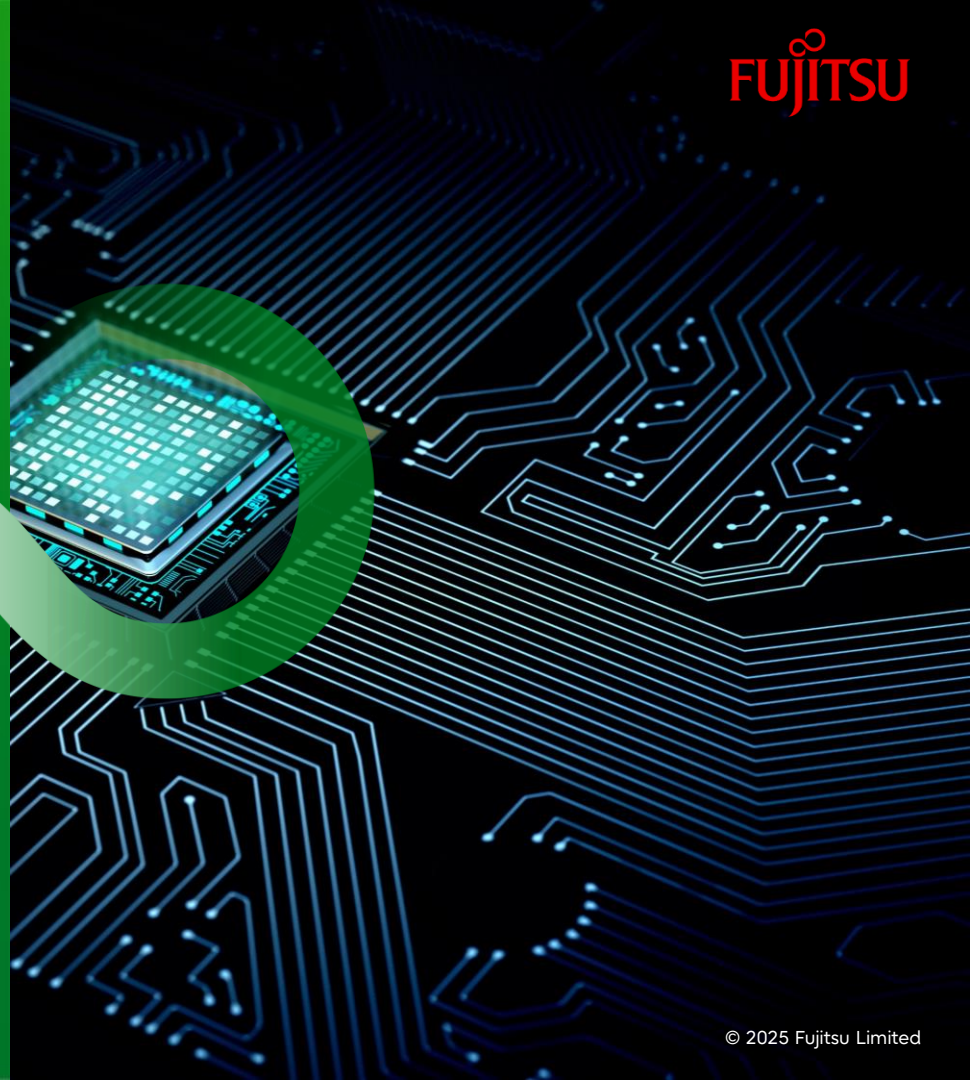
Next Generation Processor
FUJITSU-MONAKA



