Fastly AI Accelerator Helps Developers Unleash the Power of Generative AI

December 16, 2024

Fastly expands support to include OpenAI ChatGPT and Microsoft Azure AI Foundry

SAN FRANCISCO--(BUSINESS WIRE)-- Fastly Inc. (NYSE: FSLY), a global leader in edge cloud platforms, today announced the general availability of Fastly AI Accelerator. A semantic caching solution created to address the critical performance and cost challenges faced by developers with Large Language Model (LLM) generative AI applications, Fastly AI Accelerator delivers an average of 9x faster response times.[1] Initially released in beta with support for OpenAI ChatGPT, Fastly AI Accelerator is also now available with Microsoft Azure AI Foundry.

"AI is helping developers create so many new experiences, but too often at the expense of performance for end-users. Too often, today's AI platforms make users wait," said Kip Compton, Chief Product Officer at Fastly. "With Fastly AI Accelerator we're already averaging 9x faster response times and we're just getting started. [1] We want everyone to join us in the quest to make AI faster and more efficient."

Fastly AI Accelerator can be a game-changer for developers looking to optimize their LLM generative AI applications. To access its intelligent, semantic caching abilities, developers simply update their application to a new API endpoint, which typically only requires changing a single line of code. With this easy implementation, instead of going back to the AI provider for each individual call, Fastly AI Accelerator leverages the Fastly Edge Cloud Platform to provide a cached response for repeated queries. This approach helps to enhance performance, lower costs, and ultimately deliver a better experience for developers.

"Fastly AI Accelerator is a significant step towards addressing the performance bottleneck accompanying the generative AI boom," said Dave McCarthy, Research Vice President, Cloud and Edge Services at IDC. "This move solidifies Fastly's position as a key player in the fast-evolving edge cloud landscape. The unique approach of using semantic caching to reduce API calls and costs unlocks the true potential of LLM generative AI apps without compromising on speed or efficiency, allowing Fastly to enhance the user experience and empower developers."

Existing Fastly customers can add AI Accelerator directly from their Fastly accounts. To learn more and get started, visit fastly.com/ai.

About Fastly, Inc.

Fastly's powerful and programmable edge cloud platform helps the world's top brands deliver online experiences that are fast, safe, and engaging through edge compute, delivery, security, and observability offerings that improve site performance, enhance security, and empower innovation at global scale. Compared to other providers, Fastly's powerful, high-performance, and modern platform architecture empowers developers to deliver secure websites and apps with rapid time-to-market and demonstrated, industry-leading cost savings. Organizations around the world trust Fastly to help them upgrade the internet experience, including Reddit, Neiman Marcus, Universal Music Group, and SeatGeek. Learn more about Fastly at https://www.fastly.com, and follow us @fastly.

Forward-Looking Statements

This press release contains "forward-looking" statements that are based on our beliefs and assumptions and on information currently available to us on the date of this press release. Forward-looking statements may involve known and unknown risks, uncertainties, and other factors that may cause our actual results, performance, or achievements to be materially different from those expressed or implied by the forward-looking statements. These statements include, but are not limited to, those regarding the ability of Fastly AI Accelerator to help developers enhance performance, deliver faster response times, reduce costs, and improve user experience. Except as required by law, we assume no obligation to update these forward-looking statements publicly or to update the reasons actual results could differ materially from those anticipated in the forward-looking statements, even if new information becomes available in the future. Important factors that could cause our actual results to differ materially are detailed from time to time in the reports Fastly files with the Securities and Exchange Commission ("SEC"), including without limitation Fastly's Annual Report on Form 10-K for the year ended December 31, 2023 and our Quarterly Reports on Form 10-Q. Copies of reports filed with the SEC are posted on Fastly's website and are available from Fastly without charge.

Source: Fastly, Inc.

[1] Responses from Fastly AI Accelerator semantic cache were served 9 times faster on average compared to those served without the AI Accelerator, calculated using all beta customer and demo traffic between October 15, 2024 and November 27, 2024.

Media Contact
Spring Harris
press@fastly.com

Investor Contact
Vernon Essi, Jr.
ir@fastly.com

Source: Fastly, Inc.