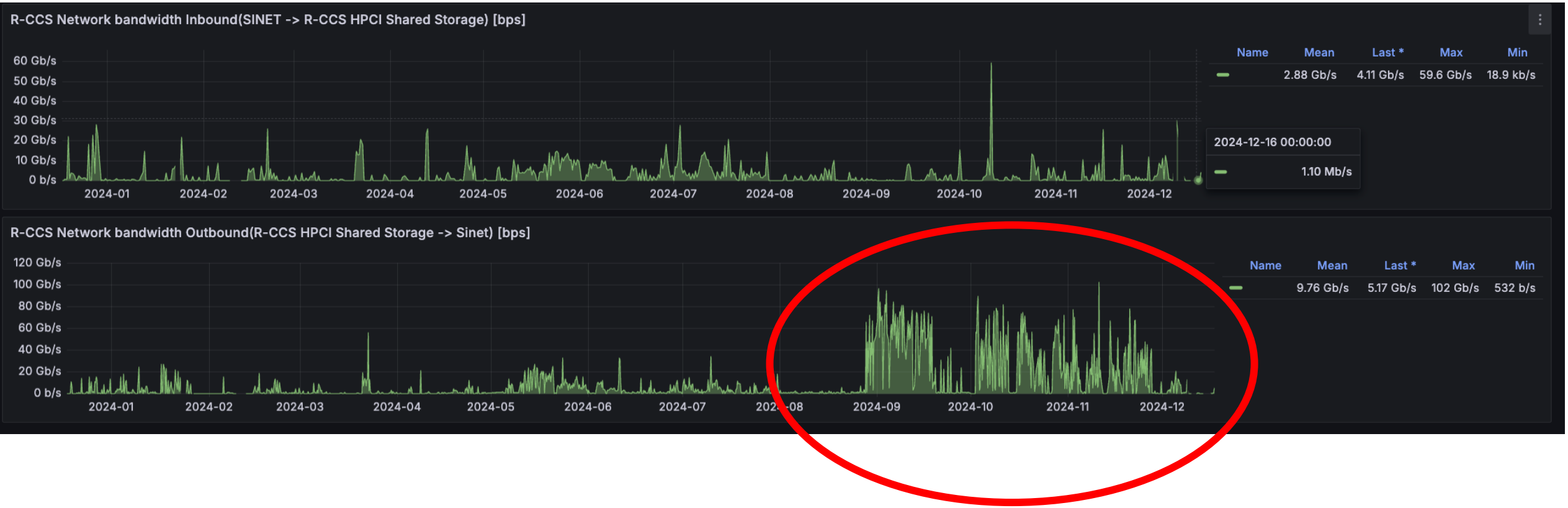


- 平均ファイルサイズ: 150MBをキープ/smallファイル需要有り/gfptar効果高い

HPCI共用ストレージの利用状況

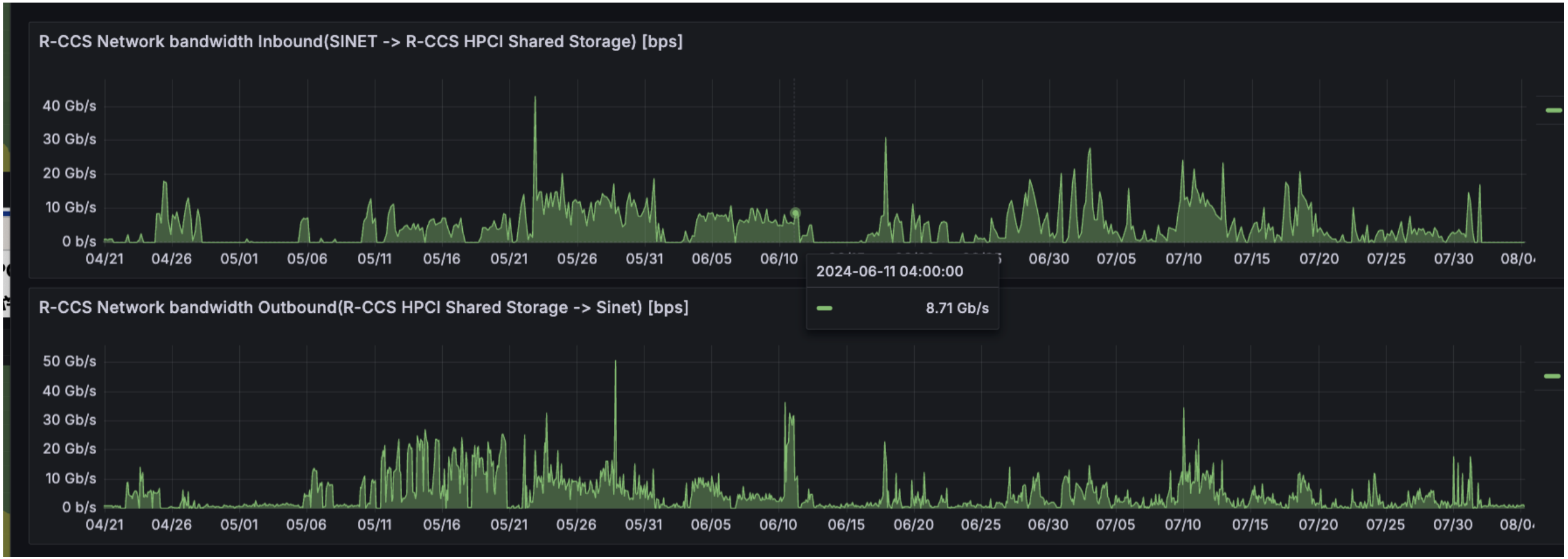
- 2024/08末からはデータ移行による東大の新機材へのレプリケーション通信が走っているため、高帯域でのデータ転送が発生していた。

  → 今後の発表？



HPCI共用ストレージの利用状況

- データ移行前の情報:
  レプリケーションがなくてもユーザアクセスが続いている傾向。



保存データ統計

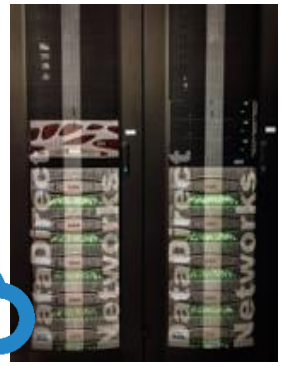
- 保存データの利活用が課題になっている。
- 2021-2023に保存されたデータが利用が止まり気味
- オープンサイエンスの取り組みなどが必須。

保存データ統計

- ファイルごとの平均は期待以上(100MB/fileで設計)
- 容量が大きなファイルが増加しているのは良いこと。
- 400Gbps x 2をメタデータボトルネックなく埋めるためには大きなファイルの転送が必須。→ gfptarの利用推進。

第3期 HPCI共用ストレージ

History



Gfarm

2011

First phase system(2012 ~ 2018)
- Capacity: 10PB → 15PB (Logical)
- HUB: Tokyo University, Tokyo Institute of Technology, R-CCS
- FileSystem: Gfarm
- Replication Count: 1 or 2
- Stotsge Type: HDD and Tape
- Network: 40Gbps
  (Max I/O Speed: 5.0GB/sec)

Gfarm

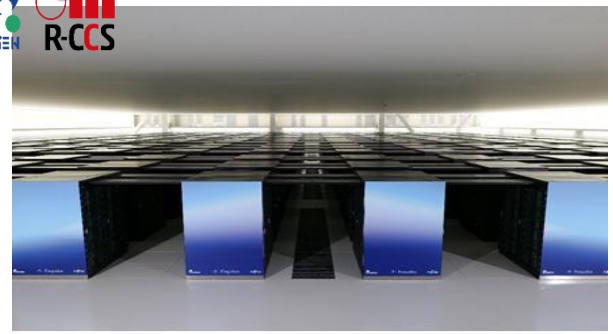
Second phase system(2018 ~ 2024)
- Capacity: 40PB → 50PB (Logical)
- HUB: Tokyo University, R-CCS
- FileSystem: Gfarm
- Replication Count: 2 or More
- Storage Type: HDD
- Network: 100Gbps → 400Gbps
  (Max I/O Speed: 50GB/sec)

Gfarm

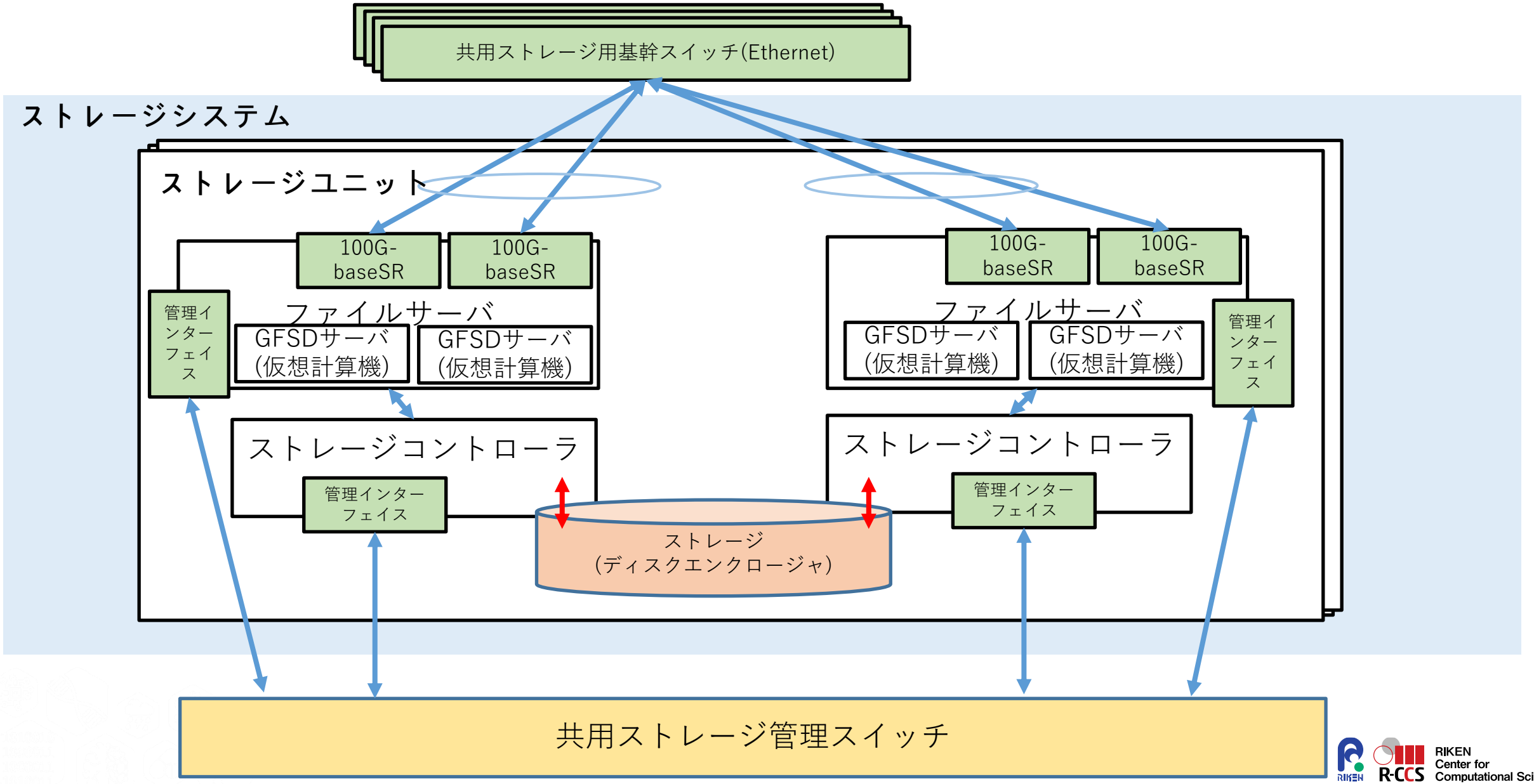
Third phase system(2024 ~)
- Capacity: 95PB+(Logical)
- HUB: Tokyo University, R-CCS
- FileSystem: Gfarm
- Network: 400Gbps
- Storage Type: HDD

