UBS Annual Technology Conference

TIMOTHY ARCURI: Good afternoon. I'm Tim Arcuri. I'm the semiconductor analyst here at UBS, and very pleased to have AMD with us and very pleased to have Forrest Norrod who is the EVP and GM of the data center group at AMD. So thank you, Forrest.

FORREST NORROD: Thanks. Good to be here.

TIMOTHY ARCURI: So Forrest, I wanted to start, you've got the broadest portfolio of data center silicon. You've got some new networking technology. You have CPU, which feeds GPU. Maybe you can start by broadly talking about your strategy around your entire data center portfolio.

FORREST NORROD: Sure. Yeah, we've been embarked over the last decade. I joined AMD, almost exactly 10 years ago now, to build out a portfolio of data center technology to really eventually take leadership in providing silicon solutions for data centers. We began, of course, with the CPU. That's where we had some heritage. We began with the epic line of the CPU and that's done very well. We're about 34% share right now.

And then we built out the GPU, and then we acquired a networking team and began building that out. And the purpose of having all of those pieces is in the end, we do believe that to really optimize the solutions, particularly in the AI era, that having all of the elements in-house is very helpful.

And so our intent is to build out reference solutions for AI systems that incorporate world-class CPU, world-class GPU, world-class networking. And the sum of the parts is greater-- the whole is greater than the sum of the parts. And we believe that we were well on the way to offering the best possible solutions. Easy to adopt for customers to adopt, and that they can source through a variety of partners.

TIMOTHY ARCURI: Can we first talk about CPU. You've been gaining share on the server side fairly consistently for years now. Your share in cloud is in the 50% range, much lower in enterprise. But you're beginning to see some success in enterprise. So can you talk about your roadmap and why you're winning on the CPU side? And if you can particularly talk about enterprise, and what you've done differently there to catalyze some recent wins there.

FORREST NORROD: Sure. And as I mentioned, we're up to I think the most recent quarter, just around 34% revenue share. And that is over 50% in cloud. And it's in the 20-ish percent for enterprise. The reason we've had this success is that we have generally focused on producing higher performance parts with better power efficiency than our competitors.

And with the cloud guys in particular, the data center is their factory. And so as they're making their decisions on what to buy, they're very focused on what's driving superior TCO, what drives the most efficiency in their data center. And so we've had relatively rapid growth there.

On the enterprise side, the equation is slightly different. There are certainly some high value workloads for which that performance and TCO benefit is critically important. But the CIO also has to balance perceived risk. And they're much more concerned about getting taken to task for a problem than getting kudos for having a slightly more efficient data center. And so there, it's been about retiring the perceived risk.

And so building out confidence in the ecosystem and our customers that AMD has not just a superior your solution. I think most folks are convinced of that, but it's low risk, it's easy to adopt. And so our focus there has been over the last few years, really working with the other partners in the ecosystem to ensure that we've got qualified solutions, that we're demonstrating case studies, that we're really retiring that perceived risk.

I think that the inflection point really has come this year, where people have enough familiarity with our solutions. We've done enough POCs that they're recognizing, hey, this is easy. There's really not a port. It's not a situation of I need to port my application.

In fact, the funny thing is the instruction set that both Intel and AMD execute is actually the 64-bit instruction set that AMD created. So the application's actually written to conform to the AMD instruction set architecture. And so it's very easy to port. And I think that that realization has really metastasized now in the market. It coupled with some of the recent concerns around Intel and the press about them. I think that's also providing an impetus for folks to consider AMD.

TIMOTHY ARCURI: Got it. I wanted to ask you about Turin and what Turin brings to the table. And how you see it stacking up with the Air Force for E-Core and for Granite Rapids. How does it stack up? And what's the early customer reaction to Turin?

FORREST NORROD: Yeah, Turin continues our strategy of providing leadership CPUs, both performance as well as power performance. And I think the third-party measurement so far, there's very limited Granite Rapids out there. But there's been some third-party benchmarking that's showing, like Phoronix showed across a very wide suite of 200 different benchmarks.

