uptime INTELLIGENCE

INTELLIGENCE UPDATE

Cloud availability improved in 2025 — but worst cases worsen

Dr. Owen Rogers 30 Jul 2026

An outage of a cloud data center does not necessarily cause an outage of a cloud application. Applications designed to operate across availability zones and cloud regions can continue to perform during outages. Indeed, cloud providers make it clear that customers should design their applications to account for occasional infrastructure failures (see *Build resilient apps: do not rely solely on cloud infrastructure*).

In 2025, Microsoft Azure, Google Cloud and Amazon Web Services (AWS) all continued to experience occasional zone and region outages. For customers who had not architected dual-zone or dual-region resiliency, applications hosted in affected locations likely experienced downtime. But some outages affected even those organizations that had diligently planned for failure. Google Cloud and Microsoft Azure both suffered from widespread, multi-region outages in 2025. In these cases, even if failover to another region had been planned, application downtime likely occurred.

In 2024, Uptime Intelligence collected availability status updates from AWS, Google Cloud and Microsoft Azure to measure historical availability rates (see *Outage data shows cloud apps must be designed for failure).* The research determined whether incidents reported in these historical status updates would have affected applications architected using virtual machines deployed in:

- The same availability zone (single zone).
- Different availability zones in the same region, balanced via a cloud load balancer (dual zone).
- Different availability zones in different regions, balanced via a cloud load balancer and domain name server (DNS) service (dual region).

In this report, the exercise is repeated using data from 2025. Analysis shows that while average availability of basic cloud services improved, customers in availability zones and cloud regions that were the worst-affected by outages were at risk of considerably more downtime. Cloud outages are rare, but when they happen, consequences for those that are unprepared are becoming worse.

Outages

An application's availability depends on the cloud services it uses and its architecture. An outage of a particular service will not affect an application if it does not use that service.

To measure availability, we use a simple application architecture that consists of virtual machines, load balancing and DNS. We assume that incidents affecting services required to operate or recover the application would materially impair its availability. This includes issues that affect latency, error rates or control planes, even if the provider does not report a full outage. We measure duration from the first time the incident is reported to its full resolution.

This approach provides a means of measuring relative performance of resiliency architectures across time periods. However, it does not provide an absolute measure of availability.

In 2025, multiple major outages were likely to have impacted applications hosted in the affected zones and regions:

- **Azure East US 2** (January 2025). A major availability zone networking failure caused by a network configuration change. The incident affected virtual machines and many other Microsoft Azure services.
- **AWS Tokyo** (April 2025). A regional connectivity incident affecting both AWS's Elastic Compute Cloud (EC2) and Elastic Load Balancing services resulted in customer-facing application disruption in the Tokyo cloud region.
- **Google Cloud** (June 2025). A global control plane outage affected numerous Google Cloud regions and services simultaneously.
- **AWS us-east-1** (October 2025). A significant region outage affecting multiple AWS services following failures that propagated through core infrastructure. EC2, Elastic Load Balancing and numerous dependent services experienced increased error rates and connectivity problems (see *AWS outage: what are the lessons for enterprises?*).
- **Azure West Europe** (November 2025). A thermal event affecting infrastructure within an availability zone in the Netherlands caused virtual machine degradation and impacted numerous dependent Microsoft Azure services.
- **Azure China** (December 2025). A broad control plane incident affected multiple cloud regions in China simultaneously, impacting virtual machines.

