FUJITSU-MONAKA: Powering AI Workloads with oneDNN, oneDAL, and OpenBLAS

SC24 Conference, Atlanta, USA

Priyanka Sharma, PhD

Director - Software Engineering & Head MONAKA SW R&D Unit

Fujitsu Research of India

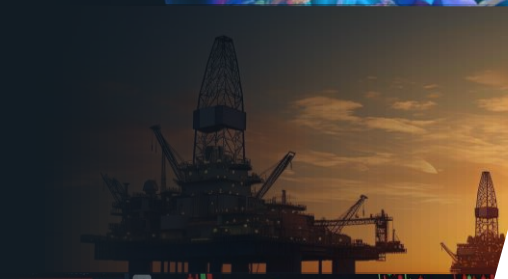
© RIKEN

Drug Discovery & Pharmaceuticals

Compute-intensive AI-based simulations of molecular dynamics and protein folding are accelerated using HPC clusters to enable drug discovery and vaccine development.

Energy (Oil, Gas & Renewable)

Using CPU-based HPC clusters, vast amounts of geological and seismic data is processed using Machine Learning to locate potential drilling sites for oil and natural gas.

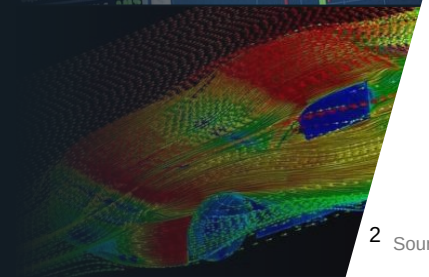


Banking, Finance & Investing

Being one of the largest adopters of AI, the BFI industry uses HPC for risk modelling, portfolio optimization and real-time market forecasting and trading.

Engineering & Manufacturing

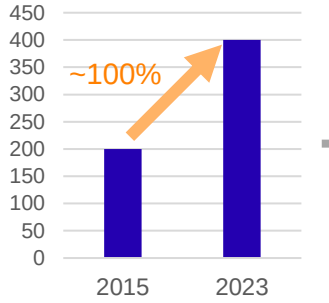
Product design, failure analysis, FEA and CFD simulations are few of many applications accelerated by HPC for the manufacturing industry.



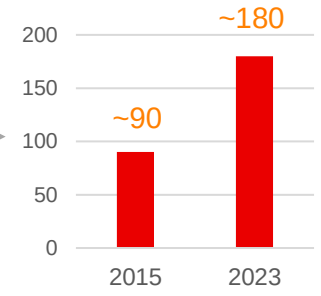
FUJITSU PUBLIC

HPC infrastructure is a key driver of the AI revolution.

With an estimated market size of $130 billion[1], developments in AI and HPC hardware, the increased workload demand for datacenters has increased power consumption by almost 100% since 2015, resulting in 2x increase in CO_2 emissions.



Estimated power consumption of datacenters in TWh in 2015 and 2023 | Source: [2]

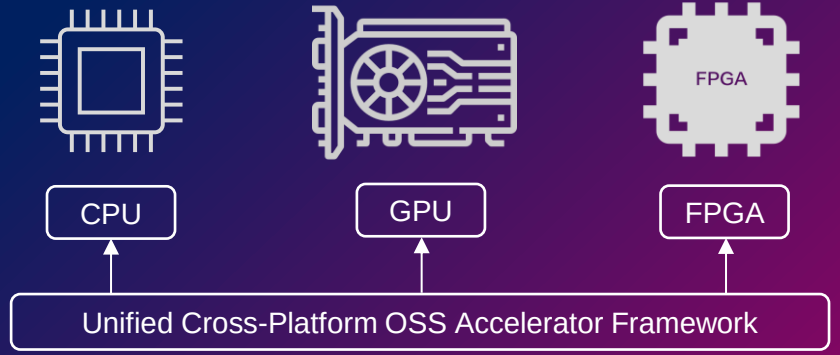


Estimated CO_2 emission due to datacenters in million metric tonnes in 2015 and 2023 | Source [3]

Sources: [1], [2], [3]

Enabling HPC workloads on an Open-Source Software Stack in a platform-agnostic manner is imperative for sustainable digital transformation!

Re-engineering software to adapt to each hardware platform is restrictive. Open-source and platform-agnostic software design enables interoperability on various hardware platforms, creating a more flexible developer ecosystem.



Trust and Transparency

Full visibility and open scrutiny of the software enhances trust in its quality, functionality and robustness.

Collaboration and Innovation

Developers from around the globe can contribute new features to the software, rapidly improving its capabilities.

Increased Flexibility and Access

Platform-agnostic design enables a wider audience to use the software, increasing access and flexibility for developers.

