**Webinar\_ AI & The NPU - Made with Clipchamp**

0:19  
Hello everyone, and welcome to our webinar on understanding AI and the NPU.

0:23  
I'm Alexa Orban, the Director of Marketing Communications at Cephable, and I'm thrilled to be hosting the session today as part of our Enterprise series.

0:31  
We'll be diving into some exciting topics around NP US, the latest advancements in accessible technology and our ongoing initiatives with Microsoft and Qualcomm.

0:41  
Today we have two fantastic guests with us.

0:44  
Brynnan Fink is a Principal manager at Microsoft with the Windows Silicon System integration.

0:49  
She and her team are responsible for the Windows developer platform and tool chain integration.

0:55  
Alex Dunn is the CEO of Cephable, an innovative, accessible software that enables users to control devices through voice commands, facial expressions, body gestures, and more.

1:06  
Before founded Cephable, Alex has held key roles at Voicify and other companies focusing on software architecture and engineering.

1:15  
His dedication to accessibility has earned him recognition, including being named Microsoft MVP and receiving the Forbes 30 Under 30 Boston Award.

1:25  
Before we begin, Please note that closed captions are available and a transcript of this webinar will be available after this session.

1:33  
Feel free to drop your questions in the chat as we go along.

1:36  
So let's get started.

1:38  
To kick things off, let's let our viewers get to know you a bit.

1:41  
Alex and Brynann, could you share a fun fact about yourselves or tell us a hobby that you enjoy outside of work?

1:49  
All right, I'll kick it off.

1:52  
So one of my favorite fun facts is that I try to spend as little time at sea level as possible.

1:57  
So when I am not at work partnering with fantastic folks like you, I like to spend time in the mountains.

2:04  
I picked up a hobby of mountaineering a few years ago and had the opportunity to actually summit Mount Rainier here in Washington state.

2:11  
And I've also really gotten into scuba diving.

2:13  
So I've had the opportunity to scuba dive all over the world and hanging out with the fishies under the water is definitely my happy place.

2:21  
So Lorraine on at sea level is in always above or always below.

2:24  
I love it.

2:26  
I love it.

2:27  
I guess my fun fact is it's something surprisingly not a lot of people know about me, but I I play guitar actually in a band that specifically does developer parody songs of classic rock songs called the Line Breakers.

2:41  
I go to Line Breakers dot bands founded by Dylan, Beatty and a handful of other people.

2:45  
But we we typically tour during developer conferences and we're usually like the after party show for those.

2:52  
So the biggest stages I've spoken on are the same stages that I've also played guitar on.

2:56  
So I get to sort of live out both my my childhood dreams of working in technology and still being a Rockstar at the same time, which is super fun.

3:03  
That's incredible.

3:05  
Yeah.

3:06  
Those are two really good fun facts.

3:10  
So Brandon, I also know you're coming to us from the Inclusive Tech Lab.

3:14  
Can you tell us a little bit more about this?

3:16  
Yeah, absolutely.

3:17  
The Inclusive Tech Lab is our space here, dedicated on the north side of Microsoft's main campus here in Redmond.

3:23  
And it's a space that's specifically set aside to engage with folks in the disability community as we consistently strive to bring mindfulness around inclusive tech and accessibility in all of the products that we design.

3:37  
That's awesome.

3:37  
I know.

3:38  
I, I think something that viewers would also love to know from both of you is something that sparked your passion for making tech more inclusive.

3:45  
So you're in the lab and you can talk about it too.

3:48  
Yeah, absolutely.

3:49  
I think working at a place like Microsoft that has the ability to create products that touch and serve folks across so many different areas, both at home and in the workplace, it's an incredible platform to be mindful of all of our end users and to find ways to make that tech work best for people.

4:08  
I think one of the great things about Copilot Plus PCs, which you may have heard in the news recently, with the integration of AI into the tech itself, it really allows folks to come to a device and have the device and meet them where they're at in ways that has not been as capable in the past.

4:24  
And so as we lean into that across the board, there's a lot of really cool applications in the accessibility space and I'm super excited to be here.

4:32  
Awesome.

4:33  
First of all, I have to say the inclusive tech lab is so cool.

4:35  
I've been there and I'm, I'm find myself lucky to have been there a few times.

4:40  
So many cool demos and things you can try out.

4:42  
And just like it's, it's such an awesome space and so inclusive of the entire community.

4:47  
For myself, getting into the accessibility space and, and really disability inclusion as a whole.

