

ABCI におけるストレージサービスの紹介

谷村 勇輔

国立研究開発法人 産業技術総合研究所



自己紹介



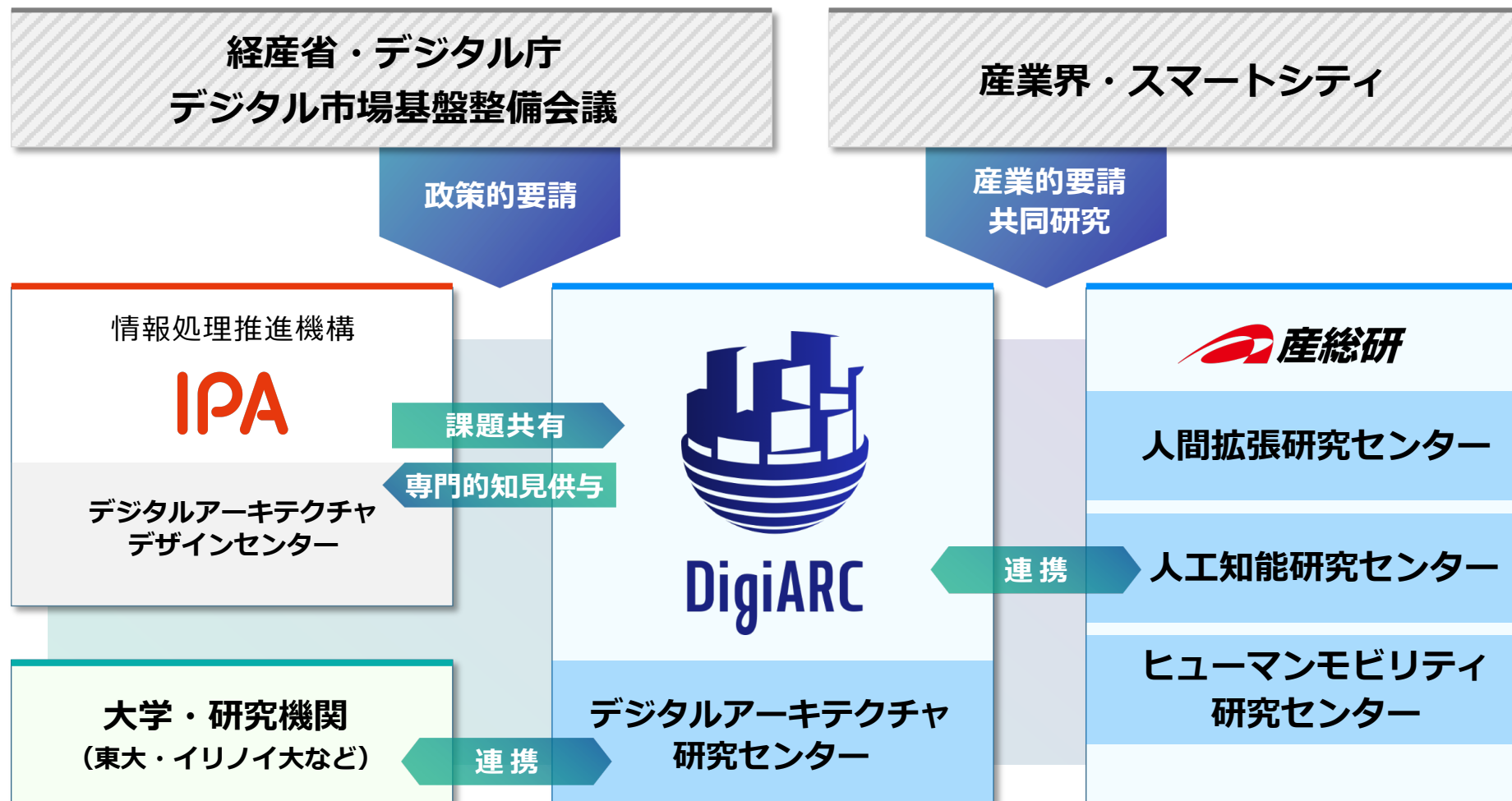
DigiARC



- 産業技術総合研究所 デジタルアーキテクチャ研究センター
 - 超分散アーキテクチャ研究チーム 研究チーム長
 - <https://www.digiarc.aist.go.jp/team/ccart/>
 - 兼務
 - 人工知能研究センター、実社会ビッグデータ活用オープンイノベーションラボラトリ等
 - 筑波大学大学院 システム情報工学研究科 准教授（連携大学院）
- 研究/業務
 - HPC, Big Data, AI 向けの並列分散データ処理、ストレージ技術
 - Continuum computing、Edge-to-cloud integration
 - 産総研 AI 計算インフラ（ABCI）の運用技術支援

