FUJITSU

Self Evolving Specialized LLMs on HPC: Efficiency via 1 bit Quantization and AI Computing Broker, Toward NPUs

Koichi Shirahata
Fujitsu Limited
June 24, 2026

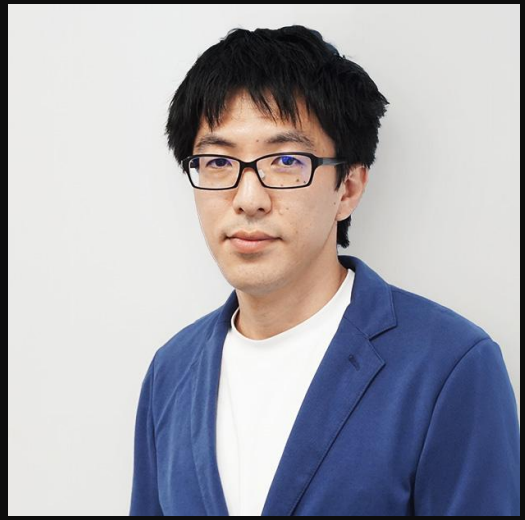
Self-introduction

FUJITSU

Koichi SHIRAHATA

Senior Project Director - Artificial Intelligence Laboratory, Fujitsu Research, Fujitsu Limited

- March 2015: Received Ph.D. at School of Information Science and Engineering, Tokyo Institute of Technology
- April 2015: Joined FUJITSU LABORATORIES LTD.
- November 2020, 2021: Achieved the world's highest performance in MLPerf™ HPC, a machine learning performance benchmark, using the Fugaku supercomputer and ABCI
- May 2024: Development of a distributed parallel training method for large-scale language models in the policy response framework of the supercomputer "Fugaku"



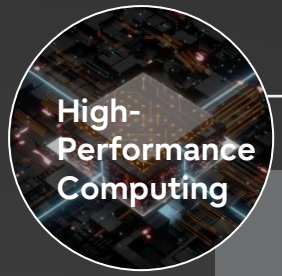
Fujitsu AI Strategy

FUJITSU

Providing innovative **sovereign** AI platforms for enterprise customers

Computing for AI – FUJITSU-MONAKA(CPU/NPU)

FUJITSU



2011 | 2020 | 2027 | 2029 | 2031

SPARC64

45nmCPU

K computer

- Achieving the world's highest performance
- Ultra-parallel SPARC supercomputer

A64FX

7nm CPU

Fugaku

World's first Arm-based No. 1 supercomputer

FUJITSU-MONAKA

2nmCPU

- High-speed data processing
- Energy efficiency /High performance
- Reduced TCO
- Reliability and Security
- Open architecture

FUJITSU-MONAKA-X

1.4nm CPU + 2 nm NPU

HPC/AI processing optimal design

HPC

Ultra-many-core, SIMD extension

AI

NPU, GPU tightly coupled

FUJITSU-MONAKA-XX

sub1.4nm CPU

- AI Agent
- Quantum integration
- CPO (Co-Packaged Optics)

Fujitsu Kozuchi AI Platform for Sovereign AI

FUJITSU

Transforming business processes and corporate culture to achieve high productivity and employee satisfaction

Domain-Specific
Generating AI technology reflecting each company's unique strengths

Flexibility
Optimizing self-evolving learning,and large-scale agent collaboration across organizations

Security
Addressing evolving LLM vulnerabilities and ensuring confidentiality and safety in cross-organizational agent collaboration

Kozuchi AI Security Platform

Addressing new cyber threats

Predictive Resilience
Anticipating the future and proactively addressing emerging risks

