Institute of SCIENCE TOKYO

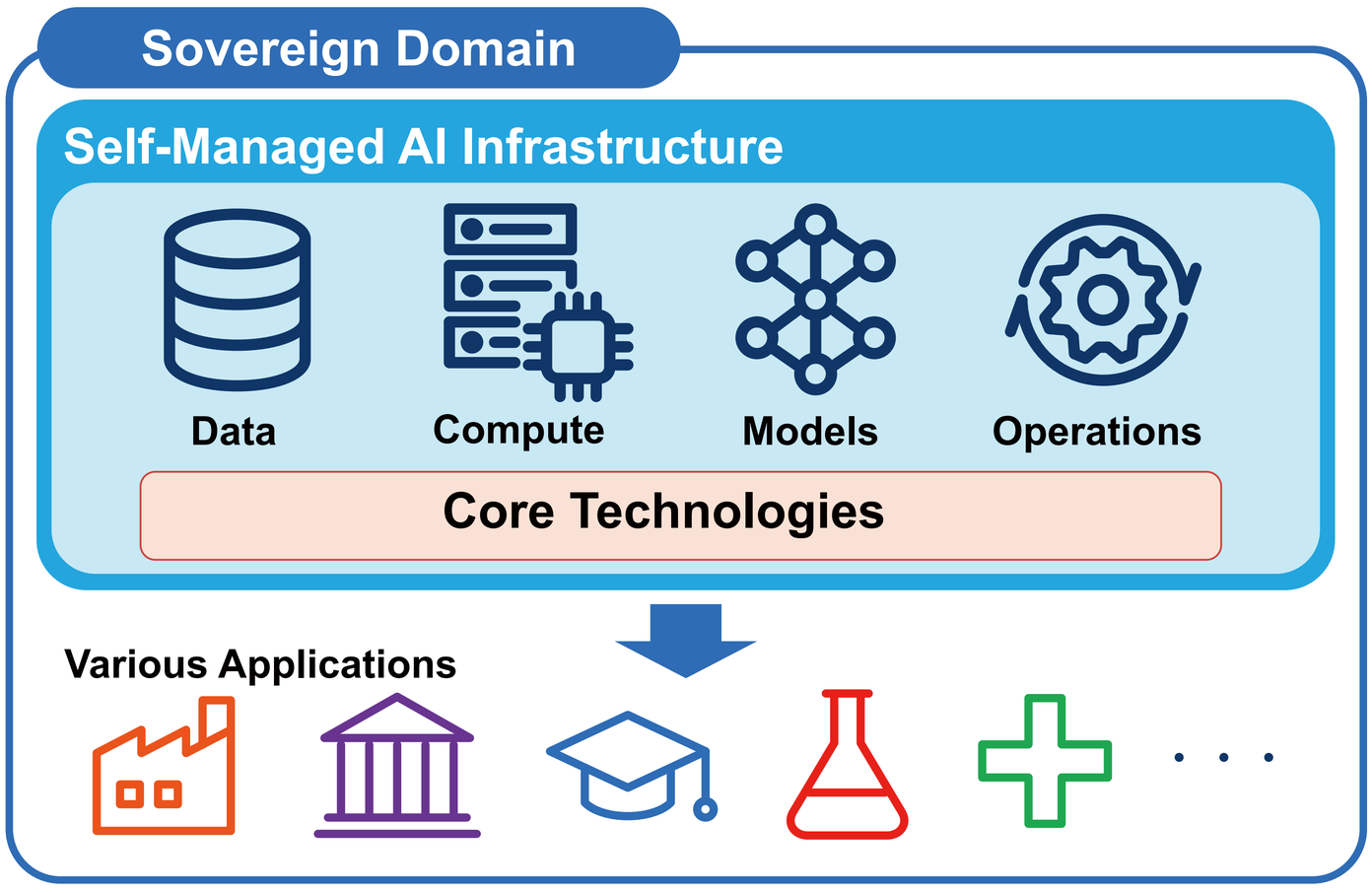
FUJITSU

Institute of Science Tokyo - Fujitsu
Quantum & HPC Infrastructure Collaborative Research Cluster

