Advancing AI 2024 Event

[MUSIC PLAYING]

(SINGING) Ain't nobody know what me and you, what me and you see. I've been--

[MUSIC PLAYING]

[INTRO MUSIC PLAYING]

[APPLAUSE]

LISA SU: Good morning.

[CROWD SHOUTING]

How is everyone this morning?

[CROWD CHEERING]

All right. Big, big welcome to Advancing AI 2024. It is so great to be here in San Francisco with all of you, so many press, analysts, customers, partners, and lots of developers today. And welcome to everyone who's joining us online from around the world.

It's been an incredibly busy year for AMD with lots of launches across new products in our PC and in our embedded portfolio. But today is a special day. Today is all about data center and AI. We have a lot of exciting news and products, and so let's go ahead and get started.

Now, at AMD, we believe high performance computing is the fundamental building block for the modern world. And we are really committed to pushing the envelope to help use technology to solve the world's most important challenges. Whether you're talking about the cloud or health or industrial or automotive or comms or PCs and gaming, AMD products today are used by billions of people everyday.

And for sure, AI is the most exciting application of high performance computing and drives the need for significantly more compute as we go forward. So let's talk a little bit about AI. Actually, I think we're going to talk a lot about AI today if that's all right with you guys.

Over the next decade, AI will enable so many new experiences that will make computing an even more essential part of our lives. If you think about it, AI can help save lives by accelerating medical discoveries. It can revolutionize research. It can create smarter and more efficient cities.

It can enable much more resilient supply chains. And it can really enhance productivity across virtually every single industry. And our goal at AMD is to make AMD the end to end AI leader. And to do that, we have four big themes.

First, it's about delivering the best high performance energy efficient compute engines for AI training and inference. And that's including CPUs, GPUs, and NPUs. And you're going to hear me say today, there is no one size fits all when it comes to computing.

The second is really to create an open, proven, and developer friendly software platform. And that's why we're so excited to have so many developers joining us here today. And it's really about enabling leading AI frameworks and libraries and models so that people can use the technology and really co-innovate together.

The third piece of AMD's strategy, and you're going to see a lot of our partners and customers, whether on stage today or throughout the show, it's about co-innovation. There's no one company that has every answer. You actually need the entire industry to come together.

And so for us, this partnership is about including the entire ecosystem, including the cloud, OEM, software, new AI companies. And our goal is to drive an open industry standard AI ecosystem, so that everyone can add their innovation on top.

And fourth, we also want to provide all the pieces needed for our customers to deliver their total solutions. And that includes not only at the chip level but really at the rack, cluster, and data center level.

So when you put all that together, we are really committed to driving open innovation at scale. And on the silicon side, and you guys know that we spend a lot of time on the hardware, this means driving the bleeding edge of performance in CPUs, GPUs, and high performance networking but also to new industry standards.

On the software side, we're going to talk a lot about software today, it's about enabling the industry's highest performance open source AI software stack. And with our acquisition of ZT systems, we're going to bring all of those elements together to really offer a complete roadmap of AI solutions.

So when you look at what business leaders are talking about, when I talk to business leaders today, everyone's focused on, number one, how do I use AI as fast as possible? But number two, how do we maximize the impact and ROI for their AI initiatives?

Now, when we think about AI, it really is about choosing the right compute for the right application. And looking across the portfolio, we see a lot of different pieces. Starting first on the CPU side, we have lots of opportunity to really think about-- today's AI is really about CPU capability, and you see that in data analytics and a lot of those types of applications.

So for instance, in an enterprise, predictive analytics are actually used very often. But you also see in generative AI applications, you actually rely on the GPU. So we're starting to see a lot of conversation about agentic AI and these new workloads.

And these are the idea that LLMs can actually be tuned to automate very difficult tasks and actually reason and help us make decisions by using natural language. So when you look at agentic AI, you actually require significantly more general purpose compute as well as AI compute.

So you see CPUs handling things in the data pipelines, and then you see GPUs for training, fine tuning, and inference. So when you look at all of these compute needs, we certainly have the best portfolio in the industry to address end to end AI.

So a little bit about our portfolio. We've built our leadership data center compute portfolio over multiple generations, starting with our EPYC CPUs. Since launching in 2017, EPYC has become the CPU of choice for the modern data center.

The largest cloud providers offer more than 950 EPYC instances and have deployed EPYC widely throughout their infrastructure on their most important services, things like Office 365, Facebook, Salesforce, SAP, Zoom, Netflix, and many more.

And on the enterprise side, numerous large customers have also deployed EPYC on prem to power their most important workloads. Today, you're going to hear from all of our largest server OEMs, and they provide over 350 EPYC platforms. If you put all that together, we're very proud to say we exited the second quarter at a record 34% revenue share.

[CROWD CHEERING]

On the AI side, we launched MI300 less than a year ago to very strong demand as the large infrastructure providers like Microsoft and Meta deployed MI300 to power their most important AI applications. Instinct platforms are now available from every major OEM, and numerous cloud providers have also launched public instances, making it easier than ever to use AMD instinct.

Now, customer response has been very, very positive. And you're going to hear from some of those leaders today. leaders like Cohere, Essential, Luma, Fireworks, Databricks and many others have been able to bring their AI workloads to MI300 very quickly. And you will hear at leadership performance and TCO.

So we have a lot of new news today, so that includes CPUs, GPUs, DPUs, NICs, and also enterprise copilot plus PCs. So let's start first with the core of the data center computing, the CPU. I've spent a lot of time with CIOs recently, and what everyone's thinking about is, how do I modernize the data center?

They want a CPU with leadership, performance, efficiency, and TCO. With Turin, that is exactly what we've delivered. Today, I'm super excited to launch our 5th gen EPYC portfolio. It all starts with our latest Zen 5 core. We designed Zen 5 to be the best in server workloads, and that means delivering an average of 17% higher ITC than Zen 4 and adding support for full AVX 512.

Turin is fantastic. It features up to 150 billion transistors across 17 chiplets, scales up to 192 cores and 384 threads. And one of the things that's very special about 5th gen EPYC is we actually thought about it from the architectural standpoint in terms of how do we build the industry's broadest portfolio of CPUs that both covers all of the new cloud workloads, as well as all of the important enterprise workloads.

And things like building 5th gen EPYC CPU at 5 gigahertz, that was because there was a new workload. We were seeing the need when you think about industry leading performance for AI head nodes, that frequency becomes really important. And that's an example of where we've broadened the portfolio with Turin.

Now, how do we do that? Turin actually uses our industry leading chiplet technology. This is an area where we have innovated ahead of everybody else in the industry, and that allows us to optimize for both enterprise scale up as well as cloud native scale out workloads.

What we do here is we use a consistent ISA and socket, which is really helpful for developers as well as our customers. And we maintain feature parity, including support for next gen memory and IO.

We've also extended EPYC's confidential compute capabilities with the addition of trusted IO that enables Turin to communicate securely with GPUs, NICs, and storage. All right. I'm going to show you some pretty chips here.

[CROWD CHEERING]

Here we have Turin. Now, this version of Turin is actually 128 cores, and it's optimized for scale up workloads. It has 16 4 nanometer chiplets. So if you look outside of the ring, you see the 16 4 nanometer chiplets and a 6 nanometer IO die in the center.

And we've optimized this for the highest performance per core because it's extremely important in enterprise workloads. When you see that software is often licensed on a per core basis, you want the highest possible performance per core. OK.

Now, this is also Turin.

[CROWD CHEERING]

This is the 192 core version of Turin. And this guy is optimized for scale out workloads. And here, we use 12 3 nanometer compute chiplets. And it's the same 6 nanometer IO die in the center. And this version of Turin is really optimized for cloud. so applications that benefit from maximum compute per socket, this is what we do.

Thank you very much. So with these two configurations, our 5th gen EPYC portfolio is the broadest portfolio in the industry. It scales down to eight cores at extremely high performance per core to up to 192 cores with extremely high performance per socket.

And you can go across a wide range of TDPS. And what that does is it just enables customers to choose the best operating point for their specific needs. Now, Turin is great, but I also want to remind everyone it actually builds on our strong track record of execution.

We've built a 5th gen EPYC CPU that runs at 5 gigahertz, delivering industry leading performance for AI head nodes in many applications. But you also see that we've gotten up to highest core count 11X performance.

So let's now take a look at Turin performance. We're going to compare many of the things in the next few slides to the competition's top-of-stack, Emerald Rapids. When you look at the competition's top-of-stack, a dual socket 4th gen EPYC server is already 1.7 times faster on Specint Rate 2017.

And with 5th Gen EPYC, the performance is fantastic. We extend that lead with 2.7 times more performance. Now, we know it's a very competitive space and we fully expect, as our competition launches their next generation CPUs and they ship in volume, that Turin will continue to be the leader.

For the enterprise space, there are many commercial software stacks that are licensed for core, and CIOs want to optimize the cost by running solutions on the fewest possible cores. When running these workloads on prem, again, 4th Gen EPYC is already the performance leader. And with 5th Gen, we deliver 1.6 times more performance per cores than the competition. That's 60% more performance with no additional licensing costs.

[APPLAUSE]

Now, let's move to relational databases. Now, these are also critical workloads for things like transaction processing and analytics. And MySQL is widely deployed in the enterprise in the cloud. This is another area where EPYC already offers leadership. And now with Turin, that performance capability increases to delivering 3.9 times more than the competition.

In video transcoding, you can also see again that 5th Gen EPYC significantly extends our lead, and we are now four times faster than the competition. Supercomputing is also another one of those really important workloads, and it has been a place where EPYC has continued to lead.

We're already the world's fastest CPU for complex modeling and simulation software. And with Turin, we extend that lead because of the Zen 5 ITC increases, as well as our full implementation of AVX 512.

And so we now deliver 3.9 times more performance than the competition. And what that means is that for researchers who are really doing the most difficult simulations in the world, they get to their answers much faster if they're using AMD.

Now, enterprises are also running many, many more of their AI applications on their CPU. And this is another area where Turin delivers leadership, three times faster performance on traditional machine learning and 3.8 times better performance on TCPX AI, which is actually an aggregate benchmark that represents end to end enterprise AI workloads.

So when you put all this together, I want to give you the business reason why people are so excited about Turin. According to IDC, nearly 75% of enterprise customers refresh their server infrastructure every 3 to 5 years.

Now, if you look at a typical data center today that an enterprise may have, it might be running something like 1,000 cascade lake servers. That was the top of the stack in the industry from four years ago.

5th Gen EPYC can do that same amount of work of 1,000 servers in just 131 Turin servers.

[CROWD CHEERING]

Now just think about what that means. It's like a huge benefit for CIOs. You can replace seven legacy servers with one single EPYC server, and that significantly reduces the power that you need in your data center. It lowers TCO by more than 60%.

And when you add on top of that the enterprise software licensing costs, that means an enterprise can break even on their investments in as little as 6 to 12 months. So that's good for CIOs. And it's also really good for CFOs who want to optimize their CapEx spend.

And on top of that, it also creates space for all of the additional compute that you need, whether you're talking about AI capacity or just adding more general purpose compute capacity. So that's the benefit of the new technology.

Now, I really love talking about our technology, but I love more having our partners talk about our technology. So to understand how EPYC delivers the most in the most demanding environments, let's welcome our first guest, a very close AMD partner who runs some of the world's largest and most advanced data centers. Please welcome Amin Vahdat from Google Cloud.

[APPLAUSE]

Hello, Amin.

AMIN VAHDAT: What's your pleasure, Lisa.

LISA SU: It is so great to have you here. I'm so excited. Google was really one of the first to adopt AMD at scale. And we have learned so much from our partnership with you. So can you tell us a little bit about our work together, and all that you're doing in this GenAI era?

