05242023 Build Panos Panay Keynote

Build 2023
Panos Panay
Shilpa Ranganathan, Pavan Davuluri, Cassie Breviu, Stevie Bathiche
Wednesday, May 24, 2023

PANOS PANAY: I love it. I love it when people get pumped about Windows. A little bit pumped about Windows a little bit. Me too.

(Applause.)

All right. I'm pumped also. I'm pumped to be here at Build. I'm honored. I have to tell you, first, I just represent this incredible team of product makers. They're in Redmond. They're around the world. They're watching right now. Let me just say thank you for all you do, for all you put together. I'm grateful. And also thank you to everybody here. Thanks for being here today. It's awesome to be at a live event together. And for those of you online, welcome. It's an exciting day. I think when you look at everything Rajesh talked about, just so, so amazing. It also happens to be an incredible time to be a developer, especially a developer on Windows.

Now, we've talked a lot about AI. You've heard a lot about it. It's clear. It is driving the largest technology shift of this generation. There is no doubt, the possibilities across industries, healthcare, finance, education, tech. Even in your own homes, in our homes, the possibilities are endless. But we're moving so fast.

Sometimes, every now and then, we can lose sight of what that actually means for us as people. What is that transformation for each of us?

Now, I'm going to do what I'm not supposed to do on stage and ask a question. You never do this in a keynote, just so you know. When you ask a question, if people don't answer it, it's a terrible moment, so just stick with me. It's so, so true. Don't do it. You always lose your audience, let me just tell you.

How many of you remember the first time or the early days of the Internet?

All right, good, we've got a lot of hands, that helps.

You heard Rajesh talk about GUI and the transformation that started to come but let me just shift it a little bit. I remember for me and some of you, I'm looking out there, and you're like, "What do you mean? The Internet's always been there?"

So I'm dating myself. And I saw you, I'm like, "Oh, right, you were born in the Internet," I got it.

For me, I remember I wanted so bad to understand what it was. It seems crazy, right? Because it seems so obvious now. But I wanted so bad to understand what were people talking about? And I

had this connection, and it made these noises, and it was this dial-up modem. And I tried so hard to get on, and I use this thing called Gopher. I was like, "What is this? Why do people think it's good?"

Then I got my first job, I got to work, and I had this thing, I think it was called DSL. This is great tech. And now I have a connection. I remember – I remember the first few days at work.

Don't get me wrong, I did a little bit of work, but I also remember being on the Internet, probably more than the actual work I was doing.

It was magic. The feeling. It was – it's indescribable. Even then -- even then, in online infancy, the opportunity felt so vast, it was almost impossible. It was impossible to comprehend. I think those first couple of weeks I read the entire Internet, for sure. I just, I read the whole ESPN catalog. There's no doubt about it. I mean, I read everything. But the level of knowledge at my fingertips, it felt transformational. It was transformational.

Then came the advent of online dating, Wikipedia, social media, the thing in your pocket, that phone, the new device form factors, streaming. And it just goes and goes and goes. But that feeling of immensity, the feeling of possibility, I had never felt it before. I had never felt anything like it. And I haven't since. Until right now. Until right now with you. Until the advent of AI.

And the truth of what LLMs are bringing forward and the opportunity that's in front of us. But it's not just about the features or the models. It's about being part of getting to witness and participate in something that will change the very fabric of how we live our lives.

Did you catch that? We are there together. And it's only the beginning.

So for the devs out there and the devs to be, some of you might just be getting started with AI, like me when I was trying to figure out the Internet. Or you might not know where to start. Or you're looking at me and you're thinking, "I've been creating AMLs for a year. I'm the expert. You're not. Get out of my way."

No matter where you are. No matter where you are on that journey, this presentation is about Microsoft and Windows being there for you.

See, I look at technology a lot like you look at a ladder, trying to just get up the side of a building or a wall or a fence. You don't go from the ground straight to the top. Some of you can. You have to take steps along the way. Windows will give you the tools. The tools each of us needs, and in this room, to be the wavemakers, the creators, the pioneers of this next generation of technology to empower yourself, to empower your customers, to change the world.

And by the end of this keynote, you're going to see a few things. You're going to see how you can increase your engagement, an ability to serve customers with the all new Windows Copilot and AI plugins.

Shilpa is going to come out here. She's going to show you how we're accelerating your ability to build seamless workflows on Windows 11. You're going to see Dev Home, Microsoft Store, widgets and other innovations we've just released. I don't know if you saw this, but we just released the full build of Windows yesterday, which is so exciting. It's so exciting. We're not going to talk in great depth about it today. Shilpa will hit some of it, but get in there, read the blog. It's crazy cool. It's crazy cool.

Then Pavan is going to come out here and I really want you to home in here. He's going to talk to you about how Windows is fundamentally—you will see it—the absolute, the absolute best endpoint for AI, and is the home for every one of your high ambitions, AI powered apps from edge compute and OSS models coming together. And you heard Kevin yesterday, he really nailed this. Pavan will bring that together for you.

Then Stevie will come out and show you innovation and how it's not slowing down. And if you are on Windows 11, whether you're building on it or using it, it is what will keep you ready for what is coming next as we pioneer our way through this opportunity together.

OK, Let's jump into Copilot.

I'm going to come down here. I have to share with you, I – you know, we've been working on this for a while because we get asked all the time. And when you work on details for a long enough time, you're a little bit too close. And so regardless of what it is, regardless if it's a pixel on the screen or one of the hundreds of scenarios that we have for the copilot coming forward for you, you look at it, you thrash through it, you try and get it right.