Creating Safety and Security in the Information Society

AI Scanner/ Guardrails | Multi-agent Security | Secure inter-agent gateway

FUJITSU

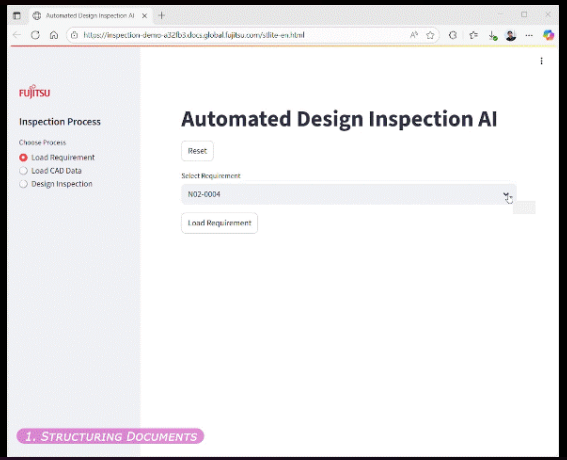
Self-Evolving Multi-AI Agent Technology

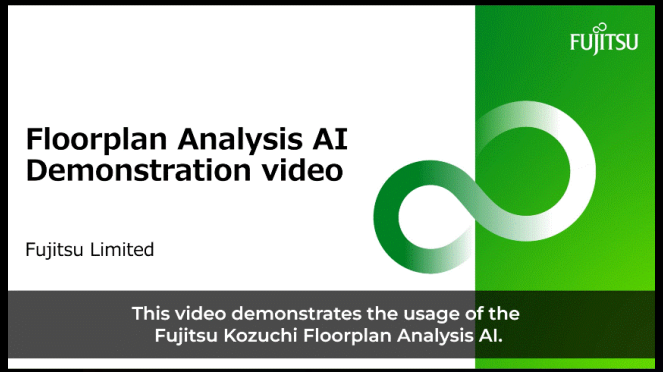
Challenges in fields such as manufacturing and construction are difficult to solve with general-purpose AI: Fujitsu has developed "Takane," a business-specific large language model

FUJITSU

Manufacturin

CAD Automatic Inspection





Constructio

Floor Plan Analysis

Proble

Because specialized design requires expert knowledge, there are limits to the precision that can be achieved by human designers.

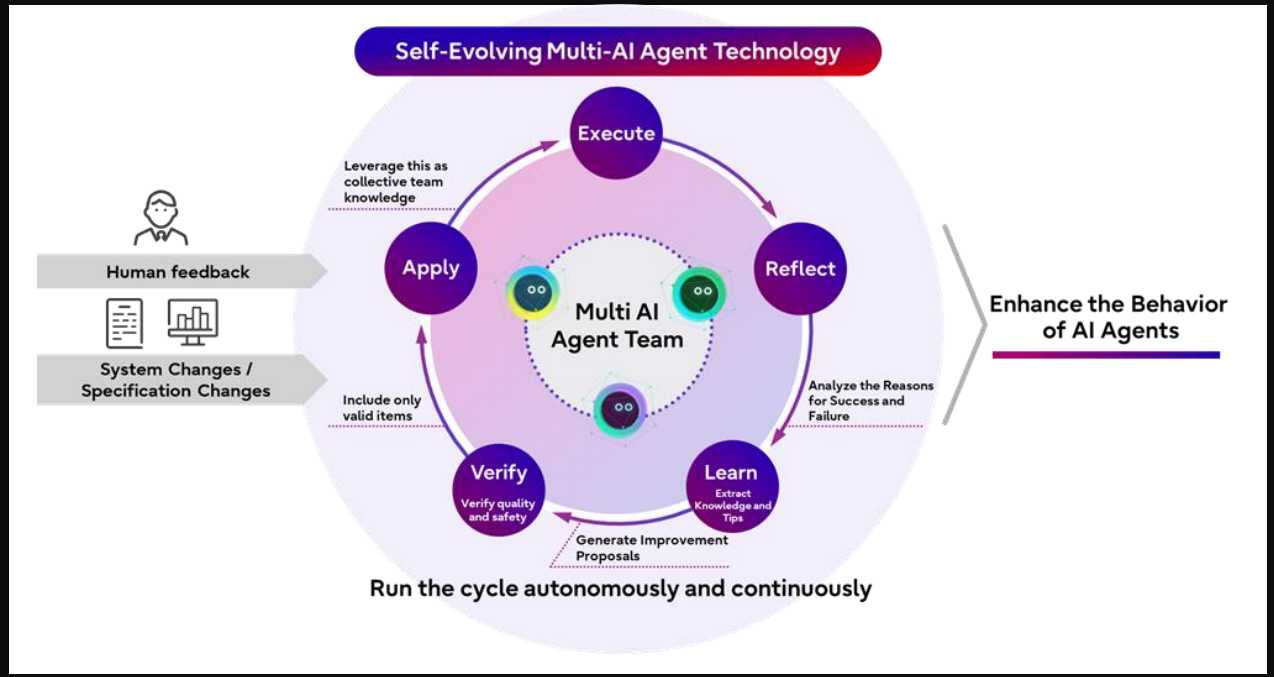
Model training, evaluation, and improvement take too much time, making it impossible to roll out the initiative across the organization

