Model-based Timing Analysis and Deployment Optimization for Heterogeneous Multi-Core Systems using Eclipse APP4MC

Lukas Krawczyk, Mahmoud Bazzal, Ram Prasath Govindarajan, and Carsten Wolff

Institute for the Digital Transformation of Application and Living Domains
Dortmund University of Applied Sciences and Arts
44227 Dortmund, Germany
lukas.krawczyk@fh-dortmund.de

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Agenda

- Introduction
- Motivational example
- APP4MC
- Integration
- End-to-End reaction latency analysis
- Optimization
- Case Study
- Conclusion and outlook

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Introduction

- Increasing demands on automotive computing platforms driven by new automotive functionalities
- Set of about 80 ECUs in todays cars will be reduced to about 10 high-performance units
- Centralized computing platforms consisting of sophisticated heterogeneous accelerators

Source: Bosch

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Introduction

- Heterogeneous Hardware
  - Different models of computation
  - Variety processing-unit specific scheduling strategies
- Heterogeneous functional Domains
  - Mixed levels of criticality
  - Freedom from interference
- COTS Hardware
  - Limited to no capabilities for adjusting hardware

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

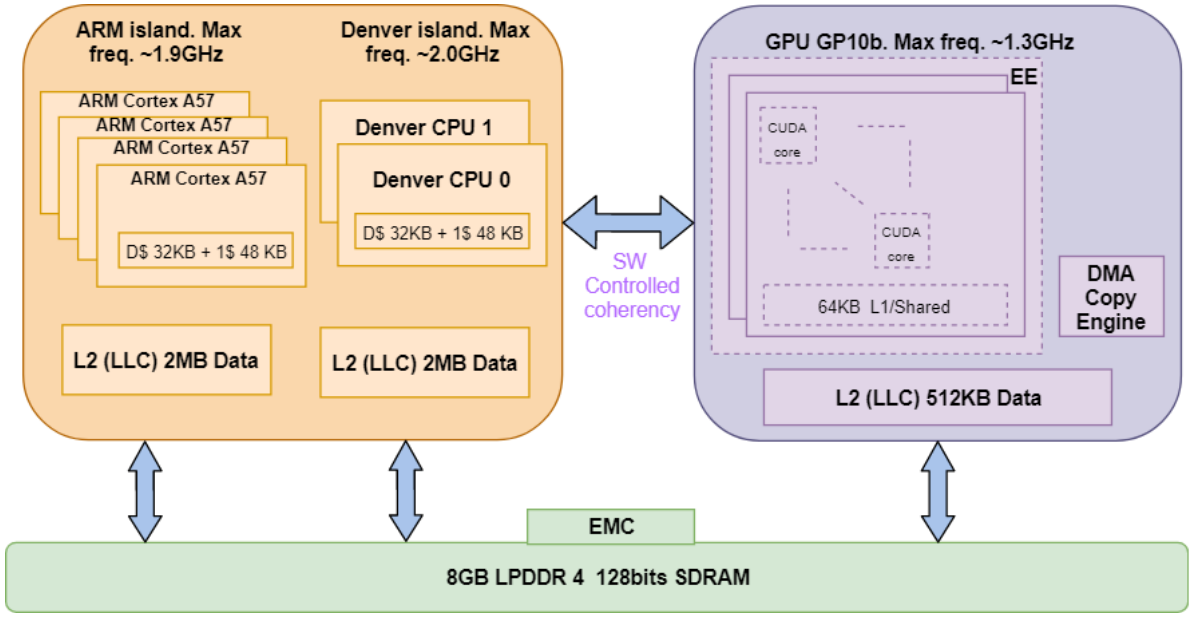
Contributions

- Model-based approach for deploying software to heterogeneous hardware while ensuring end-to-end reaction latency constraints.
  - Applicable in early design phases
  - Response Time Analysis
  - Design Space Exploration using Genetic Algorithm
  - Industrial Case Study on an heterogeneous COTS hardware platform

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Motivational Example - Hardware