デジタルアーキテクチャ研究センター

先導的なデジタル(情報処理)技術の開発、先駆的な社会実装に取り組むべく、2021.4 に設立。



Society 5.0 社会の実現に向けて

デジタル技術の発展により、
情報やデータをリアルタイムに
活用するデジタル社会が実現



サイバー空間とフィジカル空間の
「高度な融合」を目指す



詳細は <https://www.digiarc.aist.go.jp>

超分散コンピューティングコア

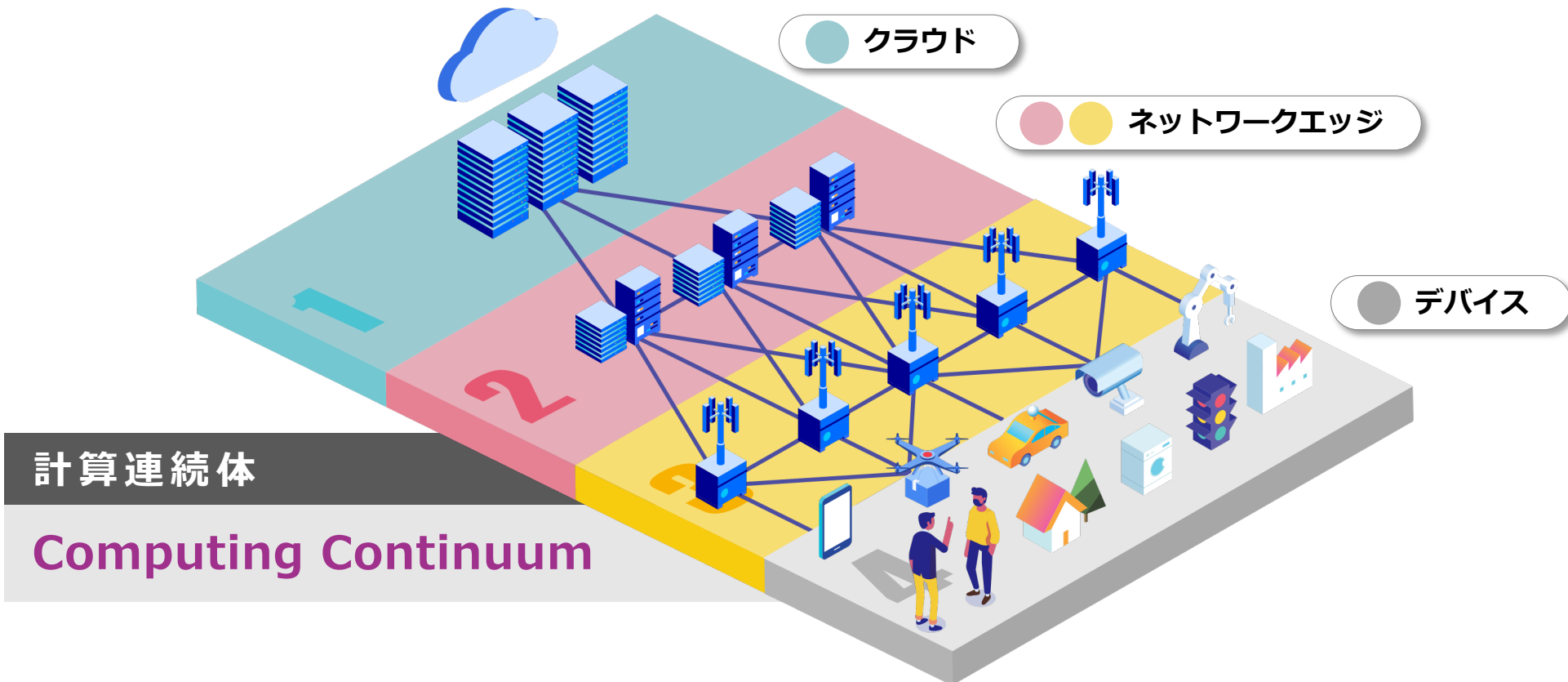
クラウド（点）ではなく、計算連続体（面）で情報処理

ユーザに近い処理による

応答性能の向上

省電力複数経路による

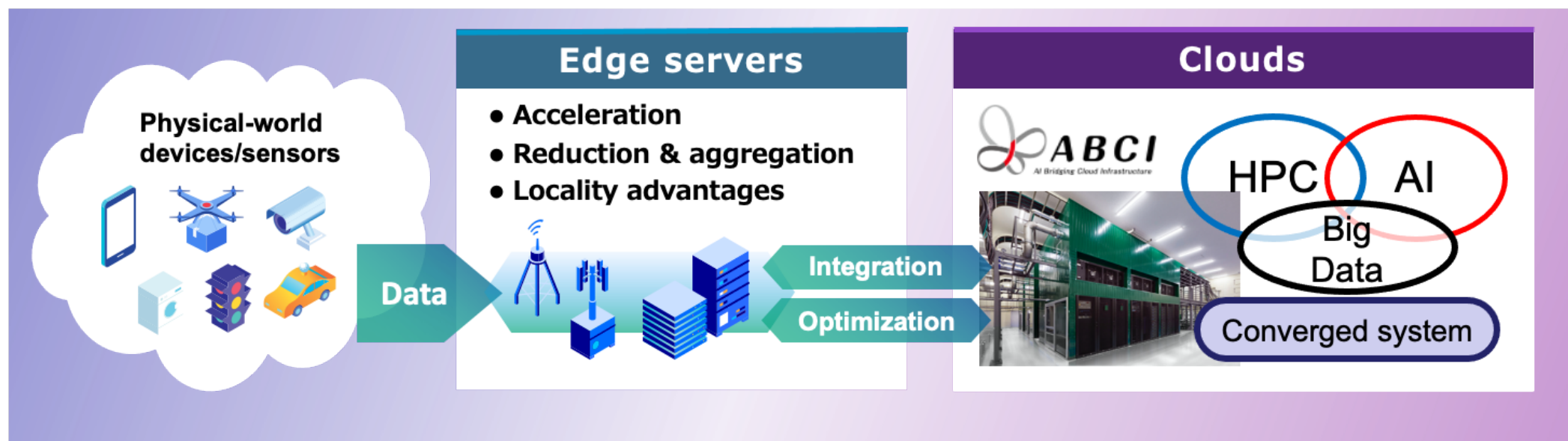
負荷分散と故障対応



超分散アーキテクチャ研究チームの取り組み

エッジからクラウドにまたがる「超分散コンピューティングコア」において

- ① エッジの活用による、アプリケーションの高度化・高性能化を促進する基盤技術
- ② エッジと連携して、莫大なAI計算・データ処理能力を提供するクラウド基盤技術の研究により、実社会での様々なデジタルサービスの実現に貢献する。



<https://www.digiarc.aist.go.jp/team/ccart/>

世界最大級・超省電力・オープンA I インフラストラクチャ



- 経産省「人工知能に関するグローバル研究拠点整備事業」(H28二次補正)の一環として整備
- 我が国における産学官によるA I 研究開発を加速するオープンイノベーションプラットフォーム
- 高い計算能力を活用したA I 技術の研究開発・実証、社会実装の推進、A I 分野の最重要課題への挑戦が目的
- 新型コロナウイルス感染症対策に無償提供
- 経産省「人工知能に関する橋渡しインフラ拡張」(R1補正)により昨年度末、大幅アップグレードを実現



2018年8月1日運用開始



AI インフラストラクチャ for everyone

Expert



ABCIグランドチャレンジ：
画期的な成果が見込まれる最重要課題への挑戦に
ABCIの全システムを最大24時間、無償提供

Advanced & Intermediate



最大2048GPUまで誰でも利用可能
すぐ使えるソフトウェア、データセット、
学習モデル等を提供

Beginner



初学者にも使いやすい統合開発環境を実現

データセンタ事業者等

企業がクラウドで個人情報等を
扱える水準のセキュリティ機構

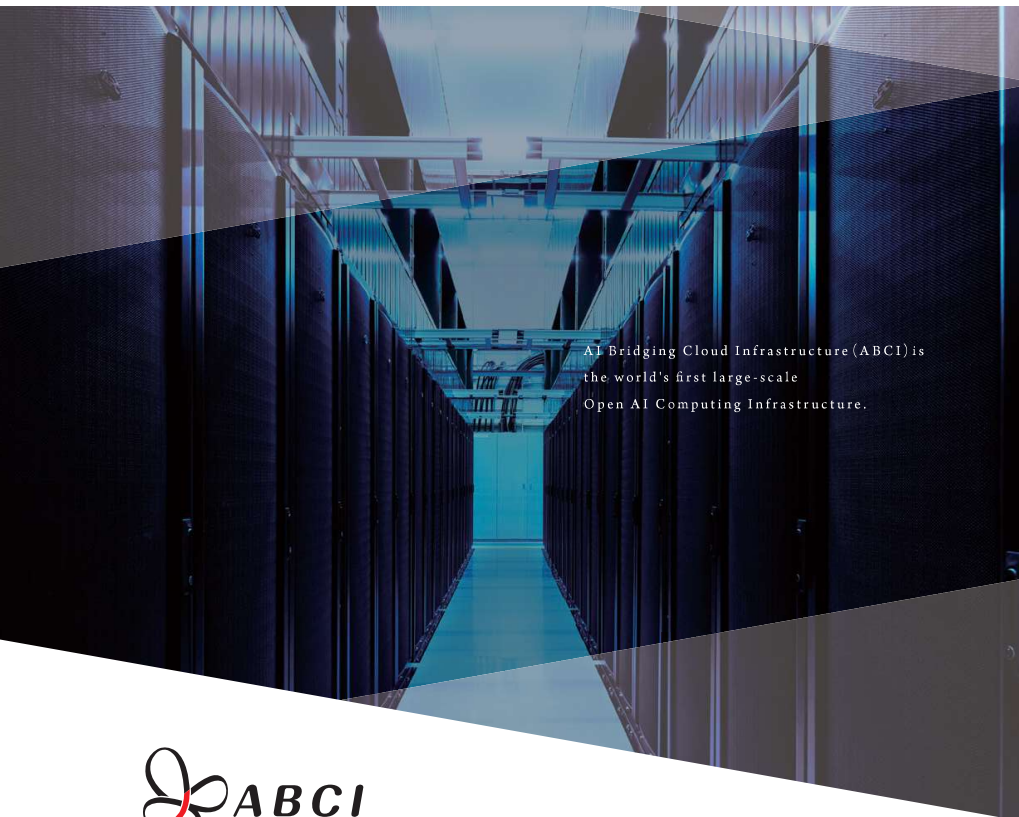
「AIを試す場」
人工知能産業のための
オープンプラットフォーム形成
最先端のAI研究から
誰でも試して使えるAIまで



数百の研究機関・大学・企業による利用・協業、
数千の研究者・エンジニアによる利用を促進

ABCI の主なユーザ企業

https://abci.ai/ja/link/use_case.html



誰もが利用できるAIクラウド計算システム「ABCI」

利用者インタビュー

事例1 ギリア株式会社 様：長年の積み重ねが開花、強力なコンピュータで実用を迎えたAI

事例2 株式会社 富士通研究所 様：スーパーコンピューティングという「商品」の開発

事例3 国立研究開発法人 日本原子力研究開発機構（JAEA）様：大規模シミュレーションと深層学習で、超高速の風予報を

事例4 株式会社 高電社 様：機械翻訳のスペシャリスト企業として30年、ABCIで実現するAI翻訳

事例5 パナソニック 株式会社 様：家電からソリューションへ、競争力ある技術をABCIで実現

事例6 アイリス 株式会社 様：インフルエンザを防ぐ、AIによる「匠の目」

事例7 Linne(リンネ) 株式会社 様：スマホからはじまる生物多様性保全の環(わ)

事例8 株式会社 トリプルアイズ様：囲碁AI研究は自動車メーカーのF1と同じ。めざせ世界の頂点！

事例9 LeapMind 株式会社様：量子化ニューラルネットワークが実現する、「どこにでもAIがある」エッジAIの世界

事例10 株式会社 リクルートテクノロジーズ様：先進技術から紡ぐ、リクルートの「未来のソリューション」

事例11 株式会社 パスコ様：光と電波で地表をキャッチ、AI解析で災害情報も迅速に

事例12 株式会社 アタリ様：人と見まがうバーチャルヒューマンが人に寄添う未来へ

事例13 オムロンサイニックエックス 株式会社様：オープンイノベーションで挑むオムロンの非連続な技術進化

事例14 株式会社 IABC様：シミュレーションとAIが切り開く地震工学の新たな地平線

事例15 株式会社 コトバデザイン様：あらゆるモノと自然な“おしゃべり”ができる世界を目指して

ABCI 2.0 (2021Q2-)

