Hi, I'm Craig Smith. And this isEye on AI. This week, I talk to Andrew Feldman, one of the founders and CEO of Cerebras Systems, who's wafer scale computer chip, the size of a square dinner plate, contains as many as 2.6 trillion transistors, all optimized for machine learning. Andrew talked about the reasons why such a chip exists, how it can be used, and about the network of chips that the company has built that has as much computing power as a human brain. I hope you find the conversation as amazing as I did.

**CRAIG:** Why don't we start, Andrew, by having you introduce yourself and a little bit about Cerebras and then we'll get into it.

**ANDREW:** How you doing Craig I'm Andrew Feldman, one of the founders and CEO of Cerebra systems. We're a Silicon valley based company. We were founded in 2016 in the winter, and we had a vision that we could accelerate artificial intelligence work, by not one or two or five or 10, but a thousand X. And then if we were able to accelerate AI compute by three orders of magnitude, we would open up entirely new areas for the ML practitioner to ply their trade.

**ANDREW:** And it would have a profound influence on our society. We attack the problem very differently than other. The problem of AI compute is unusual and is particularly poorly suited for existing compute platforms. It's a problem that has two elements. It has lots of little calculation and an abundance of communication because artificial intelligence is the feedback.

**ANDREW:** So you do some work and you move the results. You do some work and you move the results. And this moving is extremely difficult in compute. And what we saw was there was an opportunity to focus our innovation around how to move information more quickly, and that would dramatic performance improvements.

**ANDREW:** And the way we did that was to keep all the work on a giant piece of Silicon. And this was a problem that had been unsolved for the industry, how you could build a very big chip and the chip we built it's 56 times larger than the largest competitor. Our largest chip has 2.6 trillion transistors.

**ANDREW:** And the next biggest chip has 54 billion in transistors. So we are 2.5, 4 trillion transistors larger than the nearest competitor. Our chip is the size of a dinner plate, the largest competitor is the size of a postage stamp. And the reason this is valuable is because when you have a big chip, you can keep all your resources, your communication.

**ANDREW:** And your calculation engines called cores. You can keep them all in one fast domain. And that domain is fast it's high bandwidth and is very low power. If you have to break up chips in a lot of little pieces, like Humpty Dumpty and spread them out and then try and connect them over networking, it's very slow.

**ANDREW:** And so we saw this opportunity to solve a problem that had been unsolved in the 70 year history of compute, which was how to build a big chip. In August of 2019, we announced that we've yielded the largest chip ever made and set a new in the industry, the largest chip maker, who's our partner, Taiwan, semiconductor manufacturing.

**ANDREW:** Impressed with our work and our partnership with them, where they manufactured for us. They built an entire exhibit about it in their museum of innovation in Taipei as a high point and the evolution of chip design. In April of 2021, we announced that we had gone to our second generation of wafer scale in less time than most companies take to go from one generation to another, in a normal size chip.

**ANDREW:** We'd gone from 16 nanometers bypassing the 10 nanometer node all the way to seven. At the seven centimeter node, we have 850,000 cores on a chip. This is hundreds of times more compute cores than the leading competitor. Thousands of times more memory, tens of thousands of times more memory bandwidth, more than 40,000 times more compute connectivity bandwidth connecting the cores. And these are the foundational elements of high speed AI compute. And these are the reasons that we're able to deliver orders of magnitude higher performance, and to reduce the amount of time it takes to do some of the most complicated AI work from months to minutes, and weeks to seconds.

**CRAIG:** For listeners because my listeners are mL people, but not hardware people. Can you explain in very simple terms, how a CPU works, why a GPU is different than a CPU? What Google has done with the TPU and why Cerebras' solution is superior to those three. And then I want to ask you about quantum and how Cerebras relates to quantum.

