Matchmaker: Marrying Emerging AI Algorithms to Emerging AI Hardware through Compilers (3E210632)

A Data Management Plan created using DMPonline.be

Creators: Josse Van Delm, First Name Surname

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1SE7723N

ID: 199132

Start date: 10-01-2021

End date: 31-10-2026

Project abstract:

We see an enormous surge in machine learning (ML) applications within embedded “tinyML” devices, such as wearables, robots, etc. These applications require custom optimized hardware-software solutions to enable low-energy and low-latency operation in resource-scarce devices. Yet, this field is plagued by a disconnect between fast evolutions of ML algorithms, and slow availability of new hardware platforms. The long development cycles of new ML accelerators and the compilers required to exploit them, prevent the much-needed customization of tinyML platforms. This is especially problematic for smaller companies, who cannot put large processor, compiler and software development teams at work, as currently done in the large vertical companies (Google, Qualcomm, Facebook,…). Flanders has a wide range of such SME companies active in tinyML on the application side, in embedded hardware design, or system integration.

To overcome this, MATCHMAKER will develop an open extensible compilation infrastructure towards efficient development of tinyML platforms, overcoming the disconnect between HW and SW. The targeted tool will support concurrent development of programmable heterogeneous accelerators with aligned automated compiler generation. This is pursued through the creation of new hardware and compilation target description dialects in MLIR. Upon success, more companies will be able to enter this market and develop new products with a short time to market and better ease-of-use.

Last modified: 28-04-2023

Matchmaker: Marrying Emerging AI Algorithms to Emerging AI Hardware through Compilers (3E210632)

Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

This project aims to create state-of-the-art compilers. Compilers are computer programs that generate programs. As such, the following data can be expected.

- Compiler source code: The programming code that describes the internals of the compiler. This is usually written in languages like C(++), Python, Rust or others. This code can written from scratch or reused from an open-source compiler such as TVM, or LLVM.
- Compiler intermediate output: This is a program compiled by the compiler. This can be stored in many formats: E.g. C code, .mlir text format, .mlir bytecode.
- Compiler final output: This is typically binary code to run on a CPU or a compiler platform. Typical binary program formats are .ELF, .BIN, .HEX files.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

At MICAS (KU Leuven), all data are either stored in a database or in a version control system. The corresponding repositories are stored on the server clusters of the department. The data will be archived using redundant disk infrastructure with daily tape backups. Documents and data will be archived on these servers for at least five years. There is plenty of storage during and after the research. The designated responsible person is Ben Geeraerts. In addition to this, published manuscripts are stored in a central repository, i.e. the KU Leuven Lirias system.

Furthermore the research in this proposal plans to create open-source code, which will be hosted on publicly accessible git repositories (e.g. github.com) or research archival systems (zenodo.org).

What’s the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

I do not wish to to deviate from this principle.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

NA

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

A Data Management Plan (DMP) will be drawn up at the start of the project. This DMP for archiving and open access of research data will be a central aspect in the project, and in full coherence with the guidelines for open science within FWO and the FAIR Data Principles (see https://www.kuleuven.be/rdm/en/fair). The DMP will be an evolving document, and gain more precision and substance during the lifespan of the project.

Matchmaker: Marrying Emerging AI Algorithms to Emerging AI Hardware through Compilers (3E210632)

DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Matchmaker: Marrying Emerging AI Algorithms to Emerging AI Hardware through Compilers (3E210632)

GDPR

GDPR

Have you registered personal data processing activities for this project?

- No

Matchmaker: Marrying Emerging AI Algorithms to Emerging AI Hardware through Compilers (3E210632)
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)
Compiler Source Code	Programming code (e.g. python or C code that performs software/hardware compilation)	New and reused(partially based on open-source projects)	Digital	Software	.c .py,	< 100 GB
Compiler intermediate output	Intermediate compiler output in textual or binary form	New	Digital	Compiled data	.sv, .mlir	< 100 GB
Compiler output	Final compiler output in binary form	New	Digital	Compiled data	.mlir .elf (RISC-V ABI)	< 100GB
Profiling output	Program/Hardware performance measurements	New	Digital	Observational	.csv	< 100GB

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

For compiler source code:

- Reuse of TVM project open-source compiler code of https://tvm.apache.org
- Reuse of LLVM project open-source compiler code of https://llvm.org

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

The work will introduce novel methods in compilation, and digital compute architectures.
These findings will certainly have potential for valorization.