AMIN VAHDAT: It's my real pleasure, Lisa. Thank you for inviting me to join you. Our partnership with AMD goes back a long way. We've been using EPYC CPUs to power our general purpose instances in addition to our high performance and confidential computing solutions. Just last year, we launched C3D, Google Cloud's 4th Gen EPYC-based offering for general purpose workloads like web servers, databases, and analytics. In this new era, C3D is a compelling option for many AI workloads. And frankly, the demand for ML compute power is insatiable.

I see the rise of generative AI as truly the biggest transformation since the beginning of the internet. We're on the cusp of a change that will redefine industries, amplify human potential, and create unprecedented opportunities. To me, that's what makes partnerships like ours so incredibly important.

LISA SU: Thank you.

AMIN VAHDAT: We need to constantly evolve our hardware and software architectures to respond to the demand we are seeing from our customers, whether it's raw performance, cost, convenience, or energy efficiency. Lisa

LISA SU: I really look at this when I think about the work that you're doing, Amin, as pretty amazing. And I know, look, beyond AI, we're also doing a lot of work on your internal workloads as well as some of the third party workloads on EPYC. So can you just share a little bit more about what you're seeing and what customers are seeing?

AMIN VAHDAT: Absolutely. We've been using EPYC CPUs for multiple generations to serve our cloud customers and our internal properties at Google. That adoption has been driven in large part by the gen over gen performance and efficiency gains we've delivered together. For example, Snap used EPYC-based virtual machines on Google Cloud to reduce their AI inferencing costs by 40% and improved their performance by 15%.

[APPLAUSE]

So those are some big numbers. 13 months after the introduction of C3D, KeyBanc, one of the largest banks in the US, is seeing cost efficiencies modernizing to C3D . Strive works and neural magic are running inference workloads on CPUs, saving money without sacrificing speed. And that's just to name a few. Thanks to our collaboration, C3D has been one of our most successful VM instance launches to date, with up to 45% faster performance than previous generations.

[APPLAUSE]

LISA SU: I love hearing those numbers, Amin. It's wonderful to see how great C3D is doing. Now, Google's totally leading in AI innovation and I know you're doing so much. Can you talk a little bit more about your vision for what's happening in AI and the role EPYC plays?

AMIN VAHDAT: It's going to take innovation on every single front, releasing new Gemini models, software, CPUs, GPUs, and TPUs. But it's bigger than that. We're rethinking our system and infrastructure designs from the ground up. That's where our AI hypercomputer comes in. It's a supercomputing architecture that's designed to combine performance optimized hardware, open software, leading ML frameworks, and flexible consumption models to maximize the return on AI investments.

EPYC CPUs are an important part of that stack, offering a cost-effective and seamless option for AI workloads. Sustainability is also a key part of Google's strategy during this transformation, and so I really appreciate AMD's focus on delivering excellent performance per watt and increasingly critical metric for all of us. That's wonderful.

LISA: Look, we totally agree, Amin. And it's great to see all the innovation that you're bringing to the market. Now we're super excited about launching turn today. And it's another area where frankly, our teams have been partnering closely and you've given us just fantastic feedback. Can you tell us a little bit about your plans?

AMIN VAHDAT: Thank you, Lisa. Turns a beautiful chip, and we're really looking forward to the continued collaboration. It was a pleasure joining you today and congratulations on all the success. We are looking forward to partnering with you to deliver turn-based VMs early next year.

LISA SU: That's fantastic. Wonderful. Thank you so much, Amin.

[APPLAUSE]

Super exciting. Now let's turn to data center GPUs and our next generation instinct accelerators. Last December, we estimated the data center AI accelerator market would grow from $45 billion in 2023 to more than $400 billion in 2027. And at the time, I remember people asking me, that seems like a really big number. Is that true? [LAUGHS] Since then, AI demand has actually continued to take off and actually exceed expectations. It's clear that the rate of investment is continuing to grow everywhere, driven by more powerful models, larger models, new use cases, and actually just wider adoption of AI use cases.

So now as we look out over the next four years, we now expect the data center AI accelerator TAM will grow at more than 60% annually to $500 billion in 2028. For AMD, this represents just a huge growth opportunity. We launched our MI300 family last December, and we had leadership performance and very strong customer demand. And I'm very happy to say that the ramp has gone just extremely well. In the last 10 months, we've been laser-focused on ensuring our customers get up and running as fast as possible, with maximum performance right out of the box.

And to do that, we've had to significantly drive improvements and continuous improvements across our ROCm software stack. We've integrated new features, we've optimized new libraries, we've enabled new frameworks, and we've significantly expanded the third party ecosystem that supports ROCm. As a result, if you look today at MI300x performance, we have more than doubled our inferencing performance and significantly improved our training performance on the most popular models.

Today, over 1 million models run seamlessly out of the box on instinct, and that's more than three times the number when we launched in December. And we recently also completed the acquisition of Silo AI, which adds a world class team with tremendous experience training and optimizing LLMs and also delivering customer specific AI solutions, again, to help our Cloud and Enterprise customers get to market as fast as possible.

The improvements we've made with ROCm are enabling our customers to see great performance with MI300. Now, let me just show you a little bit about what customers are seeing. We've now worked with many customers across a wide range of workloads, and our experience has shown that we run out of the box actually very well. So most things run quite well, but with just a little bit of tuning MI300x consistently outperforms the competition, which is H100 in inferencing.

So, for example, using Llama 3.1 405B, which is one of the most newest and demanding models out there. MI300 outperforms H100 with the latest optimizations by up to 30% across a wide variety of use cases. And we've seen this across a lot of different customer workloads, using many different models, including Mistral, Stable Diffusion, and many others. So this allows our customers to really leverage the performance advantage to build their own leadership AI solutions. And you're going to hear some of those key use cases a little bit later on.

Now, we're always pushing the limits on performance. So let me show you what's next. Thank you. Today I'm very excited to launch MI325x, our next generation instinct accelerator with leadership generative AI performance.

[APPLAUSE]

MI325, again, leads the industry with 256 gigabytes of ultra fast HBM3E memory and 6 terabytes per second of bandwidth. When you look at MI325, we offer 1.8 times more memory, 1.3 times more memory bandwidth, 1.3 times more AI performance in both FP16 and FP8 compared to the competition. And when you look at that across some of the key models, we're delivering between 20% and 40% better inference performance and latency on things like Llama, Mistral and Mixtral.

And importantly, one of the things that we wanted to do was really keep a common infrastructure. So MI325 leverages the same industry standard OCP compliant platform design that we used on MI300, and what it does is it makes it very easy for our customers and partners to bring solutions to market. Now, when you look at the overall platform with 8 GPUs, we deliver significantly more AI compute and memory as well. The 8 GPU version features two telabytes of HBM3E memory and 48 telabytes per second of aggregate memory bandwidth, enabling our customers to run more models as well as larger models on a single MI325 server.

When you put all of that capability together, what you see is that MI325 platform delivers up to 40% more inferencing performance than the H200 on Llama 3.1. And a lot of people also are doing training. And when you look at training performance, we've made very significant progress on optimizing our software stack for training on a growing number of customers. And what you see with MI325, we have excellent training performance that's very competitive compared to the competition.

So as you can see, we're very excited about MI325. The customer and partner industry interest has been fantastic. We are engaged across all of the leading OEMs and ODMs. We're on track to ship production later this quarter with widespread system availability from Dell, HPE, Lenovo, Supermicro, and many other providers starting in Q1.

[APPLAUSE]

All right, now, let me invite my next guest to the stage. Oracle Cloud is one of our most strategic cloud partners. They've deployed AMD everywhere, including CPUs, GPUs, DPUs across their infrastructure. And to talk more about our work together, please welcome Senior Vice President, Oracle Cloud Infrastructure, Karan Batta.

[MUSIC PLAYING]

Karan, so wonderful to see you again. Thank you for being such a great partner. And actually you were also at our December event, so thank you. You talked a lot about how OCI is adopting EPYC across your platforms and services. Tell us a little bit about what's been happening.

KARAN BATTA: Yeah. Again, thank you. Thank you for inviting us again. It's very exciting to be here. Since last December, AMD and Oracle have been working together for a very long time since the inception of OCI in 2016. AMD EPYC CPUs are now deployed across 162 data centers across the globe. That covers our public cloud regions, our gov regions, our secret regions, even our dedicated regions, and alloy as well.

And we've had tremendous success on our compute platform offering bare metal instances, virtual machines on Genoa based E5 instances. And then also we also use-- at the base layer of our platform, we also use Pensando GPUs so that we can offload that logic so we can give customers the ability to get great performance instances.

LISA SU: Look, we love the work that we do together. And it's not just about the technology, but it's also about what we're doing with customers. I know that you're very active with Turin. Can you talk just a little bit about some of that.

KARAN BATTA: Absolutely, one of our largest customers today, today cloud native customers is Uber. They are using E5 Genoa instances today to actually get a lot of performance efficiency. And they've moved almost all of their trip serving infrastructure on top of AMD running on OCI compute. So that's been incredible.

LISA SU: That sounds pretty good.

[APPLAUSE]

KARAN BATTA: We also have Red Bull Powertrains that's developing the next generation of F1 engines for the upcoming seasons. And then additionally, on top of that, we have our database franchise, which is now powered by AMD CPUs. And customers like PayPal and Bank of Brazil are using Exadata powered by AMD to achieve great things for database portfolio. So that's been incredible.

LISA SU: We are super proud of the work we're doing together on Exadata. I think that's just an example of how the partnership has grown over time. So look, the momentum with customers is wonderful. We love the work that we're doing on compute. There's just a little bit of something called AI right now.

KARAN BATTA: Yap.

LISA SU: So let's switch gears to talk a little bit about AI. You just recently launched your MI300 instances publicly. Can you talk a little bit about that?

KARAN BATTA: Yeah. It's been a great collaboration between the two teams. We recently made generally available MI300x. We've had incredible reception internally, externally, and we're working with customers like Databricks and Fireworks and Luma AI to run incredible inferencing workloads on top of the AMD GPUs. Additionally, on top, we've seen incredible levels of performance for inference for running things like Llama 3.1 405B. And so we're seeing great efficiency and performance on top of AMD GPUs. And we're incredibly excited about the roadmap that you've announced. And we're excited to work together on the future of that roadmap.

LISA SU: Yeah, look, it's really, really cool to see what customers are doing. You talked a little bit about the roadmap. You talked about the importance of partnership. So what's next on the horizon?

KARAN BATTA: I mean, first and foremost, we're very excited about Turin. We're going to be working together to launch E6 instances on OCI compute later next year. So we're very excited about that. So that's our compute family. We'll continue to collaborate on the GPU, scale up the capacity for MI300x for our customers across the globe, across all types of regions. And then, again, we will continue to collaborate on the DPU architecture as well with you guys.

LISA SU: That's fantastic. Karan, thank you so much, and thank you for the partnership.

KARAN BATTA: Congratulations.

[APPLAUSE]

LISA SU: Thank you.

[APPLAUSE]

That's great to hear. Now, let me bring another partner to the stage now to talk about our partnership from maybe a different angle, for more of a user angle. Please join me in welcoming Naveen Rao from Databricks to the stage.

[APPLAUSE]

How are you, Naveen?

NAVEEN RAO: Great.

LISA SU: Thank you so much for spending some time with us today. It's been a pleasure working with you and your team. I'd love for you to share just a little bit about Databricks and what you guys do.

NAVEEN RAO: Yeah, absolutely. We're pioneering what we call the Data Intelligence Platform, which combines the best elements of data lakes and data warehouses. The platform enables organizations to unify their data analytics and AI workloads on a single platform. A crucial aspect of our mission is to democratize data and AI, and this democratization is driving innovation across sectors, enabling companies to make data-driven decisions and create AI-powered solutions.

Our team at Mosaic, which is now part of Databricks, has been pushing the boundaries of what's possible with AI. And we're not just developing models, we're actually creating entire ecosystems that make AI more accessible, efficient, and powerful for businesses across various industries.

