Bruke Kifle: 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 am your host, Bruke Kifle.

Today's episode will take a closer look at the exciting and ever evolving field of semiconductor technology and its impact on our daily lives. From smartphones to laptops, from electric cars to smart homes, semiconductor technology is at the heart of the devices that power our world. And advancements in the field have continued to the creation of smaller, faster, and more efficient electronic devices. Semiconductors are truly the building blocks of modern technology and they're really shaping the way we live, the way we work, and the way we interact with the world around us. Today we have the honor of speaking with one of the foremost experts in the field, Dr. Philip Wong.

Dr. Wong is a renowned professor of electrical engineering at Stanford University and the chief scientist of the Taiwan Semiconductor Manufacturing Company, also known as TSMC, the world's largest semiconductor foundry. Prior to Stanford, Dr. Wong was with IBM Research for 16 years. Shortly after, from 2018 to 2020, he was on leave from Stanford and served as a vice president of corporate research at TSMC, and since 2020 remains the chief scientist. He is a fellow of the IEEE and has received numerous awards for his research contributions to solid devices and technology. He's the founding faculty co-director of the Stanford SystemX Alliance, an industrial affiliate program focused on building systems, and the faculty director of the Stanford Non-Volatile Memory Technology Research Initiative. And finally, the faculty director of the Stanford Nano-fabrication Facility, a shared facility for device fabrication on the Stanford campus that serves academic, industrial, and governmental researchers across the U.S. and around the globe. Dr. Philip Wong, welcome to ByteCast.

Dr. Philip Wong: Thank you very much for the introduction and the invitation to speak with you.

Bruke Kifle: Yeah, we're very excited to have you here. I want to start off with a pretty open-ended question that I like to ask most people. You know, you have such a remarkable and a very interesting career that spans both academia, research, industry, still have deep engagements in industry. Describe some of the key points in your personal and professional career and background, that have ultimately led you into the field of computing and motivated you to pursue your field of study today.

Dr. Philip Wong: Yeah, that's a great question. Well, I came into this field kind of... It wasn't really planned. I am interested in the physical sciences, physics and electromagnetics and things like that. And so during my undergraduate years, I got interested and started to take physics and started to take electronics. But I don't want to just do a deep kind of ivory tower physics type of things. I wanted to make something that is of practical interest, so I went to electrical engineering. And in that particular area at that time, it was the beginnings of what is known now as semiconductors or microelectronics now. It's piqued my interest because it is kind of a cross between solid-state physics and the practical application of those solid-state physics, because as you mentioned earlier in this podcast, semiconductors is the heart of everything that we do. More so now than before, but even back in maybe like 20, 30 years ago, they were already indications that many of the electronic products is going to be further improved and enabled by advances in semiconductors. So that's how I got into this field.

And I was very lucky because it wasn't expected some 30, 40 years ago that semiconductors would make such a big impact in society, but it turns out to be the case, so I was really lucky.

Bruke Kifle: Oh, that's very remarkable. Were there any aspects of your personal upbringing or your personal background that motivated some of your scientific interests?

Dr. Philip Wong: Well, in semiconductor I work on device fabrication and device physics. And a lot of the device physics and also device fabrication involves materials and chemistry. And I was interested in chemistry and materials and that kind of fits pretty well with this rather interdisciplinary field. And as I move forward in my career, the fundamental materials in chemistry led to advances in devices, and advances in devices leads to new circuits and new systems and that would be a broader impact. So throughout my career, I started from really more basic physics type things and we gradually move up in terms of what the engineering people call the hierarchy abstractions, moving further up into the abstractions.

Bruke Kifle: I see. Yeah, I think there's something very interesting to be said about the cross disciplinary nature of your work, but we'll get into that shortly. One thing that I do want to touch on is the semiconductor industry relies heavily on nanotechnology to continue shrinking the size of transistors and other components on computer chips. And that's ultimately at the heart of what's leading to faster and more efficient devices. But I think many people may not actually appreciate how difficult, but also how remarkable it is to deal with materials and structures on that scale. We're talking, if I remember correctly, this is one-billionth of a meter. And I'm sure this exposes many very unique properties and functionalities that maybe you are not seeing at bulk materials. So what are some of the key challenges, but also the opportunities in scaling down electronic devices to the nanometer scale? And how do you approach them from both a scientific point of view, like you said, but also an engineering perspective?

