Microsoft Ignite Opening
Satya Nadella
Wednesday, November 15, 2023

(Applause.)

SATYA NADELLA: Good morning, and welcome to Ignite. It's great to be together in person, right here in Seattle, and all of you joining, from all over the world. Welcome. Little did we know, when we scheduled Ignite, that we would schedule it on the same day that there's a World Cup semi-final going on in cricket, and I've been up all night, but it finished five minutes ago, and so I'm glad it did, and this was a short version of the game, by the way.

And so here we are, but look, it's just been a fantastic last 12 months. It's hard to believe that it's just been a year since ChatGPT first came out, and lots has been done, and there's lots going on, in terms of the pace of innovation, which has just been astounding.

Just last week, I was at the OpenAI Dev Day, GitHub Universe, and of course, now at Ignite, but the interesting thing is, we're entering this exciting new phase of AI. We are not just talking about it, a technology that is new and interesting, but we are getting into the details of product making, deployment, safety, and real productivity games, all of the real-world issues, and that's just the most exciting thing for all of us as builders.

We're at a tipping point. This is clearly the age of copilots, from digital natives like Airbnb or Duolingo to Shopify, as well as the world's largest company, whether it's BT or Bayer or Dentsu, Goodyear or Lumen, they are all the deploying the Microsoft Copilot to companies in every industry who are building their own copilot, from LSEG in finance to Epic in healthcare, to Rockwell Automation in manufacturing and Siemens in manufacturing. It's fantastic to see people deploy their own copilot.

And today we're sharing lots of new data that shows real productivity gains that Copilot is already driving. We're taking a very broad lens to deeply understand the impact of Copilot on both the creativity and the productivity. And the results are pretty striking. With the copilots, you are able to complete tasks much faster, and that's having a real cascading effect on work and workflow everywhere.

People who use Copilot are spending less time searching for information. They're holding more effective meetings, and they're able to collaborate on work artifacts, right, whether that's Word documents, spreadsheets, emails. All of them have richer context about their role, about their organization so that they can collaborate much more and stay in focus.

And of course, we're just getting started. The way to think about this is Copilot will be the new UI that helps us gain access to the world's knowledge and your organization's knowledge, but most importantly, it's your agent that helps you act on that knowledge.
So this week at Ignite, we are introducing 100 new updates, across every layer of the stack, to help us realize that vision.

Our end-to-end Copilot stack for every organization spans the infrastructure, the foundation models, data, tool chains, and of course, the Copilot itself. And today I'll highlight five key themes across everything we're showing you this week.

So let's dive right in.

It starts with the AI infrastructure layer and our approach to Azure as the world's computer. We offer the most comprehensive global infrastructure with more than 60 data center regions, more than any other provider, and being the world's computer means that we need to be even the world's best systems company across heterogeneous infrastructure.

We work closely with our partners across the industry to incorporate the best innovation from power to the data center to the rack, to the network to the core compute, as well as the AI accelerators. And in this new age of AI, we are redefining everything across the fleet and the data center.

So let's start on how we power the data center.

As we build them, we're working to source renewable power. In fact, today, we are one of the largest buyers of renewable energy around the globe. We've sourced over 19 gigawatts of renewable energy since 2013.

I mean, just to put that in perspective, that's the equivalent to the annual production of 10 Hoover Dams.

And we're working with producers to bring new energy from wind, solar, geothermal and nuclear fusion as well. As we pursue this ambition to not just be carbon free, but to even erase our historical carbon emissions, I'm really excited to share that we are on track to meet our target of generating 100% of the energy we use in our data centers from zero-carbon sources by 2025.

(Applause.)

Now, let's talk about the network that connects our data centers. It's one of the most advanced and extensive in the world already, and to meet the demands of AI and the future workloads, we're driving up the speed. Our breakthrough hollow core fiber technology is delivering a 47% improvement in speed because photons are able to travel through these microscopic air capillaries, instead of through solid glass fiber. This is really cutting edge technology.

In fact, we are manufacturing this fiber ourselves in the world's only dedicated factory for hollow core fiber production. Our first deployment, in fact, is already live connecting our data centers in the United Kingdom. We are very excited about this.

And now, let's step right into the data center.

Today, I'm excited to announce the general availability of Azure Boost. It's fantastic to see this new system that offloads server virtualization processes onto purpose built software and

hardware. This enables massive improvements in networking, remote storage and local storage throughput, making Azure the best cloud for high-performance workloads while strengthening security as well.

Now, let's go inside of servers.