The produced compiler source code will be openly available and released under the respective permissive license of the upstream compiler code for TVM it is the Apache 2.0 License and for LLVM it is the Apache 2.0 License with LLVM exceptions.
This means that the code is available for commercial exploitation.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

The produced source code will be based on the TVM or LLVM codebase which are available under the Apache 2.0 License and the Apache 2.0 *License* with *LLVM* exceptions respectively, which are permissive open-source code licenses.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is

recorded).

Necessary tools and documentation will be provided to reproduce experiments presented in the research.
This includes in-code comments, code reference documentation, simple setup instructions and simple tutorials. The data will be kept in typical text files such as restructuredtext (.rst) and/or markdown (.md).
In case extra information is needed to guide the reader of this data, a README in markdown will be provided to guide the reader.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

There is no formally acknowledged metadata standard specific to our discipline. Instead, a markdown file will be used to describe in text what each dataset contains.

3. Data storage & back-up during the research project

Where will the data be stored?

We will use the central storage facilities of our research department.
Publicly available data will be hosted on github.com and zenodo.org.

How will the data be backed up?

The data will be stored on our servers with automatic daily back-up procedures.
Github.com: data backups are done by github inc.
Zenodo.org : Permanent research storage managed and backed-up by CERN in Switzerland.

**Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.**

- Yes

There is sufficient storage & backup capacity. Recently the storage capacity available to our research group has been expanded to 25 TB.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

At MICAS confidential data is stored on file servers which are only accessible by authorized people with specific account settings. The servers are located in a secured room with access limited to system administrators. For data related to specific, very advanced and exclusive technologies we have physically separate file servers.
Zenodo and github contain public repositories, accessible by everyone with an internet connection.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The costs for data storage are internally accounted for at departmental level. Our research group carries a proportional part of the departmental IT costs.
Github and zenodo is provided for free.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

The following data will be retained for the expected 5 year period after the end of the project:

- the data needed to reproduce and verify published research results
- the data needed to prove and increase the value of research results that have valorization potential
- all source code.

Zenodo.org and data pertained on it will exist as long as CERN exists.

Where will these data be archived (stored and curated for the long-term)?

Long term data archival will be done through Zenodo.org

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Zenodo.org is free

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository

We will publish in international journals, after careful consideration of valorization and patentability potential, during and/or after the project. We will follow the Green Open Access strategy for these scientific publications. In those cases where we do have to publish in journals that are behind a paywall (e.g. IEEE journals that are the top in the field), we will always make a digital copy of the accepted paper available through an online repository. We will ensure that every publication gets a Digital Object Identifier (DOI) and that we use our ORCID on every publication, so that the identification of the record and of the authors is unambiguous. Next to the Lirias document repository system of KU Leuven, we will also use arXiv, which is a free distribution service and open-access archive.
Data related to published results can be made available through zenodo.org.
github.com: Source code for reproducing the compiler output and experiments
Zenodo.org: Permanent back-up of github.com repository.

If access is restricted, please specify who will be able to access the data and under what conditions.

Access is not restricted.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- No

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Source code: github.com and zenodo.org
Other data: Zenodo.org
Publications will be accessible through the established channels.
The simulation, design and measurement data related to published results will be made available through KU Leuven's RDR or zenodo.org.

When will the data be made available?

Immediately after the publication based on that data.

Which data usage licenses are you going to provide? If none, please explain why.

Apache 2.0 with LLVM exceptions for source code

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

A DOI will be provided by zenodo.org where necessary

What are the expected costs for data sharing? How will these costs be covered?

The costs for data storage are internally accounted for at departmental level. Our research group carries a proportional part of the departmental IT costs.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Josse Van Delm (Researcher) + Marian Verhelst (Supervisor) as end responsible

Who will manage data storage and backup during the research project?

Josse Van Delm (Researcher) + Ben Geeraerts (IT) as support + Marian Verhelst (Supervisor) as end responsible

Who will manage data preservation and sharing?

Josse Van Delm (Researcher) + Ben Geeraerts (IT) as support + Marian Verhelst (Supervisor) as end responsible

Who will update and implement this DMP?

Josse Van Delm (Researcher) + David Maes (valorization) as support + Marian Verhelst (Supervisor) as end responsible