[UPBEAT MUSIC]

[MUSIC FADES]

[LIVELY MUSIC]

UNIDENTIFIED CO. REPRESENTATIVE 1: Trust has always been at the heart of progress. It's how ideas take flight from test to triumph. But trust also asks us to believe in one another. And that's why trust has to be earned. It's earned by relentlessly working together to solve the most important challenges and drive results. That's why as we advance into the AI future, trust has to lead the way.

So yes, we've built a roadmap we deliver on on time. Yes, we've built for the open ecosystem. And yes, we've built the broadest AI portfolio with CPUs, GPUs, and FPGAs. But of all the things we've built--

[ROUSING MUSIC]

--trust is the most important.

UNIDENTIFIED CO. REPRESENTATIVE 2: Please welcome to the stage Dr. Lisa Su, Chair and CEO.

[UPBEAT MUSIC]

[APPLAUSE]

LISA SU: Good morning. How's everyone doing?

[CHEERS]

It is great to be back here in Silicon Valley with so many friends, press, analysts, partners, and especially all of the developers who are here today.

[APPLAUSE]

And a big welcome to everyone who's joining online from around the world for our Advancing AI 2025. Now, it's been an incredibly busy nine months since our last Advancing AI event. We launched lots of new AI, data center, PC, and gaming products. But today, we have so much exciting news to share with you. I'd like to go ahead and get started.

And you guys know us well. At AMD, we're really focused on pushing the boundaries of high performance and adaptive computing to help solve some of the world's most important challenges. And frankly, computing has never been more important in the world. I'm always incredibly proud to say that billions of people use AMD technology every day. Whether you're talking about services like Microsoft Office 365 or Facebook or Zoom or Netflix or Uber or Salesforce or SAP and many more, you're running on AMD infrastructure.

And in AI, the biggest cloud and AI companies are using Instinct to power their latest models and new production workloads. And there's a ton of new innovation that's going on with the new AI startups. For example, life sciences company 310 AI uses MI300X to train a model that turns simple text prompts into novel proteins to really help accelerate drug discovery. Our versatile AI adaptive SoCs are being used to build more efficient 5G networks and improve automotive driver safety. And Ryzen is bringing AI to PCs, enabling more intuitive, responsive, and more powerful experiences.

Now, since ChatGPT launched a few years ago, the pace of AI innovation has been unlike anything I've seen in my career. And in 2025, it's only gone faster. We've seen the emergence of more powerful reasoning models, the rise of agents, really growing momentum in real-world use cases that are actually starting massive scale deployments. And it's clear that we're entering the next chapter of AI.

Now, training is always going to be the foundation to develop the models. But what has really changed is the demand for inference has grown significantly, driven by more capable models and new use cases that are increasing AI usage. We're also seeing an explosion of models. So of course, you have the new frontier models from folks like OpenAI and Google. But you also have open models from Meta and DeepSeek and many others. And we're also seeing now a surge in new specialized models that are built from everything from health care to finance to coding to scientific research.

And when you look over the next few years, one of the things that we see is we expect hundreds of thousands and eventually millions of purpose-built models, each tuned for specific tasks, industry, or use cases. And as AI does more complex tasks like reasoning, you expect agents to become more capable. It drives significantly more compute, which, frankly, is great for all of us.

Now, let me talk a little bit about agentic AI. Agentic AI actually represents a new class of user. One thing that is always on, constantly accessing data, looking at applications, looking at systems to really make decisions and really work autonomously. They need high-performance GPUs to generate insights in real time, but that's really only part of the story.

What we're seeing now is as agentic AI activity increases, all of those agents are now also spawning a lot of traditional compute going to high-performance CPUs. And just think about it. What we're actually seeing is we're adding the equivalent of billions of new virtual users to the global compute infrastructure. All of these agents are here to help us. And that requires lots of GPUs and lots of CPUs working together in an open ecosystem.

So let's talk a little bit about the market. When we were here last year, we said that we expected the data center AI accelerator TAM to grow more than 60% annually to $500 billion in 2028. And frankly, for many of the analysts and folks, at the time, that seemed like a really big number. People were like, do you really think it can be that big, Lisa? And I said, well, that's what we're seeing.

And what I can tell you, based on everything that we see today, that number is going to be even higher, exceeding $500 billion in 2028. And most importantly, we always believe that inference was actually be the driver of AI going forward. And we can now see that inference inflection point. With all the new use cases and reasoning models, we now expect that inference is going to grow more than 80% a year for the next few years, really becoming the largest driver of AI compute.

And we expect that high performance GPUs are going to be the vast majority of that market, because they provide the flexibility and programmability that you need as models are continuing to evolve. And really, algorithms are moving so fast, you want that programmability in your compute infrastructure.

Now, the other thing that we see is AI is also moving beyond the data center, from intelligent systems at the edge to PC experiences. And we expect to see AI deployed in every single device. Now, to enable all of this, you don't have any one architecture that is the right answer. So I like to say there's really no one-size-fits-all. What you need is the right compute for each use case, and that's exactly what we're focused on.

Our strategy is really focused on three key principles. First, we're delivering a broad portfolio of compute engines so customers can match the right compute to the right model and the right use case. Second, we're investing heavily in an open developer-first ecosystem. And you're going to hear us talk about open a lot today. We're really supporting every major framework, every library, every model to bring the industry together in open standards so that everyone can contribute to AI innovation.

And third, we're delivering full stack solutions. We're building. We're forging partnerships. You're going to hear from some of our partners about our ecosystem today to really put all of these elements together. So now let me just give you a little bit of color. From a portfolio standpoint, we offer the most complete suite of computing elements end to end for this vision. That includes CPUs, GPUs, DPUs, NICs, FPGAs, and adaptive SoCs. No matter where AI runs or how much compute you need, AMD has the right solution.

Next, let's talk about open. There are a lot of developers in this audience and online, so this is really talking to you. Thank you for being here. Thank you for coming today. And we believe an open ecosystem is actually essential to the future of AI. AMD is the only company committed to openness across hardware, software, and solutions.

And when you just take a look back, some of the most important breakthroughs in tech actually started out closed. If you think about things like early networking protocols, Unix operating systems, and even mobile platforms. But the history of our industry shows us that time and time again, innovation truly takes off when things are open.

Linux surpassed Unix as the data center operating system of choice when global collaboration was unlocked. Android's open platform helps scale mobile computing to billions of users. And in each case, openness delivered more competition, faster innovation, and eventually, a better outcome for users. And that's why for us at AMD and frankly, for us as an industry, openness shouldn't be just a buzzword. It is actually critical to how we accelerate the scale, adoption, and impact of AI over the coming years.

Now, we also recognize that these AI systems are getting super complicated, and full stack solutions are really critical. So to deliver full stack AI solutions, we've significantly expanded our investments over the last few years, both organically and through strategic acquisitions and investments. We're very happy to say that we recently closed our acquisition of ZT, giving us new capabilities in rack and data center scale design that are becoming extremely useful for what we're doing next.

And we've also strengthened our software stack, acquiring leaders like Nod.ai, Mipsology, Silo AI. And in the last several weeks, we announced adding the Brium and the Lamini teams to AMD. And we're also investing broadly in the AI ecosystem. Over the last year, we've actually done more than 25 strategic investments that have been a great way for us to build new relationships and also support the AI software and hardware leaders of tomorrow.

So let's talk a little bit about customers. We have tremendous momentum in the data center. Since launching in 2017, EPYC has transformed the data center. Today, EPYC is trusted by the world's largest cloud providers and businesses to run their most important workloads. EPYC powers everything from hyperscale services to enterprise data centers, supporting the most important workloads with leaders in financial services, health care, media, and manufacturing.

And our momentum is just accelerating. We exited the last quarter with a record 40% market share. And we believe with AI and high performance compute, there's a lot more room for us to grow. In AI, MI250X and MI300A enabled the exascale supercomputing error. I'm very happy to say, actually, this week, there was a new top 500 list that was released, and AMD powers the two fastest supercomputers in the world. So that's pretty cool. Thank you.

[APPLAUSE]

And with MI300X and 325, we've extended that leadership to Gen AI with large scale internal and cloud deployments at Microsoft, Meta, Oracle, and many others. And I'm happy to say we've added a lot of new Instinct customers in the last nine months. Today, seven of the top 10 model builders and AI companies are using Instinct in their data centers. Leaders like OpenAI, Meta, xAI, and Tesla. Innovators like Cohere, Luma, and Essential, and many, many more. You're going to hear from several of them. They're our guests here today. And they'll tell you a little bit about how we work together.

Now, as powerful as our hardware is, it's truly the software that enables their full potential. And I hear from lots of you as developers on what we can do better in software. I can say that I hear you and our ROCm software stack continues to make just incredible progress. We're really focused on broadening the coverage for AI models, accelerating the pace of our releases, and really setting a North Star of a developer-first mentality with ROCm.

When you hear me talk to our engineers, what I say it is all about the developer experience. It's all about what you guys say. And this is our guiding principle. So you're going to hear a lot about that from Vamsi today. And then for those of you who are going to be able to stay with us this afternoon, we have a ton of developer content to just show you how you can really use AMD and ROCm.

Now, to give you some perspective about what it's like to use AMD, I'd like to bring out my first guest. One of the newest partners who is running Instinct in their production environment, is xAI. And here to share more, please welcome Xiao Sun.

[LIVELY MUSIC]

Hello, Xiao.

XIAO SUN: Hi, Lisa.

LISA SU: How are you?

XIAO SUN: I'm good. How are you?

LISA SU: We have a decent audience today. What do you think?

XIAO SUN: Oh, that's a great audience.

LISA SU: Xiao, we are super excited about the work that we are doing with xAI. You guys are really at the forefront of developing state of the art AI models. You're going super fast. Can you share a little bit about what your team does? And how are you managing all this?

XIAO SUN: Sure. Sure. At xAI, we have a very small team, and then we are moving very fast. And we are following first principle. Basically, we advance like Grok family models. And then for maximum true thinking. And to have that, we actually, basically need to go with the first principle thinking, which is like, we always challenge the status quo. And we always ask the question like, why does things have to be done like this? And could we do it better? We also apply that into our computer infrastructure, which is very important for us.

LISA SU: Yeah, absolutely. We've been part of some of that first principle thinking and how you really are focused on speed. Look, we're super thrilled of the work that we've done together on MI300X at xAI. I asked you guys to give us a shot. Can you talk about how you're leveraging the MI300 infrastructure? Like how has it worked? How did it come up for you?

XIAO SUN: Yeah, yeah. So if I use one word, that word is basically effortless.

LISA SU: Can you say that word again?

XIAO SUN: Yeah, indeed. It's effortless to use--

[APPLAUSE]

--to use AMD GPU in our product. As I mentioned, we are a very small team moving very fast. So for us, the most valuable resource is the engineering time. So the opportunity cost is immense. So with your team's help, we do not need to spend too much time. Basically, just a few of us engineers and your team, we successfully pushed one very important product, Grok family model into product. And I remember when we first started to collaborate together, I looked at, oh, there's a meeting on Friday. I think, is this very important? Can we just meet now?

So after that, your engineers adapted to our pace. So I always get a phone call at 9:00 PM or midnight. And then my partner was asking me, who is calling? I was like, oh, vendor calling to ask about some question in kernel. And he's like a violinist. So he's like, oh, that almost never happen in the orchestra. And what is a kernel? So because of that, we collaborated very closely. And then in few months, we can push something into product that's really impressive.

