SDS PODCAST EPISODE 921: NPUS VS GPUS VS CPUS FOR LOCAL AI WORKLOADS, WITH DELL'S ISH SHAH AND SHIRISH GUPTA

Jon Krohn:	00:00:00	Welcome to another episode of the SuperDataScience podcast. I'm your host, Jon Krohn. Today I've got not one but two guests for you named Ish and Shirish. I am not making that up. Both ish and ish are senior leaders at Dell, and not only are they extremely knowledgeable, this episode is packed with fascinating actionable facts about AI hardware considerations, including cloud versus local, and what workloads are best suited to CPUs versus GPUs versus NPU neural processing units. In addition to all that knowledge that they provide, they're also both very entertaining and play off each other in funny ways. This episode is critical listening for anyone working on training or deploying ai. Enjoy this episode of SuperDataScience is made possible by AWS and the Open Data Science Conference.
	00:00:47	Welcome to the SuperDataScience Podcast. I am joined by two people today, so I've got Ish and Shirish, so a lot of isness happening today. I'm that actually
Ish Shah:	00:01:03	That's mine now. You can't have that!
Jon Krohn:	00:01:08	Yeah, and that's ish speaking right there Ish. Described Ish as his more flamboyant sounding version. So if that helps you distinguish their voices through this episode, we can try that. I'm also going to have them introduce themselves briefly to give you a bit of a sense of who they are. So some listeners may actually already be familiar with Shirish was in episode 877 of this podcast Shirish. Welcome back to the show. Where are you calling in from today?
Shirish Gupta:	00:01:37	Great to be back, Jon. I am calling from our headquarters in Round Rock, Texas.
Jon Krohn:	00:01:42	Yeah, Dell headquarters in Round Rock, Texas. And so you're a director of product management at Dell. You've been there for like 22 years or something like that

Shirish Gupta:	00:01:52	Almost.
Jon Krohn:	00:01:53	Wow. And basically in Round Rock that whole time.
Shirish Gupta:	00:01:56	Yes, I've been in the Austin area for the entire duration.
Jon Krohn:	00:02:00	Wow, wow, wow, wow. Nice stability there. Ish, you have not quite been there as long you've been at Dell for four years. You have a title that honestly, I don't understand what it means you're going
Ish Shah:	00:02:12	To have to know most days. I don't either. It's okay.
Jon Krohn:	00:02:16	Head of CTO pursuits.
Ish Shah:	00:02:18	Yes, sir.
Jon Krohn:	00:02:19	Our client devices and AI.
Ish Shah:	00:02:21	It's a really fun way of saying I get to work with people who very fortunately for me are way smarter than I am. This is a really good example. I like to joke that Shirish is the evolved Pokemon version of Ish and I think that that's true for pretty much everyone that I work with out of the CTO pursuits team because it's about getting in the field forward, deployed engineering style, roll up your sleeves, find a problem, help someone solve it, kind of go beyond the whole, we're going to sell you some laptops, vibes that maybe some people know Del four. We've got this whole new function that we've stood up and there's some really, really talented people helping solve some really, really complicated problems. So it's been a lot of fun for me.
Jon Krohn:	00:03:02	Nice. Well, I think you're downplaying it a little bit. You have achieved some things in the past that make me think that there's some intellect happening around on your side there ish. So you have an MBA from MIT and you were a consultant at BCG before being at Dell.

Ish Shah: 00:03:21 These are all my dark dirty laundry things. I don't bring up Jon here. Here's the consultant for you all. I just saw the viewership number crashed a zero. It's okay. I love saying I adored my time at Sloan, but very much kind of our fake MIT sometimes I was schooled frequently by some of our visitors from the other courses as majors are called at MIT. So yeah, it was great. I hope my mom listens to this episode. I think I'm going to get some kudos I very sorely deserve.

Jon Krohn: 00:03:51 Nice. All right, well, so let's get into, now that our listeners are hopefully familiar with the voices of Ish and Shirish, we can get into the content of today's episode. So Dell famously makes PCs, it's one of the things that if not the thing that Dell is probably best known for. And so the first question that I would like to kick things off with is for our listeners who are interested in data science, AI applications, a lot of open source, the first thing that might come to a lot of listeners' mind is Unix-based systems. And so why should somebody be considering a PC instead in particularly maybe even a Windows machine?

Shirish Gupta: 00:04:36 Yeah, well, let me go first and Ish will have his take on it as well. I would say that you have to start with what are your objectives? If you look at the data science and well the software development community at large, let's just look at the facts. Windows is still the most popular OS for software devs, right? It's about 64% the last time I checked of software devs use Windows for development, which is still the leading number out of all of the OS platforms. It is, yes, the standard in the enterprise and for consumers on PCs. So it's familiar, it's friendly, user-friendly, great for beginners in the data science field. It's compatible with popular apps who doesn't want to use productivity apps? There's other data manipulation and visualization apps that run better on Windows. And then

at the end of the day, as I said, it's a standard in the enterprise.

00:05:39 So if I'm thinking from an IT management perspective, you get the enterprise security integrations are much more easy with Windows, but that's not it because let's not kid ourselves. I think if you look at large ML and data science deployments, I think 96% of them are still running on Linux-based or Unix-based servers. So if you are doing large deployments, Linux is probably still the platform that you want to be developing in because best practices you want to be developing on the platform that is being used for production environments. But I would say at the same time, there's an alternative. There's Windows subsystem for Linux and WSL two specifically I think starts closing the gap in all of those scenarios. You can run a Linux kernel directly on Windows. So if you have Ubuntu running natively, guess what? You now have all of your Linux command line tools that you can run right there on Windows.

00:06:49 And so you have best of both worlds. You have all your productivity and other ease of use benefits on Windows, and you don't lose all of the benefits you get working with Linux for data science solutions. So I think that is something that's worth considering. I think wsls use is still pretty small out there today, but that's an alternative for someone who's on Linux used to Linux and wanting to bridge the gap. Then last but not least, this is something I touched on briefly in our last episode, Jon. If you are building for PCs, if you're building apps for PCs, software developers, it goes back to the same best practice you want to be building in the environment that you are going to have your apps in production you want to be building in Windows, right? Let's look at the number 60 million PCs with Linux, 1.6 billion PCs with Windows that's less than 3% Linux and you get all your IDs vs code, visual, studio pie, charm, IntelliJ idea, et cetera, that all run

natively on Windows. So it sounds like I'm making a case for Windows here. I'm not a Windows, I mean not a Microsoft employee, but I have to say at the end of the day it comes down to what you want to do. There's a place for Linux for data scientists and there's a place for Windows.

