Deutsche Bank 2026 Technology Conference

Qualcomm Participant

Durga Malladi, EVP, GM, Technology Planning, Edge Solutions, and Data Center

Conference Participant

Amy Sutter, Deutsche Bank Analyst

Amy Sutter: Thank you, everybody. Hi. I am Amy Sutter, and I'm sitting in for Rob Sanders, our European hardware analyst who unfortunately is not able to be here today, and he covers Qualcomm. So I will be running the fireside chat. And I want to introduce Durga Malladi, who is the EVP of Tech Planning, Edge Solutions and Data Center who's here from Qualcomm. And maybe just to get started, I know you said you've been at Qualcomm since ’98. Maybe give us a little bit of your background, kind of what different roles you've had?

Durga Malladi: Sure. Yes. Yes. So I joined Qualcomm in '98, so it's my 29th year over there. Inside Qualcomm, we have gone through like at least two different DNA mutations over the last 2.5 decades or so. So this is the third one. I ran our research and our R&D organization for the longest period of time until about like ten years back or so. That's when Cristiano, who was -- who is our CEO now, my boss now. So he said, "I want you to come over on this side and help on the business.” So for the last few years, it's been more than a few years, now over eight or nine years have been running the product planning.

And at this point in time, I run all of our technology road map across all the businesses, ranging from this lowest end IoT all the way up to a data center. Last year, when we restarted our data center business, I ran the business for one year just to kind of make sure that we're in the right direction. And now we have like a full-fledged business unit that's running by itself. So that's my role.

Amy Sutter: Awesome. Great. I think since the Analyst Day, most common questions, I think, that we've been hearing from investors is how are you going to differentiate in the data center with your newly announced high-bandwidth compute technology -- can you maybe help us understand the genesis of HBC and kind of where it fits into your total strategy?

Durga Malladi: Yes. It started off- first of all, when we unveiled our HBC product portfolio and the multiple generations that we are working on, -- it's important to understand that it's not something that we just did in one year or so. It was something that's been in the works for the longest period of time. But it was last year that I felt like we have reached a point in our R&D where now it's time to commercialize this. But working backwards from there, what is the problem that it is solving? That is something that became quite clear to us by 2021 time frame. When we were observing how the compute in each of these data center racks, and in fact, it also holds true for some of the other places. It's growing quite a bit exponentially, generation over generation, while the memory bandwidth was kind of stagnant or maybe growing more modestly, it's more of a linear growth. And it eventually will lead to a memory wall wherein you can throw as much compute as possible, but it is going to do absolutely nothing when it comes to influence decode portion of it, unless you solve the memory bandwidth as well. And so we invested quite heavily in our R&D. And we reached a point where last year where it's becoming like, yes, now it's feasible. We've scracked a lot of the solutions over there.

Keep in mind that we called it as high-bandwidth compute because it is not really a pure memory technology. It is actually about a combination of how do you have the compute right next to memory? How do you design it, codesign it in the right way and then make sure that it interfaces with the rest of the accelerator. And it was a direct take on an alternative way of doing things as opposed to HBM. And we feel very confident and comfortable with where things are right now, and that's why we unveiled our multi-generation solutions with the first generation coming out next year and the second generation after that. But that is a very unique proposition. And now it's fully endorsed by a large number of the memory vendors as well who've been our partners for a while, but now they are equally out in the open and talking about something like that.

Amy Sutter: Yes. Great. Yes. I would not have thought Qualcomm as being in the memory game. But talking about it as a compute technology is very helpful. Maybe just to dig a little deeper, when you talked about Gen 1 and Gen 2 and I think at the Analyst Day, you showcased the bandwidth per watt relative to HBM and then you also have some very high numbers relative to SRAM. Can you talk about the unique selling points versus HBM and SRAM that you're talking to customers about in terms of solving the issue with the memory wall.

Durga Malladi: Okay. So 2 things over there. First is some of the numbers and the second part is on proof points, and maybe I'll shed a little bit of light in terms of how the discussions are coming up with hyperscalers. So in terms of numbers, we always, always talk about tokens per watt because the power consumption is something that is very important. And today, with HBM, the power consumption is very high. By the way, a funny anecdote on that because we talked about 6x improvement compared to HBM in terms of tokens per second to watt. And the second part is HBM today is about, give or take, 21, 22 terabytes per second. We had like hundreds of terabytes per second. And when we compare it against SRAM, SRAM has got its own restrictions. It is the fastest interface, but it's also like it consumes more area, you will need a larger number of racks to do the same thing. So there's a TCO argument to be made against SRAM.