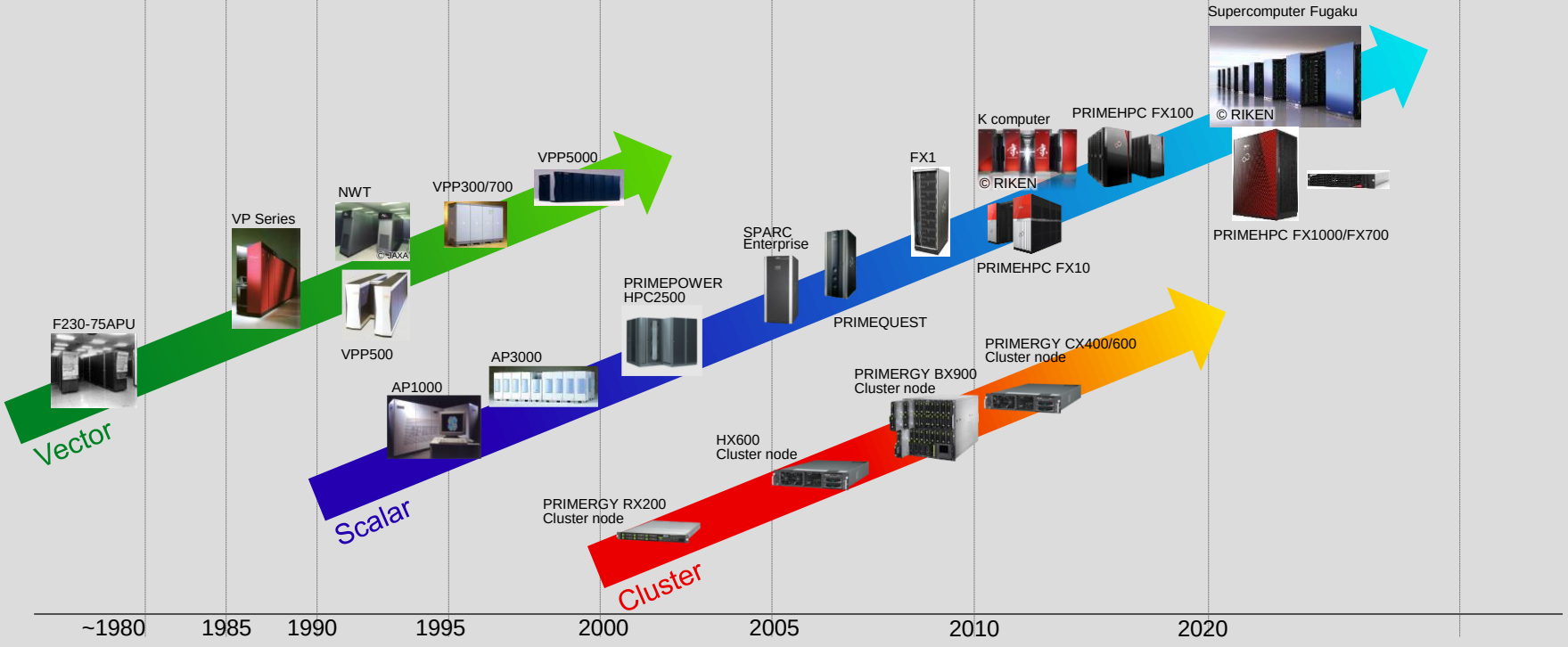
Standardization and Security

OSS enforces strict software standards, which ensures ethical and uniformized software development.

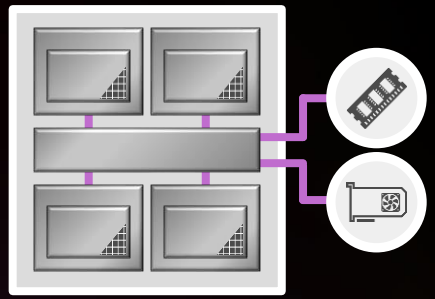
Fujitsu's Supercomputers Released to World

FUJITSU

Fujitsu has continuously developed and delivered world-class supercomputers



FUJITSU-MONAKA



- **Armv9-A Architecture**
- **Arm SVE2 for AI and HPC**
- **3D chiplet**
  - Core die — 2nm
  - SRAM die/IO die — 5nm
- **144 cores x 2 sockets (288 cores per node)**
- **Ultra low voltage for energy-efficiency**
- **Confidential Computing for security**
- **DDR5 12 channels**
- **PCI Express 6.0 (CXL3.0)**
- **Air cooling**

To be shipped in 2027

FUJITSU PUBLIC

FUJITSU

Next-generation high-performance, energy-efficient, Japan-made processor for a carbon neutral digital society

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 CO_2 emissions and power costs with high energy efficiency (2x more efficient than competing CPUs)

Goals

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

Achieved through our proprietary technologies such as self-designed microarchitecture and ultra low-voltage technology

Covering a wide range of software stacks including AI frameworks and HPC capability

FUJITSU

Applications Delivery

Customer Use Cases	Fujitsu Computing as a Service	Fujitsu Kozuchi
• Surrogate Models • LLM Software Applications	• Scikit Learn Use Cases • Hugging Face Applications	• Causal Inference • Ambient Authentication

Open-Source Contributions

API Microservices Platform

OpenMathLib OpenBLAS | Math Library, NumPy, OpenMP | UXL foundation | oneDAL oneDNN | PyTorch | Arm SW stack | Linaro | Kubernetes and OpenStack OSS

Software Delivery

PyPi | Docker | Containers | Reference Implementations | Computing Workload Broker

CI/CD Platform

Collaborations

AI Software Frameworks

Machine Learning	Deep Learning	Big Data Analytics	Data Security
Scikit-Learn, Multithreading, XGBoost, NumPy, Pandas, OpenBLAS	LLM's, Vision, NLP, Hugging Face, TensorFlow/PyTorch, OpenVINO, oneDNN, Inductor	PostgreSQL, Spark, VectorDB, Data Intelligence	Red Hat, Secured HW/SW, Attestation Frameworks, OpenShift, Confidential Computing

Cutting Edge Applications

Healthcare & Pharma	Manufacturing	Retail	Banking & Finance
❑ Drug Discovery ❑ Gene Prediction ❑ Medical robotics	❑ Defect Detection ❑ Preventive Maintenance ❑ Prescriptive maintenance	❑ Recommendation ❑ SCM Forecasting ❑ Customer experience	❑ HF Trading ❑ Fraud Detection ❑ Risk Management

