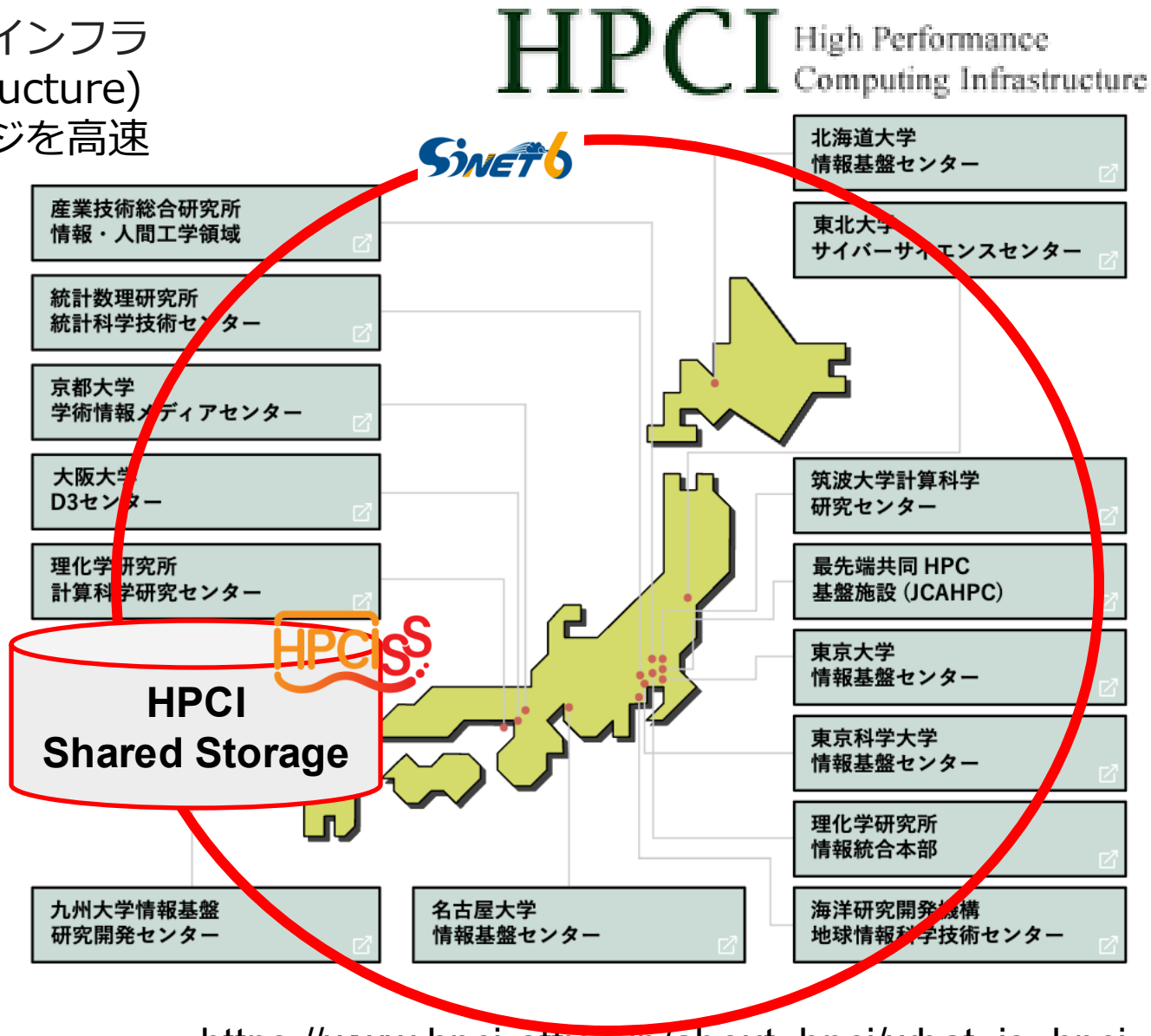
HPCI共用ストレージ
西拠点第３期システムおよび取組みについて

Gfarm Symposium FY2025
2025/09/19

RIKEN R-CCS
- Hidetomo Kaneyama
- Hiroshi harada

What is HPCI in Japan ?

革新的ハイパフォーマンス・コンピューティング・インフラ
(HPCI = **H**igh **P**erformance **C**omputing **I**nfrastructure)
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盤です。



Provide to service

- 課題選定
- アカウント管理
- シングルサインオンサポート
  - Web: Shibboleth
  - システムログイン: Oauth認証(& hpcissh)

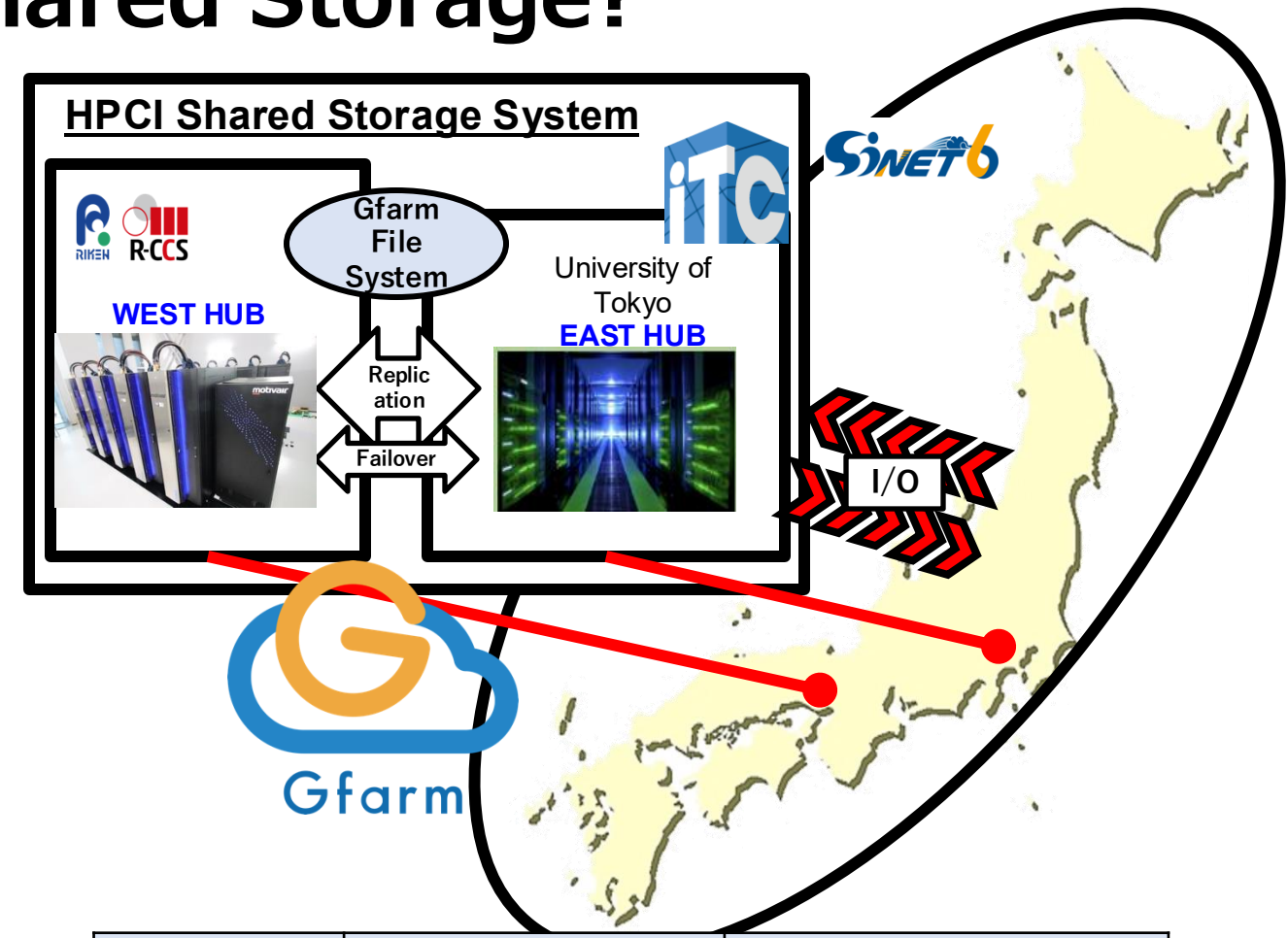
Provide to Infrastructure

- ヘルプデスクサポート - RIST
- **ネットワークストレージ**
  → HPCI Shared Storage System

https://www.hpci-office.jp/about_hpci/what_is_hpci

What is HPCI Shared Storage?

