



大型放射光施設のデータ処理を支えるシステムソフトウェア技術

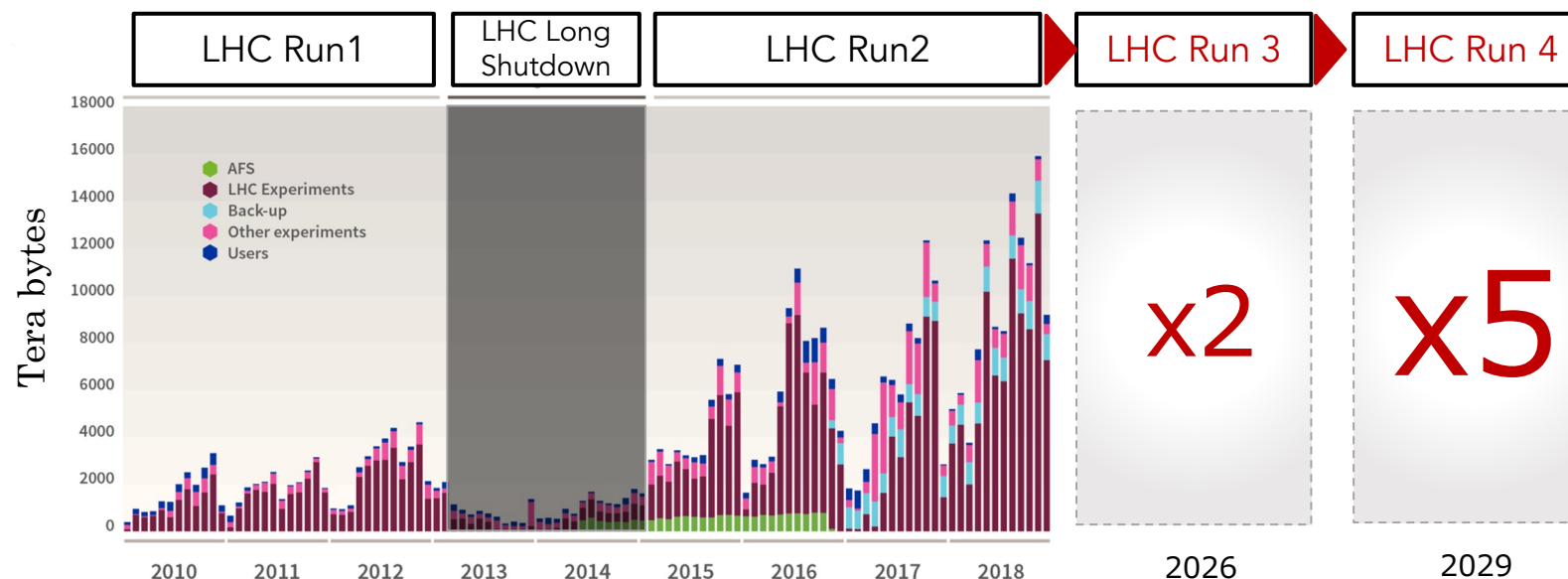
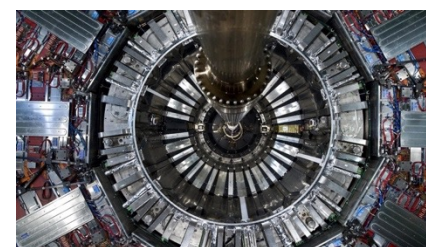
PCクラスタワークショップ in 大阪2023「ビッグデータとHPC」

2023年6月21日-22日

理化学研究所 高性能ビッグデータ研究チーム
佐藤 賢斗

● データ生成:

- CERNのLHC (Large Hadron Collider) では、2018年に約88PBのデータが生成 [1]
- “Data archival is expected to be two-times higher during Run 3 and five-times higher or more during Run 4 (foreseen for 2026 to 2029).”



● データ転送

- 生成されたデータは解析のために施設内部および外部の計算機へ転送する必要がある
 - e.g.) LHCでは 830 PB / 11億ファイルが転送された [1]
- 外部機関へデータ共有する場合, WAN (インターネット) を介してデータを転送



Sensors

Big data
transfer



Internal/External Computers

効率的なデータ転送やその管理はビッグデータ解析において重要

SPring-8 / SACLA

- 理研では、大型放射光施設(SPring-8)やX線自由電子レーザー施設(SACLA)を所有
 - Opened in 1997 in Harima, located in the same Hyogo prefecture as R-CCS
 - Managed by RIKEN, with Japan synchrotron radiation research institute (JASRI)
 - SPring-8 stands for Super Photon ring-8 GeV
 - 8 GeV (giga electron volts) is the energy of electron beam circulation in the storage ring



