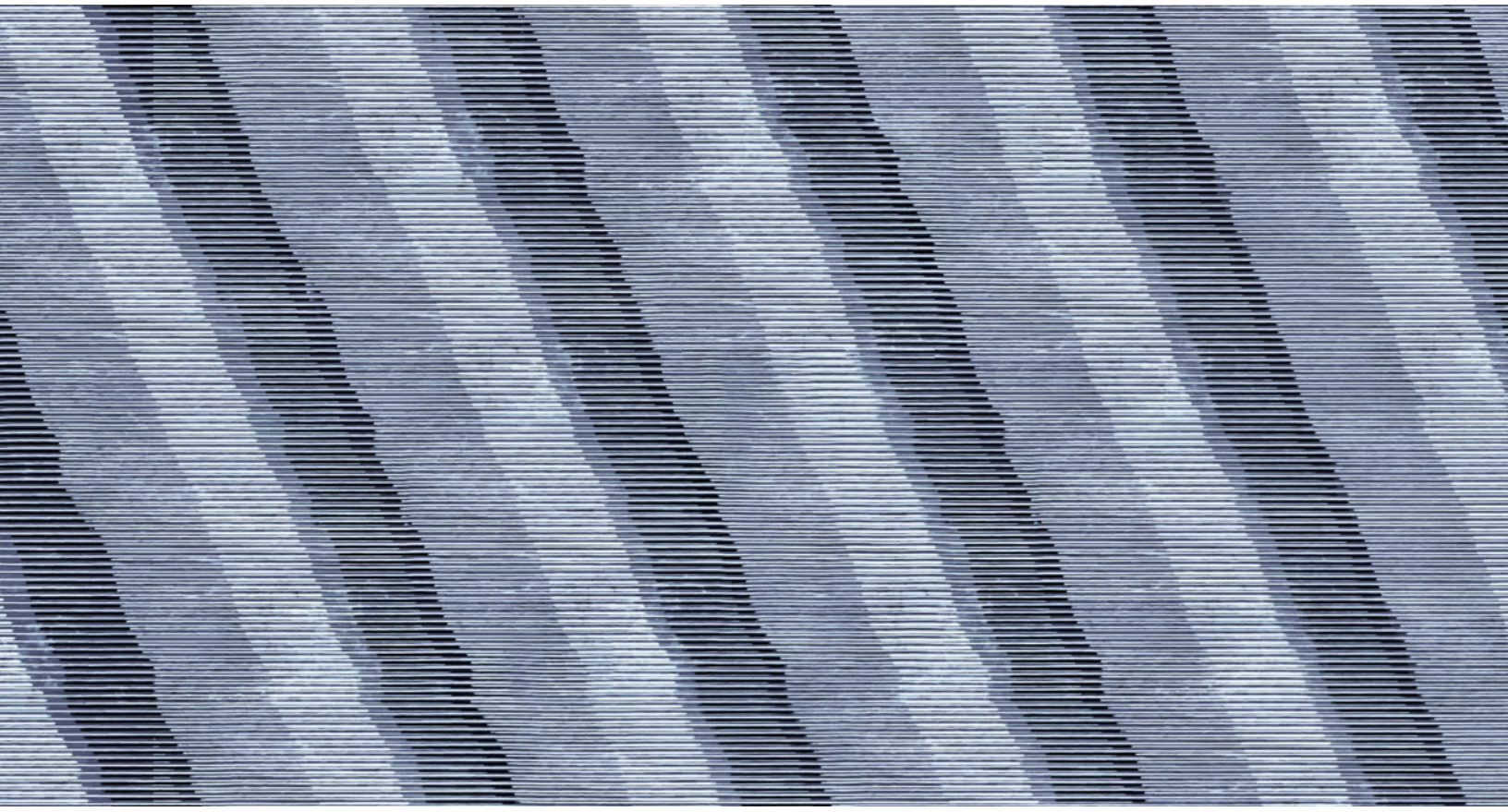
McKinsey Digital

Four ways to improve technology service resiliency

Companies can significantly reduce technology system outages and other costly tech glitches by improving incident response and how they manage change.

by Himanshu Agarwal, Ritesh Agarwal, Basel Kayyali, and Dan Stephens



September 2020

As businesses race to digitize, technology organizations have to deliver innovative services at high speed while maintaining a high level of operational stability. The pressure to act nimbly, however, often has a significant negative impact on service-resiliency issues.

While cutting-edge software companies have been leaders on resiliency issues, many traditional enterprises (those in retail, insurance, and financial services, for example) have lagged behind. System crashes, outages, and service-level disruptions have spiked over the past several years as businesses push to adopt new technologies, fast-track IT application development, and bring promising initiatives to scale quickly. And these outages and disruptions are happening even though non-tech companies have invested heavily in building service redundancy with hot standbys, real-time backup-and-restore services, stronger disaster-recovery planning, and system self-healing capabilities.

