ROSS SEYMORE: All right, everybody. Good afternoon. We'll get started with the next fireside chat. Again, I'm Ross Seymore, cover semiconductors here at Deutsche Bank. We're very pleased to have Mark Papermaster, the CTO of AMD, on stage with us today. So, Mark, thank you very much for joining here at the DB Tech Conference.

So I think it was just what? Last week, you guys announced an AI-related acquisition. So AI is obviously a topic we'll talk about in many different ways, but the ZT systems acquisition. Talk a little bit about the logic of that deal. And it seems like there's some nuance to what you were really going after in that with the engineers. So just talk a little bit about that.

MARK PAPERMASTER: Absolutely. First, Ross, thanks for having me here at your conference. Excited to be here. And an incredibly exciting time at AMD. When you think about the announcement that we made last week in acquiring ZT systems, it's the next stage in really our strategic growth of ensuring that we have the full complement of skills that we need to have, not only the best AI hardware, not only competitive and leadership AI software, but the ability to integrate it and optimize it at the system level.

And so ZT systems represents exactly that. Their 15 years of experience with building some of the most complex, heterogeneous system designs, integrating CPUs, GPUs, networking, advanced thermal management and cooling capabilities, as well as the kind of control software that you need to efficiently run these complex rack designs. So a very, very strategic addition for us. And not a typical acquisition, because the goal for AMD is indeed to leverage the 1,000 plus skills that have that design expertise.

And what we did, Ross was stated right up front that we will, for the manufacturing side, Frank the CEO of ZT systems will continue to run that 1,500 person manufacturing side. But that we will be looking for strategic partners over time, of course, once we close. So the strategic impact for AMD is really one of accelerating our time to market and optimizing that time, that solution, which is so critical going forward to have the hardware, software, and the system design co-optimized together.

ROSS SEYMORE: And is this something that will have an immediate benefit because the system design side of things can pull in the existing AMD silicon, mainly on the MI side of things I would assume, or is it something that you'll have to wait a little bit longer for the benefits to accrue because it will be dependent upon integrating the roadmap from the silicon perspective as well?

MARK PAPERMASTER: The full benefit, of course, would be upon close. We anticipate close by mid 2025. This requires a US and EU approval because that's where ZT does business. And when you think about between now and then, we've already been working with ZT systems and other ODM and OEM in terms of that system design. So we already know that we have a very good working relationship. Of course, it needs to be still a separate companies. But we expect to continue that deep dialogue and partnership as we have. And then upon close, you'll see a much deeper integration. You'll see a true co-design, from silicon to systems.

ROSS SEYMORE: And talk a little bit about the rising importance of systems. How did that come to be? And how does AMD expect to differentiate in its approach?

MARK PAPERMASTER: When you look at AI applications, they're incredibly demanding of high performance compute. And so they demand the absolute best of silicon capability. Well, in AMD, we're well armed there. We are the leaders in chiplet technology. We're the leaders in bringing together high performance CPUs, GPUs, accelerators. And we have a strong network portfolio based on our acquisitions with Xilinx and Pensando. We've been building our software skills. And so that's critically important to actually bring that computation to bear.

But again, when I go back to the demands of the most complex AI models, the foundational models, which really require massive clusters of heterogeneous systems, the optimization doesn't stop at that hardware and software of the compute engines. You need to optimize through the networking, through the connectivity, through the rack. And anything can become a bottleneck. If you have issues from wiring capabilities, if you have issues from a cooling standpoint, if you didn't fully integrate the networking in the most optimum way, then you are not delivering the most optimized system.

So we, of course, partner with our system providers today. It's a serial process. We're designing that compute complex. And then we work with our system providers and continue that optimization. So that part doesn't change. But with the ZT design capabilities coming in-house from the earliest conception of our next-generation AI-based processing complexes, we will take into account the system aspects. We'll be actually optimizing for all those facets that I just described. And that is a game changer, earlier time-to-market of an optimized system, all the way through the rack implementation.

