20-Aug-2024

QUALCOMM, Inc. (QCOM)

Rosenblatts Virtual Tech Summit

Copyright © 2001-2024 FactSet CallStreet, LLC

CORPORATE PARTICIPANTS

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

OTHER PARTICIPANTS

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

MANAGEMENT DISCUSSION SECTION

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Good afternoon and welcome to Rosenblatt's Fourth Annual Age of AI Conference. My name is Kevin Cassidy. I'm one of the semiconductor analysts at Rosenblatt, and it's my pleasure to introduce Dr. Durga Malladi. Durga is Senior Vice President and General Manager of Technology Planning and Edge Solutions at Qualcomm Technologies. Durga joined Qualcomm in 1998 as a senior engineer. And over the past 25 years, he's worked on 3G, 4G and 5G and various R&D and business roles. We don't have time to list all those accomplishments, but suffice it to say that he's a significant contributor to Qualcomm's success. He's a member of the IEEE Society, holds 578 US patents, unless that's changed, probably up to 600 by now. And we're very fortunate to have such an expert to meet with us today. As many of you know, Qualcomm is our top large cap long idea for 2024.

And just one other thing before we get started, if you want to ask a question, please submit it at the bottom of your Zoom page that says, Ask-a-Question and just type it in there, and I'll read the question for Durga.

Copyright © 2001-2024 FactSet CallStreet, LLC

QUESTION AND ANSWER SECTION

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

With that, Durga, thank you very much for joining us again today. This is Qualcomm's fourth year participating in our Age of AI Conference. The message has been that Qualcomm's Snapdragon SoCs have been and been improving their AI engines for many, many years. And this past year we've seen Qualcomm has gotten into some significant designs, like the Samsung Galaxy 24, which I'm a proud owner of. And lately, of course, the AI PCs with the Snapdragon X Elite in that they use a customized Oryon CPU cores. So, first, maybe you can just talk about some of these changes since last year, Durga, and I'll open that up to you. Thanks again.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Hi. Truly excited to be here. Thanks for hosting me here, Kevin. You know, if you just take a look at – take stock of the situation in the last 18 months or so, we are in – we're living in some – there's an unprecedented pace of innovation that's coming up utilizing AI and specifically generative AI concepts into both the smartphones and in PCs. So, I'm going to actually kind of cut to the chase in terms of what has changed and what exactly are the things that you can do today with the smartphone and the PC that you couldn't do before and where the potential is as we move forward. So, starting from over two years back when – maybe two and a half years back when ChatGPT entered, I think the – it became a buzzword in itself, practically everyone started talking about it. What became clear is that these extremely large models and generative AI, based upon these extremely large language models, is – was bringing in some truly remarkable results.

And from Qualcomm's perspective, what we were actually looking at was something slightly different, which is how do we bring that experience and try to bring those kinds of models, the really large models that can run in a very efficient way on devices that you and I can use and starting off with smartphones and PCs. They happen to be the most demanding when it comes to power consumption and overall energy efficiency, memory footprint. But at the same time, you want to have the same kind of an experience as if you are actually running with very large models. So, a lot of our work was really in making sure that these extremely large models run effectively. They fit within the memory footprint of an energy-constrained device. And performance per watt or the energy efficient, high-performance computing is where we started really focusing on. That's been our bread and butter in several – over the past several years.

But increasingly today, we focused on AI influence, which is running in this way. So there's three processes that typically are there in these phones and laptops; CPU, GPU and what we call as our NPU, which is custom designed for AI workloads. So let me actually start with the Samsung Galaxy S24 you picked up the phone. It's a nice phone. I have one myself. There are two features that – which actually, if you have not used it quite a bit, I would recommend that you kind of test it out a little. There are – there's one feature called Live Translate. Live Translate is the way in which you and I could be talking in two different languages and yet be in a position to understand each other because what I speak gets converted into a language that you understand – in real time, mind you – and what you speak then reconverts – gets converted back into a language that I understand, again in real time. And only one of us needs to have that phone, both of us don't need that.

But when you go to the nitty-gritty details of it, it involves a generative AI model that converts my voice into text. Then, there's a large language model that converts that text into a text that you understand, retransmits it back as voice and so on and so forth. And it's quite clear that it's impossible to do this on the cloud; the only way this

Copyright © 2001-2024 FactSet CallStreet, LLC

works is if you run it on device. It's a pretty powerful feature. I keep, once in a while, mentioning this; if you're a movie fan and if you remember that movie, Hitchhiker's Guide to the Galaxy, there's the Babel fish. You put it in your ear and all of a sudden it's a universal translator. You can hear any language, but it's kind of a good way of actually thinking about the kinds of use cases that come in. These are consumer-centric use cases, right? Similarly, in addition to what – this live translate that I talked about, there's live wallpaper generation. These are even more consumer-centric, image-centric use cases where you have a text to an image generation.