K Computer(2011~)

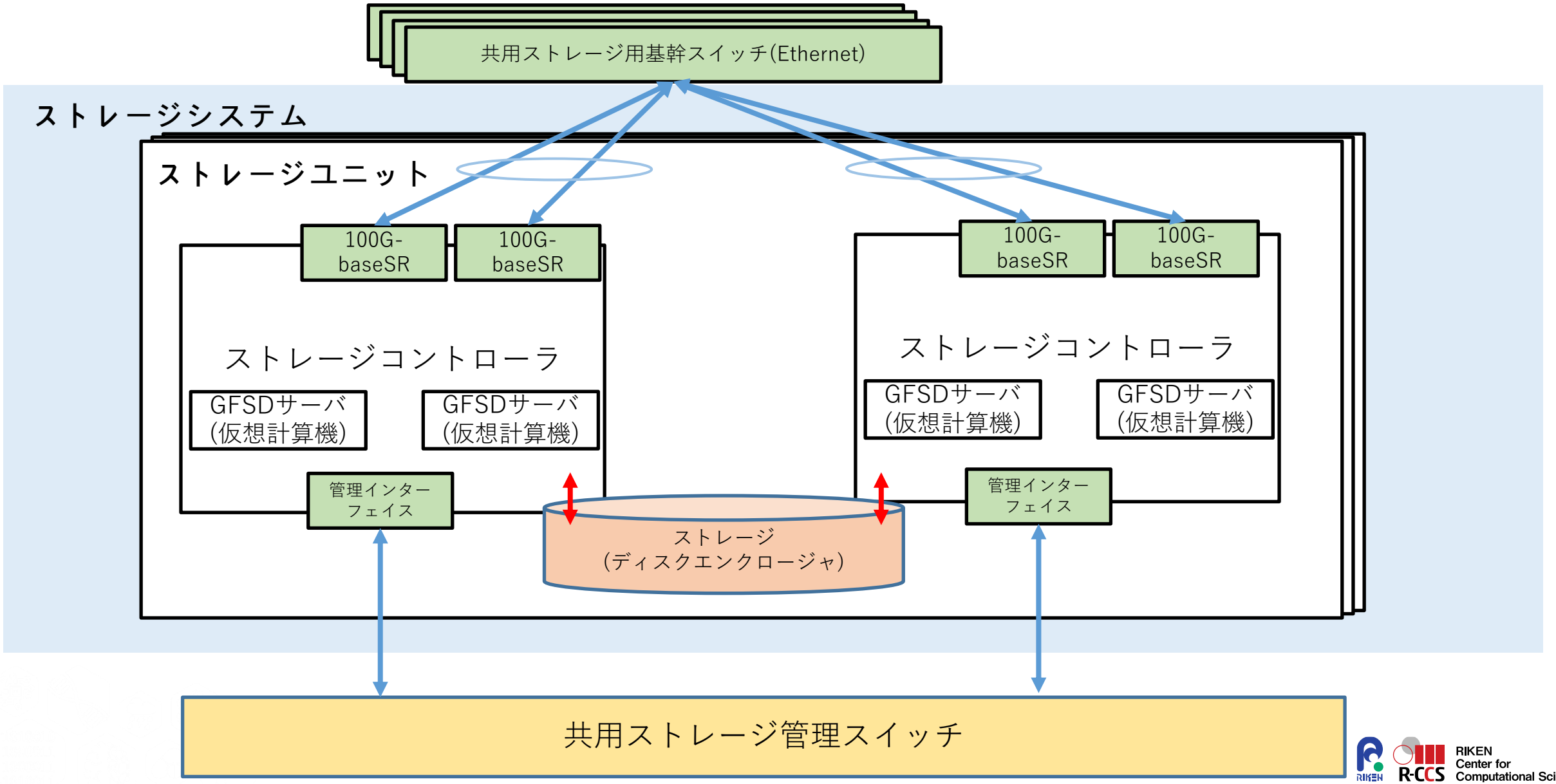
Fugaku(2020~



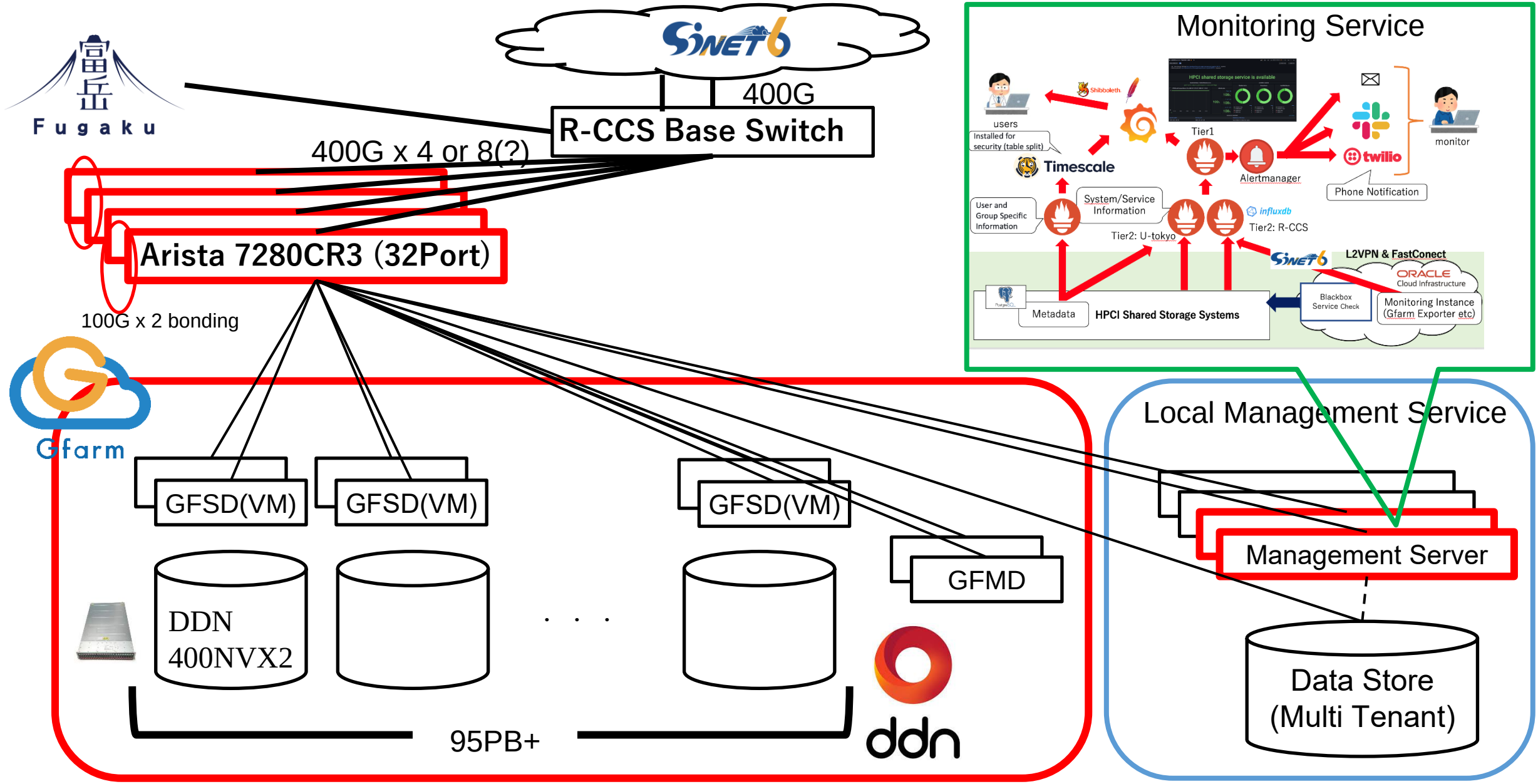
第三世代機材環境(想定1)