- **目的:**
  - スーパコンピュータ間(計算資源)間のデータ共有
  - 研究データの保存
  - 研究データの公開(公的データの利活用) [New]
- **運用機関**
  - 東京大学
  - 理化学研究所(R-CCS)
- **基幹ソフトウェア**
  - Gfarm File system
    - https://github.com/oss-tsukuba/gfarm
- **特徴**
  - HPCI資源のクライアントから並列転送によるデータI/Oが可能
  - ２拠点間でのデータレプリケーションによるディザスタリカバリ
  - 高可用性(99%の稼働率 – 運用停止/アクセス不可は1%未満)
  - 高速転送可能なネットワークストレージ(200Gbps以上での接続)

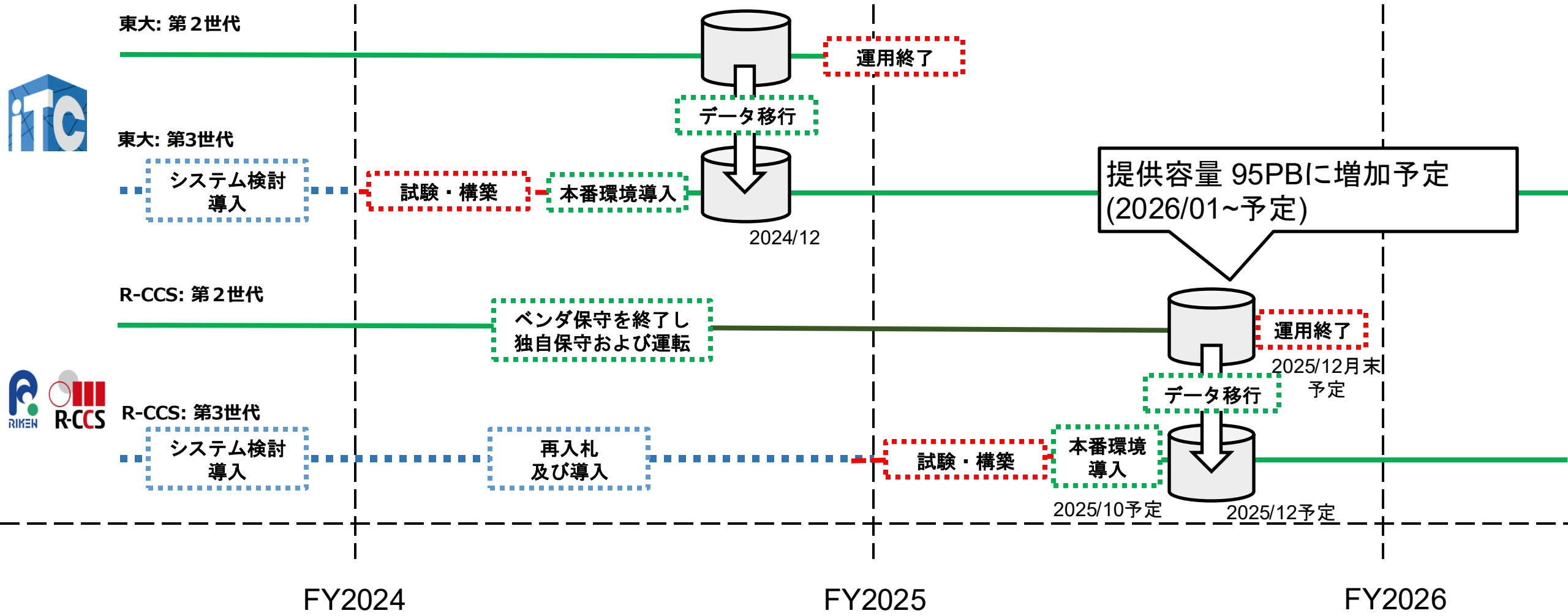


Gen	Periods	Capacity(Logical)
1st	FY2012 – FY2018	10PB → 15PB
2nd	FY2018 - FY2024	45PB → 50PB
3rd	FY2025 -	95PB+

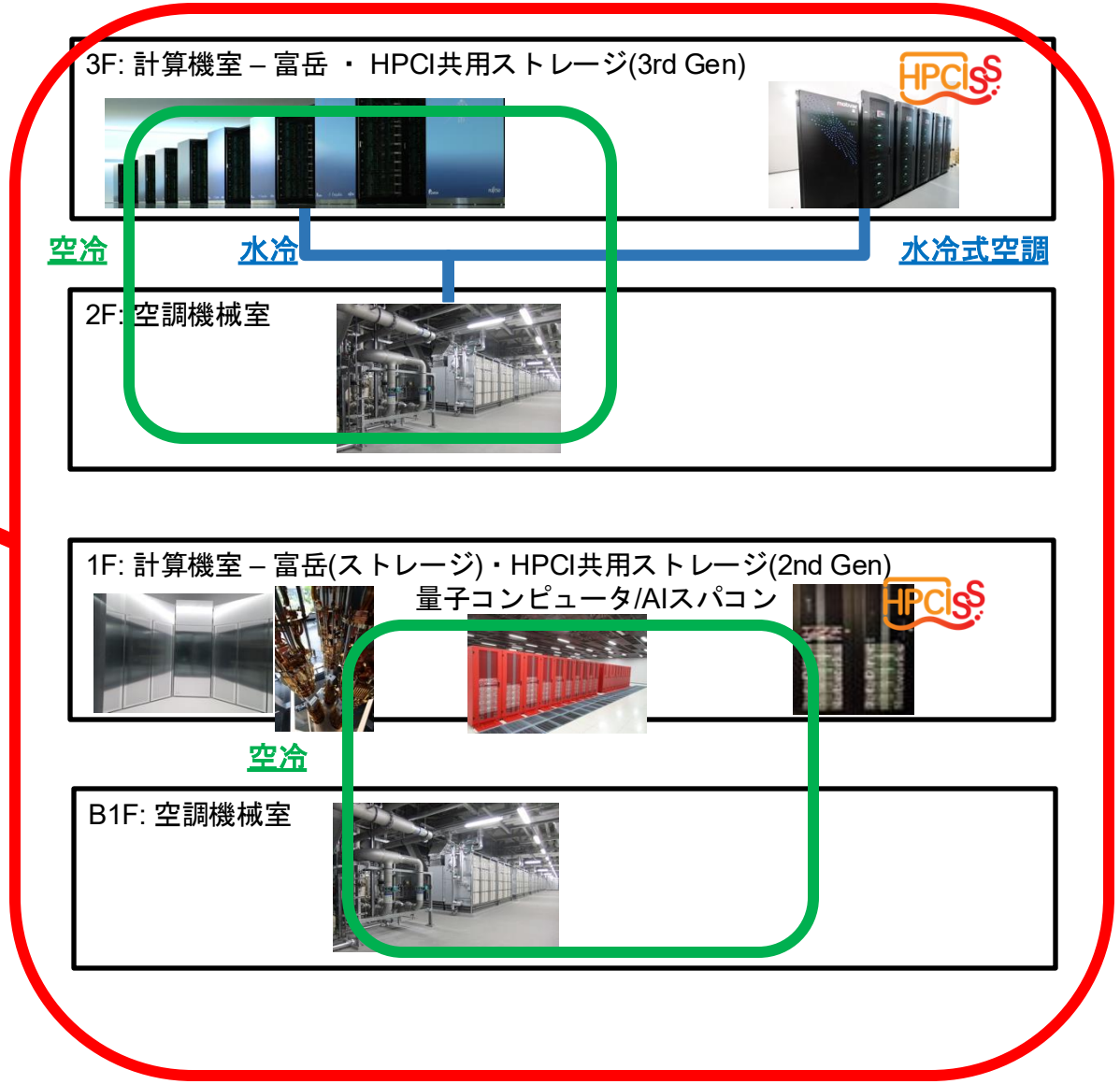
https://www.hpci-office.jp/info/pages/viewpage.action?pageId=111380786

