INTELLIGENCE UPDATE

AI uncertainty: more adoption, more caution


Andy Lawrence

28 Oct 2025

OPINION: AI uncertainty (part 1)

Uptime Intelligence published several reports in 2024 describing how enterprises and some colocation companies were planning to support AI workloads in their data centers. These reports disclosed that only a small number of colocation companies and large enterprises were investing heavily in high-density (liquid cooled) data centers. Most were plotting a course that would involve minimal disruption and expenditure now — with plans to support high density AI workloads in the future.

Now, in mid to late 2025, and drawing on our research from recent quarters — based on discussions with Uptime Network members, global clients and Uptime Institute's engineering consultants — we can identify several key trends shaping the deployment and evolution of AI infrastructure.

This report captures the main insights from these discussions. A separate report will explore some of the risks associated with a bubble in AI infrastructure, and how different organizations or data center operator may be responding or affected.

Key trends

Some of the key trends Uptime Intelligence has identified in 2025 are:

- **Financial caution.** CEOs and CIOs expect the financial returns and business benefits from AI to be strong or even transformative — eventually. At present, however, they are uncertain about how, to what extent and when to invest and build. In addition, there is widespread expectation of financial upheaval in the AI sector, with some early adopters, suppliers (and even some operators) considered to be at risk in a downward market correction. Company executives in all classes of company are taking measures to reduce their own risks.
- **The majority of operators are reluctant to commit to high density.** Operators with huge training models, large investments, or those that are planning super-dense clusters with high voltages and liquid cooling tend to stand out and draw publicity — but they are likely to be the minority. Most organizations (including both enterprises and colocation providers) are early or late adopters, and only commit to major

investments once the results are proven. In the case of AI infrastructure, much remains unproven.

Comparing late 2024 to late 2025, Uptime has found a growing hesitancy to commit at scale to high density (liquid cooled) infrastructure. The prevailing strategy is to build limited capability and then deploy flexibly, allowing for expansion and upgrades later — if and when returns are proven.

- **Experience is teaching caution.** Designer and operator understanding of AI infrastructure has moved from theoretical to early-stage practical. As a result, technical staff are becoming more aware of complications around resiliency, compliance, complexity, equipment performance, costs, commissioning and equipment supply. These issues require close attention and are slowing deployment.
- **IT is cautious too**. Just as the facilities teams are learning more about the practical problems of high-density infrastructure, IT is having to apply more governance and resilience to its AI projects as they proceed from proof-of-concept to production. This means they need to address issues around data integrity, privacy, access rights, liability, recoverability, transparency, scalability, performance, as well as compliance and reporting.

  This is neither stopping innovation, nor preventing deployment of AI — in fact, it is a prerequisite of large-scale deployment. But it is slowing innovation and deployment down – allowing infrastructure managers and colocation partners more time to assess what infrastructure to implement, and when.
- **Cloudy now, and cloudy later?** Many IT developers have found that the use of public cloud (including neoclouds) is a less risky and capital-intensive option for their initial AI experiments (than on-site deployment). Although cloud GPU services are expensive, and sometimes not available, there is a growing range of tools to adopt so that large infrastructure investments can be avoided. For this reason, a significant portion of enterprise AI will be hosted in the cloud.

  These venue decisions are not final, though. Enterprises recognize the value of building on-premises infrastructure and investment in dedicated infrastructure can be deferred until demand justifies the business case, and when the nature of that demand is better understood (see *Most AI workloads will be trained on the cloud*). At scale deployments, with inference and training, is likely to have very client-specific requirements.

