MEEP

MareNostrum Experimental Exascale Platform

D1.2 Data Management Plan (DMP)

Version 1.2

Document Information

Contract Number	946002
Project Website	https://meep-project.eu/
Contractual Deadline	30/06/2020
Dissemination Level	Public (PU)
Nature	Open Research Data Pilot (ORDP)
Author	John David Davis (BSC)
Contributors	
Reviewers	Nadia Tonello (BSC's Head of Data Management), Elisenda Rasero Rebull (BSC), Sergi Madonar (BSC)

The MEEP project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 946002. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Croatia, Turkey.

Change Log

Version	Author	Description of Change
V 1.0	John David Davis (BSC)	Initial draft
V 1.1	Elisenda Rasero Rebull (BSC), Sergi Madonar (BSC)	First review
V 1.2	Nadia Tonello (BSC's Head of Data Management)	Second review

Comments

This deliverable will be updated if new data is generated or used in this project or if different processes have to be applied to manage the data.

Index

1. Executive Summary

This deliverable presents the Data Management Plan (DMP) of the MEEP project, which describes the data management life-cycle for all datasets to be collected, processed and/or generated along the lifetime of the project. The DMP complies with the European Comission"s objective of making research data findable, accessible, interoperable, and reusable (FAIR). Concretely, this deliverable describes, among others:

- Which datasets will be generated, collected and processed, considering both, the development and execution of the MEEP application use-cases and the research activities towards the development of the MEEP emulation and software development technology.
- Which methodology and standards will be applied to MEEP datasets.
- How datasets will be stored and handled during the lifetime of the project, and after the end of it.
- How the datasets will be made (openly) accessible.

The datasets managed or created in the MEEP project are:

Dataset	Description
HPC , AI, ML, and DL benchmarking	Publicly available benchmarks like Linpack and HPCG will be used to evaluate the MEEP accelerator architecture. We envision using other publicly available high performance data analytics workloads for AI/ML/DL. We may use other workloads like GROMACS, AYLA, and other real applications and related input data.
Accelerator architecture performance modeling	The accelerator performance will be evaluated using microkernels like, DGEMM, FFT, SpMV, DCT, and others, with open data sets and synthetic data. The tests cases and the model infrastructure will be made open source.
Accelerator architecture code	When possible, we will derive components from open source and when possible, provide results that are open source.
Personal data from partners	Names and emails of the people involved in the project will be gathered and stored to ensure a proper internal communication.

Table 1: Datasets managed or created in the project

2. Data Summary

In the MEEP project, performance numbers from the HPC, and HPDA (High Performance Data Analytics, including Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL)) applications will be generated with the goal to compare the performance of key applications to the target performance of the emulated accelerator on the various applications ported to the emulated platform. It is expected that this is a small set of data, less than 10 MB.

Performance data will be collected as execution time, and energy consumption, and other metrics will be derived, such as speedup, and GFlops/Watt. There will not be any previous data used and the data will be generated from the performance of benchmarks and applications.

The result data will be useful for researchers working on similar approaches for building and accelerating software/hardware co-designed exascale systems, and in particular targeting HPC and HPDA applications. Using a common set of publicly available workloads will allow comparisons with other systems.

MEEP will generate four main types of datasets:

1. The source code for a performance model used to evaluate the architecture using small microkernels with synthetic input data.
2. The source code, RTL, for the exascale accelerator that can be mapped to the hardware emulator (FPGA system). This same or similar code may target ASICs for hardware deployment outside of the scope of this project.
3. The source code of the software components and tools that will form the software stack for the MEEP emulator.
4. Datasets generated and used for the evaluation of the accelerator software and hardware stack.

The MEEP project will also manage the personal data from the partners of the consortium under GDPR and only for project-related purposes. The personal data like names and emails will be collected by the Project Manager and stored in BSC's internal servers. This data will be accessible and updatable by all the partners through the project's intranet.

3. FAIR Data

3.1 Making data findable, including provisions for metadata

The performance data provided by the performance simulations infrastructure and the MEEP emulator will be small, less than a Gigabyte. It will be organized by application, making the results easy to find. The results will be in a standard format based on the output of the applications and thus is too small for any standard identification mechanisms. Because of the reduced size of the datasets, the identification mechanism will be based on the file names with a defined convention what will be defined once the datasets are created.

The input data for the applications, especially, the synthetic applications, will be auto-generated or from publicly available sources when possible. There are no plans to persist the data, other than the method used to generate or obtain the input data. Thus, no unique

identifiers (i.e., Digital Object Identifiers) are required to locate this data. When using public benchmark applications, the data will be generated based on theirs documentation.

Any source code for the MEEP project will be available in open source repositories provided by the BSC with a standard organization and related documentation, probably with a Gitlab server or equivalent.

3.2 Making data openly accessible

All input and output data generated in this project may follow the Open Access policy. However, when there is a huge amount of data (more than 1 GB) used as part of the application data set, only the generation and/or source of the data sets for the applications will be provided and accessible by the community through the MEEP repository (most likely BSC's Gitlab server) to demonstrate the validity of the MEEP implementation.

The source code for any components licensed as open source will be included in a Git repository. When needed, new Git projects and repositories will be created for the various parts of the MEEP project. Furthermore, Git submodules will be used to link the integrated version with the corresponding Git project of each of the MEEP components (software and/or hardware), when applicable.

For any models, codes, or other information developed by BSC, its partners, or other third parties, an agreement with the owner will be required. Ideally an open source model will be followed.

3.3 Making data interoperable

In most cases, the data will be in text format, so no specific data format will be provided to the datasets needed to evaluate the performance of the MEEP accelerator due to the small size. It will be possible to compare the performance data obtained from other sources with the MEEP accelerator to determine its relative performance. This information will be included in scientific documents to properly determine the advances of the MEEP architecture.

The simulation data will be published with enough detail in order to allow other scientists to compare their results with the ones generated in this project.

3.4 Increase data re-use (through clarifying licences)

During the project data re-use options for the data sets will be dealt on a case by case basis, aiming, when possible, to keep them public, accessible, free of charge and reusable under request, under a Creative Commons license to enable the widest reuse or inheriting the license types from the different data sources as explained in the data summary description, and taking care of each partner's' business constraints or legal limitations on them.

4. Allocation of resources

There is no additional cost of making our performance data FAIR, as it does not need a special treatment. The data will be FAIR with current BSC resources and equipment.

Performance data will be under the responsibility of BSC, the coordinator of the project.

Data will be kept for three years after the MEEP project. After three years, we consider that the data will not have value anymore, as results will be superseded by new datasets obtained from future developments. Data will still be present in project publications and public repositories when appropriate.

5. Data security

There is no need for applying data security policies in the project given that the data used and produced do not include any personal or private data that could be considered as sensitive. Regular backups for keeping the information safe will be used.

6. Ethical aspects

There will not be major ethical or legal issues directly devised for the generated data. The only detected ethical issue will be to respect the intellectual property of any possible software we could use from partners or third parties.

7. Further support in developing your DMP

No other procedures for data management for the performance data. This document will continue to be updated over the lifetime of the project and a final draft will be released at the end of the project.