Research @ Citi Podcast, Episode 73: A Potential Solution for AI's Memory Crunch?

Recorded: April 7, 2026

Published: April 9, 2026

Host: Rob Rowe, U.S. Regional Director of Research, Citi

Guest: Peter Lee, Semiconductor and IT Hardware Analyst, Cit

Rob Rowe (00:02)

Hi everyone. I'm Rob Rowe, U.S. Regional Director of Research. Welcome to the Research @ Citi podcast. Today I have on the podcast with me Peter Lee, who is our Head of Semiconductor and IT Hardware Analysis at Citi Research.

And our topic today is an interesting one: There's been a recent introduction of TurboQuant, and we're going to discuss how this new technology may have profound effects on memory and storage demand.

Welcome, Peter, and thanks for joining us.

Peter Lee (00:28)

Hey, thank you Rob.

Rob Rowe (00:30)

This has actually taken the market a little bit by storm, hasn't it? I mean, maybe what we should do first is just talk about the introduction of TurboQuant recently by Google. What is TurboQuant, and how does it work?

Peter Lee (00:42)

TurboQuant is AI memory compression, and it was designed to solve one of the biggest bottlenecks in running the large language models. So the massive amount of memory required to store the conversation context — known as the KV cache, Key Value cache — TurboQuant allows the AI memories to run faster and handle much longer conversations on existing hardware by shrinking the size of this temporary data by roughly six times without losing accuracy.

TurboQuant works by transforming the complex data vectors into much simpler. So instead of using the standard XYZ coordinates, it converts the data into radius and angles. Because the rotation makes the angle follow as a distribution, the system doesn't need to store the extra metadata to describe how to decompress the data,

saving significant space. So that's actually a very innovative approach to reduce the memory size.

Rob Rowe (01:43)

So I know that there are some views that this is negative for the markets, or negative because it could cannibalize some memory companies, or something along those lines. How do you foresee memory demand changing? Does this reduction in KV cache enabled by TurboQuant lead to a decrease in memory demand, or do you think there's something else happening here?

Peter Lee (02:07)

I think that this kind of software efficiency isn't a threat to memory. I think it's positive for demand because it makes AI cheaper and more useful, which drives a purchasing cycle of even higher demand for the advanced chips. So I think it's like Jevons paradox. I think it will be more positive for high-performance memory like HBM and Server DDR5.

Rob Rowe (02:30)

Ah, so there's essentially a cost efficiency here as well, which I know has been a concern for AI too. Can you give us other examples of what you just mentioned, which is Jevons paradox?

Peter Lee (02:42)

Yeah, sure. Maybe the cloud computing, for example. When I was in Samsung Electronics before I joined Citi, I was in the marketing division at the time. People were actually worrying about the demand correction from cloud computing because cloud computing is more efficient. So, we don't need, maybe, PCs as much. That was concerning the market at the time in the memory industry.

But the opposite thing happens, because people start to use cloud computing more and more, and then it's more efficient. Computing power, computing cost is getting cheaper and cheaper, so people use more computing systems. So that's the example of the different products.

Rob Rowe (03:21)

You mentioned that TurboQuant is a quantization algorithm that reduces KV cache memory use by six times maybe? What is KV cache? This sounds like a rather significant innovation, am I right?

Peter Lee (03:35)

Think of KV cache as kind of bookmark for every word in a conversation. So normally when an AI agent talks to you, it has to reread everything from the very beginning to decide what the next word should be. So instead of rereading everything, the AI writes

down a summary note for every word. The AI keeps these note cards on memory. And then we call it KVK, so it's kind of short-term memory.

Rob Rowe (04:04)

Maybe I can go back, Peter, to the previous question: How do you think TurboQuant will affect the future direction of the memory market in general?

Peter Lee (04:15)

I think TurboQuant could impact the edge device. So we think that the PC smartphone, for example, could be personal AI in the future. So maybe people can use the AI with smartphone or PCs because we can shrink the model size and also shrink the KV cache more. Then we have enough room to operate the AI model or AI computation.

I think that by making memory requirements smaller, TurboQuant allows the powerful AI to run on the smartphone and laptops, instead of just massive data centers, in the mid-, long term. So this move of memory demand from a few thousand servers to billions of personal devices in the future, I think.

Rob Rowe (04:55)

Let's focus on that for a second, because I think that's a key thing that you just mentioned. So, this KV cache quantization can potentially enable a lot of edge devices to run massive models that were previously restricted to data centers. Is that right? To what extent can people start running things on edge devices?

