SustainML

Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems

Annual report on collaboration - Y1

Deliverable D7.7

WP7 - Dissemination, Exploitation and Management

This project has received funding from the European Union's Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408

eProsima The Middleware Experts

DFKI

Inria

upmem

RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

IBM

Project

Title:	SustainML: Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems
Acronym:	SustainML
Coordinator:	eProsima
Grant agreement ID:	101070408
Call:	HORIZON-CL4-2021-HUMAN-01
Program:	Horizon Europe
Start:	01 October 2022
Duration:	36 months
Website:	https://sustainml.eu
E-mail:	sustainml@eprosima.com
Consortium:	**eProsima (EPROS)**, Spain **DFKI**, Germany **Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (RPTU)**, Germany **University of Copenhagen (KU)**, Denmark **National Institute for Research in Digital Science and Technology (INRIA)**, France **IBM Research GmbH**, Switzerland **UPMEM**, France

Deliverable

Number:	**D7.7**
Title:	**Annual report on collaboration - Y1**
Month:	12
Work Package:	WP7 - Dissemination, Exploitation and Management
Work Package leader:	eProsima
Deliverable leader:	DFKI
Deliverable type:	Report (R)
Dissemination level:	Public (PU)
Date of submission:	30/09/2023
Version:	v1.4
Status:	Finished

Version history

Version	Date	Responsible	Author/Reviewer	Comments
v1.0	25/08/2023	eProsima	Raúl Sánchez-Mateos	Deliverable template and first draft
v1.1	26/08/2023	UPMEM	Yann Falevoz	UPMEM collaborations
v1.2	26/08/2023	eProsima	Raúl Sánchez-Mateos	Review and eProsima collaborations
v1.3	27/08/2023	DFKI	Bo Zhou / Daniel Geißler	DFKI collaborations
v1.4	28/08/2023	eProsima	Raúl Sánchez-Mateos	Executive summary and final review

Executive summary

SustainML project aims to develop a design framework and an associated toolkit, so-called SustainML, that will foster energy efficiency throughout the whole life-cycle of Machine Learning (ML) applications: from the design and exploration phase that includes exploratory iterations of training, testing and optimizing different system versions through the final training of the production systems (which often involves huge amounts of data, computation and epochs) and (where appropriate) continuous online re-training during deployment for the inference process. The framework will optimize the ML solutions based on the application tasks, across levels from hardware to model architecture. It will also collect both previously scattered efficiency-oriented research, as well as novel Green-AI methods. Artificial Intelligence (AI) developers from all experience levels can make use of the framework through its emphasis on human-centric interactive transparent design and functional knowledge cores, instead of the common blackbox and fully automated optimization approaches.

This report corresponds to *Deliverable D7.7 - Annual report on collaboration - Y1* of the SustainML project. This deliverable covers the main collaborations between SustainML consortium members and research groups, organizations, EU projects and communities external to the project.

Contents

Acronyms

AI	Artificial Intelligence.
ML	Machine Learning.
ROS	Robot Operating System.
TSC	Technical Steering Committee.

1 Collaboration with other EU projects and initiatives

1.1 Humane-AI-Net

Humane-AI-Net (Grant agreement 952026) [1] is an EU-funded ICT48 network of AI excellence centers with over 50 partners including AI research institutes, national incubators, industry application partners, etc. across Europe. DFKI is the coordinator of Humane-AI-Net and has promoted the SustainML project in the recent networking events, including a consortium meeting in Krakow, Poland which took place together with the European Conference on AI (2023).

1.2 ALMA

Algebraic Machine Learning (AML) is a relatively new machine learning technique based on algebraic representations of data. Unlike statistical learning, AML algorithms are robust regarding the statistical properties of the data and are parameter-free. The aim of the EU-funded ALMA[2] project is to leverage AML properties to develop a new generation of interactive, human-centric machine learning systems. These systems are expected to reduce bias and prevent discrimination, remember what they know when they are taught something new, facilitate trust and reliability and integrate complex ethical constraints into human–artificial intelligence systems. Furthermore, they are expected to promote distributed, collaborative learning.

Some partners of SustainML project are also members of ALMA consortium (DFKI, eProsima, RPTU,and Inria). This will facilitate the integration and use of the results obtained from the ALMA project in the SustainML project tasks, making it possible to test the resulting framework with AML models.

1.3 Pro2Future Symposium

Within the Pro2Future Symposium on Sustainable and Cognitive Products and Production held in Linz, Austria on 22.06.23, Paul Lukowicz gave a keynote speech entitled "Why smaller is better: Edge Embedded AI for Smart Production".

