**ANDRE**We've expect this system to effectively have the same number of neurons as a human brain, with the same number of average inputs per neuron as the human brain, running that at the same number of spikes processed per second as the human brain.

**CRAIG**Hi, I'm Craig Smith and this is Eye on AI. In this episode, I speak with Andre van Schaik, director of the International Center for Neuromorphic Computing at Western Sydney University in Australia. We talked about Deep South, a brain scale neuromorphic computing system with the ability to simulate up to 100 billion neurons in real time. Van Schaik explains how Deep South uses spiking neurons and synapses to process information more efficiently than traditional AI models. Deep South aims to help researchers better understand brain computation and potentially unlock novel AI architectures. I hope you find the conversation as incredible as I did.

**CRAIG**

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested. So buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control, it's time to upgrade to the next generation of the cloud. Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has 4 to 8 times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber eight by eight and data Bricks mosaic, take a free test drive of OCI at oracle.com slash Eye on AI. That's E Y E O N A I all run together.

**ANDRE**

My name is Andre Van Schaik. I'm the director of the International Centre for Neuromorphic Systems at Western Sydney University. I've been in neuromorphic engineering almost since the beginning since, the early 90s, very early 90s. So it's getting on to thirty five years now.

**CRAIG**

And let's begin by explaining what neuromorphic computing is. I mean, I know that it's a different architecture that's meant to mimic the brain, I know that the artificial neurons in a neuromorphic computer fire on action potentials and spike in the way that biological neurons do. But can you sort of break down, before we talk about DeepSouth, the computer that you guys are building, how the basic differences between Von Neumann architecture and neuromorphic computing architecture, and specifically maybe start with action potentials and spikes, as opposed to weights and what happens in an artificial neural network?

**ANDRE**

Yeah, so to get back to neuromorphic computing, or actually neuromorphic engineering, to start with that, we're taking inspiration from neural systems in biology, and try and implement that in technology. That means we also need to understand neural technology and biology to take that inspiration from it. And there are sort of two main branches to neuromorphic engineering, one is neuromorphic, sensing, taking inspiration of the senses that we have and how biology senses information around in the world. And then neuromorphic computing is largely around how do we then process that information from such sensors like brains do in humans, but also in dragon flies and mosquitoes. Neural computation then uses neurons, neurons in biology indeed send action potentials to each other, they connect their axons onto dendrites of other neurons, with synapses, and the synapses typically have some strengths. And those strengths can be modified through learning rules. So that is similar to weights in neural networks but the spikes are the means of communication. And they lead then to pulses on the postsynaptic neuron over a certain weight, depending on the weight of the synapse. In this architecture of a brain, all neurons operate in parallel, they don't operate on a common clock. So it's asynchronous communication. It's pulse based communication. And that the memory that these systems have, is effectively in the synaptic strength, at least that's what we currently understand from how our brain works. And that's quite different from a Von Neumann machine, where memory is a separate thing to compute, and you're forever shuffling things back and forth from memory to the compute block, and then write the result back to memory, and then do the next operation in a sequential fashion.

**CRAIG**

Yep. And also on the clock in Von Neumann architecture and with the artificial neural networks, there is input information and there's calculation at the neuron level. And then the result or the output of that neuron is passed on to another neuron according to the clock when the transistor resets, right. And there has been work on building spiking neural networks. I know IBM has this TrueNorth chip. So can you talk about how a chip like TrueNorth works differently than a CPU or GPU processor?

**ANDRE**

Yeah, so the chips like TrueNorth, and that we've called this system DeepSouth, slightly in reference to TrueNorth because we're down under when we're very far south in the world. TrueNorth was sort of the first large scale neuromorphic system that was doing spiking neurons. This was part of a US DARPA funded project called Synaps that IBM did. That started a good decade ago now, that they did that. That has been followed up by a few that have been doing that in most recent years it's been Intel during their Loihiplatform, which is similar in goals. It's a different design. I don't know the actual ins and outs of TrueNorth or Loihi, in terms of what actual architecture they have on it. But they are all aiming at emulating neurons that communicate with action potentials, so spiking neural networks, and implementing those,and running those. DeepSouth is a system like that too. We're emulating how a brain works on the electrical level by how it communicates with pulses and how does communication with that. Some of these systems that have been designed, they have attempted to be fully asynchronous, just like the brain is and not gonna do it on a common clock. And that is quite difficult actually, to get to work properly. So most of these systems, they will still have a clock, DeepSouth, for example, it's built out of FPGAs (Field Programmable Gate Arrays) that's reconfigurable hardware that operates on a clock. And so our system is not asynchronous in that sense.