4:52  
I think we've shared this story a bunch of times, but for me it started with supporting my younger brother with a disability, trying to keep up with his friends in games like Minecraft and Roblox and with his schoolwork at the time too, which during the pandemic was all digital.

5:05  
All of a sudden it can become a little bit harder to keep up with and sort of started exploring alternative inputs with new forms of speech recognition and camera based inputs and gestures and stuff like that.

5:16  
Just to add to, you know, the repertoire of what you can use to control your computer on top of existing hardware devices, whether that's a keyboard and mouse, an adaptive controller or something like that too.

5:29  
Thank you both for sharing your stories today.

5:31  
And I think it definitely leads into so much of what we're talking about today.

5:35  
Brynnan, you did bring up there has been some great announcements and some new products that have come out in the recent weeks.

5:43  
So let's start with some of the most impactful advancements that we've seen inaccessible tech around some of these announcements too.

5:51  
Yeah, absolutely.

5:52  
So just kind of baselining copilot plus PCs is the new platform that is inclusive across all of our OEMs within the Microsoft and the Windows community.

6:04  
And so these devices are defined by advancements in minimum requirements for both memory and storage, but the most critically by the presence of a powerful NPU.

6:15  
So for our launch devices, these have 45 trillion operations per second worth of capability as a neural processing units.

6:23  
That's what we mean by NPU.

6:25  
And this device, this NPU has been tuned specifically for AI workloads.

6:30  
So that's really where that advancement comes into play.

6:32  
We've had CPUs and GPUs and workloads that work really well on those devices for many years now.

6:39  
But with the advancements in AI seen across the industry, those are hosted in the cloud, a lot of cloud AI applications, but the NPU is a local processor.

6:48  
So being able to run AI workloads locally, quickly, securely and without having to pay any additional cloud computing costs creates a lot of opportunity.

7:00  
We've had an opportunity to partner with a lot of our high ambition software partners throughout the Windows ecosystem.

7:06  
Cephable has been one of those that I think is just a fantastic example of what is possible with the implementation of an NPU, especially in accessibility space.

7:16  
So Alex, maybe I can toss it over to you.

7:19  
You will be able to best represent your tech, but it's been so much fun partnering with you and seeing what all y'all are able to do.

7:25  
Yeah, when we first connected with Brynnan, learning about these NPUs and these new devices there really this whole new category of compute, it was so exciting for us because, one, we do a lot of machine learning and AI work with Cephable, right?

7:39  
Like our speech recognition is running machine learning models on your device.

7:42  
Our camera pipeline for face expression controls and head movement controls, it's machine learning models.

7:47  
It's computer vision models running on your device.

7:49  
And so prior to being able to do that on an NPU and dedicating our processing to a processor that is built for these types of workloads, we either had to run it on a CPU or a GPU, which can work OK, but it also means that other apps that need the CPU and GPU might face challenges.

8:06  
So for example, like we have some folks that play games with Cephable, but when we're trying to use things like camera controls while playing a game at the same time, both the game and several might be competing for resources from the GPU in order to run the machine learning models in an accelerated fashion there.

8:22  
Which means that some of our users, especially with lower end GPUs, would have to go like into their graphic settings and like turn down their graphics.

8:29  
Or for folks using it at work, might have to like close certain applications in order to make sure that they're not like overloading their computer and and facing these sorts of performance issues that that come down to it.

8:41  
And so when we learned about the the NPUs and the opportunity to collaborate with Microsoft and with Brynnan's team and with Qualcomm as well to be able to bring Cephable's machine learning models to the NPU, it was a massive eye opener for us on the opportunity to support our users further right on these new devices.

8:57  
We can run Cephable's controls on those new NPU's.

9:01  
It means that there's no impact to the CPU or GPU.

9:04  
All your other apps can run there.

9:05  
You don't have to compromise on graphic settings or the number of apps you're running or on battery life with having to like be plugged in while you're using your computer the whole time.

9:14  
You know, it's, it's really been incredible to see what we can do with, you know, adaptive controls without compromise.

9:20  
I really want to like emphasize that point you made there at the end about the battery life benefits.

9:26  
So to your point, you can run a lot of these AI workloads on the CPU or on the GPU.

9:31  
So you are paying the penalty of then having something that's sitting on those processors you may otherwise be utilizing for your other workloads.

9:39  
But the key benefit that one of the key benefits that I'm really excited about is the NPU is optimized for AI workloads, meaning that NPU is sipping power instead of chewing through power.

9:52  
And that leads to incredible advancements in battery life and the ability to continue to keep your device, do everything you need to do on the go without being tethered to your power supply.