Turin is 40% faster than the top end of the Granite Rapids stack. And so I think that sort of continues our clear performance leadership with Turin. Again, Sierra Forest is no comparison. I mean, Sierra Forest, it's a great part. Don't get me wrong. But with the E-Core, it's much, much lower performance, both on a per thread as well as an overall socket level. And so it's really not-- it's not competitive at all.

We see great success with Turin, both at the 128 as well as the 192 cores-- the great success there. And I think that's going to continue. I think we've got another generation of absolute double digit performance leadership across every application.

TIMOTHY ARCURI: There seems to be some perception that your success on the CPU side is almost all because of process. And I think it goes way beyond just process. And you've always said that you assume that the process gap will close. That is your base assumption. Can you just double click on how much of it does relate to process?

FORREST NORROD: Certainly we've got a great relationship with TSMC, and access to their process is helpful, no question about it. And more and more, we're doing things to optimize the TSMC process and our circuit designs concurrently. We were the first, I think, as well, to shift production level silicon in their most recent process.

But it goes way beyond that. I mean, our designs I think are exceptionally good, are embrace of advanced packaging, chiplet technology, early use, of course. I mean, I think we've been at the forefront of things beyond the process that are going to be increasingly important to developing high performance chips going forward.

So we were the first to embrace chiplets. We're the first to embrace 3D stacking. We're the first to incorporate HBM in a major way in some of our data center parts. And by the way, on the Xilinx side as well, there's a rich heritage there on that technology.

And I think that's served us in very good stead in continuing to push the performance. And we have a very mature process around retiring risk on new technology development. And I think that's super important because that means that our roadmaps have been very predictable.

I think the other thing we pride ourselves on is, we put out the roadmap. We say what we're going to do. And then I hope you'll give us credit. We've done exactly that for generation after generation, definitely.

TIMOTHY ARCURI: Let me shift over to GPU. You're shipping MI300. You're just starting to ship MI325X, really more in Q1 than in this quarter. So your roadmap is really beginning to evolve. You recently extended the TAM. You took it from 427 to 528, the accelerated TAM. There's a bunch of questions around that. And really they are, how do you see your position in that TAM? And what would define success.

I've heard Lisa in the past say, well, look at our share in GPU. We're not saying that that's what the goal is, but that we're in the market to be a major player. So what would define success?

FORREST NORROD: Well, I think I'm going to keep myself out of trouble by agreeing with her. I would anyway. Look, I think we certainly aspire to be relevant to the market and to be relevant to the market, to be important to the market. I think you have to be strong double digit percentages, some people would say 20%.

I think over time that we're looking to-- and I'm not giving you an exact number-- but we're looking to make sure that we're relevant, we're important to the ecosystem.

TIMOTHY ARCURI: I mean, even today, if you look at the inference TAM, you're probably pretty close to 10% share this year, just of the inference TAM. So you're not-- sure, you're not saying that you have caught up to NVIDIA, but you're give or take 10% share of the inference TAM, which is not nothing in a very, very short amount of time.

FORREST NORROD: Yeah, I mean, we've been super pleased with the progress on MI300. I think we've said a couple of times it's the fastest growing product we've ever had. So from essentially $0 to $5 billion in a year has been great. That's certainly a waypoint, not a destination.

And it does offer demonstrably superior performance in TCO, particularly for inference. And we think it's gotten competitive on training as well. Our intent is to continue pushing that very heavily with the roadmap. So you mentioned the 325 beginning to ship this quarter, and then 355 in the second half of next year. Both of those parts are designed, just as we did on the CPU side, to be iteratively, more performant, more competitive versus the competition, more differentiated.

And our intent is to continue leadership in inference across the board, which we think is more and more important as chain of thought models and similar approach to inference TAM approaches become more important. And to continue developing ourselves on training as well.

Long-term, we don't aspire to be an inference solution. We aspire to be a provider of great AI solutions on the training side, on the inference side, on any new evolving models as well.

TIMOTHY ARCURI: So if I said that in the few years-- if you achieve 20%, is that a failure or is that a success?

FORREST NORROD: Well, look, again, I don't want to-- I don't want to validate a specific shared goal. Again, we aspire to be relevant. We aspire to be important in this market. And I think we certainly are looking in-- any similar number in such a large market is a huge revenue opportunity for us.

