The Walls and the Dark Silicon Era: An Arithmetic Perspective

Louis Ledoux
Department of Computer Architecture
Barcelona Supercomputing Center (BSC)
Universitat Politècnica de Catalunya (UPC)

December, 2nd, 2024

Memory wall, power wall, dark silicon, and even the adage "communication dominates arithmetic" are persistent challenges for computer architects and the High Performance Computing (HPC) community. The term "computer," from the Latin *computare* ("to calculate"), implies that computation is central to its purpose. However, HPC and supercomputers perform many non-arithmetic tasks, including synchronization, data movement, and OS management. Due to the general-purpose nature of these HPC tasks and the need for efficiency, particular attention must be given to arithmetic design. This need results in a high-dimensional space of parameters, which, combined with the sparsity of numerical requirements in scientific workloads, naturally expands the design space for exploring arithmetic architectures and number representations.

This presentation discusses software and hardware solutions to address these architectural choices while tackling the challenges posed by the various "walls." Specifically, we examine two trade-offs: 1) *accuracy (precision) versus energy efficiency in light of numerical requirements*, and 2) *parallelism versus latency for SIMD/vector floating-point operations*. Additionally, we may explore topics such as a custom arithmetic ASIC flow or Multi-Level Intermediate Representations (MLIR) in enhancing flexibility and efficiency across this design space.