We're tapping into the innovation across the industry, including from our partners AMD and Intel, and making that available to you. For example, organizations like Vestas use AMD on high-end compute and memory optimized servers in Azure to run simulations on massive amounts of weather data.

And the largest SAP database deployments are powered by our new Msv3 virtual machines, running the fourth generation of Intel Xeon scalable processors supporting up to 32 terabytes of memory.

In fact, Intel's putting their own SAP instances on these machines, and it's great to see.

As a hyperscaler, we see workloads, we learn from them, and then get this opportunity as a systems company to optimize the entirety of the stack, from the energy draw to the silicon to maximizing performance and efficiency.

It's really thanks to this feedback cycle that I'm thrilled to introduce our very first custom in-house CPU series, Azure Cobalt, starting with Cobalt 100.

(Applause.)

Cobalt is the first CPU designed by us, specifically for the Microsoft Cloud. This 64-bit 128-core ARM-based chip is the fastest of any cloud provider. It's already powering parts of Microsoft Teams, Azure Communications Services, as well as Azure SQL as we speak, and next year, we will make this available to customers.

(Applause.) Yeah.

When it comes to AI accelerators, we're also partnering broadly across the industry to make Azure the best cloud no questions asked for both training and inference.
It starts with our very deep partnership with NVIDIA. We have built the most powerful AI supercomputing infrastructure in the cloud using NVIDIA GPUs. OpenAI has used this infrastructure to deliver the leading LLMs as we speak.

In fact, last week, Azure was the largest submission to ML Perf Benchmarking Consortium 10,000 H100 GPUs, three times more than the previous record, delivering better performance than any other class. And in the latest top 500 list of the world's supercomputers, Azure was the most powerful supercomputer in the public cloud, and third, all up. That made news, but what didn't make news is we didn't submit the entirety of our supercomputer. We submitted only a fraction of our supercomputers. I'm thrilled to be No. 3 with that. And by the way, that's the only one that made the list as a public cloud.

And as we build supercomputers to train these leading large models, InfiniBand gives us a unique advantage. And today, we are even going further. We're adding NVIDIA's latest GPU AI accelerator, H200, to our fleet, to support even larger model instancing with the same latency, which is so important. As these models become much bigger, more powerful, the ability for us to have this new generation of accelerators is a big deal.

We are also introducing the first preview of Azure Confidential GPU VMs, as you can run your AI models on sensitive datasets on our cloud. We codesigned this with NVIDIA, and so if you're doing what is referred to as retrieval augmented generation, or RAG, you'll hear a lot about this throughout the conference. Running on this confidential GPU VM, for example, you can enrich your prompts with any place-specific knowledge from proprietary databases or document archives, while keeping the entirety of the process protected end to end. And so it's very exciting to see us not just me with GPUs, but lead with GPUs with confidential computing.

Now let's talk about AMD.

I'm excited to announce that AMD's flagship MI300X AI accelerator is coming to Azure to give us even more choice for AI optimized VMs. With 192 gigabytes of high bandwidth memory and 5.2 terabytes per second of bandwidth, the MI300X offers industry leading memory speed and capacity.

Again, this means we can serve large models faster, using fewer GPUs. We've already got GPT-4 up and running on MI300X, and today we are offering early access to select customers.

And we're not stopping there. We are committed to taking the entirety of our knowhow from across the systems, and bringing you the best innovation from our partners and us.
Today, we're announcing our first fully custom in-house AI accelerator Azure Maya.

(Applause.)

Starting with Maya 100, designed to run cloud AI workloads, like LLM training and inference, this chip is manufactured on a 5 nanometer process and has 105 billion transistors, making it one of the largest chips that can be made with current technology, but it goes beyond the chip, though. We have designed Maya 100 as an end-to-end rack for AI, as you can see right here. AI power demands require infrastructure that is dramatically different from other clouds. The compute workloads require a lot more cooling as well as the networking density. And we've designed the cooling unit, known as the sidekick, to match the thermal profile of the chip, and we added rack-level closed loop liquid cooling for higher efficiency.

This architecture allows us to take this rack and put it into existing data center infrastructure and facilities rather than building new ones. And by the way, they're also built and manufactured to meet our zero-waste commitment. So we are very, very excited about Maya. With Maya, we're combining the state of the art silicon packaging techniques, ultra-high bandwidth networking design, modern cooling, power management, algorithmic codesign of both the hardware and the software. And they're already testing this with many of our own AI services, including the

GitHub Copilot. And we will roll out Maya accelerators across our fleet supporting our own workloads first and we'll scale it to third-party workloads after that.

