00;00;00;02 - 00;00;08;16
Kara Miller
Welcome to MIT's Computer Science and Artificial Intelligence Labs Alliances Podcast series and Kara Miller.

00;00;14;05 - 00;00;25;04
Kara Miller
On today's show, as applications for AI explode, we're going to take a look at rising energy consumption, which has both environmental and financial dimensions.

00;00;25;15 - 00;00;31;07
Vivienne Sze
Moving data around is very expensive and the further you have to move it, the more expensive it's going to be.

00;00;31;20 - 00;00;40;28
Kara Miller
Vivienne Sze, an associate professor at MIT Department of Electrical Engineering and Computer Science, talks about why there's growing interest in specialized hardware.

00;00;41;08 - 00;00;57;26
Vivienne Sze
And this has come from companies like Google, I think Meta just announced that they had also built some chips. Microsoft, And these are traditionally software companies, but they realized that you need to have specialized hardware in order to give you the compute. You need to be competitive.

00;00;58;06 - 00;01;03;10
Kara Miller
But reducing energy consumption means you also have to take a look at your software.

00;01;03;22 - 00;01;16;19
Vivienne Sze
What you would like to do is that while you're designing your algorithms, obviously you care about accuracy, but it's also important to factor in energy consumption in the design process of the algorithms.

00;01;17;00 - 00;01;23;09
Kara Miller
And says Sze stakes are high as we increasingly plan for computing to show up pretty much.

00;01;23;09 - 00;01;36;20
Vivienne Sze
Everywhere, even if each of these vehicles only drove for one hour a day, the amount of energy that they would consume would be comparable to that of today's data centers.

00;01;37;05 - 00;01;53;06
Kara Miller
So coming up, the costs of AI computing and the breakthrough work on how to let AI proliferate while still addressing those costs.

00;01;56;25 - 00;02;06;04
Kara Miller
Vivienne Sze first began thinking about how crucial energy efficiency is to computing. Back when she was in grad school at MIT and she was working on video compression.

00;02;06;12 - 00;02;14;28

Vivienne Sze
And this was in the days before the iPhone. So it was a really exciting concept of being able to, you know, process and view video on the phone itself.

00;02;15;09 - 00;02;25;15
Kara Miller
The thing was, there was not a lot of battery life on phones, which was a serious problem. The hardware had to change, but that was not even good enough.

00;02;25;28 - 00;02;42;01
Vivienne Sze
We also realized that there were limits in terms of how low power you could make the hardware before you were limited by the software or the algorithm itself. So then what I started to do was look at how do we actually change the algorithms to also make them more low power friendly.

00;02;42;11 - 00;03;09;01
Kara Miller
When she returned to MIT as a faculty member in 2013, after working in the Research and Development Center of Texas Instruments, her work expanded and it changed beyond compression. She wanted to know if she could make understanding, images and video more energy efficient. She coauthored a book called Efficient Processing of Deep Neural Networks and taught classes to MIT students during the year and to those in industry during the summer.

00;03;09;16 - 00;03;19;02
Kara Miller
And now, amidst a whirlwind of chatter about AI and the proliferation of computing more generally. Szeâ€™s work has taken on a new urgency.

00;03;19;15 - 00;03;36;21
Vivienne Sze
I mean, I think there's been a tremendous amount of excitement about the many applications of AI, and I think one of the reasons for that is that it took a really long time for AI to get to where it is today, to have the level of accuracy that could actually be deployed and be useful for many things outside of the lab.

00;03;37;09 - 00;04;05;20
Vivienne Sze
And I think now is actually the time where you know that now that it has that level of accuracy that can actually be used, it now makes sense to kind of look at how to address the cost. In particular, what we're talking about here is the computing and the energy consumption cost. And also the accessibility of it. But I think like actually, I think that there's been a lot of discussions now more recently about the computing cost because the models to make these in particular the deep neural network models to make the AI work have been quite large.

00;04;05;20 - 00;04;25;25
Vivienne Sze
So I think that a lot of news has been drummed up by stuff like Chat GPT and we know that that has hundreds of billions of parameters in that model and it takes millions of dollars to train. So a lot of that cost is actually from the electricity bills and then it takes potentially even more energy to run the inference because you're going to run it many times.