HPCI共用ストレージ 第3期システム – スケジュール

運用の切り替えはサービス無停止を予定



R-CCS HPCI共用ストレージ 第3期システム



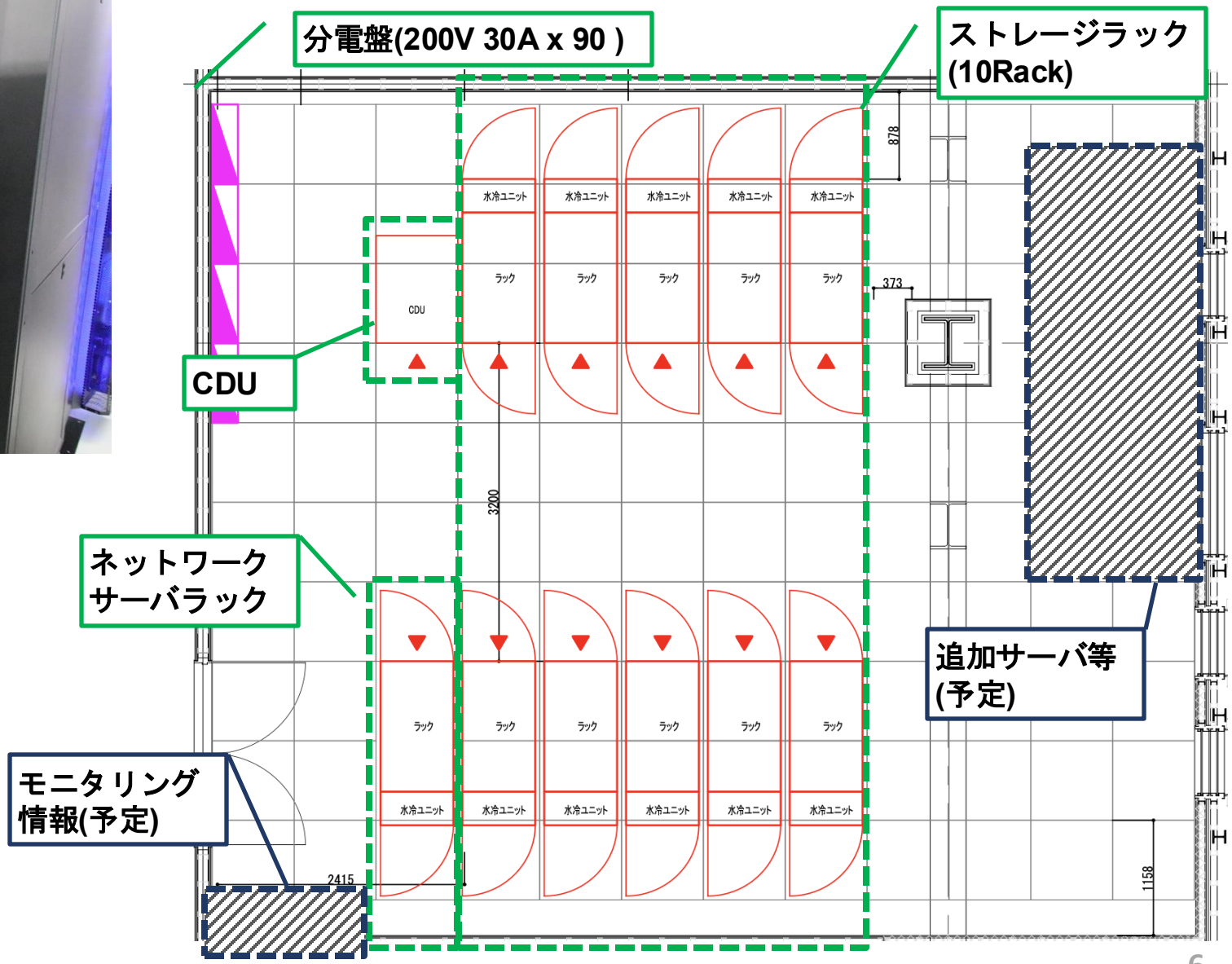
HPCI共用ストレージシステム西拠点はR-CCS計算機棟に設置されています。

- 第2期システムは計算機棟1階でしたが、第3期システムは3階に構築
- ネットワーク等は既存のものを流用せず。
- 空調機も合わせて調達。

R-CCS HPCI共用ストレージ 第3期システム

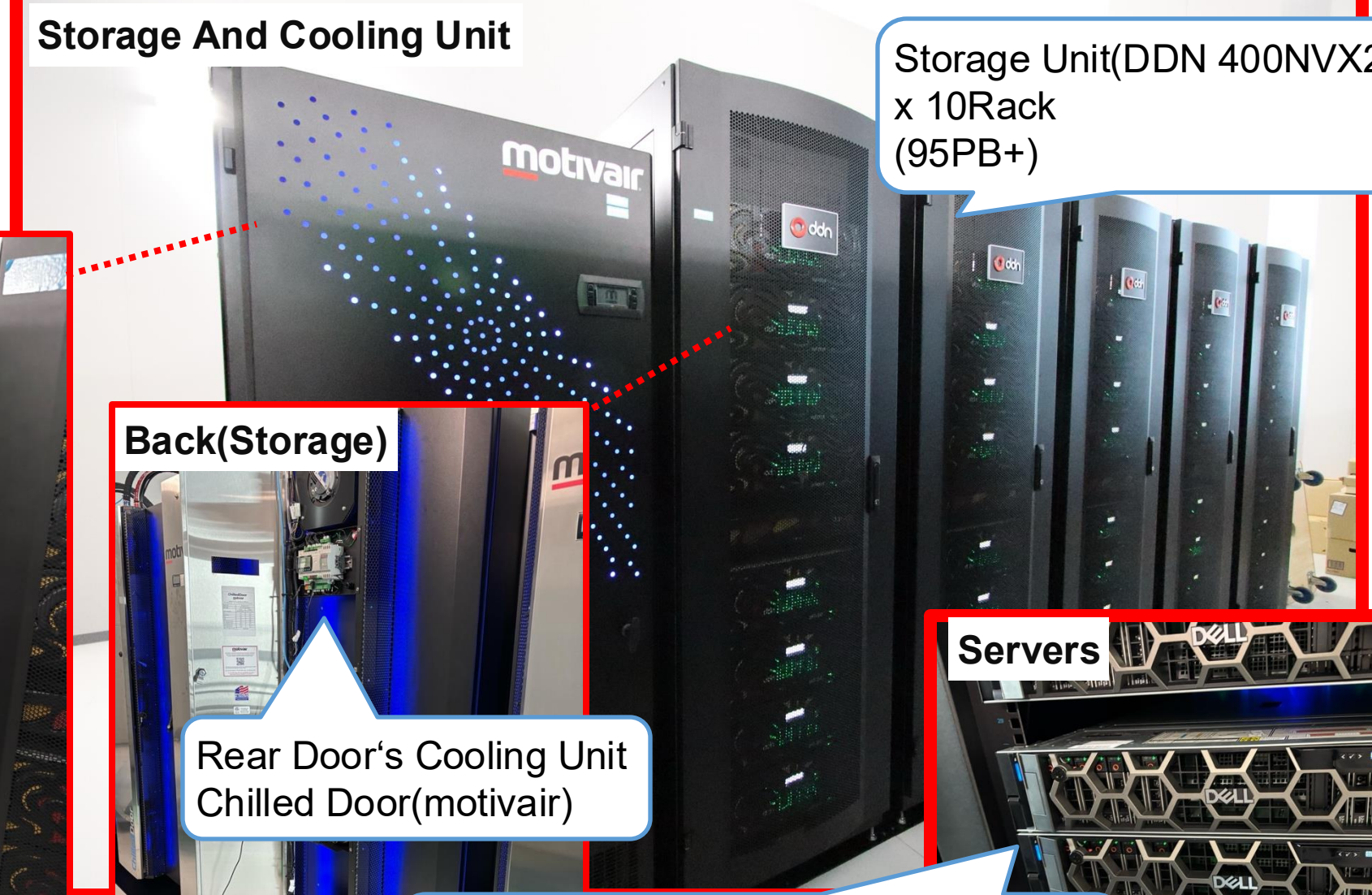
前面

背面

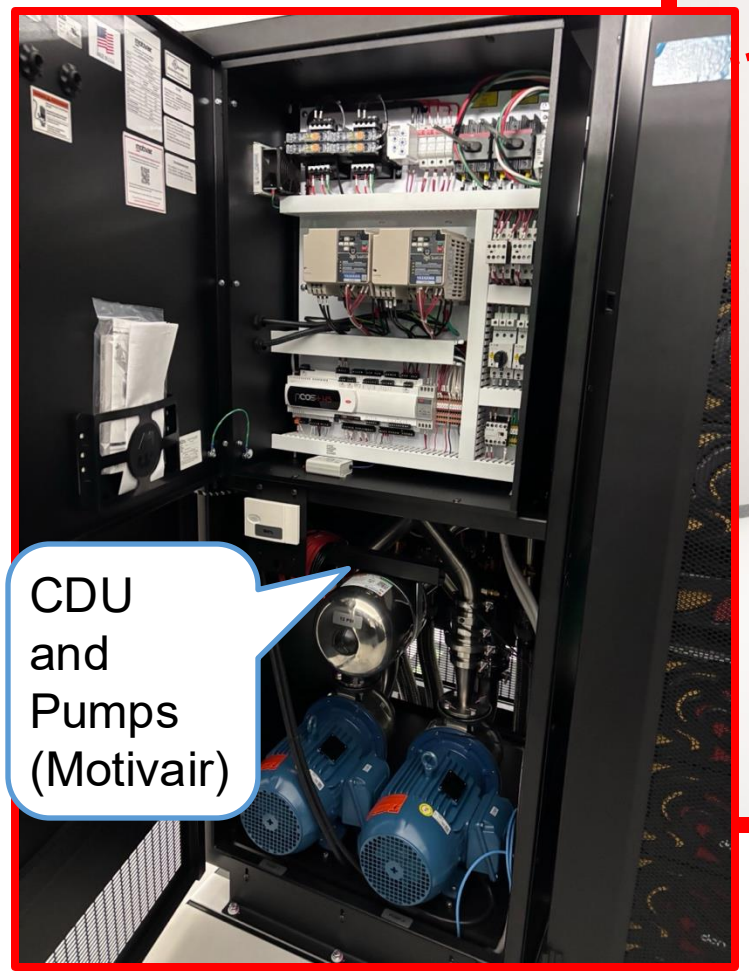


3rd floor(HPCI storage/server area)

Storage And Cooling Unit



Storage Unit(DDN 400NVX2)
x 10Rack
(95PB+)



CDU
and
Pumps
(Motivair)

Back(Storage)

Rear Door's Cooling Unit
Chilled Door(motivair)

Servers

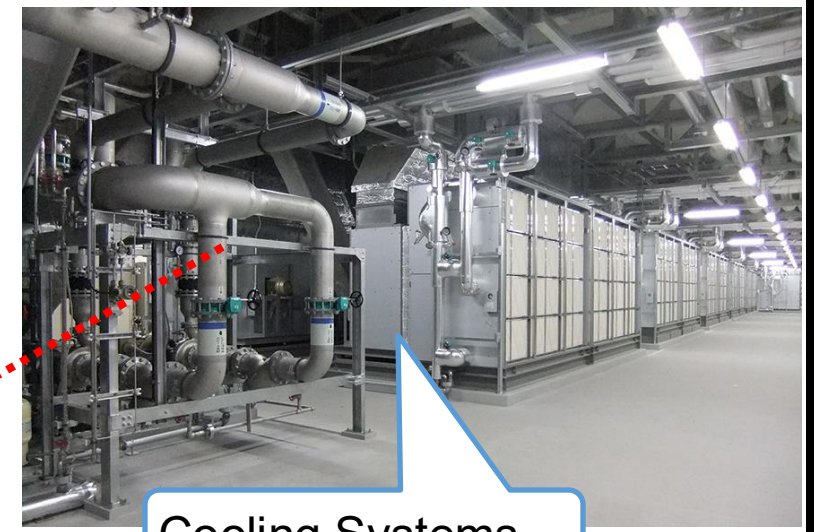
Metadata Servers (2set)
Management Servers (5set)
DELL 760R

3rd floor(HPCI storage/server area)

