Graphene-Enabled Wireless Communication for Massive Multicore Architectures

Sergi Abadal, Eduard Alarcón, and Albert Cabellos-Aparicio, Universitat Politecnica de Catalunya
Max C. Lemme, KTH Royal Institute of Technology and University of Siegen
Mario Nemirovsky, Barcelona Supercomputing Center (BSC)

ABSTRACT

Current trends in microprocessor architecture design are leading towards a dramatic increase of core-level parallelization, wherein a given number of independent processors or cores are interconnected. Since the main bottleneck is foreseen to migrate from computation to communication, efficient and scalable means of inter-core communication are crucial for guaranteeing steady performance improvements in many-core processors. As the number of cores grows, it remains unclear whether initial proposals, such as the Network-on-Chip (NoC) paradigm, will meet the stringent requirements of this scenario. This position paper presents a new research area where massive multicore architectures have wireless communication capabilities at the core level. This goal is feasible by using graphene-based planar antennas, which can radiate signals at the Terahertz band while utilizing lower chip area than its metallic counterparts. To the best of our knowledge, this is the first work that discusses the utilization of graphene-enabled wireless communication for massive multicore processors. Such wireless systems enable broadcasting, multicasting, all-to-all communication, as well as significantly reduce many of the issues present in massively multicore environments, such as data coherency, consistency, synchronization and communication problems. Several open research challenges are pointed out related to implementation, communications and multicore architectures, which pave the way for future research in this multidisciplinary area.

INTRODUCTION

Throughout the decades, technology advancements in digital circuits, i.e., precise manufacturing techniques, have enabled a steady reduction in the size of transistors. Such tendency has allowed the integration of more transistors on the same chip and resulted in a very high performance increase and cost decrease per transistor. As the level of integration approaches Ultra-Large-Scale Integration (ULSI), the intra-chip communication latency and power consumption become major barriers that prevent the continuation of the trend set by the Moore’s Law. Indeed, the main reasons for the diminishing performance returns of such downsizing trend are as follows. By reducing the width of the on-chip wires, their resistance and therefore the resistive-capacitive (RC) delay are significantly increased. Also, by taking into account the increased clock frequencies imposing reduced symbol times, the charging and discharging of the wire within the allotted time becomes a very challenging problem. Finally, the dynamic power demand of a Complementary Metal Oxide Semiconductor (CMOS) transistor grows proportionally to its operation frequency and quadratically to the circuit voltage, justifying the need for low-voltage and frequency-limited designs. Graphene, thanks to its extremely promising properties, could enable the devising of transistors with higher speed and lower energy consumption than traditional CMOS devices. However, such transistors are, thus far, projected for its application in RF circuits rather than in digital computation, due to the intrinsic absence of band gap in graphene [1]. Since better performance is no longer achievable through an increase in clock frequency due to the reasons pointed out above, the natural trend in microprocessor architecture design is to improve the performance by means of parallel architectures. Parallelization is achieved by interconnecting several independent processors forming a Chip Multiprocessor (CMP), and has led towards the recent emergence of multicore and manycore, i.e., more than 16 cores, processors. The main performance bottleneck in these systems is currently defined by the intra-chip communication requirements set by coherency or synchronization, among other common and necessary operations in multicore environments. In this context, the Network-on-Chip (NoC) paradigm was proposed to increase the performance of CMP systems by providing scalable and efficient inter-core communication through
wireline routed interconnections. This approach arose as opposed to the traditional bus-based architectures, which scale poorly in terms of delay and energy efficiency when the number of cores is increased due to their time-division multiplexing nature.

