Scott Hanselman: This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers, practitioners, and innovators, who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned, and their own visions for the future of computing. I'm your host today, Scott Hanselman.

Hi, I'm Scott Hanselman. This is another episode of Hanselminutes in association with the ACM ByteCast. Today I have the distinct pleasure to speak with Dr. Wen-Mei Hwu from NVIDIA from the University of California Berkeley. You got your PhD in the '80s, and then do you also teach at the University of Illinois at Urbana-Champaign?

Dr. Wen-Mei Hwu: Yes, I taught there for 34 years before [inaudible 00:00:49].

Scott Hanselman: Oh my goodness.

Dr. Wen-Mei Hwu: I graduated and went to NVIDIA.

Scott Hanselman: That's wonderful. So you have a wonderful paper that we'll put a link to in the show notes called, Running on Parallelism And a Few Lessons Learned Along the Way. And I love that it starts on the first slide with a list of years and people. It's so refreshing and wonderful to see the acknowledgement of the history of the space as the first slide. Was that a conscious decision to lead with that?

Dr. Wen-Mei Hwu: Yes. This actually was the set of slides that I used after I received the ACM IEEE Eckert-Mauchly Award. So basically when the committee chair called me, told me that I would be the recipient, it took me a little bit of time to accept it. If you look at the list of people that I had, these are the people who literally were building the historic machines and publishing extremely influential work in the history. Many of them are the authors of the textbook that I read when I was grad student, so I wanted to put that slide up in the context before I talk about any thoughts after receiving that award that these are the people that I was standing on their shoulder in my work.

Scott Hanselman: Yeah. That's so wonderful because it's such an amazing group and to be acknowledged in the same list and the same breath going back so many years must feel really amazing. The discussion is about parallelism and where it fits into the historical context of things, and if we think about Moore's law, throwing transistors at the problem, and maybe starting to flatten out, we're seeing computing in general start to move towards not just parallelism but massive parallelism. Is that because we're starting to hit some physical limitations, some quantum limitations? We can only throw so much power at these things. We have to start to go wide.

Dr. Wen-Mei Hwu: I actually kind of lived through the whole cycle in my professional career. When I started as a student in 1983, it was really what I would call the real beginning of

the Moore's law. Later on, we also understood scaling. These two things are incredibly important for the next 20 years. So the Moore's law basically drives the industry to reduce the transistor size, you can argue 18 months or two years. Basically, that allows people to pack in twice the amount of transistors per chip without increasing the size of the chip. That's economically very good, but for my career, Dennard scaling was even more important. Dennard scaling, as you reduce the transistor size, you will reduce the voltage by the square root. As a result, you can reduce the power consumption per chip area by one over square root. You can improve the speed of the transistor or reduce the switching time by one over square root of two and you can still have more transistors.

So that is incredibly important because for the next 15 years, people literally say that they can write a piece of code, go to sleep for five years, and wake up. That piece of code will run much faster without doing anything to it. This is great because during my grad school days, I went to University of California Berkeley, and that was a time when we really debated what the future course of microprocessors will be. That's when Hennessy and Patterson were advocating RISC. There were huge number of discussions, but the research group that my advisor, Yale Patt, and some of my fellow grad students, look sort of beyond the immediate next five years and look to the 10, 15 years and say, "Well, if we have so many transistors and these transistors are going to be running so fast, how about if we use these transistors to detect the parallelism that nobody can currently detect in their processors and use the parallelism to run things much faster?"

And we got lucky. We got lucky that by 1993 when Intel designed the first P6, the Moore's law and Dennard scaling got to the point where they can design such a processor based on some of the intellectual research we did in the 1980s and many other people did after us so that they can build the P6, which is the Pentium Pro processor. And that processor embodied all the parallelism mechanisms that we envisioned. That was history. One of the interesting lessons we learned is that when we do research, we really need to understand where the technology is going to be 10 years from now, because by the time people put any of your research ideas into this hardware, it will be at least five years, if not 10 years.

Scott Hanselman: So for folks that are listening who may not be familiar with some of these terms, I want to call out some things that you said that folks can go and read more about. So Dennard scaling, also known as MOSFET scaling. That was 1974 when that was being thought about. And that held, as you pointed out, for 25 plus years until it started to falter in the early 2000s there. And then with modern GPUs, you're looking for ways to get around it, stacking chips, coming up with new ways to cool because you may be getting higher performance without power consumption, but you're still getting power leakage. You're getting heat dissipation. You're pushing up against the laws of physics itself. You say though,

you're thinking 10 years out, is there a limit? Is there a possible theoretical limit where you will actually say, "Nope, that is as far as this is going to go?"