The COVID-19 crisis has added a new dynamic, with remote working, surges in network use, and the need to rapidly adapt to customers' digital preferences creating new strains.

The stakes are often higher than many leaders realize. A large retailer lost $5 million in sales when its systems went down for several hours on a busy shopping day. A software company suffered an 8 percent hit to its revenues after a system glitch knocked out service. A bank lost $10 million in revenue plus associated penalties when its outage stretched over multiple days. Ponemon Institute has estimated that the average cost of an unplanned outage is nearly $9,000 per minute (or $540,000 per hour).[1] In isolation, each of these costs might seem like small potatoes. But the reality is that businesses often have to deal with hundreds, even thousands, of these resilience incidents each year, and costs rapidly escalate, often without IT understanding the full scale of the issue. In addition to the financial blows, there is a cost in terms of customer satisfaction and employee productivity.

Unfortunately, the way many IT organizations in these more traditional sectors deal with these resiliency issues can resemble a technological game of whack-a-mole. When systems fail, an organization's first priority is restoring service. To do so quickly, most technology teams understandably reach for remedies that can do the job best in the moment. However, the tactical fixes that work well in one crisis will not generally work in the next, since issues vary so widely. As a result, many organizations face a continual cycle of triage, fixing one issue only to find another popping up somewhere else.

To stop that cycle, companies need to move away from a reactive emergency-response posture. In our experience, there is a set of four practices that traditional enterprises can adopt to improve technological resiliency, reduce the risk of incidents, and mitigate their commercial and customer impact.

[1] *Cost of data center outages*, Ponemon Institute, January 2016, ponemon.org.

Definitions

Availability: Time period when the service is available to customers (system uptime)

Reliability: Probability that a system will work as designed

Resiliency: Ability of the IT system to withstand certain types of failures and yet remain functional from the customer perspective.

1. Go beyond triggers to look for root causes and patterns

Surge events, such as a sudden spike in traffic, are usually a symptom of a larger issue, such as poor capacity planning, inadequate performance or load testing, or rigid architectures that are easy to break.

Companies with technological resilience take note of such triggers and then try to tease out the patterns behind them. Some of these patterns can be identified by mapping the number and type of incidents by domain (exhibit). Once the pattern is identified—be it in development, configuration, change management, or another area—companies can follow the trail, probing deeper within that area to see what is causing the persistent breakdowns.

For example, one company found that human error and rushed testing had been factors in 60 percent of its outages over the previous three years. In talking to developers, it learned that cultural factors were at play. Under pressure to hit aggressive launch targets, teams were placing too much stock in "happy path" use cases (scenarios in which everything works as planned) and paying too little attention to exceptions. Limiting use-case development to happy paths allowed code to hit production faster, but the resulting product was less stable. With a clear understanding of the root cause, the company was able to make systemic improvements, such as redesigning performance incentives to reward quality in addition to speed and beefing up its monitoring systems, which reduced subsequent incidents up to 30 percent.

Exhibit

A simple incident matrix can reveal systemic patterns.

Number of critical incidents[1]

>4 incidents | 3–4 incidents | 1–2 incidents

Responsible business domain/ infrastructure	Cause					
	Development	Configuration[2]	Capacity management[3]	Change management	Hardware failure	Total
App 1	3	3				6
App 3	2	2	1			5
App 7		2	1			3
App 8		3				3
App 10		3	1			4
Mainframe	1	5	1			7
Distributed system		4	2		2	8
Network		2		1	3	6
Storage		1	5		5	11
Database		9	1			10
Middleware	1	2	1	1	3	8
Service-oriented architecture			1		1	2
Total	7	36	14	2	14	73

[1]Severity 0/1 incidents.
[2]Configuration: incidents caused by a misconfiguration of the software/hardware parameters.
[3]Capacity management: incidents that occur due to insufficient understanding of the impact of a deployment on the infrastructure incidents resolved by increasing or changing infrastructure configuration.

2. Integrate and automate to prevent and detect issues earlier

It's not unusual for companies to have separate development, testing, and production environments. Monitoring processes can be similarly fragmented and manual. Tools may track parts of the development journey, but not all of it, and incident teams may not receive alerts in a timely way.

To resolve this issue, we recommend that teams identify their most critical customer journeys and modernize and automate the underlying processes end to end. For example, performance testing and load testing for customer channels can be made more rigorous to manage spikes in customer log-ins. One large financial institution addressed this issue by creating a single management console to track and aggregate alerts for all the journeys related to its mobile app. These alerts were integrated with its ticketing system to automatically open incident tickets as relevant anomalies were detected. This enabled earlier detection of anomalies and reduced incident volumes by roughly 8 percent within six weeks of implementation. Many companies are also investing in self-healing systems where automated scripts are executed when an anomaly appears. These scripts can perform a range of tasks, such as refreshing the servers, provisioning extra storage, or even applying the latest patch.