LISA SU: Look, I think you guys are doing amazing work. We completely agree that our goal is to make AI as broadly accessible as possible, and that's why partnerships are so important. You've also been very active in using our technology. Can you share a little bit about MI300 and what you've been seeing?

NAVEEN RAO: Yeah, we've been on this journey for a little while with you and our collaboration has been exceptional. We've achieved remarkable results, particularly with large language models. The large memory capacity and incredible compute capabilities of MI300x have been key to achieve over 50% increase in performance on some of our critical work.

LISA SU: That's a pretty good number.

NAVEEN RAO: We'll take it.

[APPLAUSE]

And that includes things like Llama and other proprietary models that we're working on. The MI300 GPUs are proving to be a powerhouse in the AI computation and efficiency. And we're excited about the possibilities that it opens up for our customers going forward.

LISA SU: Yeah, well, thank you, Naveen. Look, it's great to hear about the performance on MI300. Now, you've also been very active in giving us feedback on the ROCm stack and we so appreciate it. ROCm plays such an important role in helping people use MI300 in our MI roadmap. You have a lot of AI expertise as well. Can you talk a little bit about ROCm and what you've seen?

NAVEEN RAO: Absolutely, yeah. We've been working with ROCm since the Mosaic ML days, which was even late 2022, I believe, on instinct, MI250 and have published those results on the ease of transition from other platforms using ROCm and even scaling across many GPUs. So we observe very closely that ROCm's capabilities have expanded significantly in the last year. It now supports a wide range of features and functions for AI workloads, and the performance improvements have been substantial.

Many of our models and workflows that were originally developed for other environments can now run seamlessly on AMD hardware with no modification. Working with AMD, the AMD team has been an absolute pleasure. Together, we're optimizing at multiple levels of the software stack, and for AMD GPUs, which translates to better performance and efficiency for our customers.

LISA SU: I liked what you said, no modification. Did you say that?

NAVEEN RAO: I did.

LISA SU: Can say that again.

NAVEEN RAO: No modification.

LISA SU: Look, Naveen, we are actually thrilled with what our teams have been able to accomplish. I mean, we love working with teams like yours because it's like a very, very fast iteration and innovation cycle. What's next on the horizon?

NAVEEN RAO: Yeah, I'm incredibly excited about the future, actually. We've done a lot with MI300, but that's really just the beginning. We're looking forward to the continued optimization efforts, not just for the MI300x, but also for the new MI325x and the upcoming MI350 series. And we're excited by the compute and memory uplifts we're seeing with these products, especially new things like FP4 FP6 data types with the MI350.

So on the Databricks side, we're working on new models and techniques that will take full advantage of these hardware advancements to further improve training and inference efficiency, and making advanced AI capabilities more accessible to more organizations. The combination of AMD's cutting edge hardware and software innovations is helping to democratize AI and make it more powerful, efficient, and accessible. So we're not just pushing the boundaries of what's possible with AI. We're working to ensure that these advancements are practically and responsibly applied to solve real world problems.

LISA SU: That's fantastic. Look, Naveen, thank you so much again for joining us today. Thank you for the partnership. And we look forward to seeing all the great things you and your team are going to be doing.

NAVEEN RAO: Thank you.

LISA SU: Thank you.

[APPLAUSE]

So that gives you a little bit of a flavor of how users are seeing our instinct roadmap and our ROCm software stack. Now let's turn to the roadmap. As we announced in June, we have accelerated and expanded our roadmap to deliver an annual cadence of instinct GPUs. Today, I'm very excited to give you a preview of our next generation MI350 series. The MI350 series introduces our new CDNA 4 architecture.

It features up to 288 gigabytes of HBME3 memory, and adds support for new FP4 and FP6 data types. And again, what we're thinking about is how can we get this technology to market the fastest. It actually also drops into the same infrastructure as MI300 and MI325, and brings the biggest generational leap in AI performance in our history when it launches in the second half of 2025.

Looking at the performance. CDNA 4 delivers over 7 times more AI compute, and as we said, increases both memory capacity and memory bandwidth. And we've actually designed it for higher efficiency, reducing things like networking overhead so that we can increase overall system performance. In total, CDNA 4 will deliver a significant 35 times generational increase in AI performance compared to CDNA 3.

[APPLAUSE]

We continue our memory capacity and bandwidth leadership. We deliver more AI flops across multiple data types. When you look at the MI350 series, the first product in that series is called MI355x, and it delivers 80% more FP16 and FP8 performance and 9.2 petaflops of compute for FP6 and FP4. And when you look at the overall roadmap, we are absolutely committed to continue pushing the envelope, and we are already deep in development on our MI400 series, which is based on our CDNA next architecture that is planned for 2026.

Now, Microsoft has been one of our deepest and most strategic partners across our business, and it's played an incredibly important role in shaping our roadmap. I recently sat down with Chairman and CEO Satya Nadella to talk about our collaboration. Let's take a look.

[AUDIO PLAYBACK]

- Satya, thank you so much for being here. And thank you for being part of our advancing AI event today.

- Thank you so much, Lisa. It's a real honor and pleasure to be with you. So Satya AI is transforming our industry, and Microsoft has truly been leading the way. Can you just tell us a little bit about where are we in the AI cycle? What are you most excited about? Where are you seeing the adoption?

Now, first of all, it's always exciting, Lisa, for both of us and all of the folks when there's a new platform being born. Because in some sense, I like to say this is probably a golden age, again, for systems because of all the innovation you are doing, the system software innovation, and, of course, the application innovation in AI. And behind it are these things that all of us now call the scaling laws. It's very much like Moore's law.

Now you have these scaling laws that are really creating, I would say, abundance of compute power. And in fact, it's interesting to think about it. It's the combination of silicon innovation, system software innovation, algorithmic innovation, and even good ways to synthesize data are leading to perhaps 100x improvements for every 10x increase in compute power.

So there is clearly something afoot. It's a super exciting thing. And it's no longer, I mean, I've never seen, Lisa, this rate of diffusion ever before, which is having lived through both PC client server, web internet, mobile cloud, I would say, one, it builds on all of those previous things. And so therefore the rate of diffusion of this throughout the world is pretty exciting to see.

- Yeah. I think you're absolutely right, Satya. It's been incredible the amount of innovation that's happened in the industry. And frankly, it's been incredible, the amount that Microsoft has brought to the industry in terms of getting AI innovation out there. I know that we're personally using many of your AI tools at AMD. Now, we are so excited about the partnership that we have together. We've been long standing partners across on all aspects of our business and your business. Can you talk a little bit about the partnership, and especially our work in data center and AI infrastructure has been really accelerating.

- No, absolutely. I mean, to your point about the long standing partnership, there is not a part of Microsoft that we're not partnered with you. When I look back and think about it. We're reinventing a complete new PC category with you yet again. We historically always worked with you when comes to our gaming consoles. We started working, in fact, you and I first started working together when both of us were not CEOs when we started really doing the cloud work. And so we made progress.

And then in the last four years, in fact, it feels like it's been four years since we even started on really adopting your AI innovation for our AI cloud, as I think of it. And I think the interesting thing there is not only the silicon pieces that you brought, but even the software work that you've done. And because at some level, it's that close feedback loop between emerging workloads. In this case, these emerging workloads, which are these training and inference workloads, are unlike anything we've seen in the past.

These are synchronous data parallel workloads that require a very different way to think about the software stack and the silicon stack, and the jointly optimizing it here. We now have in our fleet. MI300 is sort of-- we did all the work to even benchmark the latest GPT models. We have seen some fantastic results. We have customers coming. Now have choice in the fleet to be able to really optimize for different considerations. People have latency, cogs, performance.

So it's fantastic to see the progress the two teams have made. And I know it's been a lot of hard work, and that's what it takes, which is to be able to see the new workload and optimize every layer of the stack.

- Yeah, absolutely, Satya. First of all, I have to say we are so proud of the work that we've done together on MI300, getting it into Azure. It was absolutely hard work, as you said, but a huge thank you to your engineering teams, hardware and software. I know that our teams couldn't be closer in how we really brought that together. So let's talk a little bit about the future. I mean, the thing about AI is I always say we're just at the beginning of what we can imagine with AI, and it requires an incredible data center infrastructure, a vision for that, and really optimization on all levels.

I think one of the things that's most unique about our partnership is, we've talked about bringing the best of each other to really form that vertically integrated stack. So can you talk a little bit about the roadmap. We're excited at this event. We're actually talking about our accelerated roadmap with MI350 coming next year, MI400 series coming in 2026. Much of that has been work that we've done together. So yeah, can you talk a little bit about?

- First of all, we are very, very excited about your roadmap. Because at the end of the day, if I go back to the core, what we have to deliver is performance per dollar per watt because, I think, that's the constraint. Which is if you really want to create abundance where the cost per million tokens keeps coming down so that people can really go use what is essentially a commodity input to create higher value output.

Because ultimately we'll all be tested by one thing and one thing alone. Which is their world GDP growth being inflected up because of all this innovation. And in order to make that happen, we have to just be mindful of the one metric that matters, which is this performance per dollar per watt. And in that context, I think, there are so many parameters, which if you think about what you are all doing. There's what's the accelerator look like for all of this. What's its memory access bandwidth? How should we think about the network?

So that's our hardware and a system software problem. So that's something that collaborating together to create, I think, the next set of breakthroughs, which create-- for every 10x, we actually get 100x benefit. That I think is the goal. And I'm very excited to see how the teams are coming together, whether it's OpenAI, Microsoft, AMD, all working together and saying, how can we accelerate the benefits such that this can diffuse even faster than what we have.

So we are looking forward to your roadmap in 350 and then the next generation after that. And the good news here is the overall change has already started and we build on it. And the fact that now all of our workloads will get continuously optimized around some of your innovation. That's the feedback loop that we've been waiting for.

[END PLAYBACK]

[APPLAUSE]

Thank you, Satya. We are so proud of the deep partnership we have built with Microsoft. And as you heard from Satya, we see even larger opportunities ahead to jointly optimize our hardware and software roadmaps. Now, Meta is another very strategic partner who we are collaborating with across CPUs, GPUs and the broad AI ecosystem. They've developed epic and instinct-- they've used epic and instinct broadly across their compute infrastructure, and share our view that open standards are extremely important. To hear more about that, please welcome Meta VP of infrastructure and engineering, Kevin Salvadori, to the stage.

[APPLAUSE]

Hello. Hello.

KEVIN SALVADORI: Hi, Lisa. How are you?

LISA SU: Thank you so much for joining us today. Thank you for the incredible partnership we've built. We are actually so honored to be part of Meta's infrastructure. Can you tell us a little bit about our partnership and how that's evolved?

KEVIN SALVADORI: Sure. Well, we first started partnering together back in 2019, but things really took off in 2020 when we started to design in the Milan CPU into our server fleet to support our planet level infrastructure, perhaps. So supporting Instagram, Messenger, Facebook, WhatsApp. That's when it really kicked off. Subsequently, our collaboration of advanced compute infrastructure has enabled us to scale our AI deployments really meeting what the seemingly insatiable demand for AI services.

And Genoa and Turin have been essential for us to optimize our workloads, and we're really excited to now be pairing AMD's compute with AMD's MI accelerators that really help us innovate at scale. So we really see our partnership with AMD as essential for us to scale AI going forward.

LISA SU: It's absolutely fantastic, Kevin. Thank you. I think one of the things that I've been super excited about is with each generation of technology, Meta has expanded your deployments. So can you talk a little bit about what drove those decisions and where we are today?

KEVIN SALVADORI: Sure. Sure. I can. You were right. With every EPYC generation, we've continued to expand our deployments. And when you serve over 3 billion people every day, which is what we do, performance, reliability, and TCO matter. And you know we're a demanding customer, but simply--

LISA SU: Just a little demanding.