Partnership with Linux Foundation's - UXL Foundation

FUJITSU

UXL
Unified Acceleration 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

arm
BROADCOM
FUJITSU
Google Cloud
GE HealthCare
Imagination
intel
SAMSUNG
Qualcomm

Image credit: UXL Foundation: Unified Acceleration

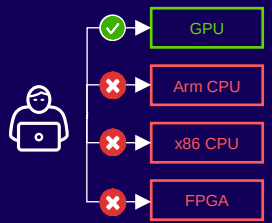
Development of an

Open Unified Accelerator Ecosystem

The Unified Acceleration (UXL) Foundation is a consortium of organizations promoting the adoption of unified acceleration: enabling the use of various accelerators with a single code. As a steering member of UXL, we aim to:

- Build a multi-arch, 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.

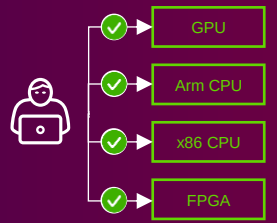
Without Unified Acceleration



- Re-write code for each hardware
- High migration costs
- Reduced maintainability

With Unified Acceleration

- Same code works optimally on all hardware
- Reduced migration costs
- Flexibility to choose optimal hardware

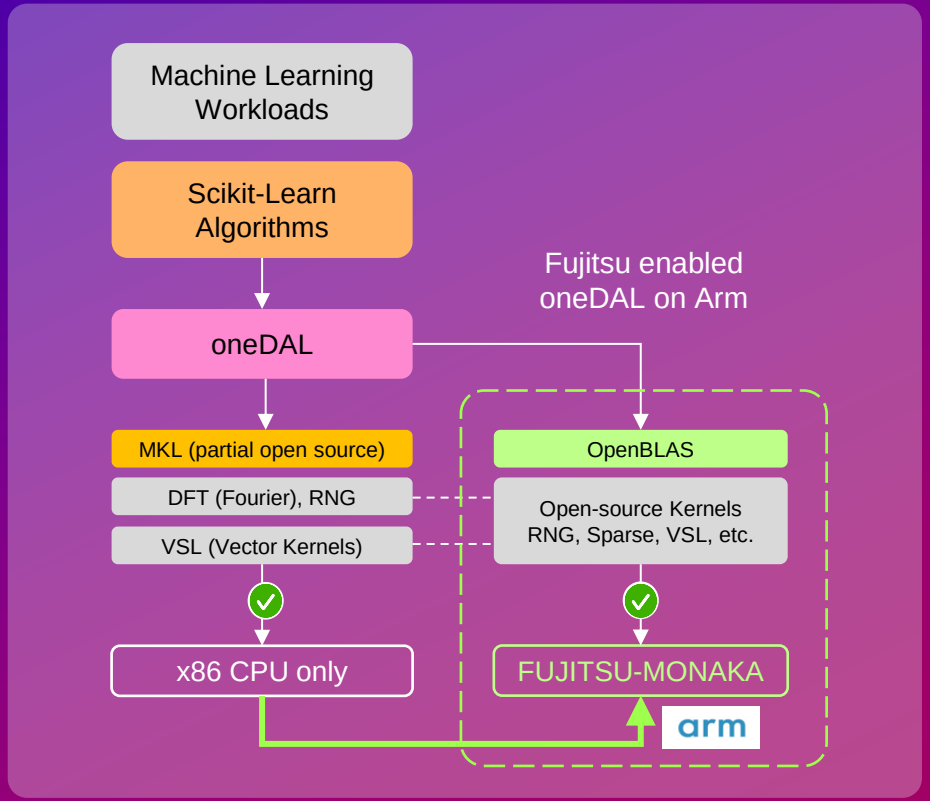


oneDAL: Arm Porting Design

Historically, oneAPI Data Analytics Library (oneDAL) could only be compiled on x86 architecture due to Math Kernel Library (MKL) binary-only backend.

To accelerate ML workloads on Arm, FUJITSU replaced MKL with open-source function calls, which resulted in oneDAL enablement on Arm.

oneDAL enablement for Arm by FUJITSU was one of the first contributions to **UXL Foundation.**



oneDAL: Open-Source Statistical Frameworks

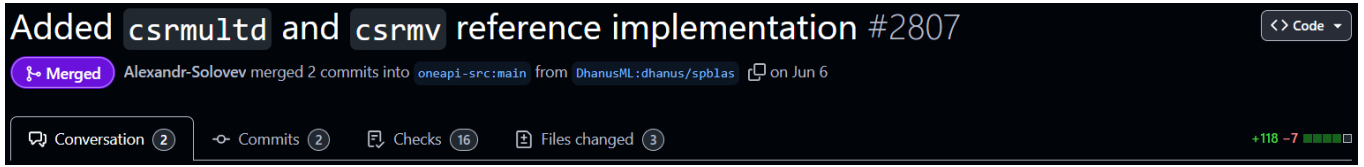
FUJITSU

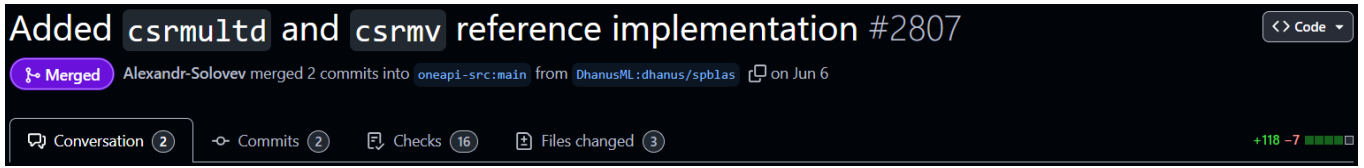
➢ Fujitsu's Recent Contributions with Statistical Kernel Development

