PERFORMANCE AND CAPACITY

VeriProof Agent Gateway performance

Throughput, latency, and evidence capture on Azure

Document ID VP-PERF-2026-01 | November 2026 | Version 1.0

For enterprise architects, platform engineers, and capacity planners

Executive summary

Moving an agent from pilot to production means enforcing your published rules and escalating exceptions to people while keeping durable evidence, without slowing routine traffic. This paper measures what one gateway sustains while it does that work.

One VeriProof Agent Gateway capped at **2 vCPUs and 512 MiB** sustained **239.76 calls per second in Policy only** and **119.05 calls per second in Capture evidence** in Azure. Each workflow made three calls, yielding **79.92** and **39.68 workflows per second**, respectively. [1]

In a separate benchmark with clustered human reviews, **Capture on escalation sustained 39.99 tool calls per second** and delivered evidence for 120 escalated actions. At the same load and review mix, it wrote **70% fewer bytes** to the journal disk than Capture evidence. [1]

Choose an evidence mode for each application's governance needs. Use dedicated journal storage and keep related protected requests on the same replica when scaling the fleet.

Contents

01 Choose an evidence mode 2
02 Where the gateway runs 3
03 Mixed-workflow throughput and latency 4
04 Tool calls with human review 5
05 Storage, memory, and CPU 6
06 Scale the gateway fleet 7
07 Measurement method 8

1. Choose an evidence mode

The gateway enforces each application's published rules in every mode. Set an organization default and apply authorized overrides by application and environment. [2]

Evidence mode	What VeriProof records	Primary benefit
Policy only	Aggregate traffic, reported model usage, and spend. Local dispatch records protect tool calls that require safe retries.	Enforce policy at high volume and monitor usage with minimal per-action storage.
Capture on escalation	Aggregate activity and case evidence for independent tool calls sent to human review before dispatch. It retains the reviewed action, decision, and recorded outcome.	Keep the evidence reviewers and auditors need for escalations while routine requests stay on the lighter path.
Capture evidence	Per-action decisions and session evidence for governed work, under the application's access and retention settings.	Investigate individual actions, follow governed workflows, and produce evidence for internal and external review.

Eligible Policy only model calls evaluate the published rules without a synchronous journal write. Tool calls with a durable retry obligation record dispatch before sending the request downstream. In the Linux restart test, the gateway stopped after dispatch and recognized the exact retry. The provider recorded one tool invocation. [1]

2. Where the gateway runs

The Agent Gateway runs in the customer's environment, between agents and model providers, Model Context Protocol (MCP) tools, or Agent2Agent (A2A) peers. The VeriProof control plane distributes published rules and coordinates reviews.

