Robert Frances Group
Business Advisors to IT Executives

Jan. 31, 2022

Summary Insights: Minimizing the Impact of Cloud Outages

RFG Perspective: One unforeseen consequence of the "cloud-first" strategy encouraged by senior business executives is the unwelcome experience of extended periods of IT downtime when an organization's cloud services go offline. Recent outages at AWS brought this topic to the surface – but outages have happened at all of the major cloud providers in recent years.

Fortunately, there are ways to minimize the damage that outages can cause: customers can opt to support applications and data on multiple "availability zones" within the cloud service provider's (CSP's) infrastructure. Alternatively, they can retain critical applications and data on-premises at their own data center, or they can use a multi-cloud strategy.

The real issue is determining how vulnerable each of their business services is to cloud outages. Given the wide adoption of cloud computing, executives should not ignore the possibility that the real-world recovery times may be longer than expected, impacting data processing and business outcomes. Business and IT executives must study their options – and take actions to avoid long periods of downtime when CSPs have a temporary outage – no matter what caused that outage.

INTRODUCTION

Recent cloud service provider (CSP) outages have surprised, and impacted, organizations that have adopted a cloud-first strategy in recent years. By using CSP services on an everyday basis, IT-related costs are contained, and data-center costs are avoided. But outages that cause minutes, or hours, of downtime can do some real damage to enterprise operations.

Expectations about high availability for the cloud have changed in recent years, as the pace of cloud migration accelerated – especially since the pandemic began in March, 2020. The number and variety of applications has changed since businesses first used the Cloud more than a dozen years ago. Back then, the focus was on application development and testing. Today, cloud service providers (CSPs) are taking on mission-critical workloads, and, consequently, the responsibility for providing high availability and quick recovery times following cloud outages.

The Robert Frances Group hosted an online conference recently to discuss the options and strategies that will help enterprises protect themselves from cloud service outages.

In December, 2021, Amazon's AWS and other cloud service providers (CSPs) endured several extended outages in recent months, negatively affecting some of its own business units and those of many of its customers. These recent disruptions hit even the largest of businesses, including Disney+, Netflix, and McDonald's Corp. These enterprises lost revenue due to the outages, angering loyal customers that had expected their preferred suppliers' services to be available all of the time.

But is 100-percent uptime a reasonable goal for all applications -- or even for one application?

In the world of high availability (HA) and disaster recovery (DR), topics long associated with data centers, uptime goals of 99.99% (four nines) to 99.9999% (six-nines) uptime for individual computer systems are often achieved – even if only on a handful of critical systems.

But even for a large enterprise that uses thousands of individual applications running on dozens or hundreds of servers, it's likely that something, somewhere in the data center is offline from day to day or from month to month. Workloads running on CSPs can experience similar problems. But limited tech issues are not the issue – widespread outages, when business applications cannot be accessed, are a key issue.

By migrating mission-critical and business-critical applications to CSPs, companies were looking for scalable resources with high uptime that won't go offline during a localized or regional outage. However, even those with a "cloud-first" strategy need to consider the following:

- Availability of alternative Availability Zones (AZs) within the CSP infrastructure.
- Avoidance of vendor lock-in, which could be caused by CSP-specific APIs and software.
- Potential to use a multi-cloud strategy, so that critical workloads can run in alternate clouds, when and if needed.
- Use of open-source software and cloud-native software (e.g., Kubernetes) that allows applications to be run on alternate CSP clouds.
- Use of cloud-native databases that can be deployed on scale-out deployments of servers, regardless of their proximity to the customer's central-site Core data centers.
- Transferable skills to develop, deploy and maintain applications and data – from the Core to the Cloud – and even to Edge locations, in case of natural disasters or outages.

Should IT executives expect to get more availability from their CSP than they get from their own data center, or does the opposing argument seem increasingly credible given hybrid cloud complexities? The tradeoffs between availability, business alignment, cost, competencies, and expertise must be continuously re-evaluated to ensure the business' needs are operationally addressed. That is why, in 2022, IT executives must anticipate that the frequency and impact of such events will increase – which is exactly why they should develop and implement plans that minimize business disruptions.