Architecture of Jetson TX2

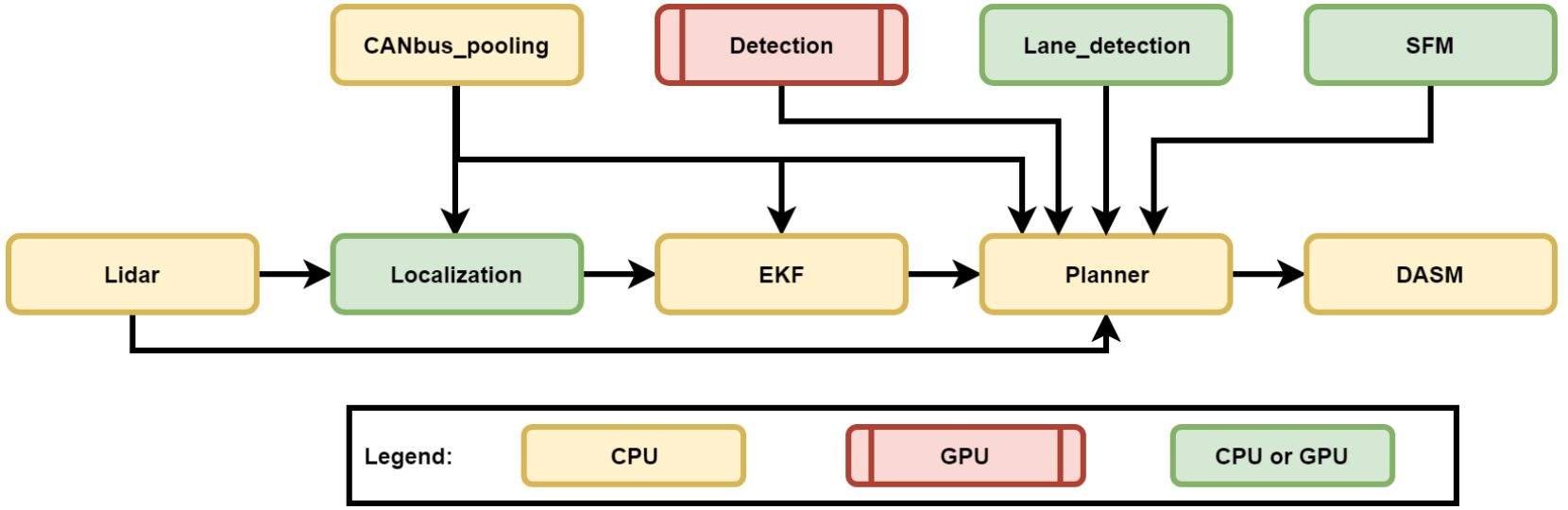


Specification

- NVIDIA Pascal™ Architecture GPU
- GPU - NVIDIA Pascal™, 256 CUDA cores
- CPU - HMP Dual Denver 2/2 MB L2 +Quad ARM® A57/2 MB L2
- Memory - 8 GB 128 bit LPDDR4

