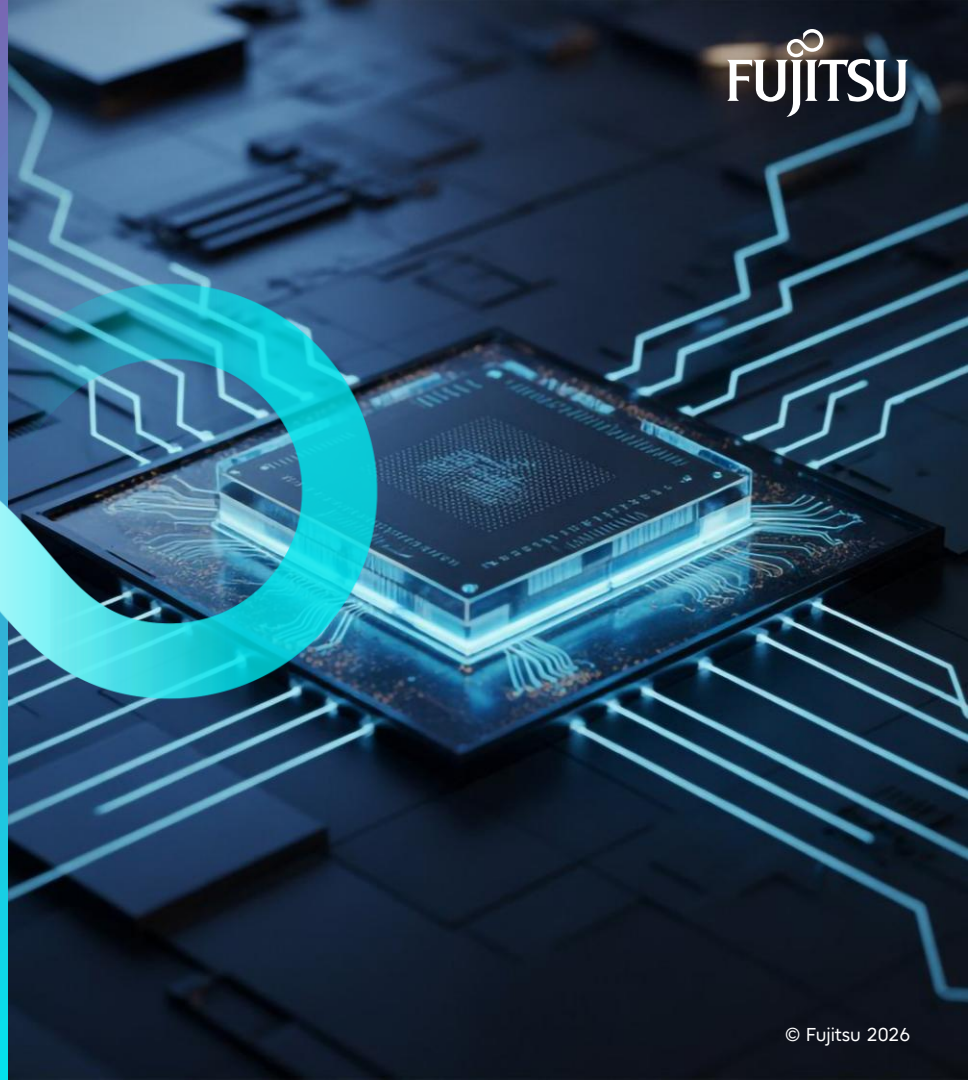
FUJITSU-MONAKA

High Performance and Efficiency for AI & HPC



* This presentation is based on results obtained from a project subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

© Fujitsu 2026

Customer Requirements for Next-Generation AI/HPC CPUs

FUJITSU

Customers increasingly evaluate AI/HPC CPUs based on real-world performance, infrastructure efficiency, and readiness & trust.

