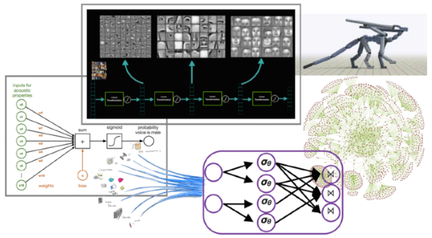
FASTER-AI
Fully Autonomous Safety- and Time-critical Embedded Realisation of Artificial Intelligence

Paris Carbone
Data Systems @ KTH

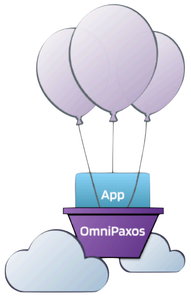
Speaker Intro

- Asst Professor at **KTH**
- Director of Data Systems Lab
  - 6 PhDs, 3 Research Engineers
  - Distributed Systems, Software & Databases Teaching
  - **Production-Grade Open-Source Projects**

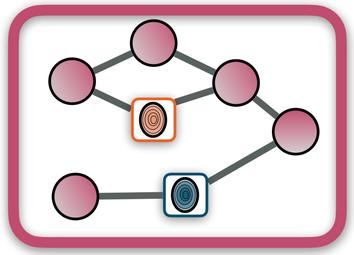
[2013-]
Apache Flink



[2018-]
ArcLang



[2021-]
OmniPaxos

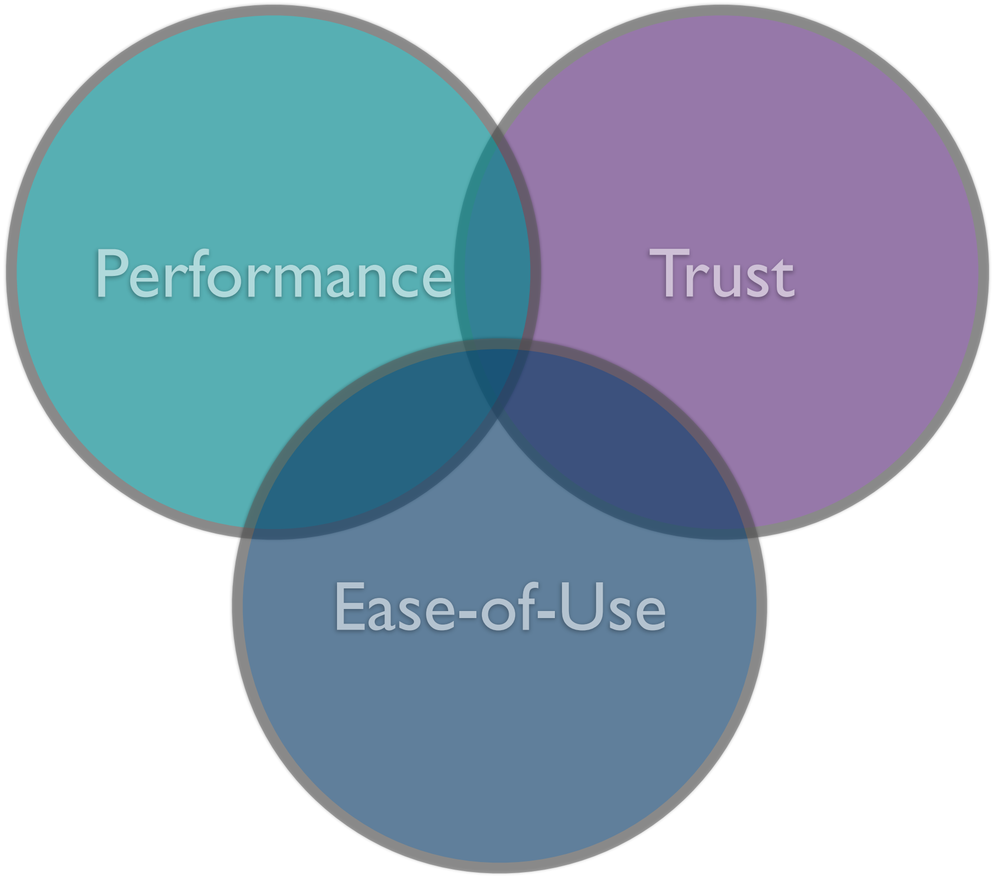


[2022-]
Portals

[2023-]
Orb DB

Key Adoption Properties for Advanced Digitalization

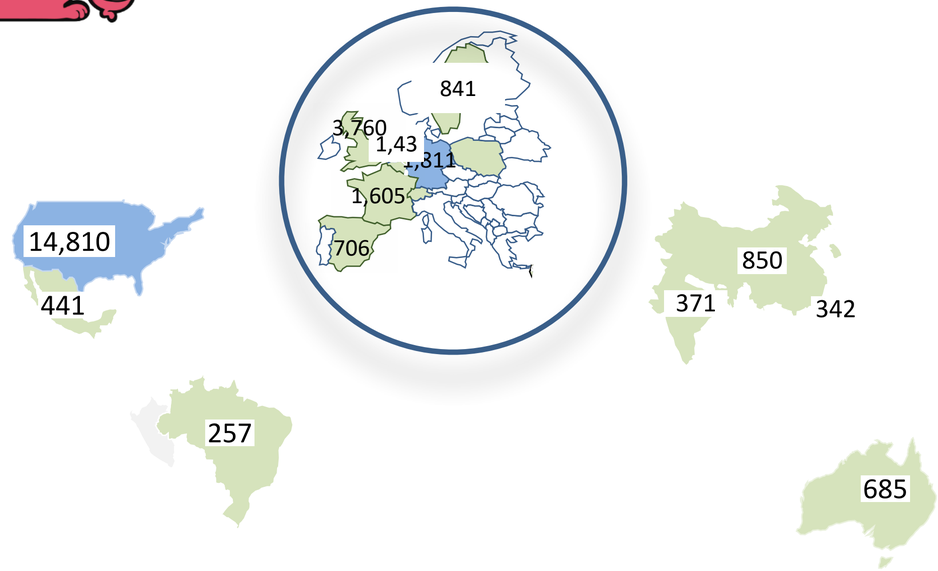
- **Trust**
  - Good Accuracy (e.g., ML)