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

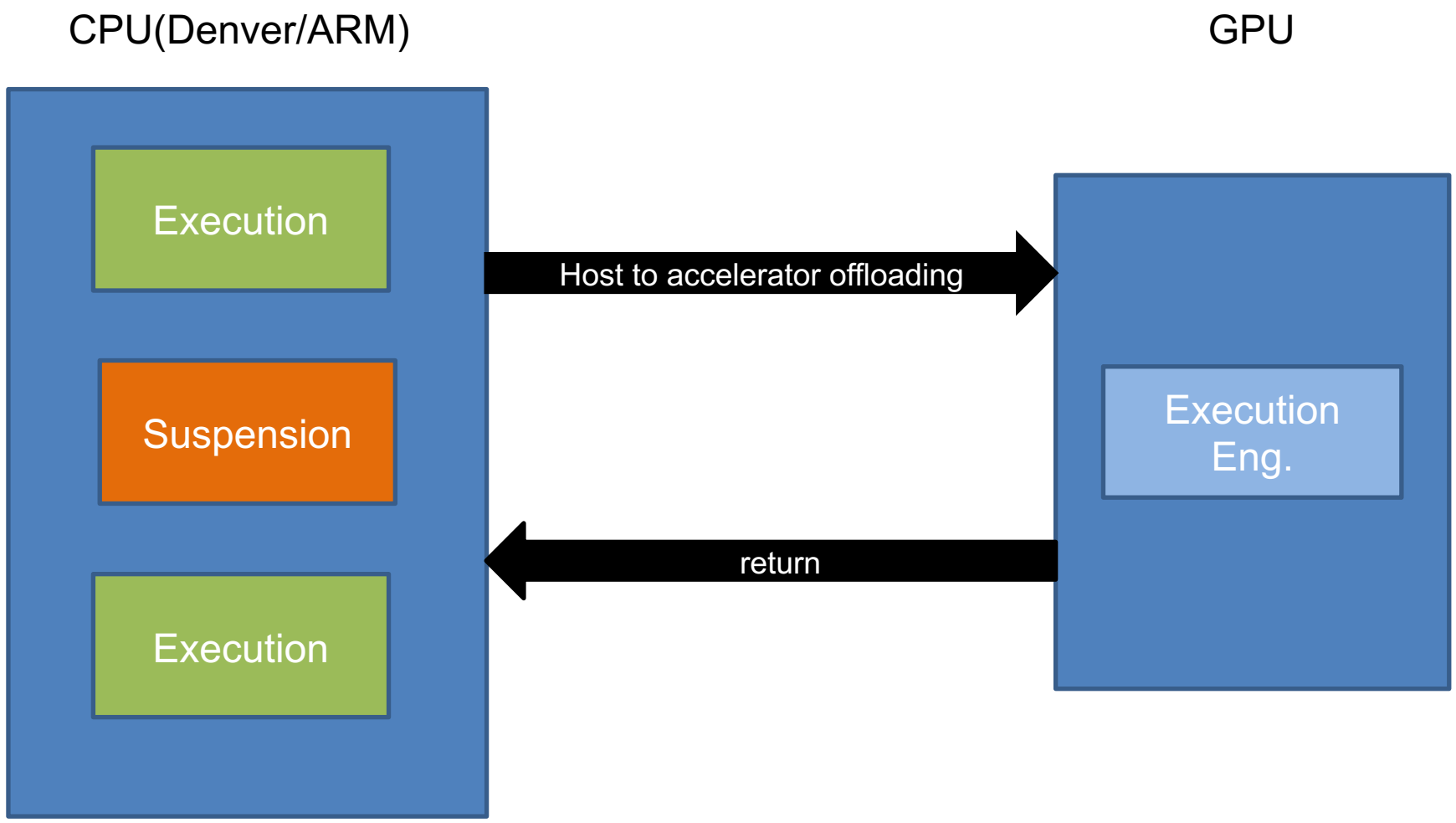
Motivational Example - Software



Task chains				
Lidar	➔ Localization	➔ EKF	➔ Planner	➔ DASM
CAN	➔ Localization	➔ EKF	➔ Planner	➔ DASM
SFM	➔ Planner	➔ DASM		
Lane_detection	➔ Planner	➔ DASM		
Detection	➔ Planner	➔ DASM		

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Motivational Example – Self suspension

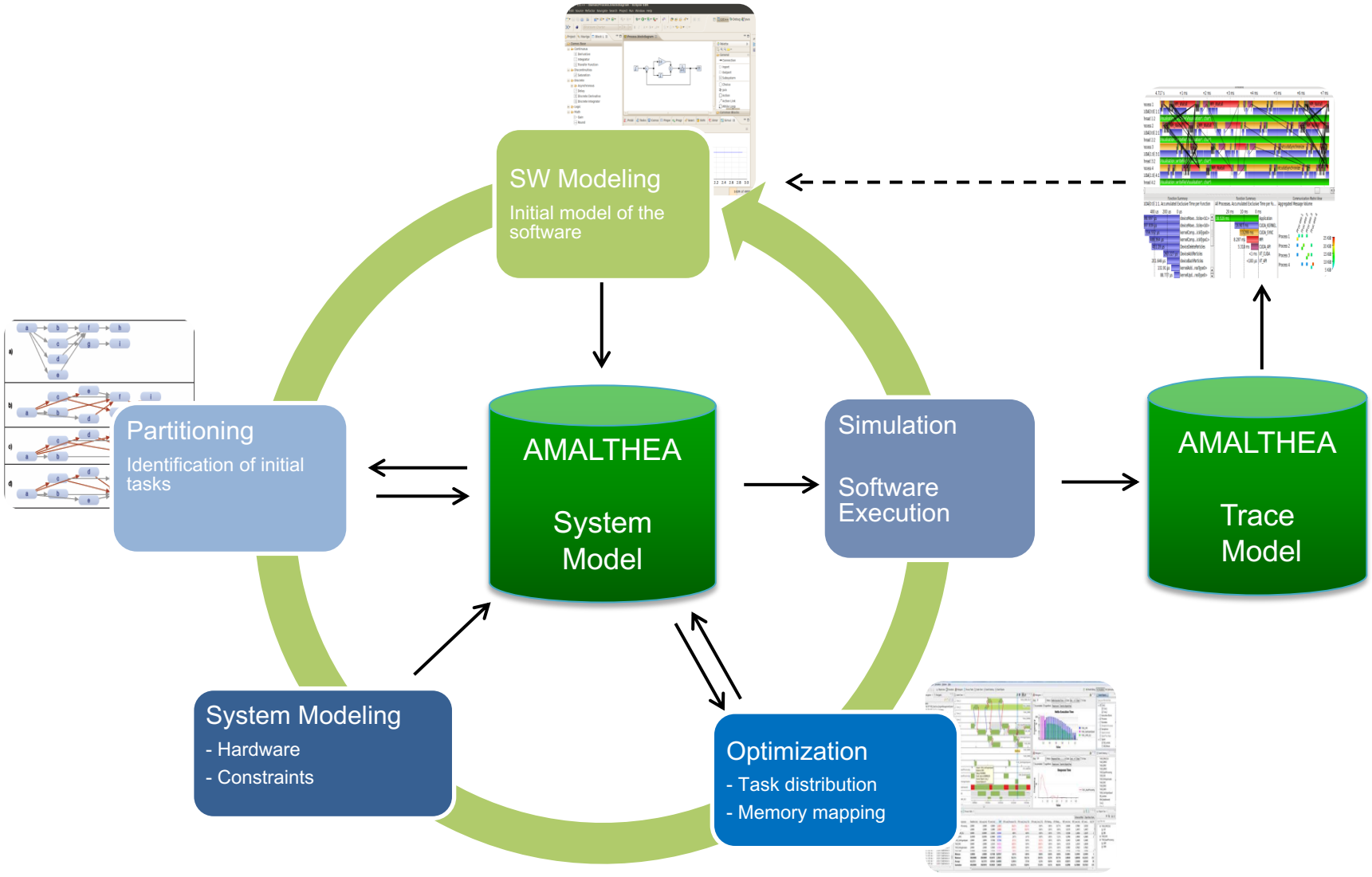


SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Fachhochschule Dortmund
University of Applied Sciences and Arts
IDiAL Institute for the Digital Transformation of Application and Living Domains

