June 6, 2024
ICS 2024

From Past to Future : The Legacy and Hypothesis of Supercomputing

Naoki Shinjo

SVP, Head of Advanced Technology Development Unit
Fujitsu Research FUJITSU LIMITED

Naoki Shinjo

SVP, Head of Advanced Technology Development Unit
Fujitsu Research
FUJITSU LIMITED

Naoki Shinjo

He Joined Fujitsu Limited in 1987. He belonged to the large-scale computer development division and was involved in supercomputer hardware development.

In 2006, he participated in the conceptual design of K computer, and from 2007, he was engaged in the development of K computer. After K computer development was completed, he participated in Fugaku feasibility studies from 2012.

In 2014, he became Head of Next Generation Technical Computing Unit and engaged in the development of Fugaku. Current position since April 2023.

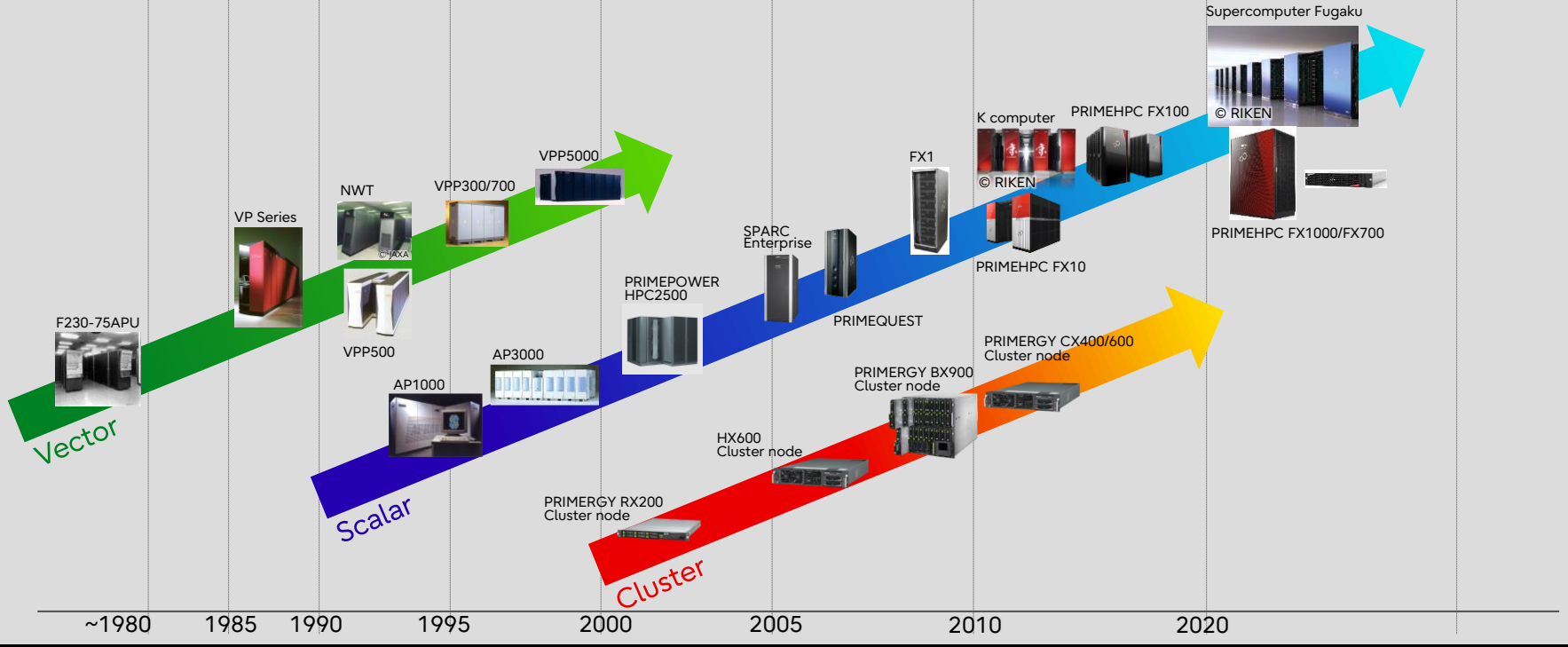
FUJITSU

Fujitsu's History in Supercomputer Development

Fujitsu's Supercomputers Released to World

FUJITSU

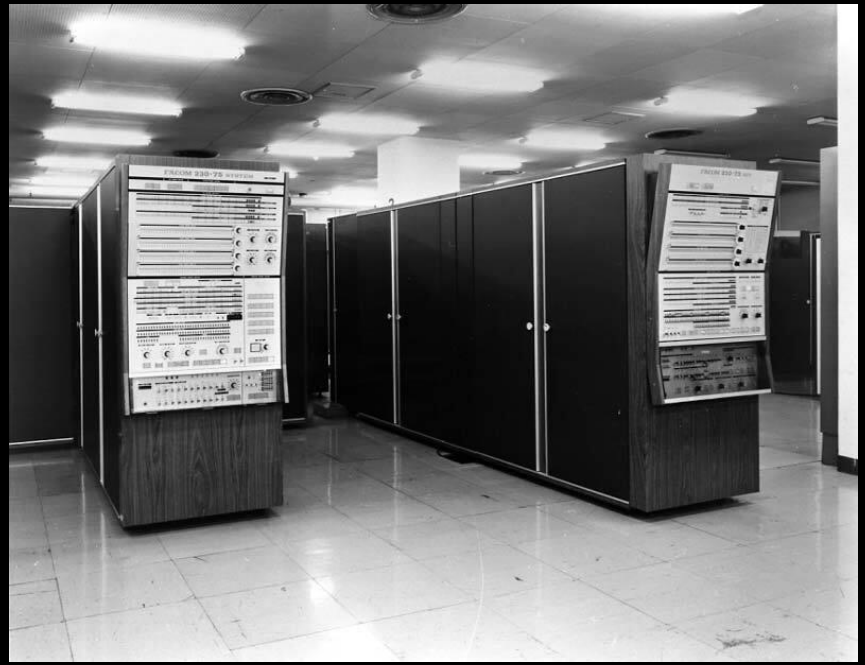
Fujitsu has continuously developed and delivered world-class supercomputers



FUJITSU

FACOM230-75 APU (1977)

Japan's first provision of supercomputers



- FLOPS :
  22 MFLOPS (Add FP32) / 11 MFLOPS (Mul FP32)
- Innovative Features (at that time) :
  - Vector Processing
  - APU (Array Processing Unit)
- Activity :

  National Aerospace Laboratory of Japan(NAL)

*NAL; a predecessor of JAXA

FUJITSU

VP series (1982~)

First “Giga scale” supercomputer in world *1

VP-200

- FLOPS :
  1.142 GFLOPS *1
- Innovative Features (at that time) :
  - Static RAM
  - Fast data sorting and conditional operations
- Activity :
  Nagoya University Plasma Research Center *2

*1 VP-400
*2 VP-100

FUJITSU

VP2000 series (1988)
Supercomputer with UNIX

VP-2100

- FLOPS :
  5G FLOPS *VP2600 maximum configuration
- Innovative Features (at that time) :
  - Virtual Machine (Enable UNIX)
  - Variety of product variations
- Activity :
  Rich industrial applications

FUJITSU

NWT (1993)

Ranked 1st in TOP500, Nov '93 and Nov '94-'95.

- FLOPS :
  280 GFLOPS (1.68 GFLOPE/PE)
- Innovative Features (at that time) :
  - GaAs LSI chips
  - distributed memory parallel vector computer architecture
- Activity :

  National Aerospace Laboratory of Japan (NAL)
  (Developed jointly by NAL and Fujitsu)

*NAL; a predecessor of JAXA

FUJITSU

VPP500 (1993)

Commercial HPC developed from NWT.

- FLOPS :
  355 GFLOPS (1.68 GFLOPE/PE) *maximum configuration
- Innovative Features (at that time) :
  - GaAs LSI chips
  - distributed memory parallel vector computer architecture
- Activity :

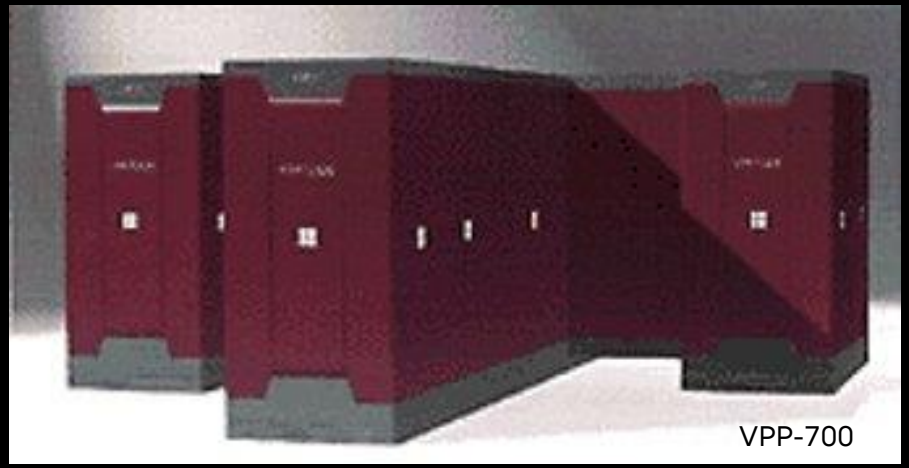
  National Laboratory for High Energy Physics in Japan (KEK)*,
  Japan Atomic Energy Research Institute(JAERI) and more.