These are the three things that we talked about at Investor Day. But the important point is, okay, these are numbers. What are the proof points? So the nice thing is that, yes, we taped out the silicon, the silicon is back in the lab. It's coming along very well. Stay tuned on that front as we start unveiling this a little bit more on what the numbers are beginning to look like. But we are pretty confident in terms of the first generation of the products.

A bit of an anecdote on this. Like right after Investor Day, we spoke about this, we said, "Hey, here's HBM", and there were like these two nice animations, one which talked about, okay, this is no good with HBM. On the other hand, with HBC, this is great. Right after that, I was in Korea with the two memory vendors, and I fully anticipated them to say, "what are you talking about? We have this awesome road map with HBM, that's like really great. It's going to be like much better than HBC." That is not the feedback that we received. Instead, what we got is, "hey, how can we work with you much closer on this." And in fact, we have been working with them because we need to make sure that we have the right supply as we do the commercialization. But in addition to that, there's been a lot of interest coming in directly from the memory community itself, which has been a very positive and pleasant experience. And that's actually great. It also kind of validates some of the points that we made that we need a different kind of a solution as we move forward in these data center racks.

The second generation of the product, so the first generation is slated to be commercial and shipping in 2027. We're already working on the second generation, which is going to be 2028. And those numbers are even higher. And one of the reasons for that as in terms of the relative comparisons against what we anticipate with HBM4E. And the reason for that is as we kind of look ahead to 2028, a lot of things are evolving in itself. On one hand, AI models themselves are evolving, which means the compute that we end up doing in the compute die or the logic die, which is underneath the DRAM stack. We actually are anticipating what we need to do, and we are throwing in even more arithmetic. It's harder to scale that with HBM, so the relative gain looks even better with that. That's where we are.

Amy Sutter: Great. Okay. One of the criticisms of it is like with the multiple layers of LPDDR on top of -- with high-performance logic that there's thermal issues. So maybe can we talk about how you're getting around these with I guess, your packaging technology. And then you mentioned Gen 1 and Gen 2, just maybe some milestones for us as investors to think about, it's for us to look for in terms of this. And then I know you're talking about talking with the Korean memory players, but also are you engaging with hyperscalers today? I don't know if that's something you've shared yet.

Durga Malladi: Okay. Short answer is yes to all of them, but I'll start with proof points. So as I said, silicon is back in the lab, looking good. It's coming along very nicely. I don't believe we have explicitly put down dates as to when this is going to be unveiled. But in the next quarter, I just stay tuned on that front. I think some of it is also within our marketing domain in terms of how we want to expose it. But we will have numbers. We will actually have the silicon validation coming in. And we are on track to ship both from the samples, the engineering and the commercial samples, and this technology validation in itself, I think it's looking good. We don't anticipate any issues for the first generation. What's happened is that as we brought this up, and we started talking to every single hyperscaler. Every hyperscaler, they have a lot of their own in-house solutions. Some of them are like, "I have my CPU." Some of them are like, "I have my CPU and I have my AI accelerator." But very few of-- they always relied upon HBM as okay, it's one of those memory bandwidth things. I will acquire that IP and then I'll build my rack on top of it.

It's always been like really good when we sat down with all the hyperscalers and they took a look at this, like this is awesome. How do I actually bring this in? And at Investor Day, we specifically talked about these are 3 different tracks, independent tracks. For example, some hyperscaler might come in and say, "I have my CPU and my AI accelerator, I need your HBC to be a stand-alone product that works with this." We have a solution for that. There might be someone else who will say, "You know what, I have my AI accelerator. That's good. I like your HBC and you know what, your CPU is looking pretty good, too. Maybe I can take both of them. How does that work with me." We have the answer for that as well. It kind of is now at a point where with hyperscalers, we're not coming in with some sort of -- it's all or nothing. No, you can take individual components, if you want, or you can take all of the above. And then we take it from there.

Amy Sutter: Got it. One of the things when I think HBC, the concern is maybe when you look at prefill versus decode, and I know memory is more focused on the decode side. That's where the bottleneck is. But can you maybe talk about how HBC does there.

