# Software Solutions for Large-scale Data Management in Scientific Research

LLSDA: Productive, Performant Software for Large Scale Scientific Data Analysis
October 22, 2025, 8:30–8:50am (Session 5)


## Kento Sato

- Team Principal, High Performance Big Data Research Team, RIKEN R-CCS
- **Unit Leader, Data management platform development unit, AI for Science division, RIKEN R-CCS**
- Unit Leader, Advanced HPC Technologies Development Unit, Next-Generation HPC Infrastructure Development Division, RIKEN R-CCS
- Visiting Professor, Kobe graduate university
- Visiting Associate Professor, Tohoku graduate university

# Our mission:
# HPC, AI & Big Data Fusion

The mission of our team is to investigate state-of-the-art techniques for efficiently running large-scale applications on HPC systems. In addition to fundamental R&D of system software in HPC, our team conducts performance analysis and software development to accelerate deep learning and big data processing on HPC systems (HPC for AI/BD), while we also apply AI techniques to solve technical challenges in HPC (AI/BD for HPC)

## HPC for AI/BD

Research and software development for accelerating AI/Big data applications by techniques in HPC

## AI/BD for HPC

Research and software development for resolving technical challenges in HPC by AI/ Big Data techniques

## Fundamental R&D for HPC

# TRIP-AGIS (Artificial General Intelligence for Science of Transformative Research Innovation Platform):
# Development and sharing of generative AI models for scientific research

- **Develop generative AI models (scientific foundation models) specialized for scientific research** by building collaborative frameworks with research institutions that have strengths in scientific research fields and using the foundation models to conduct **fine-tuning, multimodalization of scientific research data,** etc.
- **By widely opening up the use of our AI models from scientific research to industry and academia,** we aim to innovate scientific research in various fields (dramatically accelerate the scientific research cycle and expand the exploratory space for scientific research).

**Tentative Discussion Points on AI (AI戦略会議, May 26, 2023)**

AI Development Capability

- It is also almost certain that the results of AI research will contribute to the acceleration of R&D in areas other than AI.
- As the world is about to be revolutionized by generative AI, it is important to foster fundamental research and development capabilities for generative AI in Japan as quickly as possible.
- It is expected to build an environment for research and human resources development where top talents from all over the world can gather and compete with each other in friendly competition, and to strengthen the basic development capabilities of industry, academia and government.

## High quality data

- Collect and maintain high quality data for training, fine tuning, etc.
- Collaboration and joint development with related research institutions that accumulate data
- Target science fields:
  - (1) Life and Medical Sciences (e.g., predicting differences due to dynamic changes and genetic mutations caused by drugs, etc.)
  - (2) Materials/physical properties science (e.g., prediction of physical properties of novel materials)

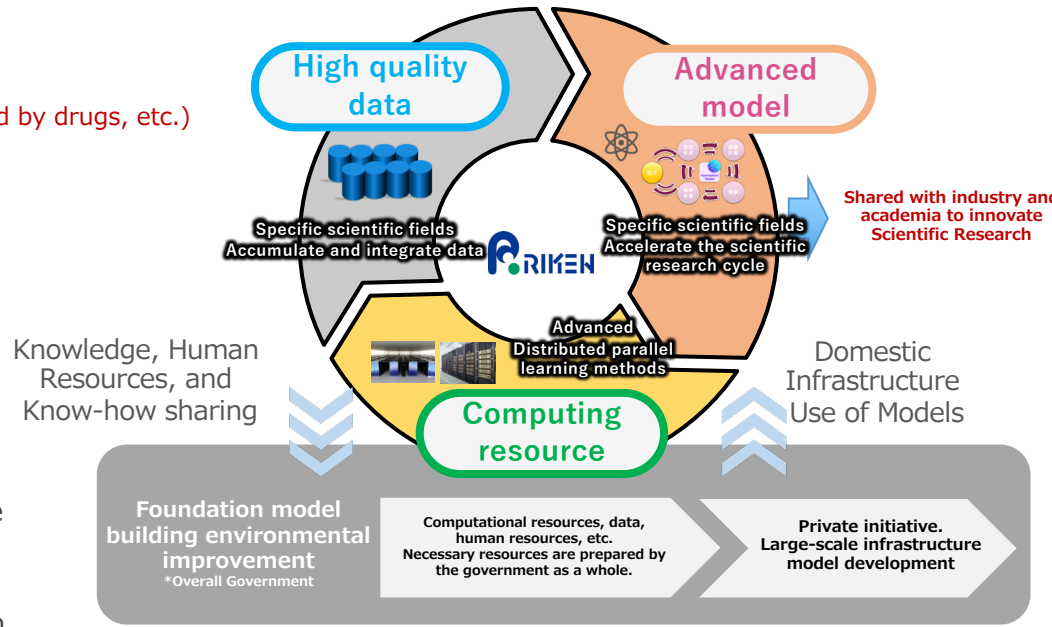
## Advanced model

- Develop, operate, and share scientific multi-modal foundation models for the target science fiealds
- In parallel, research and development necessary to read, learn, and generate multimodal data

## Computing resource

- Define requirements for AI-for-Science research, and procure and operate an AI-for-science system (GPU)
- Combine Supercomputer “Fugaku” with the AI-for-Science system through high-speed network to facilitate interplay between these systems for developing the AI models
- Develop software for accelerating fine-tuning and inference
- Research on new AI architectures (dedicated computing machines other than matrix computation) that can handle multimodal foundation model

## Innovation of Research through "Generative AI Models for Scientific Research “






* Scientific infrastructure model: An infrastructure model tailored for a specific scientific research field (domain) by additionally learning and inferring scientific research data (papers, real-time experimental and simulation data, etc.) to the infrastructure model (language, images, etc.).

# RIKEN TRIP-AGIS Organization

**TRIP-AGIS:** Artificial General Intelligence for Science of Transformative Research Innovation Platform

**TRIP-AGIS ①: Common Platform Technology**
Develop common infrastructure for creating and sharing generative AI models for scientific research

**TRIP-AGIS ②: Generative AI models for scientific research in specific fields**
Develop generative AI models for target scientific research areas (Life and Medical Sciences / Material Science)

**TRIP-AGIS ③: Innovative Computational Infrastructure**
Develop pioneering innovative computational infrastructure for AI-for-Science computation

**TRIP-AGIS ③-1: Operations for Innovative Computational Infrastructure (AI4S system)**
Advance operation technologies for the innovative computational infrastructure enabling large-scale training and inference

**TRIP-AGIS ③-2: Software technologies for the Innovative Computational Infrastructure**
Develop fundamental software for advancing the development environment of generative multimodal AI models for scientific research

**TRIP-AGIS ③-3: New AI architecture technologies**
Develop new domain-specific architectures for AI training and inference

