PhD Position

Title : CHIPCO : CHIPlet-aware cache Coherency

Start date: September 2025

Duration: 3 years

Required diploma: Bac + 5, Master/Engineering's degree or equivalent

Director:

Frédéric Pétrot,   *Professor at Grenoble INP*

UGA Supervisor:

César Fuguet,   *Researcher at Inria, Université Grenoble-Alpes (UGA)*

BSC Supervisors:

Lluc Alvarez,   *Researcher at Barcelona Supercomputing Center (BSC)*

Miquel Moreto,   *Associate Professor at Universitat Politècnica de Catalunya (UPC) and Scientist at BSC*

Keywords:

Chiplets, Cache-Coherency, Heterogeneity, Quality-of-Service, Communication Protocol, Computer Architecture, Digital Computing Systems, Simulation

Person of contact:

César Fuguet <cesar.fuguet@univ-grenoble-alpes.fr>

Acronym of the proposal	CHIPCO
Title of the proposal	CHIPlet-aware cache COherency
UGA principal investigator	César Fuguet
UGA research unit involved	TIMA, Inria, Grenoble, France
UGA research department (pôle de recherche)	MSTIC
International partner	Universitat Politécnica de Catalunya (UPC), Barcelona Supercomputing Center (BSC)

Abstract

Industry has adopted chiplets to continue scaling computing systems, particularly for applications like High-Performance Computing (HPC) and Artificial Intelligence (AI). Chiplets are smaller, cheaper, and more reusable than large monolithic chips. At the same time, modern chips are becoming more heterogeneous, composed of general-purpose (GP) processor cores and domain-specific cores (accelerators) to increase performance and energy efficiency. This heterogeneity presents design and research challenges, particularly in the communication infrastructure, which includes the Network-on-Chip (NoC) and memory hierarchy.

The memory hierarchy in modern systems contains multiple levels of cache memories, which keep copies of data close to the cores. Cache coherency, ensuring that multiple caches have a consistent view of memory, is a critical issue. Cache coherency protocols, which involve exchanging messages between caches, can significantly impact power consumption and performance. These protocols need to be re-evaluated for multi-chiplet systems.

Different questions need still to be addressed for such systems. To list a few: (1) how to organize cache memories to limit inter-chiplet traffic? (2) shall cache coherency protocols be redefined to be more heterogeneous and chiplet aware? (3) how to guarantee quality-of-service in multi-chiplet systems with multiple types of traffic?

The PhD objectives include building a simulation platform to model heterogeneous chiplet-based processors, evaluating the performance for specific applications, and proposing new techniques to increase the performance and energy efficiency of these systems, with a particular focus on the cache coherency protocol. The goal is to address the research questions above.

The PhD supervision team includes researchers from Grenoble-INP and Inria (Université Grenoble Alpes) on the one hand, Universitat Politécnica de Catalunya (UPC) and Barcelona Supercomputing Center (BSC) on the other hand. The PhD will be co-funded by UGA and BSC.

Required Skill

Personal Skills

- The candidate should be able to spend 1/3 of the PhD duration in Barcelona in the BSC facilities and 2/3 of the PhD duration in Grenoble in the TIMA facilities.
- The candidate should expect to be autonomous in developing software, experiments, and analyzing results.
- The candidate should be able to clearly express their ideas and conclusions, and to motivate their research directions.
- The candidate should be open to constructive criticism from their peers and supervisors.

Technical Skills

- Programming: C/C++. Strong knowledge and understanding of data structures, testing and debugging tools. Linux scripting: Python, bash or other.
- Hardware description: VHDL/Verilog/SystemVerilog. Knowledge and understanding of Register-Transfer Level (RTL) modeling.
- At least basic level in computer architecture (caches, virtual memory, pipelining) and Instruction Set Architecture concepts. Advanced level is a plus.
- At least basic level in cache coherence concepts (coherence protocols, snooping vs. directory)
- Strong english level (B2) is required as scientific articles are written and presented in english.

