Podcast 13

Understanding Latency in Physical AI

[speaker1]: Alright, today we're digging into something that sounds simple at first — latency — but turns out to be a complex concept in Physical AI. Latency is understood, often as faster is better, slower is worse. But in robotics and autonomous systems, it's more than one number.

[speaker2]: Yeah, exactly. In Physical AI, latency isn't a single scalar metric — it's layered. Different parts of the system have different response times, and those response times line up with completely different kinds of intelligence.

[speaker1]: Right. For this discussion, we will divide the system into three latency regimes. You've got ultra-low latency for motor control, mid-latency for navigation, and higher latency for awareness and reasoning.

[speaker2]: And what's interesting is each one has its own compute architecture, its own memory requirements, and even its own definition of what "intelligence" means.

[speaker1]: Let's start at the bottom — the fastest layer. We're talking sub-5 to sub-fifteen milliseconds.

[speaker2]: And that's not arbitrary. In robotic systems, control loops often run at kilohertz frequencies. That means the system only has microseconds to a few milliseconds to sense, decide, and act.

[speaker1]: This is where motor control comes together.

[speaker2]: And if you introduce machine learning into that loop — say learned impedance control or fine force modulation — the timing has to stay extremely tight.

[speaker1]: Because once you go past about 15 milliseconds, you start getting phase lag.

[speaker2]: Which could introduce oscillations and instability between motors in a high-degree-of-freedom robot... that can get dangerous fast.

[speaker1]: So the compute here has to be deterministic. Models run out of on-chip SRAM. Scheduling is real-time. Often integrated with MCUs or DSP pipelines.

[speaker2]: I am thinking of a biological analogy here. This layer is basically the reflex arc.

[speaker1]: Yeah —minimal processing. Just stimulus → response.

[speaker2]: You could call it the reptilian layer of intelligence. It only cares about keeping the system stable and alive.

[speaker1]: Now move up one level, and we get into the sub-50 millisecond range.

[speaker2]: This is where the system can do advanced motor control, like gait control, and also starts interacting with the environment, not just stabilizing itself.

[speaker1]: Exactly. Multi-joint coordination. Visual SLAM updates. Obstacle avoidance. Short-horizon path planning.

[speaker2]: Fifty milliseconds corresponds to about 20 Hz update rate, which lines up really well with human perception and mobile robotics.

[speaker1]: And the compute changes here too.

[speaker2]: Yeah, now you're looking at CNNs, small transformers, sensor fusion models. Data coming from cameras, LiDAR, IMUs, depth sensors.

[speaker1]: So into the deterministic loops, you've got additional processing. Often with GPUs and NPUs.

[speaker2]: Still time-constrained, but not microsecond-level anymore.

[speaker1]: And the intelligence here is situational.

[speaker2]: Right. The system understands what's happening right now, in a narrow window of time and space.

[speaker1]: It can adjust trajectory before inertia becomes a problem, but it's not doing long-term reasoning.

[speaker2]: It's reactive, but with awareness of the environment.

[speaker1]: Now the top layer is where things get really interesting. This is the sub-3 second regime.

[speaker2]: And this is where we start talking about large models — multimodal transformers, vision-language models, things that can combine video, text, audio, telemetry, all at once.

[speaker1]: This is also where the metric changes. Instead of just latency, we can talk about Time-to-First-Token, or TTFT.

[speaker2]: Which comes from large language model inference. It's basically how long it takes before the system can start producing the first meaningful response.

[speaker1]: And that matters a lot for safety.

[speaker2]: Yeah, because if the system can understand what's happening in under three seconds, that insight can still influence what happens next.

[speaker1]: That three-second window actually applies also to human factors.

[speaker2]: Like the three-second rule in driving.

[speaker1]: Exactly. It's enough time to perceive, interpret, and react.

[speaker2]: So if a robot can detect a human entering a workspace, recognize a misaligned processing object, it can still change its behavior safely.

[speaker1]: If it takes longer than that, the information becomes historical instead of actionable.

[speaker2]: Getting sub-3 second TTFT for multimodal models is not easy.

[speaker1]: You're dealing with huge parameter counts, massive KV cache data transfers, and memory bandwidth limits.

[speaker2]: So architectures that minimize data movement become really important.

[speaker1]: Compute-in-memory, memory-resident models, high-bandwidth SRAM — those kinds of designs can make a big difference.

[speaker2]: And unlike the lower layers, this one is doing real abstraction.

[speaker1]: It builds an internal model of the world.

[speaker2]: Predicts what might happen.

[speaker1]: Understands intent.

[speaker2]: This basically represents the conscious layer of the machine. Without actually being conscious, of course.

[speaker1]: One of the key design rules across all of this is separation of the latency domains.

[speaker2]: Motor control cannot wait for awareness. Navigation cannot block on reasoning. And awareness cannot destabilize control loops.

[speaker1]: So each layer runs independently, but they still coordinate.

[speaker2]: Just like biology.

[speaker1]: Reflexes happen without thinking, but thinking can still influence behavior over time.

[speaker2]: Higher cognition guides, but it doesn’t sit inside the reflex loop.

[speaker1]: This is why reliable sub-3 second multimodal TTFT is such a big milestone for Physical AI.

[speaker2]: Because it moves systems from reactive to context-aware.

[speaker1]: Robots can anticipate hazards.

[speaker2]: Interpret human intent.

[speaker1]: Adjust tasks without constant supervision.

[speaker2]: And that improves both safety and efficiency at the same time.

[speaker1]: So when we talk about latency in Physical AI, we're really talking about a cognitive stack.

[speaker2]: Sub-15 milliseconds gives you embodiment.

[speaker1]: Sub-50 milliseconds gives you navigation.

[speaker2]: Sub-3 seconds gives you awareness.

[speaker1]: And when all three work together, machines stop being just reactive systems...

[speaker2]: ...and start becoming agents that can operate safely in the real world.