耐荷増強工事を実施
水冷式空調(リアドア)向けにパイプ整備

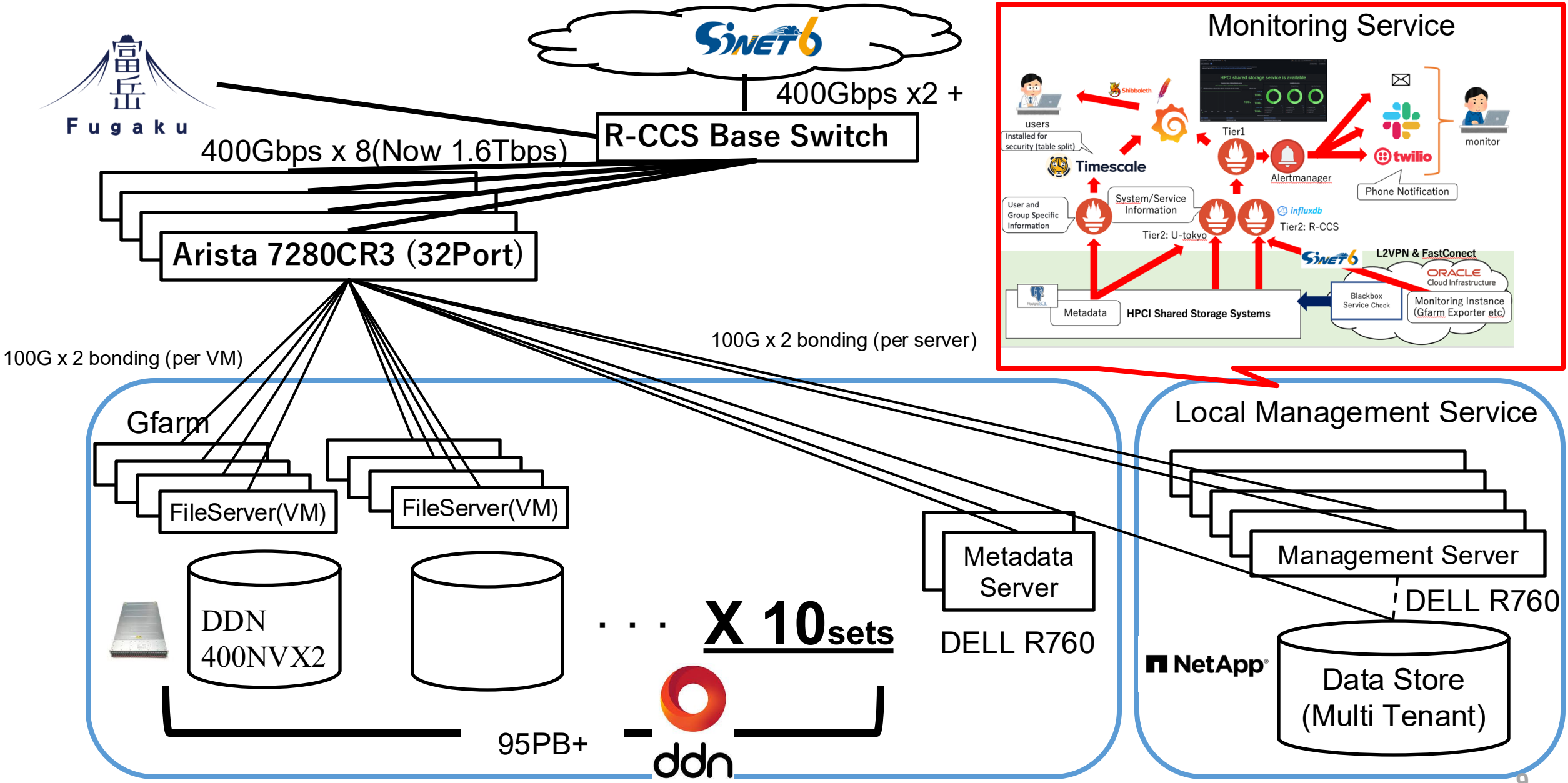
2nd floor(cooling area/water-pomp)

水冷用パイプ



Cooling Systems

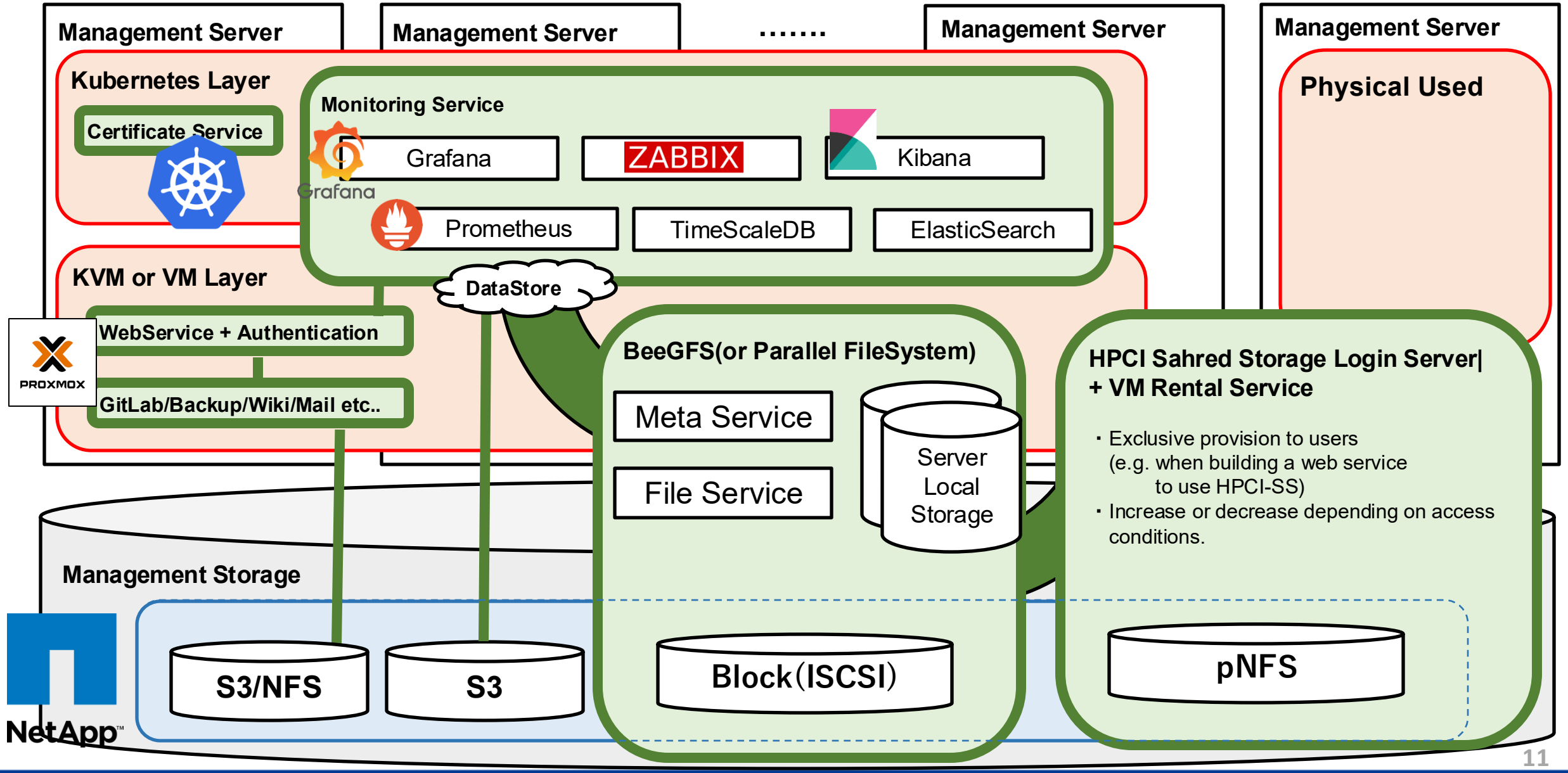
System Configuration(3rd Generation/R-CCS)



Performance

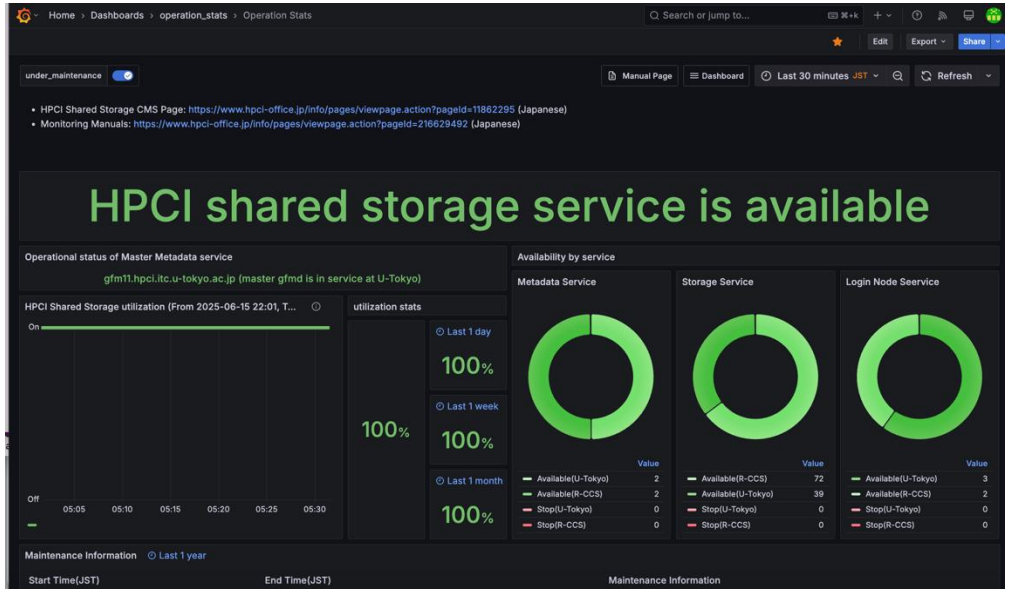
Generation	2nd (2018 ~ 2024)		3rd (2025~)	Remarks
Metadata System	Dell R730xd		Dell R760	Intel Xeon Gold 6444Y x2(CPU)
Metadata Memory	768 GB		2048 GB	
Metadata I/O	15000+ (count)		41000+ (count)	time dd if=/dev/zero of=path_to_meta bs=512 count=1000000 oflag=dsync
System(Storage)	(1) CMS Hyper STOR Flex	(2) DDN SFA14K	DDN SFA400NVX2E (VM)	
System(FIleServer)	DELL R730	Supermicro		Fileserver is Virtual Machine In 400NVX2E
Storage	HDD(10TB)	HDD(12TB)	HDD(22TB)	WD
Space(ALL/Logical)	45PB → 50PB		95PB+	Added Capacity in the future
Space(R-CCS/Phisical)	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): 320GB/sec+ (MAX Read & Write) 1set: 32GB/sec + (write/read)	For Fugaku: 200Gbps+ For Base Switch: 1.6Tbps
Network(R-CCS)	100Gbps → 200Gbps → 400Gbps		400Gbps x2 (Dual)	Access to SINET6 is redundant

