

You're Doing AI Gateway Wrong. Here's Why.

Why the Next Generation
of AI Gateways Must
Think Beyond Routing



Table of Contents

Overview	3
Why Simple Routing isn't Enough	4
Introducing Smart Routing – Routing by Intent	7
Next-gen AI GW Establishes Full Governance	9
Agentic AI Makes a Next-gen AI Gateway Required	11
What to Look for When Evaluating an AI Gateway	12
A10 AI Gateway Addresses Cost, Provides Foundation for AI Security	13
Conclusion	14
About A10 Networks	14

Overview

AI adoption is moving faster than most organizations can manage. Employees are using AI assistants. Developers are integrating LLMs into applications. Business teams are experimenting with specialized models. Agentic AI, the latest innovation and concern, can reason through tasks, call tools, access data, and interact with business systems, exponentially increasing productivity while growing the AI attack surface.

And underneath all of it is a rapidly expanding AI ecosystem

- Multiple LLM providers
- Multiple models from each provider
- Specialized models for coding, reasoning, summarizing, and other workloads
- AI applications and agents
- Different users and teams with different requirements
- Different budgets and consumption limits
- Increasingly complex security requirements

Organizations need visibility and governance across this environment. That makes the AI gateway a crucial part of the AI ecosystem. The problem is that many organizations are approaching the AI gateway as if AI traffic is still as straight forward as application or API traffic. These first-generation AI gateways are focused on routing requests, tracking usage, applying static rate limits, etc. This approach sufficed when organizations were simply connecting applications to an LLM. The reality has changed, however. AI gateways need to become the intelligent control plane for enterprise AI.

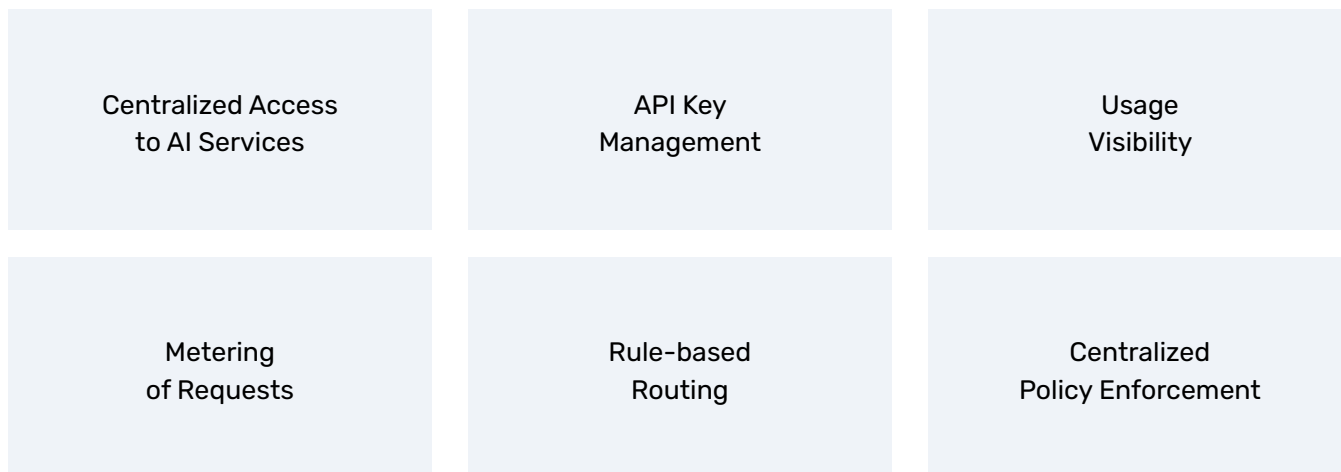


A traditional AI gateway decides where traffic goes.
A next-gen AI gateway decides where traffic belongs.