These are like the beginnings of the first generation of the AI-based use cases that you use in smartphones today. There's more, but at least in smartphones that's what we see. But it's a combination of large language models. These are like billions of parameters that are typically used in these language models that can fit within the phone. And at the same time, they're not very demanding on the power consumption, and we've done a pretty good job in terms of that. And if you kind of think about it, people have been very happy with these first generation of AI use cases. Switching gears now from there to PCs, that's where the excitement is at a very different level altogether because it's a combination of two things. First of all, when we talk of AI PCs today, AI PCs today, today you can walk into a store and buy – there's like more than 20 different AI PCs that are available.

Almost all of them, I would argue, are based upon Qualcomm SoCs to the last count that I remember. And it has – like, there's something very unique about these AI PCs. First, obviously from our partner from Microsoft, they've been quite big in terms of promoting Copilot. But their definition of Copilot+ is it runs on device, your AI is running on device. Now here, we have to separate out consumer and productivity use cases. Productivity use cases are things like document summarization, converting from one kind of a document to another, your Word document to a PowerPoint presentation, with only a little bit of a nip-tuck, not a whole lot on top of it constantly. And they haven't fully unveiled the recall feature that they talked about, but it is coming in. But it's a way in terms of you actually can just recall whatever you did at a certain point in time and do this in a much more intuitive way. But the bottom line is the human interface to the computer is changing dramatically.

And this is a place where – we also talk in terms of AI constantly running in the background. It's pervasive AI, it's not just an on-demand AI where you ask a question and you get an answer, but a lot more than that. So, bottom line is energy-efficient AI inference is probably the dominating factor for these things, and we've done a pretty good job at it. If you just take a look at the performance per watt in comparisons against what we have seen against the x86 ecosystem or even the M3 ecosystem, we've done a much better job in terms of running AI inferences. So, that's what truly excites us with AI PCs.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Yeah. And actually, when you talk about the feature in this phone, I have used it.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

You have used it.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

At Hong Kong at a presentation where I wasn't connected to the Internet, and it was impressive with the people there, that I could translate Chinese. And probably the biggest use of it is when I get a haircut; my hairstylist is Vietnamese and I can pull up Vietnamese.

Copyright © 2001-2024 FactSet CallStreet, LLC

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Immediately do that.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Amazing number of languages. I'm just amazed with that because that's all on phone and there's at least 13 different languages. How – I mean, is that part of this – is it 6 billion parameters that it can handle?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Yeah. So, today, if you were to just take a ballpark number, last year's phones that were launched, all the way up to the Galaxy S24, they can handle up to like 7 billion, 8 billion parameter models very smoothly. And what we do, by the way, is not just take the model as is, but we do a lot of work as Qualcomm in terms of taking that model, curating it in terms of compressing and quantizing it and making sure that it doesn't take up – if you just take an 8 billion parameter model as is and if you don't touch it and just implement it as is, it takes on something like 16 gigabytes of memory; that's clearly not feasible. So, you have to do a lot of work to actually make sure it runs efficiently, it's quantized and it's compressed; we do that work. But ballpark, these days in smartphones, you can go up to 7 billion, 8 billion parameters without breaking a sweat. By the time you get into the compute segment, we go up to like 13-plus billion parameter models.

And that's kind of an important trendline because we're also noticing something else and you might not be fully aware of it. And that is today's smaller models outperform last year's larger models by a large factor. As an example, if you take today, 2024, if you take today's 8 billion parameter, it's called Llama 3 from Meta, actually turns out it outperforms yesterday's, last year's Llama 2, the second generation, 70 million parameter model. So, it's a significant reduction in the number of parameters but the performance is like exactly like last year's 10x higher, which means you can build more and more sophisticated use cases and rich use cases in devices like smartphones and PCs. And that's what is actually the truly exciting journey in this whole thing.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

So, is that – they optimize them, is that how they be able to get?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

It's combination of things. I mean, there's work that Meta has done in terms of – and practically the entire industry, in terms of just the amount of sheer amount of training data that is used in these models. So, there is work that they do, but we've also been improving things in a different way in terms of making sure that it fits in even more efficiently into smartphones and PCs, occupying a smaller memory footprint, and that actually makes the big difference by the time this gets implemented. But effectively, the use cases are going to get richer as we move forward. And it's all about on device AI, by the way. And there's lots of reasons for that, of course. But one of the things that you mention even in the absence of Internet connectivity that is key, imagine let's say that you and I get used to the new normal of using Copilot+, basically using on-device AI for all sorts of bread and butter stuff that we do in productivity.