APP4MC

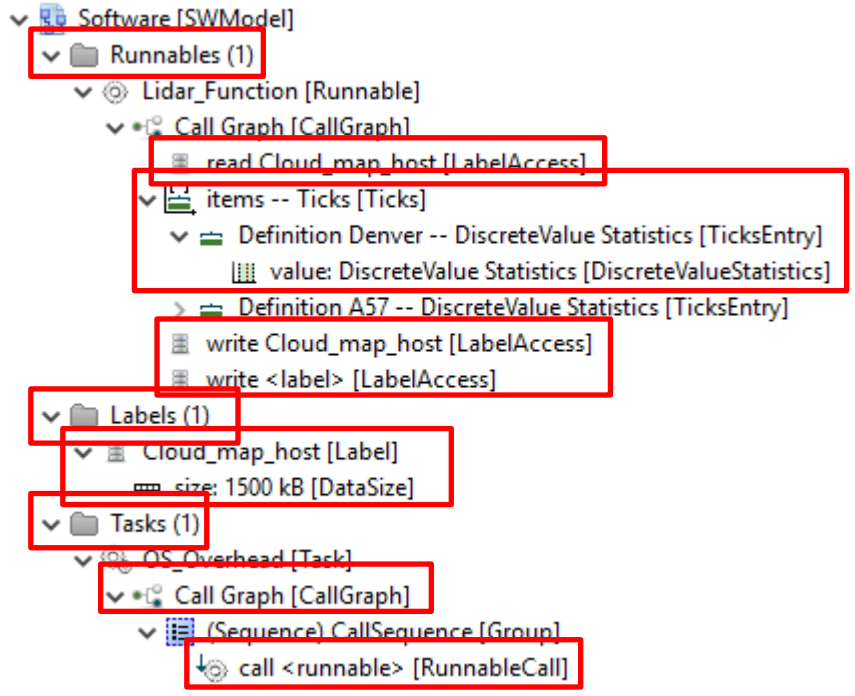
PANORAMA



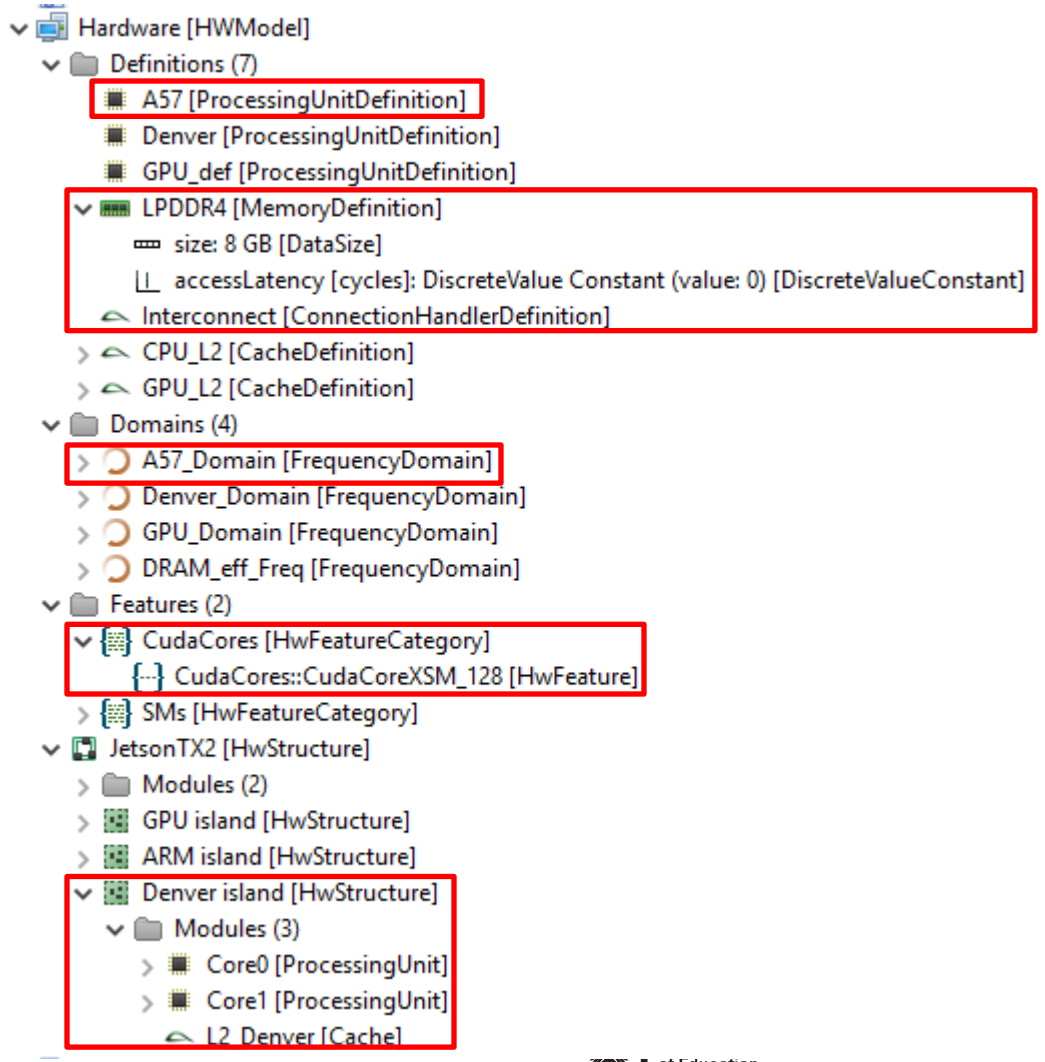
SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

