Executive summary

Solving the cost, latency, and egress bottlenecks stifling enterprise scale

When organizations attempt to scale a live, distributed portfolio of user-facing applications, legacy cloud design introduces unanticipated friction. Infrastructure leaders face a dilemma: overprovision centralized resources at an unviable cost or restrict capacity and damage the user experience.

Organizations must adopt a distributed multicloud model to achieve economic and operational efficiency. This transition allows AI projects to move from experimentation into full production. Leveraging edge networks and adopting alternative cloud platforms resolve the structural tension by prioritizing and optimizing data movement. This leads to the alignment of compute in proximity to users and data stores.

To help quantify this infrastructure transition, Akamai partnered with Omdia to survey 200 qualified IT decision-makers around the world. All respondents held direct operational visibility into their organization's AI initiatives. The respondent base spanned organizations operating across technology, financial services, healthcare, retail, manufacturing, and business services.

The data validates the core problems facing platform engineering teams today:

- **Performance:** 95% of IT leaders state that delivering exceptional, subsecond AI user experiences is critical to remaining competitive in their respective markets.
- **Centralized cloud bottlenecks:** 86% encounter major operational challenges serving distributed users from centralized cloud setups due to latency and data transport constraints.
- **Cost trade-offs:** 90% must accept operational cost trade-offs just to achieve an acceptable baseline of application performance.
- **Scale:** Organizations with live, customer-facing AI applications currently manage an average of 24.5 in production, with 59% planning to deploy 11 or more additional applications over the next 12 months.

To resolve these points of friction, enterprises are fundamentally shifting their architectural strategies. Ninety-six percent of IT leaders plan to use multiple cloud providers for customer-facing, AI-native applications, with 54% preferring an even balance between edge and centralized inferencing based on specific workload requirements. Furthermore, 88% of organizations are adopting or planning to adopt alternative, workload-specific cloud platforms. By moving away from single-cloud, centralized environments toward distributed, hybrid, and edge topologies, enterprise leaders can match compute to data locations, eliminate unsustainable egress fees, and ensure subsecond application performance at scale.