Heterogeneous Computing for Scientific Machine Learning

Laurent White

Advanced Micro Devices, Inc.

More than ever, the semiconductor industry is asked to answer society's call for more computing capacity and capability, which are driven by rapid digitalization, the widespread adoption of artificial intelligence, and the ever-increasing need for high-fidelity scientific simulations. While facing high demand, the supply of computing capability is being technically challenged by the slowdown of Moore's law and the need for high energy efficiency. This tug-of-war has now pushed the industry towards domain-specific accelerators, perhaps likely past the point of no return. The mix of general-purpose CPUs and high-end GPGPUs, which has pervaded data centers over the past few years, is likely to be expanded to a much richer set of application-specific accelerators, including AI engines, reconfigurable hardware, and even perhaps quantum, annealing, and neuromorphic devices. While acceleration and better efficiency may be enabled by using domain-specific accelerators for selected workloads, a much more holistic (i.e., system-wide) approach will have to be adopted to achieve significant performance gains for complex applications that consist of a variety of workloads where each could benefit from a specific accelerator. As an important example, scientific computing, which increasingly incorporates AI training and inference kernels in a tightly-integrated fashion, provides a rich and exciting laboratory for addressing the challenges of efficiently using highly-heterogeneous systems and for ultimately realizing their promises. Those challenges include co-designing the application, which requires domain experts to collaborate with other experts across the stack for workload mapping and data orchestration, and also adopting a decentralized strategy that embeds processing units where the data need them. Finally, the early experience of those co-design efforts should help the industry devise a longer-term strategy for developing programming models that would relieve application experts from what is often perceived as the burden of hardware-aware development and code optimization.