This silicon diversity is what allows us to power the world's most powerful foundation models, and all of our AI workloads from Copilot to your own AI applications so that when I say systems, this is the end-to-end innovation.

From glassblowing the next generation to fiber optic cables sourcing renewable energy, designing new approaches to thermal distribution, innovating in silicon, our goal is to ensure that the ultimate efficiency, performance and scale is something that we can bring to you, from us and our partners.

Now, let's go to the next layer of the stack, the foundation model.

Of course, these are only possible because of these advanced systems I talked about. Generative AI models span from trillions of parameters for LLMs that require the most powerful GPUs in Azure to a few billion parameter task-specific small language models, or SLMs. And we offer the best selection of frontier models, which you can use to build your own AI apps, while meeting your specific cost, latency and performance needs. And it starts with our deep, deep partnership with OpenAI. They're just doing stunning breakthrough work to advance the state of AI models, and we are thrilled to be all in on this partnership together.

And our promise to you is simple. As OpenAI innovates, we will deliver all of that innovation as part of Azure AI, and we are bringing the very latest of GPT-4, including GPT-4 Turbo, and GPT-4 Turbo with Vision to Azure OpenAI services.

Yeah, you can clap for that.

(Applause.)

GPT-4 Turbo offers lower pricing, structured JSON formatting, which is sort of my favorite, and extended prompting. In fact, you can now fit 300 pages of text into a single prompt. GPT-4 Turbo will be available in Azure OpenAI service this week in preview, and the token pricing for the new models will be at parity with OpenAI.

Soon, you will also be able to connect GPT-4 Turbo with Vision to Azure AI Vision, allowing you to prompt with video images and text. In fact, our customer WPP is already using this today with one of their largest clients, and take a look at the video behind me, where it's pretty amazing to see video prompts as inputs and with summaries coming out on the other end. It's fantastic to see that.

(Applause.)

Finally, we'll be introducing fine-tuning of GPT-4 and Azure OpenAI service as well, allowing you to bring your own data to create these custom versions of GPT-4. We are also all in on open

source, and we want to bring the best selection of open-source models to Azure and do so responsibly.

Our model catalog has the broadest selection of models already, and we are adding even more to our catalog, with stable diffusion which can generate beautiful immersive images, with code Llama, where you can generate code, with Mistral 7D, where you can translate and summarize text, with NVIDIA's Megatron 3 family of models, where you can build general purpose AI apps, and all these capabilities are deeply, deeply integrated with our safety guardrails.

And today, we are taking one more big step in support of open-source models. We are adding a new model as a service offering in Azure. Yeah, this is a big deal.

(Applause.)

It makes it simple, because this will allow you to get access to these large models that are all available in open source, and just focused API, right, without you as developers having to provision GPUs so that you can focus on development, not backend operations. We are to be partnering with Meta on this. It starts with Llama 2 as a service.

You can fine-tune Llama 2 with your data to help the model understand your domain better and generate more accurate predictions. We want to support models in every language and in every country. And we are partnering with Mistral to bring their premium models as a service, as well as with Group 42 to bring Jais, the world's highest quality Arabic language model, again, just as a service.

Now, when we talk about open source, there is one more very exciting thing that's happening in this space and that is SLMs. Microsoft loves SLMs. In fact, one of the best is Phi, a model that was built by Microsoft Research on highly specialized datasets, which can rival models and at even 50 times bigger. In fact, Phi 1.5 has only 1.3 billion parameters, but nonetheless, it demonstrates state-of-the-art performance against benchmarks testing with things like common sense, language understanding and logical reasoning.

And today, I am thrilled to announce Phi 2.

(Applause.)

It's a scaled-up version of Phi 1.5 that shows even better capabilities across all of these benchmarks, while staying pretty small, or I mean, relatively small at 2.7 billion parameters. In fact, it's 50% better at mathematical reasoning, and Phi 2 is open source and will be coming to our catalog as well as models of service.

Once you have these models, the next up is the tooling consideration. With Azure AI Studio, we offer the full lifecycle tool chain for you to be able to build, customize, train, evaluate and deploy the latest next generation models. It also includes built in safety tooling. Safety is the most important feature of our AI platform. It's not something we bolt on later, but we are shifting left

from the very beginning. And with Azure AI Studio, you can detect and filter harmful user-generated and AI-generated content in your applications as well as your service.