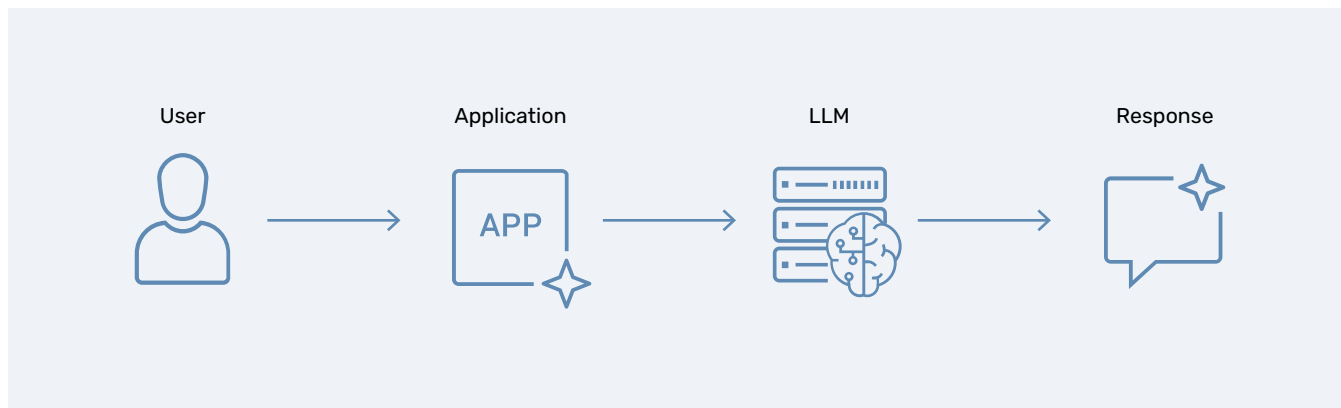
Why Simple Routing isn't Enough

The original purpose of an AI gateway was straightforward. Organizations needed a centralized way to connect applications and users to AI providers. Instead of every application maintaining its own provider credentials and individual integrations to AI services, the gateway could sit between users and AI services.

This created several immediate benefits:



These capabilities remain important, but enterprise AI has evolved. AI is no longer simply:



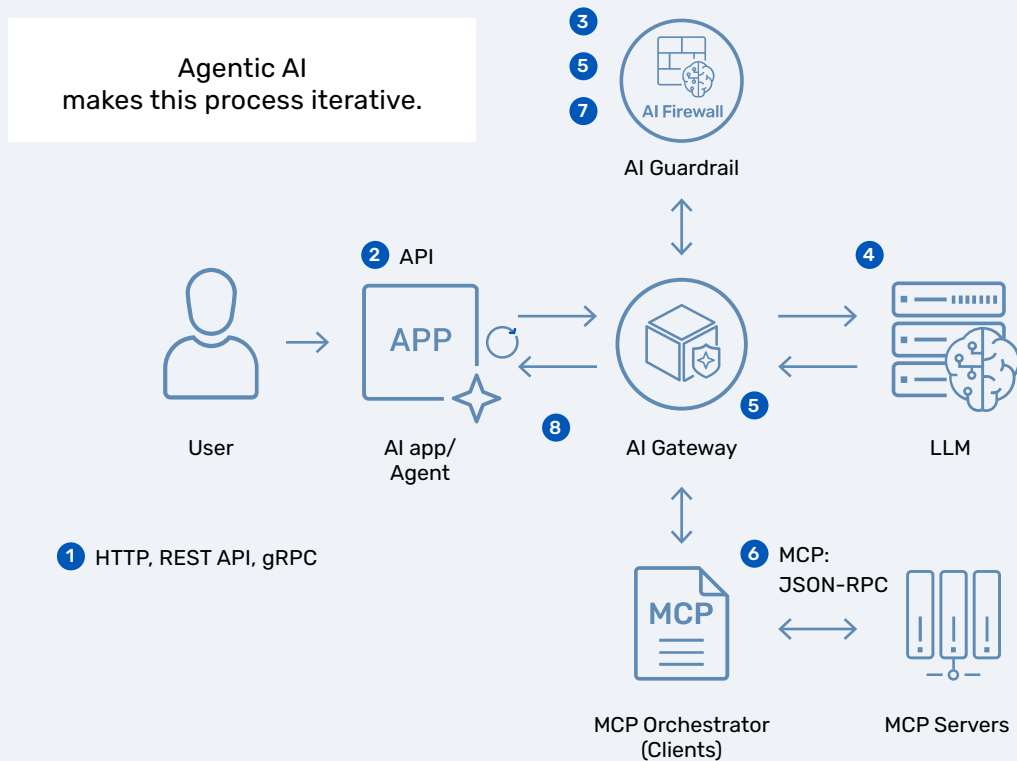
Today's AI environment, led by advances in agentic AI, can reason through tasks, call tools, access data, and iterate until the task is complete.

As a result, the AI environment has become much more dynamic.	A developer may need access to a specialized coding model	A marketing team may need a general-purpose model
An application may need a low-latency model for millions of routine requests	An AI agent may require access to multiple models and tools	A data science team may need access to a high-performance reasoning model

The next-gen AI gateway must go beyond simple request and response. It now must consider questions.

