Memory-Centric Computing

Onur Mutlu (ETH Zürich, CH)

Abstract

Computing is bottlenecked by data. Large amounts of application data overwhelm storage capability, communication capability, and computation capability of the modern machines we design today. As a result, many key applications' performance, efficiency, and scalability are bottlenecked by data movement. In this lecture, we describe three major shortcomings of modern architectures in terms of 1) dealing with data, 2) taking advantage of the vast amounts of data, and 3) exploiting different semantic properties of application data. We argue that an intelligent architecture should be designed to handle data well. We show that handling data well requires designing architectures based on three key principles: 1) data-centric, 2) data-driven, 3) data-aware. We give several examples for how to exploit each of these principles to design a much more efficient and high performance computing system. We especially discuss recent research that aims to fundamentally reduce memory latency and energy, and practically enable computation close to data, with at least two promising novels directions: 1) processing using memory, which exploits analog operational properties of memory chips to perform massively-parallel operations in memory, with low-cost changes, 2) processing near memory, which integrates sophisticated additional processing capability in memory controllers, the logic layer of 3D-stacked memory technologies, or memory chips to enable high memory bandwidth and low memory latency to near-memory logic. We show both types of architectures can enable orders of magnitude improvements in performance and energy consumption of many important workloads, such as graph analytics, database systems, machine learning, video processing, climate modeling, genome analysis. We discuss how to enable adoption of such fundamentally more intelligent architectures, which we believe are key to efficiency, performance, and sustainability. We conclude with some research opportunities in and guiding principles for future computing architecture and system designs.

A short accompanying paper, which appeared in DATE 2021, can be found here and serves as recommended reading:
https://people.inf.ethz.ch/omutlu/pub/intelligent-architectures-for-inte... [1]

A longer overview & survey of modern memory-centric computing can be found here and also serves as recommended reading:
"A Modern Primer on Processing in Memory"
https://arxiv.org/abs/2012.03112 [2]

Curriculum Vitae

Onur Mutlu is a Professor of Computer Science at ETH Zurich. He is also a faculty member at Carnegie Mellon University, where he previously held the Strecker Early Career Professorship. His current broader research interests are in computer architecture, systems, hardware security, and bioinformatics. A variety of techniques he, along with his group and collaborators, has invented over the years have influenced industry and have been employed in commercial microprocessors and memory/storage systems. He obtained his PhD and MS in ECE from the University of Texas at Austin and BS degrees in Computer Engineering and Psychology from the University of Michigan, Ann Arbor. He started the Computer Architecture Group at Microsoft Research (2006-2009), and held various product and research positions at Intel Corporation, Advanced Micro Devices, VMware, and Google. He received the Huawei OlympusMons Award for Storage Systems Research, Google Security and Privacy Research Award, Intel Outstanding Researcher Award, IEEE High Performance Computer Architecture Test of Time Award, NVMW Persistent Impact Prize, the IEEE Computer Society Edward J. McCluskey Technical AchievementAward, ACM SIGARCH Maurice Wilkes Award, the inaugural IEEE Computer Society Young Computer Architect Award, the inaugural Intel Early Career Faculty Award, US National Science Foundation CAREER Award, Carnegie Mellon University Ladd Research Award, faculty partnership

awards from various companies, and a healthy number of best paper or "Top Pick" paper recognitions at various computer systems, architecture, and security venues. He is an ACM Fellow "for contributions to computer architecture research, especially in memory systems", IEEE Fellow for "contributions to computer architecture research and practice", and an elected member of the Academy of Europe (Academia Europaea).

His computer architecture and digital logic design course lectures and materials are freely available on YouTube (https://www.youtube.com/OnurMutluLectures [3]), and his research group makes a wide variety of software and hardware artifacts freely available online (https://safari.ethz.ch/ [4]).

For more information, please see his webpage at https://people.inf.ethz.ch/omutlu/ [5].

edacentrum | Schneiderberg 32 | 30167 Hannover | fon: +49 511 762-19699 | email: info@edacentrum [dot] denach oben

---

Links:
[1] https://people.inf.ethz.ch/omutlu/pub/intelligent-architectures-for-intelligent-computingsystems-invited_paper_DATE21.pdf
[2] https://arxiv.org/abs/2012.03112
[3] https://www.youtube.com/OnurMutluLectures
[4] https://safari.ethz.ch/
[5] https://people.inf.ethz.ch/omutlu/