10:03  
That's such a good point, Brynnan, because, you know, it's not that NP US are new, it's how you're utilizing the NPU within your computer and also having softwares that are able to utilize the NPU.

10:13  
And I love that analogy of, you know, instead of it chewing, it's just taking smaller amounts of power.

10:20  
And you know, for the everyday person that will now be computing with NP US other, you know, some of the other advantages that you guys were talking about, just as you know, the all day use.

10:30  
I think this also translates to a lot of folks with disabilities that need accessible tech in their work day and assistive tech in their everyday.

10:38  
So if you guys want to talk a little bit about what this means for accessibility, I think there's a lot of tie in there too.

10:44  
Yeah, I can probably start there.

10:46  
Just given how much we work in the accessibility space and with our users with disabilities.

10:52  
I think the, the main thing that we see so often with other device users is either having to compromise on using their assistive tech with battery life issues, right?

11:04  
So being able to actually run a speech recognition or a speech to text or a dictation software, whether it's sevable or others, or having to bring other hardware into play in order to actually have these types of new inputs available typically means that they're also chewing at that battery, right?

11:20  
That you really can't leave or be mobile with your controls with you.

11:25  
And that's sort of the, I think one of the biggest like real life impacts that these new types of devices and tools like Stuffable have together is removing that tether.

11:36  
We're starting to see users excited about these new devices that are just rolling out now and what it could mean for them with being able to bring their computer and their workloads and all the things that come with that, right?

11:46  
Like their digital life with them in an easier way without having to be literally tethered to an outlet somewhere because they know that they're assistive technology, whether it's new hardware or it's new software, it's just gonna be consuming power all day long.

12:02  
So having all day battery life while running like Cephalole controls is, it's bananas how cool it is.

12:08  
And it just means that new experiences get unlocked.

12:11  
It also means like in our space, like the NPU's on client devices is very new.

12:16  
And on top of what we already do with things like computer vision models and speech recognition models, it also opens up opportunities for us to explore even further in terms of research and new types of controls and new levels of automation.

12:29  
So for example, what we're starting to look at further is how we can bring Cefable's control generation models from the cloud to the device and how we can allow a user to basically jump from application to application without ever having to configure anything.

12:43  
To just say, turn on my voice controls and turn on my camera controls.

12:47  
And then just use apps, Tell it to open an app, Have the voice controls adapt automatically to what that app is capable of doing.

12:54  
Jumping from PowerPoint to Microsoft Teams to your browser to a game without having to go through this really manual configuration process which so many folks using assisted technology are used to.

13:05  
And it's such a pain.

13:06  
And it's sort of the other half of and the, you know, this disability tax that we talk about where an individual with a disability has to spend so much more time configuring their controls, configuring their devices.

13:18  
I think with that, these new NPUs and new this whole new device category opens up the opportunity for tools like Cephable to just automate on top of that.

13:27  
Yeah, I would just like to emphasize all of those points.

13:31  
Cephable is an fantastic leader in this space of implementing tech and AI workloads on client device.

13:39  
I think for the latest version that was just released what earlier this week, you've offloaded the Vision pipeline onto the NPU and that is one element that is fantastic.

13:49  
We've got a lot of models that are accessible, both open source as well as now as a part of the Windows Copilot library.

13:56  
Or those models will allow you to leveraging models that live within the OS for developers to to implement tuned Windows controls and will allow you to do all kinds of analysis on the vision pipeline locally.

14:16  
But beyond that, like that's just one element is where I'm trying to go with this.

14:19  
There's audio pipeline, there's all kinds of workloads and AI implementation on that.

14:24  
There's also large language models, locally running small language models, lots of different layers and flavors of implementation that can be explored.

14:34  
This is just the beginning.

14:36  
That's awesome.

14:37  
And and just to break down a little bit of that for some of our viewers that you know, are not super familiar with how AI is just changing on device and how there's the localized AI as well as cloud AI as well.

14:52  
Can you explain that a little bit further on how that'll be utilized with the MP US?

14:56  
Yeah, absolutely.

14:58  
So to your point, cloud AI super powerful.

15:02  
This is often farms of GPUs that live somewhere away from where you are sitting.

15:07  
At that point you're prompting commands in your device.

15:11  
It is then connecting to that cloud where the workload is rendered and then beamed back to your device.

15:18  
The NPU, because it is local, has a few benefits and a few drawbacks which we can touch on.

15:23  
But the core benefits, I like to talk about 3 core benefits.