**ANDREW:** Sure. I think the easiest analogy is that chips are very much like cars. You want a car that's got a third row of seats, that's good for taking the kids to soccer practice can take grandma and the kids to a picnic you get a minivan. If you want something fun to drive on Saturday, don't get minivan, get a two seat Roadster.

**ANDREW:** And if you want to move bricks and garbage, get a Ford F-150 you get a truck. In each of them has made design decisions that are particularly optimized for its use. The truck has easy access in the back. It is nearly impossible without breaking your back to get 50 pound sacks of concrete into and out of a minivan. It will kill you.

**ANDREW:** And it's impossible in a two-seat Roadster to get more than your wife and your golf clubs. Can't do it. That's not what they're designed for. And each one of those has thought really hard about the work that it's going to be doing. And they've made all sorts of trade-offs in support of that work.

**ANDREW:** Now the graphics world, the GPU, the graphics processing unit is exactly like a TC Roadster. It was tuned for a very particular piece of work. And for 20 years, every single iteration, change was optimized to push pixels from your computer to the monitor. Every part of that design was optimized for that. And it's the best in the world that. Now the CPU had a different challenge.

**ANDREW:** It's more like a Toyota Camry. It was designed to be good at many things. It can run Excel, right? It can do light database work. It can, it can take a family, your dad had one and you're afraid that you have to buy one too. Now that you've got three kids in high school, it's it was the Camry of.

**ANDREW:** And then came along the specialized work workload and the workload had very different characteristics. Both the CPU and the GPU were really not well suited for it. But the GPU, the graphics processing unit was less bad at it. Now there's an old joke. I'm sure you've heard of Craig about two old guys who go camping and get attacked by a bear.

**ANDREW:** And one of the guys leads down to tie his shoes and his friend says, you can't outrun a bear. He says, Nope, you and the GPU could outrun the CPU. And that was the only game in town. And as a result, they made hay in the sunshine and they built a wonderful business and kudos to them. The fundamental machine was built for something else. It was built for graphics, and it's continued to be used for graphics and they depend on the volume of graphics to keep the cost down.

**ANDREW:** And that's different than beginning with a clean sheet of paper. Now they're 50 or 60 different companies, including all the hyperscalers who are trying to do their own parts for this work and not one of them choose an architecture like a GPU, because if you had a clean sheet of paper, you would never choose an architecture that had its origins in a different problem.

**ANDREW:** You might remember Craig that he used to be a car called an El Camino. It was like part car part truck. It was a horrendous disaster, right? That's the danger in the chip world, right, is to build the EL Camino and good for nothing, not good enough to move real bricks and garbage, really uncomfortable and ugly to seat in.

**ANDREW:** That's the challenge when you're trying to be too many things to too many people. And so a group of us who had world class, computer architecture in our teams, only a handful, began working on parts that were tuned exactly for this work to continue the analogy that understood who would be driving it and what they'd be using for.

**ANDREW:** And that involved thinking about your listener, the practitioner. What work are they trying to do? And let's be good at that. Let's say no to everything else. There's no free lunch. You can't be good at many things without dragging costs along with you. And so we began from scratch. The guys that at Google began from scratch.

**ANDREW:** The guys at Google chose an architecture from the mid eighties that was very good at one thing, large matrix multiplies. And the architecture they chose was called the systolic array. It had been discussed in the literature and they then tried to use their strength in AI to write models that we're good at big matrix multiplies.

**ANDREW:** That's a very thoughtful strategy. Our approach was different. The underpinning of neural networks, isn't dense linear algebra, it's called sparse linear algebra. And we wanted to recognize this fundamental underpinning in the design of our core. We want it to use our advantage in physical design to put more cores on a single piece of Silicon.

**ANDREW:** Use our advantage in fearlessness and to build the biggest chip ever made. And when we were done, what we had was the biggest chip ever made with cores that were optimized exactly for the type of work in artificial intelligence. And as a result is vastly faster than the competition and that's the path forward.