BDR

R-CCS

# AI-for-Science Supercomputer Overview [1]

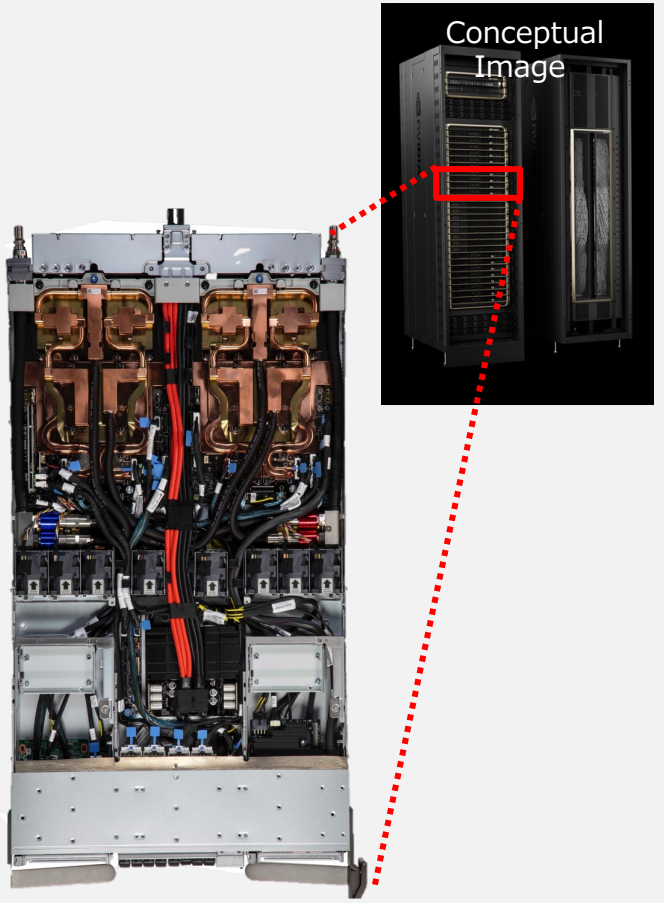
## ■ Background and Purpose
- New computational infra. to advance "AI for Science" through fusion of AI & ModSim
- Accelerating scientific research through world-leading AI performance and interplay with Fugaku
- Dramatically accelerate research cycles by scientific foundation models

## ■ System Configuration
- **Nodes**: 400 nodes equipped with NVIDIA Grace Blackwell superchips (over 1,600 GPUs)
- **Interconnect**: NVIDIA InfiniBand XDR (up to 3.2 Tbps)
- **Performance**: Over 64.16 PFLOPS (FP64), Over 15.539 EFLOPS (FP8)
- **Rack**: Super Micro servers with hot-water cooling capability adopted, achieving both high performance and energy efficiency

## ■ Comparison with "Fugaku"
- FP8 computational performance is approximately 7.23 times that of Fugaku's low-precision (FP16) unit
- Collaboration with "Fugaku," which excels at FP64, enables AI×Science fusion research

## ■ Future Plans
- Installation by fiscal year 2025, full-scale operation starting early fiscal year 2026
- Expected to develop and accelerate collaboration with institutions like the U.S. Argonne National Laboratory (MOU signed in 2024) and others

**AI Inference Performance**
15.539 ExaFLOPS or higher (FP8 performance)

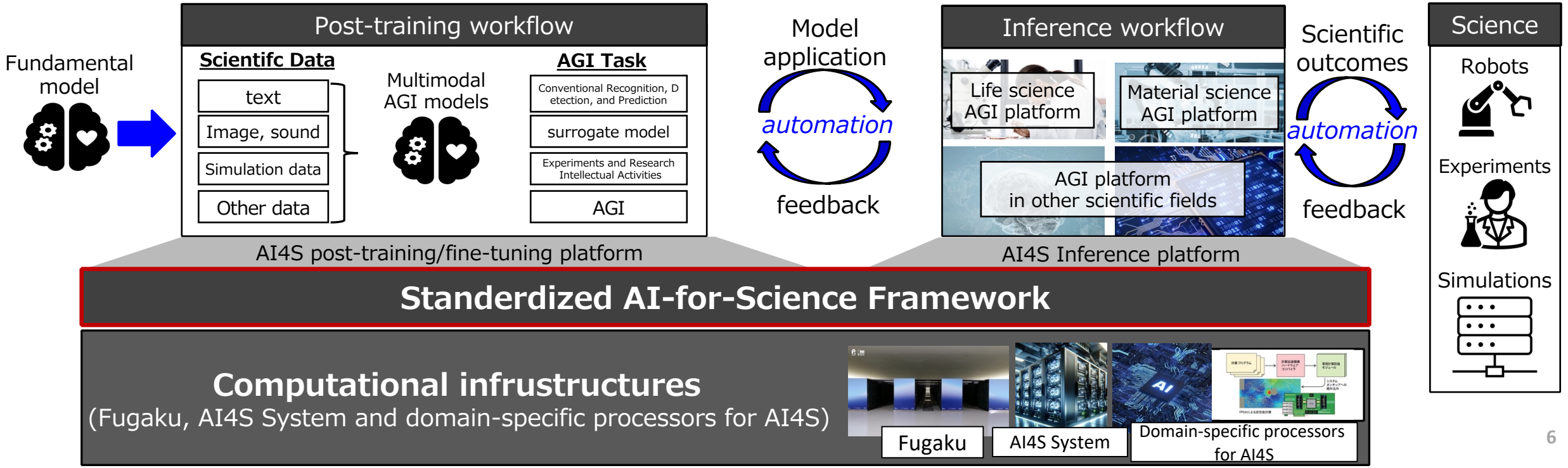



[1] System Selected for AI for Science Development Supercomputer, https://www.riken.jp/pr/news/2025/20250728_1/index.html (July 28, 2025)

# TRIP-AGIS③-2:
# Software technologies for Innovative Computational Infrastructure

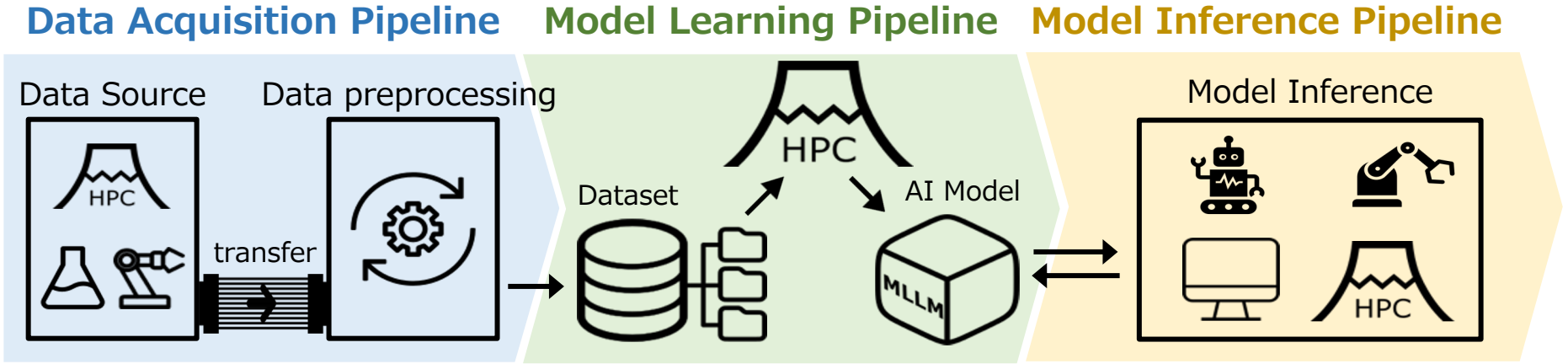
**Develop workflow infrastructures to facilitate post-training, inference and its applications**