APP4MC

SW Model

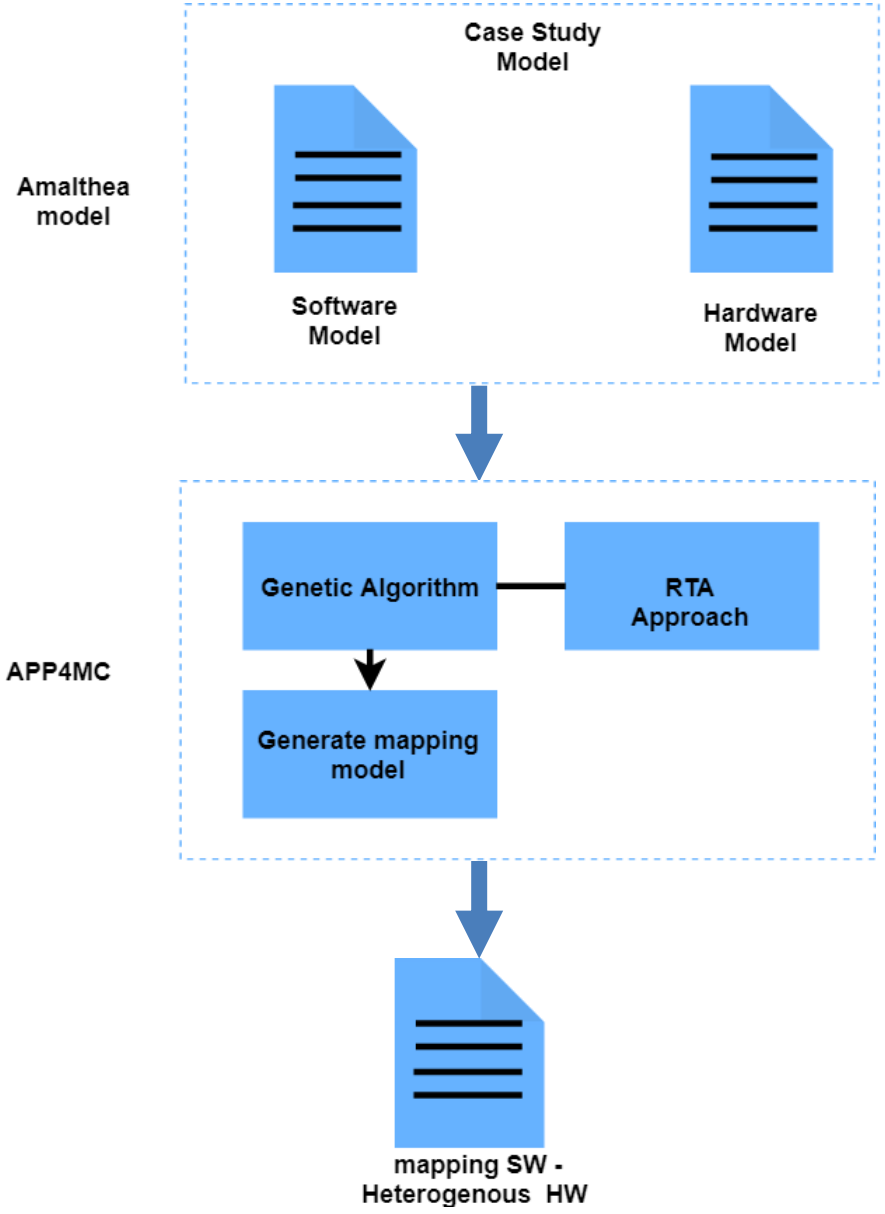


HW Model



of Education and Research
ITEA3 - 17003
EUREKA

Integrated Approach



Amalthea model is constructed out of system design information.

APP4MC is used to implement the real-time analysis and deployment optimization approach.