Durga Malladi: Yes. So if you are looking at inference in data center, typically, we split that problem into there's the prefill stage and the decode stage. The prefill stage is fully compute dominated. The more compute you throw it, the better it is. And off late, we are seeing something like a 2.5x to 3x increase in compute generation over generation. So you can always solve the problem with throwing more compute at it, HBC will not do much about that one at all. On the other hand, decode is completely memory bandwidth we dominated. You can throw as much compute as you want, it doesn't matter. Unless you have a solution like this, it's not going to work. And that perhaps is the place where in all these discussions, not just the hyperscalers with a lot of the others as well, we see this, "Hey, I have my compute solution. I can use it for prefill, perhaps for decode, as you're doing HBC, there's additional logic that goes into the bottom die. You can bring in that logic over there." So there's like even more of a mix and match that's occurring in that space. That's how we anticipate at least in the short-term things to evolve.

Now in addition to all 3 of these, CPU, our AI accelerator and HBC, at Investor Day we also talked about our custom silicon business. Now if you think about everything that I've said, you're almost getting to the point of is it like a custom silicon for something that's going into a hyperscaler. I think there is a very thin academic line between these two arguments because once you start putting it together, it does go in that direction. But there's a heavy dosage of our IP that goes in along with mix and match with their IP as well.

Amy Sutter: Got it. Got it. I think Hynix and SanDisk are also talking about high-bandwidth flash, though it seems like it's a little ways out, at least that's my take. But could hyperscalers take a pool of HBM and HBF and run the models more cheaply? Like how does your HBC defend against that kind of TCO narrative, if that's something that becomes a way that we're trying to attack the memory wall.

Durga Malladi: So when we talk to memory vendors and we kind of have gone into this in depth with them, in terms of HBC versus HBM and what's coming with the high bandwidth flash. This is my opinion over here at this point in time, HB Flash is trying to solve a slightly different problem. It's not quite exactly the same problem because one of the things that the argument over there is -- let's also use flash in addition to DRAM to actually do something else over here. That's not the same thing as increasing the memory bandwidth to the point that the decode performance becomes really good. But it has its own use case. There's still more work to be done. That's just the perspective that we have right now. But it is one of those stacks, which is kind of independent of this. Meanwhile, with HBM, comparison against HBM, the same vendors who have -- they've been working on HBM for a while, they are like, I need to open a parallel track in addition to this because it's not obvious to me that in the near future, at least for the foreseeable future, not just near future, there's anything like HBC that I can offer through my HBM. Now time may well, maybe 5 years down the stream, it's going to be a different story. But as of today, that's the only solution that actually gets us to this level of performance gains compared to what we have.

Amy Sutter: What I think has been very interesting for me coming from Hot Chips earlier this week is just the diversity of solutions that are coming out today in compute, GPUs, CPUs and memory and how every hyperscaler and customer model company is attacking it differently. When I think about HBC, can it be relevant for other end markets outside of data center, maybe auto?

Durga Malladi: It's a very good question. When we talked about HBC at Investor Day and of course, at Investor Day, we were talking a lot more in terms of our investments into beyond what we typically have done in edge devices, and that's why we talked a lot about data center. But truth be told, HBC is a generic technology. It goes all the way down into devices. And let me explain why that's the case. And we actually are working, believe it or not, we are working on HBC on devices that are around us today. We're actually working in terms of what is the technology that makes sense for that. And here's the reason why. In data center, the comparison is against HBM, and we are talking of what's the best way to solve the decode problem and the memory wall problem over there. There is no equivalent problem in devices. Let's take a smartphone or an agentic AI device, but there's a different problem over there. Especially with agent AI devices, which are like -- these are devices that are being built as we speak with those who've been in the business of thinking AI first, and I need to build a solution around it.

But it's all about ambient AI. It's constantly running. That's why we call it agentic AI. It's always running in the background and just waiting for you to say something and then it picks it up from there. But if you're always on, ambient AI always on, that means you're always burning power. Now HBC is compute right next to memory. So there's something else that you get. We've completely collapsed that transport latency between memory and the compute and not just that, the power consumption has come down dramatically. It's perfectly suited for ambient AI in all sorts of agentic AI devices that come in now. So the usage of HBC in devices is in the context of ambient AI, whereas in data center, it's about the memory bandwidth that it provides and the decode performance.

Amy Sutter: And when you say like edge cases, are we thinking about like something like robotics, where we're thinking about AI there? Or could you think of it being used in a car or a phone?

Durga Malladi: Going all the way down to smartphones, tablets, PCs, XR devices and yes, into automotive as well. It comes down all the way to those.

