Fujitsu AI-HPC Platform - Powering the Next Era of Simulation

Future Supercomputing

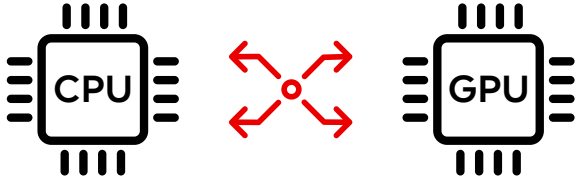
FUJITSU

Trends in HPC

- Ever-Growing data volumes and needs for computational power
- Needs for higher computational density and power efficiency
- Rapid AI evolution and expansion of applicable areas

AI-HPC platform powering the Next Era of Simulation

- Platform for fusion of AI and simulation
- CPU + GPU tightly coupled architecture
- Leverages both CPU and GPU strength
- Targeted around 2030



Revolutionizing Simulation with AI-HPC Platform

FUJITSU

AI-enhanced Simulation

Significantly improve accuracy and speed of simulation using AI

- **Faster Computation**
- **Higher Prediction Accuracy**

Simulated Data-Driven Science

Discover unknown behaviors hidden within the complexity of real-world physics

- **Massive Data Exploration**
- **Hidden Pattern Detection**

Agentic Simulation

AI agents autonomously plan, execute, and refine multi-step simulation loops

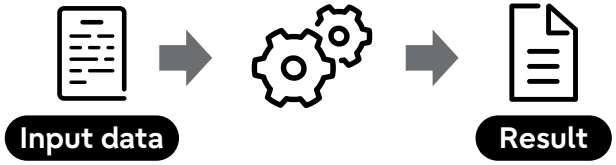
- **Autonomous Trial & Error**
- **Beyond Human Intuition**

AI-HPC Workload and CPU Requirements

FUJITSU

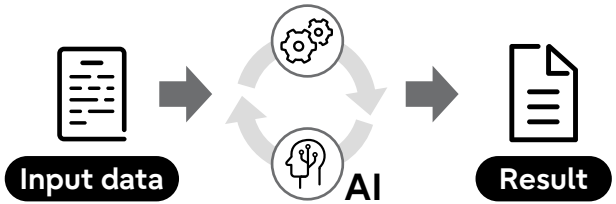
Conventional Workload

Numerical Computation



AI-HPC Workload

Numerical Computation



Demands on CPU

Fine-grained Simulation ⇄ AI Interaction
Simulation and AI exchange data at every step, creating a tightly coupled feedback loop

Agent Reasoning & Planning
Agent planning is a sequential decision chain — single-thread performance directly determines reasoning speed

Low-latency Real-time Response
Real-time inference in the control loop demands both low latency and high throughput

CPU Requirements

On-CPU AI Processing — SME2
SVE2 + SME2 and optional NPU enable AI inference directly on CPU — eliminating data transfer overhead

Tight GPU Coupling — NVLink Fusion
High-bandwidth, low-latency CPU–GPU interconnect for seamless data flow in the simulation–AI loop

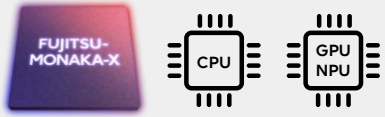
Many High-Capability Cores — >200 Cores
Massive core count with strong single-thread performance for orchestration, agent reasoning, and broad-spectrum simulation

FUJITSU-MONAKA-X Processor

FUJITSU

Arm-based Processor Enhanced for AI

Addressing CPU Requirements: Chip Evolution



FUJITSU-MONAKA-X

Scale and future Enhanced

- Armv9-A Architecture
- **>200 cores / socket**
- Ultra low voltage
- 3D chiplet
  - **Core die 1.4nm**
  - SRAM /IO die
- RAS/ Enhanced security CCA
- **Low latency cluster cache**

Use Case Enhanced

- SVE2
  - **512bit**
  - **SME2**
- **NVLink Fusion**
- **NPU** (AI)
- **Borderless Coherent UMArch**

Upscale Application thruput / performance

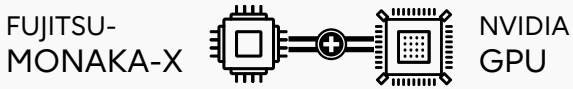
×

new

Upscale Application coverage
- Legacy HPC support
- Enhanced **GPU collaboration for** AI for science further

CPU-GPU Tight Integration technology

KEY Tech:CPU-GPU Tight Integration

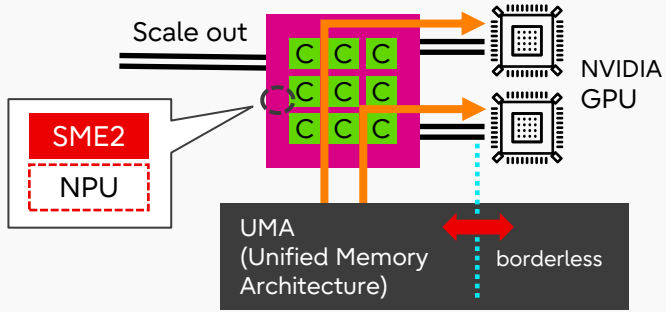


NVLink™ Fusion

CPU Strengths -Low Latency SME2
- Complex control flow, latency-sensitive tasks, and irregular memory access
  - Simulation, Realtime AI, Multimodal processing

GPU Strengths –High thruput NVLink
- Massively parallel data-parallel execution with regular memory access patterns
  - Large scale DL/ML training, Image/signal recognition

