Tackling the MPSoC Data Locality Challenge for Distributed-Shared Memory Architectures

Location: Room **Bianca A**

Scheduled time: **10:15 – 11:00**

Speaker: Andreas Herkersdorf

Andreas Herkersdorf is a professor in the Department of Electrical and Computer Engineering and also affiliated to the Department of Informatics at Technical University of Munich (TUM). He received a Dr. degree from ETH Zurich, Switzerland, in 1991. Between 1988 and 2003, he has been in technical and management positions with the IBM Research Laboratory in Rüschlikon, Switzerland.

Since 2003, Dr. Herkersdorf is the head of the Chair of Integrated Systems at TUM. He is a senior member of the IEEE, member of the DFG (German Research Foundation) Review Board and serves as editor for Springer and De Gruyter journals for design automation and information technology. His research interests include application-specific multi-processor architectures, IP network processing, Network on Chip and self-adaptive fault-tolerant computing.

Abstract

Data access latencies and bandwidth bottlenecks frequently represent major limiting factors for the computational effectiveness of multi- and many-core processor architectures. This talk introduces two conceptually complementary approaches to reduce the synchronization overheads for coherence maintenance and to improve the locality between computing resources and data: Region-based cache coherence and near memory acceleration. The presented approaches represent current work in the DFG Transregional Collaborative Research Center “Invasive Computing”.

A 2D array of compute tiles with multiple, heterogeneous RISC cores, two levels of caches and a tile-local SRAM memory serves as reference processing platform. Compute tiles, I/O tiles and globally shared DDR SDRAM memory tiles are interconnected by a meshed Network on Chip (NoC) with support for multiple quality of service levels. Overall, this processing architecture follows a distributed-shared-memory model. The limited degree of parallelism in many embedded computing applications also bounds the number of compute tiles possibly sharing associated data structures. Therefore, we envision region-based cache coherence (RBCC) among a limited working set of compute tiles over global coherence approaches. Coherence regions can dynamically be reconfigured at runtime and comprise a number of arbitrary (adjacent or non-adjacent) compute tiles which are interconnected through regular NoC channels for the exchange of coherency protocol messages. We will show that region-based coherence allows maintaining substantially smaller coherence directories (e.g., by approx. 40% reduced in size for 16 tiles systems with up to 4 tiles per region) and shorter sharer checking latencies than global coherence.

Near memory processing is an alternative concept to increase data/task locality by means of near memory accelerators (NMA). NMA positions processing resources for specific forms of data manipulations as close as possible to the data memory. The evident benefits are: reducing global interconnect usage, shortening of access latencies and, thus, increasing compute efficiency. In distributed-shared-memory architectures, where accelerator units can be affiliated with different tile-local SRAMs as well as with the globally shared DDR SDRAM, near memory acceleration requires thorough consideration of task mapping as well as task and data migration into and among compute tiles.