How are you doing?

But fundamentally, you then want to tell the story as best you can. So we built this video. We built it with bits. We then record them. We then render them. We try and bring them as perfect – I've watched this video at least a thousand times. I've edited it. I've seen it, but I haven't seen it the way you've seen it.

So we're going to watch it again and I'm going to watch it with you because this feels awesome, and I haven't seen it on this stage. So let's do it together. This is what Satya showed you yesterday.

(Video segment.)

(Applause.)

PANOS PANAY: All right. I know, it's super rad. Now, I'm going to frame back again a few things you saw, but I want to talk about them a little bit differently. I want you to come along with me on this journey. Yesterday, Yusuf showed you the demo that was in the video, but there was a lot of purpose to why we picked that. And I want to just share a few small things with you as we look at it together. And I'll show you a few new things. But let's just – let's take a look at this really quick.

So the first thing to understand is I get asked a lot, and it probably seems super obvious in this room, you know? You know how to get to AI; you know how to get to Bing. You know how to invoke ChatGPT, of course you do, but not everybody does. They hear about it. They've heard about the Internet.

So what's the first thing we do? We bring AI front and center to everybody. For you, think of this as a platform, a funnel for all those plugins that you can bring forward. The two days of plugin palooza that you got, the opportunity to take those and bring them to a billion people.

So we put Copilot right there on the taskbar. Just simple. Important. And it comes there, and you heard Satya say it yesterday, every user is a power user. You start to see that, where you can talk to Copilot, ask it what you need.

Think about 30 years of history in settings, in the understanding of a platform and the power of being able to just say, "Make it easier for me," and dark mode shows up because it's easier on your eyes. You may not even know that dark mode existed.

It cracks me up when everybody applauded yesterday for dark mode. That was rad, you know?

(Applause.)

But that power uses – think about any setting within Windows and – but people don't think of it as settings. That's just the platform. They think about it like this. I need to cast my stream to the TV. All right. That's a power user move on Windows. Not anymore. Just write it down. And watch it happen.

OK, now there's a nuance I want you to catch. In the bottom right corner here, you can see Copilot talking to you, not you thinking about what prompt you need. This is the power. This is where the plugin can just wake everything up. Just look at the nuance. It clicks. It then says, "Would you like to listen to some music?" Yes. The plugin shows up. It makes the suggestion. You picked chill vibes?

I would never pick chill vibes, but somebody does. But here's the better part. Watch the next suggestion that just happened. If you missed it, the nuance is important. At some point, Windows recognizes you've made a mess. "Let me help you organize your Windows." You don't have to find Snap, although a lot of you do. You don't have to wonder how; it'll just offer it and it cleans up your desktop. This is speed. This is staying in your flow.

All right, look at this next demo. This is a little different. OK, here is a Kubernetes manifest in Notepad. All right, in this case, I'll select the config, but watch what happens when I do. It'll copy it. It'll drop it there in Copilot for you and then simply ask you, after you decide to give it to Copilot, "What do you want to do with this text?" You clearly copied it.

So for me, I've always dreamed of being a dev. Don't worry, I practice. I try, but I'll never be able to claim my emblem. But I will tell you, the idea here, where I can get to config, is just to

ask it to explain to me. The time it saves me and what it might mean for you, now think about this on third-party applications. Think about it on anything that sits on your desktop today. What could you do with it? I'm giving you the most simple of demos. But it's powerful. Unpack that thought and work through it. It's really kind of cool, isn't it?

Now go to the audio files. OK, I record voice recordings of myself all the time. I know that sounds weird, but I do talk to myself. The best way to capture what I'm saying is just to listen to myself later. And I can't stand my own voice. Does anybody else like that? But I will listen to it. But here's what's beautiful. Just grab the file, drag it over, make a decision that you just want to hear it translate for you, transcribe for you. And just like that, you're saving time.

I only put in this demo because it's the one I wanted.

(Applause.)

It's pretty powerful. This is great. This is going to be perfect for my father when he drags over a note, a voice note or even a video in the future and he wants to translate it into Greek, he's going to do so. The power of that, the speed of that, the possibility is endless.

OK. I want to take you to this last demo, and it's probably the most important one to catch the nuances on, so I'm going to walk you through what's happening. You can see Microsoft Word here, no doubt. And you've heard a lot about plug ins. You saw the Spotify demo. That was an important demo to just understand. Yusuf showed it yesterday. You just saw it here. That plugin showed up and made sense, but here's another one, and this one is near and dear to my heart.

This is how you move creators into their flow. If you're thinking about the funnel and the opportunity that you have to the billion users on Windows, a billion-plus users on Windows, what can happen here? What you will see is when the user says, "Help me generate a logo," it literally brings up Adobe Express there in Copilot and say, "Click here, let me take you through."

Pause. Understand that. Copilot brings you a plugin and just keeps you in your flow and you're creating. When you're done creating it, you drag it back into the chat and then you can share it with all of your friends.

The steps of opening, closing, removing, they all go away. When we say we're going to help you do things differently, that AI reinvents everything on Windows, this – what you're looking at is just the beginning. There is so much power here. If you think about plugins, with ChatGPT and Bing, what you heard yesterday. Windows is your platform for high-ambition AI apps. It's sitting right in front of us.

