大阪大学のデータ集約基盤ONIONの概要

大阪大学サイバーメディアセンター
応用情報システム研究部門　伊達　進

大阪大学サイバーメディアセンター

CMC吹田本館

ITコア棟

- 大阪大学のスーパーコンピューティングセンター
  - 学内だけでなく学外の教育・研究組織や産業界と密接に連携したセンターとして機能することが求められた全国共同利用施設でもあり、その一環として、全国の大学の研究者が学術研究・教育に伴う計算及び情報処理を行うことができるよう、種々の高性能な大規模計算機システムを提供。

Applications from HPSC(High Performance Scientific Computing News)

http://www.hpc.cmc.osaka-u.ac.jp/hpsc-news/

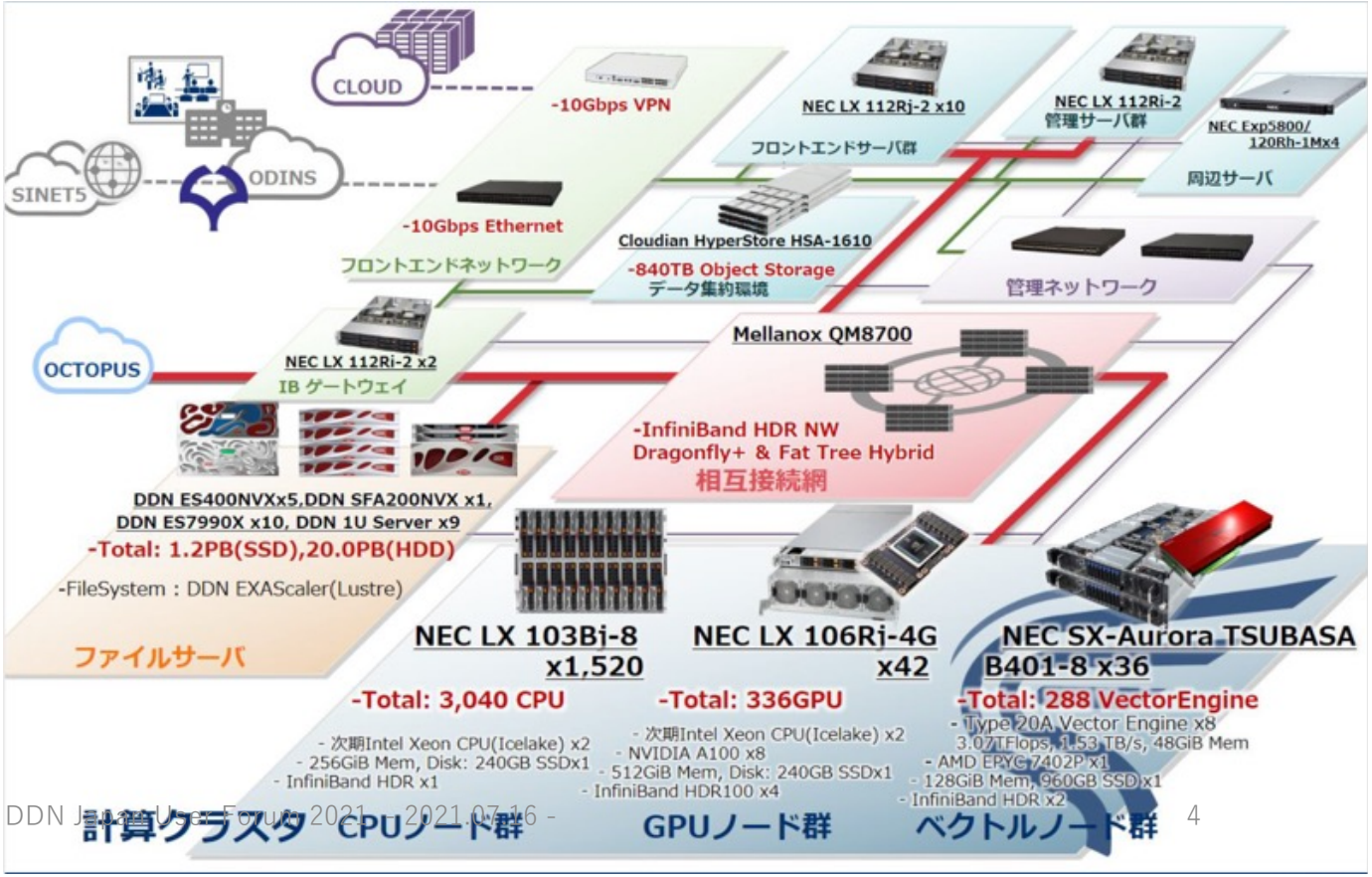
クラウド連動型HPC・HPDA用スーパーコンピュータシステムSQUID (Supercomputer for Quest to Unsolved Interdisciplinary Datascience)

Cybermedia Center
Osaka University

- わが国の学術・産業を支える研究者による未解決のデータサイエンス問題への探求を支援するため、高性能計算および高性能データ分析分野の多様な計算ニーズを収容可能なクラウド連動型スーパーコンピュータ

- 稼働
  - 2021年5月 ~

16.591 PFlops



POINT

PFlops(ペタフロップス)とは?

Flops(Floating-point Operations Per Second)は、1秒間に何回の浮動小数点演算を行えるかを示すコンピュータの計算能力を表す単位です。Peta(ペタ)は1000兆(10の15乗)を表す単位であり、1PFlopsは1秒間に1000兆回の浮動小数点演算を行うことができることを意味します。

SQUID (Supercomputer for Quest to Unsolved Interdisciplinary Datascience)　2021.05 ~

- わが国の学術・産業を支える研究者による未解決のデータサイエンス問題への探求を支援するため、**高性能計算(HPC: High Performance Computing)**および**高性能データ分析分野(High Performance Data Analysis)**の多様な計算ニーズを収容可能なクラウド連動型スーパーコンピュータ

  [特徴] 多様な計算ニーズを収容できる**3種混合(CPU, GPU, Vector)**のハイブリッド構成

ノード構成	汎用CPUノード群 1,520 ノード	プロセッサ：Intel Xeon Icelake 2基 主記憶容量：256GB
	GPUノード群 42 ノード	プロセッサ：Intel Xeon Icelake 2基 主記憶容量：512GB GPU：NVIDIA A100 8基
	ベクトルノード群 36 ノード	プロセッサ：AMD EPYC 7402P (2.8 GHz 24コア) 1基 主記憶容量：128GB Vector Engine：NEC SX-Aurora TSUBASA Type20A 8基
ストレージ	DDN EXAScaler (Lustre)	HDD：20.0 PB NVMe：1.2 PB
ノード間接続	Mellanox InfiniBand HDR (200 Gbps)	

※速報となりますので、性能に若干誤差がある場合がございます。予めご了承ください。

クラウド連動型HPC・HPDA用スーパーコンピュータシステムSQUID (Supercomputer for Quest to Unsolved Interdisciplinary Datascience)

Cybermedia Center
Osaka University

- 16.591 PetaFlopsの総理論演算性能

[**特徴**] 多様な計算ニーズを収容できる

3種混合(CPU, GPU, Vector)のハイブリッド構成

総演算性能	16.591 PFLOPS		
ノード構成	汎用CPUノード群 1,520 ノード(8.871 PFLOPS)	プロセッサ：Intel Xeon Platinum 8368 (Icelake / 2.40 GHz 38コア) 2基 主記憶容量：256GB	5.836 Tflops /node (2.918 Tflops/CPU) 0.83 kW/node
	GPUノード群 42 ノード(6.797 PFLOPS)	プロセッサ：Intel Xeon Platinum 8368 (Icelake / 2.40 GHz 38コア) 2基 主記憶容量：512GB GPU：NVIDIA A100 8基	161.836 Tflops /node (2.918 Tflops/CPU, 19.5Tflops/GPU) 5.4 kW/node
	ベクトルノード群 36 ノード(0.922 PFLOPS)	プロセッサ：AMD EPYC 7402P (2.8 GHz 24コア) 1基 主記憶容量：128GB Vector Engine：NEC SX-Aurora TSUBASA Type20A 8基	26.71 Tflops /node (2.150 Tflops/CPU, 3.07 Tflops/VEC) 3.3 kW/node
ストレージ	DDN EXAScaler (Lustre)	HDD：20.0 PB NVMe：1.2 PB	
ノード間接続	Mellanox InfiniBand HDR (200 Gbps)		