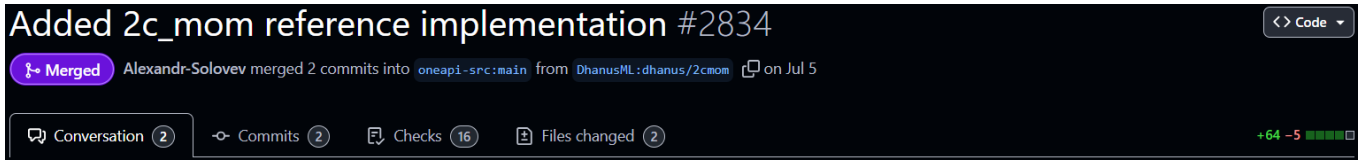
Sparse BLAS Kernel Implementations to compute covariance of CSR format data, enabled oneDAL ML algorithms.

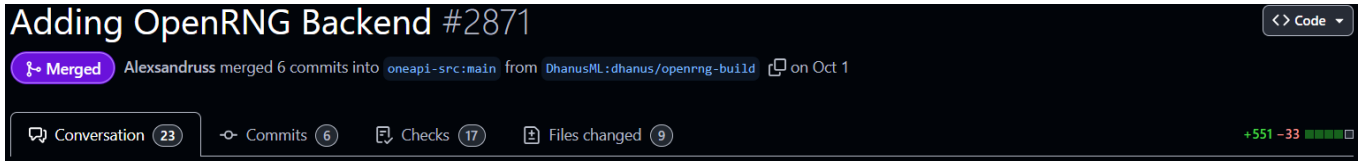
Vector Statistical Library (VSL) Kernel Implementations to compute variance and matrix of cross products, enabled oneDAL ML algorithms.

OpenRNG OSS (developed by ARM) Backend Integration in oneDAL. Conforms to MKL VSL RNG specification







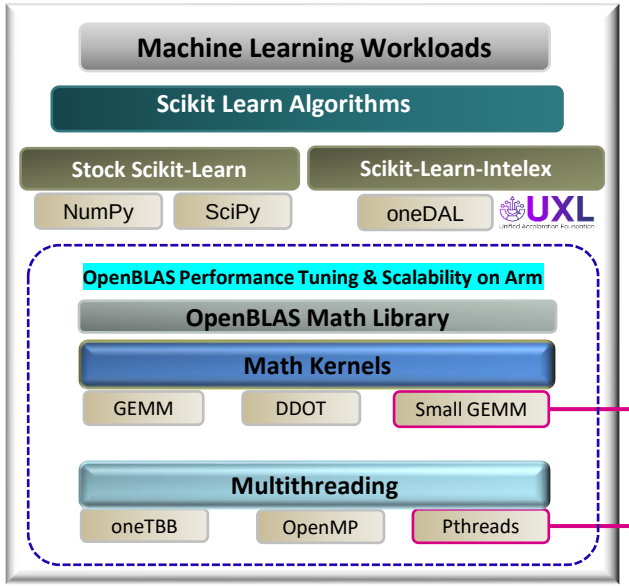


Fujitsu team presented technical session on this work during UXL Dev Summit 2024

oneDAL: Compute Library OpenBLAS Optimizations

FUJITSU

Machine Learning Workload Optimization

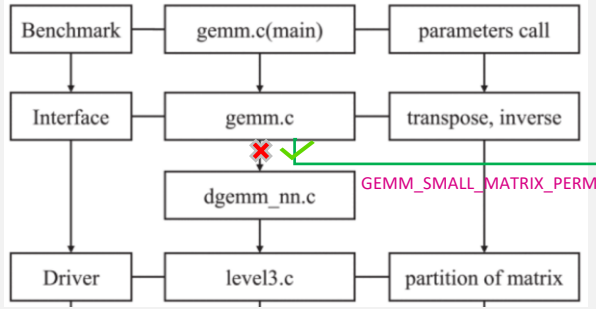


UXL Collaboration Areas:

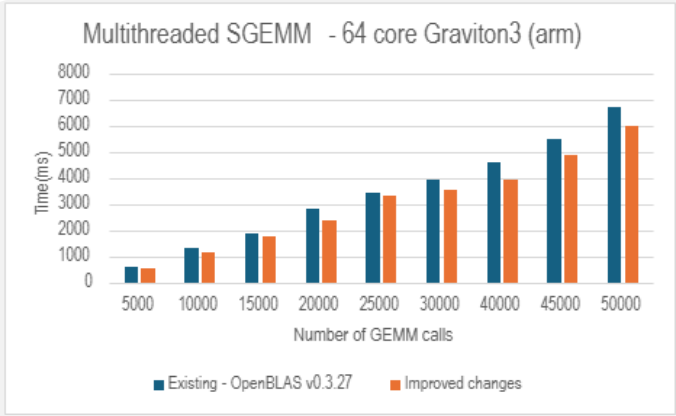
- **oneMKL or SYCL** alternatives for OpenBLAS Math Routines with High Performance on CPUs.
- **oneTBB** threading backend alternative for OpenBLAS

Optimized OpenBLAS DGEMM kernel for Small Matrix on ARM, now enabled in permit route