OK, you're going to get a ton more here in just a few moments from Pavan, how do you do it, how do you build it, how do we get into it, but let me just share this, Windows Copilot previews in June. So we're a month away. For those of you who can get on it, let me just – let me just make and ask. Go get it in preview. We're going to learn together. I say that from a true point of humility. We don't understand everything yet, but boy, do I know it changes your workflow? Do

I understand for sure the opportunity that's in front of you to bring those high ambition apps to Windows, to have that AI experience, that interjection, and will evolve together?

What did you guys think?

(Applause.)

All right, let's. Let's switch gears a bit and let's talk a little bit more about those apps that are on Windows, because Windows 11, look, it's not just about – you know, it's not only because way beyond AI, we all know that. We all know that as developers, you've known that for years. It's such an important platform.

And we announced Windows 11 nearly two years ago. Two years ago was an announcement right about this time. It was exciting. You've been so gracious. You've helped us build this platform. We have more people at speed getting on this platform than ever before on any version of Windows choosing it.

I'm so proud of it. The usage is growing. The engagement is up. We love it. There's no question, but this growth, it extends first to developers. It's awesome.

Just since the last year, 24% increase in developer monthly active devices. That means you, the group of you, more of you developing. And some of this growth is driven by an increase in segments like Python developers, which is cool because that also points to AI becoming an even more critical tool to reach our new customers. But it's also about the fact that the role of the PC has changed. It's fundamentally changed. It's more integrated into our lives than ever before.

Windows 11 has created a little bit of a transformation for each of us. The PC became the place to connect, play, work. And we also know there's no limit to what you can do from local compute to cloud compute.

And as you listen to Stevie and Pavan talk about that hybrid architecture, which is coming, it's so sweet, there's no limit. But then you look, and you go, there's apps like Spotify, Snapchat, Facebook, Messenger, WhatsApp, so many more that have recently come to the Windows Store that increase—and you'll hear from Shilpa—has been awesome.

And you have to understand, these apps, I know they seem obvious, but they were mobile, just mobile apps, only two years ago. Only mobile apps two years ago. I mean, it is incredible. Sophia, she's 17. I haven't seen her in a few days. She's traveling; I miss her. She uses these apps on her PC every day. I'm blown away. I mean, I love it, but I'm blown away. It's so powerful. You can keep your phone in your pocket, and you stay in your flow. It's almost a dream of mine. And now, the advent of AI is only making that stronger, faster, more opportunity.

WhatsApp is my favorite example. I want to share it with you. They created an amazing, an amazing Windows app. They made a native experience for the PC. People are loving it. There's no doubt it's the No. 1 app on the Microsoft Store. I use it for my fantasy baseball league. I use it for my uncles in Cyprus. It's really impressive.

The best way to understand it, we have Will here, which I'm proud of. He's here to share the news with you with what's happening with WhatsApp on Windows. Hey, Will.

(Video segment.)

WILL: Thanks, Panos. It's great to be here and share with the Build community the work we've done in partnership with Microsoft to reimagine WhatsApp on Windows.

WhatsApp is the No. 1 messaging app in the world, and it is the best way for people to communicate privately. We're always investing in new experiences, and a priority for us has been to work with Microsoft to create an amazing WhatsApp experience on large screens. Our users deserve the best experience of WhatsApp across every device and platform, especially the more than 1 billion Windows devices.

After years of development, we are thrilled to now have brought the first native desktop companion application to our users. And on new Windows 11 PCs, you can easily find it in the Start menu, just a click away. It was built from the ground up, using native technologies recommended by and in close collaboration with Microsoft.

And with the new app, our users will be able to run WhatsApp with significantly improved performance and reliability, benefiting from reduced battery, CPU and RAM usage, as well as deep OS integrations, such as background processing and notifications. Unlike before, the app does not require you to stay tethered to the mobile app after authentication, so you can use WhatsApp even if your phone is off.

Other exciting features, such as group calling, are now available. People can connect with their friends and family through eight-person video calls and 32-person audio calls. And best of all, all of this is backed by end-to-end encryption, so your personal messages and calls remain truly private.

We're working with Panos and the team to deploy cutting edge technology that will make the WhatsApp experience on Windows truly remarkable. For example, we're using ARM64 architecture to enhance app performance while minimizing power usage. We're also leveraging Windows Studio effects, which run AI models on the neural processing unit to create the best video call experiences. We know users are going to love the functionality this makes possible, like background blur and noise suppression.

We're excited to continue collaborating with Microsoft to innovate for the future of private communication. Back to you, Panos.

(End video segment.)

PANOS PANAY: Awesome. Thanks, Will. (Applause.) OK, before we move on, I want to acknowledge something. When I first took on Windows, you, I must tell you, the most passionate group, the most passionate group that gave me feedback, it was devs.

As a matter of fact, funny enough, as I was walking through the aisle, I saw a few of you. You had actually sent me PowerPoints with every single detail of what needed to be changed in Windows. It was pretty awesome. It doesn't stop. (Laughter.)

Now, I'm going to say the feedback was not gentle. (Laughter.) It's my best way of saying it, but it was necessary. And I want you to know, we're listening. It's so important. You've inspired us to improve everything from the details in the UI to foundational OS changes, like the new file system optimized for developer workloads. You continue to push us for better experiences, whether it's across Terminal, WSLs, GitHub integration. And yesterday, you saw another response to that in Microsoft Dev Box.