ABCI 1.0 (2018Q2-)

550 PF (FP16), 37.2 PF (FP64)
476 TiB Memory, 1.74 PB NVMe SSD



ABCI Expansion (2021Q2-)

300 PF (FP16), 19.3 PF (FP64)
97.5 TiB Memory, 384 TB NVMe SSD



Compute Nodes (V) x 1088

- GPU NVIDIA Tesla V100 SXM2 x 4
- CPU Intel Xeon Gold 6148 (2.4GHz/20cores) x 2
- Memory 384 GiB
- Local Storage Intel SSD DC P4600 (NVMe) 1.6TB x 1
- Interconnect InfiniBand EDR x 2 (25 GB/sec)

Compute Nodes (A) x 120

- GPU NVIDIA A100 x 8
- CPU Intel Xeon SP (Ice Lake) x 2
- Memory 512 GiB
- Local Storage Intel SSD DC P4610 (NVMe) 1.6TB x 2
- Interconnect InfiniBand HDR x 4 (100 GB/sec)

Precision	ABCI 1.0	今回追加	ABCI 2.0 (合算)	スケールアップ ^o	用途
FP32/TF32	75 PF	150 PF	225 PF	x 3	高精度DL
TF32 w/ Sparsity	↑(*1)	300 PF	375 PF	x 5	高精度DL
FP16/BF16	550 PF	300 PF	850 PF	x 1.55	低精度DL
FP16/BF16 w/ Sparsity	↑(*1)	600 PF	1.15 EF	x 2.09	低精度DL
FP64	37.2 PF	19.3 PF	56.5 PF	x 1.52	シミュレーション

(*1) V100はSparsityをサポートしないため参考値。

ABCI 2.0 全体概要



計算リソース

計算ノード(A) x 120台

- CPU: Intel Xeon Platinum 8360Y x 2
- GPU: NVIDIA A100 SXM4 x 8
- Memory: 512GiB
- NVMe SSD: Intel SSD DC P4510 2.0TB x 2
- Interconnect / InfiniBand HDR x 4

計算ノード(V) x 1088台

- CPU: Intel Xeon Gold 6148 x 2
- GPU: NVIDIA V100 SXM2 x 4
- Memory: 384GiB
- NVMe SSD: Intel SSD DC P4600 1.6TB x 1
- Interconnect: InfiniBand EDR x 2

メモリインテンシブノード x 10台

- CPU: Intel Xeon 6132 x 2
- Memory: 768GiB
- SSD: Intel SSD DC S4500 1.9TB x 1
- Optane SSD: Intel Optane SSD P4800X 1.5TB x 2
- Interconnect: InfiniBand EDR x 2

共有ファイルシステム

DDN SFA14KX

(w/ SS8462 Enclosure x 10) x 3

- 12TB NL-SAS HDD x 2400
- 3.84TB SAS SSD x 216

DDN SFA14KX

(w/ SS9012 Enclosure x 5) x 1

- 7.68TB SAS SSD x 216

DDN ES7990X

(w/ SS9012 Enclosure x 2) x 3

- 18TB NL-SAS HDD x 801

DDN ES400NVX x 3

- 7.68TB NVMe SSD x 69

ABCIクラウドストレージ

HPE Apollo 4510 Gen10 x 24

- 12TB SATA HDD x 1440

インタラクティブノード / ゲートウェイノード



ファイアウォール

SINET5

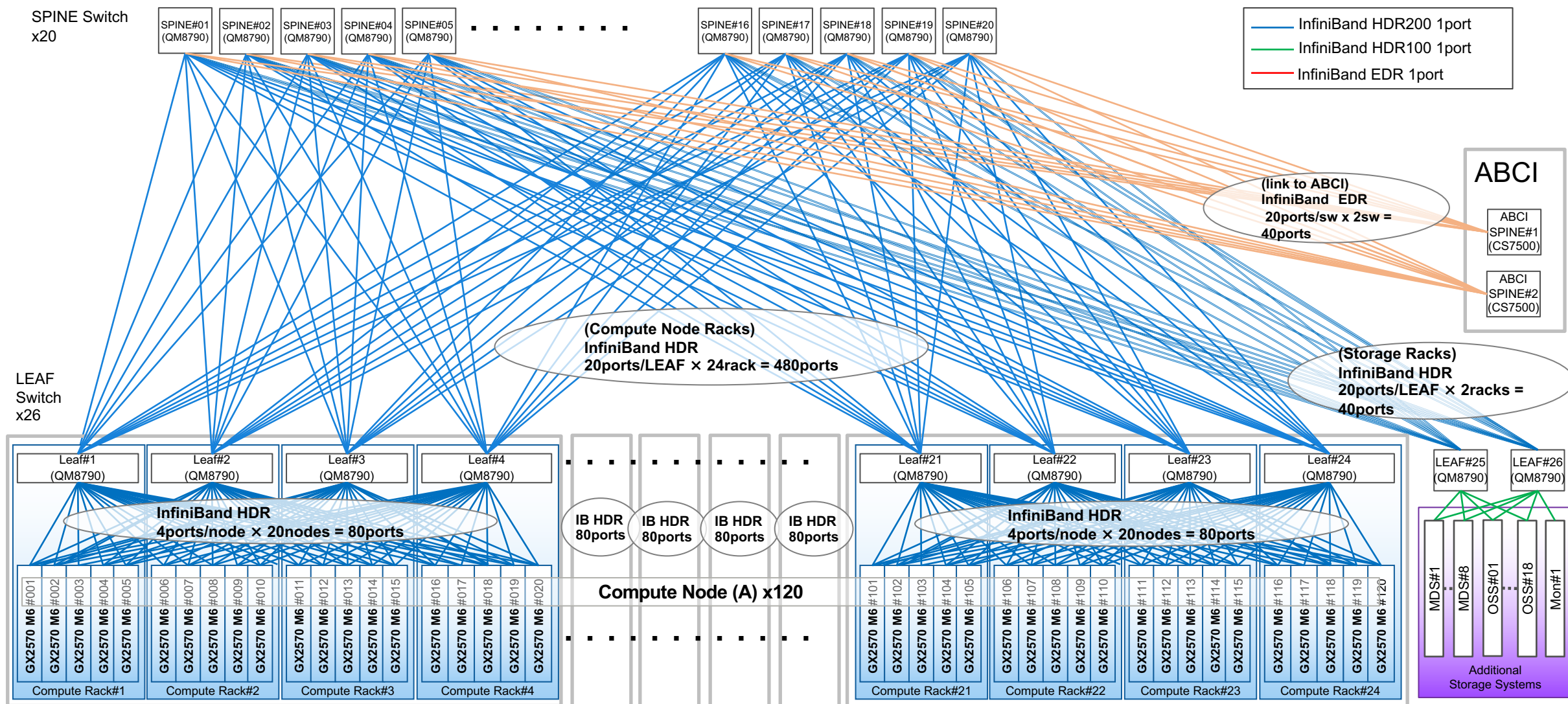
計算ネットワーク
(InfiniBand EDR/HDR)

サービスネットワーク
(10GbE)

2.0 としての増強部分

ABCI Compute Network (A)

Full-bisection & SHARPV2-enabled network



ABCI のストレージサービス

- サービスの種類
 - 計算ノードのローカルディスク
 - 共有ファイルシステム
 - クラウドストレージ
- ABCI 2.0 Upgrade に合わせて構成を変更
 - メインで利用される「グループ領域 (Groups)」に新機材を割り当て
 - 大容量の使用、アクセス頻度が低いと想定されるプロジェクト等には再構成した旧機材を割り当て

一部は作業中であるが、本日は変更完了後の構成を紹介します。

計算ノードのローカルディスク

- 用途：スクラッチ

- ジョブ実行中、**\$SGE_LOCALDIR** のパスでアクセス可能。ジョブ終了後にデータは削除される。
- 予約利用の場合、予約期間内は **\$SGE_ARDIR** のパスでアクセス可能。予約期間内であれば、各ジョブ終了後もデータは保持される。