The other thing we're doing with Azure AI Studio is extending it to any endpoint, starting with Windows. You can customize state-of-the-art SLMs, and leverage our templates for common development scenarios, so that you can integrate these models right into your applications. When you combine the power of the cloud and the edge, it unlocks super-compelling scenarios.

Let's say you want to build an NTC helper for a game. You can start with an SLM, like Phi, or as your target model in Windows. We then help you compose solutions to steer your game to do what it needs, like the retrieval-augmented generation templates to apply on your dataset to answer questions about a quest.

This can all happen locally on your Windows machine. The NTC can guide players with their quest or even generate complete new storylines, based on prompts from players. For more advanced use cases, you can adapt and fine tune the SLM on Azure, specifically for your game, using the power of even frontier models like GPT-4. It's incredibly powerful to see all this come together. And of course, we're not stopping there.

Earlier I mentioned our partnership with NVIDIA. Together, we are innovating to make Azure the best cloud for training and inference. Our collaboration extends across the entirety of the stack, including our best-in-class solutions for AI development.
NVIDIA's foundation models, frameworks and tools, as well as its DGX cloud AI supercomputing and services to provide the best end-to-end solution for creating generative AI models and custom generative models.

To share more, I would like to invite up on stage the NVIDIA founder President and CEO Jensen Huang to join me.

(Applause.)

JENSEN HUANG: Thank you.

SATYA NADELLA: Jensen, thank you so much for being here. With this partnership, I've talked a lot about all the things we've been doing on the system side, and of course, we wouldn't have been able to train the OpenAI models or to make all this progress over the last few years without sort of the unbelievable systems work we've done.

But today, we're even going a step beyond, bringing, in fact, all of the software innovation that you're doing. Do you want to share a little bit about what we're doing or you are doing on the software side on Azure?

JENSEN HUANG: I would love to. First of all, I'm so happy to be here to celebrate the amazing work of our two teams. This last 12 months, when I was just listening to you, unbelievable amount of progress for the whole computer industry, frankly, in the last 12 months.

Well, our two teams have been super busy. AI and accelerated computing is a full stack challenge, and it's a datacenter scale challenge. From computing to networking, from chips to APIs, everything has been transformed as a result of generative AI.

Now over the last 12 months, our two teams have been accelerating everything we could. Now one of the initiatives, of course, is accelerated computing, offloading general purpose computing, accelerating all the software we can, because it improves energy efficiency, reduces carbon footprint, reduces costs for our customers, improves their performance, and so on, and so forth.

We built the world's fastest AI supercomputer together. It usually takes a couple of, two, three years to plan one, easily a year to stand one up. Our two teams built two of them, twins, one in your house, one in my house. We did it, and we stood it up in just a few months. It is now the fastest AI supercomputer in the world. (Applause.) And seemingly, without barely even trying, it's the third-fastest supercomputer on the planet. It's really quite, quite amazing.

We worked on all kinds of computer breakthroughs, computing breakthroughs, confidential computing, of course, a very big deal, an invention between our two companies, all the way to deploying large language models from the cloud to the PC. The work that we did together so that Windows can now be a first-class client for large language models opens up a few hundred NVIDIA power PCs and workstations around the world.

SATYA NADELLA: Yeah, this install base on the edge of very powerful AI machines happens to be Windows PCs with GPUs from NVIDIA.

JENSEN HUANG: That's right. And now with Studio AI, unbelievable, right? Everybody could be a RAG developer. Everybody could engage large language models.

Now, we've also, and this is something that I'm so proud of, we talked about a year and a half ago, and this is such a great idea, such a great vision. And you really deserve so much credit for transforming Microsoft's entire culture to be so much more collaborative, partner oriented, that NVIDIA's platforms, and you invited NVIDIA's ecosystem, all of our software stacks to be hosted on Azure.

Today, we're announcing the two largest software stacks of our company, NVIDIA Omniverse – and, in fact, just now, you saw the WPP video. In fact, it's actually computer graphics. That computer graphics is running on Omniverse. And now, you can connect Omniverse to generative AI. And so, Omniverse is for industrial digitalization.

Today, we're announcing that Omniverse Cloud, Omniverse, which is a stack originally on-prem on large computers, now available on Azure Cloud.

The second is a brand new thing that we're announcing, and you just mentioned it. We are offering an AI foundry service. Generative AI has opened up the opportunity for every enterprise in the world to engage artificial intelligence. For the very first time, it is now useful, versatile, quite frankly, easy to use.