Who is the request for?	 Which model is appropriate? • Budgeting • Reasoning level • Organizational assignment
What is the goal or task at hand?	
How difficult is the request?	
Is this user or team authorized to use that model?	
What is the current usage traffic and cost, and how?	

These questions require next-gen AI gateways to perform more complex decision-making. And the AI traffic flow may look more like this:



1. AI Agent receives a user prompt. The agent formats an LLM request and forwards it to the AI Gateway.
2. AI Gateway intercepts the request, performs internal corporate authentication, maps the agent's specific identity, checks rate limits, and validates organizational budgets.
3. AI Guardrail executes inbound prompt check.
4. LLM evaluates the request and determines it needs external data which can be accessed through its available tool schema (provided over MCP) and returns a tool call intent back through the gateway, via AI Gateway.
5. The Gateway captures the LLM's response intent, then:
 - Runs the second guardrail
 - Verifies the agent's role-based access control for the MCP server
 - Translates the payload into standard MCP JSON-RPC protocol messages and sends
6. The request reaches the designated MCP server. The server runs the query and returns the raw context back to the gateway.
7. Post-usage of tool activates the AI Firewall again to avoid indirect prompt injection.
8. The AI Gateway passes the vetted tool data back to the AI Agent. The agent passes the new data back to the LLM to formulate the final answer or trigger the next iterative step in its plan.

Introducing Smart Routing – Routing by Intent

AI requests are different from traditional application traffic because they introduce another dimension of considerations – intent and complexity. The request itself matters.

Let's look at two requests.

"Summarize
this paragraph."

vs.

"Analyze these financial statements, identify anomalies, compare them against historical trends, explain the likely causes, and generate a five-year forecast."

Both can be requests for AI, but they can't be treated the same way. The first request is likely coming from someone looking to get a quick glimpse and executive-level summary. It would be well-suited to be directed toward a smaller, faster, less expensive model.

The second request is likely coming from someone looking for deeper analysis and decision support. Because it's asking the AI to reason through multiple considerations, comparisons, and even to generate a forward-looking forecast, it may require a more capable reasoning model.

Routing all requests to the most powerful model may deliver excellent results, but at unnecessary cost. Routing all requests to the cheapest model may reduce costs, but at the expense of quality and reliability. Striking a delicate balance between the two means implementing dynamic and intelligent routing – understand the goal, then match the right model for the right request.

A traditional AI gateway might provide routing rules such as:

Send traffic to model A

Send team B to provider C

Send 20% of requests to model D

Failover to model E if model D is unavailable

These policies are helpful but are static in nature. They tell traffic where to go but a modern AI gateway must tell traffic where it belongs.

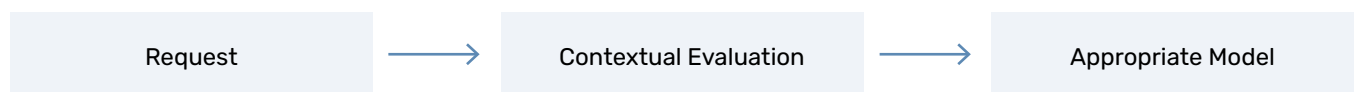
A next-gen AI gateway must use reasoning-based intelligent routing to evaluate the complexity and type of an AI request. It can then route simple requests toward more cost-efficient models while directing complex, reasoning-intensive requests toward more capable models.

This creates a fundamentally different operating model.

Instead of:



The organization gets:



Contextual evaluations can consider things like:

Request Complexity	User Identity	Budget	Security Policies
Request Type	Team	Model Availability	Other Configured Routing Context



A next-gen AI gateway is not simply deciding where traffic goes, it is deciding where traffic belongs. This is smart routing.



Next-gen AI GW Establishes Full Governance

Routing requests is only part of the problem. Full governance of AI is needed. This is controlling who can use AI, what they can access, how much they can consume, and under what conditions: identity, budgeting, security.

Traditional AI gateways may have rudimentary identity-based policies, but next-gen AI gateways will require more complex identity-based policies, akin to digital fingerprinting, which can double as a security add. AI governance becomes difficult when organizations treat everyone as an anonymous consumer of a shared AI resource. A next-gen AI gateway allows organizations to apply access and routing policies at the user and team level.

The A10 AI Gateway enables organizations to apply AI access and routing policies to individual users and groups with the capability to synchronize with existing local directory systems. For example, organizations can route engineering users to specialized coding models while directing other teams to models appropriate for their roles and needs. Individual users can also receive unique permissions when their requirements differ from those of their team.