SPring-8での典型的な実験・解析と課題



2) 大量高精細データ 解析律速の解消



先駆的研究



今回の措置後



解析律速の原因と対策

計算が遅い

→ 計算科学・リソースの確保

良い解析方法がない。新しい解析方法の探索

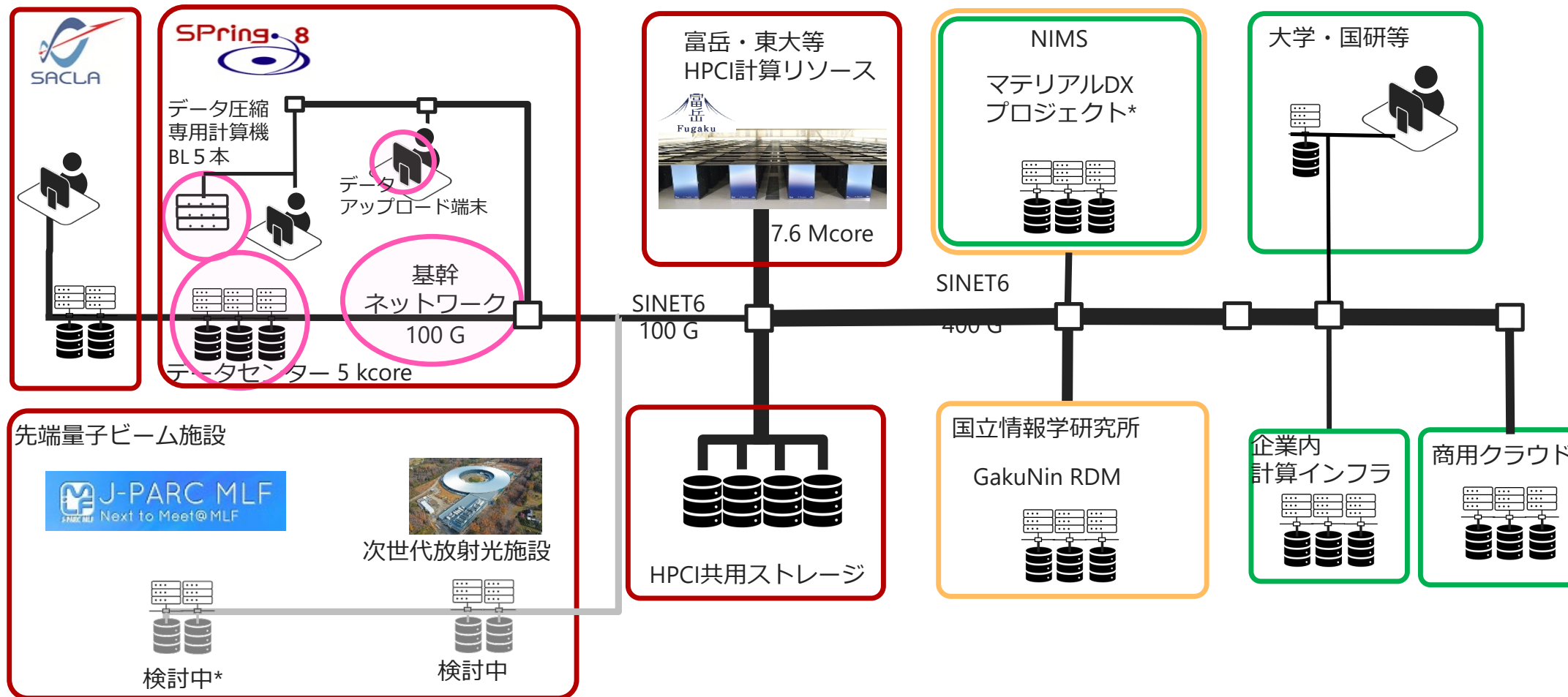
→ データのアクセス基盤の整備

→ 統計・情報科学（狭義のデータサイエンス）

T. Hatsui, RIKEN

7

SPring-8データセンター構想: インフラ



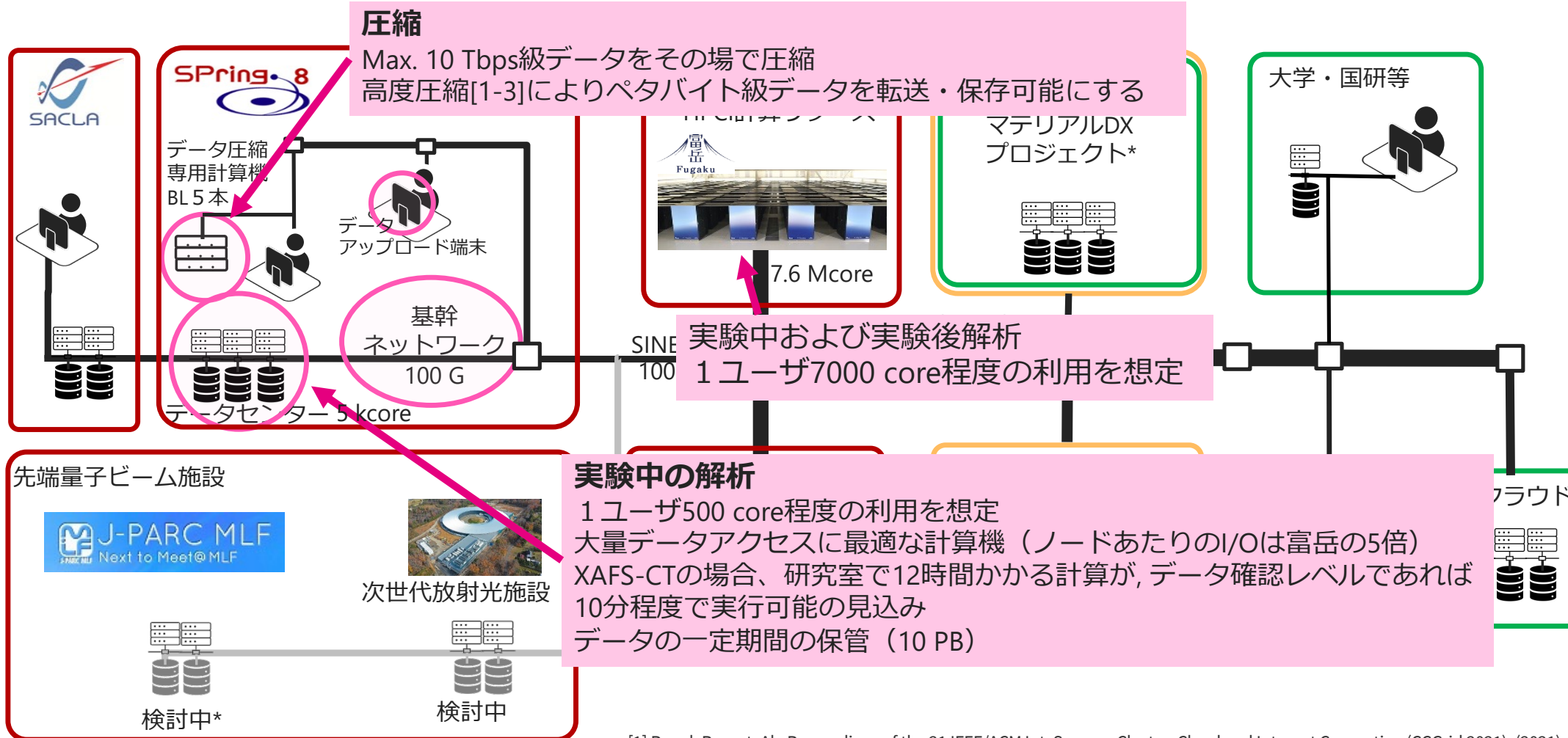
* クラウド上に設置を予定

T. Hatsui, RIKEN

9

スライド提供: 初井宇記(RSC)、城地保昌(RSC)

SPring-8データセンター構想: インフラ



* クラウド上に設置を予定

[1] Rupak Roy, et. Al., Proceedings of the 21 IEEE/ACM Int. Sym. on Cluster, Cloud and Internet Computing (CCGrid 2021), (2021).

[2] 平木俊幸、初井宇記、特願2021-175897

[3] T. Hiraki et.al., in preparation

T. Hatsui, RIKEN

10

スライド提供：初井宇記(RSC)、城地保昌(RSC)

SPring-8 Data Flow Systemのユーザー画面（検討中の案）

Login

SPring-8 SPring-8 Data Portal

User

Password

Log in

O

R

Login with XXX

SPring-8データセンターのID

富岳IDでログイン
J-PARC, 次世代放射光IDでも
ログインできるように拡張予定

データの認証・認可
先端大型施設間で連携

Upload

SPring-8 SPring-8 Data Portal

Upload

Proposal ID
XXXXXX

Sequence ID
XXXXXX

measurement metadata
XXXXXX.json

sample metadata
XXXXXX.json

Sequence Data

Drop your file(s) here, or browse file(s)

Upload

xxx MB

メタデータとデータをま
とめてアップロード

Download

SPring-8 SPring-8 Data Portal

Sequence ID	Date Created	Size (MB)	Opera
XXXXXXX	2022-01-01	10	📄 👤 🔗
XXXXXXX	2022-01-01	11	📄 👤 🔗
XXXXXXX	2022-01-01	15	📄 👤 🔗
XXXXXXX	2022-01-01	21	📄 👤 🔗
XXXXXXX	2022-01-01	21	📄 👤 🔗

Download

Notebook

メタデータを閲覧

Access Controlへのリンク

データへのリンク(REST API)

データの表示(jupyter notebook)

