Reframing Resilience: Digital Infrastructure, From Stability Assumptions to Stress Design

By Vladimir Lukic, Shoaib Yousuf, Filippo Scognamiglio, Amir Alsbih, Radu Balanescu, and Imane Rai

ARTICLE APRIL 08, 2026 12 MIN READ

Recent military strikes leading to physical damage in hyperscale commercial cloud facilities servicing the Middle East, causing regional service degradation. They have drawn attention to the broader question of digital infrastructure resilience.

The service provider's public statements have reported physical impacts and structural damage to the UAE and Bahrain's infrastructure. Multiple Availability Zones (multi-AZ) in the UAE were significantly impaired, prompting the provider to advise customers with workloads in the affected regions to reroute or switch to backup systems (failover) and, where needed, migrate to other regions.

Considerations extend beyond cloud versus on-premises decisions. Cloud regions, colocation sites, and sovereign/private data centers all sit on the same physical stack: grid power and fuel logistics, water and cooling, terrestrial and subsea fiber, key interconnection points (carriers/IXPs), satellite ground segments, and access to people and parts.

In stress scenarios, failure rarely stays neatly inside one component; the real risk is disruption that cascades across multiple layers of the stack. For example, power instability may be combined with degraded connectivity and constrained physical access. Such correlated disruption can come from conflict, natural hazards, grid instability, cable cuts, cyber-physical campaigns, or any scenario where multiple supporting systems (dependencies) fail together, and repair/access are constrained. These dynamics apply across a range of providers and digital infrastructure sites

including hyperscalers, regional cloud and colocation operators, and sovereign/private data centers, where power, connectivity, and access constraints converge.

Governments and regulated sectors that primarily operate sovereign/private data centers should not assume they are safer for not using hyperscalers. If critical services concentrate in a single campus, metro, or narrow set of power or fiber corridors, the blast radius can be larger, and recovery slower, than architectures built across multiple independent systems that do not fail together.

Leaders are invited to challenge prevalent design assumptions, especially as digital infrastructure has become a core economic backbone. Payments, telecommunications, healthcare, logistics, public services, and identity systems now depend on reliable computing and connectivity. When digital infrastructure is degraded at a regional level, the impact is not limited to IT disruption; it translates into interruptions in essential services and broader economic activity.

This shifts resilience from a technical consideration to a strategic requirement. As with power and transport systems, continuity expectations for digital infrastructure must be clearly defined, tested, and coordinated across stakeholders to ensure critical services can operate under stress. The design/planning question is the same across scenarios: what happens when a region is not just “degraded”, but operationally constrained for an extended period, and you must keep critical services running while your ability to change, repair, and resupply is limited?

— Defining digital infrastructure concentration and dual-use

Concentration: Reliance of critical digital services on a small number of providers and geographically clustered nodes (regions, key IXPs, fiber corridors, grid interconnects), creating correlated-failure and coercion risk.

Dual-use: Use of the same provider/region and its physical dependencies to support both civilian economic activity and defense/government missions, often through shared power, connectivity, operations, and supply chains, even when environments are logically (and sometimes physically) segmented.

When civilian cloud inherits military risk: the case for a “two-speed” compute posture

Isolated incidents do not imply a new, widespread military doctrine of striking digital infrastructure-related facilities. But they should change scenario planning. Taking data center design, for example, infrastructure risk, governance risk, and cyber risk are traditionally included in facility design. But kinetic security risk must now be considered as a plausible driver of regional cloud unavailability (Exhibit 1).

EXHIBIT 1

Four key types of risks to digital infrastructure requiring deliberate resilience planning and mitigations

Typically Covered in DC Design			Outside Standard Availability Design Assumptions
INFRASTRUCTURE RISK	**GOVERNANCE RISK**	**CYBER RISK**	**SECURITY RISKS**
Risk of service disruption arising from failures in physical facilities, power, cooling, network, or site design	Risk arising from weak operational controls, unclear accountability, or inadequate resilience planning and oversight	Risk of digital disruption or data compromise due to malicious cyber activity or systemic IT failures	Risk of disruption or damage from deliberate physical attacks or hostile actions against facilities
Sub-Risks	Sub-Risks	Sub-Risks	Sub-Risks
Utility power failure	Poor disaster recovery planning	Ransomware and malware	Drone/UAV[2] attack
Generator or fuel supply failure	Poor incident management	Distributed Denial of Service	Ballistic or explosive attack
Cooling system failure	Weak change management controls	Phishing and credential compromise	Armed intrusion or sabotage
Network path/carrier disruption	Vendor and 3rd party concentration	Supply chain compromise	Hostile vehicle attack
Fire, flood, or environment incident	Lack of resilience testing	Control plane or configure failure	Civil unrest/perimeter breach
Design flaws (single points of failure, poor zoning)	Insufficient regulatory compliance		Electromagnetic pulse (EMP)/kinetic shock impact