And then you just happened to be in, I don't know, a transatlantic flight or something, you still want to get some work done. Well, do you go back to your old ways or you say, no, I'm just going to go with the way I am right now

Copyright © 2001-2024 FactSet CallStreet, LLC

and with Copilot+ and basically on-device AI. It doesn't matter whether you have the best-in-class connectivity or not. They're like two different conversations altogether. You can always have the same experience regardless of your connectivity.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.
Q

Yeah. That's great. Yeah. And I know in October you're having your Snapdragon conference, but can you give us a hint of what comes next, you're at Generation 3, the Snapdragon 8 Gen 3. Then Gen 4, from the leaks and things I've previewed, it sounds like, it's at least a 30% performance increase...

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.
A

Well, I would say that if you are listening in, you really should book your tickets to Hawaii. First of all, that statement by itself, it should always be true. I mean, you should never hesitate about that. But seriously, in terms of the kinds of things that we are bringing to the table, stay tuned on that front. Lot of things that we are doing in the context of generative AI, bringing in new features, new capabilities, new partners and whatnot. So, I'm going to stop it at that. It's just a teaser, I understand, but I'm going to stop it at that. But October 23, mark your calendars for that one. We're about two months away.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.
Q

I'll get my suntan oil ready. And, maybe we talked a lot about the CPU. You talk about the, the three processors in there. And first, within that, which one – is it the NPU that's having the biggest performance increase and maybe taking more of the silicon?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.
A

Yeah. So, let me actually – so, the way that we think about it is that any of these devices typically have these three processes, but we are increasingly focused on what we call as heterogeneous computing. I'll explain it in the following way. First, we're going to start with the CPU. The CPU has – it has a day job in itself. It does a lot of things within your device. And, opportunistically, it can always be used for AI. I think that's the way that we think about it.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.
Q

Okay.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.
A

When we switch gears towards the GPU, the GPU is another one of those. And, by the way, we have the best-in-class CPU at this point in time. And then – and as we talked about it last year during our unveiling of the Oryon family of processors, the CPU processors, which are benchmarks so that in a single threaded performance we were better than the competition, every other processor that was out there, including the comparisons that we did against the M class of processors, not just x86. But then when you take a look at the GPU, the GPU has a different kind of a day job in terms of graphics rendering, whether it's gaming, but it can also be used opportunistically for AI. The third processor is our NPU. The NPU is custom designed for AI. AI inference is its day

Copyright © 2001-2024 FactSet CallStreet, LLC

job, and so it's designed just for that, extremely energy-efficient, low-power, high-performance computing all the time.

So at any given point in time whenever there is an AI-based use case, it's a combination of different models, concurrences, and whatnot, use cases stitched with a combination of things that are coming – that come together. So depending upon exactly what's happening in that laptop or in that phone, we can map it only to the NPU or only to the GPU or the CPU or a combination of these three depending upon how things go. So that's the general vision of heterogeneous computing. We believe it's the right way of actually thinking about it. But it works very well for Qualcomm because as I said, we have the best-in-class CPU, but we also have the best-in-class NPU in terms of the performance per watt, its industry leading and the GPU has always done extremely well in the way that we have actually positioned it. So, what we feel is that we have a good diversified portfolio of processors on which we can run AI workloads.

But one thing that we are doing in Qualcomm quite a lot these days, and hopefully I'll talk a little bit more through that, is focusing on the software toolchain making it very easy for any kind of a third-party developer to go from ground zero to building an app very quickly. They don't need to necessarily understand all the nuances of the CPU, GPU, and NPU. Just have the ability to interface with it at the right way because the objective is to make it easy to come out with a very rich and sophisticated set of generative AI apps and trusted to the ecosystem to build those apps. But we provide the platform along with the software around them.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Right. And I guess tying into that for the full system, the – you're still using low power DDR5 as a memory bank. Does that have to grow with all these new functions?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Yeah. That is an important question because one thing that does come up quite a lot is as the computational power increases dramatically in devices, where is the next bottleneck, where is the next choke point over here. And depending on the kind of device, very clearly, if you have an extremely large model, even after compression and quantization, you do have to pay attention to, well, what's my memory footprint and how much is my real estate that's available for me. So, it turns out that once in a while, depending on the kind of a device you might be memory bound, not really compute bound. In other words, your computation power is huge, but you just kind of store the model to begin with.

