Revolutionizing AI Inference with GSI's APU

Thanks for tuning in and joining us today! You're listening to the second episode in our exciting new podcast series, where we dive deep into cutting-edge technology that's shaping the next frontier of artificial intelligence. From breakthrough innovations to the bold ideas driving progress, we're exploring how these advancements are redefining the future of AI as we know it. Get ready for a thought-provoking journey into what's next.

Have you been following the latest developments in AI hardware? I keep hearing a lot of buzz around GSI Technology and their APU, or Associative Processing Unit.

Absolutely! It's a fascinating technology, especially when you consider its potential impact on the power consumption of AI inference. Many are calling it a truly revolutionary solution.

Revolutionary is a strong word, but I'm curious. What specifically makes GSI's APU so different in AI inference, and especially concerning power efficiency?

The core of its innovation lies in its unique architecture. Unlike traditional von Neumann architectures that separate processing and memory, the APU combines these functions then adds further close integrations to increase efficiency even more. This approach significantly reduces the need to constantly move data between the processing structures and memory, as CPU and GPU do.

So, it's essentially bringing the computation to the data, which sounds like it would inherently save power by cutting down on data transfer. Is that what's often referred to as computing in-memory?

Precisely. By performing computations directly within and very close to the memory units, the APU dramatically minimizes the energy expended on data movement, which is a major power sink in conventional processors. This architecture is incredibly efficient for AI inference and high-performance compute where massive amounts of data are processed.

That makes a lot of sense. Professor Onur Mutlu of University of E T H Zurich and other researchers, including from Carnegie Mellon, Seoul University, Google, and Samsung Research have found that data movement is indeed a huge bottleneck and can consume 61% or more of overall computation energy expended. But beyond in-memory computing significantly reducing this energy consumption, are there other features that contribute to its power efficiency?

Yes, another critical aspect is its high-bandwidth internal memory access and highly parallel processing capabilities. The APU is designed to handle parallel data processing operations with extreme efficiency, which are foundational to many AI and high-performance compute algorithms. This parallelism, combined with reduced data movement requirements, and shorter data paths when moving internally, allows it to perform tasks at much lower energy per computation.

So, it's not just about where the computation happens, but also how efficiently it can process the data.

Exactly. Take, for instance, data centers. While efficiency gains helped initially, the sheer concentration of compute now demands higher power per area. The International Energy Agency estimated data centers used 1% of global electricity in 2020, and by 2025, that could skyrocket to 20% of the world's power supply. And let's not forget, 61% of total system energy is often spent on data movement, not actual computation.

That's a staggering amount of energy.

This is where GSI Technology's Associated Processing Unit, or APU, steps in, specifically with the Gemini-II. It fundamentally breaks the von Neumann model by introducing in-memory compute. Instead of moving data back and forth between compute and memory, the GSI APU stores data and processes it in place.

So, it's about eliminating that costly data movement bottleneck?

Precisely. The GSI APU is a compute-in-memory search accelerator and a high-performance parallel compute engine. This unique approach gives it a massive performance boost. For example, the first generation APU has been shown to perform inference on one billion database entries with one tenth the total power, and one fifth the cost when compared to CPU.

That’s a game-changer for deploying AI more broadly. It also addresses one of the biggest challenges in scaling AI applications beyond data centers. It sounds like GSI Technology is really pushing the boundaries of what's possible in energy-efficient AI.

They certainly are. The focus on reducing inference power consumption is critical for the future of AI, making it more sustainable and ubiquitous. The APU represents a significant leap forward in achieving that goal, offering a high-performance, low-power solution for a wide range of AI inference workloads.

Thank you for listening to our podcast today and be sure to visit the GSI Technology website at www.gsitechnology.com to listen to future podcasts about their groundbreaking innovation, the APU.