KEVIN SALVADORI: --just a little demanding. But simply put you and your team at AMD have continued to deliver for us. So last year, we announced Meta's at scale refresh with bergamot driven by a 2 and 1/2 times performance uplift, higher rack density and energy efficiency. And that all drove up better TCO for us. And I'm happy to announce to everybody something you already know that we've deployed over 1.5 million EPYC CPUs in Meta's global server fleet.

[APPLAUSE]

LISA SU: I've like the sound of that. That's something called at scale deployment. Would you say?

KEVIN SALVADORI: That's serious scale.

LISA SU: Look, we are also so excited about our AI work together. One of the things I've been incredibly impressed by is just how fast you've adopted and ramped MI300 for your production workloads. Can you tell us more about how you're using MI300?

KEVIN SALVADORI: I can. So as you know, we like to move fast at Meta, and the deep collaboration between our teams from top to bottom combined with really a rigorous optimization of our workloads, has enabled us to get MI300 qualified and deployed into production very, very quickly. And the collective team works to go through whatever challenge came up along the way has just been amazing to see how the teams worked really well together. And MI300x in production has been really instrumental in helping us scale our AI infrastructure, particularly powering inference with very high efficiency.

And as you know, we're super excited about Llama and its growth, particularly in July when we launched Llama 405B, the first frontier level open source AI model with 405 billion parameters. And all Meta live traffic has been served using MI300x exclusively due to its large memory capacity and TCO advantage.

[APPLAUSE]

LISA SU: Thank you.

KEVIN SALVADORI: Thank you. Yeah, I mean, it's been a great partnership. And based on that success, we're continuing to find new areas where instinct can offer competitive TCO for us. So we're already working on several training workloads. And what we love is culturally we're really aligned around-- from a software perspective around PyTorch, Triton, and our Llama models, which has been really key for our engineers to run the products and services we want in production quickly. And it's just been great to see.

LISA SU: I really have to say, Kevin, when I think about Meta, I mean, we do so much on the day to day trying to ensure that the infrastructure is good. But one of the things I like to say is you guys are really good at providing feedback, and I think we're pretty good at maybe listening to some of that feedback. But look, we're talking about roadmap today. Meta has had substantial input to our instinct roadmap. And I think that's so necessary when you're talking about all of the innovation on hardware and software. Can you share a little bit about that work?

KEVIN SALVADORI: Sure, sure. Well, the problems we're trying to solve as we scale and develop these new AI experiences, they're really difficult problems to solve. And it only makes sense for us to work together on what those problems are and align on what you can build into future products. And what we love is we're doing that across the full stack, from silicon to systems and hardware to software, to applications, from top to bottom. And we've really appreciated the deep engagement of your team. And you guys do listen, and we love that.

And what that means is we're pretty excited. The instinct roadmap is going to address more and more use cases and really continue to enhance performance and efficiency as we go forward and scale. And we're already collaborating together on MI350 and MI400 series platforms. And we think that's ultimately going to be to AMD building better products. And for Meta, it helps us continue to deliver industry leading AI experiences for the world. So we're really excited about that.

LISA SU: Kevin, thank you so much for your partnership. Thank you to your teams for all the hard work that we're doing together, and we look forward to doing a lot more together in the future.

KEVIN SALVADORI: Yeah. Thank you, Lisa.

LISA SU: Thank you. Thank you.

[APPLAUSE]

All right, wonderful, look, I hope you've heard a little bit from our customers and partners as to how we really like to bring co-innovation together because, yes, it's about our roadmap, but it's also about how we work together to really optimize across the stack. So as important as hardware is, and we know that software is absolutely critical to enable performance AI solutions for our customers. So to talk more about the progress, and we've made just fantastic progress over the last year on ROCm and our broader software ecosystem. Please welcome SVP of AI, Vamsi Bopanna, to the stage.

[MUSIC PLAYING]

VAMSI BOPANNA: Thank you, Lisa, and good morning, everyone. As you just heard AMD platforms are powering some of the most important AI workloads on the planet. So today I'm excited to tell you about the tremendous progress we are making with ROCm, our AI software stack that's making all of this possible. Two years ago, when we laid out our pervasive AI strategy, we made open software a core pillar underpinning that strategy. We said we would partner deeply with the community and create an open ecosystem that is able to provide a credible alternative for delivering AI innovation at scale.

And today, we are their. AI innovators, from the largest corporations to exciting startups are delivering their most demanding workloads on our platforms. ROCm is a complete set of libraries, runtime, compilers, and tools needed to develop and deploy AI workloads. We architected ROCm to be modular and open-sourced it to enable rapid contribution by AI communities. It is designed to connect easily to ecosystem components frameworks like PyTorch, model hubs like Hugging Face.

And over the last year, we've expanded functionality in ROCm at all layers of the stack, at the lower layers, from coverage for platforms, operating systems to higher layers of the stack, where we have expanded support for newer frameworks like JAX. We've implemented powerful new features, algorithms, and optimizations to deliver the best performance for generative AI workloads. I am so proud of what our teams have accomplished this year. ROCm really delivers for AI developers.

We've also been partnering very closely with the open source community. Our deep partnership with PyTorch continues with over 200,000 tests that run nightly in an automated fashion. Our CI/CD pipelines ensure that when developers anywhere in the world commit code to PyTorch, it gets automatically checked that it works well with AMD platforms. That's what has enabled us to ship with day zero support for PyTorch. And we have expanded support for key frameworks with significant work on JAX this year, ensuring robust functionality and support for Max text. Work on Megatron LM has also been super crucial for us for our expanding training engagements.

Now, vLLM has rapidly emerged as the open source inference library of choice in our industry. We are delighted with our close collaboration with the UC Berkeley team and the open source community behind vLLM. That's been crucial for delivering the best inference solutions for our customers. Hardware agnostic languages and compilers like Triton are strategically important for our industry. Triton offers a higher level of programming abstraction with increased productivity and still delivers excellent performance.

Last year, we announced that Triton supports AMD GPUs. And we delivered on that promise. And we've continued our close collaboration with the Triton team to ensure there's expanded functional coverage and excellent performance coming out of Triton for AMD GPUs. We've continued to add coverage for emerging frameworks and technologies, and am delighted to share today that SGLang, which is an emerging inference serving framework, now offers AMD GPU support.

In fact, I'm delighted that the creators of all of these key open source technologies Triton, vLLM, SGLang, and many more are all here speaking at our developer event. And all this great work is resulting in great support for AI workloads and models on AMD platforms. Hugging Face is the largest and most important model hub in our industry. We announced our collaboration with them in June last year, with the goal that any model that's on Hugging Face should run on AMD. And today, I'm delighted to say that over 1 million Hugging Face models now run on AMD.

[APPLAUSE]

This has been made possible by our close collaboration over the last year, and effort that ensures that all their model architectures are validated on a nightly basis. And it's not just about the number of models. We've done extensive work to ensure that the most important models are supported on day-0. For example, when Llama 3.1 models came out, those ran on day-0 on AMD. And perhaps even more importantly, several of our partners like Fireworks offered services immediately thereafter on AMD platforms.

And as you just heard from Lisa, we delivered outstanding performance across a diverse set of workloads. We have been relentlessly focused on performance. From the latest public models to the flagship proprietary models, with each ROCm release, we have delivered significant performance gains. Our latest release, ROCm 6.2, delivers 2.4 times the performance for key inference workloads compared to our 6.0 release from last year.

These gains have been made possible by a number of enhancements, improved attention algorithms, graph optimizations, compute libraries, framework optimizations, and many, many more things. Similarly, ROCm 6.2 delivers over 1.8 times improvement in training performance, and, again, these gains have been made possible by improved attention algorithms like FlashAttention-3 that is supported, improved compute, communication libraries, parallelization strategies, and framework optimizations.

It is these huge performance gains that have been key to driving competitiveness and momentum for our instinct GPUs. But look, model optimization is not the only requirement for AI. AI production often requires data processing, RAG, pipeline development, and many, many more things. There's often significant effort that customers need to put in to realize the value of AI. To solve this last mile of customer AI needs, we acquired Silo AI earlier this year.

Silo AI was Europe's largest private research lab and built a stellar reputation, helping customers implement over 200 production AI solutions in the past few years. With 300 AI experts, including 125 AI PhDs with deep, deep deployment experience, we are thrilled to be able to offer our customers now the ability to implement end to end AI solutions. This exceptional team has also been behind the development of some of the most important European Open Source Language Models. And I'm thrilled to share that those LLMs have been exclusively trained on AMD platforms.

Now, I've shared a lot about our software progress, but perhaps the best indicator of our progress is what AI leaders who are using our software and our GPUs are saying. So it gives me great pleasure to invite on stage four remarkable AI leaders whose work is at the cutting edge of AI to stage. It's an honor to have them here to share their perspective on the future of AI. So please join me in welcoming these outstanding innovators.

[APPLAUSE]

Dani Yogatama, CEO of Reka AI.

[MUSIC PLAYING]

Danny. Dmytro Dzhulgakov, CTO of Fireworks AI.

[APPLAUSE]

Ashish Vaswani, CEO of Essential AI. Ashish.

[APPLAUSE]

And Amit Jain, CEO of Luma AI.

[APPLAUSE]

So great to have you all here with us today. Thank you. Dani, let me start with you. You are an AI trailblazer, having worked on groundbreaking projects like DeepSpeech and AlphaStar, and you've actually seen the potential of multimodal AI before many others. Tell us a little bit about what you're up to at Reka, and some of the exciting work that we've been doing together.

DANI YOGATAMA: Yeah, sounds great. Yeah. Thanks for the intro, Vamsi. At Reka, we provide multimodal AI that can be deployed anywhere. Our models understand text, images, video, and audio, addressing the needs of both consumers and enterprises for developing powerful agentic applications in the cloud, on-premises and on-devices. We are really, really excited that our models are optimized to run on AMD platforms, from high performance cloud GPUs to AI PCs.

VAMSI BOPANNA: That's awesome, Dani. Thanks it's an exciting vision. Now, Dmytro, you are one of the original leads that built PyTorch. You are also a co-creator of ONNX. You're very well-known in the AI ecosystem. So tell us a little bit about Fireworks AI, and how your open source contributions are shaping your work there?

DMYTRO DZHULGAKOV: Yeah. Thanks, Vamsi. So yeah, at Fireworks, we offer a platform for productionizing generative AI with key focus on inference, speed, and cost efficiency. So we help companies ranging from startups, such as Corsair, to enterprises like Uber and DoorDash to basically productionize the latest and greatest open source models across text, image, video, and other modalities.

And actually, I'm particularly excited to be here today because open source is at the center of ROCm stack. And as PyTorch maintainer, I actually worked with AMD over the past five or six years to make ROCm the first class citizen and first class backend for PyTorch and optimize multiple layers of the stack for that. And actually, this foundation helped us at Fireworks when we started 14 inference tech to ROCm. And of course, this year we work very closely with AMD to achieve the best in industry performance and leverage all the low level capabilities of MI300x.

VAMSI BOPANNA: That's a great, Dmytro. It's been great partnering with you in the Open Source over the years and now at Fireworks. Let's go to Ashish next. Ashish, you're well-known to all of us. You co-invented transformers that led to the generative AI revolution here. Tell us a little bit about how Essential is planning to change the world of AI models, and some of the work we have been doing together on MI300x.

ASHISH VASWANI: Absolutely. And, Vamsi, at Essential, we believe that in the future there should be no obstacles to knowledge work. If you're an analyst or you're a scientist, if you want to do manual work in a computer or complex research, we want to build products that you can go to that will help you get unstuck. If you're under time pressure, or if curious to help you make progress, get answers, and do your work much better.

And today we're relentlessly focused on building a tool for financial research. We're building a financial analyst that anybody can actually use to do their research in finance. And it's day-0 for knowledge work. And we're actually incredibly excited to build together with AMD with the MI300x and also the advancements that you've made on ROCm and also the entire training stack with Jackson. We're a full stack company. So pre-training, to post training, to model alignment, we're incredibly excited to work together with you on this.