LISA SU: Well, I've been impressed because our engineers are always reporting to me, where are we on the Grok model performance? And you guys have moved super, super fast. Xiao, the other thing is we're talking a lot about open ecosystems. And I know that you guys are a strong believer in open ecosystems. Can you talk a little bit about how ROCm and all of those community efforts have actually helped you?

XIAO SUN: Sure, sure. So as you know, our inference structure is based on SGLang which is like a very popular open source platform. And also the major contributors and authors are also in xAI. So while they are advancing most optimized inference system, they also contribute a lot to the open source community. We upstream our innovations to SGLang public repo.

And at the same time, we also benefit a lot from the open source community. They find bugs, they fix the code, and then we merge that into our production infra. And that really helps a lot. I think it's very essential and we will continue to commit to contribute and work together with the open source community.

LISA SU: Yeah, no. That's great. I think the SGLang progress has been just a great example of how fast things go. So look, I know you guys are always moving ahead and I have a lot of products to talk about with this audience today. Can you share a bit about your perspective of our collaboration? And what are you excited about? What do you think about MI350 series and just all the work we're doing together?

XIAO SUN: Sure. I'm actually very impressed by your yearly cadence about the new hardware. So thinking about the future. I think we will continue to go back to first principles thinking. I mean, you are a pioneer also in semiconductors. So you know that what essentially we are doing is basically like fancy waterworks here, except that it's not a water molecule. So we are basically manipulating electrons. We're pumping the electrons in very high energy level, and then we guide it through the channel of the transistor to the gate of transistor and then dissipate it to the ground. That's how we do compute.

But I think that this is not the end of it. This is the start of it. There are a way to do it, like probably 1,000 or if not 1 million times more efficiently. And also on our side, one way of thinking about the compute is basically compression of data. The data is all the text that human has ever written. And I think now and in the future, will be like all the realities, all the truths in the world. And probably even further future, there will be all the state of affairs that has not yet happened, but could happen. So we compress them all and then put them into your USB disk or something. And then when you need to use it, you retrieve it and decompress it. This is how we think about it.

But both sides have many innovation to do. But we cannot do it separately. So this is basically from my point of view, from silicon to product, this is like the largest co-design of human history. And then at xAI, we are very happy to collaborate with vendors and AMD to do this large co-design together, accelerate the iteration. And I hope that all the talents from the world should join on both sides.

LISA SU: That's fantastic. Xiao, thank you so much for joining us today. Thank you for your partnership with us on MI300, and we look forward to doing a lot more together.

XIAO SUN: Thank you, Lisa.

LISA SU: Thank you, Xiao.

[APPLAUSE]

Well look, we have a full lineup today of new announcements across hardware, software, and solutions. So let's go ahead and jump right in. Now since launching MI300 less than two years ago, we're on an annual cadence of new Instinct accelerators. With the MI350 series, we're delivering the largest generational performance leap in the history of Instinct. And we're already deep in development of MI400 for 2026, that is really designed from the grounds up as a rack level solution.

So today, I'm super excited to launch the MI350 series, our most advanced AI platform ever, that delivers leadership performance across the most demanding models. This series, you'll hear us talk about the MI355 and the MI350, they're actually the same silicon, but MI355 supports higher thermals and power envelopes so that we can even deliver more real world performance. And thank you, Drew. My favorite part.

[APPLAUSE]

Here is MI355. This is our flagship products and I'm showing this to you. It's powered by our latest 4th gen Instinct architecture. It supports new data formats like FP4. It uses the latest HBM3E memory. And it has 185 billion transistors across 10 chiplets, all integrated with our leadership 3D packaging. So what do you guys think?

[CHEERS, APPLAUSE]

All right, thank you, Drew. Look, the MI350 series delivers just a massive 4x generational leap in AI compute to accelerate both training and inference. With an industry-leading 288gb of memory, we can now run models up to 520 billion parameters on a single GPU. The MI350 series also uses the same industry standard UBB8 platform as MI300 and MI325. This is actually really important because it actually makes it super easy to deploy MI350 series into existing data center infrastructure.

Now, if you look at the specs compared to the competition, 355, supports 1.6x more memory and delivers higher flops across a wide range of AI data types. And especially if you look at FP6 and FP64, we're double the throughput. Now, what does that mean? That means that you have leadership performance at both ends of the spectrum, whether you're talking about leading edge AI models or large scale scientific simulation or engineering applications.

Now, at the platform level, an MI355X server has massive memory capacity and compute relative to the competition. We're talking about 161 petaflops of FP4 compute and 2.3 terabytes of HBM3E memory. And we have it in both air-cooled and liquid-cooled configs, giving customers the flexibility to meet their specific thermal, power, and density needs. Now let's look at some of the performance.

We set an ambitious goal with MI350 series to deliver a 35x generational increase in AI performance. And today, I'm proud to say that we delivered that. On Llama 3.1, MI355 delivers 35x higher throughput when running at ultra low latencies, which is required for some real time applications like code completion, simultaneous translation, and transcription.

We also deliver significantly higher performance across a wide range of AI applications, things like chatbots or content generation or summarization or conversational AI. We can see performance up to 4.2x higher gen on gen. And now when you look across a wide range of models, we see great performance as well. In DeepSeek and Llama 4 Maverick, we're seeing things like triple the tokens per second gen on gen. So that level of performance drives faster responses and the ability to serve more users with much, much greater efficiency.

Now, let's take a look at the competitive performance when running DeepSeek R1 or Llama 3.1, MI355 delivers leadership throughput using open source frameworks like SGLang and vLLM. We're generating up to 30% more tokens per second compared to B200, and actually, matching the performance of the significantly more expensive and complex GB200 even when the competition is using their latest proprietary software stack. This is actually pretty cool because it tells you a couple of things.

It first says that we have really strong hardware, which we always knew. But it also shows that the open software frameworks have made tremendous progress to the point where they are outperforming a closed vendor-specific ecosystem. And when you combine all of that performance with lower capex, what we're seeing is MI355 can deliver up to 40% more tokens per dollar than competing solutions.

[APPLAUSE]

40% more tokens per dollar. That means higher throughput, greater efficiency, much better TCO for cloud and enterprise. And it really makes MI355 the best choice in the industry for inference at scale. Now, one of our earliest partners to deploy Instinct broadly was Meta. To share more on our work together, please welcome Meta VP of Engineering Yee Jiun Song to the stage.

[APPLAUSE]

[LIVELY MUSIC]

YJ.

YEE JIUN SONG: Thank you for having me.

LISA SU: Hey, thank you so much for being here. We're so excited about the work that we're doing together. We have so much respect. Meta has been an incredible leader in AI across infrastructure models and services. I get to talk to YJ a lot. He gives us good feedback, good feedback. You're delivering all of this capability at amazing scale. And so it's been our privilege to be your partner across EPYC and now Instinct. Can you talk a little bit about our collaboration?

YEE JIUN SONG: First, thank you for having me here, Lisa. I'm thrilled to be here and excited to see all the progress that you and the AMD team are making. We're seeing incredible advancements that started with your EPYC products and now extending to all of your AI offerings. I think AMD and Meta has always been strongly aligned on vision, roadmap, execution.

So this means that we have very close co-engineering, performance tuning. We troubleshoot problems together, and we're able to deploy optimized systems at scale. So our teams really see AMD as a strategic and responsive partner and someone that we really rely on.

LISA SU: Well, we love working with your engineering team and you know that. We love the feedback. And you also hold a high standard. You were one of our earliest partners in AI. As your demands continue to scale, you've talked about your usage of MI300. Just what are you doing today and what are your plans with MI350 going forward?

YEE JIUN SONG: So AI has been core to many of the user experiences across all of Meta's products for a long time across Facebook, Instagram, WhatsApp. And now of course, with Llama and Meta AI, AI has become even more important than before. It's been fantastic to see our collaboration on MI300X come to fruition. So MI300X accelerators today are a key part of our infrastructure. We've deployed this quite broadly for Llama 3 and Llama 4 inference due to its high performance and excellent performance for TCO.

As we've gained experience with MI300X, we're also expanding the workloads that we run on them. So we're now today using MI300X both for training and inference of the different ranking and recommendation workloads which are critical to our business. We're also quite excited about the capabilities of MI350X. We like that it brings significantly more compute power, next generation memory, and support for FP4, FP6, all while maintaining the same form factor as MI300 so we can deploy quickly.

LISA SU: Thank you, YJ. Thank you, by the way, for your trust. I know that to deploy us on more workloads requires effort on both sides, so we really appreciate that. Meta has really been a leader developing frontier models. If you think about AI across all of your applications, tell us a little bit about what you're seeing. You're like at the front there. What are you seeing in applications? What are you seeing in your compute investments.

YEE JIUN SONG: So I think the AI work at Meta that gets the most attention is probably the Llama models. We are committed to developing frontier models with our Llama efforts. But that's really just the tip of the iceberg. We're seeing growth of AI workloads across all of our different products. AI is not only improving our existing products, but also allowing us to develop entirely new products. All of this is driving investments in compute infrastructure at a scale that's quite unprecedented.

We're building data centers and filling them up faster than I have ever seen. Our entire infrastructure team is sprinting to ensure we have the capacity to build the next great AI models, and then take advantage of those models to deliver value to our users. The result of this incredible capacity build up is that we really care about perf per TCO, and making sure that we get the best bang for the buck for our investments.

LISA SU: Well, we've been part of a few of those sprints, just a few. So, YJ, we've also worked very closely on the software side, PyTorch, ROCm, the open hardware environment. How do you think about open in your strategy?

YEE JIUN SONG: So I think our collaboration has spanned both software and hardware for many years at this point. I think most recently since 2021, we have partnered closely to enable ROCm through PyTorch to ensure that developers can leverage AMD GPUs with PyTorch's ease of use right out of the box.

LISA SU: Thank you.

YEE JIUN SONG: Beginning last year, we've worked closely on improving ROCm's communication library recall, which is critical for AI training.

LISA SU: We really appreciate that partnership, by the way.

YEE JIUN SONG: Us too. Beyond PyTorch, Meta contributes heavily to the open source community. An example here is the work on compiler frameworks such as Triton, which allows us to write code once and then run on different accelerator families. Now, of course, we also collaborate on optimizing Llama models to run well on AMD GPUs.

Operating at scale also require that our accelerators, the accelerators that we buy, be compatible with our network and data center infrastructure. Here, we rely on our common infrastructure hardware racks to be the integration point between our accelerators, the network, and the data centers themselves. This is one of the reasons why we've been able to introduce MI300 into our production environment so quickly.

LISA SU: Yes. No, absolutely. I think the OCP work is fantastic. Look, we're super excited about our partnership, everything that we're doing together. And it feels like we're just starting the ramp of MI350. But in our business, we're always talking about the future. You guys are always asking us what's next.

YEE JIUN SONG: Absolutely.

LISA SU: I'm actually going to preview MI400X a little bit later in the show. So can you just talk about where you see AI going in the future? And how does that shape what you need from partners like us?

YEE JIUN SONG: So AI is driving massive growth in infrastructure demand. But actually it's not just about the size of the demand or the amount of capacity. The type of workload is also changing very rapidly. As an example, not so long ago, as an industry, we were very focused on pre-training. But towards the end of last year, we start to see the emergence of test time inference and reinforcement learning and other new workloads that demand huge amounts of computation.