1.4 Carl-Zeiss foundation SembAI

Sustainable Embedded AI (SEmbAI)[3] is a multi-year research project funded by the Carl-Zeiss Foundation in Germany. The project focuses on novel embedded AI applications with data and energy efficiency and has shared principle as SustainML. DFKI and RPTU are both participants in SEmbAI and has started to investigate collaboration opportunities in sustainable embedded AI.

1.5 TALON

TALON[4] introduces an AI orchestrator that envisions transforming the I5.0 into an automated intelligent platform by exploiting advances in edge networks and bringing intelligence near sensors in embedded systems with limited computational, storage, and communication resources, as well as the integration of advanced and adaptive sensors and perception.

[1] https://www.humane-ai.eu

[2] https://alma-ai.eu/

[3] https://sembai.cs.uni-kl.de

[4] https://talon-project.eu/

In this direction, TALON's AI orchestrator maximizes both global and individual users' and systems' capabilities without violating the design parameters of each application. In particular, the orchestrator selects AI datasets, algorithms, metrics, and models based on the application.

eProsima has participated in the *Establishing the next level of "intelligence" and autonomy Clustering Workshop Event*[5] organised by the EU-funded project TALON in close collaboration with the projects AutoFair[6], ENEXA[7], EVENFLOW[8], REXASI-PRO[9], SAFEXPLAIN[10], SustainML[11], TUPLES[12], and ULTIMATE[13] funded under the HORIZON-CL4-2021-HUMAN-01-01 topic. This workshop aims at bringing together projects and experts for exploring synergies and identifying actions that can be pursued in common in the area of verifiable robustness, energy efficiency and transparency for trustworthy AI. Future collaboration between different European projects was one of the key topics discussed.

1.6 BioPIM

In addition to spearheading the groundbreaking developments within the framework of the EIC Accelerator, UPMEM is proud to be an active participant in several collaborative projects, fostering innovation both at the national and European scales. Our engagement spans four prominent European projects (including SustainML) and two national initiatives, forming a rich tapestry of collaborative efforts. These diverse projects not only underline UPMEM's commitment to cross-disciplinary collaboration but also hold the promise of mutually reinforcing successes. The synergy between these endeavors presents a unique opportunity for knowledge exchange, technological advancements, and the acceleration of sustainable solutions.

UPMEM is one of the consortium partners of BioPIM[14] (Grant agreement ID: 101047160). This project is dedicated to harnessing processing-in-memory (PIM) technologies for enhanced edge computing in bioinformatics. By co-designing algorithms and data structures with various PIM architectures, BioPIM aims to achieve optimal cost-effectiveness, energy efficiency, and time savings in bioinformatics applications. This project aligns seamlessly with the objectives of SustainML. The advancements in PIM technologies explored in SustainML, particularly in Task 2.3 focusing on energy-efficient hardware for machine learning, can significantly benefit BioPIM. The cross-disciplinary insights and innovations developed in SustainML provide valuable foundations for BioPIM, ensuring a symbiotic relationship where the sustainable machine learning practices pioneered by SustainML contribute to the efficiency and success of BioPIM's endeavors in bioinformatics and beyond.

[5] https://talon-project.eu/eventsestablishing-the-next-level-of-intelligence-and-autonomy-clustering-workshopevents/

[6] https://humancompatible.org/

[7] https://enexa.eu/

[8] https://evenflow-project.eu/

[9] https://rexasi-pro.spindoxlabs.com/

[10] https://safexplain.eu/

[11] https://sustainml.eu/

[12] https://tuples.ai/project/

[13] https://ultimate-project.eu/

[14] http://biopim.eu

1.7 STRATUM

STRATUM (responding to the HORIZON-HLTH-2023-TOOL-05-05 call, GA under preparation, and with UPMEM as part of the project consortium) focuses on developing a 3D Decision Support Tool for brain surgery guidance and diagnostics, integrating AI algorithms and energy-efficient computing solutions. As SustainML advances machine learning practices for sustainability, the methodologies and innovations fostered by the project could significantly benefit STRATUM. The cross-disciplinary insights and energy-efficient computing techniques developed in SustainML offer valuable foundations for STRATUM's mission to improve intraoperative diagnostic accuracy and reduce surgery time in brain tumor surgeries.

1.8 TIRESYAS

TIRESYAS (responding to the EDF-2022-RA-SENS-ART call, GA under preparation, and with UPMEM as part of the project consortium) focuses on enhancing radar capabilities to counter emerging threats. The project aims to bolster sensor resilience, sustainability, and operational efficiency. As TIRESYAS explores intensive computation, signal processing, and cyber resilience, SustainML may offer valuable perspectives on optimizing energy efficiency and sustainable practices within the radar technology domain, contributing to the overarching goal of enhancing radar performance while ensuring environmental sustainability.

