

Democratizing the use of AI with - FUJITSU MONAKA

oneAPI Dev Summit on AI and HPC 2023



Priyanka Sharma, PhD

Director, Software Engineering
MONAKA SW R&D (HPC AI) Centre
Fujitsu Research of India





Converging
Technologies



Data & Security



AI




Network



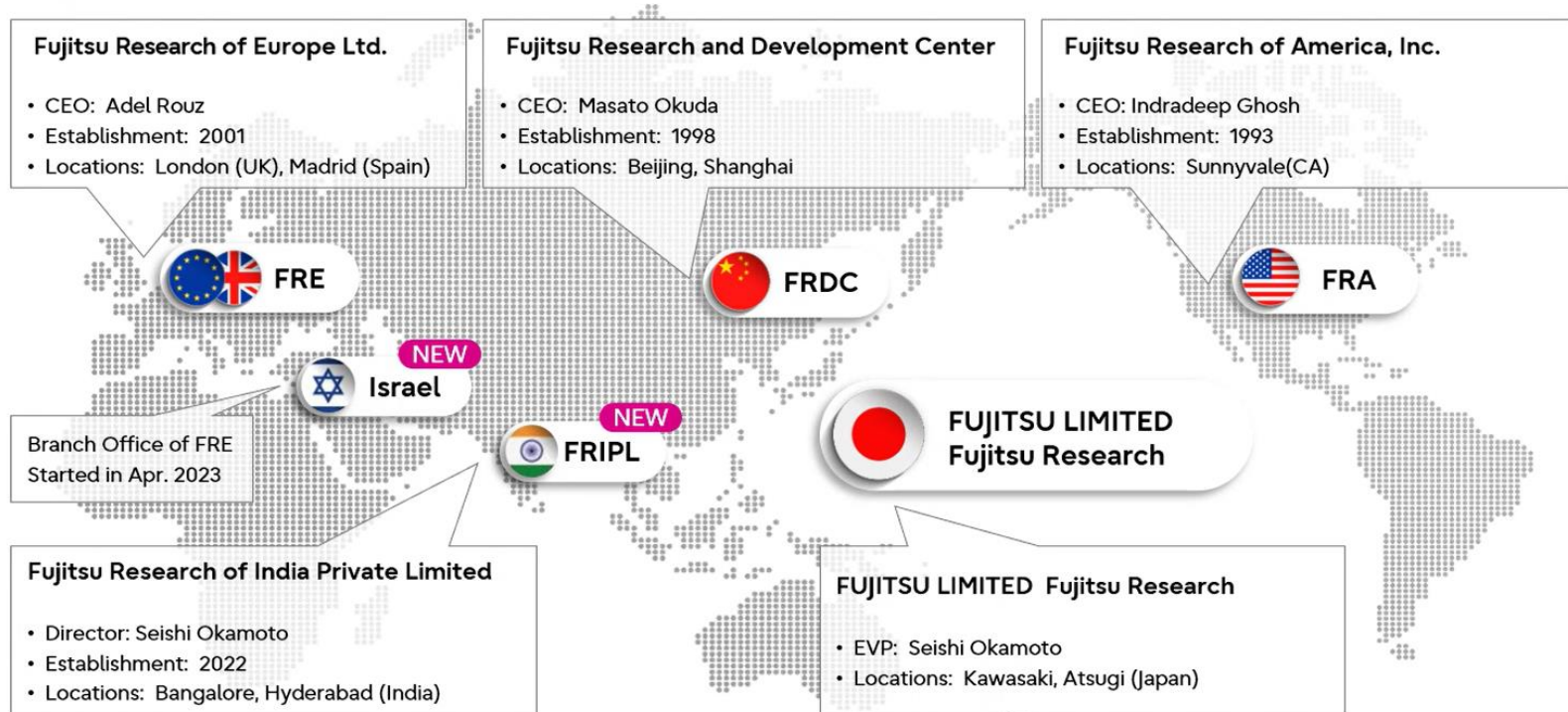
Computing

Fujitsu's Technologies



The present generation is witnessing the fastest technological revolutions of all times and AI and high-end Compute is its biggest enabler!!

Fujitsu's commitment to R&D and innovations for sustainable digital transformation



High-speed/high-precision quantum chemical calculations can be performed without expertise

Zero-emission materials



Halve the development time of new energy materials with AI x Computing technology and contribute to realizing a sustainable world

Large-scale industrial design



Fusion of computing and design optimizes designing that realizes comfortability for people and excellent functionality

Drug discovery without side effects



Using massively parallel computing power of a supercomputer, analyze numerous proteins and develop drugs with fewer side effects

Personalized healthcare



Analyze a patient's genetic information using AI that is accelerated by computing power and select the appropriate treatment for each patient

Computing Workload Broker

First in the world! Technology for easily utilizing hybrid calculation of quantum and HPC

Quantum Computer/Simulator

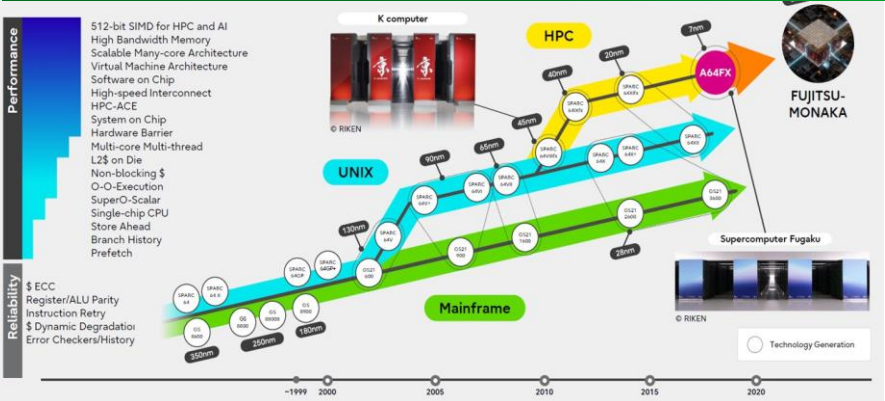
AI Accelerator

CPU (FUJITSU - MONAKA)

MONAKA is backed by Fujitsu's 60 year old legacy in supercomputing domain!