  - Consistency Guarantees and Fault Tolerance
- **Performance**
  - Speed and Energy Efficiency
  - Hardware Acceleration Support
- **Ease of Use**
  - Simple user Interface (e.g., LLMs)
  - Development Flexibility

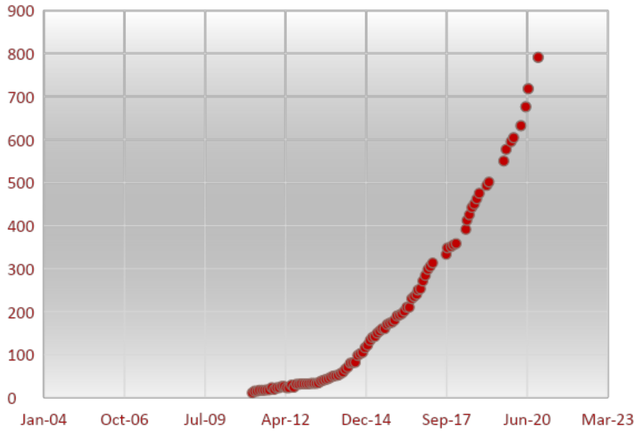


Example - Apache Flink

- **Trust**
  - First system that offered Exactly-Once-Processing Guarantees
- **Performance**
  - Ultra-Low Event-at-a-time latencies
- **Ease of Use**
  - Fluid Java-Scala-Python API with Automated State Management

Example - Apache Flink



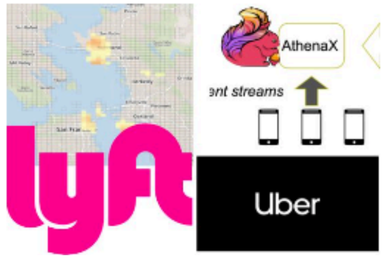


Flink Bidragsgivare

produktions installationer

Netflix, Uber, Lyft, John Deere, Microsoft, Telefonica, King, Alibaba, SK Telecom, Airbnb, Huawei, ING ...