Amy Sutter: And when do you think that happens?

Durga Malladi: Discussions are going on as we speak. Typically, we have like a lead time on this. Like if you have a commercial product in one year, then usually, it's like 3 years before that is when the commercial discussions begin. That's a question for -- once it shows up, you will realize it, but suffice it to say that we're not that far from the first generation of the commercial products coming in.

Amy Sutter: Got it. I guess then maybe kind of closing out on HBC a little bit. Do you think it's with Gen 2 or Gen 1 where you see customers really clamoring for it? And does it become a technology that you might offer to license to third parties to help them scale? Or do you look at this as like some kind of proprietary moat that Qualcomm can offer?

Durga Malladi: So there's two questions over there. So the first one in terms of going beyond like Gen 1 and as we're going into Gen 2, where exactly is the action over there. Gen 1 is next year, like that's like literally next year. There would be a few customers will go in that direction. But then as we talk to a lot of the hyperscalers and so on, there will be some separation between where the volume is going to be between Gen 1 and Gen 2. We will certainly see traction on Gen 1. I hope we see a lot more because more and more customers are lining up. So the earliest adopters are the ones who are going to go with Gen 1. But now we have so much of incoming interest coming in, I do anticipate more coming up the Gen 2 as well. So that's the way that we see it.

The second part is in terms of how we see this technology. Just keep in mind that there's actually three different components to it. We're not a memory vendor. So someone is still doing their DRAM stacking, we're not doing the DRAM stacking ourselves. What we prescribe is this is how things should be done. These are the TSVs that need to occur. This is the design for it. The compute die of course, comes from us. We do all the logic inside that. We still have to work with the likes of TSMC to tape it out together. I think we're all in our swim lanes in our own way. But we are kind of like the overall system integrator and putting it together in the right way.

So in that sense, this is our current model. We haven't said anything about how it's expected to evolve, but that's where we are now.

Amy Sutter: And so if I'm a hyperscaler customer, would I be the one who is focusing on the memory like in terms of like because memory is a big bottleneck today. You don't have to worry about that. The hyperscaler worries about that, and you just help get the solution integrated together in any-

Durga Malladi: So when a hyperscaler is interested in our HBC product, they will have certain volume commitment, some sort of a demand signal that we get, but they talk to us and they will cross check and make sure that, hey, you're all like from a supply perspective, that's great, but they talk to us. And we are like the front end for all of that conversation. And of course, they will be talking to the memory vendors to make sure and to the likes of TSMC to make sure everything else is an in the right order. But this is a product that comes from us to them. They are our customers. So that's how it starts.

Amy Sutter: Okay. Okay. Great. Switching over a little bit. Let's talk about the Dragonfly C1000 server CPU. You have some very high claims in terms of performance per watt relative to peers. But it's a second half '28, if I'm correct.

Durga Malladi: Right.

Amy Sutter: So you're kind of giving away some stuff. How do you think about what competition is doing? And does the baseline shift from the other big CPU competitors? Could it be challenged at arrival?

Durga Malladi: So last year, I remember when we decided to restart this and I was thinking about, okay, what are the differentiations that we have over there. The first one, I was like, okay, we got to go into the memory architecture and memory wall, solve the problem. Our R&D project is good to go. So that's good to go. Then I was looking at, okay, where are we with our existing CPU solutions? And what's the right way to scale that? We had something in planning already at that point because over the last three or four years, we've been investing quite heavily in our custom CPU solutions our

Orion family of CPUs. And I was paying very close attention to exactly where each of the data center class CPUs are. You can start from Neoverse V2 and take a look at the generation over generation. So what I was looking at is we're not competing against anyone else out there today. I want to compete against where they're going to be in 2028, and that was like the mindset with which we went in, and we need to beat it by a certain generation for it to be compelling. If it's like 2% better, then nobody is going to take it, compared to their own in-house solution or something else that you get; it has to be really compelling. That's the bar that we set for ourselves. And over time, as we kind of built this up, and we did all the benchmarking typically, this is done with SPEC 2000 and so on, and we have our projections. One of the hyperscalers -- the response from them was "Your numbers are too good to be true." I was like, okay, so what I'm hearing is, "if you do actually show up with that, you will like them. He was like, "Oh, yes, absolutely." So now, we made our claims. You're absolutely right. We feel very confident about those claims. We have absolutely no any sort of a second guessing on that at all. Now it's execution and crunch time. We announced Meta as a first customer for C1000. They believed our claims. Qualcomm is a company where if you say this is what is going to happen, never have we actually fallen short of it. We have always been on par or better. So they went with our reputation and said we are good to go. There are others like let's see your silicon and then we'll be there. So I'm really looking forward to that, but I feel extremely confident.

