Next Arm-based Processor FUJITSU-MONAKA and Its Software Ecosystem

Fujitsu Limited

FUJITSU

Next Gen. Arm-based Processor "FUJITSU-MONAKA"

FUJITSU

FUJITSU-MONAKA is the processor developed for HPC, AI, and Data Center

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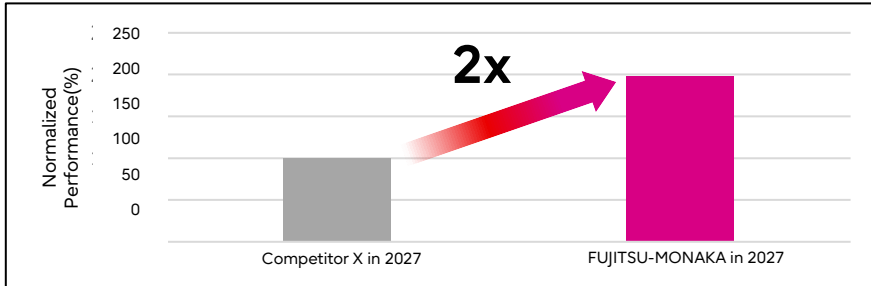- 3D many-core architecture
- Confidential Computing

FUJITSU-MONAKA

| High-performance | Energy Efficient | High Reliability & Security | Easy to Use |

Application Performance

2x

Normalized Performance(%)

250
200
150
100
50
0

Competitor X in 2027 — FUJITSU-MONAKA in 2027

Performance per Watt

2x

Normalized Performance(%)

250
200
150
100
50
0

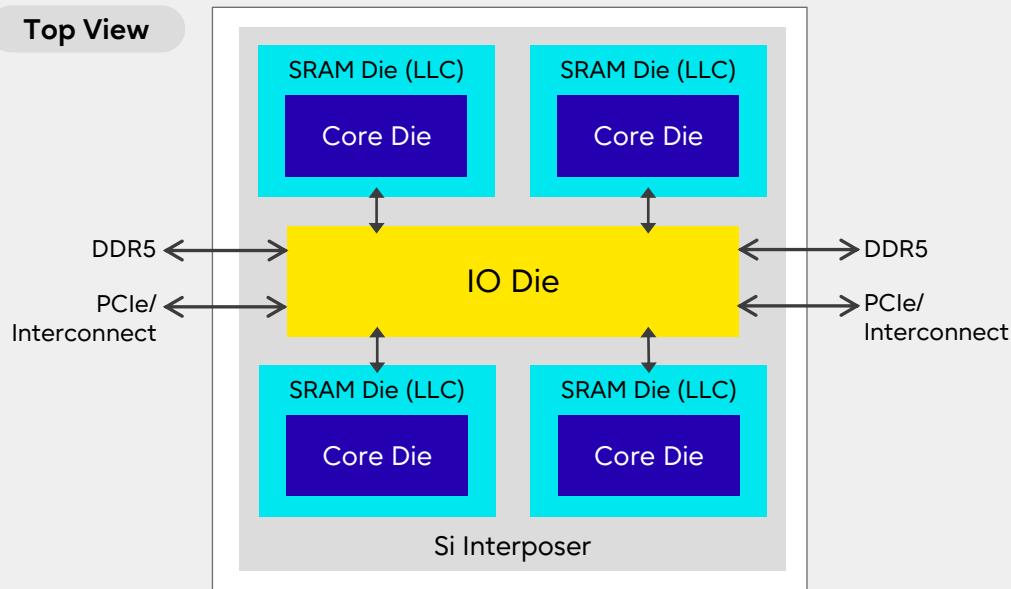Competitor X in 2027 — FUJITSU-MONAKA in 2027
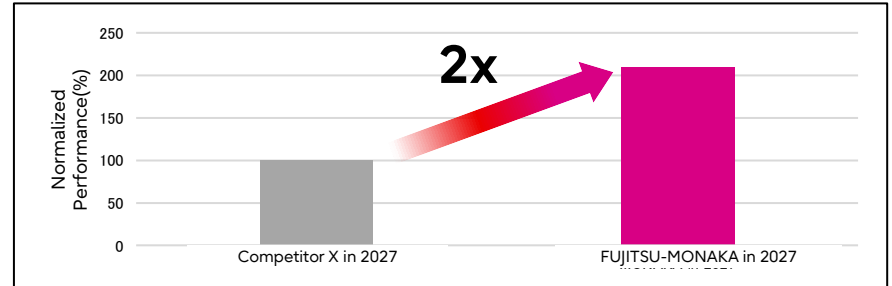
FUJITSU