We specialize in making our own microarchitecture for processor development



- Fugaku, the fastest supercomputer of the world!
- Its built using more than 152K Fujitsu's A64FX Arm CPUs



GRAPH500
Big Data Processing No.1 (*1)

HPCG
Real-world Application Workload No.1 (*1)

TOP500
Floating-point Computation No.2 (*2)

HPL-AI
Machine Learning Computation No.3 (*2)

(*1) 8 consecutive terms since June 2020

(*2) Ranked No.1 for 4 consecutive terms until Nov 2021



Our 60 year old legacy in HPC!

FUJITSU PUBLIC

Solving Social Issues with Computing



Achievements of the supercomputer Fugaku*

*Powered by A64FX, Fujitsu's world-leading Arm CPU

Disaster Prevention

Real-Time Tsunami Prediction

Medicine

Cancer Gene Network Analysis in Less than a Day

Meteorology

Largest Ever Meteorological Calculation

A finalist for the ACM Gordon Bell Prize in 2020

COVID-19

Modelling the Spread of Droplets and Aerosols

A winner for the ACM Gordon Bell Special Prize in 2021



MONAKA (2nm Arm CPU) Hardware architecture launch video
(released during SC 2023, Denver, USA)

FUJITSU

FUJITSU

NEW CPU:
FUJITSU-MONAKA





MONAKA's Delivery Focus

Arm-based 2nm CPU
FUJITSU-MONAKA



Fujitsu microarchitecture

3D many-core architecture

Confidential Computing



High-performance

- Cloud native 3D many-core design by Fujitsu-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

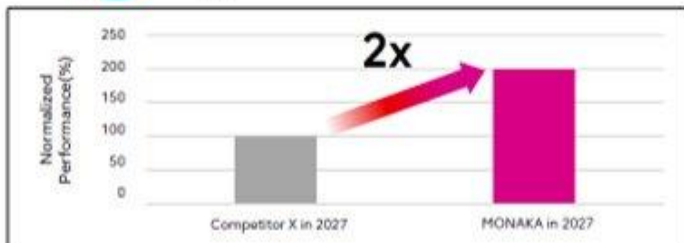


Easy to Use

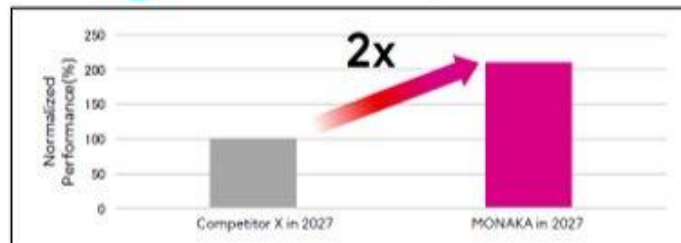
- Open & de-facto standard software stacks
- Fujitsu compiler technology
- Air-Cooling for easy deployment



Application Performance



Performance per Watt



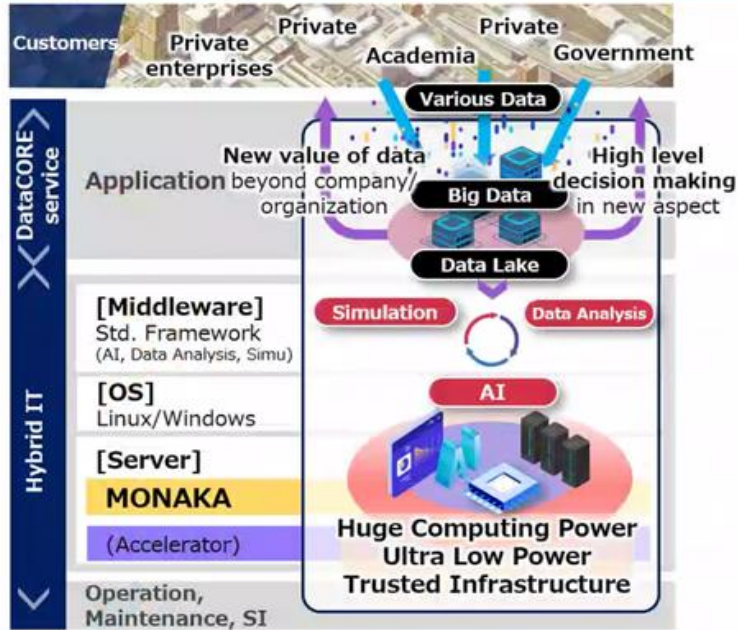
(*1) NEDO program

"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.