Nya affärsmodeller



dynamisk bilprissättning

skötsel av levande skörd

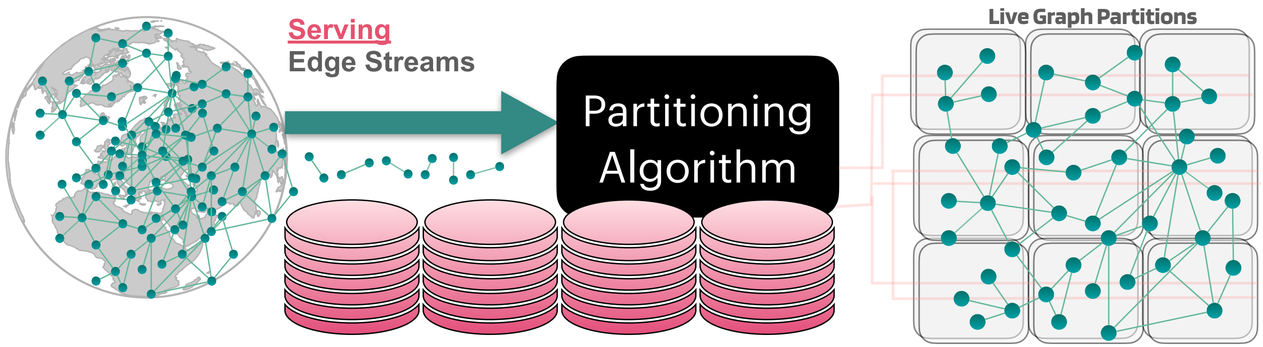
ACM SIGMOD Systems Award 2023
“Nobel” of Data Systems

Advanced Digitalization through Data Streaming

Advanced Digitalization through AI?

AI vs Best System Algorithms

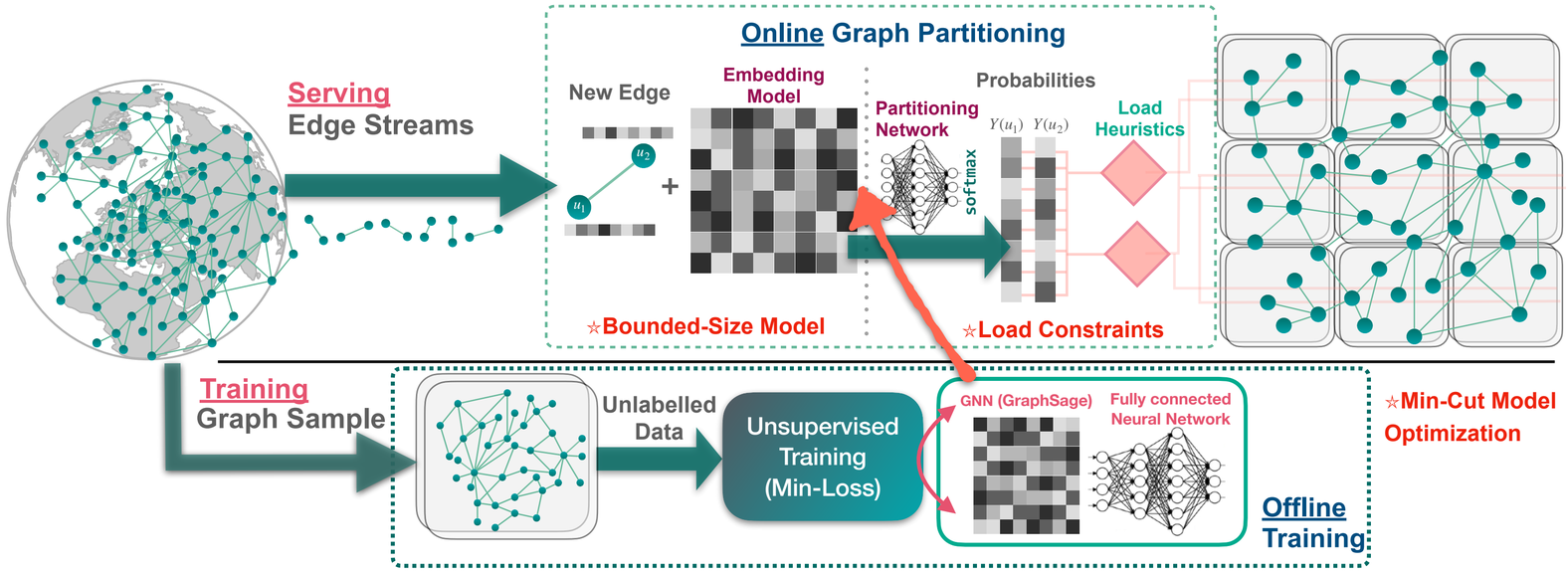
Streaming Graph Partitioning with Boundless State (state of the art algorithms)



