35th Euromicro Conference on Real-Time Systems 2023 – Vienna, Austria

Industrial Challenge

ANALYSIS AND OPTIMIZATION OF REAL-TIME APPLICATIONS RUNNING ON HETEROGENEOUS HARDWARE

Iosu Gomez, Unai Díaz de Cerio, Jorge Parra
Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), Arrasate-Mondragon, Spain

Juan M. Rivas, J. Javier Gutiérrez
Universidad de Cantabria, Santander, Spain

Challenge

- Analysis of performance bounds (Response Time and WCET Analysis)
- Optimization:
  - Data-flow analysis
  - Scheduling
  - Resource mapping
  - Shared resource interference

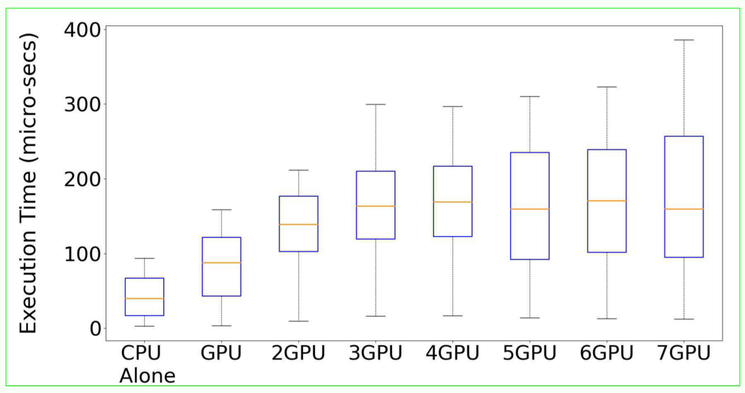
Early Stage proposal

Methodology based on two main aspects:

- Partition scheduling with strong temporal and spatial isolation.
- Modeling, schedulability analysis and optimization for distributed multipath end-to-end flows (DAGs).

Memory Interference

- WCET has a strong dependency on memory interference from CPUs-GPUs.
- Difficult to obtain precise measures of WCET values (interference may lead to very high upper bound of WCETs).
- WCET increases by a factor of 4 and the variability also widens.



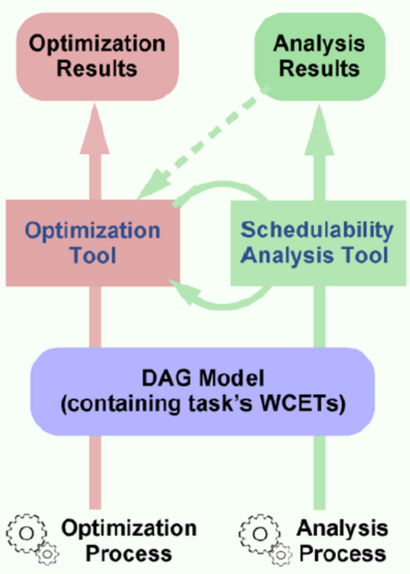
Impact of GPU on the CPU execution time on a Jetson AGX Xavier

MAST

- Modeling and analysis tool for Real-Time systems. [1][2]
- Aligned with OMG MARTE standard.
- MAST model supports offset-based response time analysis techniques.
- Available at: https://mast.unican.es/