2 Research collaborations

2.1 ASAP conference

DFKI preseneted a paper funded by SustainML titled *FieldHAR: A Fully Integrated End-to-end RTL Framework for Human Activity Recognition with Neural Networks from Heterogeneous Sensors* [1] on the 34th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2023) in Porto [15] and has received positive interests.

2.2 Diehl Aviation

Based on the gained research knowledge in the scope of SustainML, specifically on deploying machine learning models in energy-constrained environments, DFKI gained the attendance of Diehl Avionics which led to an industry-research partnership project. The accompanying neural network model training techniques on compressing model size and complexity enable research towards embedding neural network models into the safety-critical and energy-constrained environment of avionics. Currently, the guidelines of the avionics sector prohibit the embedding of any AI-based controls in safety-critical systems, though there is a strong urge for such solutions to enhance the avionic sector's potential. DFKI not only supports Diehl Avionics through providing knowledge, rather the mutual exchange of two viewpoints on different problem settings and the resulting implementation of prototype systems covering the collected requirements sums up the collaboration.

3 Collaboration with ROS community

The partnerships with the Robot Operating System (ROS) community are intended to bring the developers of the ROS community closer to the SustainML framework to leverage this framework the main

[15] `https://www.asap2023.org`

design tasks prior to ML application development. The methodology to be followed in this collaboration is based on the use of three major channels:

- ROS Discourse
- ROS 2 Technical Steering Committee (TSC)
- AI Integration Working Group

These are explained below in a section dedicated to each of them.

3.1 ROS Discourse

ROS Discourse [16] serves as a dynamic forum within the ROS community are able to share use cases, announcements and other news related to ROS, rather than technical questions. It is a place for users to interact with each other by showing the progress of their implementations, proposals for new working groups and research results in order to improve the ROS ecosystem.

EPROS is actively engaged in this channel, committed to providing regular updates on developments performed within the SustainML project, with a specific focus on the SustainML framework tool. Our collaboration with the ROS community aims to ensure that the roadmap of new ROS-compatible technologies is useful and potentially exploitable by ROS end-users and developers.

In the initial stages of the project, , forum activity has been limited, as outlined in *Deliverable 7.4 - Report on communication, exploitation and dissemination - Initial*. The phase of engaging the ROS community to use the SustainML framework in their AI/ML design and development process is scheduled for the upcoming project phase, when proven innovation results and stable and testable products resulting from the innovation will be made available.

3.2 ROS 2 Technical Steering Committee (TSC)

The ROS 2 Technical Steering Committee (TSC)[17] is the organism comprising representatives of companies, organisations, and ROS 2 community in charge of setting the guidelines for the development of ROS 2. These organisations are committed to collaborate in the development of ROS 2, and have the responsibility to define the direction of the project.

EPROS as a member of the ROS 2 TSC has the right to vote on actions taken affecting the governance of ROS 2. Therefore, EPROS has a privileged position to promote the SustainML project at dedicated TSC meetings and hence reach the ROS 2 community.

Moreover, because the representatives of the organisations that comprise the ROS 2 TSC have a wide experience in the robotics industry, and especially in ROS 2, it is possible to get first-hand feedback on all the results of the SustainML project.

3.3 AI Integration Working Group

Before the initalization of the proyect, eProsiam submitted a proposal for the creation and leadership of a new working group that aims to create a community of ROS users who apply ML technologies in systems developed with the ROS 2 framework. Thus, it would be possible to closely monitor the use cases

[16] https://discourse.ros.org

[17] https://docs.ros.org/en/rolling/Governance/ROS2-TSC-Charter.html

of AI in robotics in order to present SustainML as an innovative solution that has a number of benefits compared to traditional ML applications design processes.

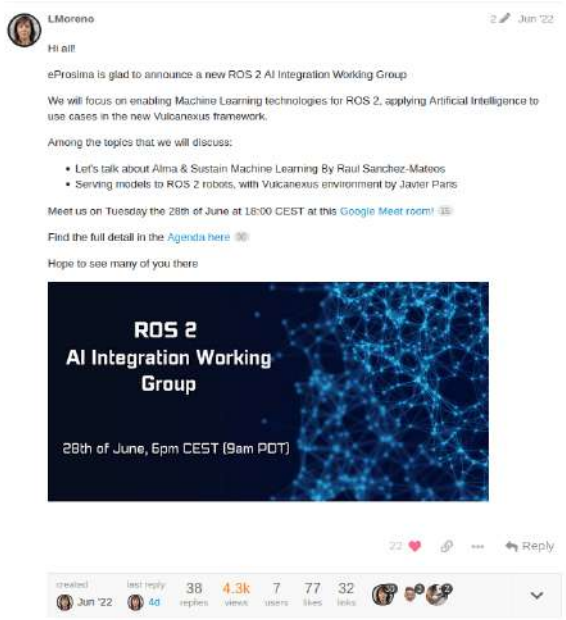
During the first phase of the project, eProsima has been leading the AI Integration Working Group. The primary objective is dedicated to empowering Machine Learning technologies for ROS 2 and applying Artificial Intelligence to various use cases. Through collaborative synergy and the exchange of insights, the group is dedicated to showcasing the potential of SustainML framework for robotics applications based on ML in the ROS community.

