

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Architecting more than Moore: wireless plasticity for massive heterogeneous computer architectures (WiPLASH)

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Klein, J., Levisse, A., Ansaloni, G., Atienza, D., Zapater, M., Dazzi, M., et al. (2021). Architecting more than Moore: wireless plasticity for massive heterogeneous computer architectures (WiPLASH). New York, NY : Association for Computing Machinery [10.1145/3457388.3458859].

Availability:

This version is available at: <https://hdl.handle.net/11585/847005> since: 2022-01-23

Published:

DOI: <http://doi.org/10.1145/3457388.3458859>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Klein J. et al., “Architecting more than Moore: wireless plasticity for massive heterogeneous computer architectures (WiPLASH)”, CF '21: Proceedings of the 18th ACM International Conference on Computing Frontiers, May 2021, Pages 191 -193, <https://doi.org/10.1145/3457388.3458859>

The final published version is available online at:

<https://dl.acm.org/doi/abs/10.1145/3457388.3458859>

Rights/License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it>)
When citing, please refer to the published version.*

Architecting More Than Moore – Wireless Plasticity for Massive Heterogeneous Computer Architectures (WiPLASH)

Invited Paper

Joshua Klein
Alexandre Levisse
Giovanni Ansaloni
David Atienza
joshua.klein@epfl.ch
alexandre.levisse@epfl.ch
giovanni.ansaloni@epfl.ch
david.atienza@epfl.ch
EPFL
Lausanne, Switzerland

Davide Rossi
Francesco Conti
davide.rossi@unibo.it
f.conti@unibo.it
Università di Bologna
Bologna, Italy

Zhenxing Wang
Kun-Ta Wang
wang@amo.de
ktwang@amo.de
AMO GmbH
Aachen, Germany

Marina Zapater
marina.zapater@heig-vd.ch
HEIG-VD/HES-SO
Yverdon-les-Bains, Switzerland
EPFL
Lausanne, Switzerland

Elana Pereira de Santana
Peter Haring Bolívar
Elana.PSantana@uni-siegen.de
peter.haring@uni-siegen.de
University of Siegen
Siegen, Germany

Max C. Lemme
max.lemme@eld.rwth-aachen.de
RWTH
Aachen, Germany
AMO GmbH
Aachen, Germany

Martino Dazzi
Geethan Karunaratne
Irem Boybat
Abu Sebastian
DAZ@zurich.ibm.com
KAR@zurich.ibm.com
IBO@zurich.ibm.com
ase@zurich.ibm.com
IBM Research Europe
Zurich, Switzerland

Mohamed Saeed
Renato Negra
mohamed.elsayed@hfe.rwth-aachen.de
negra@hfe.rwth-aachen.de
RWTH
Aachen, Germany

Akshay Jain
Robert Guirado
Hamidreza Taghvaei
Sergi Abadal
akshay.jain@upc.edu
rguirado@ac.upc.edu
taghvaei@ac.upc.edu
abadal@ac.upc.edu
Universitat Politècnica de Catalunya
Barcelona, Spain

ABSTRACT

This paper presents the research directions pursued by the WiPLASH European project, pioneering on-chip wireless communications as a disruptive enabler towards next-generation computing systems for artificial intelligence (AI). We illustrate the holistic approach driving our research efforts, which encompass expertises and abstraction levels ranging from physical design of embedded graphene

antennas to system-level evaluation of wirelessly-communicating heterogeneous systems.

CCS CONCEPTS

• **Hardware** → **Radio frequency and wireless interconnect; Emerging architectures.**

KEYWORDS

WiPLASH, On-chip antennas, In-Memory Computing

Reference Format:

Joshua Klein, Alexandre Levisse, Giovanni Ansaloni, David Atienza, Marina Zapater, Martino Dazzi, Geethan Karunaratne, Irem Boybat, Abu Sebastian, Davide Rossi, Francesco Conti, Elana Pereira de Santana, Peter Haring Bolívar, Mohamed Saeed, Renato Negra, Zhenxing Wang, Kun-Ta Wang, Max C. Lemme, Akshay Jain, Robert Guirado, Hamidreza Taghvaei, and Sergi Abadal. 2021. Architecting More Than Moore – Wireless Plasticity for Massive Heterogeneous Computer Architectures (WiPLASH)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863337 (WiPLASH).

The final publication is available at ACM via
<http://dx.doi.org/10.1145/3457388.3458859>

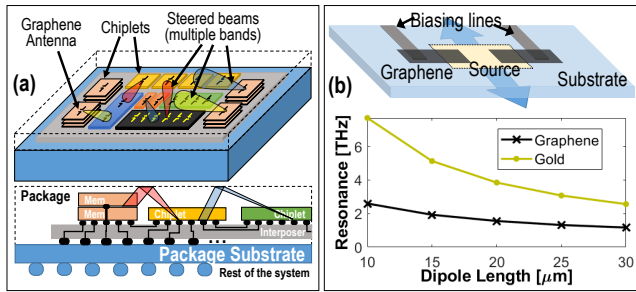


Figure 1: (a) Graphene antennas integrated within chiplets in a heterogeneous architecture, enabling reconfigurable links in frequency and direction of radiation. (b) Schematic of a simple graphene dipole and simulated proof of miniaturization (from [1]).