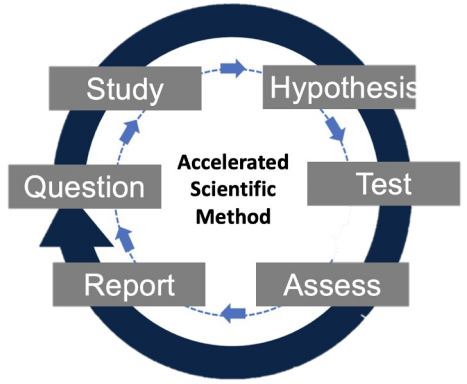
- **TRIP-AGIS ③-2-(i): Learning Optimization Platform Development Unit**
  - Analyze performance and advance system software technologies in post-training/inference for interplay between Fugaku and the AI4S system
  - Explore new software development for the new domain-specific AI processors
- **TRIP-AGIS ③-2-(ii): Data Management Platform Development Unit**
  - Analyze storage system and enhance I/O performance for post-training/inference of multimodal AI models requiring a variety of data types (e.g., text, images, sound, videos, other scientific data)
  - Explore new data management and curation (e.g., collection, compression, organization, encryption etc.)
- **TRIP-AGIS ③-2-(iii): Application Interface Platform Development Units**
  - Develop post-training/inference platforms enabling automation of model application and feedback in life and material science



# AI Platform in the FugakuNEXT era and beyond:
## HPC/AI platform enabled by state-of-the-art system software technologies

- **AI-for-Science Platform**
  - Develop AI and data processing pipelines/workflows that seamlessly integrate individual AI tools to automate a wide range of AI-for-Science research tasks
- **Acceleration of AI for Science**
  - **(1) Data acquisition pipeline**: Fast generation of training data, fast transfer of various scientific data measured and produced and management of the data
  - **(2) Model learning pipeline**: Large-scale pre-training, post-training, and fine-tuning of AI models
  - **(3) Model inference pipeline**: Automated workflow for scientific experiments by AI

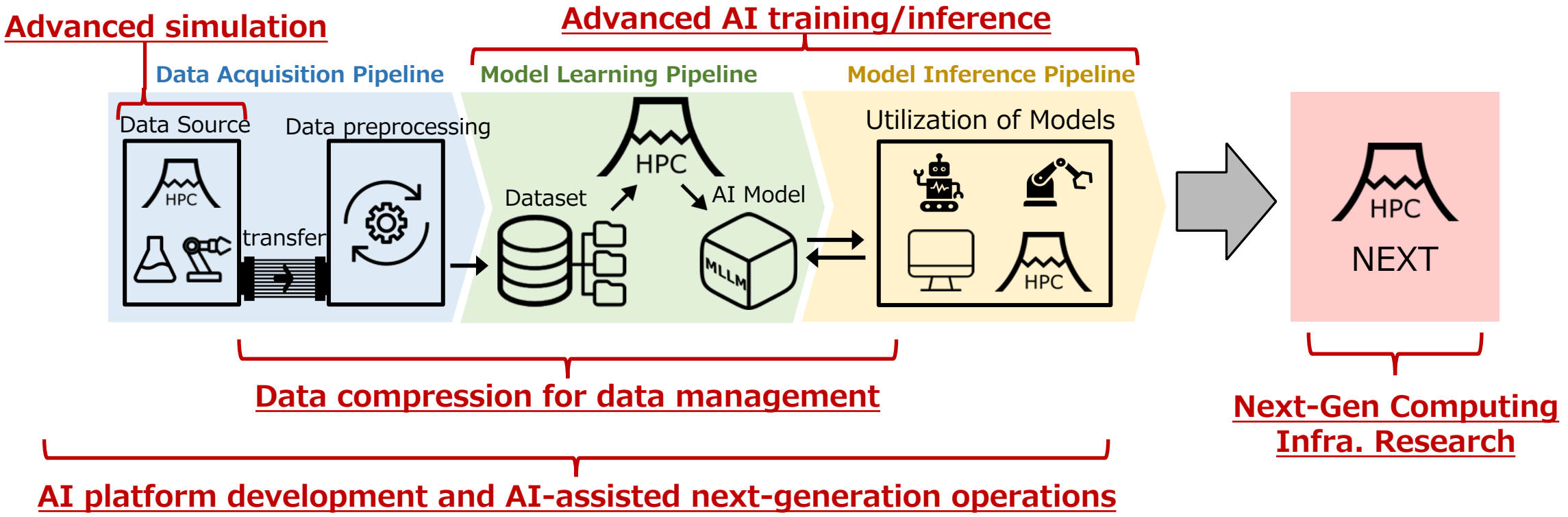



(Source: https://doi.org/10.1038/s41524-022-00765-z )

**Automate research cycles**:
program code generation, execution of program, data generation, data analysis on experimental results, proposal for the next steps

# Research Plan and End-user software adoption strategy:
## Fundamental/core technology research for AI for Science Platform



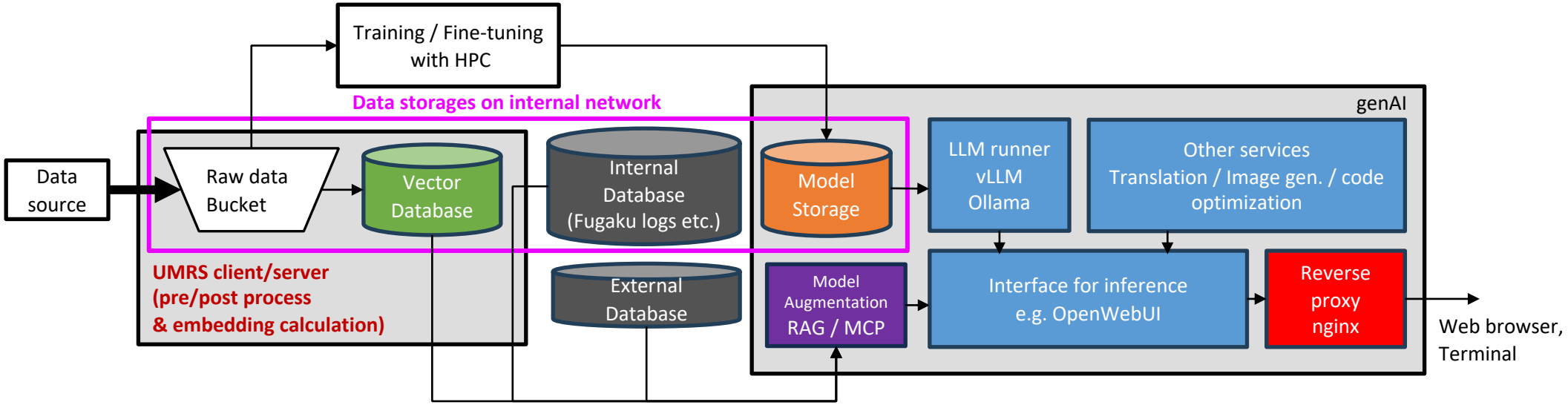

We aim to develop “**default-on**" software incorporated into the platform so that end-users adopt our software without consciously