- 各ノードの構成と資源タイプ毎の割り当て容量

- 計算ノード（A）：2.0 TB NVMe SSD x2
- 計算ノード（V）：1.6 TB NVMe SSD x1

タイプ名	CPUコア	GPU数	メモリ (GB)	ストレージ (GB)
AF (ノード占有)	72	8	480	3440 *
AG.small	9	1	60	390
F (ノード専有)	40	4	360	1440
G.large	20	4	240	720
G.small	5	1	60	180
C.large	20	0	120	720
C.small	5	0	30	180

* 1590GB と 1850GB のスペースが提供される

共有ファイルシステム

- スクラッチ利用
 - オンデマンドに構成できるファイルシステム: /beeond
 - (検討中) キャッシュとして利用可能な高速共有スクラッチ領域: /scratch
- Home 利用: /home
 - 利用者の \$HOME 領域。1 アカウントにつき、200GiB が割り当てられ、容量追加は不可。
- Groups, Projects 利用: /groups, /projects
 - 利用グループ、プロジェクト毎に割り当てられる。
 - 初期クォータ値は 0TiB。250TiB まではポータルから容量追加（即反映）が可能。
 - 1TiB のクォータ増量につき、毎月 5 ABCI ポイントが課金される。
- 特別利用: /bb
 - グランドチャレンジ 等の特定用途向けの高速領域

/beond のサービス

- 複数ノード占有時、計算ノードのローカルディスクを束ねて提供される。
 - BeeGFS の BeeOND (BeeGFS On Demand) 機能により構成
 - BeeGFS version: 7.2.3 (適宜更新)
 - 計算ノード (A) では片方のディスク (/local2) のみを BeeOND に提供
- 利用方法
 - ジョブ投入時に **-I USE_BEEOND=1** を指定
 - **\$SGE_BEEONDDIR** (/beond) でアクセス可能
 - 詳細オプションの指定
 - ジョブ投入時
 - **-v BEEOND_METADATA_SERVER=n** : メタデータサーバ数の指定 (default: 1)
 - **-v BEEOND_STORAGE_SERVER=m** : ストレージサーバ数の指定 (default: 要求ノード数)
 - ジョブ実行時にディレクトリ毎に切替可
 - **beegfs-ctl** の **-num-targets** : ストライプカウントの指定 (default: 4)
 - **beefs-ctl** の **-chunk-size** : ストライプサイズの指定 (default: 512KB)

/home のサービス

- 各アカウントに割り当てられる \$HOME 領域
 - 容量は 200GiB、追加不可
 - `/home/<username>` でアクセス
- 構成
 - 総物理容量: 1.42 PB、総実効容量: 1.09 PB
 - 5000 アカウント以上の作成が可能
 - 理論最大性能: 6M IOPS, 100 GB/s
 - Lustre File System (version 2.12.5、適宜更新)
 - MDT/OST: DDN SFA 14KX x1
 - SSD 7.68TB x 185 (All Flash)
 - MDS: 2 servers
 - OSS: 4 servers

/groups のサービス

- 利用グループ毎に割り当てられる領域
 - 容量は 250TiB まで追加可能
 - 1TiB のクォータ増量につき、毎月 5 ABCI ポイントが課金される
 - **/groups/<groupname>** でアクセス
- 構成
 - 総物理容量: 14.26 PB、総実効容量: 10.64 PB、総inode数: 18.7 billion
 - 1グループ 30TB 利用で約 350 グループをホスト可能
 - 理論最大性能: 2.4M IOPS, 150 GB/s
 - Lustre File System (version 2.12.5、適宜更新)
 - MDT: DDN SFA 200NV x1 (/bb と共用)
 - DNE (Distributed Namespace Environment) を有効にし、6 VDs に分割して運用
 - MDS: 2 servers
 - OST/OSS: DDN ES7990X x3
 - NL-SAS HDD 18TB x (264 + 3): 12 DCR pools, 12 VDs (RAID6: 8d+2p)
 - 2 OSS VMs

/projects のサービス

- 利用グループ毎に割り当てられる領域
 - 100 TiB 以上の利用、SE 作業、アーカイブ目的等の利用グループを割り当て
 - ABCI 初期導入時の機材を利用 (Spectrum Scale (GPFS) → Lustre に移行中)
- 構成
 - 総物理容量: 28.8 PB、総実効容量: 21.8 PB (見込み)
 - 理論最大性能: 18M IOPS, 144 GB/s
 - Lustre File System (version 2.12.5、適宜更新)
 - MDT: DDN SFA 200NV x1 (/scratch と共用)
 - DNE (Distributed Namespace Environment) を有効にし、6 MDT に分割して運用
 - MDS: 2 servers
 - OST: DDN SFA 14KX x3 (/scratch と共用、ディスクは分離)
 - NL-SAS HDD 12TB x (780 + 20): 78 VDs (RAID6: 8d+2p)
 - OSS: 12 servers (/scratch と共用)

/bb のサービス

- 特定用途向けの高速領域
 - ABCI グランドチャレンジ や特別実験等に割り当て
- 構成
 - 総物理容量: 507 TB、総実効容量: 395 TB、総inode数: 6.2 billion
 - 理論最大性能: 9M IOPS, 300 GB/s
 - Lustre File System (version 2.12.5、適宜更新)
 - MDT: DDN SFA 200NV x1 (/groups と共用)
 - DNE (Distributed Namespace Environment) を有効にし、2 VD に分割して運用
 - 必要に応じて Data On MDT (DOM) を有効可
 - MDS: 2 servers
 - OST/OSS: DDN ES400NVX x3
 - NVMe SSD 7.68TB x (22 + 1): 2 DCR pools, 8 VDs (RAID6: 8d+2p)
 - 4 OSS VMs

クラウドストレージ

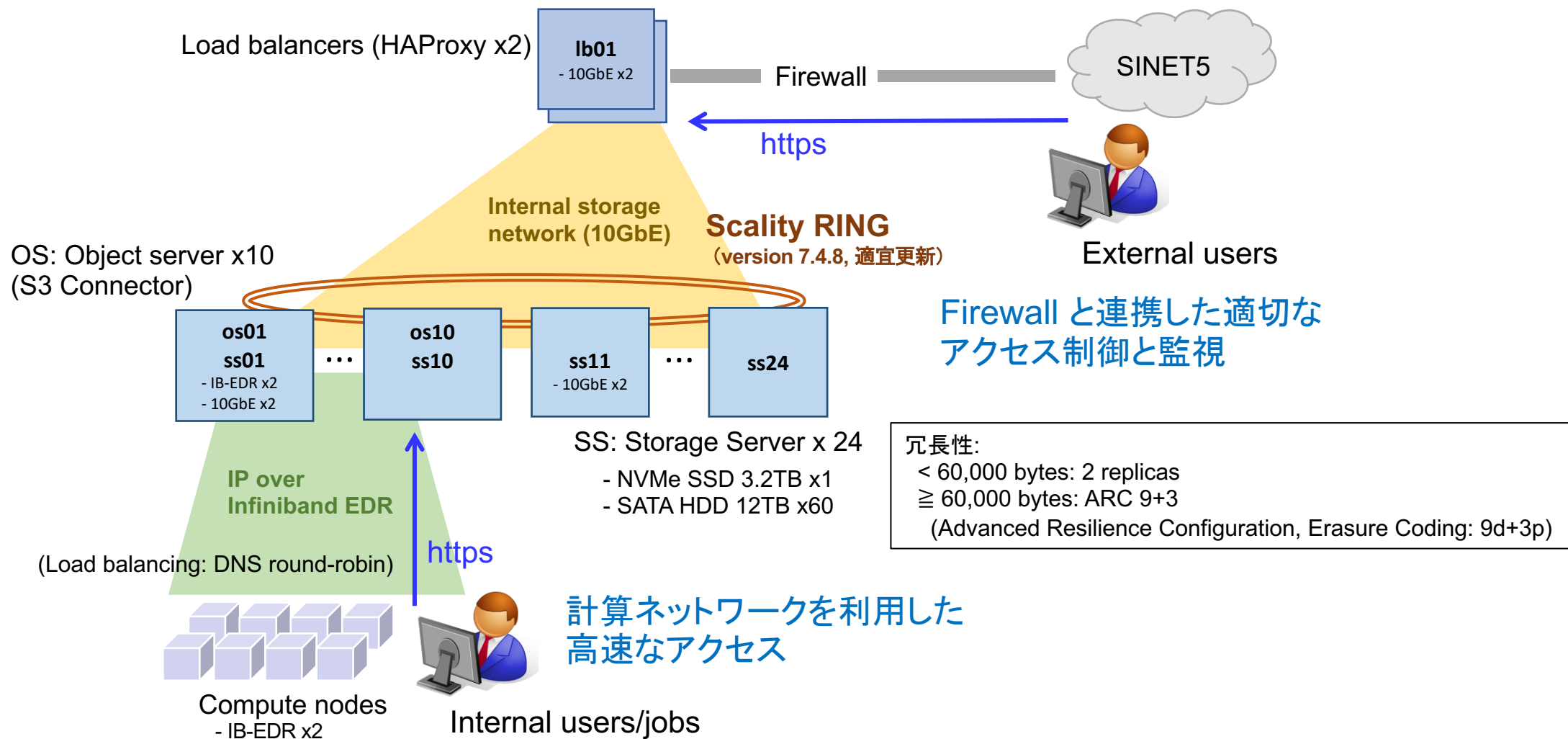
- 用途

- ABCI とのデータのインポート, エクスポート
- 任意の“データ利用グループ”の作成とその中でのデータ共有
- データの一般公開, パブリックリポジトリの提供

