oneDAL Arm SVE Enablement for Accelerated AI-ML Computing with FUJITSU-MONAKA

UXL Foundation AI SIG

14th March 2024

Chandan Sharma

Software Engineer,

MONAKA SW R&D (HPC AI) Unit,

Fujitsu Research of India

© RIKEN

Outline of presentation

- Fujitsu’s presence in OSS community and FUJITSU-MONAKA
- Design and Methodology of oneDAL Arm Porting Contribution
- Performance Results obtained with oneDAL on Arm SVE
- oneDAL Multi Architecture Collaboration and OSS Development
- Concluding Remarks, Resources, and Acknowledgment

Partnership with Unified Accelerator (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	FUJITSU	Google Cloud
Imagination	intel.	Qualcomm
SAMSUNG	vmware® by Broadcom	

Fujitsu Arm Processor "FUJITSU-MONAKA"

- Creating a new era of computing power is mandatory for the future society with massive data generation and processing
- Ever-increasing power in datacenters is critical, and the power efficiency in CPU (consists of 60%) would be the vital factor for a sustainable future
- Fujitsu shall utilize its Supercomputer success and technology for the solution

FUJITSU-MONAKA

- **Developing the new power efficient CPU “FUJITSU-MONAKA“ for datacenters, which will be shipped in 2027**
- **Targeted for wide range of usage in the datacenter including AI and HPC, and contribute to the realization of carbon-neutral society**

Software Ecosystem for AI & HPC Computing

FUJITSU

Focus on Application Performance

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

Focus on Ease-of-Use & Open-Source Software

Fujitsu

UXL Unified Acceleration Foundation

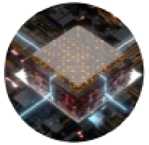
arm

Feature Development

Linaro

Quality Enhancement

Technical Support



FUJITSU-MONAKA

Open Source Community

Linux Kernel | QEMU | LLVM | Python

oneDNN | Kubernetes | OpenStack

oneDAL

Software Distributors Linux OS/System SW

Customers

Fujitsu-Public

Fujitsu’s key contributions to OSS Community

2005
Linux kernel for Mission Critical Server

2010
KVM and Virtualization

2018
OpenStack, Kubernetes

2021
Ported oneDNN to Arm

2022
Automotive Grade Linux, Yocto, Arm Linux on Supercomputer Fugaku

2024
Ported oneDAL to Arm

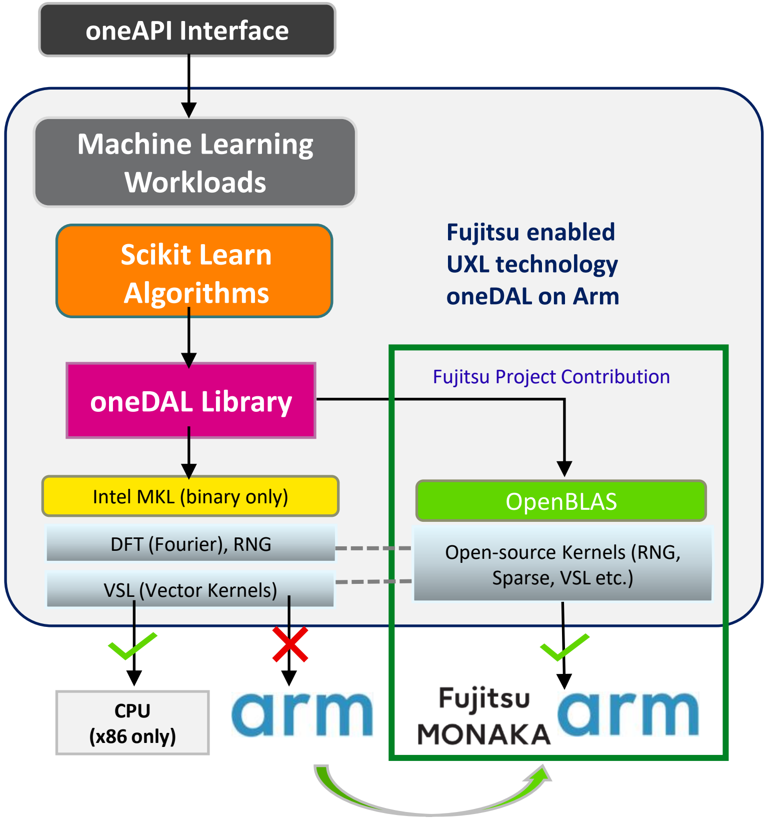
A long history of collaborating with open-source communities, via open-source development in mission-critical systems and in the supercomputer Fugaku, **continuing the legacy for FUJITSU-MONAKA**

