www.pipelinepub.com
Volume 22, Issue 9

AI for IoT Latency: the Closer the Better

By: Mark Cummings, Ph.D.

In most IoT systems, latency is an important issue. In critical control systems such as pipelines, electrical grids, factory automation, and autonomous vehicles, reaction time is critical. Latency determines reaction time. Keeping sensor, actuator, and processing very close to each other produces the lowest latency. Replacing simple hardwired logic/microcontroller processing with AI in IoT provides a significant increase in capability. But it brings latency challenges with it.

There are two ways of dealing with this challenge: finding ways to reduce the size of the LLM, and placing the LLM in the network as close to the sensors and actuators as possible. As AI hardware and software continue their rapid evolution, placing sensors, actuators, and LLMs in a single package will become possible. Therefore, IoT systems should be designed to fit today's technology capabilities and to be able to migrate easily to new emerging capabilities as they emerge.

IoT AI Challenges

Latency can be thought of as a function of distance. Distance can be thought of as a combination of propagation delay and processing delay. The lower the distance between sensors and actuators, the lower the latency. That distance is a function of the length of wires/fibers/wireless links between the sensor and the logic that interprets the sensor data and issues actuator instructions; the processing time; and the length of wires/fibers/wireless links to the actuators. Putting this whole system into a single small package tends to minimize latency.

With simple processing systems using hardwired logic or a small microcontroller, it was not so challenging to put the processing in the package with the sensors and actuators. With the advent of GenAI and the desire to have more capable processing, able to make more sophisticated data analysis and decision making, the processing hardware requirements go beyond a simple microcontroller.

AI processing requires at least a full processor and significant memory. This increases the power requirements. The increased power means increased heat that must be dissipated. Providing the increased power may also be a system challenge.

There are two ways to respond to this challenge: LLM size reduction and placing the LLM outside the sensor/actuator package, but as close as possible to it in the network.

LLM Size

Frontier LLM models are getting quite large. They run in large data centers with lots of space, power, cooling, etc. They are designed to meet a very broad range of uses and respond to a very wide range of situations. Although the ability to run these large models on Edge devices is increasing, it is still challenging to run them in the same packages as sensors and actuators.

IoT systems tend to have a very narrow set of stimulus-response requirements and operate in relatively narrow, well-defined environments. Therefore, it is possible to use smaller, more specialized LLMs.

Smaller models are improving their capabilities. This comes from two directions: more capable general capabilities, and more capable specialized capabilities. General capabilities of smaller models are coming from a combination of more engineering attention on the problem and new technology. The most dramatic of the new technologies is the use of Distillation. Distillation uses the outputs of larger models to train small models. Thus, giving the small models capabilities that approach the larger models. A second approach is the use of specialized training. This can occur after the initial training to 'fine tune' the model. Or, where economics justify the effort, training a new specialized model from scratch.

Network Processing

If it is not possible to place the AI in the same package as the sensor/actuator, then attention should turn to placing the AI in the network as close to the package as possible. It is often tempting to aggregate AI needs for a series of packages into a single, more powerful processor that can support them all. This tends to lengthen latency. Aggregating AI processing into a single processor can allow for a more powerful processor with larger memory. But the aggregated processor has to be further away from the sensors/actuators. From a latency point of view, multiple smaller AIs closer tend to be a better solution.

In some cases, there may be enough system resources in the communications processor near the package to run a small LLM there. In any case, effort should be made to avoid a hub and spoke network. That is a network where data travels from an Edge down a spoke to a central hub and then back out again on another spoke.

IoT Infrastructure Evolution

Just as the advent of the general-purpose microprocessor led to the development of the Microcontroller, it is likely that a micro-AI specialized processor will be developed. For example, a Micro-GPU/CPU combo.

At the same time, software evolution will continue. Although the frontier model companies are focusing on the largest LLMs, there is still some promising work going on with smaller LLMs.

As the quantity of sophisticated model developer talent continues to expand, it will become economical for companies focused on IoT to develop their own specialized models. For these companies, this means investments in: training existing staff and in adding advanced AI capability staff as market economics allow.

Brian Case is a microcomputer expert whose career started with early microprocessors. He has an interesting perspective on AI evolution. “When we started designing microprocessors in the 80s, we were trying to go beyond minicomputers, and we did. We failed, though, to anticipate tiny, cheap microcomputers that could bring sound and music to greeting cards; we simply failed to extrapolate the effects of Moore's law correctly. I think this time we should anticipate AI so ubiquitous that it will be in our light bulbs.”

Conclusion

Latency is a key issue for many IoT systems. Adding AI to IoT can dramatically increase capabilities. However, it can bring latency challenges. Keeping sensor, actuator, and processing very close to each other produces the lowest latency. With today’s hardware and GenAI systems, this may be challenging. There are two ways of dealing with this challenge: finding ways to reduce the size of the LLM, and placing the LLM in the network as close to the sensors and actuators as possible. As AI hardware and software continue their rapid evolution, the challenging aspects of placing sensors, actuators, and LLMs in a single package will decrease. Therefore, IoT systems should be designed to fit today’s technology capabilities and to be able to migrate easily to new emerging capabilities as they arise. That is, don’t wait. But plan for migration.