Project Detailed Description

Title: CHIPlet-aware cache COherency (CHIPCO)

Industry has adopted chiplets as the solution to continue the scaling of computing systems. The size of Systems-on-Chip (SoCs) is continuously increasing to provide the computing throughput required in applications such as scientific High-Performance Computing (HPC) or Artificial Intelligence (AI). However, because of the cost of advanced technology processes, SoCs are split in several small chiplets, which are cheaper to develop and easier to reuse in different products than a large monolithic chip. Chiplets are integrated together within a System-in-Package (SiP) as illustrated in Figure 1.
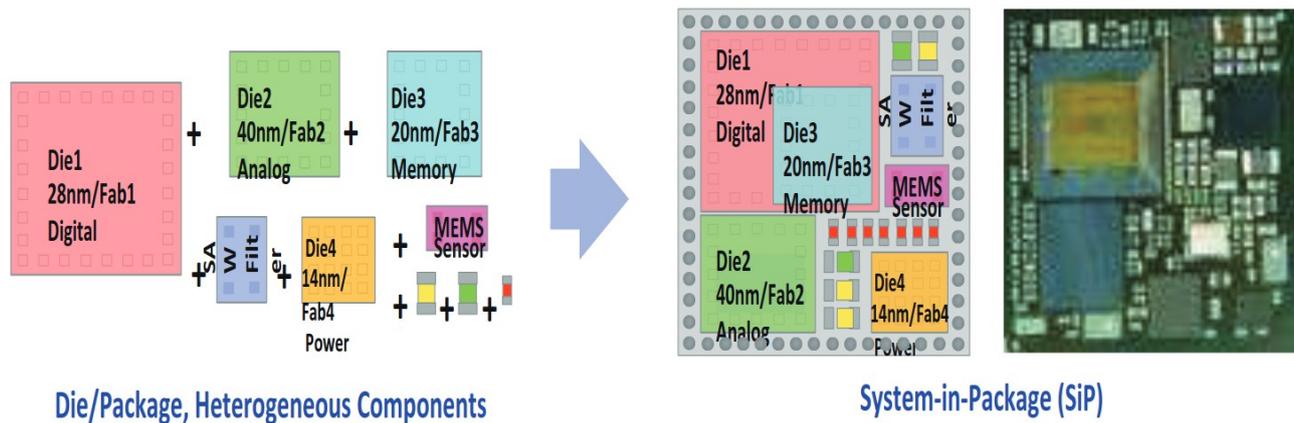

Figure 1: Heterogeneous Integration & System-in-Package (SiP) - Source: [26]

At the same time, chips are becoming more heterogeneous. They are composed of general purpose (GP) processor cores and domain specific cores (accelerators) to get the best energy efficiency (performance/watt). This new distributed and heterogeneous nature of processors' SoCs creates important design and research challenges on software and hardware. A critical component of such SoCs is the communication infrastructure, which plays a key role on their performance and energy efficiency. This infrastructure is composed by the Network-on-Chip (NoC) and the memory hierarchy as the accelerators and GP cores communicate through the memory.

The memory hierarchy in a modern system contains multiple levels of cache memories. These keep copies of data from the main memory (which is usually off-chip and implemented in dense but slow memory technology) close to the cores (both general purpose and accelerators). A common issue on those systems is the cache coherency. Multiple caches, which can be private to the different cores, may contain the copies of the same piece of data. However, if one of the cores modifies a copy, the other caches need to be informed so they can invalidate or update their local copy and keep a consistent view of the memory. This is usually achieved through hardware cache coherency protocols.

Cache coherency protocols consist of a set of messages that are exchanged between the different caches in the system. Those messages, essential to keep a consistent view of the memory to programmers, can represent a significant amount of the power consumption of the system, and may even negatively impact the performance in some cases. A high cache coherency traffic may saturate the NoC, preventing the actual application traffic to progress, and issues such as false sharing, may provoke the cache coherency protocol to invalidate copies of data unnecessarily.

Cache coherency protocols became again a hot research topic since the rise of multi-core processors around 2005. However, such research was mainly on monolithic SoCs. For multi-chiplet systems, the problems need to be readdressed. Given that the semiconductor industry is following the heterogeneous multi-chiplet path, improving the efficiency and scalability of the communication infrastructure is key.