SQUID
Supercomputer for Quest to Unsolved Interdisciplinary Datascience

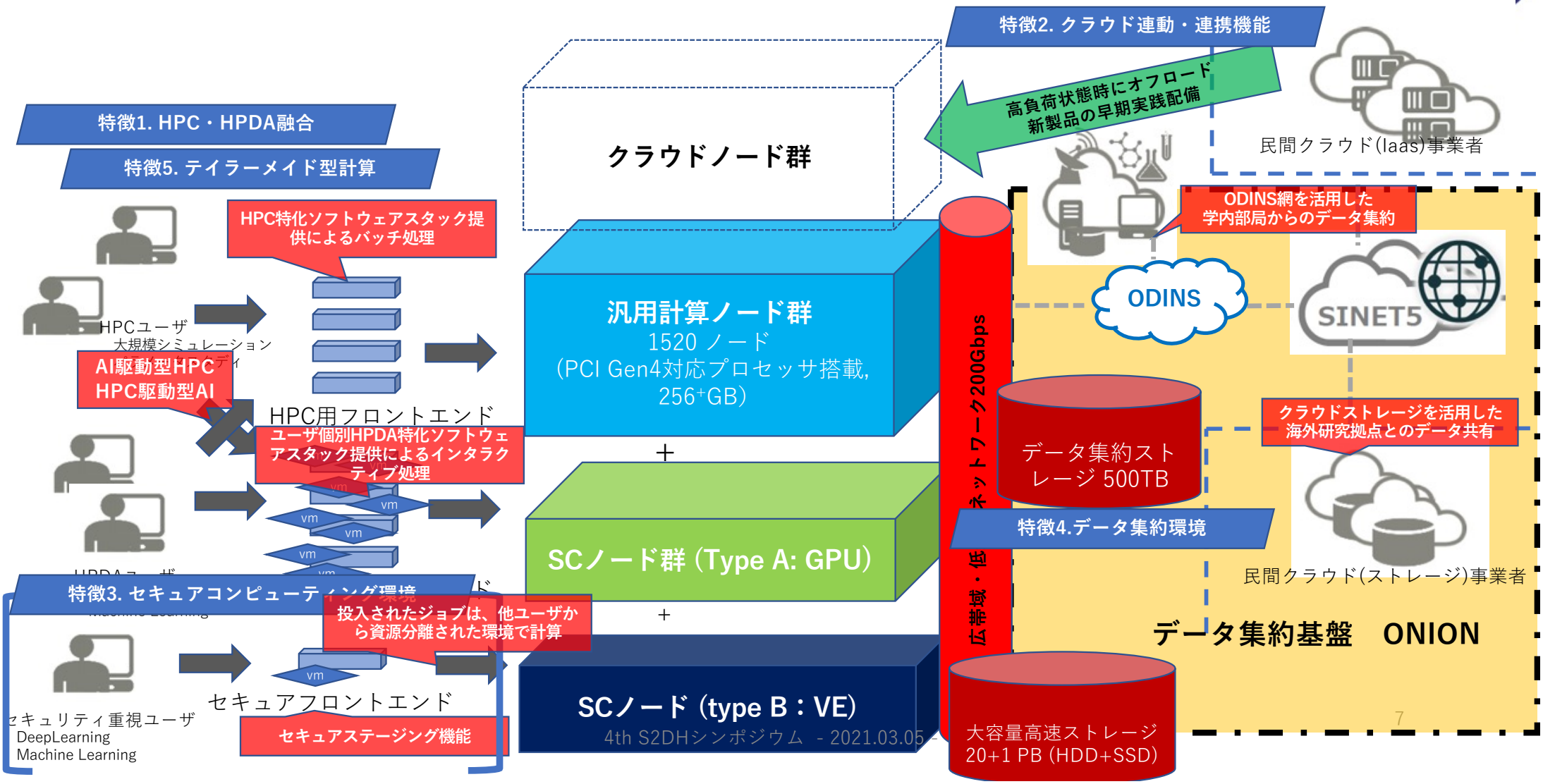
OCTOPUS
Osaka university Cybermedia cenTer Over-Petascale Universal Supercomputer

1.996 Tflops /node
(0.998 Tflops/CPU)
0.44 kW/node

5.836 Tflops /node
(2.918 Tflops/CPU,
5.3Tflops/GPU)
1.89 kW/node

276 Gflops /node
0.56 kW/node

SQUID+ONION 5つのチャレンジ



高性能計算・データ分析基盤システムの5 key concepts

Cybermedia Center
Osaka University

1. 高性能計算(HPC)/高性能データ分析(HPDA)融合

- 高性能計算という従来型の計算ニーズだけでなく、ビッグデータ、AI、ディープラーニング、機会学習等のキーワードに代表される高性能データ分析という新たな計算ニーズを収容

2. クラウド連動・連携機能

- 計算負荷のオフロードによる待ち時間の軽減
- 新規ハードウェア・ソリューションの早期実戦配備

3. セキュアコンピューティング環境

- 計算資源およびネットワーク資源の分離・統合により、高機密性・秘匿性データをより安全に解析・計算処理
- セキュアステージング

4. データ集約環境

- キャンパス内の科学計測機器や各種IoTセンサ等の各種データソースからシームレスにデータを集約し、高性能計算・データ分析基盤システムで活用
- 解析・計算の結果データを学外・海外研究拠点の研究者と共有

5. テーラーメイド型計算環境

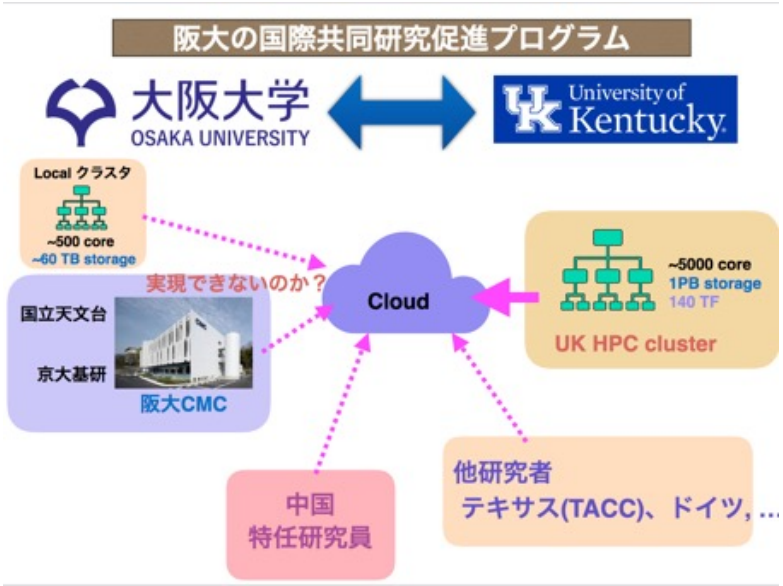
- HPC/HPDA研究者がそれぞれの必要とする計算環境（ソフトウェアスタック）を利用可能

データ集約基盤ONION

ONION導入の背景 (1)

- **学術研究の広域化・グローバル化、産学共創への期待**
  - 共通の課題解決にむけて世界の研究者と協働する国際共同研究

地理的に分散する計算資源、データ資源を駆使して、研究開発できる環境 （昔グリッドと言っていた環境？）

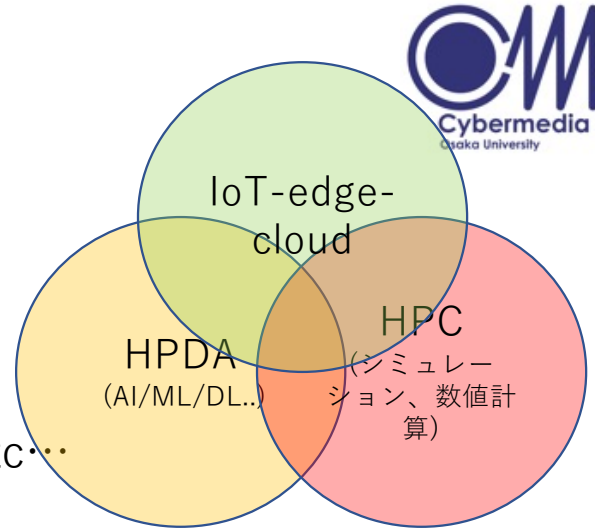


長峯健太郎, "理学研究とクラウド利用のニーズ:理論宇宙物理学の例", Cyber HPC Symposium 2019, Mar. 2019. 発表スライドより。

OUビジョン2021 - 社会変革に貢献する世界屈指のイノベーティブな大学へ -, https://www.osaka-u.ac.jp/ja/oumode/OUvision2021/y66s8j

ONION導入の背景 (2)

- **HPDA分野の計算ニーズの急速な高まり**
  - AI/MLを活用するデータ駆動型サイエンス
  - 各種IoTセンサを利活用したビッグデータ解析
  - HPC for AI, AI for HPC
    - データ同化シミュレーション, AIによる計算結果の判定, etc…



