Evan Sparks: Hi there. I'm Evan Sparks, Vice President of Artificial Intelligence and High Performance Computing here at Hewlett Packard Enterprise. I'm joined today by my colleague and longtime collaborator Ameet Talwalkar, who's a professor in the Machine Learning Department at Carnegie Mellon University, and also one of our top AI researchers here at HPE. We're really excited for the opportunity to speak with you today and share some of our thoughts around a really important and increasingly important issue that's an externality of modern AI, namely its impact on the environment.

So to provide a little context here, what we've seen over the last few years is that really computationally demanding machine learning technology has becoming increasingly prominent in the industry. And this has resulted in increasing concerns about the associated rise in energy usage and correlated, but not always really cleanly, concerns about carbon emissions and carbon footprint of some of these workloads. So Ameet, maybe with your expertise, you can share some light on some of these concrete concerns.

Ameet Talwalkar: Yeah, sure. And first, thanks for having me. It's really great to be here chatting with you, but kind of in a nutshell the big issue is that many high profile ML advances really just require a staggering amount of computation. And I think maybe the best way to talk about this is just via some examples. So first, consider open AI, which is a prominent research lab that created for instance, GPT-3, which is a popular language model that's been in the news a bunch recently. They had a blog post back in 2018 where they revealed that the amount of compute and energy required to perform their largest AI training runs, including for instance, things like training GPT-3 had increased 300,000 times since 2012. And that's a figure at this point that's four years old, but I think the trend is continuing along similar directions.

As a second example, there was an important paper back in 2019, that coined the phrases Red AI and Green AI, that introduced back of the envelope calculations about the cost of training and developing particular machine translation model using a technique called neural architecture search. And they estimated that this particular workload was responsible for over 600,000 tons of CO2, which put in perspective is roughly equivalent to five times the lifetime emissions of a car, including fuel costs.

Evan Sparks: Wow, that's staggering. So who are some of the main players in terms of driving these workloads and all this energy consumption surrounding AI?

Ameet Talwalkar: Yeah. Great question. So, machine learning is a huge active area of research, both in academia and industry, but in terms of energy consumption, it's really large enterprises and the biggest research labs that have the budget and these massive data centers that bear most of the energy responsibilities. And this is largely due to these structural issues. They have access to the compute. They have access to the data. And this includes many of HP's biggest customers. And the other thing I note here is that among these folks, what we're seeing is that these very energy hungry AI workloads are becoming increasingly common for them.

Evan Sparks: Right. In my experience, even just five years ago, these AI workloads accounted for really a minuscule fraction of the overall kind of energy consumption at a lot of these enterprises. But

now in some cases, the script has completely flipped and we're seeing greater than half of the total computational demand and energy usage associated with these AI workloads rather than the traditional workloads. So it's all happened very quickly. So I guess the next question is, in your view, now that we understand the problem a little bit better, how concerned do you think we should be and what should we be doing about this today and in the near future?

Ameet Talwalkar: Yeah. Great questions. And I think there's a bunch of answers here which I'm happy to go through one by one. The first thing I'd argue is that the fact that we're having this conversation at all is really good. I mean, both the two of us right now, but more generally the fact that the machine learning community is thinking about this is really important. It's really important to raise awareness around what's going on and this is something that we weren't really talking about as a machine learning community a few years ago. So it's good to try to get ahead of this issue, put pressure on ourselves as a community and, in particular, on those organizations potentially most responsible for these energy concerns to make sure that we're acknowledging and putting serious resources towards trying to address these concerns.

And I'd argue that this is starting to happen. As an example, a mentor of mine, Dave Patterson from Berkeley, who is a Turing Award winner, he's currently working full time at Google along with a really prominent team of researchers and engineers to actively work on some of these energy concerns, particularly in terms of what's happening internally at Google.

Evan Sparks: Yeah. Dave and the team, both at Google and at Berkeley, have been doing some really interesting work in this area and actually sharing some of the measurements from the field and helping actually quantify how big a problem this is. So I agree with you that raising awareness is really good and it'll help motivate these enterprises, not just financially but also socially, to address this situation. But I guess, in your view what are the next steps to double clicking and really understanding the problem better and getting us towards a resolution?