Dr. Philip Wong: You point out very interesting aspects, which is devices get very, very small right now. Small, we are at the atomic scale, a nanometer and atomic scale right now. And a nanometer is a billionth of a meter, so that's really, really small. And if you cut up computer chips today and look under a very powerful microscope, you can see the individual atoms. And a typical transistor today, you can actually count the number of atoms that you have in the transistor system. That's really amazing. At that particular name scale, the interesting physics will come about and the physics that are in operation for larger bulk, macroscopic scale materials and devices will change as you go down to this small scale. And that gives rise to a lot of interesting things both in the physics world and also in the practical technology world. In the physics world, some of the people who may be interested in physics knows that for the better part of the several centuries are probably people interested in high energy physics, that look into the deep physics of how the atoms behaves and how the electrons behaves and so on.

And that oftentimes involve very high energy and building up huge size of accelerators to hit particles and see how they behave when you hit them with high energy. And that gives you insights into the basic physics. But at the nanometer scale, many of the physics are really beautiful. And so in recent years, even in the physics world, a lot of the interesting physics shows up in what we call solid state physics, namely nanometer scale physics. The physics that exemplary exhibits their behavior as these very small nanometer scale devices. So for example you look at recent Nobel Prizes for example, many people got Nobel Prizes because of the study into these kinds of nano scale phenomenon. So that's kind of interesting from a fundamental physics point of view. But beyond that, in the practical application many of these nano scale physics are actually used today, every day in the electronic devices that we have. I suppose many of your audience have a cell phone or use a computer, and those cell phones and computers today typically have data storage devices called flash memory. And those flash memory operates on quantum mechanics.

And those things happen only when you are at a nanometer scale. So those deep physics do have many practical applications. And the transistors that we have today in the phones and the computers are of such nanometer scale, that we cannot possibly understand how they work unless we invoke these nanometer scale physics that is in operation.

Bruke Kifle: So lots of interesting scientific implications, but also very real practical applications to our day to day use cases. I'd be remiss if in this conversation about transistors and integrated circuits and semiconductors if I didn't bring up Moore's law, right?

Dr. Philip Wong: Yes.

Bruke Kifle: It's this observation for those who don't know that the number of transistors on a dense integrated circuit doubles every year. And it's kind of been this very interesting self-fulfilling prophecy, but quite honestly the driving force behind a lot of the rapid advancements that we're seeing in the industry for almost half a century. So one thing to call out though is as the complexity of the technology continues to increase and we see limitations in actually scaling down the size of transistors, there's actually growing concern that the end of Moore's law might be near. And at the end of the day it's not a law of physics, it's just a relationship that quite honestly may not hold forever. So in your opinion, what do you see as the next technological breakthrough that will actually drive the industry forward and hopefully prevent the end of Moore's Law?

Dr. Philip Wong: Yes, that's really a question that I've always been asked and I'm glad that I have an opportunity to kind of talk about it. There are two things I wanted to bring up. One is the Moore's law actually is a very interesting phenomenon, particularly in the semiconductor industry. The industry are able to make predictions about the future in the transistors every so year, two years and so on. And this is very unique across different industries. And there are no other industries that has these kinds of long term predictions that holds true for decades and decades. And if you look into other industries like automobiles or airplanes or aircraft and things like that, other industries don't have these kinds of predictable advancement of technology and that is very unique to the [inaudible 00:11:39] of the industry.

And as a result of that, it really propelled the entire industry toward a very rapid pace of innovation because then everybody both up and down what people call the value chain, from the material suppliers to equipment manufacturers, to people who actually design and make the chip to the users of those chips, they all have kind of a roadmap of what will happen in the future. And so therefore they could make plans for the future very accurately. And so these kinds of activities, this kind of situation need the whole industry to advance not only at regular pace but also at very rapidly because we all know what our competitors are doing. And therefore if you are a competitive company or a competitive researcher, you will try to outdo everybody else. And because now everybody knows what the general direction is, then everybody wants to try to outdo that. And therefore it leads to a very rapid evolution of the industry. And if you look at transistor miniaturization, which is one of the main driving force behind this Moore's law, doubling the transistors every so often.

If you look at main transistor miniaturization, then everybody knows what to do. And therefore we have a very well-defined path going forward for the last five decades or so. And as a result of that, the ways to go forward is clear. And everybody worked towards a common goal and the industry moved forward very fast. Now we are kind of at the end of this, so it's kind of like if I would draw an analogy it would be walking inside a tunnel. There's no other way you can go, you just go forward. And that makes you single-minded and therefore very easy to go forward because you don't have to think about something else. And the way to do its is to shrink in two dimensions. But of course as mentioned earlier, shrinking in two dimensions do have some limits. We have [inaudible 00:14:00] scale and you cut it, as I mentioned before, if you cut up a transistor you encounter a number of atoms. And if you shrink further, you cannot have half an atom.