Bringing together the best of CPU-GPU



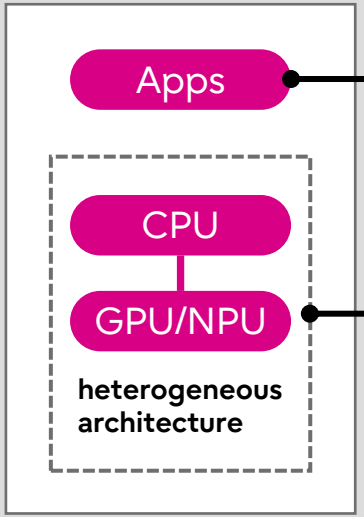
Advancing AI/HPC Through Software Evolution

FUJITSU

- Future AI/HPC platforms require both efficient hardware utilization and software-driven innovation

Co-Design–Driven Platform Optimization

Next-generation AI/HPC platforms are optimized through hardware–software co-design across system architectures and node configurations for diverse workloads and efficient execution



Accelerating Simulations with AI

Leveraging technologies such as hybrid surrogate modeling & simulation with accuracy-preserving quantization
Continuously advancing Scientific AI, Autonomous Scientific Discovery, and Autonomous Research Agents

Enhancing AI/HPC Software on FUJITSU-MONAKA-X

Continued OSS community collaboration for efficient utilization of FUJITSU-MONAKA-X CPU, GPU, and future heterogeneous accelerator architectures

1. SVE2/SME2 and microarchitecture-optimized HPC compilers and libraries
2. High-performance optimization for major AI libraries, including low-precision computing
3. AI/HPC acceleration across heterogeneous compute — CPUs, GPUs, and NPUs

FugakuNEXT – Japan's Flagship Platform

FUJITSU

FugakuNEXT development Goals

FugakuNEXT, AI-HPC platform Ecosystem

Made with Japan

Co-developed by RIKEN, NVIDIA, Fujitsu

NVIDIA | R-CCS | FUJITSU

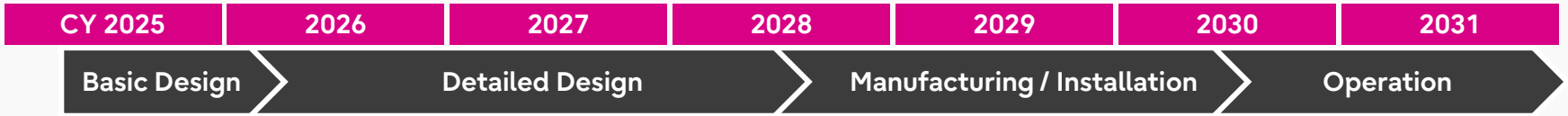
Technological Breakthroughs

- Tight CPU–GPU integration
- Modernization of applications

Sustainability and Continuity

- Sustainable software ecosystem
- Energy-efficient system

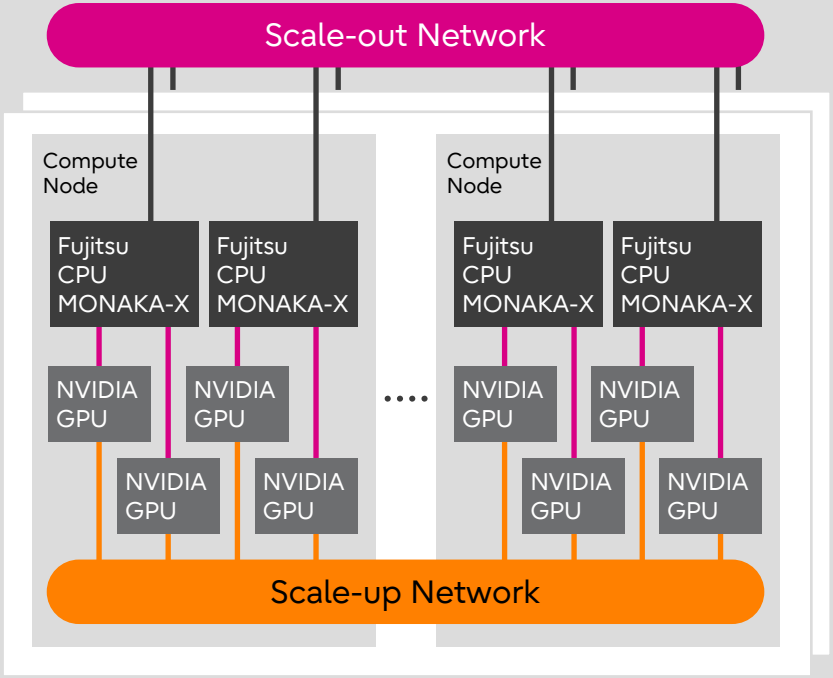
Schedule



The world's top-level AI-HPC platform

FUJITSU

- FugakuNEXT is not merely a larger supercomputer; it is a sovereign AI-HPC infrastructure designed to support both large-scale AI training and traditional scientific computing.



Specification	FugakuNEXT	
	CPU	GPU
Number of Nodes	≧3,400	
FP64 Vector perf.	≧48 PFLOPS	≧2.6 EFLOPS
FP16/BF16 Matrix perf.	≧1.5 EFLOPS	≧150 EFLOPS
FP8 Matrix perf.	≧3.0 EFLOPS	≧300 EFLOPS
FP8 Matrix sparsity perf.	-	≧ 600 EFLOPS
Memory Capacity	≧10 PiB	≧10 PiB
Memory BW	≧7 PB/s	≧800 PB/s

Thank you