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]

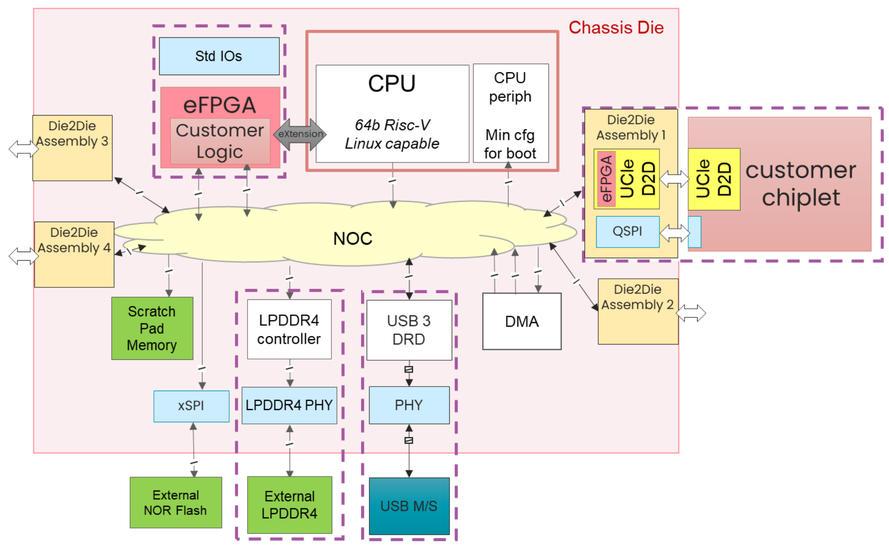
1 Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France (firstname.lastname@cea.fr)

2 Université Grenoble Alpes, CEA, List, F-38000 Grenoble, France (firstname.lastname@cea.fr)

Context

MOSAICS-LP

(MOdular **S**calable **AI** **C**omponent **S**overeign – **L**ow **P**ower)



Universal Chassis

- Complete system infrastructure with Linux-capable 64bits **RISC-V** CPU, DMA, IOs LPDDR4, and USB3 Host-Device Combo
- Reconfigurable and specializable with property IP solutions and Die-to-Die protocols thanks to embedded FPGA matrices.

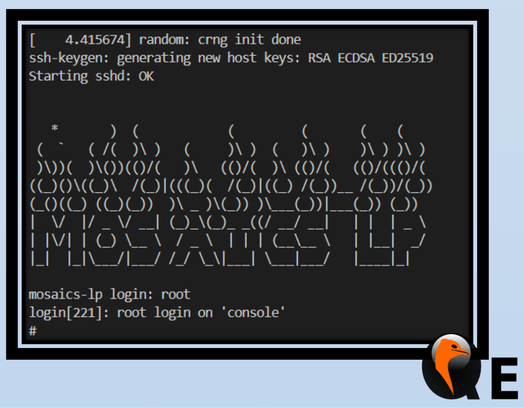
Extension Chiplets

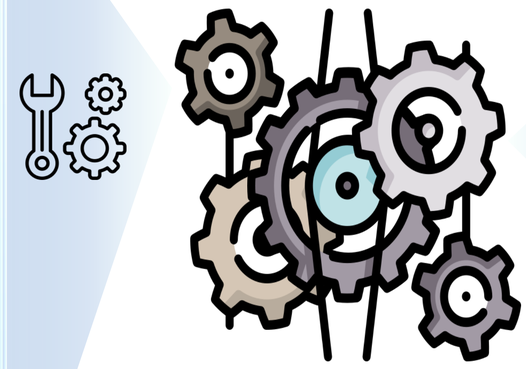
- Fully connected to the chassis NoC through UCIe Die-to-Die interface
- Access to/from all chassis resources

Proposed methodology

Functional Modeling

- Custom in-QEMU [1] machine based on the MOSAICS-LP target platform specification is being developed
- Allows users to:
  - Test the boot process with OS & firmware
  - Verify memory & peripherals interactions
  - Develop & debug device drivers

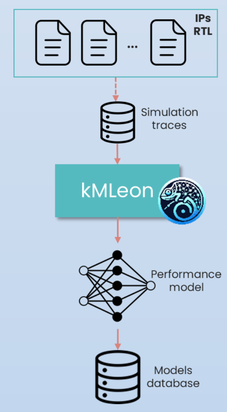




Extra-Functional Modeling

ML-based Performance Modeling

- ML-based methods to automatically build extra-functional models from accurate simulations [2]
- Generated performance models integrated with QEMU for an optimal speed-accuracy trade-off [3]

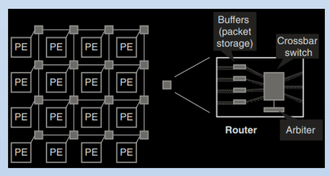


Formal-based Modeling

- An analytical NoC model is being developed in order to:
  - quantify NoC contention delays
  - provide an upper-bound estimate of end-to-end communication latency

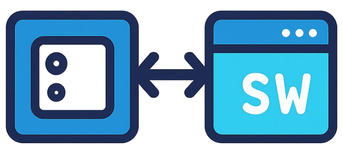
→ Ensure communication latencies do not exceed acceptable limits

[4]



Output

Platform's Virtual Prototype



HW & SW Functional Validation

Performance Evaluation

Upper-bound Latency Estimation

Conclusion

- HW/SW co-development platform for scalable chiplet-based SoC design
- Hybrid modeling approach for early software validation and performance analysis
- Custom QEMU-based platform developed for fast functional modeling
- Ongoing work on ML-driven performance modeling and formal-based delay estimation for improved accuracy

Acknowledgment

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. 2005. QEMU, a Fast and Portable Dynamic Translator. In Proceedings of the FREENIX Track: 2005 USENIX Annual Technical Conference, Anaheim, CA, USA.

[2] C. Andriamisaina, K. Trabelsi, and P.-G. Le Guay. "A Methodology for Fast and Efficient ML-based Power Modeling". In:2024 IEEE 42nd International Conference on Computer Design (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] Pasricha, Sudeep and Dutt, Nikil. "On-Chip Communication Architectures: System on Chip Interconnect". 2008. Morgan Kaufmann Publishers Inc. p439.