TIMOTHY ARCURI: Yes, for sure. What have you learned from the experience so far with MI300. You've had a lot of success with a few large customers. You've been very public about Microsoft, has been at some of your events, spoken. What have you learned that you take forward, and especially as you go to MI400 in 2026.

FORREST NORROD: Well, look, we're the challenger in this market. Even though it's a new market, obviously NVIDIA, great company, great products. And they are the de facto standard. And so as the challenger, you have to have a strategy that minimizes the friction of adoption and minimizes the effort that customers need to go through to adopt you because they're already using the alternative.

And then you also have to have differentiation. You have to give them a reason to adopt you, notwithstanding the fact that nobody really wants a market in which there's only one player. Nevertheless tragedy of the commons, if you don't give a customer a specific reason to adopt you, they won't, no matter how much they would theoretically want an alternative.

Our approach with the 300 has been to, first off, minimize the friction of adoption on the system side by making it very easy to drop MI300s into infrastructure that was, candidly, first conceived to host NVIDIA software solutions.

And then secondly, on the software side, our real thrust has been to minimize, first off, the time to functionality, so that if somebody's got a model, make sure they can immediately run it on MI. And by the time we got-- we're introducing the MI300, we had accomplished that goal.

Pretty much any model written for the standard framework, if you're running on NVIDIA, you could run it on MI300 out of the box day one and it would work. But it might not be performance. So if you roll back 12 months ago, that random model, you might get 130% of the performance of the NVIDIA solution. At the same time, you might get 50% of the performance. And that's friction. That's friction of adoption. If you don't know what you're going to get, that's a problem.

And so we've really been focused over the last year on maturing the software ecosystem. The math library is the frameworks, working with the guys that are developing the foundational models to make sure their models are AMD-aware and really working to minimize the friction of adoption, such that somebody can pick up just your random model and run it on an AMD solution today and you're going to get excellent performance.

And so those-- and you're going to get better performance on inference. You're going to be differentiated with our greater memory capacity, or greater memory bandwidth a few other things on caching. You're going to get a differentiated solution that gives you a reason for adopting.

As we've built out our roadmap going forward, we've tried to stay true to those learnings-- how do we minimize the friction of somebody adopting us for inference or training, and how do we make sure that we have, notwithstanding, you can't be too different-- how do we have some differentiation that provides that impetus the prize for adopting AMD.

TIMOTHY ARCURI: Can you talk about just the success you've had in software? One thing-- I mean, NVIDIA has done a fantastic job with CUDA and with all their transformer models. And you've been playing catch up and you made the acquisitions and you've come a long way. But where do you assess where you are today versus where you need to be? Do you have all the transformer models you need? Do you have the capabilities you need?

FORREST NORROD: Again, it's a journey, not a destination because even as we're developing NVIDIA's-- and again, great set of software resources, they're continuing to advance. But I think where we are today is that, as I mentioned a moment ago, for customers that already have a model or want to develop or fine-tune a model, we're at a point where it's very easy to do so.

We don't have the same level of vertical solutions that NVIDIA has for that. We're really focused on working with the other members of the ecosystem to develop those. But I think we're at the point where, again, it's relatively easy to adopt, relatively easy to move over to AMD, or to add AMD really more realistically. Add AMD as an alternative for your infrastructure, and then place the workloads between NVIDIA and AMD where you get most benefit.

TIMOTHY ARCURI: Can we talk about any supply constraints that you foresee in this coming year? You've gotten out, I think, pretty well in front of cores us. Of course, your allocation is not as much as NVIDIA's is, obviously, but you've done a pretty good job of getting out in front of that. Do you foresee memory being a potential bottleneck for you?

Every day, I'm sure I have five emails from somebody asking me about the challenges that one of the large companies is having and whether that's a problem for you or NVIDIA. So is that something that could constrain your growth, either of those two things or really any factor?

FORREST NORROD: Yeah, I think we're in pretty good shape. We've got an excellent supply chain team and excellent operations team. And I think we've got-- more importantly, we've got outstanding relationships with all of our partners in the ecosystem. And it's not-- look, it's not anybody's best interest-- that may be one company-- but it's not anybody else's best interest to have one customer dominating the consumption of any particular component, be it cost, be it memory, whatever.

