September 2, 2026

Equinix Accelerates AI Inference for Enterprises with NVIDIA and Together AI

Equinix Inference Exchange combines NVIDIA Enterprise Reference Architectures, Together AI's inference platform and Equinix's global infrastructure to optimize deployment speed, flexibility and cost efficiency

REDWOOD CITY, Calif., Sept. 2, 2026 /PRNewswire/ -- Equinix, Inc. (Nasdaq: EQIX), the world's digital infrastructure company®, today announced a significant expansion of its longtime collaboration with NVIDIA to deliver Equinix® Inference Exchange, a distributed AI inference program for global enterprises, alongside a new collaboration with Together AI.

As AI scales across models, providers and geographies, where inference runs is a strategic imperative that determines performance, cost and governance. Equinix Inference Exchange will give enterprises a faster path from AI experimentation to production, with secure, low-latency connectivity to the data, users and ecosystem they depend on.

This collaboration brings together NVIDIA's validated Enterprise Reference Architectures with Together AI's inference platform, supporting more than 200 open-source models. Delivered through Equinix's global data centers, it will provide connectivity to clouds, networks and AI providers through Equinix Fabric®.

The solution will be announced today at Equinix Horizon, the company's inaugural customer and partner event, alongside Equinix® Fabric One™, which will make it easier for enterprises to connect across globally distributed AI environments.

"AI is transforming enterprise technology at extraordinary speed, and the infrastructure decisions enterprises make today will define their competitive position for years to come. Equinix is uniquely positioned to deliver what this moment demands based on our nearly three decades building the trusted exchange where the world's enterprises run, connect and orchestrate their most critical workloads," said Adaire Fox-Martin, Chief Executive Officer and President, Equinix. "Our longtime relationship with NVIDIA delivers the accelerated computing foundation at the heart of modern AI, while Together AI's commitment to open ecosystems gives enterprises the flexibility to scale on their terms. Equinix Inference Exchange will enable architectures that are neutral by design, open by default and engineered for exceptional performance."

"Equinix Inference Exchange turns the world's leading digital interconnection platform into a global fabric for AI inference," said Raj Mirpuri, vice president of global AI clouds and infrastructure ecosystem at NVIDIA. "As accelerated compute becomes a strategic asset class, combining NVIDIA's infrastructure & technology with Together AI's open-model inference platform and Equinix's global reach gives enterprises a powerful, distributed foundation to bring intelligence closer to their data, applications and customers—accelerating the next generation of intelligent services."

"Together AI was built on the conviction that open, accessible AI is what will define the industry moving forward, because enterprises shouldn't have to choose between model performance and operational flexibility," said Vipul Ved Prakash, co-founder and CEO, Together AI. "What we are building with Equinix and NVIDIA proves that model choice and performance are not trade-offs. They are the foundation of enterprise AI done right."

Where Inference Runs Matters

The pace of enterprise AI adoption is outrunning the infrastructure needed to support it. As enterprise AI moves from experimentation to production, inference increasingly needs to run closer to the users, data and applications it serves across clouds, models, providers and geographies. This shift requires enterprises to determine not only how to deploy AI infrastructure, but where it should run and how it connects to the data, applications and workloads it depends on.

Managing these distributed inference deployments introduces significant operational complexity at precisely the moment enterprises need greater control and visibility.

"Performance, cost and governance have become strategic considerations as AI workloads grow more distributed across providers, data sources and environments," said Nick Patience, Vice President & Practice Lead, AI Platforms, The Futurum Group. "Organizations are increasingly focused on where inference runs and how quickly it can be deployed into production. Solutions that simplify inference deployment while preserving flexibility will become increasingly important to achieve business outcomes."

Equinix brings unmatched scale and ecosystem density to this challenge, with more than 280 data centers across 77 metros, 230 cloud on-ramps and over 10,500 businesses interconnected on its neutral exchange. Eight of the top 10 AI model providers and nine of the top 10 AI clouds are deployed with Equinix, underscoring the company's position at the center of the AI ecosystem.