We have developed technology that autonomously generates and evolves specialized AI through a cycle of operational analysis and learning

Self-Evolving Multi-AI Agent Technology

FUJITSU

- Enabling automatic enhancement of the business-specific LLM "Takane" and safe self-evolution of AI agents
- Enables multiple AI agents to perform tasks as a team, continuously and safely learning from daily execution results, human feedback, policy revisions, and specification changes.



Inspection Tasks in Manufacturing Case Studies

Check the conformity of CAD data using a test program generated from the requirements specification

Requirements Document

Requirement ID	L02-0001	Requirement Name	Constraints on the installation locations of hexagonal rivets for the rear door interior panel reinforcement					Remarks	Scope of Disclosure
Part Name	Rear Door	Category	Assembly Ease	Rank	Required	Target	All	-	Public

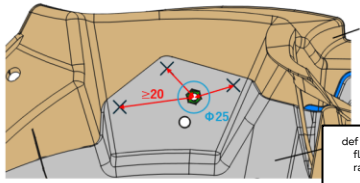
Ensure a flat surface with a diameter of 25 mm at the installation points for the hexagonal rivet nuts on the rear door inner panel reinforcement.
In addition, ensure that the distance from the hex rivet nut to the spot weld is at least 20 mm.

≥20 mm

(Reason)
1. Ensuring a flat surface
A 25 mm diameter flat surface must be secured to improve workability with the riveter.

For the installation location of the hexagonal rivet nut on the inner panel reinforcement of the rear door, ensure a flat surface with a diameter of φ25.Additionally, maintain a distance of 20 mm or more from the hexagonal rivet nut to the spot welding point.

Design example



Inner panel of rear door

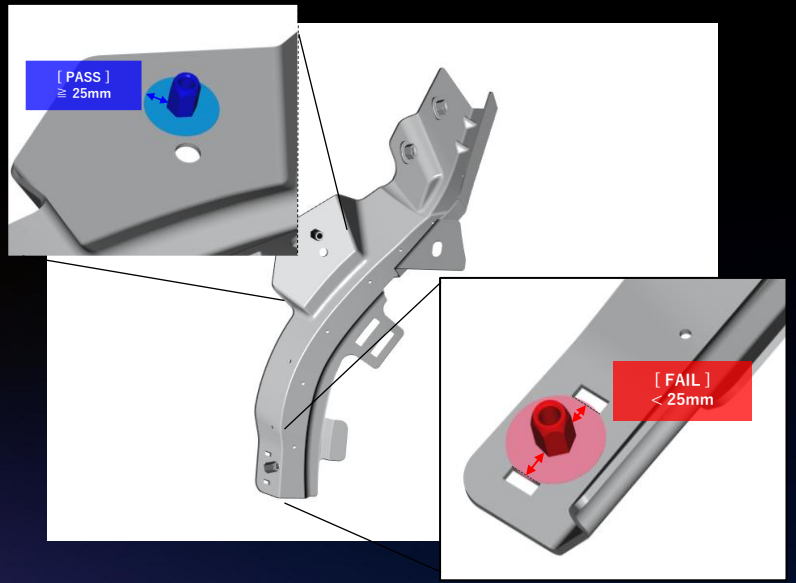
Inspection Program

```
def is_flat(surface_mesh, rivet_nut_info_list):
    flat_results = []
    radius = 12.5  # 半径12.5mm

    for rivet_nut_info in rivet_nut_info_list:
        rivet_nut_center = np.array(rivet_nut_info["center_on_panel"])

        # 半径範囲内のメッシュのポイントを取得
        in_radius_indices = np.where(np.linalg.norm(surface_mesh.vertices -
rivet_nut_center, axis=1) <= radius)[0]
        # 対応する頂点を抽出
        in_radius_vertices = surface_mesh.vertices[in_radius_indices]
```