# Secured & Extendable AI Workflow Orchestration for AI4S

Masaru Nagaso[†1], Jens Domke[†1], Eliott Jacopin[†2], Emmanuel Jeannot[†1],
Kento Sato[†1], Kozo Nishida[†2]
†1: RIKEN R-CCS, †2: RIKEN BDR

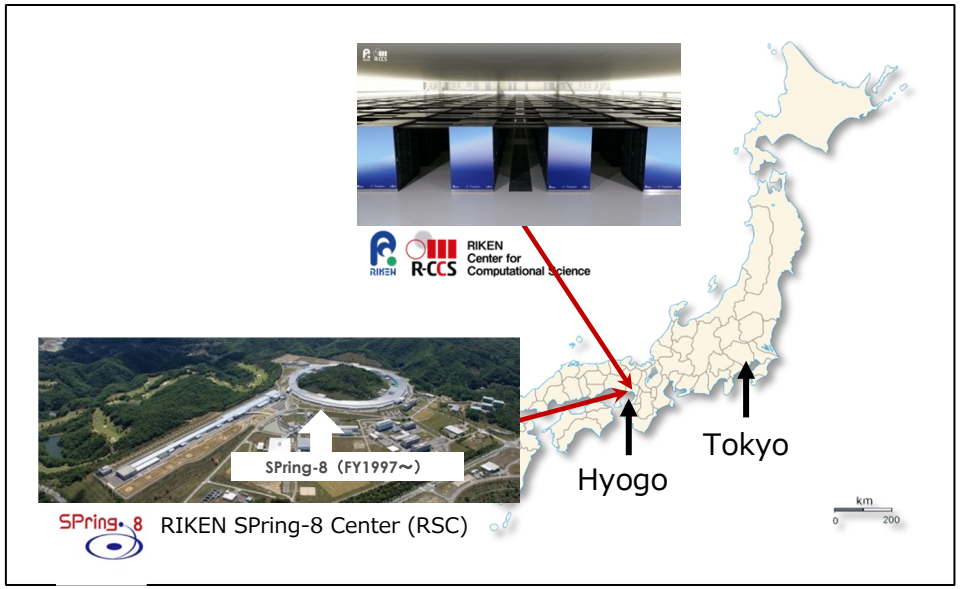
WIP

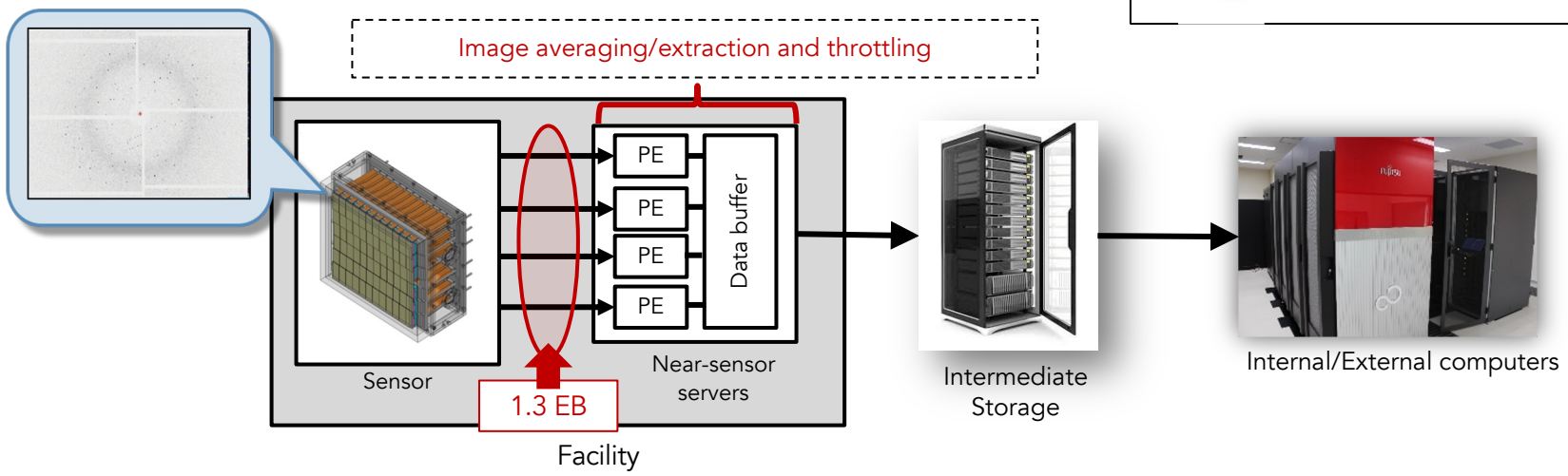
- **Background**
  - Scientific AI workflows span **data transfer, storage, model training, inference, and customized data processing & analysis**, requiring tight integration.
  - Manual handling slows research and increases risk of errors.
  - **Security is critical** in some use cases (e.g., handling pre-published or sensitive scientific data).
- **Approach**
  - **Data pipeline**: Parallel, secure transfers with/without Globus.
  - **Storage & automation**: Leverage **UMRS** (Emmanuel Jeannot) for pre/post-processing, embeddings, and data management.
  - **Training & fine-tuning**: Managed execution with monitoring of progress and resources.
  - **GenAI** services (Jens Domke): Network-isolation of models and data for security, supporting inference, RAG, MCP servers, etc.
  - **Extendibility**: Users can deploy additional services into the workflow.
  - **Centralized management layer** to coordinate data feeding, customized pre/post-processing, training status, and inference deployment.
  - Built with **scalability, automation, and secure handling of sensitive datasets** as guiding principles.



