

# Analysis and Optimization of Real-Time Applications Running on Heterogeneous Hardware

Iosu Gomez

Ikerlan Technology Research Centre,  
Basque Research and Technology  
Alliance (BRTA),  
Arrasate-Mondragon, Spain  
Universidad de Cantabria  
Santander, Spain  
iosu.gomez@ikerlan.es

Unai Díaz-de-Cerio

Ikerlan Technology Research Centre,  
Basque Research and Technology  
Alliance (BRTA),  
Arrasate-Mondragon, Spain  
udiazcerio@ikerlan.es

Jorge Parra

Ikerlan Technology Research Centre,  
Basque Research and Technology  
Alliance (BRTA),  
Arrasate-Mondragon, Spain  
jparra@ikerlan.es

Juan M. Rivas

Universidad de Cantabria  
Santander, Spain  
rivasjm@unican.es

J. Javier Gutiérrez

Universidad de Cantabria  
Santander, Spain  
gutierjj@unican.es

**Keywords**—*real-time, scheduling, schedulability analysis, optimization, distributed DAGs, heterogeneous hardware*

## ABSTRACT

This is an early stage proposal of a methodology that can be applied to the data-flow analysis of an application as the one described in [1], where software can be decomposed in task DAGs, thus it answers in part to the proposed industrial challenge. The methodology builds on two main aspects: (1) using an ARINC-like [2] scheduler, i.e., the partitioning concept can provide applications with strong temporal and space isolation (in addition, fixed priorities are allowed inside a partition at a second scheduling level), and (2) the modelling and schedulability analysis technique for distributed multipath flows (DAGs) proposed in [3]. This technology can be applied to multicore processors (shared bus for global memory plus core-local memory) where the worst-case execution time (WCET) of tasks can be measured or estimated through worst-case assumptions on the memory contention representing a bounded impact in the response times [4]. Once the WCETs have been obtained, the methodology shown in Figure 1 can be applied.

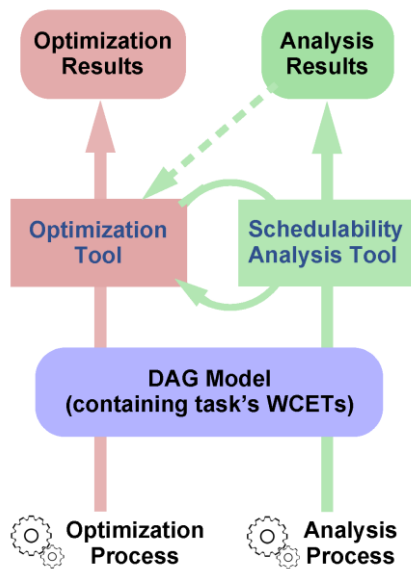


Figure 1. Traditional methodology for real-time analysis and optimization of distributed systems

However, measures of WCETs in heterogeneous hardware presents a strong dependency on the conditions of particular executions mainly due to memory interference among cores and also with GPUs [5][6][7][8][9]. This makes it difficult to obtain precise measures of WCET values or it may lead to very high values of WCET estimations by considering hypothetical worst-case situations. For instance, Figure 2 shows our measures of the impact of GPU on the CPU memory accesses for a Jetson AGX Xavier board. We can observe that the WCET increases by a factor of 4 as the number of GPU threads increases, and the execution time variability also widens.

Partitioning enables controlling inter-core and GPU interference through a proper partition windows assignment, assuming an optimization algorithm (e.g. [10]) that takes into account that these interferences may change. Other techniques can also be applied to control memory interference [11][12][13][14][15]. Thus, WCETs will no longer be fixed or known in advance, and tests to obtain or estimate these WCETs should be developed as a part of an optimization process or for the analysis of any system configuration. Figure 3 shows our proposal for this methodology.

At this moment, the analysis techniques that we propose cannot calculate response times in the GPUs, so they will be considered as a grey boxes, where direct measures of response times in the GPU as well as the memory interference can be incorporated into the model for analysis [16].