Key Industry Drivers

1 Power becomes the constraint 108 → 314GW

Worldwide data center power demand, 2025 → 2030*1

2 AI becomes a core workload 25% → 56%

Share taken by AI inference, 2025 → 2030*2

Structural shifts in AI/HPC infrastructure

Power and floor space constrain compute
- Data centers from hundreds of MW to GW scale
- Perf/Watt and Perf/Rack become evaluation criteria

HPC and AI workloads are converging
- Simulation plus surrogates and LLMs
- Real application and memory performance matter across diverse workloads

Expanding heterogeneous environments
- Diverse architectures increase software and deployment complexity
- Limits portability and infrastructure choice

Requirements of next-generation HPC infrastructure

Performance
- **Real application performance**
  OpenFOAM, LAMMPS, GROMACS
- **Memory performance and system balance**

Efficiency
- **Infrastructure efficiency**
  Perf/Watt, Perf/Rack, output per power budget
- **Deployment and operational efficiency**

Readiness & Trust
- **Software maturity and portability**

*1 Omdia, AI power capacity summary, Dec 2025 (worldwide DC power consumption).
*2 Omdia, Server Silicon Tracker 1Q2026 (GPU server units). Requirements based on discussions with US national laboratories, universities and system vendors.

© Fujitsu 2026

FUJITSU-MONAKA Processor Overview

FUJITSU



FUJITSU-MONAKA

Armv9-A | SVE2 256-bit
PCIe 6.0 / CXL 3.0
2 sockets / 288 cores per node
Core die 2 nm / SRAM & IO die 5 nm

Performance

144 cores
max per socket
(288 per node)

3.8 GHz
max frequency

8800 MT/s
max frequency

213.8 TOPS
INT8 AI inference,
2× throughput vs.
competitors

Efficiency

2×
power efficiency vs.
competitor CPUs

2nm
leading-edge process with
3D chiplet architecture

Ultra-low voltage
operation proven
in development

Liquid / Air-cooling
flexible deployment

Readiness & Trust

Confidential Computing
processor as the root of trust, protecting sensitive data from the hardware up

Mainframe RAS
reliability technologies for stable, trusted operation

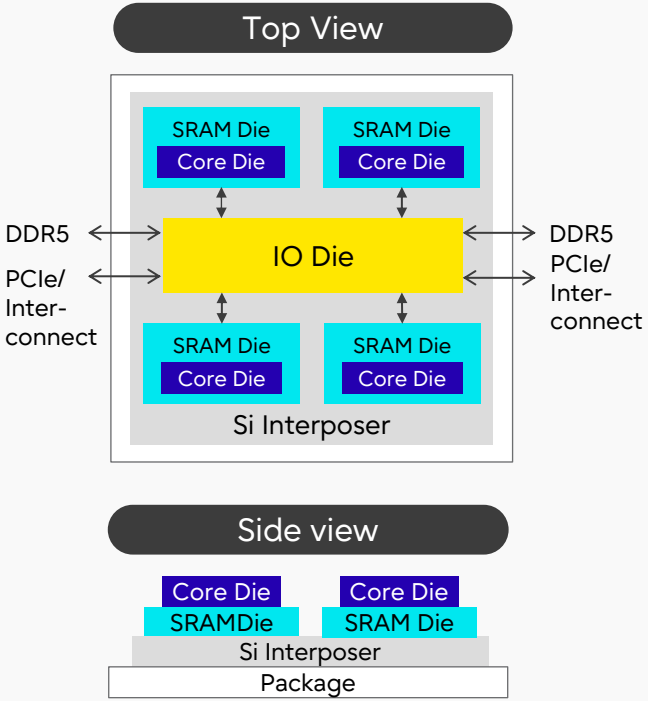
Armv9-A
Leveraging
the Arm software
ecosystem

© Fujitsu 2026

FUJITSU-MONAKA Technology Features

FUJITSU

High-Performance per Cost: Optimized Multi-Die Architecture



Core Die: *N2P process

- Die integrating multiple compute cores

* Further enhanced cutting-edge 2nm node

Heterogeneous process integration

- The costly 2 nm process is applied only to the core die
- Limiting the cutting-edge process area to less than 30%

SRAM+IO Die: 5nm process

- Die integrating last-level cache (LLC)

3D stacking

- Core die stacked on top of SRAM die
- Tightly coupled through Hybrid Bonding

➡ Enables an **optimal balance** between performance and cost efficiency.