00;04;25;25 - 00;04;41;25
Vivienne Sze
And so inference basically means how you would use the model to actually do a task after you've trained. And so there's been speculations that it's like millions of dollars a month. Just do the inference. Well, so now that there's a cost, a very clear cost associated it, now there's a clear application of it that people are trying to figure out how to cut those costs.

00;04;42;22 - 00;05;08;29
Kara Miller
You know, I want to go back to something you said before, which is like there's different ways to deal with the energy. One is to make the hardware more energy efficient, but the other you talked about making algorithms better and more energy efficient. I think to a lot of people who are less technical than you, the notion that some algorithms are less energy efficient and some are more energy efficient would be surprising.

00;05;09;07 - 00;05;13;28
Kara Miller
Is there a way, like in simple terms, to explain how that is possible?

00;05;14;17 - 00;05;39;07
Vivienne Sze
Yes, I think there's a couple of ways. We actually there's always many ways to do something. So I think the point is that what you would like to do is that while you're designing your algorithms, obviously you care about accuracy, but it's also important to factor in energy consumption in the design process of the algorithms. And so this can appear in various different ways.

00;05;39;25 - 00;06;02;16
Vivienne Sze
So, for instance, I think one thing is to understand where the energy is being consumed in the algorithm itself. So, for example, in deep neural networks, it tends to be this process of your Deep Neural network is composed of these layers, and there's kind of these like layer by layer processing you can imagine. And some layers consume more energy than the others.

00;06;02;27 - 00;06;30;25
Vivienne Sze
And so you might want to look at which layers consume the most amount of energy and then try and optimize those layers to reduce the energy consumption. So you're not changing the entire network, but only maybe a subset of it, particularly the part that consumes the most amount of energy. I think the other thing which we talk about, yeah, but like if we're if you think about the hardware itself, I mean, one question is also like, why is there energy consumption or where does the energy consumption go?

00;06;30;25 - 00;07;02;16
Vivienne Sze
And actually it turns out that energy consumption is dominated by what we call data movement. So, for example, if you want to do a multiplication, it actually takes more energy to read the inputs to that multiplier from some memory and then to write the results of that multiplication than to perform the multiplication itself. Right? So really, if you want to address energy efficiency, you want to reduce the cost or the frequency in which you're moving data around.

00;07;02;25 - 00;07;28;11
Vivienne Sze
And so, you know, the notion of when you're designing specialized hardware for DNN that's what you want. If you want to make the specialized hardware energy efficient, you really want to target this data movement. This could be by doing things like exploiting what we call data reuse. If you're going to use the same piece of data over and over again, maybe you can read it once and get close to the compute and then use it many times.

00;07;29;03 - 00;07;59;08
Vivienne Sze
You can also structure the hardware so that the storage is much closer to where you're doing the computation so you don't have to move as far. And so as a result, the data movement is much cheaper. So, you know, in collaboration with Joel Emer who is also a professor in CSAIL, and our student Eugene Chen, we had looked at this in the context of deep neural networks and how do you design these energy efficient data flows that really optimize for data reuse and minimize data movement?

00;07;59;15 - 00;08;23;06
Vivienne Sze
And that allowed us to design this chip that's called Iris, and she's like an order of magnitude more lower energy than mobile GPUs. Now, on the flip side of that, when you look go back to the software side of things, you can also say, hey, if I know that data movement is very expensive, can I design these neural networks that are, you know, require less data movement kinetics?

00;08;23;07 - 00;08;36;02
Vivienne Sze
Can we have neural networks that use data that's reused multiple times so that the hardware can exploit this? And that's kind of one form of doing the co-design of both the algorithms and the hardware together for energy efficiency.

00;08;36;06 - 00;08;54;05
Kara Miller
And when you talk about sending data around in a simple way, it sounds like, you know, if you've got a cell phone and you're trying to do four times five, it's better if it can be done there than if it's like outsource essentially like, oh, can it be done in the cloud and then sent back here? That's very costly energy wise.

00;08;54;16 - 00;09;12;27
Vivienne Sze
Exactly. So I think yeah. So I think the analogy is in general, moving data around is very expensive and the further you have to move it, the more expensive it's going to be. So obviously having to transmit something to the cloud in back that's very expensive. So honestly, locally, if you want to do it on the phone, that's going to be more energy efficient.

