Adapting Hyperscale Data Centers for AI Workloads: Challenges for Critical Power and Cooling Infrastructure

AUTHOR
Geoff Halliday
Business Consultant, Rehlko

INTRODUCTION

As the global appetite for Artificial Intelligence (AI) applications continues to surge, hyperscale Data Centers (DC) find themselves under increasing pressure to adapt their existing infrastructure.

While traditional computing environments were largely optimized for general purpose Central Processing Units (CPUs) and variable workloads, AI-driven systems, especially those involving deep learning, neural networks and large language models, place unprecedented demands on power density, thermal management and fault tolerance.

This paper explores the technical challenges that hyperscale operators encounter when upgrading from conventional computing to AI intensive workloads, focusing primarily on four critical areas. Firstly, we consider the critical power train which encompasses standby generators, uninterruptible power supply (UPS) systems and electrical distribution required to maintain optimal service performance.

Secondly, we look at the implications of AI workloads on transient loading conditions and how they could impact standby generators and UPS performance, finally considering evolving strategies and how to maintain N+1 as well as overall site resilience under these new operational regimes.

rehlko
AI Ready

THE RISE OF AI IN HYPERSCALE ENVIRONMENTS

AI applications are fundamentally different from traditional IT workloads. Instead of processing thousands of lightweight transactions or web queries per second, AI workloads involve training and inference operations that require intensive, sustained computing power, often through high-density graphical processing units (GPUs), tensor processing units (TPUs), or other accelerator architectures. These components typically consume significantly more energy than conventional CPUs.

Traditional server racks might draw 5–10 kW, AI-focused racks now routinely exceed 30–40 kW, with some configurations now regularly pushing beyond 80 kW into 100kW+ territory. Recent data center developments are incorporating rack installations exceeding 500 kW, with the potential to reach up to 1 MW within the next few years. This shift places new and significant constraints on infrastructure. Power demand is no longer sporadic or in bursts but instead exhibits prolonged high levels of utilization, increasing the baseline energy requirements per rack and hence across the entire data hall or facility.

AI workloads also alter the thermal profile of data centers. The consistent and intense processing levels generate sustained heat loads that conventional air-based cooling solutions are often ill-equipped to manage. The result is a marked need for new power and cooling strategies to meet both immediate and future demands.

Rack Type	Typical Power Density	Notes
Traditional CPU	5-10 kW	Legacy enterprise IT workloads
General Purpose	10-20 kW	Standard enterprise/mixed workloads
High-Density Enterprise	20-40 kW	Virtualization, converged infrastructure
AI/ML (early generation)	30-80+ kW	GPU clusters, moderate AI workloads
AI/ML (current generation)	80-150 kW	Modern GPU servers (A100, H100 class)
AI/LLM Training Racks	150-300+ kW	Large-scale model training, dense GPU configs
Ultra High-Density AI Racks	300-500+ kW	Advanced liquid-cooled racks, hyperscale deployments
Next Generation AI/Exascale Racks	500 - 1MW (emerging)	LLM training, generative AI, future HPC architectures

Figure 1: Power Density Comparison (kW per Rack)

IMPACT ON THE CRITICAL POWER TRAIN

UPS AND GENERATOR SYSTEMS: TRANSIENT LOADING CONSIDERATIONS

Traditional data center designs assume that power demand fluctuates within predictable ranges. UPS and generator systems are configured to handle orderly load increases, with ample time to stabilize voltage and frequency parameters. However, AI workloads introduce a new paradigm, where load ramps can occur rapidly and can without warning, particularly during training cycles or when high-density inference tasks are initiated.

Figure 2 Standby Generators at a Hyperscale Data Center

This rapid ramp-up in demand results in steep transient loading conditions, often characterized by high dP/dt (rate of change of power). For generators, this can compromise the ability to maintain voltage and frequency stability as defined within EN ISO8528 requirements. In extreme cases, it may lead to under-frequency or over-voltage events, tripping sensitive systems or triggering safety shutdowns.

Similarly, UPS systems can struggle to absorb sudden load spikes. Inverter overloads or latency in battery response can reduce the effectiveness of the UPS during these transients. Conventional valve-regulated lead-acid (VRLA) battery systems, while robust, may not discharge quickly enough to stabilize the load.