**CRAIG**

And the point of the clock in a current chip architecture is the calculation is done all at the same time across the transistors, across the chips, it's not right?

**ANDRE**

It is, and the other point of the clock for us is that DeepSouth has almost 100 FPGA boards in it. And we can do up to almost 100 billion neurons that we can simulate with that, in real time. But that doesn't mean that every 100 billion neuron has its separate implementation in the hardware in order to get to these numbers. We have to use the fact that silicon is much faster than biology. So we have one block compute, say 128 neurons, one after the other in a sequential fashion. And again, that clock is needed to schedule that basically.

**CRAIG**

Yeah. But the point of spiking, asynchronous spiking neurons in the brain is that they only spike when the action potential is reached, when they have enough stimulus to send a signal. And in an artificial neural network they all update at the same time, according to the clock. Am I misunderstanding that?

**ANDRE**

No, that is correct and that's a good point. So we do make use of that sparsity of signals that you see in biology where you only communicate when you have a spike, so when a neuron has passed a threshold. Obviously, the neuron itself is still updating, even if it hasn't spiked based on its inputs until it reaches this threshold, though, other neurons wouldn't know about that. And that gets used in this system too. So while the neuron internal updates are done on the clock, only those neurons that spike will actually communicate with the other neurons in the network.

**CRAIG**

Right, and then an artificial neural network, regardless of the input, the strength or relevance of the input to an artificial neuron, it will communicate onward. It's just a matter of how strong a signal is?

**ANDRE**

Yeah, in a artificial neural network like current deep neural networks and convolutional neural networks, all the values from all the neurons are passed on to the next layer right, and processed by the weights including zeros.

**CRAIG**

And is that why artificial neural networks consume so much energy and why neuromorphic computing consumes so little energy?

**ANDRE**

Yes, partly, at least it is not making use of the sparsity. So all the operations have to be done. Another part that is expensive is this constant shuffling of data from memory, to the compute and then back to, to memory. With the FPGAs, we can't totally avoid that, either. But we have, in our system, quite a lot of local memory, high bandwidth memory integrated with the FPGA nearby, which reduces some of the energy costs of moving the data. Because again, where, as I mentioned, where one block will calculate multiple neurons, one after the other in scheme, that's called time multiplexing. And in order to do that, you need to store the internal state of each neuron back in local memory when you're updating the next one, and then writing that result back to local memory as well. So, we still have some of that. We try and keep it as much local and distributed across all the FPGAs. And each of the FPGA blocks will have its own memory.

**CRAIG**

Yeah, can you just give us a thumbnail on what an FPGA is?

**ANDRE**

Sure. So an FPGA is a Field-Programmable Gate Array, it's a digital platform that can be programmed a bit like a CPU. But by programming it, you're actually setting hardware switches, and basically changing how one block on the hardware is routed to the next block and how things are connected to do a certain operation. So you're really configuring the hardware. And you can do that multiple times, which is why this is also called reconfigurable hardware. And so once you've programmed it, you actually have a hardware circuit that sits there that does the calculation that you have programmed whereas on a CPU, when you program it, the hardware has stayed the same. You've got your CPU that's doing the calculations. And it is looking at what instructions it now needs to do from instruction memory, computing these instructions one after the other, but with exactly the same hardware that hasn't been changed no matter what program you're sticking.

**CRAIG**

And in that reprogramming, is that dynamic, depending on the sensory input that that the FPGAs are receiving, or is that something that's physically reprogrammed by a technician?

**ANDRE**

So that's something that's physically done. So we try not to change the configuration of the hardware for the various models that DeepSouth might simulate. So the user will specify an architecture of their spiking neural network that they will want to run on it. It runs on the same hardware, we don't have to reprogram the hardware for that. We've configured it already to simulate the neurons, and the synapses, and the connections and those things are available. And users would simply have a software description of their network architecture that they want. And that then gets sent to the memory of the devices. And that can run on that hardware without having to reprogram the hardware. Where the reconfigurability does come in is if we want to add other features to it, new Learning rules, different neuron models, things like that. We can still add those after we've built this machine. And that's a big difference going back to say TrueNorth or Intel Loihi, those are chips that are custom made and what's on them is what's on them. And you can’t add things to that without doing a whole chip design and fabrication. And that's expensive and slow. If we look at Intel Loihi, to go from Loihi 1 to Loihi 2 that came out two years ago, I think there was a six year cycle. And so doing updates and adding a learning rule to that, that's not quick. Whereas for us with this hardware, we can do that.