# Big Data Transfer and management

- **RIKEN operates SPring-8 large synchrotron radiation facility**
  - Located in the same Hyogo prefecture as R-CCS' Fugaku
- **Big data generation at SPring-8 facilities**
  - In 2017, SPring-8 public beamlines (26 BLs) generated 320 TB/year
  - In 2025, with the next generation detector (CITIUS), the facility is projected to generate 100-400PB of data per year
- **Data transfer is a first-step for data analytics**
  - Such data generated by sensors need to be transferred to internal/external computers for the further analysis




**Due to the large size, transferring data from a sensor to a computer and managing it is challenging.**

# Data compression by neural network (IEEE/ACM CCGrid'21[1])

- **X-ray CT data is one of popular data in RSC**
  - Scientists periodically take snapshots of a target specimen while rotating it (→ Time evolutionary data)
  - We target such time evolutionary data for the compression
- **Prediction is one of key aspects in data compression**
  - With accurate prediction, target data can be converted data with sequence of zeros
- **We use a NN (PredNet[2]) for this prediction**
  - Pre-processing (Training): Train a NN to learn the pattern of the movement of the specimen
  - Data (De)compression (Inference): Predict future images, compute delta and apply encoding/compressor
  - *PredNet: Deep recurrent convolutional neural network

| | | | |
|---|---|---|---|
| 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 (–)

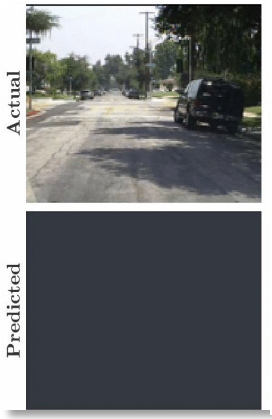
| | | | |
|---|---|---|---|
| 1.1 | 1.5 | 1.8 | 2.1 |
| 1.0 | 1.3 | 2.3 | 2.7 |
| 1.3 | 1.8 | 2.5 | 3.0 |
| 1.9 | 1.9 | 2.5 | 3.3 |

Predicted data

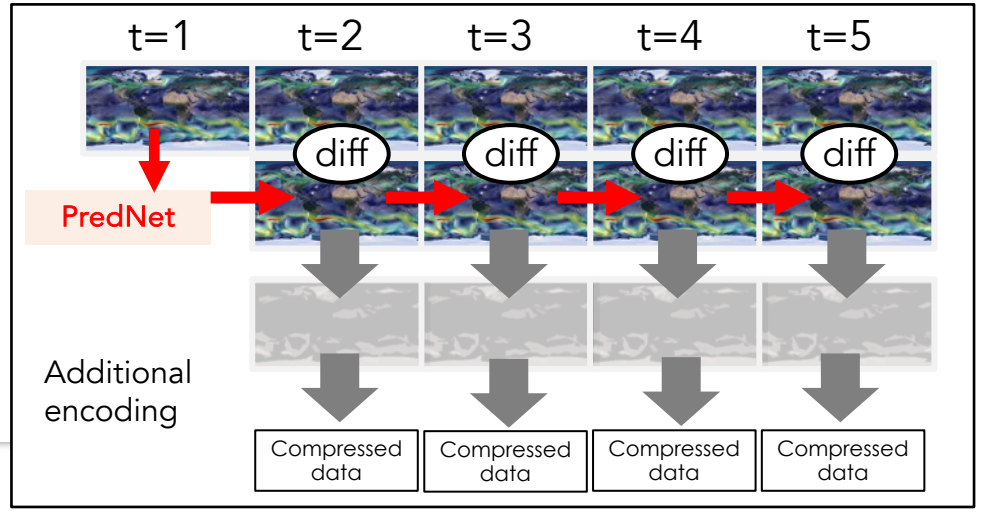
=

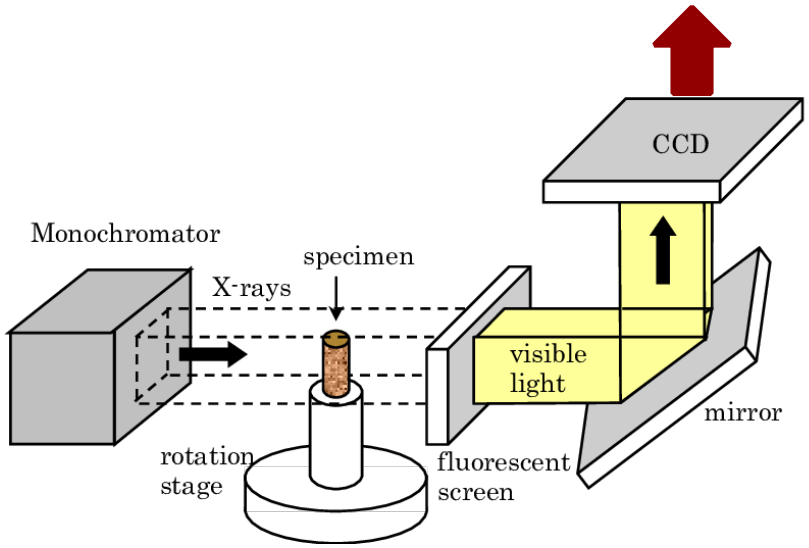
| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 0.1 | 0 | 0 |
| 0 | 0 | 0 | 0.1 |
| 0 | 0.2 | 0.1 | 0 |

Delta




PredNet (video from drive recorder)
https://coxlab.github.io/prednet/




X-ray CT system in SPring-8 [2]

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

# Compression of Time Evolutionary Image Data through Predictive Deep Neural Networks (IEEE/ACM CCGrid 2021 [1])

Rupak Roy[†1], Kento Sato[†2], Subhadeep Bhattacharya[†1], Xingang Fang[†1], Yasumasa Joti[†3], Takaki Hatsui[†4], Toshiyuki Hiraki[†4], Jian Guo[†2] and Weikuan Yu[†1]
†1: Florida State University, †2: RIKEN R-CCS, †3: JASRI, †4: RIKEN RSC,

- **Background:**
  - The next-generation detector (CITIUS) in the SPring-8 Center (RSC) generate about 100~400 PB
  - To analyze and/or train an AI model with the data, data transfer from the sensors to a large-scale computer is necessary
  - However, the transfer of large data becomes a performance bottleneck for this data pipeline
- **Approach (Figure 1):**
  - We has been developing and enhancing an AI-based data compression tool (TEZip)
  - The AI model predicts or reconstruct target images and TEZip only store the delta values
    - E.g. ) PredNet: 1st image fame –[predict]-> 2nd, 3rd,⋯ Nth image frame
    - B: Original image frames, P: Predicted image frames, D: Difference between B and P, C: Compressed image frames by a series of encoding
  - Pre-processing (by AI Training): Train a NN to learn the pattern of the movement of the specimen
  - Data (De)compression (by AI Inference): Predict future images, compute delta and apply encoding/compressor
- **Results (Figure 8, 10):**
  - TEZip gives higher compression ratio than major compression tools (e.g., X.265, SZ)
  - Lossless mode: 9 to 15 / Lossy mode (w/ a few % of errors): 40 to 50