OpenBLAS PR # **4887**



Fujitsu's Contribution in collaboration with ARM, results in speedup of ~25%



Fujitsu's Contribution results in speedup of ~15% with better core utilization

This research work accepted at 31st IEEE International Conference on HPC, Data, and Analytics (HiPC Conference 2024, Bengaluru)

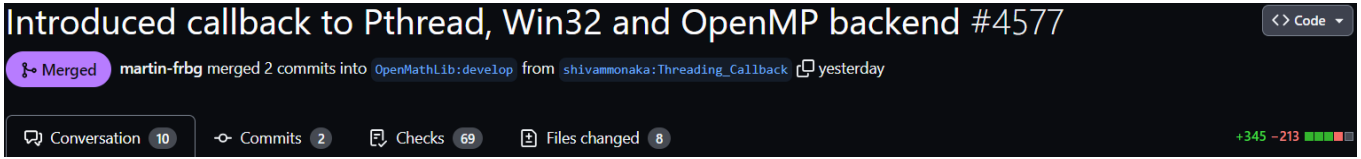
OpenBLAS PR # **4741**

Enhance core utilization in OpenBLAS with default threading backend pthreads

oneTBB: OpenBLAS Multithreading Implementation

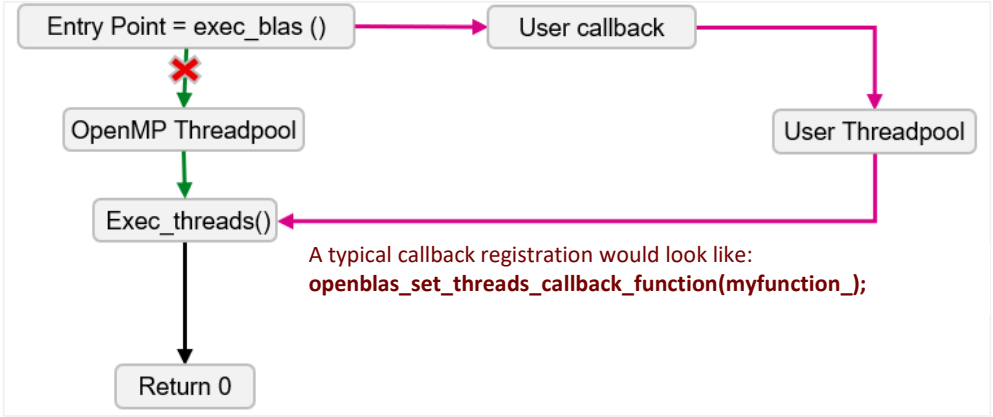
FUJITSU

Enable OpenBLAS workflow to be **composable with caller multithreading** backend e.g. oneTBB gains performance. Also supports OpenMP, pthreads and win32



UXL Collaboration Area: oneTBB integration tuning with OpenBLAS for an efficient threading backend alternative

➢ Solution Design

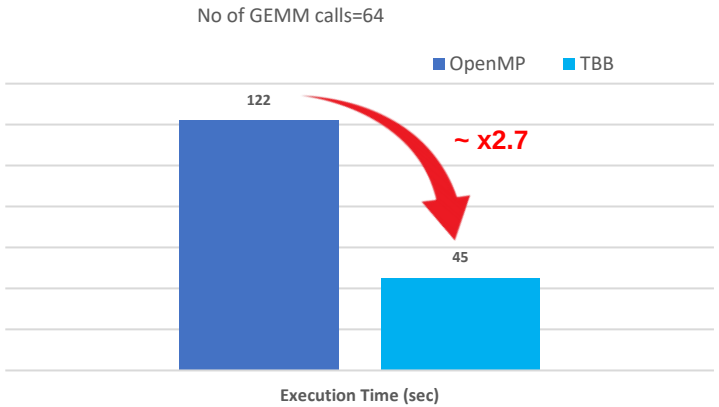


A typical callback registration would look like:
openblas_set_threads_callback_function(myfunction_);

➢ Performance Result

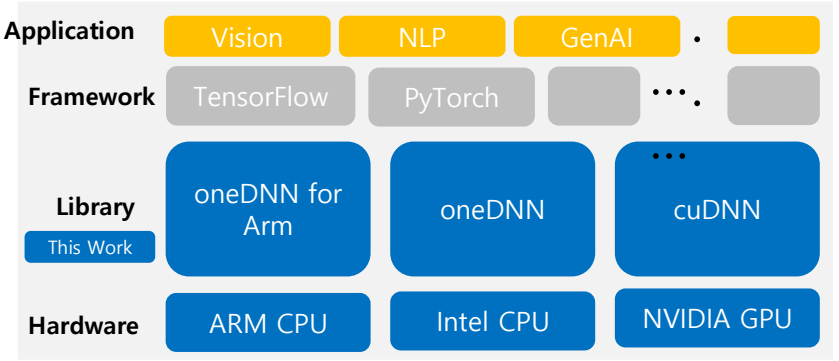
More work in Progress & Fujitsu is interested to Collaborate

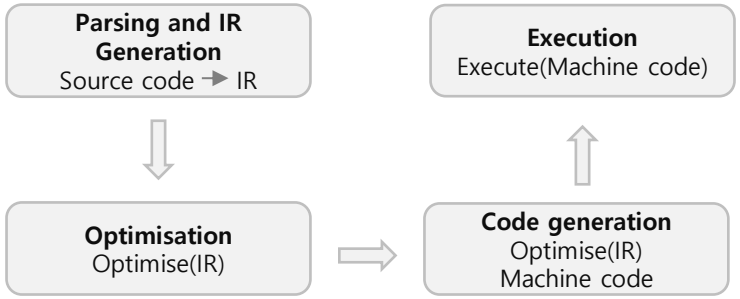
Upto 2.7x speedup with TBB v/s OpenMP in nested parallelism scenario