**CRAIG**

Yeah, that's that’s fascinating. The network works among neurons can reconfigure themselves dynamically, right? I mean they're not fixed?

**ANDRE**

Yeah. And we tend to simulate that by changing the weights of connection. So weights can go all the way to zero, which effectively means the connection has disappeared, and can grow. And connections can start as well. Typically, I've seen networks defined where the connectivity is defined by the user. It might be a probabilistic definition and saying, from this group of neurons, that group of neurons, there's a 20% probability of a connection from one neuron to the other, or something like that, rather than saying, oh, yeah, this one goes exactly to this one, this one goes exactly to this one. But that does tend to be static in most simulations at the moment. And that is, because adding this flexibility of reconfiguring all the connectivity, indeed makes the simulations even harder. It is something we can support on this machine, though.

**CRAIG**

Where are you with DeepSouth? Is it is it finished and in the testing phase, are you still building.

**ANDRE**

We're building it and actually physically, over the last few days, we've had the hardware put in the racks. And that's still ongoing, and it's being connected. It should be connected up in the next few weeks, all fully connected. And then we can run some tests. We're now looking to launch it in the beginning of June. Part of that delay is, I'm going to travel for the next seven weeks and I want to be there when we launched the machine.

**CRAIG**

Yeah. And what will you launch it on? What sort of computation are you going to ask it to do?

**ANDRE**

What is called Balanced Excitation-inhibition Networks. It's the sort of traditional setup in spiking neural networks where it's based on what we see in cortex, in the human brain, where we have about 80% excitatory neurons, so neurons that excite other neurons and 20% inhibitory neurons, neurons that inhibit or reduce the excitation of other neurons. And those are connected in networks that keep the excitation and inhibition balanced. That's why they're called Balance Excitation-inhibition Networks, and keep the activity level of the brain within a reasonable amount. If you just had excitatory neurons, then the excitation could easily run away and you end up with everything spiking all the time. And obviously, if neurons were just inhibiting each other, nothing would ever happen, so it's that balance there. That in itself is not a useful computation but it is a easy demonstration for us to show that everything works; that all the neurons are computing as they should, that we're able to achieve the balanced excitation and inhibition with a very large population, that we can run that in real, time that all the populations can communicate to each other as needed and so on. So it's a good functionality test, but it's not doing anything intelligent basically.

**CRAIG**

Yeah. Given that this is based on the biological structures of the brain and that artificial neural networks that run on Von Neumann architecture are mimicking the neural activity of the brain, do you expect them to run neural network architectures on the neuromorphic computer on DeepSouth?

**ANDRE**

We can support running artificial neural network type architectures on the machine. It is not the initial plan to do that but we specifically want this to run spiking neural networks. The reason for this is that artificial neural networks, the convolutional neural networks that underpin current machine learning and AI, they run very well on GPUs, Graphical Processing Units, and we can do that now. Pretty large networks, it still takes a fair amount to train them, but then once they're trained we can execute them in a reasonable amount of time or pretty quickly. And they run well on that. And that was, effectively a revolution or an evolution, because the GPU is was more of an evolution and a revolution, I think. That really helped machine learning and artificial neural networks to get to the power that they needed, and get to the size that they needed to be able to do the things that they can do now. I'm sort of hoping that with DeepSouth, which is aiming to be able to run spiking neural networks efficiently and lives networks efficiently, that we can do the same thing for spiking neural networks. Because spiking neural networks don't run well on GPUs, or on CPUs. They're very slow to simulate. And that means that researchers that have been trying to understand spiking neural networks at scale, they've always had to simulate small models, toy models, basically. And that's a problem because like, if you look at, say a neuron in the cortex, it can have 10,000 inputs, one neuron, from other neurons. And if you’re only simulating a network with 1000 neurons in it, because that's what you can simulate on the computer, how do you get 10,000 inputs? And you can go, well, I'll just do 100 inputs instead, and make each input 100 times stronger so they get the same total input strength but you can imagine that if you get spikes that are 100 times stronger, one spike in just means one spike out. And the dynamics of the computation gets totally different. So you can't play that trick very well. So we really hope that this will enable researchers worldwide, ultimately, in first instance, we have to get this up and running for us, but it will be available for other people to use, and they can create accounts on it, and there'll be cloud service effectively. And the other important point about DeepSouth is that it's non custom hardware, it's commercial hardware. So you can just buy the components and build one yourself and use the configuration that we've designed for the FPGAs, configure it the same, and you have a copy of the machine. And again, that's not so easy, when you use custom made chips. Often, there's only a limited volume of those chips around and only one machine is built. And that's what people have, maybe get remote access to. Here FPGAs is a commercial business, you know, these companies, they make them, they improve them, they sell millions of them, they will continue to do that, people can buy this stuff.