© Fujitsu 2026

Unlocking FUJITSU-MONAKA's AI Inference Performance with Fujitsu's Hardware and Software Expertise

FUJITSU

LLM Inference Performance

gemma4-31B

55.2 Tokens/s

batch size 2

gpt-oss-20b

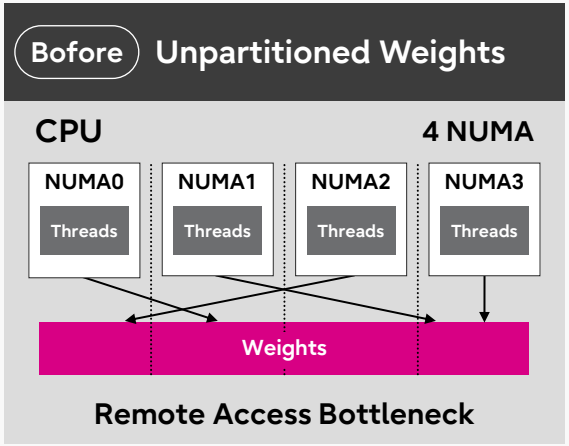
307.0 Tokens/s

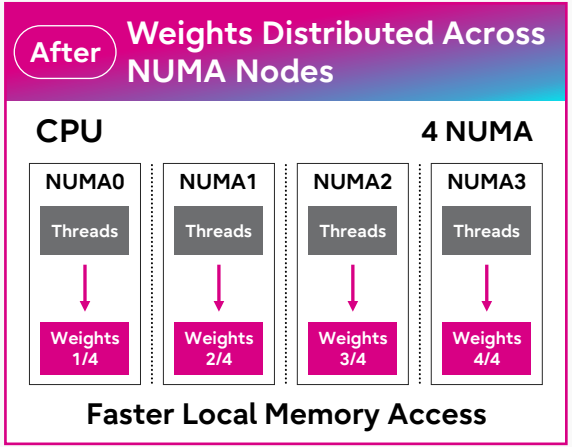
batch size 2

Tensor Parallelism Across NUMA Nodes

Distribute inference workloads across four NUMA nodes within a single CPU.

➡ Partition model weights to reduce remote memory access.





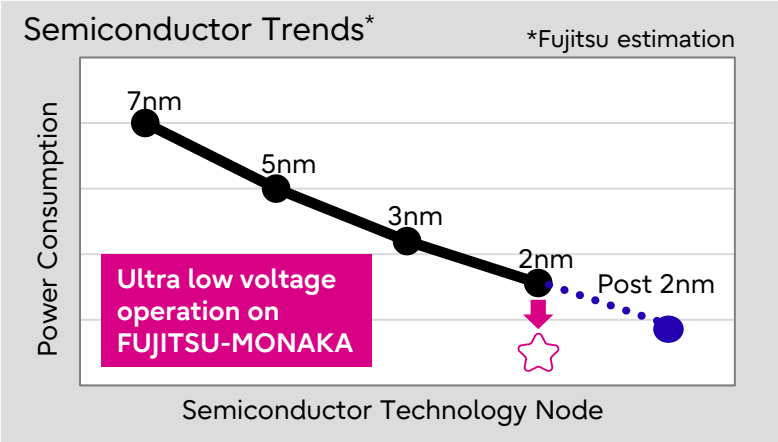
200+ Pull Requests Across 30+ Open-Source Communities

Fujitsu contributes optimizations from compilers to AI frameworks to unlock FUJITSU-MONAKA's performance.

