

FUJITSU-MONAKA series: Arm-based processor



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

"FUJITSU-MONAKA"

Power-Efficient High-Performance CPU Development Journey

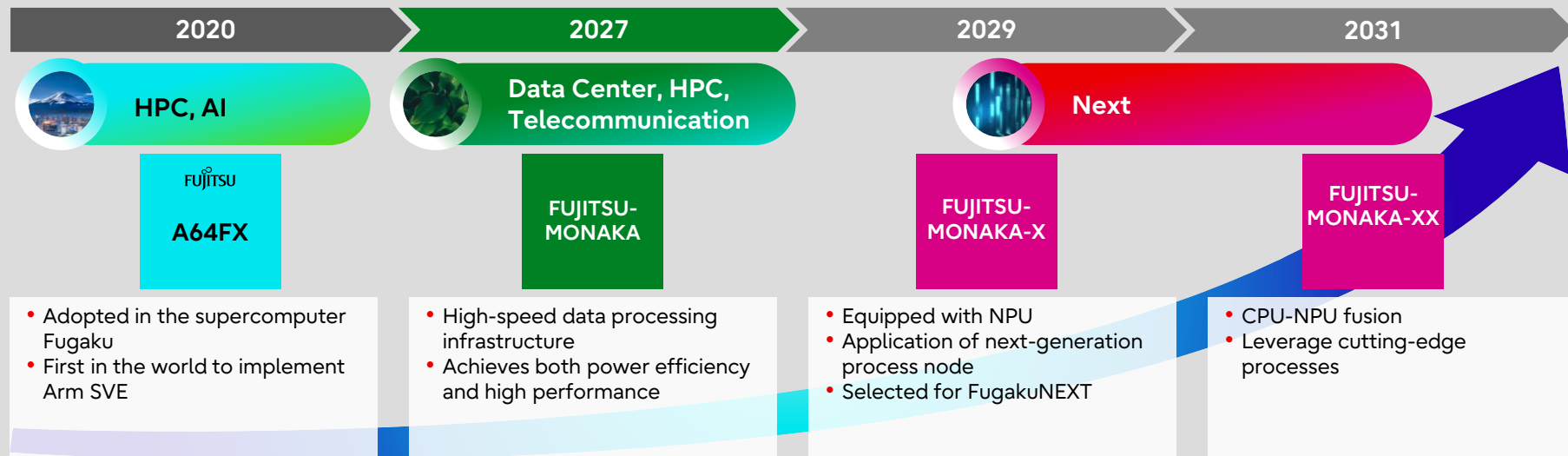


Launching in 2027, FUJITSU-MONAKA brings high performance, sustainable efficiency, and trusted security.

Development Background

- Creating a new era of computing power is mandatory for the future society with massive data generation and processing
- Ever-increasing power in datacenters is critical, and the power efficiency in CPU (consists of 60%) would be the vital factor for a sustainable future
- Fujitsu shall utilize its Supercomputer success and technology for the solution

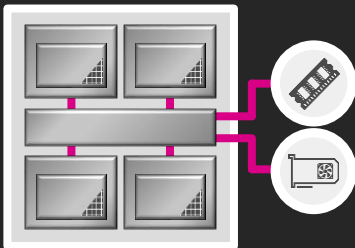
Development Roadmap



FUJITSU-MONAKA Processor Overview



FUJITSU-MONAKA



Arm v9-A Architecture



3D chiplet

- Core die 2nm
- SRAM die/IO die 5nm



Ultra low voltage for energy-efficiency



DDR5 12 channels



Liquid / Air-cooling

subject to change without notice



Arm SVE2-256bit for AI and HPC



144 cores x 2 sockets (288 cores per node)



Confidential Computing for security



PCI Express 6.0 (CXL3.0)

To be shipped in 2027

FUJITSU-MONAKA

High-Performance and Energy-Efficient CPU for a Carbon-Neutral Digital Society

High-Performance

Achieving high-speed computing centered on AI workloads (2x competitors CPUs).

Power-Efficiency

Reducing CO₂ emissions and electricity costs (2x competitor CPUs).

Safety & Security

Leveraging mainframe RAS technologies.

Ease of Use

Leveraging armv9 Software ecosystem.



3D Many Core architecture

Optimizing multi-die for heterogeneous 3D many-core architectures.

