Towards Efficient Modeling and Validation of Scalable Chiplet-based Platforms

Ayoub Mouhagir[1], Fatma Jebali[1], Oumaima Matoussi[1], Caaliph Andriamisaina[1], Anthony Philippe[2]

[1]Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2]Université Grenoble Alpes, CEA, List, F-38000 Grenoble, France

Abstract

Chiplet-based architectures offer a scalable and modular approach to SoC design. However, ensuring system modeling, validation, and performance assessment remains a challenge. To address this, a high-level modeling approach is being developed within the MOSAICS-LP project, combining QEMU-based functional modeling with ML-driven performance analysis and formal timing validation. This hybrid virtual prototyping method enables efficient design exploration and HW/SW co-validation.

Introduction

As SoC architectures grow more complex, the demand for high-performance, energy-efficient computing drives innovation in design. The MOSAICS-LP (MOdular & Scalable AI Component Sovereign – Low Power) project introduces a chiplet-based framework, integrating functional chiplets around a scalable and reconfigurable *Universal Chassis Circuit (UCC)*. At its core, the UCC, whose block digram is shown in Fig. 1, features a RISC-V CPU system supporting Linux, along with UCIe Die-to-Die (D2D) interfaces for seamless communication between heterogeneous chiplets. It also implements two types of eFPGA (embedded FPGA) matrices. The first one enables the reconfiguration and/or specialization of the circuit by integrating proprietary IP solutions. The other allows for reconfigurability and adaptability of all D2D links, thus compensating the absence of industry standards. MOSAICS-LP is targeted to address the requirements of Edge AI, industrial, vision and a variety of other embedded applications.

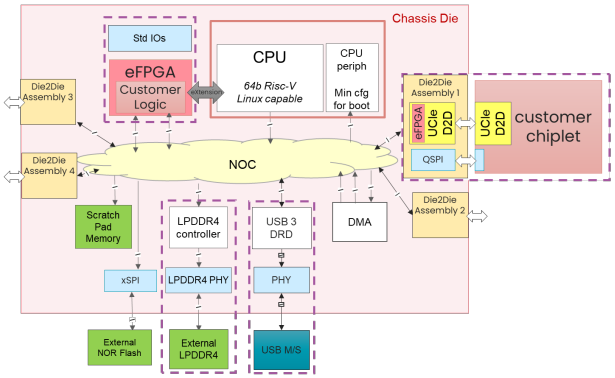

Figure 1: *MOSAICS-LP Block diagram*

Due to the inherent complexity of chiplet-based systems, high-level modeling platforms face several critical challenges. One major challenge is developing models that achieve high accuracy for early-stage validation while ensuring fast and efficient simulation performance. Another challenge involves assessing inter-chiplet communication efficiency by considering factors such as latency, bandwidth, and contention.

To address these challenges, MOSAICS-LP employs virtual prototyping, combining QEMU [1] (*Quick EMUlator*) based functional modeling with ML-driven performance analysis and formal timing validation. This approach balances speed and accuracy for efficient design exploration and system validation. This paper presents the design and ongoing development of the MOSAICS-LP high-level modeling framework, highlighting its role in performance evaluation, HW/SW functional validation, and virtual prototyping.

Proposed Framework

To model the target architecture, our framework, depicted in Fig. 2, consists of two parts: a functional part leveraging QEMU for its high simulation speed, and an extra-functional part combining ML-driven models derived from cycle-accurate simulations and formal WCET analysis for safe timing bound assessment.

Functional modeling

In modern HW & SW development, functional validation plays a crucial role in ensuring that a hardware or software behaves as expected before its final implementation. It allows users to test and debug system components and validate software compatibility (firmware, driver and OS). In the MOSAICS-LP project, QEMU is used as a functional validation tool for our platform. It provides a flexible and fast way to emulate RISC-V based platforms. Currently, a custom in-QEMU machine based on the target platform spec-

