Fujitsu AI-HPC Platform Powering AI for Science



© Fujitsu 2026

Future Supercomputing

FUJITSU

1. Infrastructure

AI Infrastructure at Scale

Growing compute demand meets real-world power and efficiency constraints

2. Computing

AI-HPC Workload

Simulation and AI increasingly operate as one continuous workflow

3. Science

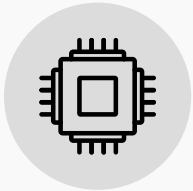
AI-Driven Scientific Discovery

AI-enhanced simulation, data-driven science, and agentic workflows

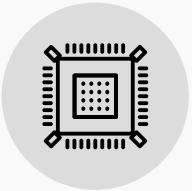
These shifts demand a new computing foundation.

Fujitsu AI-HPC Platform

- **Platform powering AI for Science**
- **Combining Fujitsu MONAKA-X CPU and GPU strengths for diverse workloads**
- **First deployment: FugakuNEXT**



MONAKA-X



GPU

© Fujitsu 2026

AI-HPC Workload Vision for the AI4S Era

AI4S = AI for Science

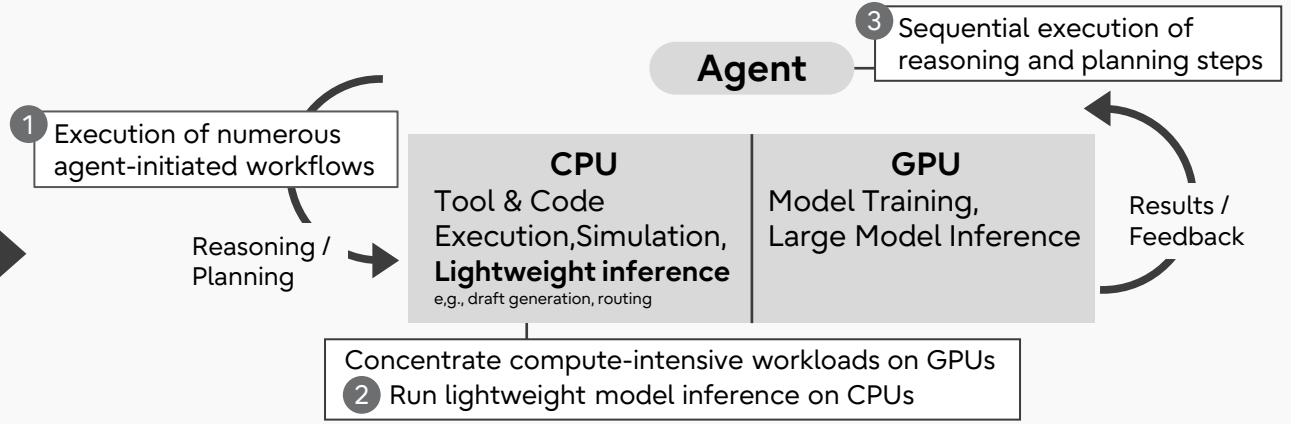
FUJITSU

As Is

User-interactive execution

CPU	GPU
Tool & Code Execution, Simulation	Model Training, Large Model Inference

To Be in the AI4S Era | An agent autonomously executes workflows in parallel and iteratively.



System Requirements for AI4S

1. Execution of numerous agent-initiated workflows → **1 Scalable parallel execution**
2. Run lightweight model inference on CPUs → **2 Low-latency Real-time inference**
3. Sequential execution of reasoning and planning → **3 Faster Tool and Generated-code execution**

To verify more hypotheses faster in the era of agentic AI,
we need systems that can quickly execute the processes above, including CPU-based inference

© Fujitsu 2026

Accelerating AI4S with Fujitsu's CPU-Optimized Software

FUJITSU

Fujitsu Software Enhancements

As AI4S evolves, CPUs play an increasingly central role in AI-HPC platforms. Our software optimizations unlock their potential.

1 Scalable Parallel Execution

Optimize task and memory placement for agents, tasks, and large-scale simulations.

CPU capability: >200 Cores

2 Low-latency Real-time inference

Optimize AI and math libraries and inference runtimes for master decisions and predictions.