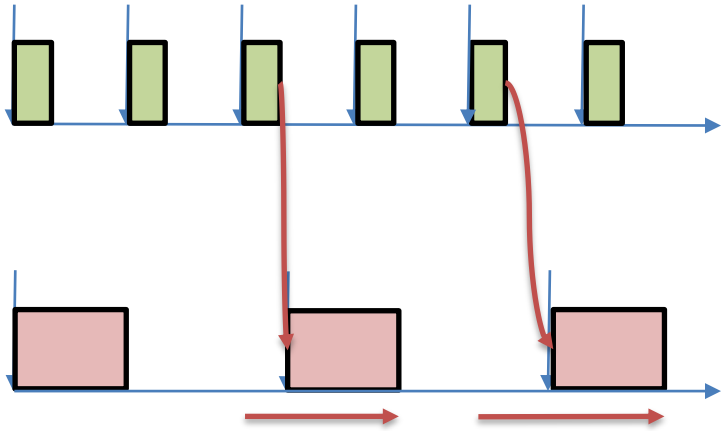
Amalthea model is updated with mapping information.

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003

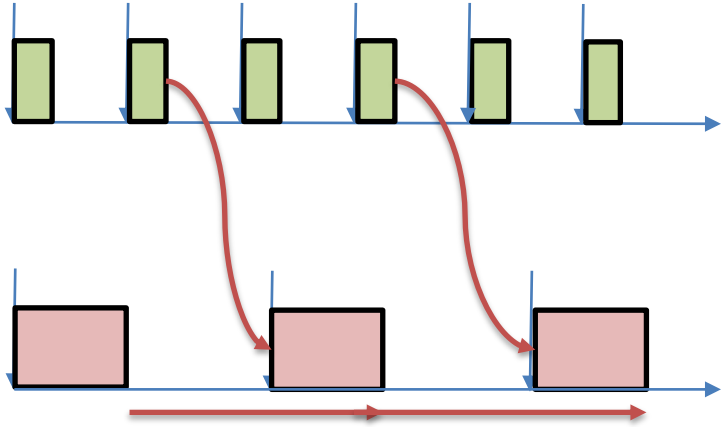
End-to-End Reaction Latency Analysis

PANORAMA

Implicit



LET (Logical Execution Time)

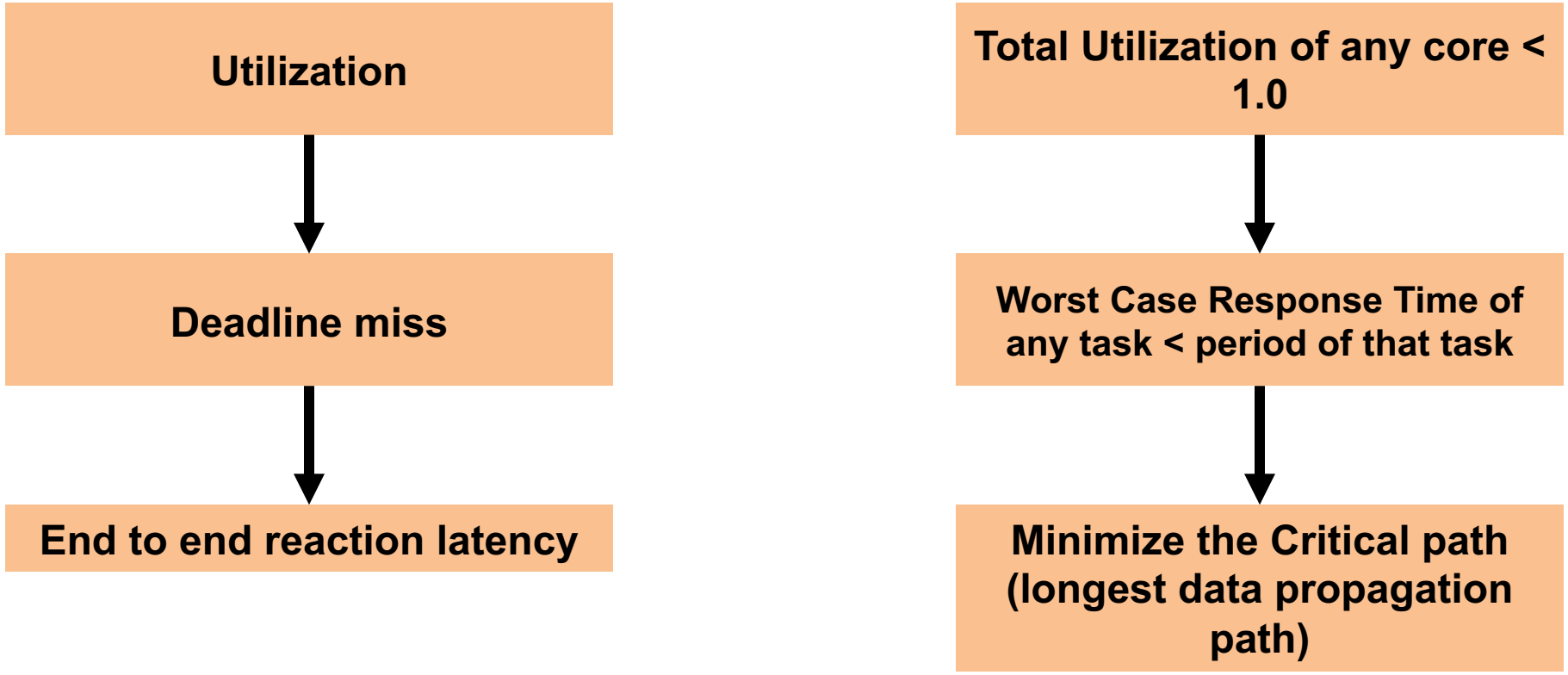


- LET communication ➔ deterministic data propagation points
- Implicit communication ➔ shorter end-to-end reaction latency