So you can naturally see there is a natural limit in there. So this tunnel that we have been walking inside is coming to an end. Now, you can think of this as two ways, cup half full or cup half empty. You can think of, oh, where is the end of tunnel with that. That's the typical kind of reaction. But if you're at the end of tunnel, that means you are going out of the tunnel. And there can be many, many possible paths going forward other than two-dimensional miniaturization. And that is really exciting for researchers and people who want to get into the field is that there are plenty of paths going forward. We don't know which one will work, but there are many, many possible paths, unlike in the past there's only one path forward. And that's where the excitement is. If you're a engineer or researcher, you would rather have a lot of options to figure things out rather than having one thing to do.

So I think this is really exciting. We're at a time right now at the cusp of a new major resolution that can have future electronic systems that are way better than what we had before. Even though of course I should say that the path going forward is unclear there are many paths forward, the opportunities are exciting.

Bruke Kifle: So certainly as you reach the saturation point, there's definitely a lot of uncertainty. But as you said, it seems like there are many new exciting open research directions that can unlock or bring a lot of value to the field. So I think it's going to be quite exciting to see how some of the current research that your group and labs across are working on will help drive the future.

Dr. Philip Wong: Absolutely. Yeah, the optimism is really based on the demand signal that we see from society. And when you say, okay, something is saturating, okay, then you know well, there's nothing more to do with it, it's saturated already. Why are you still working on it? But then the demand signal is pretty high. If you look at society demand, many of the things that we want to do from self-driving car to higher energy efficiency or AI systems and so on, really depends on continued advancement of technology. I say continued advancements, not just what we have today but you need to have continued advancement in order to fulfill our expectation of what electronic systems would do or computing systems in general would do. So the demand signal is high, and therefore when there's demand there must be innovation.

Bruke Kifle: Yes, certainly. That's very interesting. And I'm sure as part of this continued research innovation, there will be a need to draw inspiration from different fields. You talked about earlier how your personal interests have been motivated or influenced by your interest in chemistry or in biology. So within this sort of field of nanotechnology, how do you draw inspiration or insights from biological systems, for example? Or are there other fields, whether it be material science or physics or processes in those fields that guide your research on that nanoscale devices and systems?

Dr. Philip Wong: Yeah. The very insightful observation that these days many of the advances occur at the joining two or three different fields together, and capitalizing on the good properties or the advantages of different disciplines and putting them together. One of the things that I think would be exciting to do going forward is to be able to do really energy efficient computing for everything we do that involves computing. And now where is the optimism coming from? Today, if you look at computing at a data center level or running an AI training model type applications, you're talking about megawatts of power required to power up a computer. They would do these kinds of computation. But you and I know that human brain is way more energy efficient and the human brain operates on about 20 watts. So there is a million times difference in energy efficiency between what we human being do every day, and what our human designed computer can do today.

So there's a room for a million times of improvement, and that's tremendous vast space for improvement. Now, how do we get there? We don't know and that's where the exciting things are. And some people are thinking that maybe we can draw on some inspirations around how we understand the brain works and then say maybe we can design computing systems based on those principles. So this is a very active field of research in which I myself have been working on with a number of collaborators both at Stanford and also outside of Stanford. But the interesting thing about this is that our understanding about how the brain works is very little. It's more or less like you want to take apart a computer chip and look at a chip and see how the chip works. And we know that it's almost impossible.

So right now we're doing exactly the same thing with the brain. We look at the brain and see how it works and try to figure out how it is wired and what different units are doing and so on, like searching in the dark. This is the very long to go. And we are the point where we may capitalize on, we make use of our understanding about neurosciences and draw some inspirations on. Maybe we could design electronic systems and take some inspirations on how it works, and get the energy efficient computing out of that.

Bruke Kifle: I see. So I think you're referring to neuromorphic computing, correct?

Dr. Philip Wong: Absolutely. Yeah, that's what the [inaudible 00:20:20] means, neuromorphic computing.

Bruke Kifle: Yeah, I see. So beyond the energy efficiency, faster, more power efficient for existing computing tasks, are there other implications or application areas whether it be, I don't know, in medicine or in autonomous systems? Are there other use cases of this idea of neuromorphic computing?

Dr. Philip Wong: Oh, yeah. Well, in addition to neuromorphic computing, the way we understand the brain would help us design better in computing systems. But also the way we fabricate our electronic systems and the way we understand how to communicate will also help understanding about biology. Many of our colleagues are working on bio-electronic systems or for example, human brain interface, machine brain interfaces. And I have a pet project going on with putting a chip inside a living cell. And chips can be very small now, and you can make very small chips and you can actually build bunch of electronic circuits so small that you can fit inside the cell. But biology can help design better electronic systems, and also the way we understand about how electronic systems work and how electronic systems are being built can also help understand biology.