: Invited Paper. In *Computing Frontiers Conference (CF '21)*, May 11–13, 2021, Virtual Conference, Italy. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3457388.3458859>

1 THE WIPLASH VISION

The main design principles in computer architecture have shifted from a monolithic scaling-driven approach towards emerging heterogeneous architectures that tightly co-integrate multiple specialized computing and memory units [6]. This heterogeneous hardware specialization requires interconnection mechanisms that serve the architecture communication demands. State-of-the-art approaches are 3D stacking and 2.5D architectures complemented with a Network-on-Chip (NoC) or Network-in-Package (NiP) to interconnect the components [3, 11]. However, such interconnects are fundamentally monolithic and rigid, and are unable to provide the efficiency and architectural flexibility required by current and future key applications. The main challenge is, hence, to introduce diversification and specialization in heterogeneous processor architectures, while ensuring their generality and the scalability of increasingly complex computing systems.

Addressing this challenge, the main goals of the WiPLASH European project are the investigation and evaluation of key technologies for wireless communication among on-chip and in-package processing elements. We aim to introduce novel wireless communication planes able to provide architectural plasticity, this is, reconfigurability and adaptation to the application requirements, achieving very high performance without any loss of generality.

In this paper, we outline the innovations pursued by the project consortium, centering on key hardware and software enablers for wireless communication at the chip scale: integrated antennas and communication protocols. Moreover, we present solutions being developed to evaluate wireless-enabled architectures at the system level, both in ultra-low-power and in high-performance computing scenarios.

2 TECHNOLOGY ENABLERS

On-chip Antennas

We propose and demonstrate the use of graphene-based antennas enabling point-to-point and broadcasting communication modes among multiple chiplets embedded in heterogeneous systems, as illustrated in Fig. 1(a) [2]. Graphene antennas have three unique

properties that render them an excellent fit for this purpose [1]: First, they naturally support waves in the THz band, a frequency range that can potentially deliver ample bandwidth for high-speed transmission. Second, they can be miniaturized up to 100x in terms of area with respect to a metallic antenna resonating at the same frequency, thereby reducing the footprint and cost of wireless solutions (Fig. 1(b) shows an exemplary reduction in one length dimension). Third, they are electrically tunable in ways that cannot be achieved with classical metallic antennas [10], providing unique opportunities for multiple parallel frequency-space communication channels, which are highly desirable in connectivity-limited scenarios.

Communication protocols

Developing a wireless network requires, besides the antennas and other RF hardware, a protocol stack that manages the communication. Decisions typically revolve around the type of modulation, the strategy for Medium Access Control (MAC), and network-level aspects such as routing and load balancing. In terms of modulations, chip-scale networks generally resort to simple schemes to minimize the area and power footprint [5, 8]. The MAC protocol design is complex due to the constant changes in the communication patterns. Recently, we demonstrated that a hybrid variant of token passing and carrier sensing can adapt to these changes [7]. We also proposed a protocol based on deep reinforcement learning that learns to predict traffic patterns and focuses on minimizing execution time rather than maximizing wireless throughput [9]. Beyond this, the challenge is to adapt existing MAC protocols to support multiple frequency and spatial channels.

Wireless Systems-on-Chip

A core contribution of WiPLASH is the evaluation of the system- and application- level benefits deriving from wireless communication planes. To this end, we enable architectural explorations of computational and communication resources, leveraging open-source hardware and instruction sets, while providing novel functionalities and timing/energy models reflecting the capabilities of graphene antennas and dedicated MAC protocols. In this context, our effort is two-pronged.

At the low-power end of the computing spectrum, we target the PULP platform [12]. The PULP ecosystem includes a full vertical stack of components such as RTL IPs, FPGA emulators, fully layouted databases on deep scaled technology nodes and silicon prototypes, as well as a full software stack for mapping of both signal processing and AI applications including an event-based virtual platform. This vertical approach allows to calibrate the simulators on cycle-accurate emulators or RTL/layouts and plug on it the emerging technologies developed in WiPLASH.

As for High Performance Computing (HPC) systems, we instead consider event-based simulations. We base our efforts on the gem5 simulator [4], which we extended to support 64-bit out-of-order cores in full-system mode, modern Linux distributions and enhanced profiling functionalities. The resulting framework, named gem5-X [13], allows both detailed architectural explorations and the investigation of entire hardware/software stacks, including user-level processes and the operating system.