ROSS SEYMORE: Does it change at all your relationship with other partners from a system perspective?

MARK PAPERMASTER: That's what's critical about our strategy and why, as I described it at the outset, that we are partitioning off the manufacturing capability. It is not our intent to compete with any of our partners. What we want to do, though, is offer a significantly strengthened starting point, a reference design, where we have taken into account facets of the design, as I said, all the way through that system optimization. The 1,000 skills that come in from ZT that have been doing this for many, many years will allow us to actually partner better with OEMs, with ODMs, with our end customers, because as we understand those requirements, now the starting point will be optimized from the get-go.

ROSS SEYMORE: And is this something that is more cloud attuned than enterprise adopters of AI from your perspective? And I know cloud is the most aggressive adopter of it. Does it really matter. The type of customer?

MARK PAPERMASTER: Well, you said it right. The cloud, which are driving the foundational model training and inference are the most demanding of that system and rack level optimization. And so they're the first customers that will benefit from this marriage of technologies that we will achieve, upon close, with the addition of ZT systems.

But what happens in this, I will give you an analogy, to server systems today. So often the biggest server clusters are at the hyperscalers. But that learning, and that optimization comes right down into the on-prem and enterprise installations, which are deployed across the Fortune 500. You'll see a very similar trend here.

The utmost of scale up, meaning the largest of the GPU, compute nodes of the scale out, meaning the building of the largest clusters will be hyperscale. But all of that technology honing can be applied to ensure that enterprise class implementations may not be running those massive foundational models. They're going to benefit equally from scaled back systems in tune with their workload demands.

ROSS SEYMORE: And with the benefits on this, I assume everybody is going to think the MI family would benefit from this. But would it also be beneficial-- would it be a positive also for the EPYC side of things?

MARK PAPERMASTER: There is no doubt about the benefit for both our EPYC and our Instinct lines. First of all, these AI clusters are CPU and GPU based. They are indeed heterogeneous. So that integration of CPU and GPU into that rack of design will be an immediate beneficiary of the ZT system skill inclusion into the AMD family. But beyond that, this is 1,000 engineers extremely experienced in system optimization. So that will be applied to standalone GPU optimization and x86 EPYC server systems.

ROSS SEYMORE: So that is enough, at least, unless you want to talk about it more, on the ZT side of things. Let's talk about the Instinct side. The MI300, you guys have ramped that amazingly fast. People are debating what the number ends up being this year. But going from basically zero to somewhere around $5 billion over the course of a year or so is pretty impressive. Talk about what the biggest challenge is in the next generations of product. I know the race never ends. But as you go to the next gen of that, the 325, et cetera, et cetera, where does AMD start to differentiate or continue to differentiate in that roadmap?

MARK PAPERMASTER: Well, let's start with the differentiation that we have right now. How did we get that differentiation? We listen to our customers. That's a hallmark of AMD. We're able-- we are able to collaborate extremely well, listen to our customers. When they saw what we had done with the MI250, which powers the largest supercomputer in the world today with the Department of Energy Frontier system at Oak Ridge, and Lumi, and other installations, we earned the trust, and understood their requirement, and were able to show our agility, and pivot the next-generation design that we have for HPC, for supercomputers, called MI300A over to MI300X, AI optimized for the most demanding inference and training applications.

And in fact, for inferencing, we've been differentiated because we have up to 192MB of HBM3 memory on that starting point. And so in fact, by the way, I'll mention that we did publish MLPerf results. And you'll see today, this came out. And you'll see the competitive stance of MI300 on that standard benchmark.

So where do we go in our roadmap going forward? What's our strategy? One, we've embraced an annual cadence. If you think about data centers, typically, they've been on an 18 to 24 month type of cadence. But AI systems are on an annual cadence. And the reason is that the innovation on the models and the model size, the number of parameters of the models, and the capability of what those foundational AI models can do is dependent on the growth of the compute engines to support that expanded parameter growth. It has to be more efficient, or energy alone just to power would gate the advancement of AI and the industry.