Jon Krohn: 00:08:16 Yeah, that was really well argued Shirish. Thank you. It seems like I went to Oxford University for my PhD and they have something called the Oxford Union there, which is a debate club, and I went and watched some of their debates and that sounded like opening remarks. You have all the stats, very compelling argument. Ish, what do you have to say? Are you on the other side of the argument?

Ish Shah: 00:08:40 I think mine is probably a lot less articulate, basically poor K, no less dose. It's why pick one thing when you don't have to pick one thing, right? With WSL, it's like what do you need? Just kick into it. There is no dual boot. It's sort of in the same space and it's a very fluid transition back and forth. You can bop around in WSL at Windows as you need, and I'm using very technical terms here, but I think the biggest thing is you can hit a nail with a wrench, but why? If you have a hammer, right, you've got the right tool for the right job at the right time, and I think that's going to be a theme in this conversation, Jon. It's Dell's whole point of value in this whole brave new world we live in is choice. It's like you get to decide what you want to do. You want to throw Linux on your device, do it. You want to do Windows on your device, do it. You want both, do that too. Whatever works for you.

Jon Krohn: 00:09:37 Awesome. Speaking of that idea of being able to have your cake and eat it too, I'm going to butcher the Spanish that you so well said there. K,

Ish Shah: 00:09:48 K noles dose.

Jon Krohn: 00:09:49 Why not both K, no less dose. Yes. And so on the theme of that, I'd like to talk about how your machines, Dell machines are so supportive of all different kinds of hardware. So not just CPUs, not just GPUs, but npu as well, neural processing units and allowing all three of these to work together. So Shirish, in your previous episode on my show in episode 877, you began with a really brilliant explanation of what NPUs neural processing units are since that's something that maybe listeners, if they haven't listened to that episode, they should probably get at least a short intro to that. I'd love to hear a bit about PU and then maybe we can also talk about CPUs, GPUs, and the different kinds of situations where you would use those. We'll get to that next. Let's just start with what pus are.

Shirish Gupta: 00:10:45 Yeah, so I do encourage everyone to look at or listen to episode 877 where I do cover that in detail as Jon said, but for those who are coming into this cold, an NPU stands for a neural processing unit and it's integrated into, actually I'll step back. There's two ways you can have access to an NPU. You can have it either integrated into the chip set or it can be a discreet card that you either put into a slot in the PC or in a built into a mobile device, but it's still a discreet card. So you have both flavors of NPUs.

00:11:26 At the end of the day, what sets them apart from the architectures that we've had for maybe more than a decade at least, is that they're purpose built to handle AI and ML workloads. So that vector math, that is such a foundational aspect of neural networks. The NPU architectures are almost hardcoded to handle those workloads. So what you get as a result of that is a much more efficient processing engine, especially important for a mobile device where battery life is pretty important. The last thing you want is these new workloads that are running on the PC and they tank your battery life. If your

battery only lasts half of what it normally is, that's a no-go, no bueno, right? So NPUs fulfill that. It's performance per what is what you get from the NPU for AI ML processing.

Jon Krohn: 00:12:33 So it's basically it's optimized for the kinds of matrix multiplication operations, neural network architectures that are so common in everything transformer architectures. And so therefore all of the large language models, most of the AI capabilities that we have today can run more efficiently and especially as you said there in a more conservational power way, a more efficient way on mobile devices. But I mean even on a laptop, even on a desktop, a server, I'm sure that those improved efficiencies make everything run along more smoothly.

Ish Shah: 00:13:17 I think if you go back to the COMPSCI 1 0 1, think about the logic gates, think about the levels of abstraction and think about how a computer does what it does. How does a computer compute, we've been having these debates around X 86 and arm and even risk, the whole risk five project, the way that these things flow through the transistors on the device, how are you organizing those transistors on the device and on the chip? And this is the natural evolution of, Hey, I've got a CPU, it's really good at this. I've got a GPU, it's really good at this. Now there's this new thing and this new thing is an AI workload and the way that that runs through those transistors is different than the rest of the stuff. And I know there's going to be people in the comments being like, oh, it's just a current through a thing.

00:14:11 Like yes, we know, but you understand the point, which is that there is a correct tool for a particular job at a particular time. That's not to say that GPUs and CPUs can't run AI workloads. Of course they do and they do it tremendously well. And there are entire organizations and companies and startups dedicated to how can I get an AI

workload inference really, really well on the CPU? How can I get it to run more power efficiently on the GPU? And then there are people who will skip that process entirely and say, well, if I've got a purpose built chip. And the reason ish paused when he was describing what an NBU was is because Jon, the last time he talked to you, he couldn't tell you something is now true, which is that Dell has announced a device that has a discreet NPU in it.

00:14:58 So it's not just on the SOC and it is a big honking thing that we put into a laptop chassis just because we could. But I think that right now what used to be a very easy matrix and a very easy matrix for data scientists to understand it, decision makers to understand it, buyers to understand two by two, I've got this many companies making this many things and I can pick from this matrix that matrix has gone from two by two to like eight by eight and now there's this who needs what tool for what purpose? And it's a really complicated question, whereas two years ago it was like, well, what's the newest thing on the menu? I'll take that. It's not so simple anymore.

Jon Krohn: 00:15:41 Right. Okay, so this sounds pretty exciting what you're talking about there with having a discreet NPU, correct me if I completely butcher the way I'm explaining this, but so a discrete NPU on a laptop. And so if somebody wants to get that, how do they find that? How do they search for that? I can have it in the show notes. This is something that Dell sells that they could just buy. And why specifically you talk about this eight by eight matrix, why should a listener be interested in getting a laptop with a standalone with a discreet NPU?