15:26  
One is lower latency.

15:28  
If you think about having to move that data somewhere else to be processed and returned, there is inherent latency in that transfer of data.

15:35  
So being able to do it locally, it happens and is instantaneously on your device.

15:40  
It's also more secure.

15:41  
That data never leaves your device.

15:43  
So when you start to think about a vision pipeline, for example, having that being done locally means that your device and all of the stuff in your environment that your device is sensing is staying local to your device.

15:53  
So that's a really fantastic benefit.

15:56  
And then lastly, because you are doing that workload again on your device, on your NPU that you own, you're not having to pay a third party to rent space in their cloud in order to do that processing.

16:09  
The contrary of the benefit or the contrasting detriment to the local is that it's not as beefy of a processor.

16:20  
It is just an NPU that lives in your device.

16:22  
You're not getting to leverage a huge farm of devices that can do huge beefy workloads.

16:27  
So that's why it's really important to have optimized models.

16:30  
And that's why it's really important to ensure that you're leveraging the right models in the right applications.

16:37  
And that's something that Cephable has done really well in your latest release and something that will continue to partner to ensure that we have Windows optimized for developers across the board and across the inclusive tech space as well to leverage to ensure you're right sizing the workload that you're running locally.

16:55  
So you get all of those benefits and minimize any of the drawbacks.

16:59  
Yeah, that's great.

17:00  
And Alex, I'm not sure if you have anything to add on the software side of things, but I think that could tie it together explaining to our viewers the storyline with the NP US.

17:11  
Yeah, I think Bernie, you talked a lot about especially from like the user perspective on what these new, especially the optimized models running on device means.

17:19  
I honestly think it's extremely exciting from a developer's perspective.

17:23  
And of course, I say that because we're developers of software on these new models and new devices, but these workloads that are typically can be really expensive to run inference wise on cloud.

17:34  
Being able to run those offline in an optimized way, not only can it unlock new developer experience and new ideas to create, but it really can do that even for a developer that's working on a solo project, just trying to explore a new opportunity or a new thing that they can create.

17:49  
Whereas like speaking from experience, I've run up my cloud bills so high at some points because I'm just trying so many things and I don't really necessarily realize how much I'm using of these larger computes.

18:02  
But being able to now really even get into experimenting and researching with generative AI capabilities, with vision AI capabilities and so much more is, is super exciting.

18:11  
One of the ones like even specifically that us on the engineering side itself, we are so excited about is like the five, three vision models being optimized.

18:19  
So for those who don't know, 5-3 is like a family of models built by Microsoft, but there is now optimized version specifically for the Copilot Plus PCs.

18:26  
They're super easy to use with the new SDKS, but it also means like we, we can start to explore things like understanding what's going on on your screen entirely offline on your device in a private way to then help you actually execute against it.

18:39  
Thinking of like, you know, the the sort of future of like screen readers, for example, is super exciting where typically screen readers, NVDA, JAWS, narrator, things like that are so dependent on developers building well organized, well structured accessibility trees into their apps.

18:57  
We all know the web is not that accessible.

18:59  
Statistically 98.6% of websites are still not accessible if for the basic minimum requirements for screen readers.

19:06  
But being able to basically like potentially anyway jump some of that into more advanced screen readers functionality and helping user understand what's going on on the visual side of their screen all offline.

19:20  
It's like, it's so exciting.

19:22  
It's things that we're definitely going to be exploring on the Cephable side with how we can empower more folks to actually get the most out of their digital experiences without having to compromise on, hey, you know what, this website just isn't accessible for me because the developer set things up wrong.

19:35  
And so I guess I'll just move on into actually sort of creating a a true equitable experience as much as possible anyway, from the ground up.

19:44  
Yeah, yeah.

19:45  
And it's even with the integrations like Copilot, some of those new offerings that are that are able to happen, you know, people searching back in their history or searching through photos, like just allowing them to utilize their PC as even more of a search engine within their own computer, I think is just.

20:02  
Something that hasn't been as accessible for people.

20:05  
So it's great when you utilize this ecosystem, like we often say with accessible tech, right?

20:10  
There's no one-size-fits-all.

20:12  
But when you expand and you allow people to utilize many apps to come together, it allows them to create a toolkit for success, whether it be for assistive tech users or everyday computing users.

20:27  
Great.

20:27  
Well, I'd also love to talk about just the future of this tech and how tech companies should plan to balance their performance and accessibility when they're designing new tech.

20:38  
So beyond what we're seeing today, we can talk about the future a little bit as well.

