INTELLIGENCE UPDATE

AWS outage: what are the lessons for enterprises?


Dr. Owen Rogers

23 Oct 2025

On October 19 and 20, 2025 AWS experienced a significant service disruption affecting multiple services in the us-east-1 (North Virginia, US) region for nearly 15 hours. The incident resulted in increased error rates, API failures and latency across numerous AWS services.

Us-east-1 is AWS's oldest and most-used region. It is also usually the cheapest region, thereby attracting a disproportionate share of workloads. As a result, the outage affected thousands of AWS customers, ultimately impacting millions of consumers.

Major consumer platforms, including Snapchat, Fortnite, Venmo and Robinhood, experienced either complete outages or severe slowdowns. At the same time, financial institutions, government agencies and retailers also reported outages. According to online outage tracker Downdetector, the event generated more than 16 million problem reports worldwide, with businesses facing transaction failures, customer service interruptions, and data processing backlogs. Some estimates put the economic cost to customers at billions of dollars. It has reignited the debate around Europe's reliance on US hyperscalers (see *Europe will not abandon the hyperscalers*).

The root cause was a domain name system (DNS) failure. Many of AWS's back-end systems rely on the same services that AWS provides to its customers. AWS uses DynamoDB, a simple database service, to track the life cycle of virtual machine resources created on EC2, its compute service. The outage began when the DNS was unable to route data to the DynamoDB service. In turn, this prevented AWS from tracking the life cycle of virtual machines, thereby hindering their creation and management. With AWS's back-end systems relying on DynamoDB, errors cascaded to many other services across the whole us-east-1 region.

It is not yet clear what caused the initial DNS failure. Given the significant impact of such a small failure, what could have been done to prevent it?

Who was at fault?

Cloud providers, including AWS, are generally upfront that, because of the sheer scale of their

operations, services and data centers will fail from time to time. Cloud provider service level agreements (SLAs) do not promise perfection. However, in this case, AWS did breach its 99.99% dual-zone EC2 service level agreement (equivalent to 4 minutes downtime per month).

Providers such as AWS argue that developers should choose what level of failure they can tolerate, and architect their applications appropriately. They advocate that applications should be designed to span multiple availability zones and/or regions, so that they continue to operate in the event of an outage:

- An availability zone is a data center (or multiple data centers). Each zone is often understood to have redundant and separate power and networking.
- A region is a geographical location containing multiple availability zones. Each region is physically isolated from — and independent of — every other region in terms of utility, power, local network and other resources.

This distributed resiliency concept is at the heart of every cloud training certification, every reference architecture and is spelled out in freely available design documentation.

In the recent AWS outage, an entire region failed, taking down three availability zones. Applications hosted in a single zone or multiple zones in that region would have become unavailable during the outage. Applications architected across regions continued to operate when us-east-1's services were compromised.

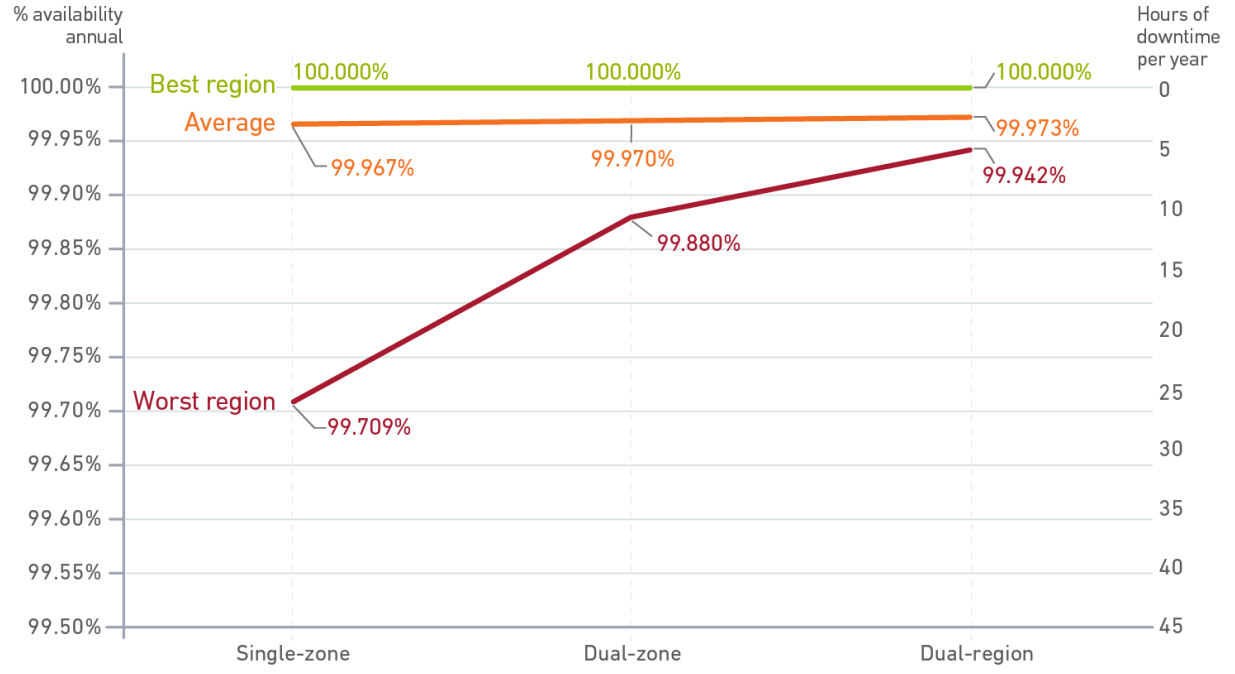
Organizations affected by the outage had most likely not architected their applications to operate across regions. These organizations were aware of the risk of an outage — they have access to SLA documentation and best-practice design guidelines. However, they chose not to architect across regions, likely due to issues relating to cost and complexity (see *Cloud availability comes at a price*). A more resilient architecture requires more resources, which translates to greater expenditure. Designing applications to work across regions requires them to be scalable, which makes implementation and management more complex.