Ish Shah: 00:16:13 Yeah, I'll tell you that the most requested device out of our CTO team right now for their next device refresh within Dell, everyone is on our boss about, I want one of those. And he's like, yeah, we really got to think about how many of these we could get for you guys, but it's the

Del Pro Max Mobile workstation and it has a discrete NPU device coming to it, and the CPU is also powered by an intel chip set. So it's a very interesting machine and it's not for everyone listening to this podcast. Well, actually maybe this is a pretty biased sample. Maybe everybody listening to this podcast is going to want one and we'll find a use for it. But think of these specialty devices, right? Think about the GB 10 and GB 300 coming out from Nvidia. Not too long in the future when we launch this Del Pro Max, it's another beast and it's like the right person will want that versus something else.

00:17:09 A very cool demo that we ran on this device at Dell Tech world not too long ago. We took a model that Northwestern Medicine had fine tuned and trained. It's their Aries model, which does imaging, CT scans, x-rays all done in-house. By the way, Dr. Mozzie and his team over there shout out to them because they're the true mad scientists. And it was a demo that we worked on with them and we ran inference on a colonoscopy video, which we blurred out, but a colonoscopy video that inference live like in front of you on that device, no internet uplink needed, no queuing, no waiting at a server somewhere to be processed and to be shipped back. So the question of who would want one of these things, hey, if you've got this sort of specialized use case or say you live in a part of the world where there are restrictions on what you can and can't send to the cloud and in which circumstances you can, this kind of device is going to change the game. It has that level of performance and speed that you just six months ago, 12 months ago, the last time shish was on this podcast, you didn't have that option, right? So it's coming out soon, not out yet, but yes, if you look for the Del Pro Max with discreet NPU, you will see the press about it.

Jon Krohn: 00:18:24 And sorry, just really quickly Shirish, the name was Depro Max something workstation, what was it?

Shirish Gupta:	00:18:29	So it'll be on the depro max plus devices. It will be, as they said, it will be available in the second half of, well, I should say Q4 now at this point.
Jon Krohn:	00:18:42	I mean both are true
Shirish Gupta:	00:18:44	More to come.
Jon Krohn:	00:18:45	Okay. Okay, that's cool. Alright, so I might not exactly be able to include a link in the show notes. I'm not sure I'll do my best, but regardless, people listening in the future will be able to either find it by looking it up or they will be able to find it soon. Cool. And sorry, Shirish, I interrupted you please.
Shirish Gupta:	00:19:03	I was going to say that NH can correct me here, but what's incredible about it is that we could also run 109 billion parameter LAMA scout speculative decoding FP six, FP 16 on this card, which is insane. So think of that running locally and the things you can do with it. This really opens up a lot of use cases that can become viable for on-device inferencing.
Jon Krohn:	00:19:34	Does this mean, are you able to give an estimate of the number of parameters of an LLM you'd be able to fit on there?
Shirish Gupta:	00:19:40	Yeah. Yes. 109 billion parameter LAMA four scout. Right, but that's a speculative decoding model at FP 16 cloud native resolution.
Jon Krohn:	00:19:49	I see. And ish, it sounded like you had something to add.
Ish Shah:	00:19:51	No, only that. Yeah, it's a pretty big honking model to put on a laptop.
Jon Krohn:	00:19:56	Oh yeah.

Ish Shah:	00:19:57	Again, me taking a very sophisticated comment from fruition, distilling it down to base components here, this is bigger than anything you've played around with in anything LLM or LM studio or llama to date. This is a today you just call it kind of a cloud quality model, right?
Jon Krohn:	00:20:14	Yeah. Not long ago, I remember sitting in meetings with data science teams that I manage where I would be trying to encourage them to be thinking about using 13 billion parameter Llama models for example, because that was about as large as we could fit on a GPU. And so that's pretty wild to be thinking about something that's almost 10 x that size running on a laptop
Ish Shah:	00:20:37	And more to come. This is sort of the progress that's been made in 2025 alone, and it's just staggering to see both the models getting smaller and the hardware getting more capable. Those two things are converging on each other and it's not going to be one or the other. They're both going to happen together.
Jon Krohn:	00:20:59	Very exciting. So you've mentioned, we've now talked about NPUs in some detail. Now you've mentioned GPUs. What is the key difference between A GPU and an NPU? It sounds like today a lot of people would be using a GPU for all the workloads that you've been describing. We should actually be using an NPU for
Ish Shah:	00:21:22	GPUs, integrated and discreet, right, Shirish?
Shirish Gupta:	00:21:25	Yeah. So I'd say that GPUs are farm over versatile today because ultimately the best at parallel processing. And so by definition, they're very performant at AI and ML workloads and we all know that because that's before the advent of the NPUs. That is the accelerator of primary value when it came to AI and ml. The point I was going to make is what really sets GPUs apart is their ability to scale and performance, right? Relative to NPUs today, if

you go back to what ish was talking about earlier, also the GB 10 and the GB 300 that have the NVIDIA GPUs that are also coming out later this year, there's some pretty staggering capabilities within those boxes. So like the GB 10 for example, is going to be capable of up to a 200 billion parameter model that can be fine tuned locally, which is game changing for your data science community.

00:22:32 Now you don't have to set up a whole cloud instance. You have this appliance and your work group in the lab and you're on business. You can really play with some of the frontier models. So that's pretty game changing and that's fairly accessible from a price standpoint point. I won't go into the details, but it's definitely more affordable than some of the alternatives. And then if you really want to scale that up, you have the GB 300, which gets you up to 500 billion parameter models at 20 petta flops, but that's at a much different price point. So again, it comes down to what you want to do, but GPUs today have far more scale in terms of performance than npu. Npu are still evolving. I think we are talking about first couple of generations. So as I said, they're going to get better. And the differentiator for NPUs is performance per what? GPUs are not that conscious when it comes to power consumption. Let me say,

Ish Shah: 00:23:34 And that's a really nuanced difference for GPUs right now. And we're talking about client devices here. There is a whole aspect to Dell that is in the server land, it lives up in the cloud on-prem hyperscale, whatever you want to do, but on client devices, the highest ceiling right now, GPU on client devices, the best efficiency right now CPU or NPU, depending on your use case, right? Speed also depends what kind of model are you running also depends, and this is that matrix I was talking about, Jon, it went from fairly evident what you needed to do to, Hey, I'm now thinking about buying a device here in August of

2025 for September of 2025, and I'm trying to figure out what the commitment of this device is going to be over four or five years. I mean you can really see that lacking something with an NPU in it or lacking something with a discreet GPU in it. You may come to realize that you're going to need to upgrade sooner rather than later if you don't make that investment now. And shish and I are not part of sales. So you buy one laptop, you buy a million doesn't really impact us, but that's the truth and that's what I'd think about if I were making a purchase right now.