For detailed information regarding the working group agenda, open public Google Meet meeting room, and meeting discussions, please visit the AI Integration Working Group topic in ROS Discourse [18]. Also note that every meeting is recorded and the session recording in publicly available in eProsima's Vulcanexus YouTube channel [19].

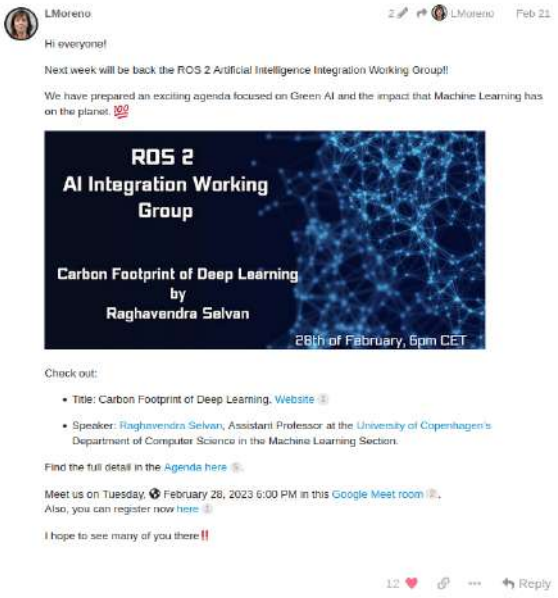
For a comprehensive overview of the group's topics and interactions, please refer to Figure 1.

[18] https://discourse.ros.org/t/ros-2-ai-integration-working-group

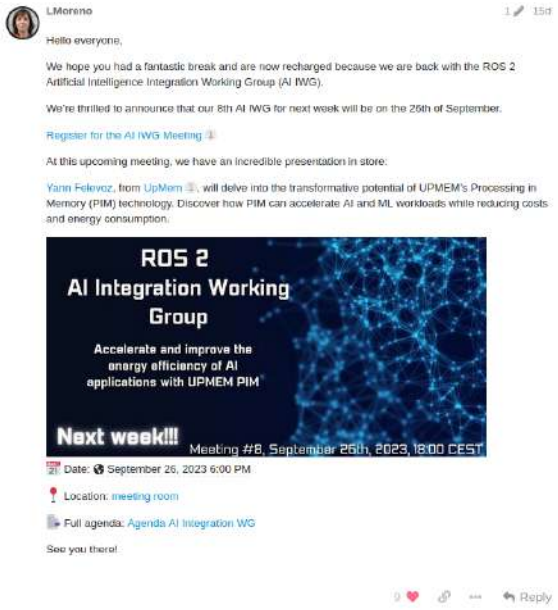
[19] https://www.youtube.com/@vulcanexus9399



(a) AI Integration Working Group first post



(b) AI Integration Working Group - meeting #4



(c) AI Integration Working Group - meeting #8

Figure 1: ROS Discourse posts

4 Conclusions

4.1 Synergizing Chip Design Evolution for Enhanced Computational Efficiency

In a collaborative convergence of cutting-edge projects, BioPIM, ODYSSAI, STRATUM, and SustainML collectively explore diverse avenues of UPMEM chip design evolution. With a shared goal of accelerating performance and minimizing energy consumption across varied applications, these projects contribute significantly to identifying optimal design options for the next generation of chips. BioPIM focuses on harnessing processing-in-memory technologies for genomics applications, ODYSSAI targets modular edge computing for embedded AI, STRATUM advances technologies in complex surgeries, and SustainML pioneers sustainable machine learning practices. Together, they form a comprehensive exploration that aims to render intensive computation faster and more energy-efficient. This collaborative effort is poised to shape the future landscape of UPMEM chip design, propelling advancements that not only optimize computational speed but also energy efficiency across a spectrum of applications.

References

[1] Mengxi Liu et al. "FieldHAR: A Fully Integrated End-to-end RTL Framework for Human Activity Recognition with Neural Networks from Heterogeneous Sensors". In: *arXiv preprint arXiv:2305.12824* (2023).