Fujitsu's contributions to oneDNN through JIT kernels

- To accelerate deep learning (DL) processes on Arm CPU's (for HPC workloads), oneDNN was ported and optimized by Fujitsu [Kawakami – San, Linaro Connect 2021]
- oneDNN is an open-source DL processing library for cross-platform architecture. oneDNN dynamically creates the execution code for the computation kernels, which are implemented at the architecture level granularity using Xbyak, the Just-In-Time (JIT) assembler.
- Just-In-Time (JIT) compilation is a technique used in compiler design where the compiler translates source code into machine code at runtime, rather than ahead of time (AOT) as in traditional compilation.
- **Major Advantages of JITed implementation:**





oneDNN: Contributions – BRGEMM Kernels

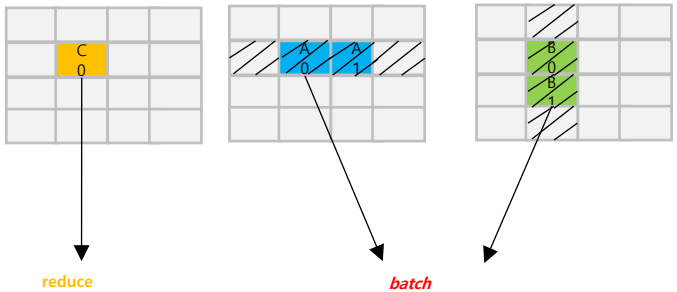
FUJITSU

BRGEMM MatMul Enablement PR (#1818) is merged, expands Arm SVE support for matrix mul. & adds BRGEMM folder for aarch64

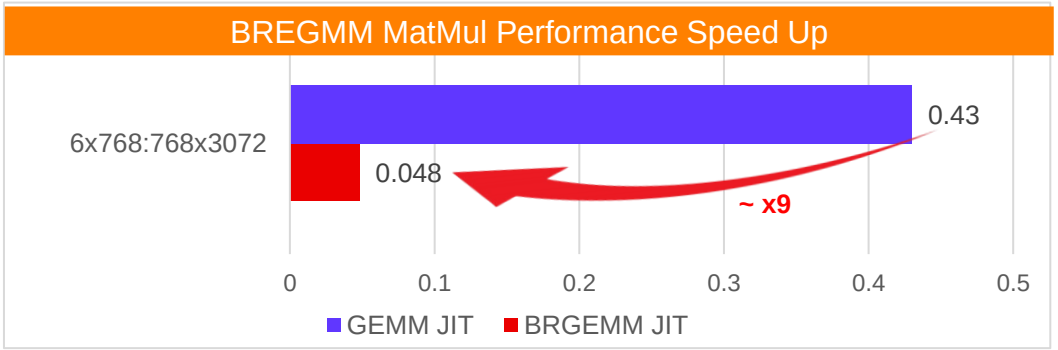
cpu: aarch64: Expand ARM SVE support for matrix multiplication #1818

Merged igorsafo merged 12 commits into oneapi-src:main from vineelabhinav:feature-sve-matmul 2 weeks ago

~9x performance gain observed as compared to oneDNN GEMM JIT Implementation



Sequence of input tensor blocks: [A0,B0], [A1,B1]
Output tensor block: C0



- **What is (Batch Reduced General Matrix Multiplication) BRGEMM ?**
  - Tensors are reduced into batches for multiplications
  - Broadcast the input matrix B values
  - Perform fused-multiply-add instruction (fma instruction) at once for multiple values

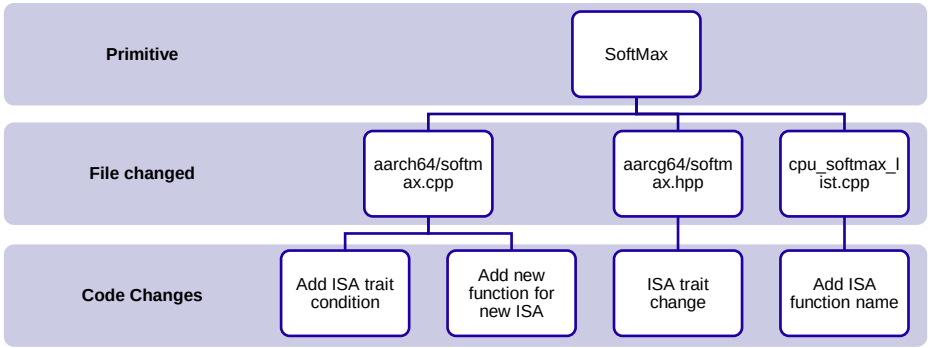
oneDNN: Contributions – Softmax Kernel

FUJITSU

oneDNN Pooling JIT Kernel PR (#1786) is merged, expands support of SoftMax for multiple ISA