00;09;13;03 - 00;09;30;19
Vivienne Sze
But even what you're saying on the phone, there's other different levels of making it local. So, for example, when we look at a computer, you have a computer chip that's doing the processing, but you also have an additional chip not on the chip itself that is storing data, the memory, what we would call, you know, DRAM, for example.

00;09;30;19 - 00;09;51;29
Vivienne Sze
So some of you buy or sample those DRAM. And so reading data from that DRAM, which is not on the chip where the processing is happening, is going to be much more expensive than reading the data from a memory that's actually built onto the same computer chip as where the processing is happening. But of course, you know, there's a limitation in terms of how much you can store on the computer chip itself.

00;09;52;07 - 00;10;07;11
Vivienne Sze
So the idea here is that if you read the data, let's say once from the DRAM you want and store it on CHIP, you want to use that piece of data as many times as possible on your computer chip. So then you can really reduce the amount of expensive data movement or data access from that DRAM itself.

00;10;08;11 - 00;10;39;29
Kara Miller
So you've talked about a few reasons why you might really want to be able to do a lot more on these edge devices like cell phones, like these things sort of on the edge of the network then maybe we do now. Do you want to talk about like what might be the motivation? Because obviously there's the energy efficiency piece of it, but there's a sort of other things that being energy efficient can allow you to do and other motives that might drive you.

00;10;40;19 - 00;10;59;08
Vivienne Sze
Right? Yes. I mean, there's several reasons why you want to do the compute locally as opposed to push it to the cloud.

So I think the first thing is that ideally you would like for the technology to not be reliant on some large communication network. You want to kind of untether yourself from the communication network and that increases accessibility.

00;10;59;08 - 00;11;20;23
Vivienne Sze
We know that, you know, around the world and even in the U.S., there are areas where you don't have a very strong communication network. Right. And so there's a lot of, you know, promising applications for AI, We want that to be accessible to everyone. So even those who are not connected to whatever network or the cloud. The other thing, of course, is actually privacy and security.

00;11;20;23 - 00;11;42;00
Vivienne Sze
So there's a lot of exciting application, again, that involve very sensitive data, be it health data or financial data. And again, it will be more secure. And for privacy reasons, you might want to, you know, bring the, you know, compute to where the data is being collected, let's say, on your device, rather than sending that very sensitive data to the cloud itself.

00;11;42;17 - 00;12;09;08
Vivienne Sze
And then finally, there's a lot of applications where you need to interact with the world, right? So, for example, self-driving cars or robotics is one and other would be, you know, AR and VR in those situations. Latency is really critical. So the reaction time is very key. And so you might not have enough time, let's say, for yourself in a in a robot to send the information to the car, wait for it to be processed and come back, especially if you're moving at very high speeds.

00;12;09;15 - 00;12;10;17
Vivienne Sze
Right, right, right, right.

00;12;11;02 - 00;12;26;25
Kara Miller
Do you want to talk about one of those? I don't know, maybe health or like a specific example where you could imagine that if you could get things more energy efficient, then possibilities would open up that maybe do not currently exist.

00;12;28;00 - 00;12;56;18
Vivienne Sze
So I think like very much the more obvious one would be the robotics and the navigation side of that, because really it is there are a lot of cases where you really both from a functionality sample, from a latency and interaction standpoint that, you know, being tethered to the cloud would not work. And also it would be very, I guess, dangerous if you're reliant on a communication to the cloud and then you know, you lose connection, you would want you to be able to, you know, be able to navigate itself.

00;12;56;19 - 00;13;15;15
Vivienne Sze
I think that's really been very critical. I think there's a lot of excitement in the health care space. I think that that's still kind of trying to figure out where is the best place to use AI and in that particular space. One thing that we've been looking at this is in collaboration with Thomas Heldt, who is also a faculty member at MIT.

00;13;15;15 - 00;13;59;08
Vivienne Sze
Here. We've been looking at the application of eye tracking, particularly as it relates to using that to detect or diagnose neurodegenerative diseases like Alzheimer's and Parkinson's. As it turns out, there is a correlation between, you know, how your AI moves and its reaction time to whether or not you have those diseases or the progression of that disease. And in the past, a lot of the tests in terms of doing, you know, the eye tracking was done in the clinic on these expensive hundreds of thousand dollars or millions of machines to do like kind of the eye movement evaluation.