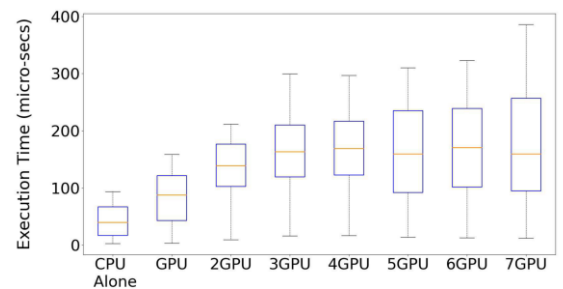


Figure 2. GPU impact (increasing number of threads) on CPU execution times for 100 memory accesses in a 10MB buffer (the yellow line represents the average value and the blue box represents the 90% of measures)

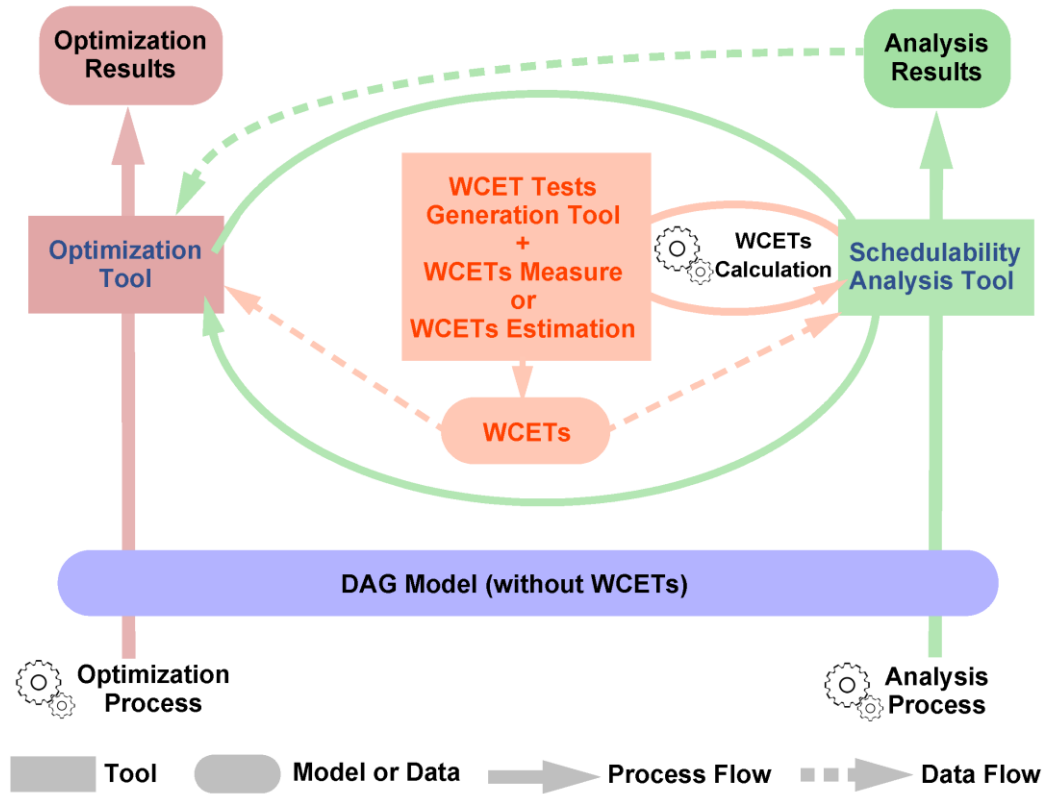


Figure 3. Methodology for the analysis and optimization of distributed DAGs running on heterogeneous hardware

#### ACKNOWLEDGMENT

This work was partially supported by MCIN/ AEI /10.13039/501100011033/ FEDER “Una manera de hacer Europa” under grants PID2021-124502OB-C42 and PID2021-124502OB-C44 (PRESECREL).

