HCW 2021 Keynote Speaker
Re-Imagining Hardware/Software Co-Design for Extremely Heterogeneous Systems

Roberto Gioiosa
High Performance Computing
Pacific Northwest National Laboratory
Richland, Washington USA

Abstract: The need of processing and analyzing extremely large amount of data with real-time and power/energy constraints has motivated the development of many highly-specialized and energy efficient architectural concepts. This trend is evident in embedded systems, such as mobile phones or smart sensors, where systems contain a myriad of small, specialized ASIC processors. Large-scale, HPC systems have also embraced heterogeneous devices to speed up computation while maintaining a strict power budget. Looking forward, one can imagine that more application-specific accelerators will be incorporated into SoC designs, which will contain general-purpose, programmable processing elements, such as CPU and GPU cores, fixed, application-specific accelerators (e.g., FFT or GEMM), and semi-programmable CGRA devices. Designing such increasingly complex device become and incredibly difficult task.

In the AI space, this revolution is already happening, with many custom hardware designs already available. The Center for co-design of ARtificial Intelligence focused Architectures and Algorithms (ARIAA) is a DOE/ASCR project lead by Pacific Northwest National Laboratory (PNNL) in collaboration with Sandia National Laboratory (SNL), Georgia Tech (GT), NVIDIA, and Qualcomm. ARIAA's objectives are to co-design novel architectures, algorithms, and programming abstractions to enable AI-based DOE applications and support sparse, explainable, and domain-informed AI models. Ultimately, ARIAA aims at understanding how AI-focused architectures can accelerate traditional, emerging, and AI DOE workloads, identifying computational kernels that can be effectively replaced by accurate AI/ML methods, identifying opportunities to leverage AI/ML methods to support computation and data analytics, and understanding and designing AI accelerators for future explainable and domain-aware AI methods. This talk will describe ARIAA's novel approach to co-design of AI/HPC accelerators, programming abstractions, and algorithms and how various technologies are integrated to form and end-to-end solution.

About the speaker: Dr. Roberto Gioiosa is a senior research scientist at the Pacific Northwest National Laboratory (PNNL) in the High-Performance Computing Group. His research interests include operating systems and runtimes, high-performance computer architectures, memory, and networks, parallel and distributed programming models, resilience, performance and power modeling and analysis, and embedded systems.

Dr. Gioiosa earned his Ph.D. in 2006 from the University of Rome "Tor Vergara", Rome Italy. He has worked at the Los Alamos National Laboratory (LANL) (2004-2005), the Barcelona Supercomputing Center (BSC) (2006-2008 and 2009-2012), the IBM T.J. Watson Research Center (2008-2009) where he contributed to the development of the Compute Node Kernel for BG/Q systems, and Oak Ridge National Laboratory (ORNL) (2017-2018).

Currently, his projects include the development of system software for extremely heterogeneous systems, system software for scalable distributed systems, DSSoC design, evaluation of emerging architecture and technologies for exascale systems and applications, and development of operating systems for exascale systems. Dr. Gioiosa leads the DOE/ASCR Center for co-design of ARtificial Intelligence focused Architectures and Algorithms (ARIAA).
He is a member of the ACM and IEEE Computer Society.