Core Die: 2nm process

- Die integrating multiple compute cores

Multi-core High performance Energy-efficient

SRAM Die: 5nm process

- Die integrating last-level cache (LLC)

Large capacity

3D stacking

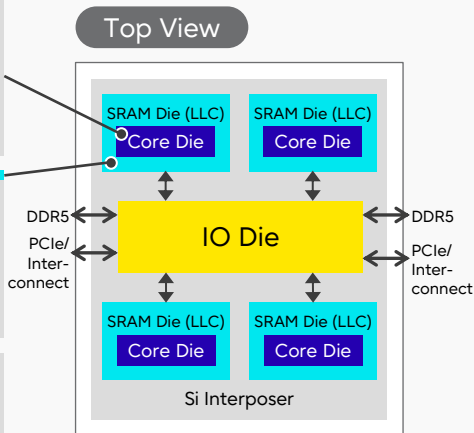
- Core die stacked on top of SRAM die
- Tightly coupled through TSV

Low latency High throughput

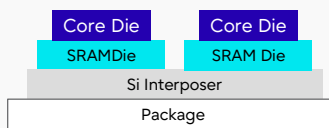
Heterogeneous process integration

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

Cost-effective



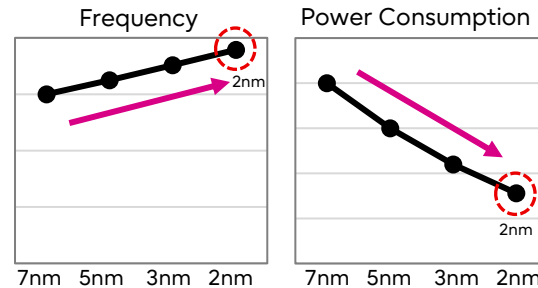
Side view



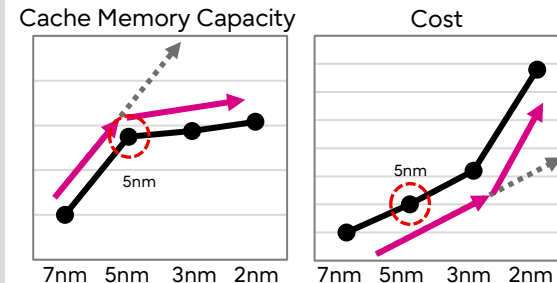
Semiconductor Trends*

*Fujitsu estimation

2nm(GAA) is better for Core Die



5nm(FinFET) is better for IO/SRAM Die



Balancing performance, power efficiency, and cost

Ultra Low Voltage Technology

Application of Ultra-Low Voltage SRAM for Energy-Efficient and Reliable Operation

Goal: Lowering the Operating Voltage of the Entire CPU

- Reducing operating voltage fundamentally cuts power consumption

Challenge: SRAM Instability Below Vendor-Specified Voltage

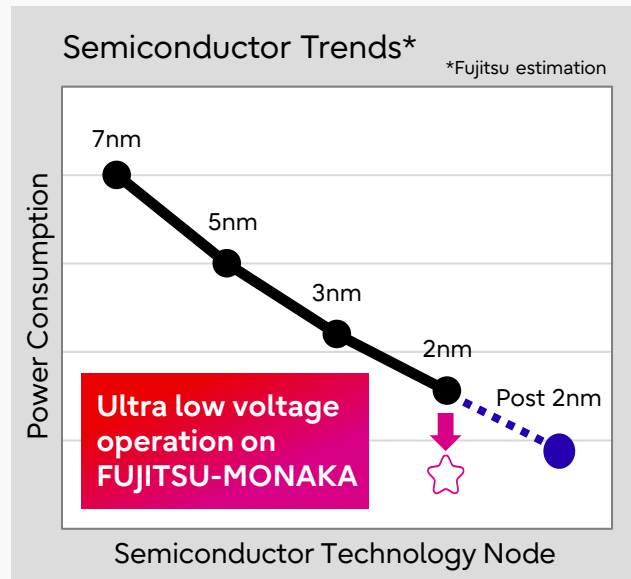
- Operating below vendor-specified voltage can cause malfunctions

Our Solution: Custom CAD Tools and Dedicated Circuits

- Designed ultra-low voltage SRAM integrated with a **single power supply**
- **Assist circuits** ensure reliable read/write operations at ultra-low voltages