FUJITSU-MONAKA specifications

- Armv9-A Architecture (SVE2[*1], CCA[*2])
- 144 cores x 2 sockets (288 cores / node)
- Ultra low voltage
- 3D chiplet
  - Core die 2nm
  - SRAM die/IO die 5nm
- DDR5 (12 channels)
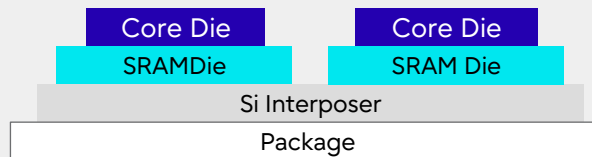- PCI Express 6.0 (CXL3.0)
- Air cooling

*1 SVE2(Scalable Vector Extension version two) is a superset of the SIMD instruction set SVE and NEON.
*2 CCA(Confidential Compute Architecture) is Arm architecture for confidential computing.

Top View

SRAM Die (LLC)	SRAM Die (LLC)
Core Die	Core Die

DDR5 ← IO Die → DDR5

PCIe/ Interconnect ← → PCIe/ Interconnect

SRAM Die (LLC)	SRAM Die (LLC)
Core Die	Core Die

Si Interposer

Side view

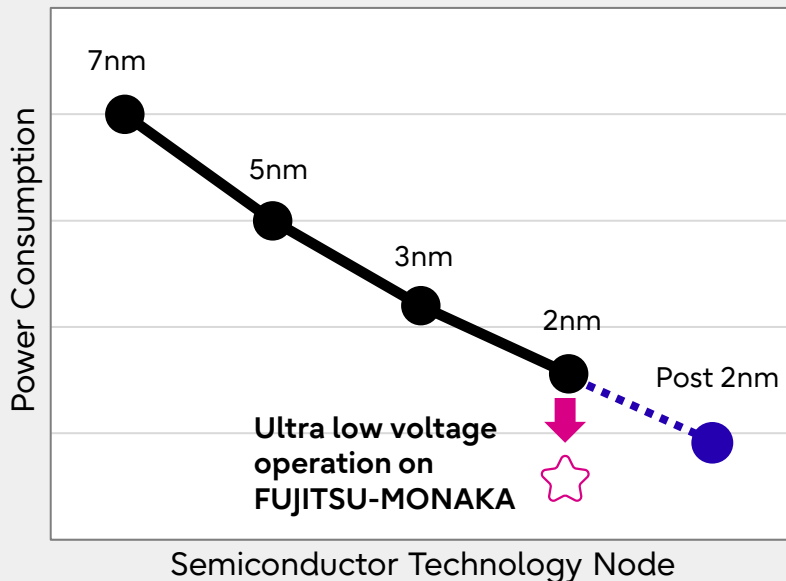| Core Die	Core Die
SRAMDie	SRAM Die

Si Interposer

Package

3

© 2024 Fujitsu

Ultra Low Voltage Technology

FUJITSU-MONAKA realizes carbon neutrality by ultra low voltage technology.

