Scaling the Many-Memory Wall for Many-Core Architectures

Nikil Dutt

Center for Embedded Computer Systems
Department of Computer Science
University of California, Irvine
Irvine, CA 92697-3435, USA
Email: dutt@uci.edu

Abstract

The move towards many-core architectures creates an inherent demand for high memory bandwidth, which in turn results in the need for vast amounts of on-chip memory space. On the other hand, many-core architectures have many (distributed) on-chip memories with limited capacities, resulting in a "many-memory wall". While efforts such as 3D stacking and smarter memory controllers try to alleviate the off-chip memory access problem, there is still a pressing need to carefully provision the limited on-chip memory budget to meet application needs. For on-chip memories, embedded systems often use both software controlled memories (e.g., scratchpad memories) and hardware-controlled memories (e.g., caches), with each having their pros and cons. Efficient on-chip memory management is extremely critical as it has a great impact on the system's power consumption and throughput. Traditional memory hierarchies primarily consist of SRAM-based on-chip caches. However, with the emergence of non-volatile memories (NVMs) and mixed-criticality systems, we expect to see heterogeneous on-chip memory hierarchies, not only in type (cache vs. scratchpad) but also in technology (e.g., SRAM vs. NVM). This talk will survey the state of the art in memory subsystems for many-core platforms, and present strategies for efficiently managing software-controlled memories in the many-core domain, while addressing emerging challenges faced by designers. I will also propose a holistic software/hardware solution to the problem of scaling the memory wall for many-core architectures.