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

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

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

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

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

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

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

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

Akshay Jain
Robert Guirado
Hamidreza Taghvaee
Sergi Abadal
akshay.jain@upc.edu
rguirado@ac.upc.edu
taghvaee@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 expertise 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

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

REFERENCES