S1 EP38 - Tackling AI Infrastructure's Challenges

Thursday, August 03, 2023 · 13:04

On this week's episode, catch Nigel Alvares, Vice President of Marketing and Business Planning and podcast host Chris Banuelos, discussing Marvell's approach to the AI revolution. Join the discussion to learn about how Marvell is poised to enable scalable solutions for the next generation of AI infrastructure. Hear about Marvell's position in the industry, the role of optical connectivity, major challenges operators of AI training networks face, and what the industry can expect from Marvell. In case you missed it, read our latest blog Scaling AI Infrastructure with High-Speed Optical Connectivity: https://bit.ly/3Dly5nb

Speakers

Nigel Alvares
VP of Marketing and Business Planning

Host

Christopher Banuelos
Senior Manager of Global Social Media Marketing

Christopher Banuelos 00:04
Welcome to the Marvell Essential Technology Podcast. I'm your host Chris Banuelos. On today's episode join me and Nigel Alvares, Vice President of Marketing and Business Planning, discussing Marvell's approach to the AI revolution. Hear about Marvell's position in the industry, the role of optical connectivity, the major challenges operators of AI training networks face and what the industry can expect from Marvell. To stay up to date on future episodes, please be sure to subscribe to the Marvell Essential Technology Podcast. Nigel, it's great to have you on today's episode looking forward to our conversation around AI. Marvell reported in its last earnings call the over 200 million in revenue is related to AI. How has Marvell positioned itself to achieve this?

Nigel Alvares 01:01
Yeah, and let me just be clear that 200 million was the revenue we achieved from AI in our last fiscal year. So that was the February of 2022, to end of January 2023. And on our earnings call, we stated that we expect our overall AI revenues for this fiscal year, which started at the beginning of February, would double again in this fiscal year, so expecting to be over 400 million this year. And this came about in a very purposeful manner. When we pivoted the company in 2016, when Matt Murphy took over as CEO, he with the leadership team, set out a strategy to focus on data infrastructure. And since that day, we've been successfully executing that strategy by forming the right team, acquiring and developing the right technologies, developing the right customer relationships to lead in this market. And now with AI becoming the ultimate data infrastructure application, Marvell is at the center of this incredible transformation. And we're super, super thrilled to be collaborating with all the leading ecosystem players be it in the cloud be it on the semiconductors be it on the software ecosystem, to enable the proliferation of AI in our day to day applications. And just to be clear, this strategy started out with acquiring and developing internally the right technologies. And with the set of acquisitions and organic investments, we've built the industry's leading data infrastructure portfolio, and that has positioned us to be a leader in cloud infrastructure and because cloud infrastructure is enabling the AI era, becoming a leader in AI infrastructure.

Christopher Banuelos 03:12
Nigel, let's talk about optical connectivity. What role does it play in AI infrastructure?

Nigel Alvares 03:19
Absolutely. I think this is a really important area to discuss. When people in the industry talk about AI, all you hear about is the compute elements GPU this, TPU this. So that's where almost all the attention is. And unfortunately, compute alone cannot satisfy the needs of developing or delivering AI applications. You need compute, in combination with conductivity and storage. And what I mean by that is, as the data workloads or the data sets for AI continue to get larger, more complex, more diverse, you need to cluster these compute elements together to process those data sets and come up with the models. Specifically, you need these compute elements, GPUs, TPUs, AI accelerators, to communicate to one another. And just to give you a feeling for how much connectivity is required today, in a cloud server application, your traditional application server, has two CPUs, so two compute elements, and the networking bandwidth coming out of that is between 100 gigabits per second and maybe 200 gigabits per second, right. That's your typical use case bandwidth needed for a traditional application server. For a GPU or a TPU or an AI accelerator type of cluster you would need at least 300x that. Let me just put that in context, one GPU today a leading edge GPU's about 3.6 terabits per second of bandwidth. While I just said, the most common compute element today is about 200 gigabits per second. So just take that multiple, right you get a 16x plus or almost 20x. But then you're clustering multiple GPUs or accelerators together, typically anywhere from 8 to 10. So multiply that by 8 to 10 and then you get this 200 to 300x multiplier. So the bandwidth required for an AI cluster is 300x of a compute server bandwidth. So that's where optical connectivity comes into play, because you need to put these GPUs or connect these GPUs together, or accelerators together, using the highest and fastest speeds to get that bandwidth in and out of the compute elements.