Most organizations also need to improve how they handle change requests. Too often, emergency change requests are bundled with routine ones, and teams are left to sift and prioritize the relative risk and urgency of requests on their own. Overwhelmed, some change-management and application-development teams resort to quick fixes that lead to high change-failure rates.

To stop this practice, we recommend not only improving risk categorization but also addressing the issue at the source by creating a scoring system that rates development teams based on change volumes and application quality. With such a system, for example, teams with low scores would not be allowed to push changes into production beyond a certain threshold.

A financial institution that implemented this approach noticed a huge improvement. Achieving a high score became a point of pride for developers and was rewarded with performance incentives. Motivated to create better-quality products, teams improved their average risk scores by 25 to 40 percent over a six-month period. Over the longer term, companies can invest in change-management systems that automate many of these processes.

3. Develop tools and expert networks to speed incident response

No matter how strong the technology organization is, incidents will happen. The goal is to reduce the frequency and severity of those incidents and minimize their impact. Businesses can do that by making it easier for teams to access needed expertise and by providing clear, ongoing communication to customers and stakeholders.

Smart solutions can include updating knowledge repositories with user-friendly tools to help teams troubleshoot more quickly. For example, intelligent dashboards and tablets could let analysts punch in a brief outage profile. Built-in analytics would then sift through the company's incident database, prompt the analyst to enter additional details to filter results, and return recommendations and contact information for contracted solution providers as well as reference cases. This can help in reducing the "mean time to resolve (MTTR)," hence reducing the impact of the outage.

Making collaboration and accessing relevant experts easier is also important. Cataloging relevant subject-matter experts—both inside and outside the IT function—and bringing groups together for occasional brown-bag discussions and tabletop exercises have helped leaders create stronger networks of experts, which has led to faster and more effective responses to resiliency issues.

Companies need to pay equal attention to keeping key stakeholders adequately informed. Most organizations have crisis communication plans, but these need to be revised periodically to ensure the chain of command for service-related messaging is up-to-date. Customers understand that outages happen, but they're less

likely to be forgiving of long wait times, outdated status reports, and clunky interfaces. Establishing a comprehensive incident-response plan that anticipates the questions of customers, investors, and other stakeholders can go a long way toward helping companies preserve strong relationships.

4. Make sure problem management has structure and teeth

To reduce the risk of repeat incidents, problem-management teams conduct postmortems and provide recommendations, but many times their suggestions are ignored, and if they are implemented, it can take a long time, increasing the chances that similar incidents will occur before they're addressed.

At a financial institution, for example, it often took problem managers up to four weeks to conduct postmortems following an incident and another six weeks before their recommendations were implemented. This issue is often tied to poorly defined service-level agreements (SLAs) on issue resolution or teams that simply don't adhere to them. To drive accountability, the SLA needs to clearly lay out agreed-on actions and ownership so management can see who is responsible for making needed changes.

Executive sponsorship is also important. On their own, problem-management teams generally lack the institutional clout to see their recommendations carried out. Strong CIO engagement can provide the needed enforcement mechanism. At one company, for instance, the CIO incorporated an issue status review into her weekly meeting with problem management. That simple step served as a forcing function to ensure that recommendations were created and implemented promptly. At the end of two months, the organization had reduced average recommendation-implementation time from 25 days to eight.

Leaders that adopt the recommendations outlined here will be able to help their organizations withstand the demands of the crisis and postcrisis period more effectively. Thinking through the following questions can help tech leaders begin that journey:

- How has digital adoption and increased change velocity impacted IT resilience over the past several years, and where has the COVID-19 crisis created the most pressure?
- How does my company's incident response compare to our peers' and to other industries' more generally in terms of the number and severity of incidents and associated costs?
- Looking across the business, what issues are driving the highest volume of incidents, and which user and customer journeys are most impacted?
- What percentage of all IT application-development issues are first detected in production, and how many human hours went into tracking down and resolving them? Would an automated process be cheaper and more effective?

CIOs should move quickly. As businesses accelerate the pace of development and adoption of tools, channels, and business models, technology service resilience will play an increasing role in overall business resilience.

Himanshu Agarwal is a partner in McKinsey's Silicon Valley office, **Ritesh Agarwal** is an associate partner in the New York office, **Basel Kayyali** is a senior partner in the New Jersey office, and **Dan Stephens** is a senior partner in the Washington, DC, office.