*KEK; a predecessor of High Energy Accelerator Research Organization in Japan
*JAERI; a predecessor of Japan Atomic Energy Agency

FUJITSU

VPP300 / VPP700 (1995~)

Successor to VPP500 using CMOS technology



VPP-700

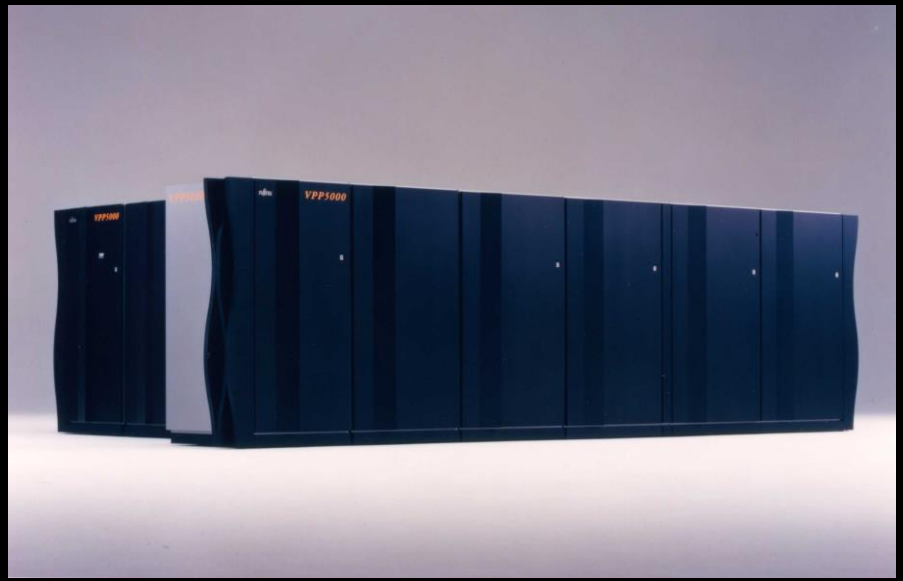
- FLOPS :
  1,126 GFLOPS (2.2 GFLOPS/PE) *VPP700 maximum configuration
- Innovative Features (at that time) :
  - CMOS technology
  - Sell widely by lowering prices
- Activity :

  LRZ,
  Kyushu University and more

FUJITSU

VPP5000 (1999)

Fujitsu's last Vector-type supercomputer



- FLOPS :
  4,915 GFLOPS (9.6 GFLOPS/PE) *VPP5000 maximum configuration
- Innovative Features (at that time) :
  - Fastest single CPU performance in the world
  - High execution efficiency(LINPACK 96.4%)
- Activity :
  Nagoya University, Europe and more

PRIMEPOWER HPC2500 (2002)

Dawn of Scalar-type supercomputer era

- FLOPS :
  85.1 TFLOPS (5.2 GFLOPS/CPU) *Maximum configuration
- Processor
  SPARC64Ⅴ: 1core, 1.30GHz, 130nm process
- Innovative Features (at that time) :
  - Scalar-type Supercomputer
  - Fusion of HPC and UNIX server technology
- Activity :

  National Aerospace Laboratory of Japan,
  Nagoya University,
  Kyoto University and more

FUJITSU

FX1 (2008)

Advanced many-core architecture

- FLOPS :
  110.6 TFLOPS (40 GFLOPS/CPU) *JAXA system
- Processor
  SPARC64Ⅶ : 4cores, 2.5GHz, 65nm process
- Innovative Features (at that time) :
  - Many-core architecture
  - highly-functional switch
  - High execution efficiency(LINPACK 91.19%)
- Activity :
  JAXA, Nagoya university

FUJITSU

K computer (2011)

Ranked 1st in TOP500, Jun 2011 and Nov 2011.

- FLOPS :
  10.51 PFLOPS (128 GFLOPS/CPU)
- Processor
  SPARC64Ⅷfx : 8cores, 2GHz, 45nm process
- Innovative Features (at that time) :
  - Tofu Interconnect
  - Water cooling
- Activity :

  RIKEN
  (Developed jointly by RIKEN and Fujitsu)

FUJITSU

PRIMEHPC FX10 (2012)

Commercial HPC that operates seamlessly 365 days.

- FLOPS :
  23.2 PFLOPS (236.5 GFLOPS/CPU) * Maximum configuration
- Processor
  SPARC64IXfx : 16cores, 1.848GHz, 40nm process
- Innovative Features (at that time) :
  - High reliability Technology
  - From academia to industry
- Activity :
  University of Tokyo, Kyushu university,
  Nagoya University and more

PRIMEHPC FX100 (2014)

Ambitious commercial HPC developed from K.

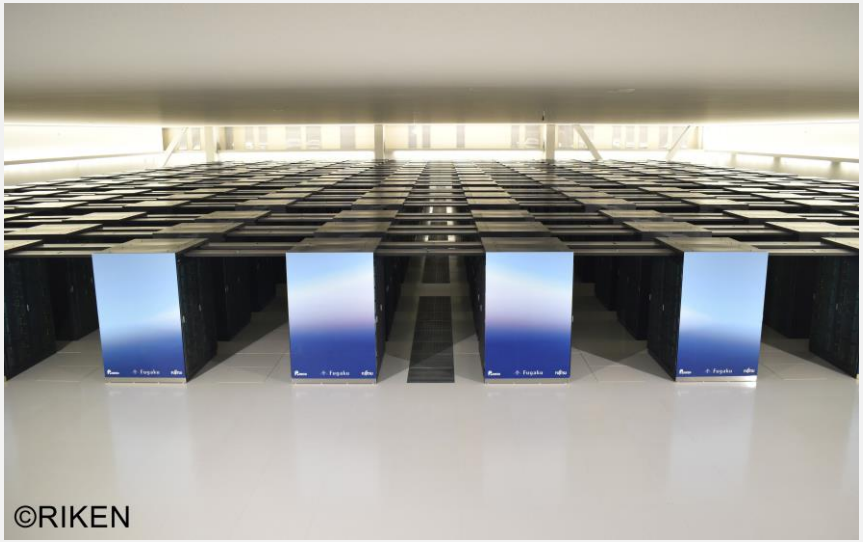
- FLOPS :
  over 100 PFLOPS (1.126 TFLOPS/CPU) *Maximum configuration
- Processor
  SPARC64XIfx : 32+2cores, 2.2GHz, 20nm process
- Innovative Features (at that time) :
  - HMC
  - Tofu Interconnect 2 (optical communication)
- Activity :

  National Institute for Fusion Science,
  JAXA, Nagoya university and more

FUJITSU

Supercomputer Fugaku (2020)

Ranked 1st in TOP500, Jun 2020 to Nov 2021.



- FLOPS :
  537.2 PFLOPS (3.38 TFLOPS/CPU)
- Processor
  A64FX : 48cores, 2.2GHz, 7nm process
- Innovative Features (at that time) :
  - Arm SVE architecture
  - HBM2
  - Energy Efficiency
- Activity :

  RIKEN
  (Developed jointly by RIKEN and Fujitsu)