**CRAIG**

Yeah. Do you need new algorithms to run on this machine? On spiking neural networks? Or are they transferable? You know, the transformer algorithm, for example, it's done so much in generative AI. Will you be able to run that algorithm on a neuromorphic computer, on DeepSouth?

**ANDRE**

So there are spiking versions of, say, a transformer architecture that have been created. So they can be trained, and they can be made, and we could run them. The thing is taking a transformer from an artificial neural network and making it spiking tends to work almost as good as the non spiking version. But there's not much point in making this the spiking version other than to say, see, I've made it spike. Our brain doesn't exactly do to transformer implementations, I believe. I believe there are indeed other architectures that are going to really unlock the advantages of spiking neural networks. And those are the ones that are yet to be discovered. And I think we need to discover them by trying things at scale with the spiking Neural Networks, just like the transformer architecture for artificial neural networks was only discovered once we could make really large, artificial neural networks. Back in the 80s and 90s, when we were doing three layer artificial neural networks, there was no transformer architecture to be found, because it just wasn't possible to make them.

**CRAIG**

Yeah, although the power consumption of the big transformer models is so enormous. They would be a lot cheaper to run in spiking neural network on a neuromorphic computer wouldn’t they?

**ANDRE**

That's the hope, yeah, so this system that we're building, it's not extremely low power. It's about 40 kilowatts, maximum power, the system, which is about the building air conditioning of a medium sized building or something like that. It's not a power that you see for data centers where it's megawatts. It's not portable power either, it's somewhere in between. But that can be improved once we know what we want to do with it; then if you really want to go to low power, then you will have to go and build some custom chips to really get to the low power. But my view on that is we still don't quite know what architectures we need, what type of features we need from the neurons, what type of learning rules we're going to need exactly. So let's explore that on this machine that is flexible first and once we've figured that out, then we can start making some chips that make it really low power.

**CRAIG**

And who were you talking to in the research community? You know, I started this podcast after meeting Geoff Hinton, and he's been on a few times. And he talked about spiking neural networks and his ambition is not so much to create AGI, but to understand how the brain works. The last time I spoke to him, he was working on forward networks. Is that what he called them? He was looking for an algorithm that could work in the brain because he didn't believe that backpropagation could work in the brain. Are you talking to people like Geoff that could use this computer to explore some of their ideas?

**ANDRE**

Not yet, at least not yet broadly. We have, as part of the application for funding that we did to get this system built, we have a team of about, at the top of my head 13 researchers largely in Australia, because this is Australian funding, but not not exclusively. So one of the people we have as part of the team is Emerson Neftci. He’s in Aachen, in Jülich, where there's a big supercomputer center. And he is one of the researchers that started Surrogate Gradient approach that is used to do the back propagation with spiking neural networks. Because the issue is spiking neural networks and error back propagation is, an artificial neural network should past the arrow back through the gradients of the activation function. But in the spiking neural network, it's a discontinuous activation function because it either spikes or not. So you don't have a gradient that you can pass things through. So you have to come up with surrogate gradients in order to do that. And so that's one of the approaches. But I would agree with Jeff, that I don't think the brain is doing backpropagation that way and there’s other architectures and learning mechanisms with local learning that need to be used in order to train such networks.

**CRAIG**

Yeah, can you talk a little bit about the number of connections in DeepSouth?

**ANDRE**

We've expect this system to effectively have the same number of neurons as a human brain, with the same number of average inputs per neuron as the human brain, running that at the same number of spikes processed per second as the human brain. The number of synaptic operations per second, so that's a spike arriving at a synapse activating the neuron; the postsynaptic neuron, is in the order of 220 trillion per second. And that is what this machine can do. Obviously, they're very simplified neural models, they don't have the full 3D structure that biological neurons have. They’re digital neurons, because we still store the states in memory. The precision of that is limited to eight bits only because we can't have the precision too high or we would need to store too much in memory, for instance. So, it's a simplified human brain-scale spiking neural network simulator but it is the first time we achieved this, I think in the world.