多様なデータをスーパーコンピューティング環境に“集約”できるデータ基盤
＝ 学術研究のデータフローを妨げないデータ基盤
＝ 利用者のデスクトップ環境、クラウド、スパコン環境を問わず
データ（計測データ、計算結果、等）を簡易に移動・管理できる環境

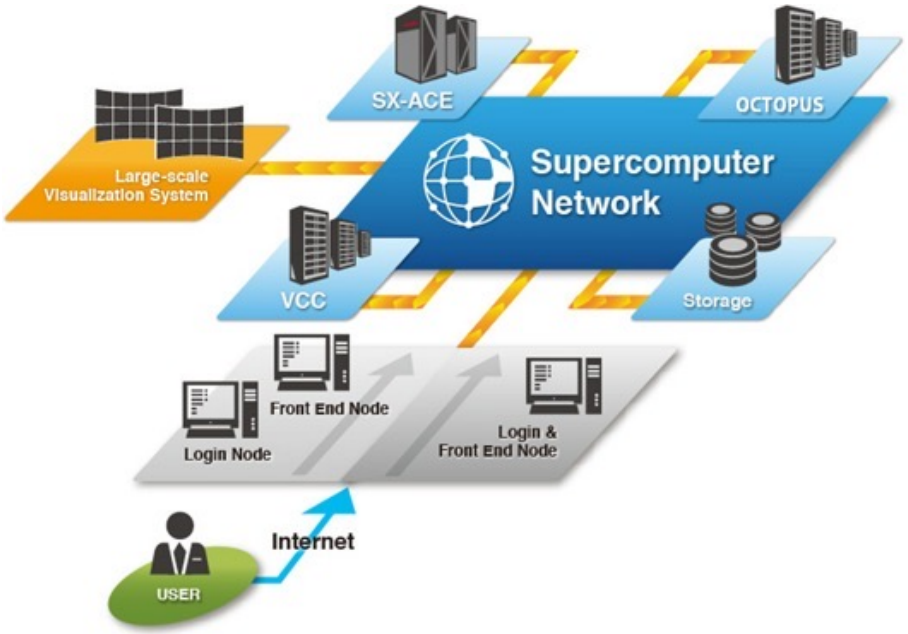
ONION導入の背景 (3)

- **研究データ保存, Reproducibilityに向けた動きの活発化**
  - 公正な研究活動の推進にむけて、研究活動に伴い作成・取得した研究データの保存期間および管理方法等についての基準を定めたガイドライン策定
  - 様々な科学データ計測機器がキャンパス内に存在しており、そのデータ管理・移動が課題
  - 大容量・大規模データを取り扱うSCセンターとしての役割・責任

近い将来、 SCセンターとしてデジタル化された研究データを適正に管理するためのデータ基盤を提供する必要がある？

現状のSC環境@CMC

- **外部からIsolatedで independentな計算環境**
  - スーパーコンピュータ群はログインノード＆フロントエンドノード経由でのみ利用
  - SC環境へのデータ移動、SC環境からのデータ移動はscp, sftp経由が基本
    ※HPCI供用ストレージは利用可能
  - SX-ACE(NEC ScateFS) – OCTOPUS(Lustre)間はNFSでデータ移動可能



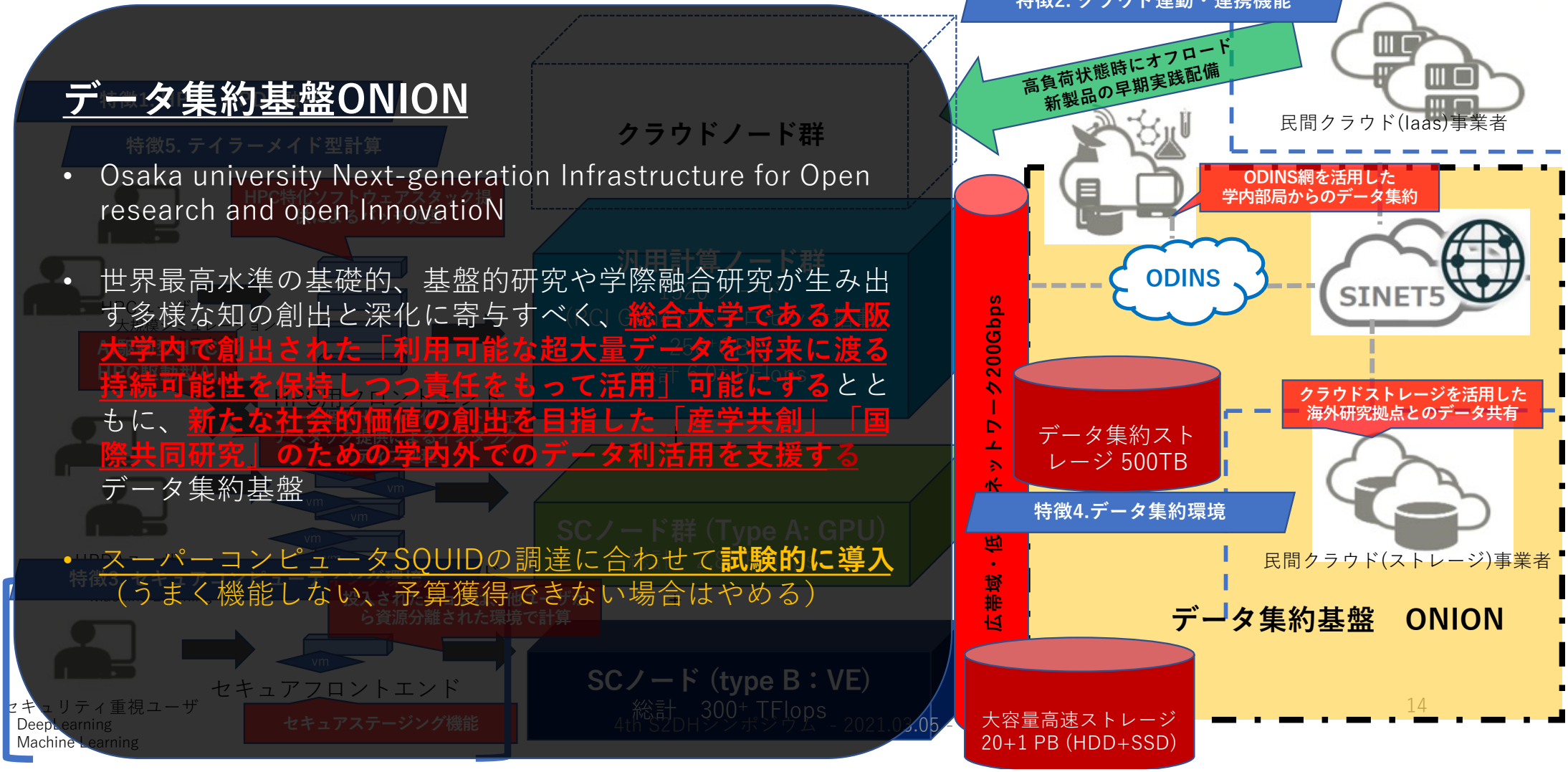
従来までのHPC分野からだけでなく、AI/ML/DLなどのHPDA分野の新しい計算ニーズの登場により、SCP, SFTPだけでなく、データ移動を自動化・遠隔利用化できる仕組みへの要望が急拡大

- 科学計測機器で計測されたデータを簡易に学内外の研究者と共有できる
- 多様なデータアクセスプロトコルを隠蔽したGUIでデータ操作できる
- クラウドストレージと連動できる。
- etc

SQUID+ONION 5つのチャレンジ

データ集約基盤ONION

- Osaka university Next-generation Infrastructure for Open research and open InnovatioN
- 世界最高水準の基礎的、基盤的研究や学際融合研究が生み出す多様な知の創出と深化に寄与すべく、**総合大学である大阪大学内で創出された「利用可能な超大量データを将来に渡る持続可能性を保持しつつ責任をもって活用」可能にする**とともに、**新たな社会的価値の創出を目指した「産学共創」「国際共同研究」のための学内外でのデータ利活用を支援する**データ集約基盤
- スーパーコンピュータSQUIDの調達に合わせて**試験的に導入**（うまく機能しない、予算獲得できない場合はやめる）