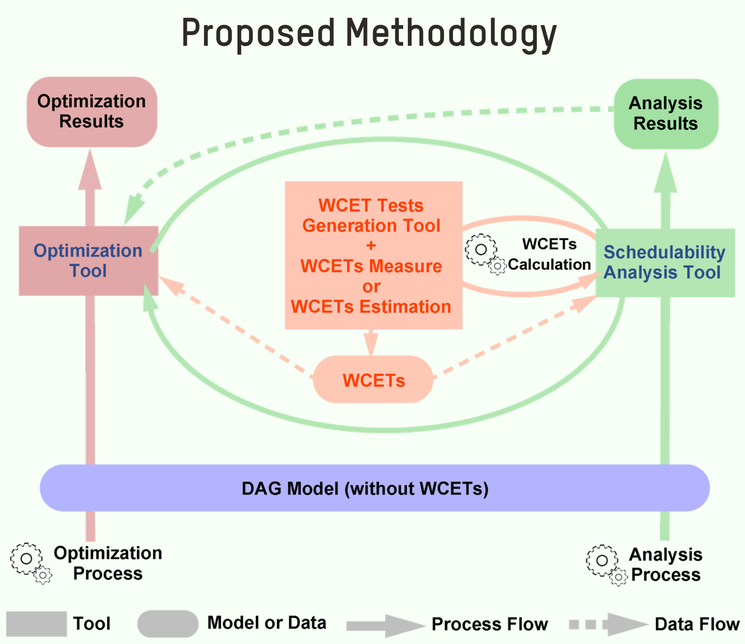
Analysis and Optimization Methodology

Traditional Methodology



WCETs are assumed known and constant

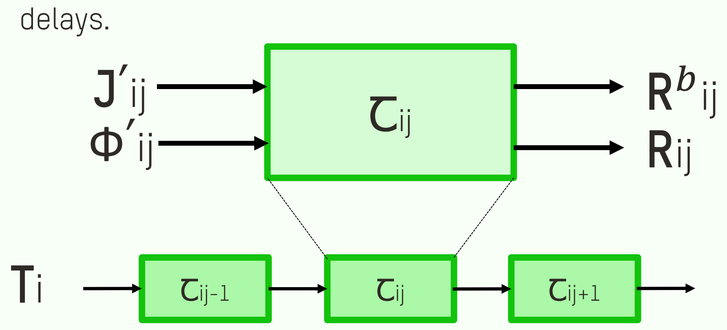
Proposed Methodology



WCETs must be obtained as part of the analysis and optimization process

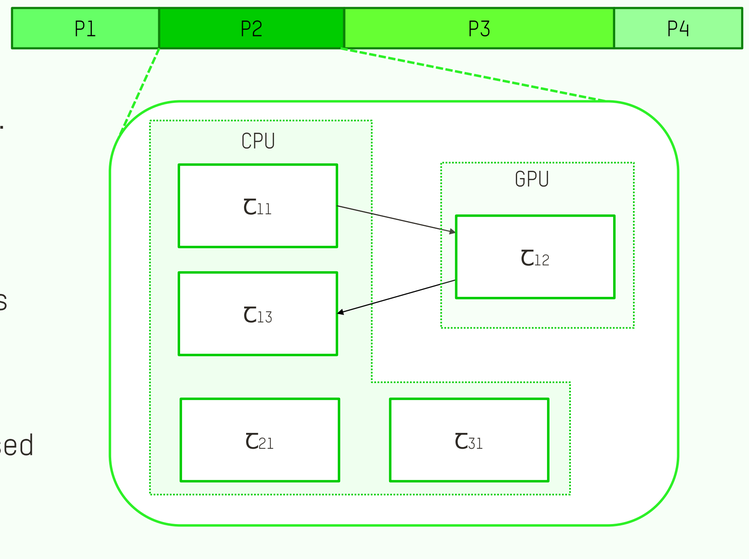
Task Model

- System composed by e2e flows, which can be activated by either periodic or sporadic events with a minimum interarrival time.
- Each task in e2e flow is activated by a minimum and a maximum release time offset Φ_{ij}.
- Equivalent jitters and offsets take into account the information from previous task in the e2e flow. [3]
- GPU operations can be modeled as activities, offsets or delays.



Model for Partitioned System

- A primary scheduler schedules temporal partitions located in different CPUs in a cyclic manner.
- A secondary FP scheduler allows optimizing RT Tasks groups.
- Assigned time window influences the accuracy of the analysis
- Assigned time window enables controlling the interference caused by GPU tasks.



Response Time Analysis

The proposed methodology is based on previous existing analysis techniques (available in MAST).

- Offset-based analysis algorithms [4][5].
- Offset-based analysis with precedence relations among tasks presented in [6].
- Offset-based analysis for hierarchical scheduling: linear [7] and multipath [8].

These techniques directly support e2e flows with GPU tasks:

- GPU is considered as grey box: Response times are estimated by any available technique, or they are directly measured and then integrated in the model as delays.
- The analysis with precedence relations [4] can only be applied to linear e2e flows while the other offset-based techniques [3][6] work also for DAGs [1].

References

[1] M. González Harbour, J. J. Gutiérrez García, J. C. Palencia Gutiérrez and J. M. Drake Moyano, "MAST: Modeling and analysis suite for real time applications", Proc. 13th Euromicro Conf. Real-Time Syst., pp. 125-134, 2001.
[2] M. González Harbour, J. J. Gutiérrez, J. M. Drake, P. López Martínez and J. C. Palencia, "Modeling distributed real-time systems with MAST 2", J. Syst. Archit., vol. 59, no. 6, pp. 331-340, Jun. 2013.
[3] J.M. Rivas, J. J. Gutiérrez, J.C. Palencia, and M. González Harbour. "Schedulability Analysis and Optimization of Heterogeneous EDF and FP Distributed Real-Time Systems," Proc. of the 23th Euromicro Conference on Real-Time Systems, Porto (Portugal), July 2011.
[4] J. C. Palencia and M. González Harbour, "Schedulability analysis for tasks with static and dynamic offsets", Proc. 19th IEEE Real-Time Syst. Symp., pp. 26-37, Dec. 1998.
[5] J. C. Palencia, M. González Harbour, J. J. Gutiérrez and J. M. Rivas, "Response-time analysis in hierarchically-scheduled time-partitioned distributed systems", IEEE Trans. Parallel Distrib. Syst., vol. 28, no. 7, pp. 2017-2030, Jul. 2017.
[6] J. C. Palencia Gutiérrez and M. González Harbour, "Exploiting Precedence Relations in the Schedulability Analysis of Distributed Real-Time Systems", Proc. of the 20th Real-Time Systems Symposium, pp. 328-339, Dec. 1999.
[7] J. Mäki-Turja and M. Nolin, "Efficient implementation of tight response-times for tasks with offsets", Real-Time Syst., vol. 40, no. 1, pp. 77-116, Oct. 2008.
[8] A. Amurrio, E. Azketa, J. J. Gutiérrez, M. Aldea, and M. González Harbour, "Response-time analysis of multipath flows in hierarchically-scheduled time-partitioned distributed real-time systems," IEEE Access, vol. 8, pp. 196700–196711, 2020.

Acknowledgement

This work was partially supported by MCIN/ AEI /10.13039/501100011033/ FEDER "Una manera de hacer Europa" under grants PID2021-1245020B-C42 and PID2021-1245020B-C44 (PRESECREL).