And so it is a necessity to have a much more rapid cadence. And we embrace that. And so what you should expect from AMD is we're leveraging all of the know-how that we've had for years. Supplying the industry leading edge, high performance computing, being the leader in chiplet, we are increasingly going to leverage that to hit that annual cycle. We have the MI325 enhancing to HBM3E, which we have coming to production next quarter. We're on track with the MI350.

And both the MI325 and the MI350 family will stay in that same socket that we have, that same base board, the UBB, so it's an easy adoption cycle for our customers. We're not demanding a new system infrastructure. But we're adding a doubling of memory capacity with 325. So that differentiation that we have on inferencing, we expect to persist with the MI325. And then with the MI350, again, dropping right into that same system infrastructure with ease of adoption.

We're going to have the next generation of our GPU compute engine. And that's going to have new math formats. That's going to support FP4 and FP6. It's going to yet leverage not only the same socket, but that same HBM3E memory. So it's a plan of how do you get customers a significant increase in capability every generation, but work closely with them, and make it easy for them to consume these new capabilities?

And then beyond that, with the MI400 family, which will be for the year following, 2026 really, really exciting advancements across the board. Again, great partnership with our customers, and even more focus on the networking capabilities and system optimization. Because by the time we come out with the MI400 family, assuming successful close of ZT system acquisitions, we'll have thumbprints all over that in terms of even further system optimization.

ROSS SEYMORE: So you and your primary competitor in this space have talked about that annual cadence. And demand is off the charts. That's great. But can the customers truly implement these solutions at the speed with which you are introducing them? Do they have a diversity of workloads, so, you know, they're not going to go whole hog with the 300 and no 325, and skip to the 350? How do we match the sheer dollars of CapEx, and the time it takes to install these systems relative to the speed that both you and NVIDIA are trying to move?

MARK PAPERMASTER: What's key in terms of adoption is making sure that there's a consistent software and software application story as you move generation to generation. When you look at MI300X, we launched in December of last year. And that was with our software base of ROCm 6.0. So we brought ROCm to production level of 6.0. And we're in production. So look at Azure as our announcements that GPT-4 or GPT-4o. We look at Llama 3.1 coming out on day one on AMD. Meta, you saw on stage with us as we announced 300. And they've been such a great partner.

It's the learning from now real-world production applications running on MI300 that we fold into a ROCm 6.1, 6.2. So we've tuned the performance. We've increased the ease of adoption. We've taken important features like the transformer optimizations with flash attention too, built in now to ROCm. And so it becomes-- rather than just hardware generation, it becomes a fluid flow, Ross, in terms of making sure that the software enhancement learning that we do is a consistent and seamless flow generation to generation.

And then with the hardware capabilities, what we do is increase the performance, and most importantly, the performance per Watt of energy. Again, this is an energy-gated sector. And so we really are very keen on the productivity that our roadmap will bring in every generation yet on the base of a consistent software platform.

ROSS SEYMORE: So to today, I believe your solutions largely have a leadership position on the inference side, a little less so on the training side. Talk about what gives that advantage or disadvantage. And perhaps more importantly, looking forward, what side are you actually focusing more on between those two? It might not be an either/or.

MARK PAPERMASTER: Well, certainly not an either/or. We've targeted the Instinct family squarely at the most demanding applications of both inferencing and training in the AI space. We did lean in, of course, upfront on inferencing with MI300. And the reason is we were the new entrant to actually bring competition to this market. And it was important that we establish ourselves where we had a clear differentiation.

Our chiplet capabilities and prowess, and the ability to have a leadership memory capacity as well as bandwidth was where we wanted to lead in the market. And it's proven out to be very beneficial. It's providing immediate total cost of ownership advantages to our customers. And it's a commitment that we'll maintain throughout our roadmap. So as I said a moment ago, you'll see us very focused on, and I'll say, staying on the torrid pace that our customers demand.

