

Unconventional Fabrics, Architectures, and Models for Future Multi-core Systems

Radu Marculescu
Carnegie Mellon University
Department of Electrical and
Computer Engineering
Pittsburgh, PA 15213-3890
radum@ece.cmu.edu

Christof Teuscher
Portland State University
Department of Electrical and
Computer Engineering
Portland, OR 97201
teuscher@pdx.edu

Partha Pratim Pande
Washington State University
School of Electrical Engineering and
Computer Science
Pullman, WA 99164
pande@eecs.wsu.edu

ABSTRACT

Massive level of integration is making modern multi-core chips all-pervasive in several domains. Hence, high performance, robustness, and low power are crucial for the widespread adoption of such platforms. However, achieving all these goals forces us to re-think the basis of designing multi-core systems at nanoscale, starting with the very *substrate* we need to use to implement such systems in the future, particularly for nanowire (or carbon nanotube) based on-chip interconnect obtained through self-assembly techniques. Due to the lack of control over these processes, such interconnects are expected to be largely unstructured [1][2][3]. While large unstructured networks are easy to fabricate, they require unconventional *architectures* and communication paradigms. For instance, by getting inspiration from many natural systems with network-based architectures [1], the future multi-core systems at nanoscale are expected to be hierarchical and heterogeneous in nature, as many powerful features such as increased performance, better resource utilization, and an increased robustness against failures of many natural networks come precisely from their heterogeneity, unstructuredness, and hierarchical nature. As such, an important performance limitation of multi-core chips designed with regular network architectures arises from planar metal interconnect-based multi-hop links, where the data transfer between two distant blocks can cause high latency and power consumption.

Different revolutionary approaches for creating low latency, long-range communication channels like optical interconnects, on-chip transmission lines and wireless interconnects have been explored, particularly in the context of network-on-chip (NoC) based communication. These emerging interconnect technologies can enable the design of hierarchical on-chip network architectures, where closely spaced cores communicate through traditional metal wires, but long distance communications is predominantly achieved through high performance specialized links. It is possible to find optimal interconnection architectures for massive multi-core chips by drawing inspiration from natural (complex)

networks that minimize resource consumption, while optimizing the relevant performance metrics, such as latency, throughput, power and area overhead.

The architecture-space exploration and optimization, however, require the incorporation of realistic *models* of network behavior depending on the region of network operation. Hence, a detailed understanding of the communication workload can be exploited to provide more performance and better resource utilization via network customization. At the same time, most application mapping or dynamic power management approaches considered so far for NoCs are based on average metrics that assume implicitly that the network is not congested and operates in stationary regimes. In practice, however, the network is used in regimes closer to congestion by most on-chip applications. Consequently, the optimization metric should also consider the communication dynamics in order to produce meaningful results and so completely new performance models are needed for successful optimization of future multi-core systems. These are just a few examples of network-related issues that require our immediate attention.

Starting from these ideas, this special session considers a holistic approach to the on-chip network paradigm and identifies a few critical issues related to the theoretical basis (e.g. graph theory, stochastic modeling and analysis), essential properties (e.g. structure, hierarchy, heterogeneity, dynamics, communication paradigm), and optimization metrics (e.g. energy, fault-tolerance, robustness, cost, performance) of designing and characterizing the communication infrastructure of future multi-core systems. Towards this end, this session consists of three forward-looking talks (going from the substrate-, all the way up to the application-level) addressing these fundamental challenges and opportunities.

The first talk by Christof Teuscher (“Self-assembled Nanoscale On-Chip Interconnect: The Good, the Bad and the Ugly”) addresses the opportunities and challenges of novel self-assembled nanoscale interconnects that are imperfect and disordered to a large extent. While such interconnects are relatively easy and cheap to fabricate, they pose many challenges for placement, routing, and QoS [1]. On the other hand, such interconnects have also many advantages, such as lower latencies because of non-local links and an increased and inherent robustness against certain types of failures. One of the challenges we address is to find network topologies that are both efficient and possible to fabricate by the chemical self-assembly of nanowires or carbon nanotubes. For that purpose, we have built a

network evaluation framework combined with simple wire-growth models, which allows us to determine the optimal points, i.e., “sweet spots,” in both the network and the physical design space. The ultimate goal is to relate the relevant network parameters to the chemical parameters of the self-assembly process, which would bring us a step closer to application-specific fabrication. The central message of this talk is that the benefits of disordered interconnects can be harnessed if we challenge the traditional communication and computing paradigms.