- 主な特徴

- Amazon S3 互換インタフェースでアクセス
 - データ転送や共有において, 既存の s3 互換ソフトウェア群を利用可能
- 通信経路、およびデータの暗号化をサポート
- ABCI 本体のアカウント／グループと統合されたユーザ管理
 - ABCI におけるグループ単位の利用料支払いの仕組みがそのまま使える
 - ただし、Groups, Projects と異なり、使用量の従量制（0.0001 ポイント/GB・日）
 - ABCI のグループ管理と, S3 によるグループを超えたアクセス制御とを両立

クラウドストレージの構成



暗号化のサポート

- クライアントとストレージ間の通信の暗号化
 - どの経路においても HTTPS を必須
- ストレージ側での暗号化（SSE: Server-Side Encryption）
 - Scality RING が提供する SSE を提供
 - Amazon S3 の SSE-S3 と類似した Bucket レベルの暗号化
 - ユーザから透過的に利用可
 - Scality RING が暗号化鍵を管理
 - 上記鍵を用いて保存時に暗号化, 読み出し時に復号化
 - 業界標準の AES256 暗号化（FIPS 140-2 certified）
- クライアント側での暗号化（CSE: Client-Side Encryption）
 - ユーザ側で AWS SDK 等を通して利用可能

アカウント管理

- 2つのアカウント管理体系の統合

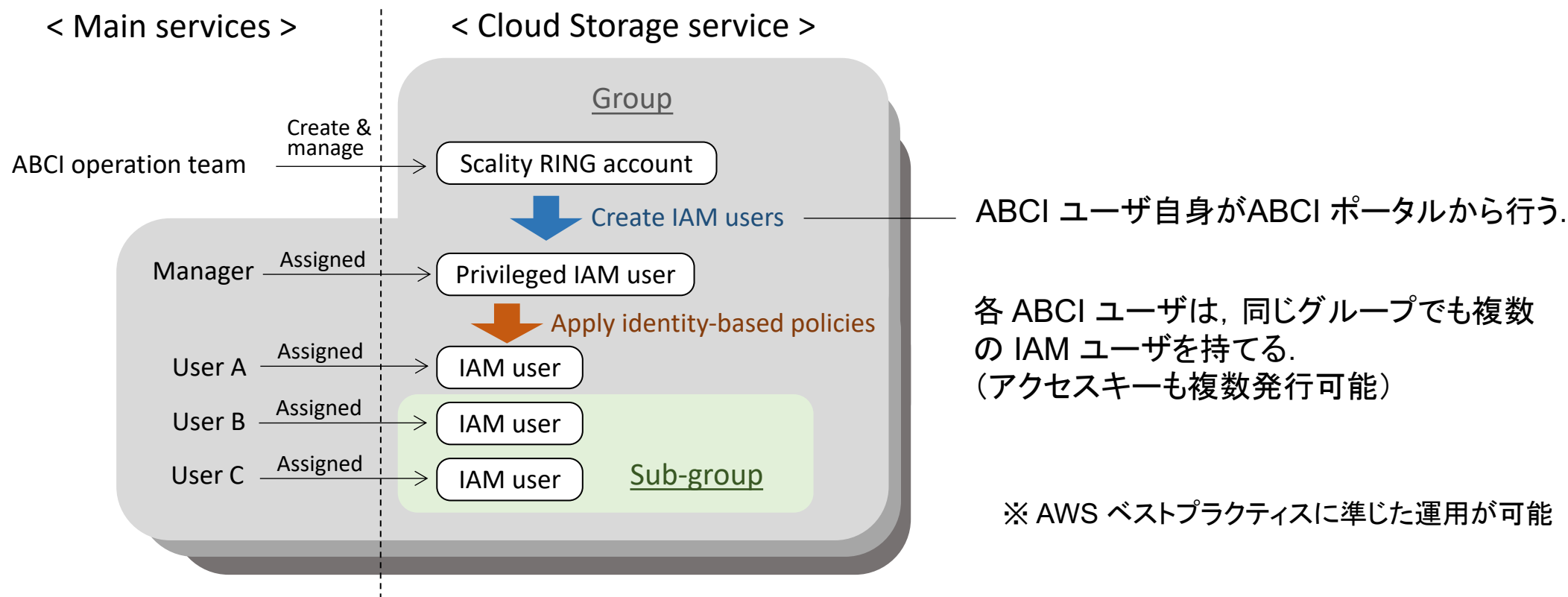
- ABCI : OS の UID/GID
 - ABCI の利用料はグループでの支払い
 - ABCI ユーザは 1 つ以上のグループに属して ABCI を利用
- Scality RING: AWS IAM (Identity and Access Management)
 - IAM ユーザ, および IAM ユーザのアクセス権の管理を行うためのサービス

- 統合管理のポリシー

- ABCI ユーザは必要に応じて, 所属グループ毎に自身の Cloud Storage アカウントを申請 (作成) できる。
- ABCI グループ管理者は勝手に Cloud Storage アカウントを発行できない。ただし, 停止 (削除) はできる。
- アクセス権の管理について最大限の柔軟性を与える。

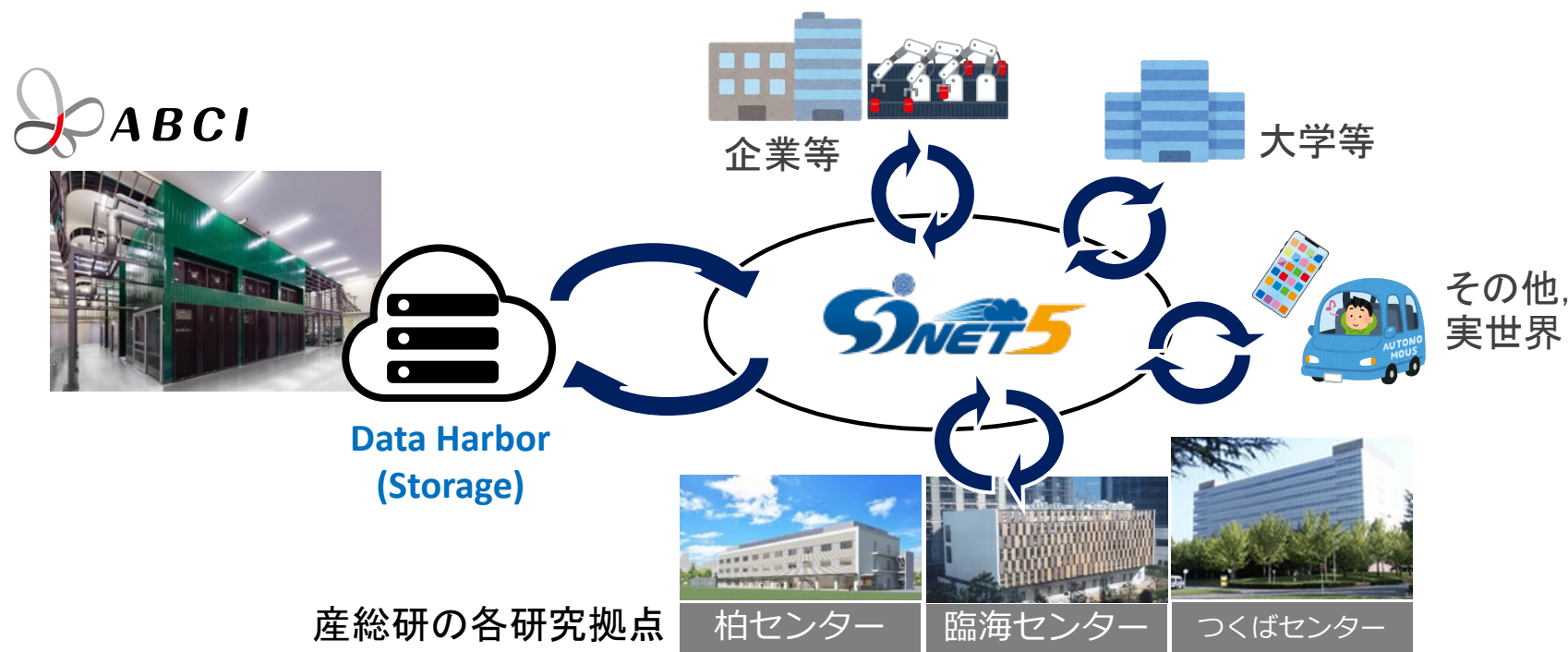
アカウント管理の実装

- ABCI 運用側で全 Scalify RING アカウントを管理
- ABCI 利用者には IAM ユーザを付与
- 一部の特権を付与した IAM ユーザをグループ管理者に付与