Problems

- Load Balancing Focused Algs (e.g., Hash) : Fail to provide Min-Cut
- Min-Cut + Balancing Focused Algs :
  - Leads to State Explosion - **Impractical**
  - Different Partitioner for different graph - **Non-Generalizable**

Streaming Graph Partitioning with Bounded State using Node Embeddings



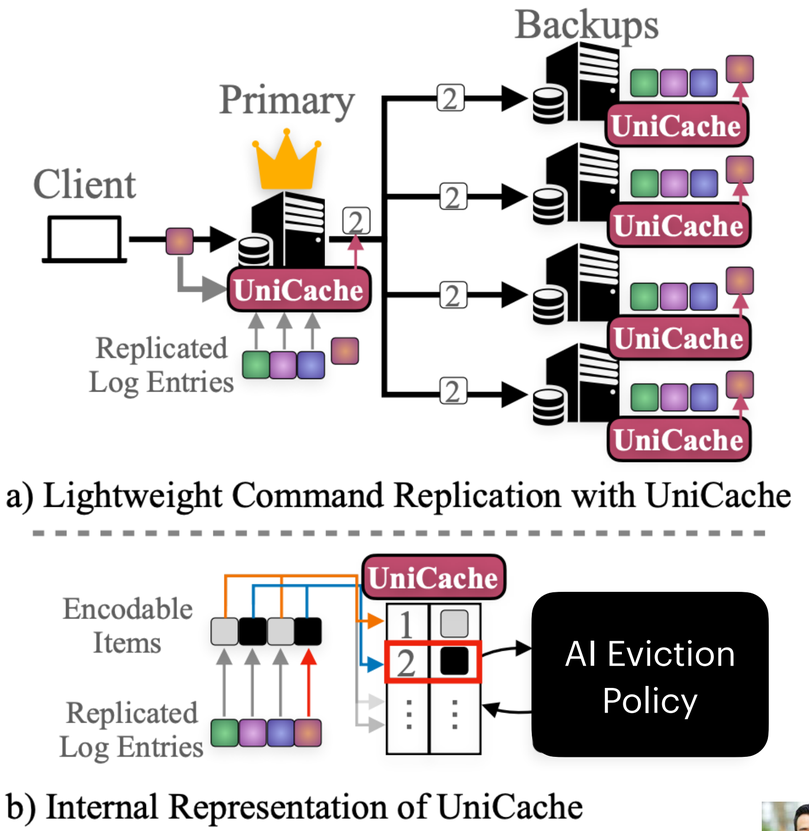
Findings

- **Practical:** Nothing more than a embedding needed
- **Generalizable:** One model can partition different graphs

Michał Zwolak, Zainab Abbas, **Sonia Horchidan**, **Paris Carbone**, and Vasiliki Kalavri. 2022. *GCNSplit: bounding the state of streaming graph partitioning*. In Proceedings of the Fifth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management (aiDM '22).

AI vs Coordination

- Reduced by 30-60% Service Replication workloads
- **Compression** achieved through an deterministic Policy
  - LeCar*, an RL-driven cache eviction policy that optimizes between using LFU and LRU uniformly across servers
  - **No coordination needed (it is inferred by taking the same AI-informed decisions)**



Ng, Harald, Kun Wu, and Paris Carbone. "UniCache: Efficient Log Replication through Learning Workload Patterns." (EDBT 2023).


* Giuseppe Vietri, Liana V. Rodriguez, Wendy A. Martinez, Steven Lyons, Jason Liu, Raju Rangaswami, Ming Zhao, and Giri Narasimhan. [n.d.]. Driving Cache Replacement with ML-based LeCaR. USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 18)

Advanced Digitalization through AI?