And occasionally, you go the other way around where in memories is not bottleneck, but you might actually land the computational bond, which historically shows up a lot more in the data center space compared to this. So, there's always a balancing act over here, and we always think in terms of some unique things that we do in terms of designing the memory subsystem so that in the end, if you're an end consumer and you've heard so much about AI, what you really care for is KPIs which are – actually you don't even care about performance per watt or NPU and so on. What you actually have is a different kind of a KPI. I input a prompt and what I get is an image that gets generated, how long does it take? So this is just pure latency. But if you kind of break down that problem, it turns out that, oh, you have to first take that model from Flash, then you have to load it on DRAM, then you have to generate the first token and then you kind of go from there. So, that's the breakdown of that.

So, our objective is that higher level KPI in terms of the latency is like significantly low. The lower, the better. And if you remember in Snapdragon Summit last year, we showed the image generated in something like 0.6 seconds. I mean, it was like pretty phenomenal if you're going to think about it. And we also showed where the

Copyright © 2001-2024 FactSet CallStreet, LLC

competition was compared to where we were, I mean, like really, really staggering in that front. The other part is the other KPIs that we really focus on is power consumption. Power matters, especially if you have AI inference constantly running all the time, 24/7 in the background, not just demand-based, as in you input a question and you get a response. Well, then you better be extremely energy efficient, and that's a place that we've done well. That's where the memory bond doesn't necessarily come in. It's not a memory question, but it's more about the processor and power consumption question.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Okay. And then, the other amazing, what you even call a processor in there is the modem...

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Yes.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

...5G modem. And maybe you can say the advancements you've had on how many generations, I've lost track, if it's 9 or 10 generations of modems now.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Well, the modems, in general, one thing that's happened is that, of course, we are in a – the original 5G launches were in 2019 and we've added – we are now well into the 5G Advanced era. But just in terms of the sheer number of features that have added, it's like every year there's like 20 features that keep getting added. So, it's like – and keep in mind, you still have a lot of 4G, and before that all of 3G and all the other modems and so on. But what I wanted to mention is that in the context of AI, actually we are in the third generation, actually, just earlier this year, the third generation of modem AI processes. That's right, we've actually fully baked in AI processing even into our modems now because we are able to extract – even though we've been doing modems for a very long – for decades now, turns out that as we started bringing in some of the learnings from AI and we brought in our own capabilities with – this is like using AI for modem processing, we're able to extract even more performance and power consumption benefits from there.

So, AI has now permeated into even our modem RF technology at this point in time. And we are already in our third generation over there, so pretty happy about that front.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Yeah. That's something – again, going back to this phone, I'm at 80% now. I've been at it on since 5:30 this morning. I think people get envious of that when I show off. So, that's part of the modem, right, because my modem is on all day long and it's just not burning up that much power.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

That's right.

Copyright © 2001-2024 FactSet CallStreet, LLC

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

And another key feature to the Snapdragon platform is – and Qualcomm's strategy for moving into other markets like automotive. And I know it was mentioned before, the dashboard and ADAS applications. Can you say how AI is playing into that as you move this along?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Yeah. So, in Qualcomm, I think it was in 2021, in our Investor Day in New York, we kind of outlined – Cristiano outlined the plan, his vision for how we want to diversify our business into all other businesses beyond just handsets. So, we talked about compute. But if I want to complete the picture over there, there's the compute segment. There's the XR segment; XR is a combination of AR and VR and some of the mixed reality portions as well. Consumer IoT and industrial IoT. Automotive, and within automotive, what used to be purely telematics, which was just the connectivity portion of it, we've expanded significantly from there into what's called as the digital cockpit which is just your user experience inside the vehicle. And, of course, ADAS, the autonomous driving stack that is needed over there.

Suffice it to say that in this entire portfolio of products, our technologies – so, I'm responsible for our technology roadmap for all of the products over here – we designed the technology in such a way that it gets very easily leveraged across all the different product lines. So, if you think of this is what I can do on a smartphone, rest assured the capabilities are approximately the same, if not even higher on compute and in automotive products, and maybe slightly lower by the time we get into the consumer IoT and maybe slightly higher when you're in industrial IoT. Why is that? Because everything that we design from an SoC standpoint, it leverages the IPs that are developed. And it's the same IP that gets leveraged, but we are very conscious of the fact that we're not simply reusing it; it's very important to understand that. It's not just about the reuse from a smartphone because it works, by the way, in a lot of ways.