Bruke Kifle: So that's remarkable. You just said we're able to have chips in a cell, that's a pretty significant milestone. What does this mean in terms of therapeutics or medicine or personalized care? This seems like actually a pretty significant achievement, I would say.

Dr. Philip Wong: Yeah, I'm not saying that we have accomplished that yet, we're moving towards that goal. And that goal is really... Because if you think about altering or monitoring cell physiology today, the thing that we can do is to make a different cell for example through biological means, or put in ways to put chemicals inside the cell that will affect the cell physiology. But you could also imagine that you can use electrical means to change the way the cell behaves. So in addition to making a new cell which is a difficult thing to me, or putting chemicals in there which is different ways, so one of the things that we think would be interesting to have is to be able to use electrical means to alter the physiology of the cell. And then use that capability to study how the cell works.

Bruke Kifle: I see. Very interesting. And when you pursue these lines of research, are you usually motivated by the potential practical applications and work backwards, or is it the scientific or research exploration eventually leads to practical applications, whether it be in healthcare or medicine or computing? What sort of drives the direction of research?

Dr. Philip Wong: Well, I guess it's a combination of both. Oftentimes when you work on something kind of totally new, there is no application yet. It's totally new because we haven't seen anything like that. And so that part would be driven by, hey, we can do this. And if we can do this, that would be kind of revolutionizing the way we see things or understand things or have a tool that could help us understand things that we were not able to understand before. And so that would be more kind of discovery type of investigation. But at the same time, these kinds of investigations... And at least because I'm an electrical engineer, only go so far because at some point you need to find an application, you need to find a use case for it to set the direction of your research because if you're just discovering, there are so many things you can discover. You can get really get lost, it's really that walking into a forest and you don't know where to go.

But if they put a target application in mind would help drive the direction of the research. And that would actually move the research forward better and faster. So you can go on for a little while with the curiosity based type of investigation. But in my opinion, eventually sooner than later you need to find an application as a driver for your research direction.

Bruke Kifle: I see. So in line with that, I know one area of work that is an important research focus for you is non-volatile memory. Maybe can you start off by describing... And it is an area of work that does have many practical applications like you alluded to earlier with smartphones or laptops. But maybe can you describe what exactly is non-volatile memory and how do you envision the use of a non-volatile memory in future computing systems?

Dr. Philip Wong: Great question. First of all for the audience, non-volatile memory, well, I'll decompose it. Volatile means it disappears, non-volatile means it doesn't disappear and memory means to store some information such as I remember what I did yesterday, to store some information. So non-volatile memory are electronic devices that stores information. Some of them store information at a shorter timescale like in less than a second or so, some store information much longer like in 10 years. So the information we want to store on our computers and phones, you want it to stay for a long time. But for example, the keystroke that I type just a second ago, I need to remember it for a second but after a second I don't need it anymore.

So there's a variety of non-volatile memory based on what we need to do. And in fact, the most numerous electronic devices that human made are these non-volatile memory because we have a lot of these non-volatile memory. Like a modern cell phone would have hundreds of gigabytes of non-volatile memory. That means hundreds of 10 to the ninth bytes times eight of these kinds of devices on your phone. There's a lot of them. So that's a very important part of the general information and communications technology that goes to kind of the electronic device. Now we're already using this non-volatile memory, and now going forward what we use this non-volatile memory for is improving the energy efficiency of computing. Computing requires you to do computing on data. Where do the data come from? Data stored in some kind in somewhere. And those data are often stored in memories. And so currently much of a lot of the memory resides on a separate computer chip than the chip that does the computing. And the act of moving the data from one chip to another chip consumes not only energy and power but also incurs time.

It takes time to move from one place to another. So this is the fact that memory chip is on a separate physical location than the computing chip, that is the situation by and large today, results in a lot of waste also in energy and also in time and reducing speed. So the research work going forward, a lot of the researchers in the field we are working on right now is how to put this memory right on top or right next to the computing chip, right on top of it. So what we are working on is building three-dimensional chips. All chips are two-dimensional right now similar to all the houses in Los Angeles. They're all urban sprawl spread out over miles. And the way we get more and more of these houses is to strengthen two-dimensional miniaturization, building smaller and smaller houses. At some point people don't want to live in smaller houses anymore. And what do you do?

You go to Manhattan and build things on top of each other and therefore you gain base to do things. So going from Los Angeles to Manhattan, this is what we are doing right now for the computer chips, try to build computer chips that are three-dimensional with multiple layers of computing devices and memory devices on top of each other. And that is the bulk of my research right now.