第三世代機材環境(想定2) → こちらになりました



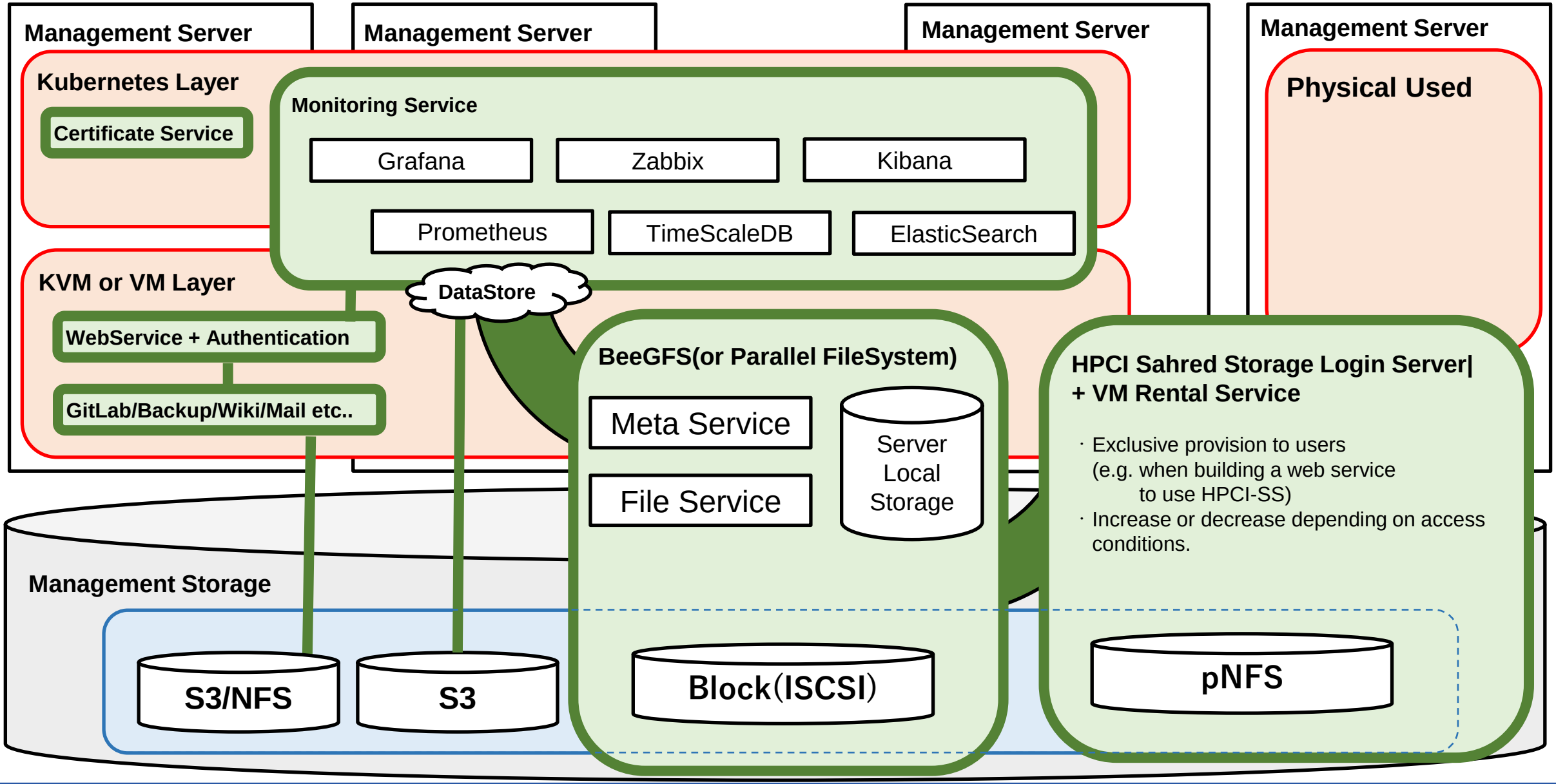
R-CCS HPCI共用ストレージ第三期システム(予定)



R-CCS HPCI共用ストレージ第三期システム(予定)

	2nd(FY2018)		3rd (予定)	Remarks
Metadata System	Dell R730xd		Dell R760	Intel Xeon Gold 6444Y x2(CPU)
Metadata Memory	768 GB		2048 GB	
Metadata I/O	15000+ (count)		41000+ (count)	time dd if=/dev/zero of=path_to_meta bs=512 count=1000000 oflag=dsync
System(Storage)	(1) CMS Hyper STOR Flex	(2) DDN SFA14K	DDN SFA400NVX2E (VM)	
System(FileServer)	DELL R730	Supermicro		400NVX2内のVMを利用
Storage(HDD)	HDD(10TB)	HDD(12TB)	HDD(22TB)	
Space(ALL/Logical)	45PB → 50PB		95PB+	第3期中の容量の追加を想定
Space(R-CCS/Physical)	45PB+	7.8PB	95PB+	
Throughput(R-CCS)	All: 180GB/sec+ (MAX Read & Write) (1) 1set: 18-20GB/sec All(9set): 160GB/sec + (2) All(1set): 20GB/sec +		All(10set): 300GB/sec+ (MAX Read & Write) 1set: 32GB/sec +	富岳内部との接続は200Gbps+で接続 R-CCS基幹スイッチとは1.6Tbpsでの接続を予定
Network(R-CCS)	100Gbps → 200Gbps → 400Gbps(Dual)		400Gbps (Dual)	SINET6との接続は大阪DC/兵庫DCと接続し冗長構成を実施

Local Management Service



導入の苦労: 空調機装置の導入と

- 弊所都合で設置場所が変更。富岳の隣の小部屋になりました。
- 冷却設備がない → 水冷式空調機の導入が必要となり…

- Motivair – ChilledDoorの導入を予定
- 水冷式空調機であるため水冷設備工事が追加
- 理研の水冷設備では戻り水温が定められているため
  CDUを用いて水量管理を行う必要あり。

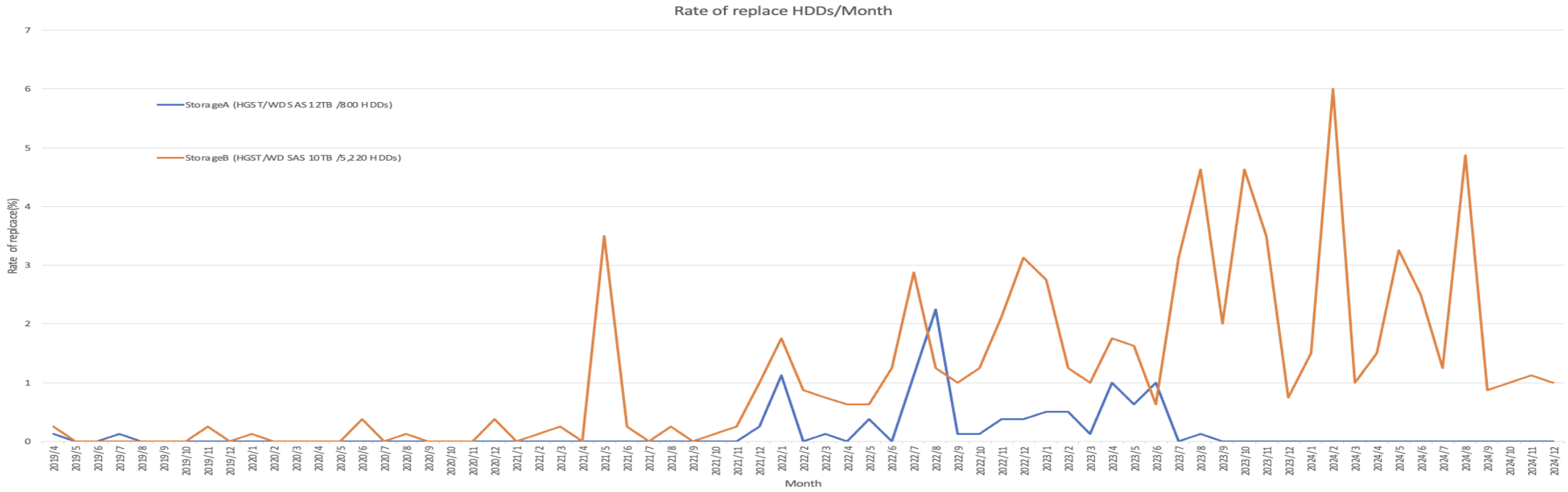
導入の苦労: 円安-SSD

- ストレージは当初100PB+を想定
- vs SSD
  - 性能的にはHDDでも問題ないが電気代や故障率を考えるとSSDメリット
  - Silent Data Corruptionなどのチェックや圧縮機能などの機能あり
    近いシステムからの高速アクセス。
    (メタデータボトルネックになる可能性があるが…)
  - 円安とSSD価格が下がらない(データ重複/圧縮してもHDDメリット)なので
    SSDは諦め。
    - 階層型も考慮したが、性能/容量を天秤にかけて容量優先を決めた。
    - データ圧縮はgfptarの利用をお願いしたい

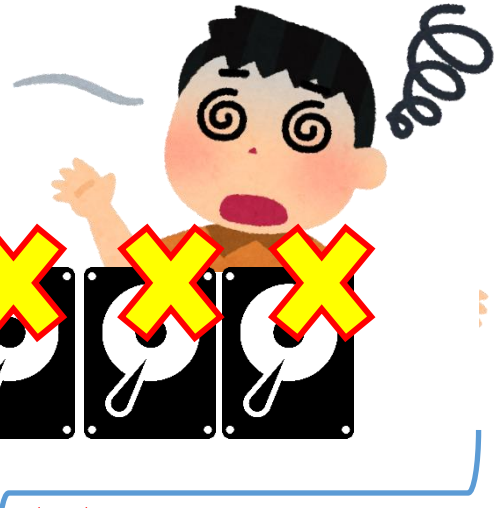
現行機材の8年目運用

HDD故障率の増加

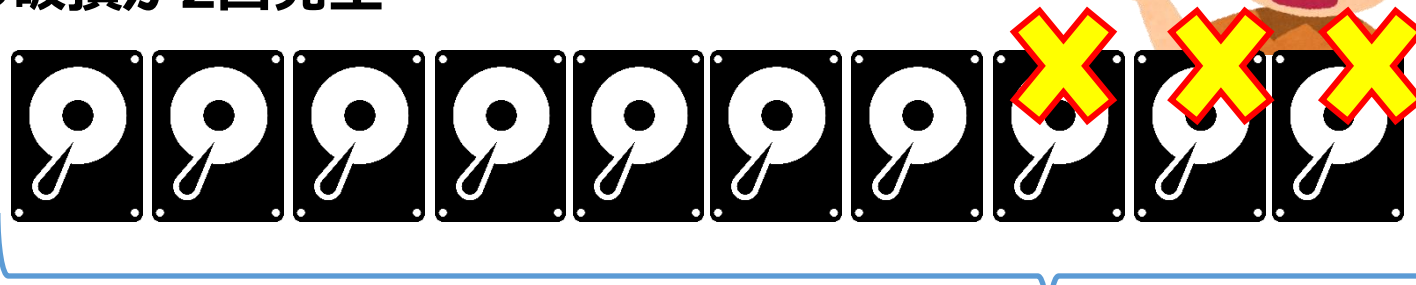
- HDDの交換数は増加中/予防交換実施
  - エラーカウンタを確認してエラーカウントやDefectListが増加しているものを優先的に交換。
  - 一定の評価のためにSMARTなどの一般的なカウンタを見れるようにしてほしい。。。