We also start to see the rise of mixture of expert models that place high demands in network interconnection speeds and the performance of the collective communication primitives we just talked about. And beyond generative AI, we are also finding that our recommendation systems are also getting more complex. The direct implication of these rapid changes is that Meta and AMD will have to work even more closely together to define our accelerator and network roadmaps for the future. I can't wait to hear what you're about to share.

LISA SU: Well, you know everything I'm about to share. But I will say that I do remember sitting in your office and asking you, YJ, tell me what's going to happen when the workloads? And you're like, well, be flexible.

YEE JIUN SONG: That's absolutely true.

LISA SU: Thank you, YJ. Thank you. We really, really appreciate the partnership. It's wonderful working with you and the entire team. And thank you for all that we're doing together.

YEE JIUN SONG: Thank you.

[APPLAUSE]

LISA SU: All right. So now let's turn to training. In addition to all the work we've done to improve inference, we've also made a lot of optimizations for training. And I'm happy to say we've seen some fantastic results. MI355 delivers significantly better training performance than MI300. And when you look at pre-training, for example, where foundational models are built from the grounds up, 355 is delivering up to 3.5x higher throughput across a range of models and data formats. And in fine tuning, we're delivering up to 2.9x more performance gen on gen, which enables just faster iteration cycles and reduces the time from model development to model deployment.

Now, comparing to the competition, MI355 pre-training performance is actually on par with B200 across a range of model sizes and data formats. And I actually think that's very good considering how new MI355 is. Now, in fine tuning, we just saw some of the latest MLPerf benchmarks that were out there. And MLPerf is largely considered the gold standard for training benchmarks. We see that MI355X actually outperforms both B200 and GB200 when we're talking about completing the benchmark up to 13% faster compared to the latest published results. So that just tells you how much progress we've made in training.

[APPLAUSE]

And now as we talk about solutions, we said we want to make this super easy to use. So with the MI350 series, OEMs and ODMs are launching racks built entirely on AMD technology for the first time. We're combining 5th gen EPYC CPUs, Instinct MI350 GPUs, and our Pensando NICs in an integrated solution. And these are all OCP-compliant designs so they drop right into existing infrastructure.

And in some of the densest environments, we have liquid-cooled racks that can scale up to 96 or 128 GPUs that deliver up to 2.6 exaflops of FP4 compute and 36 terabytes of HBM3E memory. And on the enterprise side, we can do air-cooled systems that support up to 64 GPUs and integrate all of that into an existing infrastructure. So this is the kind of flexibility and range that our customers really want.

What they want is to be able to take the technology, get it into production, get it into the data centers as quick as possible with as little work, as little disruption. And that's exactly what we can do with MI350. Now, one of our most strategic cloud partners that's building with AMD across the stack is Oracle. To share more about our work together, please welcome Mahesh Thiagarajan, Executive Vice President at OCI.

[LIVELY MUSIC]

[APPLAUSE]

Hello, Mahesh.

MAHESH THIAGARAJAN: Nice to meet you, Lisa.

LISA SU: Hey it's wonderful. Thank you for being here. We so appreciate the partnership with OCI. You guys have been with us across the board.

MAHESH THIAGARAJAN: That's right.

LISA SU: You guys are frankly at the center of AI compute. Tremendous momentum.

MAHESH THIAGARAJAN: Doing our best.

LISA SU: Powering some of the largest training and inference clusters. As you look through what's facing you in all of these deployments, what's most important to you?

MAHESH THIAGARAJAN: I'm truly honored, first, to be actually working so closely with AMD and building these AI infrastructure at relentless pace together. Fundamentally, to solve the next frontier of challenges at the intersection of cloud and AI, we need to do a deep integration across the entire stack, starting from power to compute to network to storage to truly use every last ounce of the performance that's available. So let me break that down a little bit.

So when we talk to customers about compute, what we see is that the most demanding training and inferencing workloads need the exceptional weaving of the CPU and GPU memory. This is where I think the AMD Infinity Fabric actually truly enables, offering the performance between moving the data sets really close to the accelerators at AI speeds, and we're seeing customers seeing tremendous value. The second thing, which I think is super important, and I'm very passionate about what AMD is doing here, is really around the high performance networking that comes close to delivering these large training clusters.

And fundamentally, when you think about an AI super cluster, it's operating as one giant supercomputer really looking for that ultra low latency, extreme high performance bandwidth and truly operating as one supercomputer. And so that performance across these nodes really matter to complete that task. And what we partner with and we work with AMD a lot on is on the networking technologies. And for example, the Pensando work that we've been doing for a while truly enables the security of these AI workloads and actually is powering the performance of we're seeing.

LISA SU: Yeah. No, that's fantastic. Thank you. We love that vision overall of putting all these pieces together. By the way, that's exactly our philosophy as well that you need CPUs, GPUs, networking coming together. Now OCI was one of our early adopters of MI300X. It's been great to see some of the customer response. Can you talk a little bit about how AMD Instinct looks in Oracle Cloud?

MAHESH THIAGARAJAN: MI300X on Oracle Cloud Infrastructure is a very deep integration. We've seen massive demand from both AI native companies and large frontier model companies actually doing work on top of OCI today. Now the support model is actually where we partner well together to offer that fantastic experience where a customer comes looking for a cluster. In a moment of minutes, they're able to get their AMD Instinct machines.

Truly with what AMD brings to the table, the latest and the greatest ROCm innovation, the updates. Everything is available out of the box. Step one. Two, what we try to do is also ensure that customers are getting the latest performance benefits of everything that you guys are doing on ROCm very regularly on a monthly update. So the customers actually get the best of AMD immediately.

And third is the latest addition of your PyTorch and vLLM support. We've actually seen acceleration of some of our customers who've been waiting for that support. So they're like, oh man, this is exciting. So let's go use that platform. Some of the largest names, largest customers who you've all heard about are all running AMD Instinct on OCI. And they're having a great experience.

LISA SU: It's wonderful, Mahesh. Look, I really have to thank you and your team. I know that this has very much been about getting exactly what customers need, and you guys have been super, super agile in that. Now, when you think about Oracle and your adoption of technology, we talked about 300. You're actually now leading with us on 355. I know there's a lot of interest there. Can you talk a bit about just the evolution of our partnership and what you're seeing?

MAHESH THIAGARAJAN: Look, I think our partnership probably started about a decade ago. I think we've been partnering for a long time, but I think it started heating up around a decade ago. And it started with us actually using AMD EPYC for our databases. Oracle Exadata Database machines since then, we've been using AMD EPYC. We've seen tremendous performance where we're seeing 3x higher transaction throughputs and 3.6x faster analytics queries.

And the great news about that is that's actually available not only on OCI but on premises, other cloud partners that are actually supporting Oracle databases, including our latest Oracle Autonomous Database. So it's everywhere. And then, something that's very personal to me is our Pensando partnership that truly enables a hardware-based network virtualization and our entire cloud, which is a fundamentally unique innovation that offers security and high perf for every customer that runs on OCI today. And that's with AMD.

And obviously, 18 months ago, we did the AMD Instinct partnership. Lots of customers have grown. And we think there's tremendous demand. We project about a 10x growth over the next year really trying to drive the AMD Instinct platform on OCI. And the most exciting part for me is that we're announcing our partnership on MI355, truly bringing the latest out to the cloud with support for zettascale clusters.

LISA SU: By the way, did you guys hear that, support for zettascale clusters?

[CHEERS, APPLAUSE]

MAHESH THIAGARAJAN: I'm truly excited about that. When I talked about that deep integrated compute network storage, we're going to support AMD MI355. And more importantly, we're actually going to go live in a couple of months with over 27,000 GPUs in a single cluster available on OCI in two months.

LISA SU: That's fantastic. Thank you.

[APPLAUSE]

Now, I'm asking everybody here, this is also about the future. And Oracle has actually been leading on this concept of building gigascale data centers. And we're talking about what we have to do from a system standpoint. Tell us what it means to build gigascale data centers and how we can participate in that.

MAHESH THIAGARAJAN: Yeah. I'm an infrastructure nerd, so I'll talk about power to start. Look, gigawatt scale, I think we talked about this phrase two years ago. Everybody would have been like, what are you talking about? It's insane. But I think the biggest challenge that we see is still power. Power is the fundamental bottleneck that still exists and the speed at which we can build them.

But I think Oracle's partners come here, where we've been big partners with all of the utilities and other businesses and industry verticals. So that has actually been very helpful. And second, we are making investments in sustainable energy, be it green energy around geothermal, wind, solar. And we're also looking at small nuclear reactors that can power this.

The second thing is about time to market. One of the things that OCI pioneers in actually bringing our Cloud Infrastructure in like three racks. So we spend a ton of time tuning time to market. And second, bringing that price performance value to our customers. And for me, today, we're operating at over 100 regions so we're able to reach a far-reaching audience. But I think going back to the price performance message, I really think that's why AMD and Oracle work together.

LISA SU: We're bringing value to customers. That's what we do.

MAHESH THIAGARAJAN: It's all about the price performance. And today, with that price-performance value, customers actually enjoy the AMD plus Oracle partnership. And lastly, we're really excited and looking forward to your 450X platform. I'm seeing some of the specs and I think it's going to be truly special.

LISA SU: Fantastic. Mahesh, thank you so much. Thank you for the partnership overall.

MAHESH THIAGARAJAN: Any time.

LISA SU: And I look forward to everything we're going to do together.

MAHESH THIAGARAJAN: Absolutely. Thank you so much, Lisa. Appreciate it.

LISA SU: Thank you.

[APPLAUSE]

All right. So look, customer excitement for MI350 is very strong based on the performance and cost per token advantages. I'm happy to announce that MI355 production shipments actually started earlier this month. And we have the initial wave of partners on track to launch platforms and public cloud instances here in the third quarter. So really, really excited about MI350.

[APPLAUSE]

Now, another important focus for us is sovereign computing. Around the world, we're partnering with national governments and research institutes to help build the high performance computing and AI infrastructure that is really critical for their economies. And the goal really goes far beyond just building domestic compute capacity. It's really about using AI to power public services, research, and national programs that create societal impact. To get there, governments are actually prioritizing resilient infrastructure. They want open standards, they want flexible architectures, and they want a diverse ecosystem of technology partners.

Today, we have more than 40 active engagements globally powering critical public agencies, national computing centers, and sovereign AI activities. From the world's fastest supercomputers in the US to the rapid expansion of high performance computing across Europe, Asia, and the Middle East, to a wave of sovereign AI deployments around the world. This is a growing part of the market, and we are increasingly spending more time helping nations build their computing strategy and infrastructure.

One of the best examples of our progress is in Europe with our Silo AI team. Silo is our AMD AI lab, but they're also a solutions factory, collaborating closely with governments, industries, and research institutions to develop models and applications aligned with national priorities and optimized on our hardware. Silo is working across Europe, collaborating with companies like Allianz, Nokia, Philips and Unilever, advancing open multilingual LLMs with the European Commission and pushing frontier model research on the AMD-powered Lumi supercomputer.