N+1 RESILIENCE IN THE AI ERA

The principle of N+1 redundancy is a cornerstone of data center design, ensuring that one additional unit (whether a generator or UPS module) is available beyond the minimum needed to handle normal operations. This model, however, becomes more complex in an AI environment.

Firstly, AI workloads operate at higher baseline loads, potentially reducing the available capacity margin. An N+1 configuration may no longer provide sufficient fault tolerance if the total power draw per module increases significantly. Secondly, the risk of cascading failure is elevated: if one generator or UPS fails under peak AI load, the remaining units may not have the capacity to absorb the additional peak requirements. As a result, operators are reconsidering the adequacy of N+1 designs. In some cases, N+2 or distributed redundancy strategies are being explored. These involve either adding a second layer of redundancy or segmenting the load across smaller, modular systems to contain the impact of equipment failure.

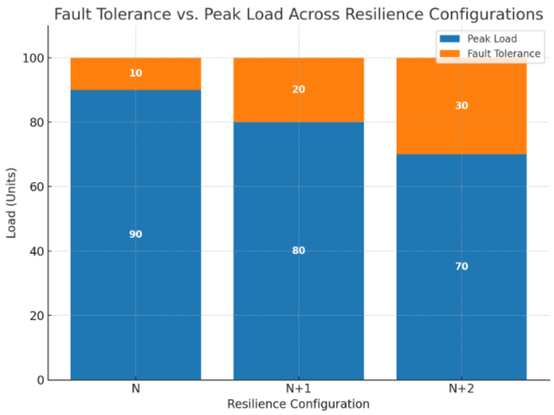

Figure 3 N+1 vs N+2 Resilience Margin

While not based on specific product data, this stacked bar chart is intended to demonstrate a general principle: as resilience increases (from N → N+1 → N+2), usable peak capacity typically decreases, whilst at the same time fault tolerance and operational headroom improve. The higher levels of redundancy result in increased effective rotating mass within the system.

For clarity and comparison, the total load is normalized to 100 units:

N: Minimal fault tolerance (10 units) with maximum peak load (90 units).

N+1: Improved fault tolerance (20 units) with a corresponding reduction in peak load (80 units).

N+2: Highest fault tolerance (30 units) with further reduced peak load (70 units).

REMEDIAL STRATEGIES AND ENHANCEMENTS

STANDBY GENERATORS

When considering generator upgrades in the context of modern load profiles, particularly those driven by highly dynamic electrical demands, it becomes clear that traditional design assumptions no longer hold the same margin of safety they once did. One of the more fundamental improvements lies in increasing the rotating mass of the generating set (over sizing the set although this is not a preferred solution).

By doing this, you are effectively adding inertia into the system, which provides a buffer against rapid load transients. In practical terms, this means the machine is better equipped to absorb short duration disturbances without excessive frequency deviation. It also enhances the generator's ability to ride through fault conditions, supporting more robust fault clearance by maintaining system stability long enough for protection systems to operate as intended. This isn't simply a theoretical benefit it is something you see immediately in reduced nuisance trips and improved overall resilience of the power system.

Alongside mechanical enhancements, the role of the generator governor becomes increasingly critical. Traditional mechanical or first-generation electronic governors often struggle to keep pace with the rate of change seen in modern electrical loads. By deploying advanced electronic governing systems, operators can achieve significantly faster and more precise responses to load fluctuations.

These systems continuously monitor operating conditions and adjust fuel input in real time, ensuring that engine speed and therefore frequency remains tightly controlled even under aggressive step loading. The result is a generator that behaves in a far more predictable and controlled manner, which is essential when supporting sensitive or mission critical infrastructure.

Equally important is the upgrade of alternators and their associated voltage regulation systems. Modern loads can introduce very steep load ramps, far beyond what older alternator designs were optimized to handle. By specifying alternators with improved electromagnetic design characteristics, coupled with high-performance automatic voltage regulators, it becomes possible to maintain voltage stability under these demanding conditions. This ensures that voltage dips are minimized, and recovery times are shortened, both of which are essential for maintaining downstream equipment operation and avoiding cascading issues across the electrical system.