And companies all over the world will use it in multiple ways, but here's the three basic ways: One, of course, public cloud services like ChatGPT; second, embedded into applications like Windows. We are very happily also a full site licensed customer of Copilot. And so, we are going to be augmented by Microsoft Copilot. And if you think that NVIDIA is moving fast now, we are going to be turbocharged by Copilot.

And then, third, of course, customers want to build their own AIs. They want to create their own, using their own data, create their own proprietary large language models and create their own RAGS.

And so, today, leveraging what NVIDIA's core assets are, our AI expertise, our AI end-to-end workflow, NVIDIA AI Enterprise, and our AI factories, which is now available on Azure, called DGX cloud. We are going to make these, built on these three pillars, help customers build their own custom large language models. We're going to do for people who want to build their own proprietary large language models what TSMC does for us, right?

SATYA NADELLA: Yeah, it's fantastic.

JENSEN HUANG: And so, we'll be a foundry for (MRs?).

SATYA NADELLA: Yeah, it's just so amazing to see, I mean, us partnering on everything on the system side, and everything up the stack on the software side, whether it's on Omniverse or DGX Cloud and this AI foundry is fantastic. I love that metaphor of TSMC for AI model development.

Talking about this arc of AI, Jensen, of course, you've been at the core of this. Long before it became fashionable to talk about it, you were talking about it. What's your arc here of AI innovation, going forward?

JENSEN HUANG: Well, generative AI is the single most significant platform transition in computing history. You and I both have been in the computer industry a long time. In the last 40 years, nothing has been this big. It's bigger than PC, it's bigger than mobile, it's going to be bigger than internet, and surely, by far.

This is also the largest TAM expansion of the computer in industry in history. There's a whole new type of datacenter that's now available. Unlike the datacenters of the past, this datacenter is dedicated to one job and one job only, running AI models and generating intelligence. It's an AI factory. This AI factory, you're building some of the world's most advanced, you're building the world's computer. That computer is now going to be augmented by AI factories all over.

The second, TAM expansion, is where our industry has focused on building tools in the past, now you have copilots that use the tools. In hardware, there's a brand new segment, AI factories. In software, there's a brand new segment, Copilot.

SATYA NADELLA: Copilot, yeah.

JENSEN HUANG: These are brand new things that the world's never had the opportunity to enjoy, big, huge TAM expansion.

The first wave is the wave that we enjoyed, which is incredible startups at OpenAI and others, who are a part of the generative AI startups, cloud internet services. That's the first wave. We're now beginning to second wave and it's really triggered and kicked off by Copilot, Office or Windows 365 Copilot, basically the Enterprise generation.

The third generation, the third wave, is the wave that I think is going to be the largest wave of all, and the reason for that is because the vast majority of world's industries run on it, which is heavy industries. And this is where NVIDIA's Omniverse and generative AI is going to come together to help heavy industries digitalize and benefit from generative AI.

So we're really, quite frankly, barely in the middle of the first wave, starting the second wave. Yeah, this is going to be –

SATYA NADELLA: I love that three waves, and all happening somewhat in parallel, but the staging of it. And I think it all accrues; it compounds across all three.

JENSEN HUANG: That's right.

SATYA NADELLA: And maybe we can close out, Jensen. You and I have worked together for decades, Microsoft and NVIDIA have worked together for decades. Partnerships are these magical things where your innovation, our innovation comes together, ultimately, to enable people in the audience. Just talk about, when you think about the Microsoft partnership, what's your vision for it? What's your expectations of it? And just any thoughts on that?

JENSEN HUANG: Well, we have a giant partnership, and many of you are our partners with Microsoft here. And I think you all agree with me that there's just a profound transformation in the way that Microsoft works with the ecosystem and the industry. We are suppliers to you, building the most advanced computers together. You're suppliers to us. And so, we're customer partners with each other.

But one of the things that I really, really love is the fact that we partner on advancing fundamental computer science, like confidential computing, and generative AI and all the infrastructure that we built together. I love that we're inventing new technologie together, but I really love that you're hosting our native stack right there in Azure. And as a result, we're ecosystem partners.

NVIDIA has a rich ecosystem of developers all over the world, several million CUDA developers. Some 15,000 startups around the world works on NVIDIA's platform. The fact that they could now take their stack, and without modification, run it perfectly on Azure, my developers become your customers. My developers also had the benefit of integrating with all of the Azure APIs and services, the secure storage, the confidential computing. And so, all of that richness amplifies NVIDIA's ecosystem.