1. DC refers to Data Center; **2. UAV** refers to Unmanned Aerial Vehicles
Source: BCG analysis

Additionally, sustained kinetic risk disrupts not only hardware but also the ability of trained operators to access sites and networks. This encompasses risks associated with disruption or damage from deliberate physical attacks or hostile actions against facilities. When assessing such risks, it is useful to identify different layers in the physical infrastructure stack and how

conflict may cause each one stress or damage. Continuing with the cloud example, the physical stack would typically include power and fuel, data centers, terrestrial and sub-sea cables, satellites and their ground infrastructure, and IXPs and core routing (Exhibit 2).

The strategic challenge is not simply that buildings can be hit. It is that digital infrastructure often concentrates very different users, including call centers, banks, hospitals, consumer platforms, and governments, all of whom are on the same underlying physical stack that may also support defense-relevant analytics and command workflows. Such dual-use does not necessarily mean that classified and consumer workloads run in the same data hall. The coupling can occur through shared providers, shared regional dependencies (grid, water, fiber corridors, IXPs), shared operational staff, or supply-chain constraints. This creates new escalation, resilience, and governance dynamics.

A common failure mode in "well-architected" environments is not the data plane, but the recovery path. For example, multi-AZ or multi-cloud does not guarantee resilience if your ability to operate depends on shared control planes and centralized management layers.

In real incidents, teams often discover that identity providers, DNS, orchestration, CI/CD pipelines, IaC tooling, key materials, monitoring, and even internal communications are centralized and become hard dependencies at the exact moment change is required. If you can't authenticate, can't provision, or can't update routing policy under duress, you don't just have an outage. You have a recovery outage.

"Last mile" recovery steps must work even when management APIs are degraded. That means break-glass procedures and credentials, replicated/printed runbooks, pre-provisioned capacity in alternate regions/providers, and regular drills that explicitly simulate control-plane constraints.

Leaders should explicitly plan for:

1. **Civilian collateral from perceived military targeting.** If a hostile actor views a region or dependency as enabling military capability, civilian services become exposed even if they are not the intended target.
2. **Military impairment through civilian-grade dependencies.** Defense users relying on commercial ecosystems inherit constraints that may not be designed for sustained threat (grid stability, fuel logistics, telecom corridors, local access rules).
3. **Escalation and attribution ambiguity.** Hybrid incidents (drones, cyber, sabotage) often have contested attribution, complicating response thresholds, insurance positions, and liability outcomes.

4. **Concentration risk becomes geopolitical.** "Single provider/single region" is no longer just an IT governance issue. Instead, it becomes an element of national resilience and potential coercion.

5. **Legal, regulatory, and reputational shock.** For cloud and other dual-use infrastructure, governments may mandate segmentation. Firms may face new disclosure, export control, and duty-of-care expectations.

A practical way to reduce these risks is to separate general elements of digital infrastructure from what must be survivable under attack. This may mean segmenting workloads and dependencies (Exhibit 3).

Leaders should take specific actions

Most organizations do not need to predict whether the recent incidents mark the start of a new military doctrine. But they do need a resilience plan that mitigates impact and ensures business continuity when events of this nature occur (Exhibit 4).

It is important to note that redundancy is different from resilience. Redundancy is designed to absorb expected component failures (a rack, a host, even a single facility). True resilience is the ability to keep operating, or recover predictably, under correlated disruption, where multiple layers fail together, and repair/access assumptions break.

This is why traditional constructs (tiering, multi-AZ, and “N+1” thinking) can create a false sense of security. They are necessary foundations, but they are not sufficient when the scenario is regional, possibly including multiple facilities impaired, grid instability, fuel logistics constraints, connectivity degradation, and denied physical access that extend recovery timelines.