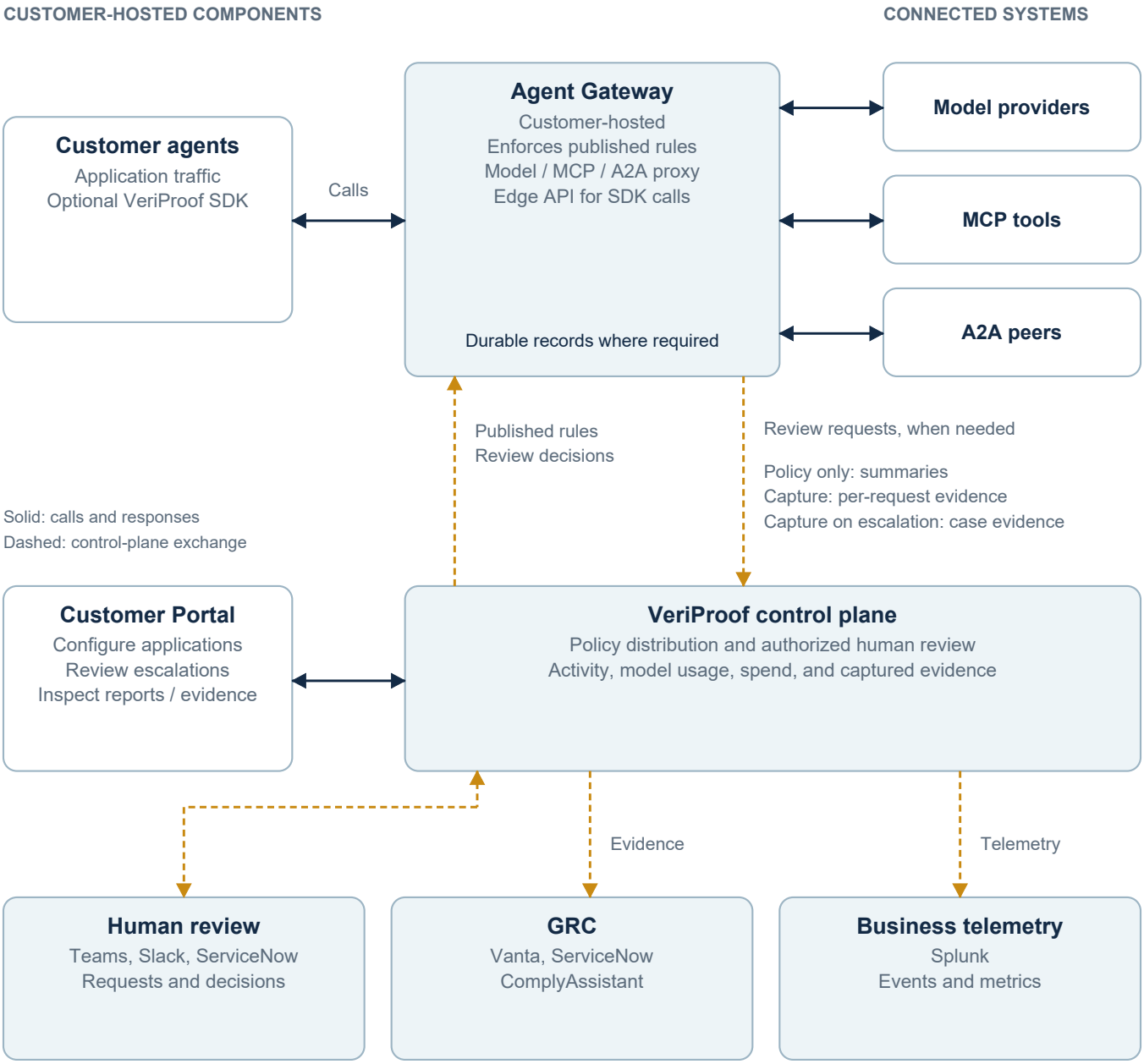

Figure 1. Single-gateway architecture. The Customer Portal provides application settings, review, and reports. The control plane connects human review, governance, risk and compliance, and business telemetry.

3. Mixed-workflow throughput and latency

Eight applications ran a three-call workflow through the gateway. Each workflow submitted an SDK policy decision through the Edge API, called a model, and invoked an independent MCP tool. These are the highest fully served points measured for that workload. [1]

Mixed-workflow mode	Calls per second	Workflows per second	Calls in window	Workflow p50 / p95 / p99 (ms)
Policy only	**239.76**	79.92	14,400	73.91 / 108.44 / 245.54
Capture evidence	**119.05**	39.68	7,200	106.05 / 485.93 / 646.74

The receiver reconciled model and tool effects in both modes. In the Capture evidence window it reconciled 12,000 decision records and 9,600 session records. The Policy only window delivered zero per-request decision or session records. The tests used the same process limits and workload mix. Policy only used a dedicated P10 journal disk, and Capture evidence used a dedicated P20. [1]

Model proxy latency

In a separate test at 50 requests per second, the Policy only JSON model proxy and direct provider control used the same request bytes, TLS, and synthetic provider. [1]

JSON model path	p50 (ms)	p95 (ms)	p99 (ms)
Direct to provider	1.01	1.32	2.04
Through the gateway	3.69	4.95	5.72
Added gateway latency	**2.69**	**3.63**	**3.68**

4. Tool calls with human review

The three-mode MCP test used the same gateway resources and dedicated P20 journal disk. Each high window ran 2,400 logical tool calls. The distributed and clustered profiles directed 5% of calls to review. Distributed reviews occurred at regular intervals. Clustered reviews arrived together in a burst. [1]

MCP evidence mode	Review demand	Tool calls/s	Routine p50 / p95 / p99 (ms)
Policy only	None	39.98	42.07 / 49.53 / 76.10
Capture evidence	None	39.96	118.75 / 154.04 / 217.64
Capture on escalation	None	39.98	41.02 / 47.78 / 74.45
Capture evidence	Distributed	39.95	120.64 / 176.92 / 330.67
Capture on escalation	Distributed	39.99	42.20 / 61.71 / 84.79
Capture evidence	Clustered	39.93	121.71 / 157.32 / 194.84
Capture on escalation	Clustered	39.99	41.73 / 50.89 / 81.43