They're also playing a very important leadership role in the open source AI community, contributing to models and partnering with leading AI innovators like Aleph Alpha, Mistral, and NXAI. Now, another extremely exciting example of our sovereign efforts is our work with Humain a new company with an ambitious vision to build advanced, locally developed AI in the Middle East. To share more about our work, please welcome Tareq Amin, CEO of Humain.

[LIVELY MUSIC]

TAREQ AMIN: Good morning Lisa.

LISA SU: Hello, Tareq.

TAREQ AMIN: Good morning. Good morning, everyone.

LISA SU: It is great to have you here. Thank you so much for joining us. We're so excited about our partnership together.

TAREQ AMIN: Well, first of all, thank you very much for inviting me here. I don't need to tell you this, but AMD is really an important partner for Humain, an important partner for Saudi Arabia, and also an important partner for the entire larger ecosystem of AI companies.

LISA SU: Look, you guys are on an exciting mission. I had the pleasure and honor to be with you in the kingdom just last month. And the vision that you're laying out, launching Humain, it's such an important moment for the kingdom and just taking sovereign AI to the next level. So can you share with us, just tell me about our vision, your plans, everything.

TAREQ AMIN: So Lisa gave me four minutes.

LISA SU: You can take more.

TAREQ AMIN: This is the biggest challenge I have. But I wish I could share with you what we have done last month. I'll take a perspective just to tell you the entire story and the partnership that we're doing with AMD to redefine the entire AI infrastructure ecosystem. In the US, I had the opportunity to build digital infrastructure across 22 cities. I moved to India, where I learned how to scale things. I moved to Tokyo, where I built technology that was in a research paper, realize it. And the story gets completed with Humain.

Humain and Saudi Arabia came together through the consolidation of various enterprises in the country, and also one government entity that was developing large language models. Our obsession is about disruption via technology. The way we pick partners is not based on what I call transactional selections. When we met Lisa and her team, we really hit it off because we both agreed that we're going to co-own the outcomes. It was very, very important that co-owning the outcomes and having skin in the game, taking a risk to build something that is good for humanity was a very, very important mission.

So today, in front of all of you, though, we've talked about this, the announcement about the joint venture with AMD, I am really thankful for your support for what we need to do. But I want to give you a glimpse of what this really means. We are committed, and when we looked at the advantage of what Saudi Arabia can really do. By 2030, the deficit in power is estimated to be around 100 gigawatt. No matter what you do, you will still need power to build the capacity that we need for AI.

This is an added advantage that we thought we could really help, and we could participate into this AI global ecosystem. We have an abundance of land, an abundance of power, mixture of renewable as well as traditional energy, and a really, very young society that is hungry to learn. So we thought this could be great. Our commitment for all AI developers and AI companies, what if we reduce your cost of ownership by 30%? From whatever you could achieve as the lowest worldwide cost, I'm committing to make that together with Lisa, the lowest cost.

LISA SU: Sounds like a good commitment.

[APPLAUSE]

TAREQ AMIN: So we're really, really happy. This is a game-changing moment. We're really privileged that this joint venture is going to be a game-changer. 2030, 1.9 gigawatt, 2034, 6 gigawatts. It starts in Riyadh, but it doesn't stop there. We will go and look at other global opportunities to build our infrastructure.

LISA SU: Tareq, I want to just point out some of the things that you said. We've talked about the need for power. We've talked about the need for speed. And we've talked about the need for efficiency in what we're doing. I think what impresses me the most about the work that we're doing together is you really have a clean sheet of paper to talk about what's next. So we've talked about a lot of AI infrastructure, both in the kingdom and outside of the kingdom. Can you just talk about some of the milestones that we have in place?

TAREQ AMIN: So I think as soon as we really crafted this agreement, the timing of launch Humain was not also coincidental. We were really happy that it was coincided with the presidential visit into Saudi Arabia to talk about relationship and partnership they're doing with technology companies such as AMD. We have already started the construction of two large campuses, 11 data centers, each one of them of 200 megawatt capacity each. I will tell you, Lisa, almost on a weekly basis, Tareq, we need to move faster. We need to move faster. So I really appreciate your spirit.

LISA SU: I have some MI350 for you that need data centers.

TAREQ AMIN: By this year, our entire build is to get our first 50 megawatt done and then we start scaling up on 50 megawatt modules every quarter. So my entire obsession now is about the infrastructure layer. One thing that I think all of you saw when Lisa was talking about the new generation, I mean, I would tell you, congratulations on MI350. I could not even be more excited about, what, 2026? I think the MI400 series is a game-changer for our industry. But realize that what we are doing with AMD is not just I'm buying chips.

Lisa and her team have enabled us to really disrupt the TCO. Second is about openness. We talked about this. We said we need an inclusivity. The world working together is a better place than us being fragmented. And the idea that we build an open ecosystem, inviting many others to participate, including AMD, including Cisco, and many other financial partners that are going to come and take this hopefully as a blueprint of what we need to do to address the gap that the world have in energy.

LISA SU: That's fantastic. Tareq, thank you again for the incredible partnership. We are super excited about what we're doing together. I think we're super excited about what we're going to do for this AI ecosystem going forward.

TAREQ AMIN: Thank you, Lisa. Thank you very much. Thank you. Really appreciate it. Thank you.

[APPLAUSE]

LISA SU: All right. So you can see there's just a lot of excitement on MI350 and our roadmap. Now, as exciting as the hardware innovation is, it is really the software that unlocks the full potential of AI. So to share more about everything that we're doing in ROCm and the developer ecosystem, please welcome AMD Senior Vice President of AI, Vamsi Boppana to the stage.

[LIVELY MUSIC]

VAMSI BOPPANA: Hey, Lisa. Thank you, Lisa. Good morning, everybody.

[CHEERS]

AI innovation is advancing at an unprecedented pace, reshaping compute and redefining what's possible. Our vision for ROCm is simple, to create an open, scalable software platform that unlocks this AI innovation for everyone, everywhere. And over the past year, we've made tremendous progress realizing this vision.

By partnering deeply with the open ecosystem, we are delivering a credible alternative that the industry can trust. ROCm is now powering AI platforms at scale, delivering some of the most demanding workloads on the planet. So today, I'm so excited to show you how far we've come and why this is just the beginning.

Now, last year, around this time, we were super focused on delivering leadership inference performance to our largest customers. Since that time, we have significantly expanded our customer base, accelerated our inference capabilities, and now added training support across key models and frameworks. We have been relentlessly focused on what matters most, making it easy for developers to build with better out-of-the-box capabilities, easy setup, more collateral, stepping up community engagements. We've been running hackathons, contests, meetups, and more.

And our customers are deploying AI capabilities at unprecedented pace. And that's why we've significantly accelerated our release cadence. New features and optimizations are now shipping every two weeks. Leading models like Llama and DeepSeek work on day zero. We've also responded to asks from the community for more industry benchmarks, starting with inference. And now, for the first time just last week, we've demonstrated leadership training performance at MLPerf.

Our collaboration with the open source community is deeper than ever before. Over 1.8 million Hugging Face models now run out of the box on ROCm. PyTorch now has a performance CI in addition to functionality. We've added vLLM, SGLang CI pipelines on our latest hardware.

A great example of our collaboration is the work we are doing with Triton. After achieving functional enablement last year, we have been laser focused on delivering performance in recent releases. And now in the last year, we've added significant support for JAX. With libraries like max text, we're seeing increasing adoption of JAX in our lead training engagements.

Now as we look ahead, the world of AI never sleeps. The pace of innovation is only accelerating at every layer of the stack, from hardware to algorithms to models and applications. And all of this is happening at scale. Our customers continue to need feature velocity and performance gains to stay at the forefront of AI. So today, I'm super proud to announce ROCm 7.

[CHEERS, APPLAUSE]

ROCm 7 is bringing exciting new capabilities to address these emerging trends and brings support for our MI350 series of GPUs. It continues our relentless focus on usability, performance, introduces the latest algorithms, advanced features like distributed inference, support for large scale training, and new capabilities that make it easy for enterprises to deploy AI effortlessly.

Within ROCm 7 inference has been the largest area of focus. We've innovated and invested at every layer of the inference stack, from the latest framework enhancements in vLLM, SGLang, implementing serving optimizations, supporting advanced data types, to delivering extremely high performance kernels, to implementing the latest algorithms like FlashAttention V3. We made it easy to author and integrate kernels with Pythonic abstractions, and we've done significant work in our communication stack. This is how ROCm 7 delivers over 3.5 times the performance of ROCm 6.

And when it comes to inference serving frameworks, it's becoming more and more clear that open source feature velocity and performance is in fact outpacing proprietary alternatives. Just look at what's happening in frameworks like vLLM and SGLang. They're actually setting the pace on commits and have both enabled FP8 optimizations and support ahead of closed alternatives. Working closely with these open source communities, MI355 is today delivering up to 1.3x better throughput on DeepSeek FP8 when compared with B200. That's the power of open collaboration, moving fast and delivering more.

[APPLAUSE]

One of our earliest partners that's innovating at scale with ROCm is Microsoft. So to talk about our work together, please join me in welcoming Eric Boyd, CVP AI platforms from Microsoft.

[LIVELY MUSIC]

ERIC BOYD: Vamsi.

VAMSI BOPPANA: Eric, so good to see you. Thank you for joining us this morning.

ERIC BOYD: Yeah, really glad to be here.

VAMSI BOPPANA: Yeah. We have been very close partners for a long time. Can you tell us a little bit about how that partnership has evolved and particularly around Instinct?

ERIC BOYD: Yeah, sure. as you know, we've been using several generations of Instinct. It's been a key part of our inferencing platform. And we've integrated ROCm into our inferencing stack, making it really easy for us to take and deploy new models on the platform.

VAMSI BOPPANA: That's great. Now tell us a little bit about the type of models, what kind of work our teams are doing together.

ERIC BOYD: Yeah. So at Microsoft, the customers that come to AI Foundry or even our internal customers are looking for the cutting edge leading models. And so models like GPT-4.0 or 4.1 from OpenAI. And the Instinct chip really gives us great performance on top of that platform, really enabling us to scale and perform at tremendous scale and low latencies that we need.

VAMSI BOPPANA: That's great to hear. And we've been super lucky to have collaborated with your team over the years. Tell us a little bit about the role AMD plays in enabling performance, efficiency, and what flexibility does it provide in your infrastructure?

ERIC BOYD: So when you're serving these large language models, one of the big challenges is taking advantage of all the memory on the chip. And so the models have tons of parameters and they have caches and things. And so the more memory you have available and the better bandwidth, the better performance you get and the better latency that you get out of it. And so the Instinct chip brings a large memory footprint, along with really dense compute across it. And that all combines to give us really great TCO benefits as we use these chips to serve our platform.

VAMSI BOPPANA: That is so great to hear because that's exactly how our engineers have been thinking about it when designing these features.

ERIC BOYD: They did a good job. Yeah.

VAMSI BOPPANA: You've expanded from the original set of models now to actually working with more open models. So can you share a little bit more about the work there?

ERIC BOYD: Yeah, of course. At AI Foundry, we're committed to making sure customers get the most advanced models from OpenAI, Mistral, Cohere, other companies like that. But we have over 11,000 models in our catalog, and most of those are open source. I think one of the interesting things over the last few months has been the emergence of DeepSeek as an open source model that provides really great quality in it.