 **Energy-efficient**

 **Stable operation**



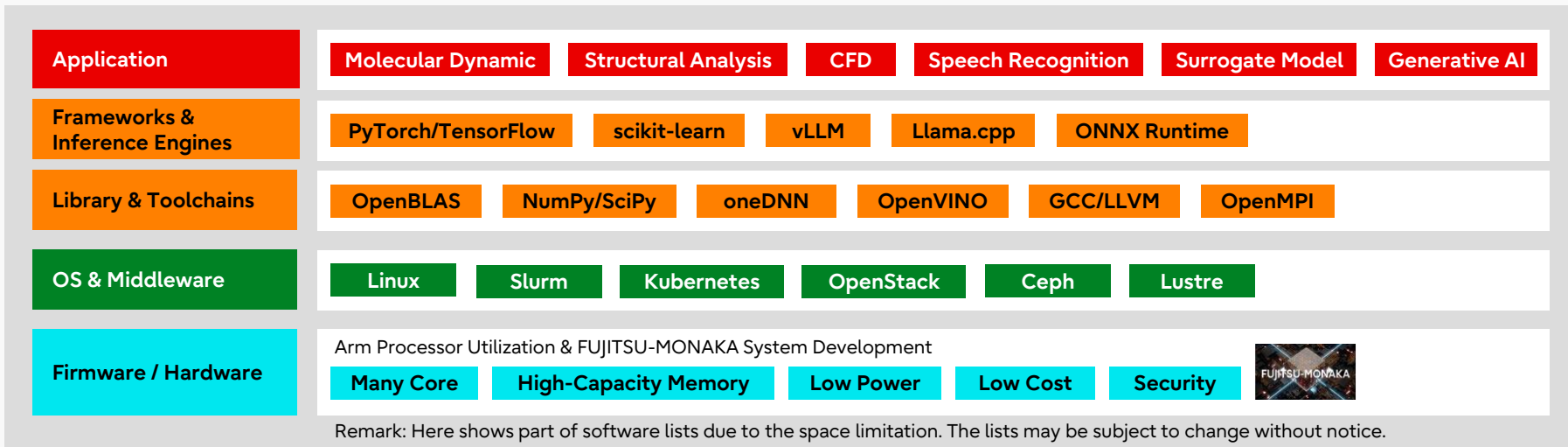
Achieved next-generation power efficiency beyond the 2nm process

FUJITSU-MONAKA Software Stack



Support for Standard OSS / ISVs per Domain

- Customers can adopt FUJITSU-MONAKA seamlessly, and enjoy its high performance & energy efficiency, reducing TCO.



Efforts to Expand AI & HPC Adoption in the Arm Ecosystem

- Driving AI & HPC performance improvement & quality enhancement of OSS by leveraging our HPC expertise
- R&D of Surrogate Models for Advanced Industrial AI

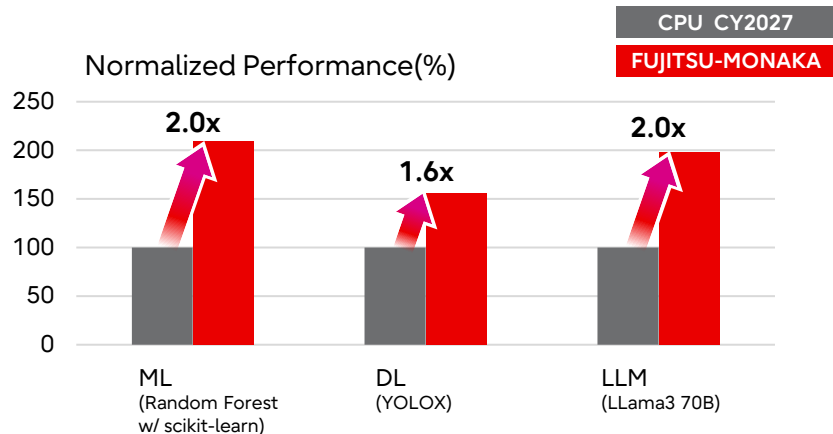
FUJITSU-MONAKA's Outperformance in AI & HPC



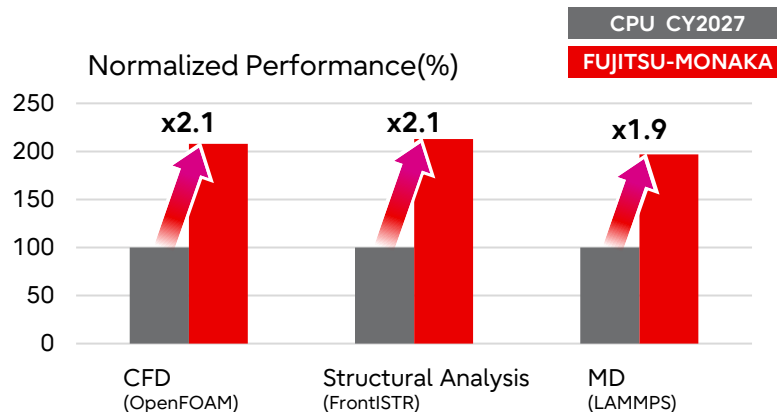
- **FUJITSU-MONAKA will deliver superior AI & HPC performance to competing CPUs in CY2027**
 - Its many-core architecture and Fujitsu's software optimization technique drive superior performance
- **Fujitsu is working on further software performance optimization to maximize FUJITSU-MONAKA's capabilities**