UPS

Turning to UPS system enhancements, the shift towards lithium-ion battery chemistries represents a significant step change in performance. Compared to traditional valve regulated lead acid batteries; lithium-ion technologies offer markedly faster discharge characteristics and a much higher cycle life. This makes them particularly well suited to applications where rapid response and frequent cycling are the norm, rather than the exception. In practice, this translates to a UPS system that is not only more responsive but also more durable over its operational life, reducing both maintenance requirements and total cost of ownership.

However, battery performance is only part of the picture. The behavior of the inverter itself is equally critical, particularly when dealing with high rates of change of power; so called high dP/dt events. By implementing adaptive inverter control algorithms, modern UPS systems can respond dynamically to these rapid load changes, maintaining output stability even under highly volatile conditions. These control strategies continuously adjust inverter operation in real time, allowing the system to accommodate sudden spikes or drops in load without compromising power quality. It's a level of responsiveness that simply wasn't achievable with earlier generations of static control systems.

Finally, the move towards modular UPS architectures reflects a broader shift in how resilience and scalability are approached. Rather than relying on large, monolithic systems, modular designs allow capacity to be built up incrementally, with individual modules sharing load dynamically. This not only improves efficiency under partial load conditions but also enhances redundancy, as the failure of a single module has a far less significant impact on the overall system. From an operational standpoint, it also provides greater flexibility, enabling systems to be scaled in line with demand while maintaining high levels of availability. In environments where uptime is non-negotiable, that combination of scalability and resilience is increasingly becoming the standard rather than the exception.

Figure 4 Containerized N+1 UPS system

INTEGRATION WITH ENERGY STORAGE

In many cases, the introduction of a battery energy storage system provides a very practical way of bridging the gap between legacy infrastructure and the demands of modern, highly dynamic loads. By incorporating BESS as an intermediary buffer within the power train, operators can effectively decouple the most aggressive transient events from the mechanical limitations of the generator and the response characteristics of the UPS. Rather than forcing the generator to react instantaneously to sharp load steps or expecting the UPS to absorb repeated high dP/dt events; the BESS can step in and handle those short duration fluctuations. The result is a much smoother load profile being presented to both systems, reducing mechanical stress on the engine, limiting thermal and electrical strain on the alternator, and preventing unnecessary cycling or stress within the UPS. From a practical standpoint, this translates into improved reliability, fewer nuisance alarms, and a noticeable extension in the service life of key assets.

There is also a very clear advantage when it comes to generator start up sequences, particularly under grid outage conditions. Traditionally, the period between a mains failure and the generator reaching stable operating conditions has been one of the more vulnerable points in any standby power system. By deploying BESS to provide ride-through support during this window, it becomes possible to maintain a continuous and stable supply to the load without placing undue stress on either the UPS or the generator as it comes online. The BESS effectively carries the load during those critical seconds, allowing the generator to start, accelerate, and synchronize under far more controlled conditions. This not only improves overall system stability but also reduces the risk of voltage and frequency excursions that could otherwise impact sensitive equipment.

Taken together, these measures offer a very pragmatic route to modernizing existing installations. Rather than defaulting to wholesale replacement of generators and UPS systems—which is often both costly and operationally disruptive—targeted integration of technologies such as BESS allows legacy infrastructure to be adapted to meet the far more demanding performance envelope associated with AI-driven and other high-intensity computing environments. It's a measured approach that balances performance, cost, and risk, and in many cases delivers the required outcome without the need for a complete system overhaul.

MICROGRIDS

There is much talk today about the use of Microgrid solutions to reduce grid dependency whilst adding resilience into the data center power eco system. Microgrids are self-contained networks that combine generation, storage and load. The use of microgrids is gaining traction as DC's face grid connection delays and capacity limits. A microgrid can integrate solar/Photo Voltaic (PV), wind, Battery Energy Storage (BESS), CHP, or diesel generating sets, operating in island or grid/parallel mode (Fig. 4). For hyperscale and regional operators alike, this improves resilience while opening the door to carbon reduction opportunities.