However, as the technology downscaling allows the integration of more cores in the same chip, initial wired NoC solutions pose several challenges in terms of delay, power requirements and chip area utilization, also referred to as area overhead. Consequently and since the inception of the NoC paradigm, many research efforts have been directed towards coping with the increasing demands of the ever-changing field of microprocessor architecture [2]. In this article, we review the state-of-the-art of this research area and introduce Graphene-enabled Wireless Network-on-Chip (GWNoC), a novel approach that relies in graphene-based nano-antennas [3] to implement wireless communication between the cores of a multiprocessor. Such proposal is expected to considerably outperform other initial Wireless Network-on-Chip (WNoC) designs and aims to provide improved scalability, flexibility and area overhead for multicore systems with hundreds or even thousands of cores. To the best of our knowledge, this is the first work that discusses the employment of nanoscale wireless communication for on-chip networks. Our main contributions are:

- We review the state-of-the-art of the field by briefly surveying different research activities, such as the application of photonics or wireless solutions to traditional NoCs, identifying their potential advantages as well as their open issues.
- We also propose the use of graphene in the WNoC context due to its excellent properties. The resulting GWNoC will provide core-level communication in the THz band by means of graphene-based nano-antennas, achieving a superior performance in terms of bandwidth and area overhead.
- We outline and analyze the open issues and research challenges regarding the combination of graphene and wireless on-chip networks, serving as a roadmap for future investigations.

The remainder of this article is organized as follows. We first detail the different emerging solutions that have been proposed to address the shortcomings of traditional NoCs. Then we present GWNoC and analyze its unique features, as well as its different research challenges from both implementation and network perspectives. Finally, conclusions are drawn.

**STATE-OF-THE-ART OF ON-CHIP NETWORKING**

Multicore processor architectures rely on complex memory systems in order to provide fast and efficient means for data access and sharing between cores. The characteristics of such systems define the communication demands that the NoC must satisfy. Indeed, memory consistency, i.e., how the memory operations are ordered in execution, and cache coherence operations, i.e., maintaining a single memory image accessible to all processors, are especially critical in terms of latency. As the number of cores on a chip increases, traditional wireline topologies become insufficient for guaranteeing such latency conditions without significantly affecting other performance metrics.

In this context, disruptive solutions are required in order to alleviate the limitations of NoC in terms of latency while providing high bandwidth and maintaining affordable power and area overhead. Several approaches have been investigated and will be briefly introduced in the following subsections.

**3D NETWORK-ON-CHIP**

The creation of tridimensional integrated circuits, wherein layers of active devices are vertically interconnected, has shown to imply significant benefits such as improved noise immunity or higher packing density, as well as several advantages related to the NoC design [4]. For instance, the average wire propagation delay is considerably reduced due to the short distance between layers, i.e., tens of micrometers. Moreover, 3D stacking enables the use of topologies which would otherwise be unfeasible in the 2D design space, potentially yielding reduced multihop latency results [4]. Since such advantages are mainly at the network level, the potential improvements are compatible with and practically independent of the underlying interconnection technology. Moreover, it is an effective way to intuitively interface different technologies in hybrid approaches, facilitating modularity by avoiding the integration of different technologies in the same layer.

It is also important to note that 3D stacking presents considerable challenges. The superposition of active layers produces an increase in the heat density that must be circumvented in order to avoid thermal effects. Also, refined techniques are needed for the manufacture of such tridimensional integrated circuits and networks, in particular, alignment methodologies for the precise positioning of the vertical interconnects.

**RF INTERCONNECTS**

As the technology downscaling further narrows the performance bottleneck imposed by the delay and bandwidth figures of traditional wireline on-chip networks, recent research efforts have been focused on finding scalable alternatives in the physical layer design. In this regard, transmitting modulated RF signals over on-chip transmission lines is proposed in order to complement traditional wireline schemes [5]. Since signals propagate at nearly the speed of light, the propagation delay is significantly reduced and becomes independent of the link length at the chip scale. Moreover, the overall available data rate can be increased by simultaneously transmitting several frequency or code-multiplexed signals through a shared transmission line (Fig. 1). Such approach allows the interconnection of multiple cores using the same transmission line by assigning each core a channel and thus reducing the number of on-chip wires. The bandwidth could be dynamically allocated depending on the needs of each core.
However, the utilization of RF interconnects entails several open challenges. The circuitual implementation of frequency or code multiplexing transceivers produces both an area and power overhead that must be controlled as the size of the network increases. Also, the physical topology must be carefully designed as impedance mismatch reflections at the terminations of the transmission line may generate interferences.