Supporting society by Computing power

Strengthen technology foundation in Horizontal Areas and contribute to the business in Vertical Areas where huge computing power is required



Promote low power consumption as an advantage in proceeding business

- Adoption widely expanded to domestic DC and contribute to environmental issues by promoting low-power DC
- Continued acquisition of HPC for private sector and academia



Expand utilization in Domestic DC

- Promote high performance, low power consumption, low cost, and security functions to DC company widely in Japan



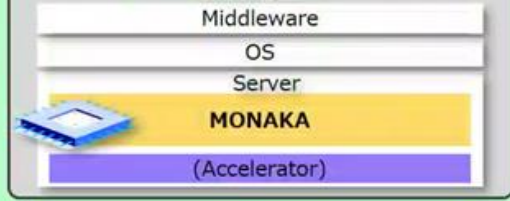
Acquisition of conventional HPC

- Development and diffusion of HPC software
- Expand from HPC to combination with AI/Data analysis

Provide system

Provide system

MONAKA platform

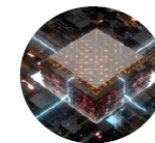
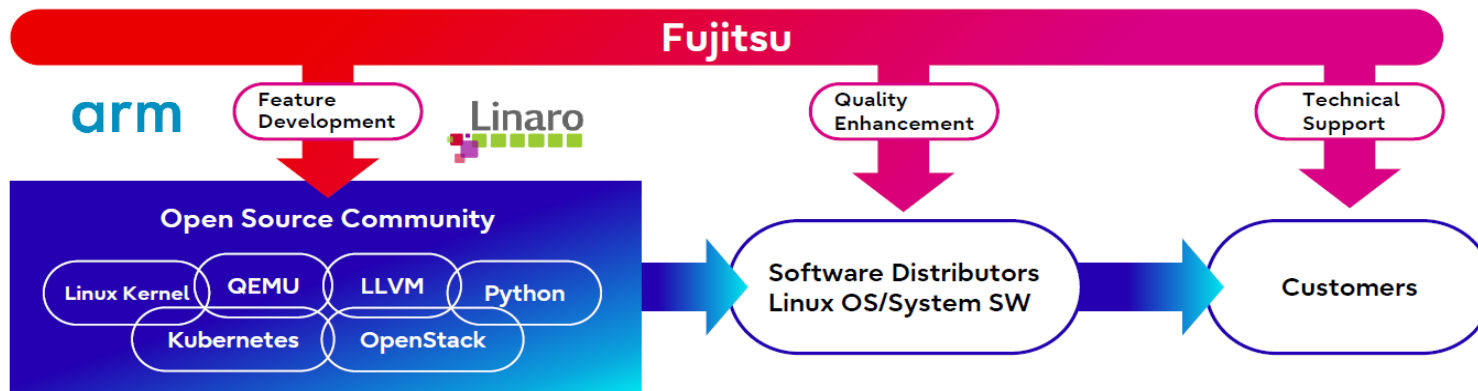


Provide private cloud service

- **Supports industry standard software**

- Standard Linux OS support and system architecture
 - Continue and expand OSS development activities for FUJITSU-MONAKA
 - OSS development achievements for Fugaku/A64FX: GCC, glibc, live-patch, papi, etc
 - Comply with standard system architecture (Arm System Ready) and support major distributions
- Arm software ecosystem
 - Working on the standard tools (Python/Java/LLVM) to provide higher performance on FUJITSU-MONAKA.
- ▶ Enabling smooth transition of customer assets and continuously enhancing performance

- Power efficient CPU for data-centers
- Contribute to the realization of carbon-neutral society
- Targeted for wide range of usage in the data-center including AI and HPC
- Will be shipped in 2027



FUJITSU-MONAKA



Comparison between A64FX and MONAKA

A64FX	FUJITSU-MONAKA
Armv8-A Architecture - SVE for HPC and AI	Armv9-A Architecture - SVE2 enhanced for HPC and AI - Confidential Computing
48 cores x 1 socket (48 cores per node)	144 cores x 2 sockets (288 cores per node)
Low voltage	Ultra low voltage
2.5D - CPU 7nm - HBM2	3D chiplet - Core die 2nm - SRAM die/IO die 5nm
HBM2 4 channels	DDR5 12 channels
PCI Express 3.0 Tofu Interconnect	PCI Express 6.0 (CXL3.0)
Air cooling and Liquid cooling	Air cooling

FUJITSU-MONAKA during SC23, Denver



Fujitsu's contributions to OSS Ecosystem and collaborations with oneAPI

Our commitment to open-source community

- We need to build cross-industry connected ecosystems, and co-create innovations that deliver environmental and social value.
- Fujitsu is committed to working with you on this journey, building a better future together.



Fujitsu's key contributions to OSS Community



2005

Linux kernel for
Mission Critical Server

2010

KVM and
Virtualization

2018

OpenStack,
Kubernetes

2022

Automotive Grade Linux,
Yocto,
Arm Linux on Supercomputer
Fugaku

A long history of collaborating with open-source communities, via open source development in mission-critical systems and in the supercomputer Fugaku

Fujitsu's key oneAPI contributions

oneAPI SPEC OPEN SOURCE COMMUNITY NEWS EVENTS RESOURCES

Developer Story: How We Ported oneDNN to Fugaku with Arm

NOVEMBER 1, 2021

[f](#) [in](#) [t](#) [✉](#)

Phase	Existing generic numeric library	oneDNN optimized for SVE-enhanced Armv8-A (This work)	Improvement
Training	9.3	85.6	x9.2
Inference	37.7	294.8	x7.8

