## Episode 5

## Compute-in-Memory Revolution—GSI Technology's Breakthrough

[speaker1]: Welcome everyone to Episode 5 of our podcast where we dive deep into the innovations that are reshaping our digital world. I'm Allison, and with me as always is my co-host David. Today we're talking about something that could fundamentally change how we think about computing - GSI Technology's Associative Processing Unit and their breakthrough in true Compute-in-Memory technology.

[speaker2]: Thanks Allison! I have to say, I'm genuinely excited about this topic. We've been hearing the term "compute-in-memory" thrown around a lot lately, but what GSI Technology has developed represents something truly different - a paradigm shift from the type of computing that is decades old, is still being used, and a shift that could solve some of the biggest bottlenecks facing AI and high-performance computing today.

[speaker1]: Absolutely! But David, let's start with the fundamentals. When most people hear "compute-in-memory," they might think it just means putting processors closer to memory. But GSI's APU is doing something fundamentally different, isn't it?

[speaker2]: Exactly, Allison! Most processors, think GPUs, CPUs, and Google's TPUs use an architecture called von Neumann that's been in use since the last century. These devices all have separate compute blocks and memory. In these you move the data you want to work on into memory, then from memory to specific individual compute units, and then move the result back to memory. You may repeat this action several times because you need to use different compute elements for different functions.

[speaker1]: This sounds like a lot of moving data and not so much processing data. Also, not all of those different blocks are used all of the time.

[speaker2]: These are definitely problems, Allison. First, the programmer has to decide which of all the compute elements they need to use. If they want to operate on 64 bits, they

won’t be using the 32-bit elements, for instance. Then they need to constantly move data from memory on the chip and between compute elements they want to do the different processes. This leads to high power consumption just from data movement, but also only provides about 50% whole chip compute utilization for workloads.

[speaker1]: I can see how this old architecture was useful when software was more dependent on hardware computation elements and less on processing massive amounts of data.

[speaker2]: Yes, and Nvidia has shown that scaling laws demonstrate that performance improves with increased data and compute resources rather than just hardware scaling alone.

[speaker1]: I’ve heard a lot lately about compute-in-memory. Aren’t many new vendors applying this to bring compute and power advantages?

[speaker2]: True. CIM has been recognized as a means to make the processing power and processing more efficient by orders of magnitude. Unfortunately, some vendors are just shrinking the von Neumann structures, putting more traditional compute units on a chip, each with local caches, and calling that CIM.

[speaker1]: That sounds like lots of smaller inefficiencies instead of a large inefficiency – that doesn’t solve the problems does it?

[speaker2]: You are correct. It sounds great when looked at from a microscope – you have a compute element with memory very close to it so the energy consumed in data movement is low, but when you look at the system-on-a-chip the inefficiency is still there and may be even greater due to all those small data movement energy losses. And this is where we need to make a crucial distinction between CIM - Compute-IN-Memory - and CNM - Compute-NEAR-Memory. This distinction is absolutely critical to understanding why GSI's approach is revolutionary. Most of what the industry calls "compute-in-memory" is actually compute-near-memory.

[speaker1]: Right, so in these architectures, you still have that fundamental von Neumann bottleneck - data has to move from memory to closer memory, then to compute units, then results to memory and on to other compute units, and so on several times. Even if they're physically on the same chip, you're still moving data around, which costs time and energy.

[speaker2]: Precisely! But GSI's APU does something remarkable - it performs computation directly on the memory bit lines themselves. This isn't compute near memory, this is compute literally IN the memory. The computation happens where the data is actually stored, at the most fundamental level.

[speaker1]: Can you tell our listeners what makes GSI's approach different from other companies' compute-near-memory?

[speaker2]: It’s actually quite elegant! The GSI APU is configured as millions of what they call bit processors. These bit processors together perform about 5 peta Boolean operations per second on the latest generation production chip. A traditional read-write on a processor becomes a read-modify-write with just moving new data in. That modify is actually an arithmetic function. If workloads require multiple iterations, then the data just loops in place in memory. Also, the architecture has a 1-bit granularity, so any instruction or next instruction can choose random widths for the memory. So, if you need to go from 32-bit to 64-bit, or you want to use 2-bit weights for AI you still use the entire array, no waste.