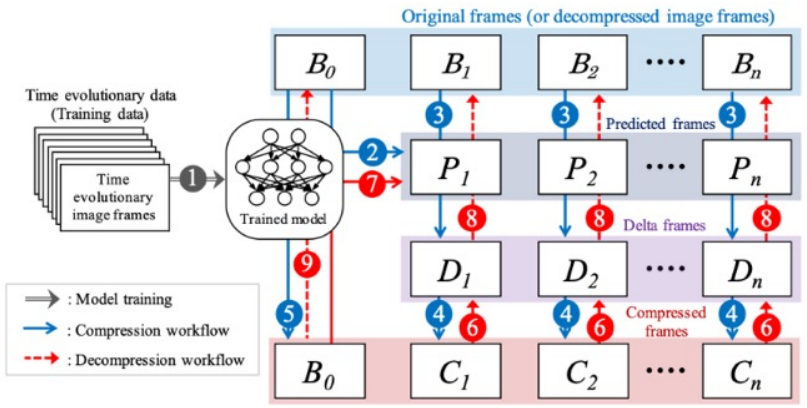



Fig 1. Workflows of TEZIP (de)compression

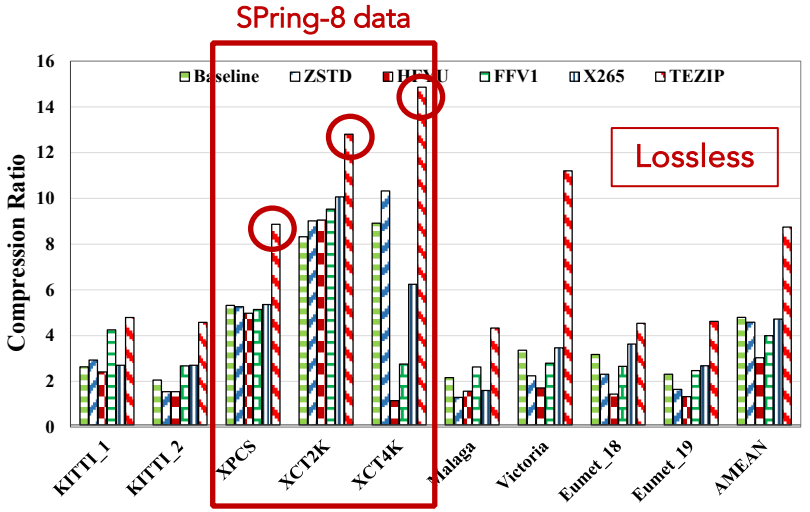



Fig. 8. Compression ratio with lossless compressors.

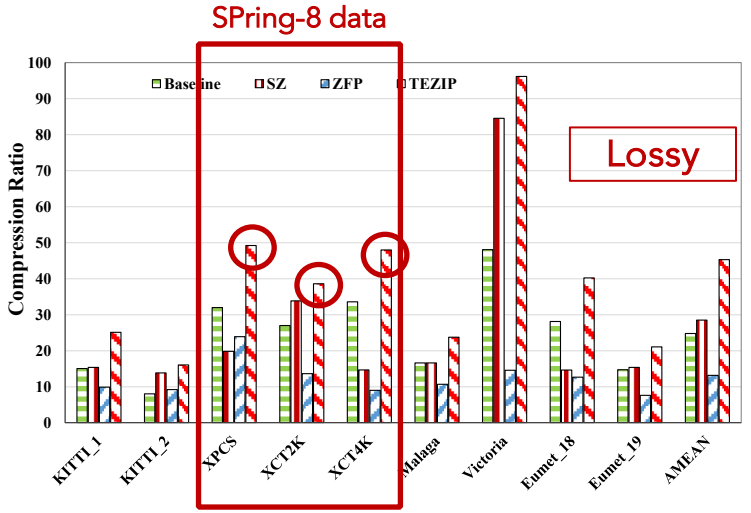



Fig. 10. Compression ratio with different lossy compressors

# Integrating TEZip into FZ: Leveraging FZ's Multi-Capabilities

Amarjit Singh[†1], Kento Sato[†1], Robert Underwood[†2], Franck Cappello[†2]
†1 : RIKEN Center for Computational Science, †2 : Argonne National Laboratory

WIP

- **Background:**
  - **TEZip** is AI-Based Predicted Time Evolutionary Model for data compression
  - **FZ framework (Figure 5)**: Comprehensive ecosystem to enable scientific users to intuitively research also compose, implement, and test specialized lossy compressors
  - From a library of predeveloped, high-performance data reduction modules optimized for heterogeneous platforms
  - **FZ Approach**: Leverage and adapt multiple existing capabilities:
    - SZ lossy compressor, LibPressio unifying compression interface, OptZConfig optimizer of compressor configurations, Z-checker/QCAT quality analysis tools, and Paraview and VTK visualization tools
- **Approach:**
  - **Extended Compression Capabilities**: Integrating TEZip with FZ expands the overall compression performance and flexibility
  - **Resource Leverage**: TEZip can directly leverage FZ's compressor resources for improved adaptability
  - **Unified Scientific Compression Platform**: FZ brings numerous scientific compression facilities together under one framework
  - **Predictive Advantage**: TEZip prediction algorithm enhances compression efficiency within FZ
  - **Interfacing via LibPressio**: LibPressio acts as the intermediate layer to connect TEZip with FZ, enabling integration
- **Results:**
  - **Earlier integration TEZip -> Libpressio efforts (Figure 1)**
  - **TEZip –> FZ Integration (In Progress):** Ongoing work to link TEZip with the FZ framework
  - **FZ Module Decomposition**: Exploring modular decomposition of FZ for integration

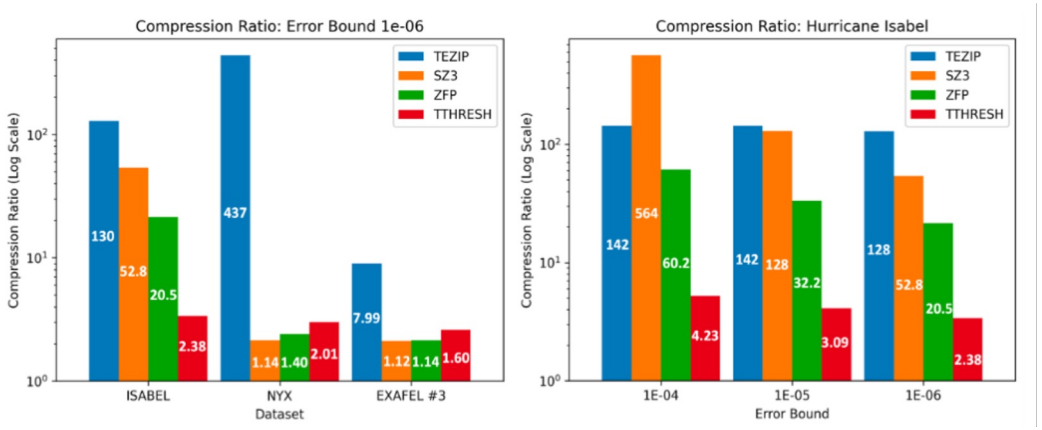



Figure 1: Compression Ratio for TEZIP, SZ3, ZFP, and TTHRESH for all three datasets at error bounds 1e-04 to 1e-06 and Compression Ratio for Hurricane Isabel data at Error Bounds 1e-04 to 1e-06

[1] I. Talukdar, A. Singh, R. Underwood, K. Sato, W. Yu. "Integrating TEZip Into LibPressio: A Case Study of Integrating a Dynamic Application into a Static C Environment" Poster @SC'23