Dr. Wen-Mei Hwu: I think there is, but we don't know exactly when this is going to hit. This is like a stock that keeps going up. Everyone says it is going to go down at some point, but nobody knows when. We have been saying that Dennard scaling will be ending, which ended. So Dennard scaling ended in early 2000, as you mentioned. For Moore's law, we can probably see at least another two generations, but we are already at the point where we're building computers very differently than scaling on the chip.

Scott, you already mentioned that we try to skip the stack memory with chip packaging, but I think most of the innovations today, if you look at the big innovations for today's research, we are already into building very big systems with reasonable latency. So we're going to see big changes in networking, especially going from electrical to optical, and we're going to see the very different system communication primitives that allow us to synchronize a huge number of machines fast. And these are all because we are already seeing that coming. We don't know exactly when this to come, but we need to run as we can so that we're ready when they come. And chances are we may not even be totally ready when it comes.

Scott Hanselman: Yeah. This might be an interesting question or a dumb question, but we hear about people being full stack engineers and you're probably not thinking about Node or Erlang or JavaScript. At the highest level when someone is a software engineer, for them, low level might be going into Google Chrome and hitting F12. But for you, you're starting to bump up against protons and electrons moving around in space, but the software stack overhead has us wondering, even as I sit here on a highly parallel machine, why does my machine feel slow? Are you thinking about software paradigms and how people are going to program in five or 10 years and applying that to how you're going to create these pieces, these chips in 10 years?

Dr. Wen-Mei Hwu: Absolutely. There are two important forces in these kind of things. One is reality. There are huge number of people programming with Python. When we deal with frameworks based on Python such as PyTorch, we see all kinds of interesting, not just overhead, but bottlenecks. For example, the Python Interpreter Lock, that fundamentally restricts the kind of parallelism we can expect at the source language level. And more subtly, all the GPU activities that are launched through the Python-based frameworks are based on APIs. And what that means is that these API calls are separated from each other. They're not adjacent to each other in terms of the compilation process or interpretation process. We're seeing these individual activities.

So one of the important things that we will begin to see is how the lower level software and hardware combination will be able to take these API calls and maybe doing things like fusing these activities on the fly and optimizing away

some of the inefficiencies because these are isolated events, but these are the kind of things that people have been working on for decades starting from the Java days. Microsoft has the runtime compilation activities for a long, long time, all the runtime, linker optimizers, but I think we will be getting to the point where some of these technologies will begin to really be deployed in the next decade or so because we are running out of room for further optimizations without deploying these kind of techniques.

Scott Hanselman: Yeah. Seems like, if I recall, when I was getting my first 386, there was the SX and the DX, and then we had ones with a coprocessor. And companies would say, "We need to figure out something for that coprocessor to do. Maybe people will buy it if they use Excel." So then we would buy this computer because I want to use Excel. When I wanted to play Quake or Doom, I would buy a 3dfx card, because I wanted that.

When I want to play a great game now, I'll buy an NVIDIA processor. I'm on a machine right now with a 4080 Super. I'm thinking to myself, "What is that really good at that my Intel chip is not very good at?" And if I look to my right here, I have a Copilot PC with an NPU from a competitor, from Qualcomm, so now suddenly it's not about a CPU and a coprocessor. I've got three very competent different kinds of processing units here. Do you think there'll only be three? Will I come into a world where there's five or six or 10 different fundamentally interesting processor units that do different tasks?

Dr. Wen-Mei Hwu: I think that's the billion-dollar question. I think personally, there will be at least one more. Scott, you hit the nail on the head. People don't want processors. People want applications. People buy processors reluctantly because they want to run their applications well. Today, I would say that the people buy GPUs because they want to train models. These models have all these training needs. People buy these GPUs reluctantly. I'll turn your question just a little bit and maybe say that I think there will be another kind of processing unit because at some point people will want to be able to deal with massive amount of data on the first-hand basis. What I mean is that we are going through the machine learning era where we reduce all that data, the training data into these models. Then we converted the query process into a computation inference through this model, and we get the information, but we kind of bypass the need of finding data in real time.