Jon Krohn: 00:25:00 So let's talk about that next. In terms of the kinds of things that people should be looking for if they want to be future-proofing for the next five years, what are the kinds of parameters? Let's go over an eight by eight matrix in an audio only podcast, but just kind of generally, let's talk about the kinds of things that people should be looking for in hardware that they're buying today. And I guess as you said, this is specifically about what you described as client devices and so I'm assuming that isn't a term that I use in my kind of day-to-day language, but it seems to me like that's distinguishing against servers. It's like laptops desktops.

Ish Shah: 00:25:39 Yeah. Most normal people aren't running around saying client devices. That is a very Dell kind of term when you think about what to buy right now, if I were starting college or if I were doing something, wow, God, that was a while back. If I was starting college today, thinking about what kind of thing do I need and there are different brands and different price points and different pursuits that you would have with this device. What's it going to be used for? Yeah, I think an NPU makes a lot of sense for a lot of knowledge work type work. And if for no other reason than to get the most out of your operating system, we know from our friends in Redmond that windows is going to start baking AI features into itself that it intends

to run on the device. This stuff is expensive to ship to the cloud and back every single time.

00:26:29 So some of this stuff like background blur on a Microsoft teams call speech to text, all of this stuff is going to look for a home somewhere on your device. And guess what? CPUs, the workload hasn't gone anywhere that CPU is still going to have to do all, it's the workhorse, it's still going to have to do all the things. It's always done. And now if you don't have an NPU or A GPU, it's also going to have to support this new kind of workload. So that's one thing to keep in mind where if you decide no NPU no GPU, well gosh, your CPU better have some slack in it, it better have some bandwidth. GPUs, I like to talk about the birth of a new persona and persona. Again being a word that people in our world think a lot about the data scientist persona is something that an IT decision maker is constantly thinking about.

00:27:19 What does that persona need? And you really have the birth of a new persona with all this AI stuff because you have people like myself who are not formally trained in that way as engineers but who know enough to be dangerous. And now with the right kind of device, I get supercharged and with the wrong kind of device I get throttle. So this is very much a productivity gains question and that is sometimes really hard to quantify. So knowledge workers NP makes a lot of sense. Knowledge worker plus maybe like these new persona at the edge of a dev and a kind of regular knowledge worker that's me. And I would ask for something like a discreet GPU because I know that's going to last me. And also if you want a device that you can use to train AI workloads during the day and give to your kid to play Fortnite later, like GPU is probably the way to go. So there's a dual use argument to be made there

Shirish Gupta: 00:28:18 And just to add to what Ish is saying, I would classify them today, and again, you have to keep in mind this is rapidly evolving, but today it's I could classify devices into three categories. You have the essential AI PCs which have what I, for lack of another moniker, call them entry level NPUs, right? Think 10 to 15 tops or trillions of operations per second. And those are great for basic workloads coming from, as I alluded to earlier, your background blur, your voice correction and other optimizations offloading that from the CPU so you have a much better experience and they can accommodate smaller models like up to maybe one to 3 billion parameters, but once you get there, now you're bringing workloads back to the CPU if you go beyond it. So that's probably the limit there. Then the second category is maybe slightly more advanced IPCs with more performant NPUs or state-of-the-art NPUs today that's about 40 to 50 tops and that really brings on device AI into focus right now.

00:29:30 You can actually bring custom workloads perhaps run up to nine to 10 billion parameter models for custom in workflow embedded use cases across a variety of verticals in addition to the copilot plus features which run locally on your PC that is talked about. So this is again a very nuanced difference here. Microsoft's copilot branding refers to everything that runs in M 365 in Azure, so that's all cloud-based, subscription based largely that's their copilot brand. Copilot plus is everything that runs locally as part of the OS itself. It's part of Windows no extra charge and as he said is going to continue and Microsoft is going to continue to add more and more capabilities that run locally on the pc. So for you to harness those capabilities and not lock yourself out of those capabilities in the future, you definitely want a pc, an AI PC with at least 40 tops on the NPU today.

00:30:44 That is my recommendation for the knowledge workers and the most common use cases. And then the third one, it's kind of self-explanatory, now it's your high performance pc. Those have your high-end CPUs from the CPU suppliers, which are much capable of much more performance, single and multithreaded processes. And then you have those augmented with discrete GPUs and discreet NPUs. Now you're talking about the persona that is talked about is you're starting to create that separation between your power users, your AI and ML and data scientists that can really now do data crunching and work with models right there on the device itself. So that's the three pronged categorization today. Forg,

Ish Shah: 00:31:37 I was the recovering consultant and here is Shirish with his three buckets, right? BCG would be proud. One thing Jon, I want to add to that is we're not a walking infomercial here and I know there's a big corner of the internet, let's be real for a second. That's like, Hey, I watched the NPU advertisement in the Super Bowl. I watched the copilot plus PC ad with the zebras and the scientists in the forest, but really what does it mean to me? And again, it's about this temporal mismatch. How long are you going to use this device, right? Oh, I'm skeptical of the features that this particular company is building. I'm never going to use any of those. Again, think about the future, think about the things that are happening at breakneck speed, breakneck pace. That's what you have to be thinking about, that temporal mismatch. So even if it's not up to your taste in this moment, there's something bigger to consider. And again, it's not about an infomercial. These are just the things that I would be thinking about if I were buying one device or if I was buying a million.