Christopher Banuelos 06:04
Nigel, my next question for you is what big challenges do operators at AI training networks face and how does Marvell help them overcome those challenges?

Nigel Alvares 06:13
The cloud infrastructure operators developing or providing the AI infrastructure are facing significant challenges. And if I had to categorize the top three, from my perspective, I would say number one is how do you continue scaling this infrastructure in an efficient manner? Meaning, keeping the power and performance at the right cadence, right? You can't, you can't increase performance 2x but 10x your power, right? So how do they do this in an effective way? That would be the number one item and challenge that the operators need to sort through. Number two would be the cost related to the services. Specifically, they need to address how they deliver these AI services in an efficient manner and not continue to grow the services, but eat into their margin models that they've currently built for their infrastructure business. And then the third one would be related to performance. How do they ensure that as they continue to scale their business, that performance is continuing to deliver these AI results via training or inference in a time efficient manner? Because every minute that passes by, or every second that passes by, is cost to them, as well as to their customer. So those are the three main challenges. And it really all comes back to how do you scale the infrastructure right, and make sure it's done in an efficient manner from a power performance perspective, and keeping the total cost down. So how is Marvell collaborating with them to overcome those challenges? It all centers around our portfolio and optimizing it in collaboration with each of the cloud operators. As I've highlighted or mentioned, in the past, every cloud is unique. So they have infrastructure that needs to be optimized for their specific environments and infrastructure. And that means they would have to customize solutions from compute elements. So really, instead of using something that standard off the shelf, optimize and customize it for their infrastructure. And that's what Marvell is doing in collaboration with the leading cloud operators. And then it's integrating at a system level, all the other different pieces, be it memory, storage, and networking with optical connectivity. And that is how Marvell today is actually working with these cloud operators and delivering AI services is we're

providing the optical interconnectivity to help them complete the training jobs in the most efficient manner by giving them the biggest pipes to interconnect between all the different compute elements in an AI cluster. And now we're working with them to continue growing that with not only scaling the optical interconnects, but also enabling optimized Ethernet switches to deliver the lowest latency meaning finish the job in the quickest time. And that is where we're working very closely with them to overcome the challenges of cost and performance. So at the end of the day, Marvell's complete portfolio provides them the opportunity to customize and optimize for their environments.

Christopher Banuelos 09:56
Nigel, my last question for you today is as the AI market expands, what can the industry expect from Marvell?

Nigel Alvares 10:05
The AI industry or the AI market is in its infancy, right. We're really in the first inning of this evolution or this new era. And what Marvell is uniquely positioned to enable and address is the spectrum of technologies and product areas needed for that expansion or proliferation. What that means is if you take a step back data infrastructure really consists of compute, networking, storage, memory, and optical interconnectivity, which you could tie it to networking, that portfolio, Marvell, is the only company out there that has all those pieces. So what the industry should expect, and we're collaborating with all the leading innovators in this space is cloud optimized solutions to help this industry proliferate and expand. If you look at cloud infrastructure, today, there's maybe a half a dozen companies, maybe up to 10 companies that are leading the charge. And each of those cloud companies or cloud data center operators are relatively unique in how they go about building their infrastructure. And the reason being is each of these cloud operators is relatively unique. They have a core business, you know, be it social media, be it search, be it infrastructure as a service. Each of these different areas requires different types of technologies and infrastructure to really deliver it at scale in an efficient manner. And that's where Marvell, given our business models and given our technology portfolio, puts us in a unique position to help these companies continue to scale their services and applications in a unique and differentiating manner, that many others do not have either the business models, or the portfolio to address. And that's what makes Marvell relatively unique in this market, and excites us at Marvell quite a bit as we're just in the early innings of this major technology disruption and era of artificial intelligence.

Christopher Banuelos 12:31
Nigel, just want to say thank you for participating on today's podcast, really excited about this topic, and I look forward to hearing more.

Nigel Alvares 12:38
Thank you, Chris. It was my pleasure. I think we're in an exciting time right now and I'm thrilled I was able to share some of my thoughts with you and the community.

Christopher Banuelos 12:48
Thank you for listening to the Marvell Essential Technology Podcast. As always, please feel free to visit our website to learn more, and we'll see you on the next episode.