And so we have outstanding support and outstanding partnerships really from all of our partners, be it substrates, be it wafers, be it memory. And I think we've done a lot to build and develop those relationships. But it's also very-- our interests are very congruent with their interests, which is customer diversity. Nobody wants to be locked down to one customer. It's way too dangerous.

And so I think that has been a great set of relationships. And as we look forward into '25 and beyond, we're very confident that we can get the support for the components that we need.

TIMOTHY ARCURI: And you feel like-- I'll just ask you directly-- so you feel like if the challenges from that main, or from that large supplier persist, that you still have other sources of supply?

FORREST NORROD: Oh, sure. We work on the memory side. For example, we work with all of the major vendors, and we've got great relationships with all of them.

TIMOTHY ARCURI: Great. Can we talk a little bit about beyond MI325 and talk about MI350. Some investors have wondered just how that product is positioned in the marketplace at that time. Will there be liquid cooled and air cooled SKUs, because the market will be significantly more shifted over to liquid cooling by then? And so how do you intersect your roadmap with the infrastructure, which is also evolving.

FORREST NORROD: Yeah. So look, we do see a rapid transition in the focus of particularly new data center builds, predominantly shifting more and more to liquid-cooled builds to support the extremely high densities that people want to get. And we're very comfortable in being able to support liquid-cooled environments, really, even from the 300, we've got solutions available today. Certainly that will be growing through the 325 and the 355.

But you'll also be able to put it into an air-cooled infrastructure as well, which I think is a real advantage for us because it gives the customers flexibility in their deployment scenarios. And we think 355 will compete very well in inference and training against the GB200 and BOM solutions. And we think we will have a broader support at fairly high densities for both liquid as well as air-cooled.

TIMOTHY ARCURI: What about from a Rack Scale point of view? Because your first product that's fully Rack Scales and not until MI400. So will that aspect hold you back? Will the market want Rack Scale so much in the back half of '25 that that's a factor that maybe could constrain 350? I'm just kind of wondering, that's like another--

FORREST NORROD: I don't think it's going to be a major issue. And that's certainly not what we're hearing from our customers that-- the question is really how well does it fit in the data centers? What's the level of performance they can get? And are they stranding anything? Are they stranding power cooling or anything else?

And I think we've been thoughtful about the design. And our customers are flexible enough in their data center deployments, so that's not an issue, that won't preclude us. We will have full Rack Scale infrastructures available, of course, in the 400 time frame. And that's the point where you're starting to get up into the 200-plus kilowatt per rack regime, in which case you really do need to have a complete Rack Scale architecture.

TIMOTHY ARCURI: Yes, totally. Just on that front, can we talk about the ZT deal? Give us an update there. And review the rationale. It's my thinking that by the time that it really impacts your roadmap, it's probably more impactful for MI400 than for MI355.

FORREST NORROD: Yeah, no, I think you've got the timing right. So first off, we were progressing very well. And we're still very confident of closing the transaction in the first half of next year. We've already gotten regulatory approval in the US and a number of other geos. We're waiting for a few, but that all looks like it's on track. And we're very optimistic that we'll close the deal in the first half of next year.

We have already started working through a set of contractual agreements, of course, that were two different companies, so we can't operate as one yet. But we can put in place strong contractual agreements that allow us to engage the ZT resources on the forward-looking products. And we have already done so on 355, on 400 series, and quite candidly beyond.

And so that has already started. You will see some contribution from the ZT system resources in the 350 series systems that our customers deploy. So I do think you'll start to see a little bit of contribution there, but certainly you'll see a major contribution from the ZT systems engineering teams on the 400 and beyond.

TIMOTHY ARCURI: Can we talk about the--

FORREST NORROD: I'm sorry. You asked me about the rationale. This is worth touching on as well. The rationale for buying ZT really is twofold. The first is that as we're designing for these 200-kilowatt-plus racks, you really do need to comprehend the requirements that the rack and cluster level as you're designing the silicon.