[speaker1]: That's brilliant! So instead of the traditional approach where you need separate arithmetic logic units that require data to be moved to them, the memory itself becomes computational. Let's talk about the massive implications for power consumption and performance.

[speaker2]: The power savings are orders of magnitude lower compared to traditional von Neumann architectures. And here's why this is so critical right now: research from UC Berkeley show AI  compute needs for LLM training has been growing at a rate of 750 times compute scaling every 2 years, but LeCun published in IEEE that traditional GPUs can only

provide about 3 times improvement in that same timeframe. We have a massive scaling crisis on our hands.

[speaker1]: Wow, 750 times versus 3 times - that's not just a gap, that's a chasm! I assume this is why we're seeing such enormous energy consumption in AI data centers and huge buildouts. I just read that Meta's new data center is the size of 70 football fields and needs a three billion dollar power upgrade.

[speaker2]: Yes, along with the micro inefficiencies, the traditional approach needs to break large problems into smaller pieces, then use complex interconnects and more cache hierarchies to align everything. It is increasing the von Neumann problem and results in diminishing returns. With GSI's approach, the compute blocks can be larger, workflows can be more diversified, and scaling is like adding more memory to a system. You can achieve high core utilization because there's less data movement at the whole system level. Imagine scaling a data center as simply as plugging in more memory cards!

[speaker1]: And I imagine the environmental implications are massive too. With data centers consuming more and more of the world's electricity, this could be a game-changer for sustainability.

[speaker2]: Absolutely! When you eliminate the vast majority of data movement energy costs and achieve orders of magnitude better power efficiency, you're not just talking about incremental improvements - you're talking about fundamentally changing the energy profile of AI and high-performance computing. This could be the key to making AI sustainable at scale.

[speaker1]: And from a business perspective, this must offer significant total cost of ownership advantages, too.

[speaker2]: Definitely! Better performance per watt per dollar, reduced cooling requirements, simpler scaling, and the ability to eliminate memory bottlenecks for applications that fit within the array. Companies won't just be saving on electricity bills -

they'll be able to do more computing with less infrastructure. It's a complete paradigm shift in how we think about the economics of computation.

[speaker1]: David, this sounds like one of those rare technologies that could really move the needle. But I have to ask - if this is so revolutionary, why do you think it hasn't been done before?

[speaker2]: That's the million-dollar question! Sometimes breakthrough innovations require the right combination of technological maturity, market demand, and visionary thinking. GSI has been working on associative processing for years, building on proven SRAM technology. But now, with the explosive growth of AI and the movement of this processing to the edge, the clear limitations of traditional architectures just cannot achieve the edge needs. The market is truly ready for this kind of paradigm shift.

[speaker1]: It really does seem like perfect timing. As ChatGPT and other AI technologies drive unprecedented demand for compute resources, we desperately need new approaches that can scale efficiently and bring that capability to the edge. GSI's true compute-in-memory technology could be exactly what the industry needs to bridge that scaling gap.

[speaker2]: Absolutely! And what I find most compelling is that this isn't just theoretical - GSI has actual hardware implementing these concepts. They're not asking the market to wait for some future breakthrough silicon. This is happening now, with production parts.

[speaker1]: Well, this has been absolutely fascinating! To our listeners, if you're working in AI, high-performance computing, or just wondering about the future of technology, GSI Technology's approach to true compute-in-memory represents something genuinely revolutionary. It's not just an incremental improvement - it's a fundamental rethinking of how computation should work.

[speaker2]: Couldn't agree more! As AI continues its explosive growth and we face increasing demands for sustainable, efficient computing, technologies like GSI's APU

could be the key to a future where we can have both unprecedented computational power and environmental responsibility. Thanks for joining us, and we'll see you next time!