And so, I think this partnership is really quite unique. I think that there's not one like it. We don't have one like it. We're incredibly proud of the partnership and incredibly proud of the work that we do together.

SATYA NADELLA: Thank you so much, Jensen. I really deeply appreciate everything that you and your team have been doing. As you said, the last 12 months have been unlike anything I've seen in my professional career. And we're obviously setting pace, and we can plan to continue to do so. Thank you so much for your partnership. (Applause.)

JENSEN HUANG: Thank you.

SATYA NADELLA: Thank you. Jensen Huang! (Applause.)

All right, so let's go one more layer up the stack to data. It's perhaps one of the most important considerations, because in some sense, there is no AI without data. Microsoft Fabric brings all your data as well as your analytic workloads into this one unified experience. Fabric has been our biggest data launch, perhaps since SQL Server, and the reception to the preview has been just incredible; 25,000 customers are already using it.

And today, I am thrilled to announce the general availability of Microsoft Fabric. (Applause.) Let's roll the video.

(Begin video segment.)

VOICEOVER: Microsoft Fabric is redesigning how we work with data by bringing all your data and analytics tools into a single experience. With Fabric's data lake, OneLake, your teams can connect to data from anywhere and all work from the same copy across engines.

Your data professionals have all the tools they need, all in one SaaS experience to reduce the cost and effort of integration. Features like direct link mode and Power BI, which provides a blazing fast, real-time connection to your data, can save you time and cost while providing up to date insights. This intelligence can then securely flow to the Microsoft 365 applications people use every day to improve decision making and drive impact, all backed by Fabric's tight integration with Microsoft Purview to govern and protect your data, no matter where it's used.

AI powered features like Copilot help everyone be more productive, whether it's creating data flows and pipelines, writing SQL statements, or building reports. And as we enter a future built on AI, you can unify, prepare and model your data to support truly game changing AI projects.

All your data, all your teams, all in one place. This is Microsoft Fabric.

(End video segment.) (Applause.)

SATYA NADELLA: Yeah, it's fantastic to see Fabric, the vision come together. In fact, today, really, it's exciting to add this new capability that we call mirroring. It's a frictionless way to add

existing cloud data warehouses as well as databases to Fabric, from Cosmos DB or Azure SQL DB, as well as Mongo and Snowflake, not only on our cloud, but any cloud to Fabric. And they're all in open source, Apache, Parquet format and the Delta Lake format that's native to Fabric.

And to bring this home, let me just kind of walk you through a simple example. Let's take an electric car charging company that wants to proactively alert its maintenance teams and crews about stations that need maintenance and servicing.

The real-time data is streaming in, so the IoT stuff is flowing right in from the charging stations into Cosmos DB. They can use mirroring to keep the Cosmos DB and Fabric automatically in sync. Inside of Fabric, they're already connecting all the other relevant data. It could be maintenance schedules, whether from Azure Databricks, AWS S3, or ADLS together into this one single lake house.

With all this data unified, you can stand, obviously, model on top of it using just data in Fabric. But you can also use, now, this new integration between OneLake and Azure AI Studio to build a preventive maintenance model that alerts maintenance teams when an EV station is likely to need servicing. And, of course, you can build a simple Power App that delivers these alerts to the maintenance crew. You can even embed a chat function into a Power App to gather more context about the alert.

This type of example is how all new data, operations, store, analytics and AI all can come together. In fact, we're integrating the power of AI across the entirety of the data stack. This retrieval augmented generation, or the RAG pattern, is core to any AI-powered application. It's what allows you to bring together your data with these foundation models.

And the first thing we did is we've added vector indices to both Cosmos DB, as well as to Postgres SQL. And we're not stopping there. We moved the management of AI-powered indices out of the app domain, into the database itself with Azure AI extensions for Postgres SQL. This makes it easy and efficient for developers to use AI to unlock the full potential of all the relational database, all the relational data in a database.

And with Azure AI Search, we've built a first-class vector search plus state of the art reranking technology, right, delivering this very high quality response much beyond what you can just get from a vanilla vector search. In fact, just last week, when OpenAI moved some of their APIs, like their agent API, from a standalone vector database, for ChatGPT to Azure AI Search, they saw unbelievable scale benefits. And it's fantastic to see this now powering ChatGPT.

Now, let's move up the stack and talk about how we're reimagining all of the core applications in this era of AI. Let's start with Teams.

Our vision for Teams has always been to bring together everything you need in one place across collaboration, chat, meetings and calling. More than 320 million people rely on Teams to stay productive and connected. It's a great milestone. (Applause.)