© Fujitsu 2026

FUJITSU-MONAKA Technology Features

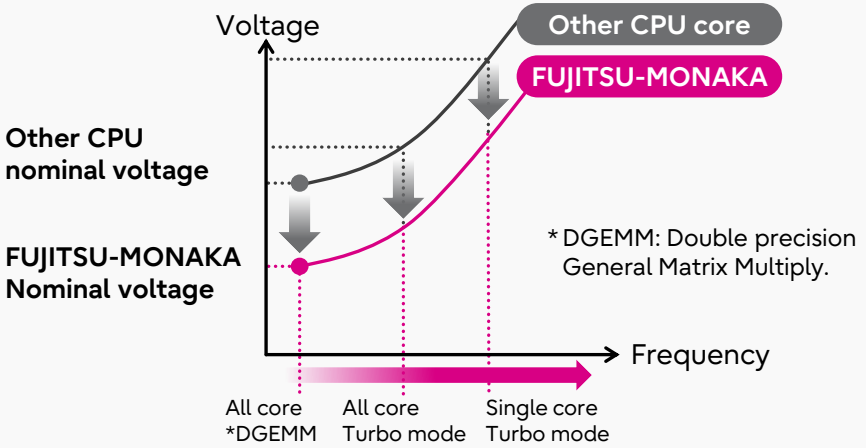
For High-Performance per Watt: Ultra Low Voltage Technology



Unique Ultra Low Voltage libraries for Core Die

- FUJITSU SRAM cells
- FUJITSU custom logic cells

➡ Achieves lower voltage operation **reduce core operating voltage.**



Advantage of Ultra Low Voltage design

- FUJITSU-MONAKA micro-architecture enables lower operating voltage for core at any operating frequency.
  (i.e. any workload)
- Since power ∝ voltage^2, FUJITSU-MONAKA core can drastically reduce power consumption, realizing **superior power efficiency.**

© Fujitsu 2026

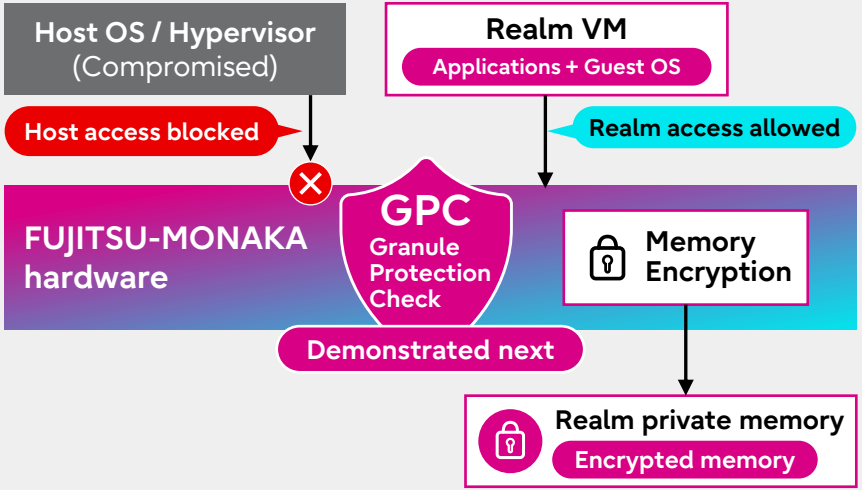
Arm CCA*1 with FUJITSU-MONAKA

FUJITSU

FUJITSU-MONAKA keeps AI models and business data confidential during processing on shared infrastructure

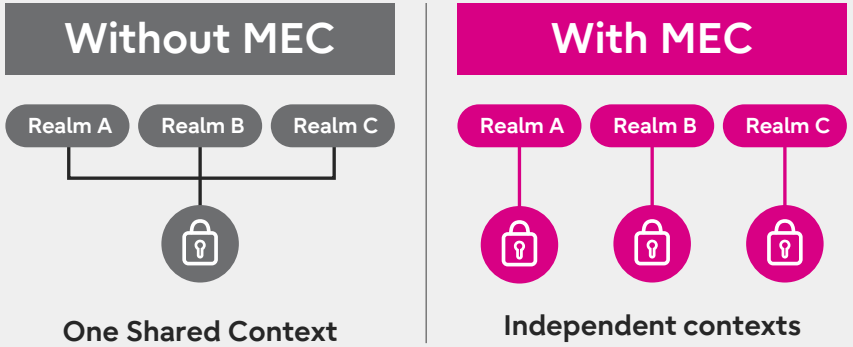
Protect In-use data
AI models
Business data