Bruke Kifle: I really love the LA to Manhattan analogy. I think that really captures the line of work in a very easily understandable way. But yeah, I think it's very exciting to see the efficiency that we'll be seen both from a performance point of view in terms of energy, in terms of time. So it certainly seems like a very exciting application for the future of computing systems. ACM ByteCast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

Looking forward, I kind of want to pivot now and talk a bit more about some of your work bridging industry and academia. One of the roles that you hold is as the faculty director of the nano fabrication facility at Stanford. And what makes this very interesting is that it's a shared facility that's serving government, industry, academia researchers globally. It seems like a very difficult undertaking. So how do you manage the demand and access to this state-of-the-art facility? And outside of your responsibilities as a researcher, as an academic, what are some of the best practices and lessons you've learned from running such a complex operation?

Dr. Philip Wong: Yeah, that's a great question. First of all, the most important ingredient is to have very highly skilled and capable staff that runs the facility. And I really appreciate that. And of course, running such a facility requires a lot of resources in terms of money and space and everything else. And the strong support from funding agencies as well as the universities are clearly instrumental in these. Our facilities are supported in parts by the National Science Foundation. And of course our universities have also invested heavily into our facilities for both capital investment as well as operational expenses and so on. So those are really kind of necessary conditions, but clearly necessary but not sufficient conditions. And one of the key things about nano fabrication or semiconductor manufacturing or fabrication, is that many of these tools are rather complex and expensive. And so it is very difficult for individual faculty or researchers to acquire enough of these tools because you need not just one, you need a collection of these things to compose a process.

Just like in the kitchen, you need an oven, you need a chopping board, you need a lot of things. So you can't just operate with one tool. So in order to have these complete set of tools, then you necessarily have to come together and share the use of those tools to number one, amortize the cost of those tools. And number two, probably more important than the cost, which a lot of people don't appreciate, the second part is that when people come in and use these shared facilities, they necessarily reside in the same facilities. They talk to each other, and that's where the innovation comes in because I have seen many, many instances in which my students would come into nano fabrication facility, do their work. They meet other students, other researchers, postdocs and other industry researchers. They talk about things when they meet each other and new ideas come about.

And that is clearly important so that a shared facility such as a nano fabrication application facility at Stanford, is not just a collection of tools. But rather is a complete community and ecosystem in which researchers would not only share their information, share their knowledge about the fabrication techniques and processes. But also is really a fertile ground for innovations for new ideas. And I would say, many, many papers have come about from students meeting each other in the nano fabrication application facilities and say, "Hey, why don't we do this together? This sounds fun." There are many instances like this.

Bruke Kifle: Yeah. I think we've seen many great successes of these kinds of collaborations across academia, industry. So all that to say, I think you're definitely advocating for the importance of collaborating between all different stakeholders in the hope of fostering innovation, but also advancing R and D and computing.

Dr. Philip Wong: Absolutely. And just creating this environment and ecosystem for innovation to occur. Apart from the cost amortization, I would say this is even more important than the cost amortization because money you can always get, but collaboration is hard to combine.

Bruke Kifle: Certainly. I think that's a great quote. I think you wear multiple hats, which I think is very remarkable. You on one end have a role in academia as a professor, as a researcher, as an advisor, but in addition to serving as faculty director for the nano fab lab or the SystemX Alliance, you also serve in your capacity as chief scientist at TSMC. So you have a deep engagement within industry as well. So how do you find your role in industry informing some of your research directions or your teaching in academia and vice versa? Do you find some of your engagements and learnings in academia as a researcher, as a professor, informing some of your roles in industry?

Dr. Philip Wong: Especially in the engineering field which I am in electrical engineering, because engineering is about practical applications of stuff, of technology. Like coming up with new technologies and understanding about basic science discovery and translating them into practical technologies. In that arena then being able to clearly understand how industry works, and what they're looking for and what is their pain points and bottlenecks in bringing new products into markets, that insight is clearly important. And that insight will bring back to not only the research at universities, but also at teaching. I was just teaching my class on transistor design yesterday, and I was reviewing with the students the latest advances in transistor design. And if I were not conversant about what industry is doing, I wouldn't be able to do that because I just don't know. But the fact that I am heavily engaged with the industry allows me to impart that knowledge to the students.

And I think that is really important in that direction from industry to academia, both for the research... Because on the research you need to know where to go. So both for the research and the teaching as well. Now, at the same time on the other direction, academic research, how would that impact industry is really as I mentioned earlier, especially in today's environment. Today, we are not quite clear what to do. In the past, we kind of know what to do. So industry knows what the next step is or even the next three steps needs to be. And so the need for academia of course is there, but not as high as it is today because if you ask people in the industry, especially in this [inaudible 00:37:20] industry, if you ask them, "Do you know what we need in three, five, 10 years?" Most of them I would say they say, "I actually don't know." Because the path ahead is less clear and we're out of the tunnel, we don't know which path to take.