Jon Krohn: 00:32:45 For sure. It's an interesting situation here because you guys obviously do both work at Dell and so it's easy to feel like, but simultaneously everything that you have

been saying so far is useful to me as somebody thinking about on the show. We have a lot of different kinds of episodes on different kinds of topics. A lot of the time we're covering open source software, but you need your software to run on something and there's no open source hardware. You can't just download free hardware. Yeah, there's no GitHub going and have all you can go do. Exactly. So by its nature it has to be in some way kind of a bit of a commercial conversation, but the hardware is something that you need for any of the work that we're doing, any computing. So speaking of software and the experience of using these devices, so if we're talking about having NPUs on our machine, having GPUs on our machine CPUs, and actually I want to dig into CPUs in a bit briefly as well, so I'm going to get your subconscious thinking about CPUs in general, but when we have, I think it's pretty easy, at least in terms of utilization, the CPU is the most general of all of these devices and so I think we're used to probably most listeners are aware that CPUs are doing the most kind of general work running your operating system.

00:34:16 And what I'm trying to get to here is it sounded pretty obvious from what you're describing that in the future or maybe even today, Microsoft will have automatic support when you're using say Windows for NPU. So you're talking about copilot plus features, text to speech, any of these kinds of things. They'll look for the best device available for running those and if there's an NPU on the machine, it gets sent there. So I get that for the kind of click and point operating system experience. But if I'm a data scientist, if I'm a software developer, if I'm an AI engineer, then what do I need to do to engage those resources? What is my experience? What software do I use to get access to NPUs GPUs or be making the decision? How does all that work?

Ish Shah: 00:35:05 Yeah, multi-part answer. I think here the first part answer is it's nothing you've not done before in some ways GPU acceleration has been around for some time, right? You mentioned all the uses for the CPU. Yeah, I know of all of you in your a hundred tabs that you refuse to clean and close. We know about you at Dell, we know you require some CPU juice, but GPU acceleration spin around solvers like Guro have GPU support, CAD software, GPU accelerated. So the sort of tried and true methods of acceleration, the abstractions are sort of similar and then you're going to start to hear about all kinds of funny name stuff and it's really going to strike you as like, well, what is all this? You've got llama CPP, you've got O LIama, you've got sort of, to a lesser extent, CUDA has been around, but you've got this Cuda X layer now that sits in the middle.

00:36:08 All of these things combine to have this plethora of stuff and maybe not all of it is deployable at scale in an enterprise environment. If you are a data scientist, sure, go pull this. Go pull Lama CPP and run a guff model. Fine, you know how to do this. You will know how to do this. If all of these words are alien to you, six months from now, maybe Jon will invite us back and they won't be as alien to people anymore. But the work that the folks at sloth are doing, the work that folks like the bloke are doing, and these are all names you see on the hugging face boards of these models getting rolled out to land. Those on particular pieces of silicon requires that software layer. And if it's not a software layer, it requires sort of this abstracted API, Dell and Jon you opened this with you are a devices company. That is what you are known for and we know that that's our identity to use the device to make the most of it. We've created something called pro AI Studio. It's a thing that lets you not have to worry about this as much if you are the developer of an application. I'll let shish talk a little bit more about studio.

Shirish Gupta: 00:37:24 Absolutely. So I think that was very insightful Ish. I learned something from that too. So that was a great monologue there. I wanted to add one more thing.

Ish Shah: 00:37:36 Oh god,

Shirish Gupta: 00:37:37 Sorry, we cut that. Experian, we got to cut that. We are

Jon Krohn: 00:37:43 Not cutting that. That was perfect. Absolutely. That is in the episode, Mario. Do not listen to the guests.

Shirish Gupta: 00:37:52 Yeah, I just wanted to add that the complexity of being able to bring your AI workloads and land them on the PC accelerators is not a trivial task, right? H painted a really vivid picture there and it's not for the faint of heart today. So if you think about someone who's beginning their journey as an AI engineer or who's trying to run AI locally on a pc, whether it's the GPU, the CPO, the NPU, they have to make sure that they have the right format of the model and if it's not the format, they have to have the right tool chain to convert it to the right format. That conversion process is different for every target CPU architecture and IHV, if it's Intel or a MD or Qualcomm or Nvidia, you are looking at a completely different tool chain to go make that conversion and then that it proliferates from there because you're talking about one model right now that's not how models are. Typically there are model families. So if you want to make your entire model family of say, LIama available and you want to quantize them or do Lauras or something, now you're really creating some tremendous, you're creating a huge tree now, massive branches. So that's a task that is daunting even for experienced practitioners. So with that said, that's why Del Pro Air Studio became almost an imperative for us to create and we focused on abstracting away all of that complexity.

00:39:41 The focus being how can we democratize access to AI on the device for the broader development community. So you don't have to be working directly with Intel Open VIO and be six months into your journey there before you can actually land something on the NPU because that's what it takes today and more on pro AI studio, that's the paradigm. Reduce complexity go faster. I'll talk about some of the speeds that we've accomplished. So we did A POC with a partner of ours, Deloitte, and there's a white paper coming out on this pretty soon. We did a comparison of a use case that was built for on-device AI and they'd already worked on it before they used Del Pro AI studio and then we normalized it to a team that was just beginning their AI journey. And so we kind of looked at it three ways and what we learned was for the team that was just starting that journey with say, Intel Open vio, it took them about three months to actually go from scratch to have an app that is fully integrated with a model that's actually running on the Intel silicon. And it took the same team with Del Pro as Studio about four days to do that.

00:41:15 And that's not like a linear simplification of time. That temporal concept is really, we just eliminated that engineering complexity for them. So that's the second reason why we built it, right? We wanted to democratize access for the developers at large to bring AI workloads to RPCs and we wanted to make it quick and easy for them to land it on a variety of silicon. And last but not the least, we also put in, excuse me, an API that actually is following the OpenAI defacto API standard for hosts. And so if you are running a web app that is pointed to either an on-prem workload or even a cloud hyperscaler workload today, but that you're using that open AI defacto spec, you can just make a quick one line configuration change, point it to the local server on a Dell machine and you are done. It is now running on a local model that's running on your device accelerator. So that's

a long-winded explanation for what Pro AI Studio is, but hopefully that resonates with...