データをまとめてダウンロード

Access Control

SPring-8 SPring-8 Data Portal

Sharing Settings

Proposal ID
XXXXXXX

Sequence ID
YYYYYYYY

Sharing

☐ Everyone ☒ Limited

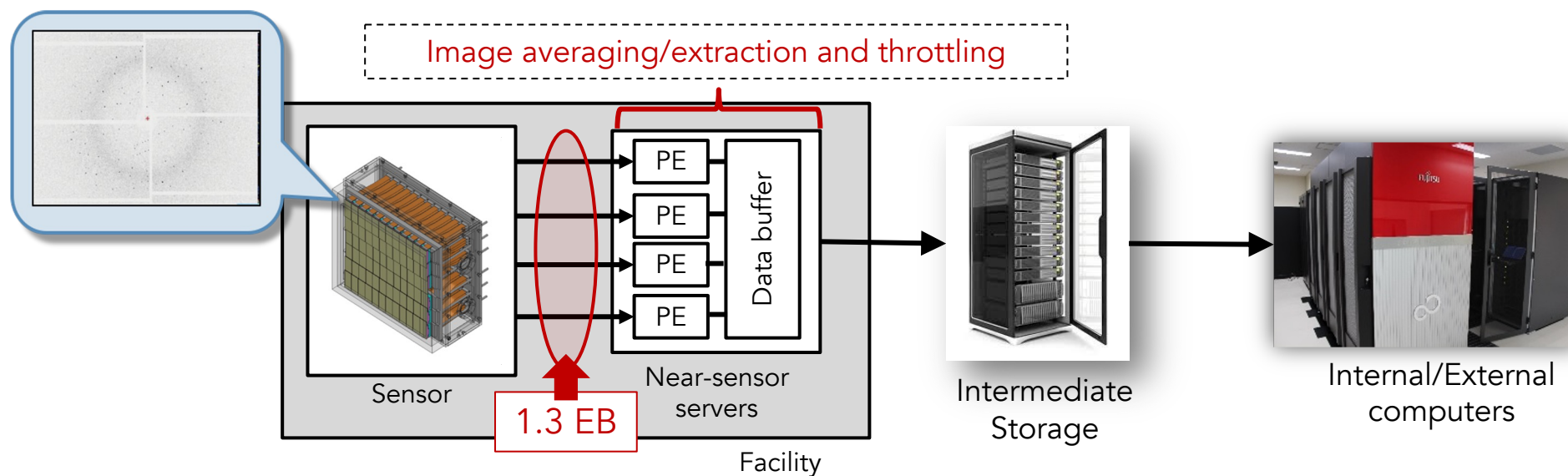
User	Access	Acti
User-A	☐ Manage ☒ Share	📄 👤 🔗
User-B	☒ Manage ☐ Share	📄 👤 🔗
User-C	☐ Manage ☒ Share	📄 👤 🔗

Close

実験責任者はデータのアクセス
権限を設定することができる。
詳細はユーザーコミュニティ・
放射光学会と協議

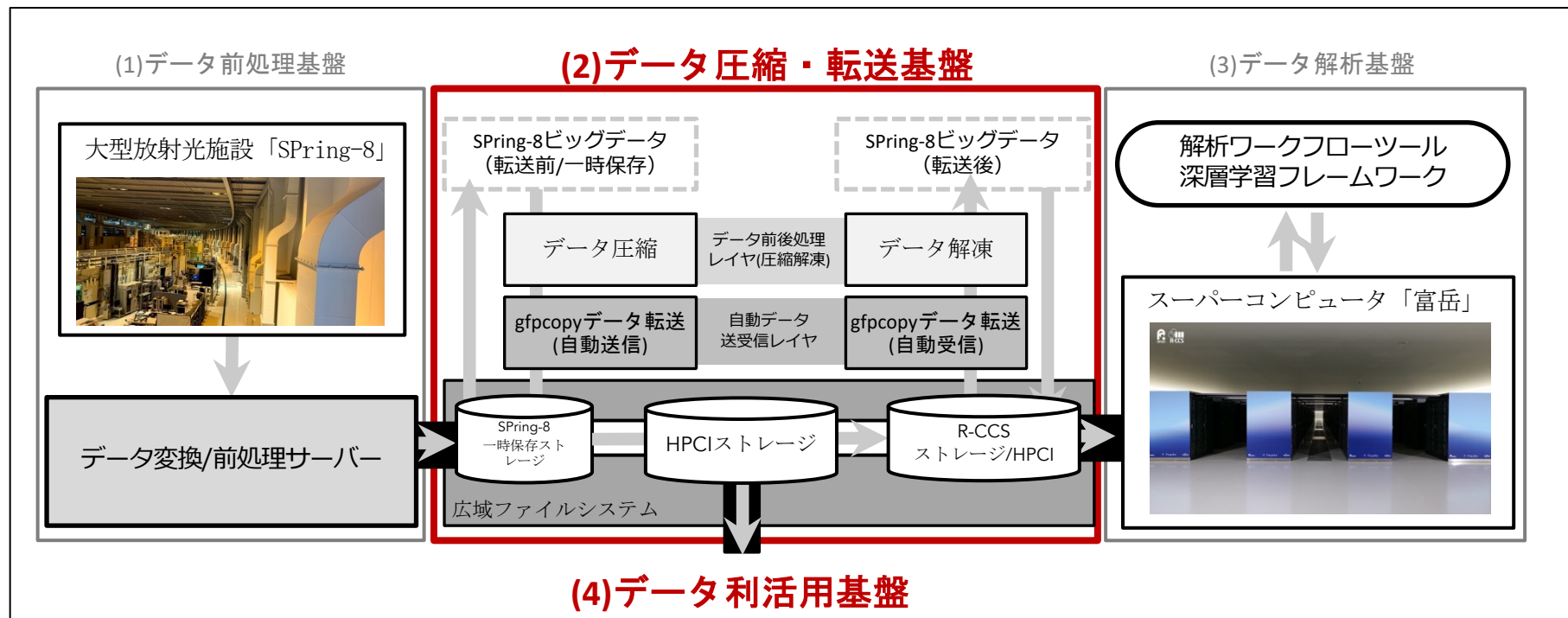
SPring-8におけるビッグデータ転送

- 2025年の次世代検出器(CITIUS)の導入により年間1.3 Exaバイトのデータが生成されると見積
- 2017年時点：SPring-8 共用ビームライン (26 BLs) において年間0.32 PBのデータ生成
- 既存の手法で100 – 400Petaバイトまで削減することが可能
 - Image averaging/extraction, Reducing duty ratio の制限



センサーから内部/外部計算システムへビッグデータの高速転送のためにデータ圧縮技術の創出

[Overview]



[Objective]

- 大規模研究施設連携によるビッグデータの収集・解析・利活用基盤（ビッグデータ基盤）をコデザインにより構築
 - (1)**データ前処理基盤**：FPGAを用いてハードウェアレベルでのデータ変換・前処理を行うことでセンサーから得られた実験データを効率よく一時ストレージへ保存する技術
 - (2)**データ圧縮・転送基盤**：一時ストレージ上の実験データを富岳上へ高速に転送するためのデータ圧縮の開発と転送基盤の開発
 - (3)**データ解析基盤**：富岳上へ集められたデータを効率よく解析するための基盤（ワークフローツールや深層学習フレームワーク）の構築
 - (4)**データ利活用基盤**：収集された一次データや解析結果などを幅広く利活用させるためのデータ利活用基盤の構築を行う。(e.g., Single sign-on authentication, GakuNin RDM etc.)

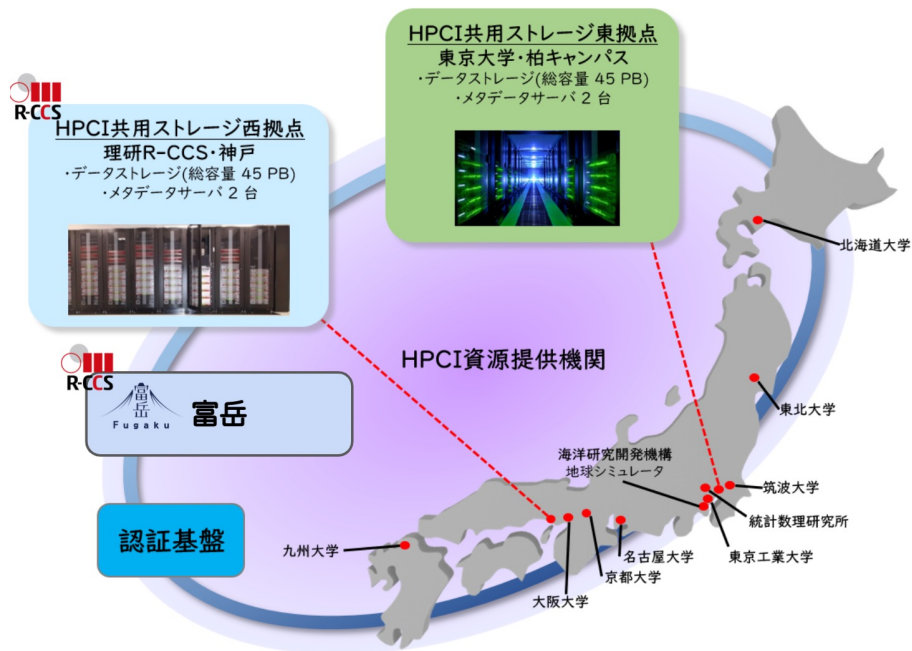
Data transfer service in SACLA (SPring-8 Angstrom Compact Free Electron Laser)