But training, where we are certainly competitive from the outset with MI300, training, no doubt is a harder task, because you need to take additional time, additional time to hone the networking capability, the scale out capability as you build clusters. You need to spend the time on the communications library, the software which is managing how the GPUs are working close knit across the build out of that cluster.

And that's exactly what we've done. So we are working with initial implementations on training. We're tuning the training. So we're not waiting for the future roadmap of Instinct. We are starting now to bring training to bear as well as inference, which of course, is the production instances that we have out there today.

ROSS SEYMORE: The last hardware-related question on this kind of AI infrastructure for you, ASICs versus the traditional, or well, GPU, traditional, otherwise. One, do you see ASICs as a threat, or is that an opportunity? You do ASICs in another part of your business. Is there any reason why you couldn't at some point choose to do it within your data center infrastructure business as well?

MARK PAPERMASTER: Well, if it's a silicon-based solution that's in need of high performance computing, we're going to view all of that as an opportunity. Because across our general offerings of CPU, GPU-accelerated computing, we have in addition to that, of course, a semi-custom division. And we've been driving our semi-custom division to be more and more efficient to get our cost structure very close to ASIC like. And so we do view that as an opportunity.

But moreover, I think the broader point is the size of the market. When you look at a TAM that we projected being $400 billion by 2027, what you're going to see is there's a need for a range of solutions. It is not one type of computing solution. Of course, the CPU and GPU combination is the most flexible. You have a programming environments out there that are very, very well established. As the models change, they can adapt immediately to new algorithms. And so that will persist.

But you will see, whether with startups or with hyperscalers, they're investing in their own custom silicon teams. Where you have an algorithm that's stabilized, and where you can come at a unique angle or tailor on a piece of market, that's very much going to be a part of that overall $400 billion TAM. And they're going to coexist together. It's going to be a key part of driving more efficiency. Efficiency of compute is the name of the game in AI. And we welcome that as both competitive pressure, to make sure our general offerings stay in the utmost of their competitiveness, as well as to participate in to our semi-custom division.

ROSS SEYMORE: So I actually lied. One more question on the hardware side is do FPGAs play any role in this, on the data center infrastructure side?

MARK PAPERMASTER: They do. Our FPGAs are adaptive compute, as we call it, are actually one of the fastest ways to tailor and to customize based on a known workload that our customers have. So we already have, and are working with, major customers in terms of hyperscaler deployments, which can allow tailoring in terms of workload flow, network management, and actually storage optimizations.

ROSS SEYMORE: So why don't we switch gears away from Instinct and over to EPYC for a bit? AMD has done an amazing job of targeting that market that forever was dominated by one of your competitors, and taking significant share, and, at least by my math, potentially a majority share in some of the markets. Talk a little bit about the GPU versus CPU crowding out dynamic. I know that's more of the business side than the technology side. But do you see that as something that we're kind of towards the end of it, and both sides are going to start to be appreciated a little bit more going forward? Or do you think more or less people are going to keep focusing on your Instinct side and not your EPYC side for some time to come?

MARK PAPERMASTER: Well, it's a great question. I think the question is born out of some of the dynamic that we saw over, I'll say the backward-looking 12 months. And there was a huge rush to strengthen the GPU infrastructure. Clearly, hyperscalers are making a huge investment, given the jump in model size. But you also saw a number of enterprises wanting to make sure they didn't get caught flat footed, and didn't have an AI compute capability. So there was some impact in terms of a GPU demand, taking, I'll say, a bigger share of the CapEx allocation on a year-to-year basis.

But I think that's normalized now. You look at going forward, guess what? The applications that always ran on a general purpose CPU still want to run on a general purpose CPU. You need to close the books every quarter. You need to have your ledger analysis. You have your customer relationship management. You may have an AI augment on top of that, but you're still running your CIM as a general purpose x86 compute. And likewise, many other workloads that don't go away, but add an AI augment.