Jon Krohn: 00:42:45 It all. Sounds great. One thing that I think might be helpful to me and to our listeners is to understand, I realize that as you mentioned there are lots of kinds of personas out there, but is there a user story that you could describe for us so that we can visualize, so that we can imagine what it's like for us to use depro AI studio? I think I get the functionality. It allows us to take advantage of all these powerful devices for whatever kind of AI workload we have, but what does it look and feel like

Ish Shah: 00:43:15 In terms of real life application And this, Jon, your question is also super salient because it's also about the bigger conversation around why would you ever run a workload locally? Why would you ever do that to begin with as a data scientist, as a knowledge worker, as anybody, why would you ever do that? Offline mode comes to mind. If you want your AI based application to continue working, even when the person's on an airplane and has a lousy connection, guess what? You're going to want to run that workload locally. Reasons related to speed, cost, security, connectivity, all of these signals are the reasons why one might run a workload locally. And examples of those Jon models are not apps and apps are not models. The data scientist universe knows this, right? Credit card fraud detection is the objective. Within that objective, there are many models doing different things.

00:44:16 It's the same for local apps today. There are a lot of them like LM Studio, anything LLM that make it easy and put a gooey on top of, I'm going to go in and I'm going to do AI work locally on my machine, look through the app, abstract away the app and start to think about you as a data scientist. What is the purpose of the thing you are building? That's what the GUI is going to look like on the backend. The user doesn't care. Is it going up to the

cloud? Is it going down in my silicon? The user literally doesn't care. They're going to care about the performance and they're going to notice that there's a difference. Yes, but that's why there's sort of this routing objective around when you run workloads locally and that's why Dell Pro AI Studio makes it as simple as a one line config change.

00:45:06 Instead of sending the workload up, you send it down, you just change where it's pointed at. And that's the whole point of what we've made. Ish mentioned open VIO a couple of times. Del Pro AI Studio is an extension of Open vio, open vios in there. We worked with Intel to bring this thing to life and to make it easy to use and as are the open VIO twins from other people who make pieces of silicon. So the whole purpose of defining the use case, it's not about running something locally just to say you did, although I enjoy that and I spend a lot of weekends doing it, right? It's also about when and why Does this make sense? Am I forcing this down someone's throat because I want to or is there genuinely a reason this data related workload, this AI related workload should happen on the device?

Jon Krohn: 00:45:59 You talked about open vno there and working on Intel with it, which is a great segue into a topic that I wanted to make sure we covered in this episode. So the Dell IPCs that we've been talking about, the hardware we've been talking about in this episode, we've talked about NPUs GPUs that are available on them, and I already alluded to this. I said that you should get your subconscious going. We've talked about the purpose of CPUs as being relatively general purpose compared to NPUs and GPUs, but I wanted to specifically because we have experts like you on the show, I don't actually know very much about CPUs or considerations about them. So for example, Intel core ultra processors that were released in 2024 last year, there's this lunar lake architecture. What is the

significance of that architecture, which I think is the standard now on IPCs? What is the significance of that architecture for my listeners for the data science persona?

Shirish Gupta: 00:46:59 Yeah, so I'll take that one a little bit about Lunar Lake. It was quite a different architecture in which you had memory on chip, so quite unique from Intel. With that integrated memory, you actually had very fast transfer rates from processor to memory MOC. They are the first Intel architecture that supports 40 plus tops on the NPU. So if you're looking for a copilot plus PC from Intel with an Intel processor, you're looking for lunar lake. That's kind of the highest level, right? Apart from everything else, well, it's still X 86 based, so there's nothing different there, nothing unique or different. It's not arm that was the biggest difference. The MOC is what allows for tremendous performance gains, and I will mention that the IGPU, the integrated GPU also achieve some step function performance gains wherein they're actually rivaling or I would say exceeding the performance of some of the entry level discrete graphics cards.

Jon Krohn: 00:48:17 Very nice. That was a clear definition as we've come to expect from you ish. And now I do feel like I understand the significance of lunar lake architectures in particular, the kind of step gain efficiency improvements that you were describing there, like memory on chip. I want to kind of bring all this together. We've talked about CPUs now most recently, obviously GPUs, npu earlier in the show, and we've also been focused on what you call client devices, local processing laptops, desktops. Let's talk about why all of this is significant in the gen AI era. So for example, at the time of recording not too long ago, a couple of weeks ago, open AI released open source LLMs for the first time in a while. Earlier in this episode we talked about llama models. So what are the kinds of things that gives us the capability to be taking these open

source models where you're talking earlier in this episode about laptops that can have 109 billion parameter models on them. You're talking about desktop appliances like the GB 10 that could have something even double that size running locally plugged into your machine. And so in this brave new world where these kinds of gigantic, highly capable open source models are available where we can be downloading them onto local advices, what are the kinds of circumstances where you'd recommend that versus doing something in the cloud

Ish Shah: 00:49:50 Use case dependent? And I know that's a frustrating answer because it's the same question that a lot of Dell customers ask us and to come in and to tell them that here's the formula you're going to use, I wouldn't be being honest with them because when I decide to rut something locally, it's because I've analyzed the situation, determined what it is I wanted to do and picked local as the thing that makes most sense. Let's say I'm working on something that contains a lot of personal information that I don't feel comfortable putting into a cloud-based AI tool. Or let's say that I am subscribed to a cloud-based AI tool, but it's close to the end of the month. I've burned a lot of my tokens and maybe I don't have many of them left. All of those are reasons why I would turn to something like what you mentioned, Jon, the G-P-T-O-S-S family of models.

00:50:39 I think it's super telling that a company like OpenAI, which has to be very intentional about what it does and where it spends, its very valuable human capital resources shows as part of their open source launch to have one of the models be a smaller model. I think that is really telling because it tells me that in this agentic world that's coming where it's not just chat back and forth anymore, it's like by the time you have your AI tool write your 50th poem, you're kind of over it and it's like, alright, data scientists or otherwise, what is the applied

practical use of this tool code gen, creating little applets to help you do the things that you do on a day-to-day basis and it's not something you want to go subscribe to. Local is a phenomenal application of that. So determining that I'm going to do my code gen locally versus in the cloud is a question of how much do I want to spend on it?

00:51:39 How fast do I need it to be? How capable do I need it to be? For me, a person like me, that local use case makes a lot of sense. I'm not burning through tokens and watching my credit card. So it is a frustrating answer because it's about choice and that agency that you have to decide where to run what and when. All the way at the top of this conversation, Jon, we talked about Windows versus Unix versus Linux versus et cetera, et cetera, et cetera. Choice has remained the theme here, choice of chipset, choice of operating system, choice of local versus cloud. There will be a right tool in a right method at the right time. So I know that that's not the answer everyone wants and it's always met with this crest falling look in the room when we talk to our customers because it's like, I thought you were going to tell me what the answer was. The answer is that this is hard. We got to do the work.