- We launched data transfer service from SACLA to “HPCI shared storage” in 2021
 - “HPCI shared storage” is a geographically distributed storage shared by Japanese supercomputing centers
 - It facilitates data analytics in HPCI systems including Fugaku
- We are planning to expand the service to SPring-8 synchrotron radiation facility and enhance the usability (Common authentication scheme, GakuNin RDM etc.)

SACLA HPC:
Data Transfer Service to HPCI Shared Storage



Source (May 14, 2021): <http://xfel.riken.jp/users/bml09-1.html>



Data Transfer Service to HPCI Shared Storage
Toward the creation of innovative achievement through SACLA

2021年5月14日
理化学研究所
東京大学

HPCI共用ストレージへのデータ転送サービス開始 —SACLA実験データの大規模解析による新たな研究成果創出に向けて—

理化学研究所(理研)放射光科学センター、理研計算科学研究センター(R-CCS)および東京大学情報基盤センターは、[X線自由電子レーザー\(XFEU\)](#)^[1]施設「[SACLA](#)」^[2]で得られた実験データの大規模解析のため、SACLAからHPCI^[3]共用ストレージ^[3]へのデータ転送サービスを5月14日より開始しました。

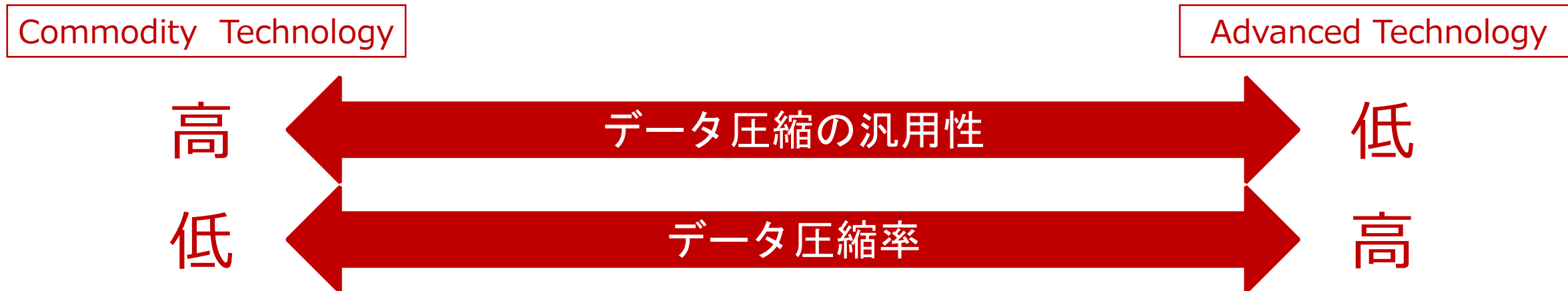
近年、SACLAで得られた大量の実験データを、外部の研究機関と迅速に共有し、高度な計算科学によって解析を行うニーズが急速に増えています。そこで本サービスでは、R-CCSと東大情報基盤センターが運用するHPCI共用ストレージを活用して、高性能・高信頼なデータ転送を実現します。HPCI共用ストレージで用いているオープンソース分散ファイルシステム「Gfarm」を活用した高速データ転送ツールを提供することで、幅広いユーザーが簡単に利用できる環境を整えました。これにより、[スーパーコンピュータ「富岳」](#)^[4]、[「Wisteria/BDEC-01」](#)^[5]をはじめとしたHPCIを構成するスーパーコンピュータの能力を活用した大規模解析が容易になり、新たな研究成果が創出されることが期待できます。

関連リンク：[放射光科学センター X線自由電子レーザー施設 SACLA](#)

Image: HPCI(<https://www.hpci-office.jp/info/pages/viewpage.action?pageId=111380786>)

Source (May 14, 2021): https://www.riken.jp/pr/news/2021/20210514_1/

- **Commodity Technology**の調査研究
 - 既存のデータ圧縮・転送技術の適用
- **Advanced Technology**の研究開発
 - TEZipの開発と高度化



- **SACLAで実際に生成されたデータを用いて、効率的なデータ圧縮方法に関する調査**
 - 画像圧縮をデータ転送の改善に活かすため、可逆圧縮と非可逆圧縮の両方に関する複数の圧縮方式の利用を検討
 - 非可逆による精度の低下の懸念より、まずは可逆圧縮を中心とした圧縮率や効率の調査

■ 既存圧縮機構の評価

- 評価対象：jpeg-xl, lrzip, jpeg-2000, webp, 7zip, xz, bzip2, png, zstd
- jpeg-xlが一番高い圧縮率
 - jpeg-xl: JPEGの次世代フォーマット。現行のJPEGよりも高い圧縮率を実現。ISO国際標準となる予定。
- 以下の懸念により、lzip, jpeg-2000, webpの採用も検討
 - 2022年10月31日時点で「Chrome」での採用を取りやめ[1]
 - jpeg-xl vs. WebP (by Google)
- 他のデータセットに対しても引き続き調査を行う

Compressor	File size(Byte)	RATE
original data (no compression)	302091064	1.000
jpeg-xl (cjxl -d 0 -e 6)	178657008	0.591
lrzip (lrzip -z -L 9 -T -p 1)	179884170	0.595
jpeg-2000 (opj_compress -r 1)	181344018	0.600
jpeg-2000 (nvjpeg2k_encode)	181343979	0.600
webp (cwebp -lossless -z 9 -m 6)	184339672	0.610
7zip (7z -m0=delta:2 -m1=ppmd -mx=9)	190544120	0.630
7zip (7z -m0=delta:2 -m1=lzma -mx=9)	195384406	0.646
xz (xz -delta=dist=2 -lzma2=preset=9e,dict=1GiB,lc=4,pb=0)	197814952	0.654
7zip (7z -m0=ppmd -mx=9)	198009567	0.655
bzip2 (bzip2 -9)	204165618	0.675
png	222057933	0.735
zstd (zstd -22 -ultra)	238981006	0.791

Table 11.1: Comparison of lossless compression ratios using the SACLA image data

データ提供：松田元彦(R-CCS)

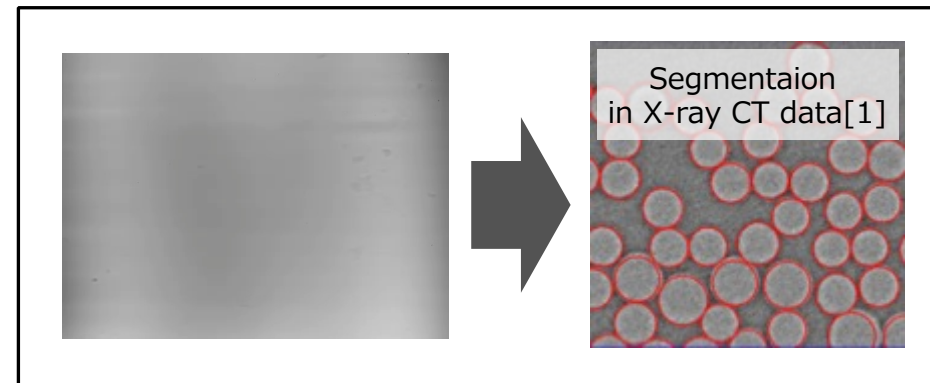
[1] FSF: Chrome's JPEG XL killing shows how the web works under browser hegemony