00;13;59;08 - 00;14;19;04
Vivienne Sze
And what we've been looking at is whether or not feasible to do these tests in home on, you know, your tablet or phone and collect a lot of eye data movement. And this allows you to provide additional information to a clinician to help make these diagnoses or actually, you know, determine whether or not certain medication works or not works.

00;14;19;13 - 00;14;30;01
Vivienne Sze
And in these particular patients, again, you might not want to share all the data of your, you know, eye movements and stuff after the phones are doing it locally might be desirable.

00;14;30;01 - 00;15;02;16
Kara Miller
You know, when we talk about the work that you're doing in trying to make computing more energy efficient, can you just give a sense of like magnitude? How big a problem is this? Like, do you feel like, boy, weâ€™re a good chunk of the way there or, oh my gosh, weâ€™re, you know, we're like 1% of the way down the road or like, you know, in order to like, let's say, have some of these applications where, yeah, you have self-driving car, you can't have latency, so you have to have things more at the edge.

00;15;03;06 - 00;15;11;04
Kara Miller
You have to have more computing power. Like I guess I just wonder, like, how far are you down the road of solving this problem?

00;15;11;24 - 00;15;46;20
Vivienne Sze
Yeah, I think there's certainly a lot more to do. So I would put it this way. I think the impact of this problem depends on two things. It depends on how much computation you have to do for a given tasks and then how many devices out there that are performing this task. So in terms of looking at how much computation you need to do for a given task, when we look at the trends and let's say deep neural networks, which is one of the kind of core techniques for A.I. that is really given this incredibly high amount of accuracy that we're seeing, we've seen over the past ten years or so, or even less than

00;15;46;20 - 00;16;12;18
Vivienne Sze
ten years, that in conjunction with the significant increase of accuracy of these techniques, we see orders of magnitude like hundreds of thousands of X fold increase in terms of the amount of compute that you actually need. And so as you know, the amount of compute rise is obviously we need to continue to push on the energy efficiency because it makes the task much more challenging.

00;16;12;18 - 00;16;38;12
Vivienne Sze
So from a both the algorithms perspective and from the hardware perspective, we need to continue to innovate to drop the energy efficiency. So from a hardware perspective, some of the new things that people are looking at are also things like can we actually move the computation into the memory itself? So rather than just putting the compute and the memory closer together, can we, you know, embed the compute into the memory or the storage element itself?

00;16;38;12 - 00;17;19;24
Vivienne Sze
And there's been a huge amount of work in that space. And so there certainly is a continued demand for that. Then if you think of the problem also from a scale point of view in terms of like the number of devices that are using this, it also can grow significantly. So some of the work that one of my myself and my collaborator, Sertac Karaman, is also a faculty at at MIT and our student, Soumya Sudhakar had looked at recently was what we have focused a lot of our research on, you know, making A.I. energy efficient for edge devices, in particular with certainly we've been looking at

in the robotics space, but looking at, you know,

00;17;20;00 - 00;17;50;02
Vivienne Sze
what we call low energy robots, where actuation is very low cost. But we were kind of curious to see, you know, what are the implications of this If we look at, you know, larger vehicles, for example, self-driving cars, which also require a huge amount of compute. And so we kind of did a study as our study and had looked really deeply in terms of what are kind of the trends in this space are, how can we kind of model what's happening in terms of compute for self-driving cars and the adoption of self-driving cars?

00;17;50;12 - 00;18;27;17
Vivienne Sze
And surprisingly, what it revealed was that, you know, if you consider the fact in a self-driving vehicle, you have many cameras in a self-driving car crossing to understand its surroundings. So it's a huge amount of data being pulled in. And if you kind of look at the trends of, you know, how many vehicles are out there, and, you know, in the ideal case, how many self-driving cars, if they were all two or significant portion of them were to become self-driving cars, let's say if you had like a billion vehicles, even if each of these vehicles only drove for one hour a day, the amount of energy that they would consume would be comparable to that

00;18;27;17 - 00;18;52;12
Vivienne Sze
of today's data centers. So it's just the scale of the number of devices and the amount of data that it needs to crunch through can be quite significant. So the other aspect of energy efficiency is to look at from the angle sustainability and the carbon footprint. And so obviously as we move forward, these, you know, very promising and exciting technologies, we also need to consider their impact on the environment.