**CRAIG**

Yeah, that's really remarkable. And what did you say? 40 kilowatts? Is that what you said?

**ANDRE**

Yeah, 40 kilowatts power usage.

**CRAIG**

As opposed to what, 20 watts in the brain?

**ANDRE**

As opposed to 20 watts in the brain, yeah.

**CRAIG**

But compared to, I don't know, what GPT four is consuming

**ANDRE**

Yeah, I don’t know either but you hear the stories, it's the power consumption of New York City for a certain extended time just to train a network like that.

**CRAIG**

Yeah, and is this an open-source architecture? You were saying that people can then build these wherever they want, at a reasonable cost?

**ANDRE**

Yes, that's the plan. So obviously, this year, we will launch it, then we will develop stuff for it internally in the lab and make sure everything works and for the remainder of this year. Then it will be opened up to the researchers that are part of the team and we're working with, directly but that are remote at all the universities around the world. And then we'll grow that community and open it up to other people and at that point in time too, our design will be made open source available so that people can copy the machine. And they don't have to do it. So for us, it's three data center racks, it's twenty four server computers with four FPGA boards in them. And that's an expensive machine, but you could scale it down to just use one FPGA and obviously much smaller neural networks, or two, or eight; or if somebody is interested and has a lot of money to burn on this, they can make a superhuman brain basically with it, or at least superhuman brain scale, I should say, because I don't know whether it's a full brain yet.

**CRAIG**

Yeah, at current prices for FPGA boards, what would it cost for somebody to build? I mean, obviously they're benefiting from all your time and effort, but to replicate this, what's the ballpark for building one of these machines now, once it's open source?

**ANDRE**

Probably around 2 million US in terms of hardware? Yeah. So it's not cheap but its not way out there either.

**CRAIG**

Yeah, I mean, it's within reach of a lot of companies. And the algorithmic, the software research, how far along is that? How far does it need to go for your computer to be useful or doing some amazing things?

**ANDRE**

Yeah, I would like to say, “Oh, we're ready to go there.”  but actually, I think we have a long way to go there. My first goal with this machine is to better understand how the brain does computation with its electrical pulses. We know some of the principles but we haven't been able to study this at any large scale yet simply because we haven't been able to simulate it. And you can look at real brains but there you can't observe everything you might want to observe; plus, you don't have full control over what is going on either. So there's a lot to be discovered still around what works and what doesn't work. And once we have that, then we can start applying that to tasks, basically make AI. That's how I think of machine learning, right, you're learning to do tasks. And once we know some of the principles, then we can start working on doing tasks with the machine. But we first have to still discover the principles that work really well. And that's going to be a big research effort.

**CRAIG**

And this is mimicking. When you say brain scale, it's the neocortex, or does it you know, include the thalamus and the hippocampus and all the other structures in a human brain?

**ANDRE**

So the numbers I used before and used to spec this system for it includes all the neurons in the brain, including cerebellum, which is actually has more neurons in it than the cortex, for instance. Now, what the user ends up doing with these neurons depends on what model they define. So the model can be defined to specifically have thalamus, or superior colliculus, or cochlear nucleus if you want to do audio processing, or all those parts in it as well, or cerebellum say you want to do some motor processing. Because it is a cloud system effectively still, in server racks, it's not ideally suited to do robotic motor processing and control loops, because the latency would become an issue.

**CRAIG**

Is your background in neuroscience? What was your educational background preceding this project?

**ANDRE**

So my background was in integrated circuit design, my education and first job. But in neuromorphic engineering, almost from the beginning. My final year, undergraduate thesis was around making electronic micro electronic circuits for artificial neural networks. And that was in 1990. And so it's always been integrated circuit design, forward neural inspired hardware. And so the education in neuroscience has been effectively self educated over the years and, you know, attending a lot of talks, and conferences, and things like that on it. And then this system, there's not really integrated circuit design that we're doing for it. As I said, we would only want to do that once we know what needs to be on there. And instead, we're using this digital hardware that is available. That is not my personal area of expertise but we have people in the group like Mark Wang, who is an expert in how to write the code to configure these systems.

**CRAIG**

Now, do you talk to people like Jeff Lichtman at Harvard who's mapping the brain? I saw an earlier interview with you in which you were showing a brainbow image that I think came from his lab.