Hereafter we list some questions that need to be answered:

- Inter-chiplet data traffic is expected to have higher latency and power consumption than intra-chiplet data traffic. How to organize cache memories to limit the inter-chiplet traffic ?

- The heterogeneous nature of modern SoCs brings the opportunity to optimize data movements in the cache hierarchy. What kind of cache coherency messages shall be supported in this kind of system ?

- Classical cache coherency protocols would be effective or shall them be redefined ?

- Classical protocols may require broadcasting messages. In that case, shall the broadcast reach all caches in a given level in a flatten way (even beyond a given chiplet) or should them be "clustered" to improve their power efficiency and reduce their performance impact ?

- On the presence of heterogeneous chiplets, shall the SoC provide a single general cache coherency protocol, or could we define different protocols for intra-chiplet and inter-chiplet communication ?

- How to guaranty a given quality-of-service for the data traffic generated by the accelerators and GP cores in the different chiplets ? It is to be expected that some traffics have higher priority than others (e.g. latency sensitive traffic such as the one in audio or video applications, or throughput sensitive traffic such as the one in scientific HPC applications).

In addition, the industry is trying to quickly adopt different standard interfaces and protocols to support chiplet-based systems. Examples of such standards are CXL, CCIX and UCIe. There are also standards for "classical" cache coherency protocols such as CHI. We would like to evaluate the more relevant protocols/interfaces to see if they are general enough to efficiently support heterogeneous multi-chiplet SoCs, and to eventually propose optimizations to improve both their performance and energy efficiency.

The objectives of this project are :

- Build a simulation platform (based on existing bricks[1], [2]) capable of modeling and evaluating a heterogeneous chiplet-based processor

- Evaluate the performance of such systems for a given set of applications

- Propose new techniques, mainly on the NoC and cache coherency protocol, to improve the performance and energy efficiency of heterogeneous multi-chiplet systems. The goal is to give an answer to some of the questions above. We will pay special attention to propose techniques that can leverage existing interfaces and standards. In some cases, we will identify their limitations and propose extensions to improve their performance.

State of the Art

Cache coherency protocols enable efficient programming of multi-core processors. There have been many coherence protocols proposed in the last four decades, but surprisingly, most of the modern processors implement variants of the « MESI » write-invalidate coherence protocol proposed by the University of Illinois in 1984 [3]. However, the introduction of multi-chiplet systems with heterogeneous accelerators invites to a re-evaluation of coherence protocols.

Regarding chiplet-aware cache coherence protocols, there are some state of the art proposals. Authors in [4] propose a write-update coherence protocol, which instead of invalidating copies in caches on a write, the protocol sends directly the updated data to the caches having a copy. They demonstrate that the extra traffic generated to guaranty those updates is hidden by the increase of the bandwidth in modern SiP

interconnects. Authors in [5] also shows similar results implementing a hybrid write-update / write-invalidate coherence protocol.

Other approaches [6], [7] propose to modify the memory hierarchy to introduce, along with caches, scratchpad memories. They use a software/hardware approach where the compiler and a modified coherence protocol works together to maintain a coherent view of the data.

While these solutions show performance improvements and good scalability in comparison to classical coherence protocols, they do not consider multi-chiplet systems (in the case of [6], [7]), or they only consider homogeneous systems with chiplets integrating general purpose cores (in the case of [4], [5]). These solutions do not consider either quality-of-service when multiple types of traffic with different priorities coexist in the system.

As mentioned previously, industry has adopted multi-chiplet, heterogeneous systems where general purpose cores and accelerators coexist in different chiplets to improve the performance and energy efficiency. We believe that coherence protocols may exploit the characteristics of the different memory access patterns from those accelerators to further improve efficiency. We also believe that considering quality-of-service is essential. Chiplets are meant to be reused in different products, but depending on the product, some traffics from a given chiplet may have higher priority than others.

