Podcast Name: *ACM ByteCast*
Episode: Wen-Mei Hwu - Episode 58

Welcome to the *ACM ByteCast* podcast, a series from the Association for Computing Machinery! The podcast features conversations with researchers, practitioners, and innovators at the intersection of computing research and practice about their experiences, lessons learned, and visions for the future of computing. In this episode, host Scott Hanselman is joined by Wen-Mei Hwu, a computer scientist and Senior Director of Research and Senior Distinguished Research Scientist at NVIDIA Corporation.

To begin, Dr. Hwu outlines his long academic career, particularly his 34-year tenure at the University of Illinois at Urbana-Champaign, before transitioning to NVIDIA. Then, Dr. Hwu reflects on the beginning of Moore's law during his graduate studies in 1983, describing how the reduction of transistor sizes and the resulting increase in computational power shaped the industry for two decades. He explains the significance of Dennard scaling, which allowed for faster, more efficient processors without increasing chip size or power consumption. During the 1980s, his research group's vision focused on using the growing number of transistors to detect parallelism and run tasks faster. This approach led to significant advancements, such as the creation of Intel's P6 processor.

As transistors reached their physical limits, industry leaders looked for alternative ways to maintain performance, such as stacking chips and improving cooling mechanisms. Dr. Hwu explains that while Moore's law may continue for another generation or two, the industry is already shifting toward building more complex systems with reduced latency. He predicts major advancements in networking, particularly with the shift from electrical to optical communications. The conversation then shifts to the limitations in software, especially when dealing with highly parallel systems. Scott recalls the days of early processors, and draws parallels to today's computing landscape, where machines often contain multiple types of processors, including GPUs and NPUs, each designed for specific tasks. Dr. Hwu agrees that this trend will continue and suggests that a new type of processing unit will emerge to handle the massive amounts of real-time data. New processing units will be necessary to handle the vast data sets generated by real-time applications.

Dr. Hwu elaborates on the challenges of handling large data sets in real time, especially in the context of machine learning. While current systems focus on compressing data into models for inference, he envisions a future where systems can access and process massive data sets directly. The next wave of innovation will likely focus on simplifying access to data, removing outdated structures, and giving users what they really want: their data, not the file system. Before wrapping up, Scott and Dr. Hwu share their predictions for the future of computing, including the advent of devices that allow users to recall names, dates, and detailed information in real time—like a personal assistant whispering details in their ear. These devices would extend human capabilities, making everyone an "infinite memory person" without the need for cloud-based services, revolutionizing personal and professional interactions. Dr. Hwu reveals

his desire to continue contributing to the field until the day he can hold a pocket supercomputer in his hand.

Key takeaways:
2:57 - The evolution of Moore's Law and scaling.
6:18 - Dennard Scaling and the future of computing.
11:09 -The rise of specialized processors and new computational units.
12:48 - Data processing challenges and the future of computing.
17:55 - Predicting the future of computing.
21:01 - Delta Project: Addressing academic GPU shortages.

Links
Learn more about Wen-Mei Hwu.
Learn more about the ACM ByteCast podcast at https://learning.acm.org/bytecast.