RAID6破損

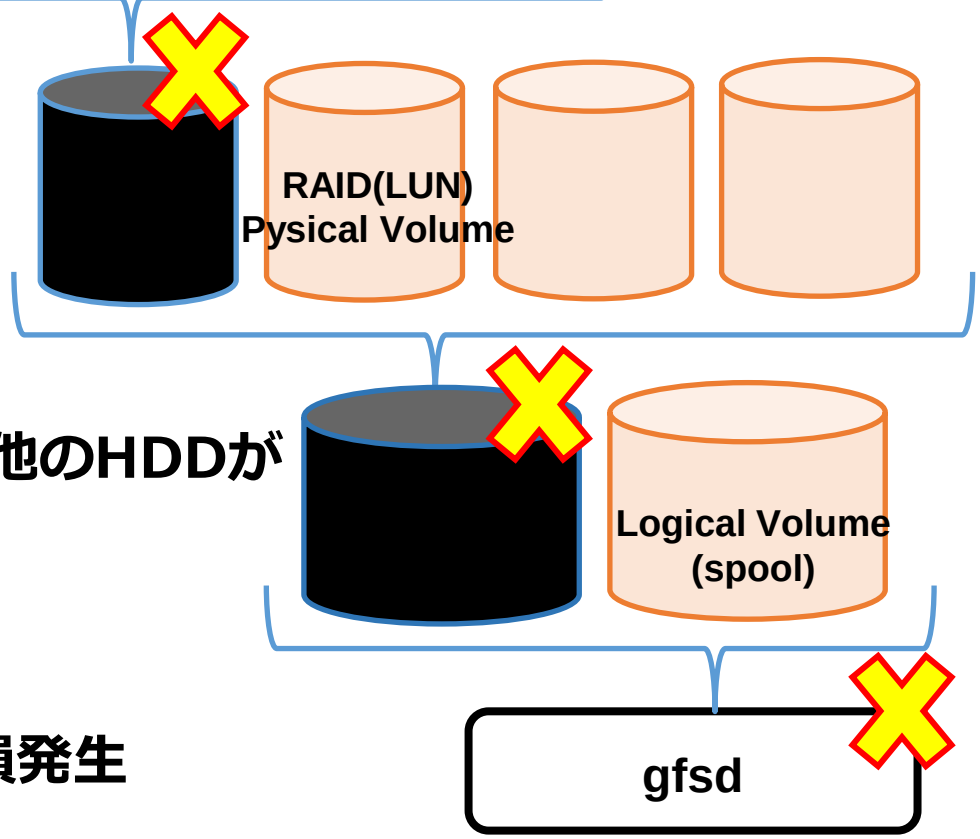


- RAID6(8D2P)の破損が2回発生



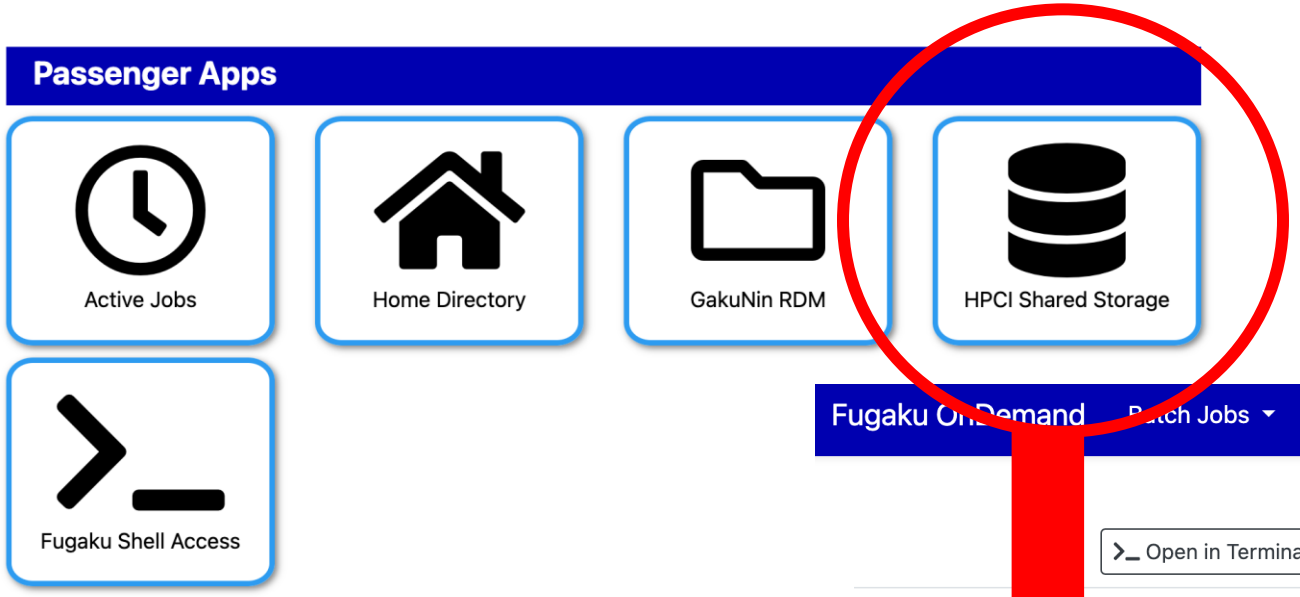
- 2回ともデータロストは発生せず
  → GfarmがWrite Throughになっているため

- 発生パターン
  - 故障によりDisk Timeout
  - Disk Timeoutのラグやキャッシュの問題などで、他のHDDが多数ディスクエラー
  - Partial Rebuildによる復元
  - 復元中の高負荷で故障ディスクが誘発
  - 同一RAIDで3台以上のディスク故障 → RAID6破損発生



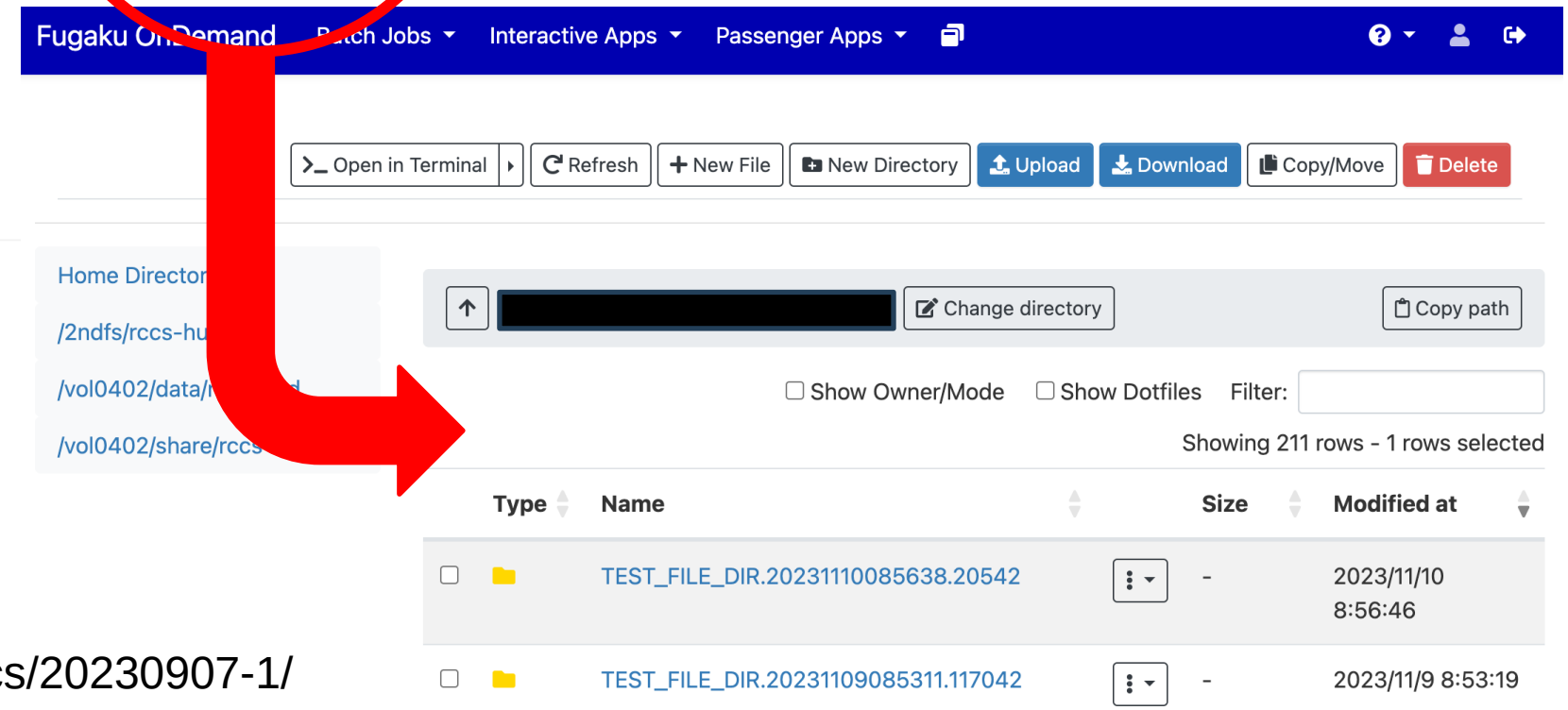
今後のHPCI共用ストレージについて

Fugaku's Open Ondemand Access



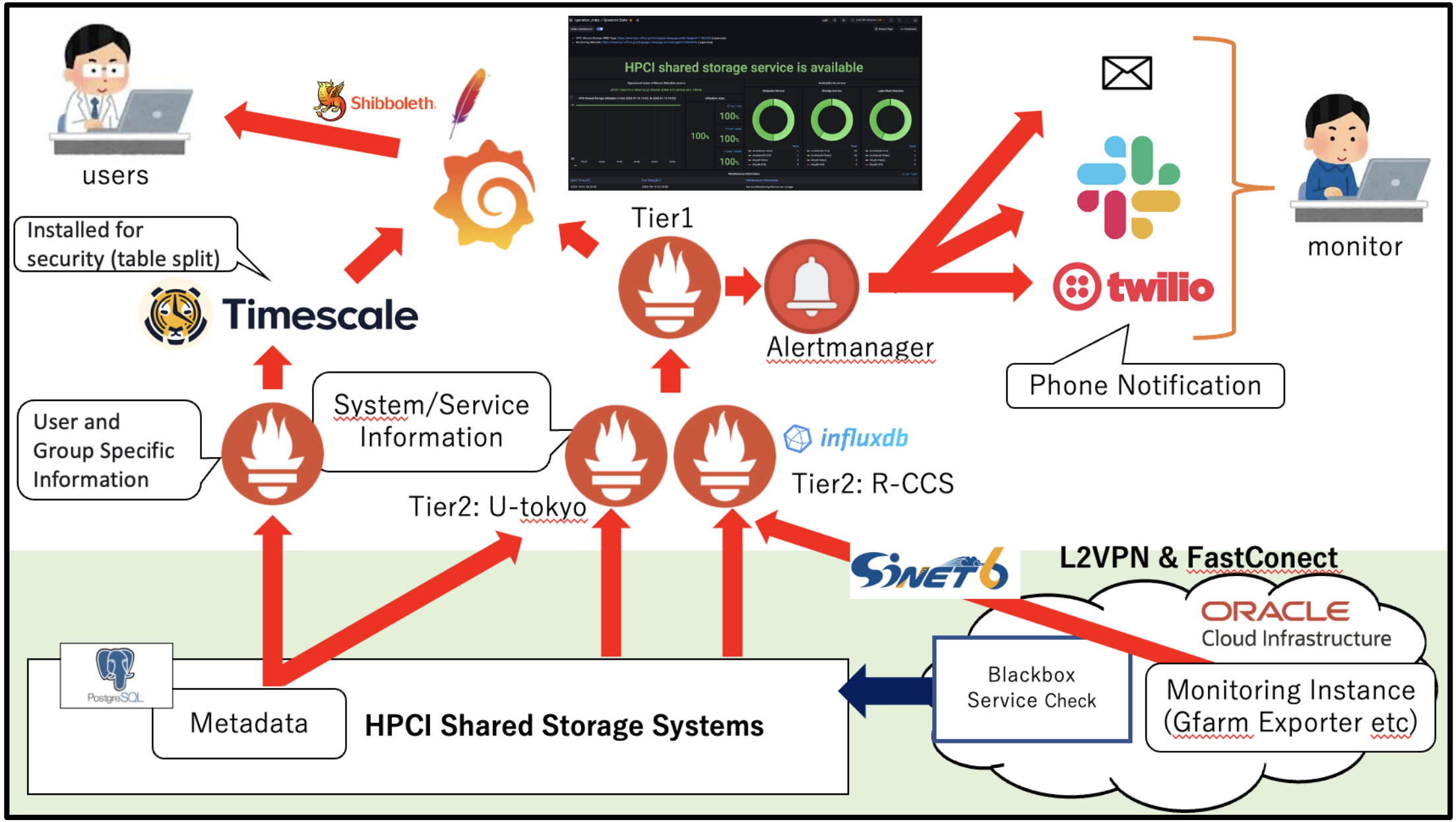
Fugaku's Open Ondemand Service allows GUI access as well.
OAuth Access is available.