Trend of Semiconductor Power[1]

7nm

5nm

3nm

2nm

Post 2nm

Power Consumption

Ultra low voltage operation on FUJITSU-MONAKA

Semiconductor Technology Node

*1 Fujitsu estimation

- Lowering voltage reduces power consumption.
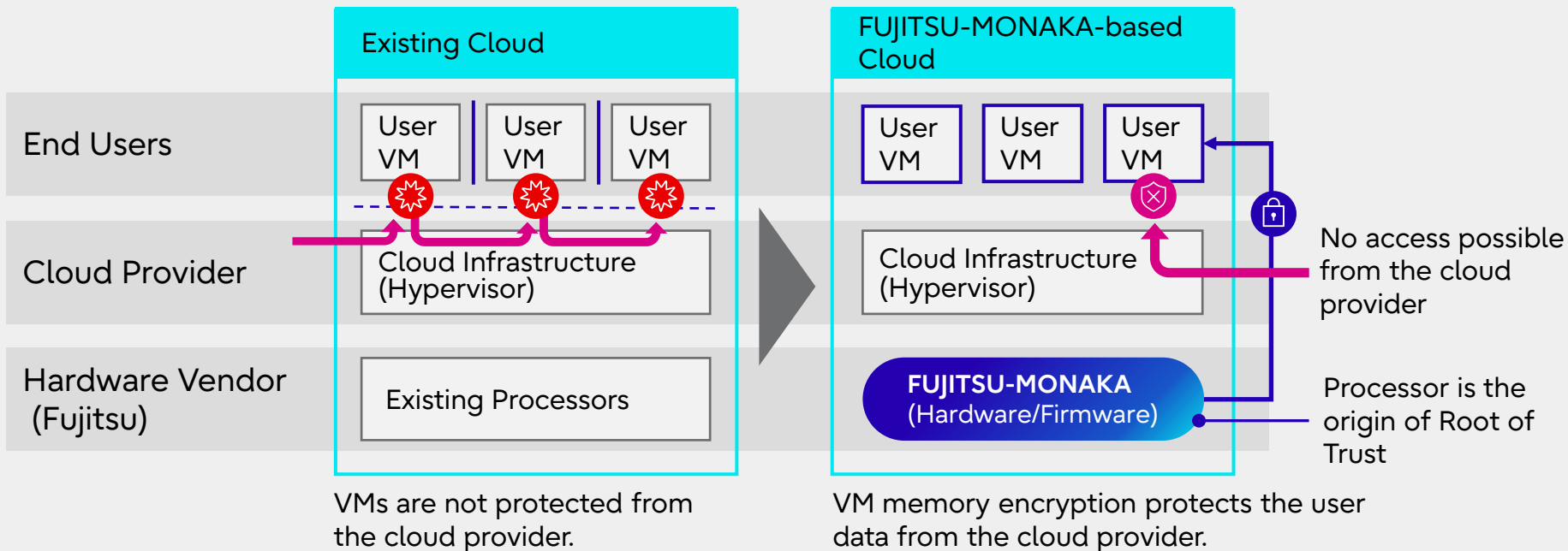
$$P \propto C V^2 f$$

C : Capacity
V : Voltage
f : Clock Frequency

- Our proprietary CAD contributes to realize stable operation at an ultra low voltage.

4

© 2024 Fujitsu

Security Enhancements

FUJITSU

- ## Confidential Computing
  - Protecting end-user data in memory by encrypting every VM
  - This is expected to be an essential technology in cloud, edge and HPC

Existing Cloud

End Users

| User VM | User VM | User VM |

Cloud Provider

Cloud Infrastructure (Hypervisor)

Hardware Vendor (Fujitsu)

Existing Processors

VMs are not protected from the cloud provider.

FUJITSU-MONAKA-based Cloud

| User VM | User VM | User VM |

Cloud Infrastructure (Hypervisor)

FUJITSU-MONAKA (Hardware/Firmware)

No access possible from the cloud provider

Processor is the origin of Root of Trust

VM memory encryption protects the user data from the cloud provider.

FUJITSU-MONAKA Target Domain

Easy to Use

Users can use OSS-based software stack without modification.

