When AI Agents Move to the Cloud

Infrastructure and Security in the Agentic Era

IEEE CLOUD SUMMIT 2026

"OpenClaw went from a weekend project to 200,000+ GitHub stars in weeks... and suddenly developers were spinning up cloud VMs to run autonomous agents 24/7. Cloud infrastructure wasn't ready."

— The Catalyst of the Agentic Era

The Architecture Mismatch

Traditional Cloud

Built for predictability and ephemeral compute:

- **Stateless:** Containers spin up, process, and die.
- **Request-Driven:** Triggered by explicit user actions or webhooks.
- **Predictable Lifecycle:** Strict timeouts (e.g., AWS Lambda 15min limits).
- **Linear Flow:** Pre-planned microservice calls.

Agentic Workloads

Built for autonomy and continuous operation:

- **Stateful:** Require persistent memory across long horizons.
- **Proactive:** Wake up on timers, unprompted by users.
- **Unpredictable Lifecycle:** Can run indefinitely evaluating goals.
- **Dynamic Flow:** Chaining APIs and browser sessions on the fly.

What Breaks in the Cloud?

When autonomous agents hit standard infrastructure, the core assumptions of ephemeral computing fail.

Because agents hold state, run continuously, and dynamically chain unmapped APIs, traditional scaling metrics (like CPU/RAM utilization) become useless. Auto-scalers thrash, serverless functions timeout, and network egress costs skyrocket unpredictably.



The New Security Surface



Supply Chain Risks

Agents rely on community "skill repos" (e.g., PyPI for agents). Malicious tools installed by the agent dynamically can compromise the entire underlying VM or container cluster.



Prompt Injection

No longer just generating bad text. When an agent has terminal access, an injection payload parsed from an external website becomes arbitrary remote code execution.



Cloud IAM Gaps

Identity Access Management is built for humans or single-purpose services. Agents act on behalf of humans but operate autonomously, breaking traditional RBAC boundaries.

OWASP Top 10 for Agentic Apps (2026)

#1

TOP THREAT

Infrastructure-Level Prompt Injection

Malicious prompts have evolved. They no longer simply manipulate LLM text outputs.

When an agent processes untrusted data, a well-crafted injection can trick the agent into executing destructive API calls, modifying cloud storage, or escalating privileges via the agent's assumed identity.

The Gap in Cloud IAM

Agents inherit broad system permissions designed for human operators, creating massive blast radiuses:

- ⚠ **Over-Provisioning:** Agents are often given admin rights "just to make it work."
- ⚠ **Lack of Context:** IAM cannot distinguish between an agent acting normally vs. acting under prompt injection.
- ⚠ **Temporal Identity:** Agents need temporary, goal-specific permissions, not permanent access keys.



Solving the Crisis

Applying Site Reliability Engineering (SRE) patterns to safely deploy and manage agentic workloads at scale.

SRE Reliability Patterns for Agents

Least-Privilege Scoping Instead of static IAM roles, use Just-In-Time (JIT) access. Agents must "request" elevated permissions for specific, short-lived tasks, which are granted via automated policy evaluation.

Container Isolation & Ephemerality Never run agents on bare VMs. Use heavily restricted, network-isolated container sandboxes (like gVisor or Firecracker) that are destroyed entirely after a task completes.

Observability for Autonomous Workflows Standard APM tools fail here. Implement semantic tracing: log the agent's "chain of thought," tool usage, and branching logic, not just HTTP response codes.

Error Budgets for Destructive Actions Calculate acceptable risk. If an agent hits its error budget for anomalous API calls, automatically quarantine the workload.

Three Pillars of Safe Deployment



Compute

Enforce strict CPU/Memory quotas. Use microVMs to prevent container escapes during unvetted code execution.



Storage

Isolate agent state. Use immutable snapshots so agents can "time travel" backwards if they corrupt their own memory.



Network

Implement Deny-by-Default egress. Agents should only reach pre-approved domains to prevent data exfiltration.

Agent Deployment Checklist

Phase	Security Requirement	Implementation Pattern
Pre-Deployment	Supply Chain Audit	Mirror "skills" to internal, scanned container registries. Block direct public internet pulls.
Pre-Deployment	Identity Scoping	Assign a minimal IAM role scoped explicitly to the agent's specific task.
Runtime	Network Egress Control	Deploy a proxy (e.g., Squid) with a rigid allowlist of approved API endpoints.
Runtime	Human-in-the-Loop (HITL)	Require explicit MFA approval for any destructive actions (DELETE, DROP, etc.).
Post-Incident	Immutable Auditing	Stream all agent thought processes and actions to WORM (Write Once Read Many) storage.

THANK YOU

Building the Future of Secure & Reliable Agent Workflows

Image Sources



https://png.pngtree.com/background/20250724/original/pngtree-cloud-computing-digital-illustration-showing-servers-and-data-processing-network-with-picture-image_16816303.jpg

Source: pngtree.com



https://static.vecteezy.com/system/resources/previews/053/716/434/non_2x/futuristic-digital-padlock-concept-with-glowing-green-circuit-design-symbolizing-cybersecurity-and-data-protection-in-technology-png.png

Source: www.vecteezy.com



https://upload.wikimedia.org/wikipedia/commons/b/b5/Cloud_computing.svg

Source: en.wikipedia.org



https://static.vecteezy.com/system/resources/previews/053/463/070/non_2x/security-digital-technology-cyber-security-and-data-protection-concept-with-futuristic-shield-and-electronic-circuit-board-on-dark-blue-vector.jpg

Source: www.vecteezy.com



https://img.freepik.com/premium-vector/abstract-futuristic-technology-network-background-with-glowing-blue-lines-nodes-dark-background-connecting-geometric-shapes-data-points-innovation-concepts_1222051-791.jpg

Source: www.freepik.com