SLAC COLLOQUIUM SERIES

High Performance Computing: Beyond Moore's Law

Rob Schreiber
Cerebras Systems Inc.

Abstract:
Supercomputer performance now exceeds that of the earliest computers by thirteen orders of magnitude, yet science still needs more than they provide. Prior to 1990, supercomputers were powered by one, or a few, sequential or vector processors. But since 1990, almost all the increased machine performance has been provided by a shift to highly parallel systems. This change, the triumph of parallelism, was made possible by Dennard scaling and Moore's law, which made low-cost commodity hardware the basis for modern supercomputers. But Dennard scaling is over, and Moore's Law is coming to an end.

Demand engenders supply, and ways to prolong the growth in supercomputing performance are at hand or on the horizon. Architectural specialization has returned, after a loss of system diversity in the Moore's law era; it provides a significant boost for computational science. And at the hardware level, the development of a viable wafer-scale compute platform has important ramifications. Other long-term possibilities, notably quantum computing, may eventually play a role.

Wafer-scale integration was tried, and failed, at a well-funded 1990s startup. It has now been brought to customers, successfully, by Cerebras Systems. Why wafer-scale? Real achieved performance in supercomputers (as opposed to the peak speed) is limited by the bandwidth and latency barriers --- memory and communication walls --- that impose delay when off-processor-chip data is needed, and it is needed all the time. By changing the scale of the chip by two orders of magnitude, we can pack a small, powerful, mini-supercomputer on one piece of silicon, and eliminate much of the off-chip traffic for applications that can fit in the available memory. The elimination of most off-chip communication also cuts the power per unit performance, a key parameter when total system power is capped, as it usually is.

I will provide some detail concerning the technical problems concerning yield, packaging, cooling, and delivery of electrical power that had to be solved to make wafer-scale computing viable. These systems are in use at several labs, and I will discuss the impact that they are having for traditional models, neural networks in science, and hybrids of AI and mathematical physics.

April 26, 2021 | Zoom | 3:30 pm - 4:30 pm