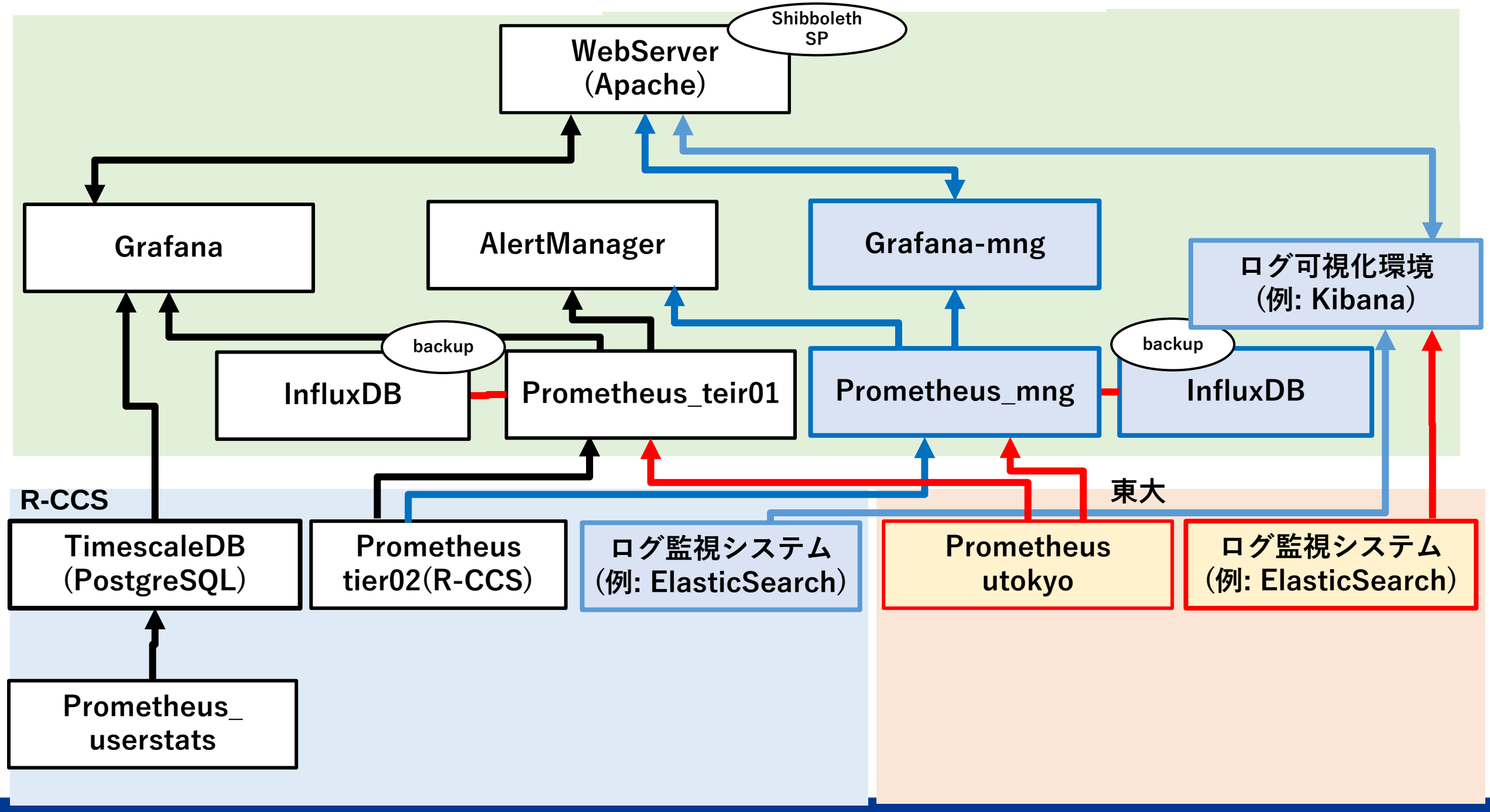
by OnDemand



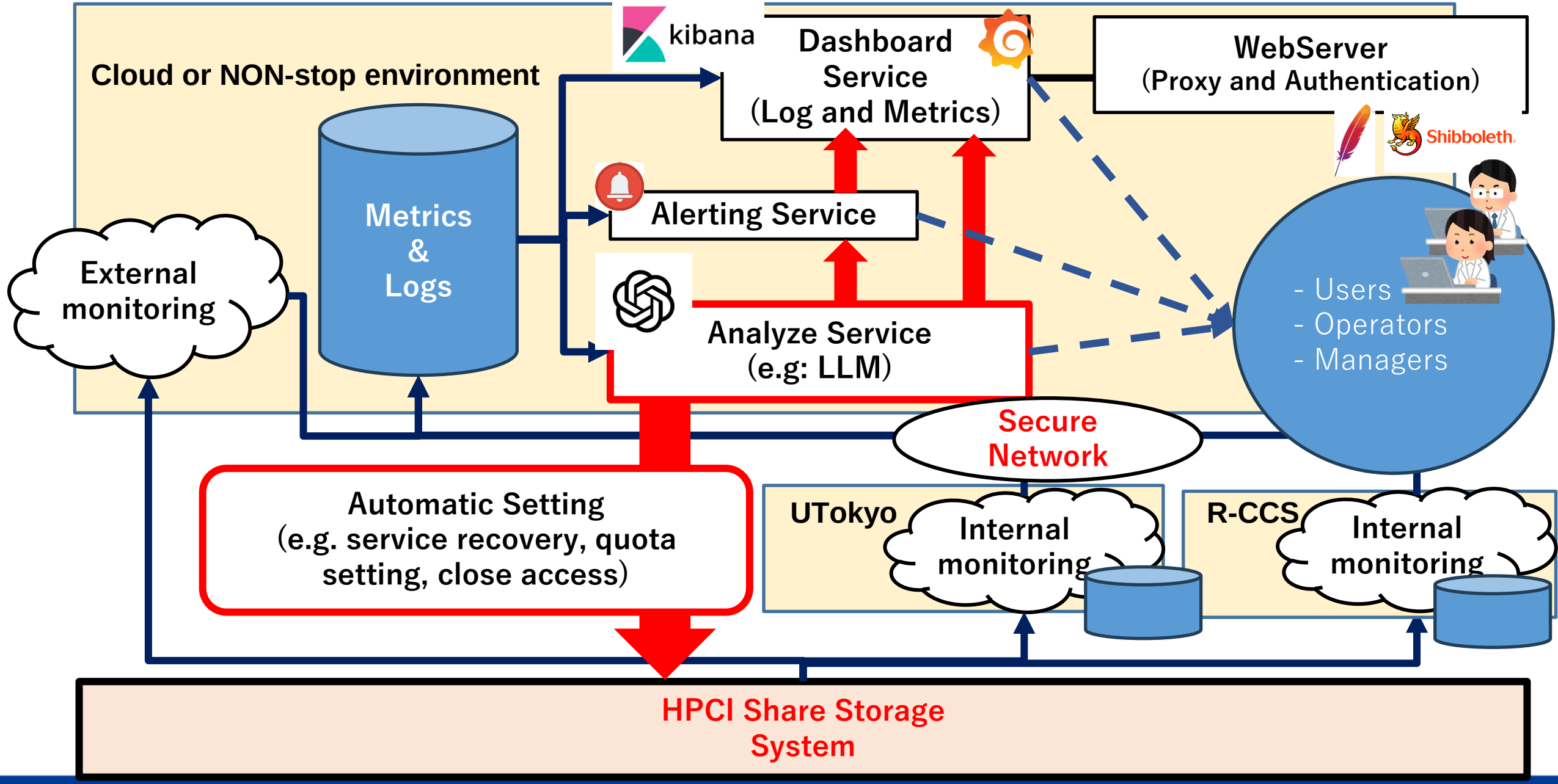
https://www.r-ccs.riken.jp/outreach/topics/20230907-1/

Monitoring


users
Shibboleth
HPCI shared storage service is available
Tier1
monitor
Installed for security (table split)
Timescale
Alertmanager
twilio
Phone Notification
User and Group Specific Information
System/Service Information
influxdb
Tier2: R-CCS
Tier2: U-tokyo
SINET6
L2VPN & FastConect
ORACLE Cloud Infrastructure
Blackbox Service Check
Monitoring Instance (Gfarm Exporter etc)
PostgreSQL
Metadata
HPCI Shared Storage Systems


WebServer
(Apache)
Shibboleth
SP
Grafana
AlertManager
Grafana-mng
ログ可視化環境
(例: Kibana)
backup
InfluxDB
Prometheus_teir01
Prometheus_mng
backup
InfluxDB
R-CCS
東大
TimescaleDB
(PostgreSQL)
Prometheus
tier02(R-CCS)
ログ監視システム
(例: ElasticSearch)
Prometheus
utokyo
ログ監視システム
(例: ElasticSearch)
Prometheus_
userstats

Future Monitoring System
-- Need to more simple monitoring, And automation setting method from analyzing results.