But then, we have to do a little bit more to make sure that it fits in within the product segment, so let me give an example of that. If you take a look at – let's take automotive, for instance. In automotive, some of the workloads are even higher compared to what we run in a smartphone just because of the nature of things over there. And, by the way, it can only run in the car and you won't necessarily go to the cloud over there. It has to run on device. But we're talking of larger models, higher computational power just in terms of the throwing in multiple cores of the same processor depending upon what we are talking of specifically over there. But then when you bring it down from there onwards, it turns out that the workloads are also not that different sometimes. So, you could be talking to your phone and what you have is a voice AI assistant, which is now increasingly voice is now the more natural UI towards your phone.

Well, voice also turns out to be the more natural UI as you are in that digital cockpit regime. For instance, instead of – let's say that, for one reason or the other, there used to be a point in time that you would have a physical manual of your car somewhere in the glove compartment and you're doing – if you have a flat or something, if you want to see what – you pull it out from there. Increasingly these days, you don't have that. What you have is maybe a PDF version of that document which is there on the screen in front of you. But even then, you have to kind of search through it. Well, imagine the next step over there where you're simply talking to – and now you have a voice assistant and say, I just got a flat, tell me what to do. And things pop up and say, okay, this is what you need to do, go here and do this, A, B, and C, and so. And so, the same kind of the user experience is gradually coming into the digital corporate.

Copyright © 2001-2024 FactSet CallStreet, LLC

It could go all the way from roll down the driver's side window. I just talk about it and sort of me trying to do something over there. These are all actually generative AI based use cases. So, the leverage from what we do in smartphones, it's not such a gigantic leap to go from something like a smartphone towards even automotive. Now, let's talk about IoT. In consumer IoT, there's a lot of instances where the AI use cases, they leverage what we do in smartphones, but then take a life of their own in a very different way. As an example, you might have a doorbell, just a completely different thing, a doorbell where – a doorbell used to be just that and nothing more, you just ring the door. But then increasingly you have like a camera on top of it and you reach a point in time where maybe the camera does face recognition. That's, by the way, AI capability over there. It's not generative AI, its classical AI, recognizes your face and actually is linked to your door and therefore opens the door.

So, it's like you don't need your key and you have to do a lot of things. But there are places where exactly the same kind of concepts, they take a life of their own and then mark themselves into a different way to these categories. So, my job is to make sure that as Qualcomm, we have the right leverage across all the product lines and we've done a pretty good job in terms of that. Ever since Cristiano actually has been emphasizing that this is the direction we're going to take, we've done a pretty good job as Qualcomm in terms of leveraging that technology and building on top of it. So, it's not just reuse of it, its leverage and then building on top of it. We just talked about compute segment, by the way, and in the compute segment, it's exactly the same thing. The processor lineup that we had, we beefed up in a very different way and now brought in the gen AI capabilities into compute as well. So, that's like a big thing for us.

And I do want to talk a little bit about software, but that's kind of the big picture in terms of how to leverage us.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Okay. Yeah. I was just going to make the point of, everyone knows about Gordon Moore and his law, but there is also a Gordon Bell that worked at DEC and he had a – his Bell's law is the simpler the interface is to the machine, the volumes go up exponentially. So, like when we got the mouse, that made PCs grow. You got touchscreens on handsets, volumes went into the billions. And now you have natural language. So, IoT capabilities that just the volumes are going to be tremendous.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Absolutely. And one of the – it was – and I want to maybe say that because this was a demo that we showed earlier in Barcelona, Mobile World Congress, just to illustrate what else is possible using generative AI. And just ways as a UI towards your phone. The demo was a one in which we – a person takes a phone and is pointing the phone towards a bunch of vegetables on a kitchen table and says, tell me a recipe that I can have based upon what you see. Now, if you're going to think about it, it's a very complex question. First of all, there's a camera that's being used. And based upon that, you have to figure out what is on the table to begin with. That's just like pure classification. What is what, which part is meat, which part is like vegetables and so on. Then you have the second thing, you have to understand the question. The question is, a recipe based upon what you see.

And that means there is a voice which you have to understand exactly what was stated. And the third part is, generate a recipe which might not have existed before. Because this could be any combination of vegetables and so on. And then actually out pops a recipe which I thought was pretty good. I've kind of tested it a little bit. I know it's a little bit of fun in that, but it does actually make a point. It's so easy. When you start sharing that interface, it is so easy to do these things. It also points to something else, which I think is going on behind the scenes as we kind of move forward. We grew up in an app-centric world in the last 15 years or so. But as voice becomes the natural UI, we are no longer navigating through a clutter of apps on our phone and saying, I'm going to click on

Copyright © 2001-2024 FactSet CallStreet, LLC

this and then start typing in something. If I'm just talking to it, the app still exists. It's there hidden somewhere else. Your UI to that is so much easier and it becomes clear to your point on the volumes increasing.

