Episode 6

Understanding Time To First Token (TTFT) in Edge AI

[speaker1]: Welcome back to the sixth episode of our podcast! I'm Alec and with me today is my co-host, Sofia. Today, we're diving into a crucial metric for large language models on edge devices: Time to First Token, or TTFT.

[speaker2]: That's right, Alec. TTFT is the delay between submitting a prompt and receiving the very first output token. It also applies when you are asking the same thing but sending new inputs, like new pictures. For multimodal LLMs on edge devices, such as in physical AI and real-time vision processing, responsiveness is absolutely critical. A slow TTFT will limit an application's usefulness in real-time use cases.

[speaker1]: So, let’s set the stage on why TTFT is so high for most edge devices. And more importantly, how is GSI Technology's Gemini-II APU flipping the script on that. It turns out that before that first token even appears, all inputs, whether text, images, or audio, have to go through what's called a 'prefill' stage. This is where the model processes the entire input request sequence. Then it moves into a 'decode' stack to generate that initial output token.

[speaker2]: Exactly. And this prefill latency is often dominated by what we call the 'memory wall' problem, which hits especially hard in transformer decoders. This is particularly crucial for multimodal models, that might combine image, sensor, and textual instruction inputs, the first thing they do is prefill. That means running a full pass through every transformer layer before producing the first token.

[speaker1]: That's a critical point. During this prefill phase, computation is heavily dominated by the widest feed-forward network—or FFN—matrix multiplication operations. For instance, in a model like Gemma3 12B, a representative case study, these FFN operations involve dimensions of 4k by 16k, even with a minimum prefill sequence of just 256 tokens from an image. That's a lot of math!

[speaker2]: And the real bottleneck isn't that math. It's data movement. Specifically, getting huge weight matrices and activations where they need to be. On embedded GPUs, everything is tiled, shuffled, and repeatedly reread from DRAM because their local memories — registers, L1, L2 — are just too small.

[speaker1]: So, if you've got something like a 4k by 16k weight matrix, that's not in a close location ready to process.

[speaker2]: Exactly. In current GPU processing it gets broken into dozens of tiles, shuffled between layers of cache, and often re-read from DRAM multiple times during prefill. For NVIDIA Jetson class devices, this can lead to 6 to 12 second TTFT under multimodal cold starts.

[speaker1]: Enter the Gemini-II APU from GSI Technology. Sofia, walk us through what makes this chip so different.

[speaker2]: It's all about compute-in-memory. GSI's production Gemini-II chip's in-memory associative computing, conducts arithmetic operations directly within its massive 1-million-bit-line SRAM array. This literally transforms the memory fabric itself into the computational substrate, enabling operations to happen right where the data is stored. It's a game-changer for efficiency.

[speaker1]: And what's remarkable are the architectural features that enable this: Each bit line contains 48 associative cells for direct computation, tightly integrated with SRAM bytes for additional temporary storage, large accumulation, and data retention. It boasts a massive 96 Megabytes of on-chip memory with an internal bandwidth of 367 Terabits per second, essentially providing a high-density, high-bandwidth shared register space, but on a massive scale.

[speaker1]: This substantial on-chip capacity is key. It allows the entire input vector to be stored and FFN weight block argument to be co-located with accumulators for the complete prefill window. This means FFN node execution happens in a single, streamed

pass with in-place accumulation, entirely eliminating duplicated DRAM access for weights. It's a huge shift from conventional methods.

[speaker2]: So instead of shuffling tiles between SMs, L1s, and L2 caches as in GPUs, this translates to eliminating redundant DRAM access, reducing DRAM traffic by over 30x during first token inference. This frees up crucial bandwidth, and drastically improves energy efficiency. By optimizing matrix operations and accumulations directly inside this high-capacity compute-in-memory, GSI's APU significantly shortens TTFT.

[speaker2]: So, what kind of edge applications would truly benefit from such accelerated TTFT? Think anything that is performing physical AI. Drones, autonomous robots, or autonomous vehicles that need to process complex sensor data and respond in real time. Any edge AI application where immediate feedback is crucial would see a huge improvement.

[speaker1]: Traditional GPU/TPU-style accelerators often struggle here due to insufficient on-chip memory capacity, often only tens to hundreds of kilobytes of fast memory per SM. This forces them into fragmented computation—think tiling, micro-batching, and constant reload cycles from off-chip memory. This not only creates duplicated off-chip memory traffic, but also results in incredibly ineffective bandwidth utilization, even with high-bandwidth memory systems like LP-DDR5 found on an NVIDIA Jetson Orin and AGX or Qualcomm Snapdragon X Elite. It's a fundamental bottleneck for fast TTFT.

[speaker2]: Precisely. The APU really expands the market for edge AI by overcoming these inherent limitations. It's not just about speed; it's about enabling entirely new classes of responsive, complex AI workloads right where the data is stored, without the need for cloud connectivity.

[speaker1]: To wrap up, the APU truly stands out as an edge AI accelerator. Its flexible precision support—through bit-granular operations—enables software-controlled selection of ultra-low formats like binary and ternary, standard integers (2, 3, 4, and 8-bit), and even floating-point (16 and 32-bit). This means it can preserve accuracy while shrinking

the memory footprint and significantly reducing both off-chip traffic and energy consumption. It's a powerful, efficient solution for the future of edge AI.

[speaker2]: And the result? This innovative approach moves Time to First Token and decisions from a frustrating several-second delay into near-real-time territory, making truly responsive AI a reality for a vast array of edge deployment scenarios.

[speaker1]: Thanks for listening and please reach out to GSI Technology for more details on the Gemini class of APU parts available now for your embedded edge applications.