There are other interesting works improving the efficiency of heterogeneous systems. For example, [8] proposes a hardware/software mechanism that improves the communication between a host processor and a accelerator. It introduces a fine-grained cache, that keeps track of data dependencies between operations and increases the reutilization of data to avoid accessing the (slow) external memory. However, authors do not consider multi-core / multi-accelerator systems, and how to address coherency in that case.

Scientific Approach, Work Plan, Expected Results and Deliverables

The 3 years of PhD will be organized as follows:

- The first year will focus on studying the state-of-the-art on the topics of chiplet-based computing systems, cache coherency and NoC. It will also include exploring existing simulators for multi-core heterogeneous systems. The student should be able at the end of the first year to model a first reference simulation platform[1], and evaluate the performance of different applications/benchmarks for parallel heterogeneous computing systems.

- The second year will focus on proposing, implementing and evaluating techniques to improve the efficiency (performance and energy) of chiplet-based, heterogeneous, high-performance computing systems. A first scientific publication on the obtained results is expected during this second year of PhD.

- The third year will focus on publishing the latest results of the PhD and writing the PhD thesis manuscript.

As requested by the doctoral school, the student will need to assist to some mandatory and very useful lectures during the 3 years of the PhD.

---

[1]We define a simulation platform as an executable model of the target system. This model includes the different components of the system such as processor cores, accelerators, network-on-chip, and the memory hierarchy. It allows to execute "real" software applications and extract performance/energy measurements.

Regarding the expected results and deliverables, we expect to publish scientific papers in high-impact international conferences and journals in the computer architecture domain. If there is also the opportunity, we could produce patents related to some specific mechanisms proposed during the PhD.

The international conferences and journals that we will target preferentially within this project are:

- International Symposium on Computer Architecture (ISCA)

- International Symposium on Microarchitecture (MICRO)

- International Symposium on High-Performance Computer Architecture (HPCA)

- Transactions on Computers (TC) journal

- Transactions on Architecture and Code Optimization (TACO) journal

In terms of number of publications, we target at least 2 publications during the 3 years of PhD. When possible, some artifacts could be published as open-source, such as the reference simulation platform. This will enable and ease future collaborations and research on similar topics.

International Positioning of the Project

The Barcelona Supercomputing Center (BSC) is one of the most prestigious research institutions in Europe in the domain of scientific high-performance computing. They are also known for hosting and managing one of the most powerful supercomputers (MareNostrum) in Europe. They are very active in the computer architecture research community, with many scientific publications on the most important international conferences and journals of the domain.

The Universitat Politècnica de Catalunya · BarcelonaTech (UPC) is a public institution of research and higher education in the fields of engineering, architecture, sciences and technology, and one of the leading technical universities in Europe. Every year, close to 6,000 bachelor's and master's students, more than 300 doctoral students and 1,600 continuing education students graduate. **It is part of the nine Unite ! alliance universities.**

Regarding the supervision team of this PhD project, we already have collaborations related to the design and development of high-performance computing architectures. We have mentored master students together resulting in scientific publications and contributions to open-source hardware projects [9], [10], [11] such as the HPDcache, CVA6 and OpenPiton.

Frédéric Petrot is a professor ("Professeur des Universités) at Grenoble INP (Institut d'Ingénierie et de management) - UGA. It mainly teaches about computer architecture and advanced lectures about processors micro-architecture. His research topics are: architecture for Artificial Intelligence accelerators, Network-on-Chips (NoCs), Systems-on-Chip (SoCs) fast simulation, multi-core architectures and cache coherency. He has many scientific publications in highly-ranked international conferences and journals of those domains (e,g, [12], [13], [14], [15]). Frédéric Pétrot will be the PhD director.

César Fuguet is a researcher ("Chargé de Recherche") at Inria (Institut national de recherche en sciences et technologies du numérique) – UGA. He teaches computer architecture at the Polytech Grenoble engineering school. His research interests are: Network-on-Chips (NoCs), memory and cache hierarchy, multi-chiplet heterogeneous SoCs, multi-core architectures and cache coherency. He has publications in highly-ranked international conferences and journals of those domains. (e.g. [5], [16], [17], [18]). César Fuguet will be a co-supervisor of the PhD.