How Arm CCA protects workloads



FUJITSU-MONAKA: MEC*2 Support

Separate encryption contexts for individual Realm VMs



Up to 4,096 encryption contexts

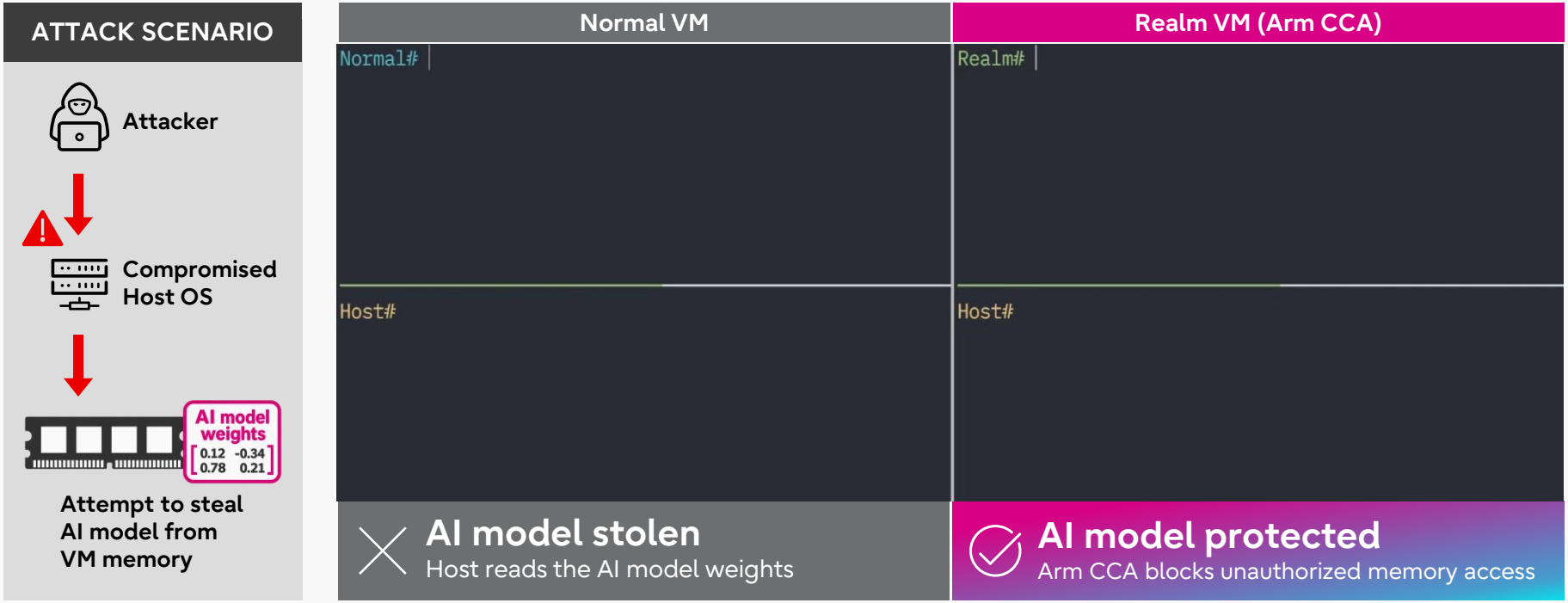
*1 CCA = Confidential Compute Architecture
*2 MEC = Memory Eancryption Context