The number of people who can get more work done using the same devices increases exponentially. A lot of us who are probably in the tech world, we understand. Even then, I think sometimes it's like do you really understand every single feature on your phone or every single feature on your laptop? Maybe about 80% there or so. But imagine someone who's never used it at all and you're just using voice, so much easier from that point onwards. The smartphones and the AI PCs and so on, I think we made the devices very sophisticated, but the people are where they are. Now, we are changing the UI to those devices, that changes things dramatically.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Yeah. Maybe I'll let you – we have a question from the audience, but first, you want to talk about software so I'll...

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Yes.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

...first talk about it.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

The reason I wanted to talk about software is that, it's a part of us as a transformation of the company. So we talk about the fact that we gradually went from, I talked about 5G and modems very little and I talk a lot more about AI and processes and computing. We have been on this journey of transformation of Qualcomm from a communication-centric company to a connectivity plus high performance computing company. That's one of the transitions that actually, we've done a pretty good job in terms of doing the transition well over the past few years. The second part that we've actually kind of truly done when it comes to software is twofold. One, software tools and SDK is to make sure that developers can use our platforms in a much more intuitive way, and they don't have like a massive hill to climb to do that. And the second part is embracing open source software as and when possible, so open source Linux and where that comes in.

Now, let me start on the first part, specifically in the context of AI. Earlier this year, we introduced the Qualcomm AI Hub. The Qualcomm AI Hub is a place where – and by the way, it has mirror site, which means you come to aihub.qualcomm.com as a developer, or you can go to Hugging Face where there's like 800-plus thousand AI models and it's also on GitHub. Step one, you come in there, you're an app developer. You could be a teenager in somewhere in Southeast Asia and you want to actually build an app. Now, you live in that world. You don't understand any of the details about the NPU, CPU and so on. But you know PyTorch very well, and you understand AI reasonably well. And you're just in the business of building an app. What do I do today? Well, as you come to the AI Hub, the first thing is there's a catalog of models. You pick one, because your use case might be text to image or text to text or text plus voice to something else.

You pick a model from there. That model is highly curated. Remember what I said earlier about taking a model and compressing and quantizing and making sure it runs very efficiently on Snapdragon platforms, that's the model that's available over there. You pick from that. Second, you write your code, you build it into your app. And

Copyright © 2001-2024 FactSet CallStreet, LLC

you're a teenager, you might not be able to afford that device that you want to test it on. No problem. As Qualcomm, we give you cloud native access to that device out there to go ahead and test it. You can come to the hub. You can then test it. Its cloud native, so there's a device form on which this is being tested. As far as the developer is concerned. You don't have to worry about it. But it's a physical device. It's not virtual there. It's a physical device. Once you do the testing, you're ready to go, take an app store of your choice, go ahead and deploy. Done.

I'll give you one example of this. Just recently, last week, one of the developers, they went through this exact exercise. They've also modified the – you don't have to only pick a model from the catalog. You can also bring your own AI model and you can build your app on top of it. So, they brought in their own AI models, this one specific app developer. In a span of one month, they went from ground zero to 1 million downloads, and they're averaging something like 100,000 downloads per day, per day, all optimized and running very efficient. It just went viral, all optimized and running very efficiently on Qualcomm platforms.

Kevin Cassidy Q
Analyst, Rosenblatt Securities, Inc.

Wow.

Durga Malladi A
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

So, it's – for the first time, we are kind of fully integrating the software with the hardware in such a way that once you like this, it works extremely well on Snapdragon platforms. How does it work on other platforms? That's really up to the developer what they want to do. But on our platforms, it runs very efficiently and they're getting a lot of traction on that. So, software-first mentality has been fully brought into Qualcomm in that sense. And as I talk about this, there's open source software which is very important by the time we get into IoT space. But this is key for us. We can build all the hardware processors and everything that we want, but investing in our software protocol stack and tools has proven to be very successful for us.

Kevin Cassidy Q
Analyst, Rosenblatt Securities, Inc.

Yeah. That was – the number you said, 800,000-plus?

Durga Malladi A
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

800,000 plus models are on Hugging Face.

Kevin Cassidy Q
Analyst, Rosenblatt Securities, Inc.

Yeah.

Durga Malladi A
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

But on this specific app developer, they brought their own model.

Kevin Cassidy Q
Analyst, Rosenblatt Securities, Inc.