Fugaku's Position in TOP500 Ranking

FUJITSU

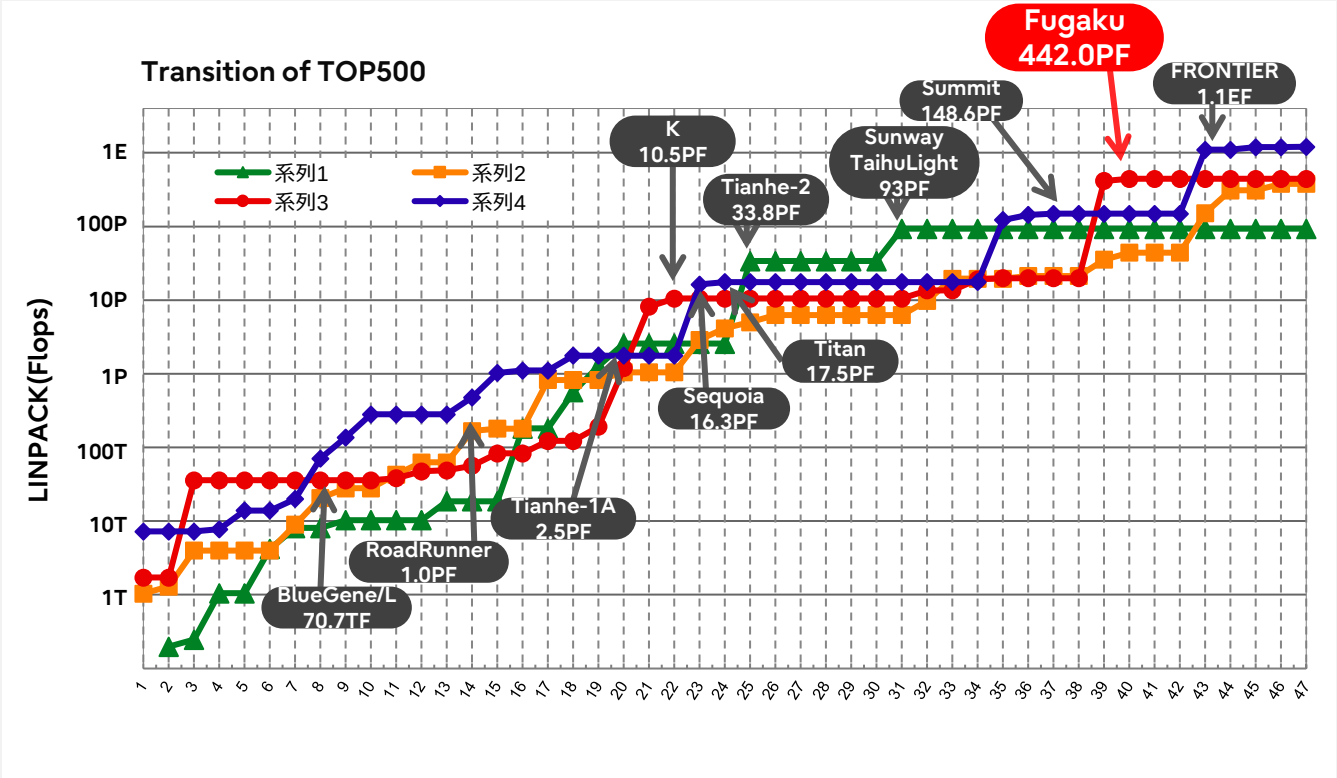
TOP500 No.1
Jun 2020 ~ Nov 2021

Graph500 No.1
Jun 2020 ~ Jun 2024

HPCG No.1
Jun 2020 ~ Jun 2024

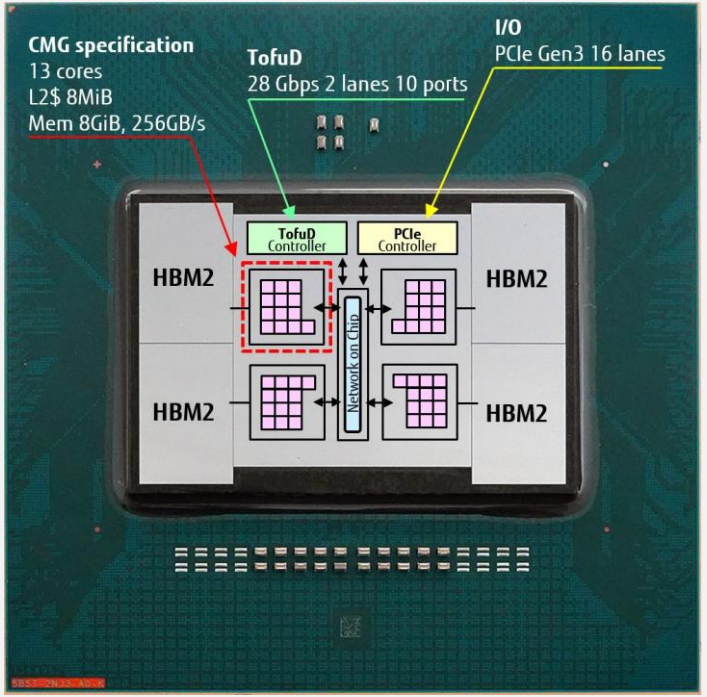
HPL-MxP No.1
Jun 2020 ~ Nov 2021

Green500 No.1
Nov 2019
* Fugaku prototype



Fujitsu Arm-based CPU A64FX

The A64FX CPU, independently developed by Fujitsu, is the heart of Fugaku, the world's fastest supercomputer. With approximately 160,000 A64FX CPUs interconnected, Fugaku achieved the highest performance at the time.



- Arm Architecture Extended for HPC
  - **Lead partner** in the development of Scalable Vector Extensions (SVE)
  - **World's first** implementation of a high-performance core supporting SVE
- Architecture for High Application Performance:
  - 512-bit SIMD
  - High memory bandwidth with HBM2
- Enhanced AI Computing Capabilities:
  - Supports FP16, INT16,INT8

PRIMEHPC FX1000 / FX700 (2019)

Commercial HPC capable of large-scale configuration beyond Fugaku.

FX1000

- FLOPS :
  1.328 EFLOPS (3.38 TFLOPS/CPU) *FX1000 maximum configuration
- Processor
  A64FX : 48cores, 2.2GHz, 7nm process
- Innovative Features (at that time) :
  - large-scale configuration (FX1000)
  - Fujitsu software Stacks (FX1000)
  - Software stacks leveraging OSS(FX700)
- Activity :

  Japan Meteorological Agency,
  University of Tokyo,
  Nagoya University and more

Fujitsu Processors Supporting HPC

FUJITSU

Performance

512-bit SIMD for HPC and AI
High Bandwidth Memory
Scalable Many-core Architecture
Virtual Machine Architecture
Software on Chip
High-speed Interconnect
HPC-ACE
System on Chip
Hardware Barrier
Multi-core Multi-thread
L2$ on Die
Non-blocking $
O-O-Execution
SuperO-Scalar
Single-chip CPU
Store Ahead
Branch History
Prefetch

Reliability

$ ECC
Register/ALU Parity
Instruction Retry
$ Dynamic Degradatio
Error Checkers/History

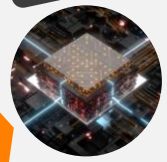
HPC

UNIX

Mainframe

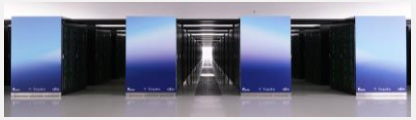
SPARC 64
SPARC 64 II
GS 8600
GS 8800
GS 8800B
GS 8900
SPARC 64GP
SPARC 64GP+
GS21 600
SPARC 64V
SPARC 64V+
SPARC 64VI
SPARC 64VII
SPARC 64VIIIfx
SPARC 64IXfx
SPARC 64XIfx
A64FX
SPARC 64X
SPARC 64X+
SPARC 64XII
GS21 900
GS21 1600
GS21 2600
GS21 3600

350nm
250nm
180nm
130nm
90nm
65nm
45nm
40nm
28nm
20nm
7nm
2nm



FUJITSU-MONAKA

Supercomputer Fugaku



© RIKEN

Technology Generation

~1999 2000 2005 2010 2015 2020

FUJTISU-MONAKA : Green Data Center Solution Rooted in Supercomputer Technology

*This section is based on results obtained from a project, JPNP21029, subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

Fujitsu Arm based Processor

FUJITSU

development goals

Performance	Power efficiency
x2	x2
Compared to other companies	Compared to other companies

To be shipped in 2027

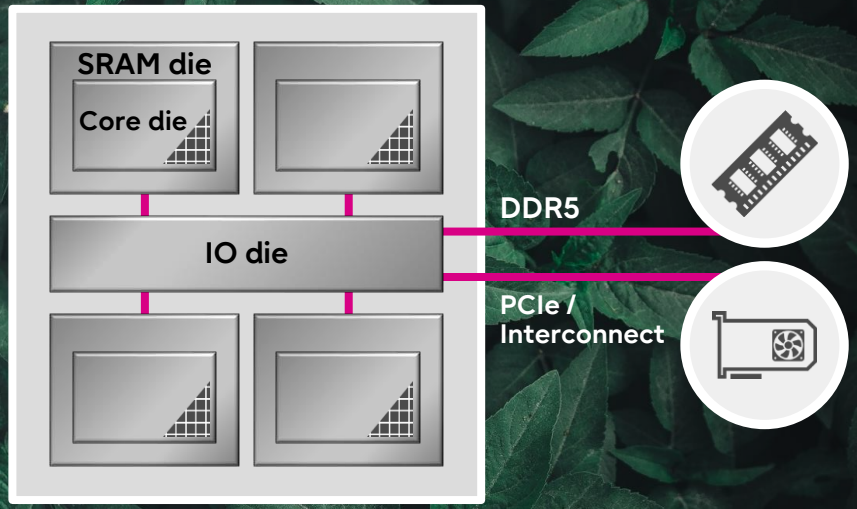
Create a sustainable digital society by developing the energy-efficient processor.

Hardware overview

FUJITSU-MONAKA adopts the innovative 3D many-core architecture

- 3D stacking of Core die and SRAM die via TSV enables more cores, low latency, and high throughput.
- By using different process sizes for Core die and SRAM die, the utilization rate of expensive fine-process technology is suppressed to 30%, achieving cost-efficiency.