, <https://arstechnica.com/gadgets/2023/04/free-software-group-decries-google-dropping-space-saving-jpeg-xl-format/>

Reconstruction of X-ray CT data

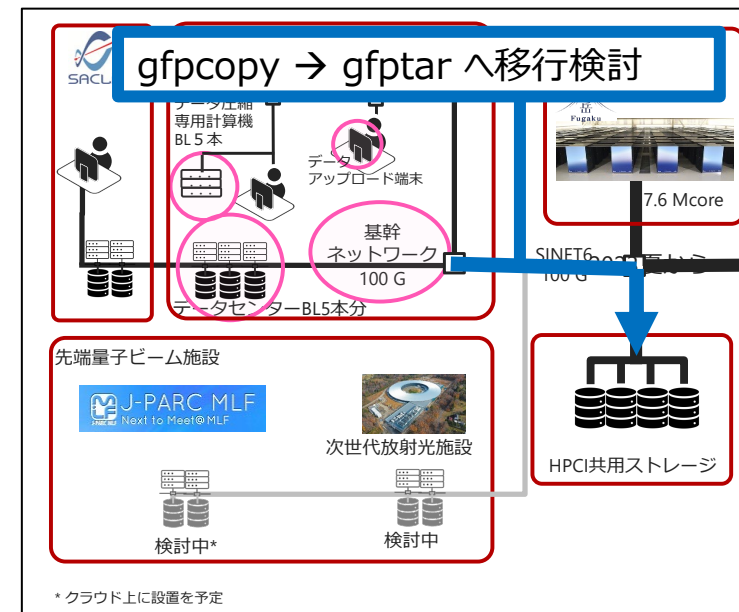
- **SACLA/SPRING-8の画像の特徴と現状のデータ転送手段**

- 現在、SACLAデータ転送サービスではGfarm*のgfpcopyを利用
 - gfpcopy: 指定されたフォルダ内のファイルを逐次転送
- 1ファイルあたり数MB～数10MBの画像データが100～1000個存在
→ Small I/O（細粒度のI/O）が発生し帯域を十分に利用できない



- **Gfarm*の新機能(gfptar)が追加、利用の検討**

- 機能：アーカイブファイルをGfarm上に作成
- -I, --use-compress-program=COMMAND
 - アーカイブを作成時または抽出時に、圧縮・伸張するための外部コマンドを指定。
 - 任意のアーカイブ・圧縮機構を使用することが可能
 - 既存のデータ圧縮機構や高度化中のTEZipを組み込むことが可能
→ SINET5/6の帯域を有効活用しさらにデータ圧縮技術と統合することにより高速なデータ転送が実現されると期待



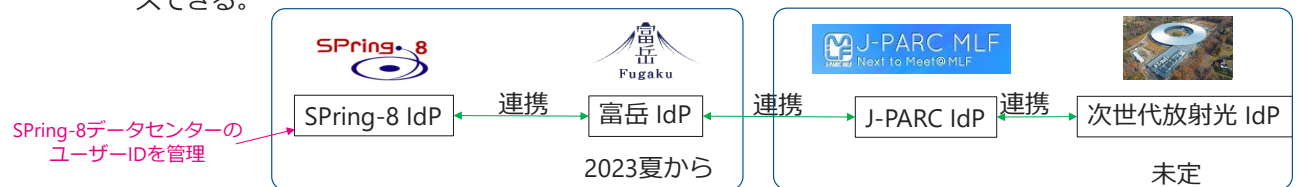
* Gfarm：HPCIストレージを管理するファイルシステム

スライド：初井宇記，“SPRING-8データセンター構想の進捗報告”，第3回データワークショップ、2022年4月18日

- R-CCS(富岳)、RSC(SPring-8、SACLA)のサービスをシングルサインオン(SSO)で利用するためのセンターのID連携についても検討
- RSCと連携して認証基盤の共通化の検討
 - Keycloakの導入を検討
 - c.f.) 富岳ではKeycloakをOpen OnDemandで採用

1B-1) 複数量子ビーム施設のデータをアクセス

- 連携が必要な機能
 - データの認証・認可：共通化することで一つのプログラムから複数施設のデータを簡単にアクセスできる。

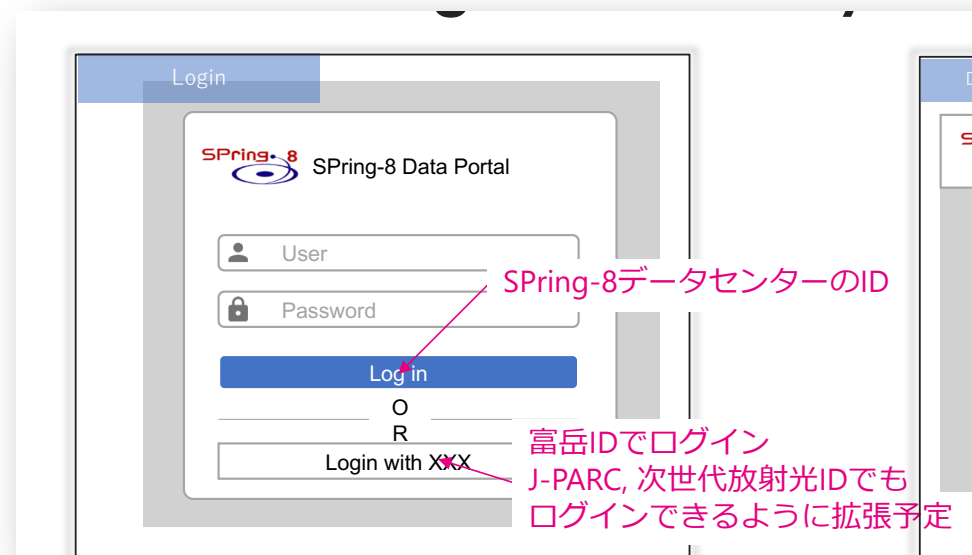


将来的にNIIの認証基盤とも連携

- 共通化が必要な機能
 - データアクセス方法の共通化
例：データアクセスの基幹部分のソフトウェアがHOST名を変えるだけでどの施設でも使える

検討中の実装

Amazon社のオブジェクトデータサービスS3と互換性のあるサービス
ユーザーから見た使い方
pythonライブラリ boto3によりSPring-8データセンター内のデータをアクセスできるようになる。



- TEZip：PredNetと呼ばれるNNを用いることにより1つの画像から後続の画像を予測し、予測画像との差分を符号化・圧縮する、時間発展型のデータ圧縮ツール

Compression of Time Evolutionary Image Data through Predictive Deep Neural Networks (CCGrid2021) [1]

Rupak Roy^{*1}, Kento Sato^{*2}, Subhadeep Bhattacharya^{*2}, Xingang Fang^{*2}, Yasumasa Joti^{*3}, Takaki Hatsui^{*3}, Toshiyuki Nishiyama Hiraki^{*3}, Jian Guo^{*2} and Weikuan Yu^{*1}

^{*1} Florida State University, ^{*2} RIKEN Center for Computational Science, ^{*3} RIKEN SPring-8 Center, ^{*4} Anhui University of Finance and Economics