And we inference the DeepSeek model on SGLang, which is an engine that's open source that we've contributed to, adding things like predictive sampling and the like to it. And being able to use that sort of open source framework has really accelerated the development in this space. And of course, ROCm's integration with open source makes all of this really easy for us to deploy at scale.

VAMSI BOPPANA: Yeah, that's been so refreshing, all the work that we've done in the open. Now, as you look ahead, you've, again, expanded the footprint of activities. And now, we're looking at training. So that's super exciting. So maybe you share a little bit about what we're doing there.

ERIC BOYD: Yeah, it's really interesting. As we look forward, we've seen such tremendous growth in inferencing. And we don't see any signs of that slowing down. And Instinct looks to be a key part of our platform on inferencing going forward. But it's also great that it works really well as a training chip. And so we've been able to train on 2100, MI300Xs a state of the art multimodal model in our research team. And really being able to use the same platform for inferencing and for training gives us tremendous flexibility in our data centers. And as we look forward, we're really excited to continue partnering with AMD on our inferencing and our infrastructure solutions.

VAMSI BOPPANA: That's awesome. And actually, this afternoon, there's more information. There's actually a nice talk on the work around training. So I encourage you to go hear about that. Thank you so much, Eric. It's been great having you and wonderful--

ERIC BOYD: Thanks so much, Vamsi.

[APPLAUSE]

VAMSI BOPPANA: Microsoft has been an incredible partner right with at scale deployments, running everything from closed source GPT models to now open source DeepSeek, and extending the work now to large scale training. So talk about training, it's an increasingly important area of focus for us, and ROCm is making big strides there too. ROCm now supports all major parallelism strategies with functionality across major frameworks and libraries, including PyTorch, JAX, torchtune and TorchTitan.

And look, we're just not enabling models, we're also building our own. Training on ROCm internally at AMD is helping us improve performance, reliability, and the developer experience. And just like inference, training performance has also taken a big leap. ROCm 7 delivers three times the performance of ROCm 6. More importantly, our users are actually telling us that they're now scaling confidently with ROCm. And one of those users is a tier one leader in AI models. Please welcome Aidan Gomez, CEO and Co-founder of Cohere, to stage.

[LIVELY MUSIC]

Aidan, thank you for joining us. It's so good to have you here.

AIDAN GOMEZ: Thanks for having me.

VAMSI BOPPANA: Tell us a little bit about Cohere and your vision for where you're heading. Actually, before I do that, I actually should introduce you. Everybody knows you as a famous AI person. But there was this seminal paper, Attention is all You Need, and Aidan was one of the authors of that paper.

[APPLAUSE]

AIDAN GOMEZ: Thank you. Thank you.

VAMSI BOPPANA: Tell us a little bit about Cohere.

AIDAN GOMEZ: Yeah, it'd be my pleasure. Thank you for having me. Cohere, what we do is we build highly secure and private AI specifically for enterprises. And our focus on security and data privacy means that we can serve large global enterprises in some of the most highly regulated industries like finance, health care, manufacturing, the public sector. And our products, in particular our AI workspace North, it gives AI agents the tools that they need to carry out extremely complex tasks securely.

And so that spans the normal stuff like emails and calendar and docs, but also the much more sophisticated stuff like ERPs, CRMs, and even custom internal tools that are secured behind firewalls. And with our models and our product North, we're giving enterprises control to really let them customize it to their needs and leverage all of their data in a secure and private environment. And most of Cohere's use cases rely on secure links to internal data. And that lets employees at large enterprises automate tasks around HR, customer support, finance, and even the supply chain.

VAMSI BOPPANA: That's great. Now you've been working with Instinct, running your models, inferring on them, and running training on them. Tell us a little bit about how things have been going.

AIDAN GOMEZ: Yeah, it's been going great. The partnership has been accelerating massively. So we were able to port our most recent model, Command A, over to the AMD platform super easily, very quickly.

[APPLAUSE]

And our stack and models are now actively deployed on AMD, and even at leading enterprise customers and global leaders like Fujitsu. And we're extremely excited to start training at scale on AMD GPUs. Instinct's compute, and memory characteristics make it a great platform for training our next model. And we're very pleased with how things are going and looking forward to all the innovation that's been announced here. And we're excited to get access.

VAMSI BOPPANA: Yeah, we're equally excited as well. Our teams are collaborating super close together. Tell us a little bit about how you're taking advantage of the memory system in Instinct, particularly as you serve large models and more complex models like reasoning.

AIDAN GOMEZ: Yeah. So for agentic systems and complex reasoning, they really depend on the context window that our models are able to support. And that can apply a lot of pressure to the memory that's necessary to serve these models. That's because for agents and for reasoning, they spend a lot of time at inference, consuming tons of external data and putting that into the context, as well as reasoning over that data and thinking in their heads before they actually respond.

So each one of these increases the computational demand on the hardware. And so the higher memory capacity and the strong memory bandwidth of AMD's chips have led us to fit longer contexts onto the GPUs. And I think most importantly for us and our customers, it helps lower the overall footprint that's needed for our models, and that drives down the total cost of ownership for our customers.

VAMSI BOPPANA: That's great. Again, super delighted that the memory system is proving to be extremely valuable for you. Now, as you look ahead, what do you foresee as the next set of things coming for enterprise AI and what breakthroughs do you envision?

AIDAN GOMEZ: So on the future, I'm extremely bullish about AI agents. I think that they're going to be deployed and used at scale. And we'll see a huge impact to both productivity and the types of work that employees spend their time on, what their day to day work looks like. So agents are going to allow people to go beyond just augmenting work and towards actually fully automating tasks which take hours, days, or even weeks.

And so an example of that would be doing research over the course of weeks to answer some sophisticated question. Can we compress that down into a matter of days or even hours? So I'm really excited about AMD's roadmap with the MI350 series and the rack scale MI400 solutions. It's a great choice and offering for our customers, and we can't wait to team up with you on it.

VAMSI BOPPANA: That's awesome. Thank you so much for joining us.

AIDAN GOMEZ: Thank you.

[APPLAUSE]

VAMSI BOPPANA: Cohere is training and serving on AMD. We are so excited that we've been able to earn their trust at every level of the stack. Now, as inference becomes more computationally intensive and gets pervasively deployed into applications across industries, it is critically important to drive down its cost. And one of the most exciting new opportunities to drive down inference costs is distributed inference. So let's talk about it. Let's talk about distributed inference.

In any LLM serving application, there are two phases. There's a prefill phase and there's a decode phase. While it's simpler to deploy, in a traditional inferencing serving applications, these two phases of the model are typically handled on the same GPU. But now, if you apply it on the same GPU, it often becomes a bottleneck for large models or when demand spikes happen and you can get limited in performance or flexibility.

We can significantly improve throughput, reduce cost, and boost the responsiveness by disaggregating the prefill and decode phases. Prefill and decode can be now assigned to specialized GPU pools, which can be independently optimized. And with sparse MOE models, and expert parallelism, there's is even more room to optimize. We have a great solution, coming for distributed inference on AMD platform.

Staying true to our strategy, we are embracing an open approach, building alongside an ecosystem of vLLM, SGLang, and llm-d. New technologies like GPUDirect access and DPP deliver significant performance gains. Together, this stack enables a truly open and performant foundation for next generation distributed inference workloads.

Now, as AI is moving into real world enterprise deployments, ROCm is evolving to meet those needs. Enterprises need more than just raw performance. They need end to end applications that helps teams hit the ground running, enabling easy and secure data integration for compliance and trust, and supporting robust workflows for ease of deployment. To make all of this possible, today, I'm excited to announce ROCm Enterprise AI.

[APPLAUSE]

ROCm Enterprise AI makes it easy to deploy AI solutions. With new cluster management software, it ensures reliable, scalable, and efficient operation of AI cluster. And our MLOps platform allows fine tuning and distillation of models with your own data. And a growing catalog of models will come for specific industries.

We partner closely with our ecosystem to deliver end to end applications that integrate existing workflows and data systems, sometimes structured and sometimes unstructured. And to show how all of this comes together in a production enterprise stack, and also discuss our strong collaboration on distributed inference, I am excited to welcome to the stage Chris Wright, CTO of Red Hat.

[LIVELY MUSIC]

CHRIS WRIGHT: Hey there.

VAMSI BOPPANA: So good to see you, Chris. Thanks for joining us.

CHRIS WRIGHT: You bet.

VAMSI BOPPANA: Now Red Hat and AMD, we've been collaborating for a long time starting with our x86 64-bit architectures, but now we're extending it to AI. Tell us a little bit about what's exciting. Where do you see AI getting traction in enterprises today?

CHRIS WRIGHT: Well, man, I love that you brought up 64-bit x86 because we started there and it's been a long time.

VAMSI BOPPANA: Yes.

CHRIS WRIGHT: We actually followed that up with virtualization. And that support and effort, these things aren't static. So fast forward to today and that virtualization support is more important than ever, as customers are looking for options to really virtualize their data center. And you guys just shared some amazing numbers at Red Hat summit a couple of weeks ago with 77% OpEx savings and 71% power reduction. AMD and Red Hat together powering the virtual data center. So that's really cool.

[APPLAUSE]

Now, as for AI, quite a few things are happening. First, you've seen it here today. We've talked a lot about it. The surge of open. And some of that is open source software, the frameworks, things that we're more familiar with but also open LLMs. And today, they have capabilities that are on par with the really proprietary large scale models, including things like reasoning. So they're there or even outperforming in some cases.

Second, the emergence of vLLM. This is something really important for Red Hat work that we're doing together. And this makes high performance inference deployments of open models easy. And then third is bringing this vLLM support to a broad set of accelerators like AMD's. And so all of this together creates this ease of use to generate real efficiency and then choice for companies today.

VAMSI BOPPANA: That's so good to hear. Now, we're not stopping there. Together, we've announced llm-d, an open source distributed inference framework. Tell us a little bit about why it is so significant for AI.

CHRIS WRIGHT: You've seen it here today already. Talking about reasoning, talking about agents, talking about token production, and driving down the cost of token production. So a key challenge for the data center today is lowering the cost of token production. It's not just tokens per dollars, but it's also tokens per dollars per watt. So really thinking about the overall efficiency to meet the Gen AI demands of reasoning models and agentic workflows.

Reasoning models literally produce more tokens as they effectively think to produce results. And so our llm-d project is trying to address this need. How do you distribute and saturate these amazing instinct processors with requests to respond to inference?

You mentioned a little bit earlier, the disaggregated prefill and decode, and these are the low level technologies that we're building into llm-d. llm-d builds on vLLM and then extends that into a distributed environment with Kubernetes. So we're so thrilled that you're joining us together in this journey and bringing your experience so that we can create this critical kind of industry initiative.

VAMSI BOPPANA: Our weight is behind vLLM and the open communities. And now with llm-d, we can get to extend that further. So let's shift a little bit to OpenShift. Openshift AI is playing a key role in simplifying AI for enterprises and making it easy to deploy. How are we working together with OpenShift and what role do our platforms play in that product?

CHRIS WRIGHT: Broadly, Red Hat AI and AMD's processors, CPUs, GPUs together bring this efficient, production-ready AI environment. So vLLM and llm-d are a key part of the Red Hat AI portfolio, which includes OpenShift AI. It includes the Red Hat Inference Server specifically. And then the AMD Instinct GPUs are fully supported within OpenShift AI. So a lot of work that goes into bringing that to life.

