Podcast Transcript: How a Low-Power Edge AI Chip Company is Driving Intelligence in Consumer Devices

Peter: [00:00:18] Hello everyone. And welcome to The Counterpoint Podcast. I'm your host, Peter Richardson. And today we're going to be talking about edge AI devices and how machine learning can be used in edge devices, all sorts. And to talk more about that and how these chips are coming to market, I'm delighted to be joined by David McIntyre vice president of marketing at Perceive.

Hello, David, how are you today?

David: [00:00:44] Great, thanks. How are you doing?

Peter: [00:00:45] Yeah, very well and indeed, very well. So, this is kind of an interesting topic. So, before we get into it, maybe you can tell our listeners a little bit about Perceive and about yourself maybe.

David: [00:00:58] Sure. So, perceive is an edge inference company. We make a chip Ergo that's really applied to consumer devices or similar things you'd find in the home or potentially at work where we're bringing intelligence to those devices.

We were created and came out of stealth in March, 2020. We're actually now about three years old from inception, and we were actually created in a public company called Experi. Some people may know them cause they own DTS and TiVo and HD radio where we kind of were a spin-out to create this focused chip company.

I'm responsible for marketing and product management and kind of new to the semiconductor world. But in kind of the marketing and strategy and tech world for my entire career. So that's kind of us and myself in a nutshell.

Peter: [00:01:51] Okay, great. Thanks for that, David. So, I'm getting into the topic of AI and specifically kind of, inferencing on edge devices. So, I guess many consumers will be familiar with products like Amazon Echo or Google Home that they can interact with in a somewhat intelligently. So maybe we can start by discussing the general need for AI chips in edge devices and maybe actually discuss exactly what an edge device is and what sort of problems that they can solve.

So how would you go about discussing that?

David: [00:02:31] Sure. So, first to clarify, when I say an edge device and edge is a term that gets bandied about a lot and people have edge of network or edge servers. When I say edge device, I use a term, some people call very edge. I mean the actual product in your hands, a phone, a laptop, a car, cameras.

Or if you were in the industrial world, a sensor on a piece of machinery. So, it's something that is at the very edge and attached to probably sensors, cameras, microphones, inertial, sensors, whatever. So that's the device we're talking about. And I come mostly from a

consumer perspective. So let's take that doorbell or one of your Google or Amazon devices, they've traditionally been powered by the cloud. The sensor data comes to the device and it is flung up to the cloud, the cloud figures out what it is, what to do with it and sends a response. That's cool. And it's created this class of devices that we all know and love. Many of us have those devices. I won't say their specific names so that everyone's devices don't wake up when they're listening to us, which is an interesting problem in and of itself.

But now that we're doing that with the cloud, we're starting to see some of the challenges, whether it's privacy, power latency, none of us wants a self-driving car that is sending sensor data to the cloud. And then returning before it knows to turn left or to brake, to avoid, a child running across the street.

So, there are a bunch of reasons why we want to do our compute at the edge. And so, the demand for edge AI chips comes from fulfilling that need to say, Hey, and you see it even in say the Amazon device. They now want to hear their wake-up signal locally or commands like play, pause, stop locally. Those shouldn't need to go to the cloud so that we can get quick response and a device that people prefer.

If you take a camera. Do I really need to send them all of the video to the cloud, or should I only send video to the cloud when something relevant happens when a burglar is in front of the camera, you want to record that, but when the leaves are blowing and waking up your camera, that's not relevant video and alert. You really don't want that. And an edge chipset would be, for example, the way you filter that out with the cloud only recording what really matters. So, I hope that kind of frames where I'm thinking and kind of the reason why there is a market for edge AI chips and edge intelligence in general.

Peter: [00:05:13] It does, I guess, what the cloud provides is huge processing power or be it that it, , it can be somewhat distant from the edge has, as we've defined it. So, and I think you've explained very well why you wouldn't necessarily want to send everything to the cloud, but there is a lot of computational power in the cloud. So, what can we expect in terms of processing power in these edge chips and what sort of things can they do?

And, and perhaps, also relevant is, cause we're talking about inferencing here. And I think one of the solutions that perceive offer is for example, recognize People that may be allowed to enter a premises or something like that. So, how do you go about training a device in that context?

David: [00:06:01] Sure. So, when we talk about edge versus cloud, I agree that the cloud has a real place. The cloud is the big horsepower in the sky and the edge is let's call it the filtering, the detecting the things that require immediate. In terms of the performance people can expect. It's been growing a lot as you'd expect the chip world and innovation in general.

