David A. Patterson
Pardee Professor of Computer Science, Emeritus
University of California, Berkeley

"A Decade of Machine Learning Accelerators: Lessons Learned and Carbon Footprint"

Wednesday, November 8, 2023

Hosted by Drs. Arash Saifhashemi and Timothy M. Pinkston

Reception 3:00 PM
Hughes Electrical Engineering Courtyard

Lecture 4:00 PM
Hughes Electrical Engineering Building (EEB 132)
or
https://usc.zoom.us/j/98706499300 passcode: 592504

Abstract The success of deep neural networks (DNNs) from Machine Learning (ML) has inspired domain specific architectures (DSAs) for them. Google’s first generation DSA offered 50x improvement over conventional architectures for ML inference in 2015. Google next built the first production DSA supercomputer for the much harder problem of training. Subsequent generations greatly improved performance of both phases. We start with ten lessons learned from such efforts.

The rapid growth of DNNs rightfully raised concerns about their carbon footprint. The second part of the talk identifies the “4Ms” (Model, Machine, Mechanization, Map) that, if optimized, can reduce ML training energy by up to 100x and carbon emissions up to 1000x. By improving the 4Ms, ML held steady at <15% of Google’s total energy use despite it consuming ~75% of its floating point operations. With continuing focus on the 4Ms, we can realize the amazing potential of ML to positively impact many fields in a sustainable way.

Biography David A. Patterson is a UC Berkeley Pardee professor emeritus, a Google distinguished engineer, and the RISC-V International Vice-Chair. His most influential Berkeley projects likely were RISC (Reduced Instruction Set Computer) and RAID (Redundant Array of Inexpensive Disks). His best-known book is Computer Architecture: A Quantitative Approach. He and his co-author John L. Hennessy shared the 2017 ACM A.M Turing Award and the 2022 NAE Charles Stark Draper Prize for Engineering. The Turing Award is often referred to as the “Nobel Prize of Computing” and the Draper Prize is considered a “Nobel Prize of Engineering.”

Questions? Contact Arash Saifhashemi <saifhash@usc.edu>