FUJITSU-MONAKA Specification



Armv9-A Architecture
- SVE2 for AI and HPC
- Confidential Computing for security

144 cores x 2 sockets (288 cores per node)

Ultra low voltage for energy-efficiency

3D chiplet
- Core die 2nm
- SRAM die/IO die 5nm

DDR5 12 channels

PCI Express 6.0 (CXL3.0)

Air cooling

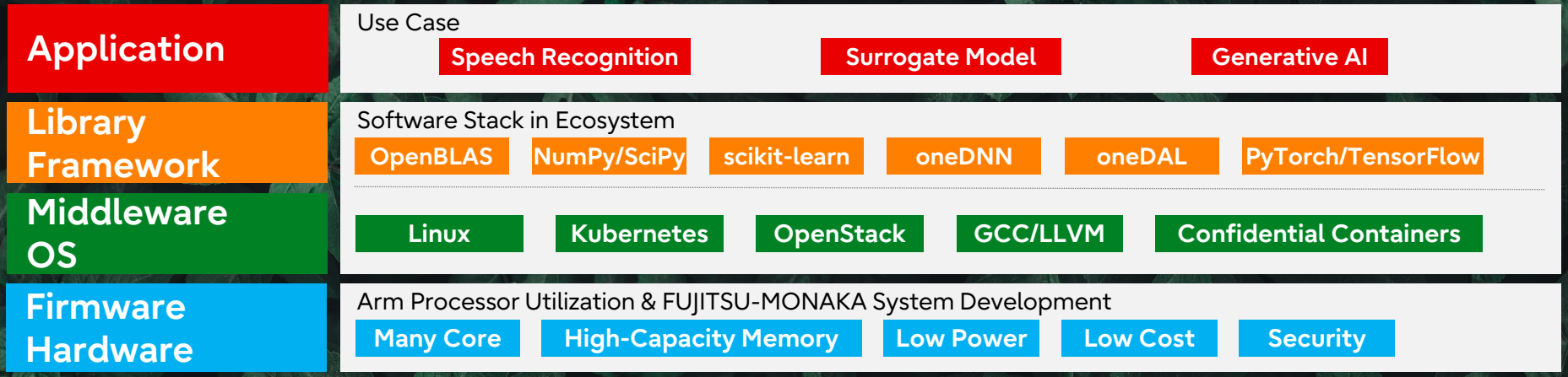
Software Overview

FUJITSU

FUJITSU-MONAKA focuses on open-source software in its software stack.

- Supports industry standard software
- Working on the standard tools (Python/Java/LLVM) to provide higher performance.
- Enabling smooth transition of customer assets and continuously enhancing performance

FUJITSU-MONAKA software stack

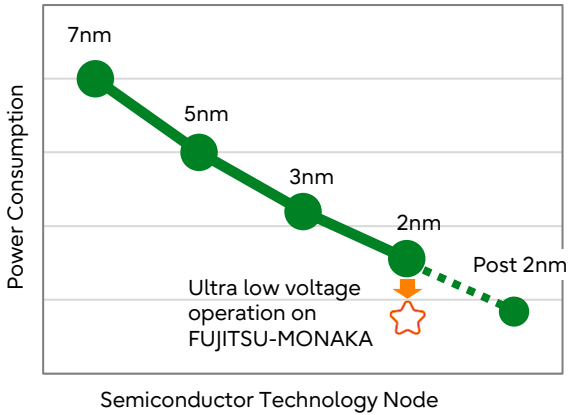


Leveraging Supercomputer Technology for Energy Efficient

Exceed vendor voltage, achieve next-gen efficiency.

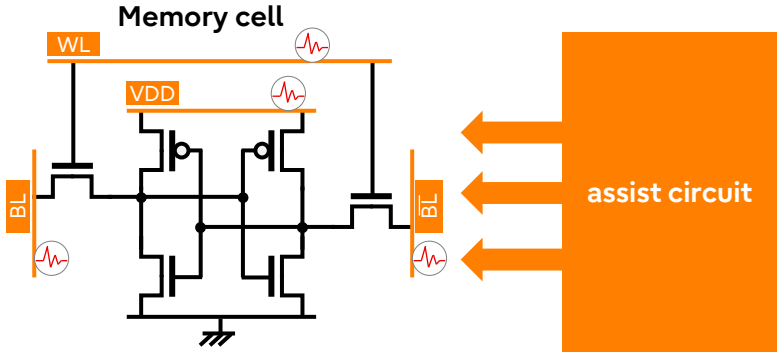
- Operate at a lower voltage than vendor-guaranteed, achieving power efficiency one generation ahead.
- This initiative alone is expected to achieve a 20% power reduction. Combined with other technologies, FUJITSU-MONAKA ultimately aims for double the power efficiency of other

power development



Energy efficient technology example: Ultra low voltage SRAM

SRAM circuits are the initial bottleneck for low-voltage operation. Fujitsu's innovative assist circuit design addresses this challenge, enabling stable operation at lower voltages.



Leveraging Supercomputer Technology for AI Workload

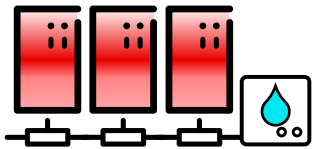
FUJITSU

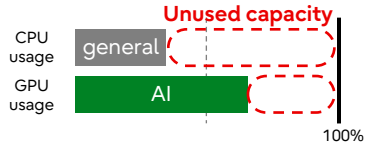
Economical and Right-sized AI infrastructure solution

- GPU clusters are costly and hinder business growth.
- FUJITSU-MONAKA, an air-cooled AI processor, offers a cost-effective and scalable solution.
- This will be achieved through AI/ML technologies cultivated by Fugaku.

AI infrastructure using only CPU

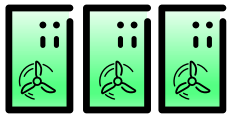
HOT AI infrastructure





High performance, but costly and cannot fully utilize resources.

COOL AI infrastructure



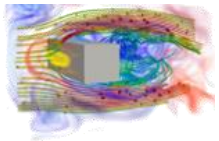
Right-sized performance with high efficiency and affordability.

AI solution example: Surrogate Model × OpenFOAM

preprocessing → post-processing

using only CPU

Fluid Surrogate Model

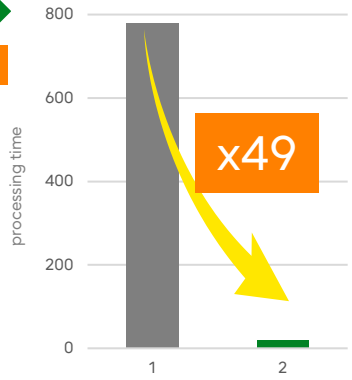


Interactive execution with visualization. Acquire approximate parameters.

Fluid Simulation

OpenFOAM

High-precision final verification using approximate parameters from the surrogate model.



Leveraging Supercomputer Technology for OSS

FUJITSU

Co-development with open-source communities to build an ecosystem

- Expand the scope of the software stack to datacenter systems
- Building software stack using real applications
- Promoting expansion of the Arm ecosystem

Key contributions to OSS Community

Open-source collaboration, including mission-critical systems and Fugaku, drives FUJITSU-MONAKA development.

- **2005** — Linux kernel for Mission Critical Server
- **2010** — KVM and Virtualization
- **2018** — OpenStack, Kubernetes
- **2022** — Automotive Grade Linux, Yocto, Arm Linux on Supercomputer Fugaku

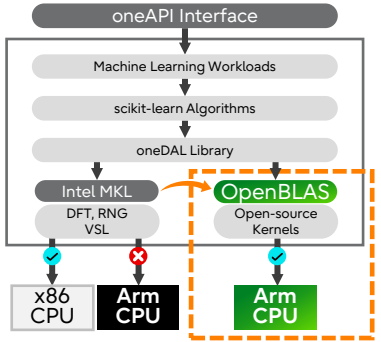
In progress

- Expand Arm software ecosystem
  - Feature development
  - Quality enhancement

with **Arm** **UXL** **Linaro**

Latest example : First successful Arm enablement for oneDAL

Successfully replaced MKL MATH functions with optimized open-source compute kernels of OpenBLAS



Fujitsu enabled core technology of oneDAL

Fujitsu's Contribution to Carbon Neutrality

The development of FUJITSU-MONAKA is under NEDO program(*1).
NEDO aims energy savings of 40% or more in datacenters by 2030 to realize a carbon-neutral society.

Fujitsu will contribute to the project by developing an energy-efficient CPU, FUJITSU-MONAKA.

- Refine Fujitsu's proven leading-edge microarchitecture.
- Cooperate with other technologies such as low-power consumption accelerator, photoelectric fusion device, wideband SSD, photonics smart NIC and disaggregation.