- **Trust**
  - ~ Model Accuracy **better** than SoA algorithms on key function
  - Challenge: Trade off **inference speed** vs **accuracy** vs **model size**.
- **Performance**
  - ~ Level of Hardware Acceleration (tensor op. execution).
  - Challenge: Support non-specialized **hardware**.
- **Ease of Use**
  - ~ Development and Runtime Costs.
  - Challenge: Seamless integration with critical logic (non-AI)

The CASTOR Research Center

Data-Driven Development

KTH: Paris Carbone

SAAB: Joakim Lindén

Ericsson: Andreas Ermedahl (S&T)

CASTOR
Software Research Centre

https://www.kth.se/castor

"Aligning Interests Across Research and Practice in Critical Industries"

Other Research Avenues

- Secure end-to-end Software Systems
- Software System Design
- Resilient Programmable Networks

Critical System Needs

- **Goal**: Bring AI capabilities to proprietary critical hardware.
  - **Multi-Sensor Processing (SAAB)**
    - Key Features: ML inference, Isolation, Situational Dynamic Reallocation
  - **Radio Resource Management (Ericsson)**
    - Key Features: Optimisation, Interoperability, Runtime Flexibility
- Meta-**Goal**: Make solution transferable to future critical hardware releases

Paris Carbone (Main PI)

Andreas Ermendal

Joakim Lindén
Karl Lunden
Emil Karlsson

Masoud Daneshtalab
Håkan Forsberg

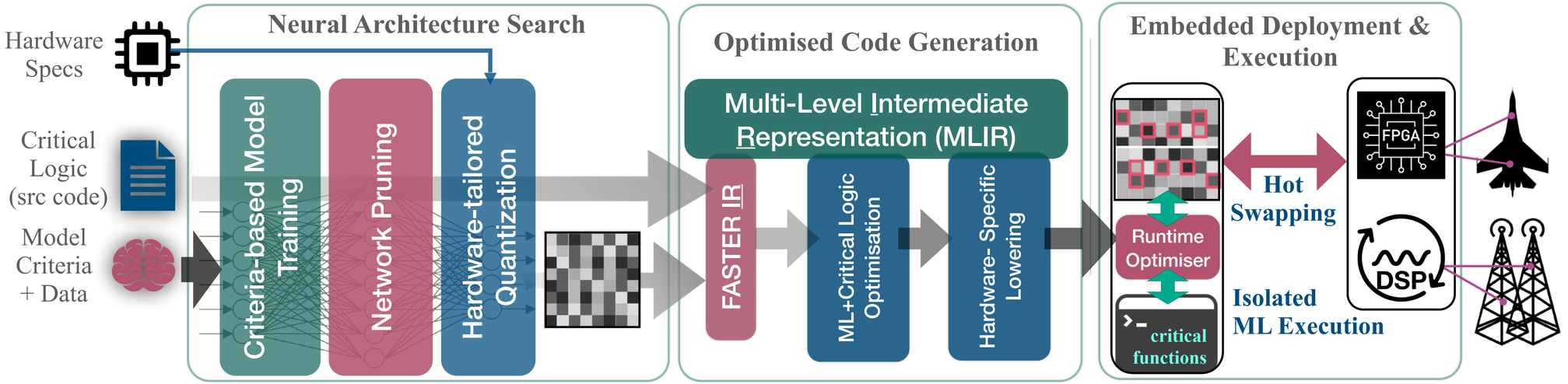
Björn Forsberg
Frej Drejhammar

Ola Tiverman
Hans Salomonsson

Advanced Digitalization through AI?

- **Trust**
  - Trade off model **speed** vs **accuracy** vs **size** (hardware constraints)?
  - **Neural Architecture Search [MDU, Embedl, Ericsson, SAAB]**
- **Performance**
  - Support critical, non-specialized **hardware**?
  - **Multi-Stage Compilation and Code Generation [RISE, KTH, Ericsson]**
- **Ease of Use**
  - Seamless integration with critical logic (non-AI)
  - **Isolated Execution for Hybrid (AI/other) workloads [RISE, MDU, Embedl, SAAB]**