3rd gen systems Local Management Service (constructing now)



Monitoring Views

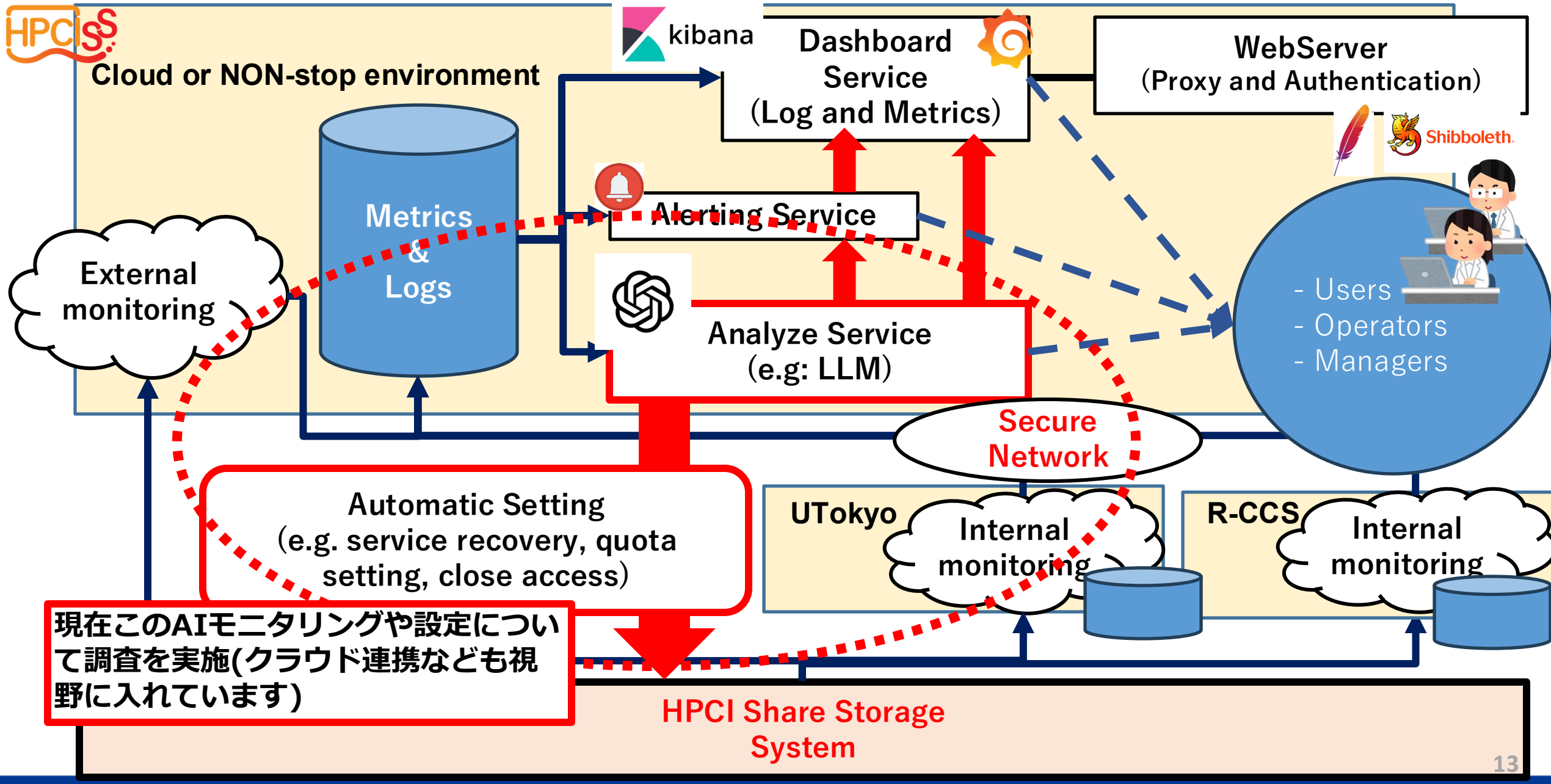
- Fault/Operating Status
- Metadata/File Server Status
- Network
- Storage Capacity
- Capacity Information by User/Group
- Node/SNMP/IPMI/Blackbox Exporters metrics
- Gfarm Exporter(original)
  - Detailed I/O Status and Service Information
- SMART Check Exporter(original)



Future Work:

- (1) Update monitoring using tools such as eBPF, gNMI…
- (2) Introduction of MPC-agent and ML/AI integration

Future Monitoring System Overview



富岳Ondemand

- **R-CCS 中尾技師がHPCI共用ストレージ向け(Gfarm Oauth認証向け)の環境を整備くださいました！！**
  - **HPCI共用ストレージを容易にWebでマウント→アクセスできる環境！**

「富岳」Open OnDemandにおけるHPCI共用ストレージとのデータ転送アプリケーションの開発

理化学研究所 計算科学研究センター（R-CCS）先端運用技術ユニットの中尾昌広技師と利用環境技術ユニットの金山秀智専門技術員は、「富岳」Open OnDemand※[1]上で動作するHPCI（High Performance Computing Infrastructure）共用ストレージ※[2]に対するデータ転送アプリケーションを開発しました。このアプリケーションにより、「富岳」とHPCI共用ストレージとの間のデータ転送をWebブラウザ上で簡単に行えるようになりました。

※[1]「富岳」Open OnDemandについては、Open OnDemandウェブサイト（英語） ならびに「「富岳」Open OnDemand の提供を開始～Webブラウザで「富岳」の操作が可能に～」をご覧ください。

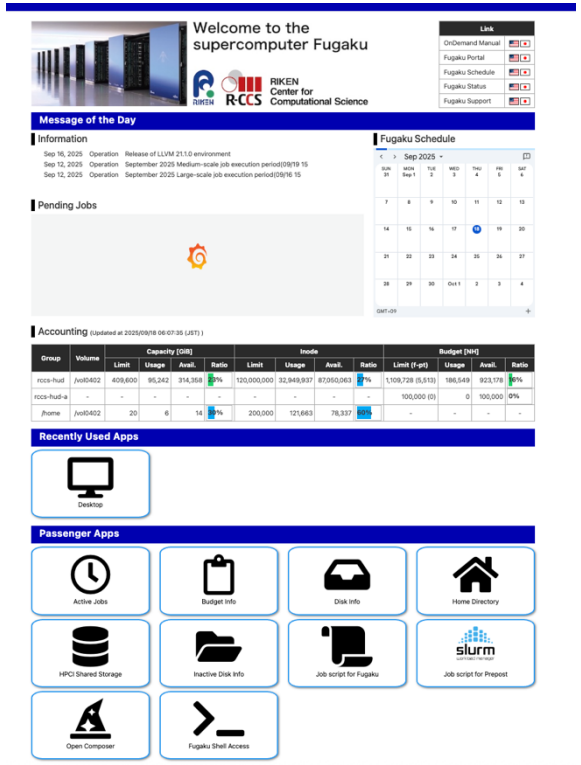
※[2]共用ストレージ（HPCIポータルサイト）

R-CCSが運用する「富岳」は、世界トップレベルのスーパーコンピュータであり、幅広い研究に活用されています。また、HPCIが運用するHPCI共用ストレージは、研究者が研究データや関連資料を管理・共有するための信頼性の高い研究データ管理サービスです。そのファイルシステムにはGfarm※[3]が用いられています。

※[3] Gfarmについては、NPO Tsukuba OSS Technical Support Center（Githubウェブサイト） 、ならびにGfarmファイルシステム（ossTsukubaウェブサイト） をご覧ください。

今回開発したデータ転送アプリケーションは、「富岳」で運用されているWebポータルOpen OnDemand上で利用することができます（図1、図2）。「富岳」とHPCI共用ストレージとの間のデータ転送をWebブラウザから行うことができるため、研究者は特別なソフトウェアのインストールや設定を行うことなく、データを直接転送することができます。本アプリケーションは「富岳」にアカウントをお持ちの方で、HPCI共用ストレージにデータ領域をお持ちの方が利用できます。「富岳」Open OnDemandにおける利用マニュアルは※[4]を参照ください。なお、本件の内容は、2023年9月8日に開催されるGfarmシンポジウム と2023年9月26・27日に開催される第191回HPC研究発表会 で発表が行われます。

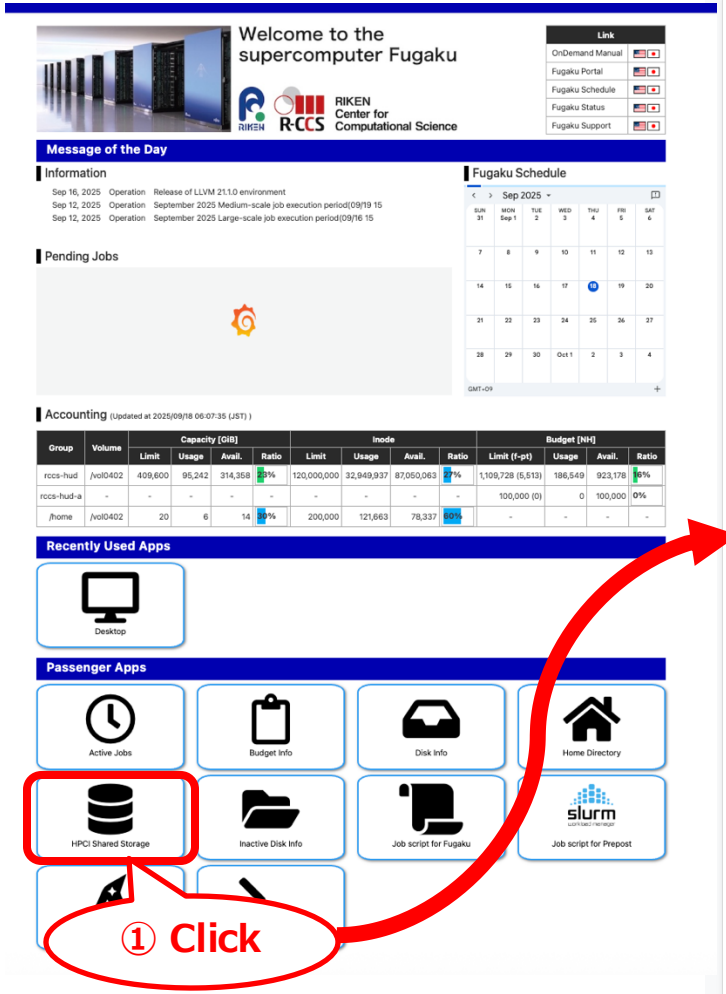
※[4]「富岳」Open OnDemandにおける利用マニュアルについては、Open OnDemandの利用方法（R-CCS Github） をご覧ください。