00;18;52;24 - 00;19;12;23
Vivienne Sze
And so their energy efficiency is also pretty key. So it's not not to say that we should not do self-driving cars, but as we're designing these vehicles, it's important to consider the energy efficiency of the computer because we're adding compute that wasn't there before. And yes, of course, I think that's also factored in in general because you want to increase the distance of the vehicle that you can drive.

00;19;12;23 - 00;19;29;05
Vivienne Sze
So you care about energy efficiency anyways. But even if we had, you know, better batteries and more energy, you would still want to account for that fact. That energy has to come from somewhere. And so accounting for energy efficiency of the compute which wasn't there before is really critical.

00;19;30;04 - 00;19;39;19
Kara Miller
Do you feel like the work you do and just more broadly, the idea of energy efficiency could help to make AI more ubiquitous?

00;19;40;28 - 00;20;10;28
Vivienne Sze
Certainly, I think that that's one of the things one of the goals at least, is to, you know, there's a lot of promising applications for air, but we certainly want it to be both accessible. So as I mentioned before, not relying on a huge amount of compute because also a lot of people can't afford, you know, huge racks of servers to run their application and then there's also, you know, really exciting applications where if you could embed at a lower cost, you can really utilize it to, you know, be used in a lot of applications.

00;20;10;28 - 00;20;30;05
Vivienne Sze

I'm not an expert at this particular situation, but I know that if, for example, in agriculture there's a lot of exciting use cases to have AI and putting these sensors around the fields to, you know, collect more useful data that can inform the farmer. But of course, you know, you don't have a lot of energy out there. You don't want to be constantly replacing batteries.

00;20;30;05 - 00;20;33;24
Vivienne Sze
So it can be useful for those particular applications as well.

00;20;34;15 - 00;20;56;10
Kara Miller
Do you feel like in your experience, the business community realizes the importance of the. They certainly recognize the importance of embedding AI in everything they do. I feel like that's a a common thing from businesses now is like, how will we thread in AI to our business? Do you think people are aware of energy efficiency in the same way?

00;20;57;00 - 00;21;23;15
Vivienne Sze
Oh, actually I do think so, in the sense that there's been a lot more interest in building, you know, specialized compute hardware, and this has come from companies like Google. I think Meta just announced that they had also built some chips, Microsoft, and these are traditionally software companies, but they realized that you need to have specialized hardware in order to give you the compute.

00;21;23;15 - 00;21;46;28
Vivienne Sze
You need to be competitive. Another lens to look at this, there's been a huge number of hardware startups out there, primarily looking at, you know, designing specialized chips for A.I., some of which are investigating this whole, you know, putting compute into the memory or even doing compute with light again, because they realized that you need a lot more computation to support A.I..

00;21;46;28 - 00;22;05;11
Vivienne Sze
And part of the thing that limits computation is the amount of power that you can supply to your heart, given limited amount of power. If you can cram more computation into there, then you can, you know, run a lot faster. So that energy efficiency is the key, you know, way to cram more compute for a given power envelope.

00;22;06;28 - 00;22;15;01
Kara Miller
What do you feel like the biggest hurdles are going forward and how are you thinking about addressing those hurdles?

00;22;16;06 - 00;22;45;00
Vivienne Sze
So I think one thing that I got are two things that I'm particularly interested in, is, as we were just talking about in terms of, you know, one way to address the energy cost of AI is to build very specialized hardware that targets AI. That being said, the more specialized you are that also limits your flexibility. And in the context of deep neural networks, this is a very fast moving and growing field.

00;22;45;15 - 00;23;19;21
Vivienne Sze
So from both the research perspective and I think it also applies in industry, is that how do you build specialized chips that are efficient but are also flexible? So every single time you introduce flexibility, usually there is going to be some form of cost in terms of, you know, energy efficiency or so on, but how do you kind of find the right places to add that flexibility where it's worth paying a little bit more cost, but you can as a result, your hardware is much more flexible, it can support a wider range of tasks.

00;23;19;21 - 00;23;39;06
Vivienne Sze
And also you try to kind of future proof it somewhat. We talked about the fact that these self-driving vehicles also need to use AI. And so you can imagine that you're unlike your phone, you're not going to swap out your car every two, three years, ideally. And so you would want some hardware, you know, whatever computer you have there to last long enough in the future.