**ANDREW:** We then quickly shrunk from 16 to seven manometer. We began shipping. We don't ship a chip. We ship a whole system we're system builders and we're system builders because you can't build a Ferrari engine and put it in a Volkswagen and expect a race. The exact same thing happens in compute. If you were to put a Ferrari engine in a Volkswagen, you go a little bit faster, but the bottleneck would just move.

**ANDREW:** Maybe it's the fuel system. Maybe it's the drive train. Now you've got a giant engine in an ordinary car and that's not the path to success. The system view says, if you're going to build a race car, you got to build the race car engine and you have to build every single thing. Craig I don't know if you like cars, but for those of your audience who love cars, think of the Porsche nine 11, every aspect of that car contemplates the fact that the engine is over the rear axle, everything that changes the weight distribution, it changes every aspect.

**ANDREW:** You have to take that into account .

**CRAIG:** Can I just interrupt on the GPU's I mean, certainly GPU's you know started as graphical processing units, but Nvidia atleast has since moved into specialized chips for AI,

**ANDREW:** they talked about it a lot, Craig, but if you look carefully you can see some things that they refuse to remove.

**ANDREW:** For example, 64 bit double precision logic, there is no use for this in AI, Yet, every GPU has it. Why do they have it? Why did they waste the space on the chip for problems that are never in artificial intelligence? Because the truth is they're trying to span multiple markets with the chip. They're trying to talk about and do marketing for AI.

**ANDREW:** And they're trying to gather revenue from high performance compute and from gaming and Bitcoin mining. It's from this collection of other markets. And for some of those, you need things that we never need. We never need a rasterage. We never need shaders. We never need all the stuff that is on a GPU today.

**ANDREW:** Even the ones they market for AI. And, we will know they're serious when they rip all that stuff out. And those parts can't be used in game when the parts can't be right, I mean that's when you're committed. When you give the trunk, you are committed to a two-seat Roadster, right? When you say, Nope, we're going to build this, not going to put it in second doors in the back.

**ANDREW:** everything's about fun. At that point, you really committed. Until end you're trying to modify a Camry with some good marketing

**CRAIG:** and google's TPU, which doesn't have all that stuff.

**ANDREW:** The TPU is a custom built part exactly for this workload. And they chose to solve part of the problem, which was the big matrix multiplies.

**ANDREW:** And that's a, an important part of the problem, but they didn't choose to solve what we think is the most important, which is the communication. And they had the added benefit that they then told their ML researchers 'write models that are good at matrix multiply.' That's not something that anybody else can do. But everybody else who's writing models.

**ANDREW:** Isn't thinking about how to write models that are good on a TPU. They're thinking about how to write models that are generally good. And OpenAI's models like GPT-3 aren't optimized for Google TPUs. Google's internal models generally are optimized for their own infrastructure.

**CRAIG:** And GPT-3. Let's talk about that.

**CRAIG:** or 70 billion parameters, I think. And that's distributed across thousands of GPU's that's right. Have you worked with OpenAI on putting GPT-3 on a single wafer scale integration?

**ANDREW:** There are a couple of important things to note. It took a long time to get GPT-3 on a cluster. The the cluster's more than a thousand GPUs.

**ANDREW:** It takes months to train. It uses megawatts. So you're using megawatts for months. The initial publication had about four months and more than a thousand GPUs.

**ANDREW:** Not very many companies on earth could do that. I mean, That was the premier algorithm guys at OpenAI with Microsoft Azure. That's it? Nobody else could do it.

**ANDREW:** We believe that with the announcements we've made, we will be able to do that on a small number of systems, a handful in a week and we'll be configurable in an hour, right? Not the four or five months, it took to get the cluster to work, but an hour. And that includes setup and cabling.

**ANDREW:** I think we are not just interested in making faster, the existing models. We're interested in creating a platform on which entirely new class of models can be built. You're right. GPT-3 is about 170 billion parameters. We announced a few weeks back that a single system could support 120 trillion parameters.