This transforms AI governance from simply controlling which models are available to controlling who is allowed to use which models, for what purposes, and under what circumstances. Such granular level of control is crucial for supporting enterprise AI use.

Cost must be governed with the same level of context. It cannot just be a reporting function telling teams how much was spent on AI. Smart routing capabilities must be able to manage AI traffic depending on the model, request volume, complexity, and other contextual factors of the workload.

Next-generation AI gateways establish budgets and enforce spending policies at a more granular level. Organizations need to be able to determine how much individual users, teams, applications, or models are allowed to consume, while also having the ability to enforce those limits. Leveraging smart routing provides these capabilities, while determining what is the most appropriate and efficient use of an organization's resources.

The A10 AI Gateway can provide real-time, per-request cost tracking, per-model and per-team token budgets, hard-limit and soft-limit alerts, and other more granular details. This allows organizations to move from understanding how much AI costs them, to controlling how AI resources are consumed.

A10 AI Gateway smart routing evaluates the complexity and type of each AI request through an 11-step decision tree and uses reasoning to route simpler tasks toward more cost-efficient models while directing complex, reasoning-intensive tasks toward more capable models.



Fundamentally, next-gen AI gateways are now taking a different approach. Given the digital fingerprint of the requester, the request, available resources, organizational constraints, and other context, the next-gen AI gateway can determine which model is the most appropriate. The AI Gateway decides where traffic goes. The next-gen AI gateway decides where traffic belongs.

The final part of AI governance is security. There is a term in the AI security industry known as the lethal trifecta. This describes three capabilities: the ability to access private data, the ability to reason over untrusted content, and the ability to take action.

Each of these capabilities creates risk. When all three are combined, the potential attack surface becomes much larger. Imagine an AI agent that has access to an organization's private data. Now give that agent the ability to reason over information coming from an untrusted source. Finally, give it the ability to act through tools and APIs. An attacker may no longer need to directly compromise an application to cause damage. They could influence the AI into accessing information it shouldn't or taking an action it wasn't supposed to do. AI security cannot not be solved by one product or solution.

Different components of the AI stack solve different problems.

.....

The AI gateway is responsible for orchestrating and governing AI operations

.....

The AI firewall addresses the natural-language security layer

.....

MCP controls govern which tools AI agents can access

.....

API and application security protect the underlying applications and APIs that AI agents ultimately interact with

.....

This paper focuses on the interactions between the AI gateway and AI firewall.

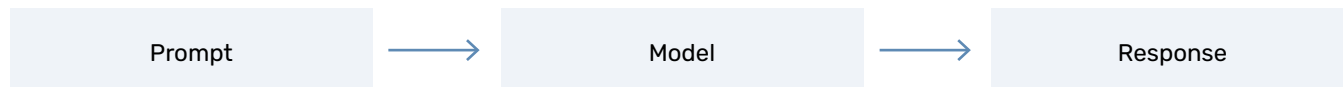
The next-gen AI gateway establishes the boundaries and context that the rest of the security stack depends on. It can associate an interaction with the identity making the request and determine what that identity is permitted to access. The AI firewall inspects the interaction itself for threats such as prompt injection, jailbreaks, sensitive-data leakage, policy violations, and other threats.

The AI gateway determines the identity of the requester and enforces the boundaries of what a requester can do. The AI firewall determines if an interaction is malicious.

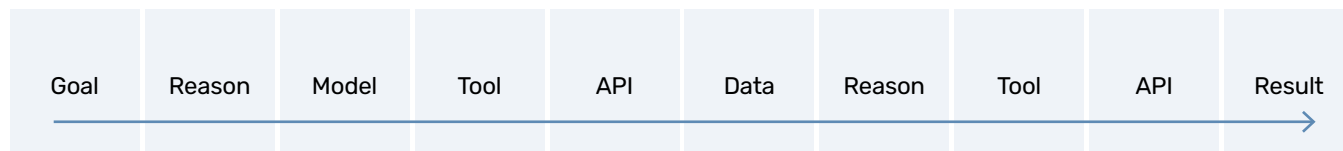
The next-gen AI gateway is an important precursor to implementing AI security by establishing the identity, access, and governance context that security controls need. By working in tandem, an AI firewall + gateway change AI security from identifying what happened, to understanding who did it and whether they were supposed to be able to do it.

