Overview of HPC Computer Architecture: A Long March Toward Exa-Scale Computing and Beyond

Guang R. Gao
ACM Fellow and IEEE Fellow
Endowed Distinguished Professor
Department of Electrical and Computer Engineering
University of Delaware

The microprocessor chip technology revolution has moved rapidly to large-scale, on-chip (heterogeneous), multiprocessing to facilitate parallel computing at an unprecedented scale. Often, this involves the incorporation of different architecture types in the same system. Growing core-count and heterogeneity will result in large-scale systems with increasingly fragmented resource pools. Achieving efficiency, reliability and high performance will require increasingly flexible, dynamic, and intelligent runtime resource management with corresponding activity scheduling for applications that is not present in existing software stacks. The needs far exceed what has existed in traditional runtimes support from current system vendors – and there is a clear *gap* in the industry-strength system software stack.

In this context, we present several research challenges we are addressing while matching toward a dynamic adaptive runtime model and related system software system for exa-scale computing.