Measurement conditions

- Framework: TensorFlow
- Benchmark: Resnet-50
- CPU: A64FX

oneAPI SPEC OPEN SOURCE COMMUNITY NEWS EVENTS RESOURCES

[←](#) BACK TO BLOG

Fujitsu and RIKEN Optimized oneDNN for Improved Performance on ARM

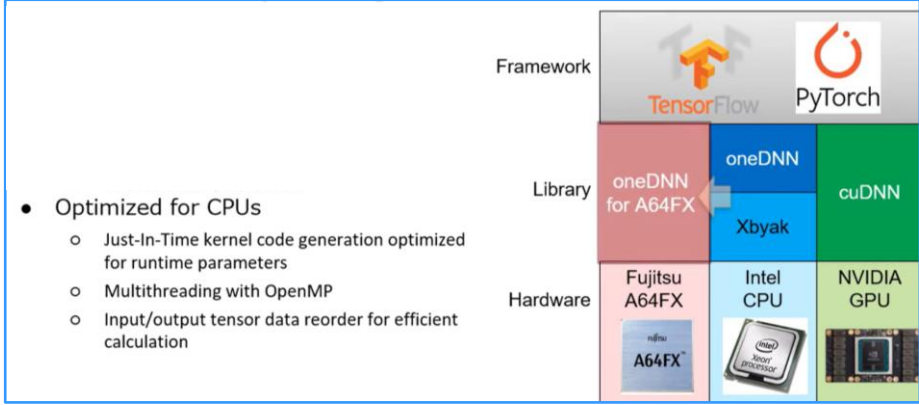
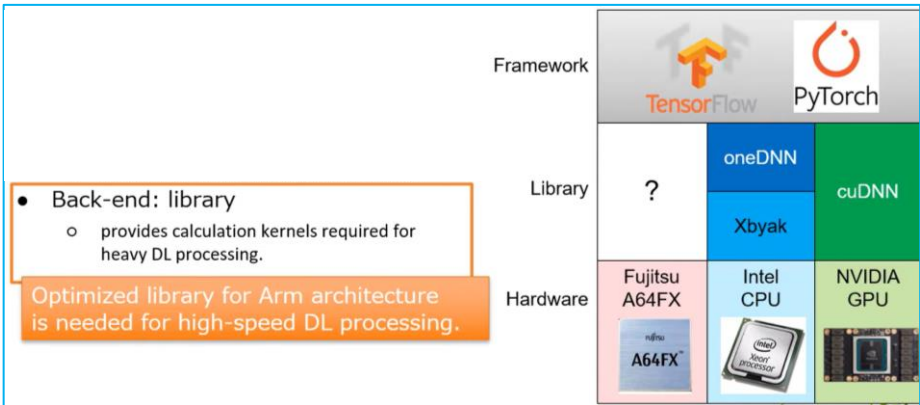
NOVEMBER 29, 2021

<https://www.oneapi.io/blog/fujitsu-and-riken-optimized-onednn-for-improved-performance-on-arm/>
<https://www.oneapi.io/blog/developer-story-how-we-ported-onednn-to-fugaku/>

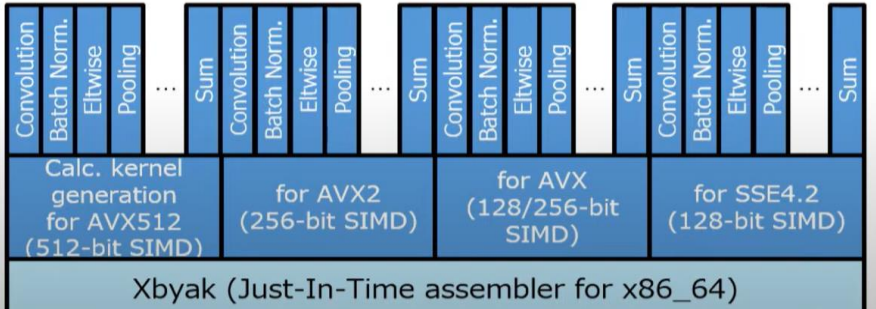


Software stack for Deep Learning processing in A64FX

[Kawakami - LinaroConnect 21]