Peter Lee (05:16)

I think the concept is very simple: We can compress the data, then we can make it smaller. Then some kind of simple AI operation could be happening in the PC and smartphones, like edge device. Maybe it could be PC or smartphone and even robots, maybe, or glasses in the future.

I think that the people could use AI stuff and then it could be more efficient. So memory is so efficient, the AI doesn't need to send your data to a cloud server. So it does all the thinking locally on your device, which is faster and more private. So maybe they could be more powerful for AI — it's big for the privacy and speed.

Rob Rowe (05:56)

Wow. So does that mean that it could also put less demand on data centers?

Peter Lee (06:01)

No, I don't think so, because they could maybe cooperate together, like AI servers could be used for high performance and high workload, maybe. But the personal device, maybe we can do some simple AI. We can operate AI models. So it could be they could help each other.

Rob Rowe (06:18)

And Peter, when we think about this innovation, what's the timeline to full deployment, do you think? Is it fully deployable now? Is it an innovation that's complete? Or is it something that still requires development?

Peter Lee (06:30)

I think it still requires development, but they are trying to work on it. In terms of personal AI, we think that could be coming in 2028. But it could be transition from AI server to personal AI in the future. I think that people are trying to use this kind of technology. But the meaningful changes, I think that could be from '28.

Rob Rowe (06:50)

And what sort of timeline do you think that would be toward, say, full deployment of this new innovation?

Peter Lee (06:56)

Maybe this year, the second half, there could be some introduction of the kind of compression technology. And then could be gradually used in '27. But maybe the edge device makers also need to prepare the new directions. And then also memory makers need to prepare for the edge device or personal AI stuff. I think that takes some time.

So the meaningful changes will come in '28, because the hardware for edge devices also needs to change to operate the models in the edge device. That's why we think that, aside from second half of this year, gradually, the meaningful changes will come in '28.

Rob Rowe (07:29)

Yeah, and I imagine there might be other adjustments to smartphones or to iPhones that would have to be made to accommodate for this kind of memory compression, no?

Peter Lee (07:41)

Yes, right. And also, we need more high-bandwidth memory, which means that we need more lanes in the road. Because to operate the AI model and AI KV cache kind of stuff, we need more IOS, which means that we require high bandwidth. So the current DRAM and NAND in the smartphones, I think that is not enough.

So, if we jump like SoCam to LPDL of six, maybe also could be the solution for the edge device to operate the AI models in the phones.

Rob Rowe (08:11)

And is there an issue where you said you could have your own personal AI on your smartphone? When I think about that phrase, it seems like it's separated from

the main AI. If you're using that, how are you using it? Separately on your phone vs. the one that's connected via the cloud?

Peter Lee (08:28)

Yeah, very good question. I think maybe that it's up to you. Maybe if you don't want to open certain information, maybe privacy information you don't want to share, then you do the models and operate the AIs in your phone or your edge device. So maybe you can choose that, but if you use more high-performance AI, then maybe you can share your information to the cloud.

So there can be some options, I think. Or you encrypt the information, maybe in a server, smartphones, your AI model, personal AI models. Encrypt the information, and then they will operate the data in the server and bring back to your PCs, smartphones, and then decompress that information. Then maybe there could be some other way to operate the models.

So I think that server and edge device is not separate. I think it'll be linked together, but you can choose the information, maybe the range, maybe that private, because some people don't want to share the information.

Rob Rowe (09:25)

Wow. I wouldn't have thought that something such as being able to have extra memory by compressing, or having this efficiency, would lead to some significant changes within the world of IT and the world of AI. But this seems to be a significant change.

Peter Lee (09:42)

I think TurboQuant is in the early stage now, but I think that will be better for cost saving. So I think that it will bring them more demand. So that's why we think that it could be possible for the AI memory demand. And also, in the future, TurboQuant could impact the edge device. So then it also will be possible for the memory demand increase from the edge device like PC and smartphone in the long term.

Rob Rowe (10:05)

So in some sense, Peter, this is more of a positive outcome, even for the memory industry, if you will.

Peter Lee (10:10)

Yes, I think there's kind of net positive for the memory industry.

Rob Rowe (10:14)

Peter, thanks again for being on the podcast. This is obviously a very intriguing innovation that we're going to have to keep an eye on going forward.

This podcast was recorded on April 7, 2026. Be sure to join us for our next Research @ Citi podcast with Xiangrong Yu, Citi's Chief China Economist, and Alicia Yap, the head of our Pan-Asia internet research at Citi.

Feel free to explore our library of previous podcasts and multiple subjects, all of which are available on this channel and other channels as well. And also, be on the lookout for our Research @ Citi Markets Edition podcast, with its 10-minute breakdown of the equity and global macro markets each and every week.

Disclaimer (10:59)