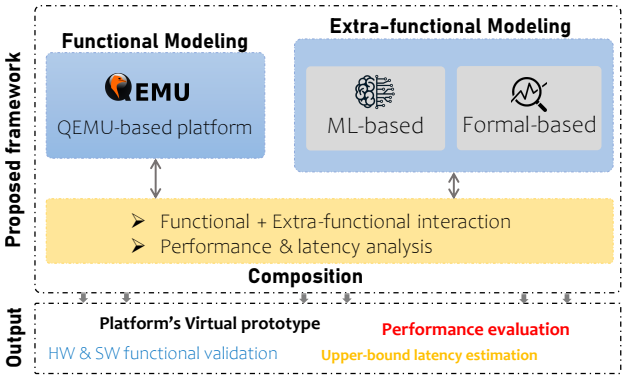

Figure 2: *Proposed simulation framework*

ification is being developed. It introduces a simulated MOSAICS-LP chassis system incorporating key hardware components necessary for functional validation. With this custom machine aligned with the target platform specification, developers can:

- Test the boot process with OS & firmware
- Verify memory & peripherals interactions
- Develop & debug device drivers

Extra-functional modeling

The QEMU functional platform will be enhanced with performance evaluation capabilities. We plan to use ML-based methods to automatically derive lightweight performance models from low-level representations like RTL. Additionally, we will conduct formal analysis of communication delays between components.

ML-based modeling When available, cycle-accurate representations like RTL models enable the automated derivation of higher-level abstractions, such as performance models. Our goal is to leverage ML methods to extract performance models from RTL simulations, and integrate them with the QEMU-based platform for an optimal speed-accuracy trade-off [2, 3]. The resulting framework aims to significantly accelerate SW performance evaluation and HW architecture exploration. We are analyzing the key subsets of the MOSAICS-LP architecture that are crucial for performance evaluation, with a particular emphasis on inter-chiplet communication. These critical subsets will serve as the foundation for our ML-based models.

Formal-based modeling The formal performance model is aimed at analyzing and bounding the communication delays between different components in the chassis. NoC latency is a critical factor in multi-chiplet systems where data must be transferred between different IPs, such as (but not limited to) AI chiplets used for tasks like image recognition and classification. We are working on an abstract representation of the NoC to analyze the worst-case latency [4] for data traversing the NoC, accounting for factors such as NoC topology, routing algorithm, arbitration policy and contention. The results of such a model will be upper-bound latency estimations, which will help ensure that communication delays do not exceed acceptable limits, even under worst-case scenarios.

Models synergy The formal model will complement the ML-based performance model by providing a level of predictability and guaranteed insights that are not easily achievable through simulation alone. While simulation-based models are more flexible and can simulate a broader range of applications, the formal model adds a layer of rigor and confidence to the design process by providing more reliable upper-bound latency estimates. Ultimately, the combination of both models (i.e. formal and ML-based) will create a more holistic performance evaluation, thus allowing the design team to make informed decisions about the allocation of resources, perform optimization strategies and minimize communication bottlenecks that could affect chiplet interoperability and overall system efficiency.

Conclusion and future work

In conclusion, this project introduces a modular chip conception framework leveraging multiple chiplets for scalable and efficient SoC design. The proposed virtual prototyping methodology, leveraging functional and extra-functional modeling, aims to provide early software validation and performance evaluation to support design optimization and validation.

Acknowledgments

This work is partly funded by the MOSAICS-LP project developed collaboratively by CEA-List and Menta with the support of France 2030. For more information about the project, please visit `https://www.menta-efpga.com/` or contact dmitriy.gusev@menta-efpga.com.

References

[1] F. Bellard. "QEMU, a fast and portable dynamic translator." In: *USENIX, FREENIX Track*. Vol. 41. California, USA. 2005, p. 46.

[2] C. Andriamisaina, K. Trabelsi, and P.-G. Le Guay. "A Methodology for Fast and Efficient ML-based Power Modeling". In: *ICCD*. 2024.

[3] I. Macanovic, F. Jebali, and C. Andriamisaina. "QEMU-based CVA6 Framework for Efficient Functional Validation and Performance Evaluation". In: *RISC-V Summit EU*. 2024.

[4] E. A. Rambo and R. Ernst. "Worst-Case Communication Time Analysis of Networks-On-Chip With Shared Virtual Channels". In: *DATE*. 2015, pp. 537–542.