Methodology



Optimisation/Tuning

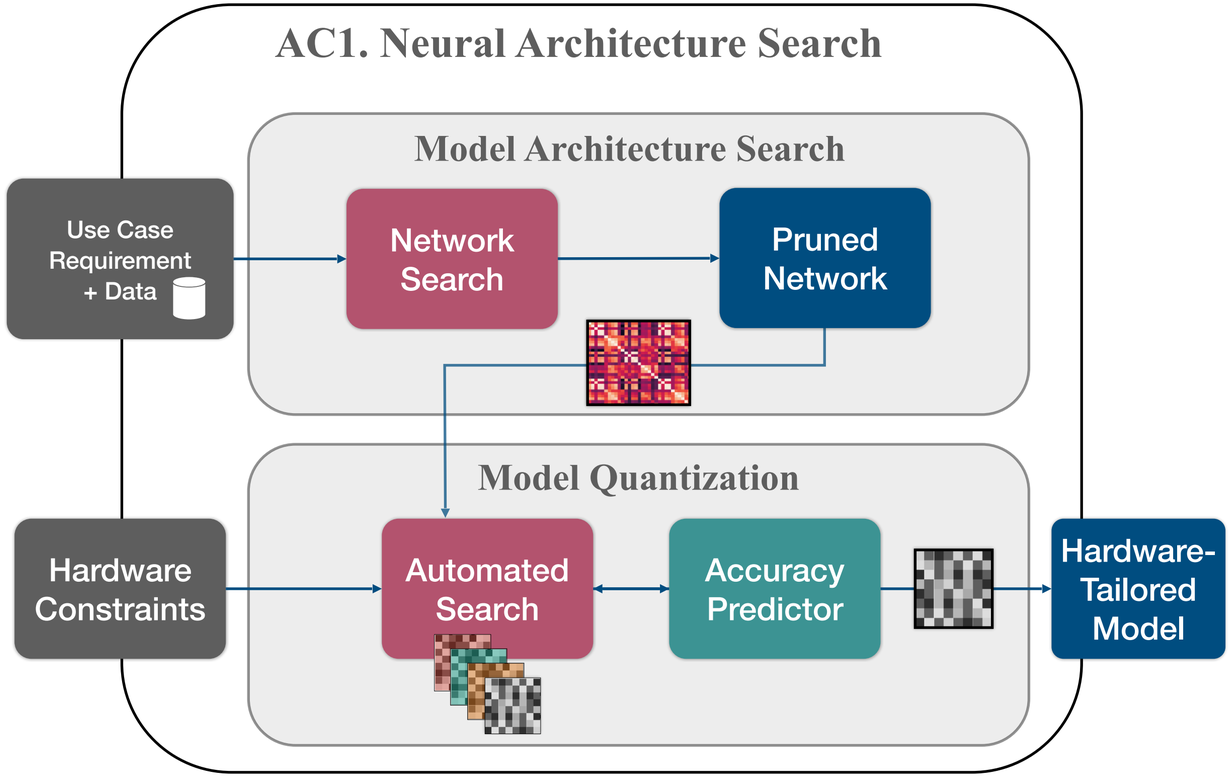
ML Inference

Isolation

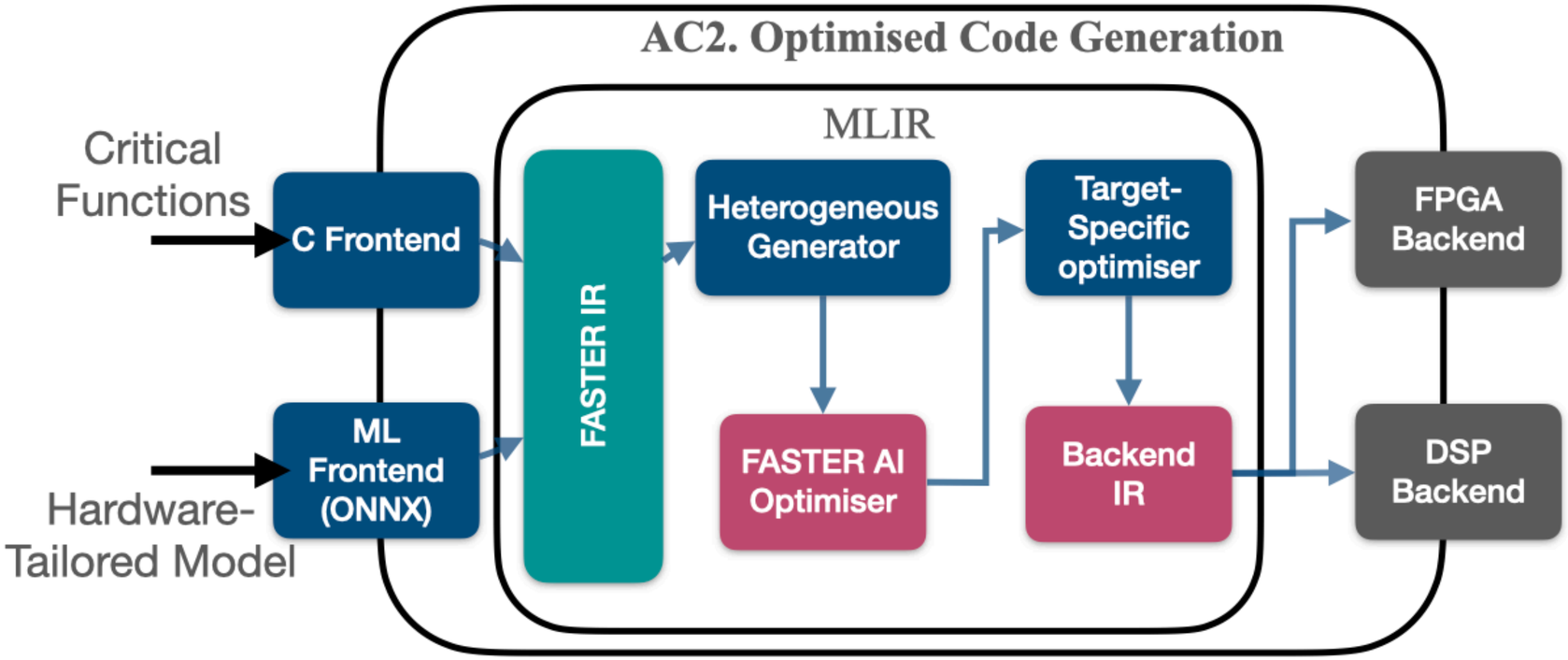
Interoperability

Dynamism

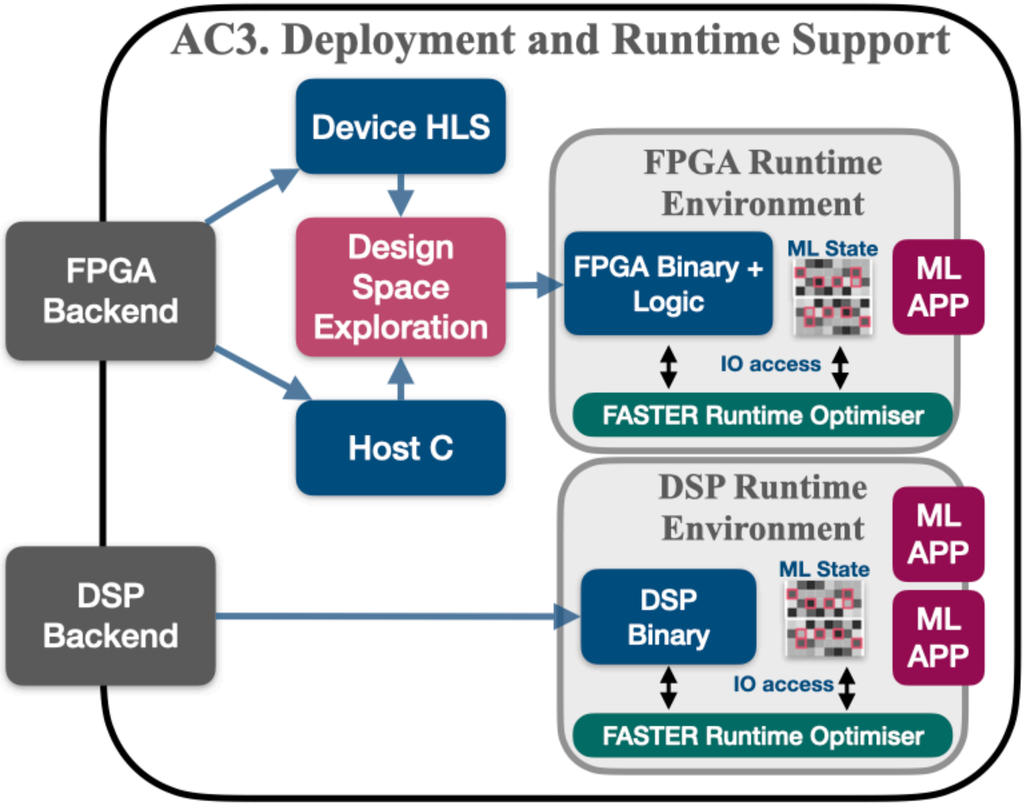
1. Neural Architecture Search



2. Optimised Code Geration



3. Deployment & Isolated Execution



More Project Details

- 2 Years
- ~16MSEK budget

- Prototyping on different FPGAs.
- Evaluation of Use-Cases on Ericsson-SAAB specialised hardware.
- Non-propriatory libs will be open sourced.