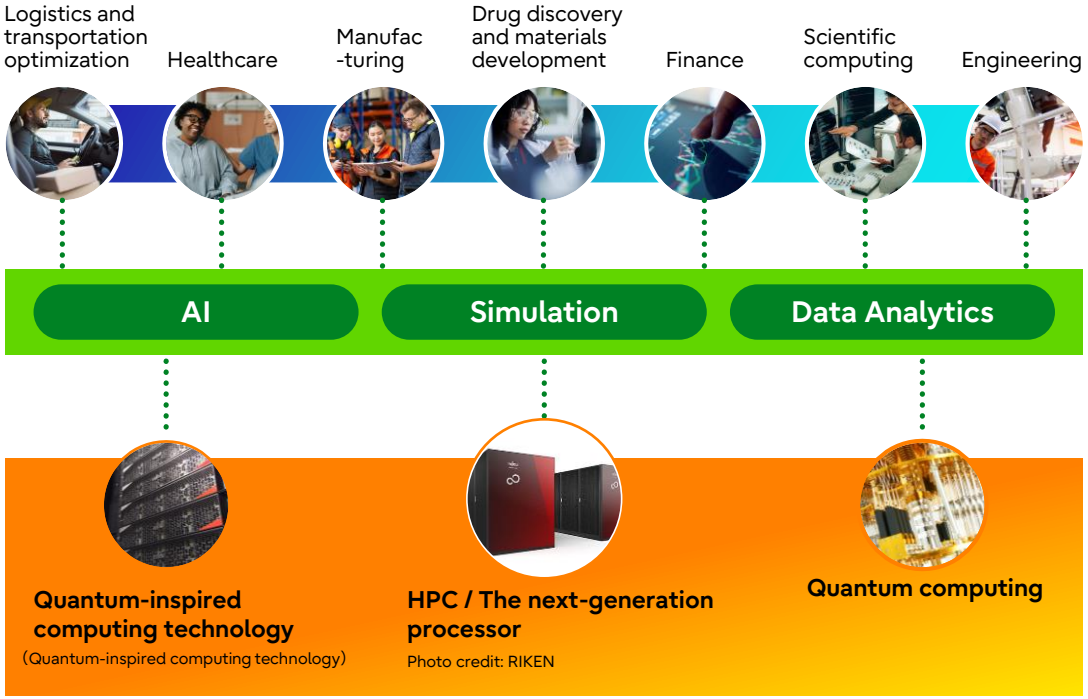
June 2025

Fujitsu Limited

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

Computing to Support the Various Needs of Enterprises

- Fujitsu develops computing solutions tailored to diverse enterprise needs, leveraging strengths in high performance, energy efficiency, and open architecture for AI, simulation, and data analytics.



Next-generation / high performance / energy efficient / AI platform

Powered by FUJITSU-MONAKA

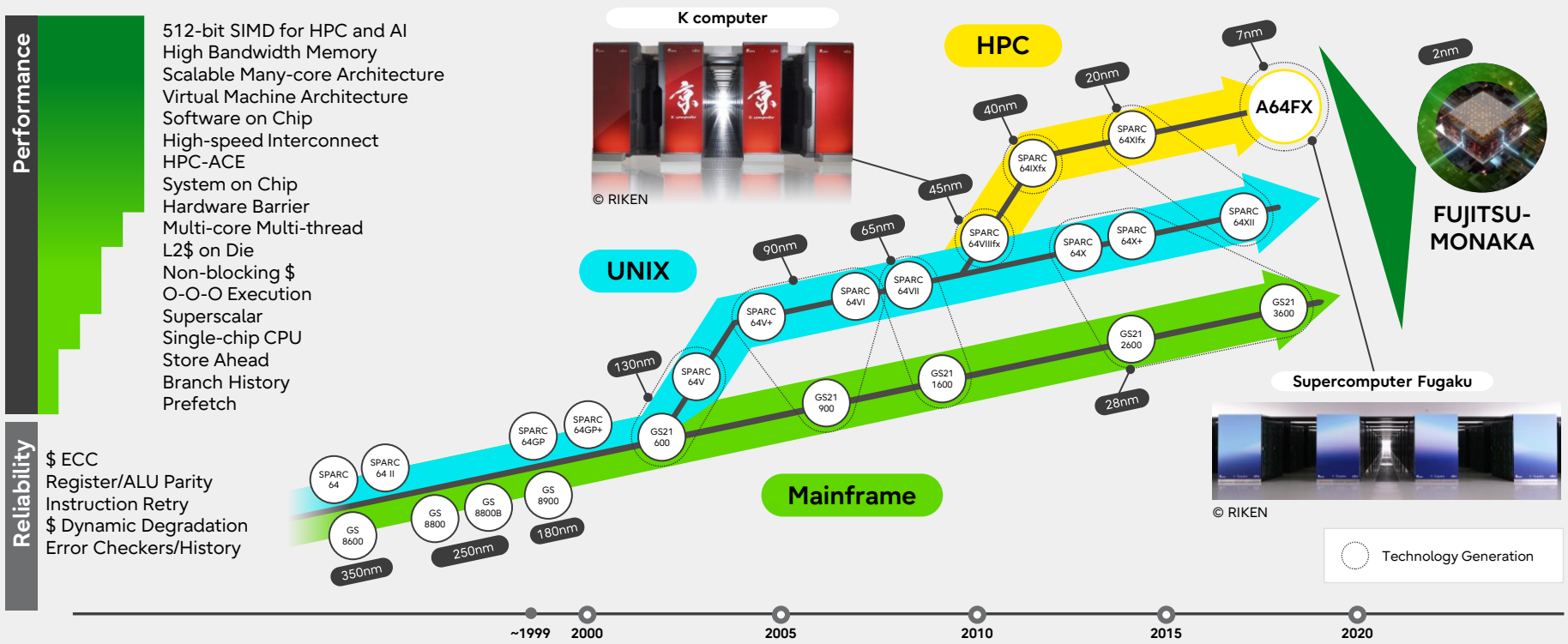
- High performance
- Energy efficient
- Expense controls
- Safe & secure
- Open architecture

