IBM Newsroom

IBM and Groq Partner to Accelerate Enterprise AI Deployment with Speed and Scale

Partnership aims to deliver faster agentic AI capabilities through IBM watsonx Orchestrate and Groq technology, enabling enterprise clients to take immediate action on complex workflows

ARMONK, N.Y. and MOUNTAIN VIEW, Calif., Oct. 20, 2025 /PRNewswire/ -- IBM (NYSE:IBM) and Groq today announced a strategic go-to-market and technology partnership designed to give clients immediate access to Groq's inference technology, GroqCloud, on watsonx Orchestrate – providing clients high-speed AI inference capabilities at a cost that helps accelerate agentic AI deployment. As part of the partnership, Groq and IBM plan to integrate and enhance Red Hat open source vLLM technology with Groq's LPU architecture. IBM Granite models are also planned to be supported on GroqCloud for IBM clients.

Enterprises moving AI agents from pilot to production still face challenges with speed, cost, and reliability, especially in mission-critical sectors like healthcare, finance, government, retail, and manufacturing. This partnership combines Groq's inference speed, cost efficiency, and access to the latest open-source models with IBM's agentic AI orchestration to deliver the infrastructure needed to help enterprises scale.

Powered by its custom LPU, GroqCloud delivers over 5X faster and more cost-efficient inference than traditional GPU systems. The result is consistently low latency and dependable performance, even as workloads scale globally. This is especially powerful for agentic AI in regulated industries.

For example, IBM's healthcare clients receive thousands of complex patient questions simultaneously. With Groq, IBM's AI agents can analyze information in real-time and deliver accurate answers immediately to enhance customer experiences and allow organizations to make faster, smarter decisions.

This technology is also being applied in non-regulated industries. IBM clients across retail and consumer packaged goods are using Groq for HR agents to help enhance automation of HR processes and increase employee productivity.

"Many large enterprise organizations have a range of options with AI inferencing when they're experimenting, but when they want to go into production, they must ensure complex workflows can be deployed successfully to ensure high-quality experiences," said **Rob Thomas, SVP, Software and Chief Commercial Officer at IBM** "Our partnership with Groq

underscores IBM's commitment to providing clients with the most advanced technologies to achieve AI deployment and drive business value."

"With Groq's speed and IBM's enterprise expertise, we're making agentic AI real for business. Together, we're enabling organizations to unlock the full potential of AI-driven responses with the performance needed to scale," said **Jonathan Ross, CEO & Founder at Groq**. "Beyond speed and resilience, this partnership is about transforming how enterprises work with AI, moving from experimentation to enterprise-wide adoption with confidence, and opening the door to new patterns where AI can act instantly and learn continuously."

IBM will offer access to GroqCloud's capabilities starting immediately and the joint teams will focus on delivering the following capabilities to IBM clients, including:

- **High speed and high-performance inference that unlocks** the full potential of AI models and agentic AI, powering use cases such as customer care, employee support and productivity enhancement.
- **Security and privacy-focused AI deployment** designed to support the most stringent regulatory and security requirements, enabling effective execution of complex workflows.
- **Seamless integration with IBM's agentic product**, watsonx Orchestrate, providing clients flexibility to adopt purpose-built agentic patterns tailored to diverse use cases.

The partnership also plans to integrate and enhance Red Hat open source vLLM technology with Groq's LPU architecture to offer different approaches to common AI challenges developers face during inference. The solution is expected to enable watsonx to leverage capabilities in a familiar way and let customers stay in their preferred tools while accelerating inference with GroqCloud. This integration will address key AI developer needs, including inference orchestration, load balancing, and hardware acceleration, ultimately streamlining the inference process.

Together, IBM and Groq provide enhanced access to the full potential of enterprise AI, one that is fast, intelligent, and built for real-world impact.

Statements regarding IBM's and Groq's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

About IBM
IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs, and gain a competitive edge in their industries. Thousands of governments and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently, and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity, and service. Visit www.ibm.com for more information.

About Groq
Groq is the inference infrastructure powering AI with the speed and cost it requires. Founded in 2016, Groq developed the LPU and GroqCloud to make compute faster and more affordable. Today, Groq is trusted by over two million developers and teams worldwide and is a core part of the American AI Stack.

Media Contact:
Elizabeth Brophy
elizabeth.brophy@ibm.com

SOURCE IBM