CAD data to be inspected



TAKANE SELF-ADAPTING MULTI-AGENT SYSTEM

Self-Evolving Agents That Automatically Build Domain-Specific LLMs

Each agent understands the nature of the task and autonomously adapts its strategy, methods, and evaluation criteria while cooperating dynamically

Self-Evolving | Dynamic Adaptation | Fully Automated | Proven in 4 Domains

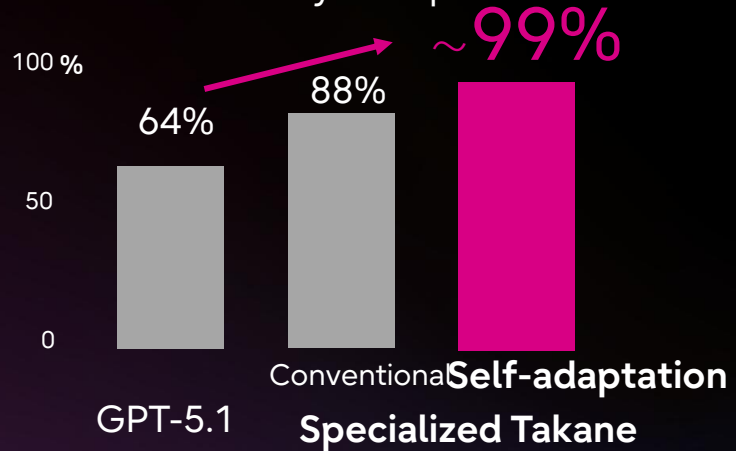
Development Technologies and Customer Value

FUJITSU

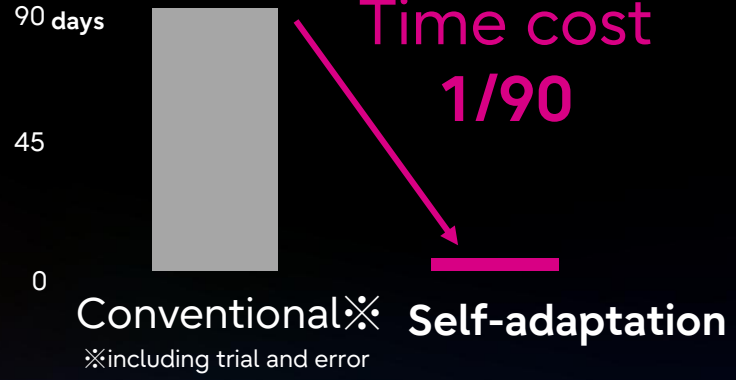
Self-evolving multi-AI agent technology

Dynamic multi-agent technology that continuously updates the weights of multiple agents through reinforcement learning, enabling the entire team to specialize and stabilize through interaction

Accuracy Comparison



Development time for specialized AI



Customer Values

Automatically acquire "task-specific AI" with higher accuracy than human-created AI

Not limited to a single task, but scalable to multiple tasks and locations

We can deliver a highly accurate, customized Takane tailored to customer needs in as little as one day to one week