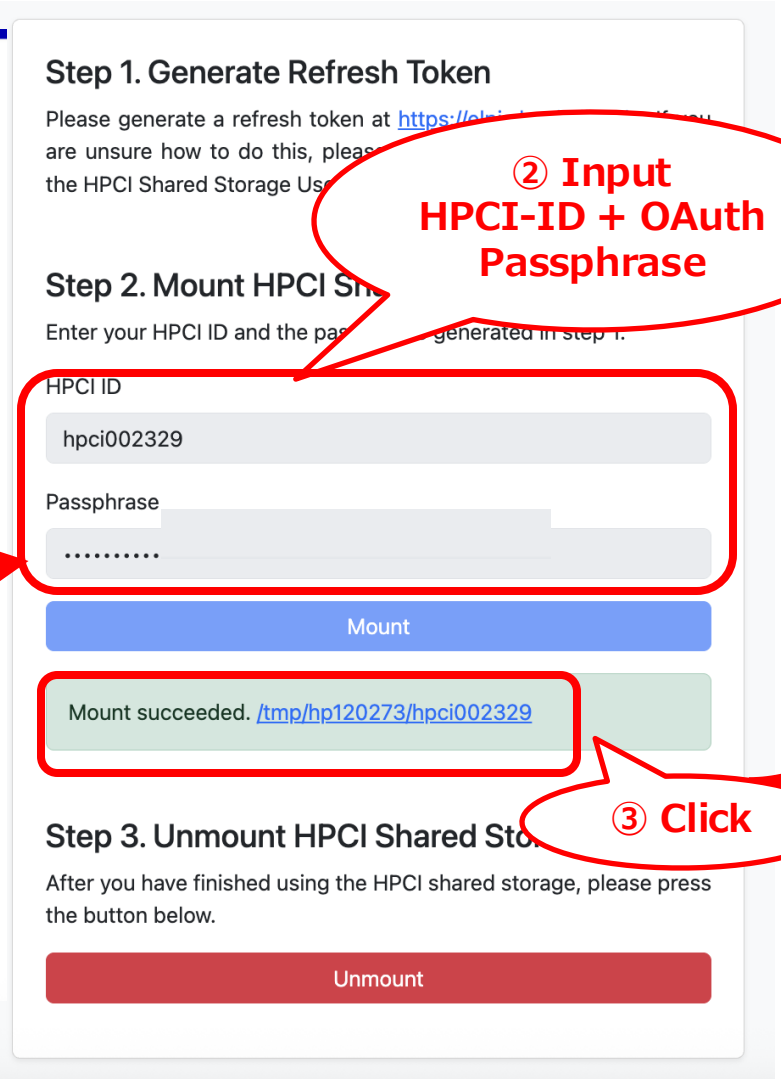


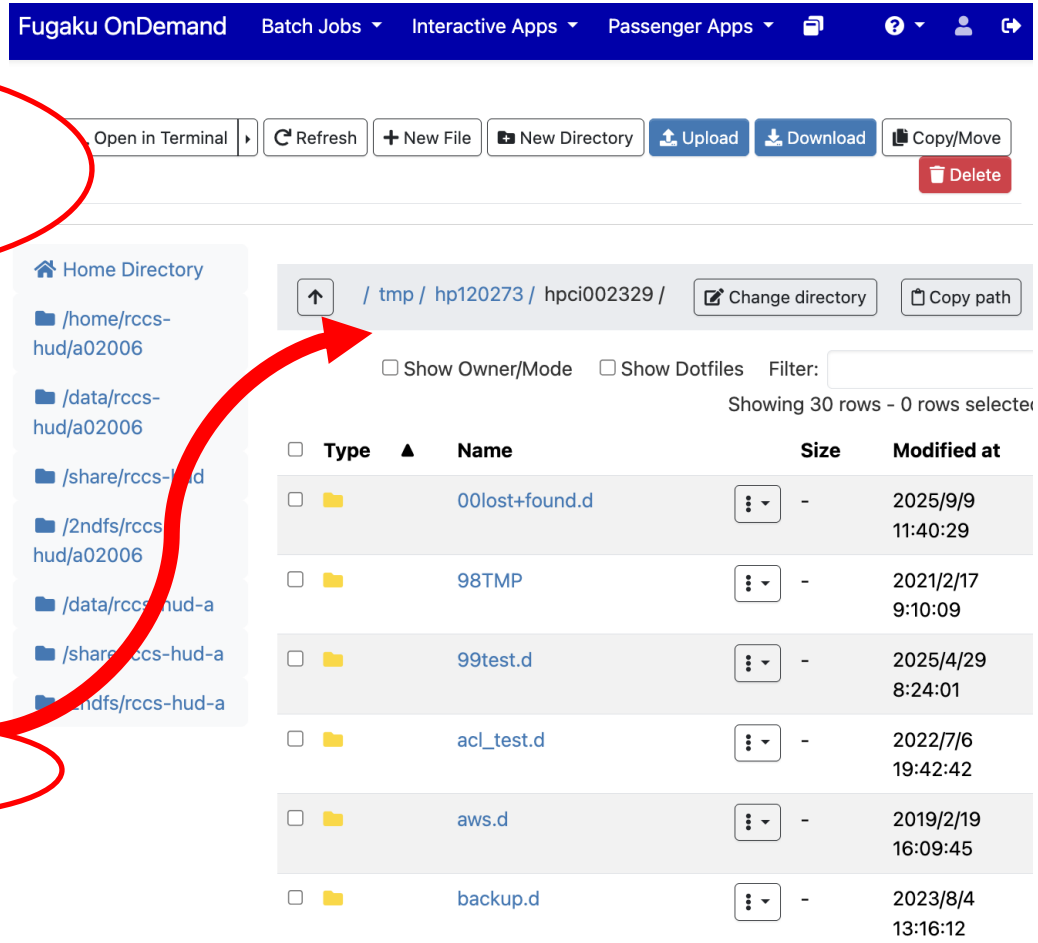
https://www.r-ccs.riken.jp/outreach/topics/20230907-1/

富岳Ondemand

利用はとっても簡単！！



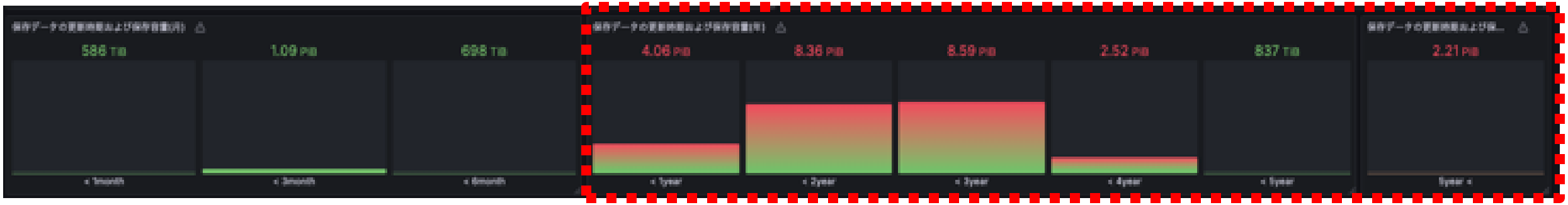




https://www.r-ccs.riken.jp/outreach/topics/20230907-1/

Cold Data in HPCI-SS

ファイル数 – 期間別利用状況



※ Coldデータを1年以上アクセスしないデータと定義した場合

容量 – 期間別利用状況

Total	Capacity	33PB
	Number of files	188 million
Cold data ratio (include Dark data)	Capacity	90.2%
	Number of files	88.8 %

HPCI共用ストレージのコールドデータ数が増加している。
一方で数PBのデータを保有するユーザなども増加している。
→ 削減をお願いしても削除すべきデータを把握する負荷が大きい
→ 利用者には監視情報を用いて詳細な情報を提供しているものの、さらなる可視化や情報付与が必要

Metadata annotation to used WorkFlow Tool(WHEEL) → SC Asia25 Poster(SG,Mar) → IDW25 Talk(AU,Oct)

CJK Project(A3)

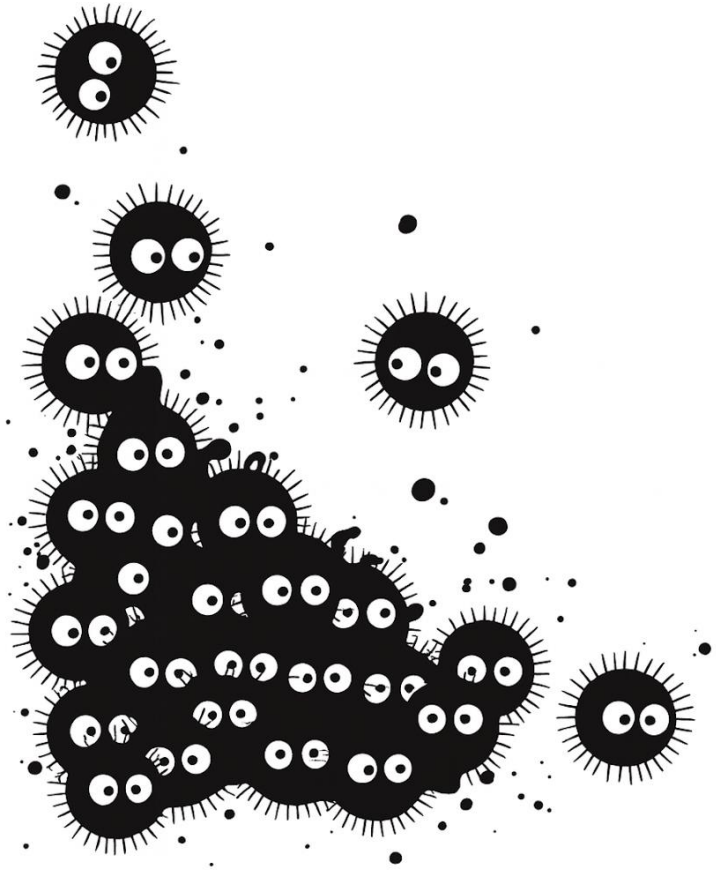
運用をしていて懸念がいくつか。。。　→ 拡張メタデータ付与環境の提供/自動的に拡張メタデータを設定する方法が必要