- Scientists take snapshots while rotating materials that they want to analyze → time evolutionary data
- We proposed new AI-driven data compressor (TEZIP) for time evolutionary data
- We achieved higher compression ratio compared to existing video encoders (Zstd, HFFU, FFV1, x.265)

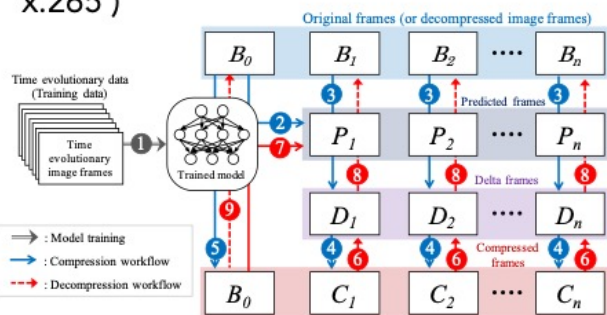


Fig. 1. Workflows of TEZIP (de)compression

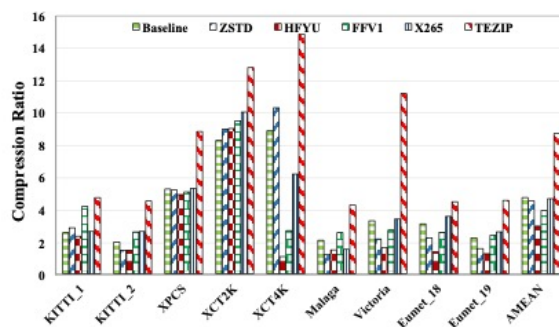
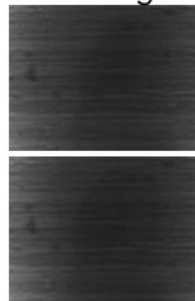
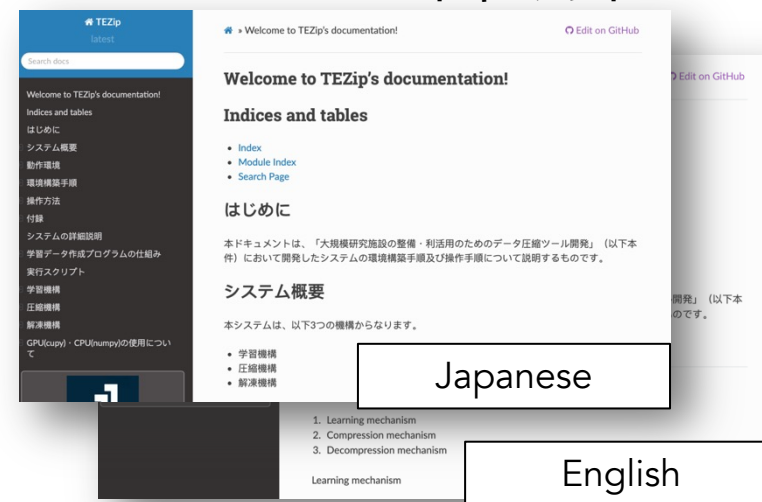


Fig. 8. Compression ratio with lossless compressors.

XCT images



Readthedoc ドキュメント



[1] Rupak Roy, Kento Sato, Subhadeep Bhattacharya, Xingang Fang, Yasumasa Joti, Takaki Hatsui, Toshiyuki Hiraki, Jian Guo and Weikuan Yu, "Compression of Time Evolutionary Image Data through Predictive Deep Neural Networks", In the proceedings of the 21 IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid 2021), May, 2021

Prediction is one of keys for good compression

Compression

1.1	1.5	1.8	2.1
1.0	1.4	2.3	2.7
1.3	1.8	2.5	3.1
1.9	2.1	2.6	3.3

Original data

diff
(-)

1.1	1.5	1.8	2.1
1.0	1.3	2.3	2.7
1.3	1.8	2.5	3.0
1.9	1.9	2.5	3.3

Predicted data

=

0	0	0	0
0	0.1	0	0
0	0	0	0.1
0	0.2	0.1	0

Delta

Sequence of same values (zero) are likely to appear many times

gzip
(LZ77 & Huffman encoding)

Compressed data

Decompression

1.1	1.5	1.8	2.1
1.0	1.4	2.3	2.7
1.3	1.8	2.5	3.1
1.9	2.1	2.6	3.3

Original data

=

1.1	1.5	1.8	2.1
1.0	1.3	2.3	2.7
1.3	1.8	2.5	3.0
1.9	1.9	2.5	3.3

Predicted data

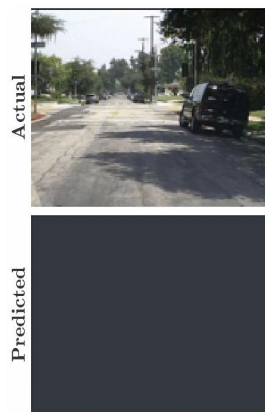
diff
(+)

0	0	0	0
0	0.1	0	0
0	0	0	0.1
0	0.2	0.1	0

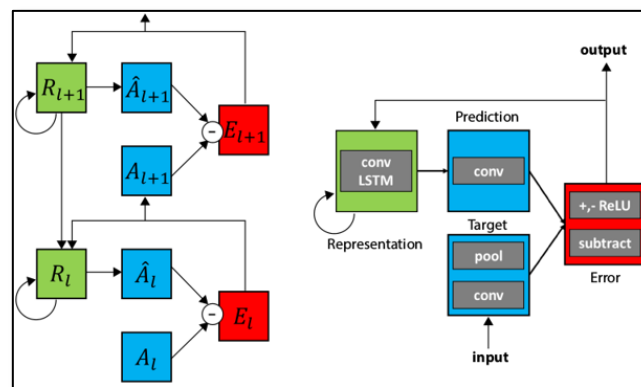
Delta

gunzip

- **PredNet [1]**
 - Deep recurrent convolutional neural network
 - Given a frame of a picture, this NN can predict multiple future frames
- **SPRING-8 data is time evolutionary data, so this type of prediction can be applied**