**ANDREW:** So what thousand pets larger, right? That's how we want to show the world that we're different.

**CRAIG:** I've got to ask how the Chinese play into this. I know Baidu just announced the large-scale a manufacturer it's Kirin I don't know if you're familiar with that. The Kirin chip, which is built for AI.

**ANDREW:** I think nobody wants to be dependent on the GPU. I think the industry and the hyperscalers in particular learned a lot from the last 25 years in which they were utterly dependent on Intel, and they do not want to be dependent again. And so they are doing internal projects. They are doing software like TensorFlow and PyTorch so that they can write models that can be moved easily across different hardware platforms, graphics processing, unit, the CPU, the TPU, our equipment, other people's equipment.

**ANDREW:** They're doing all these things to try and protect themselves from being entirely dependent on a single chip maker. Now, I think there's a reason that not everybody can make chips a specially high performance chips. And so many of the companies are starting with an inference chip

**ANDREW:** either for a consumer electronic for an edge application, it's a much simpler chip to make. It's a smaller chip. That's to put their toe in the water rather than dive on in and maybe drown, but these are our big bets chip projects, maybe a hundred million dollars in aggressive geometry. That's the starting point.

**ANDREW:** And so even for big companies, that's real money.

**CRAIG:** So the chip exists. When do you expect people to start putting systems of that size equivalent or larger than GPT-3. on the chip to see how it works?

**ANDREW:** There's Whole segments of GPT-3 running on our system, entire layers that were working and we showed a range from trillion parameter network to GPT-3, to Megatron, and data from them running on our system.

**CRAIG:** Cerebras right now, isn't built to run a single model. It's built to run multiple models simultaneously What's the biggest single model that people are using Cerebras for.

**ANDREW:** In training to completion from start to finish with results, it's probably tens of billions of parameters. Now, if you want to work on GPT-3, you're not going run a training each run. You're going to work layer by layer. Those we're running today and same for Megatron and the same for all other enormous networks, but from start to finish for a training run, tens of billions of parameters.

**CRAIG:** Does that mean that these massive transformer models are going to become commonplace?

**ANDREW:** That's a good question, Craig. Where we were in 2018, 2018, Google introduced BERT BERT had about 380 million parameters. It was the largest network at the time. Right now it's an ordinary size network. Two and a half years later, it went from being a very big network that everybody was struggling with, to a big network and where people are writing networks that are a thousand times larger. And so what was extraordinary in a two-year period became ordinary.

**ANDREW:** I expect that pattern to continue.

**CRAIG:** Super computers are massive parallel integrated computing systems. And the reason they're integrated is a lot of the speed is in the communication between the CPUs.

**CRAIG:** So you take care of that problem. You're in effect, building a supercomputer in a chip. How does the performance of a system, a layered system with multiple Cerebras chips in it? How does that compare with supercomputers, both in cost and in time, and energy

**ANDREW:** The super computers are generally built for very large simulations and those simulations have historically been 64-bit double precision.

**ANDREW:** We don't do that at all. So in that sense, we don't compete with that at all. Now, what we found is that many of these giant simulations, like our deployment at Lawrence Livermore in the simulation, there's an inference loop whe re inference is being used to inform parts of the simulation. And in fact, our CS-1 is deployed inside Lassen, which is the eighth largest supercomputer in the world.

**ANDREW:** And it's part of this giant piece of work. It's an AI part, but it's part of this huge giant workflow at Argonne national laboratory. We're being used to identify patterns in cancer patients and the interaction of drugs, the impact of different interactions with drugs to study COVID to, to look at information that's coming out of the linear accelerator at Stanford, the SLAC, Stanford, linear accelerator, all of those are different super compute applications in which we were a better fit.