MORE ON BESS

BESS is seen by many for its role in a grid reliance role. As discussed in the previous section it has a key role to play in the critical power train, but it also has a role to play in the broader DC power architecture. When combined with the critical power train and an on site Combined heat and Power system (CHP), BESS at the heart of a microgrid can reduce Scope 2 emissions, unlock participation in balancing markets, and enable renewable Power Purchase Agreements (PPAs) to align more closely with site load profiles.

Often the challenge is space. Urban campuses, such as those in the edge DC sector will struggle to host larger BESS installations.

COMBINED HEAT & POWER (CHP)

In many markets natural gas/fossil gas is a major or significant contributory source of fuel used in the production of electricity. A modern combined cycle gas power plant has an efficiency in the order of 55-65%. When a CHP system is used to its fullest extent (electricity generation plus low and high grade hot water) system efficiencies can rise above 85% (Clarke Energy), making it a "no brainer" for making a DC more energy efficient, reducing greenhouse gas emissions whilst at the same time adding resilience to an organization's operation. An on site CHP system also offers the ability to further enhance overall site efficiency by proving the capability of more closely matching on site demand with load given.

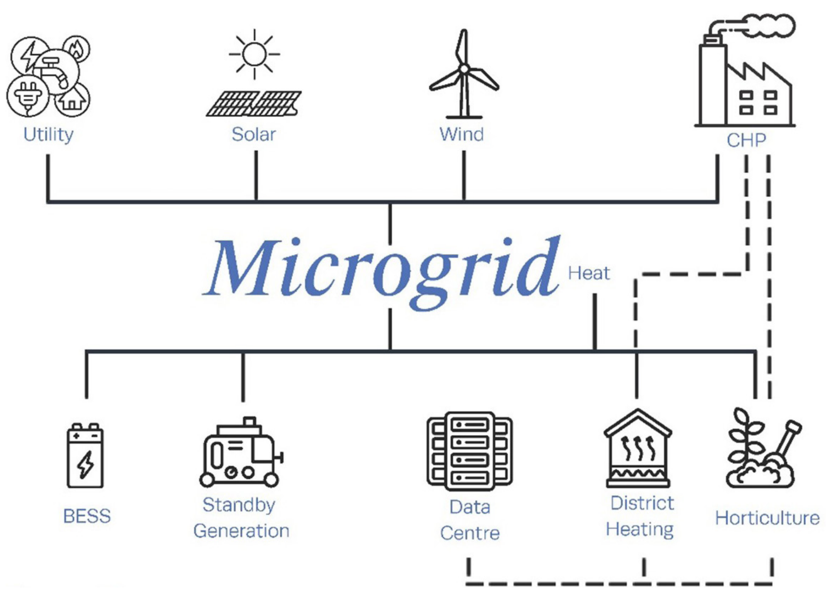

Figure 5: Typical range of Microgrid energy sources

There is also growing interest in the integration of combined heat and power (CHP) systems as part of the overall energy strategy. CHP offers the potential to enhance power resilience by providing an additional on-site generation source, while simultaneously improving overall data center efficiency through using any "surplus" capacity to recharge on site BESS or exported to the grid. When properly integrated, CHP rejected heat can be recovered and repurposed, whether for absorption cooling or for export to district heating networks, contributing to both operational efficiency and wider sustainability objectives. It's an approach that aligns well with the increasing pressure on data center operators to address both energy security and carbon performance in parallel.

CHP remains one of the most pragmatic tools for boosting power generation efficiency in a DC. It not only provides reliable onsite power but also a ready supply of heat for absorption chillers or external networks. While most CHP systems today are fueled by natural gas, engine manufacturers are designing and building for future conversion to alternative gases such as hydrogen, ammonia, or bio-methane.

FUTURE PROOFING STRATEGIES

For hyperscale operators making the transition towards AI-driven workloads, there is a growing recognition that incremental tweaks to legacy designs are no longer sufficient on their own. What's required is a deliberate shift towards a "design for AI" mindset; one that acknowledges the fundamentally different behavior of these loads and plans for it from the outset. At the heart of this approach is the concept of rack zoning. Rather than treating the data hall as a uniform environment, high-density AI racks are physically grouped together so that power delivery and cooling resources can be targeted with far greater precision. This allows operators to concentrate capacity where it is needed, avoiding the inefficiencies of over provisioning across the entire space while ensuring that the most demanding loads are properly supported. In practice, it also simplifies both thermal management and electrical distribution, as each zone can be engineered around a known and consistent load profile.