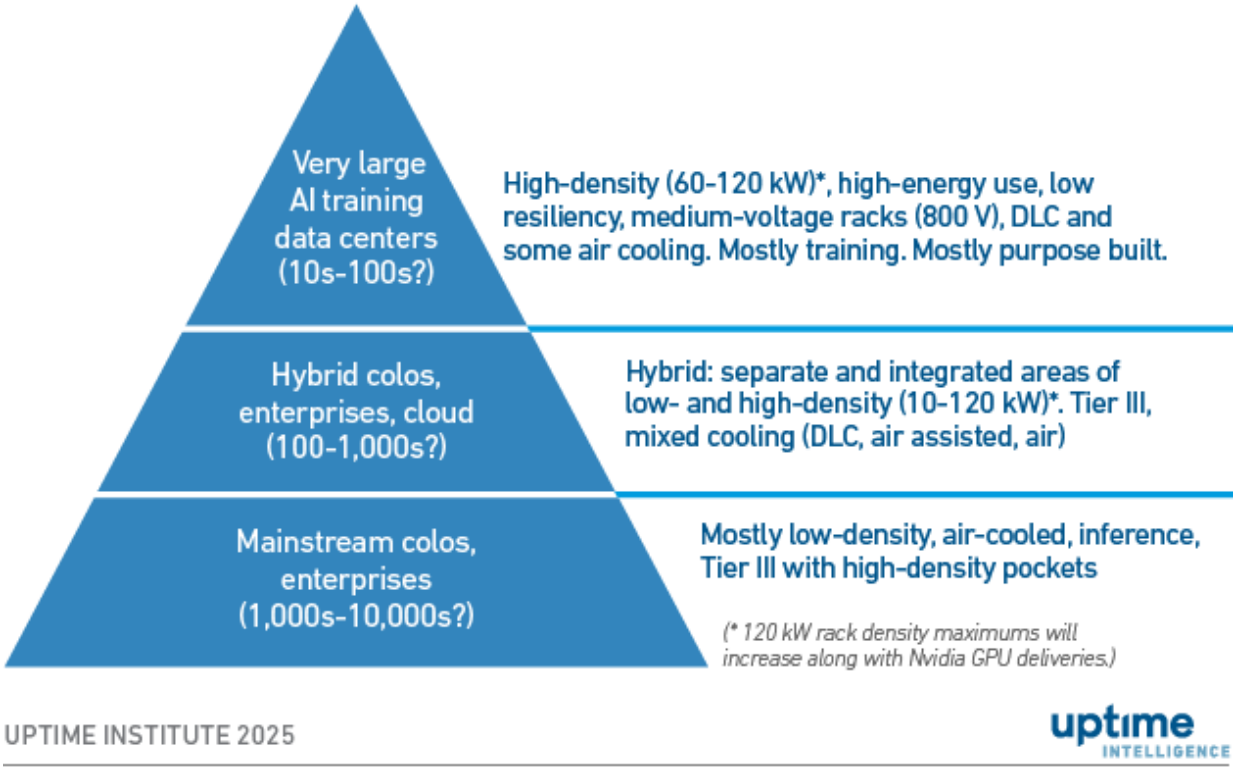
  Overall, findings from Uptime Intelligence's survey and member research suggests that AI deployment is likely to remain hybrid (much like it is today) with overall workloads continuing to grow.
- **Future proof, if possible**. Many managers say they expect significant innovations in software and hardware, which could render investments in large models, processors, and cooling and power uneconomic or obsolete. Examples of this are the impact of the DeepSeek AI model and inference processors.

  For this reason, many operators are trying to design adaptable power and cooling infrastructure, which will help them avoid a commitment to any one technology — or leave them with stranded capacity. This concern applies to the exclusive use of Nvidia GPUS and certain types of liquid cooling technologies. When making decisions, designers aim to minimize vendor or technology lock-in and allow for flexibility for future changes, with minimal cost or disruption.
- **Most GPUs sites today are air cooled.** Outside a small number of enterprises and those colocation companies building for or hosting hyperscalers or neoclouds, most data centers today are not hosting any high-density (liquid cooled) infrastructure, or have deployed it only in limited capacity. While most colocation providers do host some GPUs, it is in limited quantities. These GPUs are mostly air-cooled Nvidia

systems and are often used for pilots; enterprises are similar.
The majority of operators consulted by Uptime Intelligence, whether colocation providers or enterprises, are focusing on extending the life of air-cooled systems until demand patterns are clearer. This means using technologies, such as air-assisted liquid cooling, rear door heat exchangers — or else they are selecting Nvidia air-cooled systems, adapting temperatures and airflow as necessary. (See *Self-contained liquid cooling: the low-friction option*.)