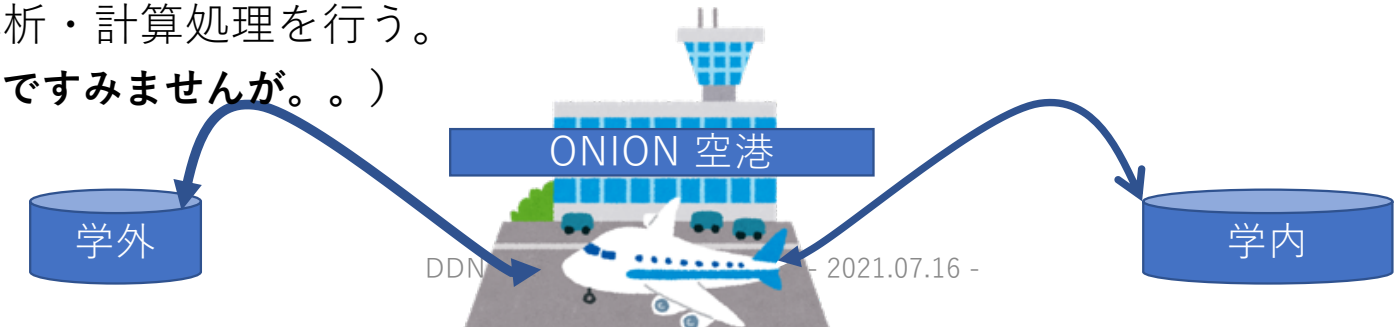


ONION

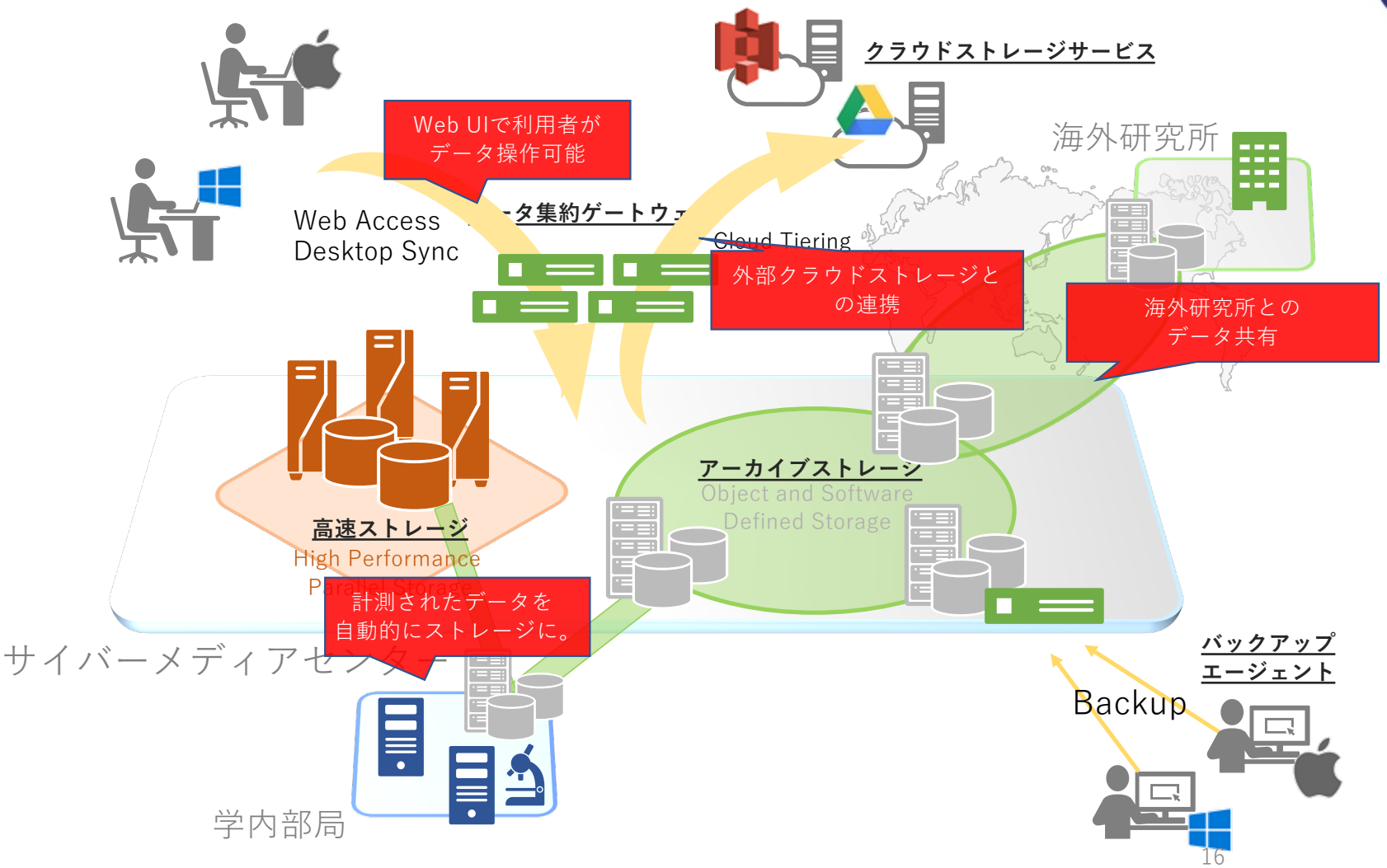
- 産学共創、国際共同研究に向けたデータ利活用を支援するデータ基盤
- SQUIDからのデータ移動、SQUIDへのデータ移動を容易にするデータ基盤
  - **高速並列ファイルシステム(21PB)** と**オブジェクトストレージ(500TB)** から構成
    - 高性能な計算・解析要求で必要とされる 高速データアクセス性能(read, write, iops)を並列ファイルシステム、データ操作の容易性をオブジェクトストレージでサポート。
  - データ操作に関するプロトコルを隠蔽した**GUI(graphical user I/F)**の提供

- 学内の科学計測機器からの計測データをSQUIDの高速並列ファイルシステムあるいはオブジェクトストレージに保存・格納する。
- SQUIDでの解析・計算結果を民間クラウドのストレージに保存・格納し、学外の研究者に公開する。
- 民間クラウドのストレージに置かれているデータを高速並列ファイルシステムに移動させ、SQUIDで解析・計算処理を行う。

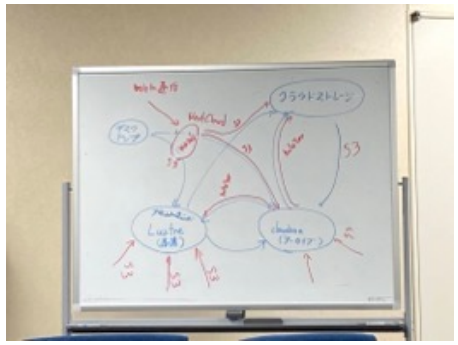
イメージ（お粗末ですみませんが。。）



データ集約基盤 ONION 概要



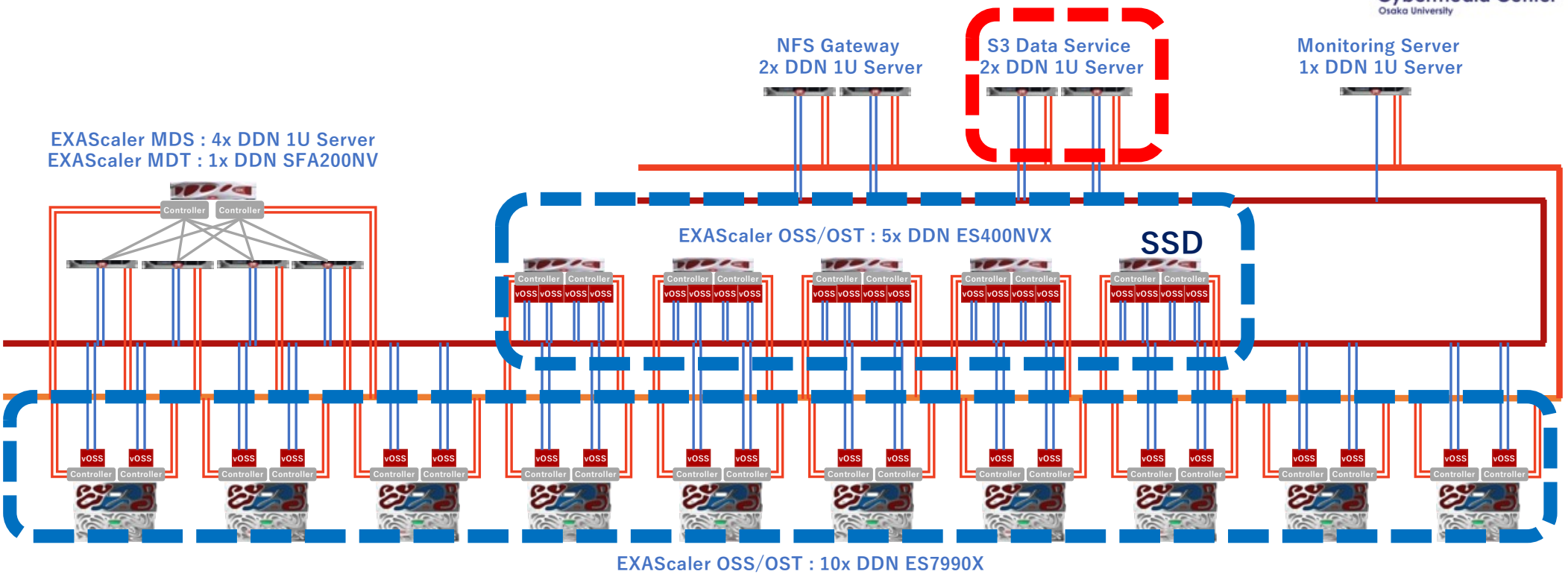
設計中のONION



http
S3
webdav
Nextcloud
S3
S3
S3
S3
NFS/CIFS?/S3
学内別拠点
S3
S3
CLOUDIAN HYPERSTORE
センサーエージェント