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Optimization - goals

- Evaluation of solutions is computationally complex
- Multi-phased optimization strategy



SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Optimization – goals

- Utilization

$$\forall \mathcal{P}_\rho \in \mathcal{PU}, \qquad \sum_{\tau_i} U_i \leq 1.0$$

- No deadline miss (Worst case response time)

$$\forall \tau_i \in T, \qquad \mathcal{R}_i^+ \leq P_i$$

- Worst case end to end latency

$$\max_{\sigma \epsilon S}(L^{TC}(\sigma))$$

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

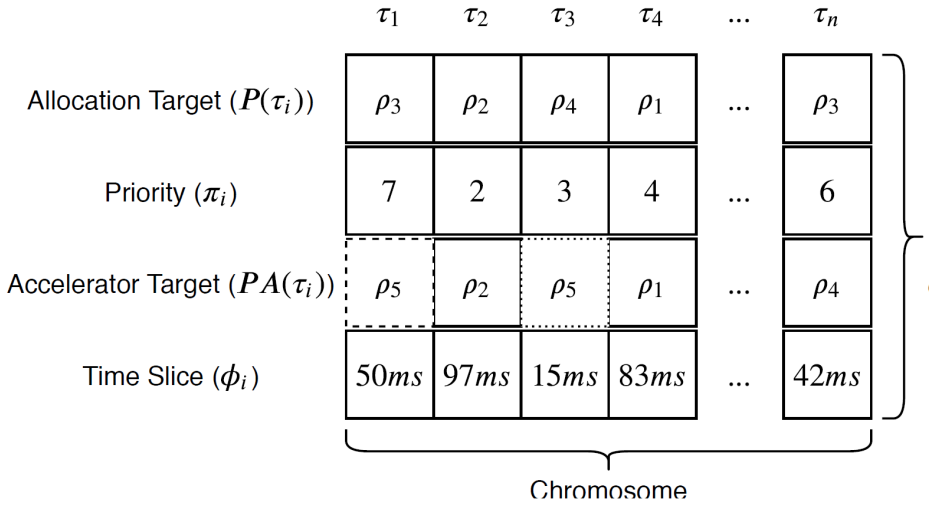
Optimization – degrees of freedom

- Allocation target
  - The processing unit executing a task.
- Priority
  - Higher priority tasks will preempt lower priority tasks
- Accelerator target
  - Defines the target which should execute the accelerable content of an executable
- Time slice
  - Amount of time given periodically to execute a task on the accelerator

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Optimization – encoding

- Allocation target
  - All processing units except accelerators
  - Offloadable tasks can also be allocated
- Priority
  - Number of priorities is the number of tasks executing on a core.
  - Priorities are unique.
- Accelerator target
  - For offloadable tasks only the runnable to be accelerated is offloaded.
- Time slice
  - Only valid for an accelerator bound executable.



SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

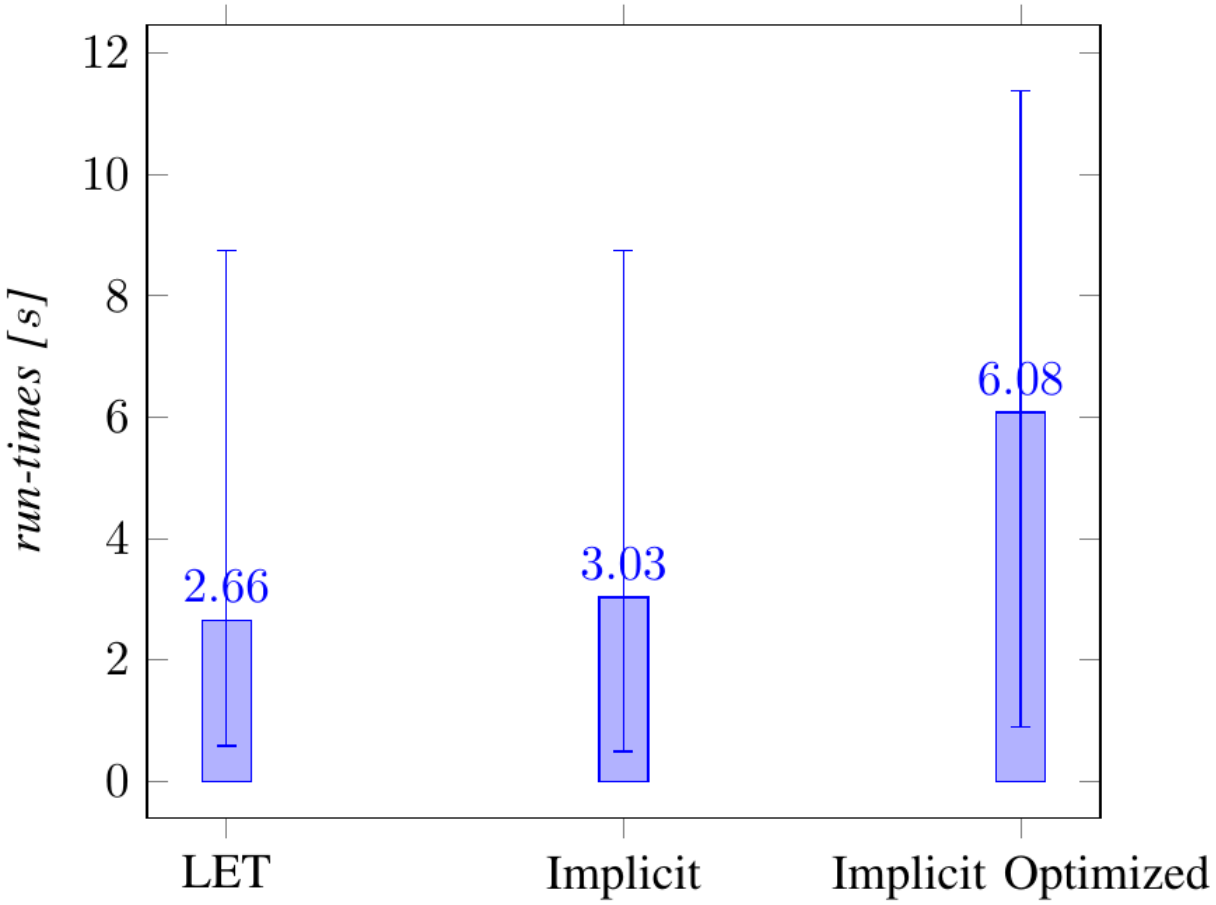
Case Study – Configuration

- GA configuration:
  - 500 initial population
  - Mutation rate of 5%
- Termination criteria:
  - Implicit communications: 2000 generations of steady fitness values.
  - LET: first feasible solution (no deadline miss)
- Hardware configuration
  - Intel Core i5-3570K quad-core CPU @ 3.4 GHz

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Case Study – run time

- LET
  - Feasible solution after ~3s (avg.)
- Implicit
  - Feasible solution after ~3s (avg.)
- Implicit optimized
  - Optimal solution after ~6s (avg.)
- Similar runtime to other approaches for the same case study.



SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Case Study – results

Task Chain	LET end-to-end	Implicit end-to-end
σ_1	876 ms	460.9 ms
σ_2	845 ms	436.9 ms
σ_3	63 ms	59.9 ms
σ_4	93 ms	83.9 ms
σ_5	225 ms	176.9 ms

Name	P	π	C^+	$\lambda \cdot \mathcal{A}^+$	O	J	R^+
Core 0 (Denver)							
Planner	12	–	11.2	0.8	0	0	12.0
Core 1 (Denver)							
SFM	33	H	27.8	2.4	0	0	30.2
Detection_Pre	200	L	3.2	2.4	0	0	65.9
Detection_Post	200	L	0.9	0.9	108.3	11.0	151.3
Core 2 (A57)							
Lane_detection.	66	–	51.0	6.9	0	0	58.9
Core 3 (A57)							
DASM	5	H	1.9	0	0	0	1.9
OS_Overhead	100	L	50	0	0	0	79.9
Core 4 (A57)							
Lidar_Grabber	33	–	13.7	12.0	0	0	25.7
Core 5 (A57)							
CAN Pooling	10	H	0.6	0	0	0	0.6
EKF	15	M	4.8	0	0	0	5.4
Localization_Pre	400	L	8.9	10.3	0	0	36.0
Localization_Post	400	L	8.7	0	117.1	240.1	372.0

Name	P	ϕ	C^+	$\lambda \cdot \mathcal{A}^+$	O	J	R^+
GP10B (iGPU)							
Localization	400	2.8	124.0	0.3	4.1	32.0	357.3
Detection	200	116.5	116.0	0.5	2.9	63.0	119.3

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Conclusion and Outlook

- A combined
  - Genetic Algorithm based Design Space Exploration approach
  - Response Time Analysis for heterogeneous hardware applying RMS and WRR scheduling
  - Fully integrated into App4MC

- Results demonstrate the applicability of our approach for industrial problems with similar run-times as other approaches while delivering better bounds.

- Future work
  - Validate Results on real hardware
  - Evaluate performance on larger problems
  - Consider further blocking factors

SPONSORED BY THE
Federal Ministry of Education and Research
ITEA3 - 17003
EUREKA

Acknowledgements

The research leading to these results has received funding from the Federal Ministry for Education and Research (BMBF) under Grant 01IS18047D in the context of the ITEA3 EU-Project PANORAMA.

SPONSORED BY THE

Federal Ministry of Education and Research

https://www.panorama-research.org

info@panorama-research.org