(details on Fujitsu press release, https://www.fujitsu.com/global/about/resources/news/press-releases/2022/0225-01.html)

(*1) NEDO program | "Technology Development of the Next Generation Green Data Center" for the "Green Innovation Fund Project/Construction of Next Generation Digital Infrastructure"

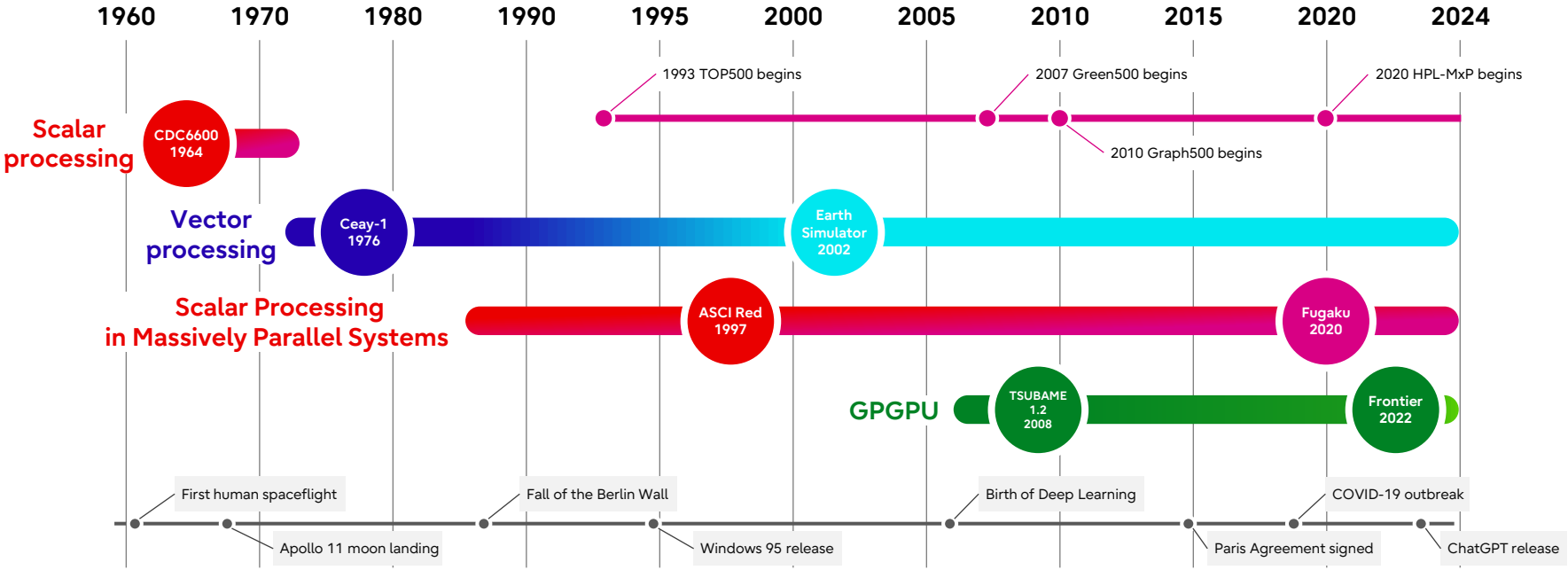
- NEDO is "New Energy and Industrial Technology Development Organization", a national research and development agency in Japan.
- Fujitsu has been selected for the national initiative along with NEC Corporation, AIO Core Co., Ltd., KIOXIA Corporation, Fujitsu Optical Components Limited and KYOCERA Corporation.

FUJITSU

Look Back at History of Supercomputer Architecture

Technology Trend Overview

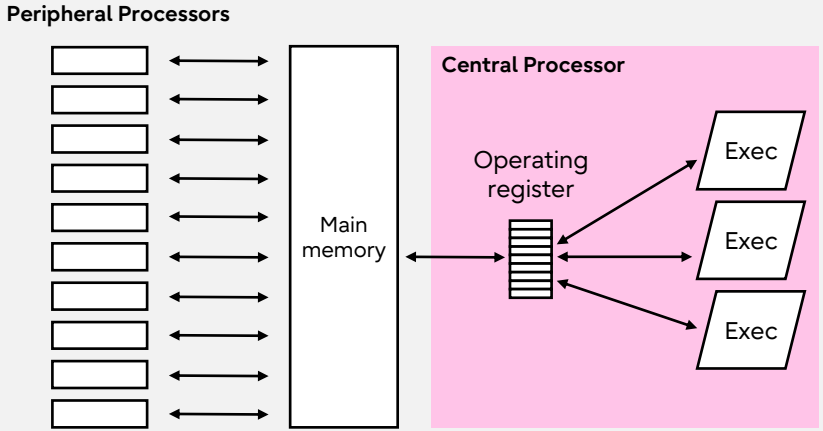
From the CDC 6600, often considered the first supercomputer (1964), to today, supercomputer architecture has been driven by the relentless pursuit of parallelization. This journey has witnessed three major turning points.



The CDC 6600 (1964), often considered the first supercomputer, was a groundbreaking machine that laid the foundation for modern supercomputing. It employed a scalar processor architecture, meaning it processed instructions one at a time, rather than in parallel like later vector processors.

- **First supercomputer (1964):**
  - Utilized a scalar processor architecture.
- **Unique architecture:**
  - Central Processor (CP) for instruction execution and calculations.
  - Peripheral Processors (PPs) for I/O operations, freeing the CP.
  - Main Memory shared by CP and PPs.
- **Efficient parallel processing despite scalar architecture.**
- **Innovative features:**
  - RISC-like design with a simplified instruction set.
  - SMT precursor for handling multiple instruction streams concurrently.
- **Paved the way for future supercomputer advancements.**

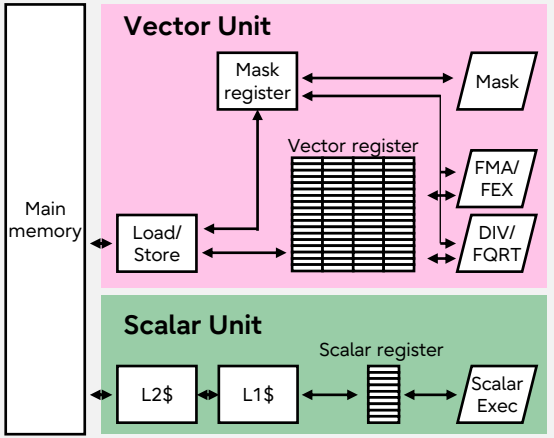
Configuration diagram



The 1970s witnessed the emergence of vector processing, revolutionizing scientific computing. Pioneered by machines like the Cray-1 (1976) and Fujitsu's FACOM 230-75 APU (1977), vector processors boasted specialized hardware for accelerating computationally intensive tasks.

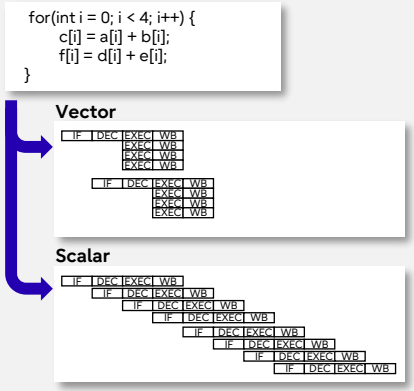
- **Vector Registers & Units:** Efficient manipulation of data elements in arrays.
- **Bulk Data Access:** Fetching entire vectors from memory with a single instruction.
- **Enhanced Parallelism:** Executing multiple instructions concurrently, boosting processing power.
- **Loop Optimization:** Vector units targeted loop-bound computations, while scalar units handled other tasks.
- **This combination of hardware and efficient data handling made vector processors ideal for scientific applications involving large datasets and repetitive operations. They paved the way for further advancements in supercomputing architecture.**

Example of configuration diagram



*VPP5000

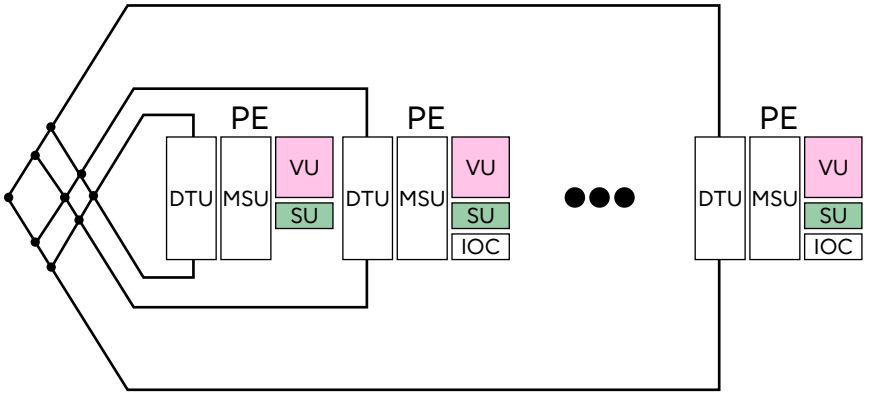
Image of instruction pipeline



- **Multi-vector-unit approach:** This approach increased parallelism, boosting performance.
- **memory exploration:** Companies experimented with unique memory configurations for optimal performance and resource utilization.

- Fujitsu's VPP Architecture: distributed memory parallel computer
  - Multiple processors with independent memory access eliminate the central memory bottleneck, enhancing scalability.
  - Flexible allocation of tasks and data among vector units for optimal resource utilization.
  - Enabled construction of powerful and efficient supercomputers.

VPP demonstrated the effectiveness of distributed memory architectures, which were a notable approach in early supercomputers.