Lustre w/ S3DS

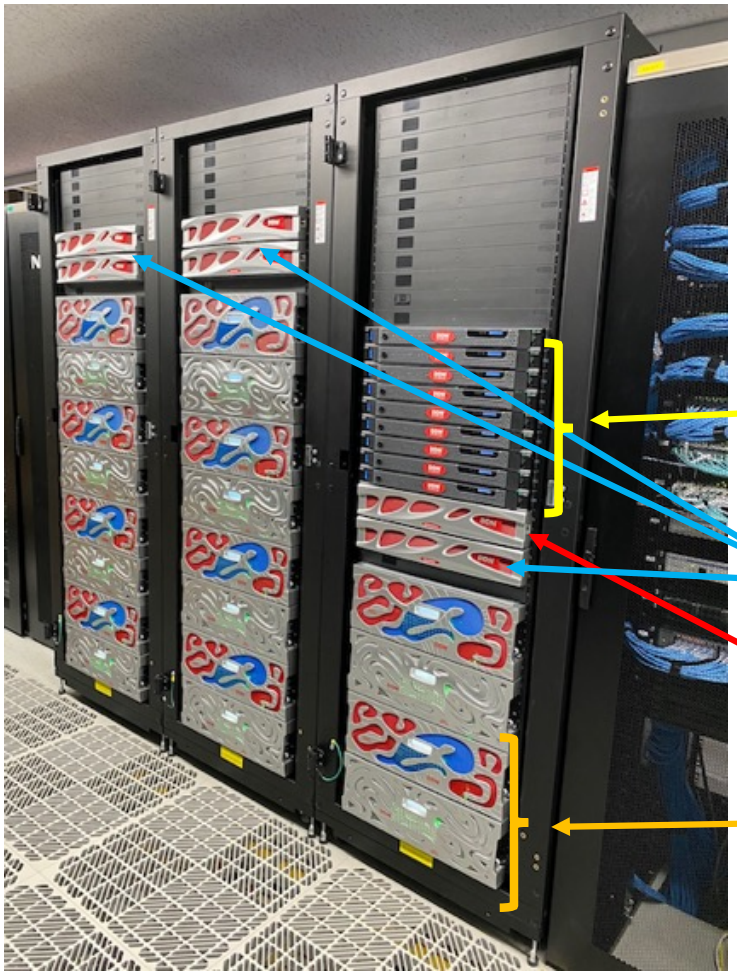
ONION構成Filesystem (DDN ExaScaler)



実効容量(HDD)	20.00PB
実効容量(NVMe)	1.20PB
最大合計inode数	約80億個
最大想定実効スループット(HDD)	160GB/s以上
最大想定実効スループット(NVMe)	Write : 160GB/s以上, Read : 180GB/s以上

ONION構成Filesystem (DDN ExaScaler) 外観

- 20PB(HDD)+1PB(SSD)PBを３ラック
  - 1700本のHDD(16TB)
  - 105本のSSD(15.36TB)



モニタリングサーバ (1個)
NFSサーバ (2個)
S3DSサーバ (2個)
MDSサーバ (4個)

SSD用ES400NVX（5個）

MDT用SFA200NV(1個)

HDD用ES7990X（10個）
HDD用SS9012 （10個）

2021年3月1日頃

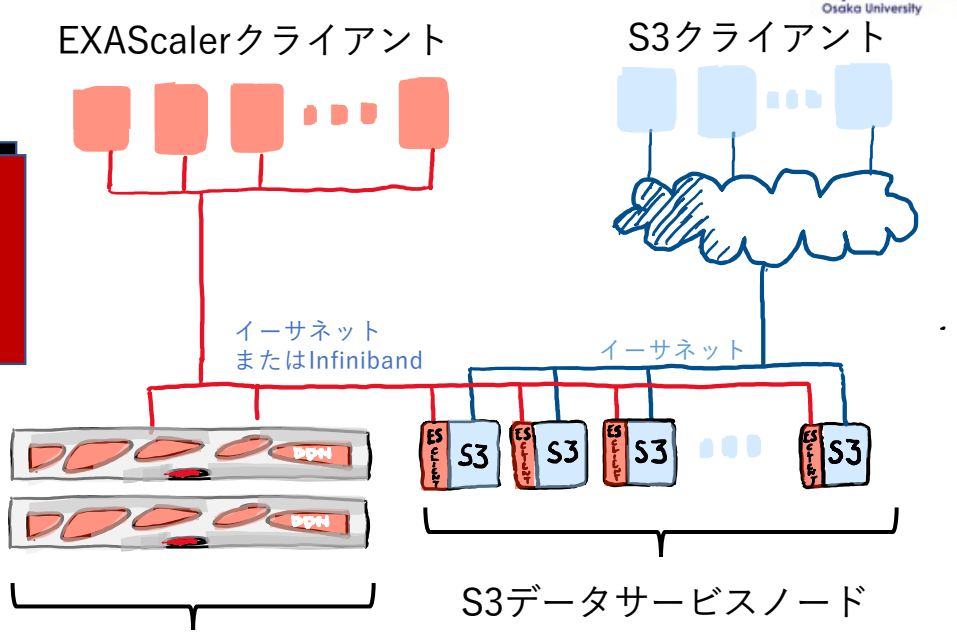
2021年3月3日頃

S3DS (S3データサービス)

- ONIONを支える基盤技術

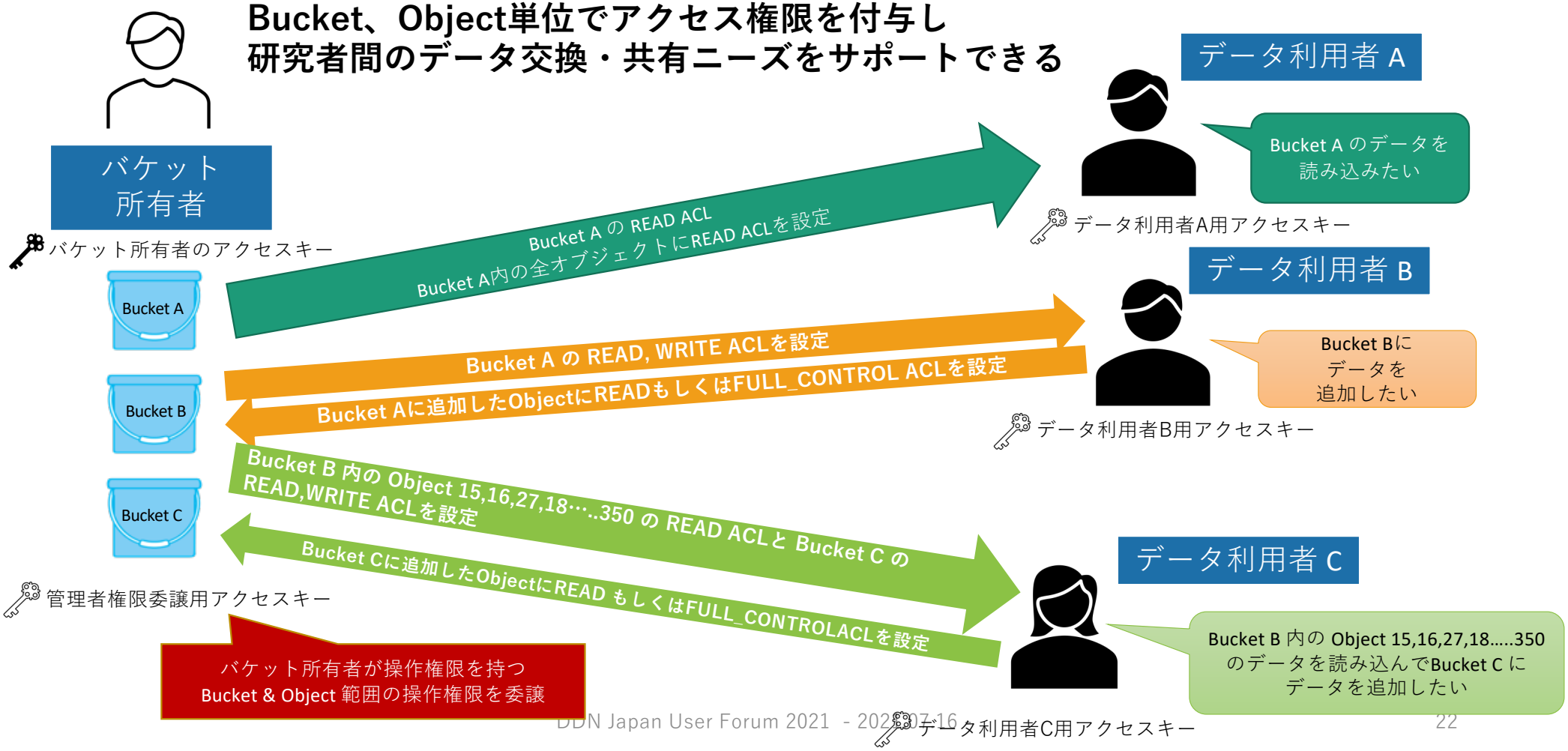
並列ファイルシステムに利用者が慣れ親しんだPOSIXのデータアクセス方法で、外部のクラウドストレージ等のS3対応ストレージのデータを扱える仕組みを補強