**ANDREW:** The Pittsburgh center for supercomputing. They have A cluster of our machines. And they have a big HP infrastructure called Superdome flex. And together they build a cloud for researchers. Dozens of research teams are banging on the machine, trying to admit new types of AI or applied AI to different novel problems. And so those are some of the use cases, but I think your question of the giant supercomputers are almost always aimed at a very particular type of problem. Just like keeping the car analogy, those giant Tonka trucks, the, earth movers right, that are in mining towns are used to move amounts of earth and are not super useful on the freeway, that have a very specific purpose.

**ANDREW:** That's what the supercomputer are being used for.

**CRAIG:** And the the individual processing units on the Cerebras wafer, those have been designed specifically for AI problems. So you've done a way with.

**ANDREW:** You might think about it, and there are different ways to think about this.

**ANDREW:** If you're only going to drive short distances in town, you might not need a fifth year. You might think about what it's like. And I owned a Fiat 508, great little car, perfect for zipping around town. My commute was 15 minutes. Perfect. Terrible car. If you want to go a long distance. Had a range of about 90 miles, but my commute was eight miles, 12 miles.

**ANDREW:** Perfect. And we thought of the core as the perfect engine. What would be the perfect engine for AI work? What can we strip away? What can we add that was only for AI and what can you take away that would make it good at anything else? And when you do that, what you get is compute core.

**ANDREW:** It's a full machine. It could run in O S it's a nine- threaded machine,

**ANDREW:** and there are 850,000 of them. Each has its own memory. They're fully independent. And what we built it for was the perfect tool for its

**ANDREW:** job.

**CRAIG:** Does it have to be that large? Couldn't you slice them up into smaller packets for different uses or do you do that?

**ANDREW:** We do that. We have customers, for example, who run four models on the wafer at the same time for small models for training, sometimes we place dozens of the same model on the wafer at the same time and run them all concurrently.

**ANDREW:** One very interesting customer has a stimulus come in and they're four different inference models, entirely different soup to nuts, and they each do their classification. And based on the correlation between the results they estimate certainty or uncertainty. If four different approaches give you the same answer, it's less uncertain than if they give you different answers.

**ANDREW:** So they're quantifying uncertainty. Extremely interesting application. And notice that we only brought in one stream and that was replicated on the wafer. If you were to do this in another approach with lots of little GPUs, you'd in fact have four independent clusters, four independent networks and four streams data.

**ANDREW:** So those are different examples, all of different ways to use.

**CRAIG:** And then you're working on this brain scale integration. Can you talk about that?

**ANDREW:** Sure. No matter how big, how much memory we could put on a single chip, right? We're a thousand times more than a graphics processing unit, there would be somebody who wanted to network that was bigger stuff. We can posit that there is a guy who wants to explore something bigger than any single piece of hardware you can make. And what we wanted to think about was, and we knew this from the beginning, we wanted to say, what can we do that would allow us to demonstrate a thousand times the status quo and that involved inventing three new pieces of technology, something we call memory x, something we called swarm axis, and something we call streaming weights, which is software. The memory system allowed us to store weights off chip and deliver the performance as if they were on chip. The swarm X allowed us to extend the unchipped fabric off chip, allowing us to connect together up to 192 systems. Now each system have 850,000 cores.

**ANDREW:** So swarm X allows us to build clusters of up to 163 million compute cores. And finally, the streaming weight software allows us to manage this infrastructure and set up those clusters quickly and easily. And one of the hardest parts about AI is trying to spread your model over dozens of small graphics processing units. It's a Brutal problem.

**ANDREW:** Often takes months. We can set up a cluster with a couple of keystrokes.

**CRAIG:** One of the issues in putting those clusters on a single wafer is heat is that why you sell them as systems rather than as individual chips, because you understand how to deal with the heat.

**ANDREW:** Of course, we are a system builders.

**ANDREW:** We want to be sure that the system we build is designed to power feed. Cool. Deliver IO. So that every ounce of this extraordinary chip that we've made delivers performance. And so we are a system builders at heart. We don't want to leave anything to chance. We don't want our sort of race car engines placed in ordinary servers.