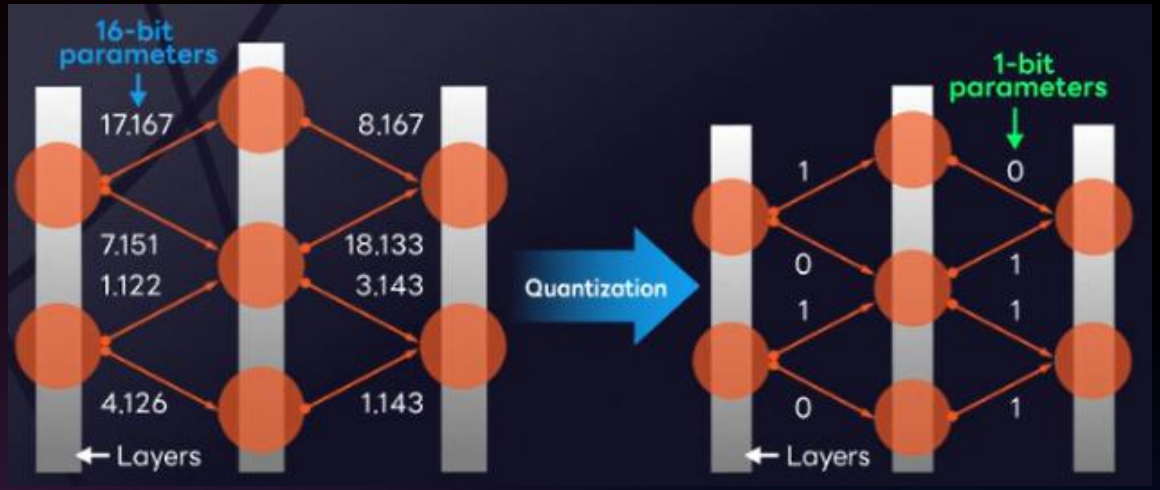
FUJITSU

1bit Quantization Technology

What is Quantization?

FUJITSU

A technology that compresses floating-point operations within the AI brain "neural network" by replacing them with simpler forms, thereby reducing the memory and computational load of AI models.

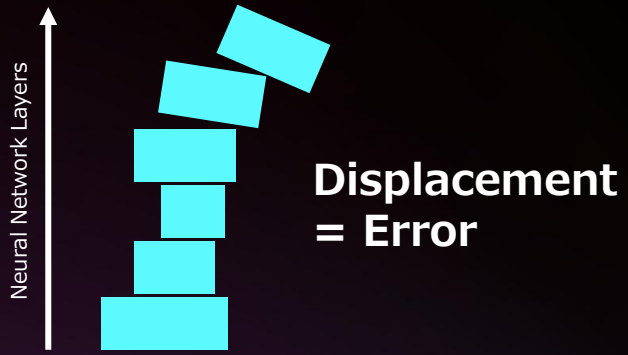


Challenge: since the accuracy decreases as the number of bits is reduced at 1 bit, no practical application exists

Our Developed technology

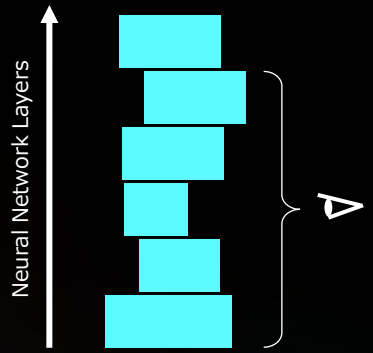
FUJITSU

Conventional technology



Displacement occurs in each layer, and as the number of layers increases, the displacement grows larger and collapses (precision breakdown).

QEP (Quantization Error Propagation)



Error propagation method

A method for adjusting while checking for misalignment to ensure the entire structure stacks straight