Agentic AI Makes a Next-gen AI Gateway Required

The transition from conversational AI to agentic AI makes the need for contextual governance even more urgent. **With traditional conversational AI, the workflow is relatively straightforward:**



An agentic AI workflow can look very different:



The agent may make multiple decisions before completing a single task, which means multiple model calls, multiple tool calls, multiple API interactions, multiple iterations of reasoning, etc. It is no longer just about lowering the cost per request. Organizations increasingly need to think about managing the cost per task, as well as the security implications of everything that happens during that task.

The more AI expands across an organization, the less practical it becomes to manage every model, application, agent, and tool independently. The gateway is the place where these interactions are brought together and governed.

Take this theoretical agentic AI workflow. A request first enters the application. The application determines that AI is required. This is where the AI gateway comes into play. It can identify the user or application making the request, determine which AI resources they are permitted to access, apply the appropriate policies, manage budgets, and route the request to the appropriate model. Additionally, by establishing identity and access control, and context, at the gateway, the underlying AI security architecture can begin answering questions such as:



Who made the request?



What were they trying to access?



Were they authorized to access it?



What model or tool was used?



Was the interaction consistent with organizational policies assigned to that identity or digital fingerprint?

As agentic AI transforms the priority from cost per request, to cost per task, and adds on the ability to iterate on a task until completion, managing every model, application, agent, and tool independently is impractical. The next-gen AI gateway brings together and governs this massive web of interactions.

What to Look for When Evaluating an AI Gateway

Here is a short checklist of questions an organization can ask when considering an AI gateway.



Can it intelligently route AI requests?

Can the gateway understand the complexity, type, and other contextual factors of a request and determine which model is most appropriate, rather than relying solely on static routing policies?



Can it tie AI access to identity?

Can organizations define different model access, routing policies, and permissions for different users and teams?



Can it control AI spending?

Can organizations establish granular budgets, rate limits, and spending controls, while also incorporating cost into the intelligent routing process?



Can it manage AI centrally?

Can organizations gain a centralized view of who is using AI, which models and providers they are using, how much they are consuming, and whether AI resources are being used efficiently?



Does it aid with AI security?

Although it's not the main player of AI security, does it establish the boundaries and permissions necessary as a crucial precursor to implementing security in this rapidly changing space?



Can it scale with the evolution of AI?

As organizations move from simple LLM interactions toward agents, tools, and increasingly complex AI workflows, can the gateway continue to provide the identity, governance, routing, and control necessary to manage that environment?

These questions help move the evaluation of an AI gateway beyond, "Can it route my AI traffic?" to the much more important question of, "Can it intelligently govern and manage how my organization uses AI?"

A10 AI Gateway Addresses Cost, Provides Foundation for AI Security

A10 AI Gateway was designed around the premise that AI traffic needs more than connectivity. It needs intelligence. It brings AI interactions into a centralized, governed environment where organizations can manage routing, costs, AI keys, access, and visibility. Core benefits include:

Identity-based AI Access – Give users and teams access to the AI resources appropriate for their roles, with individual policies available for users requiring unique permissions.