There's been a lot of focus here. If I wind back the clock to three years, people were running very small models. Maybe on an ARM-based chip set and you have a whole avenue called tiny ML, where people are taking these models and squeezing them on ARM chipsets or

microcontrollers. And now you really seeing the emergency explosion of edge AI really kind of the big case would be like our phones.

Apple just announced the iPhone 13 and the 15 Bionic chip. And it's got a large number of TOPS of processing mixed with everything else, much like Qualcomm did with the Snapdragon 888, but for more IoT devices, you're starting to see the emergence of edge chipsets that can do some of initially these basic tasks.

Detect a person detect a pet detective vehicle, right? Th those were the kind of initial, basic things. I detect a keyword like, Hey, insert a name or play or pause. That's kind of where we were, as of, I'd say about a year ago, even now, till today. What you're really starting to see with more powerful chips and ergo from perceive as an example of that.

Is I can do those basic things, but now can I do more advanced language or speech recognition? Can I do audio detection, glasses, breaking, dog barking, baby crying. And can I even go further and say, a vacuum cleaner wants to avoid obstacles. Roomba just announced their vacuum's ability to avoid messes left by your pets that had created a humorous, if not offensive set of videos on YouTube. We're seeing a lot of interest in the video conferencing around things like foreground and background or doing gestures. Recognizing there's a white board behind you and unblocking the whiteboard.

There are these kinds of things that are, are called the next generation of features beyond the basic detections. And those are now what people are looking to add to their products in a whole variety of categories. You mentioned face recognition, so face recognition, and it's a bit of a dirty word.

It's got some legal ramifications. When you're in a consumer device, face recognition is usually more of a convenience feature. Who's at the door, or we all, anyone with an iPhone uses it to unlock their iPhone and face recognition. When you're thinking about enrolling, add my face, that's not actually training.

So, from an ergo perspective, and most of these edge products, they're doing inference. Teaching a network to recognize a face, actually isn't training in the neural network sense where I'm creating a network. The network exists to create the map, the key points and thus the hash for any face. And when you add a new face, you're creating the hash for say, me, David, and all that's happening is matching the network itself.

Isn't learning David, the network is learning how to characterize a face. And then at the end, we're just matching if the characteristics of the face in front of the camera match the enrolled characteristics for David. So, you're seeing some people beginning to talk about learning at the edge, but it's very new.

I don't know if anyone's really deployed it. And with Ergo, we are not, we're all focused on inference.

Peter: [00:10:05] Okay. Got it. Thank you. Maybe we can talk a little bit about the ergo chip, because I think it's headline is a fairly significant performance at a very low power. So how

does, how would you sort of characterize that in terms of what it is able to accomplish? How does it compete with other solutions out there?

David: [00:10:27] Sure. So the headline is four tops at 55 tops per watt, doing real neural network work in 20ish milliwatts, give or take. And for sure, the interest we receive is heavily around being able to do a useful amount of work at incredibly low power.

And actually, the offshoot of incredibly low power is very low heat. So, in terms of the kinds of jobs that can do, we've talked about it, all your usual detectors that kind of foreground background segmentation. Keyword detection and some basic speech audio noise reduction, some basic kind of gesture or pose.

These are all the kinds of examples of things you can do for the technical people. Listening. You can run resonant 50 and multiple things. We've run YoloV5 at the same time as an LSTM based audio network or a TCN running audio, face recognition is actually three networks in a row.

So, you can run a reasonable amount of things that allows you to get real work done. But we're not at the scale of some of these very large chips, the chips you would find in the cloud, a hundred tops, 400 TOPS, 500 TOPS, but they're burning tens of watts. We came from the perspective of, to be in the edge and to be exciting and fit with the needs of battery powered and small products.

This has to be measured in milliwatts. You don't spending even more than a hundred milliwatts becomes a challenge and most edge chip sets. Today are 1watt, 2watts, 3watts they're single digit tops per Watts. Some people are really excited about their 10 TOPS per watt. So the reason people look at us and why we're so competitive is we're doing useful work on par with other people, but at a level of power, 10, 20, or more times lower. And that's the key differentiator for many.

Peter: [00:12:33] Yeah, I know. I think for me, that's the most startling aspect of this that you're able to, as you say, achieve, quite significant workloads, but at with a very small footprint and a. Low power consumption, which really seems to unlock a lot of potential use cases in, in these sort of edge devices.

So in, in terms of the sensors that you can hook up with this, we talked about cameras talks about audio, other things sort of mechanical accelerometers and so on. What are the limitations in terms of the type and number of sensors that you could hook up with an Ergo chip or an Ergo chip based device?