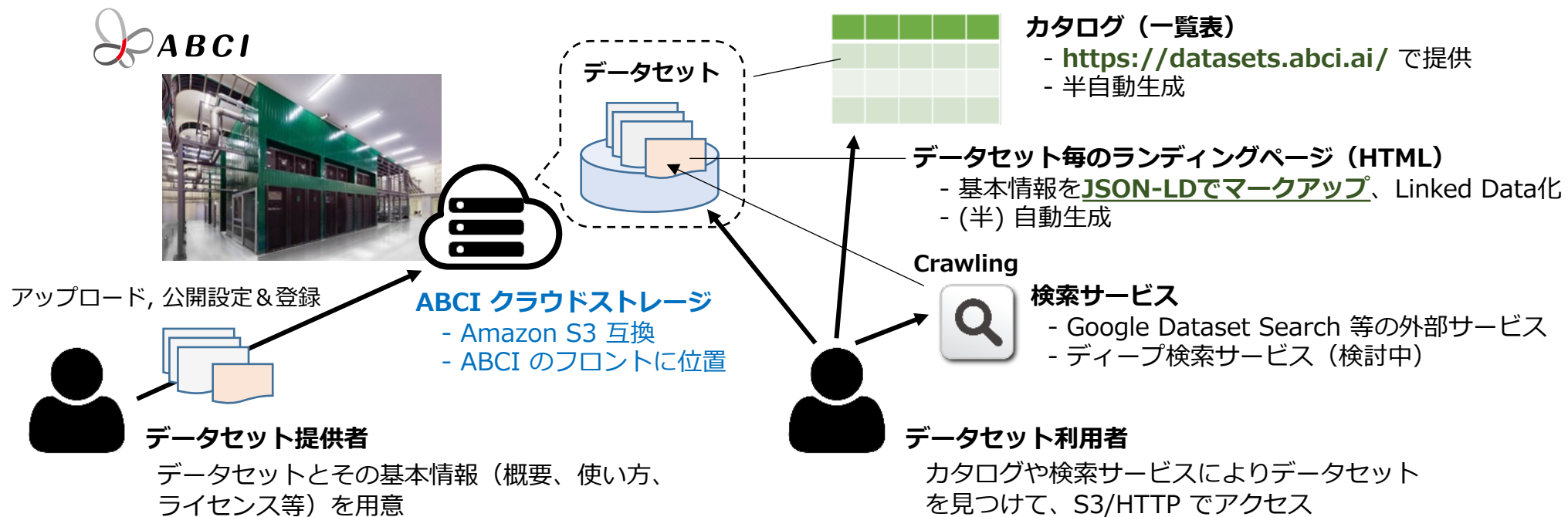
データの価値化を実践する「場」の提供を目指して

- クラウドストレージは、ABCIのフロントに位置し、SINET5に直結した「データハーバー」としての役割を担う
 - SINET5に直結した各機関から、高速・安全にデータを収集・蓄積
 - ABCI上で生成された、高性能な汎用学習モデル等の共有・配布



ABCI Datasets サービス

- データの公開・共有を支援するデータ連携基盤（サービス）
 - ABCI Datasets: ABCI 利用者によって登録されたデータセット群（公開、限定公開されたもの）
 - 公開されたデータには ABCI アカウントを持っていなくてもアクセスが可能
- 実験的に運用中
 - 2020.10 初版公開、2021.3 β版公開



ABCI Datasets



Home



Keywords



Registration(EN)



Registration(JP)



About ABCI

Helping you find and share datasets for accelerating AI R&D

search datasets

11 datasets

Order by Date Name

ランディングページ

AAIC Workload

Creator : Artificial Intelligence Cloud Research Team, AIRC, AIST

License : Custom License

Last Modified : 23-Apr-2021

job scheduling workload deep learning open data NEDO 2.0

Automatic Labeled LiDAR Data for Human detection

Creator : Wonjik Kim, Masayuki Tanaka, Masatoshi Okutomi and Yoko Sasaki

License : Custom License

Last Modified : 08-May-2019

human detection LiDAR scene deep learning open data NEDO 2.0

STAIR Actions Videos

Creator : STAIR Lab. (Software Technology and Artificial Intelligence Research Laboratory), Chiba Institute of Technology

License : Creative Commons Attribution 4.0 International license (CC BY 4.0)

Last Modified : 29-Jan-2019

videos action recognition deep learning open data NEDO 2.0

MUSIC for P3 dataset

Creator : Geoinformation Service Research Team, Digital Architecture Research Center, National Institute of Advanced Industrial Science and Technology

License : Custom License

Last Modified : 26-Jan-2018

Dataset

AAIC Workload

Version : 1.0 Last Modified : Fri, 23 Apr 2021 JST

job scheduling workload deep learning open data NEDO 2.0

This is a workload set of jobs run in GPU node queues of AIST Artificial Intelligence Cloud (AIRC). As AAIC is dedicatedly used for AI research, many jobs in the workload are expected to be run for training deep learning applications. The workload contains jobs submitted from July 14th, 2017 to December 31st, 2018.

Detailed Information : <https://github.com/aistairc/aaic-workload/>

Distribution

Encoding format : CSV

- <https://airc-aaic-workload.s3.abci.ai/workloads/annual/aaicwl-201707-201806.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/annual/aaicwl-201807-201812.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2017/aaicwl-2017Q3.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2017/aaicwl-2017Q4.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2018/aaicwl-2018Q1.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2018/aaicwl-2018Q2.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2018/aaicwl-2018Q3.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/quarter/2018/aaicwl-2018Q4.csv.gz>
- <https://airc-aaic-workload.s3.abci.ai/workloads/other/aaicwl-hpc170.csv.gz>

Creator : Artificial Intelligence Cloud Research Team, AIRC, AIST
<https://www.airc.aist.go.jp/aiprt/>Contact : shinichiro.takizawa@aist.go.jpLicense : Custom License
<https://github.com/aistairc/aaic-workload/>

Created Date : Wed, 26 Jun 2019 JST

First Publication Date : Wed, 26 Jun 2019 JST

Citation : 滝澤 真一朗, 小川 宏高, 高野 了成, 大規模GPUクラスタにおける深層学習ワークロードの傾向把握, 第170回ハイパフォーマンスコンピューティング研究会, July, 2019.

地理空間データの公開

2020.5 地表面の状態に応じて色分けされたカラーレーダー画像を公開

- https://www.aist.go.jp/aist_j/press_release/pr2020/pr20200522/pr20200522.html
- 衛星マイクロ波センサー（PALSAR: Phased-Array type L-band Synthetic Aperture Radar）が取得した SAR データ（5年3ヶ月分、200万シーン、700TB）を ABCI 上でカラー化（画像処理）し、ABCI のストレージに保存して公開
 - LandBrowser（<https://gsrt.airc.aist.go.jp/landbrowser/index.html>）からアクセス可能
 - PALSAR L1.1 以上データ（4PB）：一部公開済、全公開予定
 - PALSAR 干渉処理データ（4PB）：公開予定