**ANDREW:** Where they behave exactly like you'd expect if you put a Ferrari in Volkswagen. Not a Ferrari. We know that if you put a Ferrari in a Volkswagen, it will not behave like a Ferrari. And that's exactly by the way, white Nvidia built the DGX. Is they saw so much performance left on the table. When you were putting GPU's in Dell servers and HP servers that they were like, we need to be in the system business.

**ANDREW:** The business they've avoided for 25 years , but you can't just build a race car engines you gotta build race cars, all of it soup to nuts.

**CRAIG:** Are you a believer in quantum computing?

**ANDREW:** I am a believer. I differ with many in the quantum camp about when it will happen. I think this is an extraordinarily interesting, it's something that we should be allocating societal resources to, but it's not around the corner.

**ANDREW:** The various different types of quantum computers are interesting, and they're about six or eight different paths that you can take to create your qubit. These are interesting. I don't wake up at night in chilled sweats because they are still very much bespoke, small, not production cloogy software. They are still a decade away or more from any general use.

**ANDREW:** So is it possible that one of the things they're good at is factoring large numbers? Which is a very hard problem and is part of the essence of many of our cryptography keys.

**ANDREW:** Is there a chance that sooner than that a particular problem can be addressed by a quantum computer maybe, but it's not around the corner and solving general problems that we have.

**ANDREW:** In today's compute landscape, I don't think is where quantum will excel. I don't think it will Excel in the next decade.

**CRAIG:** It's not so much general problems, but = machine learning problems. I just had a long conversation with Baidu about this, and they're very focused on developing algorithms for quantum specific to machine learning. And then you have IBM with claims are going to have a thousand cubit computer by I think 20 25, which is right around the corner.

**ANDREW:** It sure is. You need to use your own judgment. If what they'd said on those type of claims have come true recently, right? Use your judgment.

**CRAIG:** Although the guys that are doing quantum not the guys that did Watson right .

**CRAIG:** Let's say that they do get to a thousand stable qubits and that algorithms are developed for classification problems at the quantum level. All of these quantum systems are going to be quantum classical.

**CRAIG:** You need the classical element to extract the information. Do you see a role for Cerebras in that? Is that something that you guys are looking at or is the market for what you're doing endless enough that you're not looking in that direction.

**ANDREW:** The challenge with what you described is if you are hand tuning algorithms for your hardware, by the time that hardware is ready, you're years behind ML. That's the way we used to do vision we'd hand to now algorithms defining a nose or whiskers or cool sunglasses that turned out not to be the right approach.

**ANDREW:** And those are extremely interesting vectors for a search, but, it's not clear which algorithms are going to be relevant in three years, three years ago, BERT didn't exist , the most common language processing tool today. Transformers didn't exist. If you would have taken an LSTM and focused all your energy on what was then you didn't miss the market.

**ANDREW:** And so it is hard. We tip our hat to the ambition and that the sheer challenge involved in building these quantum machines. There are challenges that have to be sort of recognize I mean you need a machine running at four Kelvin or near absolute zero. And that takes tens of megawatts to power to get to four Kelvin, let alone with colder. Say you drive latency down to zero, approximately to do a classification. You've got to move your data all the way to wherever that quantum machine is, that's running it for four Kelvin. That time and moving your information is way more than the gain you got by going from classical to quantum in the classification.

**ANDREW:** This isn't to say it's not interesting at we should be spending money. We should be spending NSF and NIH money. We should be giving grants. We should be doing research. We should be investing as a society in these things. I don't think they're going to pay dividends in the next decade.

**ANDREW:** They will pay dividends two decades, three decades down the road,

**CRAIG:** the future for Cerebras and for wafer scale. Do you see it as being able to accommodate these massive models, billions or even a trillion plus parameter models or is the future, in just providing an AI tailored hardware that can run multiple models simultaneously.