00;23;39;22 - 00;24;02;23
Vivienne Sze
Right? And so I think that's going to be on the challenges that also folds into actually also sustainability, because as it turns out, the carbon footprint of computing is not just the energy that it consumes while running the processor, but there's also a carbon footprint of manufacturing. The processor itself is a computer chip and there's been a lot of research at this for various places.

00;24;03;14 - 00;24;21;06
Vivienne Sze
And so one of the goals there is that if you can make a chip that's more flexible, hopefully that also means that it can last longer or has a longer useful life. And so then this carbon footprint will be lower as well. So I think that that is a challenge that we need to look and address and just kind of think through.

00;24;21;06 - 00;24;23;21
Vivienne Sze
What does it mean to be both flexible and efficient.

00;24;24;25 - 00;24;37;27
Kara Miller
In your experience, How have you found that your work most commonly sort of makes it out into the real world, hardware or software, Like what is the path that that generally takes?

00;24;38;19 - 00;24;57;27
Vivienne Sze
Yeah, I think there's there's been a couple of ways in which we can translate some of our research from the lab out to industry. I think the first is obviously the educational component. So I think it's always good to kind of teach. I mean, it's like I'm in this job, it's like teacher advise in terms of like, how do you think about this problem?

00;24;57;27 - 00;25;15;17
Vivienne Sze
What are the questions that you should be answering? Because, you know, it's one thing. Just be like, you know, you can evaluate different designs yourself, but it's better to for the, you know, the people to learn the principles of doing that, doing that themselves. So I think the educational component is really key, and that's why we wrote the book and we ran a bunch of tutorials.

00;25;15;17 - 00;25;37;25
Vivienne Sze
So I think the second part is obviously can do a form of like advising and consulting various companies in terms of, you know, the directions that they should head, the questions that they should be asking. And certainly there's been a lot of interest in, you know, forming companies and doing startups in this area. I myself have not pursued it, but there's a lot of colleagues of mine who are in this space as well who have pursued that avenue as well.

00;25;38;10 - 00;25;58;10
Kara Miller
Yeah. A final question for you, which is when you sort of look ahead, what do you feel like personally is something you're most excited about in terms of a problem you want to solve the the piece of this that you want to tackle or just like what? Yeah. What kind of gets you excited when you think about the years ahead?

00;25;58;21 - 00;26;39;29
Vivienne Sze
Yeah. So I think there's two things. So one is certainly trying to make I more accessible to more folks in the sense of not letting the compute costs be a barrier. And actually over the past couple of years, really trying to focus on in general, even beyond AI. So my area of expertise is energy efficient computing. We've primarily targeted maybe smaller devices, mobile devices, But really, given the challenges that we face in terms of carbon emission sustainability, trying to see if we can apply some of those techniques to aid in the challenge of sustainability, which is all one of our grand challenges of today.

00;26;39;29 - 00;27;00;14
Vivienne Sze
So whether it be applying to things like I think there's actually a lot of exciting work already going on in the data center side. But looking also at are there other areas, one of which was, for example, self-driving vehicles and other areas where we can apply these energy efficient techniques to really reduce the carbon footprint of computing because the demands for computing continue to grow so effectively?

00;27;00;14 - 00;27;14;18
Vivienne Sze
Basically, how do you get the benefits of computing without and try to minimize the cost of that? Because we know having computing has really helped a lot of people, but at the same time to mitigate the cost, the environmental costs of computing, right?

00;27;15;14 - 00;27;26;27
Kara Miller
Vivienne Sze is an associate professor in MITâ€™s Department of Electrical Engineering and Computer Science. She's a coauthor of Efficient Processing of Deep Neural Networks. Vivienne, thank you so much.

00;27;27;15 - 00;27;30;05
Vivienne Sze
Thank you very much.

00;27;36;01 - 00;27;58;23
Kara Miller
And if you're interested in learning more about the CSAIL Alliance program and the latest research at CSAIL, please visit our Web site at cap.csail.mit.edu, and keep listening to our podcast series on Spotify, Apple Music, or wherever you listen to your podcasts. Tune in next month for a brand new edition of the CSAIL Alliances podcast.

00;27;59;10 - 00;28;11;08
Kara Miller
Stay Ahead of the Curve. I'm Kara Miller Thanks for joining us.