AI Workload



HPC Workload



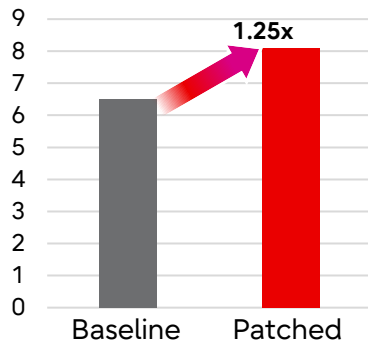
Remark: Graphs show estimated performance of softwares of each field (AI and HPC) based on an estimated performance of competing CPU @CY 2027 as 100%, subject to change without notice

Driving AI & HPC Performance on Arm Processors

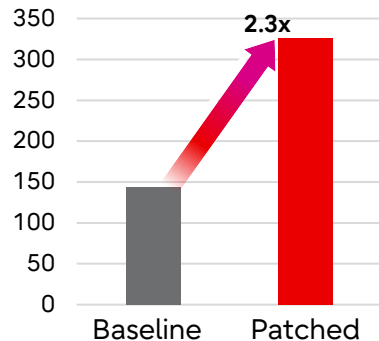
- Leveraging HPC expertise, Fujitsu leads performance optimization of AI & HPC OSS for Arm CPUs
- Our efforts uplift the entire Arm ecosystem and unleash FUJITSU-MONAKA's performance

AI Inference Software

Llama.cpp Performance Improvement (Single-batch TPS)



vLLM Performance Improvement (Multi-request TPS)



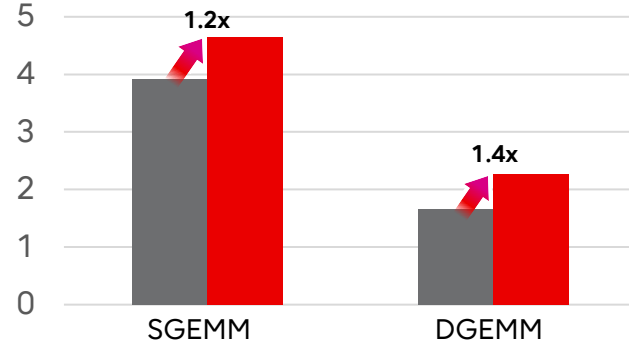
Remark: Graviton4 and Graviton3E are used for llama.cpp and vLLM evaluation, respectively.

Major Contributions to AI inference software speedup

- Llama.cpp: Introduce arm64 INT8 into GGML matrix multiplication kernel
- vLLM: Introduce SVE & threading and blocking logic into OpenVINO backend

OpenBLAS (SGEMM / DGEMM)

SGEMM / DGEMM Performance Improvement (TFLOPS)



Remark: Graviton3E is used for evaluation.

Major Contributions to OpenBLAS speedup from 0.3.25 to 0.3.30

- Optimize GEMM params considering micro architecture
- Optimize matrix block partitioning for multithread scalability

R&D of Surrogate Models for Advanced Industrial AI

Use Case: CAE Design with AI Surrogate Model

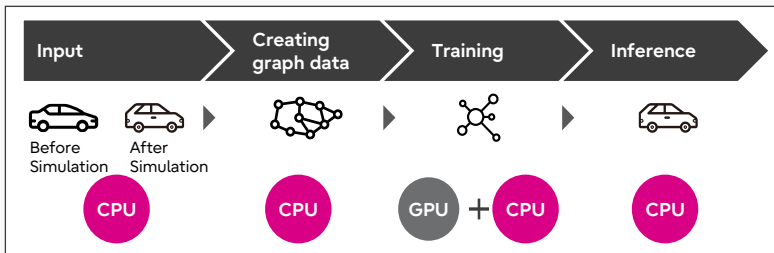
- Achieve cost-effective, high-accuracy CAE design
 - Rapid, low-cost design evaluation with surrogate models
 - Improve design accuracy in early product development stages

