The AI Boom Is Here. The Cloud May Not Be Ready.

Traditional cloud infrastructure wasn’t designed to support large-scale artificial intelligence. Hyperscalers are quickly working to rebuild it.

Isabelle Bousquette

0:00

0:09 / 5:40

Photo illustration: Annie Zhao

“There’s a pretty big imbalance between demand and supply at the moment,” said Chetan Kapoor, director of product management at Amazon Web Services’ Elastic Compute Cloud division.

Most generative AI models today are trained and run in the cloud. These models, designed to generate original text and analysis, can be anywhere from 10 times to a 100 times bigger than older AI models, said Ziad Asghar, senior vice president of product management at Qualcomm Technologies, adding that the number of use cases as well as the number of users are also exploding.

“There is insatiable demand,” for running large language models right now, including in industry sectors like manufacturing and finance, said Nidhi Chappell, general manager of Azure AI Infrastructure.

It is putting more pressure than ever on a limited amount of computing capacity that relies on an even more limited number of specialized chips, such as graphic chips, or GPUs, from Nvidia. Companies like Johnson & Johnson, Visa, Chevron and others all said they anticipate using cloud providers for generative AI-related use cases.

Google's data center in Eemshaven, Netherlands. Google Cloud Platform said it is making AI infrastructure a greater part of its overall fleet. Photo: Utrecht Robin/Abaca/Zuma Press

But much of the infrastructure wasn't built for running such large and complex systems. Cloud sold itself as a convenient replacement for on-premise servers that could easily scale up and down capacity with a pay-as-you-go pricing model. Much of today's cloud footprint consists of servers designed to run multiple workloads at the same time that leverage general-purpose CPU chips.

A minority of it, according to analysts, runs on chips optimized for AI, such as GPUs and servers designed to function in collaborative clusters to support bigger workloads, including large AI models. GPUs are better for AI since they can handle many computations at once, whereas CPUs handle fewer computations simultaneously.

At AWS, one cluster can contain up to 20,000 GPUs. AI-optimized infrastructure is a small percentage of the company's overall cloud footprint, said Kapoor, but it is growing at a much faster rate. He said the company plans to deploy multiple AI-optimized server clusters over the next 12 months.

Microsoft Azure and Google Cloud Platform said they are similarly working to make AI infrastructure a greater part of their overall fleets. However, Microsoft's Chappell said that that doesn't mean the company is necessarily moving away from the shared server—general purpose computing—which is still valuable for companies.

Other hardware providers have an opportunity to make a play here, said Lee Sustar, principal analyst at tech research and advisory firm Forrester, covering public cloud computing for the enterprise.

Dell Technologies expects that high cloud costs, linked to heavy use—including training models—could push some companies to consider on-premises deployments. The computer maker has a server designed for that use.

“The existing economic models of primarily the public cloud environment weren’t really optimized for the kind of demand and activity level that we’re going to see as people move into these AI systems,” Dell’s Global Chief Technology Officer John Roese said.

On premises, companies could save on costs like networking and data storage, Roese said.

Cloud providers said they have several offerings available at different costs and that in the long term, on-premises deployments could end up costing more because enterprises would have to make huge investments when they want to upgrade hardware.

Qualcomm said that in some cases it might be cheaper and faster for companies to run models on individual devices, taking some pressure off the cloud. The company is currently working to equip devices with the ability to run larger and larger models.

And Hewlett Packard Enterprise is rolling out its own public cloud service, powered by a supercomputer, that will be available to enterprises looking to train generative AI models in the second half of 2023. Like some of the newer cloud infrastructure, it has the advantage of being purposely built for large-scale AI use cases, said Justin Hotard, executive vice president and general manager of High Performance Computing, AI & Labs.

Hardware providers agree that it is still early days and that the solution could ultimately be hybrid, with some computing happening on the cloud and some on individual devices, for example.

In the long term, Sustar said, the raison d'être of cloud is fundamentally changing from a replacement for companies' difficult-to-maintain on-premise hardware to something qualitatively new: Computing power available at a scale heretofore unavailable to enterprises.

"It's really a phase change in terms of how we look at infrastructure, how we architected the structure, how we deliver the infrastructure," said Amin Vahdat, vice president and general manager of machine learning, systems and Cloud AI at Google Cloud.

Write to Isabelle Bousquette at
isabelle.bousquette@wsj.com