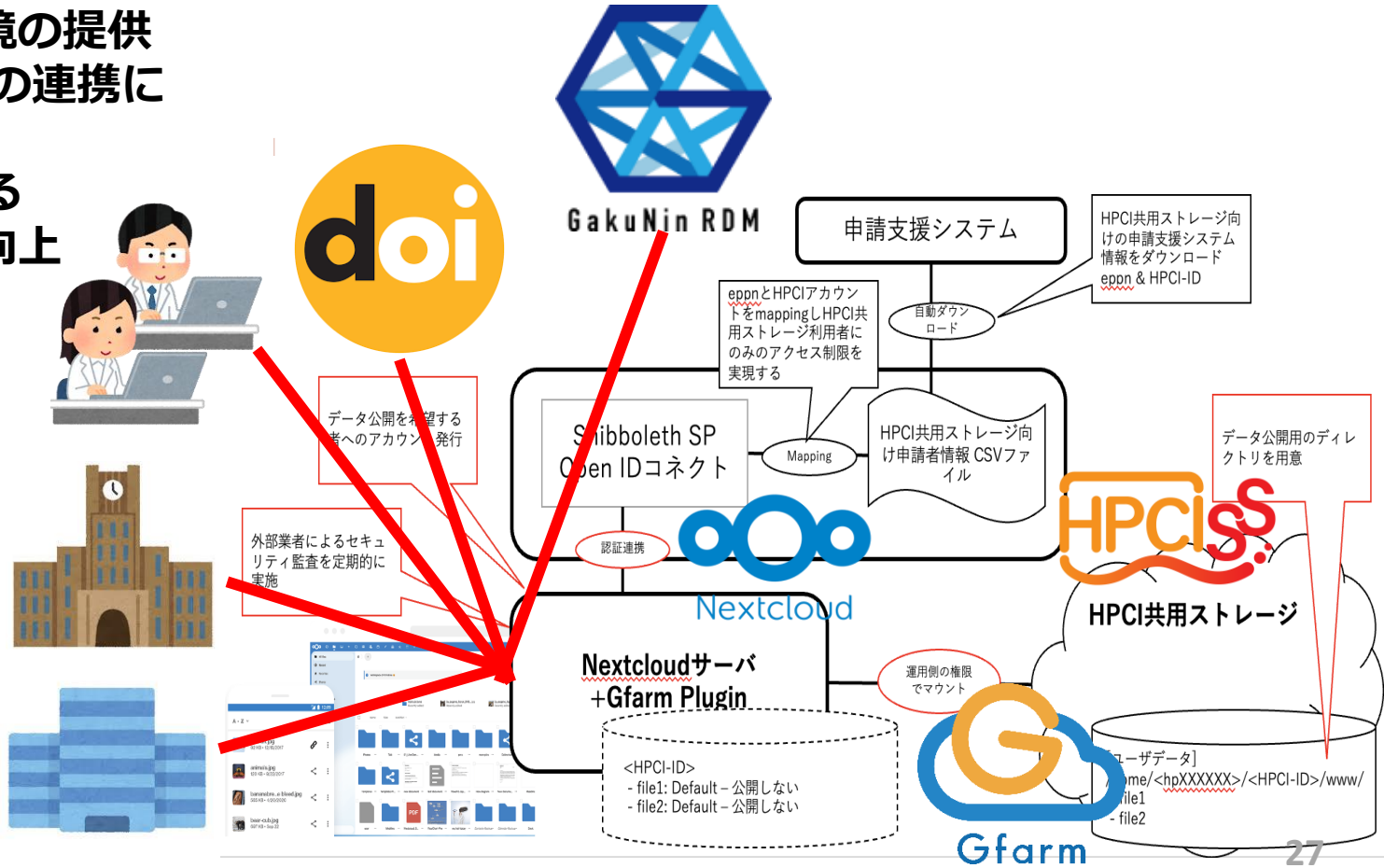


HPCI共用ストレージサービスを利用したデータ公開

HPCI共用ストレージのデータ利活用促進として、データ公開環境の整備を行います。

■ 方針
1. データ公開時の保存期間は原則無期限
2. NextCloud等の簡便なツールを利用した
   セキュアで使いやすいデータ共有・公開環境の提供
3. 他サービス(例: GakuninRDM by NII)との連携に
   よる利便性の向上
4. doi(Digital Object Identifier)付与による
   保存データのトレーサビリティ確保と価値向上

■ 課題
・他のシステムやサービスと連携するための
 プロトコルの追加
・利用者が必要とするソフトウェアと
 Gfarm/HPCI共用ストレージの連携
 (ブロックストレージとしてアクセスが
 必要なものなどの場合、常にマウント
 して貰う必要や、データ連携向けサーバ
 をgfmd/gfsdの近くに置く必要あり)



国内/国際機関とのデータ連携

国際/国内機関との研究連携では、研究データの共有および管理が必須

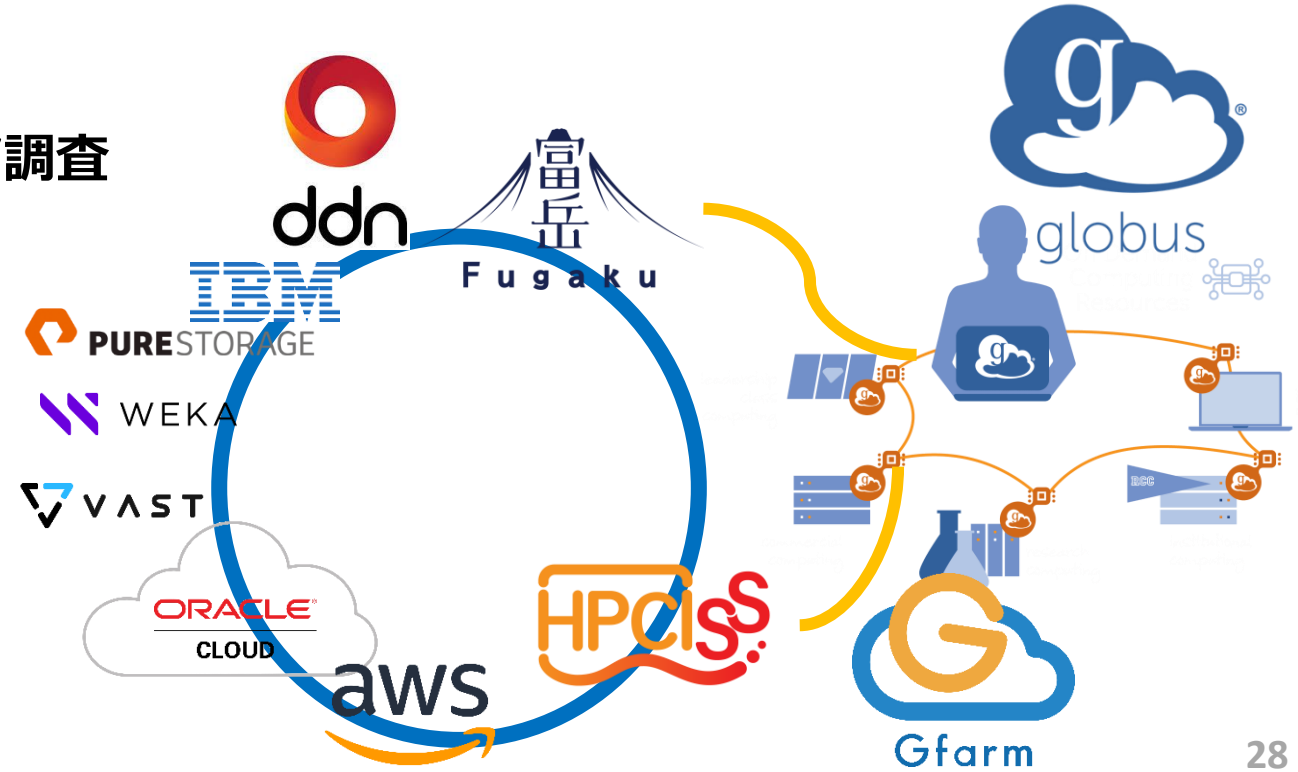
・欧州/欧米では、Globus(※) を用いたデータ共有が盛んになっています。
　※ Function as a Serviceに近いデータ管理/転送、認証などをサポートしたプラットフォームです。
・近状のAI向けストレージやソフトウェアでは、ストレージ自体に主要クラウドへのデータレプリケーションや、同ベンダのストレージであれば遠距離にあっても同一ネームスペースで利用できるようになっている。

■ 方針

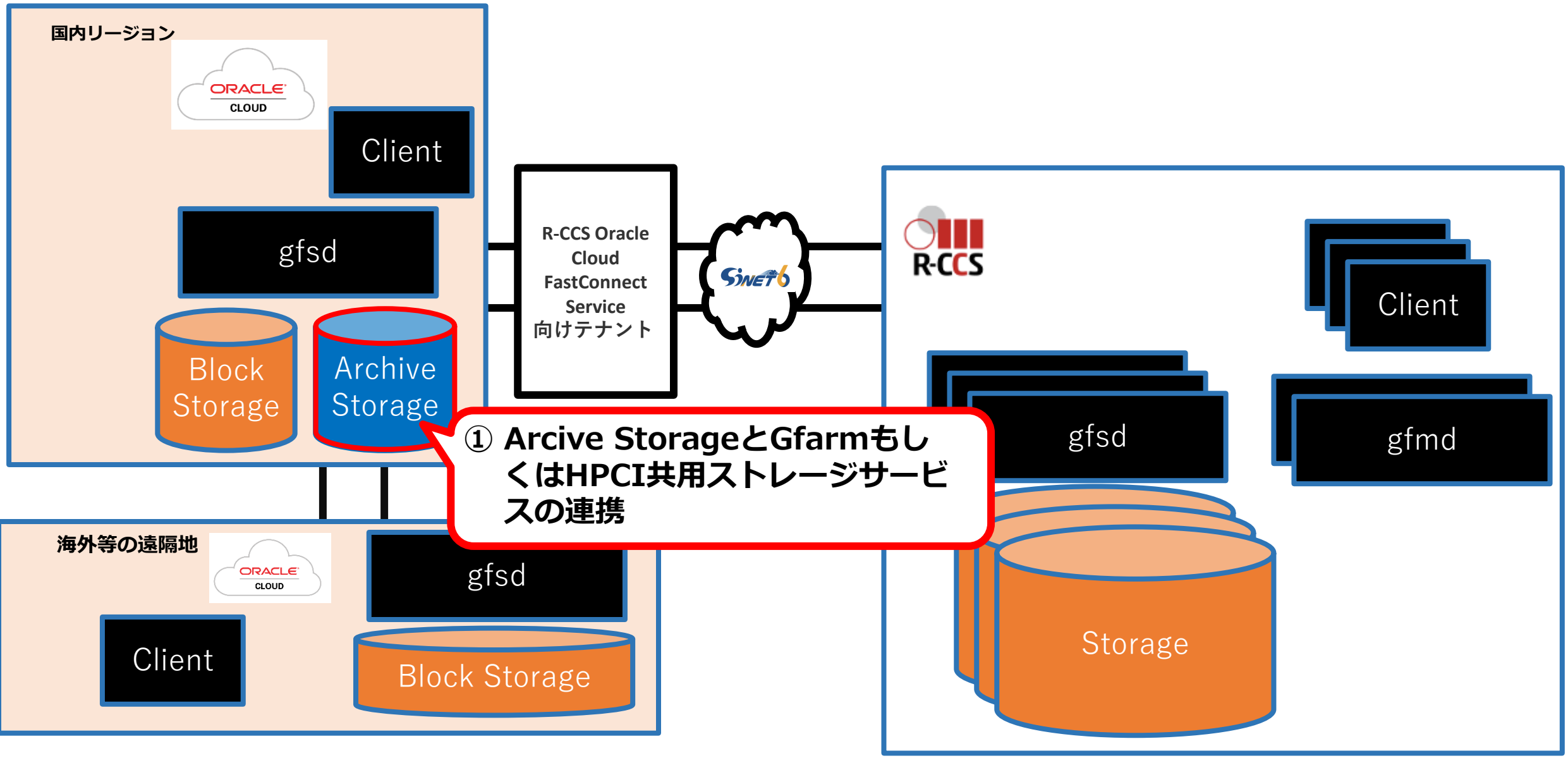
(1) 富岳/HPCIへのGlobus導入試験
(2) 実サービスとの連携/認証検討
(3) 主要AIストレージ/サービス/ソフトウェア調査
(4) 連携手法の整備、開発