And so what that means is where our customers that have a legacy x86 install base, they're finding that they're at the end of their depreciation cycle. And now when we show them what we can do with our 4th Gen EPYC, our Genoa that we've been shipping over the last year, our 5th Gen EPYC that is already in production sampling, and goes into full announce and release mode in Q4, what we're showing is that we can save dramatic, 50%, 60% of TCO cost savings. You can take that legacy data center and shrink the footprint, shrink the power, shrink the floor footprint, and make room for your next-generation workloads, whether they be CPU or CPU and GPU-based.

So I think I do see more of the last 12 months being more of that debate of CPU versus GPU. But I think we're normalizing. And frankly, the growth of AI and accelerated compute with GPU is also going to drive yet more demand on those general purpose x86 workloads. And we're very excited about the advantages we offer our customers with our EPYC product line.

ROSS SEYMORE: So the sticking with the EPYC side of things, talk about the architectural choices that your customers want you to make as far as the core count side. So kind of Bergamo versus Genoa, how do you see the market moving? And from a business perspective, we're moving towards pricing per core count, versus pricing per chip per se. I know they end up the same place. But just talk about the reasons, economically, why that can be a good thing for AMD.

MARK PAPERMASTER: Well, there's been a very significant shift in the driver of how you configure your x86 server systems. VMware did shift their pricing schemes. And it turned out that having a more performant CPU and the efficiency, energy efficiency that you get with a more performant CPU, and now the licensing costs that you get, lower licensing cost, if you have a more performant CPU, are in fact really a main major driver of our Genoa sales for those that are leveraging that type of virtualization capability.

So it moreover, underscores how demand is driven by workloads, just like data centers today are heterogeneous across CPU and GPU. When you look at CPU alone, it's not one type of server that makes sense for most customers. Again, for those virtualized environments, you want-- a Genoa's perfect with that high performance, 96-core CPU that gives you that TCO advantage.

But you mentioned Bergamo. It's really our native cloud-optimized workloads, where you have cloud native and don't need necessarily the highest performance CPU, but you want the most efficient. And so we go from 96 to 128 cores, very, very efficient. And so throughput systems, things like recommendation engines, those are workloads that really excel.

Beyond that, we've further diversified. We have stacked v-cache for where you might have a high performance or database workloads, where we really increase the cache size over the CPU. And we have telco-optimized. So again, the diversification of our EPYC CPU product line is really a result of listening to our customers and making sure that we have the diversification to match where their workloads are.

ROSS SEYMORE: Have you seen any change in the competitive intensity? Your roadmap seems to be very aggressive. And you've executed superbly on it. Whether it be from the x86 side, is that gap closing at all? Or more recently with ARM-based solutions, do you see that as a meaningful threat, or? I know the answer is it's always-- the world's always competitive. We always have to run fast, et cetera. But any meaningful change you've seen?

MARK PAPERMASTER: Well, your last comment is true. It is always competitive. And our view is actually very simple. The best defense is a super strong offense. So, I mean, our strategy is don't slow down. Don't slow down on the innovation. Don't slow down on our pace. Don't slow down on aggressive use of semiconductor leading edge technology, but doing it in a chipless-based approach, where we bring the leading edge where you need it, and not where you don't. And so that's what drives us, as well as the diversification to the customer workloads, making sure that we truly deliver TCO advantage based on the workload. And that is what positions us best to rising competition.

Of course, there are new entrants. You see ARM-based designs that are tailored for specific workloads. Had we not diversified our portfolio, we would be at a competitive disadvantages in that stance. So bottom line, we expect nothing but strong competition. This is what our DNA at AMD is all about. We thrive on that competitive spirit. And it will drive us to get the absolute best systems out for our customers.