oneDAL Porting Design for Arm

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

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

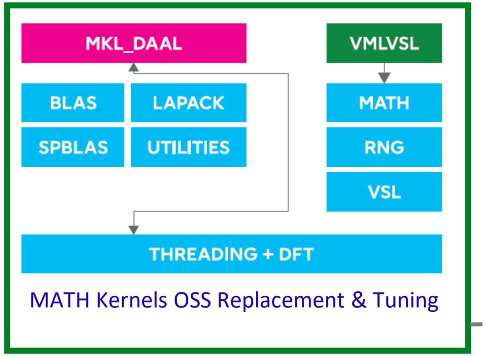
It is one of the first open-source contributions to UXL Foundation.

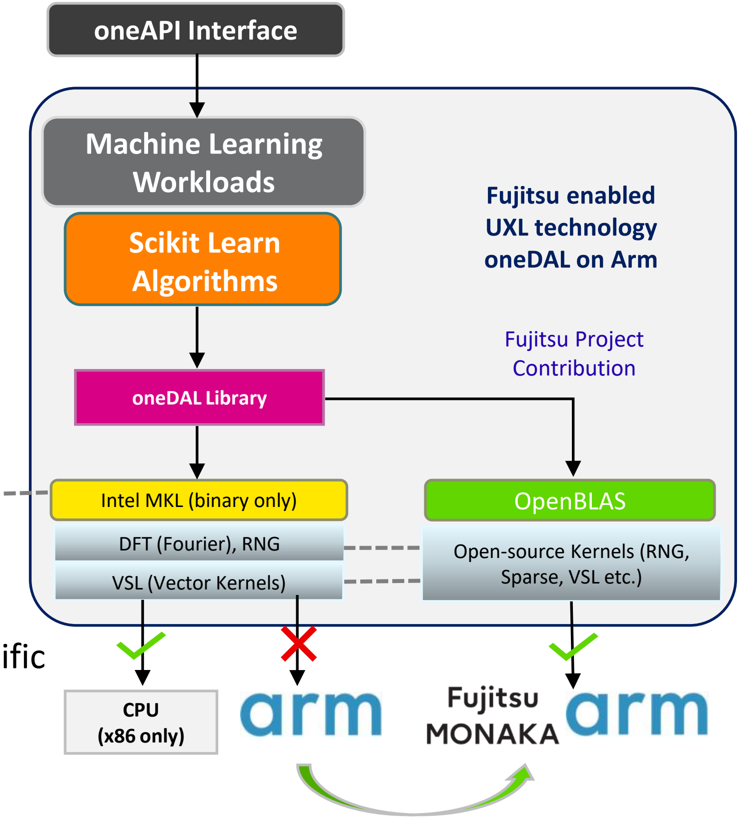


Porting Methodology with Arm SVE

- oneDAL on x86 uses MKLFPK, with functionalities

- To support these functions on Arm, open-source optimised compute kernels from OpenBLAS are used as alternatives to leverage SVE on the Arm



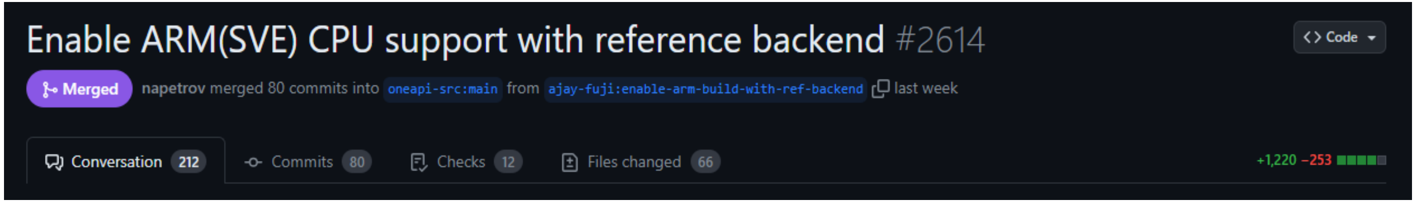


- Used reference backend & added compiler options to makefiles

- Added compiler macros throughout the code base to isolate x86 specific code chunks and handle it with arm when possible.