The second talk by Partha Pande (“Small-World Hybrid Wireless Network-on-Chip Architecture for Massive Multi-Core Systems”) addresses the performance limitations of traditional metal interconnect-based multi-hop on chip networks by inserting long-range single-hop wireless links between distant cores. Despite their advantages, an important performance limitation in regular (e.g., mesh like) NoCs arises from planar metal interconnect-based multi-hop links; wherein the data transfer between two distant blocks causes high latency and power consumption [4]. This limitation of conventional NoCs can be addressed by drawing inspiration from the human cerebral cortex. A large proportion of intra-cortical connections are established locally. Another large proportion of connections spread over longer distances, linking neurons that are located in different cortical regions. These connections ensure that the distant cortical sites can communicate fast enough in addition to the local intra-cortical communications. This network topology commonly referred to as small-world, can be incorporated in NoCs by introducing long-range, high bandwidth and low power wireless links between distant cores [5-6]. The central theme of this talk is to show the achievable performance benefit of the small-world wireless NoC (WiNoC) architecture, as well as various technological challenges and relevant design trade-offs for WiNoCs.

The third talk by Radu Marculescu (“Is the Network the Real Problem for Future Multi-core Systems?”) argues that, in contrast to traditional platforms such as PCs, servers, and parallel computers whose behavior has been well-studied, workloads have not been well characterized for multicore implementation of complex applications. Moreover, the current approaches for analyzing network performance are predominantly based on Poisson modeling which typically assumes exponentially distributed arrival processes. Unfortunately, these models are too simplistic to capture the complex behavior of the network when running real applications, particularly at high packet injection rates. Consequently, finding better approaches to workload characterization remains a fundamental open problem in multicore system design as precise workload characterization can enable better mechanisms for multi-tasking, routing protocols and priorities, power and memory management, caching - virtually all aspects of the computational infrastructure of multicore platforms [7]. The central message of this talk is that taking into account the characteristics of the workload becomes of crucial importance for performance analysis and optimization of the communication infrastructure, as well as proper resource management of future multicore platforms.

In summary, this session brings to discussion a few major challenges and appropriate design solutions that will enable multi-core systems to become the paradigm of choice when solving real problems in the future.

Categories and Subject Descriptors

J.6 [Computer-Aided Engineering]: Computer-Aided Design, C.2.1 [Computer-Communication Networks]: Network Architecture and Design, B.7.1 [Integrated Circuits]: Types and Design Styles.

General Terms

Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Theory.

Keywords

Multi-core, networks-on-chip, routing, architecture, interconnect, self-assembly, workload.

1. REFERENCES

- [1] C. Teuscher “Nature-inspired interconnects for emerging large-scale network-on-chip designs,” *Chaos*, 17(2), 026106. DOI= <http://dx.doi.org/10.1063/1.2740566>
- [2] C. Teuscher and A. A. Hansson, “Non-traditional irregular interconnects for massive scale SoC,” *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2785–2788.
- [3] C. Teuscher, N. Gulbahc and T. Rohlf, “An assessment of random dynamical network automata for nanoelectronics,” *International Journal of Nanotechnology and Molecular Computation*, 1(4):39–57.
- [4] A. Kumar et al., “Toward Ideal On-Chip Communication Using Express Virtual Channels,” *IEEE Micro*, Vol. 28, Issue 1, January-February 2008, pp. 80-90
- [5] U. Y. Ogras and R. Marculescu, “It’s a Small World After All”: NoC Performance Optimization Via Long-Range Link Insertion”, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 14, No. 7, July 2006, pp. 693-706.
- [6] A. Ganguly et al., “Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems”, *IEEE Transactions on Computers* (in press).
- [7] P. Bogdan and R. Marculescu, “Workload Characterization and its Impact on Multicore Platform Design,” in *Proc. IEEE/ACM Proc. Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, Scottsdale, AZ, Oct. 2010.