And then this delivers this powerful AI processing across hybrid clouds. You heard Lisa talking about cloud, data center, edge, even consumer devices. So that we can deliver something for our customers to efficiently use these precious resources. Openshift AI is both predictive and generative AI support needs smart CPU and GPU choices. And our work with AMD ensures this flexibility, and maximizing the customer investment so they're getting the most out of the hardware that they're procuring.

VAMSI BOPPANA: Yes, super exciting and very, very happy with the collaboration that we've had around OpenShift. So now as you look ahead, it still feels like we're in the very early innings of enterprise AI. So what excites you about what's coming and the work that we can do together?

CHRIS WRIGHT: Yeah. Early and yet moving so fast that things change fundamentally daily. Yeah. So I think it's clear that Gen AI is going to deliver huge value, both in terms of efficiencies or net new value for enterprises. I think the pressure is on each and every one of us to help get from those pilot projects, those POCs, into production. And so our mission is to make that as efficient and accessible as possible.

And much in the same way that Linux brought to life all these applications across different kinds of infrastructure, we're doing that. We're entering this same era with AI. And so to me, I think it's happening, right now, with Red Hat AI and AMD and what we're doing together to really unlock that I value for enterprises across every different kind of industry vertical.

VAMSI BOPPANA: That's so good to hear, Chris. We are super grateful for all the work we are doing together. Our teams love working with each other. Thank you for joining us.

CHRIS WRIGHT: Absolutely. Thank you.

[APPLAUSE]

VAMSI BOPPANA: With OpenShift AI and ROCm, we are now enabling enterprises with Gen AI workflows. I'm especially excited with the joint work we've done on llm-d to slash the cost of reasoning and now agentic-based inference. So now, none of this happens without developers. So let's talk a little bit about what we are doing there.

We are deeply, deeply committed to delivering an exceptional developer experience. We have significantly stepped up our efforts to make the out-of-the-box experience better and deliver great collateral. From videos, blog posts, tutorials, we're helping developers ramp up fast. And with frequent meetups, hackathons, contests, we're building a community.

I was actually so excited to see that our recent contest, Developed GPU Kernels, generated huge interest with thousands of submissions, including a high schooler who wrote high performance Triton kernels. That was just so good to see. And over the last year, as we have enabled the cloud access to AMD GPUs, there's been a big ask from the development community for an AMD developer cloud. So today, I'm super excited to announce the AMD Developer Cloud.

[CHEERS, APPLAUSE]

Instant access to AMD GPUs. No setup. Pure development velocity. Every developer in this room has a 25-hour free GPU credit email in your inbox, no strings. Just launch. Go.

[CHEERS, APPLAUSE]

Now to show it in action and to tell you about all the collateral we are going to bring to you as part of this DevCloud, please join me in welcoming Anush Elangovan and Sharon Zhou to stage.

[LIVELY MUSIC]

Hey, Anush. Hey, Sharon. Anush is responsible for a number of open source software efforts here at AMD, and is actually well known for his huge passion working with developers. He was previously the CEO of Nod.ai, a company that was famous for the open source compiler contributions.

I'm also thrilled to welcome Sharon. Sharon is also very well known in the AI community. A former Stanford faculty, she was the CEO and founder of Lamini. I am delighted that Sharon and her talented Lamini team joined us recently, with a focus of delivering rich content for developers. So Anush, tell us a little bit about the DevCloud and all the goodies.

ANUSH ELANGOVAN: Thanks, Vamsi. Developers, developers, and developers.

[CHEERS]

That is the new mantra of ROCm. We are serious about bringing ROCm everywhere and to everyone, from client to the cloud. In AI, speed is your mode. Access to compute is paramount. We've been delivering on speed, so now, let's get you access to compute. Today, we are announcing the general availability of the AMD Developer Cloud. With the Developer Cloud, anyone with a GitHub ID or an email address can get access to an Instinct GPU with just a few clicks.

All right, let's see how easy it is to get access to an AMD GPU on the cloud. Go to devcloud.amd.com. Say hello to our legal friends and sign up with GitHub. That's it. You can choose between a one-GPU VM or an eight-GPU VM, and you select the operating system that you'd like to use. One of the cool new features of ROCm 7 is that we've made it really, really easy to install. Just pip install ROCm. In case you forget, we've also printed it in a T-shirt and it's in your goodie bag.

[APPLAUSE]

We've also included a lot of easy to use frameworks like vLLM, SGLang, PyTorch, et cetera. You just select one of those frameworks, add your SSH key, and then create and you're set. We've also spent a lot of time building a lot of Jupyter Notebooks, making it easy to use. And if you've been tracking the latest attention algorithms, the log linear attention came out a few days ago. You could try something like that on the MI300X just in a few minutes. And we're just getting started. ROCm is open, proven, and now really, really accessible. Sharon.

[APPLAUSE]

SHARON ZHOU: Hi, everyone, I'm Sharon. I've taught AI to nearly a million people, many of you at Stanford as well as on Coursera with my startup Lamini. As Vamsi and Lisa just shared, I'm super excited to announce that Lamini has now joined AMD.

[APPLAUSE]

I'm personally very excited to be part of AMD's AI mission. We're just getting started, as Anush said, alongside extremely talented teammates from Lamini. We're here to make AI and AI compute easier to use and scale for you, the AI developer, you the AI researcher, you the AI leader in this audience.

What you may not know, many of you, in fact, tens of thousands of you have already run on AMD GPUs over the past year. And that's through Lamini courses with myself and Andrew Ng, who you'll hear from later today. And that's on prompting open source LLMs, LLM fine tuning, and improving LLM accuracy in partnership with Meta.

And we're going to amp that up further here by creating a huge set of intuitive, engaging courses from LLM post-training and reinforcement learning to vibe coding agents to GPU programming. All of this humming on powerful AMD Instinct GPUs on our Developer Cloud that Vamsi just announced. And there will be a hands-on tutorial in this afternoon's developer track to get you started. We'll also be out in the community, ears to the ground, listening to your feedback at top AI conferences. So whether you're at a foundation model company, an AI startup, university lab, hacker house, or just someone attending their first AI hackathon, don't be shy, come say hi.

[APPLAUSE]

VAMSI BOPPANA: Thank you, Anush. Thanks, Sharon.

SHARON ZHOU: Thanks, Vamsi.

VAMSI BOPPANA: So you just saw how easy it is to access our DevCloud. But what if you want to develop locally on your own machine, with your own data? That's where we're going next because ROCm isn't just for the cloud anymore. We are expanding ROCm to Ryzen laptops and workstations so you can build anywhere using the same software stack from cloud to client.

Whether you're on Linux or Windows, cloud or client, ROCm is there. Coming to you in the second half of this year, ROCm will be included directly in major distributions. Windows as a first class OS, fully supported and production-ready. And you can do that on the best AI client portfolio in the industry, capable of delivering breakthrough AI experiences, all locally.

[APPLAUSE]

So we've talked about all the exciting capabilities in ROCm, we've talked about empowering developers everywhere. Now it's time to hear from the builders themselves. So we have a fantastic program for later today. Join us this afternoon for the developer track, featuring leaders that are driving the shift to open and scalable AI. Hear from them how they are enabling their communities to build on AMD.

So as I close, let me leave you with this. We built ROCm to empower the world with an open software platform that unlocks AI innovation for everyone, everywhere. And we've made tremendous strides in just the last year. Our strategy of combining forces with the open source ecosystem is paying off. Together, we are delivering a credible, high performance alternative.

ROCm is delivering some of the most important AI workloads on the planet today. But this is just the beginning. We are going to push forward with urgency, with focus, and with a deep, deep commitment to developers. Because the future of AI is not closed, it is open, it is collaborative, and it is for everyone.

[APPLAUSE]

Now, to deliver AI at scale, we need to bring system level solutions together that integrate computing, networking, software into a unified AI platform. To tell you about all the progress we are making over there. It is my pleasure to invite Forrest Norrod, EVP and GM of our Data Center Solutions group to the stage.

[LIVELY MUSIC]

FORREST NORROD: Vamsi. Thank you, Vamsi. As Lisa said to start this morning, we're moving into the next phase of AI. From a period where chatbots were interesting curiosities to an era where AI drives business and innovation. And agentic AI, as we've heard, is a leading driver of that change. AI agent usage is exploding across use cases and industries, not just automating manual, labor-intensive tasks, but optimizing and automating complex workflows with planning, analysis, and creative problem solving. So not just streamlining processes, but driving innovations across business, science, and product development.

Just as information technology revolutionized the paper-based economy into a digital one, agentic AI brings about another revolution. An innovation revolution where new ideas can be implemented at an unprecedented rate. And so agentic AI has the power to impact workflows across many fields. The key in agentic AI is connecting the power of the LLM models to the business, to the organization, to its datas, tools, and applications.

Agentic flows will employ many models, including specially trained models, each performing their own roles but working together to execute complex tasks. These AI agents execute multi-step processes, many of which will need access to enterprise tools, datas, even humans. So these agents are not simply running isolated on a few GPUs.

Each agent accesses many different resources, applications, databases, unstructured data from social networks, the list could be endless. And they map onto real hardware, onto the GPUs, of course, but also onto a host of CPUs running the applications and processing data going into and out of the GPUs, and onto the network infrastructure, providing secure access to that data.

Now, agents challenge the GPU. They do more than chat. And they need higher performance inference and more memory for larger reasoning models and larger context windows, things that you've heard about earlier today. But equally, CPUs are at the heart of agentic execution, running both enterprise applications as well as managing and orchestrating AI systems. And the data fueling all of this flows across the networks, connecting everything.

But that data includes the crown jewels of any organization, and hence it must not just be accessible quickly, but above all, it must be secured. So thus agentic AI will increase the demands on every part of the data center, not just the GPU, but the CPU and networking as well.

At AMD, we build the technology powering each one of those elements. Our Pensando NICs to securely access data. EPYC CPUs, the industry's best to process the data and manage the GPUs. And of course, the Instinct GPUs to power agentic model execution. Beyond that, the scale-up and scale-out networking for AI scalability allows you to go from small enterprises to a gigawatt data center.

AMD has world class technology in all of these elements, and we have the ability to put it all together. But we also believe firmly in the principle of open. We have taken the lead on helping the industry develop open standards, allowing everyone in the ecosystem to innovate and work together to drive AI forward. We utterly reject the notion that one company could have a monopoly on AI or AI innovation. History shows the most vibrant ecosystems are open.

Now, another key belief at AMD is the principle of programmability is critical. AI is evolving so quickly that having fixed function devices or limited accelerators is the wrong approach and will slow down progress. Software innovation for many, including folks like DeepSeek, has shown time and time again the value of flexibility. Putting all of those elements together now in an open, holistic, programmable design results in the optimal platform to power the age of agentic AI. So let's look at each element.

The front end network connects the compute nodes to the rest of the world. It's the bridge to the AI node. With agentic AI, as I said before, data is ever more important and security is paramount. But security is a layered discipline. With AMD's advanced GPU technology, we support encryption, authentication, and east-west firewalling on every connection. The key to all of this is Pensando's flexible third generation P4 engine that delivers data with security and performance.

Turning to compute. Some will naively tell you that CPUs are less important in the age of AI, but that's not correct. With agentic AI, we see an explosion of autonomous agents accessing data and enterprise applications. This increases the needs for efficient, high performance x86 compute across the data center. Then, within the AI server itself, the CPU serves the demands of pre-processing and workload orchestration to keep the GPUs working efficiently.

