CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar

"Regional Coherence and Near Memory Acceleration in Distributed-Shared-Memory Architectures"

Andreas Herkersdorf

Chair of Integrated Systems at Technical University of Munich

Thursday, July 26th
11:00 a.m.- 12:00 p.m.
Donald Bren Hall 3011

Abstract: Data access latencies and bandwidth bottlenecks frequently represent major limiting factors for the computational effectiveness of many-core processor architectures. This presentation 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.
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. These compute tiles, various I/O tiles and a globally shared DDR SDRAM memory tile 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 favor region-based cache coherence (RBCC) among a limited number of compute tiles over global coherence approaches. Coherence regions are dynamically configured 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. RBCC increases the locally usable intra-tile shared SRAM memories and may reduce execution times of sample video processing applications by 30% in comparison to message passing based parallelization. However, the benefits of RBCC may strongly depend on the task and data placement among tiles in the coherency region which can affect performance by up to an order of magnitude. Near memory processing using near memory accelerators (NMA) positions processing resources for specific data manipulations as close as possible to the data memory for the benefit of shortening access latencies and increasing compute efficiency.

Biography: 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 leads 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.