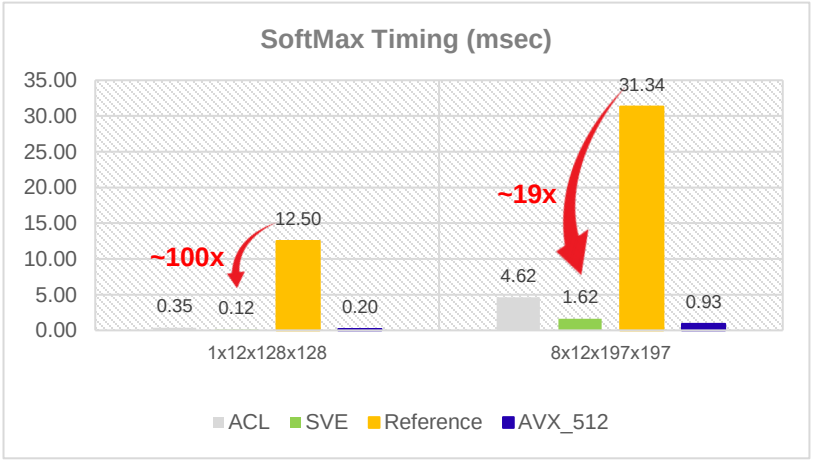
~100x performance gain observed as compared to oneDNN Reference implementation

➢ Code Changes



Conditional Statement Modification : Added OR condition to use multiple ISA
Added new function for supporting SVE in different vector length.

➢ Performance Result



oneDNN: Contributions – Pooling Kernel

FUJITSU

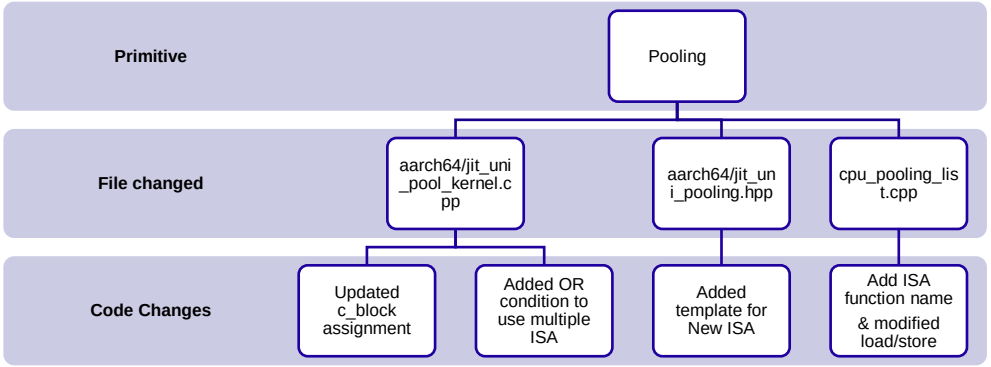
oneDNN Pooling JIT Kernel PR (#1850) is merged, expands support of Pooling for multiple ISA

cpu: aarch64: Expand ARM SVE support in jit_uni_pool_kernel #1850

Merged igorsafo merged 2 commits into `oneapi-src:main` from `vishwascm:aarch64-sve-jit-pooling` 2 weeks ago

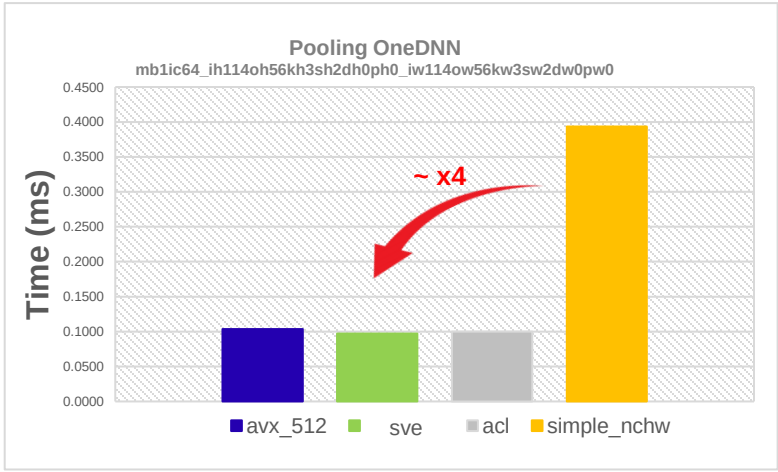
~4x performance gain observed as compared to current implementation in oneDNN

➢ Code Changes



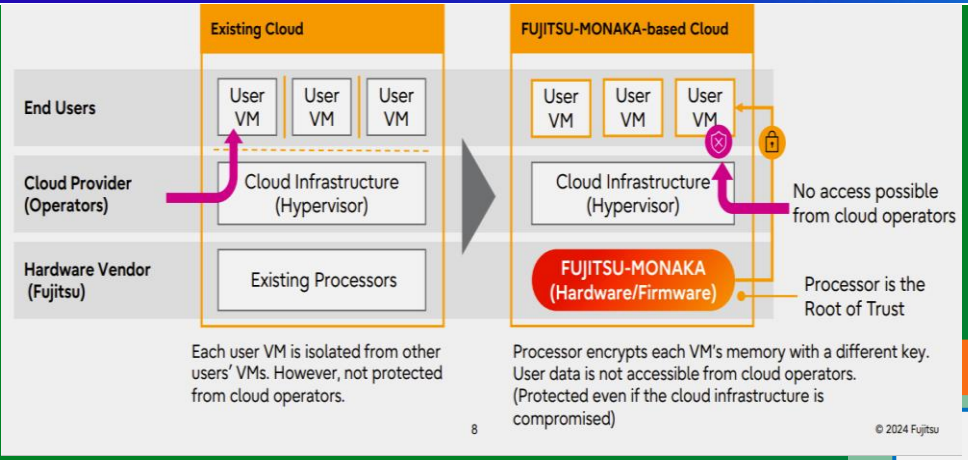
Conditional Statement Modification : Added OR condition to use multiple ISA. Modified load and store instructions to use predicate registers for correct ISA matching.