Just last month, we shipped new Teams, which we reimagined for this new era of AI. New Teams is up to two times faster, uses 50% fewer resources, and can save you time and help you collaborate a lot more efficiently. And we've streamlined the user experience. It's easier to get more done, fewer clicks. It's also the foundation for the next generation of AI-powered experiences, transforming how we work.

And with new Teams, it's also available across many places now. It's available on both Windows and Mac, of course on all the phone endpoints. But Teams is more than a communication and collaboration tool. It's also a multiplayer canvas that brings together these business processes directly into the flow of your work.

Today, more than 2,000 apps are part of the Teams store. Apps from Adobe, Atlassian, ServiceNow, Workday have more than 1 million monthly active users. Companies in every industry have built 145,000 custom line of business applications in Teams.

And when we think about Teams, it's important to ground ourselves that presence is, in fact, that ultimate killer application. And that's what motivates us to even bring the power of Mesh to Teams, reimagining the way employees come together and connect using any device, whether it's the PC, HoloLens or Meta Quest.

I'm excited to share that Mesh will be generally available in January. It's sort of been something that we've been working on diligently behind the scenes, and it's great to be bringing it. Using avatars, you can express yourself with confidence, whether you're joining a 2D Teams meeting, or a 3D immersive space.

With immersive spaces, you can connect in new ways and bring discussions all into one place. With spatial audio, for example, you can experience directionality and proximity, just like in the physical world. And with your own custom spaces, you can create a place tailored for your specific needs, like an employee event, training, guided tours, or even internal or external product showcases. Using our no-code editor or the Mesh toolkit, we are looking to see how Mesh in Teams helps your employees connect in new and very meaningful ways. (Applause.)

Now, let's move up to the very top of the stack, which is the Microsoft Copilot.

Our vision is pretty straightforward. We are the Copilot company. We believe in a future where there will be a copilot for everyone and everything you do. Microsoft Copilot is that one experience that runs across all our surfaces, understanding your context on the web, on your device. And when you're at work, bringing the right skills to you when you need them.

Just like, say, today, you boot up an operating system to access applications or a browser to navigate to a website, you can invoke a copilot to do all these activities and more, to shop, to code, to analyze, to learn, to create. We want the copilot to be everywhere you are.

It starts with search, which is built into Copilot and brings the content of the web to you. Search, as we know of it, is changing, and we are all in. Bing Chat is now Copilot. It's a standalone

destination, and it works wherever you are, on Microsoft Edge, on Google Chrome, on Safari, as well as mobile apps, coming soon to you.

Our Enterprise version, which adds commercial data protection, is also now Copilot. You simply log in with your Microsoft Entra ID to access it. It will be available at no additional cost to all eligible Entra ID users. (Applause.)

And just two weeks ago, we announced the general availability of Copilot for Microsoft 365. It can reason across the entirety of the Microsoft Graph. That means all the information in your emails, calendar meetings, chats, documents, and answer and complete tasks. It integrates Copilot into your favorite applications, whether it's Teams, Outlook, Excel and more, and it comes with plug-ins for all the enterprise knowledge and actions available in the Graph.

When it comes to extending Copilot, we support plug-ins today, and we are also very excited about what OpenAI announced last week with GPTs. GPTs are a new way for anyone to create a tailored version of ChatGPT that's more helpful for very specific tasks at work or at home. And going forward, you will be able to use both plug-ins and GPTs in Copilot to tailor your experience. And it goes beyond that. You will, of course, need to tailor your Copilot for your very specific needs, your data, your workflows, as well as your security requirements. No two business processes, no two companies are going to be the same. That's why today we're announcing Copilot Studio.

With Copilot Studio, you can build custom GPTs, create new plug-ins, orchestrate workflows, monitor in fact your Copilot performance, manage your customizations, and much, much more. It comes with a bunch of pre-built plug-ins to incorporate your own business data, as well as from applications such as SAP, Workday, ServiceNow. It can connect to databases, custom backends, legacy systems that may even be on premise. All of this allows you to extend Copilot with capabilities unique to your organization and the systems you use every day.

For example, you can have Copilot help with expense management, HR onboarding, IT services. Just take a look at Copilot Studio. (Video segment.) It's super exciting to see Copilot Studio come together. What Power Platform was for the previous generation of applications that we built and the app platform, I think Copilot Studio will be the modern equivalent for the Copilot era, and it's exciting to see this all come together. In fact, we're already using this pattern to extend Copilot across every role and function.