Google has been doing search and Microsoft has been doing search, and this search are real-time searches. If you use Google, all these things are pre-indexed data that we hope to serve the user real-time. One thing that is very interesting is that we kind of swing the pendulum so much in the deep learning models and large language models, we should begin to think about what happens? These models have their limitations. What if people want to be able to go straight to the data, but use some of the model capabilities to be able to find the data much better than the previous generation's search? So I can imagine in order to do this, we will need to have a unit that is going to be able

to tolerate latency and introduce very efficient ways to grab, let's say one kilobyte of data out of four petabytes of data, and serve it to the user in real time and go through some of the models to give user very digestible.

Scott Hanselman: Solving the needle in a haystack problem at scale, because that traditionally, you mentioned this in the paper, in the presentation, this memory storage divide. It feels like the majority of the work that my computer is doing as a user is simply moving data from one format to another. It's just transformation, and that's becoming quite tedious as I simply want to work or I want to enjoy my computer.

Dr. Wen-Mei Hwu: Absolutely. Right now, we have all these file system software stacks and so on that we built over the years that are standing in the way, and the format that you are talking about are imposed partly by these kind of software stack. What people really want is their data. People don't want the file system. People don't want the formats. How do we get people what they want? And I think that's going to be the next unit.

Scott Hanselman: ACM ByteCast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher, and TuneIn. If you're enjoying this episode, please do subscribe and leave us a review on your favorite platform.

And it seems like we're narrowing the space between our work and the data. We have that storage throughput, we have that latency that is being applied. We're setting up little islands where your parallelism and your data that it's running on are near each other so as to avoid that latency, which is now down to picoseconds, it's down to nothing. There's little latency as possible.

Dr. Wen-Mei Hwu: Absolutely.

Scott Hanselman: Yeah. I've seen you mentioned millisecond-level latency and terabytes a second throughput. It seems like a dream, but it's going to happen at some point.

Dr. Wen-Mei Hwu: I really believe that if we do the work right, when someone does this interview 10 years from now, the person may have a machine on the desktop even. I can see how we can build these kind of things in the data center at the very high cost, but the real test is that can we put this kind of capability into every person's desktop or even a small device that the person carries that has the entire lifetime's memory and information that you can retrieve real time when you talk to someone?

Scott Hanselman: It's interesting that you mentioned that because just literally yesterday, a friend of mine sent me a package, and for those of you who are listening to this podcast, you won't see this. I'm holding up a Compaq PC Companion and some of these early devices, the PalmPilot. We've always wanted our life, our family, our friends, our photos, in a tiny device like this. My first applications were on the PalmPilot. This is a pocket supercomputer with its own GPU, and now we're

going to see AI start to move its way into our pockets. The idea that what once filled a room is carried around with me with all-day battery life. To your point, the data won't be in the cloud. It could just be in my pocket.

Dr. Wen-Mei Hwu: Exactly. I can see how that can happen. I work with a lot of the SSD storage vendors on a regular basis. I work with the PCB builders, the mobile chip builders. I can see if we do the right things within a decade, we can begin to have these devices that truly extend us. I really feel that the real benefit for these devices is to make each of us a infinite memory person and then can complement some of the, let's say just like calculators complemented our way to do arithmetic, we should be able to remember the names of all our associates and colleagues, and so in real time, no problem. Right now, I can't.

Scott Hanselman: I was looking at some pictures of a former US president who was being introduced to some dignitaries, and you could tell that someone was whispering, "That's the president of France, and this is his wife, and their children's names are..." That's what I would like for my coworkers, for my colleagues, for people I haven't seen in 10 years is someone to whisper in my ear. But right now to do that, I have to go to the cloud, interact with a vendor. That's sparse access to massive data.

Dr. Wen-Mei Hwu: Yeah.

Scott Hanselman: And who knows what that cost in power consumption, as I go to Google photos today and I say, "Show me pictures of my child in the snow when they were two," you're saying that algorithms, memory capacity, the theoretical limits will happen such that we could have that in our pocket.

Dr. Wen-Mei Hwu: Not only that, you don't need to scroll through a hundred results and pick the best one.

Scott Hanselman: That's amazing. That's what I want. But that's going to require silicon changes, it's going to require algorithmic changes. That's beyond a generational shift, that's going to get the entire industry to head in a different direction.

Dr. Wen-Mei Hwu: Yes and no. We have been witnessing these kind of things. If there's enough value, these things move fast. For example, I have seen the whole GPU movement where I still remember in 2011 when we proposed to build a supercomputer called Blue Waters based on GPUs to NSF, there was a review panel. These review panelists were all supercomputing center directors who had decades of experience building supercomputers and serving communities with supercomputers. And after my presentation, one of the panelists said, "This proposal is risky to say the least and probably irresponsible if you really look at it." I said, "Why do you say this?" He said, "Compare the number of math library functions that GPUs have versus the Intel math kernel library. And you are ways away from that."

