HIGH PERFORMANCE BIG DATA RESEARCH TEAM

【機密性1（公開）資料】
[Confidentiality class 1 (Public) document]

InPEX 2026:
(April 23-25, 2026)

In-house AI pipeline infrastructure towards the FugakuNEXT era

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 professor, Tohoku graduate university
- Adjunct lecturer, Osaka graduate university

v1.7

InPex: Computing Contimuum – Use case characterization criteria

1. Access to HPC resources (as a community)
2. Interface to a federation of resources (e.g. EuroHPC)
3. Co-design for the post-exascale systems
4. Portability of the benchmarks and codes
5. Deployment issues of the workflows and applications
6. Metrics for different deployment scenarios
7. Semantics and quality of data
8. Resources provisioning/scheduling across the continuum
9. End-to-end workflow control
10. Multitenancy
11. Data logistics
12. Energy efficiency
13. Reliability
14. Real-time data streaming at high rates

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)

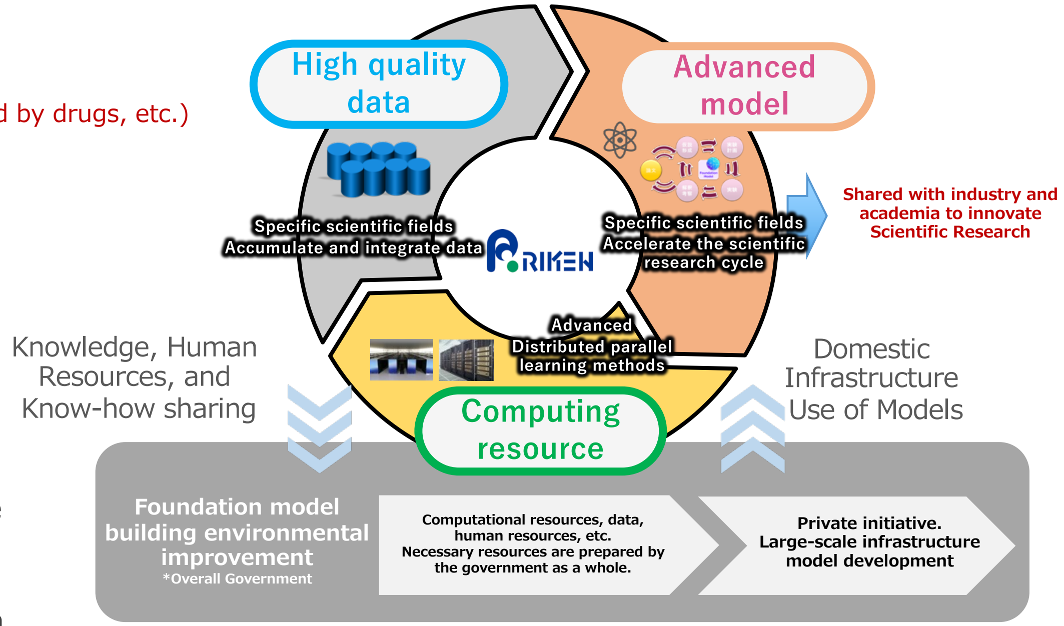
Advanced model

- 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 "



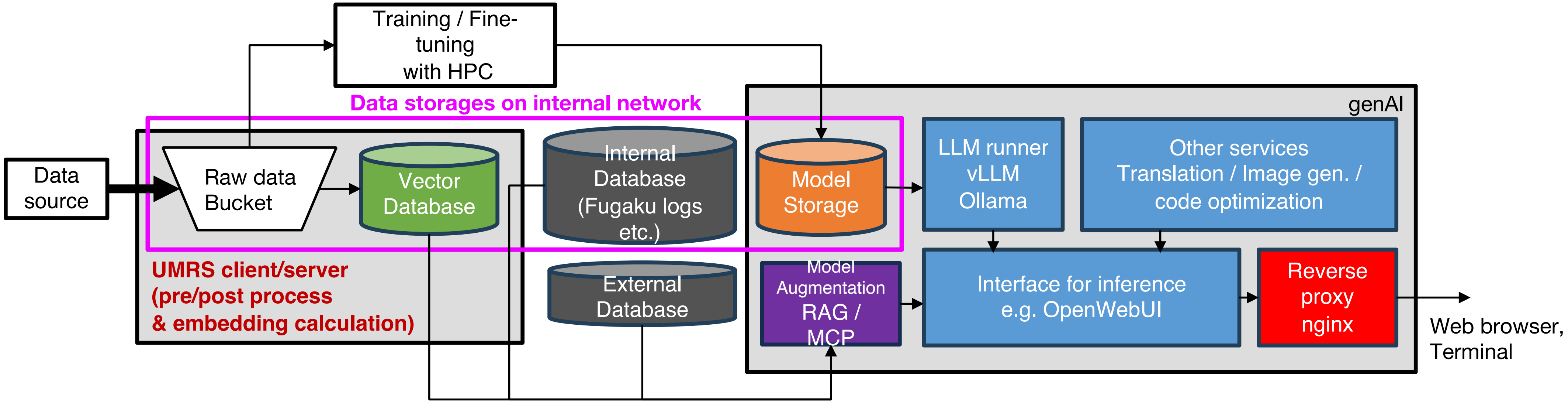
* Scientific infrastructure model: An infrastructure model tailored for a specific scientific research field (domain) by additionally learning and inferring scientific research data (papers, real-time experimental and simulation data, etc.) to the infrastructure model (language, images, etc.).

