Title	*Keynote SUSTAINABLE HIGH-PERFORMANCE COMPUTING VIA DOMAIN-SPECIFIC ACCELERATORS*
Speaker	**William J. Dally,** Stanford University and NVIDIA, United States

Abstract

High-Performance computers require continued scaling of performance and efficiency to handle more demanding applications and scales. With the end of Moore's Law and Dennard Scaling, continued performance scaling will come primarily from specialization. Specialized hardware engines can achieve performance and efficiency from 10x to 10,000x a CPU through specialization, parallelism, and optimized memory access. Graphics processing units are an ideal platform on which to build domain-specific accelerators. They provide very efficient, high performance communication and memory subsystems - which are needed by all domains. Specialization is provided via "cores", such as tensor cores that accelerate deep learning or ray-tracing cores that accelerate specific applications.

Biography

Bill is Chief Scientist and Senior Vice President of Research at NVIDIA Corporation and a Professor (Research) and former chair of Computer Science at Stanford University. Bill is currently working on developing hardware and software to accelerate demanding applications including machine learning, bioinformatics, and logical inference. He has a history of designing innovative and efficient experimental computing systems. While at Bell Labs Bill contributed to the BELLMAC32 microprocessor and designed the MARS hardware accelerator. At Caltech he designed the MOSSIM Simulation Engine and the Torus Routing Chip which pioneered wormhole routing and virtual-channel flow control. At the Massachusetts Institute of Technology his group built the J-Machine and the M-Machine, experimental parallel computer systems that pioneered the separation of mechanisms from programming models and demonstrated very low overhead synchronization and communication mechanisms. At Stanford University his group developed the Imagine processor, which introduced the concepts of stream processing and partitioned register organizations, the Merrimac supercomputer, which led to GPU computing, and the ELM low-power processor. Bill is a Member of the National Academy of Engineering, a Fellow of the IEEE, a Fellow of the ACM, and a Fellow of the American Academy of Arts and Sciences. He has received the ACM Eckert-Mauchly Award, the IEEE Seymour Cray Award, the ACM Maurice Wilkes award, the IEEE-CS Charles Babbage Award, and the IPSJ FUNAI

Achievement Award. He currently leads projects on computer architecture, network architecture, circuit design, and programming systems. He has published over 250 papers in these areas, holds over 160 issued patents, and is an author of the textbooks, Digital Design: A Systems Approach, Digital Systems Engineering, and Principles and Practices of Interconnection Networks.