NEWS RELEASE

QumulusAI Signs $71.9 Million, Three-Year Agreement With AI Inference Platform Provider

2026-07-28

The customer, a platform that helps companies deploy and serve AI models in production, will draw on NVIDIA Blackwell B300 and B200 supply from QumulusAI's AI cloud GPU capacity under a three-year agreement with renewal options.

ATLANTA--(BUSINESS WIRE)-- QumulusAI (Nasdaq: QMLS), a neocloud infrastructure provider purpose-built for the AI computing era, today announced a signed three-year agreement, valued at more than $71 million, to supply NVIDIA Blackwell B300 and B200 capacity to an AI inference platform provider. The agreement includes renewal options and ranks among the company's largest customer commitments announced to date.

The customer's platform helps companies deploy large language models, vision models, speech models and other AI applications with low latency and high reliability and runs models for customers specializing in LLMs, image and video generation. The dedicated Blackwell capacity will give the platform high-performance compute to serve production inference workloads as demand from its customers scales.

Capacity under the agreement will be served from QumulusAI's U.S. data center footprint and is expected to be ready for customer use in the third quarter of 2026. The company's demand-led deployment model places capacity into available pockets of power across a distributed network of colocation and owned facilities, enabling it to bring GPU capacity online in months, not years. The agreement adds more than $71 million in contracted, multiyear commitments to QumulusAI's book of business.

“Inference is where AI meets the real world, and the platforms serving it can't afford to wait on capacity,” said Michael Maniscalco, CEO of QumulusAI. “Our customer runs production workloads for companies that need the right combination of flexibility, access, cost, trust and speed in their infrastructure. A three-year commitment of this size reflects what our model is built to do: procure and deploy Blackwell capacity where demand already exists, and do it fast.”

The agreement extends a series of recent demand announcements. On July 23, QumulusAI announced a $32 million, two-year NVIDIA Blackwell B300 agreement with an AI inference platform provider focused on generative AI applications. On July 22, the company announced a two-year, take-or-pay Blackwell B300 agreement valued at more than $18 million with a GPU cloud marketplace serving AI teams in more than 100 regions worldwide. On June 11, it announced $124.4 million in three-year inference agreements across two customers, and on May 28, it announced two NVIDIA H200 cluster deployments with marketplace partner Shadeform. Together with today's announcement, the company has announced more than $246 million in customer agreements since early June.

About QumulusAI

QumulusAI is a distributed AI cloud platform that delivers accelerated access to high-performance GPU compute. Through an inference-first, demand-led deployment model across a network of data center sites, QumulusAI brings compute closer to customer demand, helping AI teams and enterprises scale production AI workloads with speed, flexibility and control. By combining rapid deployment with flexible private cloud infrastructure, QumulusAI gives customers a faster, more adaptable path beyond the capacity constraints of traditional centralized and hyperscale cloud models. Learn more at QumulusAI.com.

Follow us on LinkedIn and X @QumulusAI.

Forward-Looking Statements

This press release contains forward-looking statements within the meaning of Section 27A of the Securities Act of 1933, as amended, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements regarding the company's agreement, the anticipated value and term of the agreement, for Blackwell B300 and B200 capacity to serve production inference workloads, the company's ability to serve the agreement from its existing data center footprint, the expected timing of deployment of NVIDIA Blackwell capacity, and the pace at which the company can deploy capacity and bring GPU compute online. Words such as “anticipate,” “believe,” “estimate,” “expect,” “guidance,” “intend,” “can,” “may,” “on track,” “plan,” “project,” “target,” “will” and similar expressions are intended to identify forward-looking statements. These statements are based on management's current expectations and assumptions as of the date of this release and are subject to risks and uncertainties that could cause actual results to differ materially, including, among others, the company's dependence on a limited number of large customers; the availability and cost of power, network connectivity and specialized hardware such as graphics processing units; the company's substantial capital requirements and access to financing; competition and rapid technological change in the high-performance computing and AI markets; the company's limited operating history and history of net losses; and those described in the “Risk Factors” section of the company's registration statement on Form S-1, as amended (File No. 333-292514), filed with the U.S. Securities and Exchange Commission (SEC), as such factors may be updated in the company's subsequent filings with the SEC. QumulusAI undertakes no obligation to update or revise any forward-looking statement, whether as a result of new information, future developments or otherwise, except as required by applicable law.

View source version on businesswire.com:
https://www.businesswire.com/news/home/20260728505196/en/

Investor Contact
investors@qumulusai.com

Media Contact
media@qumulusai.com

Source: QumulusAI