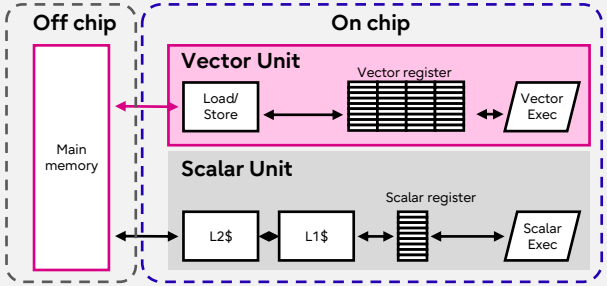
*VPP5000

Scalar Processing in Massively Parallel Systems

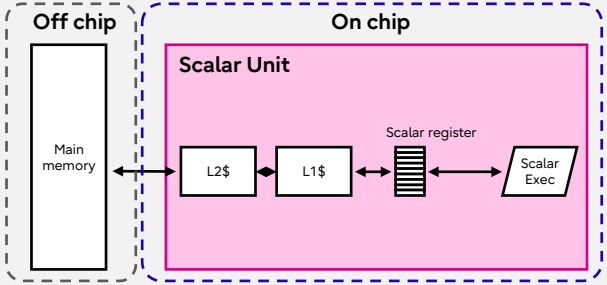
The relentless pursuit of performance in supercomputing led to the development of specialized vector processors. However, the increasing cost and complexity of vector architectures, coupled with the rapid advancements in scalar processor technology, eventually shifted the landscape towards scalar-based supercomputers.

- **Cost-effectiveness:** Scalar processors surpassed vector processors in price-to-performance ratio, becoming a more attractive option for HPC applications.
- **On-chip integration:** Unlike vector processors, which faced bottlenecks due to off-chip components, scalar processors benefited from on-chip integration of key performance elements, aligning with Moore's Law for lower cost and higher performance.
- **Large cache capacity:** On-chip caches reduced the need for main memory accesses, further boosting performance.
- These factors, in contrast to the limitations of vector processing, enabled the development of more powerful and efficient supercomputing systems based on readily available, cost-effective, and highly scalable scalar processors. This shift marked a turning point in supercomputing architectures.

Vector processing



Scalar processing



Scalar Processing in Massively Parallel Systems

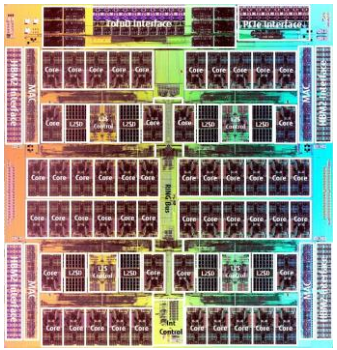
- Around 2010, the end of Moore's Law began to be discussed, marking the slowdown of performance improvements through frequency scaling. This shift led to the exploration of new parallelism approaches, as exemplified by the development of Fugaku.

Large-Scale CPU Parallelization

- Interconnecting numerous general-purpose processors through high-speed communication networks.
- Fujitsu's Tofu interconnect technology, with its 6D mesh/torus structure, enabled the connection of 393,216 nodes (1,024 racks) in the FX1000 supercomputer.
- This approach offered high inter-node communication speeds and fault tolerance.

Multi / Many Core Processors

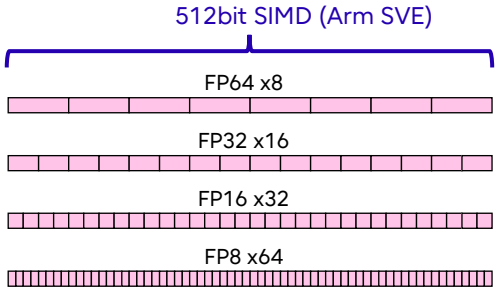
- The A64FX processor itself is a prime example of a multi-core processor, featuring 48 cores per chip. This design allows for increased parallelism within each node of Fugaku.
- This approach is driven by advancements in packaging technology, cooling solutions, and the present-day demand for cost-efficient cloud computing, where multiple virtual machines can be run on a single processor.



A64fx

SIMD Instructions

- Equipping scalar processors with vector-like instructions to enhance parallel processing capabilities.(Examples include ARM SVE, x86 AVX, and RISC-V RVV.)
- For instance, the A64FX can perform 8 FP64 operations or 64 FP8 operations simultaneously per core, demonstrating the potential of SIMD for accelerating various types of calculations, including those involving different precision levels.



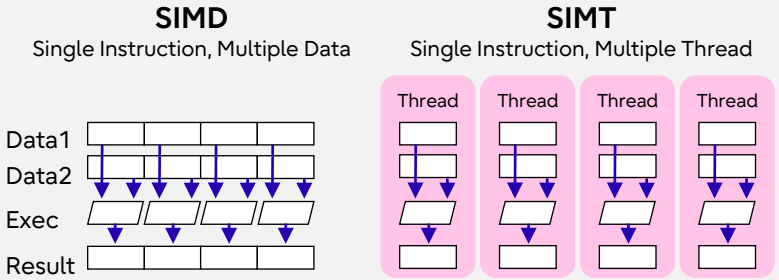
A64fx

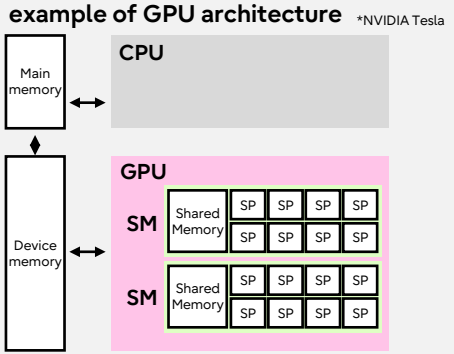
Heterogeneous Computing

FUJITSU

Heterogeneous computing, the use of multiple types of processors to achieve high performance, has become increasingly prevalent in supercomputing. Among the various forms of heterogeneous computing, the use of Graphics Processing Units (GPUs) has been particularly significant.

- **The Rise of GPGPU:**
  - GPGPU gained traction in the early 2000s, leveraging GPUs for general-purpose computations.
  - Large-scale HPC systems like TSUBAME1.2 and Tianhe-1 adopted GPUs around 2010.
- **Driving Factors:**
  - NVIDIA's CUDA platform (2007) made GPUs more accessible to HPC users.
  - Deep learning and cryptocurrency mining fueled GPU development.
- **HPC adapts to GPU:** GPU vendors have made HPC advancements like FP64 support. However, specialized architectures like NVIDIA's SIMT may not be ideal for HPC. Users may need to adapt their workflows.
- **Impact on Supercomputing :** GPUs have significantly impacted supercomputing, enabling more powerful and efficient systems.





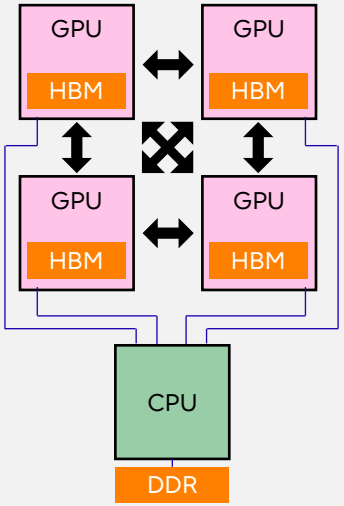
Heterogeneous Computing

FUJITSU

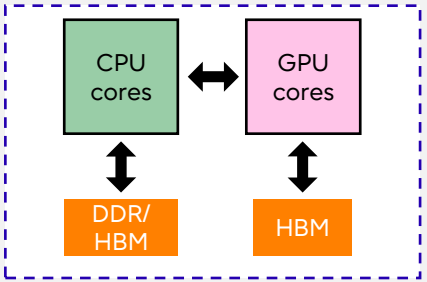
The growing demand for AI, particularly deep learning and LLMs, has driven advancements in supercomputing architectures. This focus has spurred advancements in both specialized AI architectures and traditional HPC architectures, exemplified by Frontier, the world's first exascale supercomputer with an AMD CPU+GPU architecture.

- **AI Workloads and Heterogeneous Computing:**
  - AI workloads often rely on low-precision floating-point arithmetic, which benefits from GPUs.
  - This focus on AI continues to lead to improvements in HPC workloads due to shared advancements in parallelism and memory technology.
- **Convergence of CPU and GPU:**
  - Vendors are introducing solutions that integrate CPUs and GPUs within a single package to further optimize performance.
- **Impact on Supercomputing:**
  - AI continues to significantly influence supercomputing architectures, leading to advancements in parallelism, memory bandwidth, and heterogeneous computing.
  - These advancements enable the development of more powerful and versatile supercomputers for both traditional HPC and emerging AI workloads.