Calculation kernels generation of oneDNN



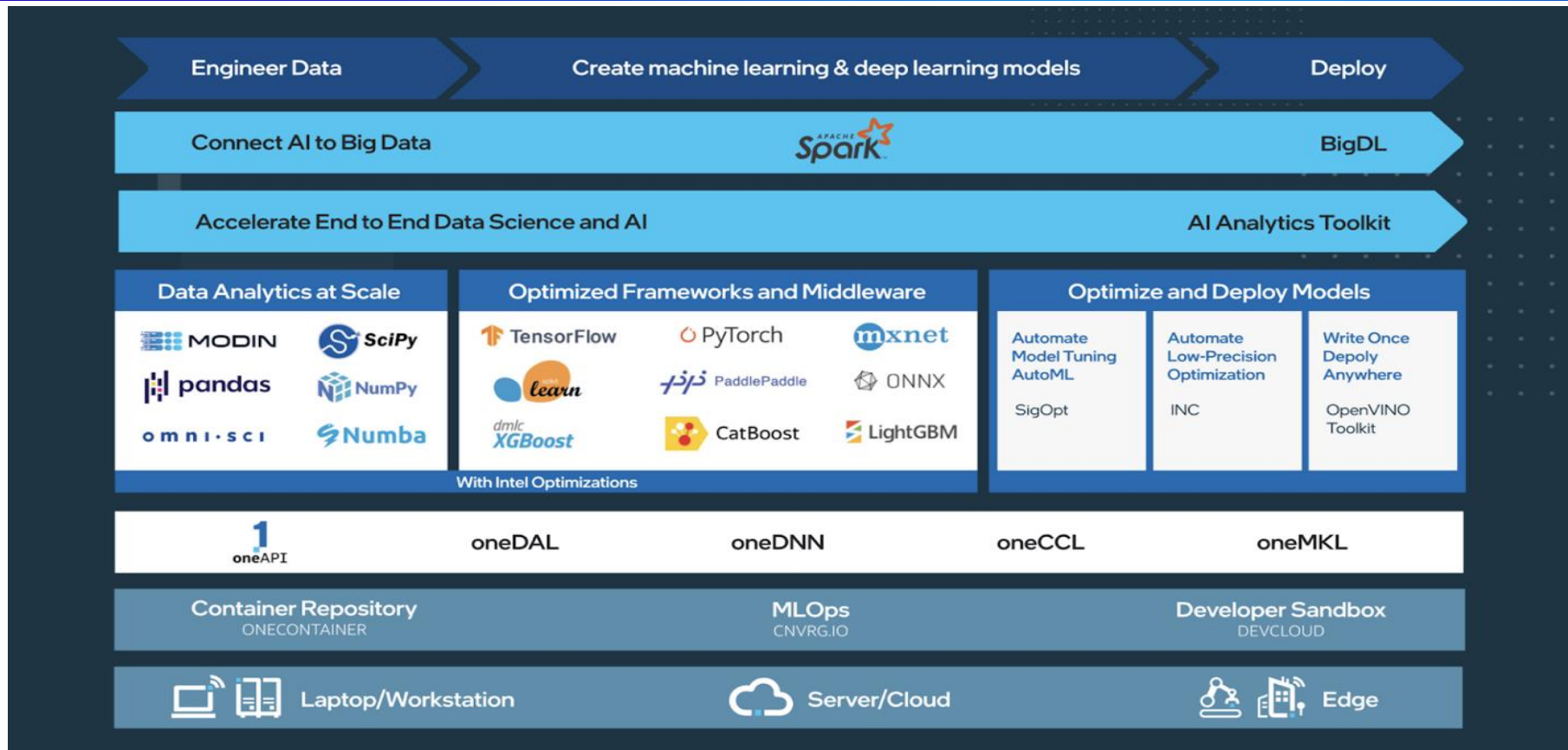
Motivation for porting oneDNN for A64FX

- oneDNN was ported through implementation at the instruction level using Xbyak JIT assembler for x86_64

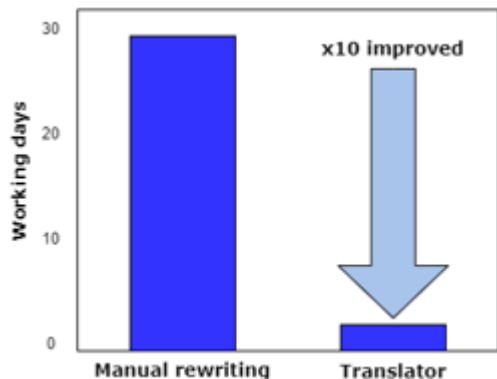
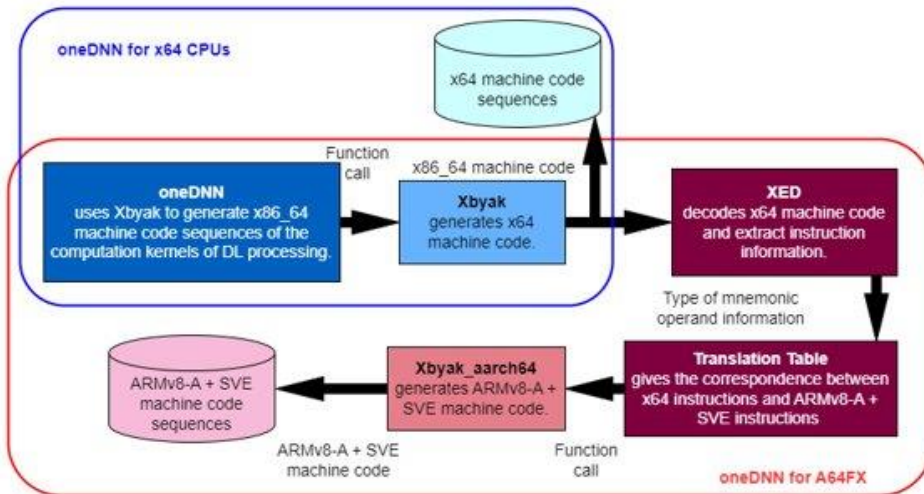
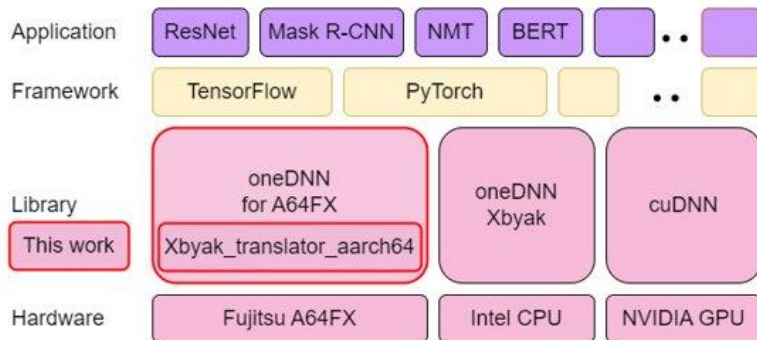
Ref: <https://www.youtube.com/watch?v=6Xn7ldLL160>

Credits: Kawakami San (Fujitsu)