Our EPYC CPUs with boost frequencies up to 5 gigahertz and the highest server CPU performance available period are perfect to feed those GPUs. But just as importantly as performance, the CPU needs to be able to seamlessly integrate into a user's environment. Our x86 EPYC CPUs not only bring trusted enterprise reliability, but provide architectural consistency across the data center, increasing flexibility and performance, enabling workloads to move seamlessly to wherever they can get the best levels of latency and throughput.

Now, let me show you a few examples of how the right CPU can make GPUs work better, and how choosing poorly can create bottlenecks that strand valuable resources. As you can see, across a range of models and use cases, our fifth generation EPYC CPUs can boost the inference performance of the entire system from 6% to 17%. That makes a huge impact on the overall TCO and performance of the AI deployment. And it's a critical element in designing the best possible AI system. So get much more out of your GPUs with the right CPU.

And as AI gets more advanced, particularly with new model architecture innovations like mixture of experts, or as MCP becomes ubiquitous, the right CPU will continue to be critical in delivering AI performance. So the CPUs drive the GPUs. And for five generations, AMD has perfected our Infinity Fabric architecture, connecting the CPUs and GPUs together in a low latency, high speed, coherent interface.

As part of our belief in open standards, we donated key IP from Infinity Fabric to the Ultra Accelerator Link Consortium. UALink expands the protocol, scaling well beyond eight interconnected GPUs, up to a thousand coherent GPU nodes, enabling AI systems to ramp, deliver GPU performance for training and distributed inference and for whatever innovations software develops next.

Ultra Accelerator Link 1.0 specification has been released. It's a modern load store architecture engineered for the demanding needs of scale-up AI systems, including low latency and high bandwidth. Now, importantly, it leverages the physical interface layers of ethernet, enabling standard components such as connectors, cables, and retimers to be leveraged by the ecosystem and drive favorable economics and reliable interconnect.

And UALink isn't just optimized for performance, it's engineered to scale. This open standard allows customers to build and support tailored systems, scaling up GPUs spread across racks, enabling pod partitioning for efficiency and security, delivering rock solid resiliency, and accelerating performance going forward with support for in-network collectives. But one of the most important features of UALink is that it is an open ecosystem. It's a protocol that can be used in a system regardless of the brand of CPU, accelerator, or switch. It is thus fully open rather than being shackled to one company's systems or technology.

Again, AMD firmly believes in the power of an open, interoperable ecosystem that accelerates innovation and protects customers' choice while still delivering leadership performance and power efficiency. The consortium is steered by some of the largest scale users and suppliers in the world, hyperscalers and leaders in the semiconductor industry. We are excited to invite some of the contributors of the Ultra Accelerator Link Consortium to the stage. Please welcome Jitendra Mohan, CEO and Co-founder of Astera Labs.

[CHEERS, APPLAUSE]

[LIVELY MUSIC]

JITENDRA MOHAN: Forrest.

FORREST NORROD: Jitendra, thank you so much for joining us. I know we're both excited about UALink. Can you tell us, from your perspective, what makes this so exciting and why Astera has chosen to focus on it?

JITENDRA MOHAN: Absolutely, Forrest. But first, those 5 gigahertz CPUs are cool. They make our chip simulations run faster.

FORREST NORROD: Fantastic.

JITENDRA MOHAN: So thank you for the partnership.

[APPLAUSE]

Really stoked to be here. We founded Astera Labs seven or eight years ago with a mission to eliminate AI infrastructure bottlenecks throughout the data center. That's what we've been doing. From the beginning. We have been laser focused on delivering solutions that meet our customers' demands. In fact, we partnered with AMD on PCIe 5 before the spec was finalized. We have a strong track record of taking cutting edge, open standards and delivering market-leading products. At Astera labs, we know an open approach works. It spurs innovation, builds robust ecosystems, and results in wide adoption.

Today, we provide a comprehensive portfolio of connectivity solutions for the entire AI era. Scale-up connectivity is a particular focus for us, because it is the most critical element of AI era architecture. And UALink is purpose-built from the ground up for scale-up. There is no baggage, no backward compatibility. UALink is designed to be efficient, fast, robust, and it combines the best of many protocols.

UALink for scale-up completely aligns with our mission, our expertise, and naturally fits into our roadmap. What is more, our customers are asking us to deliver UALink products to take the next step forward in deploying a truly open rack scale AI platform based on our vibrant ecosystem. And Forrest, in this case, I must say, the customers are coming, we just need to build it.

FORREST NORROD: Absolutely. Completely agree. I'm hearing the same from particularly the key hyperscalers. Now, what do you plan to build on UALink?

JITENDRA MOHAN: Great. Our vision is to provide complete connectivity infrastructure for the entire AI era. This includes purpose-built silicon, hardware, and software to support AI platforms based on custom ASICs and merchant GPUs, including AMD's Instinct solutions.

FORREST NORROD: Fantastic.

[APPLAUSE]

JITENDRA MOHAN: We are at the forefront of scale-up connectivity innovations with our Scorpio X-Series fabric switches and our Aries Retimers. As a UALink consortium board member, we are working with AMD and industry leaders to advance UALink. We have a close up view of the features and time frames needed by our customers to realize their vision of deploying UALink-based open rack architectures. We are working shoulder to shoulder with AMD and XPU partners.

We plan to offer a comprehensive portfolio of UALink products to support UALink deployments at scale. Smart fabric switches, signal conditioner, controllers, and many more. All of these solutions are built on our COSMOS software that provides an unparalleled view into the health of the entire rack. Our cloud scale Interop Lab provides a robust validation environment for ensuring interoperability at rack scale and accelerate time to market for our customers. Together with AMD, we are excited to bring you UALink scale-up AI infrastructure.

FORREST NORROD: Fantastic. Amazing. We're just as excited to be working alongside you and the whole team at Astera and the whole UALink Consortium to drive it forward. Thank you so much for joining us here today, and thanks for your partnership.

JITENDRA MOHAN: Thank you, everyone.

FORREST NORROD: Thanks.

[APPLAUSE]

Now, I'd like to welcome another guest and fellow member of the UALink Consortium, Nick Kucharewski, SVP and GM of Network Switching BU and Cloud Platforms at Marvell.

[LIVELY MUSIC]

NICK KUCHAREWSKI: Good morning.

FORREST NORROD: Nick, thank you so much for joining us.

NICK KUCHAREWSKI: Glad to be here.

FORREST NORROD: Marvell is well known as a leader in custom ASICs and custom solutions for hyperscalers, and you're engaged on many networking topics as well. Tell us what your customers are telling us or telling you about UALink and scale-up.

NICK KUCHAREWSKI: Yeah. As you know, Marvell is deeply involved in infrastructure technology for cloud and AI data centers, including high speed electrical and optical connectivity, switching, storage, compute, and custom silicon. And in that process, we've developed partnerships with customers who are really operating at the forefront of cloud compute infrastructure and AI technology.

And one of the questions we hear often is that what standards-based options exist for building a large scale-up AI cluster that enables high bandwidth, low latency, high reliability, and the capability to scale beyond today's rack level implementations to clusters with hundreds of connected accelerators? Now, UALink link is at the center of that conversation because it enables all of those attributes. And it also carries with it the promise of an ecosystem of interoperable components from multiple suppliers.

FORREST NORROD: Yeah, I totally agree. Now, you've got a pretty broad portfolio already. But tell us, what are your specific plans around UAL?

NICK KUCHAREWSKI: Yeah, sure. So we've been involved with UALink from the beginning. And Marvell engineers are active in the working group, supplying our expertise in high speed interconnect, low latency fabrics, high layer packet processing, and the networking software stack. This week, we announced UALink link as part of the Marvell custom cloud platform for system designs and silicon. Now, this solution can enable next generation scale-up fabrics and endpoints, offering interoperability portability between GPUs and switches for next generation AI infrastructure.

UALink joins the broader Marvell offering for custom AI silicon, which is rooted in decades of expertise in billion transistor design, and our portfolio of design IP, including networking cores, high speed SerDes for rack scale connectivity, co-packaged optics for row scale, and our family of connectivity and switching for scale-out networks. But with UALink, Marvell customers can deliver a platform comprised of their own custom vision, working literally side by side with interoperable silicon GPUs and fabrics from UALink partner companies.

FORREST NORROD: Nick, that's a compelling vision. Customers want choice, and they want the ability to innovate freely. I think together we're going to give that to them. Thank you so much and thank you for coming to visit us today.

NICK KUCHAREWSKI: Thanks very much for having me here today. Thank you.

[APPLAUSE]

FORREST NORROD: So UALink enables scaling up coherent GPUs, soon to over a thousand, but the most complex AI systems need to scale out way beyond that to truly gigawatt scale deployments. That level of scale drove the Ultra Ethernet Consortium standard. UEC leverages the complete ethernet stack, but it's more than ethernet.

The UEC standard defines a whole new transport layer, addressing the challenges of efficient data-center-wide-deployments. The result? An unparalleled scaling capacity of a shared memory fabric to over a million GPUs. UEC delivers a set of capabilities well beyond InfiniBand. AMD is proud to be a founding member of UEC, and we're excited that the UEC standard 1.0 got to full release yesterday.

[APPLAUSE]

And we're proud as well to have the industry's first UEC ready NICs. We introduced the third generation Pensando P4 engines last fall to drive front end networks. But their incredibly flexible and performant P4 packet processing technology allows them to match the rate of innovation, and is ideally suited for the unique needs of back end AI networks.

Pollara 400 supports advanced transport and congestion control innovations from multiple standards and multiple custom solutions for customers, including shortly, UEC 1.0. We've seen Pollara improve AI performance while reducing network costs for customers by up to 22% through higher fabric utilization and more uniform and simpler switch deployments, while also improving system reliability and resiliency by up to 10%. That improvement in resiliency and availability is ever more important as AI evolves into mission-critical agentic applications.

With a back end network, we complete the end to end AI platform needed to support agentic AI and drive AI forward. And at AMD, we know that agentic AI isn't just a vision or a concept, it is emerging here today. Our customers want it, the industry is demanding it, and we are enabling it with a leadership portfolio of products and our open rack infrastructure.

To develop that leadership performance at scale, again, you need more than a powerful GPU. You need a modern open rack architecture purpose-built for AI. You get that with Salina 400 DPUs for front end networks, the fifth generation AMD EPYC CPUs, the AMD Instinct 350 series GPUs, and scale-out networking solutions with AMD Pensando Pollara AI NICs, all integrated together into an industry standard OCP design fully supported with UEC NICs and offering unprecedented performance.

The industry thrives on it requires an open ecosystem. Open done right enables fully optimized rack level infrastructure without proprietary lock in, and enables innovation across the industry. To show us how we take these principles to the next level, please join me in welcoming Dr. Lisa Su back to the stage.

[LIVELY MUSIC]

Thank you, Lisa.

LISA SU: All right. So look, you've heard a lot from Vamsi and Forrest and a bunch of our customers and partners about all the momentum we have across hardware, software, and solutions. But now, let's talk about the future and how we're expanding our rack scale solutions portfolio to essentially deliver compute performance, efficiency, and density that customers need over the coming years.

