AI's hidden resilience crisis, and the colocation opportunity it creates

As inferencing shifts from PoCs and trials to scaled live deployments, what are the impacts on enterprise resilience planning, and how does this impact their colocation and cloud procurement strategy?

Joe Hurman, Principal Consultant

April 2026

Pressure on enterprise AI investment is accelerating

The pressure on enterprises to adopt AI quickly is creating a resilience problem that most organisations haven't fully recognised. Spending on generative AI grew from virtually nothing in 2022 to $37 billion in 2025, and 68% of technology executives say they are investing heavily in AI projects. A majority (61%) of the same global sample also report feeling more pressure than ever before to demonstrate a return on those investments. Caught between surging AI investment and rising ROI pressure, resilience is being overlooked.

This has consequences across the data centre ecosystem. It affects everything from sales and channel strategy in retail colocation, through to infrastructure requirements from both enterprises and the hyperscalers and other players who serve them. As enterprise attack surfaces broaden and cyber threats become more complex, uptime risks linked to poor AI resilience cannot be ignored. Mature colocation operators must take an advisory role to support their customers in navigating this challenge.

A dependency problem hiding in plain sight

One of the less visible consequences of rapid AI adoption is an expanding 'dependency footprint'. Three forces are driving this simultaneously:

1. Enterprises are deploying a growing number of AI-native applications, the vast majority of which are delivered through vendor-managed infrastructure. Everything from code generation tools (e.g. Claude Code) to vertical-specific AI agents (e.g. Harvey) must be supported by underlying infrastructure, the deployment of which introduces new infrastructure dependencies, to which the enterprise customer has limited or no visibility.
2. The democratisation of AI, coupled with the rapid pace of innovation, means that software developers, data scientists, and business teams are all deploying their own AI environments, often outside enterprise IT governance procedures, adding another level of opacity to the corporate attack surface.
3. AI agents are becoming more autonomous, creating resilience risks and infrastructure dependencies that sit beyond any one person's view, including the person tasked with resilience planning.

The result is that IT is relying on an increasingly tangled web of applications, systems, and vendors, while visibility into how they operate declines.

The average large enterprise with annual revenue over $10 billion manages over 2,400 discrete applications and APIs. Mapping the resilience of that estate end-to-end, from data centre facility infrastructure through hardware, platforms, and up to the applications themselves, is near-impossible, and one that many organisations are not approaching systematically.

The inherent opacity in procuring XaaS solutions without an accompanying IT resilience strategy at the infrastructure layer is coupled with a consolidated market for the provision of the very infrastructure on which these increasingly dependent AI solutions run. Taking Europe as an example, outside of the four main US hyperscalers (AWS, GCP, Azure and OCI), options for enterprises are the neoclouds (cheap but risky as long term partners due to limited history and widespread discourse about their financial sustainability), or just 15 colocation operators with Nvidia DGX-Ready certification. With a significant portion of these present in the Nordics where power pricing and availability can compete with the US, the options in FLAP-D markets are scarce, with only 2 or 3 providers present in each market except London. The options for potential customers of HPC colocation in metro markets are scarce.