**ANDRE**

Yeah. No. So we're not in direct contact with that lab. We do talk to various people from neuroscience over time. We have in the neuromorphic engineering community, we have two main workshops every year annually, ones in Telluride, Colorado, a terrible place to be- Not! It's very, a very beautiful town. And then the other is in Capo Caccia, which is in Sardinia, in Italy, on the beach, also a very nice place to be. And those two workshops, they're multi week. So Telluride is three weeks, and Capo Caccia a two week one. And there we combine talks from people with an engineering background and have talks from people with a neuroscience background. And that's been going on now for a long time. I think Telluride workshop started in 1994 and it's still going, so it's been running for a long time. Capo Caccia has now been running for maybe 15 years or so as well. And so that's where we get annual interactions with various between the various teams basically, in the various groups.

**CRAIG**

Yeah, and the ambition beyond understanding how the brain works by building this, presumably is to be able to solve problems that we're not able to solve today, with, with the more powerful computers. Is that right or are you really more focused on just understanding brain structures and brain activity?

**ANDRE**

The latter, I would say, first instance, I'm interested in understanding the brain, I believe that if we do that, we will also find better ways to do artificial intelligence or machine learning tasks. But that's the secondary goal that would come after we make progress in better understanding the brain.

**CRAIG**

Yeah, and what sort of timeline in understanding the brain are you looking out? I mean, you're still a young guy? Are you hoping that within the next 10 or 20 years, we will have figured out how the brain operates at a much deeper level than we understand today?

**ANDRE**

I'm definitely hoping that when. Whether we achieve that also depends on unknown progress in research and in studying the brain, right. It's like we’re making this machine available and it gives us possibilities to do these simulations and study that. How quickly will that pay off? It's very difficult to predict. The other thing that's difficult to predict is, how popular will this be? How many researchers will end up using this or a machine like this that might be a machine inspired by ours built by somebody else somewhere else, but also with hardware that people can buy and replicate, and so on. I find that very difficult to see how that will go. I would hope it'll be very popular and lots of people want to use it, and copy it, and make their own versions, and make improvements on that. And that really locks up potential to do these studies. That should then, hopefully lead to much quicker advances in understanding how brains compute with spikes.

**CRAIG**

Yeah, is there much interest from the artificial intelligence research community?

**ANDRE**

Not yet, but we haven't really pitched it to them yet either. I've only announced it at a conference that we organized here in Australia last December, the Neural Engineering conference. I've since done a podcast here and there and some news articles on this. Once we've built it and we've launched it, then we will get it all working nicely, and we can talk about it more, because then I'm more comfortable that it's all working as it should as well. And then once we're ready, then we will start talking more widely about it and trying to attract users effectively for this type of machine.

**CRAIG**

Yeah. Well, it's fascinating. As you can tell, I'm coming from a position of ignorance. Are there things that I haven't asked about that listeners would want to know?

**ANDRE**

I think we've covered the main points about what we're trying to do here quite well. That is, trying to understand the electrical computation in the brain, having hardware that makes it possible to do this at scale for researchers worldwide, using hardware that is commercially available so people can make copies, and using hardware that is flexible so that we can add other features to it as we discover that we need them.

**CRAIG**

Yeah, I mean, it would be remarkable even if it's not your ambition, if you can replicate the success of GPUs and transformer algorithms, for example, because the whole GPU bottleneck is severe.

**ANDRE**

Yeah, and that would indeed be fantastic if we could do that.

**CRAIG**

Okay, Andre, I hope we can talk again, after you've done some testing with the machine. I'd really like that. And I'll be a better interviewer at that point. We'll know more about it.

**ANDRE**

Certainly, we can do that and to get give an update on how we are going with the system.

**CRAIG**

AI might be the most important new computer technology ever. It's storming every industry and literally billions of dollars are being invested. So buckle up. The problem is that AI needs a lot of speed and processing power. So how do you compete without cost spiraling out of control, it's time to upgrade to the next generation of the cloud. Oracle Cloud Infrastructure, or OCI. OCI is a single platform for your infrastructure, database, application development and AI needs. OCI has 4 to 8 times the bandwidth of other clouds, offers one consistent price instead of variable regional pricing. And of course, nobody does data better than Oracle. So now you can train your AI models at twice the speed and less than half the cost of other clouds. If you want to do more and spend less, like Uber eight by eight and data Bricks mosaic, take a free test drive of OCI at oracle.com slash Eye on AI. That's E Y E O N A I all run together.

**CRAIG**

That's it for this episode. I want to thank Andre for his time. If you want to read a transcript of today's conversation, you can find one on our website Eye on AI. That's E Y E hyphen on AI. In the meantime, remember the singularity may not be near, but AI is already changing our world, so pay attention.