Fujitsu Contribution to oneDAL Open Source

oneDAL PR (#2614) is merged, raised by **Fujitsu**, to enable multi architecture build, extensive UXL collaboration with Intel & Arm



Enable ARM(SVE) CPU support with reference backend #2614

napetrov merged 80 commits into `oneapi-src:main` from `ajay-fuji:enable-arm-build-with-ref-backend` last week

oneDAL Contribution and Collaboration

80 Number of commits contributed by Fujitsu with 1,220 lines of code

66 Number of files modified by this pull request to enable oneDAL on ARM

05 Meetings conducted between Fujitsu, Intel & Arm, got PR approval from 3 Intel reviewers

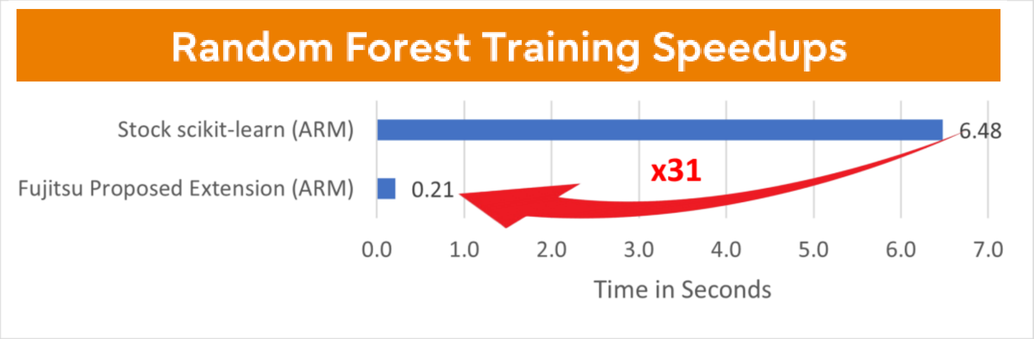
69 Number of days this pull request was OPEN and under review by UXL oneDAL team

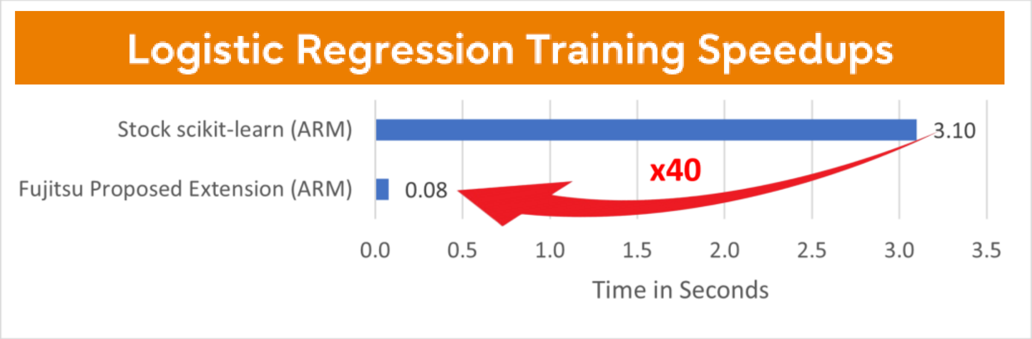
212 Number of GitHub conversations between reviewers and Fujitsu oneDAL team

oneDAL Arm SVE Performance Results

With SVE optimisation and oneDAL porting enhancements on ARM, our work showcases notable performance gains across multiple ML algorithms.

These graphs illustrate the training speedup of top two ML algorithms used by Fujitsu AutoML, which got a significant speedup of 31 times in Random Forest and 40 times in Logistic Regression.





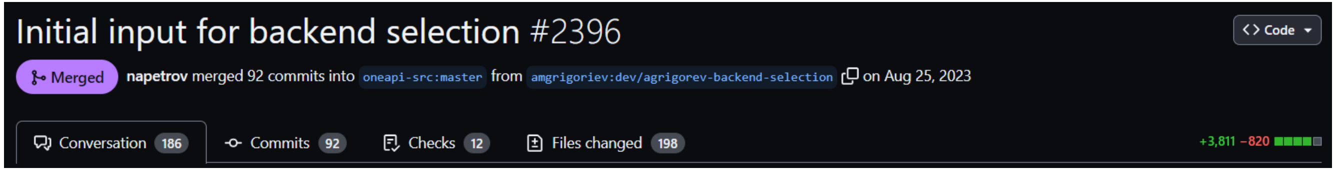
Results computed on AWS Graviton3 Arm-based CPU c7g.8xlarge 32-cores

oneDAL Multi Architecture Collaboration

FUJITSU

➢ Collaborative Active PRs

oneDAL PR (#2396) is merged, raised by **Intel** to support OpenBLAS on x86 with limitations