20:43  
Yeah, absolutely.

20:44  
I can share a little bit from the Microsoft perspective.

20:47  
This is our first wave of copilot plus PCs, which we are so excited about.

20:53  
We've been working up until launch with a few of our high ambition partners like Zephable to really get a taste of what's possible with this platform.

21:02  
But now that they are publicly released as of two days ago, I'm so excited to see what the rest of the Windows ecosystem and the rest of the community builds on top of this revolutionary hardware moving forward.

21:15  
We expect to see this continue to proliferate throughout the ecosystem.

21:20  
So Copilot plus PCs start at 999.

21:23  
They are in a more price segment.

21:25  
But as we continue to see this tech proliferate throughout the ecosystem, we're excited to see it show up in all areas of the product stack and to see the tech continue to grow and become even more powerful to leverage on device.

21:42  
Yeah.

21:42  
And I know I've sort of spoken on like where we're really excited about the future in terms of being from a developer's perspective.

21:50  
I think the more that we see this device category sort of expand and the more that it does get into the hands of developers, we're gonna see some really exciting things.

21:58  
And when it comes down to like accessibility specifically, I mean, you know, an area of what we do at Selfable is not just here's our app and here's the controls for keyboard and mouse things on your PC, but it's also a tool for developers to integrate into their apps and devices and games and services to make their experiences more accessible and empower be able to bring their personalized controls to your own apps and games.

22:18  
What's really cool about that is like, we're not just talking about like, hey, how can I make, you know, the windows on a PC more accessible as a baseline with tools like Sephable or other AI based assistive technologies?

22:29  
It's about how we can empower other developers to do the same while still not compromising, right?

22:35  
To be able to take Cephable's models, run them natively within your own applications, but also know that that's gonna be running on the MPU, which means the rest of your app experience.

22:43  
You don't have to worry about saying like, do I have to increase my package size or do I have to worry about my app being slower because I'm adding accessibility controls?

22:52  
That all just kind of goes away, and it makes it really easy to say, like, why don't we just add new levels of accessibility, new forms of controls, new ways of helping users understand what's going on that just works best for them into your own apps and your own games.

23:06  
So I'm super excited to see where developers can really take this as a whole new platform.

23:11  
Yeah and building in that, you know accessibility from the ground up because often, you know it's that afterthought to add accessibility to products that already exist.

23:21  
But with the AI capabilities, I just think that the possibilities are truly endless and the launch, you know of these new copilot plus PCs, that's what we're seeing is is the future that the NPU is here to stay and also be utilized in really interesting ways.

23:36  
So if there's anything else exciting, Brynnan, that you want to add as as far as that future, I'd love to hear.

23:43  
And then we can wrap everything up.

23:46  
Yeah, I think going back to the Windows Copilot library, especially for developers, such a cool tool.

23:52  
So this is a Windows SDK that exists for developers to pull AP is from that have been optimized for Windows.

24:00  
And so this is everything spanning the like image segmentation, object character recognition, all the way to small language models like by Silica.

24:11  
So taking advantage of what currently exists in that library and also keeping an eye on it as it continues to expand and unlocks additional capabilities that are optimized for Windows is something that I'm really excited about.

24:24  
We've got the hardware we've got now the AP is that you can build with that, take advantage of all of this cool AI tech.

24:32  
And when you wrap all of those things together with fantastic developers within the Windows ecosystem, this is going to span across the industry.

24:40  
And I'm a really decided with Windows and Microsoft's commitment in the accessibility space here in the inclusive tech lab to see all of this come together and just continue to drive the bar forward.

24:53  
Yeah, yeah.

24:54  
And I think that's a great point to Brynnan that it's, it goes from, you know, the individuals that were impacting all the way across into the industry, because I, I just don't think we realize just how many people will be able to impact with these advancements.

25:08  
And like we said, you know, it's, it's just the start with the launch.

25:12  
Well, thank you both so much for taking the time today.

25:16  
I think this webinars is just a great example of, you know, how people can start to learn about NP us and also just that future of AI, because AI really is here to stay.

25:25  
And it'll be something that's a part of every device, but also our everyday when we chat to BT as well.

25:31  
So we're looking forward to sharing more resources on the NPU.

25:36  
And then of course, our future partnership opportunities with Microsoft and Qualcomm.

25:41  
But thank you everyone, for listening in today.

25:43  
And if you had any questions, we'll be answering in the chat and our live premiere on YouTube.

25:47  
So thanks everyone so much.