**PHOTONIC NETWORK-ON-CHIP**

The advent of nanoscale silicon photonics has enabled the creation of photonic NoCs by means of the integration of CMOS-compatible optical building blocks. Such networks maintain the main advantages of RF interconnects in terms of latency, throughput and energy consumption, as shown in [8, 9].

Such benefits have favored the proposal of several preliminary and diverse WNoC designs. In principle, WNoC is usually regarded as a valid option when it is deployed in order to enhance the wired principal data NoC. In this hybrid approach, the unique advantages of wireless communication over existing wired solutions can be used to create a wireless control system [10]. Also, the insertion of wireless long-range point-to-point links is proposed in order to significantly decrease the average hop count of traditional NoC topologies, either performing fixed and regular positioning [8], explained in Fig. 2, or following the principles of small-world networks [9].

**GRAPHENE-ENABLED WIRELESS NETWORK-ON-CHIP**

The prospects of the WNoC paradigm are certainly promising. However, current implementations show several shortcomings related to the main enabler of WNoC: the on-chip antennas. For instance, the complete replacement of wired links for their wireless counterparts is proposed in [11] for the data plane, but it remains unclear how the area overhead figures will scale as the number of cores increases.
Indeed, the size of future metallic on-chip antennas, i.e., hundreds of micrometers [8], might render unfeasible the approach of integrating at least one antenna per core, as the core sizes continue to shrink with each CMOS technology generation and reach sizes of a few hundreds of micrometers. Also, since the available bandwidth is generally inversely proportional to the antenna size, metallic antennas may not be able to provide enough bandwidth in such a data intensive scenario. Such issues cannot be solved by further reducing the size of a metallic antenna, as this would impose the use of very high resonant frequencies, from the near infrared to the optical ranges. Due to the low mobility of electrons in metals when nanometer scale structures are considered, and the challenges in implementing a transceiver which will be able to operate at this extremely high frequency, the feasibility of inter-core wireless communications would be compromised if this approach would be followed.

Alternatively, we propose the employment of nanoscale wireless communication by means of graphene-based nano-antennas in order to unleash the full potential of the WNoC paradigm. Graphene-based nano-antennas just a few micrometers in size, i.e. two orders of magnitude below the dimensions of future metallic on-chip antennas, could provide inter-core communication in the Terahertz (0.1–10 THz) band. These characteristics will both enable size compatibility with each processor core and offer enough bandwidth for massively parallel multi-processors [12]. We refer to this new nanoscale wireless communication technique as Graphene-enabled Wireless Network-on-Chip (GWNoC) and we propose it as the basis of future on-chip network architectures.

Figure 3 shows a simple conceptual implementation of a GWNoC. It is important to note that all the processor cores are equipped with a graphene-based nano-antenna and a nano-transceiver, the latter of which prepares the information for outgoing transmissions and demodulates incoming transmissions. While maintaining the advantages of WNoC, the main benefits of the GWNoC approach are as follows.

**BANDWIDTH AND AREA LIMITATIONS**

Graphene-based nano-antennas support the propagation of tightly confined Surface Plasmon Polariton (SPP) waves. Due to their high effective mode index, the propagation speed of SPP waves can be up to two orders of magnitude below the electromagnetic wave propagation speed in vacuum [3]. In other words, graphene-based nano-antennas are expected to be two orders of magnitude smaller than metallic quarter wave antennas for the same resonant frequency. According to our preliminary results [3], a few micrometer wide and long nano-patch antenna could effectively radiate in the Terahertz band (Fig. 4). On the one hand, such reduced dimensions are comparable with future core sizes, i.e. few hundreds of micrometers, enabling the integration of one or multiple antennas per core and giving birth to the concept of wireless core in future generation multi-processors. On the other hand, the Terahertz band may offer enough bandwidth to accommodate the exponentially increasing requirements of multi-processors. Additionally, the employment of the Terahertz signals virtually eliminates the near-field effects between neighboring antennas since the near-field region at these frequencies is around one hundred micrometers.