© Fujitsu 2026

Arm CCA Demo: Preventing AI Model Theft

FUJITSU

Even if the host OS is compromised, **Arm CCA protects the AI model from unauthorized memory access.**



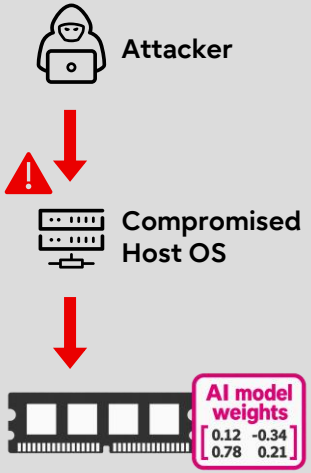
© Fujitsu 2026

Arm CCA Demo: Preventing AI Model Theft

FUJITSU

Even if the host OS is compromised, **Arm CCA protects the AI model from unauthorized memory access.**

ATTACK SCENARIO



Attempt to steal AI model from VM memory

Normal VM

```
Normal# ai-inference
Secret AI model:
    guest-transformer-v1
    weights=[-1.11,-0.48,-0.64,1.09]
Model address: 0x460e9000

Running inference...
```

```
Host# ./steal-guest-memory.sh 0x460e9000
AI model theft SUCCEEDED
Stolen model:
    guest-transformer-v1
    weights=[-1.11,-0.48,-0.64,1.09]
Host#
```

STOLEN!

AI model stolen
Host reads the AI model weights

Realm VM (Arm CCA)

```
Realm#
```

```
Host#
```

AI model protected
Arm CCA blocks unauthorized memory access

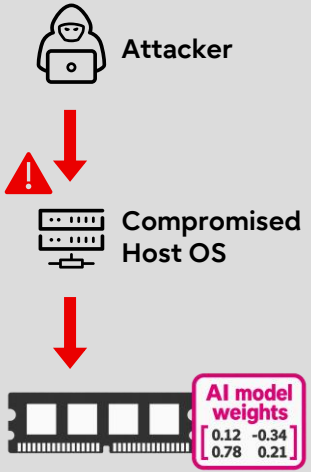
© Fujitsu 2026

Arm CCA Demo: Preventing AI Model Theft

FUJITSU

Even if the host OS is compromised, **Arm CCA protects the AI model from unauthorized memory access.**

ATTACK SCENARIO



Attempt to steal AI model from VM memory

Normal VM

```
Normal# ai-inference
Secret AI model:
    guest-transformer-v1
    weights=[-1.11,-0.48,-0.64,1.09]
Model address: 0x460e9000

Running inference...
```

```
Host# ./steal-guest-memory.sh 0x460e9000
AI model theft SUCCEEDED
Stolen model:
    guest-transformer-v1
    weights=[-1.11,-0.48,-0.64,1.09]
Host#
```

STOLEN!

AI model stolen
Host reads the AI model weights

Realm VM (Arm CCA)

```
Realm# ai-inference
Secret AI model:
    guest-transformer-v1
    weights=[2.31,-0.39,2.61,1.02]
Model address: 0x42a3c000

Running inference...
```

```
Host# ./steal-guest-memory.sh 0x42a3c000
AI model theft FAILED - blocked by GPC

Block details:
  Process: steal-guest-mem[6726]
  Fault:   Granule Protection Fault
  ESR:     0x0000000092000028
Host#
```

BLOCKED by GPC!

AI model protected
Arm CCA blocks unauthorized memory access

© Fujitsu 2026

Fujitsu MONAKA Server and Deployment Model

FUJITSU

Fujitsu MONAKA Server System: Evolving Together with AI/HPC Workloads

Three Server Models Designed to Meet AI/HPC Demands
-Designed and manufactured in Japan, enhancing traceability and supply chain transparency-

2U1N Model
for existing DC/Edge



1U1N Model
Highly Scalable



2U4N Model
High-Density
Multi-Node



Target Regions : Japan & Europe

Target Segments : Data Center Operators, Enterprises, Academia/HPC, and Defense

Server Performance: optimized for AI/HPC workload

Power Efficiency
2.26×* Perf/Watt

Performance Density
2.15×* Perf/Rack

Memory Performance
1,000 GB/s

Reliability
Proven RAS from mainframes

Security
Confidential Computing with Arm CCA

*gpt-oss-20b throughput: FUJITSU-MONAKA vs. x86. Calculated assuming the same chassis, a common rack power limit and linear CPU scaling. Subject to change.