David: [00:13:17] Yeah. So, it really comes down to the, the IOs we have on the part. So Ergo can support up to two cameras at the same time two stereo microphones at the same time. And honestly, the vast, vast majority of what we've been asked for by customers is audio and video. We do have. The connectors to support other types of sensors, ITC, I3C, you are at a series of GPIOs, so we can support a number of other sensor types.

But honestly, we haven't had occasion to really do that much and have been very focused on audio video just because that's where the market is taken.

Peter: [00:14:00] Okay. And in terms of memory, is the memory embedded in the chip itself? Or do you use some sort of external memory? How, should we think about how that would be configured?

David: [00:14:09] Yeah. So, a big difference for us versus others is we use no external DRAM, which is really important. It saves space, costs and power. So, there is SRAM on the part. A big part of our secret sauce is how we take these really interesting networks and applications and make them fit on a relatively small part. Ergo is seven by seven millimeter package and requires no external memory. And so that physical footprint and the resulting space savings, money savings, power savings is a big part of what's interesting. And probably a little hard for some people in both the audience and our customers to wrap their head around.

But it's our tools and our technology that really allows you to take. That was really interesting applications and fit them on this tiny part resulting in that low power.

Peter: [00:15:00] Hmm. And you, you talked earlier about the fact that, with many of the edge-based solutions that send data back and forth to the cloud does pose a security risk.

And I think, one of the aspects that Perceive has talked about is the. The ability to do a lot more locally and therefore contain some of that risk. Can you talk a little bit more about how you, manage the privacy aspects and security risks by keeping everything much more locally?

David: [00:15:32] Sure. So, this is a natural offshoot of what you can do locally on the device, as you were saying, Peter. I'll take the face recognition example because it's perhaps one of the ones that's more of a hot button. It is a hot button, too many people because video is going to the cloud. It's being matched in the cloud. We're recording, we're tracking. And so kind of the big brother aspect.

And yet I like face recognition on my phone and I might want face recognition on, say my door bell to let me in, when you do things at the edge, depending on how you architect the device, you could send video to the cloud, but you can also be as simple as say an a door lock or a doorbell. I enroll my face. I walk up the image sensor, captures the image. Ergo runs the network and does the match or not at all, it could pass upstream. David or unknown, or it could even just work out with the system lock or unlock. So, from a privacy perspective, the image, and in fact, potentially the naming never left the device.

In fact, depending on how you set it up could have never left even the sensor and Ergo. And so it creates a lot of possibilities to enable features while still respecting privacy, because that video data isn't sent further along or even biometric data can be kept local. When you can keep detection and the resulting application local, then you're back to face, unlock on our iPhone, which didn't have people all worried from a privacy perspective, the way face recognition is in cameras.

And so, it's this different. Local isolation that allows the privacy aspects to be respected while still getting the fun features that we want. And so, it's really how you implement things

that will determine what's possible, but allowing you to do the networks at the edge creates that capability.

Peter: [00:17:43] Right. So. Looking at how this is being applied. So, you mentioned you've got quite a few customers that are developing different solutions based on the Perceive an Ergo platform. So can you talk a bit about where the majority of those. Designs are being sort of focused.

Is it a lot in this inner security camera, domestic security area, or is it more in sort of home appliances or is the interest more in an enterprise application? Where is that going?

David: [00:18:21] So I'd say when we started out, we really focused on, I'll say home security, consumer IoT cameras. That's where we started. And what was really interesting to me was how quickly interest ballooned very broadly people came to us. Things that as a product manager, I hadn't even considered as markets. One such example, being with COVID, we've all gotten behind zoom or, teams, the video conference market, particularly hardware cameras, laptops have all shown up with their challenges and what they want to solve.

That it's not a market I even considered when, doing all that work, leading up to creating and launching perceiving ergo. So, we're seeing a pretty diverse set. I'd say the focus laptops video conference equipment. The security camera space and then the burgeoning space of advanced wearables.

I think that's going to be a space where we're going to see a lot of activity. Facebook put out their limited kind of Ray-Ban glasses with cameras, speakers. Snap has shown they're more interesting glasses. Obviously, we've had Magic Leap and Holo Lens around for a while, and I think these things we're working our way towards consumer interesting devices. And they're so focused on space, power, heat, and have a complex set of things they want to do that. I think they're a great fit for the offering we have with ergo. We haven't as Percieve really at industrial applications as much because our backgrounds are all consumer. I think there might be space for ergo there. We just haven't pursued it and really focused there. So, it's been mostly consumer products. Like the ones I was talking about.

