Podcast 10

AI at the Edge

[speaker1]: Welcome to episode 10 of our podcast. Today, we're diving into a topic that's quietly reshaping everything from autonomous cars to remote sensors—The Evolution of AI at the Edge. And to guide us, we have AI architect and expert, Eleanor. Welcome to the show.

[speaker2]: Thanks for having me, Roger. It's an exciting time for AI, and this transition is a pivotal one.

[speaker1]: So, let's start at the beginning. You've mentioned that what's happening with AI at the edge feels familiar. What do you mean by that?

[speaker2]: Exactly. Artificial intelligence is undergoing a transition we've seen before with cloud computing. In the early days of the cloud, all the heavy processing happened in huge, centralized data centers.

[speaker1]: Then it started getting distributed.

[speaker2]: Mhm. Over time, that processing started moving outward to what is called fog or near-edge. It moved closer to where the data was actually being used. This was driven by practical needs—reducing latency, saving bandwidth, and improving security and availability.

[speaker1]: And now AI is on that same path?

[speaker2]: Precisely. We're seeing an AI migration from the cloud toward the edge. But this isn't just about changing the location of the computation. What's new is it's fundamentally changing where AI systems can be used, what they can do, and how autonomous they can become.

[speaker1]: To really understand where this is all going, maybe we should break down the basics. Can you walk us through this migration?

[speaker2]: Of course. Let's start with how recent AI models and processing work. At its core, AI is built on two distinct phases: training and inference. Training is where you feed a model massive amounts of known data to build a mathematical representation that can recognize patterns. It's computationally intensive and happens on large servers with powerful GPU compute resources.

[speaker1]: Okay, so that's the 'learning' phase.

[speaker2]: Exactly. Inference, on the other hand, is using that trained model to make decisions on new, unseen data. When a camera or other sensor generates an input, inference is what allows the system to interpret that input based on what it was trained on previously from past data.

[speaker1]: So, how did this lead to the first wave of AI at the edge?

[speaker2]: Well, as AI matured, inference models became much more efficient. Through techniques like pruning and quantization, we could shrink them down without losing too much accuracy. This made it possible to run inference on CPUs and even on smaller, low-power embedded processors.

[speaker1]: And that could unlock some powerful real-world applications, I imagine.

[speaker2]: It absolutely does. Think of autonomous vehicles reacting instantly to their environment, or automatic obstacle avoidance in drones. Edge inference gave us faster responses and better reliability.

[speaker1]: Right. But these early systems had their limits, didn't they?

[speaker2]: They did. They were very task-specific. You'd have one trained network to identify a specific set of objects. If you wanted to expand its capability, you had to retrain it or deploy a whole new model.

[speaker1]: Which brings us to the next big leap... Large Language Models, or LLMs. How did they change the game?

[speaker2]: LLMs introduce a completely new dynamic. They allow a single, massive model, trained on a broad corpus of data, to generalize across many different tasks. It could interpret language, infer context, and respond flexibly.

[speaker1]: Initially, this is all cloud-based though, right? Like chatbots. But text is only one piece of the puzzle. The real world is visual, it's auditory... so what's the next step?

[speaker2]: That's the real inflection point: the move from single-modal to multimodal intelligence with Video Language Models, or VLMs.

[speaker1]: VLMs. Okay, break that down for us.

[speaker2]: VLMs extend the LLM concept to multiple sensor inputs at once—video, audio, images, structured data, and even text. Instead of building separate pipelines for each task, a VLM can interpret a complex scene using a generalized understanding of the world.

[speaker1]: So instead of training one model to find a fire, and another to find a person...?

[speaker2]: Exactly. A VLM can answer higher-level, contextual questions. Not just 'Is that a fire?' but 'Is a fire starting in this area?' or 'Is there a human present where there shouldn't be?' It enables scenario-based inference, not just simple classification.

[speaker1]: That sounds like a fundamental change in autonomy. What does that mean for systems at the edge?

[speaker2]: It means they can now not just measure their environment via sensors, but also recognize the context in the field of view, make decisions based on policies and intent, and act.

[speaker1]: You mentioned at the start that reduced latency was one reason for the move to the edge from cloud, that must be important for embedded edge applications also.

[speaker2]: Glad you brought that up. Yes, and this is one divergence from cloud computing. You might use LLMs in the cloud to create an essay on a topic. That could be thousands of tokens. The speed and cost of the effort is going to depend on how many GPU resources are dynamically allocated to the workload. On a far edge product you cannot spool up any extra resources beyond what is already in the product. But, also the workflow is less about thousands of tokens about a single input, and more about a few tokens about continuous inputs.

[speaker1]: Can you explain that further?

[speaker2]: Yes, let’s look at it from the drone that may be investigating large swaths of land after a thunderstorm. It will be continuously taking pictures, optical and infra-red, as it travels, and the workflow is to determine if there is any observed fires starting. The policy can be checked with a few tokens only, then abandoned if nothing is found, and a new sensor set interrogated. Here an important metric is time to first token. It determines how fast the identification of the policy occurs and how soon action can be taken. It’s not processing in thousands of tokens as done in the cloud.

[speaker1]: That's powerful capabilities. And I can imagine the paradigm shift of this move to the edge. It can't be that simple, though. Running these massive models on a small device must present some serious technical challenges.

[speaker2]: You're right. The biggest is the compute and memory challenge. Modern AI inference is increasingly memory-bound. LLMs in particular need fast access to huge datasets of weights, and traditional architectures are just too slow and power-hungry for the edge.

[speaker1]: Because at the edge, low SWaP, that is size, weight, and power, are critical.

[speaker2]: Absolutely. So we're seeing a shift in focus. It's not about scaling large numbers of GPUs anymore. Metrics like 'time-to-first-token'—the delay before you get a meaningful output—are becoming critical. Edge AI is about responsiveness, determinism, and radical efficiency.

[speaker1]: So, as we look to the future, what's on the horizon?

[speaker2]: Well, the cloud datacenters will increase capabilities of foundation models. The edge will use these and implement more sophisticated multimodal analyses, providing greater autonomy in remote systems.

[speaker1]: It sounds like decision making itself is becoming distributed and remote.

[speaker2]: That's the perfect way to put it. Edge systems are no longer passive data collectors. They are becoming decision-makers. The future of AI isn't just in the cloud; it's at the edge, embedded in the systems that interact directly with our physical world—right where the new data is observed.

[speaker1]: A fascinating and powerful vision for the future. Eleanor, thank you so much for breaking down this complex topic for us today.

[speaker2]: My pleasure, Roger. Thanks for the great conversation.

[speaker1]: Join us in the new year for future episodes where we will cover practical implementation of these concepts, such as how the GSI Gemini-II addresses the challenges of AI and LLM migration to the edge. Happy Holidays!