Building AI data centres at half the cost: how neoclouds are fuelling the rise of low-cost AI infrastructure

July 2026

Sylvain Loizeau

Neoclouds, the latest entrants to the AI infrastructure market, are providing graphics processing units as a service (GPUaaS) and developing purpose-built AI data centres for roughly half the capital cost per megawatt of conventional cloud facilities. This is no longer a niche experiment: CoreWeave has reported over 1GW of active power capacity and Crusoe has announced 4.9GW under contract. Their cost base shapes the price of compute across the AI ecosystem, setting a benchmark against which data centre developers, operators, investors and customers are now measured.

Denser and hotter: shrinking the building itself

The most obvious saving comes from the buildings themselves: neoclouds simply require less physical infrastructure. A conventional enterprise or co-location hall is usually designed around racks drawing 5–20kW of power. A state-of-the-art Nvidia GB300-based rack draws up to 142kW and the next-generation VR200 racks are expected to draw 220kW (with Nvidia's roadmap mentioning 1MW racks in the future). At those densities, a given IT load fits into a tenth of the whitespace or less: a 100MW deployment that would once have required more than 10 000 racks (considering 10kW racks) and several football pitches of technical space can now be housed in fewer than 1000 racks (considering 100kW racks). Land, steel, concrete and construction labour all scale with floor area. By reducing the required footprint, neoclouds can dramatically lower the cost of the building itself, which typically accounts for ~15% of total capital cost.

Direct liquid cooling (DLC), which makes such densities possible, delivers a second, less obvious source of savings: the facility cooling loop can run at much higher temperatures. Depending on the platform, secondary coolant supply temperatures range from roughly 30°C to 45°C, compared with the chilled water systems used in air-cooled facilities, which generally operate at ~20°C. At temperatures of ~40°C, depending on the facility's design and local climate, mechanical chillers can become mostly unnecessary, because dry coolers can reject heat all year round. Removing the chillers cuts costs and reduces the size of the supporting electrical infrastructure, helping to improve power usage effectiveness (PUE) from the 1.3–1.5 typically achieved by efficient air-cooled facilities to ~1.1–1.2.

Right-sized resilience: paying only for the required redundancy

The industry's redundancy conventions were designed for workloads where even a brief outage could breach a service-level agreement (SLA). Conventional cloud facilities therefore rely on extensive resilience measures, including concurrent maintainability, 2N power trains, uninterruptible power supply (UPS) coverage of the full IT load and generator fleets able to run the site indefinitely (typically Tier III- or Tier IV-equivalent designs), but this level of redundancy is very costly.

AI-training workloads, and non-critical inference jobs, are typically much less time sensitive than cloud workloads, which allows for a rethinking of the design. When a power failure interrupts an AI training job, the workload can typically resume from its most recent checkpoint rather than starting again from scratch. As a result, the cost of an outage may be limited to a few seconds or minutes of lost GPU compute time, depending on factors such as checkpoint frequency, recovery time and cluster scale.

Data centres optimised for neoclouds have therefore moved away from the traditional practice of duplicating every element of the power infrastructure. Instead of deploying fully redundant 2N power chains, they favour shared-reserve architectures, like 4N/3 configurations or 'catcher' blocks that back up several halls at once. UPS and generators are reserved for equipment whose failure could result in data loss or disrupt site operations, including storage, networking and orchestration systems. In countries with reliable grids, GPU clusters often run without a back-up diesel generator fleet sized to support the full IT load. The result is a significantly lower cost per megawatt, achieved by trading a degree of resilience that many AI workloads simply do not require.

Freedom in siting: software resilience (scale-across) and latency-insensitive use cases

What makes right-sized redundancy even more relevant is the fact that resilience has in part moved from the building to the software stack. Modern training frameworks checkpoint continuously and reschedule work around failed nodes; increasingly, they also spread a single training run across several data centres. This 'scale-across' approach turns a handful of individually less resilient sites into one large virtual cluster: if one site experiences an outage, performance may degrade but the training run can continue. Beyond improving resilience, mutualising infrastructure across locations allows operators to combine capacity across sites, supporting training runs larger than any single campus could host.

The same relaxation of constraints transforms siting economics. AI training workloads tolerate latency, so an AI data centre no longer needs to be close to a metro fibre hub where land and power are often most expensive. Neoclouds build where energy is inexpensive and grid access is available, including industrial brownfield sites, locations close to underutilised power generation assets, or repurposed cryptocurrency mining facilities with existing grid connections.

From construction project to product

Traditional data centres are bespoke construction projects: each site is engineered individually, and much of the costs tied to site-specific design, integration and commissioning. AI specialists are changing that by treating the data centre as a product that can be replicated at scale. Chip vendors and large integrators now publish reference designs that specify the rack, the power train and the cooling loop as a single coherent system, and equipment suppliers respond with standardised, prefabricated modules (e.g. skid-mounted electrical rooms, packaged cooling plant, pre-populated racks) that are assembled and pre-commissioned in a factory before they reach the site.

The savings compound. Engineering effort is amortised across many identical builds instead of being repeated for each one. Factory testing replaces months of on-site commissioning. Construction schedules compress from 24 months or more to around 12 months or less. A shorter build time lowers financing costs and accelerates time to revenue, ensuring that GPUs – assets that depreciate rapidly regardless of whether they are in use – start generating revenue earlier.

A short-term bet on training with some longer-term uncertainties

All of these design optimisations are based on the same premise: the facilities are intended to support large-scale AI training or low-criticality inference workloads.

As AI use cases diversify and scale, an increasing number of time-critical inference workloads are likely to emerge, with increased latency sensitivity and stricter service-level commitments. For these applications, some of the trade-offs described above will have to be revisited, requiring parts of the installed base to be retrofitted. In the longer run, it will be critical to assess how the shorter lifecycles of GPUs and the increasing resilience requirements of some AI workloads will influence overall facility economics. This will determine how the total cost of ownership of AI-optimised facilities compares with that of traditional cloud data centres, whose lifetime can extend well beyond 25 years.

If you would like to discuss these issues from a technical or commercial/financial perspective, please contact Sylvain Loizeau, expert in transaction support, data centres and GPUaaS.