Computing workload broker

Greater efficiency and optimization of AI computations through a combination of CPU, NPU, and GPU features

Fujitsu Processor Development

- Fujitsu has a 70-year history of processor innovation, culminating in the Arm-based FUJITSU-MONAKA, building upon the success of the K and Fugaku, both world-leading supercomputers.



“FUJITSU-MONAKA” - Fujitsu Arm-based Processor

- FUJITSU-MONAKA, set to be released in 2027, will deliver the performance, power efficiency, and security you’ll need for the future.

Fujitsu microarchitecture | **3D many-core architecture** | **Confidential Computing**

- **Fugaku technologies**

High-performance

- Cloud native 3D many-core design by Fujitsu's proven microarchitecture
- **High memory bandwidths**

Energy Efficient

- Leading-edge process technology
- **Ultra-low voltage operation**

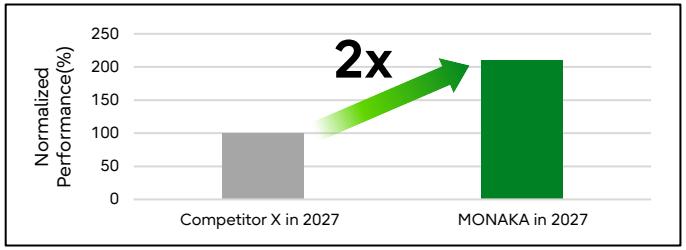
High Reliability

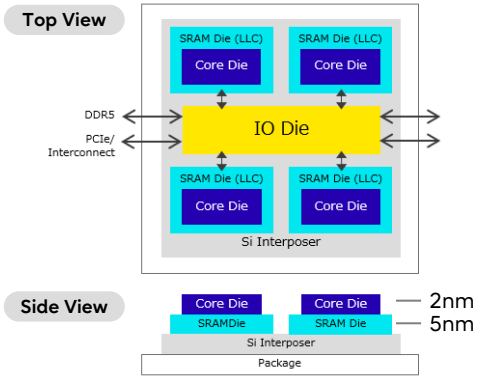
- Multiple VM Confidential Computing
- **Mainframe class RAS for stable operation**

Easy to Use

- **Open & de facto standard software stacks**
- **Fujitsu compiler technology**

Performance per Watt





subject to change

- Armv9-A architecture
  - SVE2 with 256-bit for AI and HPC
  - Confidential Computing for security
- 144 cores x 2 sockets
- Ultra low voltage
- 3D chiplet
  - Core die 2nm
  - SRAM die/IO die 5nm
- DDR5 12 channels
  PCI Express6.0(CXL3.0)
- Water cooling / Air cooling

Key Benefits and Values for AI, HPC

- FUJITSU-MONAKA delivers high-speed processing, energy efficiency, security, and ease of use — enabling reduced TCO, optimized HPC performance, and a robust architecture for AI inference.