**ANDREW:** It's a good question I think we have organized our technology and our go to market to address both of those. We are building big systems you can run small workloads on them, or you can run big workloads on them and subscribe to the system for a week to do work that otherwise taken you months and then ' I got my answer.'

**ANDREW:** So you can subscribe for a week, a month, a year, multiple years. It can be on our prem on your prem on a partner's prem you can access it by a cloud. It can be on your prem and you can own it. There's a collection of different approaches all to get at the very nature of your question, which is some people will want very big networks for a short period of time.

**ANDREW:** Other people will want ordinary size networks frequently. Other people will want very small networks that will want to iterate very quickly. And we have approaches both in technology and in, in purchase options to meet that need by the sip, by the box, in cloud, in a subsription all available,

**CRAIG:** The brain scale , how do you see that being used or is that sort of a, forgive me, but, a PR exercise.

**ANDREW:** Craig, I'm pretty old for PR exercises.

**ANDREW:** I think of the Al Pachino movie with the line where he says I'm old and I'm tired. we've been at this too long to do PR exercises. This is available in Q4. We are delivering solutions that use this we have customers in queue. We have customers doing problems today that others simply can't do that are impossible to do with other platforms.

**ANDREW:** So we're not interested in talking. We're interested in building cool stuff and we're interested in being in the computer history museum. When we founded the company, we wrote two things on the white board. The founders had all worked together before and what we wrote and we want to work together again.

**ANDREW:** The second point is we want to move an industry. And we love the computer industry. It has been where we've been passionate since we were in grade school, trying to put together a Commodore machine from the back of a popular mechanics, borrowing parts and assembled it as a hobby with our Dads.

**ANDREW:** That's who we are as people. And that's what we love to do and what it means to be an infrastructure builder is to build things on which other people's ideas take flight. You're a road builder. You take other people to places they didn't think they could go. And there's tremendous pleasure in that.

**ANDREW:** So no, this isn't a PR exercise. We were not interested in those. What we're interested in building is enabling people like OpenAir like others who are listening here and around the world to do work that couldn't otherwise be done.

**CRAIG:** And looking into the future, just indulge me for a moment.

**CRAIG:** I don't know how you feel about the AGI debate, but do you see Cerebras playing a role in that development that's largely been discussed on the data and algorithms side, not so much on the hardware side.

**ANDREW:** It was pointed out to me that there's a cartoon called Archer. And a few weeks ago, Archer had an episode in which there was an AI HVAC system that tried to take over the world called Cerebras after us, which I appreciated, I think. There are application level visionaries at open AI and others who are thinking of those thoughts. I'm trying to build hardware that will take them some or all of the way there. And right now, we're a long way from that. Now we have amazing things.

**ANDREW:** GPt-3 is, it's an extraordinary accomplishment and I'm sure they're working on whatever's next, but for me, we're not close to a sentient thing or AGI. I think what we're very good at is a class of problem that humans have never been good at before. And for me, that's equally interesting that in the hundreds of thousands of years of humans existence, there's a whole Class of problems, large data pattern matching problems that we've never been able to get our arms around.

**ANDREW:** And with these tools we can/ that's good enough for me. To do something then in human kind, hasn't been done to open up new areas for exploration. That's what we're trying to do.

**CRAIG:** My last question you mentioned, the cloud are the big cloud providers making Cerebras chips available in their clouds, or do you guys have a cloud? How do people access it?

**ANDREW:** We offer it, we have a partner Cirrascale who offers it and we're engaged with all the largest players.

**ANDREW:** You'll hear more in the years to come.

 That's it for this week's podcast. I want to thank Andrew for his time. If you want to read a transcript of this episode, you can find one on our website, eye-on.ai.

We love to hear from listeners. So, please email me at craig@eye-on.ai, and put the word ‘listener’ in the subject line so I don't miss it. Like all of you I'm inundated with emails.

And remember the singularity may not be near, but AI is about to change your world. So, pay attention.