Towards Sovereign AI

Efficient Training/Inference on Limited Computing Resources

Realizing Sovereign AI



Building and operating AI under sovereign control

Challenge 1: High Cost of AI Development and Operation

- Massive power consumption from continuous operation of large GPU clusters

Challenge 2: Limited Computing Resources

- AI compute demand continues to grow
- Compute capacity cannot scale at the same pace

Challenge 3: Limits of Conventional Optimization

- Model and system optimization have already delivered major gains
- Further gains require rethinking AI workloads and architectures

Goal: Build the Foundation for Efficient AI Training/Inference

- Accelerate AI training and inference through efficient use of computing resources
- Explore next-generation AI architectures through workload and system co-optimization

100x Lower Training Cost

Current Research Areas

Long-Context Training

Growing Demand for Long Contexts

- Source code
- Conversation history
- Domain-specific data, …

→ **Broader and more capable AI applications**

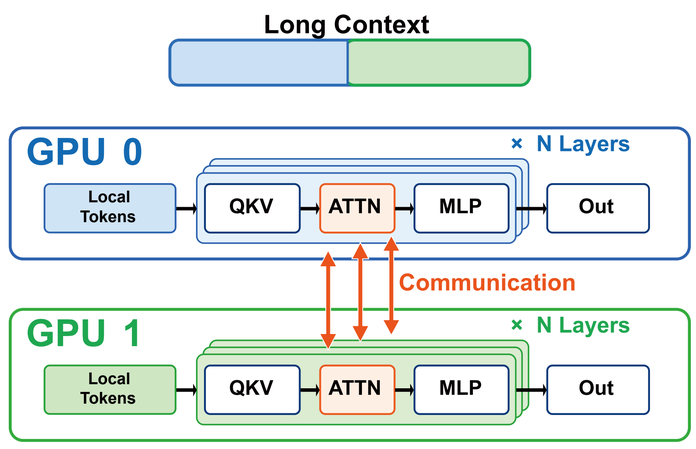
Challenge

Increasing memory and compute costs

→ **More GPUs and longer training time**

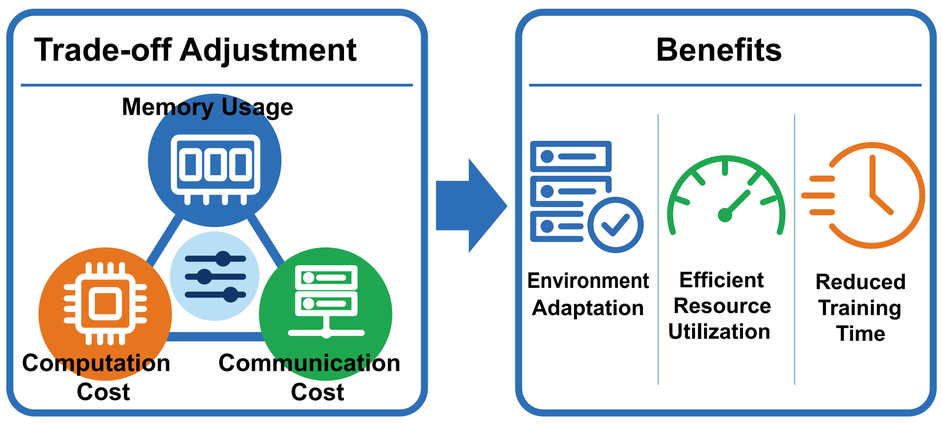
Limited parallelization support for diverse attention patterns

General-Purpose Context Parallel Implementation



- **Support Diverse Attention Patterns**
- **Skip Redundant Computation**
- **Comm-Comp Overlap in Forward & Backward Passes**

Distributed Training Optimization



- **Optimize Parallelization Strategies**
- **Flexible Activation Recomputation Control**

~40% Training Time Reduction for Long-Context Learning

Future Directions

- Integration and Further Optimization
- Automatic Tuning
- Next-Generation AI Architecture Exploration

https://www.isct.ac.jp/