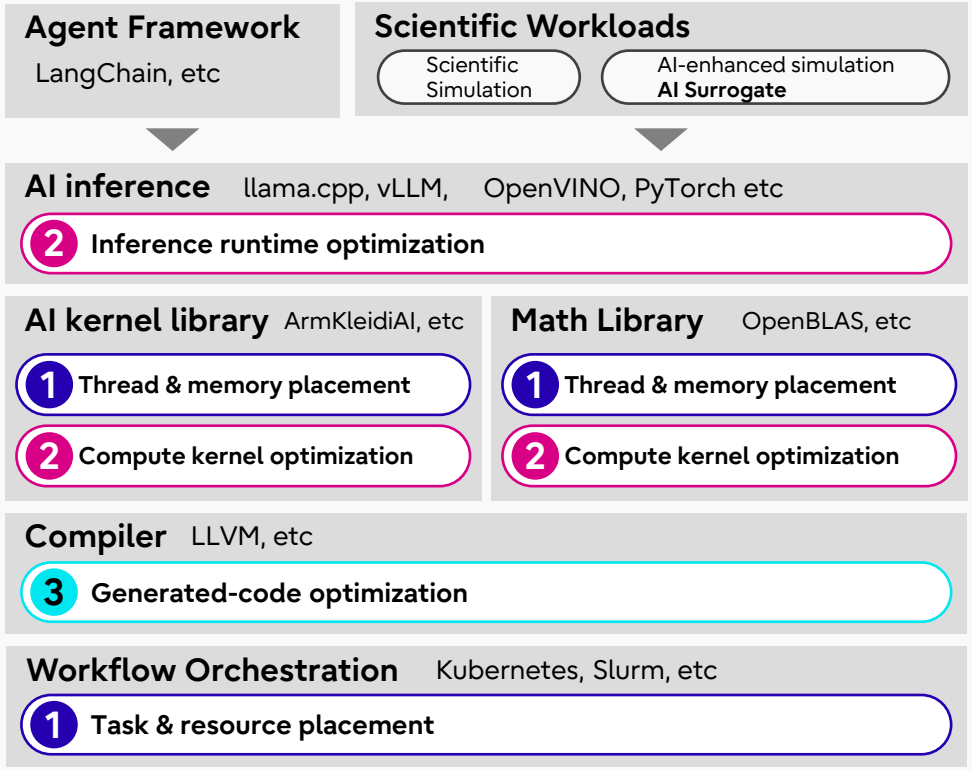
CPU capability: 512-bit SVE2 / 512-bit SME2

3 Faster Tool and Generated-code execution

Optimize compiled tools and agent-generated code for faster execution.

CPU capability : Strong single-thread performance

AI4S Software Architecture



© Fujitsu 2026

Fujitsu AI-HPC Platform

CPU-centric platform

TARGET

- Real-time AI, Agentic, AI for Science

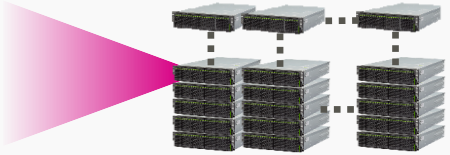
ARCHITECTURE

- MONAKA-X CPU standalone nodes
- Many high-capability CPU cores with SVE2/SME2
- Node-by-node flexible configuration

Compute Node



System



BENEFIT

- Real-time, Low-latency, energy-efficiency

CPU+GPU platform

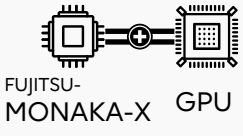
TARGET

- Large-scale AI, Agentic, AI for Science

ARCHITECTURE

- MONAKA-X CPU + GPU nodes
- NVLink Fusion for CPU and GPU tight coupling
- Rack-scale domain with a scale-up network

Compute Node



System



BENEFIT

- High performance in large-scale workloads

© Fujitsu 2026

Fujitsu Arm-based CPU, MONAKA-X

FUJITSU

Advancing MONAKA's performance, energy efficiency and trust to the next level with cutting-edge technologies

2027
FUJITSU-MONAKA

Data Center, AI, HPC



2nm

- Armv9-A Architecture
  144 Cores
  Arm SVE2
- 3D chiplet
  - Core die 2nm
  - SRAM die/IO die 5nm
- PCI Express 6.0
- Ultra low voltage for energy-efficiency
- Confidential Computing for security

2029
FUJITSU-MONAKA-X

Agentic AI,
Data Center, HPC



1.4nm

- Armv9-A Architecture
  > 200 Cores
  Arm SVE2+SME2
- 3D chiplet
  - Core die 1.4nm
  - SRAM die/IO die TBD
- PCI Express 7.0
- Enhanced Ultra low voltage
- Enhanced Confidential Computing