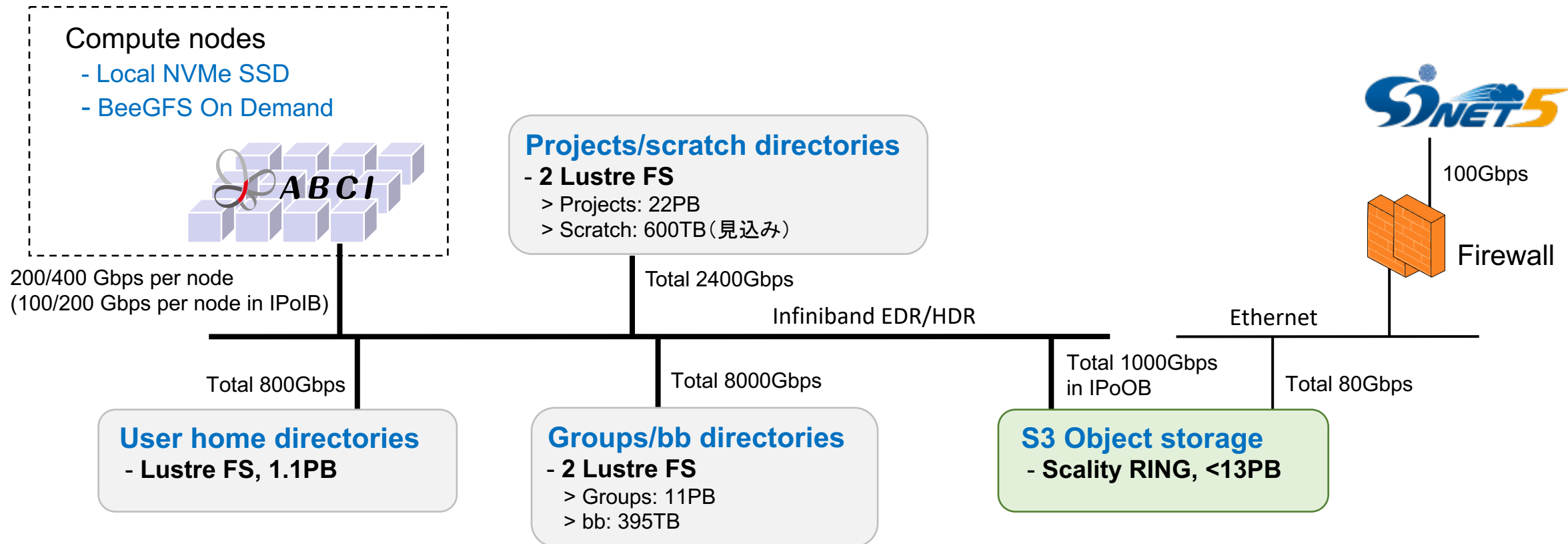
2020.8 3DDB Viewer による三次元データの公開

- https://gsrt.airc.aist.go.jp/3ddb_demo/tdv/index.html
- 地形や建物、構造物の三次元データを地図に重ねて表示

2021.2 ASTER-GDEM ファイルの全球、全数を公開

- <https://s3.abci.ai/aster-pds/index.html>
- 人工衛星搭載センサ「ASTER」による地球の陸域全てを対象とした数値地形データ
- Cloud Optimized GeoTIFF 形式で配布

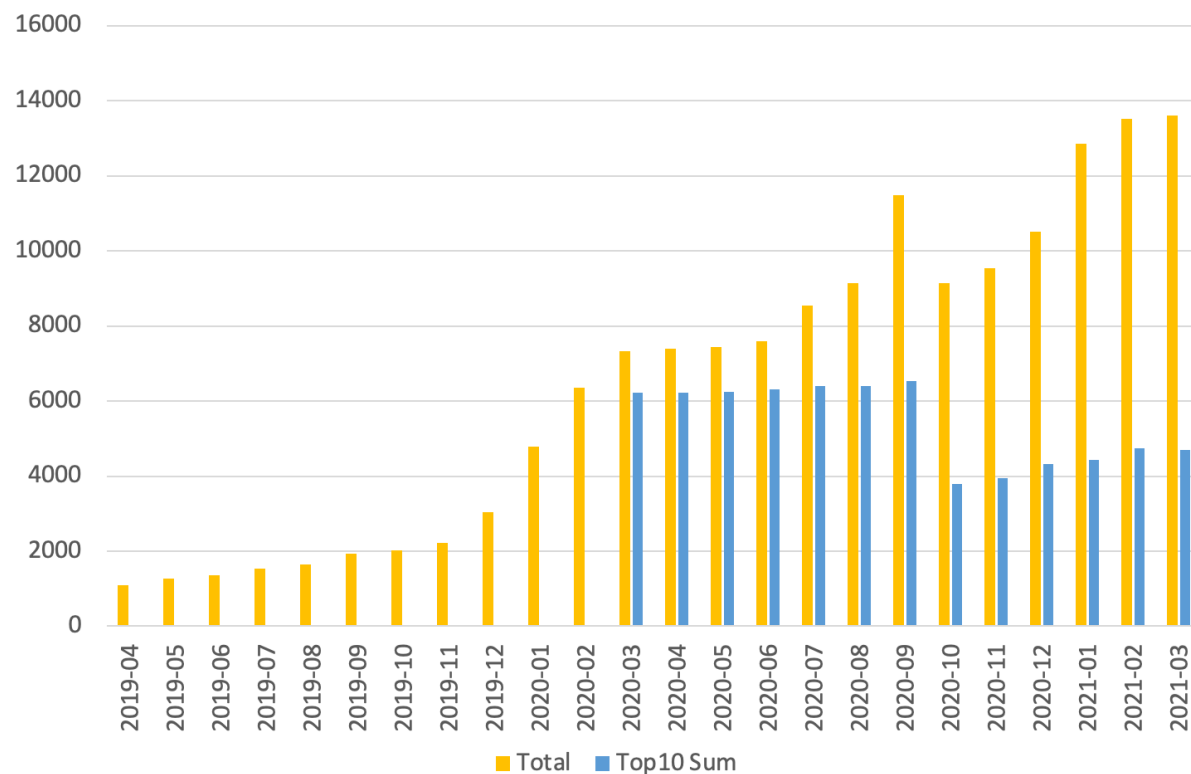
ストレージサービスの全体概要



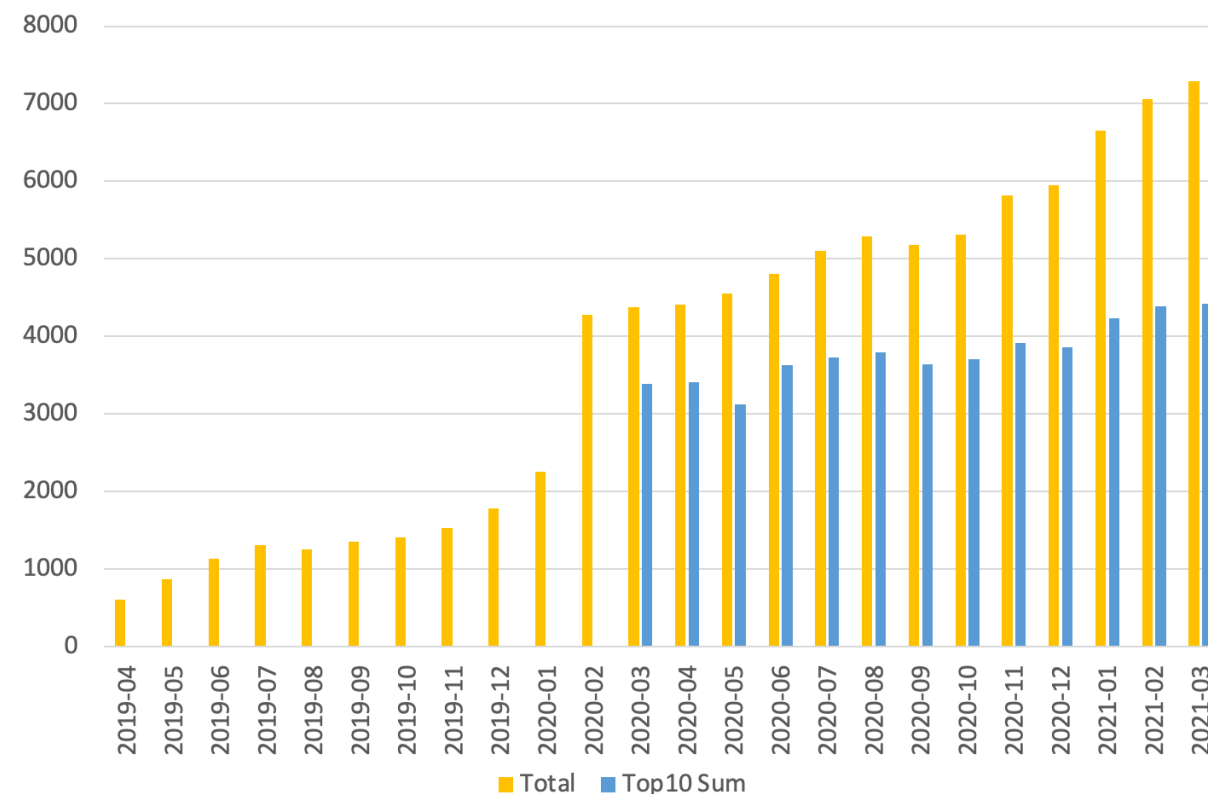
利用統計: Groups

- 基本的に増加傾向：利用者の増加、利用グループ内の使用量増加
 - 当初は特定グループの使用が大部分を占めていたが、全体的にも使用量が増えつつある。

使用容量(Tbytes)