For developers, GitHub Copilot is turning natural language into programing language, helping them code 55 times faster. For SecOps teams, Copilot is helping them respond to threats at machine speeds. In fact, this week we're adding plug-ins for identity management, endpoint security, and for risk and compliance managers as well. For sellers, Copilot is right there helping you close more deals. Whether you're responding in email or in a Teams meeting, you can enrich that customer interaction by grounding the Copilot with your CRM data, whether it's in Salesforce or Dynamics 365.

For customer service teams, today we are very excited to announce Copilot for Service to help agents resolve cases faster. It provides agents with access to the right knowledge across all the

data within the tools they use every day, whether it's Teams, Outlook, and it can be embedded directly inside the agent desktop applications. Copilot for Service includes out-of-the-box integrations to Salesforce, ServiceNow, Zendesk, as well as Dynamics 365. It's the one Microsoft Copilot with all the data, plug-ins and skills you need.

We're already seeing a new Copilot ecosystem emerge as you all extend Copilot. Dozens of ISVs, including Confluence, Jira, Mural, Ramp, Trello, all have built Copilot plug-ins for their applications, and customers are building their own line of business plug-ins, too, to increase productivity and create deeper insights. Not only can you access these in Copilot, but you can surface them across our applications.

For example, Bayer has built a plug-in so that their researchers can use natural language to ask Copilot about crop science models and their suitability for new projects right within Teams as they accelerate the development and delivery of their products to farmers. This idea that you build Copilots, you use them as plug-ins inside of the Microsoft Copilot and Teams, is going to be one of the powerful paths that will play out in the years to come.

These are just a few of the 100+ announcements we'll make during the conference, but I want to close by talking about the arc of innovation going forward in two critical areas. AI and mixed reality, and AI and quantum. AI is not just about natural languages and input. Of course it starts with language, but it goes beyond that. To see, to hear, to interpret and make sense of our intent and the world around us. I want to show you a glimpse of what's possible when the real world becomes your prompt and interface. That's what happens when you bring mixed reality and AI together.

Pay attention to how not just your voice, but your gestures, even where you look, becomes the new input and how transformative it can be to someone like a frontline worker using Dynamics 365. Let's roll the video.

(Video segment.)

SATYA NADELLA: It's pretty amazing when you bring these two powerful technologies together, and this stuff is real today. In fact, it's being deployed in preview with Siemens Energy, Chevron, Novo Nordisk, so it's great to see the power, and I think this is going to be even more powerful in the years to come. The other area I want to talk about is the convergence of quantum computing and AI. Your key to scientific discovery today is complex simulation of natural phenomena, whether it's chemistry, biology, physics, on high performance computing today.

You can think of AI as an emulation of those simulations by essentially reducing the search space. And that's what we're doing with Azure Quantum elements. In fact, we built a new model architecture called GraphFormers for this very purpose. Just like large models can generate text, you will be able to generate entirely new chemical compounds. Just imagine if you can compress 250 years of progress in chemistry and material science into the next 25 years. That's truly using the power of AI to change the pace of science.

In this example, I'm just using a Python notebook. I mean, think about it, just a Python notebook with quantum elements to discover a new coolant. A process that would have perhaps taken three years if we just used traditional computational techniques, it probably takes nine hours now. I can reason over these results with a copilot, narrow them down, find the most promising candidates. Using quantum elements, any scientist can design novel new molecules with unique properties for developing more sustainable chemicals, drugs, advanced materials or batteries.

This is just the very beginning. In parallel, we are also making progress on quantum computing, because quantum will ultimately be the real breakthrough for speeding up all these simulations. In fact, just last week, we announced the strategic collaboration with Photonics to expand this full stack quantum approach that we have taken to quantum networking. Photonics' Novel spin photon architecture natively supports quantum communication over standard telecommunication wavelengths. Combining that infrastructure and bringing it right into Azure takes us one more step closer to the promise of quantum networking and computing inside of Azure.

At the end of the day, though, all of this innovation will only be useful if it's empowering all of us in our careers, in our communities, in our countries. That's our mission. We want to empower every person and every organization across every role and business function with a copilot. Just imagine if 8 billion people always had access to a personalized tutor, a doctor that provided them medical guidance, a mentor that gave advice for anything they needed. I believe all that's within reach. It's about making that impossible possible. I want to leave you with a video of Anton, a developer from Ukraine, who shares his story of how Copilot has empowered him. Thank you all so very much. Enjoy the rest of Ignite. Let's roll the video.

END