Podcast 16

AI's Hidden Power Consumption

[speaker1]: Welcome back. Today we're talking about something most of us probably don't think about when we use AI: electricity. We type a question into an AI assistant, analyze an image, or ask a model to summarize a document, and seconds later we have an answer. From our perspective, it feels almost effortless.

[speaker2]: But behind that interaction is a lot of computing infrastructure, and all of that computation requires electricity. As AI adoption accelerates, data-center power demand is growing with it. In fact, Lawrence Berkeley National Laboratory estimated that data centers consumed about 4.4 percent of U.S. electricity in 2023. Depending on how the market develops, that could increase to somewhere between 6.7 and 12 percent by 2028.

[speaker1]: That's a remarkable increase in just five years. And this isn't only a problem for technology companies. It potentially affects utilities, grid operators, businesses, and the communities where these facilities are being built. So the question becomes: If AI continues growing at this pace, can we keep solving the problem simply by building bigger data centers and adding more processors?

[speaker2]: Or should we also look at how efficiently we're performing the computation in the first place?

[speaker1]: Let's start there. Why does AI require so much electricity?

[speaker2]: There are several reasons, but one of the most important is something we don't hear about nearly as often: data movement. In a conventional architecture, memory and processing are largely separate. The processor needs data to perform a calculation, so information must continually move between external memory and the processor. For AI, we're moving model weights, activations, context, intermediate results, and live data. Moving all of that information consumes energy.

[speaker1]: So we're not just using electricity to perform calculations. We're also using electricity to constantly move the information needed to perform those calculations in the overall system.

[speaker2]: Exactly. Multiply that activity across thousands of servers running around the clock, and power becomes one of the defining challenges of AI infrastructure. Then there is the heat. Electricity consumed by computing equipment creates heat that must be removed. Higher-density AI infrastructure can require increasingly sophisticated cooling systems, adding another layer to the power and infrastructure challenge.

[speaker1]: And this is where the issue becomes relevant even for people who may never set foot inside a data center. A large data center can become an enormous new customer on the regional electrical grid.

[speaker2]: That's important because utilities have to ensure there is enough generation, transmission capacity, substations, transformers, and other infrastructure to serve these large loads while continuing to reliably serve existing homes and businesses. That doesn't necessarily mean building a data center automatically causes everyone's electric bill to increase. Electricity markets and utility regulations vary considerably. But when large new loads are added quickly, utilities may need substantial investments to support them. If demand grows faster than available supply and infrastructure, it can create pressure on electricity prices and grid reliability.

[speaker1]: So someone living near a rapidly growing data-center market could potentially feel the effects even if they never use the services being provided inside those facilities.

[speaker2]: Right. The broader question becomes: How do we support AI growth without unnecessarily increasing the burden on the electrical infrastructure shared by everyone? And that's where efficiency becomes incredibly important.

[speaker1]: This brings us to GSI Technology's approach with the Gemini-II Associative Processing Unit, or APU. Instead of simply asking how we can make conventional processing faster, GSI is approaching the problem from a different architectural direction.

[speaker2]: Gemini-II uses compute-in-memory technology. Rather than continually moving large amounts of data between separate memory and processing resources, the architecture enables more of the data to reside where computation takes place. The idea is straightforward: if you reduce unnecessary data movement, you can also reduce the energy associated with moving that data.

[speaker1]: And that can be particularly valuable for workloads involving massive parallelism, including vector search, similarity search, certain AI inference workloads, and other applications that fit the APU architecture.

[speaker2]: Exactly. This isn't about replacing every CPU or GPU in a data center. Different processors are optimized for different jobs. The opportunity is to create a more heterogeneous data center, where workloads are directed to the architecture that can execute them most efficiently. Instead of asking a GPU to perform every operation, we can ask a better question: What is the most efficient architecture for this particular workload?

[speaker1]: And we've already seen examples of what that efficiency can look like. In GSI testing using the Gemma 3 12B vision-language model, Gemini-II demonstrated approximately three-second time-to-first-token performance while the APU subsystem consumed around 30 watts.

[speaker2]: That's important because it illustrates the larger principle. Performance alone isn't enough anymore. We also need to consider performance per watt. If a lower-power architecture can execute a particular workload while delivering the performance the application requires, moving that workload away from a higher-power processor can potentially free up both computing resources and electricity. One workload may not make much difference. Multiply that efficiency across thousands of processors operating 24 hours a day, and the impact becomes much more significant.

[speaker1]: And the benefits don't necessarily end with the processor.

[speaker2]: Exactly. If computing equipment consumes less electricity, it also generates less heat that has to be removed. That can potentially reduce cooling requirements and influence power distribution, rack density, and other supporting infrastructure. Most importantly, more efficient computing means that a data center can potentially perform more useful computational work within a given power envelope.

[speaker1]: And when electricity itself is becoming a constraint on data-center expansion, that is an important advantage. Instead of asking only, "How much computing power can we build?" the industry increasingly needs to ask, "How much useful computation can we get from every watt available?"

[speaker2]: Efficiency addresses something fundamental: how much electricity is required to perform the computation in the first place.

[speaker1]: Every watt that doesn't need to be consumed is a watt that doesn't have to be generated, transmitted, converted, and ultimately dissipated as heat. As AI expands into more applications, improving that equation becomes increasingly important.

[speaker2]: That's what makes architectures like GSI Technology's Gemini-II APU so interesting. Compute-in-memory addresses one of the fundamental inefficiencies of conventional computing by reducing the need to constantly move data between memory and processing resources.

[speaker1]: AI isn't going away. We're going to run larger models, analyze more video, process more sensor data, perform more searches, and put intelligence into more applications.

[speaker2]: Ultimately, the future of AI won't be determined only by how much computing power we can build. It will also be determined by how intelligently we use the electricity available to us.

[speaker1]: Because as data centers become a bigger part of the conversation about grid capacity, electricity costs, and community infrastructure, performance per watt isn't just a processor specification anymore. It's becoming an essential part of how we scale AI responsibly and efficiently. Thanks for listening. We'll see you next time.