Leveraging oneAPI Ecosystem



Role of oneDNN in DL SW stack with xbyak Translator

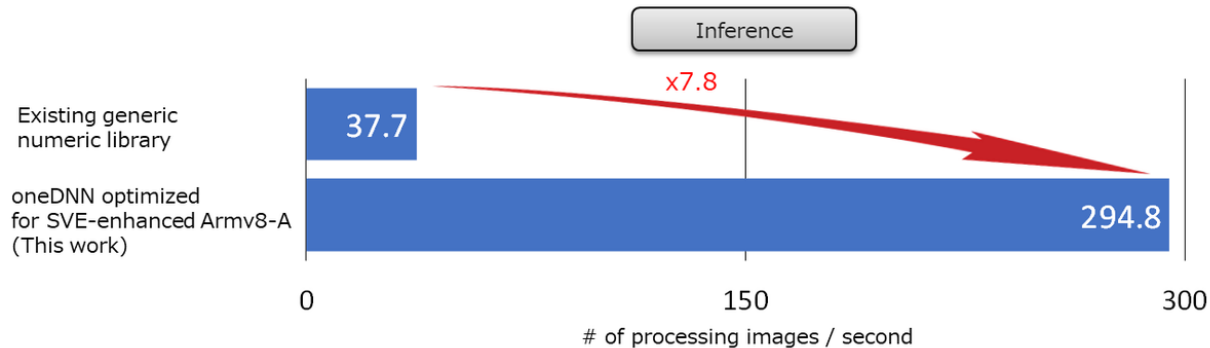
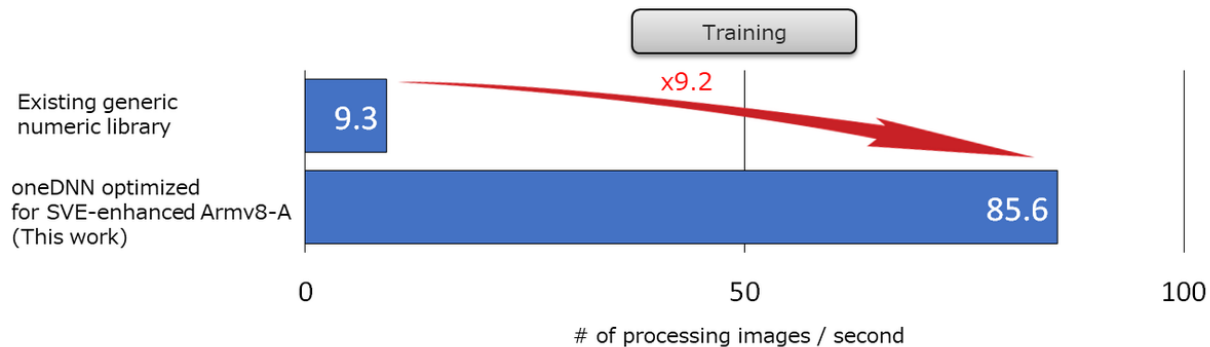


- It is a binary translator that at runtime converts oneDNN dynamically produced executable codes for the x86_64 architecture into executable codes for Armv8-A instruction set.



HPC results on oneDNN using JIT assembler Xbyak_aarch64

Improved processing speed for CPU-based DL



- Measurement conditions
- Framework: TensorFlow
 - Benchmark: Resnet-50
 - CPU: A64FX

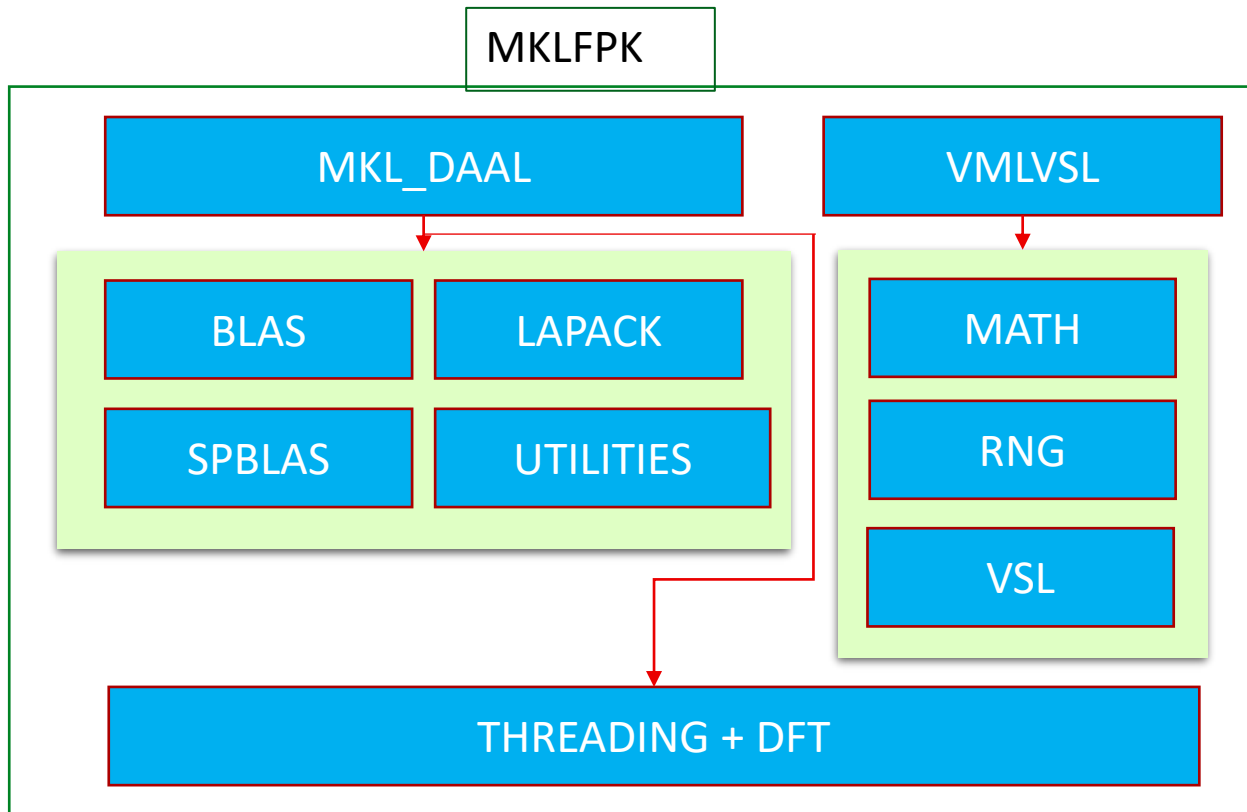
❑ With two software in our hands, Xbyak_aarch64 and Xbyak_translator_aarch64, Fujitsu did the porting of oneDNN for Armv8-A instruction set.