Jon Krohn: 00:52:36 There were some clues in there though. Things like data privacy edge you in the direction of not doing things on the edge.

Shirish Gupta: 00:52:45 I'll let us spin to it. I think the other thing to consider is apart from I was going to bring up privacy and you nailed it, Jon, that's an important consideration. We're seeing more and more of that concern like sovereign ai, enterprise ai, they're almost anonymous now where customers, they're trying to get to initial value as quickly as possible to start their AI journey. They want to show outcomes, show value, and sometimes the answer is let's just go to the frontier models in a hyperscaler environment because the easiest way, I don't have to

scale up my team, I don't have to invest a whole lot and increase my time horizon. I can get there quickly, but they're also trading off control. They're trading off, they're at risk of getting locked in to the ecosystem and they're also trading off security and privacy of data, especially ip, which is starting to become a big deal. And we hear about policy decisions being made across the world that are changing things or putting guardrails on what governments can cannot do. And then you have of course private enterprises that they have their own sets of concerns and the regulatory compliance requirements that they have to adhere to. So I do think that from that perspective, if you think about cloud versus local, to me the answer is hybrid. Really the future is hybrid because it is not practical or feasible to move all workloads to the pc.

00:54:41 And at the same time you want optionality because you don't want to be locked into just one type of node. You want to be able to run the right workload on the right compute engine at the right time. So I foresee there being intelligence in an agent future where there are either basic ML classifiers or much smarter agents that are able to decide where a workload goes and then just be able to use distributed nodes, whether it's the PC fleet, a GPU cluster on a data center or wherever. So I think that's the future to be honest.

Ish Shah: 00:55:23 Most people have not thought about computation as a resource like this in a long time as a token to be spent as a dollar to be spent. How many units of compute do you have? And now on YouTube you see every which person stringing together 50 of the same device stacked all the way to the ceiling. Look at this node I made. There's a particular Formula one team that Dell works a lot with a particular team that happens to be winning. I'll add a particular team. Dell works a lot with and they are very big on sandbox experimentation. They're very big on, I

don't want to worry about, I just want to test something quickly. I have an idea, I have a hypothesis and this is data science we're talking about. This is thousands of points of telemetry being captured off of these cars, rows upon rows upon rows of data.

00:56:16 I just want to see something real quick. I don't want to have to worry about do I have to spin up my instance? Is there a cue? Just give me a device, let me go look inside. But this file is much too big to open an Excel and scroll to row seven 50 to check a particular telemetry point. Sandboxing is another big reason why this stuff on device is going to matter. And even if the rest of it isn't clear, experimenting with local AI to see if it is part of your workflow or if it is something that's useful to you, that's the only way you're going to find out you got to do it.

Jon Krohn: 00:56:52 Really cool. And as an F1 fan, it certainly is not a bad time to be affiliated with that particular team there. Dominance this season is pretty insane.

Ish Shah: 00:57:00 Yes, powered by dust.

Jon Krohn: 00:57:05 Nice. I have one last technical question for you that I'd just like to squeeze in quickly. There's a big Windows refresh coming up in a couple of months in October, well I guess at the time of this episode release in a month. And so it sounds like a lot of people might need to be thinking about upgrading their hardware now, especially if they're using a pc. And so tell us a bit about this big refresh and that'll be kind of where we end the episode. I'll ask you both for book recommendation after, so you can start thinking about that in your subconscious. But otherwise, yeah, let's wrap up with this Windows Refresh.

Shirish Gupta: 00:57:43 Alright, so let me go first and then Ish we'll get your take, since I'll keep it audience conscious. So let's talk about Depro Max first. And if I even look at it broadly, just data

scientists and developers, something to look forward to is the new line of PCs that are equipped with NVIDIA's Blackwell, RTX Pro GPUs use. Really game changing. I think looking at some of the capabilities, right? You can run h talked about code Gen A while back. You can run a model like Devs Stroll Small, which is the 24 billion parameter one at SP 32. You can run that completely locally on the RTX 6,000 Blackwell, GPU, which is doubling the memory for VRA M. It's 96 gv, VRA M, which is allowing you to do these kind of models work with these kind of models locally. The other cool things about these devices is to think about the power that these are consuming.

00:59:02 So these are triple fan cool designs. So massive thermals investments in thermals and acoustics to give you that end experience. So if you're the queue to your listeners is if you're trying to bring some of these newest sota state-of-the-art models and run them locally and you try to do it on the previous gen product chances that you either just can't do it, or if you're able to, depending on the model of choice you're going to have pretty much like an airplane on your hands, right? That's what it's going to feel like. So strongly recommend your listeners to consider the newest line of performance desktops and laptops. One other thing I'll tell you is one of my colleagues was running some quick tests and what he found was for a video workload, just think of animation, they were working with a large studio and they were trying to create based on previous artwork and shows content.

01:00:15 They were trying to train a model that would keep the style consistency and be able to sort of stage up the content in the same milieu. And so different, they were trying to create mockups with fine tuned flux loras. And in the previous gen GPUs it was taking approximately a week. And with the Blackwell GPUs, they're getting it

done in about three days. So that's pretty measurable in terms of performance gains that you're getting. So just talking about that very specific persona, I would say take this opportunity to, with this win 10 end of life, if you're looking for a new device, definitely look for the latest and greatest because the gains to be had will set you up for the next four or five years to come.

Jon Krohn: 01:01:07 Excellent. I can't believe how you can reel through those technical stats seemingly without notes. Pretty mind blowing. Shirish. Ish. How are you going to follow that up?

Ish Shah: 01:01:17 Yeah, I'm going to do this and I'm going to whack some people ahead. Right? Do not let this deadline pass you by for the love of God. This is not a selfishly motivated thing that is coming out of my mouth right now. This is a look, when operating systems age out, there's a reason they're aging out. There's new stuff coming from a security perspective. From a manageability perspective. If you have been procrastinating, please stop procrastinating. Whether you buy a Dell device or not, just please make sure you understand the gravity of an OS refresh and make sure you're not hanging onto something that might be best retired in its 17th season. Right. So I think the only thing I can add to shish is like, Hey, October is here. It is no longer this big thing. In a future horizon, you need to solve this problem. Now it is a problem.