And that's where academia comes in because academia is a place where A, you can explore a lot of things very quickly with very low cost. And that's one. And two, academia is filled with people who have no experience. These are students. And you may wonder, well, what do these people with no experience, what can they do? Well, they do very interesting things because they have no experience, they have no preconceived notion of how things should be done.

Bruke Kifle: No constraints.

Dr. Philip Wong: Constraints, yes. And therefore they will come up with things that nobody in industry have thought of, because industry are used to think about ways and things in a certain way and these students have no idea how people were thinking about it before. So they come up with very, very interesting things that nobody had thought about, and that's where new ideas come from.

Bruke Kifle: Yeah. I think that's a very perfect way to this idea or this notion of constraints. In industry there are objectives that you have to meet, there are business targets or organizational targets. Those kinds of one may call them constraints, can stifle some of the progress or innovation or moonshot thinking. But I think you captured it perfectly where in academia where presumably some of those requirements or constraints don't exist, that's where you can see the true success from an innovation point of view.

Dr. Philip Wong: Absolutely.

Bruke Kifle: Yeah. So in terms of looking forward at future directions, I think the widely known that the Covid-19 pandemic has caused a lot of significant disruptions to the global supply chain, but particularly for the semiconductor industry. And I think it was interesting because the pandemic sort of created this dual shock to both supply side and demand side, where you see a boost in the demand for these devices and products as people are shifting to remote work. But then you also see a hit from a supply side on the global supply chain. So moving forward, what do you think needs to be done to address these challenges and ensure the industry's resiliency, but also continued growth and success?

Dr. Philip Wong: Yeah, that's a really timely discussion here in a COVID-19 and also conflict geopolitical situation today, causes a lot of disruption in global supply chain and also awareness of supply chain resiliency. And you see a lot of regions and countries who want to be able to have local industries and things like that. And that has several implications. One is that collaboration across countries, across boundaries has become a lot more difficult. And hopefully smart people will come up with policies and ways to navigate around this so that cross border collaboration can flourish because knowledge knows no boundary. There's no reason why one region knows everything. So knowledge knows no boundary. And in order to advance technology which basically benefits the entire world, we want to benefit the entire world, in order to advance technology we really need global collaboration. And I'm hoping that smart people will come up with policies and methods to enable this to continue.

Secondly is that we see very strong demand signals for continued advancement in semiconductor technology because as you mentioned earlier, is a foundation of almost everything modern society would do from solving energy sufficiency problems, to food security, to climate change. We all need electronic systems that would help us do our job better. So from that point of view, the demand signal is very strong. And so we would need to have a rapid advancement in technology. And now where do these advancements come from? They come from people because the ideas come from people, the research and developments come from people, and the manufacturing comes from people. So cultivating talent is probably one of the most important things that every country and region needs to do, cultivating talent. And the talent is clearly the driving force for technology development going forward. So the most important thing for society to do is ensure that we have a very healthy industry, so that young people who are contemplating getting into a new career would consider going into this direction because it has a healthy industry. And having that healthy industry is a necessary condition for talent and workforce development.

Bruke Kifle: I see. And yourself you do actually play a very instrumental role in that. As a professor, I'm sure you've taught and mentored many students who've gone on to become successful researchers, who've joined industry, who are entrepreneurs, who are leaders in their own fields. So what are some of the skills and qualities that you really look for, but also you try to cultivate in your students to ensure they're ready for that next stage? And more generally, what advice would you give to young aspiring engineers, scientists, who really want to make an impact in the world?

Dr. Philip Wong: First of all, technical excellence is really the necessary condition. We can think about everything about kind of a more higher level of things such as solving societal problems and things like that, but in order to solve societal problem you have to have the technical expertise to do that. So technical excellence is clearly the necessary condition but it's not sufficient, obviously. And a sense of curiosity is important and an attitude of questioning what is normally done, is that really the way to go? Some call it question of the status quo. That is important because that's where new ideas come from. But as you question the status quo, you need to be sure about your understanding about the status quo. You know what I mean?

If you come up with new things and you need to be able to say how this new thing compare with what we do things today. And in order to make that comparison, you need to know exactly what is done today. So a lot of people kind of miss that part of it in the sense that, oh, I come up with new things. But okay, new things is different but is it better? If it's not better, then why is it good? So being able to understand the status quo is important, and also being able to retain that level of curiosity is clearly important. And that's from a technical point of view, but that is only probably a necessary but not sufficient condition. Really I ask my students to maintain a broad view, a broad perspective, not only of the technology but also for the applications as impact, because those broad views will often take you to places where other people would not go or be aware of the opportunity there.