Calibration Data Utilization

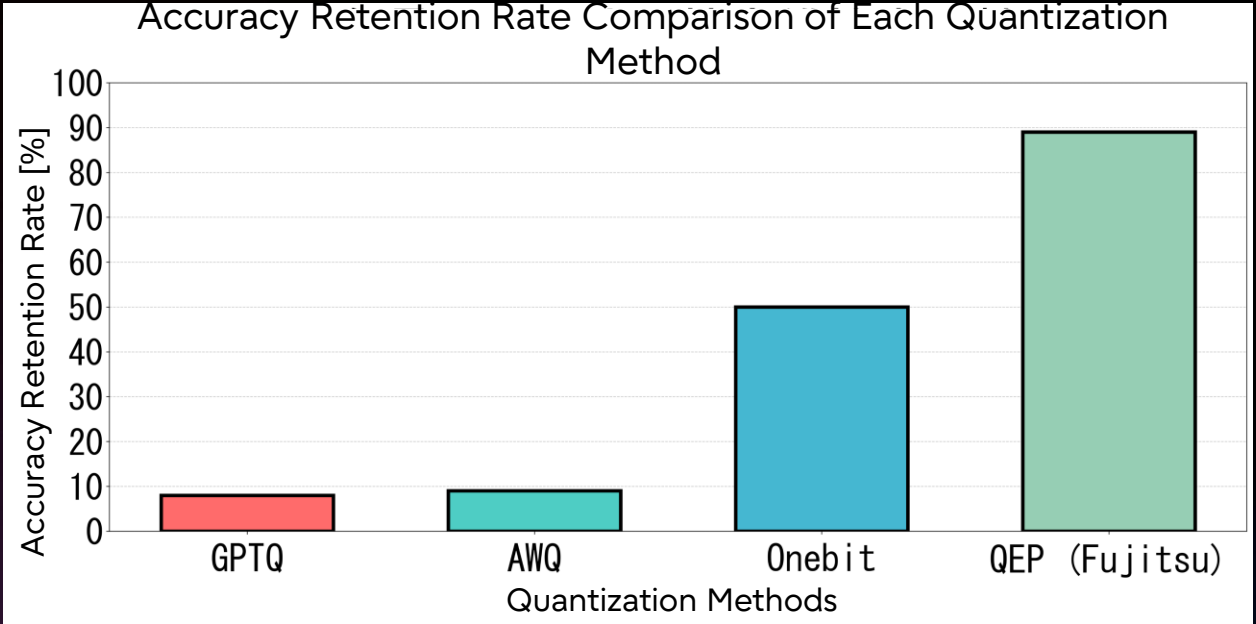
Use sample data to determine the stable position in advance.

QQA (Quasi-Quantum Annealing)

World-leading large-scale optimization algorithms developed through Digital Annealer

Achieved the world's highest precision retention rate of 89%

Accuracy Comparison with Conventional Technology



Achieved the world's highest accuracy retention rate of 89% with 1-bit quantization.
Memory usage reduced by up to 94%, enabling code generation via 32B LLM to run at the edge

We have released the core functionality as open-source software: OneCompression

FUJITSU

https://github.com/FujitsuResearch/OneCompression

FUJITSU

AI Computing Broker

Optimization for AI - Computing Workload Broker

FUJITSU

A platform combining HPC and quantum computing technologes

- Configures a quantum/HPC hybrid framework tailored to the operational characteristics of each computational type
- The framework predicts CPU/GPU/quantum behavior and optimally allocates resources based on each characteristics

Application

Finance | Materials | Energy | Drug discovery | Meteorology | Disaster prevention

Platform

AI | Simulation | Data Analytics

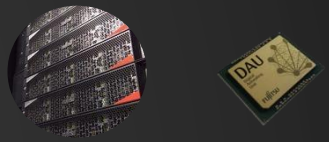
Computing Workload Broker (CWB)

Low power consumption | High computational performance | Reliable platform

Middleware
OS
Hardware

CPU
(MONAKA/-X)