Key Challenge

- Enhance model accuracy & versatility
 - Realize various evaluations with fewer models
 - Reduce training frequency for truly cost-effective CAE design

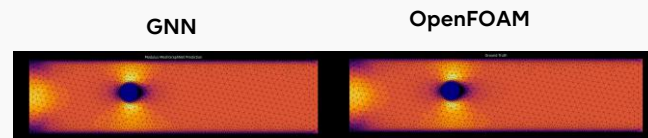
Our Solution

- Use GNN technology leveraging graph data for building accurate & versatile surrogate models

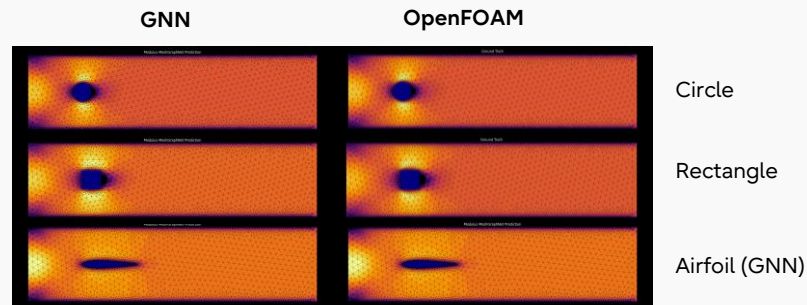


Initial Experimental Result (GNN vs OpenFOAM)

- Show high accuracy and versatility across varying object locations and shapes
 - Accuracy & Versatility for the location of objects



- Accuracy & Versatility for the difference in shape of objects



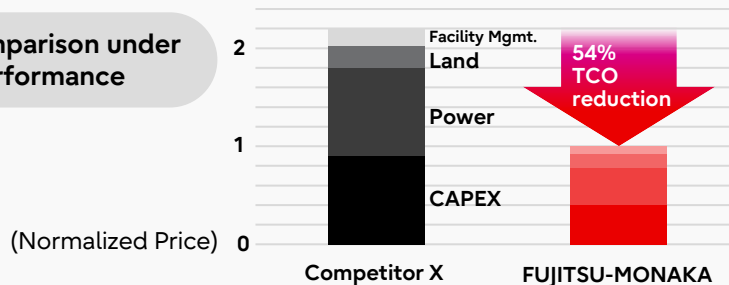
FUJITSU-MONAKA Servers



• TCO reduction in data centers

- The FUJITSU-MONAKA's high power efficiency reduces Total Cost of Ownership(TCO) in data centers while providing necessary computing power.

TCO comparison under same performance



• Versatile Solutions for Diverse Data Center Environments

- The FUJITSU-MONAKA server portfolio, offering both liquid and air cooling options, delivers optimized performance and scalability across a multitude of data center environments and use cases.

High Performance Computing

Core Data Center

High-performance, high-density liquid-cooling server

- High-density of 8CPUs per 2U in 19-inch rack
- High clock frequency to maximize FUJITSU-MONAKA performance
- Best fit to high density computing



Regional Data Center

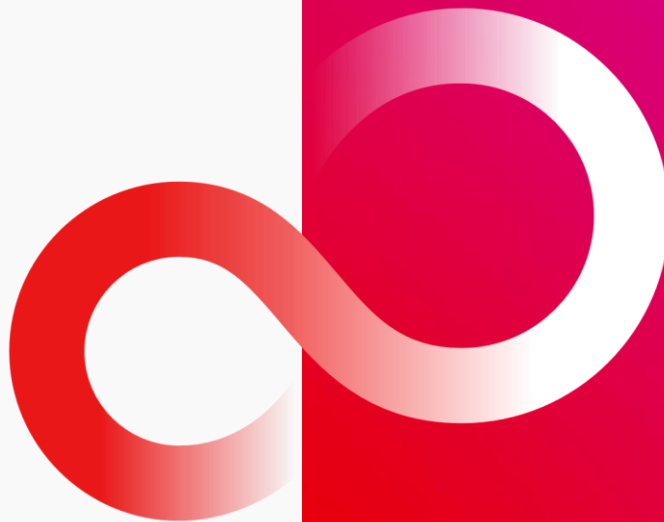
Edge Computing

Flexible air-cooling server

- Easy-to-install 2CPUs per 2U in 19-inch rack
- Many PCIe slots & drive bays for flexible configuration
- Best fit to existing or small data centers



Thank you



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