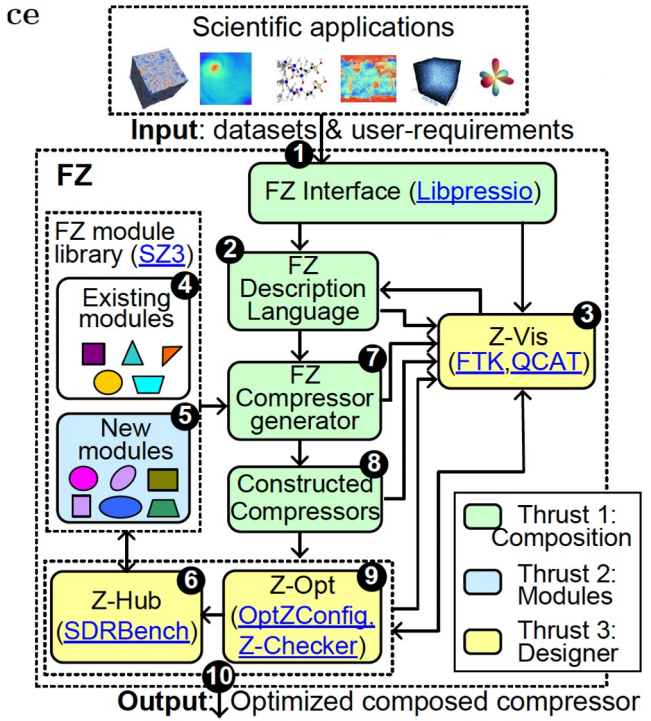



Figure 5: FZ design overview.

# AutoCheck: Automatically Identifying Variables for Checkpointing by Data Dependency Analysis (IEEE/ACM SC24[47])

Xiang Fu[†1], Weiping Zhang[†1], Xin Huang[†1], Wubiao Xu[†1], Shiman Meng[†1], Luanzheng Guo[†2], Kento Sato[†3]
†1: Nanchang Hongkong University, †2: PNNL, †3: RIKEN,

- **Background**
  - Checkpoint-restart (C/R) is one of big I/O workloads in supercomputers
  - To facilitate C/R, an application-level C/R tool (VeloC) is installed in Fugaku
  - However, while we can specify only variables needed to be checkpointed, thereby, reduce the size of checkpoint data, finding all necessary variables to restart is challenging in large programs
- **Approach**
  - To remove such manual coding, we developed AutoCheck which automatically detect variables to be checkpointed by static and dynamic data dependency analysis
- **Results**
  - We successfully detects all the necessary variables for checkpointing in micro-benchmarks and proxy apps
  - AutoCheck significantly reduces checkpoint size compared to a system-level C/R tool, BLCR
- **Github: https://github.com/zRollman/AutoCheck**

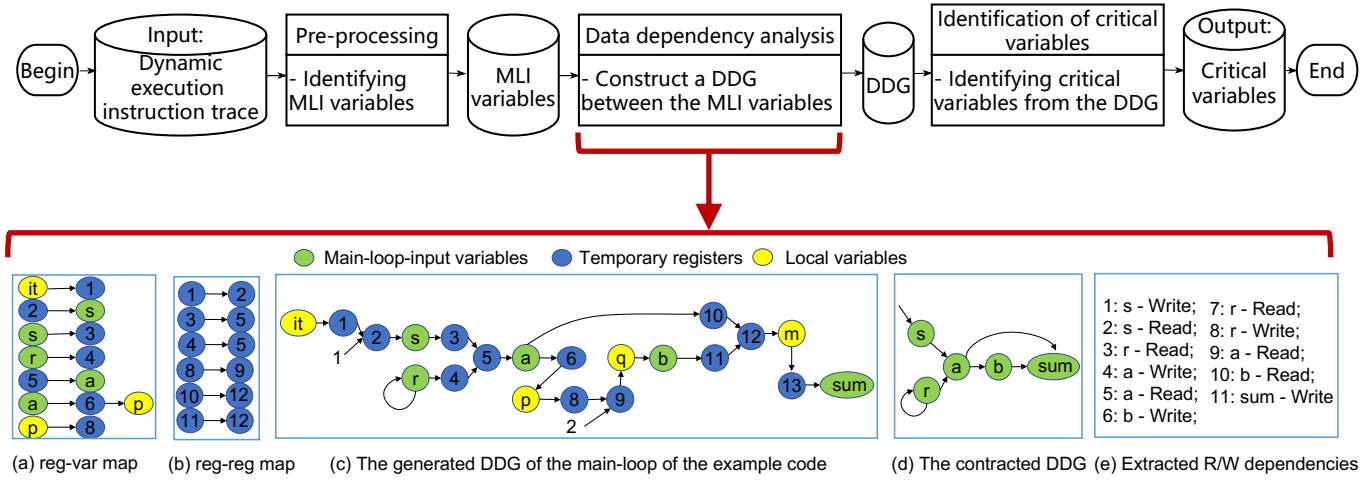



TABLE IV
STORAGE COST FOR CHECKPOINTING.

| Name | Input size | BLCR [9] (MBs) | AutoCheck (MBs) |
|---|---|---|---|
| Himeno | $129 \times 65 \times 65$ | 32550.76 | 2.53 |
| HPCCG | $64 \times 64 \times 64$ | 452202.50 | 610.9 |
| CG | S | 16569.47 | 0.16 |
| MG | S | 3220.39 | 2.84 |
| FT | S | 53616.26 | 24.6 |
| SP | S | 20068.88 | 7.81 |
| EP | S | 50061.67 | 0.03 |
| IS | S | 952.85 | 2.53 |
| BT | S | 34042.18 | 4.69 |
| LU | S | 17263.79 | 9.33 |
| CoMD | $-x$ 8 $-y$ 8 $-z$ 8 | 375798.50 | 241.71 |
| miniAMR | $-nx$ 8 $-ny$ 8 $-nz$ 8 $-max\_blocks$ 8 | 30310.90 | 0.09 |
| AMG | -problem 2 -n 40 40 40 | 647577.68 | 58.83 |
| HACC | $-N$ 8 $-t$ $16 \times 16 \times 16$ | 837533.14 | 334.93 |

[47] Xiang Fu, Weiping Zhang, Xin Huang, Wubiao Xu, Shiman Meng, Luanzheng Guo, Kento Sato, "AutoCheck: Automatically Identifying Variables for Checkpointing by Data Dependency Analysis", SC '24: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA

# An Optimization Technique for Hiding Communication Costs in 3D Parallel Training of DL (IEEE/ACM CCGrid2025 [59])

Ryubu Hosoki [†1], Kento Sato [†2], Toshio Endo [†1], Julien Bigot [†3], Edouard Audit [†3]
†1: Institute of Science Tokyo, †2: RIKEN R-CCS, †3: CEA