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, data curation, model training, inference, and customized data processing & analysis**, requiring tight integration.
  - Manual handling slows research and increases risk of errors.
  - **Security is critical** in several use cases (e.g., handling pre-published, sensitive scientific data).
- **Approach**
  - **Data pipeline**: Parallel, secure transfers with/without Globus.
  - **Storage & automation**: Leverage **UMRS** for pre/post-processing, embeddings, and data management.
  - **Training & fine-tuning**: Managed execution of model training and fintuning with monitoring of progress and resources.
  - **GenAI** services: Network-isolation of models/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.



Open challenges in the AI pipeline/workflow

- Trade-off analysis between near-sensor processing vs. data transfer to HPC and processing in HPC
- Efficient data transfer from sensors, instruments, and experimental facilities to HPC systems
- Fast and scalable search over vector databases for scientific data
- Model management, versioning, and selection of appropriate models
- Optimized deployment of inference frameworks such as vLLM and Ollama
- Secure use of open models to prevent leakage of confidential or pre-publication data
- Seamless integration of individual components in the AI workflow
- OSS deployment of a series of reusable workflow components
- Monitoring and traceability of workflow execution and model behavior
- Portability of the workflow across different HPC/AI platforms, different research communities and different domains
- Resource scheduling for mixed workloads: data processing, training, and inference

Compression of Time Evolutionary Image Data through Predictive Deep Neural Networks (IEEE/ACM CCGrid 2021 [15])

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:**
  - 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 (**Fig.1**)
    - 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 (**Fig.1**)
  - Data (De)compression (by AI Inference): Predict future images, compute delta and apply encoding/compressor (**Fig.1**)
- **Results (Fig.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

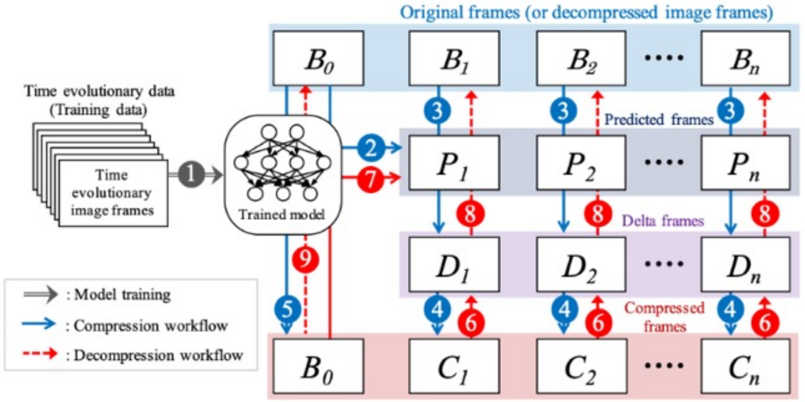


Fig 1. Workflows of TEZIP (de)compression

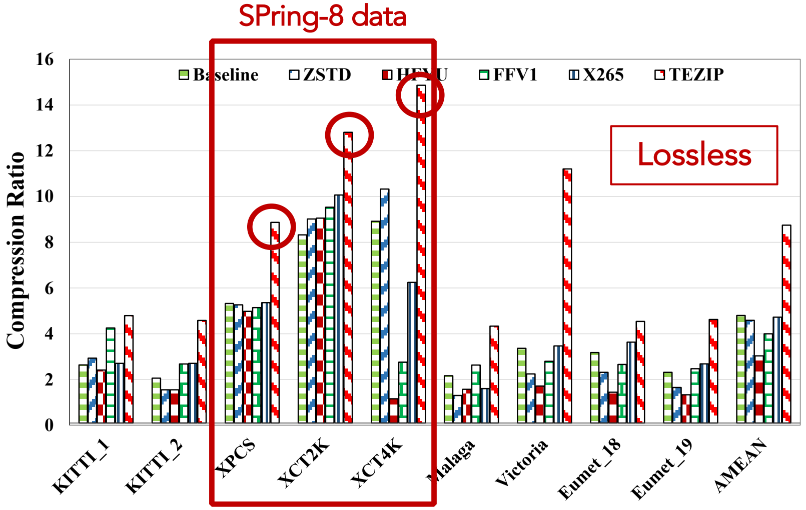


Fig. 8. Compression ratio with lossless compressors.

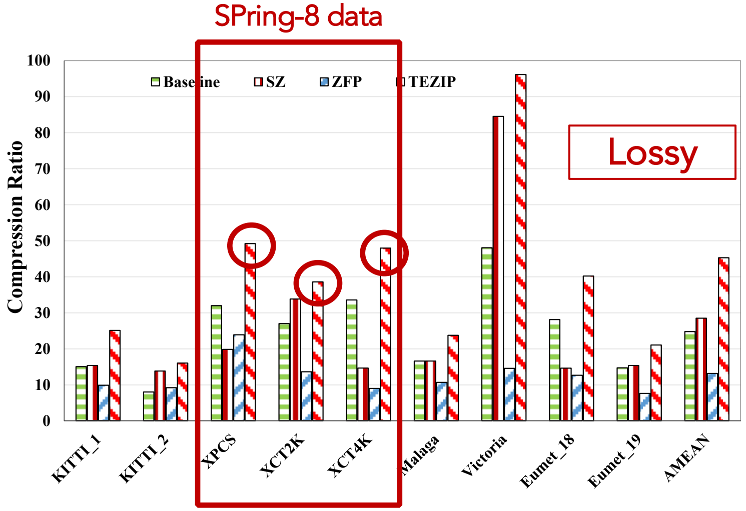


Fig. 10. Compression ratio with different lossy compressors

[15] 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