**MULTICAST AND ALL-TO-ALL COMMUNICATION**

In multicore environments, a critical part of the on-chip traffic are short, generally multicast, control messages employed for cache coherence, data consistency or synchronization purposes. Moreover, some parallel applications require the transmission of large amounts of data in an all-to-all manner, e.g., 3D FFT calculation [13]. The majority of NoC solutions do not perform well under such conditions due to the difficulty of implementing efficient all-to-all communication in general, and multicast or broadcast schemes in particular. Conversely, since the information is radiated and can be potentially received by any receiver in the transmission range, GWNoC provides native broadcasting and multicasting capabilities, and makes
data transmission transparent with respect to the location of data within the chip. To the best of our knowledge, no other interconnect technology inherently offers such options, which ensure the scalability of traditional multicores architectures and open a vast design space at the architectural level. Efficiently transmitting simultaneously from multiple sources to multiple destinations heavily alleviates the constraints of parallel architecture design, therefore reducing the complexity of parallel programming.

MODULARITY

Replicability and modularity are two properties that can be achieved with the creation of the wireless core, in which a graphene nano-antenna and a transceiver are integrated within a core processor. A library of general-purpose or specific wireless cores could be created, allowing the building of custom multicores processors by the integration and initial configuration of a set of such predesigned cores. The cores of such multiprocessors will be wirelessly interconnected, potentially eliminating the necessity of a specific wireline on-chip network. Moreover, the modularity of a GWNoC also enables communication between circuit layers in a 3D approach without including additional wiring.

OPEN ISSUES AND RESEARCH CHALLENGES

The adoption of graphene as the basis of a new generation of WNoC brings up a wide variety of research challenges, covering areas from physical implementation up to computer architecture design. Such challenges need to be addressed in order to pave the way for the development of a new breed of multiprocessors.

ANTENNA CHARACTERIZATION

The further characterization of the graphene nano-antenna is central in order to assess the feasibility of the GWNoC approach and poses a grand challenge by itself. One important aspect to evaluate is the radiation efficiency of these antennas. Recent works report a radiation efficiency of up to 25 percent [14], which is reduced to a total efficiency of 4 percent when the impedance mismatch between the antenna and a transceiver. The dimensions of impedance mismatch between the antenna and a transceiver. The dimensions of graphene nanoribbons (GNRs) required for such RF components are quite relaxed compared to GNRs for logic devices [15] and hence it seems quite feasible to eventually obtaining graphene-based circuits operating in the Terahertz band. For instance, impressive cut-off frequencies ($f_T$) of 350 GHz have been obtained in Graphene Field-Effect Transistors (GFETs) due to the high carrier mobility in the nanomaterial [1]. Graphene is also uniquely suited for Low Noise Amplifiers (LNAs) as it theoretically offers high frequency and low noise.

Precise, efficient and replicable production of graphene nanoribbons (GNR) or bilayer GFETs is essential in order to provide high quality building blocks for the creation of the nano-antenna and the transceiver. The dimensions of GNRs required for such RF components are quite relaxed compared to GNRs for logic devices [15] and hence it seems quite feasible to achieve high-yield manufacturing of graphene nano-ribbons (GNRs).

Another important issue to investigate is how to integrate the antenna with the transceiver. The fundamental challenge here is to identify suitable heterogeneous integration techniques that make possible the integration of graphene into a semiconductor circuit environment. In general, graphene device technology is compatible with silicon technology [16]. However, graphene-dielectric interfaces and metal-graphene contacts need to be optimized as they limit the overall RF performance by reducing the carrier mobility [16] and introducing parasitic resistances [17]. Ultimately, another option may be to co-integrate other
The high bandwidth, inherent multicast and broadcast capabilities and extremely low area overhead offered by this novel approach could deliver a major breakthrough in massive multicore architectures. Several implementation, networking and architecture research challenges need to be addressed in order to achieve such ambitious goals.