VAMSI BOPANNA: That's awesome, Ashish. It's been great to partner with you on your training vision.

ASHISH VASWANI: Thank you.

VAMSI BOPANNA: Let's good to Amit. Amit, Luma turns like a simple text prompt into something beautiful, stunning videos, and 3D models. Tell us a little bit about Luma and the work that we've been doing together.

AMIT JAIN: Awesome. Thanks for having me here. At Luma, we are tackling one of the hardest problems in generative AI, which is, how do you model the world and create really high quality, consistent, controllable, intelligent videos? And we do that just from very simple text and images. We have this platform called Dream Machine, and we have seen a great deal of traction from various industries like advertising, education, entertainment. But also we believe this has huge potential to change how people communicate and share their ideas in the world.

Video is the medium through which most information travels in the world today, and we think we can actually bring a whole level of democratization to that that just doesn't exist right now. And here we are very excited to push the boundaries of this very critical, but also difficult AI research and product milestones with the powerful GPUs that AMD has.

VAMSI BOPANNA: It's great. It's been wonderful partnering with you. So, Dani, let's just go to you. It's been great to watch your models come to life on MI300x. Can you talk a little bit about your efforts. How was it getting onto our stack, and what are you seeing?

DANI YOGATAMA: Yeah, thanks. We parted our models to the MI300x and even AI workstations in just a few days, including the full application stack. What's even more exciting for us is we were able to meet our performance goals within about two weeks. ROCm's open source platform really allow us to move fast, enabling seamless integration adaptation from MI GPUs to AI workstations. It's been great.

VAMSI BOPANNA: Awesome. It's been great to watch you guys bring all these models to life so quickly. It's been an absolute pleasure. Dmytro, can you talk a little bit about your experience. You serve pretty large models. How has it been on MI300x, and how you're seeing performance TCO?

DMYTRO DZHULGAKOV: Yeah, so actually, I mean, frankly speaking, when we started integrating our inference tech with ROCm, we were like pleasantly surprised how easy it was. And since this success, actually were able to offer Fireworks Inference API running on AMD platforms hosted by Oracle Cloud on the day when Llama 3.1 was released.

VAMSI BOPANNA: Yeah, that was amazing.

DMYTRO DZHULGAKOV: And actually, I mean, a lot of data tends to outstanding capabilities of AMD GPUs, and particularly the leadership in HBM capacity and bandwidth. When you serve large language models, that comes really handy. For example, if you're trying to serve Llama 4 or 5B in original precision with full context window, it allows you to still stay on a single server and can simplify complexity quite a bit.

And of course, paired with Fireworks Inference Engine, AMD GPUs can really demonstrate competitive performance compared to other GPUs and other accelerators. And overall, this basically allows us to do a really efficient deployment of large scale AI models in real world production, with higher speeds and lower costs.

VAMSI BOPANNA: That's awesome. This is exactly what our engineers dreamed of when they built MI300x Ashish, it's been super exciting to watch your team train on MI300x. Can you share a little bit about how the scaling is going and your experience with ROCm?

ASHISH VASWANI: Actually, I want to start by thanking the entire engineering team at AMD. It's been phenomenal. The advancement in such a short time has been absolutely incredible.

[APPLAUSE]

So very excited.

VAMSI BOPANNA: Lots of engineers.

ASHISH VASWANI: Yeah, I wanted to start by saying thanks. And it's been wonderful to partner. And we're seeing per device best in class performance. The linear scaling characteristics are extremely exciting to scale to our large training workloads. And of course, the advancements all through the ROCm, the training stack, and the blessing of being able to continue to use JAX has been painless. That experience has been painless. So and just seeing how much has been done in such a short time, I think we're going to do something incredible all together.

VAMSI BOPANNA: That's awesome. Ashish, when you say linear scaling, it's like music to our ears.

ASHISH VASWANI: It's so to us as well. And the team, some of which are here. Yeah.

VAMSI BOPANNA: So, Amit, you've been deploying models. We would love to get your experience perspective on how it has been getting onto ROCm as well.

AMIT JAIN: Yeah. The kind of models we're training, they are very challenging models and don't look like LLMs at all. But yeah, we've been impressed with how quickly we were able to bring or at least get the model running on the ROCm stack and on MI300x GPUs. And it took us a few days to get the pipeline running end to end, which is quite fantastic. And I think, it took minimal changes. It was we continued to still work on performance and getting them to be even better than what they were before. But yeah, I think, this is quite fantastic.

VAMSI BOPANNA: Yeah, it's been great how quickly, and actually my team showed me some of those videos. They were just stunning.

AMIT JAIN: Nice.

VAMSI BOPANNA: So it's a great, great set of products. Awesome. Hey, to just wrap things up, I'd like to hear from each of you about your vision for the future of AI. So in 30 seconds, would each of you comment on what you see as the biggest breakthrough opportunity for AI in the next five years. Dani.

DANI YOGATAMA: Yeah, thanks. I think contextual understanding across modalities coupled with superhuman reasoning will lead to self-improving systems. I'm very excited to pursue that goal with AMD in line with the roadmap that Lisa presented today.

VAMSI BOPANNA: Awesome. Dmytro.

DMYTRO DZHULGAKOV: Yeah. So I think in the next five years or so, we will see AI becoming more integrated in a lot of practical systems and a lot of actual production use cases. So focus on inference, speed, and efficiency as important as ever. And of course, I'm confident that we as an industry, as a community, we'll also continue to push the limits of what AI can do on the frontier side as well.

VAMSI BOPANNA: Awesome. Ashish.

ASHISH VASWANI: I'll end how I started. In the future, I believe our products will remove all obstacles to knowledge work and we're at day-0. Extremely excited to partner with--

VAMSI BOPANNA: Day-0 of knowledge work.

ASHISH VASWANI: It is. Yes.

VAMSI BOPANNA: Awesome.

AMIT JAIN: Yeah, I believe AI will be way more pervasive than what we actually imagine today. How wide it can go. And I think the thing I'm most excited about is that it will help us unlock and make lot more people materialize ideas that they have today, like they're stuck in their heads. I think, with the work we are doing and with the work going across the industry, a lot more people will be able to effect change in the world. Will be able to do the kinds of things they just don't imagine. And I think that will lead to rapid innovation and advancement in just human condition across the world. I think that will be really, really fun.

VAMSI BOPANNA: Well, thank you all for your insights and for your inspiring vision. So let's take a moment to thank our panelists.

[APPLAUSE]

Thank you for joining us.

[MUSIC PLAYING]

It's been so exciting to see what companies like Luma, Reka, Essential, and Fireworks are achieving with our Instinct GPUs and our software and how thrilled they are with the experiences they have had. But look, we are not stopping there. We are expanding our efforts to connect with the developer community. Today we are hosting an AI developer breakout with top developers from across the world. We are honored to have luminaries from OpenAI, Meta, Microsoft, x.AI, Cohere, Reka, creators of some of the most important AI frameworks and technologies today.

Try it in vLLM, SGLang, TensorFlow. They're all here to share their experiences, how they're enabling their communities to work with AMD. These developer sessions run from 12:30 to 5:00 later today, so please don't miss them. So to sum it up, we've made tremendous progress over the last year to provide AI developers a truly open software alternative, a software stack that's been deployed at scale, serving the world's most important AI workloads, and has great momentum in the ecosystem.

With ROCm and MI300x, innovators are advancing the AI state of the art at scale on AMD GPUs today. Look, delivering AI at scale involves putting an entire data center solution together with CPUs, GPUs, networking, and thoughtful system design. And so to tell you more about the progress we're making there, it is my pleasure to invite Forrest Norrod, EVP and GM of our Data Center Solutions Group to stage. Thank you.

FORREST NORROD: Thank you, Vamsi.

[MUSIC PLAYING]

Thank you, Vamsi, and thank you all for joining us today. Over the past decade, AMD has built the broadest portfolio of leadership data center silicon in the industry. You've seen from Lisa, the progress in leadership roadmap for both EPYC CPUs and GPUs. And now I'd like to introduce our networking technology, and show how all of these parts working together, enable our partners to build great data center solutions.

Let me start by showing how CPUs and GPUs working together can improve the performance on AI workloads. Often overlooked, the right CPU can dramatically improve GPU performance. For example, in deep learning workloads, the CPU is responsible for feeding the GPUs and performing orchestration tasks like kernel launches, moving data, consolidating results. The latency and responsiveness of the CPU is critical to driving overall performance.

The high frequency Turin 9575F CPU that Lisa mentioned speeds up the workload, enabling faster process of orchestration tasks and delivering better performance of a customer's GPU. On Llama 3.1 inference, for example, this translates into about 10% more throughput on the overall workload on the overall GPU cluster. And at a cluster level, that can be significant. 700,000 more tokens per second on a 1K cluster of GPUs, just by using the 5 gigahertz Turin CPU.

On training workloads like Stable Diffusion, picking the right CPU for the GPU host results in an impressive 20% faster time to train, and this principle extends beyond the CPU. Training is typically a clustered workload where networking is a critical piece of the performance. Now, the network for most data centers used to be relatively simple, with apologies to my networking friends in the audience, but with AI, this has changed.

Today's AI systems connect to multiple networks, each with different roles, requirements, and needs. These requirements are constantly evolving, meaning adaptability is needed to stay at the forefront of innovation in this new AI era. So AI systems first have a front end network that connects to the rest of the data center infrastructure, including storage and the WAN connections to the rest of the world. This network is for user access and data ingestion to the CPUs, setting up AI engines, kicking off queries, delivering results.

It needs storage acceleration to feed the GPUs, but most of all, it needs to be secure to protect data models and users' privacy. This evolving set of services needs to be delivered without taxing the CPU for exactly the reasons we just discussed. The AMD Pensando Data Processing Units, DPUs offload infrastructure storage, and security services from the CPU and provide secure wire rate connections to rich set of software defined networking features.

But in AI systems, there's another network, a back end network that directly interconnects the GPUs and allows them to share queries, activations results to operate as an entire unit to perform at scale training and inference. For small pods, dozens to hundreds of GPUs, this could be Ultra Accelerator Link or NVLink. But at the cluster level, networks must now scale efficiently to tens of thousands or soon hundreds of thousands of GPUs.

They've got to be resilient to failure, maintain high utilization, and have features to detect and avoid network congestion. Because congestion can hold back the progress of many GPUs, and worst case drop data can force a restart of a job or an application rollback. Now, protocols are constantly evolving to ensure that we optimize AI cluster uptime and performance. And this is important because data shows that networking is a key part of performance.

Meta showed that 30% of the elapsed training cycle time is spent in the back end network. AMD research has shown communications accounting for 40% to 75% of the time in some training and distributed inference workloads. So AI networking must manage the unique challenges and characteristics of AI workloads to unlock performance. But of course, for the last 30 years, whenever we've been talking about networking, whenever the question has been networking, ethernet has been the answer.

Why? Because it provides much better TCO, a huge scalability advantage over any competitive technology. And it has a very broad ecosystem. Quite simply, from cloud to enterprise, ethernet is preferred. And many customers have already successfully deployed ethernet as a production solution for front-end and back-end AI networks. But while ethernet is a clear winner in terms of scale and TCO, it is important to maximize utilization of the available bandwidth, particularly for the back-end network. And here's where ethernet does need to continue to evolve.

General purpose ethernet networks usually see about 50% utilization, whereas back-end AI networks really want to operate at 95 plus percent utilization. And this is only achievable if ethernet evolves to support intelligent load balancing, to fully utilize the bandwidth, provide advanced congestion management to avoid performance degradation, and is resilient with fast failover and recovery from packet loss or delay.

Now, luckily, ethernet has been evolving and will continue to do so. A distinguished group of founding members, including AMD, established the Ultra Ethernet Consortium to standardize ways to improve AI network utilization and performance. The objective of UEC is to ensure openness, interoperability, and great TCO. And we're pleased to see the community growing rapidly now, including key vendors and customers really across the industry.