PANEL:

Panelists shared their extensive experience in working with cloud service providers to improve availability of enterprise applications and data. All of the panelists have seen the impact of unexpected outages and they have examined this downtime challenge from both IT and business perspectives. The panel for the Jan. 12, 2022, discussion included:

- Brian Binovsky, senior vice president at Barclays Bank
- Evan Bauer, CEO at OpStack Inc.
- Jonathan Seelig, chairman and chief evangelist at Ridge.co

POLL:

A quick poll of the attendees on the topic of cloud outages yielded the following results:

- 65% of respondents reported using multiple availability zones from the same cloud services provider (CSP). This approach leveraged the CSP's software tools, supporting consistent management across the CSP's availability zones.
- 41% said a business case was developed and approved by IT executives for dealing with cloud services downtime and recovery. Only 18% had their business cases approved by a line-of-business (LOB) or corporate executive.
- 24% said they use a colocation (colo) provider, content delivery network (CDN) or data center as part of their hybrid cloud environment. This leverages a third-party colocation provider supporting high availability for data and apps hosted on cloud services.
- 24% said that they have one or more critical applications that do not have full redundancy for HA/DR purposes. This surprising result may be related to a prioritized hierarchy of HA/DR across many kinds of applications, including legacy applications that are aging-in-place and should be considered for updating or eventual replacement.

- 12% said they use multi-cloud solutions to increase availability in the cloud. This is an increasingly popular approach, although it is generally more difficult to deploy and manage than a CSP-specific plan that uses multiple availability zones. Even so, some CSPs are offering multi-cloud solutions for global customers who must meet data-privacy compliance regulations requiring local data storage.
- Only 6% of these cloud customers reported that their critical (mission-critical or business-critical) databases are being kept in a local colo site or corporate data center, rather than in the public cloud. Moreover, only 35% said they had DR in place for some, but not all, critical databases. This is another surprising answer, because all critical databases should have a specific, and tested, availability plan – no matter how that plan is implemented.

ANALYSIS

During the RFG 100 panel discussion, one speaker said that CSPs are great at scaling workloads, but added: "Can you fail if you scale?" This is the essential question regarding CSP outages. Now that more companies have adopted cloud-first strategies, and now that important applications need to recover quickly from any unplanned outages, can CSPs do a better job of recovery?

The answer, if course, is yes. CSPs are adding availability zones (AZ) worldwide. Where there was one, now there are two; where a region did not have multiple AZs, they are being built. This is true in Africa, South America, and Indonesia – all of them are regions where AZs were few and far between until quite recently. However, the major CSPs have been expanding the number of AZs they operate – and some large customers are also adopting a multi-cloud recovery strategy, giving them more options if a major cloud outage occurs.

The Quest for More Recovery Options

In the interim, IT executives have learned to become pragmatic about how they will recover from unexpected cloud outages when they do occur. Cloud migrations have accelerated during the COVID-19 pandemic, making reliable HA/DR and business continuity plans a "must" for most organizations. Unanticipated issues regarding natural disasters and HA/DR capabilities will affect how CSPs restore services to dozens or hundreds of customers worldwide.

Some large customers have even decided to eliminate some, or most, of their on-premises data centers to reduce capital expenditure (CapEx) costs by moving more applications to CSPs. As a result, these business operations have become highly dependent on their CSPs' ability to continuously deliver compute, storage and networking services. When a CSP's services go offline, some business units within their enterprise can come to a halt – or nearly so.

Recent cloud outages, affecting customers of multiple CSP providers, may prove to be a "wake-up call" for cloud-first advocates – and other cloud users – who had underestimated the scope and impact of extended outages on their organization's business units.

Now, in light of some very real extended downtimes that fail to meet recovery point objectives (RPO) and recovery time objectives (RTO), cloud advocates need to re-examine their HA/DR plans and CSP contracts to determine how their CSP agreements can be amended and improved, in anticipation of a future wave of CSP outages – whenever that may occur.

