Networks on Chip: A New Paradigm for *Systems on Chip* Design

Luca Benini
DEIS - Università di Bologna
40136 Bologna Italy
lbenini@deis.unibo.it

Giovanni De Micheli
CSL - Stanford University
Stanford, CA 94305
nanni@stanford.edu

This paper is meant to be a short introduction to a new paradigm for *systems on chip* (SoC) design. We refer the interested reader to an extended overview of this problem [1] and to some recent results in this area in industry [21, 10] and academia [4, 5]. The premises are that a component-based design methodology will prevail in the future, to support component re-use in a plug-and-play fashion. At the same time, SoCs will have to provide a functionally-correct, reliable operation of the interacting components. The physical interconnections on chip will be a limiting factor for performance and energy consumption.

The *international technology roadmap for semiconductors* (ITRS) [23] projects that we will be designing multi-billion transistor chips by the end of this decade, with feature sizes around $50nm$ and clock frequencies around $10GHz$. Delays on wires will dominate: global wires spanning a significant fraction of the chip size will carry signals whose propagation delay will exceed the clock period. Whereas relatively large delays can be managed with wire pipelining techniques, timing uncertainty will be more problematic for designers. Moreover, synchronization of chips with a single clock source and negligible skew will be extremely hard or impossible. The most likely synchronization paradigm for future chips is *globally-asynchronous locally-synchronous* (GALS), with many different clocks. Global wires will span multiple clock domains, and synchronization failures in communicating between different clock domains will be rare but unavoidable events [7].

SoC design will be guided by the principle of consuming the least possible energy. This requirement matches the need of using SoCs in portable battery-powered electronic devices and of curtailing thermal dissipation which can make chip operation unfeasible or impractical. Energy considerations will impose small logic swings and power supplies, most likely below 1 Volt. Electrical noise due to *cross-talk*, *electromagnetic interference* (EMI) and radiation-induced charge injection (*soft errors*) [6] will be likely to produce additional timing errors or data errors. Regardless the source of errors, the end result is that the mere transmission of digital values on wires will be *inherently unreliable* [12]

The distinguishing challenge for SoC design will be to provide adequate *quality of service* (QoS), with a limited energy budget under strong limitations of the technology. QoS requirements include, but are not limited to, performance and reliability. High performance is demanded by the increasingly complex software applications required to run even on small portable appliances (e.g., multi-media on cell phones). Reliability is mandated by the increasing reliance of consumers on electronic communication and control systems in everyday's life.

We propose to use network design technology to analyze and design SoCs. In other words, we view a SoC as a *micro-network* of components. We postulate that SoC interconnect design can be done using the *micro-network stack* paradigm, which is an adaptation of the protocol stack [19] (Figure 1). Thus the electrical, logic, and functional properties of the interconnection scheme can be abstracted.

SoCs differ from wide-area networks because of local proximity and because they exhibit much less non-determinism. Local, high-performance networks (such as those developed for large-scale multiprocessors), have similar requirements and constraints. A few distinctive characteristics are unique of SoC networks, namely, energy constraints and design-time specialization.

The ITRS [23] projects that overall power consumption can marginally scale up in future technologies. Even though voltage downscaling will be beneficial, it will not be sufficient to contrast the overall growth in complexity. Indeed, whereas computation and storage energy greatly benefits from device scaling (smaller gates, smaller memory cells), the energy for global communication does not scale down. On the contrary, projections based on current delay optimization techniques for global wires [17] show that global communication on chip will require increasingly higher energy consumption. Hence, communication-energy minimization will be a growing concern in future technologies. Furthermore, network traffic control and monitoring can help in better managing the power consumed by networked computational resources. For instance, clock speed and voltage of components can be varied according to available network bandwidth. The emphasis on energy minimization creates a sleuth of novel challenges that have not been addressed by traditional network designers [2].

Design-time specialization is another facet of the SoC network design. Whereas macroscopic networks emphasize general-purpose communication and modularity, in SoC networks these constraints are less restrictive. The communication network fabric is designed on silicon from scratch. Standardization is needed only for specifying an abstract network interface for the components, but the network architecture itself can be tailored to the application, or class of applications, targeted by the SoC design. Hence,

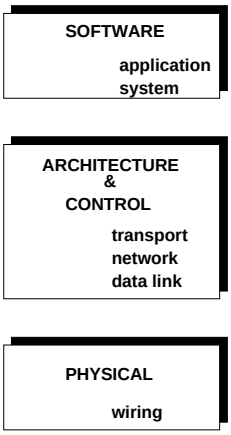

Figure 1. Micro-network stack

we envision a vertical design flow where every layer of the micro-network stack is specialized and optimized for the target application domain. Such *application-specific on-chip network synthesis* paradigm represents, in our view, an open and exciting research field. Needless to say, specialization does not imply loss of flexibility. From a design stand-point the network reconfigurability will be key in providing plug-and-play use of components, since they interact with the others through (reconfigurable) protocols.