But today, we want to bring you the best Windows development experience yet. We want to build a home for devs. So, we did, and we called it Dev Home. Take a look.

(Video segment.) (Applause.)

PANOS PANAY: All right. We've got a lot of exciting stuff to cover from here on out. Shilpa, an incredible product maker, a dear friend of mine and absolutely one of the leaders in Windows, is going to come out here and take you through every single detail, including the latest release of Windows. Shilpa? (Applause.)

SHILPA RANGANATHAN: Thank you, Panos. Thanks so much. Good morning, everyone. It's great to see everyone in person after so long. OK, so let's get into it.

When developers asked us to support Bash on Windows, we knew that meant making Linux a first-class environment with WSL. When developers asked for a modern command line experience, we created one of the most popular projects on GitHub in Windows Terminal. And when power users wanted more flexibility to customize Windows, we launched PowerToys. And together we've made it a top five community supported project on GitHub. You were right on all counts.

And when I say you, I'm referring to you, Windows developers. We love it when our developers give us feedback on how we can make Windows the best place to create software. And the best part, we're still listening.

We've heard from you that it takes too much time and too many clicks to set up Windows for coding. We've heard that disk performance slows down your inner loop productivity, and we've also heard that the Windows experience should have more options for power users. And we're going to address all those three things today.

Now, as you saw in the video, we are really excited to introduce Dev Home, a brand new experience in Windows to help developers quickly get productive and stay in the flow. And we know that when you're coding on Windows, you're working in an ecosystem of both local and remote services.

Dev Home will make it easy to connect to GitHub and set up your machine to code for the repos you care about, easily installing all the tools and packages you need to get running. And once you've connected to GitHub, all your surfaces across Windows will light up with developer functionality, all the way from File Explorer to Windows Terminal.

Dev Home can also configure your coding environments in the cloud, using both Dev Box and GitHub Codespaces. The Dev Home setup experience is powered by a brand new feature in Windows package manager called WINGET config. You can define a dev environment in a single file and apply it to your local system, to your Codespaces and to your Dev Box with a single click. (Applause.)

Oh, thank you. That's great. The teams worked really hard on this feature, and one of the things we've really prided ourselves on is helping you save time. I'm really glad that many of you liked it, and hopefully, you will use it as well.

And Dev Drive is a brand new storage solution tailored for source code packages and working folders. Now, inner loop scenarios often deal with repos containing thousands of files, and these really slow down IO constrained builds. Dev Drive is based on the resilient file system, and it's combined with a new performance mode capability in Microsoft Defender Antivirus. Dev Drive will yield up to 30% improvement in build times over what you get on WIN11 today. (Applause.)

Oh, thank you. Thank you, that's great. And obviously, the less of the amount of time you spend on builds, the more time you can spend creating value for your customers.

Now, once you set up the code, Dev Home helps you stay in the flow by removing distractions with our dashboard feature. The dashboard is a customizable surface that provides a glanceable overview of dev environment states, project information, code reviews, issues assigned to you and CI system status, alongside widgets that help you quickly track your tasks in your to-do list.

Now this dashboard we're showing you is extensible. In fact, we're collaborating with Team Xbox to bring the GDK to Dev Home to make it really easy to get started with dev creation. And because Dev Home is open source, you can contribute to and extend this dashboard as well.

Now this holiday, everyone will have Dev Home in Windows, but you right now can get Dev Home in preview from the Microsoft Store today. (Applause.)

OK, I'm going to switch gears a little bit now and talk about all the improvements we're bringing to Windows 11 to increase your efficiency and productivity. Now these will start rolling out to our Windows Insider program in the coming weeks. Obviously, you've got a lot to read in the blogs. Panos talked about it as well. I'm just going to touch on a few of these today.

We're adding native support for additional archive formats. We also have the ability to easily end a task from Taskbar without opening the Task Manager. (Applause.) Yes, yes! Yes, that is a favorite, yes. And you can quickly identify and access every single instance of each app with just

a single click. Windows Terminal will also now have tab tear outs to help you organize your different shells on Windows. (Applause.)

I'm so glad you're glad! The team's going to be thrilled. This is awesome. (Laughter.)

Now you heard Panos briefly talk about this earlier and I'm really excited to share that Windows Terminal is getting smarter with GitHub Copilot Integration. This brings the power of AI to your command line, making it really easy to run correct commands and troubleshoot errors directly in Windows Terminal. And you can also use the experimental copilot chat right where you need it. To get access to the new copilot features in Terminal, please sign up for the wait list on GitHub.

OK, so now you've used Dev Home to be productive, and you've built a great app. Let's talk about how we can help you acquire and engage customers with your apps.

Today, we have a new release for Windows 11. I know, this week has been fantastic. There's several improvements in here. I'm going to touch upon the widgets board, which enables quick, glanceable content for customers in Windows.

If you haven't developed a widget yet, now is a great time to do it. In February, we introduced support for third-party widgets on the board, including new experiences from Meta, Microsoft and Spotify.

Now the biggest feedback we heard was the desire to customize the content on this widget board. As part of the latest release, the board will now have a larger surface area and a dedicated space for widgets right next to the feed. We're giving you more space. And later this year, we will have additional board layouts and the ability to tune the board so you can display widgets only, the news feed only, or a mix of both. The board will also have space for recommended widgets, and we've added on-install and banner alerts so we can make new widgets easier to discover.

We're really excited to make widgets great for customers to stay on top of what matters to them the most and to amplify your apps and your services. So many new features today. Now I would love for you to get a chance to build your widget and have it reach millions of customers.