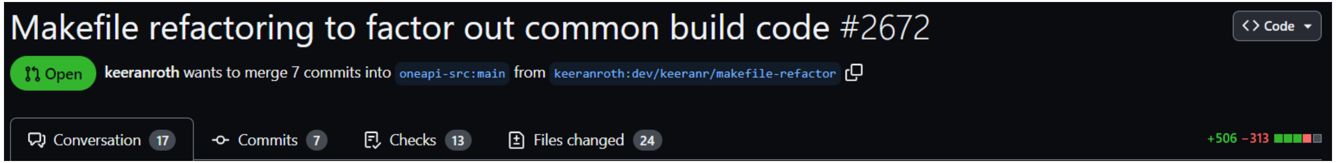
Initial input for backend selection #2396

<> Code

Merged **napetrov** merged 92 commits into `oneapi-src:master` from `amgrigoriev:dev/agrigorev-backend-selection` on Aug 25, 2023

Conversation 186 | Commits 92 | Checks 12 | Files changed 198 | +3,811 −820

oneDAL PR (#2672) is open, raised by **Arm** updating Makefile structure to ease future additions



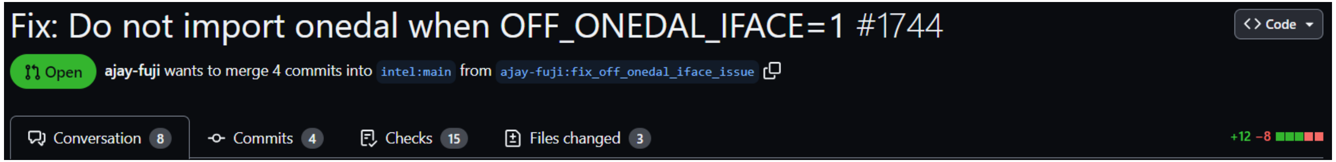
Makefile refactoring to factor out common build code #2672

<> Code

Open **keeranroth** wants to merge 7 commits into `oneapi-src:main` from `keeranroth:dev/keeranr/makefile-refactor`

Conversation 17 | Commits 7 | Checks 13 | Files changed 24 | +506 −313

Scikit Learn Intelex PR (#1744) is open, raised by **Fujitsu** to handle oneDAL usage in other packages



Fix: Do not import onedal when OFF_ONEDAL_IFACE=1 #1744

<> Code

Open **ajay-fuji** wants to merge 4 commits into `intel:main` from `ajay-fuji:fix_off_onedal_iface_issue`

Conversation 8 | Commits 4 | Checks 15 | Files changed 3 | +12 −8

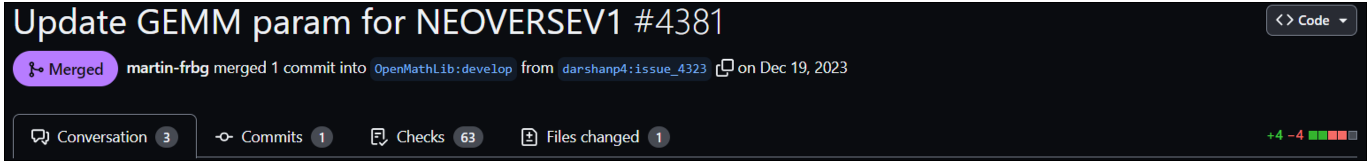
➢ Upcoming Contributions by Fujitsu

Cross Compilation of Arm on x86	Block Size Optimization for Arm	Bazel Build Support for Arm
Utilize x86 machines for testing Arm compilation and CI test suite without additional Arm instances required on Intel side	Dynamic template dispatcher to identify architecture/ISA specific optimal block size	New architectures to support bazel build system, starting with Arm

OpenBLAS Development for oneDAL

➢ Collaborative Active PRs

OpenBLAS PR (#4381) is merged, raised by **Fujitsu** to support GEMM cache size optimization