And I said, "Yes, but we know that there are only a limited number of science applications that people truly care about using supercomputers. We know that this is going to start with more special purpose, but if you look at the computational speed over wattage, this thing is a small puppy with big paws, and let's take that step." And somehow by miracle, that panel agreed, allowed us to do it. 13 years later, we're sitting here looking at all the math libraries that people already built, and if you look at the model training process, people put together multiple generations of these kind of frameworks over the past five years. If it provides enough value, if there's enough incentive, it will happen.

Scott Hanselman: That's amazing. There's the people who have the audacity to do it once, to propose a petascale supercomputer, to get the National Science Foundation to think it's a good idea to give you $200 million to do it, and now it's being done by others. That's unbelievable.

Dr. Wen-Mei Hwu: Thank you.

Scott Hanselman: And then now, is it the Delta Project that is the next step of Blue Waters?

Dr. Wen-Mei Hwu: The Delta Project, it is the next step of Blue Waters in the sense that there's a real academic need for more of these GPUs, and there's a serious shortage of GPUs available to academic researchers today. And so Delta is meant to fill that gap. But Delta is not like Blue Waters, where we know there so many known risks. We know there are so many potential failures that can become scandals for the public. We had to work with all the science teams. There are four science teams that we're eternally grateful how much work they put into their applications to move them into GPUs and prove the benefits. NVIDIA dedicated hundreds of people working on this project to make sure that it does not fail. So Delta is not at that kind of level. Delta is really about how we can enable academic researchers who really need the GPU time to be able to make some real progress in them.

Scott Hanselman: I see. So it's about access to those resources more than it is about pushing the boundaries of computational power?

Dr. Wen-Mei Hwu: Exactly.

Scott Hanselman: Okay. What projects are you excited about where folks can go and learn about supercomputing projects and audacious moonshot-type things like that are going to take us into the exaflops?

Dr. Wen-Mei Hwu: Yeah. There are several exascale projects that come in together, and I would say most of them have very interesting ideas, but I think one of the biggest challenges that all these supercomputers are trying to figure out is this whole networking, because we do have the option to build optical networks. These optical networks are so expensive. They're more than 10 times more expensive

than their copper equivalents. That's why whenever the distances are short, optical connections are not used.

If we look at the next generation, we're going to see, I would say big movements in terms of the network connection and in terms of the cooling. We're definitely at the cooling limit for all the machines. I still remember that when we're the amount of water that we need to pump through one of these data centers, and I said, "Can we really have that much water?" So I think some of the exascale machines would have very interesting cooling mechanisms, and these things will be published as soon as these machines are operational.

Scott Hanselman: I'm trying to think about what I'm going to do as I think about retirement, as I think about what are the next 50 years going to be? Are you going to be doing this until you can't walk anymore? Are you going to be a parallelism person and a computing pioneer until the end? Because you seem like you're having a lot of fun.

Dr. Wen-Mei Hwu: Exactly. And, Scott, let me tell you, when I think I'm going to retire. I will retire when I have that device in my hand and I can go around and show people, pretend to be the most intelligent person on earth, spend my time explaining to people how they can use this device.

Scott Hanselman: That's wonderful. I hope we both get to see that happen because that is the promise of a pocket supercomputer that really makes your life better, makes your relationships better. It is about the things you can do with it. It's about the people and the connections.

Dr. Wen-Mei Hwu: Absolutely. I really hope to live to see that day.

Scott Hanselman: Fantastic. Thank you so much, sir, for chatting with me today.

Dr. Wen-Mei Hwu: Thank you, Scott, for this opportunity.

Scott Hanselman: I have been chatting with Dr. Wen-Mei Hwu. He is the senior director of research and distinguished research scientist at NVIDIA Corporation. And many congratulations as well, sir, on your award, your ACM-IEEE Award for your pioneering contributions in the design of these processor architectures. We do appreciate you and all of the work that you and your colleagues have done to move us forward in computing.

Dr. Wen-Mei Hwu: Thank you. Thank you so much.

Scott Hanselman: This has been another episode of Hanselminutes in association with the ACM ByteCast, and we'll see you again next week.

ACM ByteCast is a production of the association for Computing Machinery's Practitioner Board. To learn more about ACM and its activities, visit acm.org. For

more information about this and other episodes, please do visit our website at learning.acm.org/bytecast. That's B-Y-T-E-C-A-S-T. Learning.acm.org/bytecast.