Copyright © 2001-2024 FactSet CallStreet, LLC

Yeah.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

And then, they did – they went through the whole optimization and everything that they did on AI Hub and they could test it very efficiently. And they could go ahead and deploy, and next thing you know they're averaging 100,000 downloads per day. I asked the CEO this question: Hey, tell me, let's say we didn't exist. Let's say that the AI Hub didn't exist. What would have happened? And the answer that we got was no, this wouldn't have been possible because we were able to fine-tune the performance to a point where the use case actually made sense. And so, it's not just the fact that we could test our app on the device front, but we could actually use the tools that you gave. You actually helped us go through that entire exercise. Pretty big moment for us.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Yeah, that's tremendous. I'll ask a question from the audience that says can you speak to the implications for AI demand if battery – modem battery performance increases? Can you update on any work that Qualcomm is doing with next-generation silicon battery suppliers? Typically, 100% silicon batteries work on 2.7 volts. Do you currently plan to support these voltage levels?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

That's an interesting question by itself. I think the way – maybe I'll answer the question in the following way. Our focus is quite heavy on optimizing the performance per watt, which means that we focus on energy efficiency. The reason we have to do it like that, regardless of the actual capacity of the battery in itself, is because the device portfolio that I mentioned can vary dramatically from by the time you go into doorbells. There, you're extremely limited in terms of your battery capacity versus these days, if you pick up a smartphone, you can be anywhere between 4,000 to 5,000 milliampere-hours; and on laptops, it's a completely different ballgame. So, yes, we pay attention to what the battery capacity is and in terms of what other things that we need to enable for them.

But a lot of our emphasis really is no matter what the real estate has provided to us, our job is to extract the last ounce of performance from there, so that's the performance for what we really focus on.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

Okay, great. And maybe we're down to 5 minutes left. But for the AI PC, you've had an exclusive relationship with Microsoft. And next year that opens up, I would expect you're going to have other ARM-based CPUs as a competitor. What have you learned, I guess, working with Microsoft that you probably have a head start on anybody else just coming into the market?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

So, first things first over here. I think it's been an extremely useful journey. There's lots of learnings that we had together jointly with Microsoft. There were a lot of things we didn't know what we didn't know when we started this journey. The learnings and the lessons learned were actually super-valuable for us. I mean, by the time you take a look at the launches that occurred this year, the sheer number of models, the amount of software that they had to do, that we had to do, it was like joined at the hip in terms of the development of this entire Copilot+ experience that we were talking of. So, as we go into the next generation of products, I mean, there's a lot of things that we're

Copyright © 2001-2024 FactSet CallStreet, LLC

building on on top of it. Personally, I don't pay attention to what the competition might do or not. This is a world in which you have to expect the very best from the competition, nothing less than that.

But we focus on pushing ourselves extensively, going from where we today to where we need to be. That's the bar that we set up on ourselves and that's the way that we think about it. It's no different from some of the other domains that we think of in terms of the competition. We pay attention to it but that's not the driving factor. I think we have our own goals in terms of where we want to be. And moving forward, we believe that pervasive AI, AI that's constantly running in the background is going to be increasingly important because people – just based on the kinds of use cases that are coming up, you have to have a very energy-efficient high-performance computing platform. Energy efficiency is going to be key. That's the only way that you can actually enable these pervasive AI workloads constantly running in the background. So, our focus is very heavy on that part.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.
Q

Excuse me. Yeah. Maybe if you could touch in on the AR/VR. And I think people are surprised that even the glasses are getting to be even more important than the headsets, I would say, and the goggles. But you're in all of these. So...

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.
A

Yeah.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.
Q

...maybe talk about how that market has developed?

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.
A

Absolutely. I mean, I have my Ray-Ban glasses which are pretty cool actually. These are – if you just think of all the technology that goes into that based upon what we have done, there's a camera. There's a microphone. You can take pictures from it. You can talk through it. AR workloads are actually getting pretty nice in these form factor and pretty sleek smart glasses. We are pretty bullish on that working with our partners in this space. But why is AR becoming important on that? Actually, there's a couple of things. One is that if you just think in terms of a lot of the image processing, image rendering geometric processing that is done in these glasses, actually it turns out that it's far more stringent from a KPI standpoint. The latency bounds in terms of it's called as the motion to photon latency is in the order of like a few milliseconds in terms of how you quickly render, especially by the time we get into VR.

And for AR, just in terms of the augmentation, I mean there's a lot of places where you wear your glasses, you're just looking in a certain direction and say, tell me what I'm looking at. Well, now at that point in time, there's a way in terms of you do the process because there's a question. That's a generative AI problem out there. There's a limited amount of horsepower that you can do on the glass by itself. But if you have an accompanying device, you actually have a way in terms of the inference workload gets transferred to that other device and you can actually go back and forth between your glasses and the device. There is a future at some point in time when we'll be able to do all of that just in the glasses by itself. But today, a majority of it is still based upon outsourcing, a little bit of that towards the device that you have in hand in addition to the glasses.