Peter: [00:20:14] Right. And in terms of the support that you provide to customers, I mean, you mentioned about the, secret sauce of fitting in a complex machine learning Algorithms and networks onto, what is essentially a pretty small chip. What else do you provide in terms of software and customization, reference designs? Anything like that?

David: [00:20:38] Sure. There's the two approaches you can take. Ergo is the solution and we have. Person detection, face recognition, audio events, like glass break. And that's kind of where we started with the growth of neural networks and machine learning.

Really, as a discipline, we really seen customers really want to move to what I'll call a platform usage, where they want to bring the networks and their proprietary data to do interesting things with Ergo. And for those people, we have a tool set, which is. It's about bringing your network to ergo and then an SDK for all the software related work of

plumbing, your network within ergo to the sensors, etc, as well as plumbing, ergo to the outside world.

Are you connecting to an SoC or a microcontroller? Everything from as simple as how do I turn ergo on or off, or load a network onto ergo versus how do I manage. Internally or externally frame rates and microphones and sensors. So, we have that set of tools for both. Let's call it the network piece.

And then the system piece, in terms of reference designs, we don't really have reference designs, but we do provide samples. So, code samples and examples so that people can kind of get off to a quick start. As diverse as it feels, the world of applications and neural networks are, it's actually not that many tasks.

When you look at say image classification or object detection. And so you can actually kind of have a handful of examples covering. The majority of the use cases, people are interested in to get them started in their work.

Peter: [00:22:20] All right. So just before we wrap up David looking ahead into the coming years, what are you most excited about? Where do you see this?

David: [00:22:31] Oh, so I think there's a lot of applications and I think both Perceive an Ergo and the industry in general has an incredible opportunity to really reshape how we think about and interact with devices and really make devices that work the way they should. We've talked about it.

It's like having, you want these devices to work almost like there was a person operating them for you, not something where you have to kind of be very literal. Yes, you can talk to the device, but if you have to give it a string of literal commands, it was easier just to walk up to the light switch and flip it.

So, I kind of see for us some interesting areas that I'm excited about. One are those wearables. I think it'll really transform your ability to interact with people. Things like live translation or reading signs or not staring down at your phone for maps, etc. I think drones and robots when they have real intelligence to navigate.

We gave the robot example, but if that, home vacuum is driving around and keeping track of where things are, how many people are always wondering where their shoes are or where their teenager left something in a room that. Penn or triple , having the robot vacuum going around and just knowing, keeping track of where things are, could be really valuable all the way through.

I think we've already seen home security drones that can fly around, which I think would be a great target for many, a cat leaping off taken out of the sky, or the drones, I'm into driving and cycling and, the automation of drones following get fun action photography. These are all things that people are working on and intelligence will really help guide and frame what they're doing.

And then appliances. I am a terrible cook, married to fortunately, a woman who used to be a chef. There's a lot of opportunity for automation simplification. And my favorite example is the microwave. No one knows how to use a microwave. They know the 30 second button and the popcorn button. Why can't the microwave give me my Star Trek Next generation, Picard used to say tea, Earl, gray, hot. I don't even want to do that. I want to walk up to it, holding a mug with water and it goes, there's. There's his tea. He likes his tea hot, and I just put it in and maybe hit a go button. There's a huge opportunity to improve the UI.

The interface that is needlessly complicated on a lot of devices. And for me, it's kitchen appliances to simplify things, to do what we expected, not burn the toast. Everyone burns toast and toasters because you're setting time and slice. Toaster just look and go. That's how dark you likes it. It's done, , these aren't that hard.

And so it's those kinds of things that I think really will be the next step. Not some crazy Jetson's future, but a lot more advanced than we have now. And, and it's not that far away it's I can already see how you would do it. And so I think this is within our kind of three five-year horizon.

Peter: [00:25:44] It reminds me of the I can't remember the machine was cool, but there was a machine in Hitchhiker's guide to the galaxy that always produce something that tasted almost entirely unlike tea.

Whenever the author didn't ask for something from it like tea. But anyway, David, this has been a fascinating discussion. But I think we're going to have to wrap it up here. So. That was David McIntyre, vice president of marketing at Perceive. Thanks, David. A fascinating discussion.

David: [00:26:14] Thank you, Peter. Appreciate it. It was a lot of fun.

Peter: [00:26:16] All right. And for everyone out there, thanks for listening to this edition of the counterpoint podcast, and please check back in for the next edition. Thank you. Bye now.

