June 25th , 2018

Using Cellphone Technology to Build AI Servers

Peter Hsu
EPFL University

Abstract

It is said artificial intelligence is going to change the way we live, work and play in 2018. Certainly the market for AI technology is growing rapidly. Some of us believe excessive energy consumption is holding back even more revolutionary advances in AI software. This talk begins by looking at how energy is consumed in the IBM AC922 server, marketed for enterprise AI computing and used in the world's fastest supercomputer, US DOE Summit. The AC922 is a CPU+GPU Data-Streaming architecture. I propose an alternative architecture based on Near-Memory Computational Model using low-power consumer cellphone technology. In combination with some architecture and 3D SoC/DRAM chip layout/packaging co-design ideas, I suggest it may be possible to improve energy efficiency by an order of magnitude within one process generation. This talk presents work-in-progress I hope to continue during my summer visit to EPFL University.

Using Cellphone Technology
to Build AI Servers
Peter Hsu, PhD
pete2222@mac.com
Presented at Barcelona Supercomputer Center, Spain
25 June 2018

Short bio

Peter Hsu was born in Hong Kong and moved to the United States as a teenager. He received a B.S. degree from the University of Minnesota at Minneapolis in 1979, and the M.S. and Ph.D. degrees from the University of Illinois at Urbana-Champaign in 1983 and 1985, respectively, all in Computer Science. His first job was at IBM T. J. Watson Research Center from 1985-1987, working on code generation techniques for superscalar and out-of-order processors with the 801 compiler team. He then joined one of his former professor at Cydrome, which developed an innovative VLIW computer. In 1988 he moved to Sun Microsystems and tried to build a water-cooled gallium arsenide SPARC processor, but the technology was not sufficiently mature and the effort failed. He joined Silicon Graphics in 1990 and designed the MIPS R8000 TFP microprocessor.

The R8000 was released in 1994 and shipped in the SGI Power Challenge servers and Power Indigo workstations. Fifty of the TOP500.org list of supercomputer systems used R8000 chips in 1994. Peter became a Director of Engineering at SGI, then left in 1997 to co-found his own startup, ArtX, best known for designing the Nintendo GameCube. ArtX was acquired by ATI Technologies in 2000. He left ArtX in 1999 and worked briefly at Toshiba America, where he developed advanced place-and-route methodologies for high frequency microprocessor cores in SoC designs, then became a visiting Industrial Researcher at the University of Wisconsin at Madison in 2001. Throughout the 2000's he consulted for various startups, attended the Art Academy University and the California College of the Arts in San Francisco where he learned to paint oil portraits, attended a Paul Mitchell school where he learned to cut and color hair. In the late 2000's he consulted for Sun Labs, which lead to discussions about the RAPID research project, a power-efficient massively parallel computer for accelerating big data analytics in the Oracle database. He was with Oracle Labs as an Architect from 2011 to 2016. In 2017 Dr. Hsu founded CAVA Computers, Inc. in an unsuccessful attempt to bring hyper-converged storage with integrated compute to market. He will be a visiting professor at EPFL University in Lausanne, Switzerland summer of 2018.