Built for Choice and Flexibility

Together AI is the latest addition to Equinix's expansive AI ecosystem, bringing open-model flexibility and choice to enterprises deploying AI at scale. The solution combines three complementary layers designed to simplify distributed AI inference:

- **Equinix** provides the infrastructure foundation, including power, advanced cooling and day-two operations, connected through Equinix Fabric to the clouds, networks and AI providers that inference depends on.
- **NVIDIA** anchors the build with its Enterprise Reference Architectures and AI infrastructure purpose-built to maximize AI factory throughput and minimize token cost.

- **Together AI** runs the platform on top, supporting both multitenant deployments for shared efficiency and dedicated single-tenant environments for workloads that require dedicated capacity.

Built on Equinix Fabric, the solution will connect to inference providers across major metros worldwide, cutting time-to-first-token. It also will connect to an expansive ecosystem of clouds, networks and AI providers, reducing deployment complexity.

Designed for Modern Enterprise Inference

The solution aims to support a broad range of enterprise inference scenarios, including:

- **Metro edge inference**: For organizations that need inference running closer to users and data, enabling lower-latency AI experiences while leveraging the security, operational scale and global reach of Equinix.
- **Open model migration**: For enterprises moving workloads from closed, proprietary models to open-source alternatives to control cost and avoid lock-in, the solution will provide a direct, low-friction path to run that migration in production, with Together AI's open-model platform reachable over the same interconnected fabric enterprises already use to reach their other providers.
- **Sovereign AI**: For enterprises operating in regulated industries or specific geographies, the solution will enable AI workloads to run in locations that support data residency and sovereignty requirements, providing a simpler path to deploying AI at scale while maintaining control over where data and inference are processed.

Equinix Inference Exchange will be available starting in Q1 2027.

Additional Resources

- Token Optimization Begins with Choice [Analyst Report]
- The Coordination Economy: How Enterprises Really Build AI Value [Blog]
- Equinix Inference Exchange [Product Page]
- Equinix Inference Exchange Product Release Note [Product Release Note]
- Equinix Horizon Event Page [Event Page]

About Equinix
Equinix, Inc. (Nasdaq: EQIX) shortens the path to boundless connectivity anywhere in the world. Its digital infrastructure, data center footprint and interconnected ecosystems empower innovations that enhance our work, life and planet. Equinix connects economies, countries, organizations and communities, delivering seamless digital experiences and cutting-edge AI—quickly, efficiently and everywhere.

Forward-Looking Statements

This press release contains forward-looking statements that involve risks and uncertainties. Actual results may differ materially from expectations discussed in such forward-looking statements. Factors that might cause such differences include, but are not limited to, risks to our business and operating results related to the current inflationary environment; foreign currency exchange rate fluctuations; stock price fluctuations; increased costs to procure power and the general volatility in the global energy market; the challenges of building and operating IBX® and xScale® data centers, including those related to sourcing suitable power and land, and any supply chain constraints or increased costs of supplies; the challenges of developing, deploying and delivering Equinix products and solutions; unanticipated costs or difficulties relating to the integration of companies we have acquired or will acquire into Equinix; a failure to receive significant revenues from customers in recently built out or acquired data centers; failure to complete any financing arrangements contemplated from time to time; competition from existing and new competitors; the ability to generate sufficient cash flow or otherwise obtain funds to repay new or outstanding indebtedness; the loss or decline in business from our key customers; risks related to our taxation as a REIT; risks related to regulatory inquiries or litigation; and other risks described from time to time in Equinix filings with the Securities and Exchange Commission. In particular, see recent and upcoming Equinix quarterly and annual reports filed with the Securities and Exchange

Commission, copies of which are available upon request from Equinix. Equinix does not assume any obligation to update the forward-looking information contained in this press release.



EQUINIX

WHERE OPPORTUNITY CONNECTS

View original content to download multimedia:https://www.prnewswire.com/news-releases/equinix-accelerates-ai-inference-for-enterprises-with-nvidia-and-together-ai-302867226.html

SOURCE Equinix, Inc.