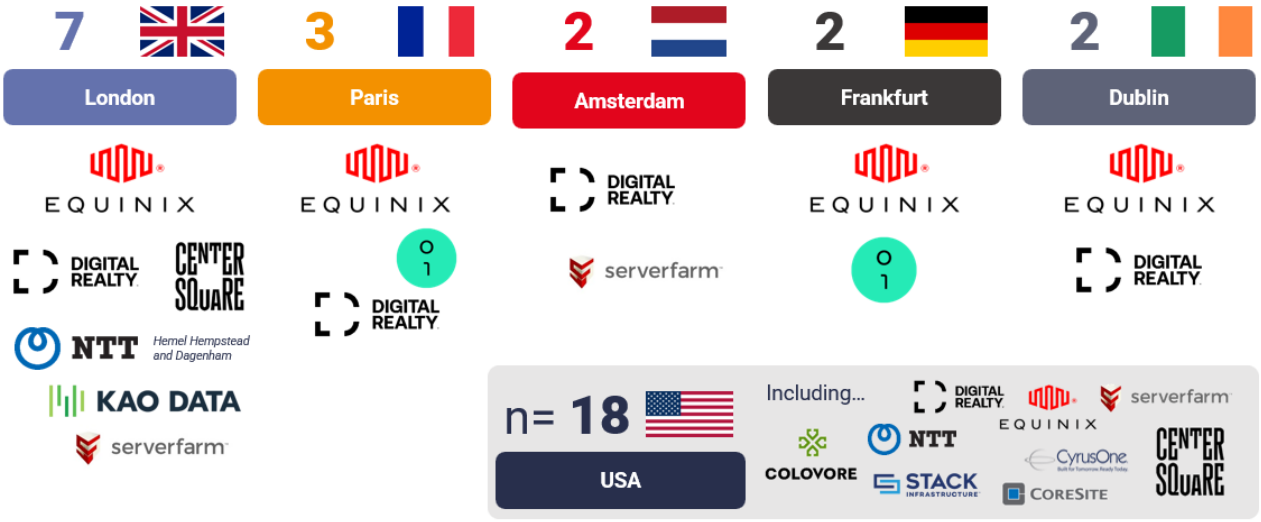

Figure 1: Nvidia DGX certified colocation providers across FLAP-D

Data sovereignty and latency requirements could compress IT estates spanning hundreds or thousands of applications into as few as two or three facilities. For now, hyperscalers are capturing much of that demand, but as AI deployments scale and become more critical, enterprises and service providers are seeking more cost-effective alternatives over the medium to long term. Even for hyperscalers, built-in backup systems do not eliminate disruption. Switching to a backup site can still take time, and even a short delay can cause serious financial and reputational damage.

The best evidence for this can be found in the very public fall-out from hyperscaler downtime. While rare, they illustrate the very real risk of running an IT estate across multiple service providers which consolidate and use the same infrastructure provider.

The AWS US-EAST-1 outage of October 2025 is a prime example of this, impacting a long list of enterprise applications including (but not limited to):

- Zoom
- Slack
- Perplexity
- Canva
- Asana
- Atlassian
- Box
- ADP
- Xero
- Smartsheet

Beyond the number of applications affected by the outage, it is also important to consider the workflows and systems they support. AI investments are increasingly tied to semi-autonomous agents, with the security risks of shadow IT agentic integrations well documented. Take Perplexity from the above list — let's assume a

financial services business has integrated it into both automated and manual workflows to increase productivity. If Perplexity goes down, so do all of those workflows across front, middle and back office functions. Widespread disruption across a business caused by improper resilience planning and an overreliance on a single vendor which is opaque to the impacted enterprise.

Every risk is an opportunity

The risk posed by rapid innovation across a small number of AI infrastructure operators is, conversely, a significant commercial and strategic opportunity for those data centre operators positioned to respond. The scarcity of Nvidia DGX-certified colocation capacity in key European metros, combined with the operational complexity of running high-density AI workloads, creates a defensible moat for operators that can demonstrate mature liquid cooling capabilities and genuine expertise in GPU fault management. Even if this moat proves short-lived as competitors build their own HPC colocation capabilities, early movers are still likely to benefit: by the time the second wave of demand arrives, they will have more operational experience and a sticky base of accounts to grow.

Enterprises actively seeking alternatives to hyperscaler lock-in need partners who can not only provide power and space, but who can elevate themselves from commodity infrastructure vendor to strategic partner, and seen by their enterprise customers as an advisor with valuable inputs to their digital infrastructure strategy.

Three investment pillars to position for colo growth

For operators seeking to grow their HPC colocation business, there are three key investment pillars:

1. **Infrastructure**: Upgrade power and cooling for HPC workloads, either to customer specifications or pre-built to Nvidia pod standards for faster deployment.
2. **Technical skills**: Develop expertise in operating HPC colocation environments, from specialised monitoring to advanced remote hands. For less experienced customers, support with managing GPU and ASIC infrastructure can be both valuable and profitable. Integrating provider and customer data, such as using thermal variance to flag GPU throttling risk, can further strengthen the offer.
3. **Sales**: Take a consultative approach that positions the colocation provider as a resilience partner, not just a capacity vendor. That requires account teams able to engage enterprise architects and CIOs on resilience and broader architecture, and to translate HPC technical capabilities into enterprise outcomes.

In short, early HPC deployments help colocation providers build the skills and reference customers needed to win as enterprise AI demand shifts from early adoption to the mainstream and HPC colocation demand accelerates.

Joe Hurman is a Principal Consultant and leads the Data Centre practice at STL Partners.

Get in touch with the author to learn more
joe.hurman@stlpartners.com

Or visit STL Partners' Data centre hub page:
https://stlpartners.com/data-centres/