Each review window reconciled **120 completed reviews** and exactly one upstream invocation per logical tool call. Capture on escalation delivered **240 session records** for the reviewed actions. Capture evidence delivered **4,800 session records** across the workload. Two records describe the dispatch and authorized response for a completed tool call. [1]

In the clustered test, reviewed calls took **225 ms p99** of active gateway time with Capture evidence and **301 ms p99** with Capture on escalation. The test's scoped case read returned the case **1.19 ms p99** and **17.71 ms p99**, respectively, after the gateway paused. The test measured synthetic approval separately from gateway work. [1]

5. Storage, memory, and CPU

Storage was the first resource to saturate in the Capture evidence disk study. Moving journal writes to a dedicated disk separated them from operating-system I/O. At the highest fully served Capture evidence mixed-workload rate, the process used 1.10 of 2 CPU cores on average. [1]

Mixed-workload resource	Policy only at 239.76 calls/s	Capture evidence at 119.05 calls/s
Mean gateway CPU	1.24 cores	1.10 cores
Peak gateway memory	512 MiB	512 MiB
Full I/O pressure	12.98 s	14.76 s
Full memory pressure	5.63 ms	4.34 ms
Dedicated journal disk	P10 Premium SSD	P20 Premium SSD

At the same clustered MCP review load, Capture on escalation reduced journal-disk traffic and I/O pressure. Device write counters include foreground and background storage activity. Peak memory is the cumulative peak through the measured window. [1]

Clustered MCP resource	Capture evidence	Capture on escalation
Mean gateway CPU	0.72 cores	0.45 cores
Peak gateway memory	349 MiB	512 MiB
Journal-disk writes	1.41 GiB	0.42 GiB
Device write operations	63,217	24,299
Instrumented SQLite commit events per call	5.67	5.18
Full I/O pressure	13.42 s	5.23 s

Size for the workload

- Give each replica a dedicated persistent journal disk. Check VM and disk I/O limits. Use journal commit latency, storage queue depth, and evidence delivery lag to choose between more IOPS and lower write latency. [3]
- Allow memory headroom for larger payloads, concurrent model streams, and bursts. Measure peak memory and pressure with your request mix.
- Add CPU where policy workers or serialization saturate. Check p95 and p99 latency after each capacity change.

6. Scale the gateway fleet

Place the load balancer in front of the gateway replicas. Each replica owns its persistent journal. Keep related requests together so the gateway can recognize the original dispatch and preserve its review state.