Schedule for Availability

2026 End of August	PoC Servers Available for Loan	2026 Nov.	Fujitsu MONAKA Server Global Sales Launch	2027 April	Fujitsu MONAKA Server Shipments Begin

© Fujitsu 2026

Supermicro FlexTwin™ Server / Rack-Scale Solution with FUJITSU-MONAKA

FUJITSU | SUPERMICRO

Server Solution

FlexTwin™ 2U 4-node (ARS-223FT-HE-LCC/ALC)



Liquid-cooled multi-node server
with 8 FUJITSU-MONAKA processors (1,152 cores), up to 24 TB DDR5 8,800 MT/s memory, with direct-to-chip cold plates

PRIORITY USE CASES

- **Finance:** Enabling trusted AI and high-performance financial computing through Risk Modeling, Quantitative Analytics, and Fraud Detection.
- **Academic:** Advancing scientific discovery through Scientific Simulation, Molecular Modeling, Quantum Computing Research, and Spatial AI & Graph Analytics.
- **Manufacturing:** Accelerating engineering innovation through Computational Fluid Dynamics, Structural Analysis, and Materials Modeling.

CPU per Node

2 FUJITSU-MONAKA with 288 cores

Memory per Node

DIMM 24 Slots
Up to 6TB
DDR5-8800MT/s

PCIe Expansion per Node

1 PCIe 6.0 x16 FHHL slot,
1 PCIe 6.0 x16 HHHL slot,
1 PCIe 5.0 x16 AIOM slot
(OCP 3.0 compatible)

Rack Solution

FlexTwin™ 2U4N EIA Rack-Scale Solution



High-density, multi-node
rack-scale liquid cooled architecture

Rack Configuration

48U EIA rack
Up to 96 nodes per rack

CPUs / Cores per Rack

Up to 192 CPUs
27,648 cores

Cooling Solution

CPU / VRM / DIMM – direct-to-chip liquid cooling (server), In-row CDU, in-rack CDU or L2A sidecar

Power Budget per Rack

Max 153.6 kW

© Fujitsu 2026

AI Evolution and FUJITSU-MONAKA Technology Roadmap

FUJITSU

AI workloads evolve from HPC-centric to AI-native computing, and FUJITSU-MONAKA addresses this through multi-generation innovation

Development Roadmap

	2020	2027	2029	2031
	 HPC, AI	 Data Center, AI, HPC	 Agentic AI, Data Center HPC	
	Fujitsu A64FX 7nm core	FUJITSU-MONAKA 2nm core	FUJITSU-MONAKA-X 1.4nm core	FUJITSU-MONAKA-XX Further Advanced Node
Hardware	• Adopted in Fugaku (Top-tier supercomputer) • World's first implementation of Arm SVE	• High-speed data processing infrastructure • High performance × power efficiency	• CPU with integrated NPU • Selected for FugakuNEXT	• CPU-NPU fusion • Cutting-edge technologies
Compute Architecture	SVE	SVE2	SME2	
			NPU	
		PCIeGen6 GPU	NVLink Fusion	
Software & Ecosystem	Fugaku HPC software stack	FUJITSU-MONAKA software stack	• NPU software stack • AI/HPC optimized stack	