Amy Sutter: And so the next milestone will be launch with your first customer?

Durga Malladi: Launch with the first customer. But even before that, the commercial launch is in '28. But with silicon back in the lab, all the proof points and so on, hopefully, we'll be able to share a few more.

Amy Sutter: The offerings we see in the market today, some of them are more focused on more cores. Some of them are more focused on single core like hyperthreading or if I get it right. So I mean I'm kind of curious from your perspective, when you look at CPUs in the data center, what you think is the right solution or maybe there isn't, like, again, because workloads are going to vary so much, but I would love to get your opinion on that.

Durga Malladi: Until about two years back, I would say that, by the way, the CPU landscape is evolving as we speak, and it's becoming like for the longest period of time, including inside Qualcomm, by the way, it's like our XPU team or the GPU and the NPU teams are like, okay, we are in the AI business. And the CPU team is not quite there. It's a little bit out there, but it's not quite like that anymore. All three processes are important, but in a different way. But what's happened in the last two years is it used to be that you had two different kinds of workloads in data center. General-purpose compute, that means you just run, you light up all the core if you have -- at Investor Day, we said that our racks have like 250-plus cores and so on. Okay, so let's say that you have 256 cores or so, then you light up all of them. It's a general purpose, it's like a work horse. Then there is a different configuration, which is an AI head node, where actually most of the AI workload is being done by some XPU that are sitting out there. The CPU's job is to do some management and kind of get out of the action and let the GPU or the NPU perform whatever it is that's needed. All the AI inference occurs over there. But that led to an asymmetry in terms of number of CPUs to XPU. It was like 4:1. It used to be 8:1, became 4:1. But with agentic AI, that's changing quite a bit. In fact, nowadays, we are more in the 2:1 configuration for every CPU, there's like two XPUs, maybe that's not four. And I wouldn't be surprised if it actually becomes even less than that, like it becomes more and more symmetric. So as these conversations evolve, we initially started, hey, your cores looked really great, excellent for general purpose compute. I think I got it for AI head node. I probably don't need you, but general purpose compute is great. But over time, now we are getting like -- let's actually talk about not just general-purpose compute, but also agentic AI because I would like to use those CPUs over there as well. So that's the evolving landscape with CPUs and the workloads themselves are shifting. They're becoming a mix and match of general-purpose compute and agentic AI, not just AI head node, but agentic AI workload, which look pretty different.

Amy Sutter: Right. Okay. And then also kind of talking a little bit about your road map overall. Like you're also using your RISC-V cores. And so you kind of have two road maps here with Arm and RISC-V. Maybe you could talk about, like why do that? What challenges does that create or what opportunities does that create?

Durga Malladi: So a couple of data points on that. The RISC-V conversation is an important one because it's usually -- I would argue that whenever one talks about RISC-V for the longest period of time, it was seen as something which is academic, which is a little bit out there. It's a good science experiment. If you're a grad student in the university, it's an excellent project that you should be working on, but not really for commercialization. That used to be the narrative till about 1.5 years back; not anymore. At this point in time, there's a lot of not just interest, but pre-commercial engagements that are occurring in terms of where do we take RISC-V cores and- but I'll get to that. Because the first question to ask is, why? What's causing this? What exactly is the issue with this?

Historically, what has happened is that if you're in the custom Arm CPU business, then you inherit a certain architecture and your innovation is within the micro architecture domain. That's where you tend to innovate and you try to differentiate. But you inherit a certain architecture in so much and that's all there is to it, which means your scope of innovation is only below a certain threshold but not everywhere else. RISC-V, on the other hand, is completely open. If you have some ideas, you actually go to the standards process and you say this is what I would like to see and that gets done. It's an open standard. So there's more innovation coming over there. It's a clean start. Every three decades or so, there's a new ISA that comes in and instruction set architecture. And we see a lot of potential in RISC-V as well. So that's where, as Qualcomm, we started investing quite heavily. First, making sure that the standards organization is like it's a professional organization that is run with commercial milestones in mind and not just for research purposes.