HPC | Batch job based HPC cluster system

AI | Inference and AI model development platform

DC | Data center workload and secure computing

Parallelization
- OpenMPI
- PMIx
- MPICH
- xCCL
- UCC/UCX
- Libfabric

Analysis tools
- Linaro Forge
- perf

Job/system management
- Slurm
- Warewulf
- Ansible

REST API
- FastAPI
- Flask

Serving
- KServe
- MLServer
- Triton

Framework
- scikit-learn
- PyTorch
- TensorFlow
- ONNX
- LangChain
- llama.cpp

LLM LangServe Milvus

MLOps
- MLflow
- Jupyter Notebook
- MinIO
- GitLab

CCA
- CoCo
- RA-TLS
- Veraison

Service Development
- PostgreSQL
- Keycloak
- pypiserver
- Quetz

Monitoring
- Grafana/Loki
- AlertManager
- Prometheus

Toolchain | LLVM | GCC | ArmCL | UXL | OpenBLAS | ArmPL | FFTW3 | OpenMP | oneTBB | Arm Optimized Routines

File System / Storage | Ceph | Ceph CSI | Lustre

Container Infra | Kubernetes | Harbor | Cilium | Istio | Knative

Virtual Infra | OpenStack

OS | Linux (Major distribution) | QEMU/KVM | libvirt | containerd/CRI-O/Kata Container | PAPI/libpfm

Hardware : Arm/FUJITSU-MONAKA

* Software list may be changed.

6

Collaboration with OSS Community

FUJITSU

Fujitsu enhances software ecosystem with OSS communities

Meta

15% speed up
(Arm SIMD support)

PyTorch

CCC*[1]

MIT

OpenBLAS

30% speed up
(Thread control improvement)

Fujitsu

arm

oneDAL   PyTorch

UXL
Unified Acceleration Foundation

oneDNN   oneDAL

31x speed up ML performance on Arm

Stock   6.48

0.21   **x31**

Random Forest Training

linaro™   LLVM

33% speed up SWPL*[2] and quality improvement

disable SWPL

enable SWPL   **x1.33**

*1 Confidential Computing Consortium
*2 Software Pipeline

Use Case Exploration with Customer

FUJITSU

- **Explore use cases with customer to expand Arm ecosystem**
- **Expand the FUJITSU-MONAKA application areas**

Use Case 1 — LLM + RAG*1

- FUJITSU-MONAKA improves operational efficiency of the generative AI tasks (e.g. LLM+RAG).

Lang Chain · CPU · CPU · GPU · LLM · Vector DB · CPU
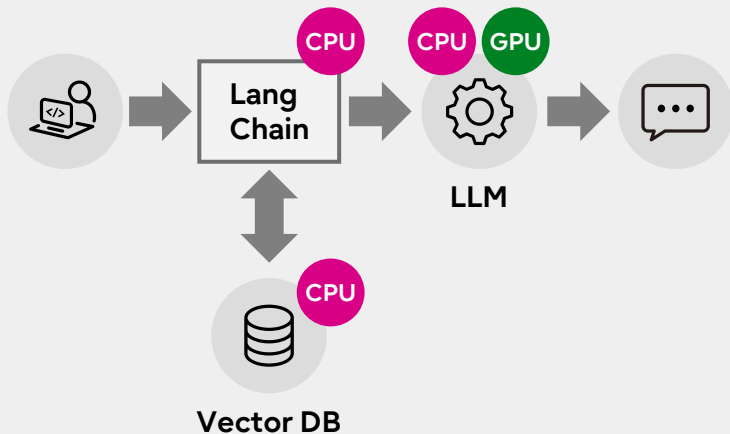
Use Case 2 — AI Surrogate Model

- FUJITSU-MONAKA accelerates the physics simulation with AI surrogate model.

PINNs*2	Simulations
✔ **High speed inference**	✖ Slow execution
✔ **Simple process**	✖ Complex process
✔ **High versatility**	✖ Problem-specific

VS

Accelerating CAE design with PINNs

Further acceleration with **FUJITSU-MONAKA**

*1 Retrieval-Augmented Generation
*2 Physics Informed Neural Networks

8

© 2024 Fujitsu

Conclusion

- Fujitsu develops the processor called FUJITSU-MONAKA with 3 architectural innovative key technologies.

  (Our Own Microarchitecture) (3D many-core) (Confidential Computing)

- FUJITSU-MONAKA supports open-source based software stack for 3 domains.

  (HPC) (AI) (Data Center)

- FUJITSU-MONAKA contributes carbon neutrality by ultra low voltage technology.

- Fujitsu enhances software ecosystem with OSS communities and co-creates use case with customers.

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

Thank you