© Fujitsu 2026

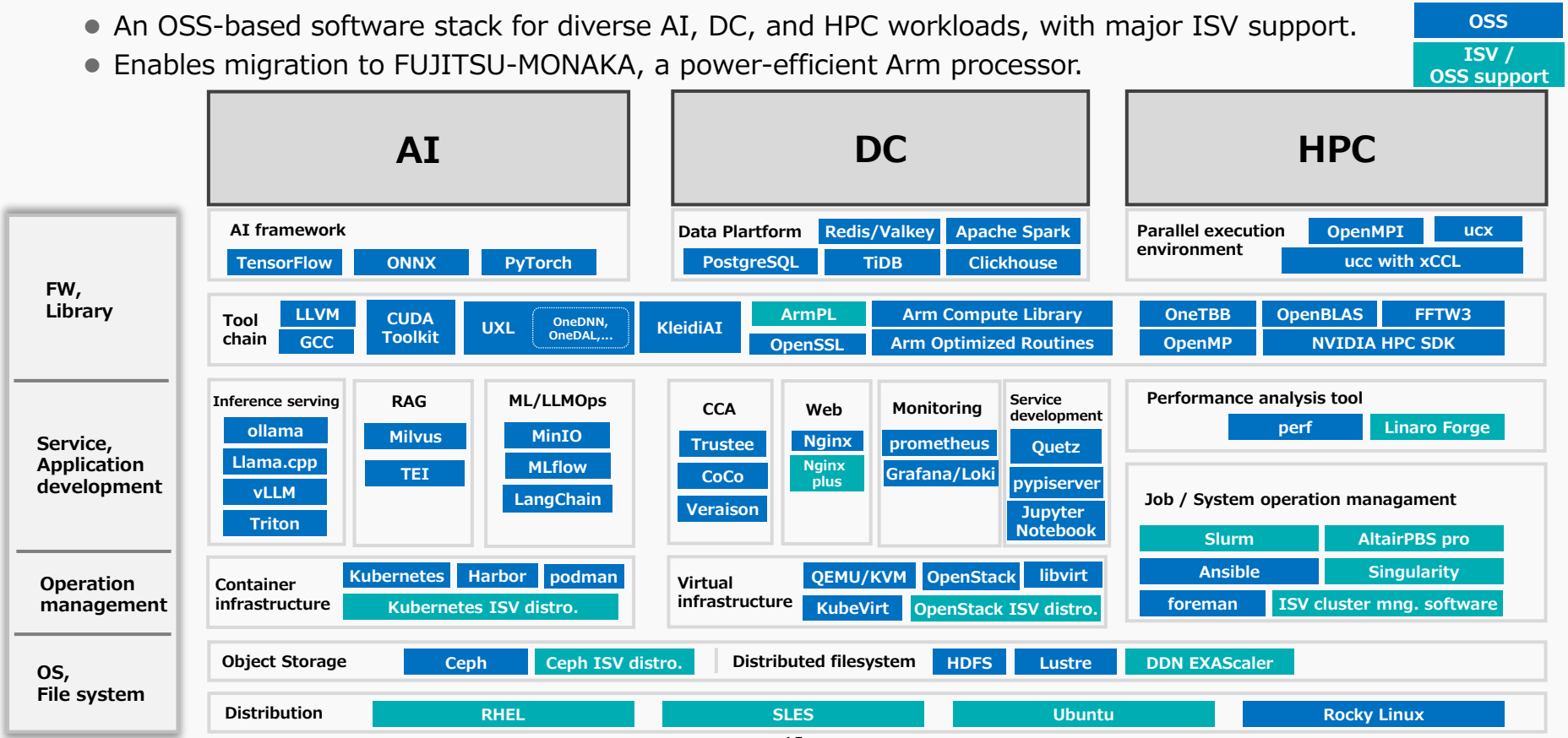
FUJITSU

Appendix

© Fujitsu 2026

Software Stack Optimized for FUJITSU-MONAKA

- An OSS-based software stack for diverse AI, DC, and HPC workloads, with major ISV support.
- Enables migration to FUJITSU-MONAKA, a power-efficient Arm processor.



© Fujitsu 2026

FUJITSU

Thank you

© Fujitsu 2026