Copyright © 2001-2024 FactSet CallStreet, LLC

But already taking pictures and doing voice calls and any other kind of calls and so on using the smart glasses is here to stay, and I personally like those workloads quite well.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Q

You know, it just maybe not too much of a coincidence but we had the CTO from Adeia yesterday. You know, so they're writing software or intellectual property, and he used the idea of that – like you had mentioned that you don't have to use the apps on your phone. But he's saying the glasses would know that you're at the airport and you're getting off the airplane and call the Uber for you. So, the advancements of all these things being intertwined.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

A

Perfect. Yeah.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Yeah. So, we're just about out of time. Maybe if you want to make any conclusions. It's so exciting to be able to talk to you about all these different things you're working on.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

Yeah. I mean, I was just going to say that hopefully I gave a glimpse into where we're heading with AI. One final note that I wanted to mention is that as we are increasingly going down this path, there's -as a technology, AI is evolving quite rapidly and we're always going – we're on a mission to be always ahead of the curve in terms of what to expect and therefore bake it into our product lines as we move forward. And increasingly, we are also focusing a lot more on the personalization aspect of AI, trying to bring in, like a lot of the things that we do are very personal in nature. And so, that makes it even more sophisticated, actually. The way that we build the software within the products that we offer. So, hopefully, I gave you a glimpse of where things were, where things are today with AI. And what we are doing as we move forward and how we are able to leverage the investments across the entire portfolio of products.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

Great. Thank you very much for your time, Durga. And yeah, I'll book my flight to Honolulu.

Durga Malladi
Senior Vice President & General Manager-Technology Planning & Edge Solutions, Qualcomm Technologies, Inc.

All right. Thanks, Kevin.

Kevin Cassidy
Analyst, Rosenblatt Securities, Inc.

All right. Thank you again.

Copyright © 2001-2024 FactSet CallStreet, LLC

Disclaimer
The information herein is based on sources we believe to be reliable but is not guaranteed by us and does not purport to be a complete or error-free statement or summary of the available data. As such, we do not warrant, endorse or guarantee the completeness, accuracy, integrity, or timeliness of the information. You must evaluate, and bear all risks associated with, the use of any information provided hereunder, including any reliance on the accuracy, completeness, safety or usefulness of such information. This information is not intended to be used as the primary basis of investment decisions. It should not be construed as advice designed to meet the particular investment needs of any investor. This report is published solely for information purposes, and is not to be construed as financial or other advice or as an offer to sell or the solicitation of an offer to buy any security in any state where such an offer or solicitation would be illegal. Any information expressed herein on this date is subject to change without notice. Any opinions or assertions contained in this information do not represent the opinions or beliefs of FactSet CallStreet, LLC. FactSet CallStreet, LLC, or one or more of its employees, including the writer of this report, may have a position in any of the securities discussed herein.

THE INFORMATION PROVIDED TO YOU HEREUNDER IS PROVIDED "AS IS," AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, FactSet CallStreet, LLC AND ITS LICENSORS, BUSINESS ASSOCIATES AND SUPPLIERS DISCLAIM ALL WARRANTIES WITH RESPECT TO THE SAME, EXPRESS, IMPLIED AND STATUTORY, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, ACCURACY, COMPLETENESS, AND NON-INFRINGEMENT. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, NEITHER FACTSET CALLSTREET, LLC NOR ITS OFFICERS, MEMBERS, DIRECTORS, PARTNERS, AFFILIATES, BUSINESS ASSOCIATES, LICENSORS OR SUPPLIERS WILL BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR PUNITIVE DAMAGES, INCLUDING WITHOUT LIMITATION DAMAGES FOR LOST PROFITS OR REVENUES, GOODWILL, WORK STOPPAGE, SECURITY BREACHES, VIRUSES, COMPUTER FAILURE OR MALFUNCTION, USE, DATA OR OTHER INTANGIBLE LOSSES OR COMMERCIAL DAMAGES, EVEN IF ANY OF SUCH PARTIES IS ADVISED OF THE POSSIBILITY OF SUCH LOSSES, ARISING UNDER OR IN CONNECTION WITH THE INFORMATION PROVIDED HEREIN OR ANY OTHER SUBJECT MATTER HEREOF.

The contents and appearance of this report are Copyrighted FactSet CallStreet, LLC 2024 CallStreet and FactSet CallStreet, LLC are trademarks and service marks of FactSet CallStreet, LLC. All other trademarks mentioned are trademarks of their respective companies. All rights reserved.

Copyright © 2001-2024 FactSet CallStreet, LLC