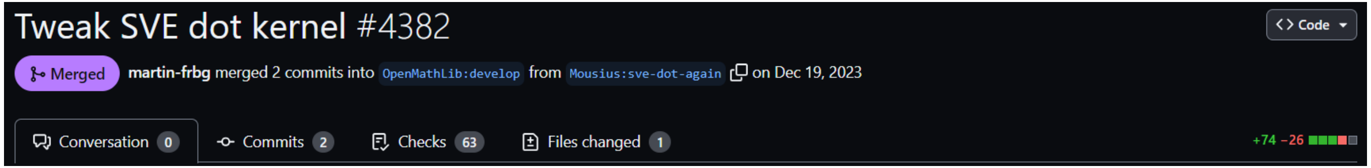


Update GEMM param for NEOVERSEV1 #4381

Merged **martin-frbg** merged 1 commit into `OpenMathLib:develop` from `darshanp4:issue_4323` on Dec 19, 2023

Conversation 3 · Commits 1 · Checks 63 · Files changed 1 · +4 −4

OpenBLAS PR (#4382) is merged, raised by **Arm** to streamline SVE predicate & DOT kernel assembly

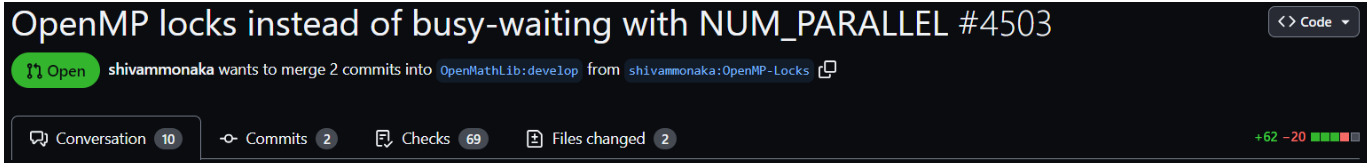


Tweak SVE dot kernel #4382

Merged **martin-frbg** merged 2 commits into `OpenMathLib:develop` from `Mousius:sve-dot-again` on Dec 19, 2023

Conversation 0 · Commits 2 · Checks 63 · Files changed 1 · +74 −26

OpenBLAS PR (#4503) is open, raised by **Fujitsu** to improve OpenBLAS threading performance



OpenMP locks instead of busy-waiting with NUM_PARALLEL #4503

Open **shivammonaka** wants to merge 2 commits into `OpenMathLib:develop` from `shivammonaka:OpenMP-Locks`

Conversation 10 · Commits 2 · Checks 69 · Files changed 2 · +62 −20

➢ Performance Results with PRs

Update GEMM param for NEOVERSEV1	Tweak SVE DOT kernel	OpenMP locks instead of busy-waiting with NUM_PARALLEL
Performance for SGEMM improved by ~ 2-5% and DGEMM improved by ~2-12%	The benchmarks indicate perf improve by ~33%.	Improved OpenMP with OpenBLAS to have controlled parallel execution and consistent design with Pthreads and Win32 backend.

Resources

oneDAL Pull Request Contribution

- Enable ARM(SVE) CPU support with reference backend

Scan the QR code to know more

Additional oneDAL Pull Requests

- Initial input for backend selection #2396
- Makefile refactoring to factor out common build code #2672
- Fix: Do not import onedal when OFF_ONEDAL_IFACE=1 #1744

OpenBLAS Pull Request Links

- Update GEMM param for NEOVERSEV1 #4381
- Tweak SVE dot kernel #4382
- OpenMP locks instead of busy-waiting with NUM_PARALLEL #4503

FUJITSU-MONAKA Reference Links

- FUJITSU-MONAKA Next Arm Processor
- Democratizing the use of AI: FUJITSU – MONAKA
- FUJITSU leads development of energy-efficient CPUs and photonics smart NIC for next-generation green data centers under NEDO program

Scan the QR code to know more

Concluding Remarks

Contribution

❑ **Fujitsu successfully contributes to UXL OSS** enabling oneDAL on ARM, showcasing significant AI-ML algorithm speedups with SVE optimization and porting.

Fujitsu Vision

❑ **FUJITSU-MONAKA aligns software acceleration** commitment with green data center goals and aims to democratize AI for sustainable digital transformation.

Applications

❑ **Use Case performance spans multiple domains like** healthcare, retail, smart city, manufacturing, finance, defect detection, recommendation, banking, digital twin, data generation etc.

Collaboration

❑ **Advancing our broader vision together with UXL**, Fujitsu looks forward to actively collaborate with OSS community for accelerated computing ecosystem.

UXL vision for open standard accelerator software ecosystem & evangelize OSS community efforts

Acknowledgement

NEDO Project | “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, AIOCORE Co., Ltd., KIOXIA Corporation, FUJITSU Optical Components Limited and KYOCERA Corporation.

- 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