The key innovations introduced in UEC ensure that the AI back-end network can package and deliver data efficiently, leverage all available paths for data transfer, maximizing throughput, and minimizing congestion in the network. And it also avoids hotspots in the network while sustaining maximum throughput between heavily loaded GPUs. Importantly, it will also include features to detect and recover fast network failure and packet loss.

And the great news is that UEC ready RDMA is just about here, and has demonstrated already five to six times better performance than legacy RoCE v2 on critical AI model parameters. So networking is evolving in an open way to meet the needs of AI. But that role of that rate of innovation stresses the industry to keep up and stresses us to deliver optimized chips to meet the new standards. AMD believes we have cracked the code on the problem of delivering high performance, evolving networking chips that can evolve at the speed of AI.

The secret. The AMD Pensando team has now delivered the third generation P4 engine, featuring over 200 fully programmable match processing units, a bunch of table engines. What it means is a super high performance, fully programmable datapath engine that can deliver 400 gigabits per second line rate performance, while multiple advanced services run concurrently on top of that engine. These services can be coded and changed at the speed of software while matching the performance of hard wired solutions.

This architecture makes AMD uniquely positioned to deliver to the future of AI networking and a principal motivation for us bringing the Pensando team into AMD. Leveraging this third-generation engine, we continue our leadership in the DPU product line. Today, I'm excited to announce Salina 400, the third-generation DPU for SmartNICs and smart infrastructure, and Pollara 400 are first AI NIC product. Salina offers 400 gigabits per second throughput while running SDN security encryption services simultaneously.

It provides greater than 2x the performance of our previous generation while being fully backward compatible. And that fully programmable pipeline means we can continuously deliver innovations and features to our customers, and they can add their own innovation on top. Salina will power high performance front-end networks for AI systems and meet the increased demands of general compute clouds that are powered by Turin CPUs that need to be fed some data.

But I'm equally excited to announce AMD's Pollara 400. It uses the same third-generation P4 engine to enable what we expect will be the industry's first Ultra Ethernet Consortium ready AI NIC. It will deliver the performance benefits of UEC ready RDMA and ensure that AMD's customers can continue to innovate at a rapid pace and achieve the fastest time to production. The AMD networking teams are delivering extremely well, and I'm pleased to announce that Salina and Pollara will both be available early next year.

[APPLAUSE]

Now, putting together all of these world class components designed to complement each other makes for great solutions. Our OEM and cloud customers design full solutions and craft platforms greater than the sum of their parts, even these great parts. Together with our partners, we have delivered over 350 server platforms and more than 950 cloud instances powered by AMD data center technology. And so now I'm excited to introduce you to several of our fantastic partners to talk about how we're driving innovation together. First up, please join me in welcoming Ihab Tarazi, Dell SVP and CTO, ISG AI, compute and networking.

[APPLAUSE]

Ihab, it's great to see you. Now your role is broad and gives you a unique perspective on the end to end data center. Can you give us a view of what's happening in Enterprise AI, and how our collaboration is evolving.

IHAB TARAZI: Thank you, Forrest. It's good to be here. Thank you for having me. You're right. Our partnership with AMD is built on co-innovation with an ear for customer feedback. More businesses than ever are trusting AMD with their critical enterprise applications. And together we see transformative impact across many industries. So I want to give you an example here OSP, OSF HealthCare. They operate 145 locations across Illinois and Michigan, and they leveraged Dell servers and AMD processors. They were able to reduce system downtime and speed up database access by about 75%

So you can imagine that's a game changing improvement because it speed up their ability for the clinics to spend time with patients, improving overall quality of care, and even delivering seven figure savings over five years. These are the kind of meaningful outcomes we're seeing when businesses trust and adopt Dell and AMD together.

FORREST NORROD: That's an incredible outcome, Ihab. As engineers, we're always gratified by the technology, but it's more important to see the impact it can provide. In healthcare super, super impressive. So I think if we look at the impact the fourth-generation has because that was fourth-generation. It's been great. But maybe more importantly, particularly after what you heard before on Turin. Tell us about what's coming up next for the PowerEdge lineup.

IHAB TARAZI: Yeah. One of the most exciting launches this year for us is the next generation PowerEdge portfolio that is powered by AMD EPYC processors. We have four new R series servers and we see 66% performance improvement. And also this is enabling seven to one server consolidation for our customers. These servers have hit world record benchmarks for virtualization, database performance, and just as important, AI workloads.

The launch also features for us PowerEdge 7745. That is the first server that is exclusively designed for PCIe AI accelerators. So it's designed to take 16 accelerators, and it has also space for eight slots for network and AI fabric connectivity, as well as for it's just talked about networking is so critical for the AI space.

FORREST NORROD: Yeah.

IHAB TARAZI: And with this in mind, these servers are also optimized for energy reduction, which is another very critical component for AI. So we've designed these with variable fan speeds that dynamically change based on the changing workload. And with that we've seen 65% reduction of energy.

FORREST NORROD: That's fantastic. That's incredible. Energy efficiency and scalability are both essential as businesses grow and explore their AI capabilities. But tell us more. What is Dell seeing in the broader enterprise AI landscape, and how is Dell simplifying AI adoption?

IHAB TARAZI: We're definitely seeing a continued increase in Gen AI from our customers. And I personally talked to hundreds of customers. What they say is two things-- how do I get started, and how can you minimize my risk? And to add is that we're simplifying the implementation of AI and securing the data. So we're bringing AI directly to the data, is hard to move that data. And then to do so we're doing a couple of things. We're adding to our Dell AI Factory AMD solutions.

And the second thing we're going to do is also which I'll talk about, is we're adding AMD to the Dell Enterprise Hub on Hugging Face. So let me take that just one step down. We have the Dell AI Factory with AMD, includes compute network storage, completely validated, tested with all the software, the entire key ecosystem of partners and logo and enablers that you heard about today. And then we've taken that and done extensive testing, fully optimized the configurations, the software, the performance. So our customers don't have to worry about it. And then on top of that, we added services for deployment and strategy so they can get started very quickly.

FORREST NORROD: That's great news for the customers. Making it easy for them is so important. Now you mentioned Dell Enterprise Hub, and I think you've got an association with Hugging Face, as well. Tell me more about that.

IHAB TARAZI: This year at Dell Technologies World, we announced Dell Enterprise Hub with Hugging Face. And we're excited today to add AMD MI300 accelerators to our hub as of today.

[APPLAUSE]

And what we have on that Dell Enterprise Hub is optimized containers with all the models fine tuned, ready to go with few clicks. So things like Llama, Mistral, some of the other models now ready to go with the AMD platforms.

FORREST NORROD: So clicks and get ready to go.

IHAB TARAZI: Big day for our customers, for us. Thank you.

FORREST NORROD: Thank you, Ihab. It's exciting to see how our solutions are making AI adoption for enterprise customers much more accessible. Thank you so much for the partnership. Thank you for joining us today.

IHAB TARAZI: Thank you.

[APPLAUSE]

FORREST NORROD: Now, AMD has a legacy of developing the world's most advanced and energy efficient supercomputers and the silicon for them. And I'm thrilled to bring to the stage one of our long time collaborators, Hewlett Packard Enterprises, Krista Satterwaite.

[APPLAUSE]

Krista, welcome. Now, I may have already foreshadowed this, but can you tell us how the AMD, HPE partnership is driving innovation to address some of the biggest challenges out there.

KRISTA SATTERWAITE: Sure. I sure can. And I'm going to leverage something you said in a meeting that we had a few months ago. You said that we designed and deployed and serviced the world's most complicated systems ever built. And it's true. We broke the exascale barrier with Frontier. It's still the number one supercomputer in the world. And it's not just about performance, oh, thank you very much.

FORREST NORROD: Over two years.

KRISTA SATTERWAITE: Yeah, I know it's amazing. And it's not just about performance. It's also about energy efficiency. And I'm really excited that HPE has the majority of the top 10 most energy-efficient supercomputers with 100% direct liquid cooling, and HPE and AMD partner closely delivering those. And our customers are already seeing benefits. For example, Carnegie Clean Energy. What they're doing is they're extracting energy out of ocean waves and turning that into electricity.

And then we have the European Commission. And what they're doing is Destination Earth, and they're making a digital twin out of the planet so they can simulate extreme weather conditions. We definitely need that, don't we? And GE Aerospace, they're leveraging that frontier exascale system. And they're trying to design a new fan for jet engines that could deliver up to 20% better emissions. So it is inspiring to help organizations that are trying to tackle the world's biggest challenges.

FORREST NORROD: Yeah, that is great to see. And it is clear that our partnership is addressing and helping our customers address these critical issues. But all of those were supercomputing examples. Now, how can organizations that don't require the supercomputers apply these same breakthroughs to help resolve their own challenges?

KRISTA SATTERWAITE: Yeah, well, luckily we have taken our IP and our decades of experience, and we've used those to create AI training systems, and these would be used for training Gen AI models. And today, I'm excited we are launching our latest model. It's our HPE ProLiant Compute XD685. And it's an eight-way accelerator platform, and it accepts eight of the AMD MI325x accelerators. And it's not just more performance, it's also more optimized density. It features trusted security by design, and it has our iLO chip in it, which has security built into the silicon.

It also has a direct liquid cooling option. It comes with our world class support, dedicated on site, and remote for the life cycle of the server. So we're really excited about our new announcement today.

FORREST NORROD: That's great. That's great. And again, you're making it easy for customers to adopt these new technologies. But as we look to the future, what new innovations can enterprises expect from our partnership, particularly in edge computing and hybrid cloud environments?

KRISTA SATTERWAITE: Yeah. So a few weeks ago, we launched a brand new platform. It's an edge platform. It's called the ProLiant DL145 Gen 11. And I have to tell you, we launched this platform because there was a sales leader at HPE who was calling our retail accounts. And she said you know, the edge is changing, and I don't know if we have the server we need to fill those needs. And I'm thinking, I'm sure we do. We didn't, but we do now.

We work with her. We work with her customers to design this platform. We chose AMD because the power efficiency and the price performance, it is half the depth of our most popular rack servers. It is, you can put it on a wall. You could put it on a table. You could put it in a cabinet. We've optimized the acoustics, we've optimized it for environmental conditions, and it features our HPE GreenLake Compute Ops Management software. Where it's cloud based, you can see your servers wherever they are in the world, securely and then automate tasks. So we're really excited about that.

And then today, we are announcing that we're refreshing all of our ProLiant Gen 11s with the new fifth-generation AMD EPYC processors. And our results in terms of performance, we are blown away. Really, our Gen 11 servers already had high performance, but what we're seeing is up to 35% more performance and up to 25% more efficiency. And we're releasing dozens of winning benchmarks today on these platforms. And thank you very much.

[APPLAUSE]

And all of this is audible today. And HPE and AMD are committed to continuing to drive breakthrough innovations from edge to exascale. And stay tuned because we have more announcements coming in the next couple of months.

FORREST NORROD: That's what it's all about addressing the world's most challenges from the most complex to the local McDonald's.

KRISTA SATTERWAITE: That's right.

FORREST NORROD: Well, I'll tell you. Thank you so much for joining us today, Krista. The partnership has just been incredible. And the systems that you guys are building, incorporating our technology and others is just incredible to see. So thank you for all those.

KRISTA SATTERWAITE: Thanks for having me.

FORREST NORROD: Thank you.

KRISTA SATTERWAITE: Bye. Bye.

[APPLAUSE]

I feel like in these sessions there should be-- have people coming around and handing out order forms for the systems here. But another key partner of ours that also builds some incredible systems for many years is Supermicro. And Supermicro and AMD have had an amazing partnership that spans not just the data center, but also workstations and AI more broadly. And so I'd like to invite Vik Malyala to this SVP of technology and AI at Supermicro to the stage. Welcome to Vik.