So for example, combining different technical areas and make a progress by combining the two areas. And also having a broad view of the application space because the application will drive your research direction. So being able to have a broad perspective is important. Going deep is good, but going deep in and of itself is not enough.

Bruke Kifle: I see. So technical depth and excellence, the intellectual curiosity, the ability to question things, but to do that with a good understanding of the status quo. And then finally having this broad perspective or this broad view of what are the practical applications, what are research collaborations, what are ways to intersect this line of work with other fields.

Dr. Philip Wong: Absolutely. And also I should mention that, maybe quote my former dean of engineering at Stafford, Jim Plummer. He said, "Engineering is a team sport. If you are a loner, you won't be big on our progress." It's a team sport. So you need to collaborate.

Bruke Kifle: Excellent way to capture it. So looking ahead, what do you see as some of the most exciting research opportunities in the field of semiconductor technology? I know we mentioned the dark tunnel and finally reaching the light. You also discussed some of the work with 2D shrinking. Moving towards this LA to Manhattan analogy, what are some of the exciting opportunity areas that keep you up at night?

Dr. Philip Wong: Well, two things. One is that I mentioned earlier building 3D chips, and how to build 3D chips and how to come up with a variety of device technology that I would call application domain specific device technology. Let me explain, for the past few decades we have one device technology, silicon transistors. And that does everything from storing the data to doing the computing, to running your radio for the cell phones and things like that. So one technology does everything. Now, of course we have one thing that does everything that's wonderful, but also there's inherent inefficiency. It's more or less like you drive a 18 wheeler truck every day because you expect to move from East coast to west coast one day and say, therefore, I drive this truck even though I just drive the truck to buy groceries. That is not efficient.

So this is what we have today, but going forward we demand extreme energy efficiency so therefore we need different device technology to do very specific things that makes it very energy efficient. If I want to just go to campus and teach a class, I ride my bike because it's way faster. I can pop my bike right in front of the classroom, but I can't ride my bike to move my house so you need a truck. So we need to develop very specific, what I call domain specific device technology to make these 3D chips because then we need extreme energy efficiency. So that's from the kind of base fundamental device technology level. At the higher level than going into how to build a system point of view, then the people who develop semiconductor device technology really need to work very closely with people who develop the applications, because how you develop the technology because of the efficiency requirement has to be closely coupled to how you're going to use it.

Whether you use this chip for an automobile or a medical device or a wearable device, have a very different design point. You've got to design the technology very differently. And in order to achieve the highest energy efficiency, you need to co-design these kinds of systems from device technology all the way to how you're going to use it in the system. So that kind of co-op optimization requires people who would understand across what we call the system stack, across different levels of extraction from device technology to system design to even software design. So that is a big challenge for people in this field is how could somebody comprehend so many things? And that's where the team sports come in. You need people in a team who can talk to each other, who understand each other's languages. So that's where the research direction comes in. Apart from the basic device technology, the other direction will be the more and more closer and closer coupling between the user application and the fundamental device technology.

Bruke Kifle: I see. So domain specific technologies and this idea of co-op optimization. One thing that I would love to just raise quickly is obviously we're in an AI arms race.

We're seeing the rapid evolution and growth with generative technologies. And I'm sure one of the biggest practical application areas is accelerating or improving the efficiency of how we run these large scale language models. So where do you see the implication or the impact of some of the advancements from semiconductor technologies on accelerating this AI development that we're seeing in recent years?

Dr. Philip Wong: The advancements that we've seen in recent years in AI actually are really enabled by three things. First of all, new AI algorithms and new architecture of doing the computation required for AI certainly. Secondly, availability of a large amount of data because many of the AI, machine learning models trained on data. So second thing is availability of large amounts of data, enormous amount of data such as all the data you find in internet and say the data that you collect wearing your iWatch or wearable devices and so on. They're collecting data all the time. So availability of these huge amount of data to train AI model. The third is that you need to have really powerful computer to crunch this AI model. To train an AI model, the ChatGPT takes months of computer crunching numbers, 24 hours, seven days a week. So three things, new algorithms and architecture. Second, availability of a large amount of data and third, very energy efficient and high speed computers.

Out of these three things, two of them rely on semi-conductive technologies. Well, of course, energy efficient computing relies on semi-conductive technology. That's very obvious. But the availability of large amount of data also depends on semiconductor technology because where did the data come from? They are collected by devices that operates on chips. So without these ubiquitous deployments of chips, you wouldn't have these big data that is available to all of us today. So in that regard, the potential advancement of AI will necessarily be dated by advancement in some graphic technology. Take an example, you wouldn't be able to run ChatGPT using computers that are 20 years old. There's no way you can do that. So that is very important to realize. And going forward, the AI revolution will revolutionize many things that we do. As I mentioned before, one thing can influence another and then another one thing would come back and influence the other thing.