Getting to the Bottom Line

The bottom line is this: business and IT executives must reassess their current HA/DR readiness in a cloud-first environment that is supporting multiple business-critical and mission-critical applications. Most probably they will be inclined to keep their current CSP services – but will likely add third-party partners and additional data-management software programs to "mind the gaps" that may exist, silently, just underneath their everyday cloud-fueled business operations.

In many cases, business units were the ones who voiced the strongest arguments advocating cloud-first strategies. Now, they must step up their dialog with their company's IT organizations – and meet with IT managers to discuss the impact of CSP outages on their daily compute, storage, and networking capabilities and the funding necessary to improve various levels of added uptime.

Key Considerations and Takeaways

Here is a quick list of how-to's for customers looking to protect their organizations from extended outages associated with CSP cloud services, grouped by topic:

Network Outages Are Often the Cause:

- Network outages can be the cause of widespread outages (e.g., damaged under-ocean cables, and disruptions caused by back-hoes in street construction can trip off a localized CSP outage.

Executive Responsibility

- Executives need to determine whether a cloud solution is truly usable for each of their mission-critical and business-critical workloads. These are the workloads that contain sensitive data and/or can cause business units to go offline for hours, or more.
- Executives must know the architecture and risk profile as it exists and as-planned before issuing any approvals to move forward for major CSP deployments. A cloud model for HA/DR needs to be developed if one is not already in-place.

Stakeholders:

- Final decisions about cloud migrations and CSP deployments should be made by all the stakeholders, not just a subset. Stakeholders need to determine or know the cost of downtime, the mean time between failures (MTBF), the acceptable mean time to recover (MTTR) levels, and the cost/benefits for the different options.
- Stakeholders must calculate the opportunity costs of cloud outages, as well. The central question here is: Can you scale when you fail? Make sure you know the hidden costs. CSP contracts may include overprovisioning and include unnecessary added costs.

Availability Zones for Recovery:

- Use of multiple availability zones (AZs) is the least expensive option, but it is not necessarily the best one. One key determinant of which options to use is whether or not a data-export component for a given application could initiate costly data-egress charges and drive up costs if another CSP is used for HA/DR purposes.
- Always have reserve capacity available somewhere in your infrastructure as a backup.
- Avoid single points of failure and, if possible, avoid vendor lock-in that makes it hard to "exit" by moving your application or datasets to another CSP provider.

Partnerships:

- Most CSP offerings are complex – as are their multi-tenant environments. A strategic partnership with one's network service provider can help simplify the solution.
- All parties need to understand both Day 1 and Day 2 architectural and workload impacts. Many companies look at Day 1 and do not understand the scope of their Day 2 exposures.

Taking Stock of Application Requirements

- Executives should take into consideration the entire lifecycle of each application.
- Questions to ask include the following:
  - Is it steady state?
  - Is it cyclical?
  - Is it compute-intensive or data-intensive?
  - What are the throughput requirements of that application?
  - What are the dependencies and the rate of technology refresh?
  - What are the RPO and RTO requirements?

RFG POV: Recent CSP outages are causing business and IT executives to re-examine how their organization will recover from future outages caused by CSP outages, network outages, power outages and other causes of unexpected downtime.

One hopeful sign is that the major CSPs themselves are investing heavily in constructing and deploying more availability zones (AZs) around the world. Most of the major CSPs have 20 or more AZs worldwide.

These investments will provide several types of benefits to the organizations that have a cloud-first strategy – supporting uptime for critical applications; supporting regional security and compliance regulations; and increasing customer loyalty when lengthy downtime is avoided, as workloads are resumed on other Availability Zones.

In 2022, as more and more workloads move to the cloud, the impacts of CSP outages must become a top-of-mind priority for IT organizations. That's why reducing downtime and avoiding outages has risen to the top of the priorities list for cloud computing. Business and IT executives must jointly consider which of their revenue-producing and customer-facing applications are most vulnerable to unplanned CSP outages – and they must take steps now to ensure their organization's business continuity.

Additional relevant research and consulting services are available. Interested readers should contact Client Services to arrange further discussion or interview with Cal Braunstein, CEO and Executive Director of Research. Jean S. Bozman, president of Cloud Architects llc, co-authored this report.