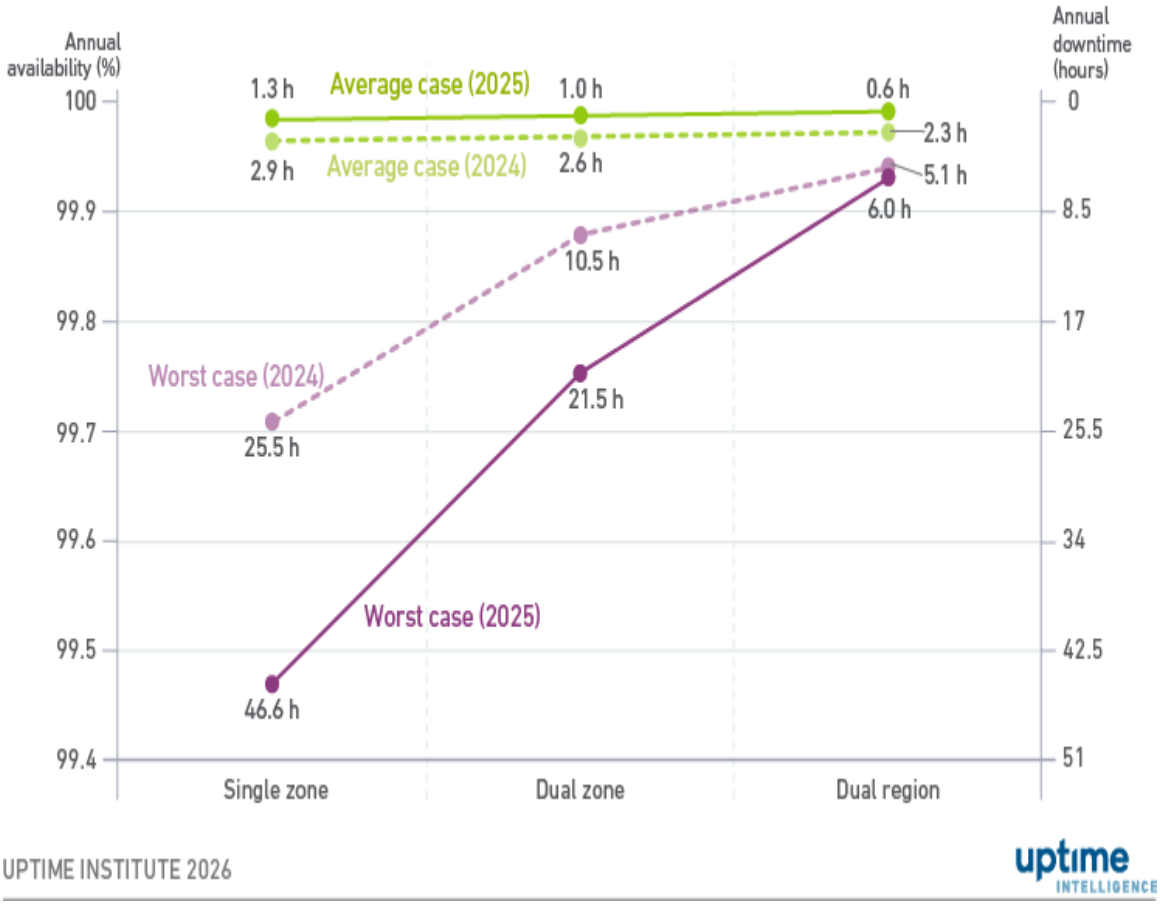
There were also significant outages that would not affect the example architecture but are noteworthy for their wide-reaching impact. The largest was the Azure Front Door global outage in October 2025 (see *What the Azure outage revealed about internet fragility*).

Results

Figure 1 shows the availability of the worst-performing regions in 2025, along with the average, across the different application resiliency architectures. Most regions would have provided 100%

application availability over the period.

Figure 1 Average and worst-case availability by architecture in 2024 and 2025



On average, applications hosted in single and dual zones experienced less than half the annual downtime of 2024. Applications hosted across regions experienced around three-quarters less downtime in 2025 than in the previous year.

However, average availability paints an optimistic picture. Most availability zones experience little or no downtime in a typical year, making single-zone deployment appear reasonable. But major outages are usually concentrated in a small number of locations. Organizations with applications that happen to be deployed in one of these zones can experience dramatically worse availability.

An application hosted entirely within a zone that suffered a substantial outage (such as Azure's East US 2 failure in January 2025) would have experienced more than 46.6 hours of downtime — nearly twice as long as in 2024.

An application hosted in two zones also saw more downtime when the worst case is examined. An application downtime of 21.5 hours would have occurred in 2025, double that of 2024. Whole-region failures, such as AWS's October 2025 failure in us-east-1, affect several availability zones and can have a significant negative impact even on the more resilient applications.

Even with major multi-region outages, as in the case of Google Cloud's June 2025 incident, applications distributed across regions (engineered to a very high level of resiliency) generally performed well. In the worst case, application downtime was 6 hours per annum, slightly above the previous year's 5.1 hours.

Unfortunately, there are cost and carbon implications to multi-region and multi-zone availability

architectures. A single zone, statistically, provides a good level of availability. But if that zone fails, the impact of an outage can be long-lasting and costly, with little ability to recover beyond waiting on the cloud provider to restore service. Compensation from cloud providers is not designed to offset losses from downtime (see *Cloud SLAs punish, not compensate*).

As with all workload placement decisions, organizations need to balance availability, cost and carbon against other qualitative requirements to ensure they protect their business without paying more than necessary (see *Cloud: when high availability hurts sustainability*).

Table 1 Annual availability by architecture and year (with downtime in parentheses)

Zone or region	Year	Single zone	Dual zones	Dual regions
Best	2024	100% (0 h)	100% (0 h)	100% (0 h)
	2025	100% (0 h)	100% (0 h)	100% (0 h)
Average	2024	99.967% (2.9 h)	99.970% (2.6 h)	99.973% (2.3 h)
	2025	99.985% (1.3 h)	99.989% (1.0 h)	99.993% (0.6 h)
Worst	2024	99.709% (25.5 h)	99.880% (10.5 h)	99.942% (5.1 h)
	2025	99.471% (46.6 h)	99.755% (21.5 h)	99.932% (6.0 h)

UPTIME INSTITUTE 2026 uptime INTELLIGENCE

The Uptime Intelligence View

Generally, cloud providers continue to deliver good availability across zones and regions. However, when a zone does suffer an outage, the downtime can be significant. Which zones will suffer serious outages is largely a matter of chance. The data show that duplication reduces the impact of a cloud outage. However, there are cost, carbon and technical implications for building resiliency. Organizations need to balance application availability against these considerations.

ABOUT THE AUTHOR

Dr. Owen Rogers

30 Jul 2026

Dr. Owen Rogers is Uptime Institute's Senior Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a chartered engineer, product manager and industry analyst. Rogers covers all areas of cloud, including AI, FinOps, sustainability, hybrid infrastructure and quantum computing.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com