ROSS SEYMORE: So just to front-run the AI PC topic and get the competition side out of the equation in both ways. Is there something inherent in the ARM architecture you believe, versus the x86 architecture, that is an advantage, whether it's on power consumption, performance, whatnot, that changes the game versus an x86 to x86 competitive environment that we've seen more predominantly over the last, say, 20 years?

MARK PAPERMASTER: We don't see that the instruction set architecture is a differentiator at the end of the day. I mean, we love x86 because there's a massive install base. It's a CPU architecture that we've been able to hone micro-architectural implementations over decades to eke every bit of performance out of it. And the software tool chain that is out there, and the familiarity of applications development make it a superb play.

Yet we offer ARM across our portfolio. It's in our adaptive embedded compute. It's in various controllers we have. At the end of the day is we're about providing the best solutions for our customers. You look at Zen 5 that we're rolling out now, we just brought it out in our PC market with a Strix Point. What did we do with Zen 5? We widened the execution pipeline. We added AVX-512 full physical implementation. And you might think, well, that's for more like server, and that's for those vector calculations and floating point calculations.

But what we have is a capability to also excel with AI, because the core of the AI is in fact a multiply accumulate, which we perform very, very well. And we added FP16 in Zen 5. So we'll continue to evolve the x86 architecture to excel at the workloads needed across PC, embedded, and our data center customers.

And we also, where an Arm ecosystem, not the ARM ISA, there are differences there, but more it's the ARM ecosystem, where does it play, and where that's an advantage. Like our embedded roadmap will persist and continue to partner closely with Arm and those markets. And the PC, we'll see. Now, it's a dual-- it's a dual ISA play. So we're going to leverage x86 where that's best. But again, we are not married to ISO. We're married to the best experience we can provide our customers.

ROSS SEYMORE: How does AI change the roadmap of your PC CPUs?

MARK PAPERMASTER: AI changes dramatically in terms of the horsepower needed to operate AI workloads in a incredibly restricted power environment. So if you think of what does an PC do, it's allowing you to have, at your fingertips on the device you're using for the bulk of your content creation, the ability that to run incredibly efficiently AI applications, yet you won't accept if your battery life degrades significantly. So the demand upon us, and our competitors, frankly, for the AI PC, was how to provide AI acceleration without impacting that experience.

And this will all come to light as we close 2024. But as we ramp into 2025, and you start seeing the amazing applications that are going to be available to you in the PC space. So Copilot plus, which we now support and run on our AMD systems, will bring a plethora of new capabilities. In two or three years, you won't remember what it was like when you didn't have those AI tools to assist you in all of your day-to-day work tasks and your personal life.

And so it's added a whole new tier because we now have high-end PCs, which have significant investment in the AI capability. And AMD in our new Strix Point, we added 50 TOPS of AI acceleration on top of the Zen 5 CPU, and on top of a Radeon 3.5 GPU that was optimized based on the partnership we had with Samsung on that GPU technology that went in Galaxy. So an incredibly honed combination of IPs to deliver-- honed combination of IP to deliver a superior AI PC experience.

ROSS SEYMORE: And do you think there's any, from a CTO perspective, any kind of cost tradeoffs that you have to hit? You just said you're going to add, on top of everything, you're going to add all this functionality. Is the price of the end device just going to go up, and it's going to be a higher end, establish a new higher end PC? Or when you're thinking about the roadmap, does cost come into the equation as well?

MARK PAPERMASTER: Well, at the outset, it will be a higher price tier for the highest capability of AI PC. So where our Strix Point with 50 TOP capability is positioned in the higher tier from an ASP Strata. But it won't take long before the AI enablement is really end-to-end across the PC portfolio, across the industry. That's how quickly, I believe, the adoption will be of these AI-enabled PC-based applications.

ROSS SEYMORE: All right. Well, Mark, we are officially out of time. Thank you so much for all your insights, and joining us here at the conference. I appreciate it.

MARK PAPERMASTER: Ross, thanks for having me.

[APPLAUSE]