

Envoy Today

*What's New in Managing Cloud-native and AI
Workload Traffic*

10 Years of Envoy · KubeCon + CloudNativeCon Europe 2026 ·
Project Lightning Talk



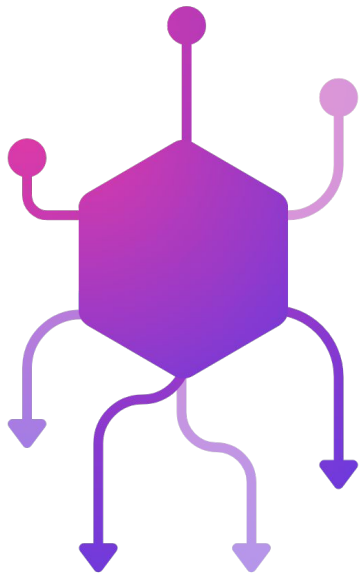
KubeCon



CloudNativeCon

Europe 2026

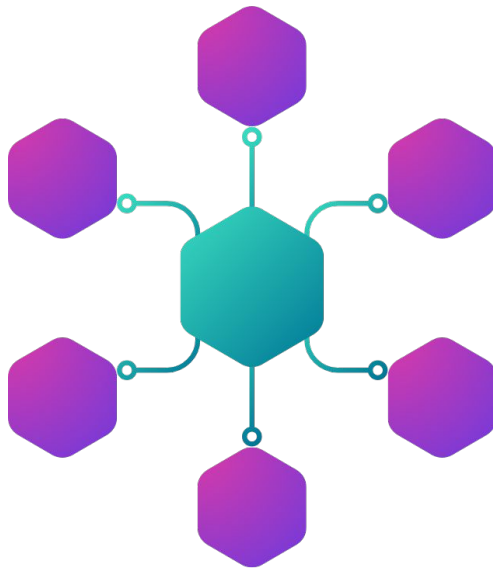
The Envoy Ecosystem



Envoy Proxy

The Data Plane

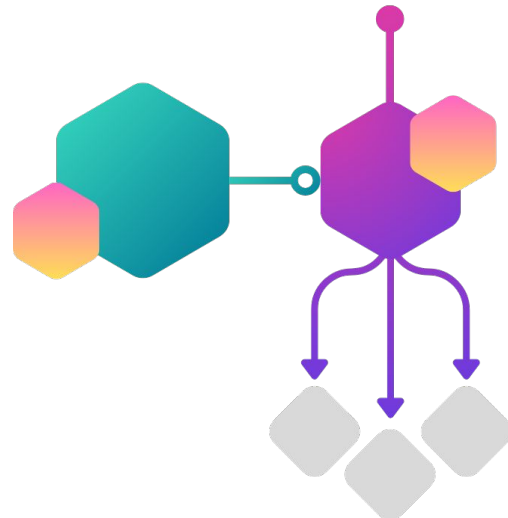
High-performance L7 proxy. The foundation for 10 years.



Envoy Gateway

The Kubernetes Control Plane

Kubernetes-native gateway. Makes Envoy Proxy adoption effortless.



Envoy AI Gateway

AI Traffic Extension

Data + control plane extension for AI/LLM traffic management.



KubeCon



CloudNativeCon

Europe 2026

Why AI Traffic Is Different



- **Streaming inference:** long-lived connections, not request-response
- **Token-based economics:** single request can cost wildly different amounts depending on the model and prompt size
- **Multi-hop agentic workflows:** one request triggers chains of tool calls across backends
- **Per-model observability:** not just “this endpoint is slow” but which model, how many tokens it uses and at what cost

What's Next Across the Ecosystem



Envoy Proxy

Rust Extensions: Broader support for new Envoy extensions in Rust.

Inference improvements Request buffering & cross-service failover.

Native Protocols: Core MCP and A2A awareness.

Connectivity: Reverse tunnels & xDS federation

Performance improvements and filter ecosystem growth

Envoy Gateway

Expanding Gateway API (v1.5) conformance to serve as a reference implementation for Kubernetes-native ingress traffic.

Improved traffic management across distributed environments with new load balancing features

Richer extensibility points allowing for more custom logic and filter integrations.

Envoy AI Gateway

API graduation (alpha → beta), marking the transition to stable, production-ready interfaces.

v1.0 (GA) milestone & API stability to provide long-term reliability

Open Responses API support for out-of-the-box compatibility with OpenAI-standard agents.