❑ The chart here shows the measured processing speed of Resnet-50 when TensorFlow was used as a framework software.

❑ Our oneDNN optimized for Armv8-A allows for a significant speedup of **9.2 times** in the training process and **7.8 times** in the inference process

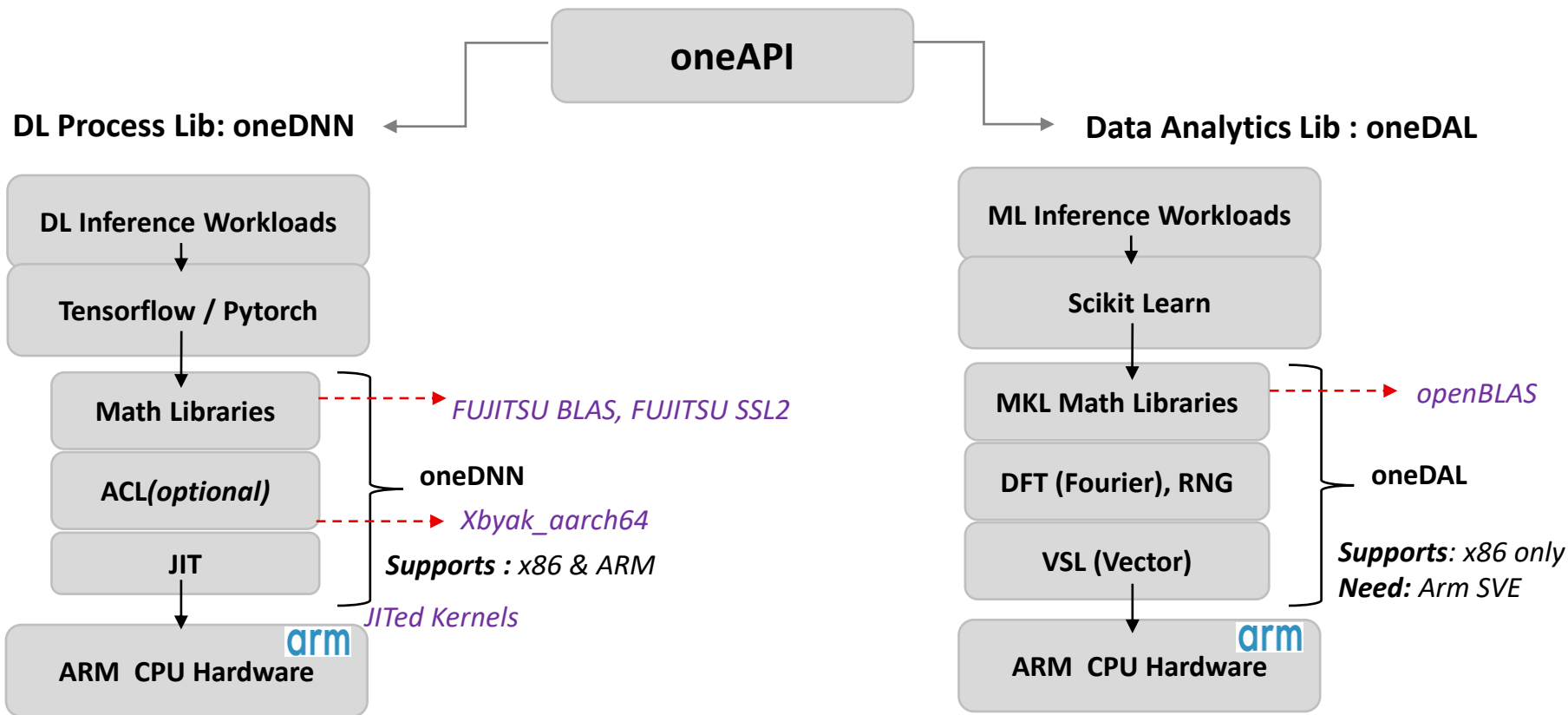
❑ **Contributed to oneDNN open source project**

oneAPI MKL for Optimized Math Functions



MKL Functions
BLAS (Basic Linear Algebra Subprograms)
LAPACK (Linear Algebra Package)
SPBLAS (Sparse BLAS)
RNG (Random Number Generator)
MATH
DFT
VSL (Vector Statistical Library)
UTILITIES
Atomic calls
ASM Checks
Compression Algorithms

Leveraging oneAPI for ARM ecosystem



Fujitsu's partnership with Unified Accelerator (UXL) Foundation



- Build a multi-architecture multi-vendor software ecosystem for all accelerators
- Unify the heterogeneous compute ecosystem around open standards
- Build on and expand open-source projects for accelerated computing

Steering Committee Members



The logo for ARM, consisting of the word 'arm' in a lowercase, blue, sans-serif font.



The logo for FUJITSU, featuring the word 'FUJITSU' in a red, uppercase, sans-serif font with a red infinity symbol above the 'J'.



The logo for Google Cloud, with 'Google' in its multi-colored font and 'Cloud' in a grey, sans-serif font.



The logo for Imagination, featuring a stylized 'i' icon followed by the word 'Imagination' in a black, sans-serif font.



The logo for intel, with the word 'intel' in a lowercase, black, sans-serif font.



