DAPHNE

Press Release – Jan 11, 2021

EU H2020 Project DAPHNE on building a Next-Gen System Infrastructure for Integrated Data Analysis Pipelines (DM, ML, HPC)

Context: Increasing digitization efforts, sensor-equipped everything, and feedback loops for data acquisition lead to increasing data sizes and a wide variety of valuable, but heterogeneous data sources. Modern data-driven applications leverage these data collections for finding interesting patterns and building robust predictive models. In this context, we observe a trend toward complex integrated data analysis pipelines that combine data-parallel computing frameworks, distributed machine learning (ML) systems, and high-performance computing (HPC) libraries and systems. While data management, ML, and HPC share many underlying techniques, and the characteristics of their, increasingly heterogeneous, hardware environments are converging; the respective programming paradigms, cluster resource management, and data formats differ substantially.

Challenges: Developing and deploying such integrated data analysis pipelines, faces two major challenges. First, the large computational requirements are impacted by fundamental HW challenges. The end of Dennard scaling (power stays proportional to the area of the transistor), causes increasing power density, and thus, stagnating frequency and dark silicon (only parts of a chip active). Additionally, Moore's law (number of transistors per chip doubles every two years at constant costs) also comes to an end. While these scaling limitations can be addressed with increasing parallelism; by Amdahl's law, even small sequential regions, fundamentally limit the potential speedup. Second, developing complex integrated data analysis pipelines still requires substantial manual effort for orchestrating these different systems and environments. The system boundaries further cause unnecessary overhead and utilization challenges due to separate, statically provisioned clusters, lack of interoperability, and coarse-grained file exchange.

DAPHNE Overview: The recently launched DAPHNE project (Integrated **D**ata **A**nalysis **P**ipelines for Large-Scale Data Management, **H**PC, and Machi**ne** Learning) aims to address these challenges by defining and building an **open and extensible system infrastructure** for integrated data analysis pipelines. Addressing the hardware and utilization challenges requires specialization for heterogeneous hardware, dedicated data representations (e.g., sparsity-exploitation, compression, and indexing), and a tailor-made data flow from storage to compute and along the different pipeline tasks. However, this increasing specialization necessitates good abstractions in order to ensure productivity.

The DAPHNE project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 957407 for the time period from Dec/2020 through Nov/2024.

To this end, specific technical objectives include appropriate APIs and language abstractions, an MLIR-based intermediate representation, dedicated compiler and runtime techniques, hierarchical scheduling and task planning, techniques for near-data processing and HW acceleration (e.g., CPU, GPU, FPGA, vector devices), and a seamless integration with existing data processing frameworks, HPC libraries, and ML systems.

Excellent Consortium: A key strength of the DAPHNE project is an excellent, inter-disciplinary consortium of partners from the areas of data management (DM), machine learning (ML) systems and algorithms, high-performance computing (HPC), and their applications. In detail, the consortium consists of the Know-Center (DM, ML systems), AVL (automotive use case), DLR (earth observation use case, computational storage), ETH Zurich (ML systems and algorithms), HPI (DM, benchmarking), ICCS (HPC, distributed runtime), Infineon (semiconductor use cases), Intel (HPC, FPGAs), ITU Copenhagen (DM, computational storage/performance models), KAI (material degradation use case), TU Dresden (DM, HW accelerators), University of Maribor (EuroHPC center, optimization use cases), and the University of Basel (HPC, scheduling). The kickoff meeting in Dec 2020 already brought together more than 45 researchers from all partners, who are all very excited to get the project started.

Use Cases and Benchmarks: A variety of real-world application use cases grounds the research activities on a next-generation system infrastructure, and allows evaluating the improvements of productivity and end-to-end performance under different metrics like runtime, accuracy, and energy consumption. These use cases include local climate zone classification using earth observation data, semiconductor manufacturing analysis and optimization, material degradation modeling, and automotive development and simulation. Challenging characteristics include (1) petabyte-scale ML pipelines, (2) combined simulation, optimization and machine learning, (3) the integration of multi-modal, heterogeneous datasets, and (4) a broad spectrum of analysis techniques from traditional query processing and HPC, over statistical modeling and tests, to complex integrated data analysis and ML pipelines. Finally, inspired by the diversity of use cases, we will devise a new benchmark for integrated data analysis pipelines to foster the comparison of system alternatives and future improvements by industry and academia.

DAPHNE Facts & Figures

- **Web:** https://daphne-eu.github.io/
- **Project Consortium:** 13 partners from 7 European countries
- **Project Coordinator:** Know-Center GmbH
- **Project Duration:** 48 months, Dec 2020 – Nov 2024
- Open source DAPHNE reference implementation and benchmark
- **EU Funding:** € 6.6 Mio

The DAPHNE project is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 957407 for the time period from Dec/2020 through Nov/2024.