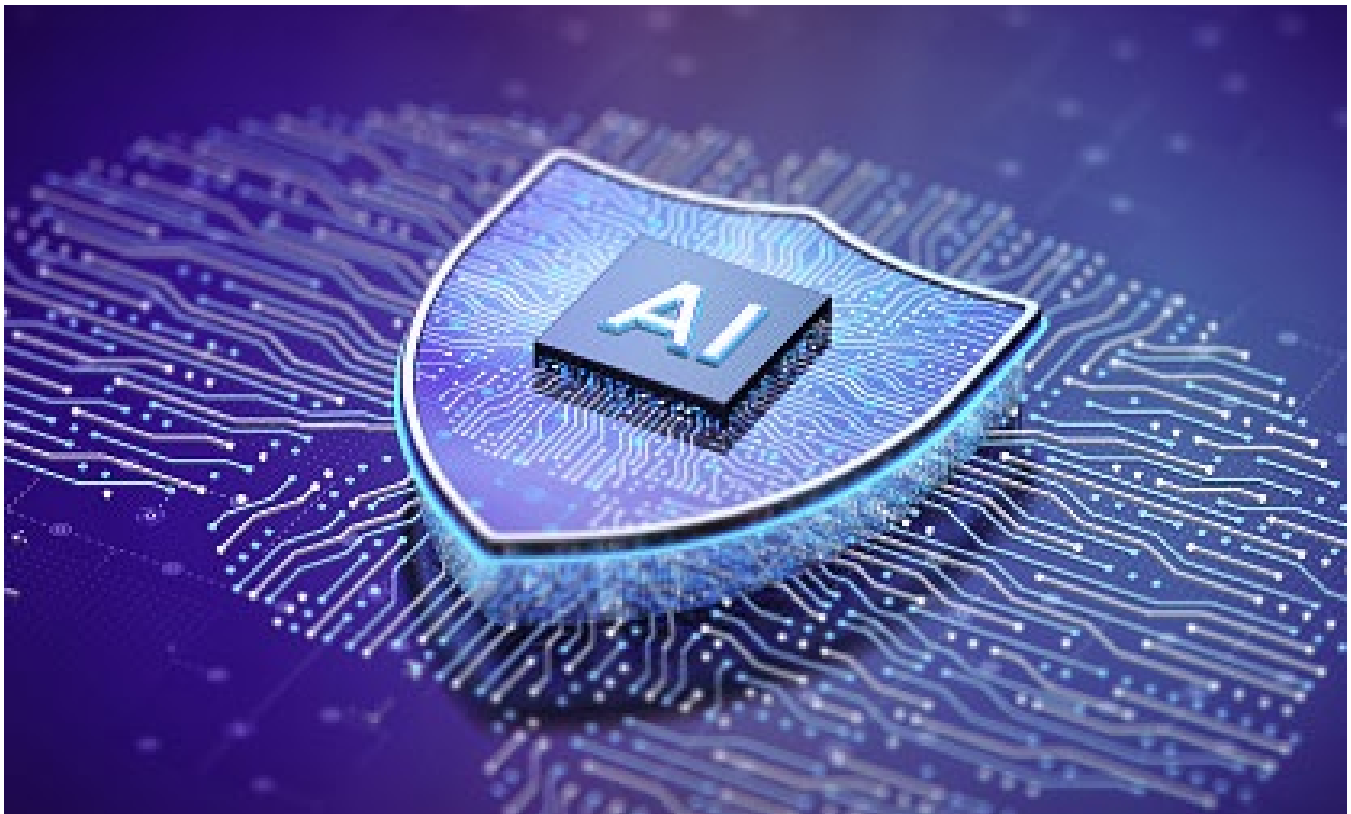
Reasoning-based Intelligent Routing – Evaluate request complexity and type through an 11-step decision tree and route requests toward models appropriate for their requirements.

Intelligent Cost Management – Track costs per request while establishing budgets and rate limits at the model, team, and key levels.

Centralized AI Control – Manage AI traffic across users, applications, agents, models, and providers through a centralized control plane.

The result is a gateway that

- Determines where requests belong, rather than where they go
- Optimizes cost per task, rather than just cost per request
- Enables the underlying AI security architecture to function



Conclusion

An AI gateway must decide where traffic belongs, not just where it goes.

The first generation of AI gateways helped organizations connect to AI. The next generation must help organizations intelligently manage AI.

The AI gateway allows organizations to make intelligent decisions about every AI interaction and answer questions like:

- Who is asking?
- What question/goal is being asked?
- Which model should handle it?
- Is the user authorized?
- How should this request be processed based on budget?
- How should security protocols be implemented?
- How is the overall AI workflow being governed?

When the AI gateway knows where the request belongs next, you're directing the AI traffic, lowering cost per task, and in control of the most uncontrollable growth since the internet boom.

About A10 Networks

A10 Networks provides security and infrastructure solutions for on-premises, hybrid cloud, and edge-cloud environments. Our 7000+ customers span global large enterprises and communications, cloud and web service providers who must provide business-critical applications and networks that are secure, available, and efficient. Founded in 2004, A10 Networks is based in San Jose, Calif. and serves customers globally.

For more information, visit A10Networks.com and follow us at [@A10Networks](https://twitter.com/A10Networks).



About A10

A10Networks.com

Contact Us

A10Networks.com/contact

©2026 A10 Networks, Inc. All rights reserved. A10 Networks, the A10 Logo, A10 Control, A10 Defend, A10 Harmony, Harmony, A10 Thunder, Thunder, ACOS, A10 SSL Insight, SSL Insight, SSLi, vThunder, ThreatX, ThreatX Protect, and TrojAI are trademarks or registered trademarks of A10 Networks, Inc. or its affiliates in the United States and other countries. All other trademarks are property of their respective owners. A10 Networks assumes no responsibility for any inaccuracies in this document. A10 Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice. For the full list of trademarks, visit: A10Networks.com/a10trademarks.