The availability dilemma

Why did those companies fail to consider the impact of a regional failure? Data from the Uptime Intelligence report *Outage data shows cloud apps must be designed for failure* helps to explain the dilemma architects face when designing cloud applications for failure.

Uptime Intelligence obtained status updates from AWS, Google Cloud and Microsoft Azure for 2024 to measure historical availability. Figure 1 shows the availability of the best and worst performing regions, alongside the average, for the different architectures (a full methodology can be found in the report referenced above). The difference between average and worst region availabilities reveals that most regions experienced high levels of uptime in 2024, while some regions encountered serious incidents.

Figure 1 Availability for different application architectures



UPTIME INSTITUTE 2025

uptime INTELLIGENCE

In general, cloud availability zones and regions have very high availability. Of the 116 cloud provider regions examined in this study, 29 experienced no issues (green line).

Average availability across all regions is also high (orange line). Crucially, the average is always high regardless of architecture. Architecting an application to work across zones or regions is often not worth the cost and complexity — the improvement in availability is negligible for the incremental cost and effort. Many customers affected by AWS's outage likely thought that a regional outage would be unlikely. They assumed that, even if it did occur, it would be unlikely to take place in their region. Such reasoning is reasonable, based on the general stability of cloud availability zones and regions.

However, averages can be misleading. For those unlucky organizations whose applications happen to be located in a region experiencing a significant outage, the value of resiliency is clear. In the worst performing regions (red line), architecting across availability and zones makes a substantial difference to availability. Those who suffered during AWS's recent outage likely decided that multi-region was not worth the cost and complexity, because the worst-case scenario was unlikely to occur.

Is multi-region enough?

So far, there have been no incidents of an entire hyperscaler public cloud suffering an outage. However, the risk (albeit small) remains. The recent AWS outage demonstrates how minor issues can propagate from a backend process to multiple customer-facing services across various locations.

It also demonstrates concentration risk: how the failure of a region can affect many customers who are reliant on that region. If those applications had been distributed across a wide range of

locations and providers, a failure of any one of them would have had a reduced impact compared with the outage of a single, centralized cloud.

If an entire public cloud were to fail because of a cascading error, a vast number of companies would be affected. The concentration risk is high, even if the probability of a failure appears low.

Cloud providers take significant steps to ensure that regions operate independently, so that errors or issues do not spread. However, there are some single points of failure, notably DNS. Cloud provider DNS services direct traffic to the appropriate region across global regions and a failure of DNS could render a whole public cloud unavailable. In this recent AWS outage, DNS failed to route to a single endpoint on a single service in a single region. A significantly larger DNS issue could have a more widespread impact.

Some companies architect applications to run across multiple cloud providers, or across on-premises and cloud services. These implementations are complex and expensive (see *Cloud scalability and resiliency from first principles*). As a result, few organizations are keen on multi-cloud in practice, considering the low likelihood of a whole cloud provider outage. However, a multi-cloud-architected application would not have suffered issues as a result of the recent AWS outage and, depending on its architecture, could stand a good chance of surviving a full AWS failure.

Given that a public cloud provider failure has yet to occur, it remains unclear how a cloud provider outage might impact the broader internet and other hyperscalers. If one cloud provider were to fail, would other providers also experience other issues due to sudden spikes in data center capacity demand or network traffic, for instance?

The risk of an outage can never be eliminated, even in a multi-cloud or on-premises implementation. More layers of resiliency may not necessarily translate into better availability, due to the complexity of the implementation. Even with the best planning, there may be points of failure that are hidden from view, within colocation providers, network operators or power companies. Enterprises cannot realistically assess and mitigate all these risks.

Nevertheless, a regional failure is not a rare, unpredictable event. Architecting across regions is more expensive and complex than a non-resilient architecture, or one distributed across availability zones. However, it is cheaper and significantly simpler than a multi-cloud implementation. For many of those affected by the outage, the small incremental cost of regional resiliency would have easily offset the losses caused by downtime.

The Uptime Intelligence View

Ultimately, AWS's customers are responsible for the failure of their applications. Cloud providers have a duty to deliver services that are available and performant. But they are also upfront that data centers will fail occasionally — and they will fail again. AWS customers knew they were exposed to the failure of a region. They took the chance but, this time, it did not pay off. Such

gambles may be acceptable for some workloads, but not for mission-critical applications. Organizations should assess the risk and impact of the failure of each of their applications. Greater resiliency requires greater cost — if an outage is going to have financial repercussions, paying for greater resiliency is worthwhile. Cloud resilience is as much an architectural discipline as it is a service level guarantee.

ABOUT THE AUTHOR

Dr. Owen Rogers

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. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.