[APPLAUSE]

VIK MALYALA: Thanks Forrest. I'm excited to be here. And it's not the caffeine that's talking.

FORREST NORROD: Well, I mean, we're always working on pretty exciting stuff. So can you tell us a little bit more. It's been a long journey. Tell us a little bit more about our partnership from your perspective.

VIK MALYALA: Absolutely. So, Forrest, from the very first EPYC and Instinct products, until what we launched today, we have been working and closely collaborating with AMD to enable customers with kickass solutions. Can I say that?

FORREST NORROD: You can say kickass. Absolutely.

VIK MALYALA: OK. Sorry about that.

[APPLAUSE]

More politically right, I guess, it's been an incredible journey working with AMD in this tremendous growth. Together we have and we can further grow the entire ecosystem in terms of bringing the very best from Supermicro and AMD. Our building block approach. It enables us to go from very small to a very large cluster, and we have been using that to enable customers with the most efficient, performant, and cost optimized solutions that brings the best to the customers.

By working closely with AMD, one of the things that we do is to further stretch the performance limits. So we have a lot of performance records with EPYC today, 45 to be precise. 45.

FORREST NORROD: 45. Wow, that's great.

VIK MALYALA: And this is the partnership that actually can make it happen.

FORREST NORROD: That's fantastic. Vik, I know we're all really proud of the solutions that we've built together. What has the experience been like for your customers deploying and managing Supermicro AMD solutions?

VIK MALYALA: That's the most important thing. First of all, customers want to focus on their business not in validating a solution or a system or whatever every few months or few years. So the way AMD is launching the products, which are making it future proof and it also makes the existing platforms to support, which is phenomenal. Current Supermicro platforms are the most extensive product portfolio in the world today.

At the launch, we are having the most products supporting it, whether it is the EPYC, whether it's the Instinct MI300x, MI325x, and including all the storage and networking optimized for it. For example, you mentioned about Pensando 400 gig, right? So you think about Pensando's 400 gig Pollara. We actually are validating with this. This actually will help the customers to look at AI training overall end to end, and whether it is going to be on performance wise or making it seamless integration into that and making it efficient, we are able to deliver together.

Today, Supermicro has at least 5,000 racks per month capacity to bring end-to-end solutions right here in Silicon Valley, whether it's air-cooled or liquid-cooled. So just think if customers want to deploy at any scale, we are ready and we are able to bring that into their data center. So they can just focus on their work while we optimize the time it takes to deploy, as well as the time it takes for them to recover the investment. That's important. So this is all in the hardware. It's exciting part for me.

But ultimately, it's all about applications. So when you look at the software that is going to be run on this and how people are going to recover. A good example would be like Nutanix, Red Hat, and VMware. By having this open ecosystem and be able to run these different applications on these platforms, customers actually can focus on their applications without having any vendor lock in, and more importantly, a seamless migration. Businesses need to quickly deploy their VMs, whatever the workload may be, and with confidence, they can actually focus on Supermicro AMD solutions. And that's what excites the most about this whole thing.

FORREST NORROD: So that is great to hear. I mean, we agree ease of migration and deployment is critical. I mean, it's really all about supporting the applications and making it seamless. And an open ecosystem there is so, so critical. Now, we've been referring to your extensive portfolio of AMD EPYC and Instinct-based solutions. I know you've got some more news to share with us today. So what's coming?

VIK MALYALA: So first of all, we call our products like H11, H12, H13 like that. So the H14 products today. Right now we are announcing this supports the entire spectrum of EPYC 9005 series processors, as well as the support for Instinct MI325x accelerators to design the best performance for AI, HPC, enterprise, and virtualized workloads for some benchmarks because ultimately it's all about how the performance is going to be. We ran different benchmarks. World records is great, but in some specific workloads and benchmarks, we are seeing 77%, no, 73% something like that.

So it's an amazing performance. I think this is something to be happy about. And we are able to deliver that. And not just that. The platforms when you combine with the Instinct MI325x, whether it is in an 8U system, air-cooled, or for audio system, liquid cooled, we have that basically to enable the customers for unprecedented speed efficiency for AI and deep learning. And we want to lead with time to market and time to enable customers. And we are doing that with both air and liquid-cooled options.

So the last but not least is that all this is great. But we have announced and we are making it ready. The H14 platforms are available for our customers for remote testing using our JumpStart program for POC today at the launch.

FORREST NORROD: Fantastic.

[APPLAUSE]

VIK MALYALA: And last but not least, together with AMD, we are committing to help our customers to transform their business and as well as accelerating the growth. Super exciting time, and thanks for having me here.

FORREST NORROD: Oh, thanks, Vik. We really appreciate the partnership with Supermicro and with you. Thanks so much.

VIK MALYALA: Thank you.

[APPLAUSE]

FORREST NORROD: So for the last session, I want to dig in a little bit more on the growing demands of AI also down in the SMB space, as well as the enterprise and cloud markets. And so who better to join us on stage to address that spectrum than Lenovo? And so I'd like to welcome to the stage the senior vice president, ISG for Lenovo, Vladimir Rozanovich.

[APPLAUSE]

VLADIMIR ROZANOVICH: How are you?

FORREST NORROD: Great, Vlad.

VLADIMIR ROZANOVICH: Good to see you, Forrest.

FORREST NORROD: Good to see you, man. Glad you could join us here today.

VLADIMIR ROZANOVICH: Thrilled.

FORREST NORROD: And so I'd love if you could start by sharing with the audience a little bit about how you see the partnership between AMD and Lenovo, how it's gone and how it's evolving?

VLADIMIR ROZANOVICH: Yeah, well, for us, as you know, it's a long standing partnership. And when we look at what AMD and Lenovo have done together, it's all based on technology. It's all based on bringing that right set of solutions to our customers so that our customers get that maximum value for what we're doing out there, especially in their AI workloads. And we are delivering this exceptional business value by integrating AMD's latest EPYC and Instinct Accelerators into a wide area across all of our product range.

In fact, one of the things that Lenovo we talk about is smarter AI for all. It's a vision that we have that's really helping organizations and the industry differentiate on how do we bring these most innovative, general purpose solutions out to market, but also how do we service our large cloud service providers using this technology. And I think that's something that AMD and Lenovo have done extremely well. And you mentioned it. From Lenovo's perspective, we look at this from enabling all the way down from SMB, all the way to enterprise, and to those cloud service providers, many of which we saw today.

So with the AMD EPYC launch today, we believe that we are better positioned not only from a time to market standpoint, but also better positioned with the most robust feature set in the market. And I'll actually talk a little bit about performance as well in a little bit. But one of the things I think I'm really proud of is the relationship we've had with AMD is really service to this tier 2 cloud service provider in an unbelievable way.

And one of the things I was going to share with you Forrest us today is over the last fiscal year, we have actually grown our AMD footprint in that tier two cloud service provider by 50% last year.

FORREST NORROD: Wow.

VLADIMIR ROZANOVICH: And in fact, this year, we expect that's going to grow about 70% year on year.

FORREST NORROD: That's fantastic.

[APPLAUSE]

VLADIMIR ROZANOVICH: And not only that, but actually over the last three years, we've actually grown our AMD footprint within Lenovo by 3x. And in the cloud and HPC segment, you'll be happy to know that AMD is now the leader from an x86 competition standpoint in the amount of volume we sell to HPC and cloud.

FORREST NORROD: That's fantastic. That's fantastic. Those are some very impressive stats. And thank you so much for your partnership there. But let's dive a little bit deeper into the enterprise data center. Now when customers are at various stages of their AI journey, tell us what Lenovo is doing to integrate AMD and to simplify AI for the enterprise data center?

VLADIMIR ROZANOVICH: Yeah, sure. Two things there, Forrest, that I really want to dig into. So when we think about the ThinkSystem servers that we are integrating AMD technology into, first and foremost, I want to mention some of the things like peak performance and the engineering work that Lenovo does into our platforms. And so when we think about it, we want to make sure we're designing these platforms for optimal compute, faster data insight, and most cost effective business operations.

In fact, some of those benchmarks that you saw earlier from Lisa, not only those, but there's also 200 additional benchmarks where Lenovo systems were the ones to give AMD that number one benchmark record performance.

FORREST NORROD: 200 world records. That's great.

VLADIMIR ROZANOVICH: Yeah, so I think from a performance standpoint, Lenovo is really top of class. But the other place that we excel is really when it comes down to reliability and security. In fact, ITC has actually named Lenovo ThinkSystem servers as the most reliable servers globally for the last five years and the most secure. And that's something we're extremely proud of. But not to end just there. We all know that this AI Buzz is really creating a draw on power to really get this performance level.

And something that Lenovo has been doing now for six generations is our Neptune warm liquid cooled servers. And this is really something that we think we can add tremendous market advantage not only for AMD Instinct in the market, but also the CPU technology and EPYC that you're bringing. So I think these are just a few of the areas that together, Lenovo and AMD are really working to drive AI into the best customer experience possible.

FORREST NORROD: That's fantastic. That's fantastic. And tell us a little bit more about your ThinkSystem portfolio with the new EPYC processors.

VLADIMIR ROZANOVICH: And so, as everybody today, we're also talking about new products that we're launching off the AMD EPYC CPU. And it really is it's not just the CPU side, but it's also the instinct accelerator side that we mentioned today. So for instance, today we have introduced our portfolio of products around Instinct, the MI300 class. In fact, we are shipping today. We are also announcing that we will provide support for MI325x from accelerator standpoint.

And we're also collaborating on future products and future roadmaps for your future products, and making sure that we are looking at what cooling solutions need to be integrated there. And so the SR685a V3 is going to be our prime system. That's really going to take advantage of all of these GPU capable systems. But we're not stopping there because, Forrest, it's also it's not just AI in the data center to do large language models and training, but it's also about what are customers going to do at the edge.

FORREST NORROD: Great.

VLADIMIR ROZANOVICH: And this is really an area where Lenovo exceeds. When you look at the strategy we have around retail and healthcare and manufacturing environments, one of the things we've done is we've introduced the ThinkEdge SE455 V3, and it is really the most versatile edge platform, quiet, cool, power efficient, that's really being deployed in things like telco deployments, security system deployments, quality control deployments in manufacturing environments.

And we've partnered with Microsoft on offering things like Microsoft Premier, which is your Azure stack HCI on the ThinkAgile MX455, that's being used in some of these vertical markets, like retail like healthcare, and manufacturing. And so these are all the areas you're going to see Lenovo products across AI in the industry.

FORREST NORROD: That's fantastic. Very broad portfolio from edge to cloud. But I know that Lenovo is also advancing your AI capabilities in your PC lineup down into the commercial client. What new opportunities do those innovations offer?

VLADIMIR ROZANOVICH: Well, this is exciting. So not only is Lenovo a tremendous partner on the data center side, as you know, Lenovo is also really the only pocket to cloud provider out there. And what Forrest is referring to is actually something we heard from Satya earlier today, which is Lenovo is announcing and introducing the new line of Copilot plus PCs that is really that next generation of enterprise class AI PC to the marketplace, all using the latest AMD Ryzen AI pro processor.

And we're going to first launch it in our ThinkPad T14s Gen 6. And then we're following on that with our ThinkBook product Gen 7 products. And they are going to be the most sleek, sexy products you're going to see out in the industry utilizing AMD technology and Lenovo.

FORREST NORROD: Oh, that's fantastic. I can't wait to take a look at the New lineup. Vlad, thank you again for joining us.

VLADIMIR ROZANOVICH: Oh, it's great. Thank you. Thank you so much.

[APPLAUSE]

FORREST NORROD: So now you've heard it from us and you've heard it from our partners. AMD delivers end-to-end infrastructure solutions for all data center workloads in AI, from the cloud, to the data center, to the PC, we are very proud of the close collaboration with our partners to deliver leadership, performance, and efficiency needed to embrace and thrive in the era of AI. With that, please join me in welcoming back to the stage Dr. Lisa Su.