Alongside this, there is a clear move towards more modular power architectures. The traditional model of large, centralized generator plant and UPS systems is giving way to smaller, distributed units that can be deployed incrementally. This approach brings a number of advantages. It allows capacity to be scaled in line with demand, rather than committing to significant upfront investment, and it provides a more flexible route to achieving redundancy. With distributed systems, redundancy can be managed at a much more granular level, reducing single points of failure and improving overall system resilience. From an operational perspective, it also makes maintenance and upgrades far less intrusive, as individual modules can be taken offline without impacting the wider system.

The integration of battery energy storage systems also plays a key role within this design philosophy. As AI workloads introduce increasingly volatile and unpredictable power demands, BESS can be used to smooth out sudden, large-magnitude variations in real time. By acting as a buffer within the electrical system, it limits the extent to which these rapid fluctuations are seen by upstream infrastructure such as generators and grid connections. This not only stabilizes the overall system but also allows existing assets to operate within a more manageable and predictable envelope, reducing wear and improving long-term reliability.

While new build facilities have the advantage of incorporating these principles from day one, retrofit projects present a more complex challenge. Upgrading an operational hyperscale environment requires careful sequencing of works, detailed planning, and a clear understanding of system interdependencies. The priority at all times is to avoid unplanned downtime or any degradation in service, which means that changes must often be introduced in phases, with appropriate temporary provisions and risk mitigation measures in place. It's not simply a technical exercise—it's an operational one, where engineering decisions must be balanced against the realities of a live, revenue-generating environment.

CONCLUSION

The shift toward AI workloads in hyperscale data centers is not a minor evolutionary step, it is a foundational transformation that affects nearly every aspect of facility design and operation. In particular, critical power systems must be re imagined to handle the high density, high power utilization, and high transience demands of AI computing.

From enhancing UPS and generator systems to adopting advanced liquid cooling technologies, operators must embrace a proactive and technically nuanced approach. Resilience strategies such as N+1 must be revisited, and modular, scalable designs should be prioritized to future proof investments.

Ultimately, meeting the infrastructure demands of AI is not just a matter of capacity, it is about agility, efficiency, and resilience in a rapidly evolving computational landscape.

Not every solution fits every site. But the direction we need to go in is clear. The data centers we design today must be efficient, sustainable and integrated into the wider energy system.

It is good engineering practice, makes good business sense, and it is responsible citizenship. In the end it will be commercial logic and pragmatism that will likely drive the adoption of these solutions.

ABOUT THE AUTHOR

Geoff Halliday started his career as an apprentice working for Square D (later part of Schneider) before moving into the critical power sector where he has now worked for over 40 years, splitting that time equally between both the UPS and standby diesel generation sectors.

During this period Geoff has held several roles ranging from Customer Service Engineer, Project Manager, Technical Director, Sales Director through to Managing Director.

The Critical Power market exposes the individual to a wide and diverse range of market sectors ranging Health Care, Life Science, Water Treatment, Banking and Finance, Military, Manufacturing, Process Control through to Data Centers of all sizes. Drawing on his management skills, product knowledge and vast application experience amassed throughout his career Geoff now enjoys sharing his knowledge with others.

ABOUT REHLKO

A global leader in energy resilience, Rehlko delivers innovative energy solutions critical to sustain and improve life across home energy, industrial energy systems, and powertrain technologies, by delivering control, resilience and innovation. Leveraging the strength of its portfolio of businesses – Power Systems, Home Energy, Uninterruptible Power, Clarke Energy, Heila Technologies, Curtis Instruments, and Engines, and more than a century of industry leadership, Rehlko builds resilience where and when the grid cannot, and goes beyond functional, individual recovery to create better lives and communities, and a more durable and reliable energy future.

rehlko

REHLKO.com