[1] ダークデータの取り扱い

前スライドの通り、HPCI共用ストレージには多数のコールドデータが…

コールドデータ内「分析や意思決定に活用されていない/存在しているが価値もリスクも不明な」データを
ダークデータと呼ぶそうです。
参照: https://www.gartner.com/en/information-technology/glossary/dark-data

コールドデータのうち特に古いデータについては、利用されていないなら削除やアーカイブしてほしい！！
一方、ダークデータのような管理されずに何のデータか不明なものについては、
ユーザの削除にかける負荷が高く難しい…　(ユーザへの情報提供も行ったり改善しているが…)



[2] AIやオープンサイエンス連携

AIやオープンサイエンスではラベルやメタ情報を付与が必須

現状の主なスパコンのストレージやHPCI共用ストレージではラベルやタグのような仕組みを提供していない傾向
(Object Storageが整備されていなかったり、あっても活用されていないことが多い)

HPCI共用ストレージのデータについては、「誰が」「何をつかって」「どのように取得したか」などの
情報が保持されていない傾向
データの引き継ぎなどでデータの重複や、不明なデータを保有したままになっていることも。
(実際は個別にエクセルなどで管理されているようだが管理の負荷も大きい…)

https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

Metadata annotation to used WorkFlow Tool(WHEEL) → SC Asia25 Poster(SG,Mar) → IDW25 Talk(AU,Oct)

CJK Project(A3)

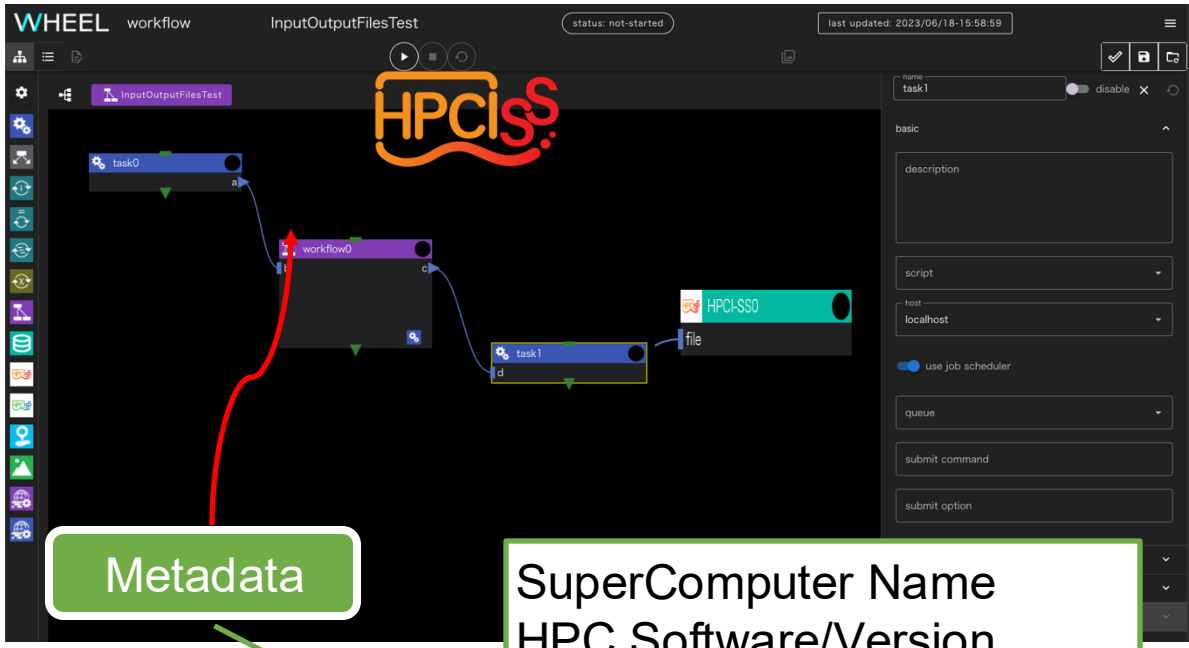
[A] HPC環境での自動拡張メタデータ付与方法
スパコンのジョブスケジューラやHPCストレージ側への機能追加や、
Starfish等の拡張メタデータ管理等ができるライセンスソフトとの連携は、
我々の権限外や費用負担が大きい(HPCI等、全体で検討すべき内容)

[B] HPCI-SS側への管理ソフトウェアの導入について
第三期システムの設計が完了している時点であり(容量優先の方針)
[A]と同様に費用負担が大きい。
VAST Catalog/IBM AFMなどを利用する環境に大きくシフトは難しい状況

[A]/[B]より直接システムに影響を与えない方法で改善案を検討
→ 富岳等のスパコンで利用されるWHEEL Workflow tool(R-CCS 川鍋上級技師)と連携
WorkFlow実行の際にHPC環境の拡張メタデータ情報を自動取得し、
拡張メタデータ情報をまとめてHPCI共用ストレージへ自動保存する
研究・開発を開始(最終的にはDOIなどとの連携を目指す)
WHEEL: https://github.com/RIKEN-RCCS/OPEN-WHEEL

(FY2024)
WHEEL にHPCI共用ストレージ(GfarmAPI)へのI/Oを整備
HPCやプリポスト計算後のデータをHPCI共用ストレージへ書き込めるようになった

(FY2025)
WHEEL にHPCでの計算時やプリポスト時に自動的に拡張メタ情報を収集し、
HPCI共用ストレージへ書き込む仕組みを整備中



SuperComputer Name
HPC Software/Version
JobID
Tags
Purpose
UserName(GroupName)
Create Time(HPC Using)
Etc…

https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

Gfarmへの拡張メタデータ情報の付与

Gfarmにはgfxattr/gffindxmlattrというメタ情報格納・検索ができるコマンドが！！(建部先生に教えていただいた)

■ メタデータ情報を作成(XMLファイル)

```
$ cat test.xml
<test_metadata>
    <write_user>kaneyama</write_user>
    <write_data>2025/09/18</write_data>
    <descripts>試験用のテストファイル</descripts>
</test_metadata>
```

→ **手動で管理は少し複雑**
- **GUIやラベル管理などと組み合わせられるととても便利？？**
- **メタDBから情報を検索する(?)ので高速**

■ 「test」属性およびメタ情報をmetadata_testfile.txtに付与

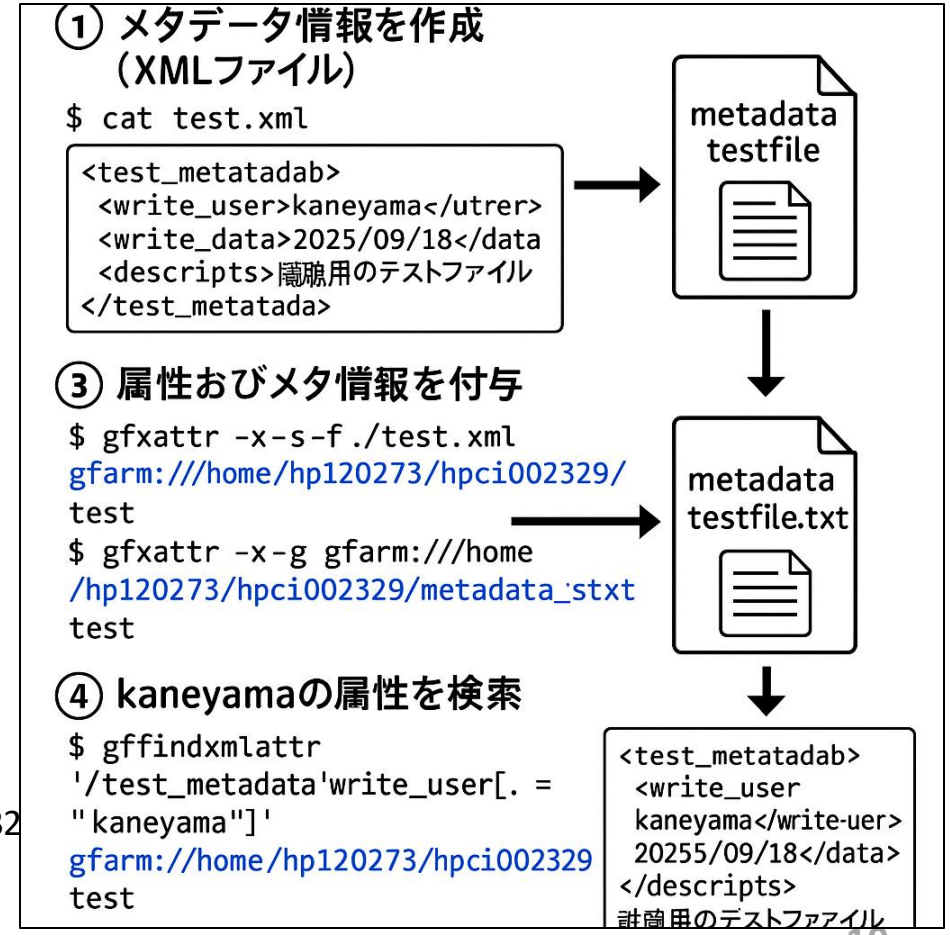
```
$ gfxattr -x -s -f ./test.xml ¥
gfarm:///home/hp120273/hpci002329/metadata_testfile.txt test
```