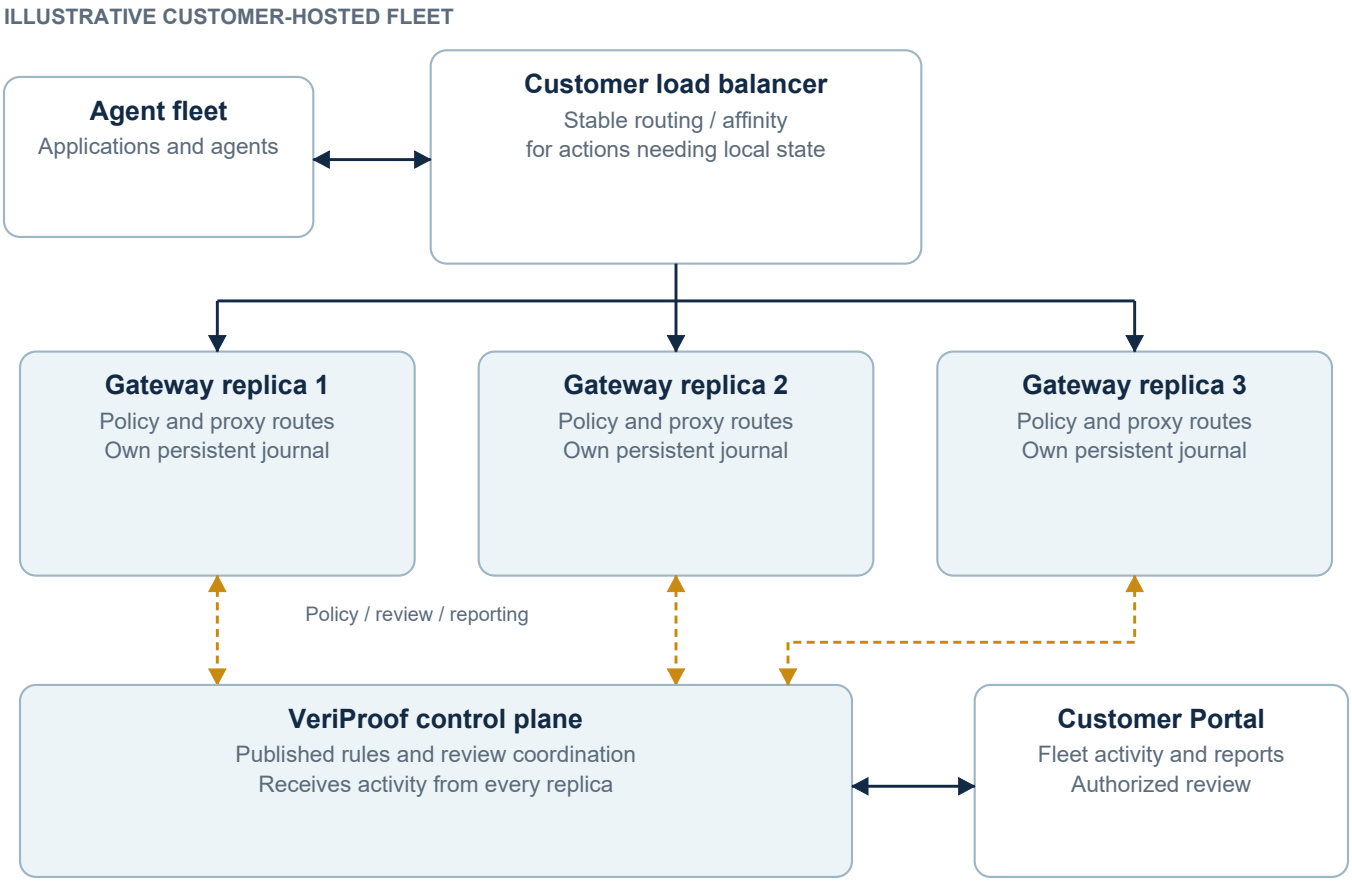

Figure 2. Load-balanced gateway fleet. Every replica exchanges published rules, review state, and reporting with the control plane. The Customer Portal presents fleet activity and authorized reviews.

Send new work to additional replicas as demand grows. During scale-down, drain protected work on the replica that owns it. Verify retry routing during a planned failover. Azure Load Balancer supports source-IP affinity. Select a scheme that distributes the workload when agents share an address and keeps each protected action's requests together. [4]

7. Measurement method

Test setting	Configuration
Gateway and driver	Azure Central US. The 8-vCPU service VM capped the gateway at 2 vCPUs and 512 MiB with swap disabled. A separate 2-vCPU load driver sent HTTPS requests over the private network.
Workloads	Eight applications shared the gateway in mixed and MCP tests. Four applications ran the model proxy comparison. Model and tool providers were controlled fixtures.
Throughput	Successful calls divided by elapsed time through completion. The high traffic windows were scheduled for 60 seconds. Workflow latency spans all three calls.
Model latency	The comparison subtracts corresponding percentiles from the direct-provider control.
Review and recovery	The harness verified callbacks, release, exact replay, provider effects, and evidence delivery. Synthetic approval time was measured separately.
Resources and sources	Linux counters measured CPU, memory, and pressure. Linux device counters and Azure metrics measured storage. The reviewed and zero-review MCP runs used separate builds on the same infrastructure. The performance record identifies both builds.

References

1. VeriProof. Agent Gateway Azure performance baseline and qualification record. November 2026.

2. VeriProof. Set up a gateway and choose an evidence mode.
https://docs.veriproof.app/customer-portal/gateway-setup/

3. Microsoft. Azure virtual machine and disk performance.
https://learn.microsoft.com/en-us/azure/virtual-machines/disks-performance

4. Microsoft. Azure Load Balancer distribution mode.
https://learn.microsoft.com/en-us/azure/load-balancer/load-balancer-distribution-mode

Plan a gateway with an engineer

Measure one production-shaped workflow under its target evidence mode and review demand, using realistic payload sizes. A technical review with our engineers can set initial capacity from its completion rate and tail latency.

Ask for a technical review
veriproof.app/contact