Miquel Moreto is an associate professor at Universitat Politécnica de Catalunya (UPC) and team leader at BSC. Miquel Moreto has participated in many relevant research projects in the area of computer architecture, leading to numerous top-tier highly-cited publications in topics such as SoC design [19], cache hierarchies [20], NoCs [21] and resource management [22]. Miquel Moreto will be co-supervisor of the PhD.

Lluc Alvarez is a researcher at Barcelona Supercomputing Center (BSC) who has participated in multiple highly-ranked international scientific publications and journals in the area of computer architecture, and specifically some of these works are related to the multi-core architectures and cache coherency (e.g. [6], [7]). He also has publications that address the efficient execution of parallel applications in multi-core systems through both software and hardware approaches (e.g. [24], [25]). Lluc Alvarez will be a co-supervisor of the PhD.

References

[1] J. Balkind *et al.*, "OpenPiton: An Open Source Manycore Research Framework," *SIGPLAN Not*, vol. 51, no. 4, pp. 217–232, Mar. 2016, doi: 10.1145/2954679.2872414.

[2] N. Binkert *et al.*, "The gem5 simulator," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, May 2011, doi: 10.1145/2024716.2024718.

[3] M. S. Papamarcos and J. H. Patel, "A low-overhead coherence solution for multiprocessors with private cache memories," in *Proceedings of the 11th annual international symposium on Computer architecture*, in ISCA '84. New York, NY, USA: Association for Computing Machinery, Jan. 1984, pp. 348–354. doi: 10.1145/800015.808204.

[4] G. Chirkov and D. Wentzlaff, "Seizing the Bandwidth Scaling of On-Package Interconnect in a Post-Moore's Law World," in *Proceedings of the 37th International Conference on Supercomputing*, Orlando FL USA: ACM, Jun. 2023, pp. 410–422. doi: 10.1145/3577193.3593702.

[5] P. Vivet *et al.*, "IntAct: A 96-Core Processor With Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management," *IEEE J. Solid-State Circuits*, vol. 56, no. 1, pp. 79–97, Jan. 2021, doi: 10.1109/JSSC.2020.3036341.

[6] L. Alvarez *et al.*, "Coherence protocol for transparent management of scratchpad memories in shared memory manycore architectures," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, Portland Oregon: ACM, Jun. 2015, pp. 720–732. doi: 10.1145/2749469.2750411.

[7] M. Casas *et al.*, "Runtime-Aware Architectures," in *Euro-Par 2015: Parallel Processing*, vol. 9233, J. L. Träff, S. Hunold, and F. Versaci, Eds., in Lecture Notes in Computer Science, vol. 9233. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 16–27. doi: 10.1007/978-3-662-48096-0_2.

[8] S. Chen, H. Zhang, and T. Austin, "Zipper: Latency-Tolerant Optimizations for High-Performance Buses," in *30th Asia and South Pacific Design Automation Conference (ASPDAC '25)*, Tokyo, Japan: ACM, New York, NY, USA, Jan. 2025, p. 8 pages. doi: https://doi.org/10.1145/3658617. 3697546.

[9] D. Million, N. Oliete-Escuín, and C. Fuguet, "Breaking the Memory Wall with a Flexible Open-Source L1 Data-Cache," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Mar. 2024, pp. 1–2. doi: 10.23919/DATE58400.2024.10546547.

[10] "Enhancing OpenPiton+Ariane With a High-Performance Data Cache," Google Summer of Code. Accessed: Feb. 25, 2025. [Online]. Available: https://aileonn.github.io/projects/gsoc/

[11] "Transforming the OpenHW High Performance Data Cache into a High Performance Instruction Cache," Google Summer of Code. Accessed: Feb. 25, 2025. [Online]. Available: https://summerofcode.withgoogle.com/archive/2024/projects/zTms74CY

[12] H. Alemdar, V. Leroy, A. Prost-Boucle, and F. Pétrot, "Ternary neural networks for resource-efficient AI applications," in *2017 international joint conference on neural networks (IJCNN)*, IEEE, 2017, pp. 2547–2554. Accessed: Feb. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7966166/

[13] F. Dubois, A. Sheibanyrad, F. Pétrot, and M. Bahmani, "Elevator-First: A Deadlock-Free Distributed Routing Algorithm for Vertically Partially Connected 3D-NoCs," *IEEE Trans. Comput.*, vol. 62, no. 3, pp. 609–615, Mar. 2013, doi: 10.1109/TC.2011.239.