We've talked about how we can help acquire more customers for your app. Now let's talk about retention.

Today, for the top 100 apps on Windows, we see a 40% retention after users set up a brand new PC. Now I want to make sure you can retain your customers even if they choose to switch to new devices. Windows is making it easier than ever with your improved restore experience. As soon as the user sets up a brand new PC, they will automatically land on a desktop. This will include all of their store apps right where they need them, and it's on the Start menu, as well as on the Taskbar. (Applause.)

Yes, yes! Yes, yes. No sense in losing customers when they set up new devices. We're happy to have this capability roll out in preview, starting today in Windows Insiders.

Now we just talked about widgets. We talked about restoring apps. And now, it's really important to complete this picture by sharing what's new with the Microsoft Store on Windows.

Now, in the last 12 months, the Microsoft Store has had over a billion unique visitors. Now what is really energizing is that Windows 11 users are really engaged with our store. They come at nearly twice the rate of Windows 10 users. And we're really proud of our open store principles, allowing developers to bring any app type with industry leading revenue share.

And this has led to outstanding success. We've seen double the amount of apps since Windows 11 launched, and more Android apps as well. And if you're around, we have an on-demand session for the Windows subsystem for Android. I'd love for you folks to drop in and see the team there.

In March, we released Microsoft Store Ads in the U.S. market. Now one of our partners who use Store Ads shared feedback that they experienced a 25% increase in their app installs. Next month, we're going to be expanding Microsoft Store ads to 150 regions. (Applause.) Awesome.

Now we begin the next chapter for the Microsoft Store on Windows as we embark on leveraging AI to help you be more productive and expand your ability to reach your customers. We're introducing a new feature called AI Generated Review Summary. What this does is summarizes thousands of app reviews into a few sentences, making it easier than ever for your fans to inspire and influence future customers of your app, and also to provide summarized insights back to you.

Now one of the biggest challenges in getting your app discovered starts with using the correct search terms. Let me show you how AI-generated keywords will recommend which keywords you should be using, based on AI insights that are uniquely available in the Microsoft Store.

And I'm really excited to share that we will soon be introducing the new AI Hub in the Microsoft Store in the coming months. It's a space that will showcase the best of AI apps for our customers. It will educate customers on how to get started on their AI journey and inspire them to use the AI as part of their everyday workflows.

Now you heard me about increasing your productivity with Dev Home, leveraging widgets, getting your app restore experience so that you can engage and retain your customers. And now you can take advantage of AI Hub to enhance your app with AI in the store. You're about to hear Pavan Davuluri share more about the tools that can help you build your AI-powered app and have it show up here in the AI Hub.

Please welcome my friend and colleague, Pavan. Thank you. (Applause.)

PAVAN DAVULURI: Thank you, Shilpa. That was fantastic. Thank you, thank you. That was amazing. It's really exciting to be here with all of you here today.

Yesterday, Kevin shared that Windows is the best client for AI development. And as we just heard from Panos and Shilpa, there are more developers now on Windows than ever before. I'm

here to show you how AI can work for every developer here, whether you're just getting started with the AI or you're really wanting to take it to the max.

Here's what you're going to hear today. If AI is new to you, Windows can get you started with inbox models and APIs. If you're wanting to center your apps and experiences around new AI features, we have ONNX Runtime to help you take it to the next level. And if you're really ready to take AI to the limit, optimize your models, create differentiated experiences, we make that easier than ever with ONNX Runtime, the Olive toolchain, and of course, the Hybrid Loop.

Now I want to talk to those of you who are just getting started or even don't know where to get started. If you're on Windows 11, we have AI models for you to integrate into your apps right now, no heavy lift, no extra work.

You know the Windows Studio effect you saw earlier in Will's keynote? The WhatsApp team was able to integrate those with just a small shift in settings, and you can, too. We're building a Windows AI library for you with Windows Studio effects and a range of in-box models and APIs coming soon. (Applause.) Thank you.

OK, let's take it to the next level. If you're familiar with AI, and you want to be able to tweak your existing models, bring your own models, ONNX Runtime is going to be your best friend. And to me, the best way you can show the impact of ONNX Runtime is to see a developer using this technology.

Camo is an app that enables any camera to produce incredible video, and you're going to see how they've used ONNX Runtime, Windows devices and modern silicon. Aiden from the Camo team is here to show us more.

(Video segment.)

AIDEN: Thank you, Pavan. Camo gets you more than great video quality. It also adds next gen effects like spotlight, AI auto framing, beautiful bouquet, privacy mode, color grading, and drag and drop overlays. These features are great for content creators of all types, whether planning a big meeting, producing YouTube videos or live streaming.

And we're not done making Camo the best third-party native Windows app. With the advent of NPUs on Windows, our roadmaps got even more ambitious. I'd love to show a sneak preview of what we're working right now to bring that vision to life.

Camo's Emoji Hands feature is a visible and natural way to signal a reaction when you're on a video call or stream. Perhaps you want to politely interject with an idea. Lift an index finger and a light bulb appears or raise a hand if you need to. And when you need to drop, peace out. How's this for a hand-wavy demo?

Obviously, there's a lot of ML powering this. Today, Camo uses CPU and GPU to segment the user from the background, to track hand gestures and to spot those hand gestures, all in real time. Up to four models are in play for this.

