Podcast 11

Edge AI Realized: A Technical Deep Dive into GSI's APU Architecture

[speaker1]: Welcome back to episode 11 of our podcast. In our last episode, we talked about how AI is migrating from centralized cloud infrastructure to the far edge—and why that shift changes everything from latency expectations to system autonomy.

What's interesting is that this shift isn't theoretical anymore. At CES 2026, we saw a flood of AI systems that are clearly ready for real-world deployment. Autonomous cooking robots. Four-wheeled handler robots. Even bipedal humanoids. SERVE was there—already delivering food with autonomous wheeled robots—and ZOOX was actively transporting people up and down the Las Vegas Strip in fully autonomous vehicles.

So today, we're continuing that conversation, but we're going much deeper technically. We're going to break down how GSI Technology's Associative Processing Unit—Gemini-II—actually delivers on the edge AI requirements these systems are exposing.

To help us unwrap the details we welcome back Eleanor to the talk.

[speaker2]: Thanks, Roger. CES was a perfect snapshot of where edge AI really is today. These weren't demos—they were deployed systems making decisions in real time, in uncontrolled environments.

[speaker1]: And what really stood out to me was that none of those systems were measuring TOPS. They were all about real-world performance and responsiveness.

[speaker2]: Exactly. That was a big theme in the technical sessions at CES as well. During the Intel and Edge AI session, low inference times was highlighted for robotics, as well as the adoption of agentic and physical AI for edge control uses. One concept that kept coming up was time-to-first-token, or TTFT. It's becoming a primary performance metric for LLMs in edge AI.

[speaker1]: Which is a shift from how we've traditionally measured AI performance!

[speaker2]: It is. Historically, we focused on peak FLOPs or tokens per second which is key for batch processing in the datacenter. But edge products don't work that way. They need to react immediately. Whether it's a robot avoiding a human, a vehicle recognizing a hazard, or a security system flagging an anomaly, latency matters more than sustained batch throughput.

[speaker1]: That ties into something you mentioned last episode—that modern AI inference isn't really compute-limited anymore. It's memory-limited.

[speaker2]: Right. With modern transformer-based models—LLMs, VLMs, multimodal systems—the arithmetic is relatively cheap. The real cost is moving data. Every token generation requires repeated access to large weight matrices, and most of the time is spent waiting on memory, not doing math.

[speaker1]: So even if a processor advertises massive compute capability, it can still be slow.

[speaker2]: Exactly. GPUs are phenomenal at throughput, but they're designed around shuttling data back and forth between DRAM and compute units. That memory traffic dominates both power consumption and latency—especially when workloads have constantly changing inputs, which is exactly what you see at the edge.

[speaker1]: And edge deployments span a wide range of power environments.

[speaker2]: They do. Edge use cases range from wall-powered systems—like small workstations, compact servers, and network appliances—to power-over-Ethernet devices such as traffic systems and security cameras, all the way down to battery-powered platforms. This is a range of 1500 watts, to 100 watts, and down to sub-20 watts.

High volume battery-operated products include drones, portable inspection tools, and systems that can literally be carried by a person. Across all of those categories, the common requirement is low-latency inference—and very fast frame-by-frame time-to-first-token for LLM-based queries.

[speaker1]: But today, the best TTFT numbers usually come from very power-hungry systems.

[speaker2]: That's the tradeoff. Wall-powered edge applications often rely on processors that deliver excellent TTFT, but they draw 250 watts—and in some cases, up into the kilowatt range. That's simply not viable for many edge deployments, especially mobile or thermally constrained ones.

[speaker1]: So let's talk about how GSI fits into this picture.

[speaker2]: GSI's APU takes a fundamentally different architectural approach called compute-in-memory. Instead of moving data out of memory to be processed, the computation happens directly inside the memory array itself.

[speaker1]: And that provides more energy efficiency compute?

[speaker2]: Precisely. The model weights reside in large on-chip SRAM arrays, and operations like comparison, accumulation, and vector matching all happen right there in memory.

[speaker1]: So the weights don't move much.

[speaker2]: Right. That's one use case. And that data movement is the biggest bottleneck in LLM processing. Data motion is the most energy-expensive part of AI. When you reduce it, you dramatically reduce both power consumption and latency.

[speaker1]: How does that impact time-to-first-token specifically?

[speaker2]: On GPUs, the hierarchy of cores and SMs means that even with several hundred kilobytes of register memory, individual engines have very little memory to hold weights. So every step in node processing requires getting model weights for every matrix operation. On the APU, the model weights can be read once per node and remain resident for the column processing. This significantly reduces memory transfers, particularly for larger models which the APU supports at the edge.

[speaker1]: So physical AI scenario interpretation is faster.

[speaker2]: Exactly. That's why Gemini-II can deliver up to three times faster time-to-first-token compared to GPU-based edge solutions. And this is for a multi-modal VLM, not a single dimensional text LLM. The type of data fusion processing imperative for physical AI at the edge.

[speaker1]: Which becomes critical in autonomous systems.

[speaker2]: Absolutely. Whether it's drones, robots, or vehicles.

[speaker1]: Like what we saw from ZOOX at CES.

[speaker2]: Yes, when you are looking for contextual awareness, you aren't generating thousands of tokens. Edge use cases require evaluating policies continuously and making fast, safety-critical decisions. They need results fast to operate in real time.

[speaker1]: Which enables understanding environment rather than just recognizing pre-selected objects.

[speaker2]: Exactly. Instead of a security camera looking for specific faces, objects or weapons, an edge security installation can be looking for “people acting suspiciously” That’s the kind of reasoning edge systems are increasingly being used for. Human-AI collaboration is not about giving people more highlights to overwhelm them. It’s about making autonomous identification and put the human-in-the-loop for critical exception decisions.

[speaker1]: Let’s put some numbers on this. What does the performance actually look like?

[speaker2]: Gemini-II delivers first response awareness via time to first token in multimodal LLM processing in less than 3 seconds at 30 watts. This is less than a quarter the power consumption of Nvidia Jetson Thor for the same workload and performance.

[speaker1]: That’s a massive gap.

[speaker2]: It is—and it reinforces why TTFT per watt matters more than peak FLOPs for edge AI.

[speaker1]: So to wrap it up, CES showed us that edge AI is here—and the bottleneck isn’t models. It’s architecture.

[speaker2]: That’s right. The APU isn’t just another accelerator. It’s an architecture designed for how edge AI actually behaves: memory-bound, latency-sensitive, multimodal, and power-constrained.

[speaker1]: Thanks for joining us. Be on the lookout for our next episode. We’ll see you next time.