Target Area **AI, HPC** *Data-intensive workloads* (Enterprise, R&D, Academia)

Key benefits of FUJITSU-MONAKA

High-speed data processing platform
Achieve high-speed processing of computing workloads, particularly AI workloads (2x faster than competing CPUs)

Balance of energy efficiency and performance
Significantly reduce CO2 emissions and power costs with high energy efficiency (2x more efficient than competing CPUs)

High security & reliability
Stable operation technology cultivated in mainframes and high security for cloud utilization

Easy to use
Utilize the Arm software ecosystem, and collaborative design across services, software, and hardware

FUJITSU-MONAKA's Value Proposition

- **TCO reduction**
- **Optimized for HPC workloads**
- **Suitable architecture for AI inference**

OSS Activities for FUJITSU-MONAKA

- Fujitsu actively contributes to OSS communities, enhancing performance and quality across diverse software stacks and fostering collaborative innovation

AI Frameworks/Libraries

Submitted 100> PRs to 15> OSS communities,
contributing to performance improvements for FUJITSU-MONAKA

ML **DL**

OpenBLAS **PyTorch** **oneDNN**

By introducing SVE, Int8, thread throttling and others,
Improved MM/BRGEMM/Tanh kernel performances by **1.2-3x**

LLM

vLLM **llama.cpp** **OpenVINO**

By supporting quantization and enhancing framework interoperability,
Improved inference performance by **2-3x**

DP

Milvus **PostgreSQL**

By using SVE and OpenMP,
Improved similarity search by **1.3x** and encoding/decoding by **17x**

System Software

Collaboration with Arm, Red Hat, SUSE

Linux **PAPI/libpfm**

Working on community development for FUJITSU-MONAKA and ARM v9.
Patch merged: 6, Review&Test: 60>

GCC **LLVM**

Working on performance improvements(SVE, instruction scheduling),
Fortran quality enhancements. **Submitted bug report: 280>, Patch merged: 10>**

Confidenial Computing (ARM CCA)

OpenStack **CoCo**

Started **Arm Confidential VM** and **Arm Confidential Containers** developments.
Working on design/implementation/reviews in the community

Libvirt

Developed **libvirt CCA support** patches and started discussion in the community

Collaboration with Linaro

Veraison

Tested **Remote Attestation** patches and provided feedback to the community

Activities around OSS compilers

FUJITSU

- **Enhancing GCC/LLVM compilers for FUJITSU-MONAKA**

Towards Faster and More Reliable Execution of C/C++/Fortran Applications on FUJITSU-MONAKA

- **GCC/LLVM support for FUJITSU-MONAKA (-mcpu=fujitsu-monaka)**
  - GCC: 15.1 (released in Apr. 2025)
  - LLVM: 20.1 (released in Mar. 2025)

  supported feature: ARMv9.3+α
  with 256-bit SVE vector

- **Flang quality enhancement (Flang: LLVM's Fortran frontend)**

Bug reporting (total 269 reports)

- ~2025: Fortran77/90/95
- 2025~: Fortran2003/2008

Building Flang CI

- Using Fujitsu Compiler Test Suite
- 49 regressions were detected

- **Ongoing performance improvement for FUJITSU-MONAKA**
  - **Flang run-time functions**
  - **Loop optimization in LLVM backend**

AI x HPC : CAE Design with AI Surrogate Model

- **FUJITSU-MONAKA targets AI × HPC workloads, such as the use of AI surrogate models to accelerate real-world simulations**
- **The effectiveness of these models is currently being validated through customer co-creation. Advanced GNN techniques and hybrid CPU/GPU usage are also being explored to meet future AI demands**

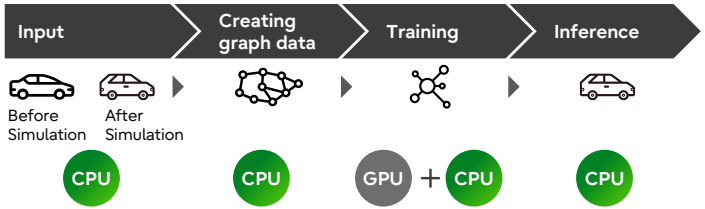
Use Case Overview

Problem

- In product development, such as in the automobile industry, complex structural simulations are performed under various material properties and conditions. **There is a need for faster evaluation in early design stages and improved design accuracy**

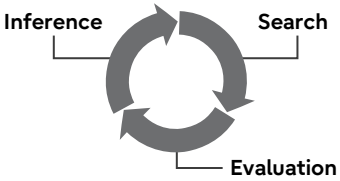
Solution

- Build a surrogate model with comparable accuracy using GNN technology leveraging graph data (hundreds of thousands of nodes).