COMMUNICATION AND NETWORKING

Being the radiation frequency known, a channel model is fundamental in order to evaluate the available on-chip communication bandwidth. Such models must analyze the presence of molecular absorption, which is due to the internal excitation of certain molecules at high frequencies. While preliminary results point out the possibility to utilize very large transmission bandwidths for short-range Terahertz communications [12], the multipath effects of within-package reflections are not taken into account.

The study of the propagation and reflection of the EM waves inside the chip package is also necessary due to the following. Generally, the radiation efficiency of planar antennas is extremely low in the coplanar direction and substantially higher in the transversal direction. Line-of-sight communication becomes very challenging, while communication through reflected EM waves might be feasible since the distance between the antenna and the chip package is in the millimeter range. This and other peculiarities of the on-chip wireless scenario, e.g., the constraints in terms of energy and chip area, require us to rethink the entire protocol stack with respect to classical wireless networks, including the following aspects:

- Special coding and modulation schemes need to be investigated in order to achieve efficient communication through reflected waves while avoiding, or even taking advantage of, multipath propagation.
- Addressing strategies for GWNoC must be carefully considered, as multicast and broadcast are envisaged to be essential for next generation multicore architectures.
- The flexibility offered by wireless communication allows to significantly reduce the multihop latency, at the expense of adding interference and medium access contention due to the existence of simultaneous transmissions in the same medium. A Medium Access Control (MAC) protocol could be designed and taken into account the design tradeoff between contention delay and multihop latency when the transmission range of the antennas is adaptively modified. The MAC protocol should also consider possible high-contention phases present in parallel programs.

MULTICORE ARCHITECTURE

Given the specifications of the physical implementation and protocols which enable intercore wireless communication and, on its turn, broadcast and multicast capabilities, a radically new paradigm in multicore architecture can be envisaged. Such all-to-all communication capabilities open a vast range of possibilities in terms of architecture design. For instance, at the memory level, cache coherency protocols could be modified and in a few cases even eliminated due to the possibility of implementing all-to-all on-chip communication. In this regard, we aim to explore the new multicore architecture design space opened by GWNoC in our future work.

CONCLUSIONS

In this position article, we present the vision of a novel and multidisciplinary research area in which nanoscale techniques enable core-level wireless communication for massive multicore processors. Although the concept of wireless on-chip networks is discussed in the literature, the size of the proposed on-chip antennas prevents such paradigm from fully exploiting its potential. We propose the employment of graphene-based nano-antennas for the design and development of flexible and scalable wireless on-chip networks. The high bandwidth, inherent multicast and broadcast capabilities and extremely low area overhead offered by this novel approach could deliver a major breakthrough in massive multicore architectures. Several implementation, networking and architecture research challenges need to be addressed in order to achieve such ambitious goals.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from the European Research Council through a Starting Grant (InteGraDe, No. 307311), the German Research Foundation (DFG, LE 2440/1-1), the European Commission (GRADE, 317839), the Generalitat de Catalunya under grant SGR 2009-1140, and the Spanish Ministry of Science and Innovation under grant EXPLO-RA-TEC2010-10440-E and projects TEC2010-15765 and RUE CSD2009-00046 (Consolider-Ingenio 2010). The authors would also like to thank Ignacio Llatser for his valuable comments that improved the quality of this article.

REFERENCES


**BIOGRAPHIES**

SERGI ABADAL (abadal@ac.upc.edu) received the B.Sc. and M.Sc. degree in Telecommunication Engineering from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2010 and 2011, respectively. From September 2009 to May 2010, he was visiting researcher at the Broadband Wireless Networking Lab, Georgia Institute of Technology, Atlanta. Since 2011, he is pursuing his Ph.D. at the NaNoNetworking Center in Catalunya (N3Cat, http://www.n3cat.upc.edu) at UPC. His current research interests are graphene-based wireless and nanophotonic communications for on-chip networks.