- **Direct liquid cooling can be deployed over time.** Few enterprises and colocation providers are convinced that liquid cooled infrastructure needs to be deployed as rapidly or speculatively as some AI leaders have been advocating. Many of Nvidia's current — and future — products can be deployed using air. Even hyperscalers are, where they can, deploying cabinets with liquid-cooled side cars, cooled by air at a facility level.
  Although only a small number of operators are building, or have the capacity to deploy liquid cooling to support up to 130 kW per rack, almost all are actively exploring the technology. Most colocation providers, in particular, say they are prepared to offer this capability if requested by clients (see below).
- **Inference is still a wild card.** Enterprises (and their colocation partners) are uncertain about the compute and electrical power (and hence, density and cooling) that will be needed for inference — and have consistently queried this over the past two years. ~
  Companies such as Nvidia have argued that GPUs are required — if not essential — for inferencing, and that GPU clusters can do both training and inferencing work, helping to drive up their utilization. But others disagree: they say that even low-end CPUs can do most of this work, and that inferencing is a wasteful and expensive use of power-hungry GPUs. To complicate the decision, a new class of inferencing application specific integrated circuit (ASIC) can outperform both in processing tokens and power consumption.
  However, there is some agreement: inferencing needs to be resilient, and is best positioned in most (but not all) cases near the edge where a wide range of hardware options will be available, from small, low-cost systems to larger, high-density and expensive systems. Operators have disclosed to Uptime Intelligence that the uncertainty around inferencing — which is predicted to eventually use more compute and power than training — is making it difficult for them to know how much data center capacity and what density they will need, and where it should be sited.
- **Varied workloads, varied designs.** Although current thinking suggests foundational AI models will become larger and denser — with potentially hundreds in development — there will also be tens of thousands of smaller domain-specific models. Many companies are already deploying these.
  These smaller models can be deployed without using the highest performing GPUs, with smaller clusters and at lower densities. This will also enable greater re-use of the same hardware for inferencing.
  Given the wide variation in technical and business requirements for AI, it is likely that operators and colocation providers will build or retrofit pragmatically in response to immediate opportunities, and according to siting and power opportunities. Uptime Intelligence expects to see a much more varied infrastructure landscape than in recent years, dominated by commercial and standardized cloud (see **Figure 1**).

Figure 1 Rack density in 2025. Racks will become yet more dense in the years ahead.



- **Contracts are being rewritten.** The need to support higher density infrastructure, and accommodate higher power demands, has changed the business risk profile for many operators, especially colocation providers, that have significant capital investments and commitments to power companies. This is prompting a review of contracts with both suppliers and tenants.
  Colocation providers, in particular, are seeking to put constraints and conditions on tenants, to ensure they are able to meet their financial and usage obligations. They are also often being forced to negotiate new contracts with power suppliers, who themselves are exposed to new risks caused by large power demands.

Summary

Across the IT and infrastructure sector, executives and managers are trying to balance two conflicting priorities. The first of these is that they need to understand, invest in and exploit AI, which most believe will be transformational for their business, the business environment and the world. But AI infrastructure is also expensive, complex and technically immature — and full of deployment traps which open up new business and legal risks.

This uncertainty may not be of much concern to the leading evangelistic pioneers, but most AI will be deployed by mature businesses who are more risk and cost aware. It is not inertia or AI scepticism that is slowing adoption, but the need to do it well — after which, AI in all its forms is likely be deployed on an even larger scale.

(*The Uptime Intelligence 2024 report, Six AI infrastructure conundrums, described the uncertainty over demand, power availability and instability, resiliency for training and inferencing, rack density and cooling. Also in 2024, AI: enterprises are active – but cautious reported the gradual step-by-step approach being taken by many operators.*)

ABOUT THE AUTHOR

Andy Lawrence

Andy is a founding member and the Executive Director of Research for Uptime Institute Intelligence, which analyzes and explains trends shaping the critical infrastructure industry. He has extensive experience analyzing developments in IT, emerging technologies, data centers and infrastructure, and advising companies on technical and business strategies.

alawrence@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.