So Qualcomm has a lot of leadership positions in the RISC-V forums. In addition to that, we work with our hyperscaler partners and say, hey, it can’t be just us. There's got to be like an entire software ecosystem that needs to come in, and they started participating quite heavily. And we have like very good partners over there, a few hyperscalers, both in U.S. and in China. Which brings me to the other point, because the other narrative that comes up is RISC-V is something that is going to be done in China, but I don't know about the rest of the world. Not true. In fact, we will start seeing that in the U.S. as well because there's a lot of incoming interest.

So as a part of this, we have started investing quite a bit in our RISC-V portfolio even towards data center, especially towards data center. And last year, we made an acquisition of Ventana Micro Systems. And that team is now fully integrated and we are kind of on our way.

Amy Sutter: What is the advantage of RISC-V then?

Durga Malladi: So a couple of things come in. First of all, it's going to be a parallel track, which is occurring at the same time. And we see the demand is going to be like a mix and match of both over time. We expect to see that just because of the features that are coming in on the RISC-V road map as we look ahead and it's to be seen as to how that actually -- where does this end? Are we going to have two tracks. We had x86 around for the longest period of time. And so one shouldn't be surprised to see even these parallel tracks come in. It's not A or B, it's A and B depending upon their own customer needs and choices.

Amy Sutter: Why did China -- it seems like China has been the early adopter. Is there a reason for that?

Durga Malladi: That's a good question for China.

Amy Sutter: All right, I'll ask them. And then maybe to move on a little bit on the software side of things. Obviously, as we all know, I think NVIDIA dominates with CUDA. And changing those programs to be able to adapt to non-NVIDIA chips, it's hard. And Modular, that software, how does it make it possible if I'm in NVIDIA to maybe use for example, maybe your CPU or what have you. So

how does that happen? And then how do you reassure developers over time that these compilers will be open and available, so that they don't get locked at Modular?

Durga Malladi: Absolutely, yes. By the way, that's -- there's like two or three questions over there, but I'm going to actually start back with -- it's okay, but it's a very big investment that we've made, and it's kind of an important concept. So I'll start with -- at Investor Day, we talked about our acquisition of Modular. We hadn't closed yet at that point in time. And so Chris Lattner actually talked about, okay, where this comes in. And there was this one flagship slide. For those of you who might have either attended or seen it on video. And if not, I'll just explain where we put together, with Chris and Tim, we were literally on the floor and saying, okay, we should put it like this, like literally put them right next to each other. So there's Mojo and then we put in CUDA on one side, that's from NVIDIA. Then we said MAX, that's with the compiler layer, and said this is Triton with the compiler layer over there and Modular cloud, and that's Dynamo on top of it. So it's like this is the CUDA stack, if you will, and this is the Modular stack. Underneath that, from a hardware perspective, it's NVIDIA hardware only. And here, anyone -- it can be any third-party hardware and any third-party processor. It's a bold claim. And then right below that, we put in a footnote. We said, when you take Modular stack, and you take your existing rack as an example. And you have your native stack and then you say, okay, instead of that, you use Modular stack. The performance is going to be on par or better -- and when we said better, we said up to 50% better. That was the fine print over there. It took some time to actually write that down. And I remember telling okay, Chris, okay, now that we've thrown this gauntlet, so we're going to get a lot of questions in, and it's going to be fun. And it is. In fact, one of the first things that we did, the moment we did that we got some really good feedback from a lot of the others saying, “This is awesome. I would really like to actually try this out and take a look at it.” Because others have tried. Others have tried exactly the same thing, and it's not quite panned out that way. It seems like too good to be true.

You have true disaggregation of third-party software that can run on any third-party hardware. And as Qualcomm, we are saying -- here is a software product, try it out and it can run on your hardware as well. And second, it's going to be open. There's quite a few new things from Qualcomm's perspective. This is- that's why I said it's our third mutation over here. Last week, at ModCon, then we went one step further. Then we said it's going to be open source for Mojo. And on MAX, by and large, it's open source. It's an Apache 2.0 license, anyone can take it. They will have full source access of this. There's some portion of it that's going to be licensed, and we go from there onwards. And we showed even more numbers.

The best part, and we had a person from AMD who showed up on stage and said, “This is awesome. I'm going to try this out.” That is like a game changer. I mean, it seems -- yes, that's okay. But actually, at least for us, we were like it's quite something to have a third party come on stage and say, “I'm going to actually start taking your software, and I'm going to try this out.” So everyone is interested in now kicking the tires and taking a look at what the performance is going to be. And we really want third parties to reach that conclusion themselves. What we don't want is, yes, we are doing our own analysis. We have our benchmark. But if I were to show up over here and say, this is how much better it is, but it's all done by Qualcomm, no, I want AMD to do it. I want someone else to do it. I want third-party artificial analysts to do it and reach the same conclusion. That is what is happening as we speak.