Ameet Talwalkar: Yeah, absolutely. I think that the first thing we need to do is better understand the problem we're facing, right? At a high level we realize that this is a problem still, but I think we first need to do a better job of accurately measuring exactly the degree to which this is the problem, both in terms of the energy requirements of current AI workloads, as well as coming up with quite accurate projections of what we expect future requirements to look like. To double click on this a little bit, going back to that Red AI paper that I mentioned earlier, they presented a bunch of back of the envelope calculations, which by definition were approximate. They relied on a bunch of assumptions. They didn't really have access to the workloads themselves.

They were kind of just estimating how they imagined these workloads were being performed based on whatever details were shared in papers and so on. And of course these estimates were really helpful and raising awareness, but in some cases they were off by a little bit like estimates always are, and they have the potential to be a bit misleading if we're not careful. Take, for example, that neural architecture search figure that I talked about, which just to remind everybody, that was the workload that when put in context had roughly the same energy or the same carbon footprint as roughly five cars during their entire lifetime.

That's a staggering number but it turns out that the estimate of, there's a few things that are a bit misleading about that figure. I won't go into all of the weeds here, but one issue is that the estimate itself is a bit inflated. And there are good technical reasons to talk about that, but I'm not going to go into details. But the other thing is just better understanding of how that workload is actually being performed and what it's actually doing. And here the idea is that this particular workload, we were conflating two things, this problem of neural architecture search and this problem of training a model together and saying, wow, this is super computationally expensive. And while that might be true, in reality what you typically do is perform neural architecture search very infrequently, and then train models many, many more times, much more frequently.

And so the expensive part of this workload, neural architecture search, that cost is really actually amortized over many, many training workloads and conflating these two numbers or joining these numbers together is potentially a bit misleading, especially when projecting costs in the future.

Evan Sparks: Yeah, it makes sense. You're not necessarily going to run that big, expensive neural architecture search job a million times. You might use its output a million times I think is kind of what you're getting at. So I imagine that most engineers would agree with you, as I do, with the claim that we need to more accurately measure the problem and quantify it in order to fix it. But to be clear in your opinion, the current estimates while rough, still correctly, point to a pretty serious concern with respect to energy consumption.

Ameet Talwalkar: Yeah. Absolutely. This is a real problem. It needs to be taken seriously. And that kind of leads me to my next point, which is that there are clear strategies that we, and in particular, these, these largest organizations that are using a lot of this energy, that can be leveraged in the short term to address, or at least mitigate some of these concerns. So let me go through some of them. The first one is just thinking about the profile of the model training workflow. Model training is one of the most expensive AI workloads and it has a unique computational workload. So a lot of traditional workloads, you have strict SLAs in terms of, I need to serve a webpage within a hundred milliseconds or something like that.

Training of ML model is quite different in terms of its computational profile. It's fundamentally very compute intensive and has a very long turnaround time, say on the order of days or even weeks. And as a result of that we don't really expect to get a result very quickly, but that also gives us flexibility in terms of figuring out how to most efficiently say from an energy perspective, allocate our computational energy resources towards this problem. So for instance, it's fairly straightforward to adapt the workload to allocate computational resources that take advantage of compute when it's really cheap or energy when it's particularly abundant or particularly green. And so, being dynamic in how we're actually performing this training, even while performing the same workload, can potentially lead to much greater efficiencies. And we're already starting to see some of the largest organizations take advantage of that.

There's another aspect of strategies that we can take, which involve leveraging advances in the research community to make training and deployment more efficient. And that's both in terms

of hardware and fundamental kind of advances in software modeling. So in the context of hardware performance, we're already seeing this. We're actively in the midst of hardware proliferation in terms of specialized hardware design, specifically for training and or deployment of machine learning models. And this is what initially motivated Google to work on the TPU. They realized that their computational demands for this particular set of workloads involving AI was going to get larger and larger over time. And they wanted to be more efficient, both in terms of cost and energy in terms of this particular workload.

But at this point, in the next few years what we're going to see is Nvidia and AMD and Intel and SambaNova and Mythic, Graphcore, Cerebra, a bunch of other companies introduce new hardware, specialized hardware, into the market that will make things much more efficient, or that's the hope. And on the other hand, there's this idea of better software, better models. It's an active area of research right now to devise models that are smaller, sparser, and in some way more efficient in terms of training or deployment. And we've seen a bunch of success stories here, but we're still kind of already scratching the surface. And what I mean by that is that as a research community, we still don't really understand exactly how or why deep learning is working. And so I would argue that with a better understanding of these underlying mechanisms behind deep learning, whether theoretical or empirically motivated, we should be able to take advantage of that knowledge to develop more efficient methods.