- Lustreストレージ上でS3 APIを提供
- ファイルとして管理
- S3とPOSIXの名前空間を共有
  - S3とPOSIXのデータアクセスの統合
  - 双方向からのデータアクセス
  - S3もしくはLustreからのデータ書き込み、読み込み



Nobu Hashizume, "DDN Update @PCCC20", 第20回PCクラスタシンポジウム, Dec. 2020.

S3の細粒度でのデータアクセス特性(2)

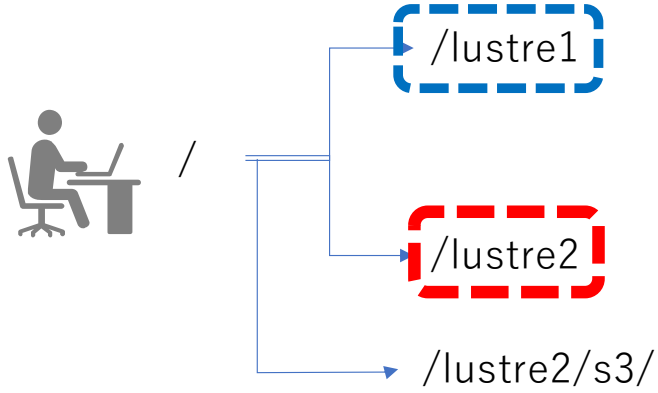


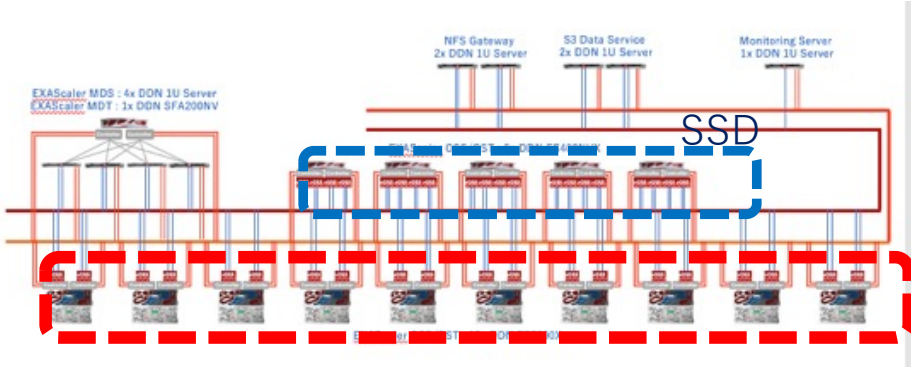
ONION構成Filesystem

- Lustreファイルシステム構成

領域	ファイルシステム名	MDS	MDT数	ファイル数	DNE	OSS数	OST数	容量	機器
SSD	lustre1	#1-2	1	約20億	No	40	40	約1.2PB	ES400NVX
HDD	lustre2	#3-4	3->2	約60億	Yes	80	80	約20PB	ES7990X

- SSD, HDD毎にファイルシステムを構成
- HDDは5TBをデフォルトで利用者申請グループ単位で提供
- SSDを希望される利用者には1TB単位で提供
- ディレクトリの違いでSSD/HDDを利用者は区別



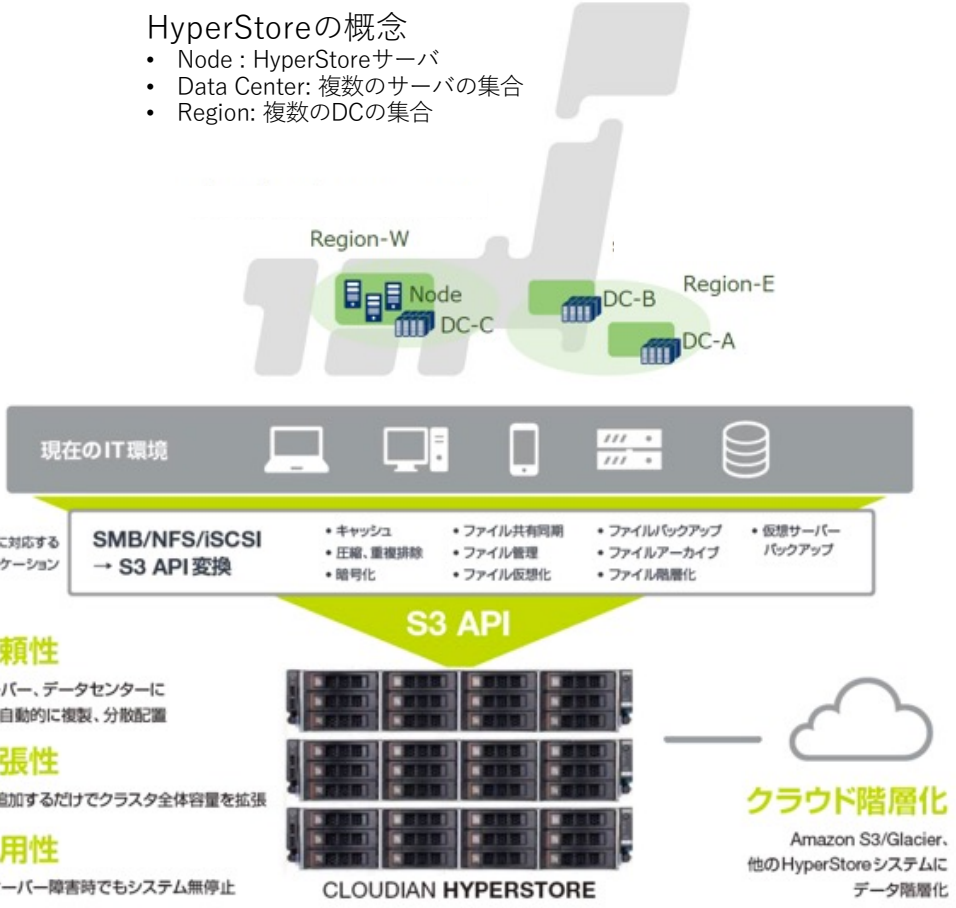


ONION構成ObjectStorage (Cloudian HyperStore)

- ONIONを支える基盤技術

S3対応型オブジェクトストレージ
高い拡張性
単一障害点のない完全分散処理
高い保護強度
　イレージャーコーディング(DC内)
　レプリケーション(DC内, DC間)

HyperStoreの概念
- Node : HyperStoreサーバ
- Data Center: 複数のサーバの集合
- Region: 複数のDCの集合



ONION ゲートウェイ： NextCloud

- ONIONを支える基盤技術

ONION構成ファイルシステム(DDN)、ONION構成オブジェクトストレージ(Cloudian)にデータをウェブブラウザを通じてアップロード/ダウンロード可能な機能を補強。外部ストレージ連携機能により、クラウドストレージにも対応。