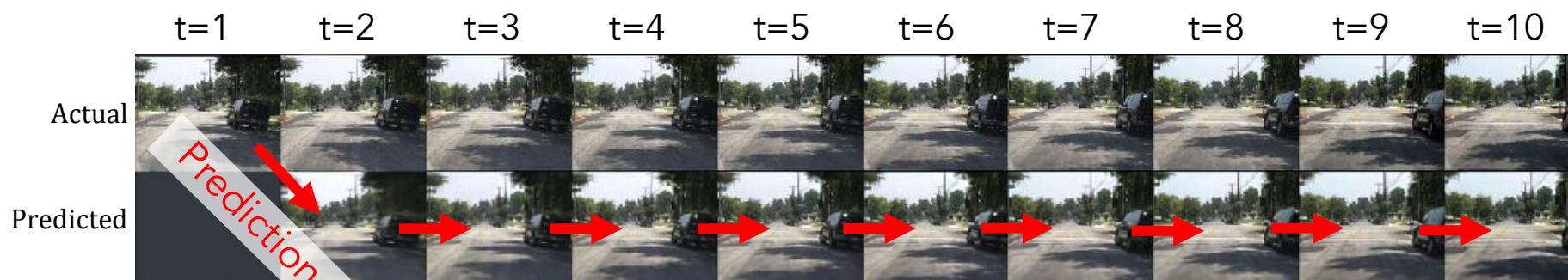
<https://coxlabs.github.io/prednet/>



PredNet neural network

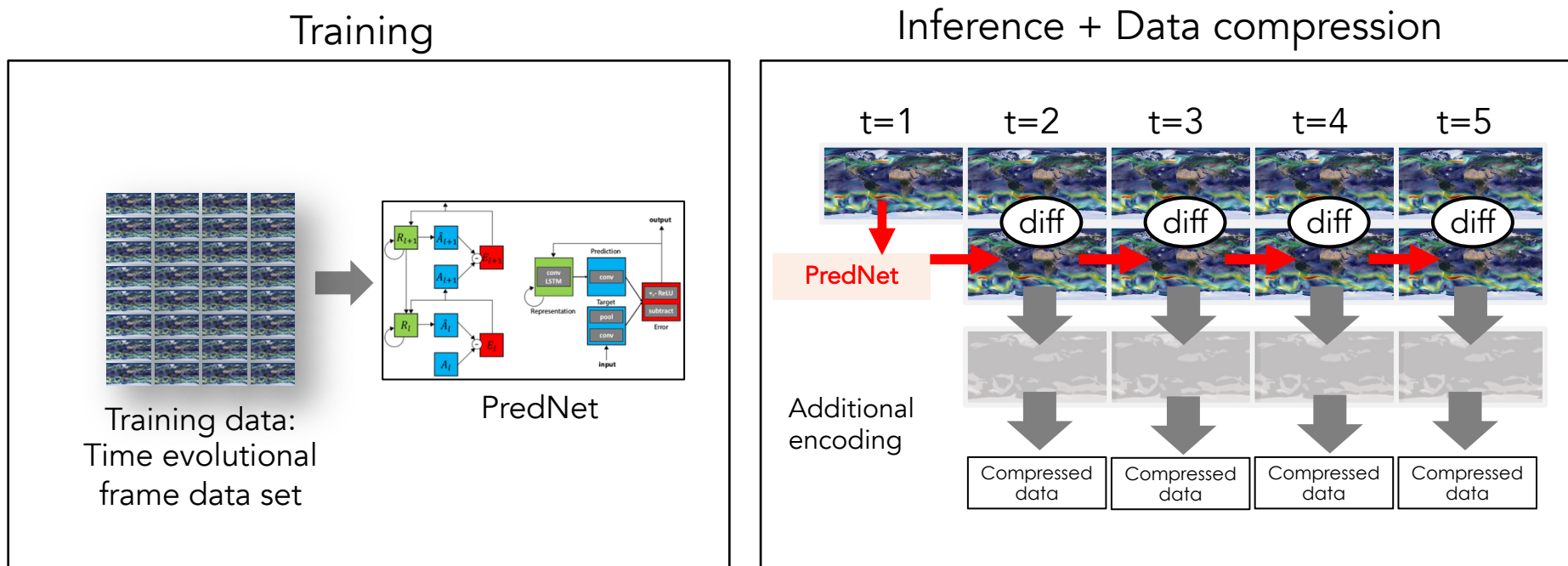


X-ray CT images



[1] Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)

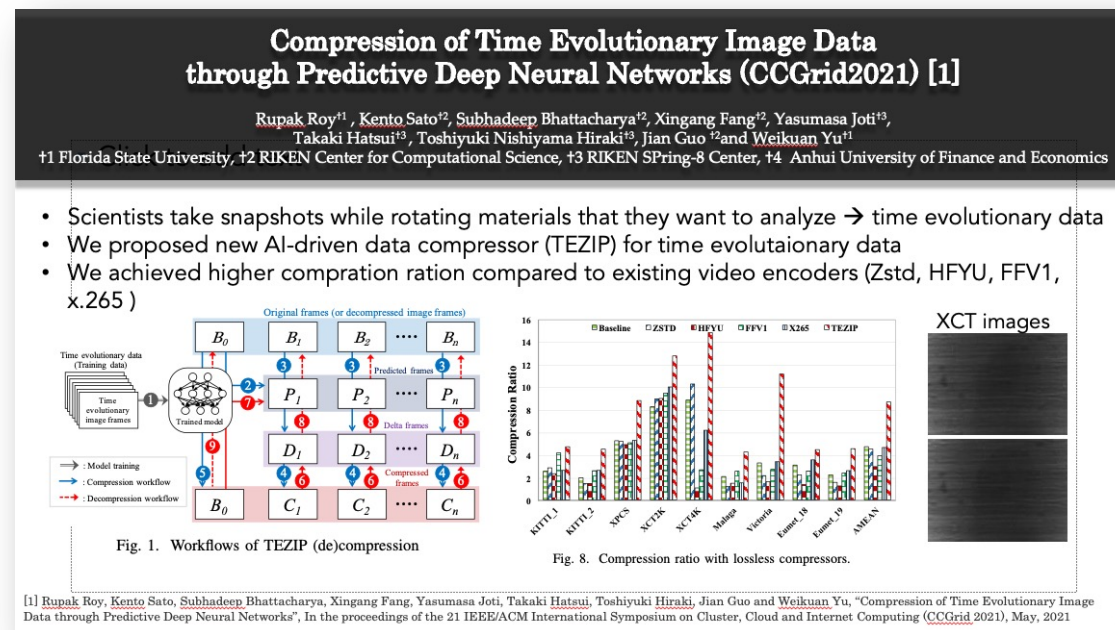
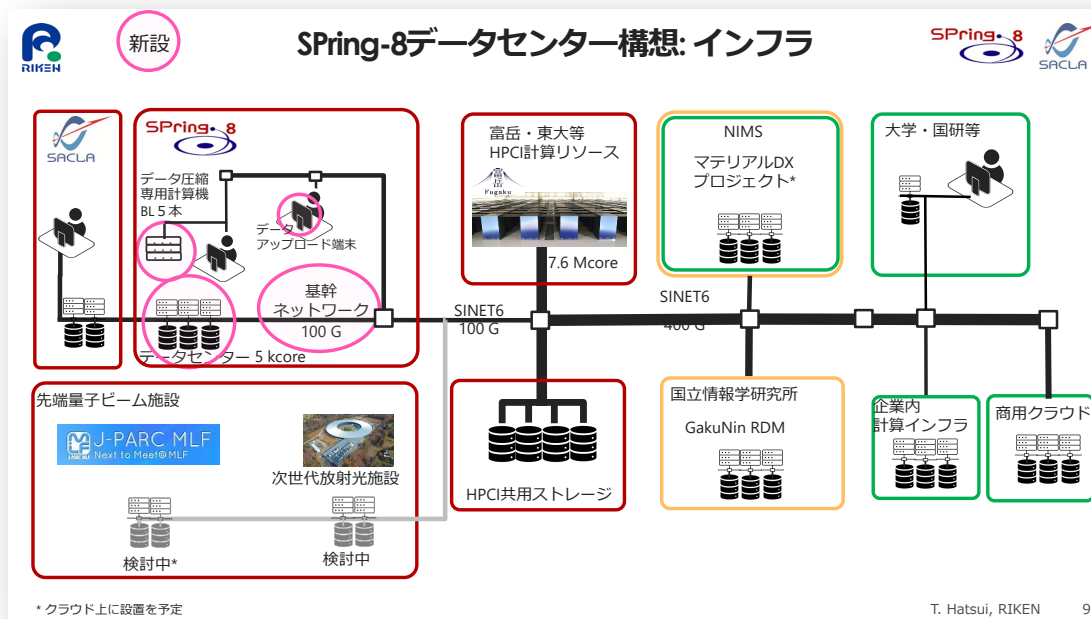
- We train PredNet to learn how pixels move and how fast
 - Giving a series of time evolutionary frames to PredNet
- When compressing frames from $t=2$ to $t=5$, we predict future frames from original data ($t=1$)
- We compute diff, apply series of encoding
- We only store (1) base frame data and (2) compressed data