EDUARD ALARCÓN [S’96, M’01] (eduard.alarcon@upc.edu) received M.Sc. (national award) and Ph.D. degrees in EE from UPC Barcelona, Spain, in 1995 and 2000, respectively, where he became Associate Professor in 2001, and has been visiting Professor at University of Colorado at Boulder, USA (2003) and KTH Stockholm (2011). He has authored more than 250 scientific publications, 4 book chapters and 4 patents, and has been involved in different national, EU and US R&D projects. Research interests include the areas of on-chip energy management circuits, energy harvesting and wireless energy transfer, and nanonetworks. He was elected IEEE CAS society distinguished lecturer, elected member of the IEEE CAS Board of Governors (2010-2013), recipient of Best paper award at IEEEWNSCAS’98, co-editor of 4 journals special issues, 5 conference special sessions, TPC co-chair and TPC member of 15 IEEE conferences, and Associate Editor for IEEE TCAS-I, TCAS-II, JETCAS, JOLPE and Nano Communication Networks.

MAX C. LEMME (lemme@kth.se, max.lemme@uni-siegen.de) received the Dipl.-Ing. (M.Sc.) and Dr.-Ing. (Ph.D.) degree in Electrical Engineering from RWTH Aachen University in Germany. He is currently Heisenberg-Professor for Graphene-based Nanotechnology at the University of Siegen, Germany and Guest-Professor at KTH Royal Institute of Technology, Sweden since 2010. Before joining KTH, he was a research fellow at Harvard University from 2008-2010 and worked for 10 years at nanotechnology start-up AMO GmbH, Germany, as Head of Technology Department. His research interests include non-conventional nano-CMOS devices, novel high-k materials for gate stacks and, since 2005, graphene technology, devices and circuits. He received a NanoFutur Award by the German Federal Ministry for Education and Research (BMBF) in 2006, a Humboldt-Fellowship in 2007 and an ERC Starting Grant in 2012. He is a Senior Member of the IEEE.

MARIO NEMIROVSKY (mario.nemirovsky@bsc.es) received the Telecommunications Engineering degree from the National University of La Plata, Argentina, in 1980, and his Ph.D. in Electrical and Computer Engineering from the University of California, Santa Barbara, in 1990, where he was an adjunct professor from 1991 to 1998. After being chief architect in companies such as Apple Inc., National Semiconductors or General Motors (GM), he founded several renowned start-ups including FlowStorm Networks, Xstream Logic, ConSentry Networks or Miraveo. In 2007, he became an ICREA Senior Research Professor at the Barcelona Supercomputing Center (BSC). Mario holds more than 60 issued patents; he pioneered the concepts of Massively Multithreading (MMT) processing for the high performance processor and the by now well-established Simultaneous Multithreading architecture (SMT). He also architected the GM engine control being used in all GM cars for over 20 years. His current research interests include multithreaded multicore systems, high performance systems, network processors and Big Data.

ALBERT CABELLOS-APARICIO (acabello@ac.upc.edu) received a B.Sc. (2001), M.Sc. (2005) and Ph.D. (2008) degree in Computer Science Engineering from the Universitat Politècnica de Catalunya (UPC, http://www.upc.edu), Barcelona, Spain. In 2004 he was awarded with a full scholarship to carry out Ph.D. studies at the Department of Computer Architecture, UPC. In September 2005 he became an assistant professor of the Computer Architecture Department and a researcher in the Broadband Communications Group (http://cbb.upc.edu). In 2010 he joined the NaNoNetworking Center in Catalunya (http://www.n3cat.upc.edu) where he is the Scientific Director. He is an editor of the Elsevier Journal on Nano Computer Network and member of the Project Management Committee of the LISPMob open-source initiative (http://lispmob.org). His main research interests are future architectures for the Internet and Nano-networks.