➢ Performance Result

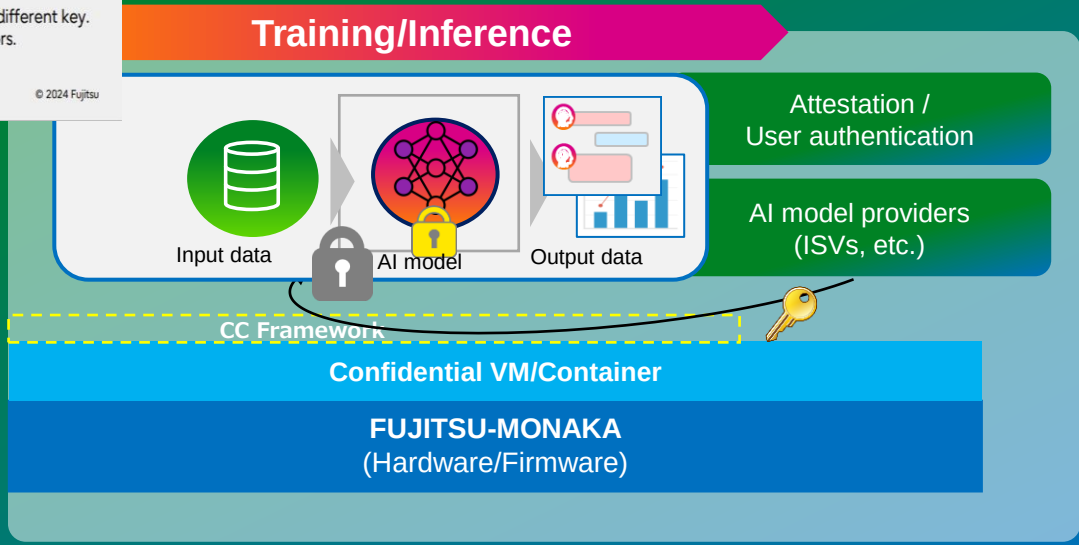


ARM CCA for Confidential Computing

FUJITSU

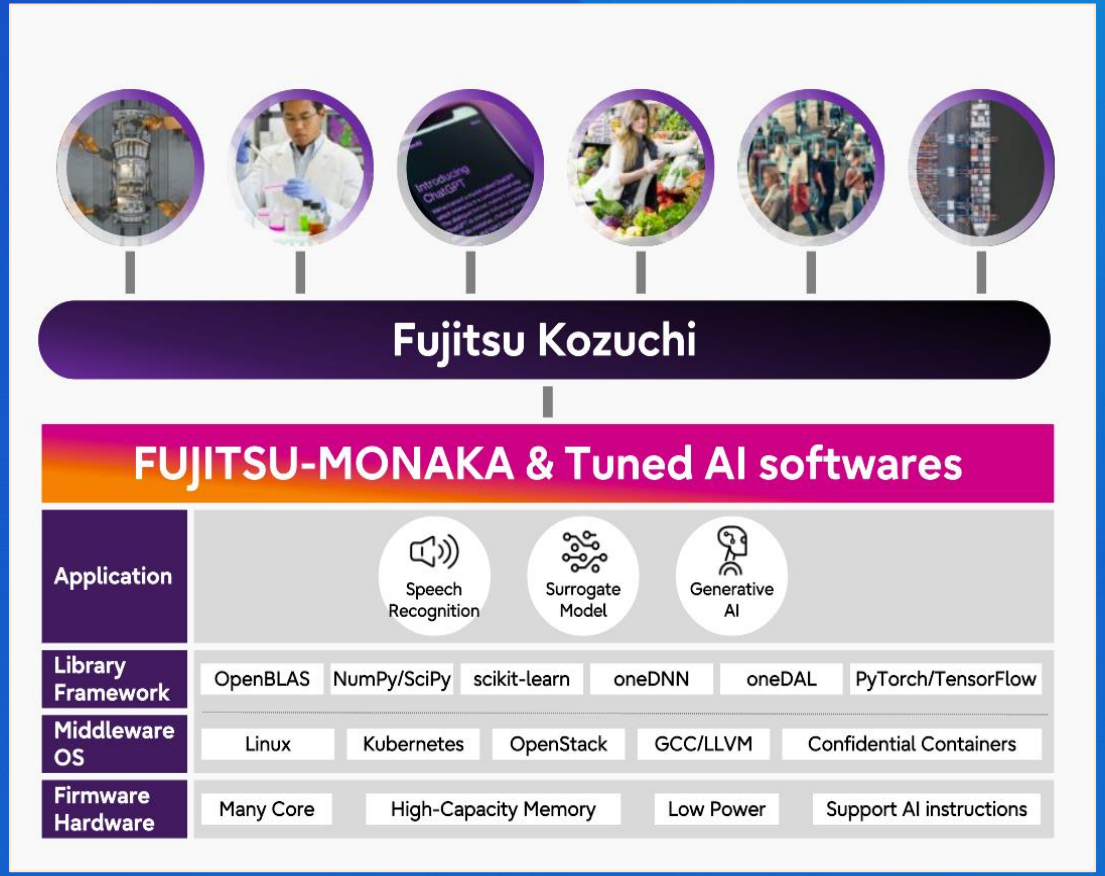


Protect end-user data in memory by encrypting every VM with a different key generated by the processor hardware and firmware



Expected to be an essential technology in cloud, edge and HPC environments which deals with sensitive data

FUJITSU-MONAKA will solve customer issues as an AI infrastructure platform that can be utilized in a wide range of domains



Acknowledgement

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

Thank you

Together we need to work towards democratizing the use of AI!