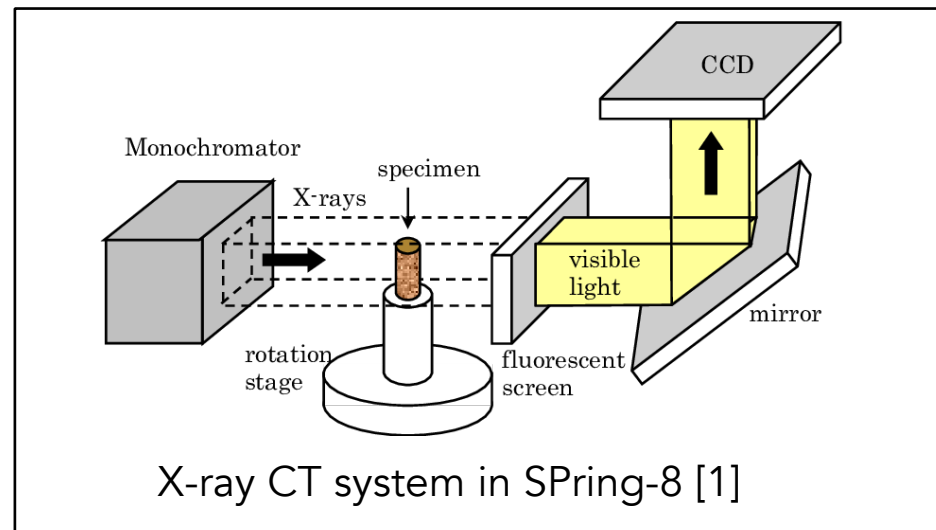
- ハード面：SPring-8データセンター構想
 - データ圧縮計算機、データ解析計算機、データアップロード端末、SINET6
- ソフト面：ビッグデータ収集・解析・利活用基盤の研究開発
 - データ転送サービス、データ圧縮技術の調査研究・研究開発

SPring-8 データセンター構想

データ圧縮の研究開発(TEZip)



- **Commodity technology の調査研究**
 - 別のデータでの調査研究・性能評価やシステム導入の検討
- **Advanced technologyの研究開発**
 - TEZipの高度化：予測機構の改良
 - X-ray CTデータ特化型のデータ圧縮の研究開発
- **データ圧縮共通化のための連携 (led by ANL)**
 - 共通API LibPressio/Z-checkerとTEZipの統合
 - 共通データ基盤 (SDRBench) でのデータ共有



Data sets:

Name	Type	Format	Size (data)	Command Examples	Link
CESM-ATM Source: Mark Taylor (SNL)	Climate simulation	Dataset1: 79 fields: 2D, 1800 x 3600 ; Dataset2: 1 field : 3D, 26x1800x3600. Both are single precision, binary	Dataset1 (raw): 1.47GB Dataset1 (cleared): 1.47GB Dataset2: 17GB (cleared data zeroed all background data)	SZ(Compress): sz -z -f -i CLDHGH_1_1800_3600.f32 -M REL -R 1E-2 -2 3600 1800 SZ(Decompress): sz -x -f -i CLDHGH_1_1800_3600.f32 -2 3600 1800 -s CLDHGH_1_1800_3600.f32 -z - 8 ZFP: zfp -f -i CLDHGH_1_1800_3600.f32 -z CLDHGH_1_1800_3600.f32 zfp -o CLDHGH_1_1800_3600.f32.zfp.out -2 3600 1800 -a 1E-2 -s LibPressio: pressio -b compressor=SCOMP -i CLDHGH_1_1800_3600.f32 -d 3600 -d 1800 -f float -o nh=1e-2 -m time -m size -M all where SCOMP can be sz, zfp, sz3, mgard, etc... Z-checker-installer: ./runZCCase.sh -f REL CESM-ATM raw-data-dir f32 3600 1800	Dataset1 (raw) Dataset1 (cleared) Dataset2 (raw) Metadata Dataset1's property Dataset2's property
EXAALT Source: EXAALT team This dataset has been approved for unlimited release by Los Alamos National Laboratory and has been assigned LA-UR-18- 25670.	Molecular dynamics simulation	6 fields: x,y,z,vx,vy,vz, Each field stored separately, Single precision, Binary, Little-endian	Dataset1: 60 MB Dataset2: 973 MB Dataset3: 2.4 GB	SZ(Compress): sz -z -f -i xx.f32 -M REL -R 1E-2 -1 2869440 SZ(Decompress): sz -x -f -i xx.f32 -1 2869440 -s xx.f32.sz -a ZFP: zfp -f -i xx.f32 -z xx.f32.zfp -o xx.f32.zfp.out -1 2869440 -a 1E-2 -s LibPressio: pressio -b compressor=SCOMP -i xx.f32 -d 2869440 -f float -o nh=1e-2 -m time -m size -M all where SCOMP can be sz, zfp, sz3, mgard, etc... Z-checker-installer: ./runZCCase.sh -f REL EXAALT raw-data-dir f32 2869440	Dataset1 Metadata1 Property1 Dataset2 Metadata2 Property2 Dataset3 Metadata3 Property3

<https://sdrbench.github.io>

Customized compressors based on SZ Framework

We created multiple effective compression pipelines for different applications/use-cases.

Interp-SZ (default setting in SZ3): Optimized for seismic imaging (oil and gas)

- > Predictor: Multi-dimensional spline interpolation based prediction [IEEE ICDE2021](#)

Pastri-SZ: Optimized for quantum chemistry.

[IEEE CLUSTER2018 \(best paper\)](#)

- > Predictor: Scaled-pattern based prediction

MMD-SZ (MDZ): Optimized for material molecular dynamics.

[IEEE ICDE2022](#)

- > Predictor: Multi-peak distribution based prediction in space + time-based prediction

Roibin-SZ: Optimized for X-ray crystallography (instrument).

[Under going](#)

- > Pipeline: Pixel-binning + SZ2.1 + preserving region-of-interests

Quality-oriented SZ (QoS): Error-bounding + tuning user-specified quality metric.

[IEEE/ACM SC2022](#)

- > Predictor: quality-oriented predictor

Constraint-supported SZ: Error-bounding + meeting dynamic constraints

[IEEE TPDS 2022](#)

- > Pipeline: Preprocessing + traditional prediction + improved quantization bins

Feature-preserving based SZ: Toward Feature-Preserving Vector Field Compression.

- > Pipeline: Multi-dimensional spline interpolation + adjusted quantization bins.

- **SPring-8データセンター構想**
 - **RSC**：初井宇記、城地保昌，矢橋牧名，中嶋享，本村幸治，平木俊幸，杉本崇，山鹿光裕，中町将貴、渡邊一輝
 - 連携企業の皆様
- **HPCI連携、データ圧縮、Workflowツール、認証**
 - **理研計算科学研究センター(R-CCS, 富岳)**：松田元彦，原田浩、金山秀智，山本啓二，佐藤賢斗，佐野健太郎，庄司文由，松岡聡
- **研究データ管理**
 - **理研情報統合本部(R-IH)**：實本英之，富士健太郎，小林紀郎，美濃導彦
 - **国立情報学研究所(NII)**：込山悠介，山地一禎，喜連川優
 - **物質・材料研究機構 (NIMS)**：谷藤幹子，吉川 英樹，松田 朝彦
 - **放射光学会**：横山利彦 会長
- **施設間連携**
 - **J-PARC**
 - **KEK 物質・生命科学ディビジョン**：大友季哉、稲村泰弘、鈴木次郎
 - **CROSS中性子科学センター**：伊藤崇芳、岡崎伸生
 - **次世代放射光施設**
 - **QST**：中谷 健
 - **PhoSIC**：高田昌樹、中村哲也
 - **東北大学**：青木孝文、滝沢寛之