Fujitsu's Unique Technology: Performance, Accuracy, and Usability Improvements

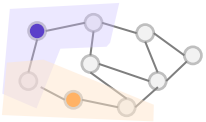
Optimized Automatic Parameter Search



Search for parameters with high accuracy while predicting AI accuracy

Applying High-Precision AI Models to Large Graph Data

- Create multiple graph data points by extracting portions of the original graph
- Perform this operation for every node to create the dataset



an example of multiple graph with neighborhood 1

GOAL Establish real-world applications and use cases for AI surrogate models using CPU/GPU for FUJITSU-MONAKA

FUJITSU-MONAKA Servers

- **The FUJITSU-MONAKA server portfolio, offering both water and air cooling options, delivers optimized performance and scalability for diverse workload requirements.**

High-performance, high-density server

- Water cooling increases clock frequency to maximize FUJITSU-MONAKA performance
- High density implementation of 8 CPUs per 2U of 19inch rack
- Arm SystemReady SR* standard for Arm servers
- Uses InfiniBand/Ethernet for node-to-node interconnects

Specification Overview

Subject to change

Processor	FUJITSU-MONAKA CPU 144cores
Core / node	288 cores (144cores x 2 sockets)
PCIe Slot / node	Low Profile x 2 (PCIe6.0/CXL3.0 x 16 lane) OCP3.0 NIC x 1 (PCIe6.0/CXL3.0 x 16 lane)
Memory / node	RDIMM (DDR5) x 24 slots, up to 6TB (21GB/core)
Chassis Config.	4 Dual CPU nodes (8CPUs / chassis)
Chassis Size	19inch(W) x TBD mm(D) x 2U(H)
Cooling	CPU : Water cooling, Others : T.B.D.

Scalable air-cooling server

- Easy-to-install air cooling system
- Many PCIe slots and drive bays for future system expansion
- Arm SystemReady SR* standard for Arm servers
- Uses InfiniBand/Ethernet for node-to-node interconnects

* Arm SystemReady SR standardizes the interface between base software such as operating systems, operational management software, and hardware, making system construction easier.

Specification Overview

Subject to change

Processor	FUJITSU-MONAKA CPU 144cores	
Core / node	288 cores (144cores x 2 sockets)	
PCIe Slot / node	Low Profile x 5 (PCIe6.0/CXL3.0 x 16 lane) * Up to two full size GPGPU cards(Optional)	
Memory / node	RDIMM (DDR5) x 24 slots, up to 6TB (21GB/core)	
Storage drive bays	2.5“ SAS SSD x 24, up to 368.64TB	EDSFF E3.S SSD x 4, up to 61.44TB
Chassis Config.	1 Dual CPU nodes (2CPUs / chassis)	
Chassis Size	19inch(W) x 873.1 mm(D) x 2U(H)	
Cooling	Air cooling	

Potential Collaboration with NVIDIA

- **NVIDIA announced their next-generation custom AI computing infrastructure.**
- **Leverage the power of FUJITSU-MONAKA, seamlessly connecting to NVIDIA's expansive ecosystem to drive AI supercomputing advancements.**

FUJITSU-MONAKA Commentary in NVIDIA's NVLink Fusion Press Release

Vivek Mahajan
Corporate Vice President, CTO, System Platform

"Combining Fujitsu's advanced CPU technology with NVIDIA's full-stack AI infrastructure delivers new levels of performance,"

"Fujitsu's next-generation processor, FUJITSU-MONAKA, is a 2-nanometer, Arm-based CPU aiming to achieve extreme power efficiency. Directly connecting our technologies to NVIDIA's architecture marks a monumental step forward in our vision to drive the evolution of AI through world-leading computing technology — paving the way for a new class of scalable, sovereign and sustainable AI systems."

Citation: NVIDIA Corporation - NVIDIA Unveils NVLink Fusion for Industry to Build Semi-Custom AI Infrastructure With NVIDIA Partner Ecosystem. (2025, May 18). https://investor.nvidia.com/news/press-release-details/2025/NVIDIA-Unveils-NVLink-Fusion-for-Industry-to-Build-Semi-Custom-AI-Infrastructure-With-NVIDIA-Partner-Ecosystem/default.aspx

Thank you

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