To date, the physical layer design for SoCs has been addressed by complex and expensive CAD tools. We think that there will be a diminishing return in attempting to design fail-safe layouts in deep submicron technologies, because problems just get increasingly harder. We believe that appropriate interconnection architectures and control protocols can cope with the unreliability of the physical layer. Techniques such as bus encoding [5] and packetization [4, 11] have been proposed, but much research has still to be carried out to determine which networking techniques are most appropriate for SoCs and which parameters fit best this new context. Eventually we think that software will play a major role in network control at different layers, from protocol implementation to network reconfiguration and to management of component service levels and power consumption. Overall, we think this new perspective opens an exciting area of research in the years to come.

References

[1] L. Benini and G. De Micheli, "Networks on Chip: A New SoC paradigm," *IEEE Computer*, January 2002.

[2] L. Benini, G. De Micheli, "Powering Networks on Chip," *ISSS - International System Synthesis Symposium*, October 2001, pp. 33-38.

[3] L. Benini, G. De Micheli, "System-Level Power Optimization: Techniques and Tools," *ACM Transactions on Design Automation of Electronic Systems*, vol. 5, no. 2, pp. 115-192, April 2000.

[4] W. Dally, "Route packets, not wires: on-chip interconnection networks," *DAC - Design Automation Conference*, pp. 684–689, June 2001.

[5] D. Bertozzi, L. Benini and G. De Micheli, " Low-Power Error Resilient Encoding for On-chip Data Buses," *DATE - International Conference on Design and Test Europe*, 2002.

[6] N. Cohen, T. Sriram, N. Leland, D. Moyer, S. Butler and R. Flatley, "Soft Error Considerations for Deep-Submicron CMOS Circuit Applications," *IEDM, Proceedings of IEEE International Electron Device Meeting*, pp. 315-318, 1999.

[7] W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge University Press, 1998.

[8] A. Deutsch, "Electrical Characteristics of Interconnections for High-Performance Systems," *Proceedings of the IEEE*, vol. 86, no. 2, pp. 315–355, February 1998.

[9] J. Duato, S. Yalamanchili, L. Ni, *Interconnection Networks: an Engineering Approach*. IEEE Comp. Society Press, 1997.

[10] K.Goossens, E. Rijpkema, P. Wielage, A. Peters and J. van Meerbergen, "Networks on Silicon: Combining Best Effort and Guaranteed Service," *DATE - International Conference on Design and Test Europe*, 2002.

[11] P. Guerrier, A. Grenier, "A generic architecture for on-chip packet-switched interconnections," *Design Automation and Test in Europe Conference*, pp. 250–256, 2000.

[12] R. Hedge, N. Shanbhag, "Toward achieving energy efficiency in presence of deep submicron noise," *IEEE Transactions on VLSI Systems*, pp. 379–391, vol. 8, no. 4, August 2000.

[13] R. Ho, K. Mai, M. Horowitz, "The Future of wires," *Proceedings of the IEEE*, January 2001.

[14] J. Montanaro et al., "A 160-MHz, 32-b, 0.5-W CMOS RISC Microprocessor," *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1703–1714, Nov. 1996.

[15] C. Patel, S. Chai, S. Yalamanchili, D. Shimmel, "Power constrained design of multiprocessor interconnection networks," *IEEE International Conference on Computer Design*, pp. 408-416, 1997.

[16] W. Remaklus, "On-chip bus structure for custom core logic design," *IEEE Wescon*, pp. 7–14, 1998.

[17] D.Sylvester and K.Keutzer, "A Global Wiring Paradigm for Deep Submicron Design," *IEEE Transactions on CAD/ICAS*, Vol.19, No. 2, pp. 242-252, February 2000.

[18] T. Theis, "The future of Interconnection Technology," *IBM Journal of Research and Development*, Vol. 44, No. 3, May 2000, pp. 379-390.

[19] J. Walrand, P. Varaiya, *High-Performance Communication Networks*. Morgan Kaufman, 2000.

[20] W. Weber, "CPU Performance Comparison: Standard Computer Bus Versus SiliconBacplane," *www.sonics.com*, 2000.

[21] J. Williams, N.Heintze and B. Ackland, "Communication Mechanisms for Parallel DSP Systems on Chip," *DATE - International conference on Design and Test Europe*, 2002.

[22] R. Yoshimura, T. Koat, S. Hatanaka, T. Matsuoka, K. Taniguchi, "DS-CDMA wired bus with simple interconnection topology for parallel processing system LSIs," *IEEE Solid-State Circuits Conference*, pp. 371-371, Jan. 2000.

[23] http://public.itrs.net/