And computationally, it doesn't come cheap. In an app like Camo, we need to use as little resource as possible, because when you use the live streaming games on Twitch or presenting in Teams, we can't afford for them to see dropped frames, reduced performance or jumpy video. In fact, the only way Camo can run the Emoji Hands models in a performant way today is by combining the power of a PC with an attached phone.

But thanks to the NPU, Camo can now afford to do much deeper image analysis on every frame of live video, using any camera or webcam, all this without touching the CPU or the GPU. It's not just much faster, but it helps overall system performance and battery life, too.

We're confident there's a lot of great things we can build using the ONNX Runtime in the new Olive toolchain. You'll be able to check out Emoji Hands and other great additions later this year when we ship Camo with NPU support on Windows 11.

(End video segment.) (Applause.)

PAVAN DAVULURI: It's great to see the Camo team democratize amazing video experiences through AI. Cassie Breviu from the ONNX Runtime team is going to join us shortly and show us how all of this works behind the scenes.

Before that, to get to the next level, you need ONNX Runtime, the Olive toolchain and the Hybrid Loop. This is a great way for you to optimize your models and create performances and experiences that are tailored for your customers.

Cassie, super happy you're here today. Glad you could join us.

CASSIE BREVIU: Thank you. (Applause.)

PAVAN DAVULURI: Can you tell us a little bit more about ONNX Runtime and the Olive toolchain?

CASSIE BREVIU: Yeah, so sure. Let's start with ONNX Runtime, which makes it straightforward to deploy and run your AI models on any platform, including the web.

PAVAN DAVULURI: Can you tell us a little bit more about what that means?

CASSIE BREVIU: Yeah. ONNX does all the heavy lifting of dealing with diverse platforms and hardware, freeing developers up to focus on delivering delightful AI-enabled solutions. And ONNX Runtime also enables you to target the cloud, which we'll talk about later.

PAVAN DAVULURI: That's fantastic. One thing to note, this is the same development pattern and set of tools we use internally at Microsoft for apps like Office, Edge and Teams. And this really inspires us to build the best product we can before we share it with you.

Now, if I'm a dev who really wants to optimize performance of AI models. I'm going to be targeting the rich and diverse Windows ecosystem, and that can be a lot of work at times.

CASSIE BREVIU: Yes, it is. And that's where Olive toolchain comes in and really shines. Olive brings together all the optimization steps and tools to make this process simpler for developers, from hardware to data processing to inferencing in multiple languages. If you're just dipping your toes in this space, Olive has walkthroughs to get you started. If you're a data scientist that has been creating and optimizing models manually, Olive simplifies your workflow so you can focus on finetuning your model.

PAVAN DAVULURI: OK, to summarize, ONNX Runtime helps you run across platforms, and the Olive toolchain really streamlines your workflows. Let me tell you just a quick story about how Olive came about.

Devs and data scientists at Microsoft working on models were dealing with the same model optimization complexity, I want to say about a year ago or so. And this last year, I think they just sort of threw a fit and they were like, "This is just stupidly complex. We have to simplify this." And so, it was Yuan on the AI Platform team and Stevie, who led this effort to go solve the problem, and that is how Olive was born.

Cassie, let's show them what this looks like in action.

CASSIE BREVIU: Yeah, I'm going to take you on a quick tour of what this optimization process looks like and show you the results. We'll use Stable Diffusion for this demo, which is a hugely popular open source image generation model. Let's look at how to use Olive to optimize Stable Diffusion for better performance with the direct ML execution provider.

PAVAN DAVULURI: All right.

CASSIE BREVIU: First, you'll see here, this is a JSON config, and this is how I tell Olive what I want to do to optimize my model. Now let's take a look at the generation in the C-sharp WPF application. This is going to run my Stable Diffusion models locally, and it watched this progress bar a lot while building this out.

PAVAN DAVULURI: It's taking a little bit.

CASSIE BREVIU: Yeah. (Laughter.)

PAVAN DAVULURI: OK. Coo, excellent. That's the image landscape in about 10-ish seconds.

CASSIE BREVIU: That's right, but we can do better than that.

PAVAN DAVULURI: Boom, two seconds. Fantastic, excellent. (Applause.) Can we do that again?

CASSIE BREVIU: Why not? It's so fast.

PAVAN DAVULURI: It's awesome.

CASSIE BREVIU: Can do it all day. (Laughter.)

PAVAN DAVULURI: (Laughter.) The (inaudible) improvement in performance. That is impressive. If you'd like to go check out and get a deeper view for ONNX Runtime, the Olive toolchain or really, the direct ML execution provider. Please take a look at our workshops this afternoon. Cassie and our ONNX experts are going to be there. Yeah?

CASSIE BREVIU: Yeah.

PAVAN DAVULURI: OK, great.

Now, at last year's Build, one of the concepts we introduced that was new was this idea of the Hybrid Loop. That is about seamless computing across the cloud and the client. It brings together the benefits of the edge, things like performance, and responsiveness and privacy, with the power of the cloud to run large models, to house massive data sets, be able to invoke cross-platform capabilities.

And while the concept of Hybrid Loop can be simple, in practice, it can be complex. Pulling this off requires managing a bunch of services, virtual machines in the cloud. And the beauty of the ONNX Runtime is you don't have to deal with any of this complexity. Azure shows up as a co-processor in Windows, just like an NPU or a GPU.

Cassie, let's show us what this looks like.