Discrete GPU



Packaged GPU

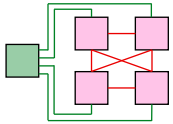


TOP10 Architectures (Jun 2024)

FUJITSU

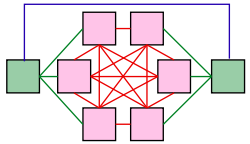
1.Frontier (2022~)

- 1,206.0 PFLOPS (70.3%)
- 22.79 MW
- 9,408 nodes
- 1 cpu + 4 gpu / node
  - AMD EPYC
  - AMD MI250X



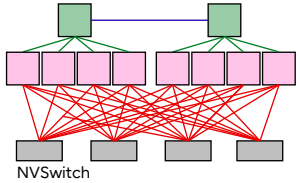
2.Aurora (2023~)

- 1012.0 PFLOPS (51.1%)
- 38.7 MW
- around 11,000 nodes *1
- 2 cpu + 6 gpu / node
  - Intel Xeon CPU Max
  - Intel Data Center GPU MAX



3.Eagle (2023~)

- 561.2 PFLOPS (66.2%)
- N/A MW
- around 1800 nodes *1
- 2 cpu + 8gpu / node
  - Intel Xeon Platinum
  - NVIDIA H100

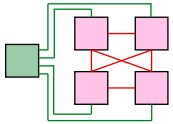


4.Fugaku (2020~)

- 442.0 PFLOPS (82.2%)
- 29.9 MW
- 158,976 nodes
- 1 cpu / node
  - A64FX

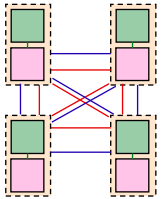
5.LUMI (2022~)

- 379.7 PFLOPS (71.4%)
- 7.11 MW
- 2,978 nodes
- 1 cpu + 4 gpu / node
  - AMD EPYC
  - AMD MI250X



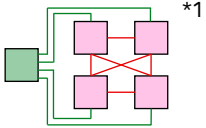
6.Alps (2024~)

- 270.0 PFLOPS (76.3%)
- 5.19 MW
- 2,688 nodes
- 4 x (cpu + gpu) / node
  - NVIDIA GH200



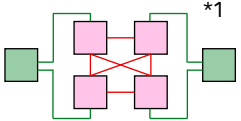
7.Leonardo (2022~)

- 241.2 PFLOPS (78.7%)
- 7.49 MW
- 3,456 nodes
- 1 cpu + 4 gpu / node
  - Intel Xeon Platinum
  - NVIDIA A100



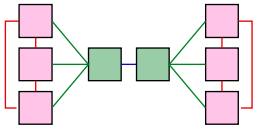
8.MareNostrum 5 ACC (2023~)

- 175.3 PFLOPS (70.2%)
- 4.16 MW
- 1,120 nodes
- 2 cpu + 4 gpu /node
  - Intel Xeon Platinum
  - NVIDIA H100



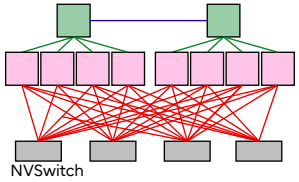
9.Summit (2018~)

- 148.6 PFLOPS (74%)
- 10.1 MW
- 4,608 nodes
- 2 cpu + 6 gpu / node
  - IBM POWER9
  - NVIDIA GV100



10.Eos NVIDIA DGX SuperPOD (2023~)

- 121.4 PFLOPS (64.3%)
- N/A MW
- 576 nodes *1
- 2 cpu + 8 gpu / node
  - Intel Xeon Platinum
  - NVIDIA H100



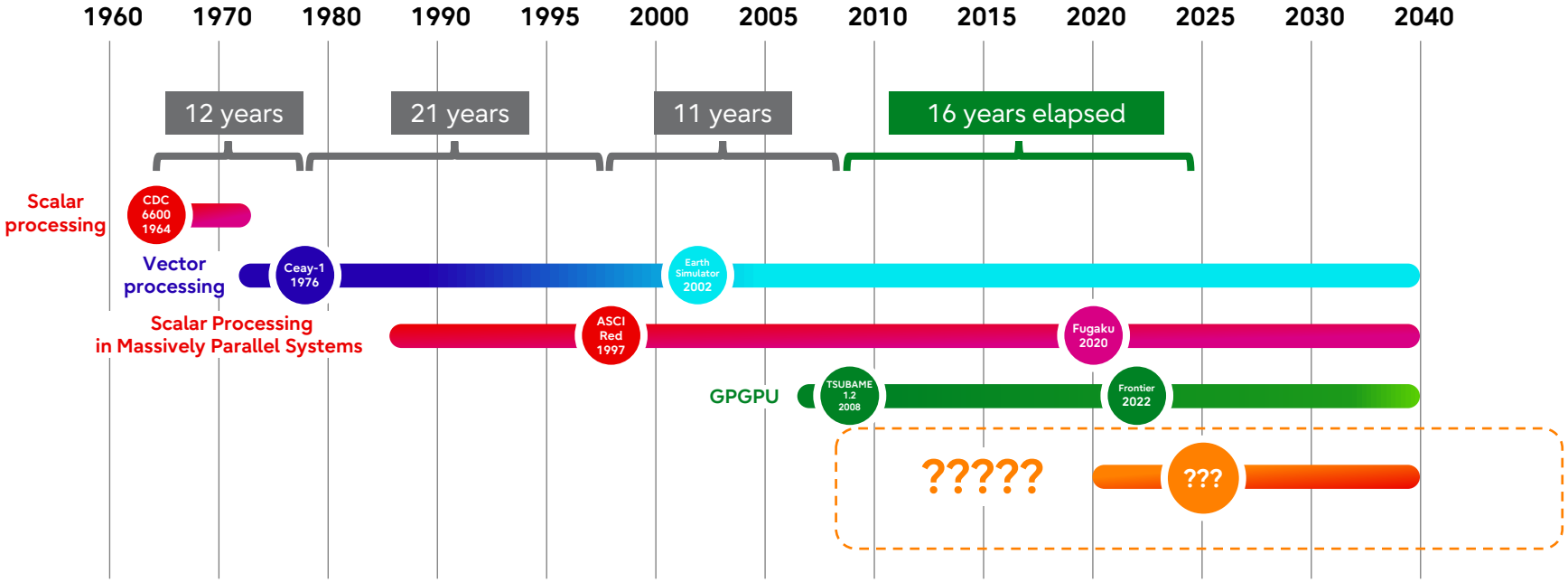
*1 : The details have not been made public, so this is my speculation.

FUJITSU

From Historical Review to Future Speculation

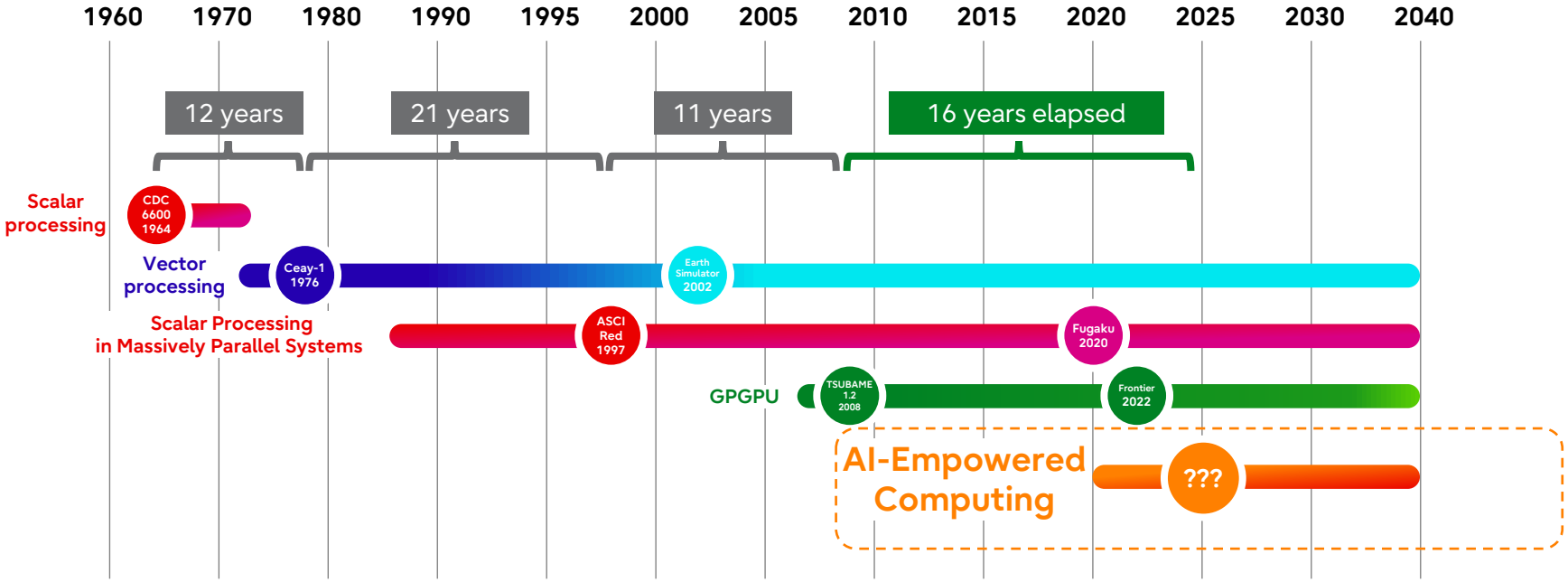
Unveiling the Next Era: What Will It Be Called?

Supercomputers have seen three major turning points, ushering in new eras every 10-20 years. Historically, this timeframe has proven sufficient for significant shifts. Sixteen years have passed since GPGPUs entered large-scale configurations, indicating that we are likely on the cusp of another major transition.



Unveiling the Next Era: What Will It Be Called?