Token budget & rate limit policies for fine-grained control over LLM usage and costs.

Richer telemetry for LLM backends including per-model metrics and detailed latency tracking.



KubeCon



CloudNativeCon

Europe 2026

Adoption That Speaks for Itself



2.1M+

Helm chart pulls
per 30-day window

5M+

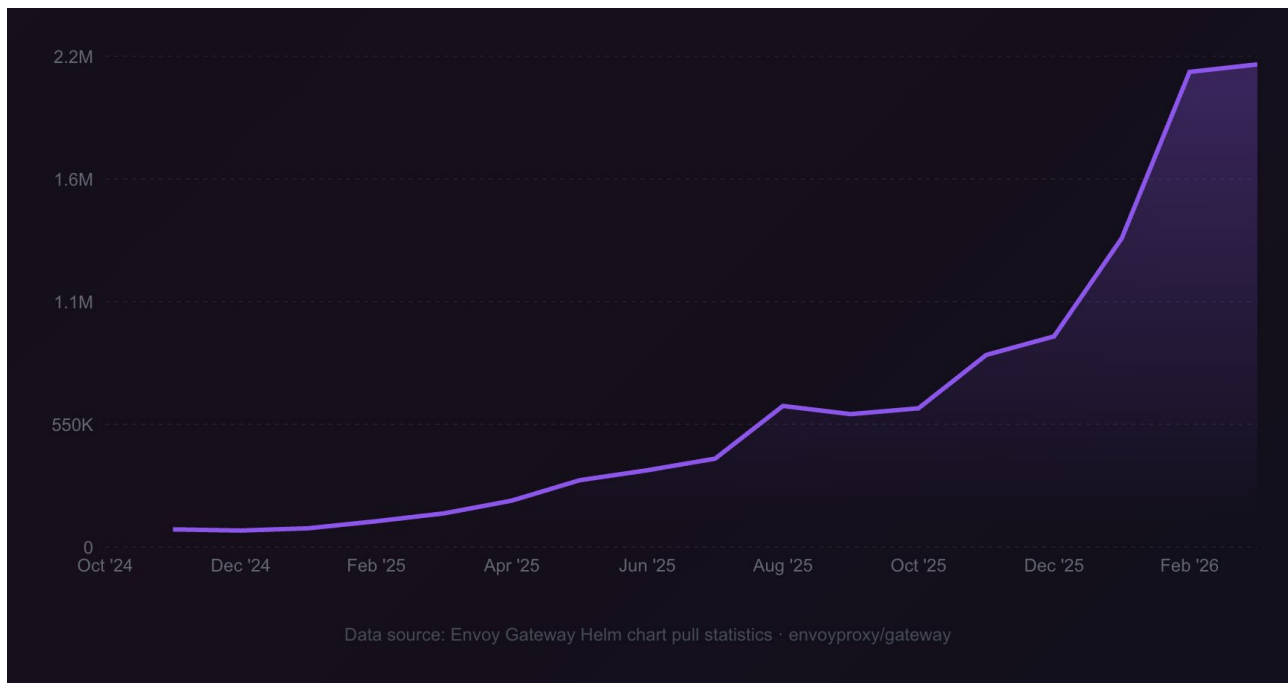
Docker image pulls
Gateway total

27K+

GitHub stars
Envoy Proxy

300+

contributors
Gateway project



KubeCon



CloudNativeCon

Europe 2026

Find Envoy at Kubecon



Mar 24 Tuesday

- | | | |
|-------|--|----------|
| 14:30 | Enterprise-Scale Migrations Using
Agentic Workflows with Human-in-the-
loop - Alvaro Saurin & Jose M Navarro,... | External |
| 16:15 | Smart Routing at Scale: How Spotify's
XDS Control Plane Cut 75% of Cross-
Zone Traffic - Yannick Epstein & Anya... | External |
| 16:15 | Celebration of Newly Graduated Projects
+ Project Birthdays 🎉 | External |

Mar 25 Wednesday

- | | | |
|-------|----------------------------------|----------|
| 11:00 | Cloud Native Theater EnvoyCon: | External |
|-------|----------------------------------|----------|



Meet maintainers and contributors at
the Project Pavillion

Show up by the Envoy Kiosk from
Tuesday to Thursday

P-11A

Join us on the Envoy Slack



KubeCon



CloudNativeCon

Europe 2026

Thank you!



karol@kubermatic.com



cnvergence



Karol Szwaj



KubeCon



CloudNativeCon

Europe 2026