GPU
NVIDIA, AMD



Quantum-inspired technology
Quantum simulator/Digital Annealer

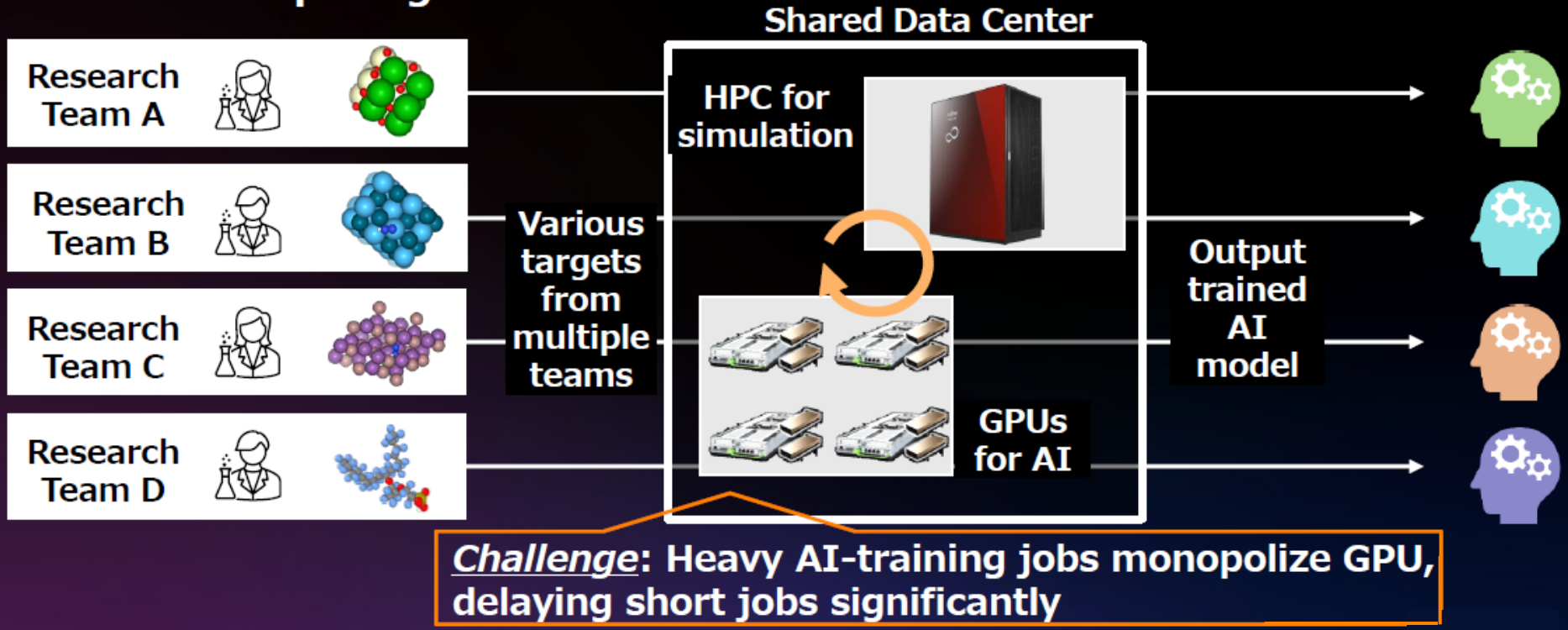
©RQC

Quantum computer
Superconducting/Diamond spin

Practical Use Case of NNP Generator

FUJITSU

Multiple AI training for each target are running simultaneously on limited computing resource

