THE ORIGINAL TECH AUNTIES

Computing Needs of AI and Its Impact on the Environment

Episode 7 · Technology Track · May 7, 2023

Building an AI system is not just a software problem — it's a hardware, architecture, networking, and energy problem. In this second conversation with Tony Foster, Gina Rosenthal and Angelia McFarland dig into what it actually takes to run AI at scale: what servers and GPUs are required, why your network can't be an afterthought, and what the carbon footprint of training a model really looks like. They also tackle the business question every organization faces: invest in infrastructure now, or wait for the next generation of more efficient hardware?

SPEAKERS

Angelia McFarland
Co-host; tech marketing veteran and founder, EOP Media

Gina Rosenthal
Co-host; technologist, product marketer, and founder, Digital Sunshine Solutions

Tony Foster
Guest; Senior Principal Technical Marketing Engineer, Dell Technologies; Adjunct Professor of Technology, Kansas State University. Find him at wondernerd.net or on LinkedIn and Twitter at @wonder_nerd.

TRANSCRIPT

ANGELIA MCFARLAND
Welcome to the Tech Aunties podcast, where we're bringing you industry context and vision from myself, Angelia McFarland, and Gina Rosenthal. On each podcast, we will share our marketing and technology industry experiences along with the team. Listen to us as we explain the past so you can have context to understand and create your own version of the future. So let's get into it.

GINA ROSENTHAL
Nice to see you again, Angelia, and nice to see you again, Tony. Today we're going to pick your brain some more about the computing needs of AI, but also the impact on the environment of those needs. But first, let me introduce you again. Our guest is Tony Foster — Senior Principal Technical Marketing Engineer at Dell Technologies, Adjunct Professor of Technology at Kansas State University, or as we like to call him, the Wonder Nerd. He describes himself as the VDI, EUC, and GPU fanatic bringing deep learning, machine learning, AI, and HPC to the virtual world. Did I get it right again?

TONY FOSTER
You got it right again. It's a mouthful. Let's just stick with Wonder Nerd. Keep it high level.

GINA ROSENTHAL
Last time we had you on, we had you explain what AI is. Today, what we want to dig into is the architecture needed to support one of these AI systems — and also the carbon footprint, the environmental impact, and other things we should be aware of. Are you up for that?

TONY FOSTER
I'm up for that. Let's go.

ANGELIA MCFARLAND
I'm up. Ready. Let's do it.

GINA ROSENTHAL
So let's level set with architecture. AI is not some magical futuristic thing that only a select few can understand. It's a technical domain with smart people building the algorithms, and equally smart people architecting the computers those algorithms are going to run on. What does that architecture look like?

TONY FOSTER
The architecture for AI is not that far off from what you'd expect for any high-performance computing environment. You need very powerful processors — and specifically for AI, you need GPUs. Graphics Processing Units are very good at doing the kind of math AI requires — lots of parallel calculations happening simultaneously. The more GPUs you can throw at a problem, the faster you can train a model and the faster you can run inference — actually using the trained model. You also need significant amounts of memory, both system memory and GPU memory. The bigger the model, the more memory you need. And you need very fast, very large storage. You're dealing with massive data sets, and you need to feed data to those GPUs fast enough to keep them busy. If you're starving your GPUs of data, you're wasting money.

GINA ROSENTHAL
I'm not saying you need to go bleeding edge or have the latest and greatest. But you definitely want to be running on newer equipment.

TONY FOSTER
I would slightly disagree. The technology moves so fast. If you have the money, get the best that's out there — because if you get something less powerful, the software will outpace the hardware within two years. You can buy 20 servers today, or you can take an opportunity loss and wait for the next generation — where maybe that shrinks down to 15 or 10 servers. What's that opportunity loss for your AI program? If you can run it on 20 systems today, go for it. If you can run it on fewer tomorrow and save on systems and operating costs, is the wait worth it? Those are the questions every organization has to answer.

GINA ROSENTHAL
Your architect needs to be talking to the data scientists to find out what's required — and that's not even that hard, because most neural networks have a map of what hardware you need to run them on.

TONY FOSTER
Correct. And it's not difficult.

GINA ROSENTHAL
Let's talk about networking. Most of our audience won't need to go deep on networking, but they have to have a clear understanding of where they need throughput.

TONY FOSTER
With AI, you need the throughput. That is one place you absolutely cannot skimp. You can't buy a switch at a big box store and expect to run your AI on a one-gig network. You can't even really do it on a ten-gig network anymore. You've got to have a solid network — because you're moving a lot of data around. If your AI spans more than one system, you're pulling data in and out of your data lake and sending it to multiple systems. All of that traffic is large. Today's modern data centers are about 25-gig networks at the bottom end. You'll see 40, 80, 100-gig networks. 100 gig is where a lot of AI runs right now. But 200 gig is fast approaching, and 400 gig is what the big supercomputer systems use — because that becomes the bottleneck. The systems are delivering answers in sub-seconds, and if they can't move data across the network fast enough, that's what increases your time to discovery.

ANGELIA MCFARLAND
Absolutely.

GINA ROSENTHAL
A lot of companies now have ESG — Environmental, Social, and Governance — requirements. What about the carbon footprint of AI? And does it vary at different stages of a model's life cycle?

TONY FOSTER
It does vary. With AI you typically have three areas. First is the development or training stage — and that's the biggest draw on power and resources. You're running nonstop: training, tweaking, running again, until you get the results you want. That uses a lot of power, a lot of resources, a lot of data — and water for cooling, depending on how you're cooling your data center. Next is the validation phase — shorter, and not quite as intense. You take new data and make sure the model returns valid results. Then you have the actual usage phase — where everything's trained and you're just taking new requests and running them. That's the least intensive. And a trained model can actually run on something about the size of a deck of cards. I'm holding up a Jetson Nano right now — it has a GPU on it, it's tiny, and a lot of models once trained and optimized can run on something that small. They become very energy efficient.

GINA ROSENTHAL
They have to be small because they put them in autonomous cars. I wanted to ask about CO_2 emissions estimates from training. GPT-3, to train it, took 502 tons of CO_2 emissions — compared to someone traveling from New York to San Francisco being about 110 pounds of CO_2. Is that kind of what you're seeing with training models in general?

TONY FOSTER

It's similar — though I'm not an emissions expert and everything I say here is generalized. They can be significant. And they can also be lower, depending on how you do things. The carbon impact of training models is more than just how much power it took to power the servers. You have to figure in cooling costs, the power used for the switches, the power for storage. If you do liquid cooling in your data center — where servers have liquid cooling packs that plug into ports on the rack and go out to a water condenser — you're more efficient, because it's a direct heat transfer as opposed to processors giving off heat into the air in the server room.

GINA ROSENTHAL

I think this was really good — these last couple of episodes with you, making this plain talk for people who need to work in this new space. Tony, one more time — where can people follow you?

TONY FOSTER

You can find me on LinkedIn at linkedin.com/in/wondernerd. You can find me on Twitter at @wonder_nerd. You can find me at wondernerd.net. Or if you're on the K-State campus, you can find me wandering around the computer labs.

GINA ROSENTHAL

All right, Angelia — this was great. Another one in the can.

ANGELIA MCFARLAND

Thank you, Tony. Thank you for joining us today on the Tech Aunties podcast. If you have a topic you would like us to cover, please connect with us on LinkedIn and Instagram. You can also find this episode and others at Tech Aunties dot com. Until next time, y'all be sweet.