That's just like biology influencing electronics and electronics influencing our understanding of biology. And this two-way street also exists, the electronics and fundamental semiconductor technology will help propel AI to go forward because then you can train even more powerful model, even more complex algorithms and so on. That's for sure in one direction. In the other direction, the application of AI and machine learning would revolutionize the way we fabricate and manufacture these semiconductor chips. Today, you probably hear about building semiconductor fabs and so on. We need a lot of people to run the fabs and so on. Why do we need that many people to run the fab? We don't need that many people to run the fab. If we need to produce 10 times more chips, we couldn't afford to have 10 times more people to run the fab. We need to then be more efficient in running the fab and being able to run the fab with 10 times less people.

And how do we do that? Well, AI and machine learning would be able to help us on that going forward. And so this is kind of symbiotic relationship I see going forward will be very important to have.

Bruke Kifle: I see. Yeah, I think this triad of data compute and algorithms maybe some folks, I myself included, failed to realize the importance of compute on the data as well. Not just the computing technologies for training and running inference, but also how important those computing technologies are for enabling the massive amounts of data that are powering a lot of these models. I have maybe one more question. Thinking about some of the potential risks or challenges, I know for instance with the emergence of LLMs there's growing discussion around this idea of responsibility of ethics. So with emerging technologies in semiconductor technologies, how do you think about some of the ethical implications, for example of... I don't know, neuromorphic computing or the environmental impact of semiconductor manufacturing? Like in general, what are some of the potential risks or challenges associated with some of these emerging technologies and what's the best way to think about addressing them?

Dr. Philip Wong: Yeah. And I'm glad you bring up these environmental aspects of the semiconductor manufacturing. So the industry has already been moving toward what we call green manufacturing, recycling things and so on. For example, a new modern fab built today, 99% of the water is recycled. So no one drop will be wasted, everything is recycled in terms of water, for example in terms of power consumption, energy consumption, that is also heavily invested. Companies are also very heavily invested in reducing power consumption and so on. But one thing that I point out in an interesting study that I saw just recently, that for every unit of electricity that a semiconductor manufacturing fab used, they would produce chips that would save four units of energy that would not otherwise have been saved. And so for every unit... It is a great investment, for every unit of electricity they use to produce a chip, that chip will in turn save four units of energy.

Bruke Kifle: That's a pretty good trade off.

Dr. Philip Wong: Yeah, that's good trade off, good return on investment.

Bruke Kifle: I see. I want to just end with one final question from the perspective as a visionary in your field, but just looking more broadly into the field of computing, what's maybe one or two emerging technologies that you're excited about? And then what is maybe one or two grand challenges or open questions that you'd like to see addressed by the computing community?

Dr. Philip Wong: Yeah. So for the first one, the exciting technologies, I am very optimistic that we will go beyond the use of silicon as the only technology for computing and computation. I think going in the next decade or so, we will see the emergence of other materials that will be used in a computing system. So I'm pretty optimistic about that. In terms of your second question, the kinds of impact that we will be making is that I think there will be broader and broader use of the computing systems going forward. And that would really bring about a sea change in the way we operate, because many of the things that we wanted to do from self-driving cars to energy efficient electric grid and so on, and energy storage and so on, they're all gated by advances in the basic semiconductor technology. And going forward, I'm expecting that the society will continue to progress really propelled by advancement in semiconductor technologies.

The semiconductor technologies are kind of marginally invisible to people. We see the phones and we touch the screens of the phone, we listen to things, we watch videos, but oftentimes we don't see what it is powered by. And with the recent attention and supply chain resiliency, pandemic and so on and so forth, I'm expecting that the general public, the general society would recognize more than before the [inaudible 00:59:57] of what drives these things and therefore recognize the importance of this basic technology. And I think that would be useful for everybody and helpful to propel the advancement of technology going forward.

Bruke Kifle: And I think certain people certainly have. So the future is bright. And I think we're certainly, as you alluded to earlier, nearing the end of the tunnel. So very excited to see where the future of the field will continue to be, especially with individuals like yourself helping drive the future directions. So Dr. Philip Wong, thank you so much for taking the time to join us. It was really a fruitful conversation.

Dr. Philip Wong: Thank you for the opportunity to speak with you. Thank you.

Bruke Kifle: All righty. Thank you so much, Professor. ACM ByteCast is a production of the Association for Computing Machinery Practitioner Board. To learn more about ACM and its activities, visit acm.org. For more information about this and other episodes, please visit our website at learning.acm.O-R-G/B-Y-T-E-C-A-S-T. That's learning.acm.org/bytecast.