[MUSIC PLAYING]

LISA SU: All right, we've covered so much today. Thank you, Forrest. You've heard from Vamsi and Forrest and all of our great customers and partners about all of the progress we've made on data center platforms. Now let's spend a few minutes on PCs. We actually believe AI will revolutionize how we interact with and what we expect from a PC. Transforming PCs into much more intelligent, truly personalized devices, that will become even more important and more essential to everything that we do.

Let me just give you a couple of examples. AI is going to boost enterprise productivity, and it will allow models to be tailored for each individual business. Collaboration will become much, much better with things like instant meeting transcriptions, real time translations, and automated summaries. For creators, AI will enable much richer content to be developed in less time. And AI will actually give all of us our very own personal assistant and managing every aspect of our lives, which would be wonderful. And all of this can be done locally to protect your privacy. This is the promise of the AI PC.

Now, a few years ago, we recognized how important this area would be and we invested in on chip AI accelerators called NPUs. From that standpoint, we were the first to integrate an NPU into an x86 CPU when we launched Phoenix in January of 2023. Since then, we've been aggressively driving the overall rise in AI roadmap to add more compute, and we launched our strict CPU actually earlier this year in consumer notebooks that deliver the highest industry tops at 50 plus tops of AI I compute.

Now today, in going with the data centers theme, we're actually going to talk mostly about enterprise PCs. When you look at our platforms, we call them the Ryzen AI pro platforms. They're really built to deliver the best performance, multi-day battery life, security, reliability, manageability, all of the things that we need in large enterprises. With Ryzen AI, we've actually enabled hundreds of different AI functions, and our latest software stack makes it really easy for developers to optimize thousands of their pre-trained models for Ryzen.

This collaboration is driving the next generation of enterprise applications, and you can see just great performance on things like Microsoft, Zoom, Webex, Blackmagic, Adobe, SolidWorks, and many, many more. And our enterprise software developers actually want all of the AI compute we can give them, which is why we are so excited to launch our Strix pro processors. The AMD Ryzen AI pro 300 Series resets the bar for what a business PC can do.

We combine our high performance Zen 5 CPU, our new RDNA 3.5 graphics, our new XDNA 2 NPU with 50 plus tops of AI performance, and all of this is within Copilot+ PCs. So if you just look at some of the things that are important, there are lots of PC applications that really need performance. So multithreaded performance we see Zen 5 gives us 1.4 times more performance compared to the competition in the enterprise PC category.

And we also offer fantastic battery life due to Zen 5. So we're seeing things like 23 hours of battery life on video playback and up to 9 hours when using Microsoft Teams. Now for us to really accelerate AI adoption for PCs, we've been working very closely with a broad set of ecosystem partners. By the end of the year, we'll have more than 150 ISVs developing for rise in AI platforms.

[APPLAUSE]

Now, Microsoft is really the leader in this space, and we've been collaborating very closely with them to bring the hardware and software for Copilot+ PCs. So my last guest of the day is Pavan Davuluri, corporate vice president of Windows and devices from Microsoft.

[APPLAUSE]

PAVAN DAVULURI: Hey, Lisa.

LISA SU: Hello, Pavan. Great to see you. Thank you so much for being here. By the way, congrats on all of the incredible work that you and your team have been doing. We're so excited about it.

PAVAN DAVULURI: Absolutely.

LISA SU: So tell us a little bit about what's been going on.

PAVAN DAVULURI: Sure. First, thank you for having me, Lisa. We are excited to be here today, especially as Microsoft and AMD bring powerful new AI experiences to enterprise customers across the world. As Satya shared earlier today, I think we are truly in the golden age for innovation across silicon software, algorithms, and applications with AI. And we're just delighted to be able to partner with AMD to harness all of this opportunity going forward.

LISA SU: We're too.

PAVAN DAVULURI: I want to just actually start by taking a moment to congratulate you, Lisa, and the entire AMD team on your amazing, incredible new performance Ryzen AI pro processors. Thank you very much. It's amazing work.

[APPLAUSE]

LISA SU: Thank you. Look, we couldn't have done it without our close partnership. We're super excited about Strix pro. I'd like to say that all of those tops are for you. So can you share a bit more about some of the new AI experiences that you're bringing.

PAVAN DAVULURI: For sure. We've been working on this for a little bit. In May, we announced Copilot+ PC, a new class of devices that introduced AI-enabled experiences to Windows customers. Copilot+ PCs are the fastest, most intelligent, most secure Windows PCs we've ever built. And quite frankly, we're just energized with the feedback we've received from customers and from reviewers. And we're excited to bring Copilot+ to Strix pro powered designs.

Just last week, we introduced a couple of new AI experiences to Copilot+ PCs built to empower customers to solve everyday tasks and more complex challenges. So how about we just take a look, Lisa.

LISA SU: Fantastic.

[MUSIC PLAYING]

[APPLAUSE]

PAVAN DAVULURI: Thank you. Those experiences are just more ways for us to bring AI to more customers, to empower them to do things not possible on any other PC.

LISA SU: Yeah, look, I think those experiences look great. I know we've been working on them for quite some time with you. Can you just go a little bit deeper into what's going on there?

PAVAN DAVULURI: I'd love to. One of the things we spent time doing in our PCs is looking for content. Whether it's a document summarizing a project plan, a hotel receipt, a picture someone sent to you. Today, there are billions of searches on Windows, and quite frankly, far too many of them end with no results. Now with improved windows search, powered by the NPU and Ryzen Strix pro, it saves you time and energy by helping you find what you're looking for without quite frankly, having to remember the file, its name, its location, quite frankly, even the setting. And I think that's incredible.

And recall is an entirely new way to instantly find what you've previously seen on your PC, with a scrollable timeline that helps you go back and forth. That feature really takes advantage of all the tops you're giving us, Lisa, on the Strix processors. In fact, I am self-hosting recall on my Strix pro machine at work. And it is a truly magical experience.

[APPLAUSE]

Today, as customers are working across multiple applications trying to get things done, it's important for us to stay in our flow. Click to do connects you to your tools and Copilot actions based on the context of what's on your screen, making it faster and easier to get things done.

LISA SU: It's wonderful, Pavan, to see these Copilot+ experiences really lit up on Strix. I know we talked about it for quite some time. You're also bringing a lot of new AI capability to Microsoft 365. Do you want to talk us through that?

PAVAN DAVULURI: I'd love to. Over the past 18 months, working with Copilot has become a daily habit everywhere, helping people complete tasks faster, hold more purposeful meetings, collaborate more efficiently, streamline business processes. Just a few weeks ago, we announced the next wave of Microsoft 365 Copilot, a new design system for knowledge work bringing together web, work, and pages. Copilot Pages is a dynamic, persistent canvas designed for multiplayer AI collaboration.

Improvements to Microsoft 365 apps with Copilot-powered advanced data analysis in Excel, dynamic storytelling, and PowerPoint, managing your inbox and outlook, and many more. And of course, agents making it faster and easier to automate and execute business processes on your behalf, enabling you to scale like you've never done before. And today, Microsoft 365 experiences are cloud and client, and in the future we will take advantage of hybrid connections. And of course, AMD is an important partner, enabling all of this by use of Instinct GPUs in the Azure infrastructure.

LISA SU: That's fantastic to hear. Now I love these experiences. We're all dying to try them out. Now, performance is also super important for the enterprise. So can you talk a bit about the performance that you're seeing on Strix pro.

PAVAN DAVULURI: Absolutely. Speed matters, Lisa. And Copilot+ PCs, these are fast. Copilot+ PCs in Windows 11 powered by the Strix pro processors for us deliver unparalleled performance tailored for enterprise needs. They're up to 38% faster than the latest MacBook Air on sustained multi-threaded performance with all day battery life.

LISA SU: I like that.

PAVAN DAVULURI: And the thing that we love, an impressive 50 NPU tops on these systems. Very, very powerful.

LISA SU: That's fantastic. Look, Pavan, we talk a lot about how do we bring people closer together, and collaboration is super important. Talk a little bit about how those experiences are enhanced with Copilot+ PCs.

PAVAN DAVULURI: For sure. Our teams do a lot of work together on Teams. Copilot+ PCs are built with AI features to make you look and sound your best and collaborate and communicate effectively on Teams. For example, live captions with translations helps you bridge across language barriers, translating audio from over 40 languages to English from any audio on your PC. Thanks to our co-engineering efforts, Lisa. Battery life when using Teams on Strix just got better.

Together, we implemented further hardware offloading of more camera and video and processing effects onto the latest Radeon graphics for Teams. And so together, they're just pretty amazing when it comes to collaboration on Teams.

LISA SU: Yeah. Look, I love what we've done together to bring those experiences to life. Security is also super important. And we've spent a lot of time talking and collaborating on security across cloud, Xbox, and PCs.

PAVAN DAVULURI: Of course.

LISA SU: What are you planning there?

PAVAN DAVULURI: You know, for us, Copilot+ PCs are the most secure PCs. They're secure core PCs that are built with pluton enabled capabilities on top of the platform. Copilot+ PCs provide data security and privacy by enabling AI tasks to be processed locally, like you mentioned earlier on Strix pro laptops. And, Lisa, we're also partnering on increasing hardware support for additional interfaces on Strix pro. We've improved Windows Hello, making it easier to set up. We've extended its support to passkeys.

We've hardened it with VBS support, for example, to isolate credentials. And we do that by default, and quite frankly, most importantly, Windows Hello enhanced sign in security is also now enabled by default. So as you can see, we are taking our commitments around security and privacy to a meaningful step with customers on Strix.

LISA SU: Yeah. Pavan, look, I am super proud of the work that we're doing together. I think what we're bringing to the enterprise here is something truly special. And I can't thank you and your team enough for all of the partnership.

PAVAN DAVULURI: Thank you, Lisa. Grateful for our partnership over the years. Thank you.

LISA SU: Thank you,

PAVAN DAVULURI: Thank you.

LISA SU: Thank you.

[APPLAUSE]

OK, as you can tell, we're really excited to bring Ryzen AI PRO 300 to the market to really bring those experiences to life. We've been working very closely with our OEM partners to accelerate solutions to the market. HP and Lenovo are on track to launch more than triple the number of Ryzen AI PRO platforms in 2024, and we expect more than 100 Ryzen AI PRO platforms in market in 2025. So lots of activity in the AI PC space.

So it's been a big day for us and it's time for me to wrap things up now. We've launched a tremendous number of new products today that extend our portfolio of leadership across data center, AI, and PC solutions from 5th Gen EPYC to MI325 to our next generation Instinct roadmap with MI350 and MI400 series. You also heard from Vamsi and the panel of brilliant AI leaders about the significant progress we've made in strengthening the ROCm stack, and how we're really committed to building out that open software ecosystem so that everyone can use and deploy Instinct at maximum performance out of the box.

You also heard from Forrest on how we're extending our solutions capability with industry standard networking technologies and bringing new technologies to market. And a very huge special thank you to all of our customers and partners who joined us today Google, Oracle, Databricks, Microsoft, Meta, Dell, HPE, Supermicro, Lenovo, Esential, Fireworks, Luma, and Reka. Thank you guys so much.

[APPLAUSE]

I hope you get a glimpse of how we view innovation in the industry. I really believe that these deep partnerships, where we're co-innovating with our customers and partners, is something that's truly differentiating and really allows us to bring amazing solutions to market by bringing the best minds together in the industry. I'll just finish on a personal note. So this week I actually celebrated my 10th anniversary as AMD CEO.

[APPLAUSE]

Huge shout out to all of my AMD colleagues across the world. You guys are really the best. There is no better way I could spend the week than launching AMD products. I feel incredibly fortunate to be part of this industry where everything that we do means so much and is essential to every aspect of our daily lives. You can count on all of us at AMD to continue to push the envelope on high performance computing and AI, and we are just getting started. Thank you all for joining us today.

[APPLAUSE]

[MUSIC PLAYING]