Jon Krohn: 01:02:15 Nicely said. And by the way, for our listeners, so ish when he was hitting you on the side of the head, he was flicking his microphone, which he probably presumed would create some kind of loud sound or whatever, but there was nothing,

Shirish Gupta: 01:02:27 Man.

Jon Krohn: 01:02:30 Just kind of a pause. So yeah, fantastic. Noise

Ish Shah: 01:02:35 Tion. Yeah,

Jon Krohn: 01:02:36 Exactly. Something like

Ish Shah: 01:02:37 That too. Good. The AI features have gotten too good guys.

Jon Krohn: 01:02:41 Yeah. Ish and ish. This has been a fantastic episode, but yeah, before I let you go, I would love to get a couple book recommendations ish. Let's go with you first.

Ish Shah: 01:02:51 There's a great book called Deep Learning Illustrated, Jon. Jon's book of course, of being a great one. I took a course at Sloan called The Analytics Edge, and it took me as a person from, didn't know anything about anything in terms of data science, and it gave me what I thought was a great starter foundational toolkit for a business and strategy oriented person to become familiar with applied data science and how to use it. In

Jon Krohn: 01:03:24 My book.

Ish Shah: 01:03:25 No, the analytics, no, no, you're talking about the analytics analytics,

Jon Krohn: 01:03:29 So that's a course that people can take.

Ish Shah: 01:03:32 It was,

Jon Krohn: 01:03:33 I thought maybe you were saying that my book was the companion text for that course or something.

Ish Shah: 01:03:38 No, honestly, it should be though having looked at both, I think it should be, so we should have that conversation and see who we can bother over at MIT to get that used. But no, it's about feeling like this is within reach for a lot of people. You don't have to feel like AI and data science in 2025. Like gosh, there's these 16-year-old Wonder kids who are out there for me to start now mid-career trying to

learn this stuff. Why even bother? So on the technical side, like go check out the analytics edge, it made me a somewhat competent human being, which I was less so before I read it and before I took the course. The second thing I'll say is I have a renewed interest in fiction in this world because what these models are doing for creativity and what these models are doing for content and storytelling more broadly is a really fascinating element to this that I spend a lot of time thinking about, which is if everything is derivative of something older, is that really a new phenomenon or has it always been that way? So I'm actually blazing my way right now through the Reacher series, the Jack Reacher series, and inspired by the Amazon show. I had never actually picked up those books and read through them, and as I'm reading them, I'm also thinking about, gosh, what is the future of fiction going to look like? So that's my plug for Lee Childs.

Jon Krohn: 01:05:01 Very nice. I like the fiction recommendation there. And just so that I'm understanding, so the Analytics edge is both a course and a book?

Ish Shah: 01:05:09 Correct.

Jon Krohn: 01:05:12 Okay. Yeah. Great. Perfect.

Ish Shah: 01:05:13 We will take a look at that, Jon, after we hang up just to make sure that I'm not making something up. But yes is the short answer.

Jon Krohn: 01:05:21 We'll have something in the show notes for you listeners and Shirish, do you have another book recommendation for us? I know we,

Shirish Gupta: 01:05:27 I do. I'm actually reading Seven Powers by Hamilton Helmer. Fantastic. I think it's a very fresh take on business strategy and sustainable, long-term profitable companies, so I highly recommend it. I'm still reading it.

I'm not done with it, but it's brilliant. The second one is actually one that I picked up based on your recommendation. It's the Hands-on Machine Learning Guide by Aurelian, which I haven't started yet, but it's sitting on my desk, so it's going to be my companion as I revisit some Python after years.

Jon Krohn: 01:06:14 So that will actually be, so this episode that we're filming right now, it will be a week after we just had Aurelien Geron on the show. He's the bestselling machine learning author of all time with his Hands-on machine learning series and yeah, so thank you for that nice cross-reference there, Shirish. Fantastic. Thank you both so much for being on the show. This is another wildly informative episode from you Shirish and Ish. I really enjoyed getting to meet you as well. You do have so much color to add and you can't hide that intellect. I'm going to tell your mom that you do have it.

Ish Shah: 01:06:53 Coming from you, Jon. That's going to mean a lot. Also, confirmed Analytics Edge is a book not Making it up. I wasn't hallucinating the years 2018 through 2020.

Jon Krohn: 01:07:03 Nice. Alright, thanks so much both of you and I wouldn't be surprised if we're hearing from you again on the show sometime soon.

Ish Shah: 01:07:10 It's been awesome, Jon. Thank you.

Jon Krohn: 01:07:12 Thanks

Shirish Gupta: 01:07:12 Jon. Fantastic.

Jon Krohn: 01:07:17 What a fun and informative episode with Shirish Gupta and Ish Shah. In it, we covered how Windows is used by 64% of software developers and WSL two bridges the gap by running Linux kernels directly on Windows. We talked about how with modern discreet NPUs, we can locally

train models of up to 200 billion parameters, how NPUs are optimized for AI workloads with superior performance per wat for battery efficiency, while GPUs offer higher absolute performance and scalability today, and lots of other considerations around cloud versus local AI work and what hardware you might want to use for either situation.

01:07:55 As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for my social media profiles, as well as my guests at superdatascience.com/921. Thanks of course to everyone on the SuperDataScience podcast team, our podcast manager, Sonja Brajovic, media editor, Mario Pombo partnerships manager, Natalie Ziajski, researcher Serg Masís, writer Dr. Zara Karschay, and our founder Kirill Eremenko. Thanks to all of them for producing another super episode for us today for enabling that super team to create this free podcast for you. We're deeply grateful to our sponsors. You can support the show by checking out our sponsors links in the show notes and yeah, otherwise you can help us out by sharing the episode with people who would value it, reviewing the episode wherever you listen to or watch podcasts subscribe if you're not already subscriber. But most importantly, I just hope you'll keep on tuning in. I'm so grateful to have you listening and I hope I can continue to make episodes you love for years and years to come. Till next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.