- **Background**
  - DNN models have grown rapidly in accuracy, but this progress has come with a training cost (e.g., 100Bs-T params)
  - Training large models takes tremendous time and memory capacity → Parallel training, but challenging in decomposition
  - Auto-parallelization (e.g., Alpa) finds the optimal balance of DP/TP/PP without expertise for parallel training in HPC systems
  - An XLA (Accelerated Linear Algebra) compiler in Alpa produces inefficient all-reduce stages in PP backward computation [**Fig.1**]
  - XLA is a domain-specific compiler, XLA: high-level computation graph (e.g., JAX, PyTorch) → Optimized machine code
- **Approach: Comm-Shift Optimization at the level of an XLA compiler used in Alpa**
  - We analyze computation graph and shift gradient-averaging communication from backward to parameter update [**Fig.2**] → eliminate synchronization time from one pipeline stage from another → Reduce overall training times
- **Results**
  - Comm-Shift improves the training performance across various models up to 27% at maximum (GPT-J-6B) [**Table 1**]
  - Improvement becomes more significant with more communication in larger models

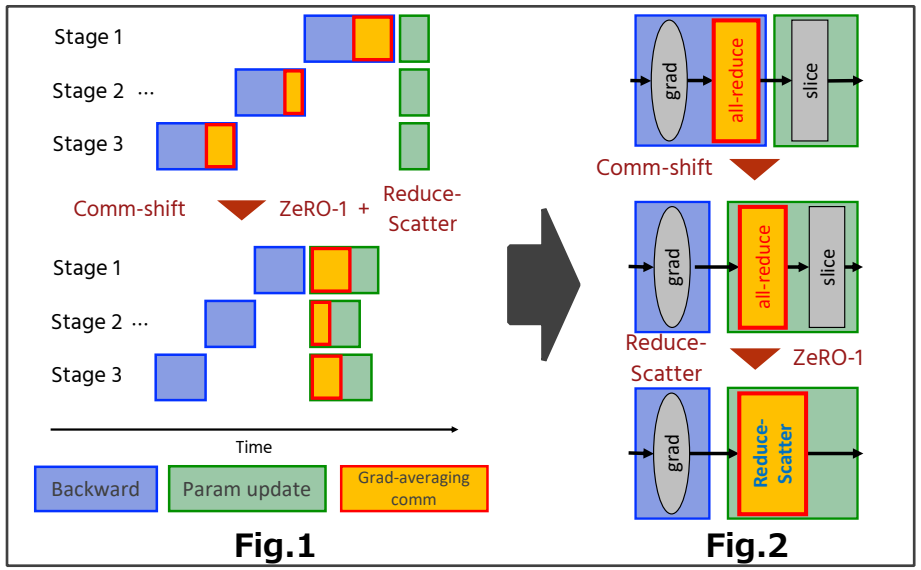



**Fig.1** **Fig.2**

**Table 1: Experimental results on TSUBAME4.0 (NVIDIA H100 SXM5 /InfiniBand NDR200 x4 (Fat Tree))**

| Task | Model | # of GPUs | Strategy | Throughput w/o comm-shift opt. | Throughput w/ comm-shift opt. | Speedup |
|---|---|---|---|---|---|---|
| NLP | GPT-2 Small | 16 | [(8x1), (8x1)] | 2019190 token/s | 2073480 token/s | +2.69% |
| | GPT-2 Large | 16 | [(8x1), (8x1)] | 403880 token/s | 419156 token/s | +3.78% |
| | GPT-2 XL | 32 | [(16x1), (16x1)] | 477236 token/s | 508616 token/s | +6.58% |
| | GPT-J-6B | 32 | [(8x1), (4x2), (4x2), (4x2)] | 143158 token/s | 181812 token/s | +27.00% |
| | Mamba-1.4B | 16 | [(2x1), (2x1), (2x1), (2x1), (2x1), (2x1), (2x1), (2x1)] | 29786 token/s | 30913 token/s | +3.79% |
| Image Classification | ViT-base | 8 | [(2x1), (2x1), (2x1), (2x1)] | 1192 image/s | 1306 image/s | +9.55% |
| | SwinV2-L | 16 | [(4x1), (2x1), (4x1), (4x1), (2x1)] | 1827 image/s | 1893 image/s | +3.60% |
| | CoAtNet-7 | 16 | [(8x1), (8x1)] | 384 image/s | 444 image/s | +15.36% |

[59] Ryubu Hosoki, Kento Sato, Toshio Endo, Julien Bigot, Edouard Audit, "An Optimization Technique for Hiding Communication Costs in 3D Parallelism", In the proceedings of The IEEE International Symposium on Cluster, Cloud, and Internet Computing (CCGrid 2025), May, 2025

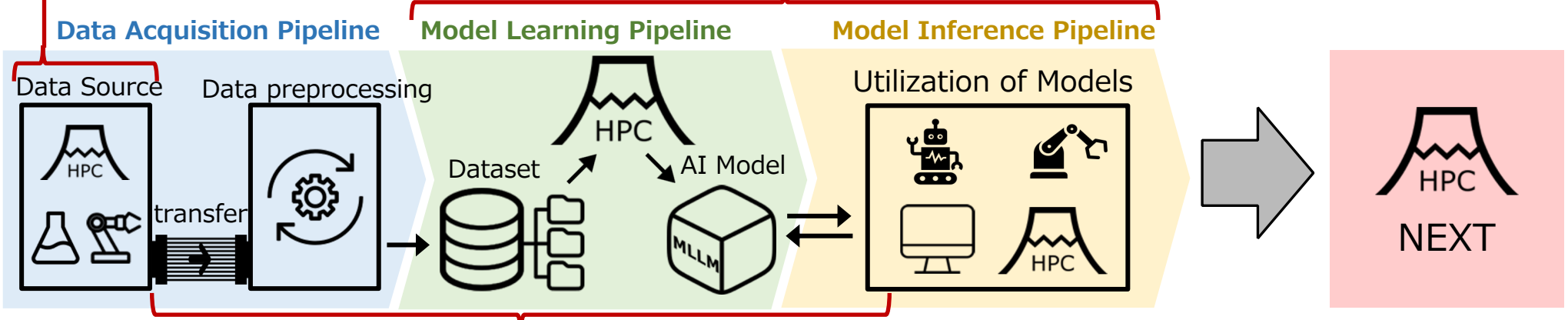
# Summary: Research Plan and End-user software adoption strategy

(T2) Data aggregation | (T2) MPI I/O | (T2) Debugging tools | (T2) AutoCheck

**Advanced simulation**

(T2) 3D parallelisms | (T2) C/R for AI | (T2) Data plancement

**Advanced AI training/inference**




**Data compression for data management**

(T1) Text data compression | (T1) TEZip enhamcement

(T1) Image data compression | (T1) XCT data compression

**Next-Gen Computing Infra. Research**

(T3) DSE

**AI platform development and AI-assisted next-generation operations**

(T1) AI Pipeline Platform | (T3) Advanced Operaton by AI | (T2) Energy-aware Job Scheduling

We aim to develop “**default-on**" software incorporated into the platform so that end-users adopt our software without consciously