Today I am super excited to give you a first look at the next big step for our AI roadmap, the Instinct MI400 series. You may hear us call it MI400 series. You may hear us call it MI450. MI400 series is really bringing together everything we've learned across silicon, software, and systems to deliver a fully integrated AI rack platform. And this guy was built from the grounds up for leadership for both large scale training and distributed inference. Let me now introduce you to our Helios AI rack.

[CHEERS, APPLAUSE]

Helios is truly a game-changer. For the first time, we architected every part of the rack as a unified system. That's combining our CPUs, our GPUs, our Pensando NICs, and our ROCm software all together in one platform. And it's really purpose-built for the most demanding AI workloads, from training to the largest frontier models to scaling inference across thousands of nodes.

But Helios has more than just lots of compute. We also have leadership memory capacity, leadership memory bandwidth, and leadership interconnect speed. And all of that is delivered in an open, OCP-compliant rack that supports both Ultra Ethernet and UALink. And when Helios launches in 2026, we believe it'll set a new benchmark for AI at scale.

So think of Helios as really a rack that functions like a single massive compute engine. It connects up to 72 GPUs with 260 terabytes per second of scale-up bandwidth. It enables 2.9 exaflops of FP4 performance. And that is a great number. But Helios goes even further.

Compared to the competition, we support 50% more HBM4 memory, memory bandwidth, and scale-out bandwidth. And these are big advantages. I mean, this is our sweet spot. We've always had this memory architecture. And what this translates in is faster training, higher inference throughput, and the ability to really handle massive models.

Now, let's take a look at each of the components that make Helios possible. Starting with our next generation EPYC processor, code named Venice. Venice extends our leadership across every dimension that matters in the data center. More performance, better efficiency, and outstanding total cost of ownership. It's built on TSMC's 2 nanometer process and features up to 256 high performance Zen 6 cores.

And it delivers 70% more compute performance than our current generation leadership Turin CPUs. And to really keep feeding MI400 with data at full speed, even at rack scale, we've doubled both the GPU and the memory bandwidth and optimized Venice to run at higher speeds. And you heard from Forrest how important the CPUs are. Now, we just got Venice back in the labs and it is looking fantastic.

[APPLAUSE]

Now, at the heart of Helios, though, is the MI400 series. This is truly the most advanced accelerator we've ever built. It's really the engine for the next generation of AI, and it's designed to run trillion-plus parameter models. We deliver up to 40 petaflops of FP4 performance. We have 432 gigabytes of HBM4, and supports 300 gigabytes per second of scale-out bandwidth to connect across racks and clusters.

And now, as you've also heard from Forrest, we need a high performance networking fabric to connect all of that. And that's why we're also introducing Vulcano, our next generation scale-out AI NIC. Vulcano is fully UEC 1.0 compliant. It supports PCIe and UAL interfaces to connect directly both CPUs and GPUs. And it delivers 800 gigabits per second of line rate throughput to scale for the largest systems. Now, with Helios, every GPU in the rack is connected through the high speed, low latency UALink tunneled over standard ethernet.

Now, when you look at our AI roadmaps, every generation is always special, but Helios is truly a giant step forward. With MI355, we're taking a big step forward. You've heard some of that this morning. We're delivering 3x more performance across a broad range of workloads, extremely competitive versus state of the art today. And with Helios, we're bending that curve further. The MI400 series is expected to deliver up to 10x more performance for the most advanced frontier models, making MI400 the highest performing accelerator.

[APPLAUSE]

I think 10x is a good number. Is it a good number?

[CHEERS, APPLAUSE]

Look, if I sound excited, that's because I am excited. And as you might expect, customer excitement for the MI400 series and Helios is really high. Like, these are the types of programs you don't just start today. I mean, we have been working with customers for the past few years to really like just jump ahead of the curve and see what our customers really need.

One of those customers who has been a very, very early design partner, who has given us significant feedback on the requirements for next generation training and inference is OpenAI. And we have a very special guest today. I am so happy to say that this person is a great friend, someone who is really an icon in AI. To hear more about our work, please welcome OpenAI founder and CEO Sam Altman to the stage.

[LIVELY MUSIC]

Can I call you an AI icon?

SAM ALTMAN: I don't think so, but that's OK. You know what? It's your show. You do whatever you want.

LISA SU: Sam, look, we are truly so happy and excited to be your partner. OpenAI has truly been at the center of the universe. Everyone listens to what Sam Altman has to say when it comes to Gen AI.

SAM ALTMAN: I think they just listen to ChatGPT at this point.

LISA SU: Actually, I listen to ChatGPT.

SAM ALTMAN: We'll take it.

LISA SU: Some of the numbers I've seen, like over 500 million weekly active users. Just amazing growth. Give us a little bit of a landscape. Where are we today? What's the state of play? What are you seeing? What's most exciting right now?

SAM ALTMAN: It's definitely been, for us and many other people, just an explosion of usage over the last year. I think the models have gotten good enough that people have been able to build really great products. Text, images, voice, all kinds of reasoning capabilities. We've seen extremely quick adoption of the enterprise now. Coding has been one area people talk a lot about.

But I think what we're hearing again and again in all these different ways is that these tools have gone from things that were fun and curious to truly useful. Work, people's personal lives. And the fact that you can now ask a system like Codex to go off and do some work for you autonomously over minutes or hours, it's pretty remarkable.

LISA SU: Yeah. I think the key point that you said is really enterprises are seeing lots and lots of value. I think the other thing that's been amazing is, man, the rate and pace of what you guys are putting out. It seems like every week you have a new model. Workloads are just changing so fast. What are you seeing? How are things changing? And most importantly for us, how are you seeing compute demands changing?

SAM ALTMAN: Tons of changes all the time. But one of the biggest differences has we've moved to these reasoning models. Sc we have these very long rollouts where a model will go off and think about a problem and come back with a better answer, or in some cases, like a whole PR ready to go. But this has really put pressure on model efficiency and long context rollouts. We need tons of compute, tons of memory, tons of CPUs as well.

LISA SU: I've seen that, actually.

SAM ALTMAN: And our infrastructure ramp over the last year, and what we're looking at over the next year has just been a crazy, crazy thing to watch.

LISA SU: Is there ever enough GPUs?

SAM ALTMAN: Theoretically, at some point. You can see that a significant fraction of the power on Earth should be spent running AI computes. And maybe we're going to get there.

LISA SU: Yes. Yes. That's definitely true. Look, we have been honored. We've really, really appreciated the partnership and collaboration between OpenAI and AMD over the last few years. Working together in Azure, working on some of your research stuff, and particularly the deep design on MI450. You guys were really early in just some of the important insights. Can you just tell us a little bit about how that's evolved and how we can do more for you?

SAM ALTMAN: It's been amazing working with you all, obviously. We're already running some work on the 300X. But the MI450 series, and the work we've been able to do there is, you've worked on that over the last couple of years, and we're very grateful for you listening to our input. Hopefully it will be a good representative for what the industry as a whole needs. But we are extremely excited for the MI450.

The memory architecture is great for inference. We believe it can be incredible option for training as well. And when you first started telling me what you were thinking about for the specs, I was like, there's no way. That just sounds totally crazy. That's too big. But it's really been so exciting to see you all get close to delivery on this. And I think it's going to be an amazing thing.

LISA SU: First of all, thank you for saying that. I appreciate that very much. One of the things that really sticks in my mind is when we sat down with your engineers, they were like, whatever you do, just give us lots and lots of flexibility because things change so much. And really, that framework of working together has been phenomenal.

Sam, look, this is a moment here where we have lots of folks in AI wanting to know what's next. So help us with big picture. What do you see in the future? Perspective on where things go. How do the workloads evolve? What happens with, quote unquote, "AGI?" And really, how do we as AMD and we as the computing industry help enable all of that for you?

SAM ALTMAN: At the beginning of the 2020s, we didn't kind of have AI as we think of it today. We had a bunch of other systems, but that was still the pre-GPT-3 era just by a little bit. Now, as we sit at this sort of halfway mark through the decade, it's really been remarkable progress from not even a GPT-3 model to a GPT-4.5 and all of these models that really feel smart and helpful and can give these real utility experiences where people would look at this if they could go back in time and say, that feels almost impossible.

Like if you went back to 2020 said, by halfway through the decade, we're going to be at this system that you can talk to and it's really smart. It's like a smart person that can do work for you. I think we're going to maintain the same rate of progress, rate of improvement in these models for the second half of the decade as we did for the first. I wasn't so sure about that a couple of years ago. There were new research things to figure out. But now it looks like we'll be able to deliver on that.

So if you think forward to 2030 and the systems that we can have, these systems will be capable of remarkable new stuff. Novel scientific discovery, running extremely complex functions throughout society, and things that we just couldn't even imagine as possible before. To get there, to be able to deliver on this, it's really going to take, these are huge systems now, very complex engineering projects, very complex research. And to keep on this curve of scaling, we've got to work together across research, engineering, hardware, how we're going to deliver these systems and products. And this has gotten quite complex.

But if we can do on that, if we can deliver on that, if we can drive this collaboration across the whole industry, we will keep this curve going. And so we're tremendously excited about the work that we're doing with AMD and what you all are going to deliver. We'll keep delivering great models.

LISA SU: Sam, I can say that we really, really appreciate the work with OpenAI. You guys push us. You guys push us hard. But at the end of the day, we all want to deliver that vision. So thank you so much for being here.

SAM ALTMAN: Thank you very much for having me. And thank you for the partnership too. See you. Thank you.

[CHEERS, APPLAUSE]

LISA SU: All right. So as you can tell, we are super excited about what MI400 brings to the market. There are lots of active customer engagements already. This is about really co-optimizing together. But it really doesn't stop there. We're already deep into development of our 2027 rack that will push the envelope even further on performance, efficiency, and scalability with our next generation Verano CPUs and MI500 GPUs. So lots and lots of stuff to come from AMD.

Now, that brings us to the close. It's truly been an amazing day. We've covered a lot from the launch of MI350 series, to our next generation MI400, to the Helios rack scale solutions, to all of the incredible momentum that we have building our open software and hardware ecosystems. And I really want to say a big thank you to all of our partners who joined us today on stage. There are a number of partners who have helped us with putting together this event. There are a number of breakout sessions I hope you guys get to later on in the day. And hopefully, what you've gotten from today is that we're moving faster than ever before to deliver the best AI solutions for the market.

But let me just end with a few personal thoughts. When I think about this past year, it's really redefined what progress in AI looks like. It's really moved at a pace unlike anything that we have seen in modern computing. Frankly, anything that we've seen in our careers. And frankly, anything that we've seen in our lifetime. We, in this community, I call this community the AI ecosystem, we're really at the center of everything that matters. And isn't that just an incredibly phenomenal place for us to be?

I think of it as a journey, I've always said. This would be a journey, and I'm incredibly proud of how far we've come. But more than that, I'm actually really proud of how we're bringing together the technology, the talent, and the partners needed to make AI more powerful, more accessible, and more useful for everyone.

The future of AI is not going to be built by any one company or in a closed ecosystem. It's going to be shaped by open collaboration across the industry. It's going to be shaped because everyone is bringing their best ideas. And it's going to be shaped because we are innovating together. So on behalf of all of us at AMD, we look forward to changing the world with you together. Thank you for joining us today.

[CHEERS, APPLAUSE]

[LIVELY MUSIC]