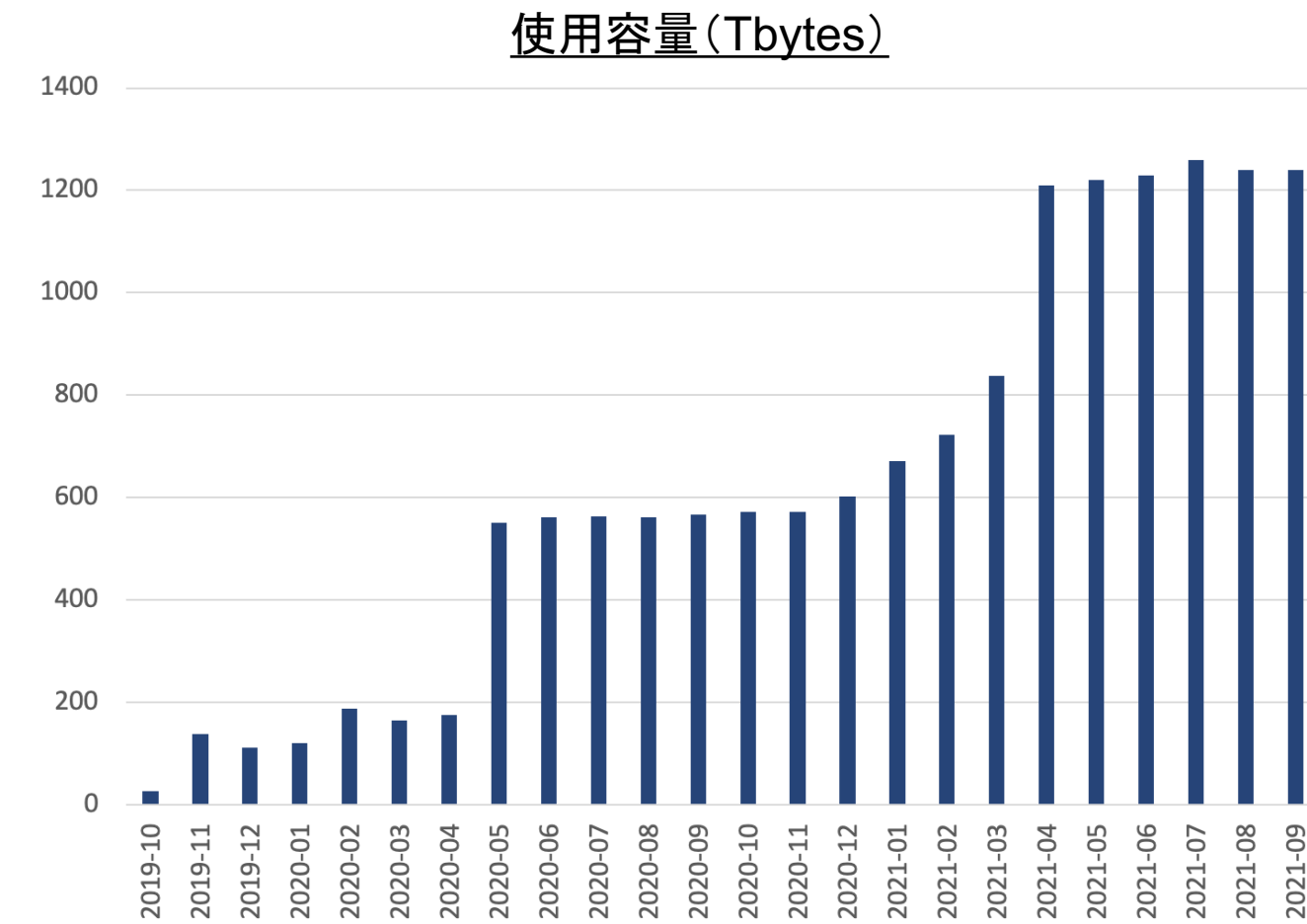


使用 inode 数 (million)



利用統計: Cloud Storage

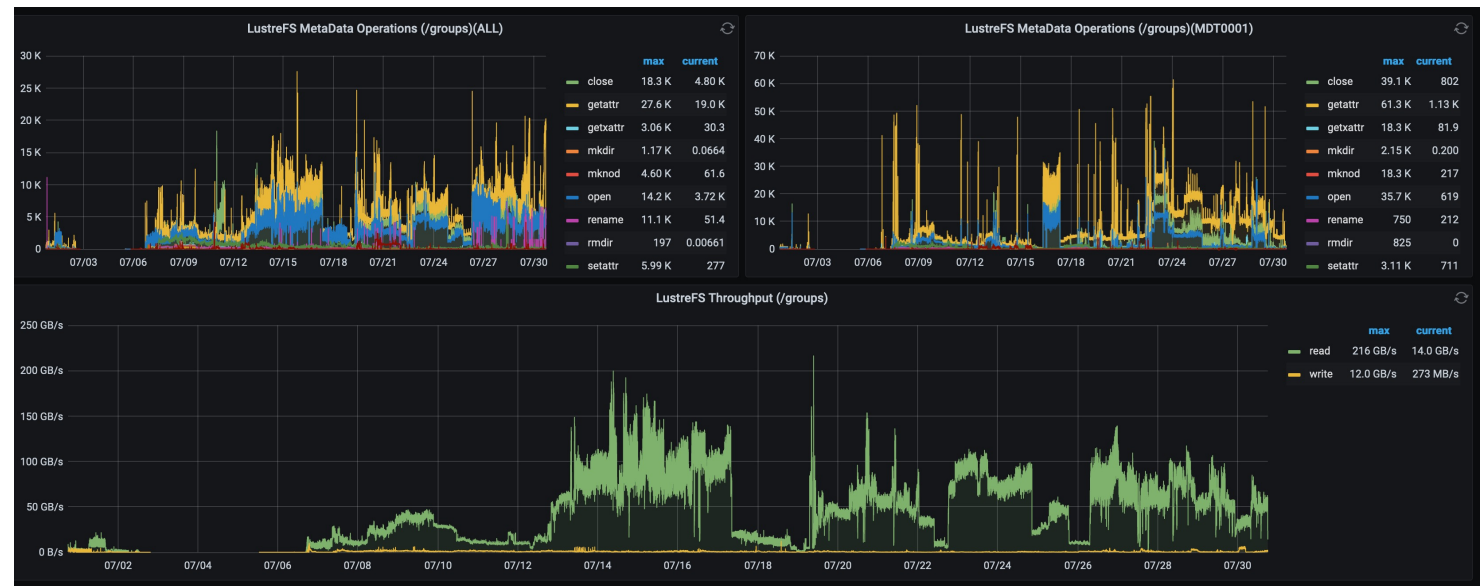
- 増加傾向。共有 FS に比べるとまだ利用は少ない。



モニタリング: 共有 FS

- 2021.4 以前は Spectrum Scale GUI により、Groups, BB をモニタリング
 - Read/Write の転送 Bytes、OPS 等をクライアントやストレージサーバのノード単位で確認可能
- ABCI 2.0 へのアップグレードにおいて DDN Monitoring Tool (LustrePerfMon) を導入
 - ファイルシステム単位でユーザ毎、ノード毎、ジョブ毎の I/O 情報を1 分間隔で取得
 - I/O 情報：open, close, stat 等のメタデータ操作、read, write 操作
 - Grafana による時系列での可視化
 - 利用者には未公開

運用チームによる異常検知は、
他の手段と組み合わせて実施



今後に向けて

- データや処理結果（学習済みモデル等）の共有支援
- 外部アクセスインタフェース
 - s3 以外のサービスのサポート、s3 の上位サービス
- セキュリティ
 - 各種のデータ保護規則要件に対応できる機能拡充（暗号化、隔離、トレース等）
- 共有ファイルシステムへのアクセス性能問題の改善
 - ローカルディスクの活用
 - 利用推奨、使い勝手の改善
 - 負荷集中を避ける仕組みや高速化の仕組みの導入
 - /scratch、GPUDirect Storage