Evan Sparks: Yeah, totally agreed. I think it's tempting just to say, throw more hardware at the problem, but I think simultaneously as a research community, we're really pushing the envelope as well on the algorithmic advances. There was a follow up blog post to that open AI post from 2018 that we mentioned that detailed a lot of the algorithmic improvements that have happened to increase that efficiency of training and getting these large models out. So really need to make sure that we're pushing forward on both angles. Switching gears and thinking forward a little bit more, many of the concerns surrounding this topic are motivated really by these future projections. We've seen this tremendous growth the last five or 10 years, but the billion dollar question is have we saturated the amount of compute required here or do we think it's going to go in a direction of ever more kind of energy requirements? What's your take on this Ameet?

Ameet Talwalkar: That's a great question and I don't think anyone knows for sure. And I think you and I might have slightly different opinions here, but I would say I personally, somewhat of an optimist here in that we're certainly in the early days of AI progress, seems like we're seeing new applications showing up daily that are pretty amazing and out there that seemed like science fiction just a decade ago or whatever. These are really bleeding edge applications and a big aspect of what we're doing in the research community right now is still really just exploring what is possible rather than really focusing on how to perform a particular task efficiently. And I would argue that as this technology continues to mature, we'll focus not just on what is possible, but as we throw more of our resources into thinking about not just what's possible, but how to do whatever it is that is possible more efficiently, we will improve on the efficiency side of the house.

So I think as the technology matures we'll naturally get efficiency gains. And I think there is precedent for this. So I think an analogy in a fairly different area is in genomics with the human genome project, which happened a couple decades ago. So this is a project that, nearly $3 billion

of funding and over 13 years to map the human genome. There was a lot of fanfare behind it when it started. Thirteen years later the outcome was to some extent viewed as being mixed. It was really expensive. We didn't immediately have personalized medicine as an outcome of having sequenced the human genome, but it did provide the foundation for subsequent technology, which really has pushed things forward very quickly.

So now today it's possible to sequence a human genome for a hundred dollars, say, and in a couple of hours using this next generation sequencing technology, basically technology at companies like Illumina. And that technology itself very heavily relies on the main artifact of a human genome project, which is the so-called reference genome. And so the argument here, the analogy connecting it back to AI is that while the human genome project lacked efficiency, it provided the groundwork for the foundation for subsequent solutions solving similar or the same problems that were significantly more efficient.

Evan Sparks: That's a really interesting perspective. And I definitely see your point, but I'd also argue that there are other interesting historical analogs and also trends that we're seeing in the marketplace that maybe suggests the other direction. So, from a historical perspective, I think about things like particle accelerators over the course of the 20th century, went from in the twenties, desktop particle accelerators that felt really, really expensive at the time to, towards the end of the century, we had projects like CERN taking up multiple kilometers of space and literally moving mountains to build these accelerators and solve some of humanity's most pressing challenges and our understanding of the universe.

So it's hard to know exactly where we are on that spectrum between desktop accelerators and moving mountains in Switzerland, but I think it's clear to me that the expansion of this technology is really improving the lives of many around the world. So in terms of specific trends that we're seeing with our customers, is this move towards unsupervised or semi-supervised or self supervised modeling where one of the key bottlenecks of driving these large models has been the amount of labeled data that's available to them. As we move beyond that paradigm, suddenly the amount of data that these models can be trained on goes up. The models get bigger. The amount of computation required gets much bigger. And so I think at least for the next few years, we're going to see a lot more, not a lot less, in terms of technology.

At the same time. There's great algorithmic advances for things like pruning and quantization, distillation to help shrink these models when we go to put them in production and we're learning as we go along the way here. So it's going to be really fascinating to see what the next few years hold.

Ameet Talwalkar: Yeah, I agree. I can see both sides, and I think there's a tension between things that should make things more efficient and things that are going to make the energy concerns just grow. And I think that kind of further emphasizes the need to, in the short term, really focus on this problem and execute on strategies to sort of mitigate concerns as soon as possible.

Evan Sparks: Thanks so much for that, Ameet, and really appreciate your time here. So for folks who want to learn more about these issues do you have any pointers for them?

Ameet Talwalkar: Yeah, absolutely. There's been a bunch of interesting discussion, various articles, and other resources online. And I think we're going to post some particularly interesting or relevant papers and resources on the screen. So I'd encourage you guys all to take a look at them. And that's including some of the work that we're doing here at HPE to address or think about these issues.

Evan Sparks: All right. Great speaking with you all.