NextCloudを通じて、
オンプレミスストレージと
クラウドストレージを利用可能

S3対応ストレージ

ONION ユースケース(1) : 共同研究者からのデータ共有

- (阪大CMCの利用者アカウントのない)共同研究者に計算に必要なデータをアップロードしてもらう

https://onionweb.hpc.cmc.osaka-u.ac.jp/index.php/s/egbjLMTpmzkwk3H

アクセス&アップロード

password: zpWBzxEqEqY9

共同研究者A

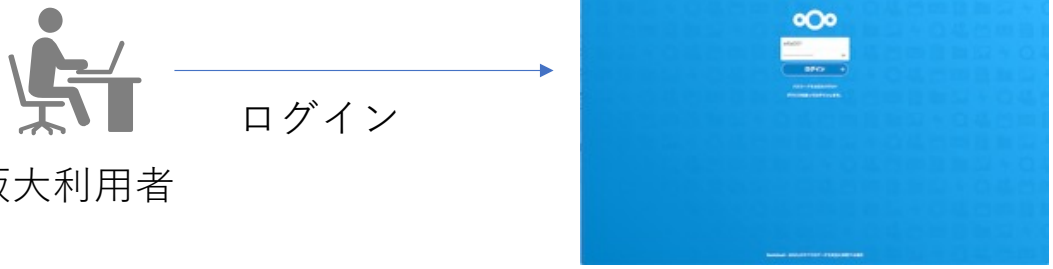
password: zpWBzxEqEqY9

共同研究者A

共同研究者B

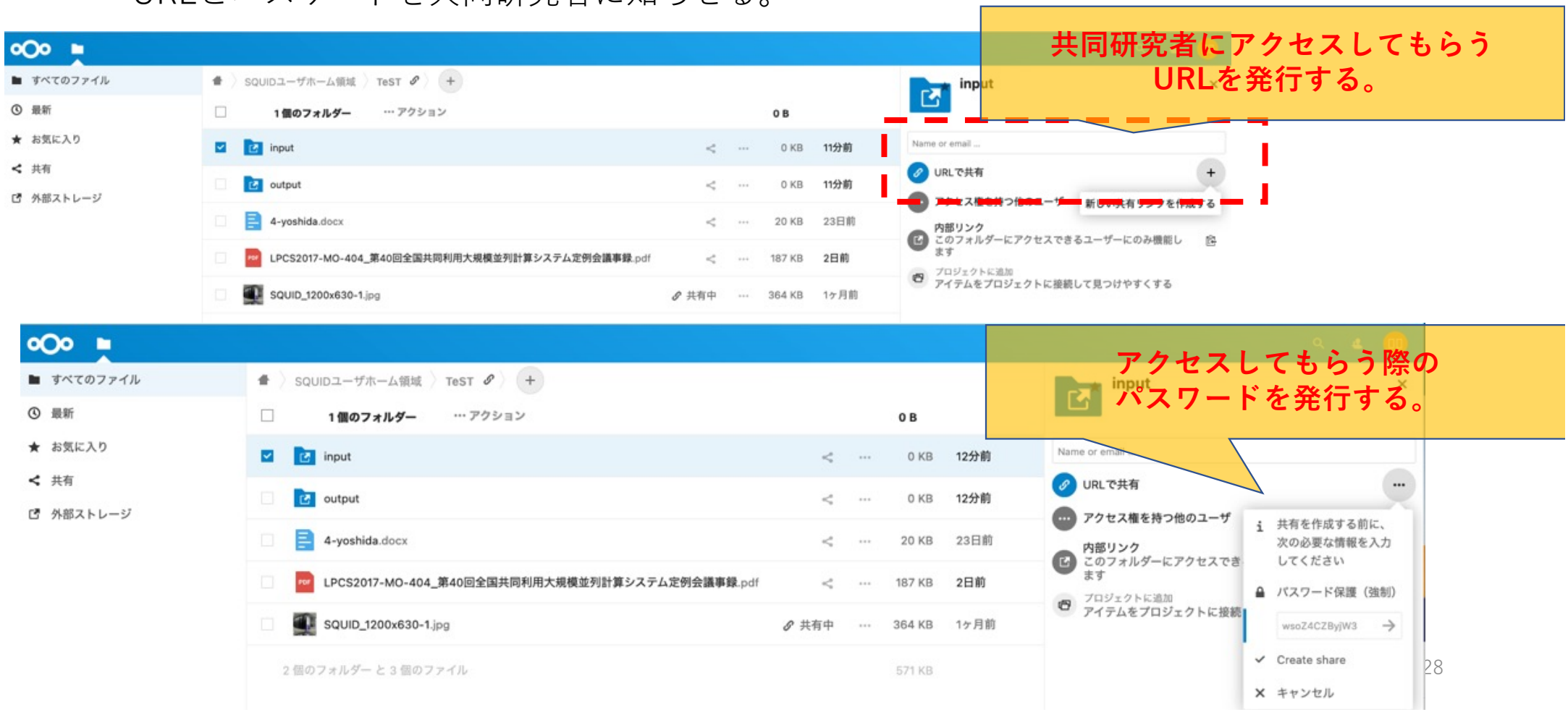
ONION ユースケース(1) : 共同研究者からのデータ共有

1. NextCloudのウェブI/F (https://onionweb.hpc.cmc.osaka-u.ac.jp)にログインして、共同研究者にデータをアップロードしてもらうディレクトリを作成する



ONION ユースケース(1) : 共同研究者からのデータ共有

2. アップロードしてもらうディレクトリの共有設定を行い、ディレクトリにアクセスするためのURLとパスワードを共同研究者に知らせる。



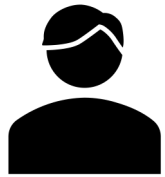
ONION ユースケース(1) : 共同研究者からのデータ共有

3. 共同研究者にアクセス＆アップロードしてもらう。

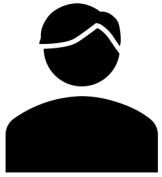
https://onionweb.hpc.cmc.osaka-u.ac.jp/index.php/s/egbjLMTpmzkwk3H

password: zpWBzxEqEqY9

アクセス&アップロード



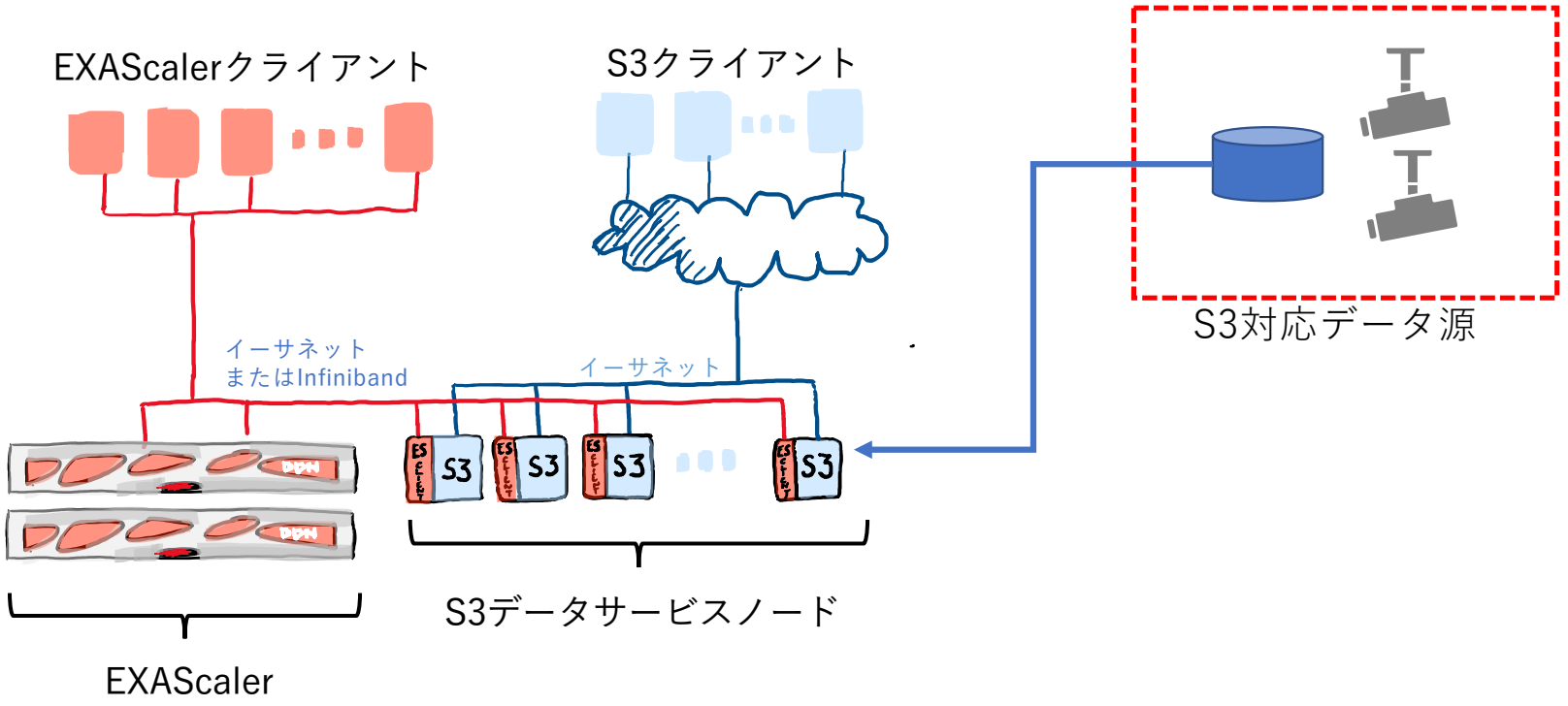
共同研究者A



共同研究者B

ONION ユースケース(2): S3対応データ源からPFSへのデータ移動

- 外部のS3対応データ源からデータをLustreファイルシステム(/sqfs/s3/)に格納する。



ONION ユースケース(2): S3対応データ源からPFSへのデータ移動

1. S3DSを使うためのアクセスキーとシークレットキーを作成する。

阪大利用者
on フロントエンドノード

```
[w6a001@squidhpc2 ~]$ s3dskey create --group=xxxxxxxx
```