Amy Sutter: It's interesting you say AMD, I mean, does it compete with their ROCm.

Durga Malladi: That's something that they are kind of -- that's a question for them, by the way, but it was nice to actually have them say, “This is something we want to try out.” Where that takes them, that's a different story, and we'll take it from there. And the final part that I wanted to mention, which is kind of a humorous anecdote because the Modular -- I was talking to Chris, they had brought up Modular software on everyone else's hardware, except Qualcomm, which is kind of so odd. And so last week, for the first time, we also unveiled running on our platforms. including one of the

older generation AI 100, that's on the data center side. And we even showed it on one of the laptops, the X2 Elite platform.

So it's the beginning of -- and we said any hardware, it, of course, must include Qualcomm as well. So we're beginning to do that.

Amy Sutter: Got it. Can you talk about the partnership you have with Hugging Face?

Durga Malladi: Yes. So that was the second part of that presentation at Investor Day. Hugging Face: 15 million developers flocked to the website on a daily basis, largest repo of AI models, open weight models out there, 3-plus million models. And on the other hand, we have some data that indicates that the most popular ones are probably like far more concentrated, not all 3 million are the same and some of them are derivatives. But clearly, it's the place to go if you're a developer, you're in the business of building apps or writing agents or building agents. First, you go to shop. You actually -- that's like the place that you land up with and say, what do I have to work with, like all the things that are available out there. So in our partnership with Hugging Face, we want, first of all, agentic onboarding of any of those models. Any of those models, agentic onboarding of those models onto our platforms. What does it mean? Today, I go to Hugging Face, I'm going to click, I have to manually click on something and then it gets downloaded. Then I have to write something with that and then have to write some specific commands, and then I can start building the applications. I don't want to do that.

I have one small thing. They actually have something called Hugging Chat. I would like to pick some model that does object detection and classification or a text to video generation model, pick one of the good ones and make sure that it runs on this platform from Qualcomm. That's your prompt. Everything else happens behind the scenes. You can actually see the code being written down out there and it gets done and it comes on to your platform. That makes it extremely easy for developers to start adopting our platforms over there.

Second piece of the puzzle, we threw in Modular as a part of that as well. And that means if I'm a developer, I would say like instead of in PyTorch, maybe I want to actually use Modular. A developer might say, do I have to now learn another language. Mojo just happens to be very similar to Python. But these days, increasingly, developers themselves are not necessarily writing all the code. They're using coding agents. So I'd rather say using Mojo, I would like to see an application being done -- that's actually written because it does all the tool calling, looks at all the documentation getting done. This is where Hugging Face has emerged as a really good partner for us to do a direct engagement with developers while at the same time, those developers get exposed a lot more with the Modular stack.

Amy Sutter: Got it. It sounds like a great partnership.

Durga Malladi: It is.

Amy Sutter: Yes, for sure. And then you kind of touched on this when you're talking about running Modular on some of your older technologies, Qualcomm technology that you just announced. But can you talk about how it might then support your cloud-to-edge strategy? And can it be like a universal layer to write an enterprise agent AI agent all once and kind of trust that it can run on these different devices, whether it be a laptop or data center CPU.

Durga Malladi: So that's the, I think, the final piece of the puzzle. We talked about it even in our Hugging Face engagement. If you take these emerging agentic AI devices that we are talking of, a good fraction of them actually, the user interface is directly to an agent. You ask something you want a specific task to be done. The agent then decides what needs to be done. Do I run some inference on the device. First, I need to shop around inside the device and see what models do I have to work with? Is this good enough for what I need? Or do I need something better? If it's good enough, I use that, but you might still need something else that's running in the cloud, so you go there and you do a second inference instance that's running there. That's two parallel inference instances running

concurrently. It's not one or the other, but it's both. And meanwhile, there might be some sort of a tool calling with a web call in which you're extracting information from somewhere else, that's the third part. When you put it all together, we have reached the conclusion that it's never going to be one or the other. It's going to be all of the above, and the agent is running several inference instances. So hypothetically, if you were to picture a world in which you have Mojo and MAX, the Modular stack actually running on the device, and it happens to be running on the cloud instances where the inference is running, does it make it more efficient? And the answer is yes. So that actually might be -- and by the way, it could be any rack. It could be an AMD rack, it could be a Qualcomm rack, it could be someone else's rack. It doesn't matter. But does that make it more efficient? We do believe the answer is yes, but that's what we are working on.

