Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M20-163P
Published: 1/15/2021

Inventors

Michael Riera
Fengbo Ren
Masudul Quraishi
Erfan Bank Tavakoli

Contact

Shen Yan
shen.yan@skysonginnovations.com

Framework for Extreme-Scale Heterogeneous High-Performance Computing (HPC)

In recent years, new accelerator choices for heterogeneous HPC have emerged, such as field-programmable gate arrays (FPGAs, considered as reconfigurable accelerators) and tensor processing units (TPUs, considered as domain-specific accelerators). Although these options offer flexible or customized hardware architectures with excellent capabilities for exploiting temporal/pipeline parallelism efficiently, their adoption in extreme-scale scientific computing is still in its infancy and is expected to be a tortuous process (as was the adoption of GPUs) regardless of their superior performance and energy efficiency benefits.

The fundamental challenge to the adoption of these and other new accelerators is that each accelerator's programming model, communication model, and virtualization stacks are developed independently and are specific to the respective hardware architecture. As a result, the pervasive adoption of a new accelerator in HPC not only requires a hardware vendor to develop complicated virtualization stacks to abstract away the accelerator hardware from HPC applications, but also demands domain experts to produce code or reengineer legacy codes using different programming and communication models. Both of these requirements, unfortunately, are artificial barriers that make the adoption of new accelerator technology in scientific computing a very slow and laborious process, impeding the realization of extreme-scale heterogeneous HPC.

Researchers at Arizona State University have developed a framework for extreme-scale heterogeneous HPC. Included are three critical enabling technologies: (1) a computation-centric message passing interface (C2MPI), (2) a hardware-agnostic accelerator orchestration (HALO) software framework, and (3) a standalone accelerator protocol (SAP). These technologies unify the programming and communication models, and the virtualization stacks of various accelerators for extreme-scale heterogeneous HPC through a systematic, scalable solution-as-a-service approach. The resulting unification of system environments delivers true accelerator interoperability with intelligent, adaptive resource management, which is the key to realizing true performance portability and scalability for extreme-scale heterogeneous HPC. This framework allows quick and effortless adoption of existing accelerator technologies (e.g., GPUs, FPGAs, and TPUs) and any future accelerator technologies for the joint and synergetic acceleration of complex scientific computing tasks in a hardware-agnostic fashion.

Potential Applications:

- High-performance computing, cloud computing
- GPUs, FPGAs, TPUs, and other domain- and application-specific accelerators

Benefits and Advantages:

- Hardware-agnostic programming against a unified application programming interface (API) and true accelerator interoperability.
- True performance portability and scalability without needing to change application codes.
- Quick and effortless adoption of all types of accelerators in HPC.
- Standardized protocol and IPs for empowering accelerator or computing hardware vendors to enter the HPC market with no more barriers.
- Highly bindable API to other programming languages.

Related Publication (PDF):

HALO 1.0: A Hardware-agnostic Accelerator Orchestration Framework for Enabling Hardware-agnostic Programming with True Performance Portability for Heterogeneous HPC