

Demand isn't about to let up, but prices will start plunging by factors of 10.



INSIDE VIEW

How the AI Boom Ends

I can tell you how the artificial-intelligence datacenter boom will probably end, but I can't tell you when. Implementing large language models has been a brute-force approach: Big, fast and expensive GPUs from Nvidia crunch a trillion parameters using massive amounts of memory housed in liquidcooled, gigawatt-gobbling data centers, all with brute-force pricing. It sure smells like IBM mainframes of old.

Be careful about extrapolating anything from today's pricing. With AI, like all new technologies, the refinement stage is kicking in—just as mainframes gave way to cheaper minicomputers, microprocessors, web servers and now AI architectures. AI token prices have been dropping but are set to plunge faster than most anticipate. Tech growth is never a straight line.

Silicon Valley's mantra for decades has been smaller, cheaper, faster, better. I asked venture capitalist Steve Jurvetson to translate that into numbers. "We often see a technology shift when the new approach is 10 times better. It can be 10 times more performant or 10 times cheaper, sometimes both." He calls it the "factor of 10."

And it's happening. Last week, according to OpenRouter, OpenAI and Anthropic were charging a premium for some of their models—\$1 to \$2 per million input tokens. Comparable Chinese openweight models like Z.ai's GLM and Alibaba's Qwen charged 6 to 12 cents, around 90% less—that factor of 10 inflection point. American AI labs have gone from 70% market share last year to 30% today. All are still growing. But prices matter.

AI architecture is evolving. Instead of a trillion parameters, efficient small language models might use 70 billion and run on an Apple Mac. These local models have no token costs beyond the cost of hardware and electricity.

But they must be less capable than the large language models, right? Not really. Stanford researchers show that a locally run small language model "can successfully answer 88.7% of single-turn chat and reasoning queries." No question we'll always have high-end models running in data centers. IBM still sells mainframes. But growth will happen elsewhere.

Where? Another Silicon Valley saying I've lived by: Intelligence moves out to the edge of the network. The real breakthrough will happen when robust AI models run on smartphones. It's already starting. In September, California-based PrismML released a compressed version of the popular open-source Qwen3.8 27B AI model. The company got it down to under 6 gigabytes. That means it fits not only on PCs but also on high-end smartphones. The memory size is reduced by roughly 9 to 10 times. There's that factor of 10.

Technology doesn't stand still. Qualcomm recently announced a new Snapdragon neural processing unit, or NPU, for smartphones. Using what the company calls a "mixtureof-experts" model, it could run a 30 billion parameter AI model in 3 gigabytes of memory. According to the Deep View, the largest AI model on a phone has been four billion parameters on an Apple A20 Pro chip. Qualcomm provides a 7.5 times boost.

NPUs, like Qualcomm's new Snapdragon and Google's Tensor chips in its Pixel phones, are multiplying. Same for the latest Apple M- and A-series chips, engines that can run AI models locally on iPhones—maybe why Apple hasn't joined the AI data-center spending delirium.

2026-10-12

ECO

AI

Note

Note

A Nvidia-backed startup, Reflection AI, claims it matches leading Chinese openweight AI models with “3- 4x less inference compute.” That’s it? Only 3 or 4 times? Wake me up when it’s 10 times. At the pace AI is moving, that could be year-end.

Agents are all the rage, from Meta’s Muse to xAI’s Grokbot. They run on virtual machines in data centers but raise all sorts of privacy concerns. Startup Perplexity recently teamed up with Nvidia to run agents on PCs. Zero token costs. And you own your data.

I often hear, that if tokens get cheaper, demand will go up. That’s Jevons Paradox. Sure, elasticity works. But if prices drop 90%, demand must increase 900% just to stay flat. Remember when internet demand grew while debt-laden Global Crossing spectacularly failed because prices dropped too fast? Historically, semiconductors got 30% cheaper every year. When prices dropped by more than 50%, disaster struck with fabrication plant closings and company failures.

Will 2027 see peak data center? Again, timing is dicey. Current demand is gangbusters, but spending is a game of chicken until multiple factor-of-10 improvements kick in and prices collapse. This could set off a cascade of data-center halts and debt defaults, even as token demand increases.

Yes, model training, drug discovery and winning prizes like the recently solved Navier-Stokes equations require hot chips in data centers. But chatbots and agents and other tasks may increasingly be handled on PCs and phones and smart glasses and whatever new gizmos Silicon Valley invents over the next decade. I’ve heard this called the “Great Migration.” Mr. Jurvetson’s factor of 10 is visible now, as it often is, but ignored. Change doesn’t happen overnight. It happens gradually, then suddenly.

Write to kessler@wsj.com.

By Andy Kessler

The following is a digital replica of content from the print newspaper and is intended for the personal use of our members. For commercial reproduction or distribution of Dow Jones printed content, contact: Dow Jones Reprints & Licensing at (800) 843-0008 or visit djreprints.com.