© Fujitsu 2026

MONAKA and MONAKA-X Features

FUJITSU

MONAKA-X brings more high-capability cores to HPC, AI, and agentic workloads

Workload	MONAKA Features	MONAKA-X Features
HPC Kernels & AI Data Processing Bandwidth-sensitive HPC AI	• 256-bit SVE2 • Strong FP64 performance • 12-ch DDR5-8800	• 512-bit SVE2 • Strong FP64 performance • Higher memory bandwidth
Sparse, Graph & Vector Search Memory-latency- and capacity-sensitive HPC AI	• 3D-stacked LLC • 12-ch DDR5-8800 • CXL 3.0 memory expansion	• Enhanced memory subsystem • Tight coupling with GPU
On-CPU Inference Inference-latency-sensitive AI Agentic	• 256-bit SVE2 • INT8DOT, INT8MM	• 512-bit SVE2 & 512-bit SME2 • INT8DOT, INT8MM • 2-/4-bit LUTI, 1-bit BNN
Orchestration & Agent Control Latency-sensitive Agentic	• 144 cores • Up to 3.8 GHz	• >200 cores • Higher frequency and IPC

SVE : Scalable Vector Extension SME : Scalable Matrix Extension

© Fujitsu 2026

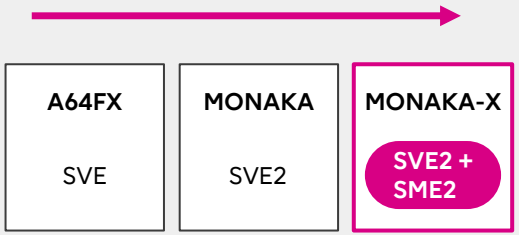
SME2 adoption in MONAKA-X

FUJITSU

1. Portability of SVE assets

Reuse designs and optimizations built on VLA (vector-length agnostic).

Preserving VLA



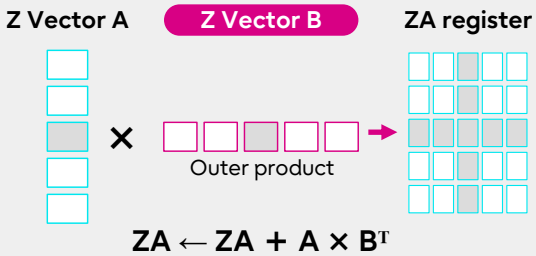
Reuse existing software

Lower migration costs

2. Matrix operations with ZA

Accumulate outer products in ZA with less data movement.

Outer product → ZA



Faster matrix operations

Higher computational density

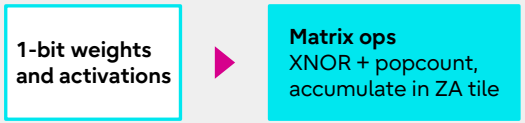
3. Low-precision, quantized AI

Provide two SME2 instruction types for sub-8-bit models.

2/4-bit lookup-table decoding



1-bit BNN acceleration



Faster low-precision inference

Support for quantized models

© Fujitsu 2026

First deployment: FugakuNEXT

FUJITSU

FugakuNEXT development Goals

FugakuNEXT, AI-HPC platform Ecosystem

Made with Japan

Co-developed by RIKEN, NVIDIA, Fujitsu

NVIDIA | R-CCS | FUJITSU

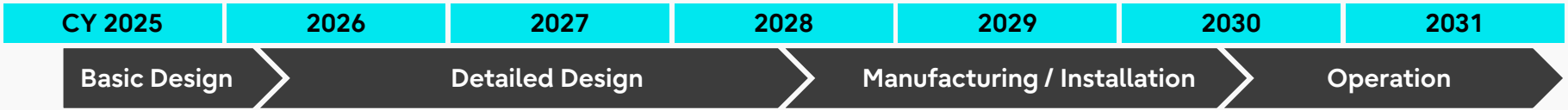
Technological Breakthroughs

- Tight CPU–GPU integration
- Modernization of applications

Sustainability and Continuity

- Sustainable software ecosystem
- Energy-efficient system

Schedule



© Fujitsu 2026

FUJITSU

Thank you

© Fujitsu 2026