Amy Sutter: Maybe it's a basic question, but like how did Modular -- like what was the secret sauce? Because you said so many people have tried this. Why were they able to do this when no one else could? Because I think a lot of people would like to break the CUDA moat.

Durga Malladi: It's a combination of things, in my opinion, but one of them happens to be that if you kind of think about the existing stacks that are there, including CUDA, for example, CUDA itself is about 15 years old. I mean what was initially written for graphics and specifically focused, it evolved over time to this. So there's a lot of legacy as well. So one thing that's certainly been done is when it comes to the development of Mojo and MAX, it's kind of a fresh look at it, and it's turning to a fundamentally different way, which makes it cleaner, simpler and for lack of a better phrase, more modular, like literally, it is modular, the objective.

The second part of it is that as you go through the compilers and this goes one level below into the technology in itself, compilers themselves have been gradually evolving. There is notion of using AI within compilers itself, and that's emerging as well. So a lot of these things have been built into it. We believe that's the real differentiation. There's certainly far more to it, but I would actually reserve that for a tech deep dive.

Amy Sutter: Got it. And then maybe just to talk a little bit about prefill and decode in terms of Modula’s tool stack. How does their software -- how does that software handle that? Can it do it dynamically?

Durga Malladi: So just in terms of the Modular stack, the way that it's currently envisioned and constructed, it doesn't-- it kind of sits on top of like I would say, prefill and decode are like system-level concepts sitting upon certain hardware, but the software layer itself is kind of independent of each other. So everything that I said earlier about Modular equally applies to both stages.

Amy Sutter: Got it. Okay. And then also, when you think about the memory wall that we've been-- and also the power wall, is there ways to-- does Modular also kind of attack some of this in terms of like harnessing idle devices, et cetera, that might be able to help improve the efficiency or performance?

Durga Malladi: I think it's a little too early to come to that stage. We are still doing a lot of our analysis in terms of what else we can do in this space. But at this stage, I think it's a little too early.

Amy Sutter: Got it. And when investors are thinking about Modular, like what kind of milestones and proof points should we look for? Are you talking about like number of developers using or-

Durga Malladi: We have a lot of things actually planned out over the next quarter or two. As we start bringing in speeds and feeds and data points in terms of, okay, here's an existing rack, here's with the native stack, this is the performance based upon all these workloads. And you compare and contrast with what happens when you have Mojo and MAX running on top of it. That's like the first set of-- tons and tons of data coming up on that one. The second proof point of that is, how well does it actually scale into all the other platforms beyond data center. Keep in mind that as Qualcomm, from a device perspective as well, we are in the business of any framework, any run time, any operating system. And we provide all the debuggers, the compilers, the tools and whatnot that actually go along with it. And there is a very rich set of run times that already exist. I mean you have

Executorch/PyTorch from Meta, we have LiteRT from Google, we have Win ML coming in from Microsoft. They're all great partners of us. We will continue to support them while at the same time, also bringing in our stack as well. So the next step would be to how does it work on a given platform, when you already have support for one and where does Modular come in? That's the other part that you will see over the next couple of quarters.

Amy Sutter: Got it. I mean we've kind of touched on a lot of different things. HBC, CPU and Modular, a lot of exciting new opportunities for Qualcomm. I guess, like -- do you think there's something we didn't cover right now that you'd want to point out to investors? I know you just had your Analyst Day, but I thought you might wrap with that in the last.

Durga Malladi: I think yes. I think the one part that I would like to say is that there was a fourth piece of this. We talked about GPU and XPU and HBC, but custom silicon, I think, is going to be important for us. So post our acquisition of Alphawave, we've retained the custom silicon business. Increasingly, all of our discussions with hyperscalers involved some level of customization. It's a mix and match of everything that we bring to the table, along with their IP. And keep in mind that we are in the networking business now. We have SerDes, this is what we acquired from Alphawave, and that will continue to evolve. And that's one place where I would like everyone to be reminded of the fact that prior to acquisition, Alphawave already had these hyperscaler customers. So we'll continue to go in that direction.

Amy Sutter: Great. Well, thank you so much. I appreciate you.

Durga Malladi: All right. Thank you.