Supercomputers have seen three major turning points, ushering in new eras every 10-20 years. Historically, this timeframe has proven sufficient for significant shifts. Sixteen years have passed since GPGPUs entered large-scale configurations, indicating that we are likely on the cusp of another major transition.

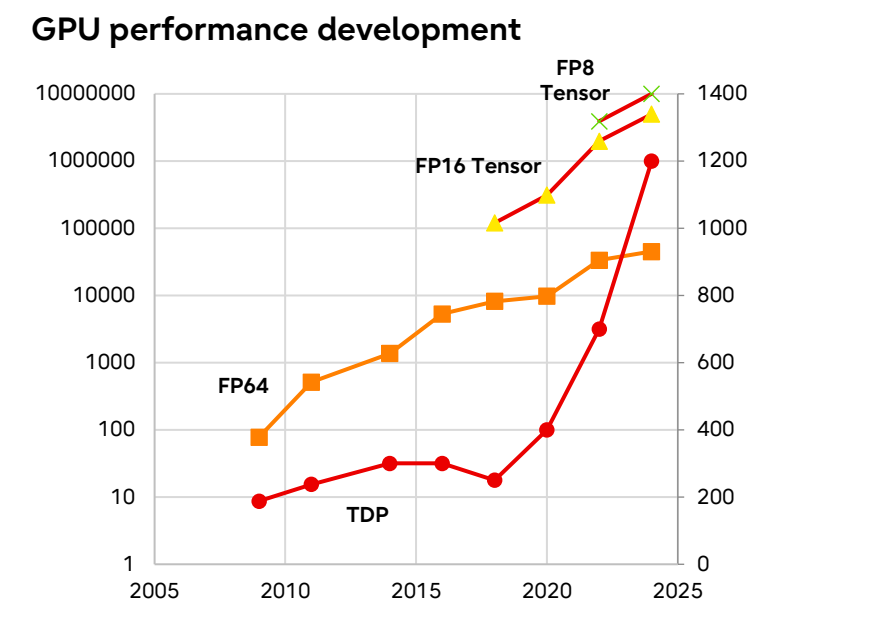


The focus of GPGPU is shifting towards AI.

AI is a lucrative market, driving industry focus towards low-precision computing. Double-precision computing, essential for scientific research, is neglected. Scientists must adapt to this shift by exploring alternative solutions.

1. **Slowdown in FP64 Performance development**
2. **Rapid TDP Increase**
3. **Increased Complexity due to Mixed Vector and Matrix Units**

While low-precision computing advances rapidly, FP64 stagnates, and rising TDP limits its future.
This necessitates exploring alternative solutions, including leveraging low-precision instructions for high-precision scientific computing.



What Should HPC Users Do?

FUJITSU

- **Software**
  - Explore the development of new libraries that use low-precision computing to achieve high-precision results.
  - Develop applications like HPL-MxP that inherently use mixed-precision computing.
  - Explore new AI solutions like surrogate models for high-precision computing.
- **Hardware**
  - Focus on the improved FP64 performance of CPUs and consider a return to CPU-based computing when necessary.
  - Pressure GPU vendors to improve the FP64 performance of GPUs.

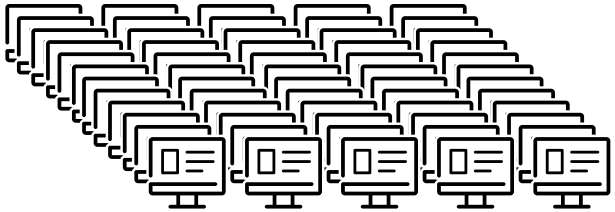
Practical Examples of AI-Empowered Computing

FUJITSU

I believe that AI for HPC, which incorporates AI workloads into simulations, will become mainstream. In particular, I see potential in combining surrogate models with simulations, where the majority of the processing is done with a rough pre-processing using deep learning, and the parameters obtained there are used to improve the accuracy with FP64 instructions. With numerous successful precedents emerging, this approach is likely to become widely adopted in the near future.

GPGPU

- Conducts countless scientific simulations with high-precision calculations until convergence is achieved.
- Requires high-fidelity models to be computed with high precision to obtain promising results.

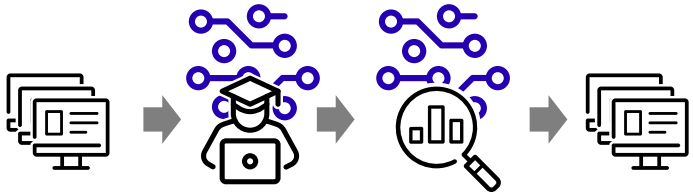


Simulation

High-precision simulation

AI-Empowered Computing

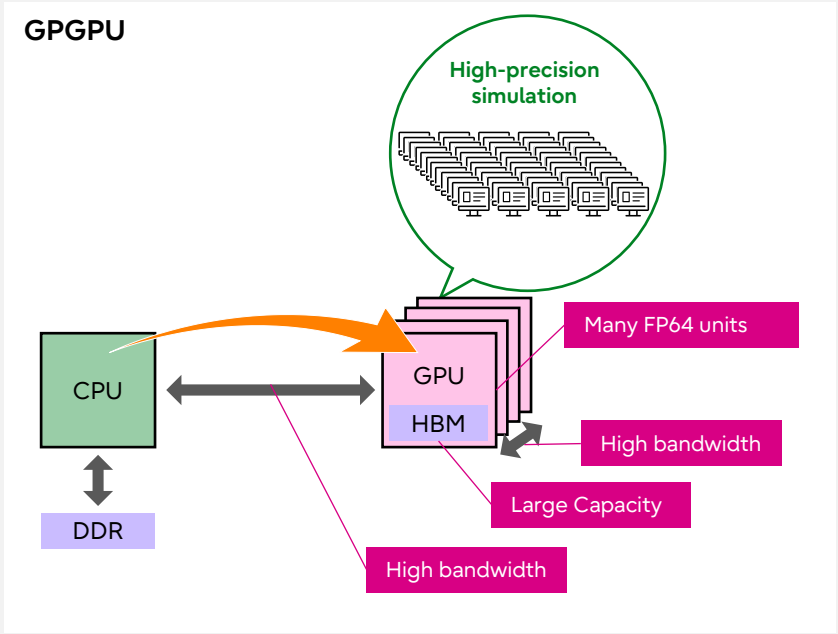
- Leverages low-precision calculations in a preprocessing stage to refine approximate values, followed by high-precision scientific simulations in a post-processing stage.
- Performs the majority of calculations with low precision using a simplified model (surrogate model), leading to enhanced computational efficiency.

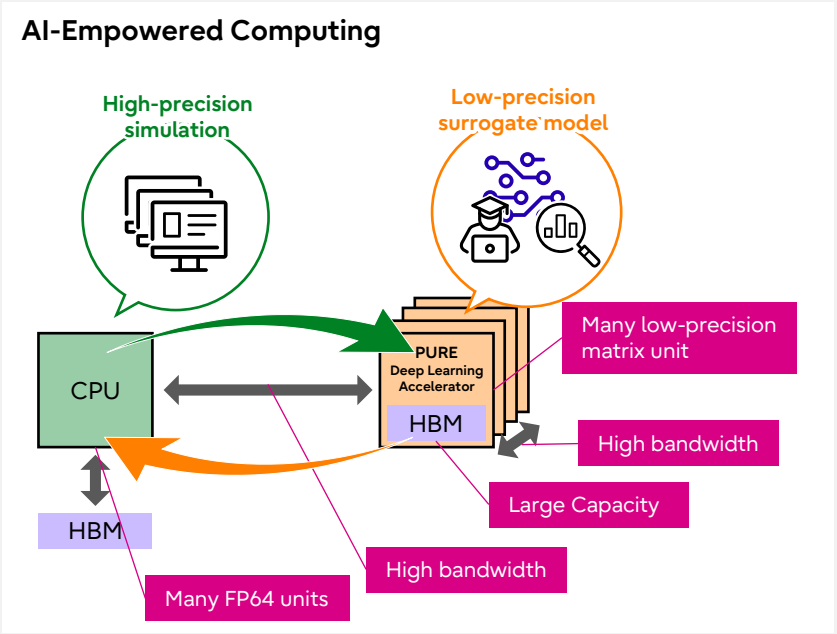


Simulation	AI estimation	AI Inference	Simulation
High-precision simulation	Low-precision create surrogate model	Low-precision use surrogate model	High-precision simulation

Architectures suitable.

In pursuit of high-performance and energy-efficient architectures, the ideal form will diverge from the present. The market will demand PURE Deep Learning Accelerators, stripped of unnecessary AI features. Additionally, reduced high-precision computations will enable efficient execution on CPUs, leading to a resurgence in CPU performance requirements.





Conclusion

FUJITSU

- Fujitsu's supercomputing legacy, from early machines to Fugaku, continues to influence data center technology like the FUJITSU-MONAKA.
- Supercomputer architecture has undergone several major transitions, driven by the need for increased performance and efficiency.
- The market is changing, and supercomputing is once again at a major turning point. It presents both challenges and opportunities.
- Looking ahead, advancements in AI and other technologies will drive exciting new possibilities for supercomputing.

Thank you