Massively-Parallel Heterogeneous Computing for Solving Large Problems

Wen-mei Hwu, Mert Hidayetoğlu, Carl Pearson, Simon Garcia, Sitao Huang, and Abdul Dakkak

Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

w-hwu@illinois.edu

Since the beginning of the 21st century, we have been witnessing a major paradigm shift in science and industry innovations. While major innovations in the 20th century were primarily driven by physical instruments such as light sources, transceivers, and satellites, the high-valued innovations in the 21st century has been primarily driven by computation methods. Examples include computational microscopy, deep-space telescopes, telepresence, and self-driving cars. Tremendous amount of resources have been invested into innovative applications such as first-principle based models, deep learning, and cognitive computing. Many application domains are questioning the conventional "it is too expensive" thinking that led to inaccuracies and missed opportunities. As a result, the size of the problems being solved has increased rapidly in order to meet the fidelity requirements.

Consider a time-domain molecular dynamics simulation: its scaling can be investigated in two ways. Scaling out includes more number of computational nodes, where each node has a fixed amount of memory and CPU power. Spreading the problem among nodes and allows accessing more amount of memory, which provides solution of problems with larger sizes with greater number of grid points. Scaling up improves the single-node computational power. This speeds up the simulation of each time step, which is inherently not parallel because of causality relationships of the steps. Speaking of scaling up, the saturation of single-core CPU performance in the last decade forced the computing industry to move into the direction of multi-codes CPU and graphics processing units (GPUs). The hierarchical memory architecture of GPUs overcomes the memory-speed bottleneck of multi-core CPUs, providing four to five times speedup for many algorithms, compared to multi-core CPUs.

This talk is on the development and future trends of heterogeneous computing, not only with multicore CPUs and GPUs, but also with FPGAs, many-core CPUs, flash disks, etc. Additionally, the talk stipulates that it is imperative to move to reduced-complexity solvers as we continue to pursue innovations by increasing the size of the computing systems into exascale.