SORS: Understanding the Potential of Real Processing-in-Memory for Modern Workloads

Objectives

Abstract: Processing-in-memory (PIM) is becoming a reality which promises to overcome the data movement bottleneck (i.e., the waste of execution cycles and energy due to frequent movement of data between memory and compute units) by equipping compute systems with compute-capable memories. Several major vendors and startups have prototyped and announced their PIM architectures. Among them, the UPMEM company commercializes the first publicly-available real-world PIM architecture, which combines traditional DRAM memory arrays with general-purpose in-order cores, called DRAM Processing Units (DPUs), integrated in the same chip. In this talk, we will provide an overview of our experiences with several important primitives and modern applications (genomics, machine learning, sparse linear algebra, transcendental functions, etc.) on a real memory-centric computing system with the UPMEM PIM architecture. Outcomes of our work are several key observations, takeaways, and recommendations for software designers, and suggestions and hints for hardware and architecture designers of future PIM systems.

Short bio: Juan Gómez-Luna is a senior researcher and lecturer at SAFARI Research Group @ ETH Zürich. He received the BS and MS degrees in Telecommunication Engineering from the University of Sevilla, Spain, in 2001, and the PhD degree in

Computer Science from the University of Córdoba, Spain, in 2012. Between 2005 and 2017, he was a faculty member of the University of Córdoba. His research interests focus on processing-in-memory, memory systems, heterogeneous computing, and hardware and software acceleration of medical imaging and bioinformatics. He is the lead author of PrIM (https://github.com/CMU-SAFARI/prim-benchmarks), the first publicly-available benchmark suite for a real-world processing-in-memory architecture, and Chai (https://github.com/chai-benchmarks/chai), a benchmark suite for heterogeneous systems with CPU/GPU/FPGA.

Speakers

Speaker: Juan Gómez-Luna, senior researcher and lecturer at SAFARI Research Group @ ETH Zürich
Host: Toni Peña, Accelerators for High Performance Computing Group Manager, CS, BSC

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---