■ 課題

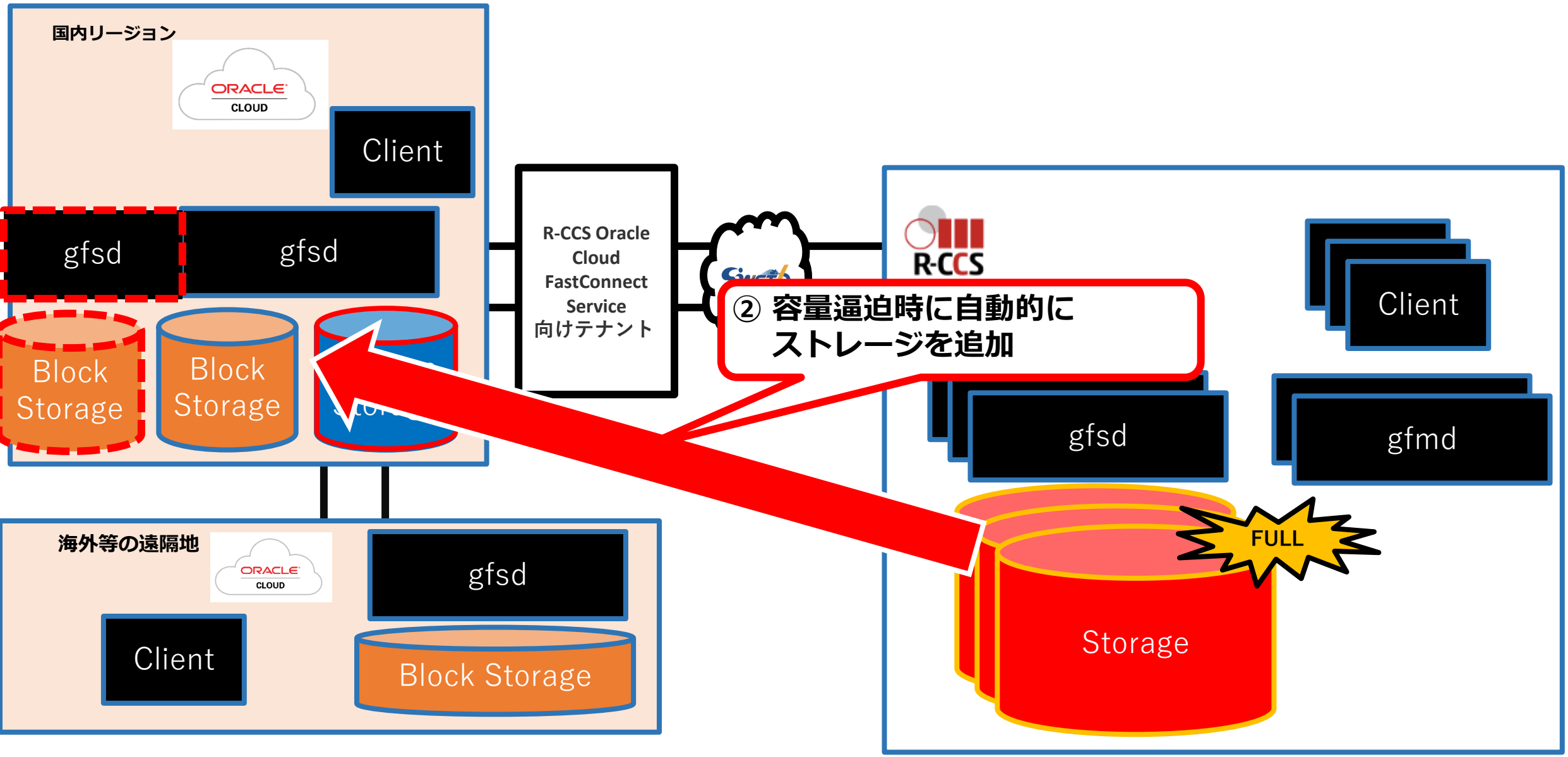
・認証連携
・ファイルシステム連携
・メタデータ管理およびメタデータ付与方法の検討