CASSIE BREVIU: Yeah. Now, let's show you a simple logic flow like this in your code, where you can control whether your model runs locally or in the cloud. And for the data scientists in the audience, Olive enables this Hybrid Loop by creating a single end-to-end model to reduce the amount of code you need to write.

The first thing I'm going to do is select "cloud inference on," and then I'm going to load my file. This is a quote from a book that I like. I'm going to click "transcribe." You can see in my breakpoint that I'm hitting the Azure execution provider, and I'm using the OpenAI cloud endpoint for the Whisper large model running in the cloud.

PAVAN DAVULURI: Yep.

CASSIE BREVIU: It's going to send that up, and I'm going to get back my transcription.

PAVAN DAVULURI: Fantastic.

CASSIE BREVIU: Now, if I want to run locally, I'm going to turn the cloud inference off, and we're going to run the Olive optimized Whisper tiny model on my tiny, cute, adorable laptop here.

PAVAN DAVULURI: That is a lovely machine, for sure. (Laughter.)

CASSIE BREVIU: And I'm going to choose a quote from the Kevin keynote yesterday. And then I'll click "transcribe." Now you see I'm hitting my local execution provider.

PAVAN DAVULURI: Yep, I see that. And here we go.

CASSIE BREVIU: There you go.

PAVAN DAVULURI: Fantastic, excellent. Thank you. Thank you very much. (Applause.)

CASSIE BREVIU: Thank you.

PAVAN DAVULURI: OK, let's take a minute to process that. What we just saw is how you can use ONNX Runtime across platforms, silicon, cloud and client to take your AI to the next level. I know many of you are web developers. We see you and we love you. With ONNX Runtime, or ORT, you are covered for AI development on the web. ORT Web can run your ML models in the browser using web-based tools. It natively integrates constructs like WebAssembly and Web GPU with web in support coming soon. This gives you the flexibility to target a whole range of hardware accelerators and Azure in your web apps, just like you saw in the ST demo and the Whisper demos.

But really, how do you accelerate these models and apps across the range of the hardware ecosystem? Well, GPUs and CPUs are two ways, two great ways, actually, going forward. Let me start with GPUs.

Windows 11 has a large and powerful ecosystem of GPUs for AI-accelerated workloads. Running behind me, you should be able to see two LLMs in the class of 10 billion parameter models, the DALL-E model you saw in action yesterday in Kevin's keynote and NVIDIA's Nemo model fully optimized running on Windows Client with Nvidia client GPUs. You can reach more than 200 million discrete GPU customers today for acceleration of AI models and Windows. (Applause.)

That's right. It's kind of amazing, actually. It is kind of amazing.

Partnering with NVIDIA, we're making these and other cutting edge open source models available to you soon.

Now, let me switch gears a little bit and talk about neural processing units, or NPUs.

NPUs are powerful accelerators, purpose built to accelerate machine learning workloads. And as we look to the future, NPUs in the Windows ecosystem will continue to scale, and grow and get you plenty of opportunities.

AMD recently announced a Ryzen Mobile 7040 series along with the Ryzen AI execution provider for ONNX Runtime. Intel's KeemBay accelerators are going to be in market shortly, and the Meteor Lake platform comes to market this holiday, featuring Intel's first integrated neural processor with OpenVINO support for ONNX Runtime. This means by the end of the year, you'll be seeing a lot more NPU devices in market.

There are already some incredible NPU devices out there showing us the power of AI on modern silicon. The Qualcomm Snapdragon compute platform and devices like the Windows Dev Kit 2023 that we, in fact, announced at Build last year, and the Surface Pro 9 5G are empowering developers to create some amazing AI experiences.

Luminar Neo is a great example, in my mind. Luminar Neo is about making creativity accessible to everyone, and empowering photographers of all levels with AI-assisted editing features. A single feature in Luminar Neo, like automatic background removal, relies on more than 20 models that are running behind the scenes, so the speed at which these models run is critical to the user's experience.

With the latest NPU-powered devices, this kind of a workload in Windows is drastically accelerated. Like their super sharp AI feature, it uses AI to fill in and sharpen blurred details. On a CPU, this can take about two minutes, but that same operation moved to an NPU takes only about eight seconds. That is extra time for all of us to get creative. (Applause.)

Thank you. That's awesome. I agree, it was pretty phenomenal.

(Applause.)

Like Shilpa mentioned, there will be some incredible apps and tools for you in the Microsoft Store AI Hub, including Luminar Neo and other amazing AI apps. With Windows AI enabled silicon, ONNX runtime and the Olive toolchain, we are empowering every developer to be an AI developer, and we can't wait to see what you will create on Windows.

Last year, Stevie Batiche shared a vision for the hybrid loop and how your models can run on cloud and client. And what you saw with Cassie in the Whisper demo is that this vision is now a reality.

Looking forward, the hybrid loop will continue to get more seamless, providing you with richer capabilities. We believe we're on the infancy of what the transformation can look like. And Panos is going to come back to you to tell us what this opportunity looks like in the future.

Thank you very much.

(Applause.)

Sorry, that was Stevie.

STEVIE BATHICHE: I'm not Panos. He's a little shorter. I've got to admit, I might be feeling a little vulnerable. See, last year, I had a bunch of really cool demos to lean on. This year, Panos handed me a blank piece of paper and asked me what I should write and say to this audience. And so I did. And I'm here with no demos.

(Applause.)