The following question offers a practical litmus test. *If your primary region is unavailable for an extended period, what critical functions still operate, and what can you recover without relying on the affected region’s management dependencies?*

Below is a pragmatic agenda, framed around decisions that are actionable even when disruptive events lack perfect attribution. These actions can be implemented within existing architectures, without requiring a fundamental shift away from current platforms. We consider immediate and longer-term steps for leaders of three key stakeholder groups: **critical enterprises; hyperscalers** and **colocation providers; and governments and regulators.**

Critical Enterprises (finance, energy, telecom, health)

In the next 30-90 days, enterprise leaders should: validate that "multi-AZ" is not in name only (shared power, shared control plane, shared identity); run a live failover test to a different region (or provider) with a measured RTO/RPO, after ensuring that data residency and sovereignty requirements are still fulfilled; map key SaaS dependencies to physical hosting regions; and identify single-region vendor exposures. In the coming 12-24 months, they should: move the most critical workloads to active-active or warm-standby multi-region designs; negotiate contractual rights to exit/portability, including data egress and escrow; build dual connectivity (terrestrial plus satellite where justified); and test routing resilience.

Hyperscalers and colocation

Within the next 90 days, leaders at these organizations should: improve "break-glass" migration tooling and communications that work under duress; stress-test correlated-failure assumptions (multiple facilities/AZs; local access denial); and review physical security and emergency response procedures with local authorities. In the next 12-24 months, they should: expand options for segmented, higher-assurance offerings (including dispersed footprints); harden critical dependencies, including grid interconnect diversity, water resilience, and fuel logistics; and treat interconnection (carrier hotels/IXPs) as first-class resilience design objects.

Governments and regulators

In the next 30-90 days, these leaders can: identify public services that are single-region/single-provider in conflict-adjacent theaters; set minimum continuity expectations for critical services (tested failover, not paper plans); and establish channels for rapid coordination with providers during incidents. Within 12-24 months, they have an opportunity to: clarify policy for segmentation of defense-critical compute (dispersion, hardening, sovereign controls); invest in strategic enablers like cable repair capacity, protected corridors, and grid resilience; and update resilience standards to include multi-layer conflict scenarios (power, connectivity, and compute).

Bottom line

The March 2026 events challenge the assumption that cloud infrastructure is insulated from military conflict. In an AI-enabled world where commercial compute underpins both civilian and defense-relevant capabilities, leaders should act as if conflict could render regional cloud unavailable, and design for it.

Hyperscalers remain the best default platform for many mission-critical workloads because they offer depth of services, operational maturity, and economies of scale. The shift is away from the assumption that regional availability is guaranteed. Leaders should ensure the highest-consequence services have a tested survivability path across independent regions and dependencies. This path should include identity and management, so recovery does not depend on the very systems that are degraded during a crisis.

A useful lens is to consider how long critical functions can operate without reliance on the primary region's power, connectivity, or control plane. Where continuity cannot be sustained over an extended disruption window, from several days to weeks, resilience remains incomplete.

Ultimately, digital infrastructure resilience will depend on the ability of leaders to decouple critical services from single points of failure across the relevant physical stacks.

Authors

Vladimir Lukic

Managing Director & Senior Partner; Global Leader, Tech and Digital Advantage
Boston

Shoaib Yousuf

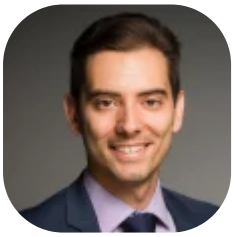
Managing Director & Partner
Dubai

Filippo Scognamiglio

Managing Director & Partner; BCG Henderson Institute Functional Leader; Global Cloud Advisory Business Leader
New York

Amir Alsbih

Partner and Associate Director, Data & Digital Platform
Munich



Radu Balanescu

Principal
Dubai

Imane Rai

Project Leader
Dubai

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group partners with leaders in business and society to tackle their most important challenges and capture their greatest opportunities. BCG was the pioneer in business strategy when it was founded in 1963. Today, we work closely with clients to embrace a transformational approach aimed at benefiting all stakeholders—empowering organizations to grow, build sustainable competitive advantage, and drive positive societal impact.

Our diverse, global teams bring deep industry and functional expertise and a range of perspectives that question the status quo and spark change. BCG delivers solutions through leading-edge management consulting, technology and design, and corporate and digital ventures. We work in a uniquely collaborative model across the firm and throughout all levels of the client organization, fueled by the goal of helping our clients thrive and enabling them to make the world a better place.