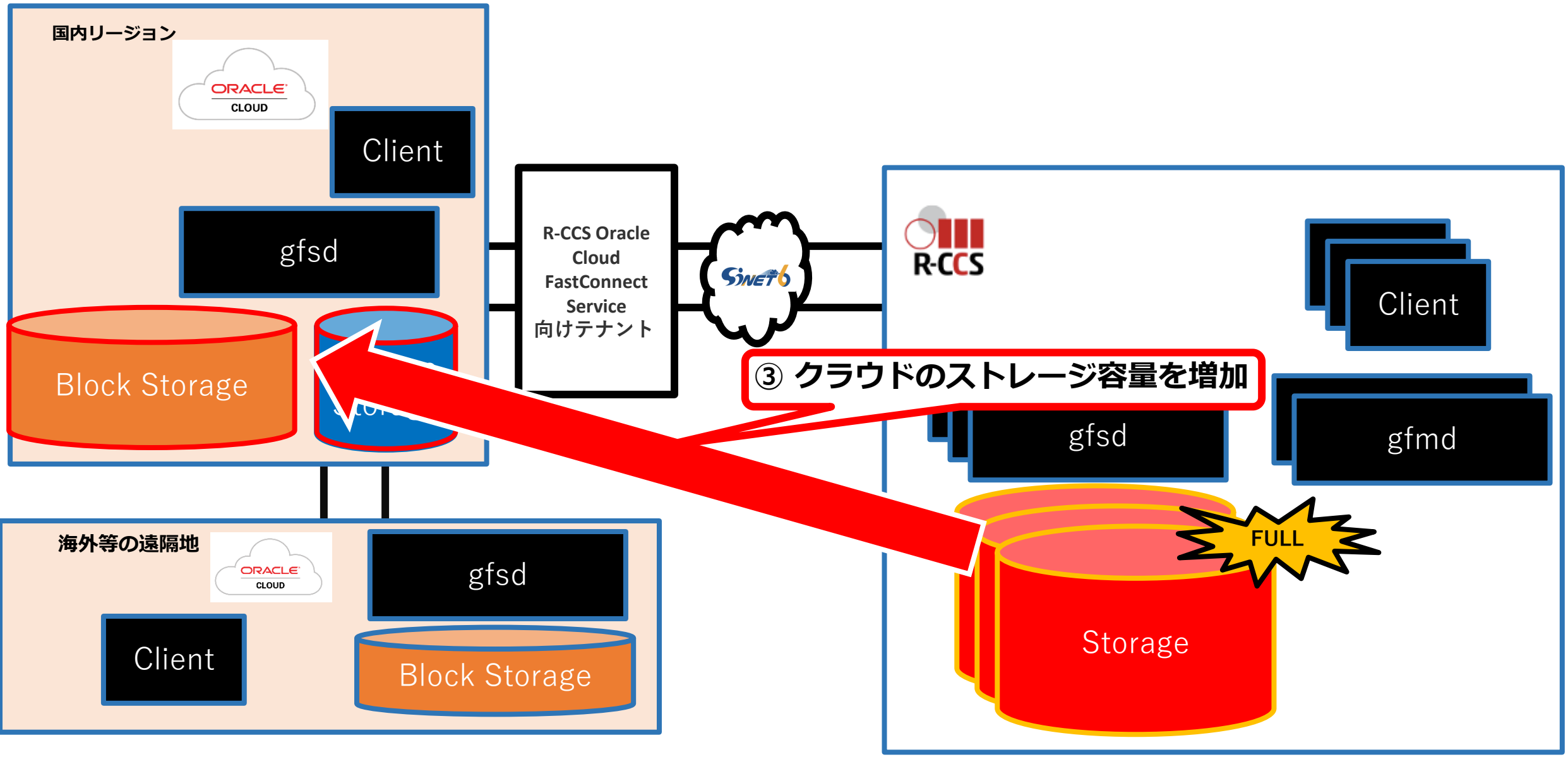
クラウド連携/利用状況による可変型システムの実現



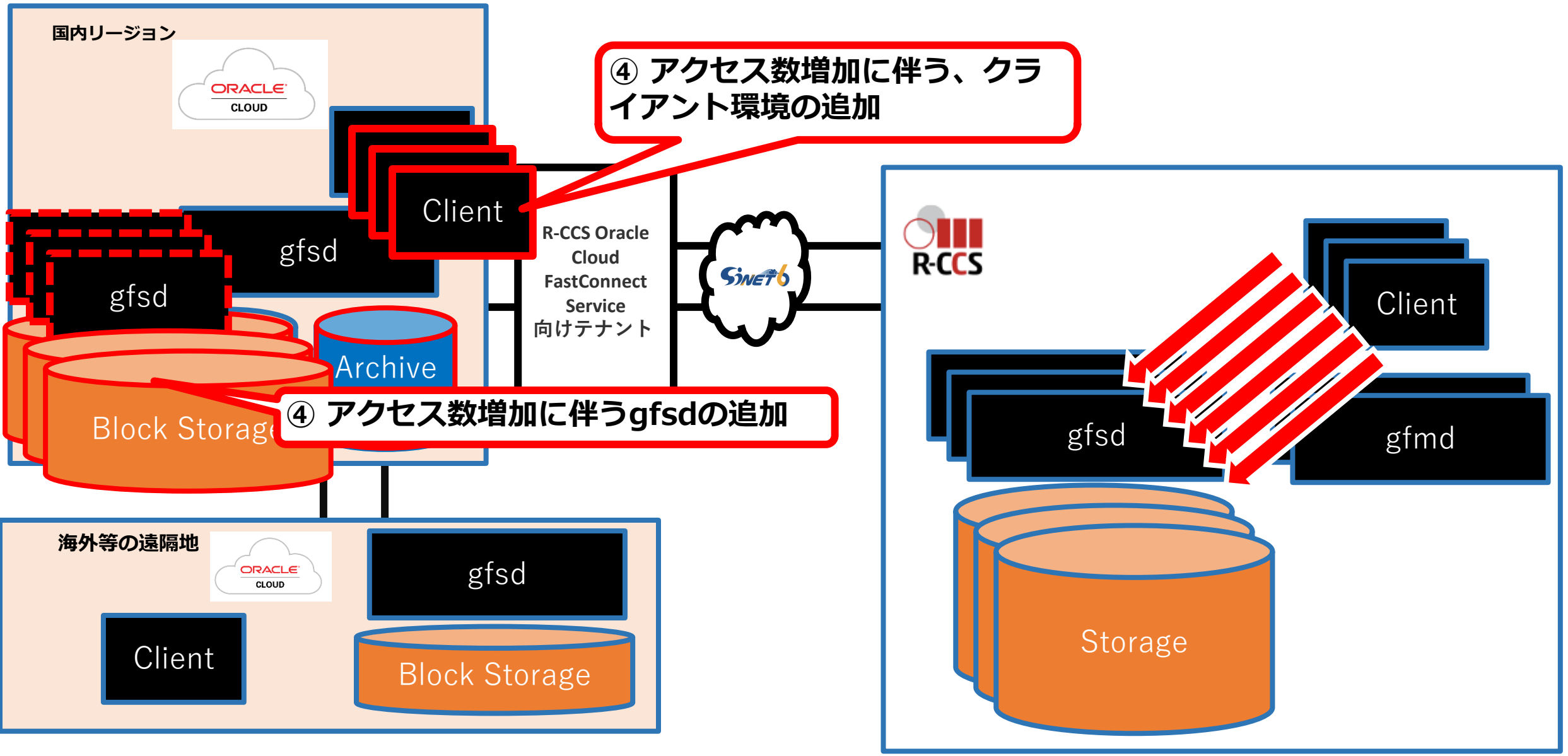
クラウド連携/利用状況による可変型システムの実現



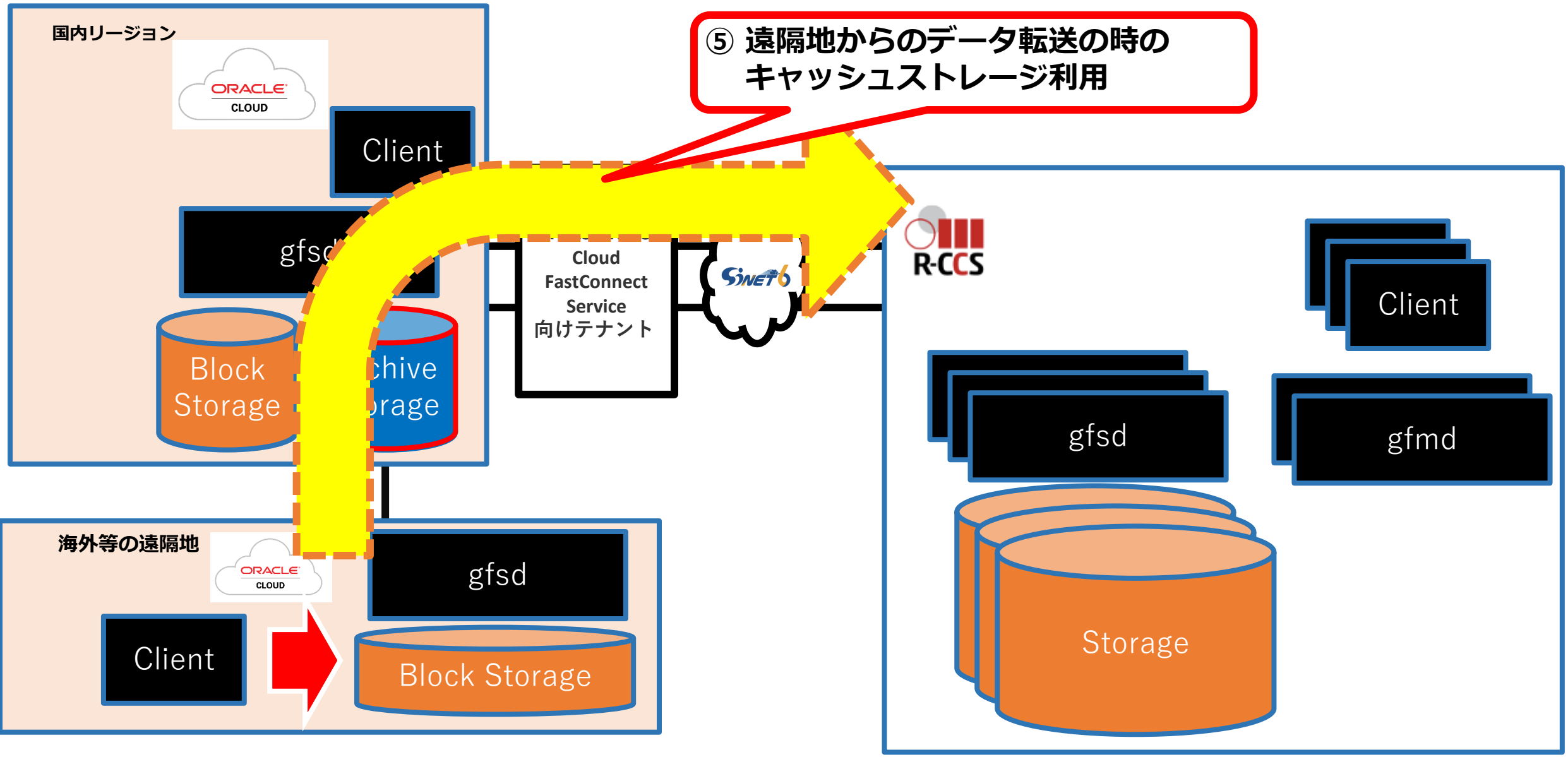
クラウド連携/利用状況による可変型システムの実現



クラウド連携/利用状況による可変型システムの実現

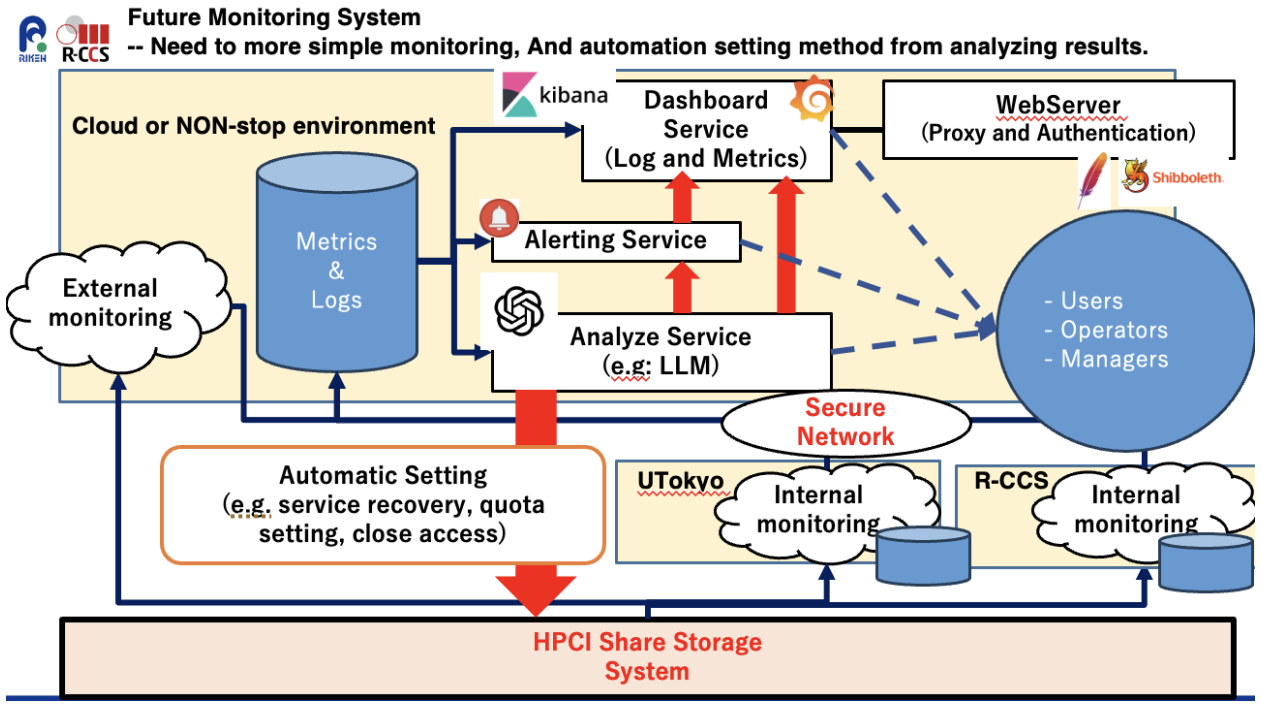


クラウド連携/利用状況による可変型システムの実現



LLM連携

- 生成AI(LLM)やDeep Learningと連携して、より高度な運用管理の実現
  - 障害アラートへの調査アシスト(新規メンバ向けのAI学習)
  - モニタリング環境への分析環境の追加やLLMによるグラフ等の出力
  - システム設定などをAIを用いて自動化・効率化



まとめみたいなもの

- HPCI共用ストレージは第三期システムとして継続
  - もちろんGfarmが採用
  - R-CCSでは導入作業開始 2025年度早々のサービス投入を目指す

- HPCI共用ストレージは他システムとの連携が求められている
  - オープンサイエンスやデータ公開への舵切り
  - Gfarm + @ の環境整備が今後課題

- いっしょにHPCI共用ストレージを運用・改善・開発
  いただける方を募集しております!!

ありがとうございました。