And so being able to do that design that system and cluster level design very early on allows you to define and design a better piece of silicon to fit into it. And so that's a big part of it.

And then the second part is, look, we want to support the ecosystem and adding value to our solutions. And so we don't want to take the approach that we have a one-size-fits-all. We're not trying to take the Henry Ford Model T approach of can have your hyperscale data center rack any way, any color you want, as long as it's this color black. We're not taking that approach.

So we're investing in enough systems engineering not only to produce a great set of base designs and elements, but also to allow others in the ecosystem to do variations to add their own value. That actually takes a little bit more engineering up front to add the hooks and design the components such that others can do that. But we think by doing so, we better harness the engineering talent really across the industry, to accrete value to our ecosystem.

TIMOTHY ARCURI: Great. Can we talk about the back-end part of the network? I think on a recent earnings call, you said this is 30% of the cycle time, back-end part of the network. And we all know how strong NVLink is to drive true cache coherency. You have UALink. But that's for a pretty small cluster. Can you talk about how you're positioned for these larger clusters as we get to a 100,000 and potentially a million cluster?

FORREST NORROD: Yeah, so our focus-- our focus has clearly been with, we think the preponderance of the industry on how do we evolve Ethernet to address those challenges. And we think there's no impediment for doing so.

We think in the end, an Ethernet is a superior solution to InfiniBand or any of the other alternatives that have been proposed in dealing with high performance, high scale, scaling out behind some of the physical limits. One of the things people don't realize is there's physical limits on an InfiniBand network, that limits you to a certain size.

We've helped form the ultra-Ethernet consortium. That consortium is going to publish their first specification, we believe, in Q1.

And we think we're well on track to be one of the first to offer an ultra-Ethernet networking solution that really will unlock million plus node clusters with very high performance, very high resiliency, very high manageability, very high debuggability, which are all critically important aspects of not just how do you build a cluster, but how do you keep it running and keep it running efficiently. And those are key attributes we think of what ultra-Ethernet and evolved Ethernet are going to provide.

TIMOTHY ARCURI: So bottom line question that I'm asked all the time, when do you think-- given what you know about your roadmap and given what you know about NVIDIA's roadmap, when do you think you're going to have a product in the marketplace at the same time as them that is equivalent or better, that you've caught up?

FORREST NORROD: From a day one perspective, I would say right now, the products in the market for NVIDIA are H100 and H200. I think we're very well-positioned with 300, 325 against them. But you're saying, when are we going to introduce it at the same time?

Look, we're taking the same approach on the GPU side as we did on the CPU side, which is build a multi-generational roadmap, put in place the engineering discipline to retire technology risk during the development cycles in a predictable way, and run them down. And so that's what we're doing.

So we're doing the same general approach on the GPU that we did on the CPU. And I think that we're-- by the time you get to the middle of next year, GB200, I think, really will be deployed in volume at that point. That's when it's really going to be starting to ramp up in volume.

I think we're going to be there with 355. And I think there's no questions, no asterisks on our MI400 generation. We aspire to be there with a leadership training and inference and chain of thought solution with MI400.

TIMOTHY ARCURI: Great, and with the last minute we have for us, how do you give investors confidence that you can keep up with this annual product cadence? Because they really have accelerated the roadmap to this annual cadence. Really it's a new platform every two years, not every year, with a memory upgrade every year. So do you feel like AMD is a big enough company to keep up with this annual cadence?

FORREST NORROD: Yeah, I think we are. And we've already adjusted our resourcing to do so some time ago. So I think wait and see. We'll see. But we're very confident of being able to hang on to this annual cadence. And very importantly, we are extremely experienced in critical elements of technology that we think will be increasingly important around chiplets, 3D stacking very large body and substrate devices. And we know how to retire that risk. We know how to deliver those without surprises.

I think that as others have to go down that path, they're going to potentially encounter problems. I think they already have. And so we're all about execution, fidelity, and meeting our commitments. I think we've shown that on the CPU side. And I'm confident we'll deliver on the GPU as well.

TIMOTHY ARCURI: Thank you, Forrest.

FORREST NORROD: Thanks a lot.

TIMOTHY ARCURI: Appreciate it.

FORREST NORROD: All right.