#### REFERENCES

- [1] M. Andreozzi, G. Gabrielli, B. Venu, and G. Travaglini. “Industrial Challenge 2022: A High-Performance Real-Time Case Study on Arm,” *Leibniz International Proceedings in Informatics (LIPIcs)*, Vol. 231, 34th Euromicro Conference on Real-Time Systems (ECRTS 2022), July 2022.
- [2] Airlines Electronic Engineering Committee, Aeronautical Radio INC. Avionics Application Software Interface, required Services, ARINC Specification 653-1, 2010.
- [3] A. Amurrio, E. Azketa, J. J. Gutierrez, M. Aldea, and M. G. Harbour, “Response-time analysis of multipath flows in hierarchically-scheduled time-partitioned distributed real-time systems,” *IEEE Access*, vol. 8, pp. 196700–196711, 2020.
- [4] J.M. Rivas, J.J. Gutiérrez, J.L. Medina, and M. González Harbour. “Comparison of Memory Access Strategies in Multi-core Platforms Using MAST,” 8th International Workshop on Analysis Tools and Methodologies for Embedded and Real-time Systems (WATERS), Industrial Challenge 2017, Dubrovnik (Croatia), June 2017.
- [5] R. Cavicchioli, N. Capodiecci, and M. Bertogna, “Memory Interference Characterization between CPU cores and integrated GPUs in Mixed-Criticality Platforms”, 22<sup>nd</sup> IEEE International Conference on Emerging Technologies and Factory Automation, Limassol (Cyprus), September 2017.
- [6] R. Cavicchioli, N. Capodiecci, and M. Bertogna, “Contending memory in heterogeneous SoCs: Evolution in NVIDIA Tegra embedded platforms”, *IEEE 26<sup>th</sup> International Conference on Embedded and Real-time Computing Systems and Applications (RTCSA)*, Gangneung (South Korea), August 2020.
- [7] I. S. Olmedo, N. Capodiecci, and R. Cavicchioli, “A perspective on safety and real-time issues for GPU accelerated ADAS”, 44<sup>th</sup> Annual Conference of the IEEE Industrial Electronics Society (IECON), Washington DC (USA), October 2018.
- [8] D. Shingari, A. Arunkumar, and C. Wu, “Characteriation and Throttling-Based Mitigation of Memory Interference for Heterogeneous Smartphones”, *IEEE International Symposium on Workload Characterization*, Atlanta (USA), October 2015.
- [9] F. Rehm, D. Dasari, A. Hamman, M. Pressler, D. Ziegebein, J. Seitter, I. Sañudo, N. Capodiecci, P. Burgio, and M. Bertogna, “Performance modeling of heterogeneous HW platforms”, *Microprocessors and Microsystems*, Vol. 87, November 2021.
- [10] A. Amurrio, J.J. Gutiérrez, M. Aldea, and E. Azketa. “Partition window assignment in hierarchically scheduled time-partitioned distributed real-time systems with multipath flows,” *Journal of Systems Architecture* 130, Elsevier, September 2022.
- [11] H. Kim, P. Patel, S. Wang, and R. Rajkumar, “A server-based approach for predictable GPU access with improved analysis”, *Journal of Systems Architecture*, Vol. 88, pp. 97-109, August 2018.
- [12] B. Forsberg, L. Benini, and A. Marongiu, “HePREM: A Predictable Execution Model for GPU-based Heterogeneous SoCs”, *IEEE Transaction on Computers*, Vol. 70, pp. 17-29, January 2020.
- [13] S. Kim, C. Jung, and Y. Kim, “Comparative Analysis of GPU Stream Processing between Persistent and Non-persistent Kernels”, 13<sup>th</sup> International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island (South Korea), November 2022.
- [14] H. Aghilinasab, W. Ali, H. Yun, and R. Pellizzioni, “Dynamic Memory Bandwidth Allocation for Real-Time GPU-Based SoC Platforms”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 39, pp. 3348-3360, November 2020.
- [15] R. Li, T. Hu, X. Jiang, L. Li, W. Xing, Q. Deng, and N. Guan, “ROSGM: A Real-Time GPU Management Framework with Plug-In Policies for ROS2”, *IEEE 29<sup>th</sup> Real-Time and Embedded Technology and Applications Symposium (RTAS)*, San Antonio (USA), May 2023.
- [16] 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.