アクセスキー/シークレットキーが生成される

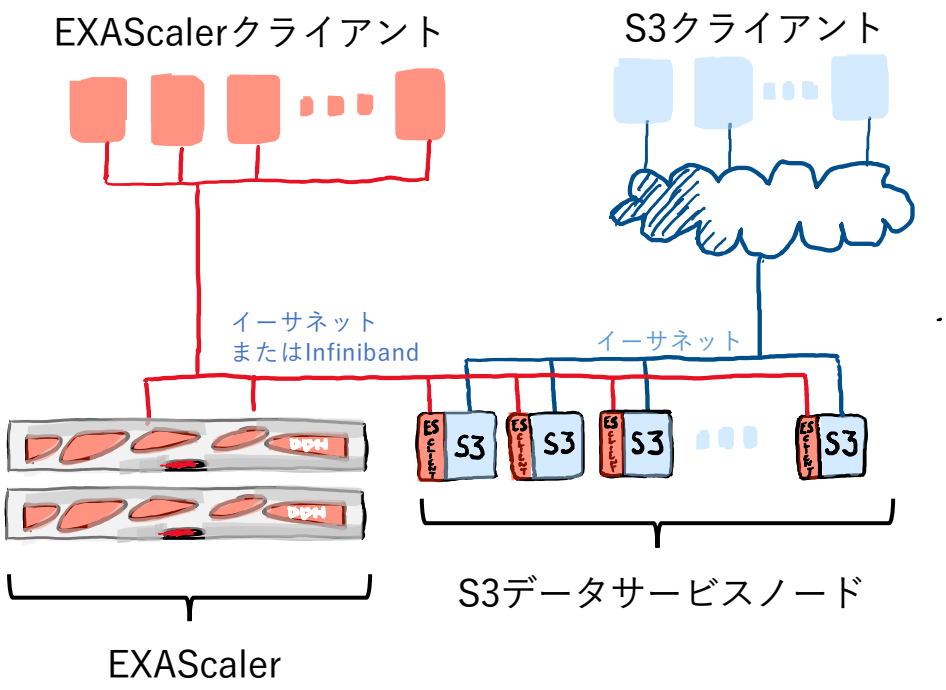
```
+-----------+------------------------------------------+
| accesskey | AKIA7X1KXXXYYYZZZ000                     |
| enabled   | True                                     |
| fspaths   | None                                     |
| fsuid     | 60101:10                                 |
| secretkey | ****************************************** |
| tag       | (連番):(ユーザ名):(グループ名)             |
| uuid      | ea60f00a60hufaweofapo12813nfawe9506e1216 |
+-----------+------------------------------------------+
```

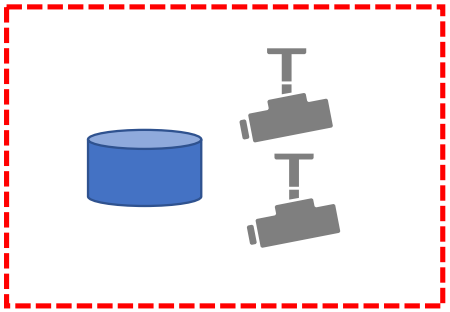
アクセスキー

シークレットキー

ONION ユースケース(2): S3対応データ源からPFSへのデータ移動

2. S3対応データ源側の設定を行う





S3対応データ源

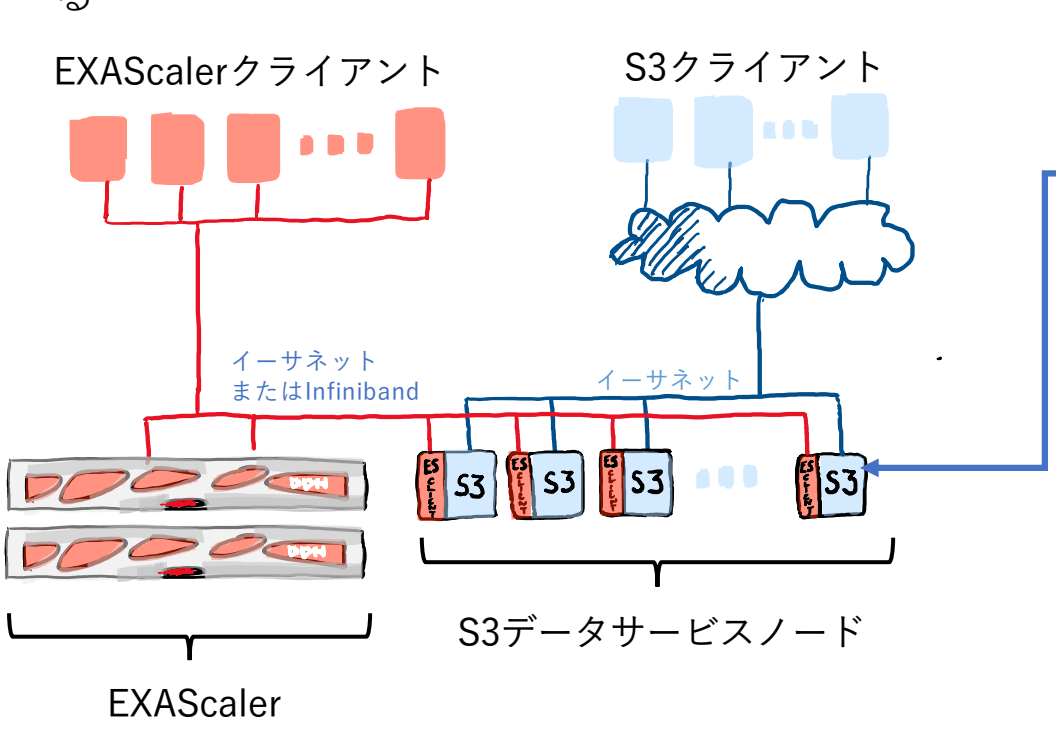
% vi ~/.s3curl

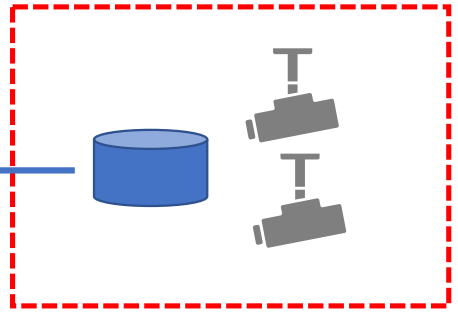
```
%awsSecretAccessKeys = (

username => {
            id => 's3dskeyコマンドで発行したAccess Key',
            key => 's3dskeyコマンドで発行したSecret Key',
            },
);
push(@endpoints, 'squidgw.hpc.cmc.osaka-u.ac.jp')
```

ONION ユースケース(2): S3対応データ源からPFSへのデータ移動

3. SQUID GW (squidgw.hpc.cmc.osaka-u.ac.jp/)にアクセスし、S3クライアントツール(s3curlなど)を用いてデータ(datafile)を置きたいバケット(bucketT)を作成し、データをアップロードする





バケットの作成

% s3curl.pl --id **username** --createBucket
-- -s https://squidgw.hpc.cmc.osaka-u.ac.jp/**bucketT**

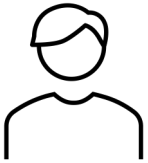
データのアップロード

% s3curl.pl --id **username** --put **datafile** --
-s -v https://squidgw.hpc.cmc.osaka-u.ac.jp/**bucketT**/
datafile

ONION ユースケース(3): ジョブ終了後に計算結果をスマホで確認

- ジョブ終了後に計算結果をスマホで確認し、即座に共同研究者と共有

①スケジューラから
ジョブ完了通知メール受信



②ONION webにログイン
(ブラウザ or アプリ)

③対象ファイルを確認

④ 共有設定する

まとめ

- **データ集約基盤ONION**
  - 産学共創、国際共同研究に向けたデータ利活用を支援し、大阪大学OU Vision 2021のOpen Research、Open Innovationを実現するためのデータ基盤となるよう、**Osaka university Next-generation Infrastructure for Open research and open InnovatioN**と命名
  - DDN Lustreファイルシステム、Cloudian HyperStoreオブジェクトストレージ、NextCloudを基盤技術として構成
  - SQUID調達に合わせて試験導入