The logo for Qualcomm, with the word 'Qualcomm' in a blue, sans-serif font.



The logo for SAMSUNG, with the word 'SAMSUNG' in a bold, blue, uppercase, sans-serif font.



The logo for vmware, featuring a stylized 'v' icon composed of two overlapping squares (one orange, one blue) above the word 'vmware' in a blue, lowercase, sans-serif font.



The founding companies are seeding the project with highly valuable contributions to open source libraries



Working Groups

Specification – defining an open standard for accelerated libraries

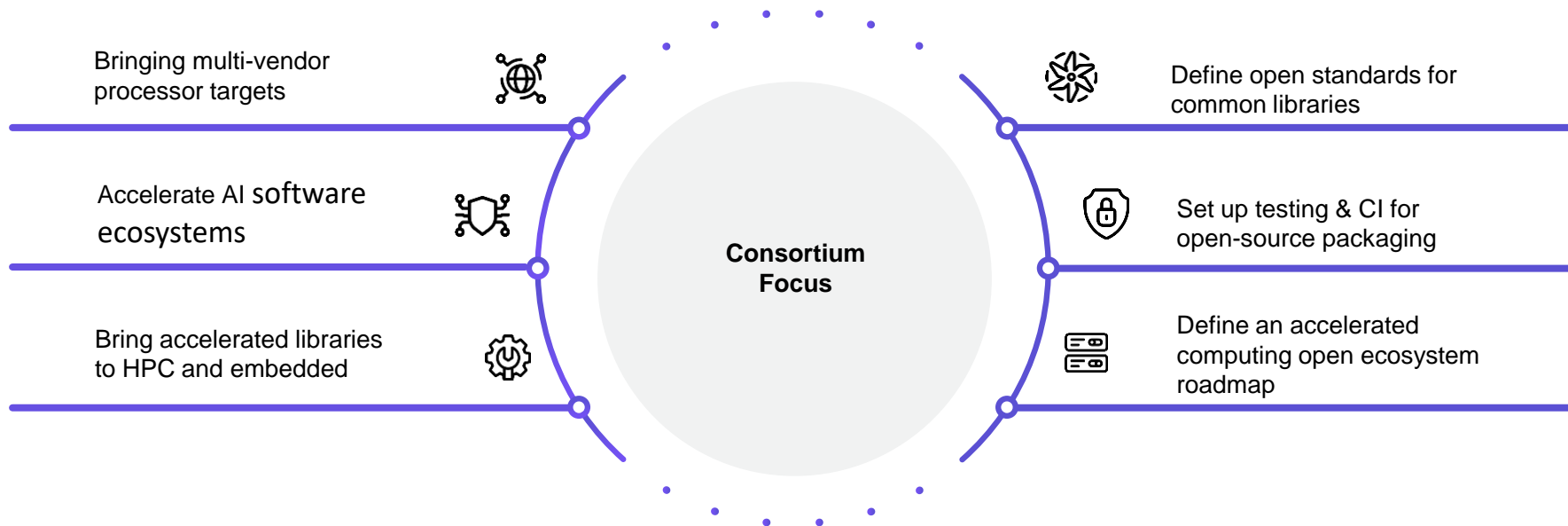
Open Source – delivering value to the community through collaboration



The group will work to drive the development of an open ecosystem for accelerated computing based on the fundamentals of open standards and open source.

Project governed by the Joint Development Foundation (JDF), a part of the Linux Foundation

Open specifications, APIs, open source for AI and HPC, Edge Compute and Edge AI



Join us: www.UXLFoundation.org

The background image shows a modern office environment. A man in a red t-shirt is leaning over a desk, holding a laptop and high-fiving a woman in a black t-shirt who is sitting at the desk. Another man in a white t-shirt is sitting at the desk behind them, smiling. The desk is cluttered with various items like a water bottle, a mug, and papers. The office has large windows in the background, a clock on the wall, and a whiteboard with sticky notes.

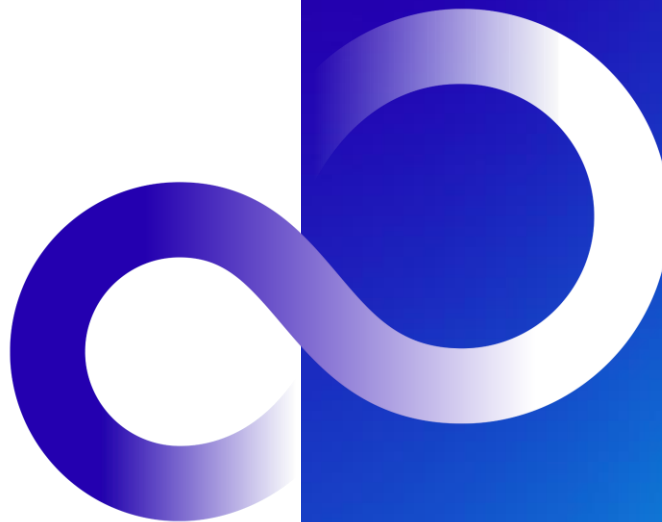
Conclusion

- ❑ Fujitsu develops high performance and energy-efficient processor called FUJITSU-MONAKA using our own microarchitecture and innovative 3D many-core architecture
- ❑ We continue and expand software development with communities and our partners for easy-to-use
- ❑ This processor will meet future computing demand of performance, power, reliability and usability for wide range of usage in the data-center including AI and HPC
- ❑ We will contribute to the realization of carbon-neutral society by our computing technologies and collaboration with users and partners

Acknowledgement

This presentation is based on results obtained from a project, JPNP21029 subsidized by the New energy and Industrial Technology Development Organization (NEDO)

Q&A



Thank you