For both low-power and HPC scenarios, we also provide validated models of analog in-memory computing (AIMC) accelerators for supporting deep learning workloads [14]. The emerging AIMC paradigm leads to significant energy and performance gains for

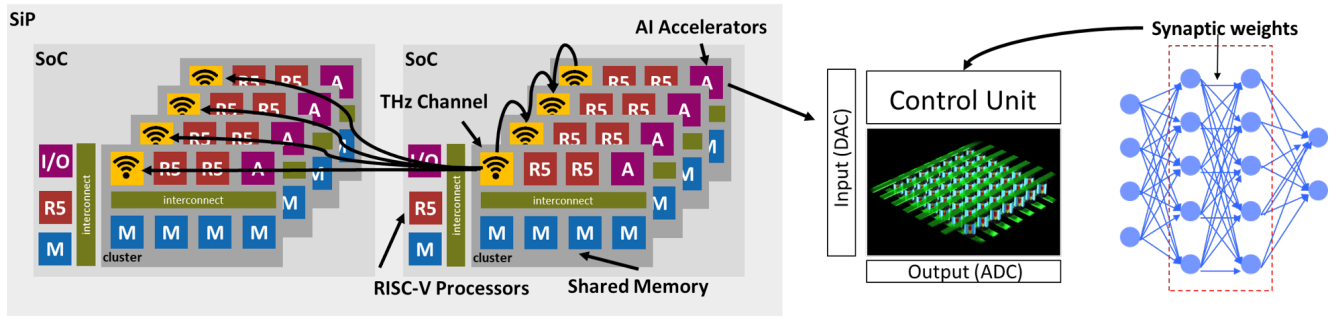


Figure 2: The WiPLASH European project aims at developing Massively Parallel Heterogeneous Architecture, exploiting Analog In-Memory Computing for acceleration AI workloads and Wireless THz Channels for Cluster-to-Cluster and Chip-to-Chip communication.

various applications, by reducing the local memory traffic and by offering massive parallelism.

REFERENCES

- [1] Sergi Abadal et al. 2017. Graphene-based terahertz antennas for area-constrained applications. In *Proceedings of the TSP'17*. IEEE, 817–820.
- [2] Sergi Abadal et al. 2020. Graphene-based Wireless Agile Interconnects for Massive Heterogeneous Multi-chip Processors. *arXiv preprint arXiv:2011.04107* (2020).
- [3] Davide Bertozzi et al. 2015. The fast evolving landscape of on-chip communication. *Design Automation for Embedded Systems* 19, 1 (2015), 59–76.
- [4] Nathan Binkert et al. 2011. The gem5 simulator. *ACM SIGARCH computer architecture news* 39, 2 (2011), 1–7.
- [5] Chul Woo Byeon et al. 2020. A 2.65-pJ/Bit 12.5-Gb/s 60-GHz OOK CMOS Transmitter and Receiver for Proximity Communications. *IEEE Transactions on Microwave Theory and Techniques* 68, 7 (2020), 2902–2910.
- [6] Francesco Conti et al. 2017. An IoT Endpoint System-on-Chip for Secure and Energy-Efficient Near-Sensor Analytics. *IEEE Transactions on Circuits and Systems I: Regular Papers* 64, 9 (2017), 2481–2494.
- [7] Antonio Franques et al. 2021. Fuzzy-Token: An Adaptive MAC Protocol for Wireless-Enabled Manycores. In *Proceedings of the DATE'21*.
- [8] David Fritzsche et al. 2017. A Low-Power SiGe BiCMOS 190-GHz Transceiver Chipset With Demonstrated Data Rates up to 50 Gbit/s Using On-Chip Antennas. *IEEE Transactions on Microwave Theory and Techniques* 65, 9 (2017), 3312–3323.
- [9] Suraj Jog et al. 2021. One Protocol to Rule Them All: Wireless Network-on-Chip using Deep Reinforcement Learning. In *Proceedings of the NSDI'21*.
- [10] Long Ju et al. 2011. Graphene plasmonics for tunable terahertz metamaterials. *Nature nanotechnology* 6, 10 (2011), 630–634.
- [11] Jinwoo Kim et al. 2019. Architecture, chip, and package co-design flow for 2.5 D IC design enabling heterogeneous IP reuse. In *Proceedings of DAC-56*.
- [12] Antonio Pullini et al. 2019. MrWolf: An Energy-Precision Scalable Parallel Ultra Low Power SoC for IoT Edge Processing. *IEEE Journal of Solid-State Circuits* 54, 7 (2019), 1970–1981.
- [13] Yasir Mahmood Qureshi et al. 2019. Gem5-X: A Gem5-based system level simulation framework to optimize many-core platforms. In *2019 Spring Simulation Conference (SpringSim)*. IEEE, 1–12.
- [14] Abu Sebastian et al. 2020. Memory devices and applications for in-memory computing. *Nature Nanotechnology* (2020).