[14] M. Gligor, N. Fournel, and F. Pétrot, "Using binary translation in event driven simulation for fast and flexible MPSoC simulation," in *Proceedings of the 7th IEEE/ACM international conference on Hardware/software codesign and system synthesis*, Grenoble France: ACM, Oct. 2009, pp. 71–80. doi: 10.1145/1629435.1629446.

[15] F. Petrot, A. Greiner, and P. Gomez, "On cache coherency and memory consistency issues in NoC based shared memory multiprocessor SoC architectures," in *9th EUROMICRO Conference on Digital System Design (DSD'06)*, IEEE, 2006, pp. 53–60. Accessed: Feb. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1690020/

[16] Y. Thonnart *et al.*, "POPSTAR: A robust modular optical NoC architecture for chiplet-based 3D integrated systems," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2020, pp. 1456–1461. Accessed: Feb. 25, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9116214/

[17] C. Fuguet, "HPDcache: Open-Source High-Performance L1 Data Cache for RISC-V Cores," in *Proceedings of the 20th ACM International Conference on Computing Frontiers*, in CF '23. New York, NY, USA: Association for Computing Machinery, août 2023, pp. 377–378. doi: 10.1145/3587135.3591413.

[18] E. Guthmuller *et al.*, "Xvpfloat: RISC-V ISA Extension for Variable Extended Precision Floating Point Computation," *IEEE Trans. Comput.*, pp. 1–14, 2024, doi: 10.1109/TC.2024.3383964.

[19] K. Asanović *et al.*, "The Rocket Chip Generator".

[20] H. Cook, M. Moreto, S. Bird, K. Dao, D. A. Patterson, and K. Asanovic, "A hardware evaluation of cache partitioning to improve utilization and energy-efficiency while preserving responsiveness," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, in ISCA '13. New York, NY, USA: Association for Computing Machinery, Jun. 2013, pp. 308–319. doi: 10.1145/2485922.2485949.

[21] C. Martínez, R. Beivide, E. Stafford, M. Moretó, and E. M. Gabidulin, "Modeling Toroidal Networks with the Gaussian Integers," *IEEE Trans. Comput.*, vol. 57, no. 8, pp. 1046–1056, Aug. 2008, doi: 10.1109/TC.2008.57.

[22] K. J. Nesbit, M. Moreto, F. J. Cazorla, A. Ramirez, M. Valero, and J. E. Smith, "Multicore Resource Management," *IEEE Micro*, vol. 28, no. 3, pp. 6–16, May 2008, doi: 10.1109/MM.2008.43.

[23] P. Caheny, L. Alvarez, S. Derradji, M. Valero, M. Moretó, and M. Casas, "Reducing cache coherence traffic with a numa-aware runtime approach," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 5, pp. 1174–1187, 2017.

[24] P. Caheny, L. Alvarez, M. Valero, M. Moreto, and M. Casas, "Runtime-Assisted Cache Coherence Deactivation in Task Parallel Programs," in *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*, Dallas, TX, USA: IEEE, Nov. 2018, pp. 454–465. doi: 10.1109/SC.2018.00038.

[25] E. Castillo *et al.*, "Architectural support for task dependence management with flexible software scheduling," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, IEEE, 2018, pp. 283–295. Accessed: Feb. 21, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8327016/

[26] W. Chen and B. Bottoms, "Heterogeneous Integration Roadmap: Driving Force and Enabling Technology for Systems of the Future," in *2019 Symposium on VLSI Technology*, Jun. 2019, pp. T50–T51. doi: 10.23919/VLSIT.2019.8776484.