■ 付与されている属性を確認 → test属性のメタ情報を確認

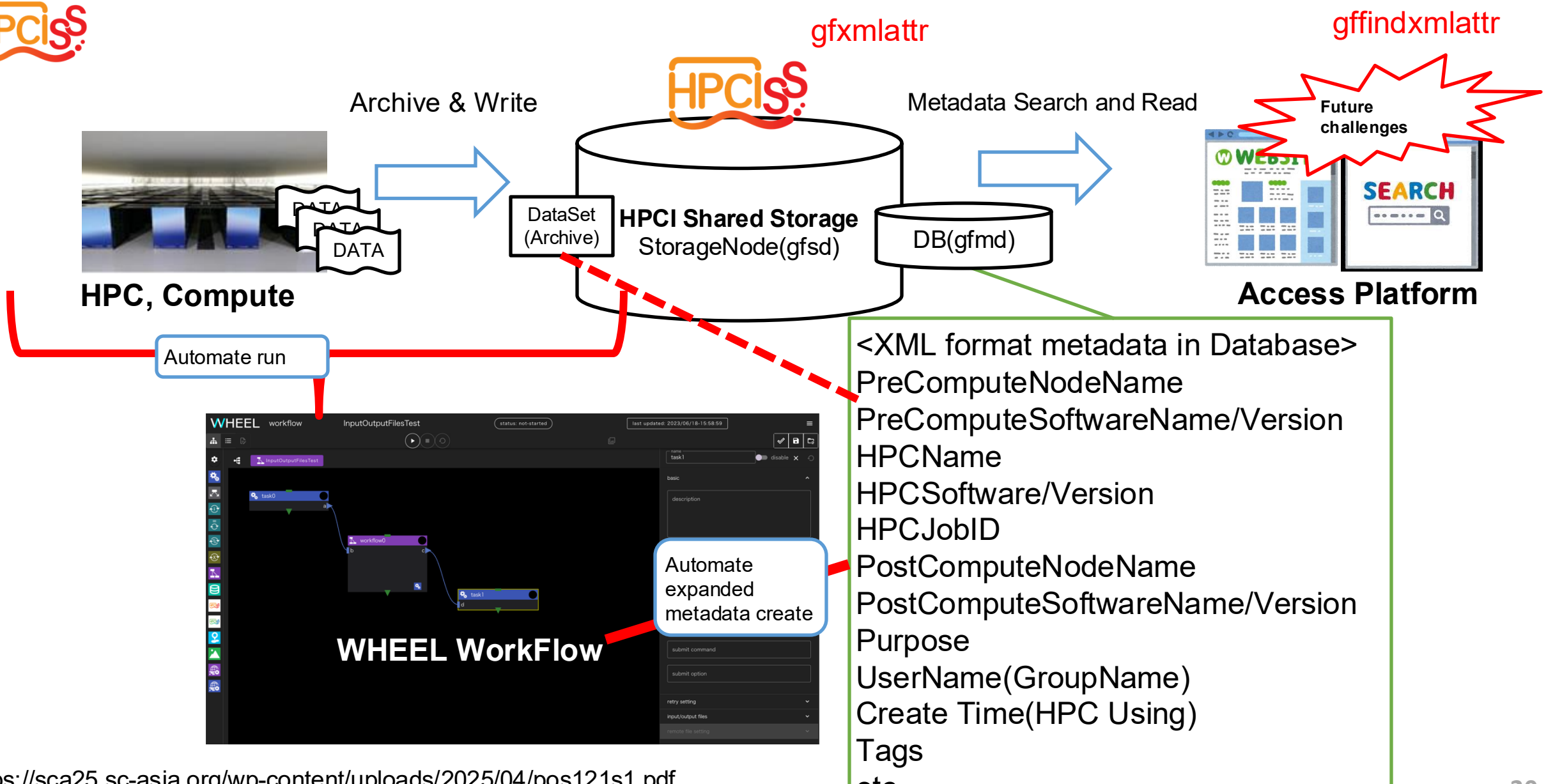
```
$ gfxattr -x -l gfarm:///home/hp120273/hpci002329/metadata_testfile.txt
test
$ gfxattr -x -g gfarm:///home/hp120273/hpci002329/metadata_testfile.txt test
<test_metadata>
    <write_user>kaneyama</write_user>
    <write_data>2025/09/18</write_data>
    <descripts>試験用のテストファイル</descripts>
</test_metadata>
```

■ write_userがkaneyamaのファイルを自身のホームディレクトリから取得

```
$ gffindxmlattr '/test_metadata/write_user[ . = "kaneyama"]' gfarm:///home/hp120273/hpci00232
gfarm:///home/hp120273/hpci002329/metadata_testfile.txt test
```



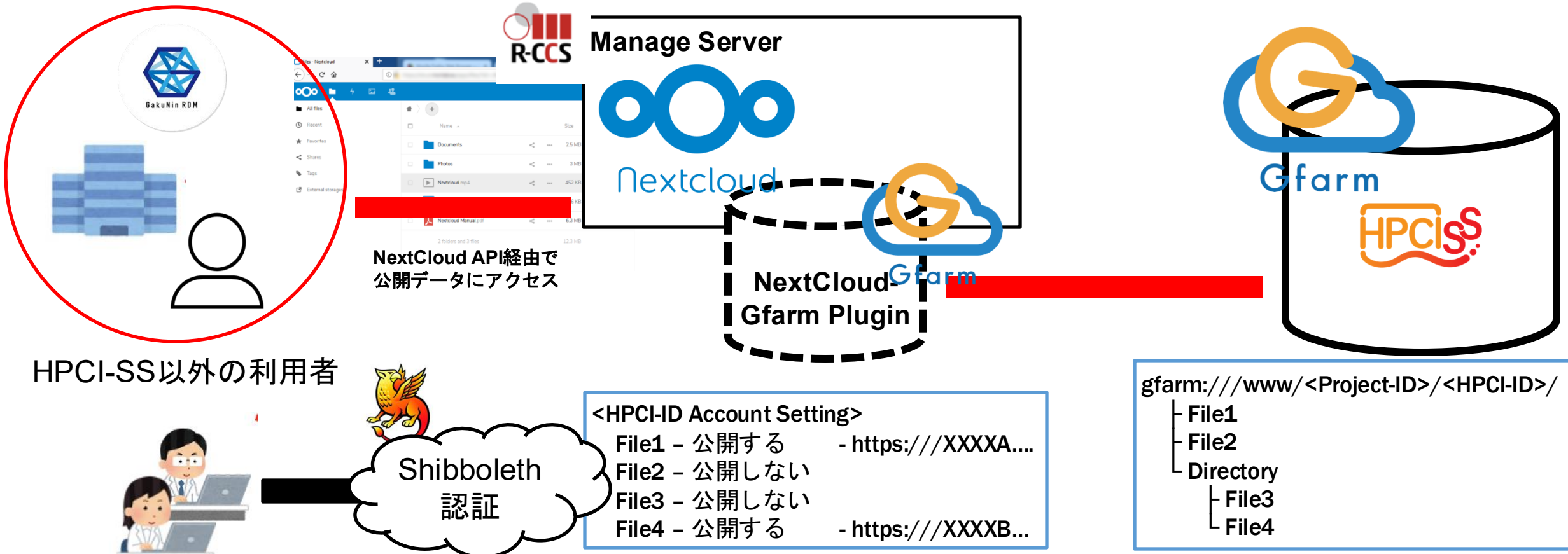
Metadata annotation to used WorkFlow Tool(WHEEL)



https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

OpenScience / データ公開

- HPCI共用ストレージを用いたデータ公開を進めます。
  - 公共データの利活用ユーザへ優先した割当容量/割当数を提供
  - 2025/10より環境整備、その後一部の課題グループにご協力いただき試験および改善検討を行う。



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

SCA/HPCAsia 2026: Call for Submissions

- Event Overview:
  - Date: January 26-29, 2026
  - Venue: Osaka International Convention Center (Osaka, Japan)
  - Theme: “Everything with HPC –AI, Cloud, QC, and Future Society”
- Call for Submissions: Papers, Posters, Workshops, BoFs, and Tutorials

Papers	Posters	Workshops	Birds of a Feather	Tutorials
Paper abstracts: 29 Aug 2025 Submissions close: 5 Sep 2025 Result notification: 20 Oct 2025	Submissions close: 27 Oct 2025 Result notification: 14 Nov 2025	Submissions close: 30 Jun 2025 Result notification: 31 Jul 2025	Submissions close: 1 Sep 2025 Result notification: 1 Oct 2025	Submissions close: 11 Jul 2025 Result notification: 15 Aug 2025

For more details, please visit our website:
https://www.sca-hpcasia2026.jp/

たいせつなこと

- HPCI共用ストレージ第三期サービスが開始します。
  - 東大はすでに開始。R-CCSが遅れており申し訳ございません(2025/10 ~)
  - 論理容量は45 → 95PBまで増加
  - メタ性能は向上 2.7倍+
  - ストレージ I/O性能向上 320GB/sec 1.7倍+
    - R-CCS 内部ネットワークも増強予定 3.2Tbps / SINET接続 800Gbps
  - I/O性能やネットワーク帯域が広いため、メタ・ネットワークボトルネックの影響は受けやすい。
    大きなファイルに固めて送るほうが効率が良いです → gfptarコマンド
- データ利活用にむけたサービス拡充を目指します
  - データ公開環境の提供開始、今後もサービス拡充を目指す
  - WHEEL WorkFlowツール連携/拡張メタデータ対応
- ひとでぶそく・えんじにあ不足
  - 一緒に運用してくださるメンバーを募集しています！！


Thank you