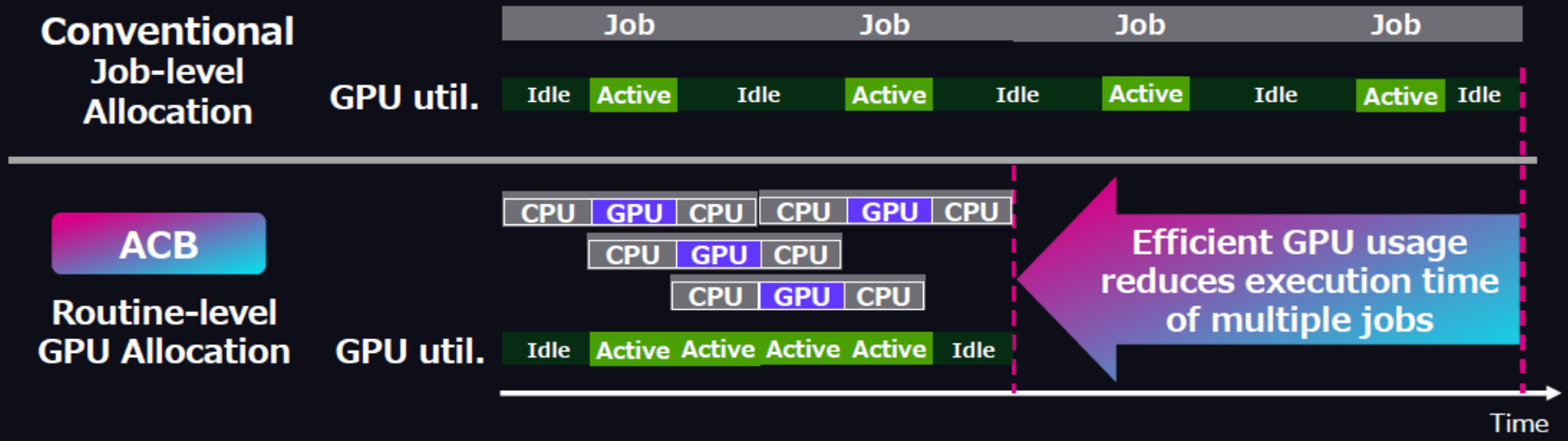


Feature of AI Computing Broker

FUJITSU

Best-in-class GPU utilization efficiency

- “Routine-level” allocation that detects actual GPU parts of jobs and dynamically allocates GPU accordingly



Feature of AI Computing Broker

FUJITSU

Enabling full GPU memory for each job

- Allocate GPU to only one job at a time (Temporal-sharing)
- Data of other jobs on GPU is automatically swapped to CPU

Conventional: Spatial-sharing

Memory is divided among jobs
Limited to small AI models

Memory

Job A | Job B | Job C

Computing Unit

Sharing among Job A/B/C

ACB Time-sharing

Memory is occupied by each job
Large AI models can run

swap memory

swap memory

Memory	Memory	Memory
Job A	Job B	Job C
Comp. Unit	Comp. Unit	Comp. Unit
Job A	Job B	Job C

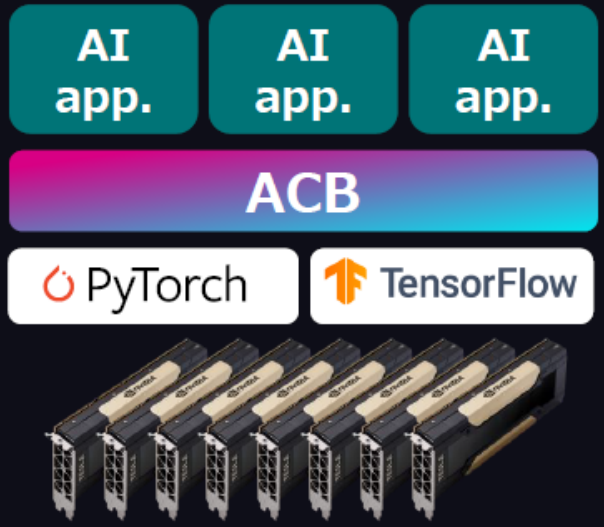
AI computing broker (ACB)

FUJITSU

A middleware to share GPUs among AI apps.

Key Features

- Best-in-class GPU utilization efficiency
- Optimized memory management across various AI applications



"Doubled model training throughput!"

"Deploying multiple jobs beyond physical GPU memory capacity!"

Success stories from ACB users
Scan here for more detail!

Key Takeaways

FUJITSU

- AI is evolving from static models to self-evolving systems
- Fujitsu's self-evolving multi-AI agent technology automates the full lifecycle: data generation, training, and optimization
- Deployable efficiency technologies (1-bit quantization, distillation, AI Computing Broker (ACB)) enable AI from HPC to edge environments
- Looking ahead, these techniques will be integrated with emerging NPUs through hardware–software co-design
- Extending specialized AI into resource-constrained and sovereign environments

FUJITSU

Thank you