So I have thoughts for you, and I think it's worthwhile. All right, everyone's already said this, but really, AI is bringing unprecedented change. And at times I really do feel like I'm an intern all over again. And it can be daunting on where to start and even how to think about its impact on your apps. But here are three thoughts for you.

First, between Windows, M365 and Azure ML, we're giving you the latest and most powerful tools to help. Pavan and Cassie just showed this to you. They showed the ONNX runtime, the Olive toolchain. Use them, use them to compile, deploy and run your AI efficiently across the most diverse ecosystem of devices on the planet. But it doesn't just stop there.

Yesterday, Scott talked about new tooling that allows you to finetune large models with techniques like LoRA, with just one click. I mean, jeez, taking something at the scale of GPT-3.5 and retraining it, making it your own, that's groundbreaking. Something that was so difficult just last year is now so easy.

Second, contextualize every interaction. Earlier, Rajesh spoke about the importance of the person being at the center. One of the most powerful ways to do so is use a Microsoft 365 Graph and its APIs to help ground the API and personalize every interaction, every prompt. This contextualization is a key differentiator that will enable you to go broad and deep with your customers.

And third, and finally, Panos just mentioned it. AI is the new interaction technology. But really, what does that mean? Fifty years ago, the industry had a milestone event. Take a look at this demo with Doug Engelbart, showing off the mouse and keyboard interaction. That impact that impact of those innovations revolutionized application design. And amazingly, that application structure hasn't changed much since.

It's wild. Until now. For the next 50 years, this direct, very explicit interaction model will be completely transformed by what's happening today, and we already see it. Our interfaces are transforming from being exact to being more implicit and fuzzier, less programmatic, more piloted.

To seize this opportunity to build the next generation of apps and services, I want to take just a brief moment, because I got to catch a plane actually right after this, so just a brief moment to share the patterns we're seeing and how people infuse AI into their experiences.

Three new AI application structures are emerging, shaped by how AI functions relative to your applications. Is the AI beside your app, inside or outside? It's a simple frame to use alongside what Kevin Scott spoke about yesterday.

In the first application structure, the AI is beside your application, helping. Helping your tasks. Being a helper. It's like a copilot. It is a copilot. It's very appropriate that the first types of significant AI experiences are copilots because it enables us to get in the game quickly. It keeps the original app architecture definition and is minimally disruptive to what our customers already know.

Yet, this new application structure delivers immensely capable tools and experiences that did not exist before. We're excited for the new Windows Copilot and all the category specific copilots like M365, Bing, and even the ones you will create. Use them. Like plugins, integrate with them.

In the second application structure the AI is inside as the main scaffolding of the app. It's the main input loop. Here, you use AI to completely redefine the application interaction model and even its purpose. The interaction model will be less dependent on point-and-click commands. Things will become much more automatic.

We see glimpses already happening in applications like Designer, Bing Chat and Luminar Neo that take pro-level skills and turn them into one-click, slider-driven intents and much more intuitive interactions, all without compromising the result.

Here there are fewer toolbars, fewer deep menus, simply because you don't need them. You want to just intuitively direct the app with what you're managing, and this task is accomplished from within the context of the application.

And this brings us to the third and final application structure, where AI goes from executing from within the context of the application frame to AI being outside, executing globally. Here, the AI will orchestrate across multiple apps, plugins and services, functioning more as an agent.

This structure will bring code to the person rather than the person going to the code, allowing the agent to connect, orchestrate and keep context across entire workflows, across devices and even across vastly different time scales.

You see these ideas already emerging in agents and orchestrators like Microsoft Jarvis, Semantic Kernel and the Bing Orchestrator. In fact, if you take a step back, the Windows Shell itself is an orchestrator. In fact, maybe one of the most powerful orchestrators across apps, across content, across the Graph.

Imagine with AI and natural language, you start to see glimpses of the opportunity with the Windows Copilot. And here when you get intelligence that's functioning, not at just the granule details, but at the higher levels, where you get a mixing of both tactics and strategy, you get both vision and execution. It's like a copilot of copilots. A very powerful application structure.

And with the plugin model from Bing, Windows and M365, we already start to see these outside structures emerging and you can start creating them today. Each of these three new structures offers unique advantages and purposes. And with the solutions we're providing you from our

copilots to our plugins, from our foundational models to our AI runtimes, you can start building on one, if not all three, of the new AI application structures.

With all the new amazing technologies and opportunities, I feel like I'm learning all over again. And maybe so are you. But Microsoft is here alongside you, providing the easiest tools to build, the latest AI to delight, and the broadest platform to deliver those cutting-edge experiences and revolutionary interaction models.

Each of our solutions build on each other. This is so important for you to understand each of them. So use the tools we're building for you to help you optimize and customize your AI for your applications. And that enables you to personalize and contextualize every interaction, which helps you reinvent your interaction model, further enabled by embracing one of the three new AI application structures I just talked about: beside, inside and outside.

Look for ways to use AI and agents to achieve your customer's overall goals, not just their individual tasks. Doing so will make your work more intuitive yet functional, more natural yet powerful. All this to enable and reach and empower more people.

With that, I'd like to bring Panos back on stage to help us close it out.

(Applause.)

PANOS PANAY: Stay with me. So on behalf of the entire Windows team, I hope you love what you saw today. What did you think?

(Applause.)

Just remember, the Windows Copilot is at first high-ambition workload for the client and remains at the center of the new AI app platform on Windows. We are grateful for your time. Thank you for all you do. Go change the world.

(Applause.)

END